Title
Inferring evolutionary history from whole genomes: outlier tests and methods based on the coalescent with recombination

Author
Yoshihara Caldeira Brandt, Débora

Publication Date
2022

Peer reviewed|Thesis/dissertation

Inferring evolutionary history from whole genomes: outlier tests and methods based on the coalescent with recombination

by

Débora Yoshihara Caldeira Brandt

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Integrative Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Rasmus Nielsen, Chair
Professor Doris Bachtrog
Professor Yun S. Song

Summer 2022

Inferring evolutionary history from whole genomes:
outlier tests and methods based on the coalescent with recombination

Abstract

Inferring evolutionary history from whole genomes:
outlier tests and methods based on the coalescent with recombination

by

Débora Yoshihara Caldeira Brandt

Doctor of Philosophy in Integrative Biology

University of California, Berkeley

Professor Rasmus Nielsen, Chair

Whole genome sequences allow for new, powerful inferences of evolutionary history. In addition to increasing the scale of application of traditional population genetic inferences by using them on an enormous amount of loci, genomic data also provide valuable information about the location of those sites and correlations among them, which can be leveraged for inference of evolutionary parameters. In this dissertation, I explore three types of evolutionary inferences that are thriving with the increasing availability of genomic data and new methods. In Chapter 1, I use simulations to evaluate methods for inference of ancestral recombination graphs. In Chapter 2, I apply a new method based on the pairwise sequential Markovian coalescent to infer population split times and migration rates from a pair of diploid genomes. In Chapter 3, I use population-level genomic data from the population of a rural village in Ecuador to infer signatures of selection possibly related to the beneficial effects of diet on their cardiovascular health.

To Daniel

for the companionship in the best adventures of my life,
including this one.

Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I thank my advisor Rasmus Nielsen for the constant support in all the work for this dissertation and for opening so many doors to other stimulating research collaborations and teaching opportunities.

I thank my committee members Yun Song, Doris Bachtrog, and especially Jeff Wall for supporting me in the beginning of this Berkeley journey. I also thank Monty Slatkin for hosting me in his lab in my first year here, and convincing me that I should learn Python.

I thank Monica Albe for the efficient and warmhearted administrative support.

I thank Işın Altınkaya and Thorfinn Korneliussen for all the help with ANGSD.

I thank my lab mates, past and present, for all that you taught me, for the friendship, for the fun moments in the lab and outside. I am so lucky to have been surrounded by such a smart, kind and fun bunch of people over these years!

I am particularly grateful to Sandra Hui, server wizard and American politics tour guide, and Lenore Pipes, command line wizard and running inspiration, who shared the most time with me in the lab. Extra thanks to Sandra for the all the check-ins and exchange of tips on dissertation formatting in the final weeks. It's been great to sprint to the finish line beside you! I thank Diana Aguilar-Gómez and Joana Rocha for the NGSocorro bioinformatics and emotional support group. I am also so proud and grateful to have collaborated with Diana Aguilar-Gómez, Emma Steigerwald, Yun Deng, Andrew Vaughn, Vladimir Shchur and Hongru Wang on such interesting and diverse projects!

The amazing friends I made in the Bay Area made these years even more joyful and fulfilling. I am deeply grateful for the Brazilian friends I met so far from home and for the friendships from many different places that I am taking with me to wherever the next home is. Very special thanks to Ixchel, Diana and Mauri for the regular check-ins, for feeding us from time to time and for the amazing trips we took together. Extra thanks to Ixchel for putting up with the final hours of dissertation writing.

I also thank my Brazilian-biologists-abroad friends pelas trocas e confidências da vida de cientistas brasileiros pelo mundo. E claro, aos amigos do Brasil que mesmo longe, continuam próximos. Aquelas conversas de vez em nunca foram essenciais pra aguentar a distância, e mais importantes do que talvez eu seja capaz de demonstrar.

My family who managed to support and encourage from afar, always with such gentle but powerful and uplifting support. Agredeço aos meus pais, Célia e Fernando e minha irmã Beatriz pelo carinho, apoio e torcida. Vocês são meus maiores exemplos e fonte inesgotável de inspiração e motivação.

Daniel for being the best partner in everything I do. Obrigada por recarregar minhas baterias nos melhores e piores momentos. Por cuidar tão bem de mim e me deixar cuidar de você. Não teria sido possível sem você ao meu lado.

Obrigada!

Introduction

The use of whole genome sequences opens new possibilities for inference of evolutionary history that were not available with traditional population genetics methods developed mostly for single loci. For decades, the field of population genetics developed without the possibility of obtaining any DNA sequence data (Pool et al., 2010), and during this time, obtaining genome-wide population data as easily as it is today was unimaginable. The possibility of extending population genetics analyses to whole genomes expanded the field in many directions, and I address three of them in the chapters of this dissertation. First, the availability of data from continuous regions of the genome allows for the explicit modeling of correlations between sites, through the inclusion of recombination as a parameter of these models. Second, genome-wide data allows inference of population demographic history from a single diploid individual. Third, sampling genome-wide data from several individuals in a population allows the use of genome-wide distributions of statistics as null distributions, from which outliers can be pinpointed as candidates of selection.

1. Explicit modeling of correlations between sites with ancestral recombination graphs

An important advance brought about by population genomics is tightly linked to an extension of the standard coalescent model: the coalescent with recombination (Hudson, 1983). The data structure that represents the outcome of this stochastic process is called an ancestral recombination graph (ARG). Recent computational methods allow powerful and scalable inference based on the coalescent with recombination. These methods can leverage all the information about linked sites along a genome for inferences about recombination rate and its variation along the genome, as well as understanding how recombination rate varies between populations or species and through time. Explicitly taking recombination into account also informs the models about correlations between sites, which itself improves further inferences of population genetic parameters.

In Chapter 1, I, along with collaborators, use standard neutral coalescent simulations to benchmark the estimates of pairwise coalescence times from three popular ARG inference programs: ARGweaver, Relate, and tsinfer+tsdate. In addition to inferring the ARG, some of these methods can also provide ARGs sampled from a defined posterior distribution.

Obtaining good samples of ARGs is crucial for quantifying statistical uncertainty and for estimating population genetic parameters such as effective population size, mutation rate, and allele age. We compare 1) the true coalescence times to the inferred times at each locus; 2) the distribution of coalescence times across all loci to the expected exponential distribution; 3) whether the sampled coalescence times have the properties expected of a valid posterior distribution. We find that inferred coalescence times at each locus are most accurate in ARGweaver, and often more accurate in Relate than in tsinfer+tsdate. However, all three methods tend to overestimate small coalescence times and underestimate large ones. Lastly, the posterior distribution of ARGweaver is closer to the expected posterior distribution than Relate's, but this higher accuracy comes at a substantial trade-off in scalability. The best choice of method will depend on the number and length of input sequences and on the goal of downstream analyses, and we provide guidelines for the best practices.

2. Inference of population history from a pair of individuals

A second advance of population genomics is the ability to infer evolutionary parameters of a population from a single genome. These methods take advantage of the fact that a single diploid genome contains fragments of many ancestors. In Chapter 2, I leverage PSMC, a method that is able to infer past effective population sizes (N_e) from a single genome, to infer split times and migration rates from a pair of diploid genomes with ancestry from two populations.

In this chapter, I, along with collaborators, also explore the various definitions and applications of the concept of effective population size. The estimation of effective population sizes through time with methods such as PSMC became widespread in genetics, and this method has been applied in several organisms. Effective population size is a fundamental parameter in population genetics, but the interpretation of N_e as the effective number of breeding individuals in the population is challenged by the effect of population structure. In fact, variation in N_e reported in many studies may be a consequence of changes in migration rates between populations rather than changes in actual population size. We address this long-standing problem here by constructing joint models of population size changes, migration, and divergence that can adjust temporal estimates of N_e and estimate the actual N_e of a local deme connected to another population through migration.

We also develop a method for estimating divergence times and migration rates taking into account complex scenarios of changing population sizes. We apply the method to previously published data from humans, and show that, when taking migration and changes in N_e into account, the estimated divergence between the San and Dinka populations is approximately 108 kya, and not 255 kya as reported in a previous study. Using simulations, we demonstrate that the previously reported and surprisingly old estimates of divergence between San and Dinka is in fact caused by a quantifiable estimation bias due to changes in N_e through time.

3. Inference of selection with outlier-based neutrality test

Finally, I highlight what was perhaps the first advance of population genomics to be widely applied: outlier tests to detect natural selection. This type of neutrality test was first proposed by Lewontin and Krakauer (1973), but with the increasing availability of genome-wide sequences, it has been more widely applied. The rationale is simple: the effect of genetic drift on the distributions of alleles frequencies is expected to be the same across all loci. What Lewontin and Krakauer (1973) called "heterogeneity between loci in their inbreeding coefficients" could then be interpreted as evidence for selection. Since then, this rationale has been applied to statistics other than the original inbreeding coefficients, to detect selection.

In Chapter 3, I investigate the genetic ancestry and evidence of natural selection on a rural population from the coastal region of Ecuador (Atahualpa village). For that, I use an outlier scan of selection using the population branch statistic (PBS), a measure of population differentiation of one population relative to two other closely related populations. This study was motivated by a prior hypothesis of natural selection in response to a diet rich in oily fish in the population from Atahualpa. The consumption of oily fish among people from the Atahualpa village has been associated with several positive effects on their cardiovascular health. Indeed, we find evidence of selection in the Atahualpa population on genes that are related to the metabolism of lipids. Therefore, these gene variants may mediate the benefits of fish consumption in this population.

Chapter 1

Evaluation of methods for estimating coalescence times using ancestral recombination graphs

This chapter is co-authored by Xinzhu Wei, Yun Deng, Andrew H. Vaughn and Rasmus Nielsen and has been published in Genetics, Volume 221, Issue 1, May 2022, iyac044, https://doi.org/10.1093/genetics/iyac044

1.1 Introduction

The full ARG is a structure that encodes all coalescence and recombination events resulting from the stochastic process of the coalescent with recombination. Hudson (1983) first described a stochastic process that combines recombination and coalescence to generate genealogies. At each given site, the genealogy resulting from this process is equivalent to the one generated by the single-locus coalescent model (Kingman, 1982), but because recombination breaks loci apart (Figure 1.1A), the local genealogies can differ between sites.

Representations of the ARG

The full ARG can be represented as a directed graph with two types of nodes: 1) coalescence nodes, where two or more edges merge into one (backwards in time) and 2) recombination nodes, where one edge splits in two (backwards in time) (Figure 1.1B). Alternatively, the full ARG can also be represented as an ordered collection of marginal coalescence trees, annotated with the recombination nodes. These marginal trees are embedded in the graph representation (Figure 1.1B,C).

In some representations, the collection of trees may or may not contain all the information from the full ARG, depending on whether the times of recombination events (red crosses in Figure 1.1) are stored with the trees (Rasmussen et al., 2014), and whether the

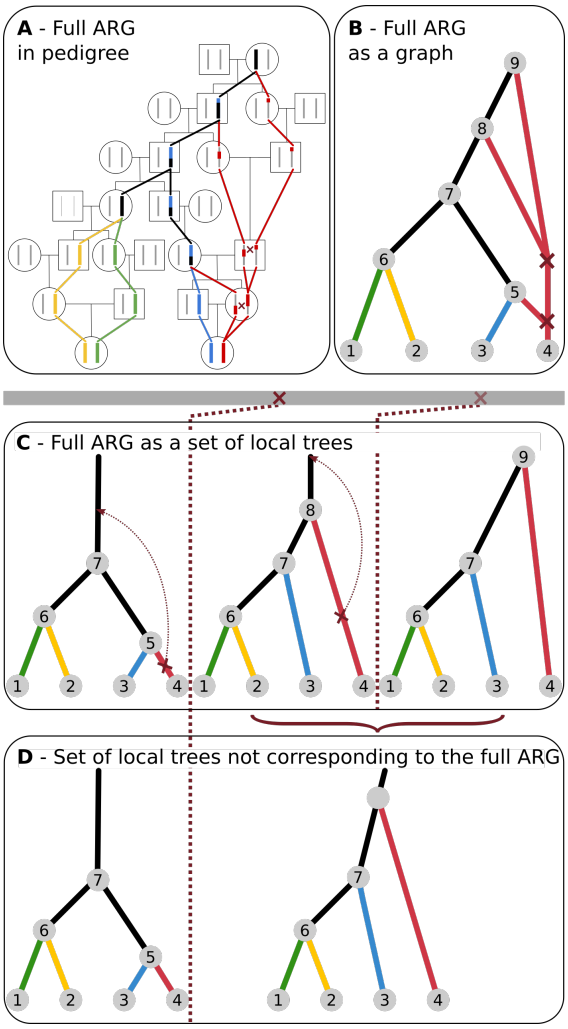

Figure 1.1: Schematic representations of the genealogy of a sample of two diploid individuals. Colors denote the four haplotypes sampled, and black lines indicate lineages or sequence tracts where at least one coalescence has occurred. Dark red crosses indicate recombination events. (A) The genealogy embedded in a pedigree. (B) An ancestral recombination graph (ARG) that fully represents all genealogical relationships shown in A, assuming that recombination events are annotated with the sequence coordinates. (C) An equivalent representation of the full ARG as a set of local trees separated by a single recombination event. (D) A set of trees that does not correspond to the full ARG. Instead, the second tree is an average of the local trees at that region.This set of trees is missing a recombination event that does not change topology, but changes the coalescence time. Other types of recombination events that could me missing in a partial ARG are: i) recombination followed by coalescence in the same branch, which does not change topology or other coalescence times and ii) topology changing recombination events.

internal nodes of the tree are labelled so they can be explicitly shared between adjacent trees. Furthermore, in some cases only topology changing recombination events are represented, and thus information regarding recombination events that do not lead to topology changes can be lost (Kelleher et al., 2019). Finally, some representations of ARGs as a collection of local trees allow more than one recombination event between trees (Speidel et al., 2019). In the latter two cases, each tree will potentially be an average of multiple coalescence trees. Figure 1.1D shows an example of a collection of local trees that does not correspond to the underlying full ARG, since one of its local trees is an average of two adjacent trees with identical topologies.

Collections of local trees with labelled internal nodes, regardless of whether they represent a full ARG or not, can be represented efficiently in computer memory by noting that each branch is part of many marginal trees (note repeated node numbers across trees in Figure 1.1C)). This property has been explored in the "tree sequence" format (Kelleher et al., 2018).

The full ARG contains all the information in a sample of DNA sequences regarding demography. Specifically, for a set of demographic parameters θ, parameters of the mutational process μ, sequence data x, and ARG G, $p(x|\theta, \mu, G) = p(x|\mu, G)$, *i.e.* if G is known there is no more information in the data about θ. A similar statement can be made for recombination and selection, if the leaf nodes of G are augmented with the allelic state at the selected loci. Therefore, the ARG is necessarily at least as informative as the combination of any and all summary statistics traditionally used to infer evolutionary processes (such as F_{ST}, π, Tajima's D, or EHH). Knowledge of the ARG is key for constructing powerful methods for extracting population genetic information from DNA sequencing data.

Inferring ARGs

Unfortunately, ARGs cannot be directly observed but must be inferred from the data. Together with an estimate of the ARG, it is desirable to quantify the uncertainty around the inferred ARG, for example by obtaining samples of ARGs according to their posterior probabilities under a given model (we discuss examples of these models in the next section). Such samples can be used to quantify uncertainty regarding ARG inferences in downstream analyses. Accurate sampling from the posterior distribution is especially relevant for downstream methods that rely on importance sampling to infer evolutionary parameters from ARGs. In essence, these methods weight parameter inference under each sampled ARG by the ARG probability and therefore require that the samples of ARGs accurately reflect their probability distribution. These types of methods can be used to infer population size history, selection (Stern et al., 2019), migration (Osmond and Coop, 2021), mutation rates and recombination rates.

Inferring full ARGs and quantifying inference uncertainty by sampling from the posterior distribution is a challenging problem computationally. It requires navigating a high-dimensional distribution of ARGs, which are themselves a complicated data structure. For this reason, inferring ARGs and sampling from their posterior distribution seemed like a nearly impossible endeavour some years ago, but important methodological developments

now allow us to do so. Today, there are several methods available to estimate the full ARG or approximations of it, including ARGweaver (Rasmussen et al., 2014), Relate (Speidel et al., 2019) and tsinfer+tsdate (Kelleher et al., 2019; Wohns et al., 2022).

Approximations of the coalescent with recombination

The classical way to include recombination in coalescence models is to consider the temporal process of lineage splitting caused by recombination and lineage merger caused by coalescences as one moves backwards in time (Hudson, 1983; Griffiths and Marjoram, 1997) (Figure 1.1A,B). Wiuf and Hein (1999) considered instead the spatial process of recombination along a sequence. In this formulation, the ARG is constructed as a sequence of local coalescent trees along a genome, where each tree is separated from adjacent trees by recombination events (Figure 1.1C). At each recombination breakpoint, a new tree is formed from the immediately preceding tree. To form the next tree, first one of the branches in the current tree is detached. Next, a point earlier than the detachment point is randomly chosen from any of the branches in **any of the previous trees** in the sequence. Finally, the detached branch coalesces to this chosen point.

To improve the computational efficiency in simulations, McVean and Cardin (2005) proposed approximating the spatial process as a Markovian process called the Sequentially Markovian Coalescent (SMC). In the SMC, when a lineage is detached from a tree at a recombination event, it can only coalesce back to one of the other lineages present **at the current tree**. Marjoram and Wall (2006) proposed an improved approximation, the SMC', in which the detached lineage can coalesce to any branch in the current tree, including the one it was detached from. This means that some recombination events in this model do not generate a different local coalescent tree. This simple modification significantly improves the model in terms of approximating the full coalescent (Marjoram and Wall, 2006; Wilton et al., 2015).

A heuristic approximation to the coalescence with recombination proposed by Li and Stephens (2003), extending ideas from Stephens and Donnelly (2000), approximates the coalescent with recombination using a copying process where one sequence is modeled as a copy of other sequences in the sample, with errors representing mutations and switches in the copying template representing recombination events. While this model has disadvantages, such as a dependence on the input order of sequences, it has proven computationally convenient for many purposes, including demography inference, introgression detection, and more (Sheehan et al., 2013; Steinrücken et al., 2018, 2019).

The formulation of the coalescent with recombination approximated as a Markovian process generating tree sequences in the SMC (McVean and Cardin, 2005) and SMC' (Marjoram and Wall, 2006) and as a copying process of individual sequences by Li and Stephens (2003), paved the way for more scalable ARG inference methods. Notably, ARGweaver (Rasmussen et al., 2014) based on the SMC or SMC' model, and Relate (Speidel et al., 2019) and tsinfer+tsdate (Kelleher et al., 2019; Wohns et al., 2022) based on the model by Li and Stephens (2003).

Table 1.1: Genome-wide genealogy inference programs compared.

Program	Samples topologies	Samples coalescence times	Supports demographic model	Scalability (number of genomes)	Outputs full ARG	Supports unphased data
ARGweaver	Yes	Yes	No	∼50	Yes	Yes
ARGweaver-D*	Yes	Yes	Yes	∼50	Yes	Yes
Relate	No	Yes	No	$\sim 10^3$	No	No
tsinfer+tsdate	No	No	No	$\sim 10^5$	No	No

* Hubisz et al. (2020)

ARGweaver

ARGweaver uses Markov Chain Monte Carlo (MCMC) to sample ARGs from the posterior distribution under the SMC or SMC'. It relies on a discretization of time (such that all recombination and coalescence events are only allowed to happen at a discrete set of time points) which makes the state space of ARGs finite countable and allows the use of discrete state-space Hidden Markov Models (HMMs). It then uses a lineage threading approach, which is a Gibbs sampling update, to sample the history of a single lineage or haplotype from the full conditional posterior distribution given the rest of the ARG connecting all other haplotypes.

Relate

Relate simplifies the problem of ARG inference by inferring marginal coalescence trees, instead of full ARGs. Inference is divided into 2 steps. First, the Li and Stephens (2003) haplotype copying model is used to calculate pairwise distances between samples in order to infer local tree topologies. Next, it uses MCMC under a coalescent prior to infer coalescence times on those local trees. Relate is able to output samples of coalescence times from the posterior distribution using this MCMC approach, but it does so for the same fixed sequence of tree topologies. This is different from the ARGweaver MCMC sampling, which also samples the tree topology space (Table 1.1).

tsinfer, tsdate, and the tree sequence framework

Tsdate (Wohns et al., 2022) is a method that estimates coalescence times of tree sequences. Here, we used this method to date tree sequences inferred by tsinfer (Kelleher et al., 2019). Similarly to Relate, tsinfer is also based on the copying process from Li and Stephens (2003). A key innovation of tsinfer is a highly efficient tree sequence data structure which stores sequence data and genealogies (Kelleher et al., 2016, 2018, 2019; Ralph et al., 2020). Tsinfer performs inference in two steps. First, it recreates ancestral haplotypes based on allele sharing between samples. Next, it uses an HMM to infer the closest matches between

ancestral haplotypes and the sampled haplotypes using an ancestral copying process modified from the classical Li and Stephens (2003) model to generate the tree topology. Finally, nodes in tree sequences inferred by tsinfer can be dated by tsdate. Tsdate uses a conditional coalescent prior, where the standard coalescent is conditioned on the number of descendants of each node on a local tree. Like ARGweaver, tsdate also discretizes time for computational efficiency. This framework infers a fixed topology and coalescence time, but it has the potential to sample coalescence times.

Benchmarking of ARG inference methods

Here, we use standard neutral coalescent simulations to benchmark coalescence time inferences in ARGweaver (Rasmussen et al., 2014), Relate (Speidel et al., 2019), and tsinfer+tsdate (Kelleher et al., 2019; Wohns et al., 2022). We focus mainly on ARGweaver and Relate because they report measures of uncertainty in inference by allowing the user to output multiple samples from the posterior distribution. Sampling from the posterior is not currently implemented in tsdate (Table 1.1), but we include it in this evaluation because it is a promising framework for very fast tree-sequence inference, and it will likely provide an option to output samples from the posterior distribution of tree-sequences in future updates.

We focus our analyses on coalescence times not only because they are a very informative statistic about evolutionary processes, but also because they can be fairly compared across all methods. More specifically, ARGweaver and tsdate allow for polytomies (*i.e.*, more than two branches coalesce at the same node). Relate, on the other hand, does not allow polytomies. Comparing topologies with and without polytomies could bias our results depending on how we chose to deal with polytomies, so we decided to focus on coalescence times only.

We run coalescent simulations on msprime (Kelleher et al., 2016) and compare the true (simulated) ARGs to the ARGs inferred by ARGweaver, Relate, and tsinfer+tsdate. We compare the ARGs with respect to their pairwise coalescence times using three different types of evaluation (Figure 1.2). First, we compare the true pairwise coalescence time at each site to the inferred time. Second, we compare the overall distribution of pairwise coalescence times across all sites and all MCMC samples to the expected distribution. In Bayesian inference, the data averaged posterior distribution is equal to the prior. Since data are simulated under the standard coalescent with recombination the data averaged posterior should be exponential with rate 1 in coalescence time units ($2N_e$ generations, where N_e is the effective population size). Third, we used simulation-based calibration (SBC) (Cook et al., 2006; Talts et al., 2020) to evaluate if the posterior distributions sampled by ARGweaver and Relate are well calibrated (see details in Methods).

1.2 Methods

Simulations

We simulated tree sequences and SNP data with msprime version 0.7.4 (Kelleher et al., 2016). For simulations with Jukes and Cantor (1969) mutational model, we used msprime version 1.0.2 (Baumdicker et al., 2021) to add mutations to trees simulated under msprime 0.7.4, because the Jukes and Cantor (1969) model option was not available in msprime 0.7.4. Unless otherwise noted, simulations were done under the standard neutral coalescent (Hudson model in msprime) and using the following parameters: 4 diploid samples (*i.e.* 8 haplotypes), total map length $R = 20000$ and mutation to recombination rate ratio $\mu/\rho = 1$. In practice, we used the following parameter values in msprime: effective population size of 10,000 diploids ($2N_e = 20,000$), mutation rate and recombination rate of 2×10^{-8} per base pair per generation and a total sequence length of 100Mb.

We varied these standard simulation scenarios in several ways: using SMC and SMC' models, different numbers of samples (4, 16, 32 and 80 haplotypes), a 10-fold increase and 10-fold decrease in the mutation to recombination ratio (in each case changing either the mutation or the recombination rates), and changing the total length of input sequence from 100Mb to 5Mb and 250kb. These simulated sequences were then divided into 20 equally sized segments, so that ARGweaver could be run on each in parallel (see below). The minimum length of total simulated sequence (250kb) was chosen such that the average number of pairwise differences between each of the 20 segments was 10, given a mutation rate of 2×10^{-8}.

We extracted coalescence times at all sites in the simulated trees in BED format (columns: chromosome, start position, end position, coalescence time), with one BED file for each pair of samples. Figure 1.2 shows an overview of the metrics extracted from simulated ARGs and from ARGs estimated by tsinfer+tsdate or sampled from the posterior by ARGweaver and Relate.

ARGweaver

VCF files from msprime were converted to ARGweaver sites format using a custom python script. We ran ARGweaver's *arg-sample* program to sample ARGs. This was done in parallel on 20 segments of equal size, using the *–region* option. We used the same values used in the msprime simulations (*–mutrate* and *–recombrate 2e-8* and *–popsize 10000*) and except where otherwise noted, we ran ARGweaver using the SMC' model (*–smcprime* option). We ran ARGweaver with 1200 or 2200 iterations (*–iters*) (with burn-in of the first 200 or 1200 iterations, respectively), depending on how long it took to converge. Assessment of convergence is described below and in the Appendix of Chapter 1, Evaluating MCMC Convergence. We extracted 100 MCMC samples from every 10th iteration among the last 1,000 iterations (default *–sample-step 10*).

We extracted all pairwise coalescence times in BED format using options *–tmrca* and *–subset* in the program *arg-summarize*, and we used bedops (version 2.4.35 (Neph et al.,

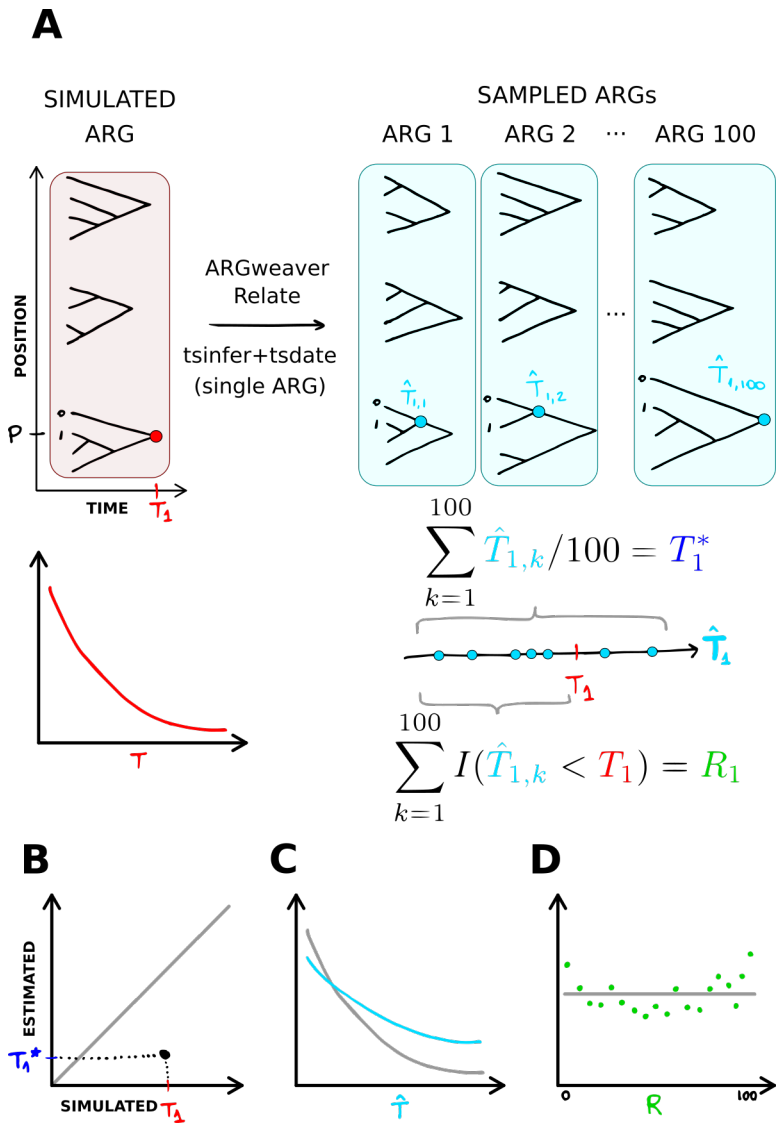

Figure 1.2: Methods overview. (A) Data (ARGs and DNA sequences) were simulated from the coalescent with recombination. In the model and simulated data, pairwise coalescence times (CT) are exponentially distributed (Figure A.3). T_1 represents the CT between samples 0 and 1, at position P in the simulated data. $\hat{T}_{1,k}$ is the CT between samples 0 and 1 at position P, in each ARG sample k. Point estimates T_1^* are obtained as the mean of $\hat{T}_{1,k}$, and the rank statistic is computed as the number of $\hat{T}_{1,k}$ that are smaller than the true value T_1. (B) We compare estimated to simulated values of the CT of each pair of samples, at each position of the genome. (C) We compare the distribution of sampled CT across all sampled ARGs, all sites and all pairs of samples to the expected exponential distribution. (D) We compare the distribution of ranks to the expected uniform distribution.

