ARCHAIC ADMIXTURE IN MODERN HUMANS

by

Ryan James Bohlender

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Anthropology

The University of Utah

December 2015

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Ryan James Bohlender**

has been approved by the following supervisory committee members:

Alan Rogers	, Chair	**8-5-2015** Date Approved
Henry Harpending	, Member	**8-5-2015** Date Approved
Chad Huff	, Member	**8-5-2015** Date Approved
Dennis O'Rourke	, Member	**8-5-2015** Date Approved
Jon Seger	, Member	**8-5-2015** Date Approved

and by **Leslie Knapp**, Chair/Dean of

the Department/College/School of **Anthropology**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

Recent advances in the study of archaic hominin DNA have resulted in the rapid development and application of new methods designed to test our relationship to our nearest relatives. These methods have been applied with archaic samples in contexts with ghost admixture, sparse sampling, ascertainment bias, and poorly understood historical events. They have also been applied to modern samples with complex relationships which will exacerbate the same problems. Here, we introduce a new method for estimating the admixture fraction, and test it and several previous methods to determine their sensitivity to the above problems. Finally, we apply these methods to Single Nucleotide Polymorphism (SNP) microarray and whole genome data, and compare our estimates to those published previously.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank Laura Bohlender, Ronald Bohlender, Lyndsey Bohlender, and Lauren Capoccitti for their tireless support through this project. Thank you to Michael Lewis, Samantha Davis, Gregory Smith, Nathan Harris, Justin Tackney, and Jesse Burns for the support and friendship I needed to succeed, both in graduate school and in life. Finally, thank you to Alan Rogers for your support, guidance, and most of all, patience. Life is hard, and your mentorship made mine just a little easier.

CHAPTER 1

$\mathcal{L}$: A COMPOSITE LIKELIHOOD ESTIMATOR OF ARCHAIC ADMIXTURE

1.1 Introduction

With the publication of several archaic hominin genomes, substantial progress has been made on the study of our relationship to our distant relatives [1]–[5]. With new data, and a new set of problems to resolve, new methods have been developed to directly address a long standing debate: whether ancestors of modern non-Africans completely replaced, or interbred with, archaic hominins as they left Africa [6]–[8]. One result of the draft Neanderthal genome paper [1] was direct evidence that non-Africans shared 1–4% of their genome with Neanderthals. Shortly afterward, a distinct archaic species, now referred to as Denisova for the cave in which the fossil was discovered, was sequenced [2] and shown to contribute an additional 3–6% archaic DNA to modern Oceanians. Further studies have since reaffirmed and refined these estimates [3]–[5].

Archaic introgression has had significant effects on modern populations. Estimates of the admixture fraction or mixture proportion have been relatively small, but we can still see signs of this introgression in the genes of modern non-Africans. There is evidence that selection has acted on these introgressed segments, conferring benefits at high-altitude [9], changing our skin and hair [10]–[12], and affecting our immune system [13]. There is also evidence that selection has removed substantial amounts of introgressed DNA. The evidence for this is two-fold. First, there are wide stretches of the modern genome that lack evidence of archaic introgression, implying that selection may have purged introgressed alleles in those regions [10], [11]. Additionally, a human specimen that was a close relative of a Neanderthal was sequenced, and contained a significantly greater admixture fraction than today's populations [14]. This also implies that selection acted against some portion of archaic introgression.

The mixture proportion—the amount of ancestry shared by a modern and an archaic

human—has been estimated directly using an approach called F_4-ratio estimation [15]. These methods assume a phylogenetic tree relating samples from different populations. In our case, those populations are Africa, Eurasia, Neanderthal, and Denisova. In the absence of what we call ghost admixture—introgression into a focal population from a secondary archaic hominin population [16], [17]—these methods generate unbiased estimates of archaic admixture in modern Eurasians. However, when there is ghost admixture in the Eurasian population, as would be the case in Melanesia and perhaps East Asia, results are biased for several methods in the F_3 and F_4-ratio statistic family, including $\hat{f}$, $R_{\text{Neandertal}}$, and Q [1], [2], [15], [18].

Once corrected for ghost admixture, the methods used to generate these results require strong assumptions about the number of archaic populations, the number of admixture events, population size, and event ordering [1, Supp. 18, 2, Supp. 8, 3, Supp. 11, 4, Supp. 14, for corrections see 18]. We have previously shown that these biases depend on time, population size, and the quantity of ghost admixture, but differ between methods, so a single correction for all F_3, or F_4-ratio statistics is not possible [18]. Given that some Melanesian populations have been shown to carry both Neanderthal and Denisovan admixture [2], [3], that mainland Asian populations may carry some Denisovan DNA [4], [19], that Yorubans have detectable levels of Neanderthal ancestry [3], and that Denisova carries some signal of non-Neanderthal archaic admixture [4], it is apparent that having a full understanding of the consequences of ghost admixture is essential.

Here, our purpose is two-fold. First, we present a composite likelihood estimator of the archaic-modern mixture proportion and the date at which the modern and archaic populations separated. This new method generates simultaneous estimates of the genetic contribution of two archaic populations, making it useful for generating direct estimates of the mixture proportion in populations where ghost admixture is suspected to be present. Second, we compare the newly proposed method against methods used previously, with and without corrections for ghost admixture, across varying sample sizes, and with uncertainty in estimates of parameters present in those corrections. In doing so, we shed light on the effect of model misspecification for several of the methods that have been used so far, and to evaluate the newly proposed method.

1.2 Methods

The estimators for Neanderthal admixture used in the original publications of the Neanderthal and Denisovan genomes, $\hat{f}$ and $R_{\text{Neandertal}}$, use the same underlying logic [15],

[20]. We define a site pattern as a subset of the populations we are studying which share the derived allele. If there has been no admixture between any of the populations of interest and the archaic, then site patterns in which one or more of the modern populations shares the derived allele with an archaic population should be the result of lineage sorting before the populations split, and therefore should be roughly equal in frequency.

A significant excess in site patterns shared between a modern sample and a Neanderthal is indicative of admixture between Neanderthals and that population. Alternatively, that excess may be the result of ancient population structure [21], [22]. Analysis of linkage disequilibrium surrounding putatively admixed alleles indicates a time of last gene exchange roughly 56kya [23], which does not fit with the ancient population structure hypothesis. Additionally, an ancient modern genome also appears to carry a significantly larger quantity of archaic admixture, implying that he was recently descended from a Neanderthal [14].

The amount of excess due to admixture is a function of the mixture proportion, which we call m. If we include the bias corrections presented by Rogers and Bohlender [18], then there are several parameters for which we need estimates. These parameters are listed in Table 1.1. So long as we have estimates for these other parameters, this fact allows us to generate method-of-moments estimates of m, as is done in $\hat{f}$, $R_{\text{Neandertal}}$, and Q.

Alternatively, we can generate estimates in a likelihood framework, as is shown below.

1.2.1 Composite Likelihood Estimator $\mathcal{L}$

The method derived here is a composite likelihood estimator of the following: m_N, the mixture proportion or quantity of shared allelic states between an archaic and modern sample; m_D, the mixture proportion between a second archaic and that modern; λ, the date for the separation of the modern-archaic ancestral population into the modern and archaic lines. The composite likelihood, hereafter referred to as $\mathcal{L}$, takes the form shown in Section 1.2.1. Following Rogers and Bohlender [18], we will use uppercase X, Y, N, D, and O to refer to populations, and lowercase x, y, n, d, and o to refer to haploid

Table 1.1. Labels for each of the population level events we simulate and model.

α	date of Denisovan admixture
γ	age of older Neanderthal fossil used in $\hat{f}$
δ	date of Neanderthal admixture
ζ	separation date for Africans and Eurasians
κ	separation date for Neanderthal and Denisova populations
λ	separation date for modern and archaic populations

genomes sampled from those populations. We will indicate site patterns using lowercase characters for the samples containing the derived allele. Below, we combine site patterns xy and xyd because prior methods did not distinguish when sites shared the derived state with Denisova. As a result, all sites fitting the xyd pattern were counted as part of the xy pattern. We do not derive the probability of the xyd pattern separately here, but we do make the distinction between xy and xyd clear, so we sum xy and xyd for clarity. The same is true for patterns xn and xnd.

$$\mathcal{L}(m_N, m_D, \lambda) = \left(\frac{P_{xy}+P_{xyd}}{S}\right)^{I_{xy}+I_{xyd}} \left(\frac{P_{xn}+P_{xnd}}{S}\right)^{I_{xn}+I_{xnd}} \left(\frac{P_{yn}}{S}\right)^{I_{yn}} \left(\frac{P_{yd}}{S}\right)^{I_{yd}} \left(\frac{P_{ynd}}{S}\right)^{I_{ynd}} \quad (1.1)$$

Here, P_{ij} is the probability of site pattern ij; S is the sum of the probabilities of the site patterns of interest; I_{ij} is the observed count of site pattern ij in a dataset. Consider a set of populations related as $(((X, Y), (N, D)), O)$ in Newick notation, where populations grouped by parentheses are more closely related to one another than to populations outside those parentheses. A diagram of the population tree we envision is presented in Fig. 1.1. X and Y represent modern populations, N and D represent archaic populations, and O the outgroup used to polarize the data. Because O is only used to polarize the data and does not appear in any of the following calculations, it will be omitted from here on. We distinguish between P_{ij} and P_{ijk}, where the former indicates that sample i and sample j share the derived allele, while k carries the ancestral, and the latter indicates that all three samples are carrying the derived allele.

We consider sites at which the derived allele is shared by one or both of the archaic samples and the focal modern lineage y, or just the modern lineages. In a prior publication, Rogers and Bohlender [18] derived formulas for the site patterns: P_{xy}, $P_{yn} + P_{ynd}$, and $P_{xn} + P_{xnd}$. In Appendix A, we derive the probabilities for P_{yn}, P_{yd}, and P_{ynd} separately.

The above site patterns do not constitute all of the possible site patterns in our sample space. For example, we do not include the site pattern x. Because we are using only a subset of the possible site patterns, we need the conditional probability of each pattern, given that the pattern is one of our patterns of interest. In all cases, the conditional probability of the site pattern is its unconditional probability divided by the sum of the probabilities of the site patterns which we define as:

$$S = P_{xy} + P_{xyd} + P_{xn} + P_{xnd} + P_{yn} + P_{yd} + P_{ynd}$$

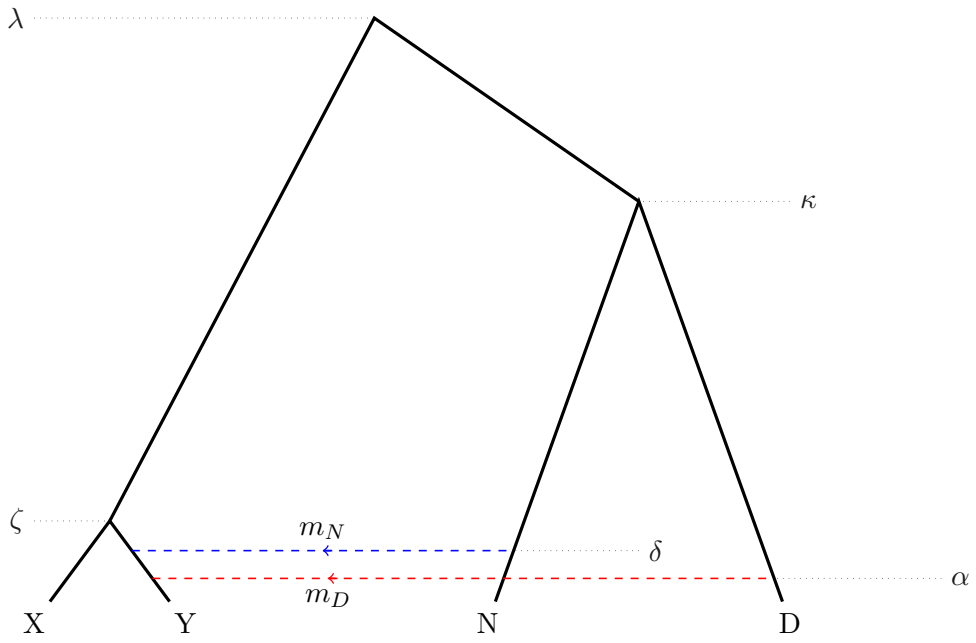

Figure 1.1. The assumed relationships between populations in the corrected versions of $\hat{f}$, $R_{\text{Neandertal}}$, and Q, as well as the relationships for $\mathcal{L}$. Within each branch, populations are assumed to be randomly mating. The red and blue lines indicate alternate paths that the y sample may follow when derived from an archaic population. m_N and m_D represent the fraction of all y lineages that derive from Neanderthals and Denisovans, respectively.

The observed count, or frequency, of a site pattern ij in a data set is denoted I_{ij}. Rogers and Bohlender show that the unconditional probabilities of the site patterns xn, yn, and xy may be written as shown below in Eqs. (1.2) and (1.4) [18]. Note that Rogers and Bohlender did not distinguish between xy and xyd etc. and we make this distinction explicit here.

$$P_{xn} + P_{xnd} = m_N(1-m_D)S_N^{(\delta,\kappa)}S_{ND}^{(\kappa,\lambda)} + m_D S_{ND}^{(\kappa,\lambda)} + (1-m_N)(1-m_D)S_{XY}^{(\zeta,\lambda)} \tag{1.2}$$

$$\begin{aligned} P_{yn} + P_{ynd} = {} & m_N(1-m_D)\big\{\lambda - \delta + (1-K_N)F_N^{(\delta,\kappa)} + (1-K_{ND})S_N^{(\delta,\kappa)}F_{ND}^{(\kappa,\lambda)}\big\} + \\ & m_D\big\{\lambda - \kappa + (1-K_{ND})F_{ND}^{(\kappa,\lambda)}\big\} + P_{xn} + P_{xnd} \end{aligned} \tag{1.3}$$

$$P_{xy} + P_{xyd} = (1-m_N)(1-m_D)\Big[\lambda - \zeta + (1-K_{XY})F_{XY}^{(\zeta,\lambda)}\Big] + P_{xn} + P_{xnd} \tag{1.4}$$

$S_N^{(\delta,\kappa)} = e^{-(\kappa-\delta)/K_N}$ is the survival function in population N for the interval (δ,κ); $F_N^{(\delta,\kappa)} = 1 - S_N^{(\delta,\kappa)}$ is the CDF for that interval; K_N is N_N/N_0, or the ratio of the population size in the current population to the size of the ancestral population. Other combinations of subscripts and superscripts are likewise defined.

Rogers and Bohlender's formula for the yn site pattern can be partitioned into two mutually exclusive events, P_{yn} and P_{ynd}, where the derived allele may be shared by yn but not d, or shared by all three populations. This allows us to define yn and ynd

separately. Because the proposed population tree in Fig. 1.1 is symmetrical with respect to the relationship between Y and either N or D, it is trivial to convert an expression for yn to an expression for yd.