2012)) to match the times sampled by ARGweaver to the simulated ones at each sequence segment. Finally, we used a custom Python script to calculate the ranks of simulated pairwise coalescence times on ARGweaver MCMC samples per site.

Time discretization

In ARGweaver, time is discretized such that recombination and coalescence events are only allowed to happen at a user-defined number of time points, K (default value is 20) (Rasmussen et al., 2014). These time points s_j (for $0 <= j <= K-1$) are given by the function

$$s_j = g(j) = \frac{1}{\delta}\left\{e^{\frac{j}{K-1}log(1+\delta s_{K-1})} - 1\right\} \tag{1.1}$$

where δ is a parameter determining the degree of clustering of points in recent times. Small values of δ lead to a distribution of points that is closer to uniform between 0 and s_{K-1}, and higher values increase the density of points at recent times (default value is 0.01) (Hubisz and Siepel, 2020). Equation 1.1 ensures that s_0 is always 0, and s_{K-1} (or s_{max}) is user defined by the parameter *–maxtime* (default value is 200,000).

Rounding of continuous times into these K time points is done by defining bins with breakpoints between them, such that the breakpoint between times s_j and s_{j+1} is $s_{j+\frac{1}{2}} = g(j+\frac{1}{2})$. All continuous values in the bin between $s_{j-\frac{1}{2}}$ and $s_{j+\frac{1}{2}}$ are assigned the value s_j. We note that for the first and last intervals, the values assigned (s_0 and s_{K-1}) do not correspond to a midpoint in the time interval but rather to its minimum ($s_0 = 0$) or maximum ($s_{K-1} = s_{max}$)

Here, when reporting results in bins, we use the same time discretization as defined by the ARGweaver breakpoints ($s_{j+\frac{1}{2}}$). However, we change the value assigned to times in these bins: instead of using s_j, we define t_j as the median of the exponential distribution with rate 1 at the interval between $s_{j-\frac{1}{2}}$ and $s_{j+\frac{1}{2}}$. To this end, we first calculate the cumulative probability of the exponential distribution with rate 1 up to the median of the jth interval

$$\begin{aligned} p_j = & \int_0^{s_{j-\frac{1}{2}}} e^{-x}\,\mathrm{d}x + \frac{1}{2}\int_{s_{j-\frac{1}{2}}}^{s_{j+\frac{1}{2}}} e^{-x}\,\mathrm{d}x \\ =& 1 - \left(\frac{e^{-s_{j-\frac{1}{2}}} + e^{-s_{j+\frac{1}{2}}}}{2}\right) \end{aligned} \tag{1.2}$$

We then take the inverse CDF of the exponential distribution with rate 1, at the point p_j, to find the time $t_j = -ln(1-p_j)$ corresponding to the median value for the interval.

This step is relevant for the simulation-based calibration (see below), where we take the rank of true (simulated) coalescence times relative to the values sampled by ARGweaver. If we used s_j, coalescence times in the first or last ARGweaver time interval would not be

represented by a midpoint. We correct for that by using t_j, so that all time intervals are comparable.

Relate does not use time discretization, and tsdate uses a discretization scheme where the time points are the quantiles of the lognormal prior distribution on node ages (Wohns et al., 2022). Here, we always apply the ARGweaver time discretization scheme when comparing results in bins.

Relate

VCF files generated with msprime were converted to Relate haps and sample files using *RelateFileFormats –mode ConvertFromVcf* and Relate's *PrepareInputFiles* script. We ran Relate (version 1.1.2) using *–mode All* with the same mutation rate (*-m 2e-8*) and effective population size (*-N 20000*) used in the msprime simulations, as well as a recombination map with constant recombination rate along the genome, with the same rate used in msprime (*2e-8*).

We used Relate's *SampleBranchLengths* program to obtain 1000 MCMC samples of coalescence times for the local trees inferred in the previous step in anc/mut output format (*–num-samples 1000 –format a*). Similarly to the ARGweaver analysis, we also performed this step in 20 sequence segments of 5Mb, and we thinned the results to keep only every 10th MCMC sample. Finally, we extract pairwise coalescence times and calculate the ranks of true pairwise coalescence times relative to the 100 MCMC samples. Due to the large number of pairwise coalescence times, for the simulations with 80 and 200 samples, we extracted coalescence times from a subset of 210 pairs of samples. We extracted coalescence times for every 4th vs. every 4+1th sample in the case of 80 samples, and 10th vs. every 10+1th sample in the case of n=200.

tsinfer and tsdate

VCF files generated by msprime were provided as input to the python API using *cyvcf2.VCF* and converted to tsinfer *samples* input object using the *add_diploid_sites* function described in the tsinfer tutorial (`https://tsinfer.readthedocs.io/en/latest/tutorial.html#reading-a-vcf`). Genealogies were inferred with tsinfer (version 0.2.0 (Kelleher et al., 2019)) with default settings and dated with tsdate (version 0.1.3 (Wohns et al., 2022)) using the same parameter values as in the simulations (*Ne=10000*, *mutation_rate=2e-8*), with a prior grid of 20 timepoints.

Pairwise coalescence times were extracted from the tree sequences using the function *tmrca()* from tskit (version 0.3.4 (Kelleher et al., 2018)), and output in BED format, with one file for each pair of samples. Finally, coalescence times at each site, for each pair of samples were matched to the simulated ones (also in BED format) using bedops (Neph et al., 2012).

MCMC convergence

We evaluated MCMC convergence of Relate and ARGweaver through 1) visual inspection of trace plots, 2) autocorrelation plots, 3) effective sample sizes and 4) the Gelman-Rubin convergence diagnostics based on potential scale reduction factor (Gelman and Rubin, 1992; Brooks and Gelman, 1998). Trace plots were also used to determine the number of burn-in samples, and autocorrelation plots were used to determine thinning of the samples. See Evaluating MCMC Convergence in Appendix of Chapter 1 for details.

Point estimates of pairwise coalescence times

We estimated pairwise coalescence times from the MCMC samples from Relate and ARGweaver by taking the average of 100 samples at each site (Figure 1.2). Since tsdate does not output multiple samples of node times, we use its point estimate of pairwise coalescence times directly. Point estimates of coalescence times were compared to the simulated values for each pair of samples, at each site along the sequence.

Mean squared error (MSE) of point estimates was calculated from each point estimate of coalescence time (for each pair of samples, at each site), as well as per bin of size 0.1 of the simulated coalescence times (in units of $2N_e$ generations) for Figure A.2. We also report Spearman's rank correlation (r_2) of the point estimates of pairwise times in each tree against the simulated tree, averaged over all positions in the genome.

Simulation-based calibration

In addition to comparing MCMC point estimates to the true simulated values, we use simulation methods proposed by Cook et al. (2006) and Talts et al. (2020) to assess whether Bayesian methods are sampling correctly from the true posterior distribution. (Cook et al., 2006) proposed simulating data using parameters sampled from the prior. The posterior, when averaged over multiple simulated data sets, should then equal the prior.

In our case, we sample ARGs, G, from the full coalescence process with recombination with a known implicit prior of pairwise coalescence times, $P(t) = e^{-t}$. We simultaneously simulate sequence data, x, on the simulated ARGs from the distribution $p(x) = \int p(x|G)\mathrm{d}P(G)$. The distribution of the averaged posterior of G, $p_{ave}(G) = \int p(G|X)\mathrm{d}P(x)$ should then equal the prior for G (Talts et al., 2020), and hence the prior distribution for the pairwise coalescence times, t, should equal the averaged posterior distribution for t. Here, all population parameters relating to mutation, effective population sizes, etc., are kept fixed and suppressed in the notation. One way we will examine the accuracy of the posterior inferences is, therefore, to compare the average of the posterior of t to the exponential distribution. In practice, we simulate data using msprime (Kelleher et al., 2016) and pipe the data to the MCMC samplers (ARGweaver and Relate) for inference of the posterior distribution. ARGweaver uses an approximation (SMC') of the model (coalescent with recombination) used in the data simulations, and Relate uses a heuristic method based on

the Li and Stephens model. Thus, inadequacies of the fit of the posteriors could potentially be caused by this discrepancy between the model used in simulations and the models used for inference.

However, even if the averaged posterior resembles an exponential, the inferences for any particular value of t may have a posterior that is too narrow or too broad. For a closer examination of the accuracy of the posterior, we use a method proposed by Cook et al. (2006) and Talts et al. (2020) that compares each posterior to the true value. To this end, we compare each true (simulated) pairwise coalescence time to the corresponding posterior for the same pair of haplotypes. If the posterior is correctly calculated, the rank of the true value relative to the samples from the posterior should be uniformly distributed (Cook et al., 2006; Talts et al., 2020). We use 100 MCMC samples from ARGweaver and Relate for each data set, meaning our ranks take values from 0 to 100. Deviations from the uniform distribution of ranks quantifies inaccuracies in estimation of the posterior. For example, an excess of low and high ranks indicates that the inferred posterior distribution is underdispersed relative to the true posterior.

Data availability

All the code and data related to this work are available in GitHub `https://github.com/deboraycb/ARGsims`.

1.3 Results

Comparison of simulated to estimated coalescence time per site

We compared coalescence times estimated by ARGweaver, Relate and tsinfer+tsdate to the true values known from msprime simulations. In all three methods, estimates of coalescence time per site are biased (Figure 1.3 and A.2). Small values of coalescence times are generally overestimated, while large values tend to be underestimated (Fig A.2). In tsinfer+tsdate, point estimates are apparently bounded to a narrow range (Figure 1.3G). The mean squared error (MSE) of point estimates is larger in Relate (MSE=0.625) and tsinfer+tsdate (MSE=1.631) than in ARGweaver (MSE=0.397), showing that point estimates of pairwise coalescence times at each site are closer to the true value in ARGweaver. Spearman's rank correlation is also highest in ARGweaver (r_s=0.761), but in this metric tsinfer+tsdate (r_s=0.705) perform better than Relate (r_s=0.669).

For ARGweaver and Relate, the point estimates of coalescence times are obtained as the means of samples from the posterior. These Bayesian estimates are not designed to be unbiased and unbiasedness of the point estimator is arguably not an appropriate measure of performance for a Bayesian estimator. Therefore, we also evaluate the degree to which the posterior distributions reported by ARGweaver and Relate are well calibrated, *i.e.* represent distributions that can be interpreted as valid posteriors, and the degree to which

the data-averaged posterior distributions of coalescence times equals the prior exponential distribution.

Posterior distribution of coalescence times

We simulated data under the standard coalescent model, where the distribution of pairwise coalescence times (in units of $2N_e$ generations, where N_e is the diploid effective population size) follows an exponential distribution with rate parameter 1 (Figure A.3). As argued in the Methods section, the same is true for the data-averaged posterior.

We compared the expected exponential distribution of coalescence times to the observed distribution of coalescence times across all sites inferred by ARGweaver, Relate, and tsinfer+tsdate (Figure 1.4). For ARGweaver and Relate, we output 100 MCMC samples from the posterior distribution and plot the distribution of pairwise coalescence times across all sites and MCMC samples.

To facilitate visual comparison of the distributions between methods, we discretized Relate and tsinfer+tsdate coalescence times into the same bins as ARGweaver (Figure 1.4D,G, see distributions without discretization in Figure A.4 and see Methods for a description of ARGweaver time discretization). Because the time discretization breakpoints are regularly spaced on a log scale, we use a log scale on the x-axis for better visualization.

Distributions of coalescence times from ARGweaver and Relate (Figure 1.4A and D) show an excess around 1, when compared to the expected exponential distribution. However, that bias is more pronounced in Relate than ARGweaver. In tsinfer+tsdate, the distribution is truncated at 1.6, and it deviates more strongly from the expected exponential distribution (Figure 1.4G). We note that the plots from ARGweaver and Relate are not directly comparable to those of tsinfer+tsdate, since there are 100 coalescence time samples at each site from the former two programs but only one from tsdate.

Simulation-based calibration

In this section, we use simulation-based calibration to evaluate whether ARGweaver and Relate are generating samples from a valid posterior distribution of coalescence times (see Methods). To that end, we simulated coalescence times at multiple sites following the standard coalescent prior distribution, and we generated 100 MCMC samples from the posterior distribution using both ARGweaver and Relate. Finally, we analyse the distribution of the ranks of the simulated coalescence times relative to the 100 sampled values at each site.

In the previous section, we showed that the posterior distributions of ARGweaver and Relate are similar to the theoretically expected exponential distribution. However, in that analysis we have not evaluated the distribution of MCMC samples relative to each simulated value. The results of simulation-based calibration are informative about that distribution and can reveal if the posterior distribution is well calibrated.

The distribution of ranks from ARGweaver (Figure 1.5A, Kullback-Leibler Divergence (KLD) =0.027) is closer to uniform than that of Relate (Figure 1.5D, KLD=0.602). However,

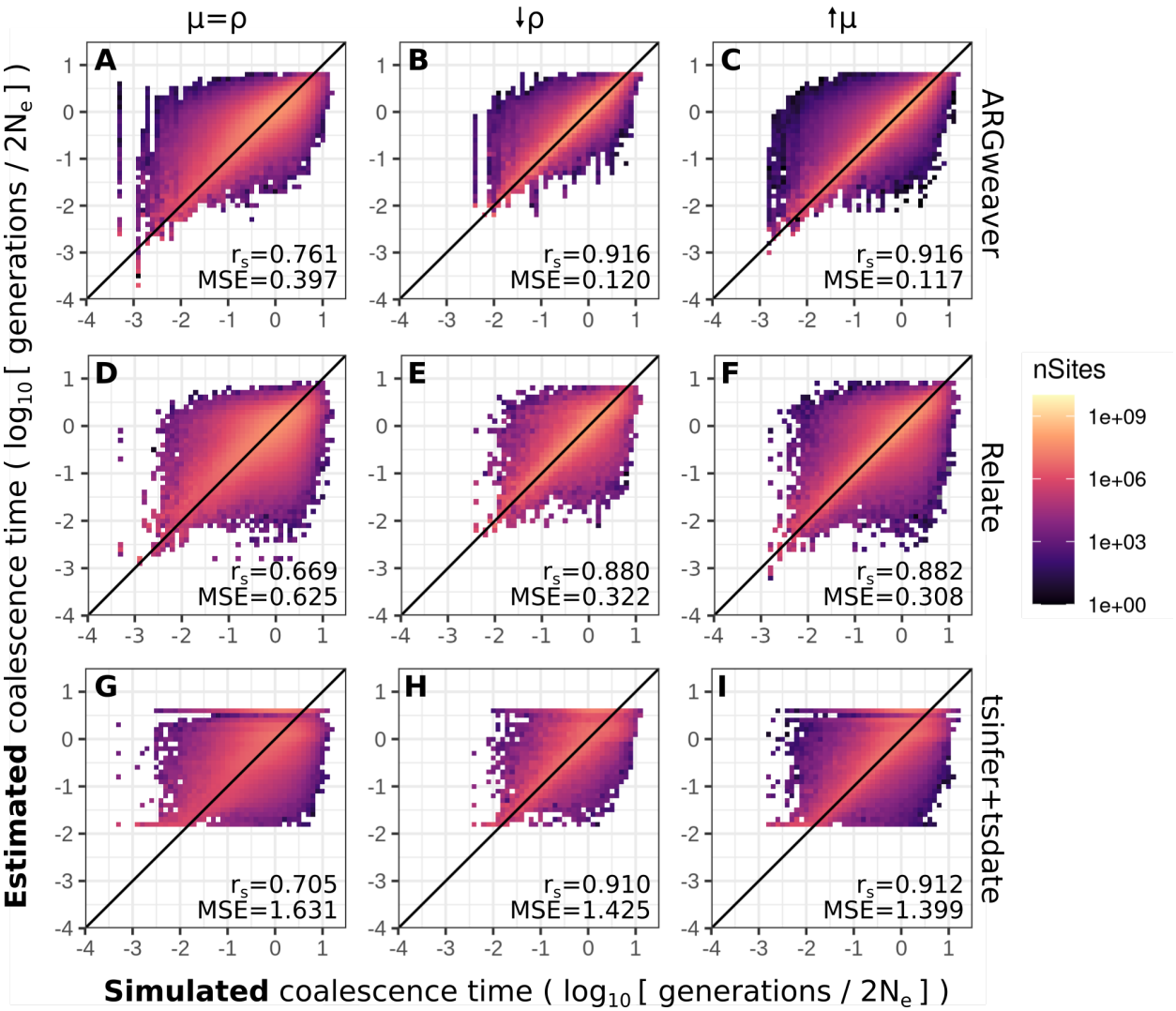

Figure 1.3: Point estimates of coalescence times in ARGweaver (A-C), Relate (D-F) and tsinfer+tsdate (G-I). Left column: $\mu = \rho = 2 \times 10^{-8}$; middle column: $\mu/\rho = 10$, $\rho = 2 \times 10^{-9}$; right column: $\mu/\rho = 10$, $\mu = 2 \times 10^{-7}$. For ARGweaver and Relate, point estimates are the means of 100 MCMC iterations. Note that axes are in log scale. See Figure A.1 for the data in plots A,D,G plotted in linear axes. Diagonal line shows x=y. MSE: Mean squared error; r_s: Spearman's rank correlation.

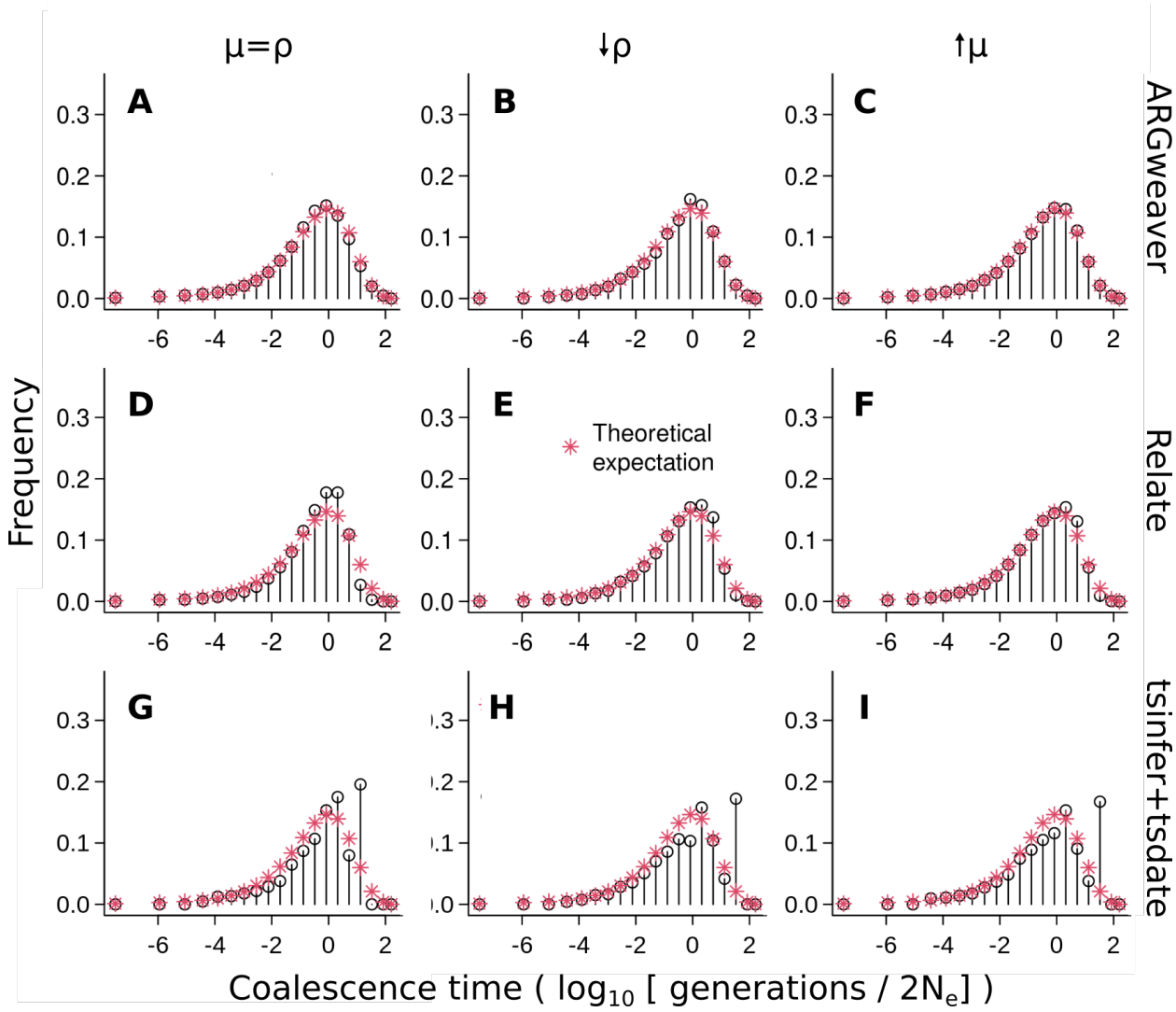

Figure 1.4: Distribution of coalescence times inferred by ARGweaver (A-C), Relate (D-F) and tsinfer+tsdate (G-I). Left column: $\mu = \rho = 2 * 10^{-8}$; middle column: $\mu/\rho = 10$, $\rho = 2 * 10^{-9}$; right column: $\mu/\rho = 10$, $\mu = 2 * 10^{-7}$. Plots D and G show the same data as in Figure A.4, using different binning.

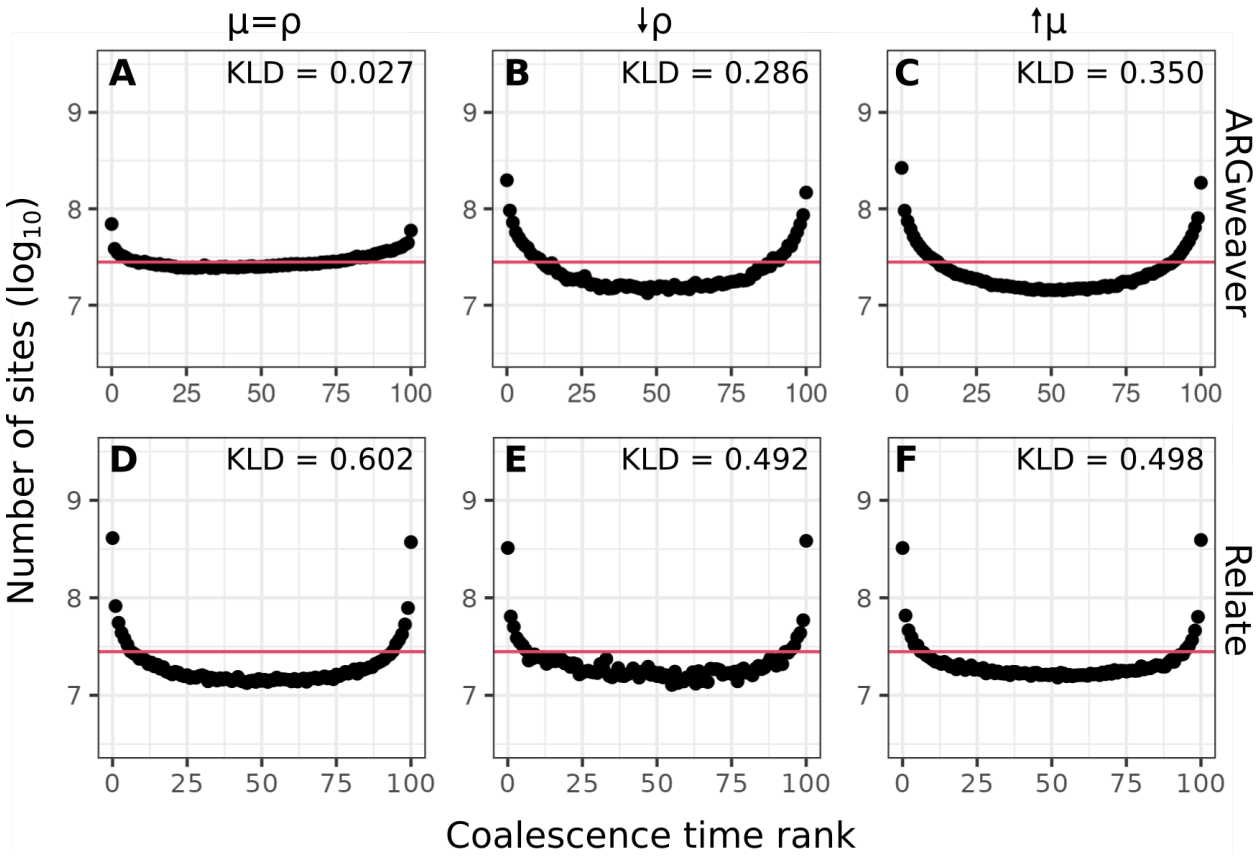

Figure 1.5: Counts of ranks from simulation-based calibration in ARGweaver (A-C) and Relate (D-F). Horizontal line shows expected uniform distribution. Left column: $\mu = \rho = 2 * 10^{-8}$; middle column: $\mu/\rho = 10$ decreasing recombination rate ($\rho = 2 * 10^{-9}$); right column: $\mu/\rho = 10$ increasing mutation rate ($\mu = 2 * 10^{-7}$). Horizontal line shows expected uniform distribution.

both show an excess of low and high ranks. The excess of low and high ranks indicates that the sampled posterior distribution is underdispersed (Talts et al., 2020), *i.e.* the posterior has too little variance and does not represent enough uncertainty regarding the coalescence times.

One possible cause for this type of deviation from the uniform distribution could be MCMC convergence, *i.e.*, samples being autocorrelated, resulting in effective sample size is lower than the number of samples taken, the MCMC chain not mixing well and/or the MCMC chain not being run long enough to achieve convergence.

We show detailed results for MCMC convergence in Relate and ARGweaver in the Appendix of Chapter 1. Briefly, we have not found these types of convergence issues in ARGweaver or Relate with simulations of 8 haplotypes and mutation to recombination ratio of 1.

Potential scale reduction factor (PSRF) from Gelman-Rubin convergence diagnostic statistics are all close to 1 (Tables A.2, A.3), and effective sample sizes are almost all larger than 100. Therefore, MCMC convergence does not seem to explain why the rank distributions are not uniform.

Increased mutation to recombination ratio

When inferring an ARG from sequence data, the information for inference comes from mutations that cause variable sites in the sequence data. The lower the recombination rate, the longer the span of local trees will be and the more mutations will be available to provide information about each local tree. More generally, an increased mutation to recombination ratio is expected to increase the amount of information available to infer the ARG.

In our standard simulations presented so far, the mutation to recombination ratio is one ($\mu = \rho = 2 * 10^{-8}$). We increased the simulated mutation to recombination ratio to 10, both by decreasing the recombination rate (ρ) tenfold and also by increasing the mutation rate (μ) tenfold. We expected that these scenarios would improve inference of ARGs, and consequently the estimates of pairwise coalescence times. Point estimates are better with increased mutation to recombination ratio in ARGweaver (Figure 1.3B,C), Relate (Figure 1.3E,F) and tsinfer+tsdate (Figure 1.3H,I).

The coalescence times distribution in Relate (Figure 1.4E,F) are closer to the expected with $\mu/\rho = 10$ relative to $\mu/\rho = 1$ (Fig 1.4D), and the simulation-based calibration also improved (Figure 1.5D-F, KLD=0.492 and 0.498 compared to KLD=0.602).

The results from ARGweaver with $\mu/\rho = 10$ were more surprising, with the simulation-based calibration showing a more pronounced underdispersion of the posterior distribution (Figure 1.5B,C, KLD=0.286 and 0.350, compared with KLD=0.027 for $\mu/\rho = 1$). The overall distribution of coalescence times, however, showed little change (Figure 1.4B,C). One possible explanation for ARGweaver results being worse with higher mutation to recombination ratio might be that MCMC mixing is worse under those conditions, leading to convergence issues not observed for the previous scenario. Examining convergence diagnostics seems to confirm this with more coalescence times showing low effective sample size, and with a potential scale reduction factor showing evidence of lack of convergence of some coalescence times (see Evaluating MCMC Convergence in Appendix of Chapter 1).

We show additional simulation results in the Appendix of Chapter 1, including simulations with reduced μ/ρ, which could be a realistic scenario around recombination hotspots (Figures A.5 and A.6) and ARGweaver results on simulations with intermediate values of μ/ρ (2 and 4), under the SMC and SMC' genealogy models, and with the Jukes-Cantor mutation model in the Appendix of Chapter 1.

Number of samples

Next, we evaluate ARG inference with simulations with different sample sizes. Our standard sample size used so far was 8 haplotypes, and here we change it to 4, 16 and 32. For Relate

and tsinfer+tsdate, which are scalable to larger sample sizes, we also evaluated inference with 80 and 200 sampled haplotypes.

For ARGweaver, increasing sample sizes decreased the MSE of point estimates (Figure 1.6A-C), distributions of coalescence times remained similar (Figure 1.7A-C), but underdispersion of the posterior distribution increased (Figure 1.8A-C). As mentioned in the previous section, this could be caused by an MCMC mixing problem. In particular, a larger number of samples will contribute to an increasing number of states for ARGweaver to explore, possibly leading to poor MCMC convergence (see Evaluating MCMC Convergence).

With a smaller sample size (n=4 haplotypes), the coalescence time distribution from Relate showed an excess around the mean value (coalescence time of 1) (Fig 1.7D). With increasing sample sizes, it became more similar to the expected distribution (Fig 1.7E-H). Calibration of the posterior distribution improved with increasing sample sizes up to 32 haplotypes (Figure 1.8D-H).

Both the point estimates and posterior distribution of coalescence times in tsinfer+tsdate do not consistently improve or worsen with increasing sample sizes in the range tested here (Figures 1.6I-M and 1.7I-M).

Length of input sequence

Point estimates of all programs remained similarly accurate when a much shorter input sequence was provided (5mb and 250kb, Figure A.7A-C and A.8A-C, compared with 100Mb in previous analyses). The distribution of coalescence times with 5Mb input sequence remained similar to the ones inferred with 100Mb input sequence (Figure A.7D-F). However, distributions from simulations with only 250kb input sequence are visibly more deviated from the expected exponential distribution (Figure A.8D-F). Distributions of ranks are noisier with decreasing input sequence length, but KLD remained similar (Figure A.8 and A.7H,G).

Runtime

We point out that runtimes differ widely among the programs compared here, and this factor should be taken into account for users making decisions on what method to use for their applications. For example, in the simulations with mutation rate equal to recombination rate, with sample size of 8 haplotypes and taking 1000 MCMC samples, ARGweaver took a total of 641 computing hours while Relate took 17 hours. The clock time was reduced by running both programs in parallel for segments of 5Mb of the total 100Mb sequence, meaning that ARGweaver took approximately 35h. However, this still could be a significant amount of time for the user, depending on their utilization of the algorithm. For a systematic comparison of runtimes between Relate and ARGweaver, see Speidel et al. (2019). Impressively, tsinfer and tsdate took only 5 minutes.

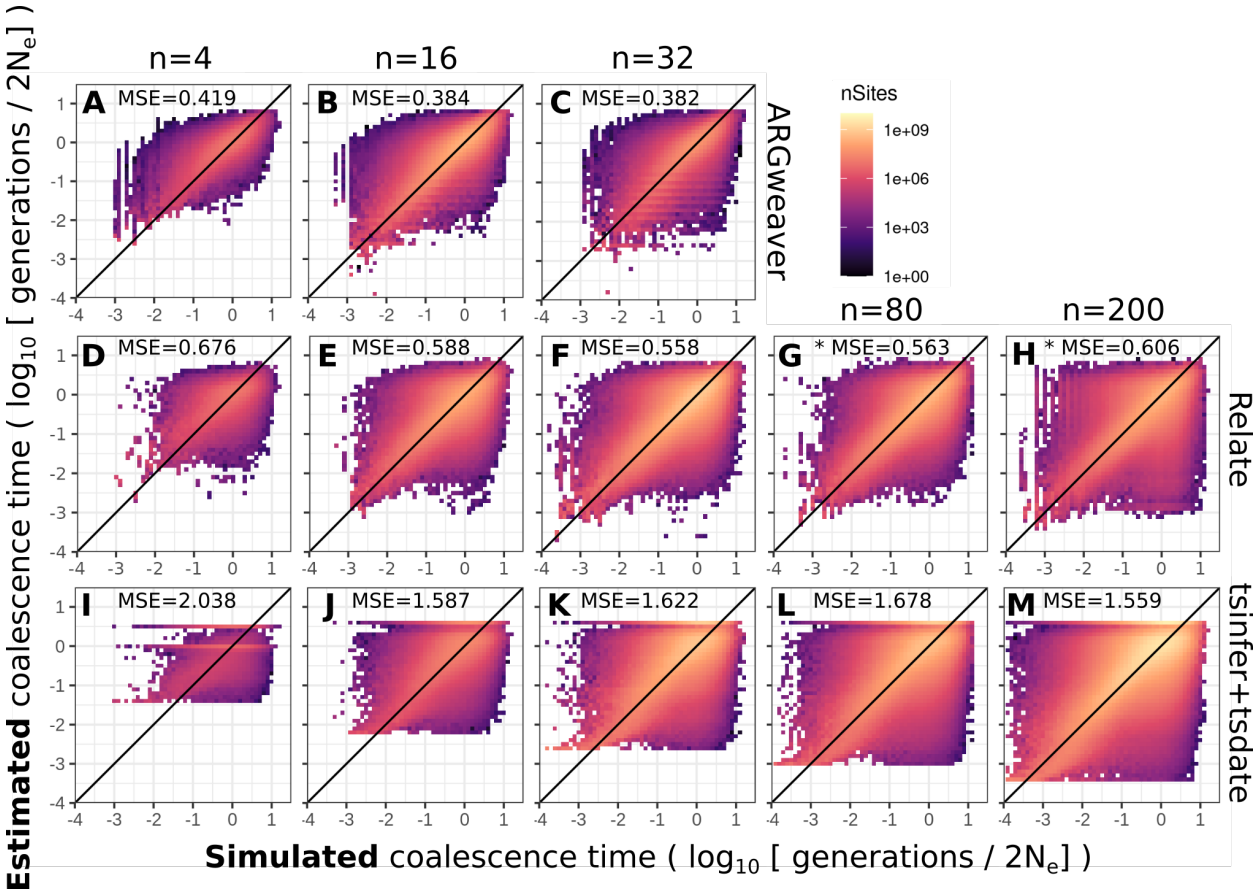

Figure 1.6: Point estimates of ARGweaver (A-C), Relate (D-H) and tsinfer+tsdate (I-M). Columns show different number of simulated samples 4, 16, 32, 80 or 200 haplotypes. Mean squared error (MSE) is shown for each plot. Note that ARGweaver is not scalable for simulations with larger sample sizes. * indicate results for a subset of 210 pairs of samples, instead of all pairwise coalescence times.

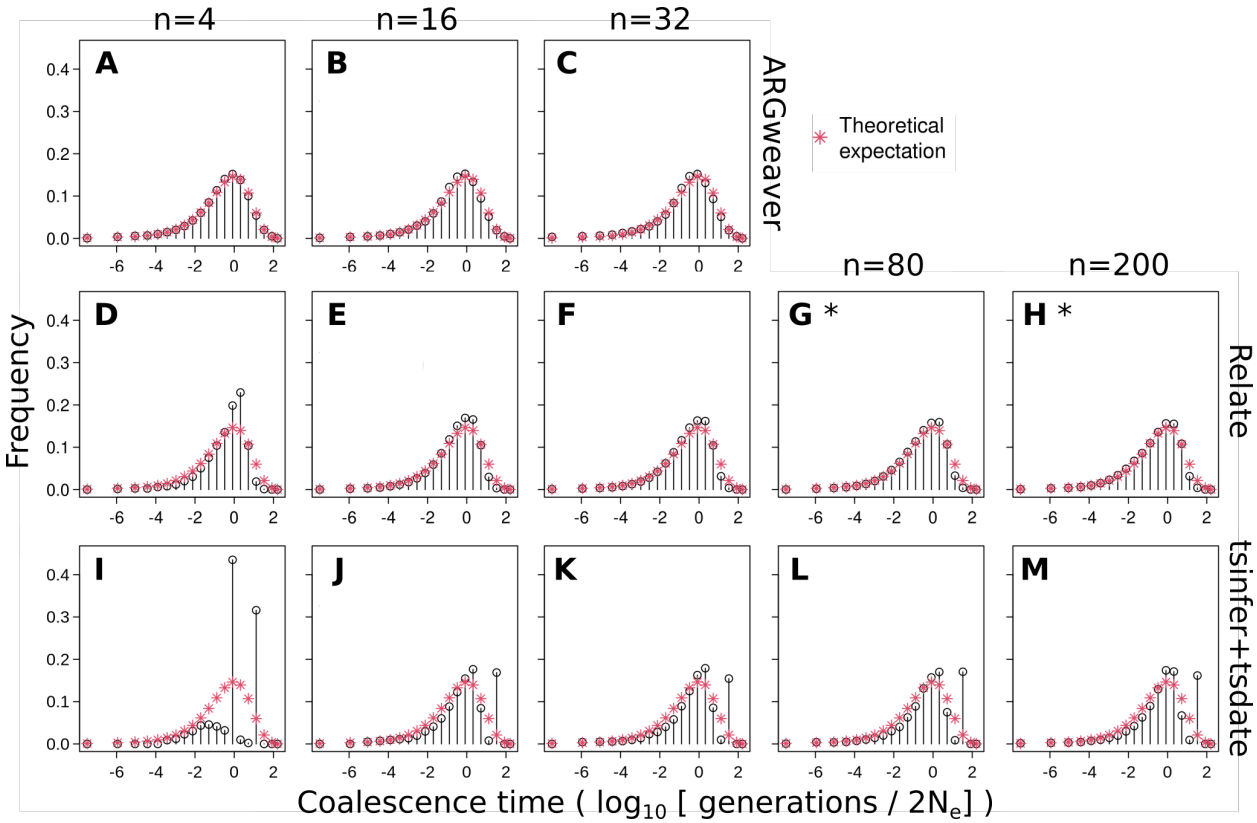

Figure 1.7: Distribution of coalescence times in ARGweaver (A-C), Relate (D-H) and tsinfer+tsdate (I-M). Columns: sample sizes of 4, 16, 32, 80, 200 haplotypes. * indicate results for a subset of 210 pairs of samples, instead of all pairwise coalescence times.

1.4 Discussion

ARG inference promises to be a tremendously useful tool for inferences of evolutionary history, such as natural selection or demography. However, it is also a very hard computational problem. We compared methods that use different approaches to this problem and evaluated their accuracy using simulated data and comparisons of three aspects of coalescence time estimates: 1) individual point estimates of each pairwise coalescence time; 2) the overall distribution of coalescence times across all sites; 3) the calibration of the reported posterior distributions.

Ancestral recombination graphs are extremely rich in information, including topological information of individual coalescence trees and information regarding the distribution of recombination events. We have not evaluated these aspects of inferred ARGs but have instead only focused on pairwise coalescence times. However, pairwise coalescence times are extremely informative statistics about many population-level processes and pairwise rela-

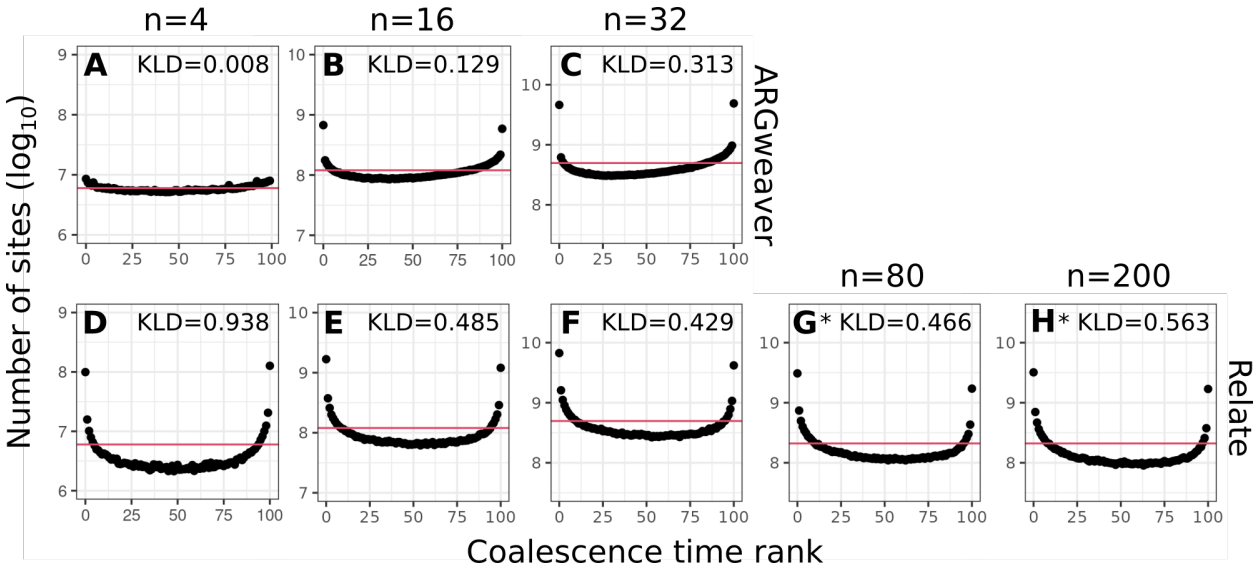

Figure 1.8: Simulation-based calibration for ARGweaver (A-C) and Relate (D-H). Columns: sample sizes of 4, 16, 32, 80, 200 haplotypes. Horizontal line shows expected uniform distribution. Note that the y-axis is centralized on different values but always has the same length. * indicate results for a subset of 210 pairs of samples, instead of all pairwise coalescence times.

tionships between individuals, and they are also indirectly informative about tree topologies. Other research has compared the accuracy of tree topology inference (Rasmussen et al., 2014; Kelleher et al., 2019) and recombination rates (Deng et al., 2021) among ARG inference methods. We opted to focus on coalescence times not only because they are a very informative statistic about evolutionary processes, but also because they can be fairly compared across all methods. As described in the Introduction, comparisons of tree topologies could be confounded by the presence of polytomies in ARGweaver and tsinfer+tsdate and the absence of polytomies in Relate.

We found a strong speed-accuracy trade-off in ARG inference. ARGweaver performs best in our three tests: point estimates, the overall distribution of coalescence times, and the quality of sampling from the posterior. Importantly, it is also the only method we compared that resamples both topologies and node times (Table 1.1). This likely leads to a better exploration of ARG space and is one reason why it provides better samples from the posterior. On the other hand, it also contributes to making ARGweaver much slower than the other methods and not scalable for genome-wide inference of 50 or more genomes.

Relate largely undersamples tree topologies (Deng et al., 2021), and thus every marginal tree estimate is only as good as an average over a series of true trees (Figure 1.1D). This will naturally lead to a more centered, under-dispersed distribution, as shown by the larger deviations from the uniform distribution in simulation-based calibration (Figures 1.5 and

1.8, where ARGweaver KLD values range from 0.008 to 0.350, and Relate range from 0.429 to 0.938). Despite not performing as well as ARGweaver in our evaluation criteria, Relate seems sufficient for comparisons of average trees across different regions in the genome.

Additionally, we showed that Relate's inferences generally improve with sample size (Figures 1.6, 1.7, 1.8). This is expected from inference using the Li and Stephens (2003) copy algorithm, which tends to better approximate the genealogical process with larger samples sizes (Hubisz et al., 2020). Because Relate is fast enough, even for thousands of samples, it is preferred for large numbers of genomes - not only because ARGweaver is not scalable for such large sample sizes but also because Relate inference tends to improve with larger sample sizes (Hubisz and Siepel, 2020).

The framework of tsinfer and tsdate is also based on the Li and Stephens (2003) model, and it additionally takes advantage of the succinct tree sequence data structure that makes it scalable to even larger sample sizes than Relate, and at least an order of magnitude larger than tested here (Wohns et al., 2022). Although we did not find an improvement of tsinfer+tsdate estimates with increasing sample sizes in the range we tested (4 to 200 haplotypes), our analyses cannot rule out the possibility of better tsinfer+tsdate inference at larger sample sizes.

Increasing the mutation to recombination ratio in simulations improved point estimates from ARGweaver but did not improve posterior calibration (Figure 1.5). This lack of improvement of the posterior sampling can be explained by lack of convergence and could potentially be improved by increasing the number of MCMC iterations. Although the statistics recorded by ARGweaver at each iteration (likelihood, number of recombinations, etc.) show convergence (Figure A.9, Table A.1), we observed that certain pairwise coalescence times did not converge in the simulations with increased mutation to recombination ratio (Table A.2 see more discussion in ARGweaver in Appendix of Chapter 1).

Limitations of our analyses and future directions

The focus of this study is the inference of coalescence times under the standard neutral coalescent, assuming all parameter values of this model are known and correctly provided to the programs performing inference. In other words, our goal was to investigate the performance of the ARG inference methods when the underlying assumptions are met. We have not explored how the methods perform under more complex demographic models and in the presence of natural selection, when the underlying assumptions are not met, but this is clearly an important future direction.

We also restrict our analyses to small sample sizes relative to what is possible for Relate and tsinfer+tsdate. However, increasing sample sizes up to 200 samples does not consistently improve performances of these methods. We also note that interesting discoveries have been made by applying ARG-based methods with similarly small sample sizes, *e.g.* Hubisz et al. (2020) analysed gene flow between archaic and modern humans using five genomes: two Neanderthals, one Denisovan and two modern humans.

Other factors not explored here could also be relevant for applications to real data. For example, sequencing or phasing errors could reduce the performance of all methods. Each of the methods compared here deal with these problems in a different way. Both Relate and tsinfer require phased data. While Speidel et al. (2019) argue that Relate is robust to errors in computational phasing, Kelleher et al. (2019) acknowledge that phasing errors could reduce the performance of tsinfer. ARGweaver is the only method of the three that supports unphased data, by integrating over all possible phases. However, the performance of the program on unphased data has not been evaluated in this study.

Relate takes sequencing errors into account by allowing some mutations that are incompatible with the tree topology in its tree building algorithm. Some robustness to error is shown in Speidel et al., Figure S3 (Speidel et al., 2019). Tsdate also uses heuristics in the ancestral haplotype reconstruction stage to increase its robustness to genotyping errors (Kelleher et al., 2019), and its newest version also accounts for recurrent mutation. ARGweaver can deal with genotyping errors statistically, using genotype likelihoods and integrating over all possible genotypes (Hubisz and Siepel, 2020). In addition, it can take into account local variation in coverage and mapping quality, all of which are features not tested here. ARGweaver can also incorporate a map of variable mutation rates. ARGweaver, Relate and tsinfer can all incorporate maps of variable recombination rates across the genome, a feature which was not used in our constant rate simulations.

In our standard simulations, we use mutation rate equal to recombination rate, which is believed to be approximately true for humans. In reality, even if average recombination and mutation rates are similar, the average recombination rate is not distributed equally along the genome in humans and other mammals but is concentrated in recombination hotspots. Therefore, it is possible that ARG inference could be more accurate with real data, since local trees could span longer sequences separated by recombination hotspots.

Recommendations for usage

Given that ARGweaver provides the most accurate coalescence times estimates and the most well-calibrated samples from the posterior distribution of coalescence times, we recommend using it whenever computationally feasible. However, it is highly computationally demanding and its usage can become unfeasible with sample sizes close to 100. Running ARGweaver on small segments of sequence (5Mb or 250kb Fig. A.8, A.7) gave similar results to applications on 100Mb segments, making the program highly parallelizable, at least for the purpose of estimating pairwise coalescence times.

When ARGweaver is computationally prohibitive, Relate and tsinfer+tsdate are viable alternative options. However, we emphasize that we have only examined coalescence time estimates, and for other downstream uses of ARG inference that do not rely mostly on coalescence times, the tradeoffs between these methods could be different. See Deng et al. (2021) (Deng et al., 2021) for a comparison of these methods in the context of estimating recombination rates.

Chapter 2

Estimating population split times and migration rates from historical effective population sizes

This chapter is co-authored by Vladimir Shchur, Anna Ilina and Rasmus Nielsen.

2.1 Introduction

Effective population size is one of the main characteristics of the demography of a population, and an important parameter determining evolutionary processes. Although it is related to census population sizes, it does not correspond to it in most cases. For example, for humans, effective population size is estimated to be on the order of ten thousand, which is much smaller than actual population size, in the order of billions (Henn et al., 2012).

Effective population size has been formally defined in many different ways. The concept was originally introduced by Wright (1931) in the context of describing causes of random variation in allele frequencies through time (*i.e.* genetic drift). In that context, effective population size is *"the number of individuals in a theoretically ideal population having the same magnitude of* ***random genetic drift*** *as the actual population"* (Hartl and Clark, 2007). Common models for this ideal population are standard Wright-Fisher (Fisher, 1930; Wright, 1931) or Moran (Moran, 1958) models. In an actual population, a combination of population size, inbreeding, unequal sex ratios and variance in offspring number all contribute to genetic drift, and thus to the effective population size. In population models where these parameter values are known, effective population size can be calculated directly using well-established equations (Hartl and Clark, 2007).

In practice, because effective population size measures genetic drift, and drift determines the amount of genetic diversity in the idealized population, effective population size is often interpreted as a measure of genetic diversity. The definition of effective population size then changes slightly to *the number of individuals in a theoretically ideal population having*

the same amount of ***genetic variation*** *as the actual population.* Under this definition, effective population size can be calculated directly from measures of genetic diversity of real populations. In a real population, however, the level of genetic diversity is affected by processes other than drift, such as migration and natural selection. When the effective population size is estimated from genetic diversity, it is affected by these other processes. In other words, when we define effective population size as a measure of genetic diversity, it no longer reflects only drift.

The Wright-Fisher and Moran population models mentioned earlier, as well as many others, converge to Kingman's coalescent model when population sizes are large (Kingman, 1982; Sjödin et al., 2005; Wakeley and Sargsyan, 2009). In this model, effective population size is a parameter determining the probability distribution for coalescence times (time to a common ancestor) of two lineages. More specifically, the **coalescent** effective population size is *the inverse of the coalescence rate.* Because coalescence times predict most measures of genetic diversity in a population, the coalescent effective population size has been argued to be the most general definition of effective population size (Sjödin et al., 2005), and it is the definition we will use here.

Methods to estimate historical effective population size

Many methods were developed to estimate the variation of effective population size through time. The most widely used one is PSMC (Li and Durbin, 2011), which uses a single diploid genome (or a pair of haploid genomes) to infer past effective population sizes. PSMC is based on a sequentially Markovian coalescent model (McVean and Cardin, 2005), where states (coalescence times, discretized) change along a DNA sequence. A Hidden Markov Model (HMM) approach is used to infer the distribution of coalescent times between a pair of lineages. The emission probabilities of this HMM are the probabilities of observing a site that is variable (heterozygous) or non-variable (homozygous), given a coalescence time. The underlying intuition is that the probability of observing a heterozygous site increases with coalescence time. The effective population size is then given by the inverse of the coalescence rate. There are many other HMM-based methods for inference of historical effective population size: coalHMM (Hobolth et al., 2007; Dutheil et al., 2009), diCal (Sheehan et al., 2013), MSMC (Schiffels and Durbin, 2014), SMC++ (Terhorst et al., 2017) (see (Spence et al., 2018) for a review).

Previous studies have shown how these and other methods can infer past effective population sizes that are very different from simulated census population sizes when populations are structured (Heller et al., 2013; Mazet et al., 2016; Chikhi et al., 2018). These studies showed how population structure can lead to misinterpretation of past effective population size plots. While it is true that PSMC plots can be often misinterpreted, we argue that the change in past effective population size due to population structure is an expected and desirable behaviour, since these methods infer the coalescent effective population size, and not the census population sizes.

Migration, for example, is a phenomenon that generally increases the coalescent effective population size of the population receiving migrants, since incoming migrants will likely increase genetic diversity. Interestingly, if we consider the alternative definition of effective population size as the change in allele frequency through time due to drift, the effect of migration on effective population size is the opposite: migration introduces sudden changes in allele frequency, which can be interpreted as strong drift, and thus small effective population size (Wang and Whitlock, 2003). This effect is expected only in the short term, and it is reversed in the longer term as populations approach an equilibrium (Wang and Whitlock, 2003). Here, however, we are focusing on the coalescent effective population size, which tends to increase with immigration. We emphasize that the effective population size inferred with PSMC should be interpreted as the amount of genetic variation in a population through time. Therefore, PSMC results are informative about both population size and migration. Nonetheless, as we will show, inferences of effective population size from PSMC can in some cases be biased when the transition probabilities of the HMM underlying PSMC inferences cannot adequately fit the true transitions in coalescence times along the genome in the presence of migration.

In this work we make an effort to disentangle the effects of migration from effective population sizes. We discuss the case of two fully exchangeable populations (*e.g.* Wright-Fisher populations) with migration between them. A sample from any of the two admixed populations contains footprints of historical effective population size of both parental populations. We formalize the concept of *local* effective population size, which is the effective population size of the parental populations, after accounting for the effect of migration. We show that the local effective population size can be determined from the the ordinary effective population size estimated by PSMC if migration rates are known or inferred.

We also develop a method called MiSTI (for "migration and split time inference"), which infers split time and migration rates under a model of two populations that exchange migrants after their split from a common ancestor. To do so, MiSTI combines information from the joint site frequency spectrum (SFS) of two diploid samples with the ordinary historical effective population sizes (as inferred by PSMC). MiSTI also uses the inferred migration rates to recover the local effective population size, *i.e.* to "correct" the PSMC curves for the effect of admixture. By applying this method to simulated data we show scenarios where PSMC finds a good approximation of the simulated effective population size, and scenarios where PSMC results do not correspond to the true effective population size. We also show that MiSTI appropriately corrects PSMC curves for the effect of migration, when migration rates are known and PSMC estimates are close to the true effective population size. Next, we apply MiSTI to data from humans and show i) How MiSTI can correct the effect of Neanderthal admixture on the historical effective population size of a human genome of European ancestry (CEU) and ii) What split times and migration rates best fit a model of split time and migration between pairs of human populations.

Other methods that infer population split times and/or migration rates from a pair of diploid genomes

SMC++ (Terhorst et al., 2017) is a method that infers historical effective population sizes, combining the coalescent HMM (similar to PSMC) with information from the SFS. When provided data from two populations, SMC++ can also jointly infer split times under a model of a clean split, *i.e.* without migration between populations after the split.

Song et al. (2017) fit an isolation-migration model to infer population split times from PSMC results. Their approach differs from ours methodologically and conceptually. In terms of methodology, Song et al. (2017) use an ABC approach to fit parameters, while we compute the composite likelihood of parameter values based on equations derived analytically. Conceptually, we formalize the distinction between the ordinary effective population size of admixed samples (often inflated by migration) and the local effective population size of its parental populations, which we can recover in the presence of migration, while Song et al. (2017) does not make this distinction.

Wang et al. (2020) developed a method, MSMC-IM, that also infers migration rates from historical effective population sizes. MSMC-IM fits an isolation-migration model with continuous symmetric migration to the inverse coalescence rates inferred by MSMC2. Instead of explicitly modelling a split time point, they model population split as a continuous process. The event of two populations merging backwards in time is represented as an increase in migration rates, to a point where both populations exchange migrants freely. In contrast to MiSTI, MSMC-IM does not aim to recover the local effective population size. Other differences worth pointing out are that MiSTI allows for asymmetric migration between populations and it uses PSMC instead of MSMC, which requires phased genomes.