1.2.2 Simulations

All methods were compared against data simulated in fastsimcoal 2.1 [24], [25]. Admixture events occurred over a single generation in all cases. Simulated admixture quantities, and simulated λ, are shown as dashed lines in all relevant figures, and all parameter values are summarized in Table 1.2. Archaic samples were not introduced to the simulation until the ages estimated for their corresponding samples in Green, Krause, Briggs, *et al.* [1] and Reich, Green, Kircher, *et al.* [2]. All simulations in Fig. 1.2 assume constant and equal size for all populations. Additional simulations with varying population size and event dates, but fewer repetitions, were conducted to verify methods and results are presented in Appendix A. The results with varying population size were identical to those with constant population size shown below, so we do not include them here.

1.2.3 Parameter Sensitivity

The initial proposals for $\hat{f}$ and $R_{\text{Neandertal}}$ concluded that they were unbiased, and they are nearly unbiased, so long as the assumptions of the models are met. However, Rogers and Bohlender [18] have recently presented alternative derivations showing that the methods tested here and others are biased when used without correction for ghost admixture, and then rely on several other parameters, some of which lack high-quality estimates. The results below are in agreement.

To study the extent of this effect, we have collected estimates and ranges for these parameters from the literature. When there were ambiguities due to uncertainty about mutation rate estimates, we chose an estimate based upon a mutation rate of 0.5×10^{-9}/yr.

Each estimate drew values from a truncated normal distribution with mean equal to the values in Table 1.2, and variance taken from their associated publications, presented in Tables 1.3 and 1.4. The ranges in this paragraph are stated as the width of the mean $\pm 1.96\sigma$. Values for λ and κ were chosen between 550–765kya and 381–473kya, respectively, following the estimates published in Prüfer *et al.* [4]. Values for ζ were drawn between 100–120kya matching Veeramah and Hammer [29]. Values for δ were drawn between 47–65kya to match the most likely range published by Sankararaman *et al.* [23]. Finally, γ was taken as 60–70kya, the center of the dates published for the Mesmaiskaya specimen [27] used in

Table 1.2. Values of parameters used in all simulations. All time parameters are scaled by $2N_0$ generations. The length of a generation is assumed to be 29 years to match [23], [26]. $2N_0$ is the size of the modern-archaic common ancestral population.

Parameter	Value
$2N_0$	20,000
α	0.0431
γ	0.1121
δ	0.0966
ζ	0.1897
κ	0.7362
λ	1.1328
m_N	5%
m_D	2%

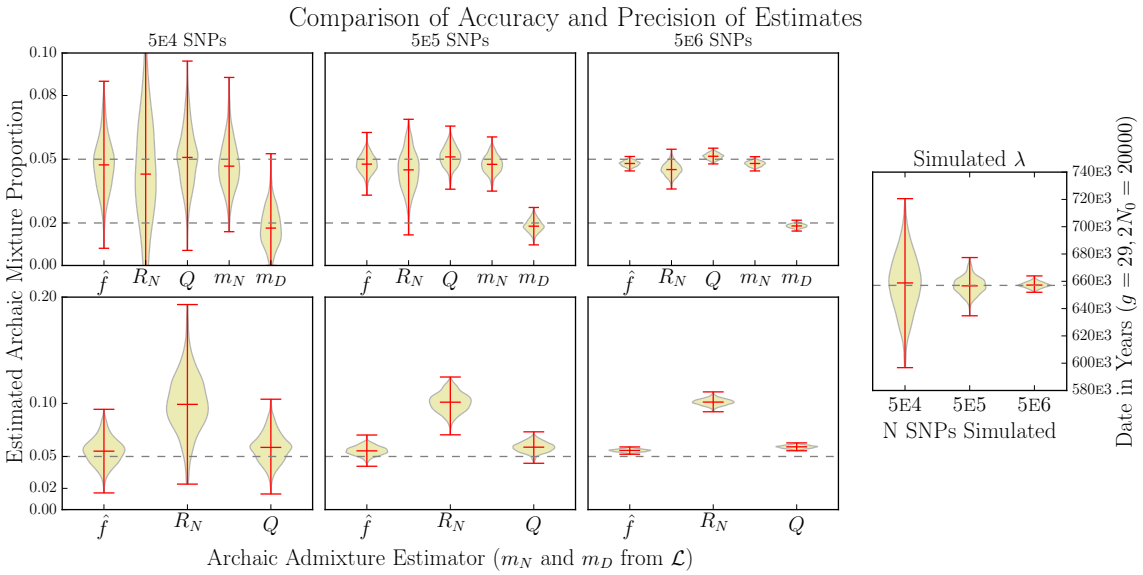

Figure 1.2. A comparison of estimates from each method studied, with and without correction for ghost admixture. Estimates with corrections are given in the top row, while estimates without correction are in the bottom row. The average number of SNPs across all runs is listed at the top of each column. True values are given as gray dashed lines, while the median estimate for each method is a red horizontal line. The simulated values of m_N, m_D, and λ are given in Table 1.2.

Table 1.3. A summary of the tested range of dates around each event, in years, and the corresponding source for the range. The ranges for m_D and α were chosen by the authors to reflect a reasonable level of uncertainty. All other values are $\pm 1.96\sigma$ where σ was available.

Parameter	Range	Source
α	20000–30000	**
γ	60000–70000	[1], [27]
δ	47000–56000	[5], [23]
ζ	100000–120000	[28]
κ	381000–473000	[4]
λ	550000–765000	[4]
m_D	0.015–0.025	**

Table 1.4. Previous estimates of the value of λ. We have assumed the higher range from Prüfer *et al.* for all of our analyses. [4]

Publication	Separation Date
Green *et al.* 2010 [1]	270,000–440,000
Meyer *et al.* 2012 [3]	170,000–700,000
Prüfer *et al.* 2014 [4]	270,000–383,000 550,000–765,000

Green *et al.* [1]. $2N_0$, the ancestral haploid population size, was assumed to be 20000, and each population value was perturbed ±20% during the sensitivity analysis.

No reliable estimates were available for α, so we chose to use $\alpha = 25$ and $\alpha = 25 \pm 5$ for the simulations and sensitivity analysis, respectively.

1.2.4 Estimation

The values of the site pattern counts were generated with simulated SNPs. This can be done with allele frequencies using the following:

$$I_{XY+XYD} = \sum_s p_X p_Y q_N$$

$$I_{XN+XND} = \sum_s p_X q_Y p_N$$

$$I_{YN} = \sum_s q_X p_Y p_N q_D$$

$$I_{YD} = \sum_s q_X p_Y q_N p_D$$

$$I_{YND} = \sum_s q_X p_Y p_N p_D$$

Here, p_A is the frequency of the derived allele in population A, and $q_A = 1 - p_A$ is the frequency of the ancestral allele in that population. The sum is across all SNPs s.

The I values were calculated from simulated data. The estimates for $\hat{f}$, $R_{\text{Neandertal}}$, and Q were easily calculated from this point. For $\mathcal{L}$, we used a combination of the basinhopping algorithm [30] and the TNC optimization algorithm [31], for bounded optimization. All optimizations used SciPy [32] and NumPy [33].

1.2.5 Sequencing Error

To evaluate the effect of sequencing error on each method tested, we took our simulated data and randomly swapped bases from ancestral to derived and vice versa. We did so at two rates, 1% and 0.1%. The 1% error rate represents an extreme case, as all bases switched will move from ancestral to derived, ignoring the possibility of switching to a third allele and being ignored in our analysis. Thus, 1%, which is higher than the estimated sequencing error and contamination in Prüfer, Racimo, Patterson, *et al.* [4], clearly demonstrates the effect, while 0.1% is meant to give a more realistic approximation of the effect in practice. Both cases are shown in Figs. 1.3 and 1.4.

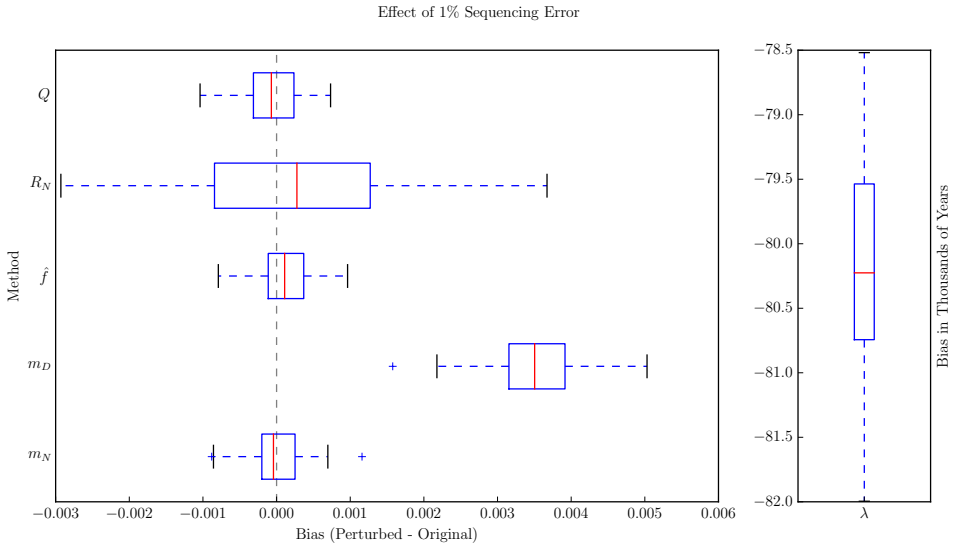

Figure 1.3. Effect of swapping 1% of all bases from ancestral to derived, simulating sequencing error. A more extreme effect than we would see in practice.

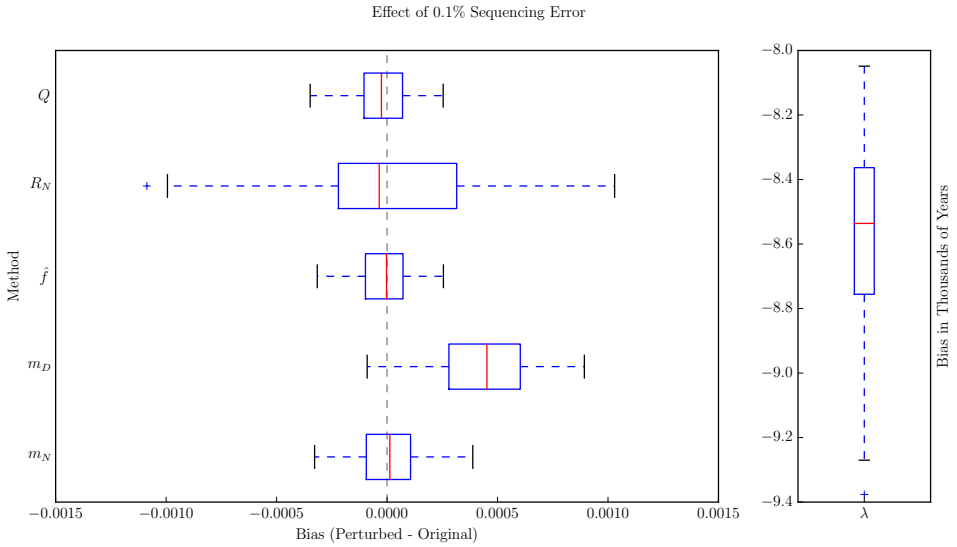

Figure 1.4. Effect of swapping 0.1% of all bases from ancestral to derived. A more realistic representation of the effect of sequencing error in practice.

1.3 Results

1.3.1 Accuracy and Precision

Each method we compared was tested against the same datasets as all other methods, and their outputs are compared in Fig. 1.2. For clarity, the individual estimates of m_N and m_D from $\mathcal{L}$ will be referred to as m_N and m_D, while the corresponding estimates of m_N from the other estimators will be referred to by the name of the estimator.

Fig. 1.2 shows the sampling distributions for all methods with and without correction for ghost admixture. $\hat{f}$, Q, m_N, and m_D all share similar levels of precision and accuracy, but $R_{\text{Neandertal}}$ is both less precise, and less accurate, at all sample sizes, with and without correction for ghost admixture. Additionally, while $\hat{f}$ and Q both show some upward bias when ghost admixture goes uncorrected, it is apparent that estimates from $R_{\text{Neandertal}}$ are shifted upward severely, with a median estimate twice as large as the simulated value. This confirms the theoretical results of Rogers and Bohlender [18] regarding the expected values of $\hat{f}$, Q, and $R_{\text{Neandertal}}$, and shows that the sampling distribution for $R_{\text{Neandertal}}$ is much wider than all of the other tested methods. This also reiterates the fact that individual corrections are necessary for each method derived from this framework.

Results from $\mathcal{L}$ demonstrate that it is an accurate estimator of both Neanderthal and Denisovan admixture. Given corrections, the simulated parameter values, and a large sample size, all methods perform reasonably well.

1.3.2 Estimate of λ

Our estimate of λ in simulated data 1.2 shows $\mathcal{L}$ is an accurate estimator of the archaic-modern separation date. Precision increases with sample size, and the median estimate across 500 independent simulations is accurate for all sample sizes, similar to the mixture proportion estimates from $\mathcal{L}$. In this case, our estimator was provided with accurate values for all other date parameters at each of the sample sizes used throughout.

1.3.3 Sensitivity to Error in Parameter Estimates

In general, all methods are affected to some degree by model misspecification, but to different extents and for different parameters. Figs. 1.5 and 1.6 are plots of the percentage change in the estimate as a function of the percentage change in a single parameter value, noted at the base of a column. The slope in each plot is the elasticity of the method noted at the left, with respect to the parameter at the base. The y-axis is normalized for all plots within a column, but not between columns. For any method the parameters α, δ, and ζ have relatively little effect, relative to κ, λ, and m_D. Looking at the ranges tested in Table 1.3,