Arredondo et al. (2021) use yet another approach. Their method, SNIF (Structured Non-stationary Inferential Framework) fits the curves of coalescent effective population size through time (which they denominate inverse instantaneous coalescence rate, IICR), as inferred by PSMC, to island models with symmetric migration and constant deme size. This method allows to infer the number of demes and migration rates among demes from IICR curves alone.

Schlebusch et al. (2017) introduced the TT-method (Sjödin et al., 2021), which can also infer population split times from two diploid genomes representing each of the two populations. The TT-method uses the joint site frequency spectrum of these two genomes to analytically calculate split times. It relies on two assumptions that are relaxed in MiSTI: 1) the effective population size of the ancestral population remains constant and 2) there is no migration between populations after the split. We compared MiSTI and TT-method inferences of split times between human populations, and we show through simulations that the first assumption of TT-method leads to large errors in the inferred split times for historical effective population sizes similar to those of human populations, even in the absence of migration.

2.2 Methods

Historical effective population size.

As previously discussed, effective population size can be defined as the average time to coalescence of two lineages, measured in number of generations (Wakeley and Sargsyan, 2009). Under the standard coalescence model with a single homogeneous population (Kingman, 1982), the interpretation is simple. For an effective population size $N \gg 1$ the rate of coalescence is $\lambda = 1/N$ per generation, and the expected waiting time to coalescence is $\lambda^{-1} = N$ generations. This definition can be naturally extended in order to define the historical effective population size. Consider the coalescence rate at time t, $\lambda(t)$, between the pairs of lineages from a population. The time $t = 0$ corresponds to the present and t increases toward the past. The coalescent rate $\lambda(t)$ determines an inhomogenous Poisson process which describes the distribution of coalescent times. Hence, the probability distribution of coalescent times T_c is

$$P(T_c = t) = \lambda(t)e^{-\int_0^t \lambda(s)ds}.$$

We define the inverse of $\lambda(t)$ as the ordinary *historical effective population size*

$$N(t) = \frac{1}{\lambda(t)}.$$

This quantity depends on population structure and demography, and it is a parameter which allows mapping of a real population with complex demography and structure on a single idealized population (e.g., a Wright-Fisher population) that is similar with respect to some property, as described in the Introduction. We note that this concept of historical effective population size is useful for interpreting the results of methods such as PSMC that allow inferences of varying effective population size through time (Li and Durbin, 2011; Spence et al., 2018). We will show that though for some scenarios PSMC indeed infers a good estimate of the ordinary effective population size, in other scenarios with migration, PSMC infers a biased estimate of effective population size.

Using standard population genetic theory, it is possible to explore the effect of population structure on effective population size (Mazet et al., 2016; Chikhi et al., 2018). Assume, for example, that an observed (modern) population S_m is formed by admixture of several parental populations. To determine its historical effective population size, we need to trace pairs of lineages from S_m back in time, until their coalescence. The lineages can switch between parental populations (Fig. 2.1), hence the genetic variation of S_m has a footprint from each of these populations.

We here develop this concept for the case of two parental populations with admixture (continuous or pulse). We denote the two parental populations by $S_1(t)$ and $S_2(t)$. At any time t, a lineage ancestral to the observed population S_m is either in population $S_1(t)$ or in population $S_2(t)$, due to migration. Within populations $S_1(t)$ and $S_2(t)$ lineages are fully exchangeable, which means that every pair of lineages from the same population has the same probability of coalescence. Effective population sizes of $S_1(t)$ and $S_2(t)$ are $N_{L1}(t)$

and $N_{L2}(t)$ respectively. $N_{L1}(t)$ and $N_{L2}(t)$ are what we define as *local effective population sizes*, i.e they represent the effective number of individuals in the populations at time t after discounting for the effect of migration. In other words, this is the rate of coalescence of two lineages conditional on both of them being in the given population.

If two lineages are in the same population $S_i(t)$ $(i = 1, 2)$ at time t, they can coalesce with rate $1/N_{Li}(t)$. If they are in different populations, coalescence is not possible between them. Conditional on two lineages having not coalesced by time t, let $P_1(t)$ and $P_2(t)$ be the probabilities that two lineages are in the population S_1 and population S_2 respectively. Let $P_0(t)$ be the probability that the two lineages are in different populations. Then the coalescence rate between a pair of lineages at time t is

$$\lambda(t) = P_1(t)\frac{1}{N_{L1}(t)} + P_2(t)\frac{1}{N_{L2}(t)} + P_0(t) \cdot 0, \tag{2.1}$$

and the *ordinary effective population size* is

$$N(t) = \frac{1}{\lambda(t)} = \frac{1}{P_1(t)\frac{1}{N_{L1}(t)} + P_2(t)\frac{1}{N_{L2}(t)}}. \tag{2.2}$$

The condition that the sampled population S_m is $S_1(0)$, is equivalent to setting the initial conditions of probabilities P_i $(i = 0, 1, 2)$ to

$$P_1(0) = 1, P_2(0) = P_0(0) = 0.$$

The dependence of P_i on the time of observation is natural, because probabilities of migration might change over time, and even if they are constant, the cumulative amount of migration changes over time.

So, as shown above there is a clear difference between the *local* effective population size ($N_{L1}(t)$ and $N_{L2}(t)$) of parental populations and the *ordinary* effective population size of an observed admixed population ($N(t)$). The estimates of effective population size obtained by PSMC and similar methods are estimates of the ordinary effective population size ($N(t)$) and not local effective population size ($N_L(t)$).

Continuous and pulse migration

Equation 2.2 defining the ordinary effective population size, depends on the probabilities $P_i(t)$ $(i = 1, 2, 0)$ of two lineages being in population i at time t, given they have not coalesced at t. These probabilities can be found by solving a set of differential equations. The past dynamics of two lineages is described by a coalescent model, which is a Markovian process going back in time. There are four possible states for this process:

- both lineages are in the first population at time t with probability $p_1(t)$,
- both lineages are in the second populations at time t with probability $p_2(t)$,

- the lineages are in different populations at time t with probability $p_0(t)$,
- the lineages have coalesced by time t with probability $p_c(t)$. This is an absorbing state.

Transitions between the first three states are possible through migration (either continuous or pulse). Transitions into the last absorbing state occur through coalescences, and $p_c(t) = 1 - p_1(t) - p_2(t) - p_0(t)$. By definition of conditional probabilities,

$$P_i(t) = \frac{p_i(t)}{1 - p_c(t)} = \frac{p_i(t)}{p_1(t) + p_2(t) + p_0(t)}, \tag{2.3}$$

for $i = 1, 2, 0$.

The continuous migration rate $m_{ij}(t)$ (for $i = 1, j = 2$ or $i = 2, j = 1$) is the rate with which a single lineage from population $S_i(t)$ moves into population $S_j(t)$, backwards in time. Considered forward in time, these migration rates correspond to the fraction of population S_i made of lineages from population S_j ($i \neq j$), when scaled in units of a reference effective population size $N_0 \gg 1$. This is the same definition as used in standard coalescence models with migration (Slatkin, 1982, 1987; Notohara, 1990; Wilkinson-Herbots, 1998) including Hudson's ms simulator (Hudson, 2002).

Henceforth, we omit the dependence of these functions on t in our notations to improve readability. From standard definitions of the coalescent with migration, we then have the following system of differential equations:

$$\begin{cases} p_1' = -\left(2m_{12} + \dfrac{1}{N_{L1}}\right) p_1 + m_{21} p_0, \\ p_2' = -\left(2m_{21} + \dfrac{1}{N_{L1}}\right) p_2 + m_{12} p_0, \\ p_0' = 2m_{12} p_1 + 2\mu_{21} p_2 - (m_{12} + m_{21}) p_0. \end{cases} \tag{2.4}$$

where p_i' indicates the derivative of p_i with respect to t.

Pulse migration acts instantaneously at time t_π. Backwards in time, it drags a lineage from one population to the other with a certain probability π. Forward in time, π is the proportion of a recipient population made up of individuals from the donor population due to pulse admixture. Assume that the donor population is population 1 and the recipient population is population 2. We write t_π^+ to indicate the time right before the pulse migration and t_π^- to indicate the time right after the pulse migration, forward in time (see Figure 2.1 for clarification). Then the probabilities p_i ($i = 1, 2, 0$) change as follows

$$\begin{cases} p_1(t_\pi^+) = p_1(t_\pi^-) + \pi^2 p_2(t_\pi^-) + \pi p_0(t_\pi^-), \\ p_2(t_\pi^+) = (1 - \pi)^2 p_2(t_\pi^-), \\ p_0(t_\pi^+) = (1 - \pi) p_0(t_\pi^-) + 2\pi(1 - \pi) p_2(t_\pi^-). \end{cases} \tag{2.5}$$

The parameter π is equivalent to the parameter $1 - p$ of the -es switch in Hudson's ms simulator (Hudson, 2002).

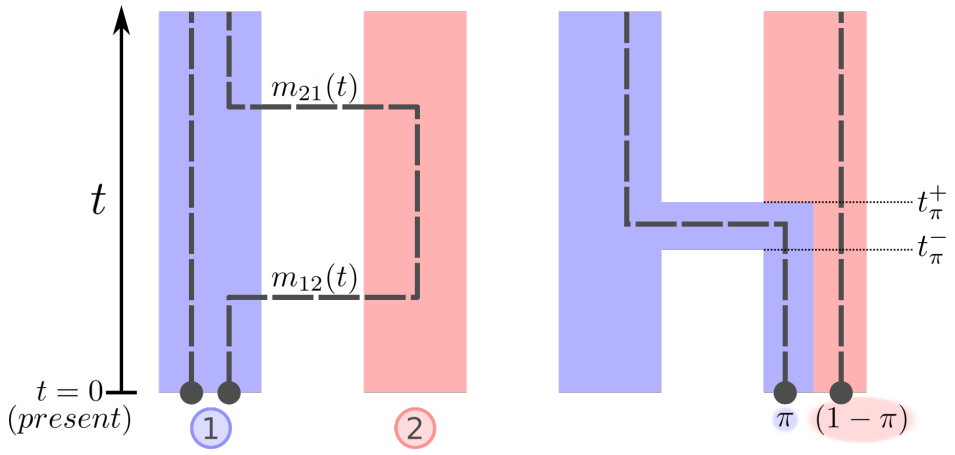

Figure 2.1: Notations for continuous migration (on the *left*) and pulse migration (on the *right*) models. In the continuous migration case, m_{ij} is the migration rate from population i to population j, backwards in time. In the pulse migration case, π is the probability of migration of a lineage, t_π is the instantaneous time of migration, t_π^+ and t_π^- are the times right before and right after the migration pulse.

Disentangling the effect of migration on effective population size.

Assume that we observe two populations $S_m^{(1)} = S_1(0)$ and $S_m^{(2)} = S_2(0)$, which had ancestral admixture with each other. Writing equation 2.2 for samples from both populations, we get the system of equations relating the *ordinary effective population size* of $S_m^{(1)}$ and $S_m^{(2)}$ (N_1 and N_2) with the *local effective population size* of each of the two parental populations (N_{L1} and N_{L2}).

$$\begin{cases} N_1(t) = \dfrac{1}{P_1^{(1)}(t)\frac{1}{N_{L1}(t)} + P_2^{(1)}(t)\frac{1}{N_{L2}(t)}}, \\ N_2(t) = \dfrac{1}{P_1^{(2)}(t)\frac{1}{N_{L1}(t)} + P_2^{(2)}(t)\frac{1}{N_{L2}(t)}}, \end{cases} \tag{2.6}$$

where $P_i^{(j)}$ is the probability that both ancestral lineages from population $S_m^{(j)}$ are in population i (see equation 2.3). These functions can be derived from equation 2.2 by setting initial conditions to $P_1^{(1)}(0) = 1$ for the first populations and $P_2^{(2)}(0) = 1$ for the second population.

As we already mentioned, PSMC or similar methods can be used to estimate the ordinary effective population size, $N(t)$. Given samples from two admixed populations, the underlying local effective population size (N_{L1} and N_{L2}) of their parental populations $S_1(t)$ and $S_2(t)$ can be estimated from equation 2.6.

Unfortunately, there is no closed form solution of equations 2.4 and 2.6. PSMC approximates historical effective population size with a piece-wise constant function. In our

inference, we assume that migration rates are constant in each time interval, and similarly to PSMC, we approximate local effective population sizes with a piece-wise constant trajectory. So, instead of solving equation 2.6, we calculate piece-wise constant functions $N_{L1}(t)$ and $N_{L1}(t)$ such that the probabilities to coalesce within each time interval is the same as inferred by PSMC. In more details, for two lineages from population i the probability of coalescence $\hat{p}_c^i$ within $[t_1, t_2]$ inferred by PSMC is

$$\hat{p}_c^i = p_{nc}\left(1 - e^{-\frac{t_2 - t_1}{N_i}}\right),$$

where p_{nc} is the probability that two lineages have not coalesced by time t_1.

From equation 2.4 the probability p_c^i that two lineages from population i coalesce within the interval $[t_1, t_2]$ is

$$p_c^i = p_c(t_2) - p_c(t_1).$$

And we fit N_{L1} and N_{L2} so that

$$p_c^i = \hat{p}_c^i.$$

Estimating migration rates and split time

In the previous subsection we show how one can calculate local effective population sizes for given values of migration rates and split time. Of course, it is also desirable to estimate these parameters, because they are often unknown. Our method fits the joint site frequency spectrum (SFS) of two diploid individuals representing two populations. Let $f_{i,j}$ $(i, j = 0, 1, 2)$ be the probability that a variable site has i derived alleles in the first individual and j derived alleles in the second individual. Notice that $f_{0,0}$ and $f_{2,2}$ are excluded because they correspond to non-variable sites. Then the probabilities $f_{i,j}$ define a multinomial distribution.

Let $\mathbf{n} = \{n_{0,1}, n_{1,0}, n_{1,1}, n_{1,2}, n_{2,1}\}$ be the site frequency spectrum from the data, i.e. $n_{i,j}$ is the non-normalised counts of sites with i and j derived alleles in the first and second individual, respectively. We consider the composite likelihood function (which ignores possible correlations between the sites) given by the multinomial distribution:

$$\mathcal{L}(\mathrm{SFS}|\mathbf{n}) = \hat{\prod}_{i,j} f_{i,j}^{n_{i,j}}.$$

The theoretical SFS for a given set of parameters and PSMC trajectories can be computed by numerically solving a set of linear differential equations describing a Markov process with 44 states. These states describe all possible ways in which lineages of a coalescent tree with four tips (two samples from each of two populations) can be distributed among populations. These states include the possibility of coalescence between lineages and migration between populations through time (details are given in the Appendix B.1).

Software implementation

Our method for estimating underlying local effective population size is implemented in Python 3 under the name MiSTI. The implementation is available at `https://github.com/vlshchur/MiSTI` and distributed under GNU GPL3.

2.3 Results

Obtaining local effective population size from the ordinary effective population size estimated by PSMC

In this section we demonstrate the effect of migration on effective population sizes, and we will qualitatively assess the PSMC inference of historical effective population size trajectories.

We used ms (Hudson, 2002) to simulate two population size trajectories: one trajectory (population 1) has constant size through time after the population split, and the second trajectory (population 2) has a bottleneck after the split, followed by a recent population expansion. Using these population size trajectories, we simulated symmetric migration between populations, as well as unidirectional migration. For each simulation, we show: 1) the ordinary effective population size of both populations (N_1 and N_2) calculated using Equation 2.6, 2) the ordinary effective population size estimated using PSMC ($\hat{N}_1$ and $\hat{N}_2$), and 3) the local effective population size ($\hat{N}_{L1}$ and $\hat{N}_{L2}$) obtained with MiSTI by correcting the effective population size trajectories for the effect of migration.

Continuous, bidirectional migration between populations 1 and 2 from the present until the split time generally increases the ordinary effective population size (N_1 and N_2) relative to the simulated local population size (N_{L1} and N_{L2}, Figure 2.2A). However, notice that population 1 has a decreased effective population size relative to its simulated local size, during the population 2 bottleneck (Figure 2.2A). This decrease in genetic variation observed in population 1 is caused by the possibility that lineages from population 1 go through the bottleneck in population 2, where coalescence rates are increased.

In this scenario, PSMC generally estimates the ordinary effective population size (Figure 2.2C) well, despite the smoothing of instantaneous populations size changes that has been described previously (Li and Durbin, 2011). We also note that in Figure 2.2A, the historical effective population size of populations 1 and 2 coincide before the bottleneck, but PSMC trajectories do not coincide (Figure 2.2C). We discuss this possible bias in PSMC below.

Continuous, unidirectional migration from population 1 to 2, generates an increase in the ordinary effective population size of population 2 (Figure 2.2B), which is detected by PSMC (Figure 2.2D). In this scenario, the inferred PSMC trajectory underestimates effective population size of population 2 (the one receiving migrants) during the bottleneck, and overestimates it before the bottleneck. We hypothesize that this effect, as well as the discordance of PSMC curves prior to the bottleneck in Figure 2.2C, could be due to the violation

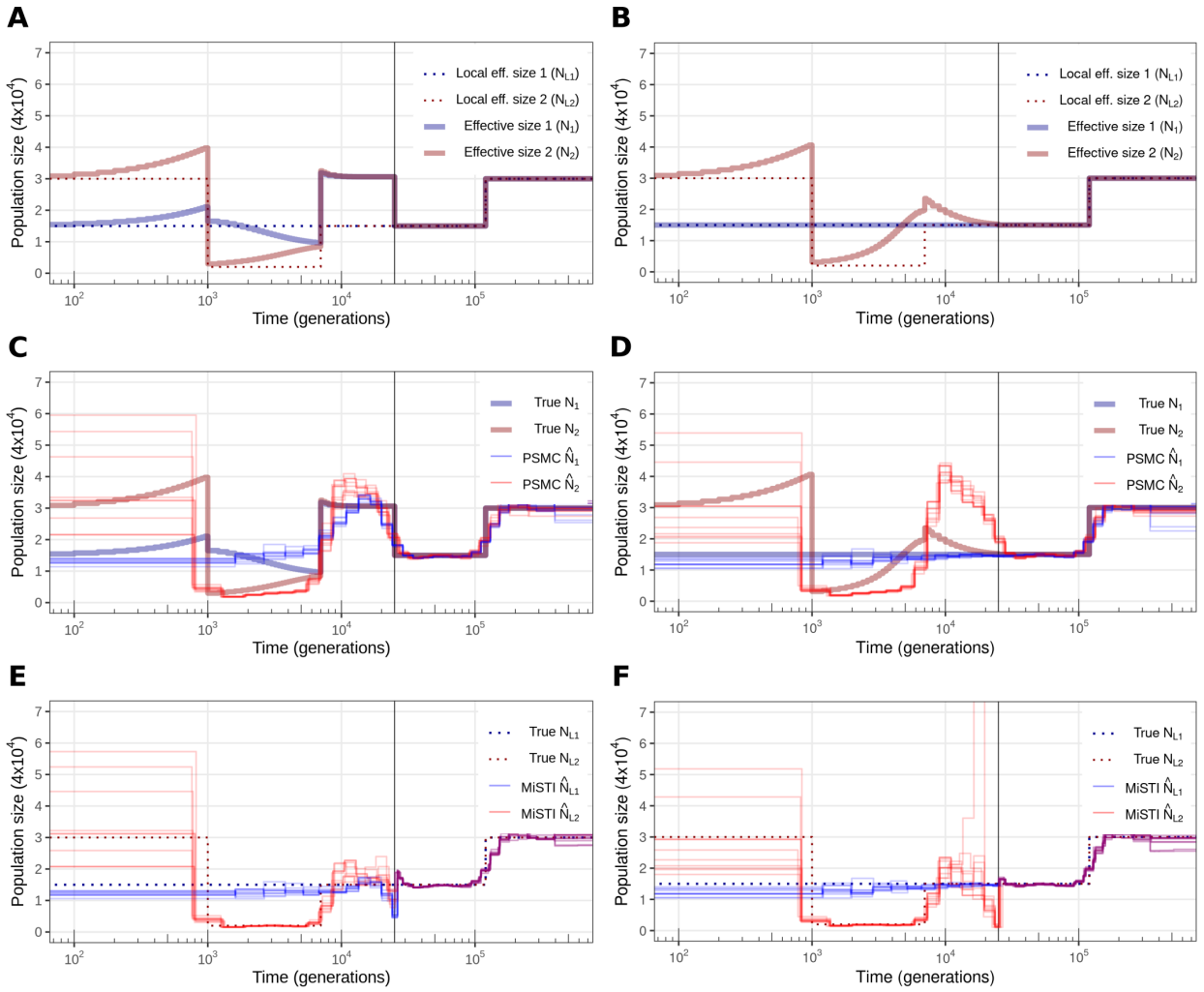

Figure 2.2: Continuous migration (ms parameter $M_{ij} = 2$) from the present to split time (25000 generations, indicated by vertical bar). (A,C,E) Bidirectional migration. (B,D,F) Uni-directional migration from population 1 to population 2. (A,B) Simulated local effective population sizes and ordinary effective population sizes calculated according to equation 2.6. (C,D) True ordinary effective population size from A,B, and estimated by PSMC. (E,F) True local population sizes from A,B, and estimated by MiSTI.

of the SMC model assumption that samples come from a single panmictic population, as we explore further in the next section.

Applying MiSTI correction of PSMC curves with the known migration rates and split times used in the simulations, to estimate local population sizes (N_{L1} and N_{L2}), recovers trajectories similar to the simulated ones (Figures 2.2E,F).

Pulse migration also increases historical effective population size. A single pulse of migration at time zero will cause effective population size to increase monotonically backwards in time until the time when populations split (Figure 2.3A,B). Similar to the continuous migration case discussed above, PSMC detects this increase in historical effective population size, although it underestimates the extreme peak of effective population size preceding the population split (Figure 2.3C,D). This underestimation is due to an smoothing effect of the PSMC method, which has been described (Li and Durbin, 2011). MiSTI recovers the local effective sizes of populations 1 and 2, slightly underestimating it when the PSMC smoothing underestimated the peak in effective population size (Figure 2.3E,F).

Transition matrices and the assumption of a single panmictic population in PSMC

PSMC assumes a model of a single panmictic population. When data come from a structured population, PSMC can often find a best fitting transition matrix that has a stationary distribution equivalent to that of a single population with the same effective population size (Chikhi et al., 2018). However, in the previous section, we showed one example where the ordinary effective population size inferred by PSMC was strongly biased ("population 2" in Figure 2.2D). This reveals that the stationary distribution of the transition matrix fitted by the HMM underlying PSMC is different from the true distribution.

To investigate this bias in PSMC, we simulated a single panmictic population with the same effective population size as population 2 (*i.e.* following the trajectory of N_2 in Figure 2.2D). In other words, this population has the same stationary distribution of the coalescence time transition matrix as population 2, but it did not receive any migrants. Let us call this population P, for panmictic. We found that the empirical transition matrix of population P (Figure 2.4C) differs more from the empirical matrix of population 2 (Figure 2.4E) than the transition matrix inferred by PSMC (Figure 2.4D). This indicates that the matrix inferred by PSMC is a better fit of the empirical matrix, and therefore the PSMC bias we detected is not due to an optimization problem.

Next, we compare the empirical transition matrix from population 2 (Figure 2.4A) to the transition matrix inferred by PSMC (Figure 2.4B). One difference between those matrices is that the PSMC matrix (Figure 2.4B) shows mostly vertical bands, while the true transition matrix (Figure 2.4A) shows a horizontal band corresponding to the time between the bottleneck and the split of populations (7-25k generations).

The horizontal band in the true transition matrix in Figure 2.4A is caused by a correlation in coalescence times between adjacent sites that can not arise in a panmictic model, assumed by PSMC. If two lineages coalescence during the bottleneck at a site, then there is an increased probability that both these lineages are in the bottlenecked population at other sites. Furthermore, if recombination happens during the bottleneck period, and both lineages are in the bottlenecked population, then there is an increased probability that the two lineages again will coalesce in this time interval in the next site after recombination. This

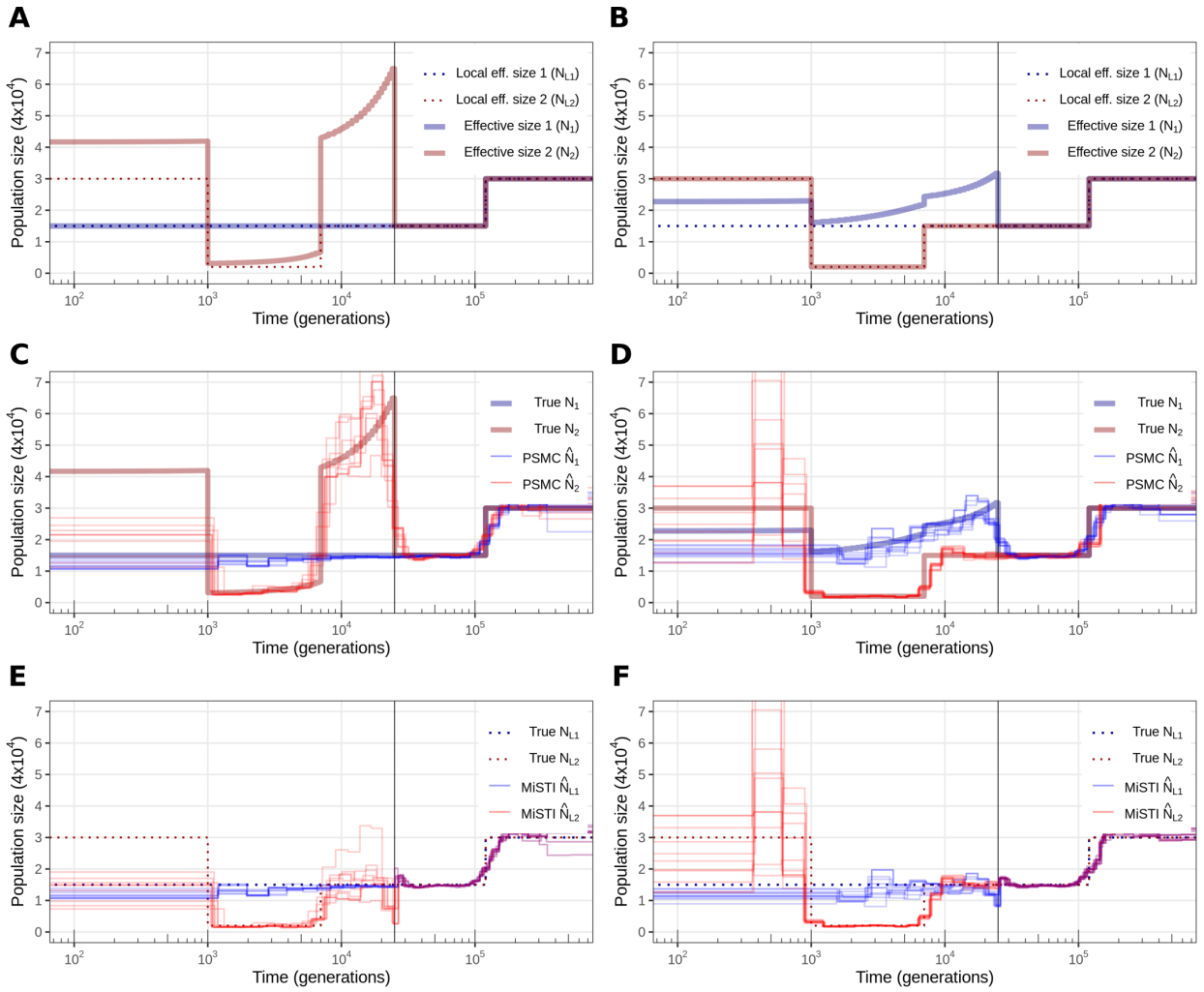

Figure 2.3: Pulse migration of 20% at the present time. (A,C,E) pulse from population 1 to population 2. (B,D,F) pulse from population 2 to population 1, forward in time. (A,B) Simulated local effective population sizes and ordinary effective population sizes calculated according to equation 2.6. (C,D) True ordinary effective population size from A,B, and estimated by PSMC. (E,F) True local population sizes from A,B, and estimated by MiSTI.

contrasts with a standard SMC coalescence model in a panmictic population, in which the time of coalescence in site $i+1$ is independent of the coalescence time in site i, conditional on it being older than the time of recombination between the sites. A structured model with a bottleneck, therefore, creates a correlation structure that cannot be modeled by the standard SMC model used in PSMC.

The bias in PSMC can therefore be explained by the fact that PSMC fits the transition

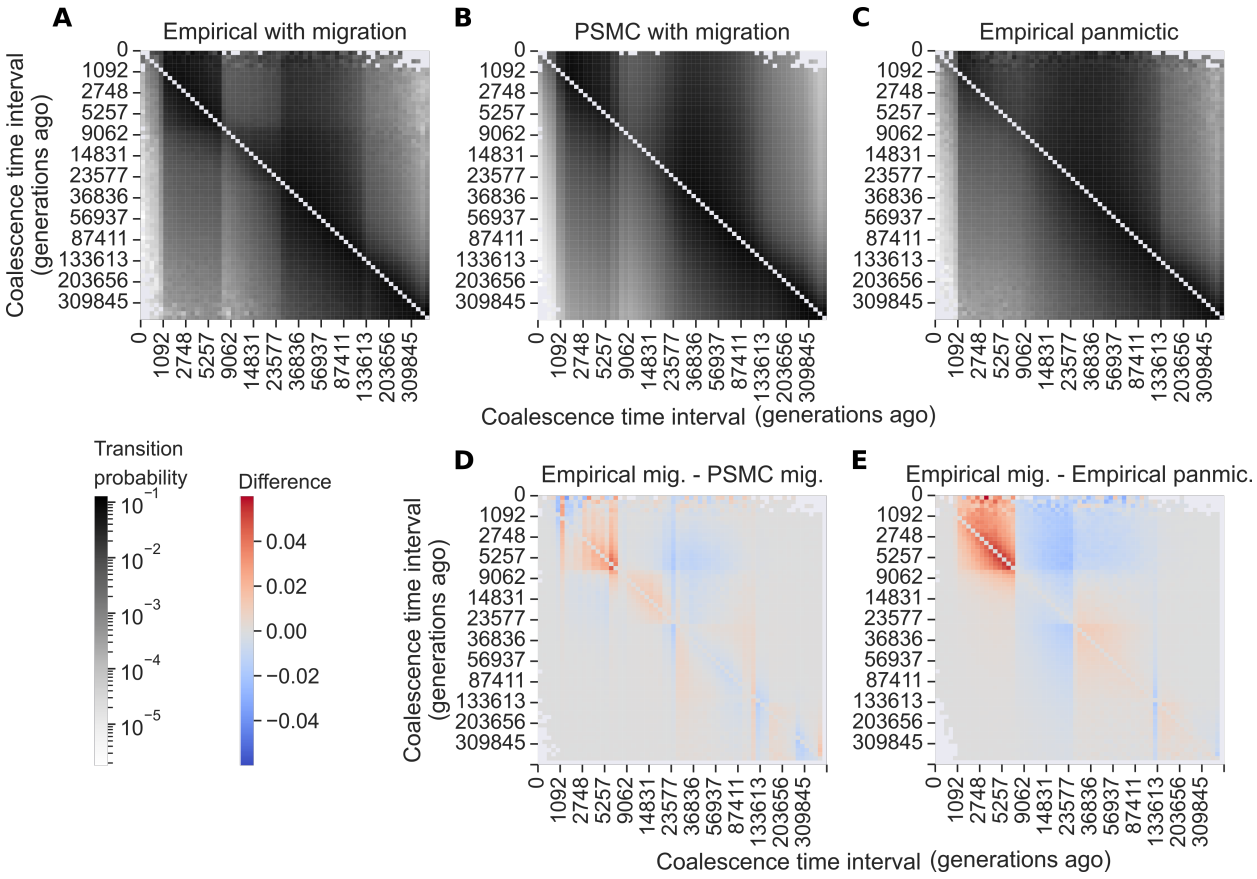

Figure 2.4: Matrices representing the (empirical or estimated) probability of transitions from one coalescence time to another along a sequence. (A) Transition matrix from simulated data from a population receiving migrants (population two of Figure 2.2B). (B) Transition matrix estimated by PSMC. (C) Transition matrix from simulated data from a single population (no migration) with historical effective population size equal to population two of Figure 2.2B. (D) Difference between matrices A and B. (E) Difference between matrices A and C.

matrix, and not its stationary distribution. Importantly, it fits the transition matrix of a panmictic model. As we mentioned previously, in many cases, fitting the best panmictic transition matrix also fits the best stationary distribution, but in this case, the best fit of a panmictic transition matrix by PSMC leads to a very different stationary distribution.

We also note that the single population model assumed by PSMC differs from the structured model with migration in the way that recombination rate scales with effective population size through time. In a single population model, an increase in effective population size (N) increases the effective recombination rate ($\rho = 4Nr$, where r is the rate of recombination per locus per generation). In a model with two populations, an increase in

effective population size that is due to migration would decrease the effective recombination rate, since two lineages can only recombine when they are in the same population. In other words, in a model with a single population, the effective recombination rate scales with the effective population size, while it is not necessarily true in a model with two populations and migration, where it can in fact scale inversely with effective population size.

Correcting European effective population size for the Neanderthal component

In this section, we show an application of the MiSTI correction of PSMC curves using known split time and admixture rates.

Non-African human populations admixed with Neanderthals 52-58 thousand years ago (Prüfer et al., 2014). Villanea and Schraiber (2019) recently reported that it is likely that there were multiple admixture events, but for simplicity we consider the case of a single pulse admixture event. This admixture would result in a number of very old coalescences which would increase the overall estimate of effective population size. In order to estimate the local effective population sizes of non-African populations, we need to correct for this admixture. One way of doing that is to call and mask the Neanderthal introgressed regions in a modern genome before running PSMC. Alternatively, one can estimate the proportion of Neanderthal ancestry in a modern genome, and use MiSTI to correct its PSMC trajectory for that admixture proportion. We compare these two approaches to verify that MiSTI's correction of effective population size is consistent with masking of known tracts of introgression.

We removed the Neanderthal tracts in an European genome (CEU population) and confirmed that the PSMC trajectory inferred from the Neanderthal-masked genome has lower effective population sizes than the non-masked, original genome (Figure 2.5). Next, we used MiSTI to correct the PSMC trajectories of the non-masked European genomes assuming 1.5% Neanderthal introgression, which has been reported by Steinrücken et al. (2018), and 3.0%, which was the previously reported estimate (Green et al., 2010) (Figure 2.5). Correcting the CEU PSMC trajectory for 1.5% Neanderthal admixture using MiSTI gives very similar estimates of local effective population size as masking the known regions of Neanderthal ancestry from that same genome. This suggests that MiSTI, at least in this case, correctly recovers the effective population size of parental populations, when applied to real data.

In the Appendix B.2, we show another application of MiSTI to obtain local effective population size by correcting PSMC curves for the effect of migration, using known parameters from *Puma concolor* populations, and we discuss limitations of applying MiSTI to that case.

Estimating split time in human-like simulations

Most often, split times and migration rates are unknown, and MiSTI can be used to estimate these parameters from PSMC curves combined with a joint (2D) site frequency spectrum

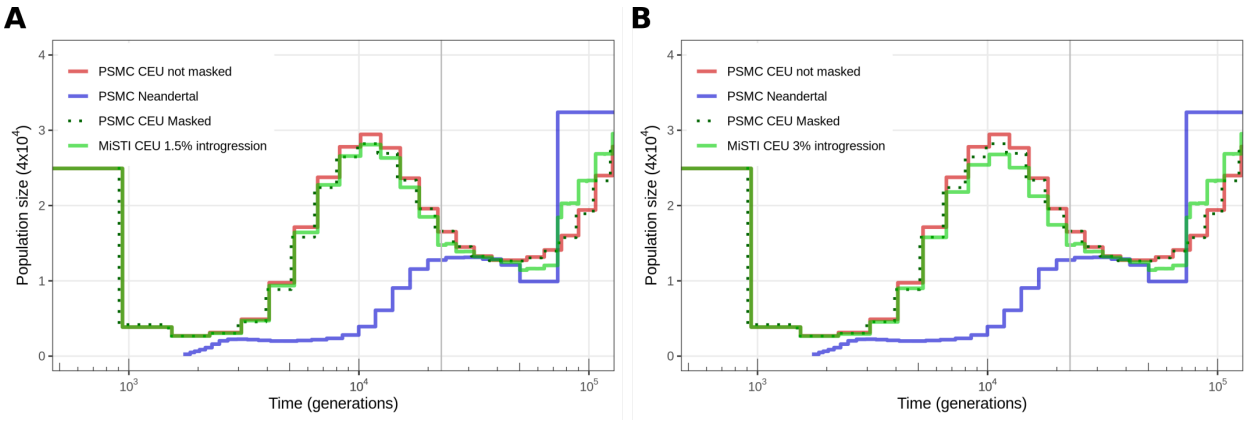

Figure 2.5: (A) MiSTI correction of PSMC effective population size trajectory of a genome from the CEU population assuming 1.5% introgression and (B) 3% introgression from Neanderthals.

for the pair of samples. In this section, we estimate split times from simulations replicating effective population size trajectories similar to those of human populations.

We simulated populations approximating the historic effective population size of human populations (Dinka, San, Sardinian, French and Han) (see "Simulations" in Appendix B.1). Briefly, we simulated historic effective population size similar to the estimated by PSMC for each of those populations (Figures 2.6A,2.7A,2.8A). We simulated population splits at various times, with no migration following the split. We then estimate these split times using MiSTI and the TT-method (Sjödin et al., 2021) (Figures 2.6B, 2.7B, 2.8B). We found that the TT method estimates negative split time in simulations where the split time happens during or immediately at the end of the bottleneck (Figure 2.8B), as has been previously described (Sjödin et al., 2021). In other scenarios of intermediate split times, the TT method largely overestimates the split times, due to violations of the assumption of constant effective population sizes in the ancestral population (Sjödin et al., 2021). In contrast, MiSTI provides substantially less biased estimates.

Estimating split time and migration rates from human data

Here, we estimate split times and migration rates from real data from the same populations we simulated in the previous section. We used MiSTI to estimate split-times and migration rates between the Han Chinese and French (Table 2.1), Dinka and Sardinian (Table 2.3), San and Dinka (Table 2.2) and San and Sardinian (Table 2.4) populations. From MiSTI, we recorded the maximum composite likelihood values of three models: no migration, unidirectional migration in each direction, and bidirectional migration. In all models with migration, we assumed a constant rate of migration between the split time and the present.

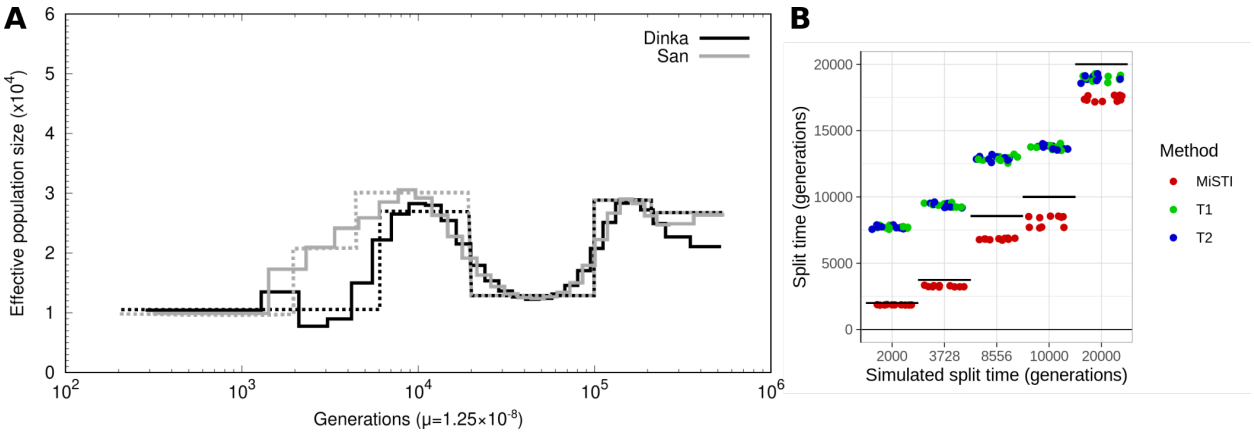

Figure 2.6: Simulations of San-Dinka split, no migration. (A) Continuous lines show effective population size inferred by PSMC from real data, dotted lines show simulated population sizes that approximate the inferred trajectory. (B) Inferences from MiSTI and the TT method for ten replicate simulations of each split time. 3728 generations was the split time inferred by MiSTI from real data; 8556 was the split time inferred by the TT method (see Table 2.2) - note that when we simulate 3728 generations, the TT method infers close to 8556 generations.

For Han-French divergence, the model with the highest composite likelihood was one with a split time of 1505 generations (*i.e.* 43,645 years ago assuming 29 years per generation) and a mostly unidirectional migration rate of 2.92 from Han to French (Table 2.1). We also replicate the results from Sjödin et al. (2021), in which the TT method infers nonsensical negative split times between Han and French. The unidirectional migration inferred from Han to French is in line with current models of the peopling of Europe through waves of farmers coming from central Eurasia (Haak et al., 2015).

The best fit model for the San-Dinka population pair includes a split time of 3729 generations ago (*i.e.* 108,141 years ago assuming 29 years per generation), and mostly unidirectional migration from Dinka to San. For the same data, the TT method infers a much larger split time (over 8500 generations ago) (Table 2.2, see also Appendix B.3 for a validation of this result with simulations).

The Dinka-Sardinian split time inferred by MiSTI is approx. 3963 generations ago, with bidirectional migration between these populations. The migration rate detected from Dinka to Sardinian is in line with previous results indicating migration from sub-Saharan Africa to South Europe (Moorjani et al., 2011). In this case, in contrast to the previous case, the TT method infers a more recent split time than MiSTI (2550 generations ago, see Table 2.3). The split time between San and Sardinian is older (approx. 4484 generations ago, see Table 2.4). We notice that these estimates are not strictly compatible with a population tree, which likely is a consequence of complex ancestral population structure and migration

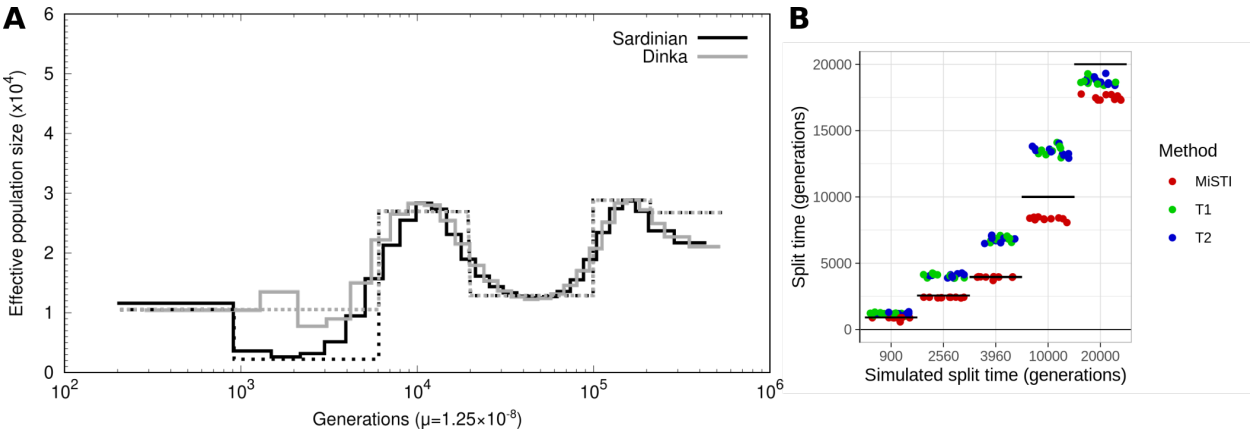

Figure 2.7: Simulations of Dinka-Sardinian split, no migration. (A) Continuous lines show effective population size inferred by PSMC from real data, dotted lines show simulated population sizes that approximate the inferred trajectory. (B) Inferences from MiSTI and the TT method for ten replicate simulations of each split time. 3960 generations was the split time inferred by MiSTI from real data; 2560 was the split time inferred by the TT method (see Table 2.3).

Table 2.1: MiSTI estimates of split times and migration rates between the Han Chinese and French populations in models with bidirectional migration (top row), unidirectional migration, or no migration (bottom row).

MiSTI				TT
m1	m2	split time		split time
French to Han	Han to French	(generations)	log(lik)	(generations)
$8.37 \times 10^{-}9$	2.92	1505	-2331	-
-	2.92	1505	-2331	-
3.84	-	1505	-2373	-
-	-	1505	-2787	T1 = -3587 T2 = -3545

between populations that is not modeled here, including archaic admixture into Sardinians. We note that archaic admixture will tend to inflate divergence time estimates, so the true divergence times might be smaller than our estimates, particularly for the splits between Sardinians and the two African populations.

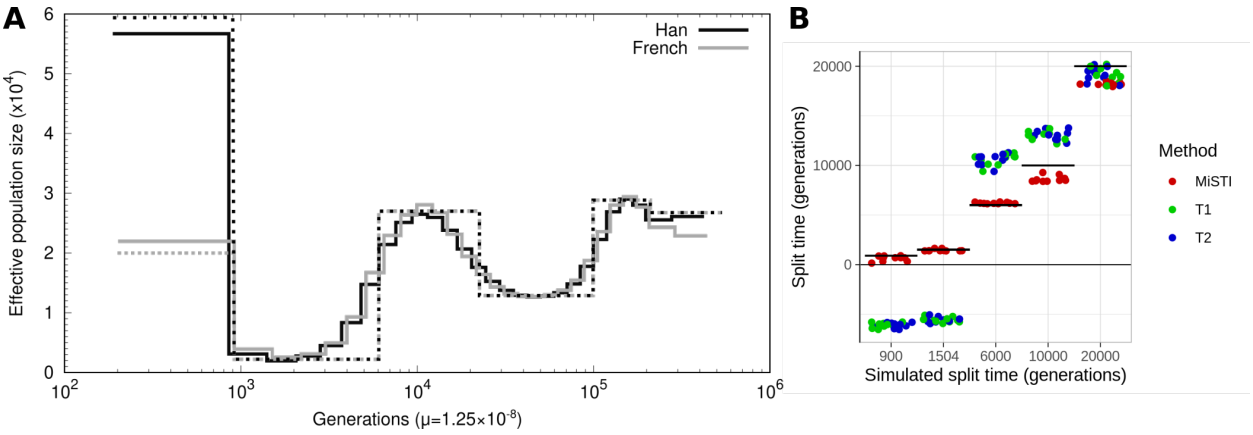

Figure 2.8: Simulations of Han-French split, no migration. (A) Continuous lines show effective population size inferred by PSMC from real data, dotted lines show simulated population sizes that approximate the inferred trajectory. (B) Inferences from MiSTI and the TT method for ten replicate simulations of each split time. 1504 generations was the split time inferred by MiSTI from real data; -3566 was the split time inferred by the TT method (see Table 2.1).

Table 2.2: MiSTI estimates of split times and migration rates between the San and Dinka populations in models with bidirectional migration (top row), unidirectional migration, or no migration (bottom row).

MiSTI				TT
m1 Dinka to San	m2 San to Dinka	split time (generations)	log(lik)	split time (generations)
2.5	2.03×10^{-9}	3729	-4381	-
2.5	-	3729	-4381	-
-	1.49	3210	-4582	-
-	-	3001	-4607	T1 = 8582 T2 = 8527

2.4 Discussion

The MiSTI method

The coalescent effective population size, defined as the reciprocal of the coalescence rate, is proportional to the census population size in a panmictic model, but can be very different from it when there is migration. The idea of disentangling the effect of migration on effective population size has been explored before. For example, (Wang and Whitlock,

Table 2.3: MiSTI estimates of split times and migration rates between the Dinka and Sardinian populations in models with bidirectional migration (top row), unidirectional migration, or no migration (bottom row).

MiSTI				TT
m1	m2	split time		split time
Sardininan to Dinka	Dinka to Sardinian	(generations)	log(lik)	(generations)
2.63	9.96	3963	-5337	-
6.92	-	3484	-5819	-
-	14.40	2286	-10877	-
-	-	2264	-13166	T1 = 2529 T2 = 2577

Table 2.4: MiSTI estimates of split times and migration rates between the San and Sardinian populations in models with bidirectional migration (top row), unidirectional migration, or no migration (bottom row).

MiSTI				TT
m1	m2	split time		split time
Sardininan to San	San to Sardinian	(generations)	log(lik)	(generations)
1.68	7.2×10^{-9}	4484	-3377	-
1.22	-	3963	-3497	-
-	1.59	3484	-4359	-
-	-	3483	-4604	T1 = 8269 T2 = 8253

2003) introduced methods to jointly estimate the local effective population size and migration rates from samples taken over time and space. Here, we are motivated by the same idea of disentangling migration and effective population size, and we do so in the context of inferring changes in effective population size through time from present-day samples of different populations, with methods such as PSMC (Li and Durbin, 2011).

We defined the ordinary effective population size of an admixed population as a function of the local effective population size of its parental populations. The local effective population size corresponds to the effective population size of unadmixed individuals from the parental populations. We developed a method, MiSTI, that uses the ordinary effective population sizes (*e.g.* estimated by PSMC (Li and Durbin, 2011)) of two samples from different populations that exchanged migrants, together with their joint SFS, to estimate the local effective population sizes and migration rates.

We note that MiSTI depends on the results of PSMC, and will be subject to its biases. MiSTI relies on a model of a population split possibly followed by migration, which itself

violates the assumption of panmixia made in PSMC and similar methods. We have shown that PSMC estimates are particularly sensitive to this violation when there is asymmetric migration (Figure 2.2D). In other scenarios, we (Figures 2.2C,2.3C,D) and others (Chikhi et al., 2018) have shown that PSMC provides good estimates of historical effective population size in the presence of population structure. Other studies have shown, using simulations, that estimates from PSMC and other methods can be biased even in the absence of population structure (Spence et al., 2018). We have not extensively explored those biases here, but we note that as less biased methods are developed, MiSTI can be adapted to use those and thus improve its inference of local effective population size, split times and migration rates.

When applied to infer population split times and migration rates in human populations, MiSTI helps settle a previous controversy. Schlebusch et al. (2017) found surprisingly deep divergence times for some Southern African populations, including the San. Their estimate for the split time between Dinka and San is 255 ± 5 years ago (Figure 3C in (Schlebusch et al., 2017)). We applied the TT method used in Schlebusch et al. (2017) (Schlebusch et al., 2017) to estimate the San-Dinka split time in our data and we found a similar result (split time 8554 generations ago, or 248 thousand years with 29 years per generation, Table 2.2), replicating their results. However, with MiSTI, we estimate a much more recent split time around 3729 generations ago (108 thousand years ago with 29 years per generation, Table 2.2). Our estimate is similar to the estimates of the earliest population divergence among modern human populations obtained with methods such as MSMC (Pagani et al., 2016; Fan et al., 2019), and momi2 (Kamm et al., 2020) (see (Bergström et al., 2021) Figure 2C for a synthesis of estimates from various studies).

The TT method makes a strong assumption that there are no changes in population size in the ancestral population, before the population split (Schlebusch et al., 2017; Sjödin et al., 2021). We simulated historical effective population size similar to the one estimated by PSMC for San-Dinka, and showed that the TT method strongly overestimates split times in this scenario (Figure 2.6). Notably, when we simulated data assuming a split time of 3728 generations ago (as inferred by MiSTI), the TT method estimated a split time close to the one it estimated from the real data (8556 generations ago), showing that the previously reported deep split time estimated by the TT method is in fact likely an estimation artifact. The TT method can be highly biased because of the assumption of constant population size and should not be applied to populations that may have experienced changes in effective population size over time.

The application of MiSTI to human data also illustrates the importance of including migration a the model is used to infer split times. In all cases (Tables 2.1-2.4), composite likelihoods were higher in models that allowed migration, and a difference of 1000 or more generations is seen in some split times inferred with models that include migration (Tables 2.3-2.4). Inferring asymmetric migration is also an interesting feature of MiSTI aimed at determining the direction of gene-flow.

Why estimate local effective population size?

Finally, we would like to highlight the broader relevance of disentangling the effect of migration on effective population size. The ordinary effective population size is important for understanding patterns of neutral genetic variability. It is a good predictor of summary statistics of neutral genetic variation such as the expected heterozygosity and average number of pairwise differences. However, for questions related to the efficacy of selection or genetic drift, the effective population size defined in terms of the number of individuals and their variance in offspring number is what matters most, not the effective population size inflated by migration. Since the ordinary effective population size is generally increased by migration, recovering the local effective population size after accounting for the effect of migration will recover values that are more informative for selection dynamics and predictions regarding the efficacy of selection, such as the rate of purging of deleterious alleles.

Local effective population size is also often more relevant for conservation genetics than the ordinary effective population size, which reflects overall genetic diversity of a meta-population. For example, a meta-population of an endangered species which occupies a fragmented habitat might have increased effective population size if considered as a whole. Their apparent high levels of neutral genetic diversity might be misleading regarding their fragile conservation status. If there is weak migration between isolated subpopulations, the ordinary effective population size of each subpopulation will be inflated by migration and will not be representative of the actual size of the local population. Correcting for the effect of population structure and migration provides measures of local effective population size that are closer to the effective number of breeding individuals in the population and thus more informative for conservation efforts.

Chapter 3

Genetic ancestry and signatures of natural selection in the people from Atahualpa village, Santa Elena, Ecuador

This chapter is co-authored by Oscar Del Brutto and Rasmus Nielsen.

3.1 Introduction

The region of Santa Elena province in Ecuador has long been inhabited by humans. Its location at the extreme West of South America suggests that it could have been home to the first humans of South America, since it is likely that the first people to invade the continent arrived using a Pacific coastal route (Dillehay et al., 2008). Early evidence of human settlements in the extreme Southern tip of South America supports the Pacific coastal route hypothesis (Dillehay et al., 2008). More concretely, archaeological studies in Santa Elena revealed a rich history of human life in the region with evidence of plant (squash) domestication as early as 11k years ago and the presence of diverse cultures, including Las Vegas (8500-4600 B.C.E.) (Raymond, 2008), Valdivia (4400-1450 cal B.C.E.), Machalilla (1430-830 cal B.C.E.) and Chorrera (1300-300 cal B.C.E.) (Zeidler, 2008). Later, the Manteño and Huancavilca cultures became predominant in the North and South (respectively), until the expansion of the Inca empire towards the region (around 1470 C.E.) (McEwan and Delgado-Espinoza, 2008). Most recently, in 1532, Spanish people arrived to Ecuador and spread through the country.

Atahualpa is a rural village located in Santa Elena. There is historical evidence that this village has been settled in the same area since before the Spanish arrival, and there is little migration to or from the village, which suggests its inhabitants are likely to have a large proportion of indigenous ancestry (Del Brutto and Zambrano, 2017). The 2010 Census

reports 3532 inhabitants, 90.3% of whom self report as Mestizos, 4.9% Afro-Ecuatorians, 1.3% Montubios, 1.3% White, 0.5% Indigenous and 1.7% as other categories. The proportion of self-reported Mestizos in Atahualpa is higher than in the country as a whole, where 71.9% of the population identify as Mestizos. Although the Mestizo ethnicity suggests admixture, previous studies have shown that people who are ethnically Mestizos in Ecuador can have a high proportion of indigenous genetic ancestry (Nagar et al., 2021).

People from the Atahualpa village consume high amounts of oily fish as part of their traditional diet, and oily fish intake in this population has been associated with several positive outcomes on their cardiovascular health, including: low blood pressure levels (Del Brutto et al., 2016), reduced arterial stiffness (Del Brutto et al., 2018), reduced severity (Del Brutto et al., 2021) and progression (Del Brutto et al., 2022) of white matter hyperintensities (a biomarker for cerebral small vessel disease). Oily fish, and marine animals in general, are rich in omega-3 fatty acids, which have been implicated in positive cardiovascular effects in many studies (although not replicated in broad scale studies, see Manson et al., 2019). For this reason, omega-3 was also proposed as a mediator between the positive cardiovascular effects and the oily fish rich diet in the Atahualpa population.

In another population that consumes a diet extremely rich in omega-3 fatty acids, the Greenland Inuit, previous studies found a strong signature of natural selection in FADS genes (Fumagalli et al., 2015). Variants of FADS genes present in the Greenland Inuit regulate metabolic pathways to compensate for the high dietary intake of omega-3, which indicates that this population is genetically adapted to its high omega-3 intake (Fumagalli et al., 2015). A posterior study showed that the same genes had strong signatures of selection in many Native American populations (G. Amorim et al., 2017). This result indicates that the selective pressure on FADS genes could have acted in the ancestors of all Native American populations, possibly during the Beringia standstill (G. Amorim et al., 2017).