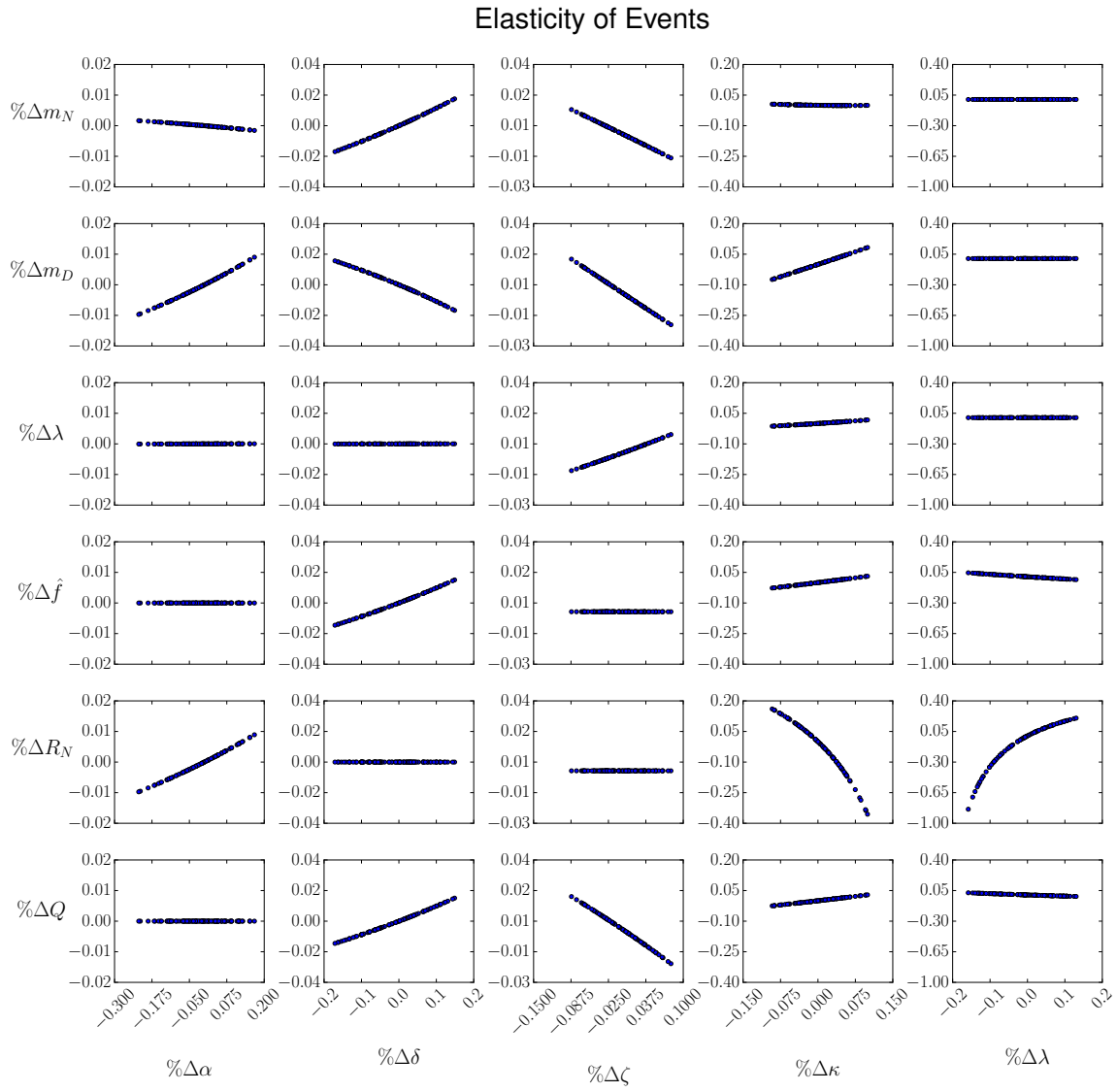

Figure 1.5. The relationship between each method and all date parameters we assume. Each subplot shows the percentage deviation from the true value of a parameter on the x-axis, and the resulting percentage change in the estimate on the y-axis. The slope in each of these graphs is the elasticity of the method with respect to a parameter. The parameter being modified is listed at the base of a column. Rows 1–3 correspond to the estimates generated by $\mathcal{L}$, while rows 4–6 correspond to $\hat{f}$, $R_{\text{Neandertal}}$, and Q, respectively.

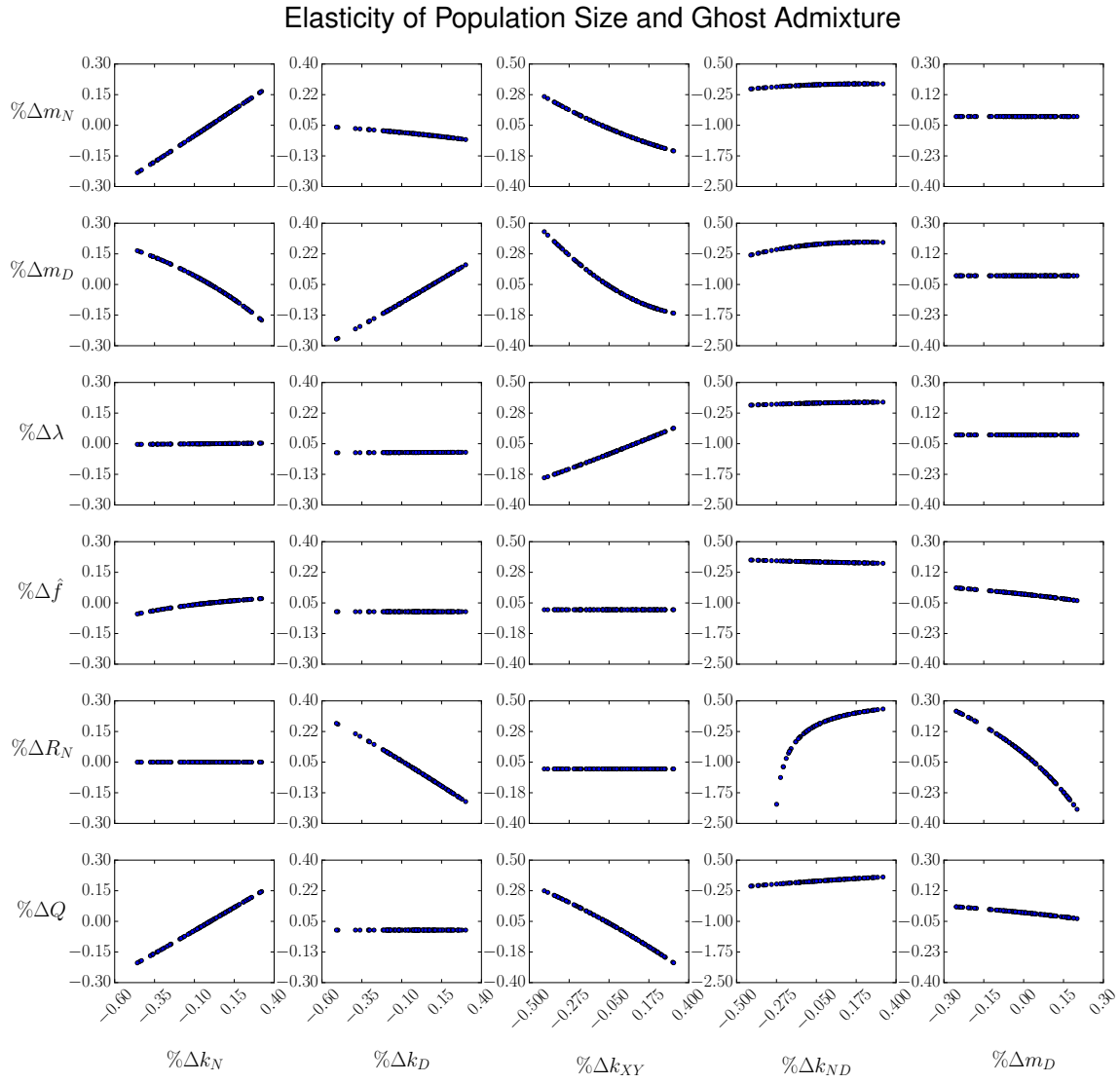

Figure 1.6. The relationship between each method and all population size parameters and the quantity of ghost admixture we assume. Each subplot shows the percentage deviation from the true value of a parameter on the x-axis, and the resulting percentage change in the estimate on the y-axis. The slope in each of these graphs is the elasticity of the method with respect to a parameter. Rows 1–3 correspond to the estimates generated by $\mathcal{L}$, while rows 4–6 correspond to $\hat{f}$, $R_{\text{Neandertal}}$, and Q, respectively.

both the values and ranges are much smaller than those being used for κ and λ, so it is unsurprising that they have a small effect. It is also clear that m_N is relatively insensitive to variation in all date parameters, while estimates of m_D are biased by a small amount when κ is misspecified. All methods have some sensitivity to population size, though $\hat{f}$ is clearly the least sensitive; the other methods are all quite sensitive to misspecification of population size, shown in Fig. 1.6.

For each column in Figs. 1.5 and 1.6, the scale of the y-axis is set by the most affected method, often $R_{\text{Neandertal}}$, where we see that estimates are severely modified by the three least-well-known parameters: κ, λ, and m_D. The estimate for λ shows little sensitivity to variation in any parameter. As with m_N and m_D, this is likely due to the fact that two of the parameters with the greatest uncertainty, λ and m_D, are being estimated directly in $\mathcal{L}$. The result is estimates of m_N, m_D, and λ with less bias contributed by date misspecification. It is clear however, that high-quality estimates of model parameters are mandatory for this family of methods.

1.3.4 Sequencing Error Sensitivity

λ is affected more heavily than any of the estimates of the mixture proportion from any method. In practice, this should not cause a large deviation from the true value in λ as well, but in extreme cases of sequencing error, we could see a significant deviation from the true value. This can also be interpreted as a demonstration of what happens to estimates when the ancestral and derived allele are reversed due to polarization errors. In that case, the effect on λ may be significant.

1.4 Discussion

It is clear from the results above that we must address the presence of ghost admixture when generating estimates of archaic mixture proportions. Previous work has identified multiple possible sources of archaic admixture into single modern populations [2]–[4], [19], evidence for archaic admixture from an unknown source into populations potentially lacking Neanderthal admixture [34], [35], and now there is evidence of non-Neanderthal admixture into Denisova [4]. As a result, understanding the effect of violating the assumptions about tree shape is essential for every method used to estimate the archaic mixture proportion. The estimated range of mixture proportion values, 1–4% Neanderthal in non-Africans [1], 3–6% Denisova in Melanesia [2], is widely cited, but the analysis here indicates that we should be skeptical of those estimates.

Though initially, there was some ambiguity about whether or not there was any Denisovan admixture in East Asia [2], [3], [19], there is increasing support for a low level of Denisovan admixture in East Asia [4]. Due to the shared history of the Neanderthal and Denisovan genomes, and our limited sample of Neanderthal genetic diversity, it is likely that additional Neanderthal admixture can be falsely identified as a smaller amount of Denisovan admixture, when in fact there was no Denisovan introgression at all. For this reason, and in light of the results above, any estimate that does not address the possibility of multiple

sources of admixture should be viewed with skepticism.

Even with the corrections presented previously [18], and demonstrated in Fig. 1.2, there remains a problem where an estimate of archaic admixture from Neanderthals depends on the estimate from Denisova, and vice versa. Furthermore, any such estimate where there is potentially additional admixture from a species we are unaware of can be difficult to correct within the current framework. $\mathcal{L}$ addresses the first part of this problem by simultaneously estimating the contribution of two archaic species, typically Neanderthal and Denisova. The latter problem has not been resolved, and the extent of the bias introduced will depend largely on the way the tree assumptions are violated.

λ, being one of the oldest population parameters with which we are concerned, and with the widest confidence interval [4, Supp. 12], may contribute significantly to error in mixture proportion estimates when misspecified. A widely repeated estimate of λ (i.e., archaic-modern separation) depends on linkage disequilibrium (LD) to generate their estimate of the same value [4, Supp. 12]. This estimate was generated using data that were computationally phased [4, Supp. 4], which introduces phasing errors [36, 37, 4, Supp. 4], breaks down LD, and is likely an overestimate of the population separation date. $\mathcal{L}$ does not use any LD information, but rather fits the best value of λ given a set of site patterns. We show in Figs. 1.5 and 1.6 that this estimate is vulnerable to misspecification of tree parameters and population size, but given parameter values within reasonable ranges, we generate accurate estimates that are robust to phasing errors.

We have presented a new composite-likelihood method for archaic admixture and separation date estimation, $\mathcal{L}$. We tested $\mathcal{L}$ under a variety of assumptions, and violations of assumptions, which has also provided additional verification of prior results. Previously, we presented Q, which uses the same approach as that used by Patterson *et al.* in each of the major archaic genome publications discussed [1]–[4], [18]. Q uses a ratio of expectations to calculate the mixture proportion directly, in much the same way that $\hat{f}$, $R_{\text{Neandertal}}$, etc. do, but requires only a single archaic individual.

While Q needs only a single archaic, both a Neanderthal and a Denisovan are necessary for $\mathcal{L}$, but it generates an estimate for the contribution of both species simultaneously. $\mathcal{L}$ generates reliable estimates of m_N and m_D for event dates within reasonable ranges and given reliable estimates of $2N$ for the populations ancestral to those used in the analysis. $\mathcal{L}$ addresses several concerns about the amount of Denisovan admixture present in populations that also contain Neanderthal admixture, and also provides an alternative estimate of λ,

the archaic-modern separation date, that does not rely on LD and thus is robust to phasing errors.

CHAPTER 2

ARCHAIC ADMIXTURE IN DATA CONTAINING ASCERTAINMENT BIAS

2.1 Introduction

There have been several recent large-scale publications using data from SNP microarrays to estimate admixture proportions in modern human populations, e.g., [19], [38]–[40]. These studies have used f-statistics, D-statistics, and F_4 ratio estimators to detect admixture, and estimate the contribution of one population to another [15]. Though f-statistics and D-statistics are supposed to be robust to ascertainment bias, this is not necessarily the case for F_4 ratio estimators.

Three of the previous publications, Reich, Patterson, Campbell, *et al.* [38], Lazaridis, Patterson, Mittnik, *et al.* [39], and Qin and Stoneking [40], used data from the Human Origins data set, which uses the Human Origins array, a SNP panel designed for population genetic studies [15], [41]. The ascertainment procedure for this SNP microarray was designed to make correcting for ascertainment bias simple, and as a result, the array is designed so that the ascertainment bias is limited to pq bias. That is, the dataset is biased toward SNPs with relatively high heterozygosity in many human populations, which tend to be polymorphisms with deep gene trees. This makes simulating the same type of ascertainment bias a simple procedure, and allows us to verify the sensitivity of some previously used methods to this type of ascertainment bias.

Though f-statistics and D-statistics may not give false positives in the face of ascertainment bias, there is no such guarantee for F_4 ratio estimators. In fact, f-statistics are only shown to generate values equal to 0 when the true value is zero, but the magnitude of the f-statistic is affected by ascertainment bias [15, p. 1073]. In this case, F_4 ratio estimators may well be biased, as they are intended to generate a point estimate — where magnitude is important — rather than test a null hypothesis of zero admixture.

Here we test four estimators: $\mathcal{L}$, $\hat{f}$, $R_{\text{Neandertal}}$, and Q. Three of these are F ratio

estimators, while $\mathcal{L}$ is a composite likelihood derived similarly to the F ratio estimators. We generate ascertainment bias in simulated data and compare estimates of admixture from the prior estimators to determine their sensitivity to ascertainment bias. Finally, we apply $\mathcal{L}$ and Q to data from the International HapMap Project [42] and the high coverage Neanderthal and Denisovan genomes [3], [4].

2.2 Methods

2.2.1 Simulations

Simulations were conducted using fastsimcoal 2.1 [25], with parameters chosen to fit the tree in Fig. 1.1. These parameters are given in Table 1.1. In all scenarios, the simulated mixture proportions into Y were 5.0% for Neandertal and 2.0% for Denisova.

Ascertainment bias was added to the simulated data using importance sampling. A site k was sampled with probability:

$$P_k = \frac{2p_k q_k}{\sum_i 2p_i q_i}$$

where p_k is the frequency of the derived allele at site k in population Y, and q_k is the frequency of the ancestral allele. Sites were drawn from the simulated dataset with replacement up to the number originally in the dataset, generating pq-bias. The analysis corresponds to the analysis in Bohlender and Rogers, with ascertainment bias now included [43].

2.2.2 Samples

We used the HapMap phase 3.2 release for our SNP data and both the Altai Neandertal [4] and the high coverage Denisova [3] extended vcf files for our archaic samples. In one set of tests, HGDP01029 — the San sample sequenced as part of Meyer et al. [3, Supp. 9] — was used as population X, i.e., the population expected to have no archaic admixture. In the other set, the HapMap YRI sample was used for population X. Results for both analyses are in Tables 2.1 and 2.2.