These observations motivated us to search for signatures of natural selection in the people from Atahualpa, and investigate whether natural selection has also acted in this population on genes related to fatty acid metabolism. We hypothesize that selection may have acted in response to their traditional diet rich in omega-3 fatty acids, and that selected variants could mediate the beneficial effects of this diet on their cardiovascular health.

Here, we describe the genetic relatedness of the people from the Atahualpa village to populations from the Americas and other parts of the world. We also perform a genomic scan for natural selection and report several regions that show genetic signatures of selection, including some genes related to fatty acid metabolism.

3.2 Methods

Participant consent

Participants of this study were informed and signed an informed consent document attesting that they agree with using their blood samples for DNA extraction and using their

anonymized genetic data for research and publications. The Institutional Review Board of Hospital Clínica Kennedy, Guayaquil, Ecuador (FWA: 00030727), approved the study.

DNA extraction

DNA was extracted from 50 blood samples from individuals with the five most common last names in Atahualpa as part of the Atahualpa Project (Del Brutto et al., 2014). There were no first-degree relatives in this sample. DNA samples were numbered and no identifiable information about these samples was provided to the authors of this study by the Atahualpa Project team.

Library preparation

Five samples were excluded due to low concentrations of DNA in the extractions, and we proceeded to prepare libraries from the remaining 45 samples for short-read massive parallel sequencing.

Extracted DNA was fragmented using Covaris m220 Focused-ultrasonicator for a target fragment size of 350bp to 400bp. Then, we proceeded to library preparation for 150bp paired-end sequencing on an Illumina HiSeq 4000 sequencer.

Fragment ends were repaired with NEBNext® End Repair Module (Catalog num. E6050): 21.25μL of DNA extract, 2.5μL of 10X end repair buffer (E6052) and 1.25μL of end repair enzyme mix (E6051), with a 20min incubation at 12°C and 15min at 37°C. Next, DNA fragments were purified with MinElute® PCR purification kit (5X volume of PB, 2min centrifugation at 8g, 700μL of PE, 2min centrifugation at 8g, discard flow-through, centrifuge for 1min at 8g, elute DNA with 10 μL EB, 15min incubation at 37°C followed by 2min centrifugation at 16g). The eluate containing end-repaired DNA fragments was then directed to adapter ligation using NEB quick ligation module (Catalog number E6056) following the product protocol except for the incubation, which was done at 20°C for 30min. Next, another round of purification with MinElute columns was done (10X volume of PB, 2min centrifugation at 8g, 700 μL of PE, 2min centrifugation at 8g, discard flow-through, centrifuge for 1min at 8g, add 25 μL EB, 15min incubation at 37°C followed by 2min centrifugation at 16g).

Next, adapter fill-in was performed with Bst DNA polymerase large fragment (M0275) with a 20min incubation at 65°C and 20min at 80°C. Finally, indexing PCR was done with Invitrogen Platinum Taq DNA Polymerase High Fidelity, for dual indexing with P5 and P7 indices. PCR was performed with an initial 60s at 94°C (60s), followed by 8 cycles of 30s at 94°C, 30s at 55°C and 30s at 68°C, and a final period of 5min at 68°C.

The PCR product was then submitted to size selection using AMPure magnetic beads to remove fragments smaller than 150bp or larger than 1000bp.

Four samples were excluded from further steps due to low concentrations at the expected library size distribution, measured with BioAnalyzer. The 41 libraries with good concentration at the library target size (350-400bp) were pooled into two pools with 22 and 19 samples

each. Each pool was sequenced in two lanes for 150 paired-end reads on an Illumina HiSeq 4000 instrument.

Read processing

The ends of raw sequencing reads were trimmed for adapter sequences and low quality bases, and filtered for minimum length after trimming using trimmomatic v. 0.38 with parameters ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:75.

Next, reads were mapped to the human reference genome `human_g1k_v37.fasta` downloaded from the 1000 Genomes Project `https://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/human_g1k_v37.fasta.gz`, using bwa mem with default options. Mapped reads were filtered for a maximum edit distance of 7 (taken from the NM tag of sam files, using a custom script `NMfilter.py` available on `https://github.com/deboraycb/`). Mapped reads were also filtered for a minimum mapping quality score of 15.

Next, we sorted bam files, added sample and lane tags, and merged reads of the same sample sequenced in different lanes into a single bam file per sample, using samtools. We marked and removed duplicated reads with picard, and remapped reads around potential indels using GATK IndelRealigner. We used samtools to filter out unmapped reads, reads with an unmapped mate, alignment not primary and reads that failed platformQC (sam flag $4 + 8 + 256 + 512 = 780$). Finally, we recalibrated base quality scores with GATK using dbSNP151 know sites.

Eight samples were excluded from further analyses due to average coverage below 0.5X. The remaining 33 samples kept for further analyses had an average coverage of 1.94539X.

Site filters

We used snpCleaner v2.4.3 `https://github.com/tplinderoth/ngsQC/` to filter sites for coverage and various types of bias. Mapped reads were pre-filtered for minimum base quality of 20 and proper pairs of reads using samtools options `-Q 20 --rf 2` before generating unfiltered genotype calls for snpCleaner. Sites were then filtered for a minimum of 10 individuals covered by at least 1 read (`-k 10 -u 1`), showing no excess of heterozygous genotypes on an exact test (`-H 1e-6`), no strand bias (`-S 1e-4`), no base quality bias (`-b 1e-10`), no mapping quality bias (`-f 1e-4`), and no end distance bias (`-e 1e-4`). A total of 2,561,742,893 sites passed these filters, including variable and non-variable sites within the sample.

We downloaded genome accessibility `http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/accessible_genome_masks/20140520.strict_mask.autosomes.bed` and mappability `http://hgdownload.soe.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeCrgMapabilityAlign100mer.bigWig` masks and selected sites that pass those masks (with mappability score $>= 0.5$). The intersection of those sites with the ones that passed the previous filters contains 2,029,003,071 sites.

With the goal of analyzing data from the Atahualpa population in the context of other populations from the region, we merged our dataset to the dataset from (Crawford et al., 2017). That dataset contained 59,568,964 sites that passed the accessibility and mappability masks above. The intersection of those sites with the ones that passed our filters contained 58,059,354 sites.

Genotype likelihoods

We did SNP calling on the 58M sites described above using ANGSD (Korneliussen et al., 2014). We calculated genotype likelihoods (GATK method, with `-GL 2`) from the bam files using ANGSD with parameters `-remove_bads 1 -only_proper_pairs 1 -uniqueOnly 1`, filtering for minimum mapping quality of 30 and minimum base quality of 20. We output the genotype likelihoods in Beagle format and convert it to vcf using a custom Python script. We then proceeded to merge this vcf with the dataset from (Crawford et al., 2017) containing genotype likelihoods calculated with the same filters and methods. The VCF files were merged using bcftools merge for a total of 375 samples, whose locations are shown in Figure 3.1. After filtering for biallelic SNPs, we obtained a final dataset with 21,423,891 SNPs.

Population genetics analyses

Since samples were sequenced at low coverage, we take advantage of population genetics methods that use genotype likelihoods and thus take into account uncertainty in genotypes in all downstream analyses. We use the program PCAngsd (Meisner and Albrechtsen, 2018) for principal component analysis, and the program Ohana (Cheng et al., 2017) to infer population structure and perform a selection scan after correcting for admixture. These programs required the input to be in Beagle format, which was obtained using vcftools `--BEAGLE-GL` option (Danecek et al., 2011).

For Ohana, we prepared a subset of the data containing only SNPs with minor allele frequency (MAF) over 0.05. MAF was calculated from the merged VCF file using ANGSD, and the merged VCF files were filtered for sites with $MAF > 0.05$ using bcftools (Danecek et al., 2021). We ran Ohana three times for each value of k, for 1000 iterations. We report results from the replicate number and iteration number with the best likelihood for each value of k.

We performed population branch statistic (PBS) selection scan using ANGSD (Korneliussen et al., 2014). PBS is calculated from pairwise F_{ST} values among three populations. PBS for a focal population i is given by $PBS_i = (F_{STi,j} + F_{STi,k} - F_{STj,k})/2$, and measures the length of the branch connecting population i to populations j and k, at a position or window of the genome. The selection scan was performed on windows of 50kb, slid by 10kb. Candidate peaks were selected as those that had at least 6 windows within the top 0.1% of PBS values. We also performed PBS scans with windows of 1kb and slide of 500bp in candidate regions identified in the scan with 50kb windows, to show more detailed plots of

Figure 3.1: Location of populations included in this study: people from Atahualpa in Santa Elena province (Ecuador) sequenced in this study (ATA), Aymara from Tiwanaku and La Paz (Bolivia) from (Crawford et al., 2017) (AYM), and the following populations from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015): Colombians from Mendellín (CLM), Peruvians from Lima (PEL), Puerto Ricans from Puerto Rico (PUR), Utah (USA) residents with Northern and Central European ancestry (CEU), Los Angeles (USA) residents with Mexican ancestry (MXL), and Yoruba from Ibadan (Nigeria) (YRI). (Google, 2022)

these regions. Plots of selection scan peaks with UCSC RefSeq genes were generated using R package Gviz (Hahne and Ivanek, 2016).

To infer recent demographic history of the Atahualpa population, we ran GADMA (Noskova et al., 2020) with the moments engine (Jouganous et al., 2017), using a joint site frequency spectrum (2D SFS) of Atahualpa and Aymara generated with ANGSD (Korneliussen et al., 2014). We initiated 8 GADMA runs with structured models for 2 populations. The two-population models were specified using an initial structure of `[1,1]` and a final structure of `[1,2]`, *i.e.* in each run we tested models with one time interval before the split and one after, and with one time interval before the split and two after. In each time interval of these models, a single dynamic of effective population size is maintained for each population and migration rates are constant. The allowed dynamics of change in effective population size are: constant size, sudden change in size, linear change or exponential change. We report the model with the highest likelihood among all runs.

3.3 Results

Population structure

Atahualpa samples cluster with other Native Americans in a Principal Component Analysis (PCA) including European and African populations (Figure 3.2). The people from Atahualpa are closest to the Aymara from Bolivia (Crawford et al., 2017) and the Peruvians from Lima (The 1000 Genomes Project Consortium, 2015) along the two main axes of genetic variation that together account for 14% of total genetic variation (Figure 3.2). Some individuals show evidence of admixture with European and African ancestry components, which is also observed in structure plots (Figure 3.3).

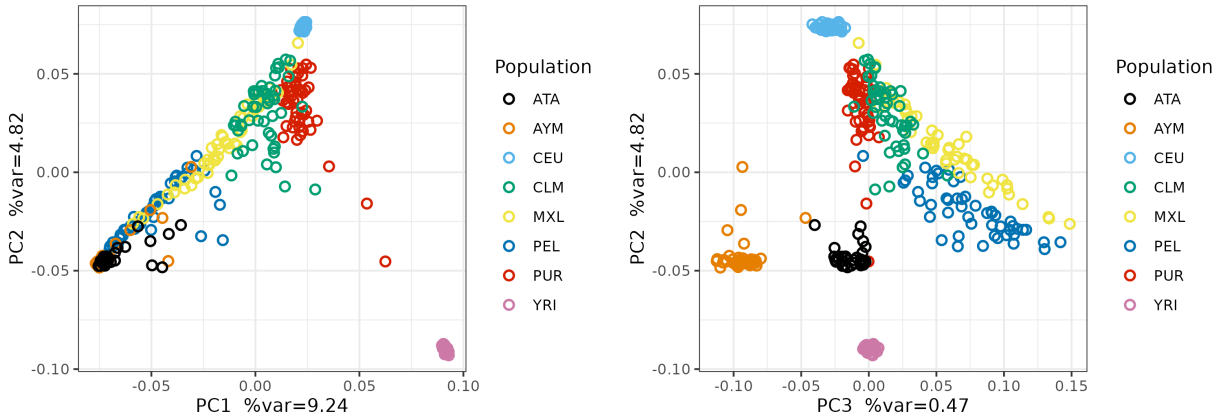

Figure 3.2: Principal component analysis. Two main axes of variation show three clusters of populations at the extremes of the distribution corresponding to African, European and Native American ancestries. The third main axis of variation separates the people from Atahualpa from other Native American populations.

The Ohana structure results with three clusters (k=3) show European and African individuals (CEU and YRI) are best described by a single component, while the other populations from the Americas are composed of a mixture of those ancestry components and a third component that likely reflects Native American ancestry (Figure 3.3). On average, the people from Atahualpa are composed of 94.1 % of this Native American ancestry, which is the second highest proportion among our sampled populations, only lower than the Aymara (Table 3.1).

Clustering with four components (k=4) splits the Native American component from k=3 into two. The Atahualpa samples are composed predominantly of one of these Native American sub-ancestries, while Aymara and the other populations from the Americas are predominantly composed of the other Native-American sub-ancestry. Increasing the number

of clusters to 5 and 6 reveals components that are prevalent in the Mexicans (MXL) and Puerto Ricans (PUR), respectively (Figure 3.3).

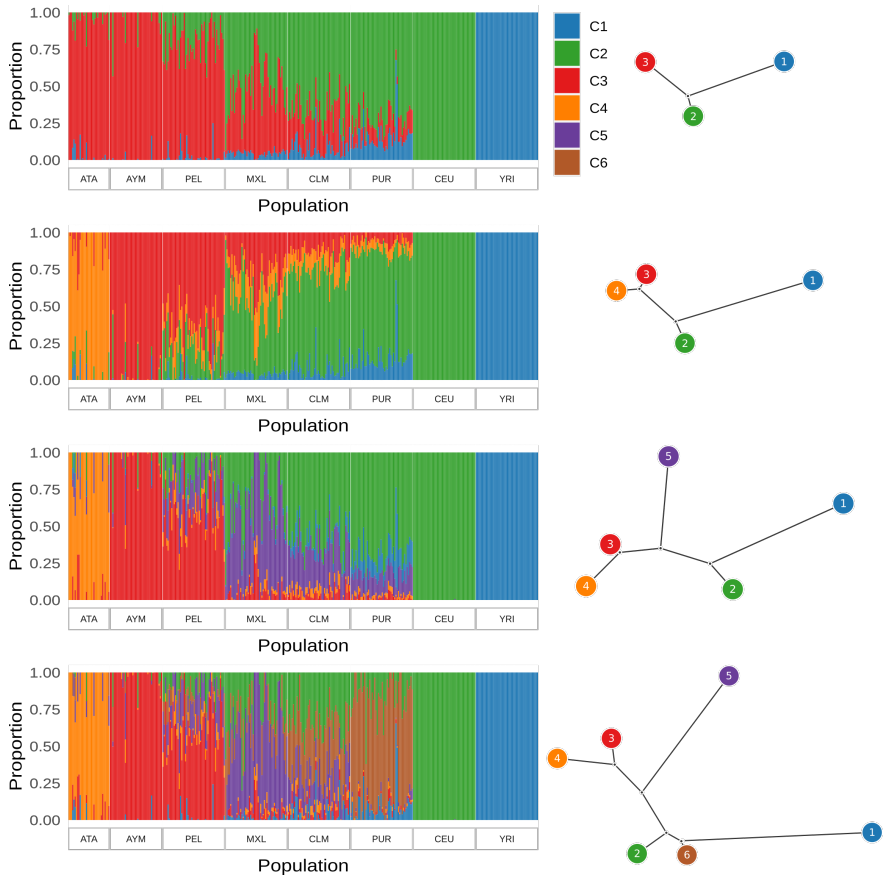

Figure 3.3: Population structure. Best clustering of genetic variation into 3, 4, 5, and 6 groups, and corresponding trees illustrating genetic covariance among clusters.

Demographic model

We inferred the split time between Atahualpa and Aymara, the population the most closely related to Atahualpa in our dataset. Our inference was based on the 2D SFS of this pair of populations and allowed for population size changes and migration between populations.

The best inferred demographic model has a population split 94.77 generations ago (2748 years ago, considering 29 years per generation), followed by a bottleneck in both populations (with the size of the population from Atahualpa decreasing to 21% of the ancestral population size, and the size of the Aymara decreasing to 12% of the ancestral population size) (Figure 3.4). The best model also includes a high and constant migration rate between populations

Table 3.1: Percentages of ancestry components (k=3) reflecting Native American, European and African ancestry in the populations from the Americas sampled in this study.

Population	Native American (C3)	European (C2)	African (C1)
Atahualpa	94.1	3.8	2.1
Aymara	96.6	2.8	0.6
Peruvians	78	19.8	2.2
Mexicans	45.2	50.4	4.4
Colombians	26.1	65.7	8.1
Puerto Ricans	12.8	71.6	15.6

from the split time to the present (rate of 10), and a population size increase in both Aymara and the population from Atahualpa 66.75 generations ago (1936 years ago), where Aymara increases to 12 times its population size following the split, and Atahualpa increases to 100 times its population size following the split.

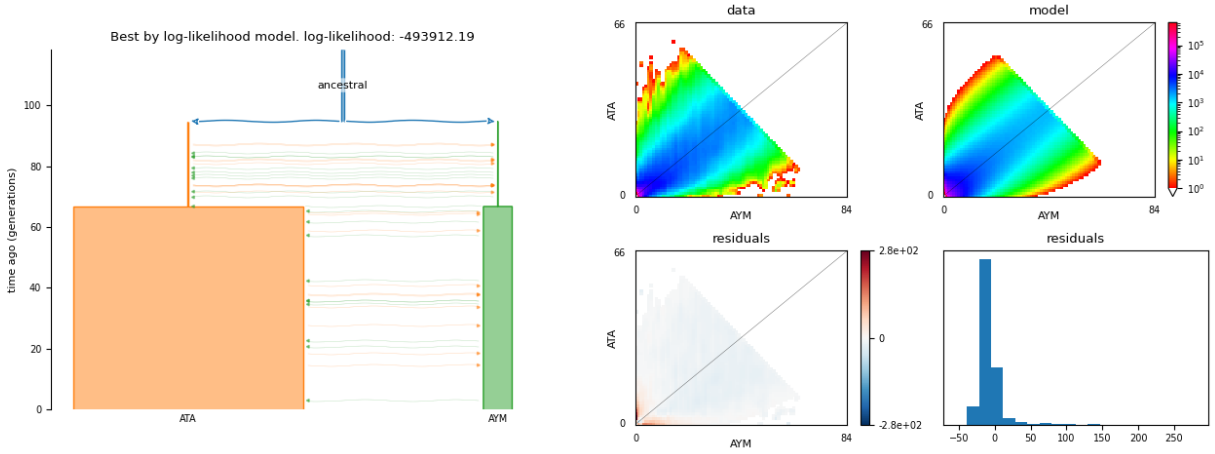

Figure 3.4: Best two-populations model fitted to the 2D SFS between Aymara (AYM) and the population from Atahualpa (ATA).

Population differentiation and signatures of selection

Genome-wide differentiation measured by F_{ST} is 0.044 between Atahualpa and Aymara, 0.040 between Atahualpa and Peruvians, and 0.016 between Aymara and Peruvians. We use this trio of closely related populations to perform a population branch statistic (PBS) genome-wide scan for natural selection. Sites with high values of PBS demonstrate high

genetic differentiation between the focal population (in this case, Atahualpa) and the other two populations, which is a signature of natural selection. Figure 3.5 shows PBS values for 50kb windows distributed along the genome, sliding by 10kb. The genome-wide average value of PBS for Atahualpa is 0.034. We identified seven peaks that show more than six windows with values of PBS on the 0.1 percentile of the genome-wide distribution (Table 3.2). We describe the candidate genes within those peaks in more detail next.

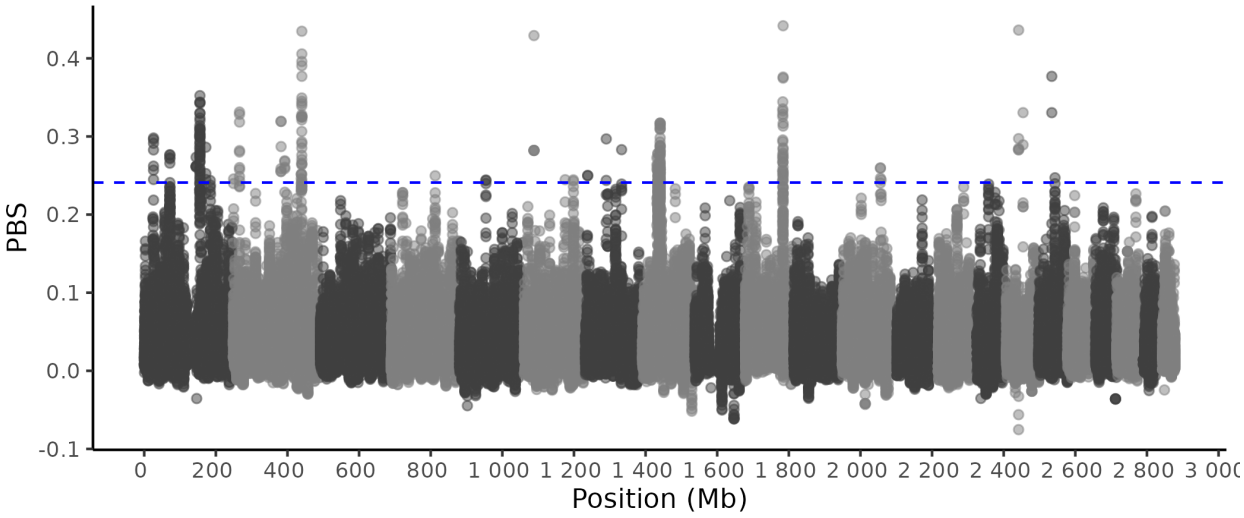

Figure 3.5: PBS scan for selection in the population from Atahualpa. Dashed blue line shows 0.1 percentile of PBS.

Table 3.2: Top selection candidate peaks from a PBS scan in the Atahualpa population relative to Aymara and Peruvians. The scan was performed with windows of 50kb, slide of 10kb. Only the windows with highest PBS values within 1Mb are listed in the table. Other candidate windows within the 0.1 percentile of the genome-wide distribution and within 1Mb of a window with higher PBS value are counted in the column "Windows".

Chromosome	Position (Mb)	PBS	Windows
10	105.185	0.441709	14
2	190.915	0.434677	27
1	155.545	0.352242	46
2	16.865	0.331273	9
2	132.665	0.319347	6
8	49.045	0.317198	51
1	25.965	0.298016	7

The most striking peak is at position 105Mb of chromosome 10. Figure 3.6 zooms into this region and reveals that it is a wide peak spanning almost 1Mb and including at least 20 genes. We list the functional information about each of those genes in Supplementary Table C.1. We highlight the *SUFU* gene, which is a repressor of Hedgehog pathway signaling. Activation of the Hedgehog pathway was recently shown to be involved in preventing obesity in adult mice under a high-fat diet (Shi and Long, 2017). Knockdown of SUFU led to lower triglyceride levels in Drosophila and decreased the mass of white adipose tissue in mice (Pospisilik et al., 2010). Therefore, this gene is clearly involved in fat metabolism and thus could play a role in the positive effects of a diet rich in oily fish on the cardiovascular health of people from the Atahualpa village (Del Brutto et al., 2016).

There are three peaks on chromosome 2: at 191Mb, 17Mb and 133Mb. The peak at 191Mb is also wide, spanning approximately 500kb and overlapping with at least 6 genes (Figure C.1). At position 17Mb, there is a sharper peak upstream of the gene *CYRIA* (Figure C.2). When we zoom into the region at 133Mb with windows of 1kb, we find a minor peak with only three windows on the gene *ANKRD30BL* (Figure C.3). Next to this region in chromosome 2, at 143Mb, two new peaks arise when we use windows of 1kb instead of 50kb (Figure 3.7). These peaks are upstream of the genes *LRP1B* and at *KYNU*. Interestingly, *LRP1B* encodes "low-density lipoprotein (LDL) receptor related protein 1B", and variants of this gene have been associated with childhood obesity (Lee, 2019). Due to its function, this gene is also a good candidate to mediate the relationship between diet and cardiovascular health in the people from Atahualpa.

We identify two peaks at chromosome 1: at 155Mb and 26Mb. The peak at 155Mb spans 700Kb and at least 29 genes (Figure 3.8), which we list and describe in Supplementary Table C.2. Among those, we highlight *FAM189B*, which has been associated with Gaucher disease, a disease that results from a buildup of fatty substances mainly in the liver and spleen. This disease association suggests that this gene could also be a good candidate related to fat metabolism.

The peak at 26Mb of chromosome 1 contains several windows with high PBS values in a narrow region of 100Kb that is next to two genes: *LDLRAP1* and *MAN1C1*. *MAN1C1* is related to the metabolism of proteins and *LDLRAP1* encodes "low-density lipoprotein (LDL) receptor adapter protein 1", a protein that helps remove cholesterol from the bloodstream. Thus, we also highlight this peak as a candidate of selection driven by the diet rich in oily fish in the people from Atahualpa.

Lastly, the wide peak on chromosome 8 can be attributed to a region of increased mutation next to the centromere, which has been described by Logsdon et al. (2021) (Figure C.4). Therefore, we will not discuss this region further as a potential candidate of selection.

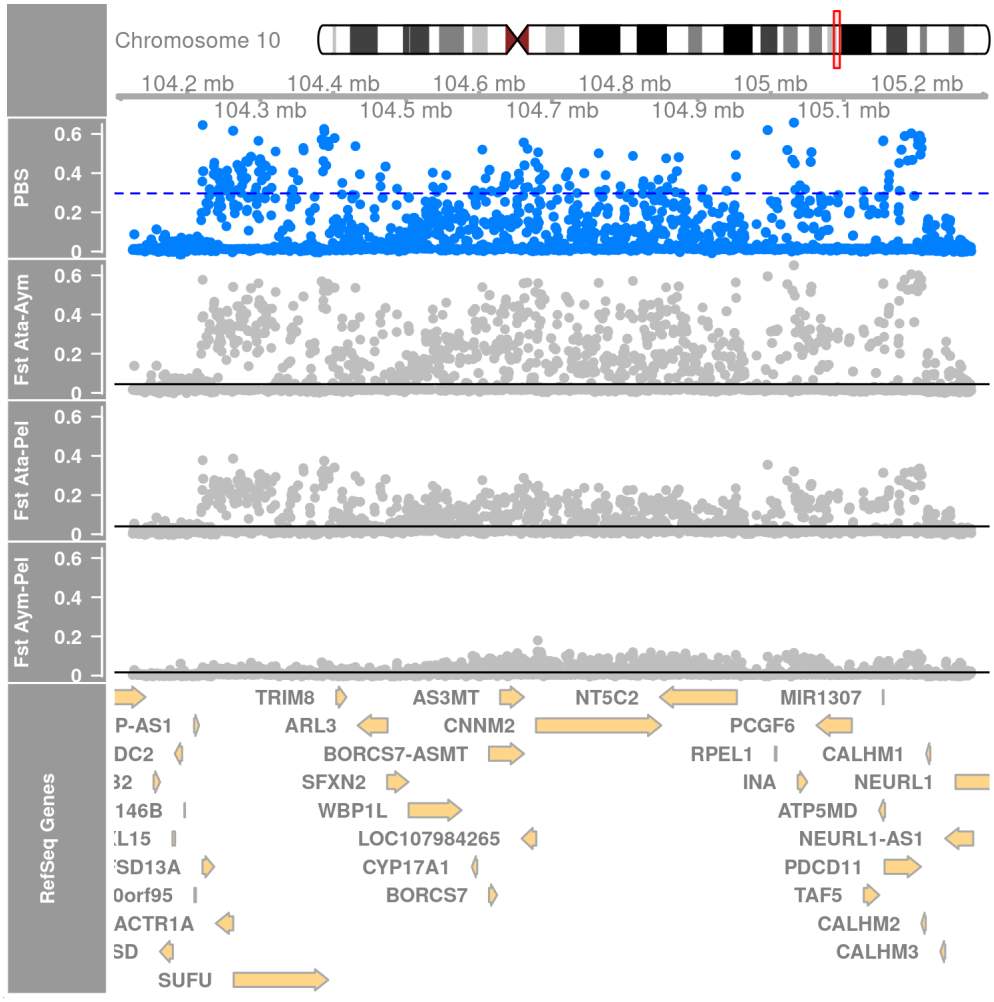

Figure 3.6: PBS scan peak on chromosome 10. We highlight the *SUFU* gene, which has been implicated in fat metabolism. Scan performed with windows of 1kb, slide of 500bp. Dashed blue line shows 0.1 percentile of PBS. Black lines show genomewide F_{ST} values for each population pair.