2.2.3 Data Conversion

Datasets from the HapMap were converted to VCF [44] using the Genome Analysis Toolkit (GATK) [45]. The data from the HapMap was aligned to human reference GRCHv36, while the whole genomes generated by Meyer, Kircher, Gansauge, *et al.* [3] were aligned to UCSC's hg19. The HapMap VCFs were converted to hg19 using the LiftoverVariants tool provided by the GATK.

Table 2.1. Estimates from $\mathcal{L}$ and Q assuming $\lambda = 440ky$ (only applicable to Q). Estimates of λ are given in years, assuming $2N_{\text{ancestral}} = 4616$ and generation length is 29 years. The column labeled Q is the estimate of m_N from Q. The first set of rows uses the San genome from Meyer, Kircher, Gansauge, *et al.* [3] as the non-admixed human group, while the second set uses the HapMap YRI population.

X	Y	m_N	m_D	λ	Q
San	ASW	0.0431	0.0390	447171	-0.0136
San	CEU	0.0456	0.0379	445674	-0.0147
San	CHB	0.0462	0.0391	445800	-0.0120
San	CHD	0.0460	0.0388	445667	-0.0135
San	GIH	0.0446	0.0381	445745	-0.0160
San	JPT	0.0461	0.0392	445584	-0.0138
San	LWK	0.0420	0.0388	447250	-0.0145
San	MEX	0.0461	0.0384	446309	-0.0100
San	MKK	0.0419	0.0381	447048	-0.0143
San	TSI	0.0453	0.0381	445761	-0.0144
San	YRI	0.0423	0.0393	447308	-0.0153
YRI	ASW	0.0327	0.0280	448665	0.0003
YRI	CEU	0.0355	0.0291	447199	-0.0033
YRI	CHB	0.0357	0.0299	447348	-0.0021
YRI	CHD	0.0355	0.0297	447201	-0.0030
YRI	GIH	0.0348	0.0290	447259	-0.0037
YRI	JPT	0.0356	0.0299	447118	-0.0032
YRI	LWK	0.0324	0.0280	448804	0.0000
YRI	MEX	0.0358	0.0294	447839	-0.0003
YRI	MKK	0.0331	0.0284	448627	-0.0003
YRI	TSI	0.0353	0.0293	447309	-0.0030

Table 2.2. Table of estimates from $\mathcal{L}$ and Q, assuming $2N_{XY} = 5 \times 10^4$. Estimates of λ are given in years. The column labeled Q is the estimate of m_N from Q. The first set of rows uses the San genome from Meyer *et al.* as the non-admixed human group, while the second set uses the HapMap YRI population.

X	Y	m_N	m_D	λ	Q
San	ASW	0.0392	0.0346	447352	-0.0094
San	CEU	0.0415	0.0336	445800	-0.0101
San	CHB	0.0421	0.0346	445940	-0.0082
San	CHD	0.0419	0.0344	445801	-0.0092
San	GIH	0.0405	0.0337	445872	-0.0110
San	JPT	0.0421	0.0347	445720	-0.0095
San	LWK	0.0382	0.0345	447430	-0.0103
San	MEX	0.0420	0.0340	446455	-0.0068
San	MKK	0.0380	0.0338	447207	-0.0098
San	TSI	0.0413	0.0337	445889	-0.0099
San	YRI	0.0385	0.0349	447499	-0.0105
YRI	ASW	0.0293	0.0245	448633	0.0002
YRI	CEU	0.0318	0.0254	447155	-0.0023
YRI	CHB	0.0321	0.0262	447314	-0.0014
YRI	CHD	0.0319	0.0260	447163	-0.0021
YRI	GIH	0.0312	0.0253	447211	-0.0025
YRI	JPT	0.0320	0.0262	447079	-0.0022
YRI	LWK	0.0289	0.0245	448777	0.0000
YRI	MEX	0.0322	0.0257	447817	-0.0002
YRI	MKK	0.0296	0.0249	448606	-0.0002
YRI	TSI	0.0317	0.0256	447268	-0.0020

The data from Meyer, Kircher, Gansauge, *et al.* [3] were in "Extended Variant Call Format", where each site is annotated with additional information, including the base from the inferred human-chimp common ancestor, the CAnc field. Bases in the CAnc INFO field were assumed to be ancestral and used to polarize the data.

2.2.4 Filtering

All analysis included only the sites that passed the Map35_100% criteria, used by Prüfer, Racimo, Patterson, *et al.* [4, Supp. 5b, Supp. 14]. The minimum filtered site list was downloaded from `<http://bioinf.eva.mpg.de/altai_minimal_filters/>`. This set of sites was further restricted to SNPs that are biallelic — have only a single alternate allele — across all samples, on chromosomes one through twenty-two. We additionally excluded all sites flagged as in a CpG context, all systematic errors, and all sites with missing data in any individual.

2.2.5 Bootstraps

$\mathcal{L}$ is a composite likelihood, which assumes independence between sites. Because our data are not independent, we use a moving blocks bootstrap to produce confidence intervals around our estimates [46]. We do 100 replicates with a block size of 100,000 SNPs. Bootstrap confidence intervals for all estimates are provided in Appendix B.

2.2.6 Parameters

Parameter estimates were taken from the literature where available, and their assumed values are provided in Table 2.3. Because our estimate of λ — the date of separation between modern and archaic populations — differed so much from the published values, we decided to run our data analysis twice. When we collected population size estimates for the XY population, we calculated the harmonic mean over the interval from $\lambda = 653,000$ years to $\zeta = 110,000$ years from Li and Durbin [47], and again from $\lambda = 440,000$ years to $\zeta = 110,000$ years. The former case corresponds to the estimate from [4], while the latter matches our own estimate. This gives two different population sizes for the XY population, to which both $\mathcal{L}$ and Q are sensitive, and a different value of λ, to which Q is sensitive [18], [43].

2.3 Results

Simulations demonstrate a mild upward shift in the estimate of Neanderthal admixture in $R_{\text{Neandertal}}$, and essentially no change in $\mathcal{L}$ and Q. However, $\hat{f}$ generates estimates that

Table 2.3. Assumed parameter values for all estimates made on the HapMap data set. Population sizes, $2N$, are given as the number of haploid individuals in a population indicated by the subscript. Dates are given in units of $2N_0$ generations, with generation length assumed to be 29 years, and $2N_0$ equal to $2N_{\text{ancestral}}$. The separate values for λ and $2N_{XY}$ are given because our estimate of λ diverged so much from the estimate in Prüfer, Racimo, Patterson, *et al.* [4].

Parameter	Assumption	Source
$2N_{\text{ancestral}}$	4616	Prüfer, Racimo, Patterson, *et al.* [4]
$2N_N$	1154	Prüfer, Racimo, Patterson, *et al.* [4]
$2N_D$	1154	Prüfer, Racimo, Patterson, *et al.* [4]
$2N_{XY,\text{young}}$	19184	Li and Durbin [47]
$2N_{XY,\text{old}}$	17425	Li and Durbin [47]
$2N_{ND}$	1154	Prüfer, Racimo, Patterson, *et al.* [4]
α	25,000	*A guess
δ	56,000	Sankararaman, Patterson, Li, *et al.* [23]
ζ	110,000	Henn, Cavalli-Sforza, and Feldman [28]
κ	427,000	Prüfer, Racimo, Patterson, *et al.* [4]
λ_{young}	444,000	*Approximate $\mathcal{L}$ estimate
λ_{old}	653,000	Prüfer, Racimo, Patterson, *et al.* [4]

are more than twice as large as the simulated value. This result is shown at each of 3 sample sizes in Fig. 2.1 and supports the argument that F_4 ratio estimators need to be independently evaluated for biases, made previously in [18], [43]. We also find that estimates of λ from $\mathcal{L}$ tend to be slightly younger than the simulated value, shown in the rightmost panel of Fig. 2.1.

Estimates of archaic admixture into all HapMap populations — including YRI — were made with San as population X. Estimates using YRI as population X did not use San at all. The results are presented in Tables 2.1 and 2.2. Confidence intervals for the estimates are given in Appendix B, and are generally narrow around each estimate, ranging $\pm$ 2% of the point estimate. When San is used as population X — the population assumed to have no archaic admixture — all other populations show evidence of archaic admixture. This fits with previous results from [34], which found evidence for archaic admixture in Africa without the use of an archaic genome. It also indicates that analyses using YRI as an outgroup should be viewed as lower bounds on archaic admixture, as F_4 ratio estimators produce estimates relative to the modern outgroup population, X.

Notably, estimates of λ from $\mathcal{L}$ are consistent across population size assumptions and regardless of the modern outgroup population used. This is encouraging, as we would expect λ to be the same for all modern human populations, so switching the modern outgroup from

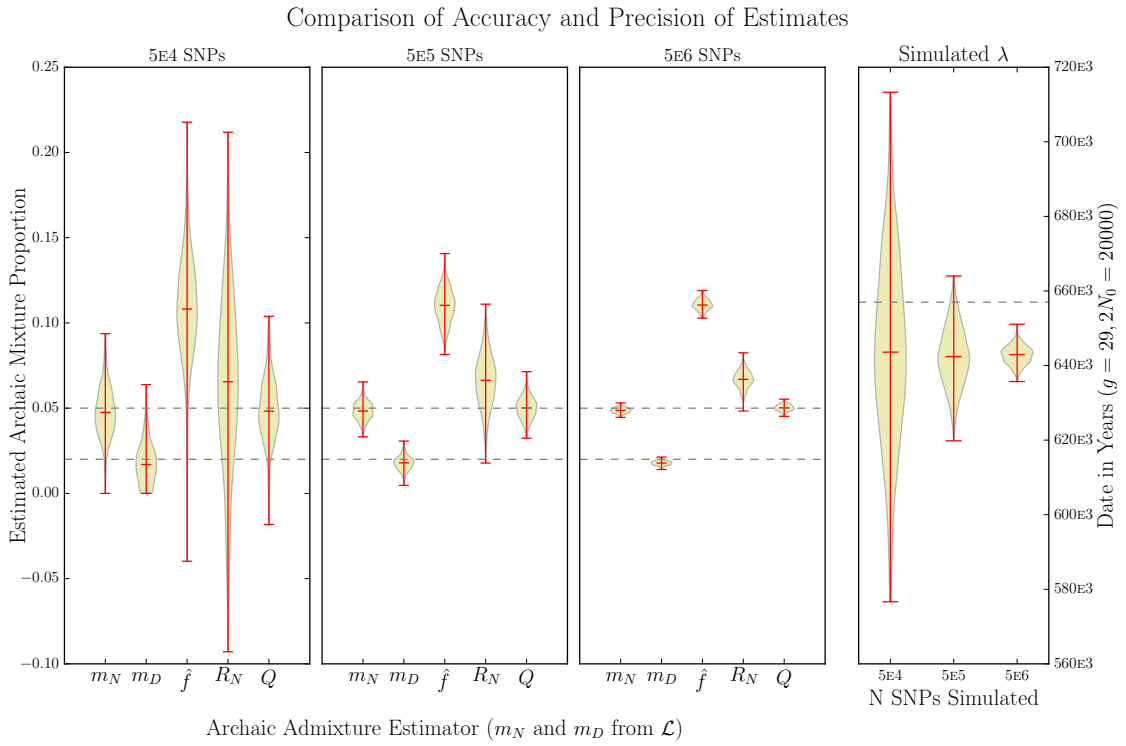

Figure 2.1. Violin plot with the median (red line in violin) and distribution of 500 point estimates from replications of the ascertainment bias resampling procedure on independent simulations. In the three left figures, the upper dashed line corresponds to the simulated value of m_N while the lower dashed line is m_D. In the rightmost figure, the dashed line corresponds to the simulated value of λ, and the violins show the distribution of estimates of λ from $\mathcal{L}$.

San to YRI should not change the estimate, as we see in Tables 2.1 and 2.2.

2.4 Discussion

Our simulations support the use of some admixture estimators — $\mathcal{L}$ and Q — on data with ascertainment bias. It also shows that estimators must be evaluated on an individual basis, as $\hat{f}$ and $R_{\text{Neandertal}}$ are more sensitive to ascertainment bias, but to different degrees. This is problematic for studies like [39], [48], which do many comparisons between modern human populations using methods in the same family as the methods tested here. These analyses suffer from a much more complicated set of relationships between populations as modern human populations are more likely to have admixed recently, so model misspecification is more likely. Adding the ascertainment bias effect on top of the ghost admixture effect means that the admixture point estimates in [38], [39], [48] are likely inaccurate, potentially by double or more.

We conducted two sets of analyses, using different assumptions for λ, reflecting differences in our archaic-modern split estimate and the estimates used more broadly in the literature [4]. In each of these analyses, we used both YRI and San as the outgroup. We found elevated admixture estimates for all Y populations when using San as the outgroup, implying less archaic admixture in San than in other populations. This also documents some variability in levels of archaic admixture in Africa, supporting previous analyses by [34], [35], [49]. This also raises concerns for analyses that lump African populations, as they will miss the variability in African populations. They will also tend to underestimate the amount of archaic admixture outside of Africa, because these methods generate estimates relative to population X. This is particularly true for approaches that exclude all sites with the ancestral base in Africa, e.g., "enhanced" D-statistics in [2].

The difference between our estimate of λ and the estimate in [4] could be due to additional model misspecification that we are not accounting for here. There are also some parameters to which $\mathcal{L}$ is sensitive, and which may reduce the estimate of λ. Those are shown in Figs. 1.5 and 1.6. These may cause a lowered estimate of λ, but unless the estimates in the literature corresponding to each of the parameters used in the model are drastically incorrect, we should not be seeing an underestimate of more than 110ky of the lower bound of the estimate from [4]. An alternative is that the estimate in [4] is biased upward due to its use of PSMC to estimate the separation date [4, Supp. 12]. The data there were computationally phased, which may introduce phasing errors. These errors may break down LD, and thus artifically increase the apparent separation date. $\mathcal{L}$ does not use any LD information, and is therefore immune to that form of bias. Instead, the estimates are generated based upon an assumed mutation rate, $\mu = 0.5 \times 10^{-9}$, inherent in the values we use for the events listed in Table 2.3.

We have shown here that further analyses using data with ascertainment bias, or more generally data that has biased heterozygosity through filtering or other means, should use methods in the F_4 ratio estimator family carefully. These methods differ in their sensitivity to various forms of bias, and as the relationships between the populations of interest become more complicated, there is more opportunity for ghost admixture to generate misleading results. Though they are still sensitive to many different parameters, $\mathcal{L}$ and Q are both largely insensitive to ascertainment bias, making them ideal choices for work where the population relationships are well understood, and the data come from SNP arrays.

CHAPTER 3

ARCHAIC ADMIXTURE IN WHOLE GENOME DATA

3.1 Introduction

F_4 ratio estimators have been used frequently since their introduction [20]. Unfortunately, many of these statistics are biased by admixture from a second archaic population [18]. These estimators typically model relationships between populations as a tree representing the relationships between populations, and allow only a single source of admixture in a binary comparison [15]. Though corrections have been proposed, they are largely inappropriate for use in contexts where we suspect multiple sources of admixture, due to a reliance on an initial estimate of ghost admixture to use as a correction [18]. In an archaic admixture context, this may mean using the estimator in populations with both Neanderthal and Denisovan admixture [4]. In modern populations, this will be true for many comparisons, particularly between populations that have been historically geographically close, e.g., Europe and the Near East, New World and European populations, etc. [20], [38]–[40]. All current corrections for ghost admixture depend on an estimate of admixture from a ghost source, which presents a clear problem for generating an initial estimate of admixture in a population.

A recent example of this problem was raised by Rogers and Bohlender [18]. In their Figure 9, they demonstrate that if the estimate of Denisovan admixture into Melanesia from [2] is accurate, the estimated value from $R_{\text{Neandertal}}$ — an F_4 ratio estimator — should be much larger than the true value [2]. This would be true even if there were no admixture into Melanesia from Neandertals at all.

To address this issue, we have proposed a maximum likelihood estimator of the archaic mixture proportion for a population, $\mathcal{L}$, which generates estimates for two archaic populations simultaneously [43]. Here, we use this new method and Q to generate estimates of Neandertal and Deniosovan admixture into the human genomes published by Meyer, Kircher, Gansauge, *et al.* [3].

3.2 Methods

All methods discussed assume a relationship between populations matching Fig. 1.1. Population names reflect the tips of that tree. $\mathcal{L}$ is a composite likelihood estimator of three values: m_N, the mixture proportion for Neandertals into population Y;m_D, the mixture proportion for Denisovan into Y;λ, the date of separation of the Archaic and Modern populations [43]. Estimates of all three values in $\mathcal{L}$ are sensitive to parameter assumptions. Assumed values for each parameter are given in Table 2.3.

Q is an F_3 ratio estimator created by Rogers and Bohlender as an alternative to previously published methods [18]. It is sensitive to misspecification of population size in earlier epochs but is otherwise a relatively stable estimator that requires only a single archaic sample to generate an estimate.

Our results for the date of population separation between modern and archaic populations, λ, are significantly different from the published estimates of the same [4]. We run our analyses twice, to contrast the implications of the different λ estimates. First, we assume the older date, and use the population size calculated as the harmonic mean of the population sizes between λ and ζ for population XY, taken from [47]. Second, we assume our estimated value of λ, and acquire the population size for population XY in the same way. $\mathcal{L}$ and Q are both sensitive to the population size assumed for population XY, but only Q is sensitive to the assumed value of λ.

3.2.1 Bootstraps

We assume independence between sites, but this is not the case in our data. As a result, we cannot use standard likelihood methods to generate confidence intervals around our estimates. Instead we use moving blocks bootstraps to estimate confidence intervals [46]. We do 100 replicates with a block size of 100,000 SNPs. Confidence intervals around estimates are given in Appendix B.

3.2.2 Data

The data used for analysis were sequenced as part of the Meyer, Kircher, Gansauge, *et al.* [3] analysis of the high coverage Denisovan genome. We used the extended vcf files provided at `<http://cdna.eva.mpg.de/denisova/VCF/>` (Accessed 12–17–2014) for our modern whole genome samples and the high coverage Denisovan genome. We also use the high coverage Neandertal genome from Altai [4], downloaded from `<http://cdna.eva.mpg.de/neandertal/altai/AltaiNeandertal/VCF/>`.

3.2.3 Filtering

All analysis included only the sites that passed the Map35_100% criteria, used in [4, Supp. 5b, Supp. 14]. The minimum filtered site list was downloaded from <http://bioinf.eva.mpg.de/altai_minimal_filters/>. This set of sites was further restricted to SNPs that are biallelic, i.e., have only a single alternate allele, across all samples, on chromosomes one through twenty-two. We additionally excluded all sites flagged as in a CpG context, all systematic errors, and all sites with missing data in any individual.

3.3 Results

As in our analysis of the data from the International HapMap Project, we find significant divergence between estimates of m_N between $\mathcal{L}$ and Q. These results are shown in Table 3.1.

3.4 Discussion

As in our analysis with ascertainment bias [50], we find a divergence between estimates from $\mathcal{L}$ and Q in Tables 3.1 and 3.2. This divergence indicates some difference between the assumed parameters and the history that lead to our data. Alternatively, there may be additional sources of admixture that are further biasing these methods. There has been some evidence of additional archaic admixture from a non-Neanderthal into Denisova, but simulations of this pattern result in essentially no change in the estimates from $\mathcal{L}$ and Q. If the issue is another source of ghost admixture, then it is unlikely to be due to admixture into Denisova.

It is important to note the distinction between estimates using San as population X and using Yoruba as population X. Previously, Africa has been treated as a unit for some analyses [2] [3]. The results here indicate two things: diversity in Africa should be considered when conducting future analyses, and the uneven distribution of Neanderthal and Denisovan admixture indicates a history of archaic admixture in Africa. The first point fits with arguments that have been made previously regarding genetic structure in Africa [51], [52]. The second supports prior studies that found evidence for archaic admixture in Africa without an archaic genome for reference [34], [35]. We find a larger signal of Neanderthal ancestry in Yoruba than in other African populations tested, which seems consistent with previous results indicating gene flow from Europe into Yoruba [49], [53].

All estimates of both Neanderthal and Denisovan admixture are larger worldwide when using San as the outgroup, implying a lower overall level of archaic admixture in San than other populations. Using $\mathcal{L}$, we find higher levels of Neanderthal admixture than has been reported using F_4 ratio estimators, but lower levels of Denisovan admixture in Melanesia

Table 3.1. Table of estimates from $\mathcal{L}$ and Q, assuming $\lambda = 653,000$ years. Estimates of λ are given in years, assuming a generation length of 29 years, and a haploid ancestral population size of 4616. The first set of estimates uses the San genome as the non-admixed human group, while the second set uses the Yoruba genome. All estimates from Q assume no Denisovan admixture.

X	Y	m_N	m_D	λ	Q
San	Dai	0.0468	0.0366	448876	0.0069
San	Dinka	0.0384	0.0363	448005	-0.0156
San	French	0.0453	0.0349	448408	0.0000
San	Han	0.0493	0.0359	449324	0.0147
San	Karitiana	0.0466	0.0365	448621	0.0042
San	Mandenka	0.0393	0.0385	448331	-0.0143
San	Mbuti	0.0400	0.0376	448743	-0.0078
San	Papuan	0.0485	0.0417	450824	0.0191
San	Sardinia	0.0451	0.0352	448359	-0.0002
San	Yoruba	0.0399	0.0373	448180	-0.0127
Yoruba	Dai	0.0387	0.0285	449792	0.0145
Yoruba	Dinka	0.0296	0.0272	448813	-0.0022
Yoruba	French	0.0369	0.0267	449122	0.0093
Yoruba	Han	0.0409	0.0279	450250	0.0202
Yoruba	Karitiana	0.0379	0.0279	449591	0.0124
Yoruba	Mandenka	0.0293	0.0277	448993	-0.0010
Yoruba	Mbuti	0.0383	0.0352	450287	0.0042
Yoruba	Papuan	0.0405	0.0334	452115	0.0237
Yoruba	Sardinia	0.0365	0.0268	449210	0.0091

Table 3.2. Table of estimates from $\mathcal{L}$ and Q, assuming $\lambda = 444,000$ years. Estimates of λ are given in years, assuming a generation length of 29 years, and a haploid ancestral population size of 4616. The first set of estimates uses the San genome as the non-admixed human group, while the second set uses the Yoruba genome.

X	Y	m_N	m_D	λ	Q
San	Dai	0.0428	0.0325	449064	0.0047
San	Dinka	0.0347	0.0322	448169	-0.0107
San	French	0.0413	0.0309	448571	0.0000
San	Han	0.0452	0.0319	449516	0.0101
San	Karitiana	0.0426	0.0324	448805	0.0029
San	Mandenka	0.0356	0.0342	448534	-0.0098
San	Mbuti	0.0363	0.0334	448936	-0.0053
San	Papuan	0.0445	0.0374	451124	0.0131
San	Sardinia	0.0412	0.0312	448522	-0.0001
San	Yoruba	0.0361	0.0331	448362	-0.0087
Yoruba	Dai	0.0350	0.0250	449797	0.0099
Yoruba	Dinka	0.0263	0.0237	448761	-0.0015
Yoruba	French	0.0332	0.0233	449085	0.0064
Yoruba	Han	0.0371	0.0245	450259	0.0138
Yoruba	Karitiana	0.0342	0.0244	449583	0.0085
Yoruba	Mandenka	0.0260	0.0242	448942	-0.0007
Yoruba	Mbuti	0.0347	0.0313	450458	0.0029
Yoruba	Papuan	0.0369	0.0298	452261	0.0163
Yoruba	Sardinia	0.0329	0.0234	449175	0.0063

[1]–[4], [40], [54][cf. 55]. It is important to note that, given our limited sample availability, we do not have a clear picture of Neanderthal and Denisovan genetic diversity. As a result of the close relationship of Neanderthals and Denisovans, some populations with little or no actual Denisovan admixture may have inflated estimates simply because we have relatively little Neanderthal data. The reverse is also true. This is just a restatement of the problem for these modeled approaches when confronted with ghost admixture, though in this case the "ghosts" are Neanderthals that we haven't sampled. The issue of sparse sampling and its effects on D-statistics and similar methods has been presented previously in other literature [56], [57].

Also of note, using San as population X results in significantly higher estimates for m_N and m_D with $\mathcal{L}$, while estimates of λ remain stable. We find that our λ estimates from whole genome data match our results from the Internation HapMap Project [42]. This result makes sense, as all modern populations should have separated from archaics at the same time. It also supports prior results indicating the robusticity of $\mathcal{L}$ to ascertainment bias. The difference between our estimate of the archaic-modern separation date and that of [4] can be explained by a lack of dependence on LD information in $\mathcal{L}$. The data used to generate the estimates of λ in [4, Supp. 12] were computationally phased [4, Supp. 4]. This computational phasing introduces phasing errors, which may break down LD and artificially increase the estimated age [58]. $\mathcal{L}$ requires no phasing, nor does it use any information from that process. The age of the separation of archaic and modern humans is instead inferred from the distribution of shared derived states. While $\mathcal{L}$ and the λ estimate are sensitive to model misspecification (see Figs. 1.5 and 1.6), it should not diverge from the estimate in [4] by as much as 110ky from the bottom of their estimated range. The archaic-modern separation estimate from [4] has been used frequently since its publication, so this divergence in estimates should provoke additional research on this split.

We suggest caution when interpreting results from F_4 ratio estimators and similar methods. Without a clear understanding of how misspecifying relationships between populations affects estimates of admixture, we should be skeptical of results depending on them. The problem of specifying the tree correctly is manageable with archaic samples, as Neanderthals and Denisovans together clearly form a clade relative to modern humans. This problem is exacerbated with modern human samples, as in the Human Origins dataset. Recent analyses relied on F_4 ratio estimators to make pairwise population comparisons in these large data sets [38]–[40]. Given the results of our analyses — both analytically [18] and in simulation [43] — the complexity of relationships between populations and the unknown

magnitude of the biases that result make it difficult to believe any of the point estimates and confidence intervals that resulted. This concern is made clear in our estimate of the amount of Neanderthal and Denisovan admixture in Papua New Guinea. Here, we used the same data as [3], but we find a pattern opposite theirs, greater Neanderthal admixture than Denisovan. This pattern is similar to the rest of the world, though the admixture fraction is greater for both Neanderthals and Denisovans in Papua New Guinea, indicating another wave of introgression into this population, and a yet more complicated history of introgression.

APPENDIX A

MATRIX COALESCENT $\mathcal{L}$ DERIVATION AND TESTS

A.1 Introduction

Here we use the matrix coalescent as derived by Wooding and Rogers [59], and originally due to Griffiths and Tavaré [60], to derive a model of admixture between two archaic hominin populations and a modern human population. This method can be used to derive the models used for the family of statistics typically referred to as F_4 ratio estimators [15], [20], [61]. Our method, $\mathcal{L}$, assumes a model that is similar to the models used for the F ratio estimators, though it is unique in the sense that we model contributions from Neandertals and Denisovans simultaneously. As a result, we have intervals in the past where three or more lineages may reside in a single population, complicating the probability of a particular pattern in the data, and increasing the number of cases where we may observe these patterns. The matrix coalescent gives the probability of having a number of remaining lineages at the end of an epoch — or a number of coalescent events within an interval — going backward in time. Using this, we can calculate the expected length of specific combinations of lineages. These lengths, an assumed mutation rate, and the number of possible pairings between lineages give us the probability of seeing a specific site pattern.

The tree we envision for $\mathcal{L}$ and Q is provided in Fig. A.1. We will refer to the populations at the tips of the tree with capital letters, and the samples from those populations in lowercase letters.

Outside of the ND population, the derivation of the site patterns for $\mathcal{L}$ is essentially identical to the derivations for Q in Rogers and Bohlender [18]. Because $\mathcal{L}$ also follows admixture from D into Y, there are up to three lineages in ND, and four lineages in the ancestral population. We only include this for the site patterns that are "aware" of D: yn, yd, and ynd.

Using the matrix coalescent — see [59] for full derivation — the probability of having i lineages left from an original three at the end of an epoch with length v is given in the vector:

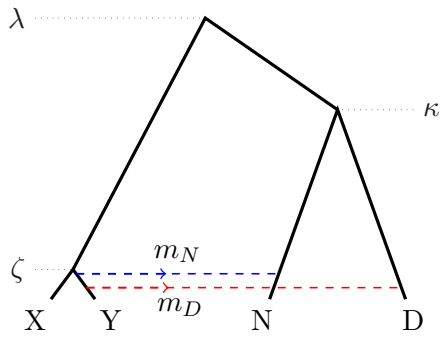

Figure A.1. The tree being described throughout the paper. Events are labeled with Greek letters, populations are labeled with capital letters. The direction of admixture moving backward in time is indicated with a blue line and arrow for Neandertals and a red line and arrow for Denisova.

$$\begin{pmatrix} \frac{3}{2}e^{-v} - \frac{3}{2}e^{-3v} \\ e^{-3v} \end{pmatrix}$$

The above vector describes the probability of having two lines of descent, and three lines of descent, at the end of the epoch. The probability of a single line of descent at the end of the epoch is then one minus the sum of the elements.

The expected length of the branches for each count of lineages in the interval is:

$$P_{ND} \begin{pmatrix} 1 - \frac{3}{2}e^{-v} + \frac{1}{2}e^{-3v} \\ \frac{1}{3}(1 - e^{-3v}) \end{pmatrix}$$

Again, the expected length for a single remaining lineage is the length of the interval less the sum of the elements of the above vector.

Now that we have these quantities, the rest is assembling each of the intervals and their associated probabilities. For example, the above probabilities are for three or two lineages at the end of an interval. But in ND, with y, n, d all present, yn only forms 1/3 of the time, so we must divide the probabilities by the number of possible pairings to get the probability of a single site pattern.

A.1.1 Coalescent Simulations

We used a custom coalescent simulation to confirm theoretical results. We simulated four populations, and varied all date and population size parameters in Table 1.1. The same values were given to the theory and the results are presented in Fig. A.2. We find good agreement between the simulated and predicted values, indicating that the theoretical results above fit well with any history matching the assumed tree structure, given appropriate dates and population sizes.