3.4 Discussion

Genetic ancestry

The people from Atahualpa show a large proportion of Native American ancestry (94.1%, Table 3.1), which is even higher than the proportion of Native American ancestry among members of the officially recognized Ecuadorian indigenous group Tsáchila (87.12%), in the study by Nagar et al. (2021). Although other populations such as the Aymaras and the Peruvians share a similarly high proportion of Native American ancestry, the population from Atahualpa is genetically differentiated from them, with a distinct ancestry component.

Previous studies have shown a signal of East-West structuring of populations in South America, mainly separating populations from the highlands of the Andes from the populations from the Amazon lowlands (most recently Borda et al., 2020; Nakatsuka et al., 2020). These recent studies also showed that in the Northern Andes (North of Northern Peru), populations are not as differentiated between East and West as in the Central Andes (the region starting from central Peru and stretching South through Bolivia, Chile and Argentina). The Northern Andes reach lower altitudes than the Central Andes, and it seems plausible that lower altitudes would allow more gene flow between the coastal region and the Amazon region (Borda et al., 2020). Indeed, coastal populations from Northern Peru (Tallanes and Moche) are genetically similar to the Chachapoyas from the Amazon Yunga, a transitional ecoregion in the Eastern slope of the Andes, between the highlands and the lowland forests (Borda et al., 2020).

The patterns of population structure along the Andes mentioned above were described based on Peruvian populations. However, the Ecuadorian highlands are also part of the Northern Andes, and the coastal region of Santa Elena province belongs to a similar dry forest ecoregion as the location of the Tallanes and Moche in Northern Peru, across the Gulf of Guayaquil. Therefore, it is possible that the genetic component that is almost exclusively present in the Atahualpa population (Figure 3.3, $k >= 4$) could be related to the component found in coastal populations from Northern Peru (Tallanes and Moche) in Borda et al. (2020). In addition to the proximity and environmental similarity, there is archaeological evidence of ancient contact between the people of Northern Peru and Southern Ecuador (Guffroy, 2008). Two possibilities then arise for the origins of this coastal ancestry component: i) it could be the result of East-West gene flow with Amazonian populations through the Northern Andes or ii) it could be an old component related to the first humans that invaded South America through the Pacific coast. These possibilities remain to be tested, but the results from Borda et al. (2020), who found similarities between populations from the coast and from the Eastern Yunga, suggest the former.

The estimated split time between the population from Atahualpa and the Aymara (2748 years ago) corresponds to the period of the Machalilla (1430-830 cal B.C.) and Chorrera (1300-300 cal B.C.) material cultures from coastal Ecuador (Zeidler, 2008). The Chorrera culture maintained contact with the Andean highlands through trade (Zeidler, 2008), which might explain the constant high migration rate we found between the people from Atahualpa

and the Aymara from the Bolivian highlands.

Selection

The results from our PBS scan suggest different processes driving the differences in allele frequencies between the people from Atahualpa and two closely related populations (Aymara and Peruvians). On one hand, four PBS peaks are sharp and indicate the action of selection at specific genes (*LRP1B*, *LDLRAP1*, *CYRIA* and *ANKRD30BL*). On the other hand, there are three wide peaks (encompassing between 500Kb and 1Mb), which contain several genes. The wide PBS peaks make it difficult to pinpoint specific sites that were targets of selection. Nevertheless, the long stretches of divergent sequence in Atahualpa also raise an interesting hypothesis: that these haplotypes were selected in this population after introgression from a diverged population. More specifically, the fact that the PBS signal spans a long sequence indicates a recent process, since the haplotype has not yet been broken by recombination. Further, the fact that the divergence remains high and decreases abruptly at the edges of the block indicates that the haplotype could have been inherited as a whole divergent unit from another relatively distant population. The latter scenario differs from the signature of a haplotype hitchhiking on a new mutation that recently underwent positive selection. In this case, we would expect the signature of high PBS to gradually decrease with distance from the selected mutation. Testing the hypothesis of introgression could include the use of methods that explicitly model recombination, such as those based on the ancestral recombination graph (Chapter 1). Further investigations could explore the possibility of introgression with archaic humans, such as Denisovans and Neanderthals, as possible sources of these haplotypes.

Our main motivation for investigating signatures of natural selection in the genomes of the people from Atahualpa came from the observed health benefits associated with the ingestion of oily fish as part of their traditional diet. Therefore, we highlight genes present in regions with high PBS scores that have been previously implicated in lipid metabolism. Interestingly, two out of the four sharp PBS peaks include genes (*LRP1B* and *LDLRAP1*) that encode cholesterol receptors. Two other genes (*SUFU* and *FAM189B*) also encode proteins that have been associated with lipid metabolism and represent promising candidates of selection within two wide peaks of high PBS scores. Therefore, these genes represent a starting point for future investigations to understand the physiological mechanisms that mediate the beneficial effects of diet on the cardiovascular health of the people from Atahualpa.

Interestingly, Fumagalli et al. (2015) recovered a signature of selection on *FADS* genes associated with a diet rich in omega-3 polyunsaturated fatty acids. It is important to note that we did not find signatures of selection in these same genes. However, this result is expected since it has been demonstrated that FADS genes show signatures of selection in many Native American populations (G. Amorim et al., 2017), suggesting that selection on these genes occurred prior to the peopling of the Americas. Considering that we compared the people from Atahualpa to the closely related Aymara and Peruvians, this selection scan is tailored to find candidates of recent selection acting specifically in the population from

Atahualpa, and would not find signals of selection that are shared with Aymara and Peruvians. Therefore, the alleles of the candidate genes that we encountered in this study have the potential to represent new adaptations, additional to the FADS genes, to a diet rich in oily foods.

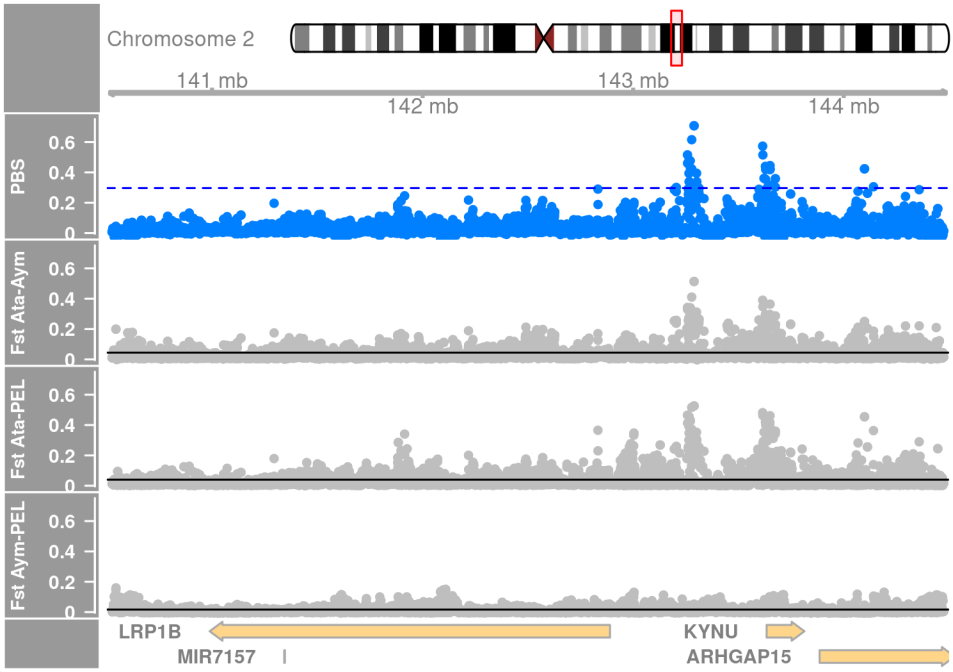

Figure 3.7: PBS scan peak on chromosome 2 at 143Mb. Scan performed with windows of 1kb, slide of 500bp. Dashed blue line shows 0.1 percentile of PBS. Black lines show genomewide F_{ST} values for each population pair.

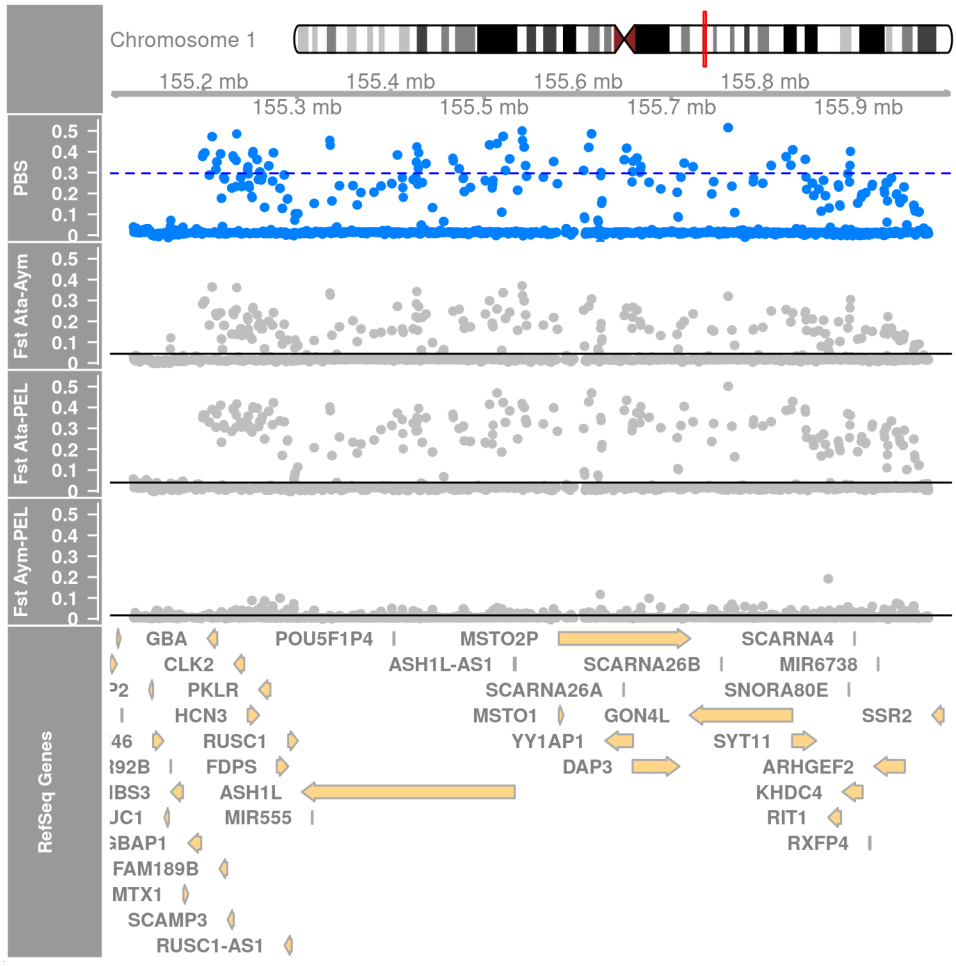

Figure 3.8: PBS scan peak on chromosome 1 at 155Mb. We highlight the gene *FAM189B*, which has been associated to a disease (Gaucher disease) that results from buildup of fatty substances. Scan performed with windows of 1kb, slide of 500bp. Dashed blue line shows 0.1 percentile of PBS. Black lines show genomewide F_{ST} values for each population pair.

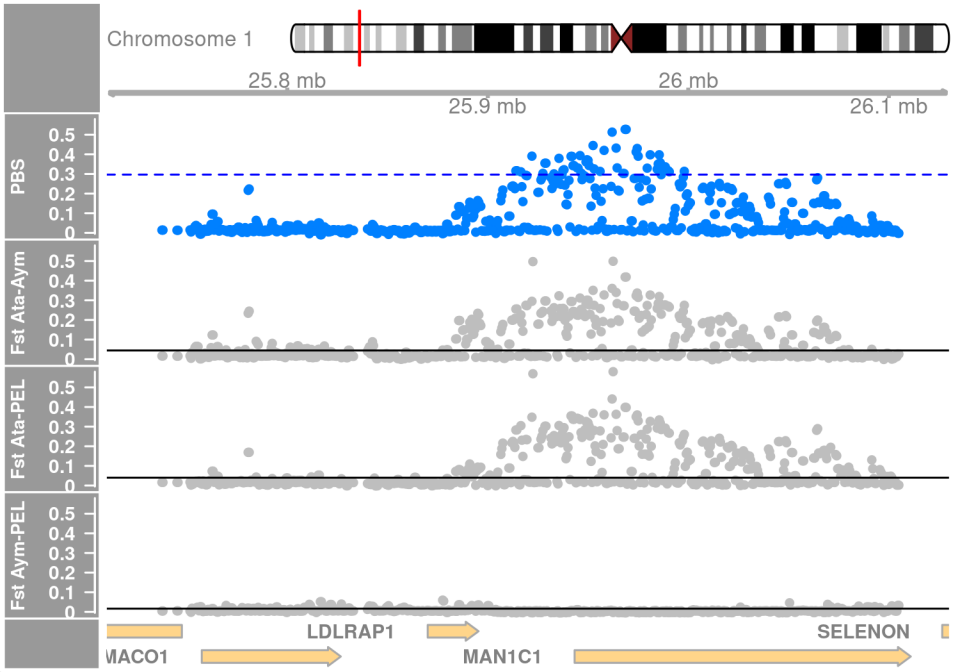

Figure 3.9: PBS scan peak on chromosome 1 at 26Mb. Scan performed with windows of 1kb, slide of 500bp. Dashed blue line shows 0.1 percentile of PBS. Black lines show genomewide F_{ST} values for each population pair.

Bibliography

A. Arredondo, B. Mourato, K. Nguyen, S. Boitard, W. Rodríguez, O. Mazet, and L. Chikhi. Inferring number of populations and changes in connectivity under the n-island model. *Heredity*, 126(6):896–912, 2021.

F. Baumdicker, G. Bisschop, D. Goldstein, G. Gower, A. P. Ragsdale, G. Tsambos, S. Zhu, B. Eldon, C. E. Ellerman, J. G. Galloway, A. L. Gladstein, G. Gorjanc, B. Guo, B. Jeffery, W. W. Kretzschmar, K. Lohse, M. Matschiner, D. Nelson, N. S. Pope, C. D. Quinto-Cortés, M. F. Rodrigues, K. Saunack, T. Sellinger, K. Thornton, H. v. Kemenade, A. W. Wohns, Y. Wong, S. Gravel, A. D. Kern, J. Koskela, P. L. Ralph, and J. Kelleher. Efficient ancestry and mutation simulation with msprime 1.0. *bioRxiv*, 17:2021.08.31.457499, 2021.

A. Bergström, C. Stringer, M. Hajdinjak, E. M. Scerri, and P. Skoglund. Origins of modern human ancestry. *Nature*, 590(7845):229–237, 2021.

V. Borda, I. Alvim, M. Mendes, C. Silva-Carvalho, B. S. S. Giordano, T. P. Leal, V. Furlan, M. O. Scliar, R. Zamudio, C. Zolini, G. S. Araújo, M. R. Luizon, C. Padilla, O. Cáceres, K. Levano, C. Sánchez, O. Trujillo, P. O. Flores-Villanueva, M. Dean, S. Fuselli, M. Machado, P. E. Romero, F. Tassi, M. Yeager, T. D. O'Connor, R. H. Gilman, E. Tarazona-Santos, H. Guio, and E. Tarazona-Santos. The genetic structure and adaptation of Andean highlanders and Amazonians are influenced by the interplay between geography and culture. *Proceedings of the National Academy of Sciences of the United States of America*, 117(51):32557–32565, 2020.

S. P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, 1998.

J. Y. Cheng, T. Mailund, and R. Nielsen. Fast admixture analysis and population tree estimation for SNP and NGS data. *Bioinformatics*, 33(14):2148–2155, 2017.

L. Chikhi, W. Rodríguez, S. Grusea, P. Santos, S. Boitard, and O. Mazet. The IICR (inverse instantaneous coalescence rate) as a summary of genomic diversity: insights into demographic inference and model choice. *Heredity*, 120:13–24, 1 2018.

S. R. Cook, A. Gelman, and D. B. Rubin. Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, 15(3):675–692, 2006.

J. E. Crawford, R. Amaru, J. Song, C. G. Julian, F. Racimo, J. Y. Cheng, X. Guo, J. Yao, B. Ambale-Venkatesh, J. A. Lima, J. I. Rotter, J. Stehlik, L. G. Moore, J. T. Prchal, and R. Nielsen. Natural Selection on Genes Related to Cardiovascular Health in High-Altitude Adapted Andeans. *American Journal of Human Genetics*, 101(5):752–767, 2017.

P. Danecek, A. Auton, G. R. 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 (Oxford, England)*, 27(15):2156–8, 8 2011.

P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, and H. Li. Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2), 2 2021.

O. H. Del Brutto and M. Zambrano. Atahualpa, una población rural ideal para la práctica de estudios epidemiológicos. *Revista Ecuatoriana de Neurología*, pages 88–94, 2017.

O. H. Del Brutto, E. Peñaherrera, E. Ochoa, M. Santamaría, M. Zambrano, and V. J. Del Brutto. Door-to-door survey of cardiovascular health, stroke, and ischemic heart disease in rural coastal Ecuador - the Atahualpa Project: Methodology and operational definitions. *International Journal of Stroke*, 9(3):367–371, 2014.

O. H. Del Brutto, R. M. Mera, J. Gillman, P. R. Castillo, M. Zambrano, and J.-E. Ha. Dietary Oily Fish Intake and Blood Pressure Levels: A Population-Based Study. *The Journal of Clinical Hypertension*, 18(4):337–341, 2016.

O. H. Del Brutto, R. M. Mera, E. Peñaherrera, R. Peñaherrera, and A. F. Costa. The relationship between oily fish intake and arterial stiffness in older adults living in rural coastal Ecuador. *Nutrition, Metabolism and Cardiovascular Diseases*, 28(11):1173–1174, 2018.

O. H. Del Brutto, B. Y. Recalde, and R. M. Mera. Dietary Oily Fish Intake is Inversely Associated with Severity of White Matter Hyperintensities of Presumed Vascular Origin. A Population-Based Study in Frequent Fish Consumers of Amerindian Ancestry. *Journal of Stroke and Cerebrovascular Diseases*, 30(6):1–8, 2021.

O. H. Del Brutto, R. M. Mera, B. Y. Recalde, D. A. Rumbea, and M. J. Sedler. Life's simple 7 and all-cause mortality. A population-based prospective cohort study in middle-aged and older adults of Amerindian ancestry living in rural Ecuador. *Preventive Medicine Reports*, 25:101668, 2022.

Y. Deng, Y. S. Song, and R. Nielsen. The distribution of waiting distances in ancestral recombination graphs. *Theoretical Population Biology*, 2021.

T. D. Dillehay, C. Ramírez, M. Pino, M. B. Collins, and J. Rossen. Monte Verde: Seaweed, Food, Medicine, and the Peopling of South America. *Science*, 320(May):784–786, 2008.

J. Y. Dutheil, G. Ganapathy, A. Hobolth, T. Mailund, M. K. Uyenoyama, and M. H. Schierup. Ancestral population genomics: The coalescent hidden Markov model approach. *Genetics*, 183(1):259–274, 2009.

S. Fan, D. E. Kelly, M. H. Beltrame, M. E. B. Hansen, S. Mallick, A. Ranciaro, J. Hirbo, S. Thompson, W. Beggs, T. Nyambo, S. A. Omar, D. W. Meskel, G. Belay, A. Froment, N. Patterson, D. Reich, and S. A. Tishkoff. African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. *Genome Biology*, 20 (1):82, 2019.

R. A. Fisher. *The Genetical Theory of Natural Selection*. Clarendon Press, 1930.

M. Fumagalli, I. Moltke, N. Grarup, F. Racimo, P. Bjerregaard, M. E. Jørgensen, T. S. Korneliussen, P. Gerbault, L. Skotte, A. Linneberg, C. Christensen, I. Brandslund, and T. J. Greenlandic Inuit showgenetic signatures of diet and climate adaptation. *Science*, 349(6254):1343–1348, 2015.

C. E. G. Amorim, K. Nunes, D. Meyer, D. Comas, M. C. Bortolini, F. M. Salzano, and T. Hünemeier. Genetic signature of natural selection in first Americans. *Proceedings of the National Academy of Sciences*, 114(9):201620541, 2017.

A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

Google. Atlantic, 2022. URL https://www.google.com/maps/@18.6090502,-56.5691989,2.92z.

R. E. Green, J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. J. 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, Z. Kucan, I. Gusic, 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. E. Reich, and S. Pääbo. A Draft Sequence of the Neandertal Genome. *Science*, 328(5979): 710–722, 2010.

R. C. Griffiths and P. Marjoram. An Ancestral Recombination Graph. In P. Donnelly and S. Tavare, editors, *Progress in Population Genetics and Human Evolution, IMA Volumes in Mathematics and its Applications, vol. 87*, pages 257–270. Springer, 1997.

J. Guffroy. Cultural Boundaries and Crossings: Ecuador and Peru. In H. Silverman and W. H. Isbell, editors, *The Handbook of South American Archaeology*, chapter 44. Springer, 2008.

W. Haak, I. Lazaridis, N. J. Patterson, N. Rohland, S. Mallick, B. Llamas, G. Brandt, S. Nordenfelt, E. Harney, K. Stewardson, Q. Fu, A. Mittnik, E. Bánffy, C. Economou, M. Francken, S. Friederich, R. G. Pena, F. Hallgren, V. Khartanovich, A. Khokhlov, M. Kunst, P. Kuznetsov, H. Meller, O. Mochalov, V. Moiseyev, N. Nicklisch, S. L. Pichler, R. Risch, M. A. Rojo Guerra, C. Roth, A. Szécsényi-Nagy, J. Wahl, M. Meyer, J. Krause, D. Brown, D. Anthony, A. Cooper, K. W. Alt, and D. Reich. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, 522(7555):207–211, 2015.

F. Hahne and R. Ivanek. Visualizing Genomic Data Using Gviz and Bioconductor. In E. Mathé and S. Davis, editors, *Statistical Genomics: Methods and Protocols*, pages 335–351. Springer New York, New York, NY, 2016.

D. L. Hartl and A. G. Clark. *Principles of Population Genetics*. Sinauer Associates, Sunderland, 4th edition, 2007.

R. Heller, L. Chikhi, and H. R. Siegismund. The Confounding Effect of Population Structure on Bayesian Skyline Plot Inferences of Demographic History. *PLoS ONE*, 8(5):e62992, 5 2013.

B. M. Henn, L. L. Cavalli-Sforza, and M. W. Feldman. The great human expansion. *Proceedings of the National Academy of Sciences of the United States of America*, 109(44): 17758–64, 10 2012.

A. Hobolth, O. F. Christensen, T. Mailund, and M. H. Schierup. Genomic Relationships and Speciation Times of Human, Chimpanzee, and Gorilla Inferred from a Coalescent Hidden Markov Model. *PLOS Genetics*, 3(2):1–11, 2007.

M. Hubisz and A. Siepel. Inference of Ancestral Recombination Graphs Using ARGweaver. In Julien Y. Dutheil, editor, *Statistical Population Genomics*, volume 2090, chapter 10, pages 231–266. Humana, New York, NY, 2020.

M. J. Hubisz, A. L. Williams, and A. Siepel. Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph. *PLoS genetics*, 16(8):e1008895, 2020.

R. R. Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23(2):183–201, 1983.

R. R. Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.

J. Jouganous, W. Long, A. P. Ragsdale, and S. Gravel. Inferring the Joint Demographic History of Multiple Populations: Beyond the Diffusion Approximation. *Genetics*, 206(3): 1549–1567, 7 2017.

T. H. Jukes and C. R. Cantor. Evolution of Protein Molecules. In H. N. Munro, editor, *Mammalian Protein Metabolism*, chapter 24, pages 21–132. Academic Press, 1969.

J. Kamm, J. Terhorst, R. Durbin, and Y. S. Song. Efficiently inferring the demographic history of many populations with allele count data. *Journal of the American Statistical Association*, 115(531):1472–1487, 2020.

J. Kelleher, A. M. Etheridge, and G. McVean. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Computational Biology*, 12(5):1–22, 2016.

J. Kelleher, K. R. Thornton, J. Ashander, and P. L. Ralph. Efficient pedigree recording for fast population genetics simulation. *PLOS Computational Biology*, 14(11):1–21, 2018.

J. Kelleher, Y. Wong, A. W. Wohns, C. Fadil, P. K. Albers, and G. McVean. Inferring whole-genome histories in large population datasets. *Nature Genetics*, 51(9):1330–1338, 2019.

J. F. C. Kingman. On the Genealogy of Large Populations. *Journal of Applied Probability*, 19:27–43, 1982.

T. S. Korneliussen, A. Albrechtsen, and R. Nielsen. ANGSD: analysis of next generation sequencing data. *BMC bioinformatics*, 15(1):356, 2014.

S. Lee. The genetic and epigenetic association of LDL Receptor Related Protein 1B (LRP1B) gene with childhood obesity. *Scientific Reports*, 9(1):1815, 2019.

R. C. Lewontin and J. Krakauer. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, 74(1):175–95, 5 1973.

H. Li and R. Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496, 7 2011.

N. Li and M. Stephens. Modelling Linkage Disequilibrium using Single Nucleotide Polymorphism Data. *Genetics*, 2233(December):2213–2233, 2003.

G. A. Logsdon, M. R. Vollger, P. H. Hsieh, Y. Mao, M. A. Liskovykh, S. Koren, S. Nurk, L. Mercuri, P. C. Dishuck, A. Rhie, L. G. de Lima, T. Dvorkina, D. Porubsky, W. T. Harvey, A. Mikheenko, A. V. Bzikadze, M. Kremitzki, T. A. Graves-Lindsay, C. Jain, K. Hoekzema, S. C. Murali, K. M. Munson, C. Baker, M. Sorensen, A. M. Lewis, U. Surti, J. L. Gerton, V. Larionov, M. Ventura, K. H. Miga, A. M. Phillippy, and E. E. Eichler. The structure, function and evolution of a complete human chromosome 8. *Nature*, 593 (7857):101–107, 2021.

J. E. Manson, N. R. Cook, I.-M. Lee, W. Christen, S. S. Bassuk, S. Mora, H. Gibson, C. M. Albert, D. Gordon, T. Copeland, D. D'Agostino, G. Friedenberg, C. Ridge, V. Bubes, E. L. Giovannucci, W. C. Willett, and J. E. Buring. Marine n-3 Fatty Acids and Prevention of

Cardiovascular Disease and Cancer. *New England Journal of Medicine*, 380(1):23–32, 1 2019.

P. Marjoram and J. D. Wall. Fast "coalescent" simulation. *BMC Genetics*, 7(1):16, 2006.

O. Mazet, W. Rodríguez, S. Grusea, S. Boitard, and L. Chikhi. On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference? *Heredity*, 116(4):362–371, 4 2016.

C. McEwan and F. Delgado-Espinoza. Late Pre-Hispanic Polities of Coastal Ecuador. In H. Silverman and W. H. Isbell, editors, *The Handbook of South American Archaeology*, chapter 26. Springer, 2008.

G. A. T. McVean and N. J. Cardin. Approximating the coalescent with recombination. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 360(1459):1387–93, 2005.

J. Meisner and A. Albrechtsen. Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data. *Genetics*, 210(2):719–731, 10 2018.

P. Moorjani, N. J. Patterson, J. N. Hirschhorn, A. Keinan, L. Hao, G. Atzmon, E. Burns, H. Ostrer, A. L. Price, and D. E. Reich. The history of african gene flow into Southern Europeans, Levantines, and Jews. *PLoS Genetics*, 7(4), 2011.

P. A. P. Moran. Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, 54(1):60–71, 1958.

S. D. Nagar, A. B. Conley, A. T. Chande, L. Rishishwar, S. Sharma, L. Mariño-Ramírez, G. Aguinaga-Romero, F. González-Andrade, and I. K. Jordan. Genetic ancestry and ethnic identity in Ecuador. *Human Genetics and Genomics Advances*, 2(4):1–12, 2021.