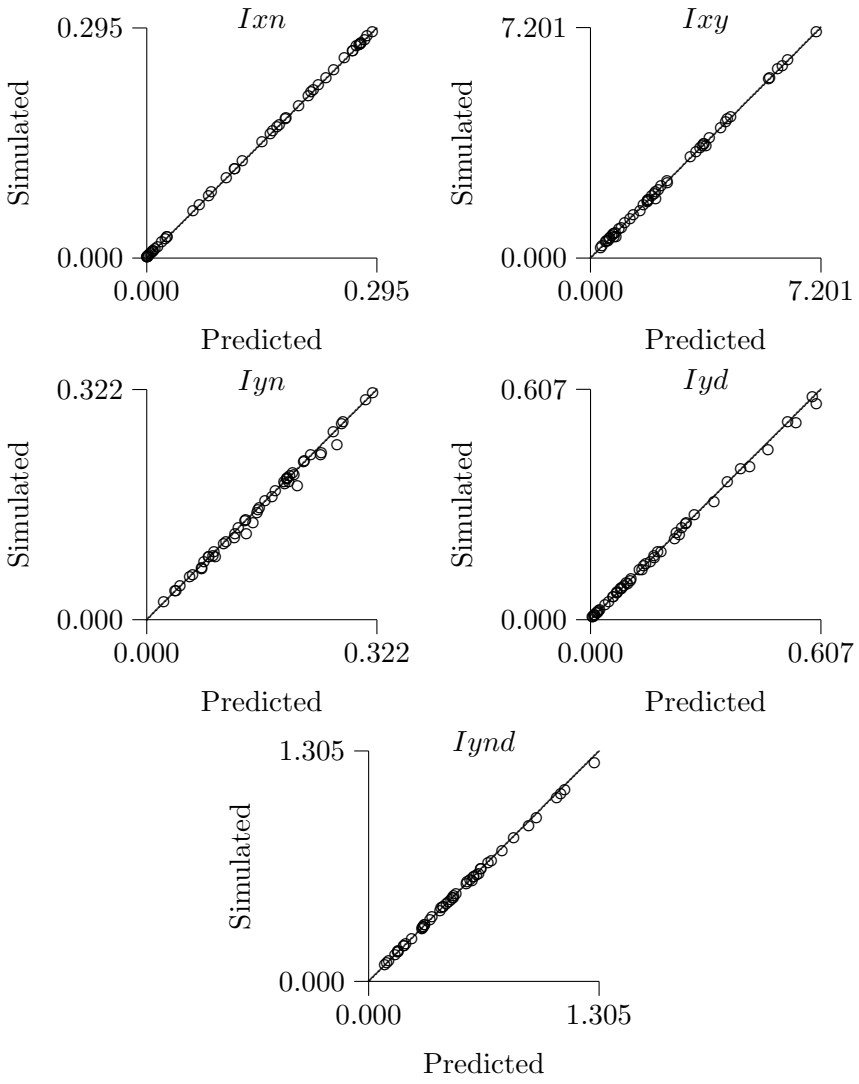

Figure A.2. Predicted versus observed site pattern counts from a coalescent simulation. All parameters in the model were varied in the simulation. The theory and simulations match well, supporting the derivations above.

APPENDIX B

CONFIDENCE INTERVALS

Table B.1. Table of confidence intervals for $\mathcal{L}$ and Q. Estimates of λ are given assuming generation length is 29 years and $2N_0$, the ancestral population size, is 4616. Q here assumes lambda is the midpoint of the estimates from Prüfer, Racimo, Patterson, *et al.* [4], 653 thousand years. The first set of rows uses the San genome from Meyer, Kircher, Gansauge, *et al.* [3] as the non-admixed human group, while the second set uses the HapMap YRI population.

		m_N		m_D		λ		Q	
X	Y	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%
San	ASW	0.0418	0.0443	0.0382	0.0402	446641	447615	-0.0185	-0.0105
San	CEU	0.0434	0.0476	0.0371	0.0392	444973	446317	-0.0239	-0.0065
San	CHB	0.0437	0.0480	0.0381	0.0403	444898	446659	-0.0233	-0.0018
San	CHD	0.0436	0.0477	0.0379	0.0397	444896	446469	-0.0228	-0.0059
San	GIH	0.0425	0.0461	0.0371	0.0394	445036	446586	-0.0263	-0.0096
San	JPT	0.0442	0.0482	0.0385	0.0405	444675	446180	-0.0271	-0.0061
San	LWK	0.0410	0.0434	0.0380	0.0399	446818	447681	-0.0185	-0.0130
San	MEX	0.0442	0.0481	0.0374	0.0397	445598	446884	-0.0189	-0.0030
San	MKK	0.0408	0.0431	0.0371	0.0393	446620	447364	-0.0178	-0.0111
San	TSI	0.0436	0.0473	0.0372	0.0394	444977	446901	-0.0239	-0.0051
San	YRI	0.0412	0.0437	0.0385	0.0405	446958	447669	-0.0181	-0.0125
YRI	ASW	0.0317	0.0339	0.0276	0.0289	448387	449053	-0.0008	0.0014
YRI	CEU	0.0341	0.0373	0.0287	0.0302	446676	447888	-0.0072	0.0015
YRI	CHB	0.0337	0.0375	0.0294	0.0308	446488	448267	-0.0087	0.0028
YRI	CHD	0.0337	0.0372	0.0294	0.0306	446425	448092	-0.0091	0.0017
YRI	GIH	0.0334	0.0362	0.0285	0.0301	446603	447935	-0.0084	-0.0002
YRI	JPT	0.0344	0.0378	0.0295	0.0307	446323	447963	-0.0080	0.0016
YRI	LWK	0.0316	0.0335	0.0276	0.0292	448432	449188	-0.0007	0.0007
YRI	MEX	0.0340	0.0376	0.0289	0.0304	446989	448490	-0.0064	0.0038
YRI	MKK	0.0325	0.0345	0.0283	0.0296	448255	449079	-0.0014	0.0011
YRI	TSI	0.0341	0.0371	0.0289	0.0302	446598	448135	-0.0071	0.0013

Table B.2. Table of confidence intervals for $\mathcal{L}$ and Q. Estimates of λ are given assuming generation length is 29 years and $2N_0$, the ancestral population size, is 4616. Q here assumes λ is 440 thousand years, matching our estimates from $\mathcal{L}$. The first set of rows uses the San genome from Meyer, Kircher, Gansauge, *et al.* [3] as the non-admixed human group, while the second set uses the HapMap YRI population.

		m_N		m_D		λ		Q	
X	Y	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%
San	ASW	0.0377	0.0402	0.0339	0.0356	446945	447865	-0.0119	-0.0073
San	CEU	0.0394	0.0433	0.0325	0.0347	445116	446424	-0.0159	-0.0058
San	CHB	0.0397	0.0437	0.0337	0.0357	445107	446666	-0.0162	-0.0025
San	CHD	0.0392	0.0439	0.0335	0.0354	445066	446516	-0.0167	-0.0042
San	GIH	0.0386	0.0420	0.0327	0.0348	445125	446682	-0.0180	-0.0050
San	JPT	0.0400	0.0438	0.0338	0.0357	445085	446592	-0.0165	-0.0032
San	LWK	0.0370	0.0393	0.0337	0.0356	447053	447850	-0.0122	-0.0086
San	MEX	0.0399	0.0435	0.0331	0.0349	445895	447292	-0.0121	-0.0006
San	MKK	0.0371	0.0389	0.0330	0.0348	446800	447572	-0.0126	-0.0077
San	TSI	0.0397	0.0429	0.0328	0.0351	445115	446665	-0.0167	-0.0047
San	YRI	0.0375	0.0395	0.0340	0.0359	447096	447889	-0.0128	-0.0085
YRI	ASW	0.0285	0.0306	0.0243	0.0255	448271	449084	-0.0007	0.0009
YRI	CEU	0.0303	0.0337	0.0250	0.0264	446391	448011	-0.0055	0.0016
YRI	CHB	0.0302	0.0339	0.0258	0.0272	446487	448246	-0.0049	0.0018
YRI	CHD	0.0298	0.0335	0.0257	0.0269	446455	448003	-0.0063	0.0007
YRI	GIH	0.0298	0.0326	0.0249	0.0264	446471	447936	-0.0057	-0.0001
YRI	JPT	0.0304	0.0337	0.0257	0.0271	446228	447819	-0.0062	0.0009
YRI	LWK	0.0282	0.0301	0.0243	0.0254	448458	449144	-0.0005	0.0005
YRI	MEX	0.0307	0.0339	0.0254	0.0266	447204	448572	-0.0038	0.0033
YRI	MKK	0.0289	0.0308	0.0247	0.0258	448229	449012	-0.0010	0.0007
YRI	TSI	0.0302	0.0334	0.0252	0.0267	446490	448106	-0.0050	0.0010

Table B.3. Old λ confidence intervals for $\mathcal{L}$ and Q. Estimates of λ are given in units of 2N generations. The first set of rows uses the San genome from Meyer et al. as the non-admixed human group, while the second set uses the HapMap YRI population.

		m_N		m_D		λ		Q	
X	Y	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%
San	Dai	0.0443	0.0494	0.0356	0.0380	448195	449864	-0.0019	0.0164
San	Dinka	0.0373	0.0398	0.0352	0.0377	447231	448906	-0.0213	-0.0091
San	French	0.0439	0.0472	0.0336	0.0363	447750	449115	-0.0064	0.0066
San	Han	0.0466	0.0524	0.0344	0.0371	448451	450153	0.0046	0.0240
San	Karitiana	0.0448	0.0481	0.0351	0.0381	447723	449529	-0.0045	0.0115
San	Mandenka	0.0380	0.0404	0.0372	0.0397	447731	449038	-0.0204	-0.0092
San	Mbuti	0.0385	0.0410	0.0364	0.0391	447948	449577	-0.0145	-0.0029
San	Papuan	0.0463	0.0512	0.0401	0.0444	449711	451786	0.0089	0.0282
San	Sardinia	0.0433	0.0473	0.0341	0.0363	447738	449259	-0.0078	0.0084
San	Yoruba	0.0386	0.0415	0.0361	0.0388	447523	449023	-0.0172	-0.0061
Yoruba	Dai	0.0367	0.0404	0.0275	0.0298	448820	450908	0.0082	0.0207
Yoruba	Dinka	0.0286	0.0310	0.0258	0.0283	448228	449739	-0.0052	0.0017
Yoruba	French	0.0350	0.0385	0.0256	0.0282	448279	450054	0.0031	0.0140
Yoruba	Han	0.0388	0.0432	0.0270	0.0290	449325	451071	0.0140	0.0262
Yoruba	Karitiana	0.0363	0.0403	0.0268	0.0294	448646	450579	0.0071	0.0193
Yoruba	Mandenka	0.0283	0.0307	0.0267	0.0293	448352	449797	-0.0046	0.0027
Yoruba	Mbuti	0.0368	0.0395	0.0340	0.0361	449568	451019	-0.0012	0.0076
Yoruba	Papuan	0.0382	0.0432	0.0318	0.0353	451051	453118	0.0165	0.0317
Yoruba	Sardinia	0.0347	0.0385	0.0259	0.0279	448456	449966	0.0037	0.0139

Table B.4. Young λ confidence intervals for $\mathcal{L}$ and Q. Estimates of λ are given in units of $2N_{\text{anc}}$ generations, where generation length is 29 years and $2N_{\text{anc}} = 4616$. The first set of rows uses the San genome as the non-admixed human group, while the second set uses the Yoruba genome.

		m_N		m_D		λ		Q	
X	Y	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%
San	Dai	0.0404	0.0457	0.0314	0.0339	448162	450054	-0.0013	0.0118
San	Dinka	0.0338	0.0359	0.0314	0.0333	447521	448865	-0.0140	-0.0077
San	French	0.0397	0.0435	0.0298	0.0328	447741	449336	-0.0046	0.0054
San	Han	0.0430	0.0481	0.0306	0.0331	448945	450487	0.0055	0.0165
San	Karitiana	0.0405	0.0445	0.0312	0.0336	448071	449733	-0.0026	0.0084
San	Mandenka	0.0344	0.0369	0.0332	0.0359	447858	449383	-0.0130	-0.0055
San	Mbuti	0.0351	0.0377	0.0324	0.0344	448269	449715	-0.0091	-0.0018
San	Papuan	0.0426	0.0470	0.0360	0.0396	450166	451995	0.0065	0.0196
San	Sardinia	0.0393	0.0430	0.0302	0.0323	447849	449354	-0.0054	0.0053
San	Yoruba	0.0349	0.0373	0.0320	0.0340	447631	449213	-0.0122	-0.0041
Yoruba	Dai	0.0333	0.0380	0.0242	0.0263	448917	450863	0.0063	0.0151
Yoruba	Dinka	0.0250	0.0278	0.0229	0.0249	447891	449465	-0.0041	0.0009
Yoruba	French	0.0315	0.0349	0.0224	0.0246	448260	449941	0.0028	0.0093
Yoruba	Han	0.0348	0.0394	0.0234	0.0257	449327	451063	0.0093	0.0181
Yoruba	Karitiana	0.0321	0.0366	0.0233	0.0260	448520	450919	0.0045	0.0131
Yoruba	Mandenka	0.0246	0.0272	0.0232	0.0252	448267	449600	-0.0033	0.0014
Yoruba	Mbuti	0.0336	0.0361	0.0301	0.0326	449783	451173	-0.0002	0.0057
Yoruba	Papuan	0.0347	0.0391	0.0281	0.0312	451645	453499	0.0127	0.0212
Yoruba	Sardinia	0.0313	0.0349	0.0222	0.0245	448305	449891	0.0026	0.0095

REFERENCES

[1] R. E. Green, J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H.-Y. Fritz, N. F. Hansen, E. Y. Durand, A.-S. Malaspinas, J. D. Jensen, T. Marques-Bonet, C. Alkan, K. Prüfer, M. Meyer, H. A. Burbano, J. M. Good, R. Schultz, A. Aximu-Petri, A. Butthof, B. Höber, B. Höffner, M. Siegemund, A. Weihmann, C. Nusbaum, E. S. Lander, C. Russ, N. Novod, J. Affourtit, M. Egholm, C. Verna, P. Rudan, D. Brajkovic, Ž. Kucan, I. Gušic, V. B. Doronichev, L. V. Golovanova, C. Lalueza-Fox, M. de la Rasilla, J. Fortea, A. Rosas, R. W. Schmitz, P. L. F. Johnson, E. E. Eichler, D. Falush, E. Birney, J. C. Mullikin, M. Slatkin, R. Nielsen, J. Kelso, M. Lachmann, D. Reich, and S. Pääbo, "A draft sequence of the neandertal genome," *Science*, vol. 328, no. 5979, pp. 710–722, May 7, 2010. DOI: 10.1126/science.1188021. [Online]. Available: http://www.sciencemag.org/content/328/5979/710.short (visited on 08/19/2015).

[2] D. Reich, R. E. Green, M. Kircher, J. Krause, N. Patterson, E. Y. Durand, B. Viola, A. W. Briggs, U. Stenzel, P. L. F. Johnson, T. Maricic, J. M. Good, T. Marques-Bonet, C. Alkan, Q. Fu, S. Mallick, H. Li, M. Meyer, E. E. Eichler, M. Stoneking, M. Richards, S. Talamo, M. V. Shunkov, A. P. Derevianko, J.-J. Hublin, J. Kelso, M. Slatkin, and S. Pääbo, "Genetic history of an archaic hominin group from denisova cave in siberia," *Nature*, vol. 468, no. 7327, pp. 1053–1060, Dec. 23, 2010, ISSN: 0028-0836. DOI: 10.1038/nature09710. [Online]. Available: http://www.nature.com/nature/journal/v468/n7327/full/nature09710.html (visited on 08/19/2015).

[3] M. Meyer, M. Kircher, M.-T. Gansauge, H. Li, F. Racimo, S. Mallick, J. G. Schraiber, F. Jay, K. Prüfer, C. d. Filippo, P. H. Sudmant, C. Alkan, Q. Fu, R. Do, N. Rohland, A. Tandon, M. Siebauer, R. E. Green, K. Bryc, A. W. Briggs, U. Stenzel, J. Dabney, J. Shendure, J. Kitzman, M. F. Hammer, M. V. Shunkov, A. P. Derevianko, N. Patterson, A. M. Andrés, E. E. Eichler, M. Slatkin, D. Reich, J. Kelso, and S. Pääbo, "A high-coverage genome sequence from an archaic denisovan individual," *Science*, vol. 338, no. 6104, pp. 222–226, Oct. 12, 2012, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1224344. [Online]. Available: http://www.sciencemag.org/content/338/6104/222 (visited on 08/19/2015).

[4] K. Prüfer, F. Racimo, N. Patterson, F. Jay, S. Sankararaman, S. Sawyer, A. Heinze, G. Renaud, P. H. Sudmant, C. de Filippo, H. Li, S. Mallick, M. Dannemann, Q. Fu, M. Kircher, M. Kuhlwilm, M. Lachmann, M. Meyer, M. Ongyerth, M. Siebauer, C. Theunert, A. Tandon, P. Moorjani, J. Pickrell, J. C. Mullikin, S. H. Vohr, R. E. Green, I. Hellmann, P. L. F. Johnson, H. Blanche, H. Cann, J. O. Kitzman, J. Shendure, E. E. Eichler, E. S. Lein, T. E. Bakken, L. V. Golovanova, V. B. Doronichev, M. V. Shunkov, A. P. Derevianko, B. Viola, M. Slatkin, D. Reich, J. Kelso, and S. Pääbo, "The complete genome sequence of a neanderthal from the altai mountains," *Nature*, vol. 505, no. 7481, pp. 43–49, Jan. 2, 2014, ISSN: 0028-0836. DOI: 10.1038/nature12886.

[Online]. Available: http://www.nature.com/nature/journal/v505/n7481/abs/nature12886.html (visited on 08/20/2015).

[5] Q. Fu, H. Li, P. Moorjani, F. Jay, S. M. Slepchenko, A. A. Bondarev, P. L. F. Johnson, A. Aximu-Petri, K. Prüfer, C. de Filippo, M. Meyer, N. Zwyns, D. C. Salazar-García, Y. V. Kuzmin, S. G. Keates, P. A. Kosintsev, D. I. Razhev, M. P. Richards, N. V. Peristov, M. Lachmann, K. Douka, T. F. G. Higham, M. Slatkin, J.-J. Hublin, D. Reich, J. Kelso, T. B. Viola, and S. Pääbo, "Genome sequence of a 45,000-year-old modern human from western siberia," *Nature*, vol. 514, no. 7523, pp. 445–449, Oct. 23, 2014, ISSN: 0028-0836. DOI: 10.1038/nature13810. [Online]. Available: http://www.nature.com/nature/journal/v514/n7523/full/nature13810.html (visited on 08/19/2015).

[6] C. B. Stringer and P. Andrews, "Genetic and fossil evidence for the origin of modern humans," *Science*, New Series, vol. 239, no. 4845, pp. 1263–1268, Mar. 11, 1988, ISSN: 0036-8075. [Online]. Available: http://www.jstor.org/stable/1700885 (visited on 08/19/2015).

[7] M. H. Wolpoff, "Multiregional evolution: The fossil alternative to eden," in *The human revolution: Behavioural and biological perspectives on the origins of modern humans*, Princeton, NJ: Princeton University Press, 1989, pp. 62–108.

[8] M. H. Wolpoff, J. Hawks, and R. Caspari, "Multiregional, not multiple origins," *American Journal of Physical Anthropology*, vol. 112, no. 1, pp. 129–136, May 1, 2000, ISSN: 1096-8644. DOI: 10.1002/(SICI)1096-8644(200005)112:1<129::AID-AJPA11>3.0.CO;2-K. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1096-8644(200005)112:1%3C129::AID-AJPA11%3E3.0.CO;2-K/abstract (visited on 08/20/2015).

[9] E. Huerta-Sánchez, X. Jin, Asan, Z. Bianba, B. M. Peter, N. Vinckenbosch, Y. Liang, X. Yi, M. He, M. Somel, P. Ni, B. Wang, X. Ou, Huasang, J. Luosang, Z. X. P. Cuo, K. Li, G. Gao, Y. Yin, W. Wang, X. Zhang, X. Xu, H. Yang, Y. Li, J. Wang, J. Wang, and R. Nielsen, "Altitude adaptation in tibetans caused by introgression of denisovan-like DNA," *Nature*, vol. 512, no. 7513, pp. 194–197, Aug. 14, 2014, ISSN: 0028-0836. DOI: 10.1038/nature13408. [Online]. Available: http://www.nature.com/nature/journal/v512/n7513/full/nature13408.html (visited on 08/19/2015).

[10] B. Vernot and J. M. Akey, "Resurrecting surviving neandertal lineages from modern human genomes," *Science*, vol. 343, no. 6174, pp. 1017–1021, Feb. 28, 2014, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1245938. [Online]. Available: http://www.sciencemag.org/content/343/6174/1017 (visited on 08/20/2015).

[11] S. Sankararaman, S. Mallick, M. Dannemann, K. Prüfer, J. Kelso, S. Pääbo, N. Patterson, and D. Reich, "The genomic landscape of neanderthal ancestry in present-day humans," *Nature*, vol. 507, no. 7492, pp. 354–357, Mar. 20, 2014, ISSN: 0028-0836. DOI: 10.1038/nature12961. [Online]. Available: http://www.nature.com/nature/journal/v507/n7492/abs/nature12961.html (visited on 08/20/2015).

[12] Y. Hu, Q. Ding, Y. Wang, S. Xu, Y. He, M. Wang, J. Wang, and L. Jin, "Investigating the evolutionary importance of denisovan introgressions in papua new guineans and australians," *BioRxiv*, p. 022 632, Jul. 15, 2015. DOI: 10.1101/022632. [Online]. Available: http://www.biorxiv.org/content/early/2015/07/15/022632 (visited on 08/19/2015).

[13] L. Abi-Rached, M. J. Jobin, S. Kulkarni, A. McWhinnie, K. Dalva, L. Gragert, F. Babrzadeh, B. Gharizadeh, M. Luo, F. A. Plummer, J. Kimani, M. Carrington, D. Middleton, R. Rajalingam, M. Beksac, S. G. E. Marsh, M. Maiers, L. A. Guethlein, S. Tavoularis, A.-M. Little, R. E. Green, P. J. Norman, and P. Parham, "The shaping of modern human immune systems by multiregional admixture with archaic humans," *Science*, vol. 334, no. 6052, pp. 89–94, Oct. 7, 2011, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1209202. [Online]. Available: http://www.sciencemag.org/content/334/6052/89 (visited on 08/20/2015).

[14] Q. Fu, M. Hajdinjak, O. T. Moldovan, S. Constantin, S. Mallick, P. Skoglund, N. Patterson, N. Rohland, I. Lazaridis, B. Nickel, B. Viola, K. Prüfer, M. Meyer, J. Kelso, D. Reich, and S. Pääbo, "An early modern human from romania with a recent neanderthal ancestor," *Nature*, vol. 524, no. 7564, pp. 216–219, Aug. 13, 2015, ISSN: 0028-0836. DOI: 10.1038/nature14558. [Online]. Available: http://www.nature.com/nature/journal/v524/n7564/full/nature14558.html (visited on 08/19/2015).

[15] N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich, "Ancient admixture in human history," *Genetics*, vol. 192, no. 3, pp. 1065–1093, Nov. 2012, ISSN: 0016-6731. DOI: 10.1534/genetics.112.145037. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3522152/ (visited on 08/19/2015).

[16] P. Beerli and J. Felsenstein, "Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach," *Proceedings of the National Academy of Sciences*, vol. 98, no. 8, pp. 4563–4568, Apr. 10, 2001, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.081068098. [Online]. Available: http://www.pnas.org/content/98/8/4563 (visited on 08/20/2015).

[17] P. Beerli, "Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations," *Molecular Ecology*, vol. 13, no. 4, pp. 827–836, Apr. 1, 2004, ISSN: 1365-294X. DOI: 10.1111/j.1365-294X.2004.02101.x. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1365-294X.2004.02101.x/abstract (visited on 08/19/2015).

[18] A. R. Rogers and R. J. Bohlender, "Bias in estimators of archaic admixture," *Theoretical Population Biology*, vol. 100, pp. 63–78, Mar. 2015, ISSN: 0040-5809. DOI: 10.1016/j.tpb.2014.12.006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0040580914001087 (visited on 08/19/2015).

[19] P. Skoglund and M. Jakobsson, "Archaic human ancestry in east asia," *Proceedings of the National Academy of Sciences*, vol. 108, no. 45, pp. 18 301–18 306, Nov. 8, 2011, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1108181108. [Online]. Available: http://www.pnas.org/content/108/45/18301 (visited on 08/19/2015).

[20] D. Reich, K. Thangaraj, N. Patterson, A. L. Price, and L. Singh, "Reconstructing indian population history," *Nature*, vol. 461, no. 7263, pp. 489–494, Sep. 24, 2009, ISSN: 0028-0836. DOI: 10.1038/nature08365. [Online]. Available: http://www.nature.com/nature/journal/v461/n7263/full/nature08365.html (visited on 08/20/2015).

[21] A. Eriksson and A. Manica, "Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins,"

Proceedings of the National Academy of Sciences, vol. 109, no. 35, pp. 13 956–13 960, Aug. 28, 2012, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1200567109. [Online]. Available: http://www.pnas.org/content/109/35/13956 (visited on 08/19/2015).

[22] ——, "The doubly conditioned frequency spectrum does not distinguish between ancient population structure and hybridization," *Molecular Biology and Evolution*, vol. 31, no. 6, pp. 1618–1621, Jun. 1, 2014, ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msu103. [Online]. Available: http://mbe.oxfordjournals.org/content/31/6/1618 (visited on 08/20/2015).

[23] S. Sankararaman, N. Patterson, H. Li, S. Pääbo, and D. Reich, "The date of interbreeding between neandertals and modern humans," *PLOS Genet*, vol. 8, no. 10, e1002947, Oct. 4, 2012. DOI: 10.1371/journal.pgen.1002947. [Online]. Available: http://dx.doi.org/10.1371/journal.pgen.1002947 (visited on 08/20/2015).

[24] L. Excoffier, I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa, and M. Foll, "Robust demographic inference from genomic and SNP data," *PLOS Genet*, vol. 9, no. 10, e1003905, Oct. 24, 2013. DOI: 10.1371/journal.pgen.1003905. [Online]. Available: http://dx.doi.org/10.1371/journal.pgen.1003905 (visited on 08/20/2015).

[25] L. Excoffier and M. Foll, "Fastsimcoal: A continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios," *Bioinformatics*, vol. 27, no. 9, pp. 1332–1334, May 1, 2011, ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btr124. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/27/9/1332 (visited on 08/19/2015).

[26] J. N. Fenner, "Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies," *American Journal of Physical Anthropology*, vol. 128, no. 2, pp. 415–423, Oct. 1, 2005, ISSN: 1096-8644. DOI: 10.1002/ajpa.20188. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/ajpa.20188/abstract (visited on 08/19/2015).

[27] A. R. Skinner, B. A. B. Blackwell, S. Martin, A. Ortega, J. I. B. Blickstein, L. V. Golovanova, and V. B. Doronichev, "ESR dating at mezmaiskaya cave, russia," *Applied Radiation and Isotopes*, Proceedings of the 6th International Symposium on ESR Dosimetry and Applications, vol. 62, no. 2, pp. 219–224, Feb. 2005, ISSN: 0969-8043. DOI: 10.1016/j.apradiso.2004.08.008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0969804304004555 (visited on 08/19/2015).

[28] B. M. Henn, L. L. Cavalli-Sforza, and M. W. Feldman, "The great human expansion," *Proceedings of the National Academy of Sciences*, vol. 109, no. 44, pp. 17 758–17 764, Oct. 30, 2012, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1212380109. [Online]. Available: http://www.pnas.org/content/109/44/17758 (visited on 08/20/2015).

[29] K. R. Veeramah and M. F. Hammer, "The impact of whole-genome sequencing on the reconstruction of human population history," *Nature Reviews Genetics*, vol. 15, no. 3, pp. 149–162, Mar. 2014, ISSN: 1471-0056. DOI: 10.1038/nrg3625. [Online]. Available: http://www.nature.com/nrg/journal/v15/n3/abs/nrg3625.html (visited on 08/20/2015).

[30] D. J. Wales and J. P. K. Doye, "Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms," *The Journal*

of Physical Chemistry A, vol. 101, no. 28, pp. 5111–5116, Jul. 1, 1997, ISSN: 1089-5639. DOI: 10.1021/jp970984n. [Online]. Available: http://dx.doi.org/10.1021/jp970984n (visited on 08/19/2015).

[31] S. G. Nash, "Newton-type minimization via the lanczos method," *SIAM Journal on Numerical Analysis*, vol. 21, no. 4, pp. 770–788, Aug. 1, 1984, ISSN: 0036-1429. [Online]. Available: http://www.jstor.org/stable/2157008 (visited on 08/20/2015).

[32] E. Jones, T. Oliphant, P. Peterson, *et al.*, "Scipy: Open source scientific tools for python," 2001.

[33] S. v. d. Walt, S. C. Colbert, and G. Varoquaux, "The NumPy array: A structure for efficient numerical computation," *Computing in Science & Engineering*, vol. 13, no. 2, pp. 22–30, Mar. 1, 2011, ISSN: 1521-9615. DOI: 10.1109/MCSE.2011.37. [Online]. Available: http://scitation.aip.org/content/aip/journal/cise/13/2/10.1109/MCSE.2011.37 (visited on 08/20/2015).

[34] J. D. Wall and M. F. Hammer, "Archaic admixture in the human genome," *Current Opinion in Genetics & Development*, Genomes and evolution, vol. 16, no. 6, pp. 606–610, Dec. 2006, ISSN: 0959-437X. DOI: 10.1016/j.gde.2006.09.006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0959437X06001997 (visited on 08/19/2015).

[35] J. D. Wall, K. E. Lohmueller, and V. Plagnol, "Detecting ancient admixture and estimating demographic parameters in multiple human populations," *Molecular Biology and Evolution*, vol. 26, no. 8, pp. 1823–1827, Aug. 1, 2009, ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msp096. [Online]. Available: http://mbe.oxfordjournals.org/content/26/8/1823 (visited on 08/19/2015).

[36] H. Tang, M. Coram, P. Wang, X. Zhu, and N. Risch, "Reconstructing genetic ancestry blocks in admixed individuals," *American Journal of Human Genetics*, vol. 79, no. 1, pp. 1–12, Jun. 2006. DOI: 10.1086/504302.

[37] A. M. Andrés, A. G. Clark, L. Shimmin, E. Boerwinkle, C. F. Sing, and J. E. Hixson, "Understanding the accuracy of statistical haplotype inference with sequence data of known phase," *Genetic Epidemiology*, vol. 31, no. 7, pp. 659–671, Oct. 5, 2007. (visited on 09/02/2015).

[38] D. Reich, N. Patterson, D. Campbell, A. Tandon, S. Mazieres, N. Ray, M. V. Parra, W. Rojas, C. Duque, N. Mesa, L. F. Garcia, O. Triana, S. Blair, A. Maestre, J. C. Dib, C. M. Bravi, G. Bailliet, D. Corach, T. Hunemeier, M. C. Bortolini, F. M. Salzano, M. L. Petzl-Erler, V. Acuna-Alonzo, C. Aguilar-Salinas, S. Canizales-Quinteros, T. Tusie-Luna, L. Riba, M. Rodriguez-Cruz, M. Lopez-Alarcon, R. Coral-Vazquez, T. Canto-Cetina, I. Silva-Zolezzi, J. C. Fernandez-Lopez, A. V. Contreras, G. Jimenez-Sanchez, M. J. Gomez-Vazquez, J. Molina, A. Carracedo, A. Salas, C. Gallo, G. Poletti, D. B. Witonsky, G. Alkorta-Aranburu, R. I. Sukernik, L. Osipova, S. A. Fedorova, R. Vasquez, M. Villena, C. Moreau, R. Barrantes, D. Pauls, L. Excoffier, G. Bedoya, F. Rothhammer, J.-M. Dugoujon, G. Larrouy, W. Klitz, D. Labuda, J. Kidd, K. Kidd, A. Di Rienzo, N. B. Freimer, A. L. Price, and A. Ruiz-Linares, "Reconstructing native american population history," *Nature*, vol. 488, no. 7411, pp. 370–374, Aug. 16, 2012, ISSN: 0028-0836. DOI: 10.1038/nature11258. [Online]. Available: http://dx.doi.org/10.1038/nature11258.

[39] I. Lazaridis, N. Patterson, A. Mittnik, G. Renaud, S. Mallick, K. Kirsanow, P. H. Sudmant, J. G. Schraiber, S. Castellano, M. Lipson, B. Berger, C. Economou, R. Bollongino, Q. Fu, K. I. Bos, S. Nordenfelt, H. Li, C. de Filippo, K. Prüfer, S. Sawyer, C. Posth, W. Haak, F. Hallgren, E. Fornander, N. Rohland, D. Delsate, M. Francken, J.-M. Guinet, J. Wahl, G. Ayodo, H. A. Babiker, G. Bailliet, E. Balanovska, O. Balanovsky, R. Barrantes, G. Bedoya, H. Ben-Ami, J. Bene, F. Berrada, C. M. Bravi, F. Brisighelli, G. B. J. Busby, F. Cali, M. Churnosov, D. E. C. Cole, D. Corach, L. Damba, G. van Driem, S. Dryomov, J.-M. Dugoujon, S. A. Fedorova, I. Gallego Romero, M. Gubina, M. Hammer, B. M. Henn, T. Hervig, U. Hodoglugil, A. R. Jha, S. Karachanak-Yankova, R. Khusainova, E. Khusnutdinova, R. Kittles, T. Kivisild, W. Klitz, V. Kučinskas, A. Kushniarevich, L. Laredj, S. Litvinov, T. Loukidis, R. W. Mahley, B. Melegh, E. Metspalu, J. Molina, J. Mountain, K. Näkkäläjärvi, D. Nesheva, T. Nyambo, L. Osipova, J. Parik, F. Platonov, O. Posukh, V. Romano, F. Rothhammer, I. Rudan, R. Ruizbakiev, H. Sahakyan, A. Sajantila, A. Salas, E. B. Starikovskaya, A. Tarekegn, D. Toncheva, S. Turdikulova, I. Uktveryte, O. Utevska, R. Vasquez, M. Villena, M. Voevoda, C. A. Winkler, L. Yepiskoposyan, P. Zalloua, T. Zemunik, A. Cooper, C. Capelli, M. G. Thomas, A. Ruiz-Linares, S. A. Tishkoff, L. Singh, K. Thangaraj, R. Villems, D. Comas, R. Sukernik, M. Metspalu, M. Meyer, E. E. Eichler, J. Burger, M. Slatkin, S. Pääbo, J. Kelso, D. Reich, and J. Krause, "Ancient human genomes suggest three ancestral populations for present-day europeans," *Nature*, vol. 513, no. 7518, pp. 409–413, Sep. 18, 2014, ISSN: 0028-0836. DOI: 10.1038/nature13673. [Online]. Available: http://www.nature.com/nature/journal/v513/n7518/full/nature13673.html (visited on 08/19/2015).

[40] P. Qin and M. Stoneking, "Denisovan ancestry in east eurasian and native american populations," *Molecular Biology and Evolution*, msv141, Jun. 23, 2015, ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msv141. [Online]. Available: http://mbe.oxfordjournals.org/content/early/2015/07/08/molbev.msv141 (visited on 08/19/2015).

[41] A. Keinan, J. C. Mullikin, N. Patterson, and D. Reich, "Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans," *Nature Genetics*, vol. 39, no. 10, pp. 1251–1255, 2007. DOI: 10.1038/ng2116.

[42] R. A. Gibbs, J. W. Belmont, P. Hardenbol, T. D. Willis, F. Yu, H. Yang, L.-Y. Ch'ang, W. Huang, B. Liu, Y. Shen, P. K.-H. Tam, L.-C. Tsui, M. M. Y. Waye, J. T.-F. Wong, C. Zeng, Q. Zhang, M. S. Chee, L. M. Galver, S. Kruglyak, S. S. Murray, A. R. Oliphant, A. Montpetit, T. J. Hudson, F. Chagnon, V. Ferretti, M. Leboeuf, M. S. Phillips, A. Verner, P.-Y. Kwok, S. Duan, D. L. Lind, R. D. Miller, J. P. Rice, N. L. Saccone, P. Taillon-Miller, M. Xiao, Y. Nakamura, A. Sekine, K. Sorimachi, T. Tanaka, Y. Tanaka, T. Tsunoda, E. Yoshino, D. R. Bentley, P. Deloukas, S. Hunt, D. Powell, D. Altshuler, S. B. Gabriel, H. Zhang, I. Matsuda, Y. Fukushima, D. R. Macer, E. Suda, C. N. Rotimi, C. A. Adebamowo, T. Aniagwu, P. A. Marshall, O. Matthew, C. Nkwodimmah, C. D. M. Royal, M. F. Leppert, M. Dixon, L. D. Stein, F. Cunningham, A. Kanani, G. A. Thorisson, A. Chakravarti, P. E. Chen, D. J. Cutler, C. S. Kashuk, P. Donnelly, J. Marchini, G. A. T. McVean, S. R. Myers, L. R. Cardon, G. R. Abecasis, A. Morris, B. S. Weir, J. C. Mullikin, S. T. Sherry, M. Feolo, M. J. Daly, S. F. Schaffner, R. Qiu, A. Kent, G. M. Dunston, K. Kato,

N. Niikawa, B. M. Knoppers, M. W. Foster, E. W. Clayton, V. O. Wang, J. Watkin, E. Sodergren, G. M. Weinstock, R. K. Wilson, L. L. Fulton, J. Rogers, B. W. Birren, H. Han, H. Wang, M. Godbout, J. C. Wallenburg, P. L'Archevêque, G. Bellemare, K. Todani, T. Fujita, S. Tanaka, A. L. Holden, E. H. Lai, F. S. Collins, L. D. Brooks, J. E. McEwen, M. S. Guyer, E. Jordan, J. L. Peterson, J. Spiegel, L. M. Sung, L. F. Zacharia, K. Kennedy, M. G. Dunn, R. Seabrook, M. Shillito, B. Skene, J. G. Stewart, D. L. V. (chair), E. W. C. (co-Chair), L. B. J. (co-Chair), M. K. Cho, T. Duster, M. Jasperse, J. Licinio, J. C. Long, P. N. Ossorio, P. Spallone, S. F. Terry, E. S. L. (chair), E. H. L. (co-Chair), D. A. N. (co-Chair), M. Boehnke, J. A. Douglas, R. R. Hudson, L. Kruglyak, and R. L. Nussbaum, "The international HapMap project," *Nature*, vol. 426, no. 6968, pp. 789–796, Dec. 18, 2003, ISSN: 0028-0836. DOI: 10.1038/nature02168. [Online]. Available: http://www.nature.com/nature/journal/v426/n6968/abs/nature02168.html (visited on 08/19/2015).

[43] R. J. Bohlender and A. R. Rogers, "Estimating archaic admixture and the modern-archaic split," *In Prep*, 2015.

[44] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, and R. Durbin, "The variant call format and VCFtools," *Bioinformatics*, vol. 27, no. 15, pp. 2156–2158, Aug. 1, 2011, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr330. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3137218/ (visited on 08/22/2015).

[45] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo, "The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data," *Genome Research*, vol. 20, no. 9, pp. 1297–1303, Sep. 2010, ISSN: 1088-9051. DOI: 10.1101/gr.107524.110. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2928508/ (visited on 08/22/2015).

[46] H. R. Kunsch, "The jackknife and the bootstrap for general stationary observations," *The Annals of Statistics*, vol. 17, no. 3, pp. 1217–1241, Sep. 1, 1989, ISSN: 0090-5364. [Online]. Available: http://www.jstor.org/stable/2241719 (visited on 08/20/2015).

[47] H. Li and R. Durbin, "Inference of human population history from individual whole-genome sequences," *Nature*, vol. 475, no. 7357, pp. 493–496, Jul. 28, 2011, ISSN: 0028-0836. DOI: 10.1038/nature10231. [Online]. Available: http://www.nature.com/nature/journal/v475/n7357/full/nature10231.html (visited on 08/19/2015).

[48] P. Qin, Y. Zhou, H. Lou, D. Lu, X. Yang, Y. Wang, L. Jin, Y.-J. Chung, and S. Xu, "Quantitating and dating recent gene flow between european and east asian populations," *Scientific Reports*, vol. 5, Apr. 2, 2015, ISSN: 2045-2322. DOI: 10.1038/srep09500. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4382708/ (visited on 08/20/2015).

[49] J. D. Wall, M. A. Yang, F. Jay, S. K. Kim, E. Y. Durand, L. S. Stevison, C. Gignoux, A. Woerner, M. F. Hammer, and M. Slatkin, "Higher levels of neanderthal ancestry in east asians than in europeans," *Genetics*, vol. 194, no. 1, pp. 199–209, May 1, 2013, ISSN: 0016-6731, 1943-2631. DOI: 10.1534/genetics.112.148213. [Online]. Available: http://www.genetics.org/content/194/1/199 (visited on 08/19/2015).

[50] R. J. Bohlender and A. R. Rogers, "Ascertainment bias in estimators of archaic admixture," *In Prep*, 2015.

[51] F. Gomez, J. Hirbo, and S. A. Tishkoff, "Genetic variation and adaptation in africa: Implications for human evolution and disease," *Cold Spring Harbor Perspectives in Biology*, vol. 6, no. 7, a008524, Jul. 1, 2014, ISSN: , 1943-0264. DOI: 10.1101/cshperspect.a008524. [Online]. Available: http://cshperspectives.cshlp.org/content/6/7/a008524 (visited on 08/19/2015).

[52] M. C. Campbell, J. B. Hirbo, J. P. Townsend, and S. A. Tishkoff, "The peopling of the african continent and the diaspora into the new world," *Current Opinion in Genetics & Development*, Genetics of human evolution, vol. 29, pp. 120–132, Dec. 2014, ISSN: 0959-437X. DOI: 10.1016/j.gde.2014.09.003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0959437X14000987 (visited on 08/20/2015).

[53] S. Wang, J. Lachance, S. A. Tishkoff, J. Hey, and J. Xing, "Apparent variation in neanderthal admixture among african populations is consistent with gene flow from non-african populations," *Genome Biology and Evolution*, vol. 5, no. 11, pp. 2075–2081, Jan. 1, 2013, ISSN: , 1759-6653. DOI: 10.1093/gbe/evt160. [Online]. Available: http://gbe.oxfordjournals.org/content/5/11/2075 (visited on 08/19/2015).

[54] D. Reich, N. Patterson, M. Kircher, F. Delfin, M. R. Nandineni, I. Pugach, A. M.-S. Ko, Y.-C. Ko, T. A. Jinam, M. E. Phipps, N. Saitou, A. Wollstein, M. Kayser, S. Pääbo, and M. Stoneking, "Denisova admixture and the first modern human dispersals into southeast asia and oceania," *The American Journal of Human Genetics*, vol. 89, no. 4, pp. 516–528, Oct. 7, 2011, ISSN: 0002-9297. DOI: 10.1016/j.ajhg.2011.09.005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0002929711003958 (visited on 08/19/2015).

[55] K. Lohse and L. A. F. Frantz, "Neandertal admixture in eurasia confirmed by maximum-likelihood analysis of three genomes," *Genetics*, vol. 196, no. 4, pp. 1241–1251, Apr. 1, 2014.

[56] D. A. R. Eaton and R. H. Ree, "Inferring phylogeny and introgression using RADseq data: An example from flowering plants (Pedicularis: Orobanchaceae)," *Systematic Biology*, vol. 62, no. 5, pp. 689–706, 2013. DOI: 10.1093/sysbio/syt032.

[57] D. A. R. Eaton, A. L. Hipp, A. González-Rodríguez, and J. Cavender-Bares, "Historical introgression among the american live oaks and the comparative nature of tests for introgression," *Evolution*, Aug. 24, 2015. DOI: 10.1111/evo.12758.

[58] I. Pugach, R. Matveyev, A. Wollstein, M. Kayser, and M. Stoneking, "Dating the age of admixture via wavelet transform analysis of genome-wide data," *Genome Biology*, vol. 12, R19, Feb. 25, 2011. DOI: 10.1186/gb-2011-12-2-r19. [Online]. Available: http://genomebiology.com/content/12/2/R19 (visited on 09/02/2015).

[59] S. Wooding and A. Rogers, "The matrix coalescent and an application to human single-nucleotide polymorphisms," *Genetics*, vol. 161, pp. 1641–1650, Aug. 2002.

[60] R. Griffiths and S. Tavaré, "The age of a mutation in a general coalescent tree," *Communications in Statistics. Stochastic Models*, vol. 14, no. 1-2, pp. 273–295, 1998. DOI: 10.1080/15326349808807471. eprint: http://dx.doi.org/10.1080/15326349808807471. [Online]. Available: http://dx.doi.org/10.1080/15326349808807471.

[61] E. Y. Durand, N. Patterson, D. Reich, and M. Slatkin, "Testing for ancient admixture between closely related populations," *Molecular Biology and Evolution*, vol. 28, no. 8, pp. 2239–2252, Aug. 1, 2011, ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msr048. [Online]. Available: http://mbe.oxfordjournals.org/content/28/8/2239 (visited on 08/20/2015).