N. Nakatsuka, I. Lazaridis, C. Barbieri, P. Skoglund, N. Rohland, S. Mallick, C. Posth, K. Harkins-Kinkaid, M. Ferry, E. Harney, M. Michel, K. Stewardson, J. Novak-Forst, J. M. Capriles, M. A. Durruty, K. A. Álvarez, D. Beresford-Jones, R. Burger, L. Cadwallader, R. Fujita, J. Isla, G. Lau, C. L. Aguirre, S. LeBlanc, S. C. Maldonado, F. Meddens, P. G. Messineo, B. J. Culleton, T. K. Harper, J. Quilter, G. Politis, K. Rademaker, M. Reindel, M. Rivera, L. Salazar, J. R. Sandoval, C. M. Santoro, N. Scheifler, V. Standen, M. I. Barreto, I. F. Espinoza, E. Tomasto-Cagigao, G. Valverde, D. J. Kennett, A. Cooper, J. Krause, W. Haak, B. Llamas, D. Reich, and L. Fehren-Schmitz. A Paleogenomic Reconstruction of the Deep Population History of the Andes. *Cell*, 181(5):1131–1145, 2020.

S. Neph, M. S. Kuehn, A. P. Reynolds, E. Haugen, R. E. Thurman, A. K. Johnson, E. Rynes, M. T. Maurano, J. Vierstra, S. Thomas, R. Sandstrom, R. Humbert, and J. A. Stamatoyannopoulos. BEDOPS: high-performance genomic feature operations. *Bioinformatics*, 28 (14):1919–1920, 2012.

E. Noskova, V. Ulyantsev, K.-P. Koepfli, S. J. O'Brien, and P. Dobrynin. GADMA: Genetic algorithm for inferring demographic history of multiple populations from allele frequency spectrum data. *GigaScience*, 9(3):giaa005, 3 2020.

M. Notohara. The coalescent and the genealogical process in geographically structured population. *Journal of Mathematical Biology*, 29(1):59–75, 1990.

M. Osmond and G. Coop. Estimating dispersal rates and locating genetic ancestors with genome-wide genealogies. *bioRxiv*, page 2021.07.13.452277, 2021.

S. J. O'Brien, M. E. Roelke, N. Yuhki, K. W. Richards, W. E. Johnson, W. L. Franklin, A. E. Anderson, O. L. Bass Jr., R. C. Belden, and J. S. Martenson. Genetic introgression within the Florida Panther Felis concolor coryi. *National Geographic Research*, 6(4):485–494, 1990.

L. Pagani, D. J. Lawson, E. Jagoda, A. Mörseburg, A. Eriksson, M. Mitt, F. Clemente, G. Hudjashov, M. DeGiorgio, L. Saag, J. D. Wall, A. Cardona, R. Mägi, M. A. W. Sayres, S. Kaewert, C. Inchley, C. L. Scheib, M. Järve, M. Karmin, G. S. Jacobs, T. Antao, F. M. Iliescu, A. Kushniarevich, Q. Ayub, C. Tyler-Smith, Y. Xue, B. Yunusbayev, K. Tambets, C. B. Mallick, L. Saag, E. Pocheshkhova, G. Andriadze, C. Muller, M. C. Westaway, D. M. Lambert, G. Zoraqi, S. Turdikulova, D. Dalimova, Z. Sabitov, G. N. N. Sultana, J. Lachance, S. Tishkoff, K. Momynaliev, J. Isakova, L. D. Damba, M. Gubina, P. Nymadawa, I. Evseeva, L. Atramentova, O. Utevska, F.-X. Ricaut, N. Brucato, H. Sudoyo, T. Letellier, M. P. Cox, N. A. Barashkov, V. Škaro, L. Mulahasanovic´, D. Primorac, H. Sahakyan, M. Mormina, C. A. Eichstaedt, D. V. Lichman, S. Abdullah, G. Chaubey, J. T. S. Wee, E. Mihailov, A. Karunas, S. Litvinov, R. Khusainova, N. Ekomasova, V. Akhmetova, I. Khidiyatova, D. Marjanović, L. Yepiskoposyan, D. M. Behar, E. Balanovska, A. Metspalu, M. Derenko, B. Malyarchuk, M. Voevoda, S. A. Fedorova, L. P. Osipova, M. M. Lahr, P. Gerbault, M. Leavesley, A. B. Migliano, M. Petraglia, O. Balanovsky, E. K. Khusnutdinova, E. Metspalu, M. G. Thomas, A. Manica, R. Nielsen, R. Villems, E. Willerslev, T. Kivisild, and M. Metspalu. Genomic analyses inform on migration events during the peopling of Eurasia. *Nature*, 538(7624):238–242, 2016.

M. Plummer, N. Best, K. Cowles, and K. Vines. CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*, 6(1):7–11, 2006.

J. E. Pool, I. Hellmann, J. D. Jensen, and R. Nielsen. Population genetic inference from genomic sequence variation. *Genome Research*, 20(3):291–300, 2010.

J. A. Pospisilik, D. Schramek, H. Schnidar, S. J. F. Cronin, N. T. Nehme, X. Zhang, C. Knauf, P. D. Cani, K. Aumayr, J. Todoric, M. Bayer, A. Haschemi, V. Puviindran, K. Tar, M. Orthofer, G. G. Neely, G. Dietzl, A. Manoukian, M. Funovics, G. Prager, O. Wagner, D. Ferrandon, F. Aberger, C.-c. Hui, H. Esterbauer, and J. M. Penninger. Drosophila Genome-wide Obesity Screen Reveals Hedgehog as a Determinant of Brown versus White Adipose Cell Fate. *Cell*, 140(1):148–160, 2010.

K. Prüfer, F. Racimo, N. J. 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. K. 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. E. Reich, J. Kelso, and S. Pääbo. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481): 43–9, 2014.

P. Ralph, K. Thornton, and J. Kelleher. Efficiently Summarizing Relationships in Large Samples: A General Duality Between Statistics of Genealogies and Genomes. *Genetics*, 215(3):779 LP – 797, 7 2020.

M. D. Rasmussen, M. J. Hubisz, I. Gronau, and A. Siepel. Genome-Wide Inference of Ancestral Recombination Graphs. *PLoS Genetics*, 10(5), 2014.

J. S. Raymond. The Process of Sedentism in Northwestern South America. In H. Silverman and W. H. Isbell, editors, *The Handbook of South American Archaeology*, chapter 44. Springer, 2008.

V. Roy. Convergence diagnostics for Markov Chain Monte Carlo, 3 2020.

M. Safran, N. Rosen, M. Twik, R. BarShir, T. I. Stein, D. Dahary, S. Fishilevich, and D. Lancet. GeneCards - the human gene database, 2022. URL www.genecards.org.

N. F. Saremi, M. A. Supple, A. Byrne, J. A. Cahill, L. L. Coutinho, L. Dalén, H. V. Figueiró, W. E. Johnson, H. J. Milne, S. J. O'Brien, B. O'Connell, D. P. Onorato, S. P. D. Riley, J. A. Sikich, D. R. Stahler, P. M. S. Villela, C. Vollmers, R. K. Wayne, E. Eizirik, R. B. Corbett-Detig, R. E. Green, C. C. Wilmers, and B. Shapiro. Puma genomes from North and South America provide insights into the genomic consequences of inbreeding. *Nature Communications*, 10(1):4769, 2019.

S. Schiffels and R. Durbin. Inferring human population size and separation history from multiple genome sequences. *Nature genetics*, 46(8):919–25, 2014.

C. M. Schlebusch, H. Malmström, T. Günter, P. Sjödin, A. Coutinho, H. Edlund, A. R. Munters, M. Vicente, M. Steyn, H. Soodyall, M. Lombard, and M. Jakobsson. Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science*, 358:652–655, 2017.

S. Sheehan, K. Harris, and Y. S. Song. Estimating variable effective population sizes from multiple genomes: a sequentially markov conditional sampling distribution approach. *Genetics*, 194(3):647–62, 7 2013.

Y. Shi and F. Long. Hedgehog signaling via Gli2 prevents obesity induced by high-fat diet in adult mice. *eLife*, 6, 12 2017.

P. Sjödin, I. Kaj, S. Krone, M. Lascoux, and M. Nordborg. On the meaning and existence of an effective population size. *Genetics*, 169(2):1061–70, 2 2005.

P. Sjödin, J. McKenna, and M. Jakobsson. Estimating divergence times from DNA sequences. *Genetics*, 217(4), 2021.

M. Slatkin. Testine neutrality in subdivided populations. *Genetics*, 100(3):533–545, 1982.

M. Slatkin. Gene Flow and the Geographic Structure of Natural Populations. *Science*, 236 (4803):787–792, 1987.

S. Song, E. Sliwerska, S. Emery, and J. M. Kidd. Modeling human population separation history using physically phased genomes. *Genetics*, 205(1):385–395, 2017.

L. Speidel, M. Forest, S. Shi, and S. R. Myers. A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, 51(9):1321–1329, 2019.

J. P. Spence, M. Steinrücken, J. Terhorst, and Y. S. Song. Inference of population history using coalescent HMMs: review and outlook. *Current Opinion in Genetics and Development*, 53:70–76, 2018.

M. Steinrücken, J. P. Spence, J. A. Kamm, E. Wieczorek, and Y. S. Song. Model-based detection and analysis of introgressed Neanderthal ancestry in modern humans. *Molecular Ecology*, 27(19):3873–3888, 10 2018.

M. Steinrücken, J. Kamm, J. P. Spence, and Y. S. Song. Inference of complex population histories using whole-genome sequences from multiple populations. *Proceedings of the National Academy of Sciences*, 116(34):17115–17120, 2019.

M. Stephens and P. Donnelly. Inference in molecular population genetics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):605–635, 2000.

A. J. Stern, P. R. Wilton, and R. Nielsen. An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. *PLoS Genetics*, 15 (9):1–32, 2019.

M. Taboga. *Markov Chain Monte Carlo (MCMC) diagnostics*. Kindle Direct Publishing, 3rd edition, 2017.

S. Talts, M. Betancourt, D. Simpson, A. Vehtari, and A. Gelman. Validating Bayesian Inference Algorithms with Simulation-Based Calibration. *arXiv*, pages 1–19, 2020.

O. Tange. GNU Parallel - The Command-Line Power Tool. *;login: The USENIX Magazine*, 36(1):42–47, 2 2011.

J. Terhorst, J. A. Kamm, and Y. S. Song. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, 49(2):303–309, 2017.

The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

F. A. Villanea and J. G. Schraiber. Multiple episodes of interbreeding between Neanderthal and modern humans. *Nature Ecology & Evolution*, 3(1):39–44, 1 2019.

J. Wakeley and O. Sargsyan. Extensions of the coalescent effective population size. *Genetics*, 181(1):341–5, 1 2009.

J. Wang and M. C. Whitlock. Estimating effective population size and migration rates from genetic samples over space and time. *Genetics*, 163(1):429–446, 2003.

K. Wang, I. Mathieson, J. O'Connell, and S. Schiffels. Tracking human population structure through time from whole genome sequences. *PLoS Genetics*, 16(3):1–24, 2020.

H. M. Wilkinson-Herbots. Genealogy and subpopulation differentiation under various models of population structure. *Journal of Mathematical Biology*, 37(6):535–585, 1998.

P. R. Wilton, S. Carmi, and A. Hobolth. The SMC' is a highly accurate approximation to the ancestral recombination graph. *Genetics*, 200(1):343–355, 2015.

C. Wiuf and J. Hein. Recombination as a point process along sequences. *Theoretical Population Biology*, 55(3):248–259, 1999.

A. W. Wohns, Y. Wong, B. Jeffery, A. Akbari, S. Mallick, R. Pinhasi, N. Patterson, D. Reich, J. Kelleher, and G. McVean. A unified genealogy of modern and ancient genomes. *Science*, 375(6583), 2 2022.

S. Wright. Evolution in Mendelian Populations. *Genetics*, 16(2):97–159, 3 1931.

J. A. Zeidler. The Ecuadorian Formative. In H. Silverman and W. H. Isbell, editors, *The Handbook of South American Archaeology*, chapter 24. Springer, 2008.

Appendix A

Appendix of Chapter 1

A.1 Evaluating MCMC Convergence

To evaluate MCMC convergence in ARGweaver and Relate, we run these programs five independent times for the same simulated sequence of 5Mb. We do this for each simulation scenario and evaluate convergence by analysing various statistics extracted at each iteration. For ARGweaver, we analyse statistics from in the *.stats* file, described below. Relate does not generate a similar output, so we extract a subset of the pairwise coalescence times at each MCMC iteration to evaluate convergence. We also evaluate convergence based on selected pairwise coalescence times in ARGweaver, for comparison. Using these statistics extracted at each iteration, we evaluate MCMC convergence by analysing 1) trace plots, 2) autocorrelation plots, 3) effective sample sizes Taboga (2017); Roy (2020), and 4) potential scale reduction factor (PSRF) Gelman and Rubin (1992). Analyses and plots were done in R using the function *acf* for autocorrelation, and R package *coda* Plummer et al. (2006) for effective sample sizes and potential scale reduction factor. These results were used to inform our decisions on burn-in and thinning for MCMC, as well as interpreting results of our evaluations of the methods under different simulated conditions.

ARGweaver

Convergence of likelihoods ARGweaver's *arg-sample* program outputs a *.stats* file containing several statistics for each MCMC iteration: log probability of the sampled ARG given the model ("prior", in Table A.1), log probability of the data given the sampled ARG ("likelihood"), total log probability of the ARG and the data ("joint"), number of recombination events in the sampled ARG ("recombs"), the number of variant sites that cannot be explained by a single mutation under the sampled ARG ("noncompats"), total length of all branches summed across sites ("arglen") Hubisz and Siepel (2020). We generated trace plots and calculated autocorrelation between consecutive samples using the likelihood per iteration (Figures A.9 and A.11). Following visual inspection of these plots, we chose a burn-in consisting of the first 200 samples in most simulations, except in simulations with 10

times higher mutation rate (Figure A.9C,F) or sample sizes larger than 8 haplotypes (Figure A.11B,C,E,F), where we chose a burn-in of 1200 samples since those chains took longer to converge. In both cases, we ran MCMC for 1000 iterations after burn-in. Based on auto-correlation plots (Figure A.9, A.11) and on effective sample sizes (Table A.1), we thinned ARGweaver samples by recording every 10th MCMC iteration, thus retaining a total of 100 MCMC samples.

Results of the potential scale reduction factor suggested convergence of ARGweaver in simulations with mutation rate equal to recombination rate, with decreased recombination rate and with increased mutation rate (Table A.1) - see section below on convergence of individual coalescence times.

Convergence of coalescence times For comparison with Relate, which does not output statistics for each iteration, we also analyse convergence of pairwise coalescence times in ARGweaver. To this end, we extract from each MCMC iteration the values of coalescence times between two pairs of samples at 100 sites equally spaced by 50 kb along the 5Mb simulated sequences. We use those 200 values for convergence diagnostics. Figure A.12 shows trace plots of 10 of those sites, for one pair of samples. To evaluate convergence, we calculate potential scale reduction factor (PSRF) for each of the 200 coalescence times, and compare their mean, variance and range (Table A.2) among different simulations. In Table A.2 we also compare the number of coalescence times that have effective sample sizes lower than 100 (which is our MCMC sample size). These results also lead us to conclude that ARGweaver runs with mutation rate equal to recombination rate have converged. However, in contrast to the results on convergence for statistics recorded in the ARGweaver *stats* files (Table A.1), the evaluation of convergence based on coalescence times does not support a conclusion of full convergence for the other simulated data sets. In particular, simulations with mutation to recombination rate ratio of 10 had a large number of coalescence times with effective sample size smaller than 100. The same was true for simulations with 16 and 32 haplotypes. The maximum values of PSRF in those simulations are also further from one, thus indicating a lack of convergence for some coalescence times.

Relate

Relate estimates branch lengths using an MCMC algorithm with built in burn-in (Speidel et al. (2019) Supplementary Note on Method details 4.2, p. 13). To obtain samples from the posterior distribution, the tree sequence estimated in this first step was used as a starting point. Therefore, we did not implement any extra burn-in to obtain samples from the posterior. Visual inspection of traces plots also suggested that additional burn-in was not necessary (Figure A.13).

We evaluated Relate's MCMC convergence by running it 5 times for each sequence of 5Mb simulated under each set of parameters. We then extracted a subset of pairwise coalescence times to calculate the potential scale reduction factor and effective sample sizes as described above for ARGweaver. We extracted coalescence times for two pairs of samples at 100

equally spaced sites along the sequence (*i.e.* separated by 50kb). Table A.3 shows these results, which indicate convergence of all Relate runs in all simulated datasets.

A.2 Tsdate prior grid

We ran tsdate with different prior grids, using the function tsdate.build_prior_grid(). The observation that dates inferred by tsdate seem to be bounded to a low maximum value still holds when changing prior grids to have more points (timepoints=100, Figure A.14) or when manually specifying time slices with a maximum value of 12 (timepoints=np.geomspace(1e-5, 12, 50), Figure A.15).

A.3 ARGweaver subtree sampling acceptance rates

As suggested by ARGweaver authors (Melissa Hubisz and Adam Siepel, personal communication), we have verified that acceptance rates of subtree sampling steps of ARGweaver are within a range that indicates good mixing of the chain, between 10% and 90% (Table A.4). All simulations except for the one with reduced recombination rate were within that range. For a visualization of the spread of the values of acceptance rate, Figure A.16 shows the acceptance rates for subtree sampling steps of ARGweaver in one 5Mb region of each simulation.

Additional simulations results for ARGweaver

SMC and SMC' modes in ARGweaver

In all results shown in the main text, we simulated under the standard Hudson (1983) coalescent with recombination, and did inference in ARGweaver under SMC'. Here, we asked whether deviations observed in the posterior distribution of ARGweaver can be explained by differences between the models used for simulation and inference. For this, we simulate sequences in msprime under the SMC and SMC' models, and run ARGweaver inference using the same model used in the simulation. We simulated 8 haplotypes with mutation rate and recombination rate 2×10^{-8}. Results improve when simulating under SMC' and inferring under SMC' (Figures A.17B, A.18B). Surprisingly, simulating and inferring under SMC (Figures A.17A, A.18A) is not better than simulating under the full coalescent with recombination model and inferring under SMC (Figures 1.4, 1.5).

Intermediate values of mutation to recombination rate ratio

Rasmussen et al. (2014) mention in their Figure S5 that the quality of ARGweaver estimates generally improved in their simulations with increased mutation to recombination rates ratio (μ/ρ), but only up to $\mu/\rho = 4$. Motivated by this observation, we additionally ran simulations

with values of μ/ρ in between the ones shown in the main text (μ/ρ=1 or μ/ρ=10), including μ/ρ=2 and 4. We summarize our results under these conditions in Table A.5. We observed a similar pattern for these intermediate values of $\mu/\rho = 2, 4$ as we had observed from 1 to 10, *i.e.* point estimates improve with increased ratio (shown by lower MSE in Table A.5), and calibration of the posterior distribution worsens with an increased ratio (show by higher KLD in Table A.5).

Jukes-Cantor mutational model

In all results shown in the main text, we simulated mutations using an infinite sites model. ARGweaver, on the other hand, uses a Jukes and Cantor (1969) mutational model. Therefore, we hypothesize that differences in the mutational model between simulations and inference could explain deviations in the posterior distribution of ARGweaver, especially in simulations with increased mutation to recombination ratio (μ/ρ). We found that ARGweaver results with simulations under the Jukes and Cantor (1969) model are very similar to the results under the infinite sites model and follow the same pattern under increased μ/ρ (Table A.5, Figures A.20, A.21).

A.4 Supplementary figures

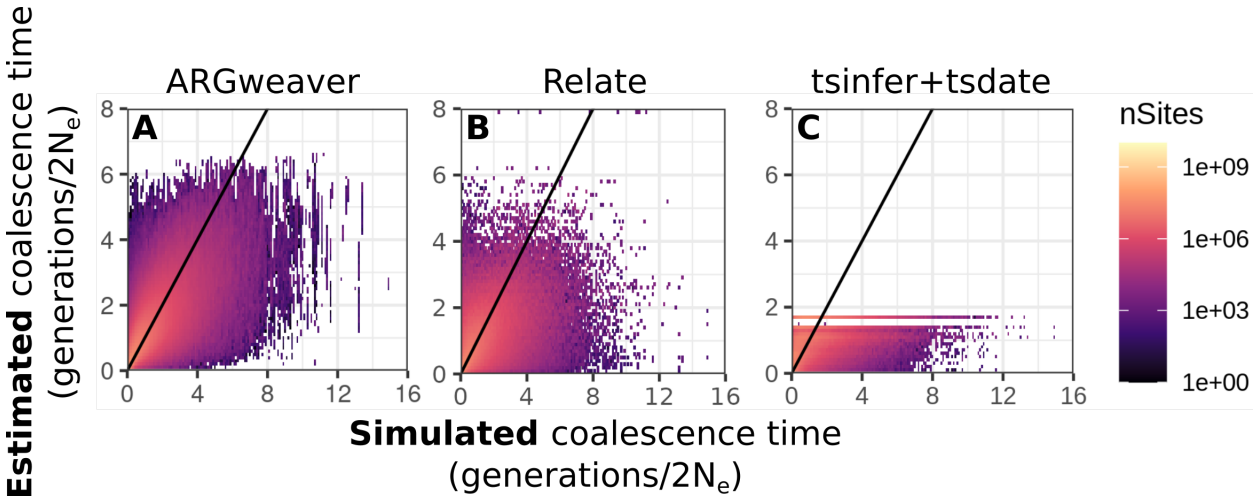

Figure A.1: True pairwise coalescence time from msprime simulations compared to inferred coalescence time from (A) ARGweaver (B) Relate (C) tsdate. Note that axes are in linear scale. See Figure 1.3A, D, G for these data plotted on a logarithmic scale. These results are for simulations with n=8 samples (haplotypes), mutation and recombination rates of 2×10^{-8}. Diagonal line shows x=y, points show the mean inferred coalescence time within a true coalescence time bin.

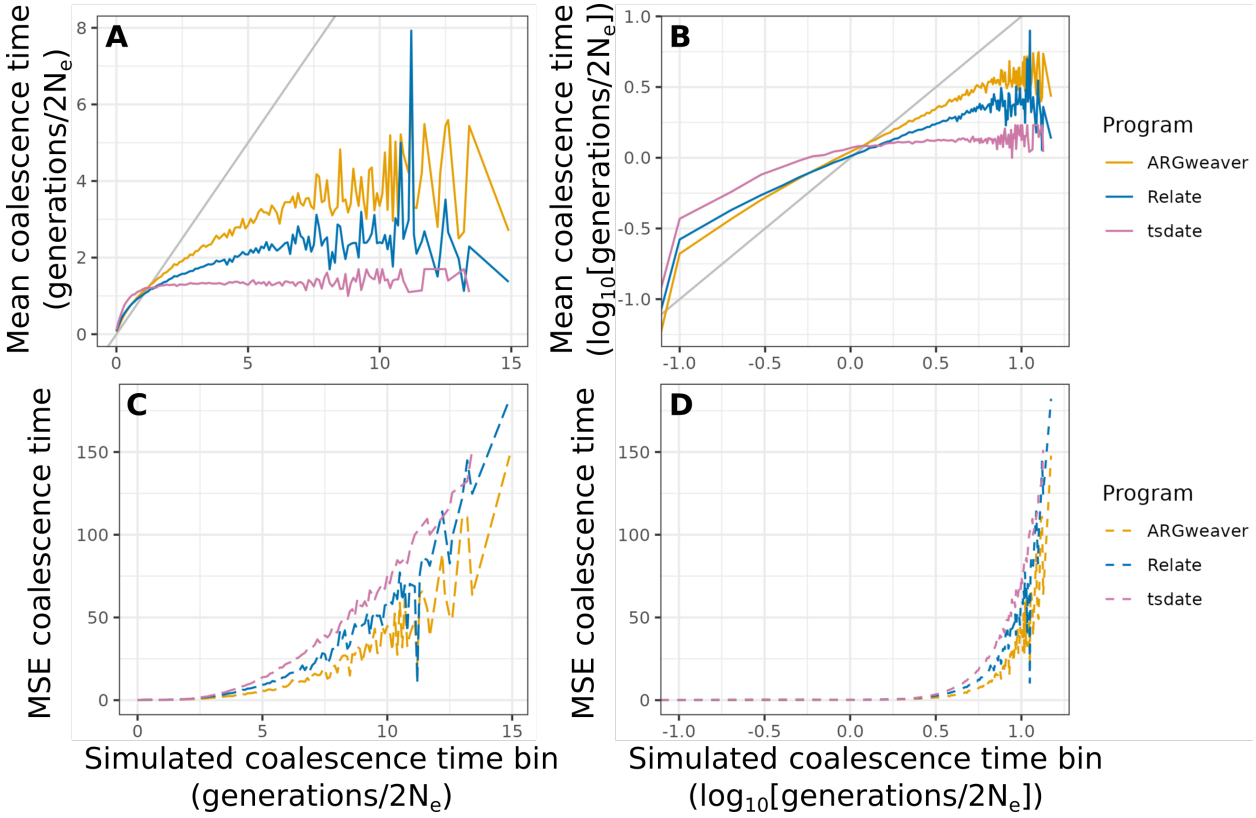

Figure A.2: Mean (A,B) and mean squared error (C,D) of point estimates of pairwise coalescence times by ARGweaver, Relate and tsdate in each bin of size 0.1 of simulated coalescence times. Diagonal gray line in plots A and B show 1:1 line. These results are for simulations with n=8 samples, mutation and recombination rates of 2×10^{-8}. Plots B and D are in log scale to highlight small values of coalescence times, which are the most abundant. Note that estimates are best (*i.e.* means in plots a and b are closer to the simulated value) at values near the expected mean coalescence time under the coalescent *(i.e.* 1 in the coalescent units of $2N_e$ generations).

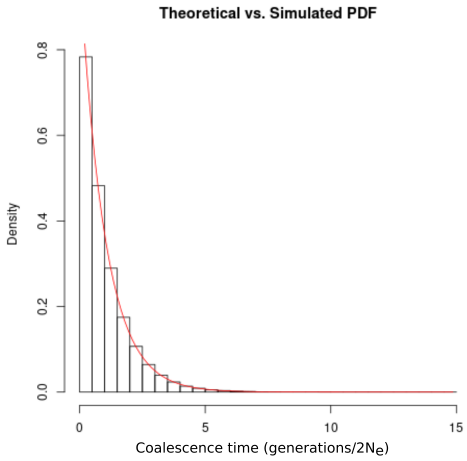

Figure A.3: Histogram of the distribution of coalescence times in msprime simulations. Red line shows expected exponential distribution with rate 1.

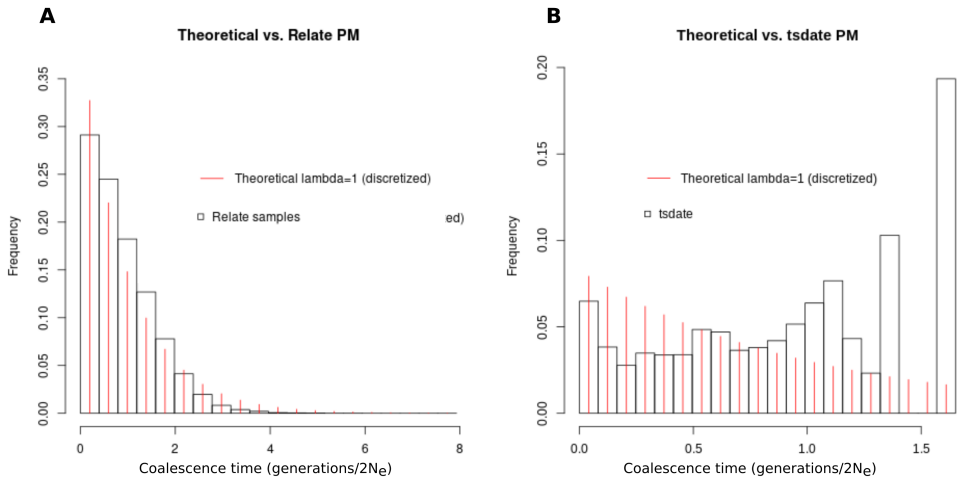

Figure A.4: Distributions of pairwise coalescence times in Relate and tsdate without ARG-weaver time discretization. These results are for simulations with n=8 samples, mutation and recombination rates of 2×10^{-8}. (A) Relate, (B) tsdate, both with 20 equal size bins .

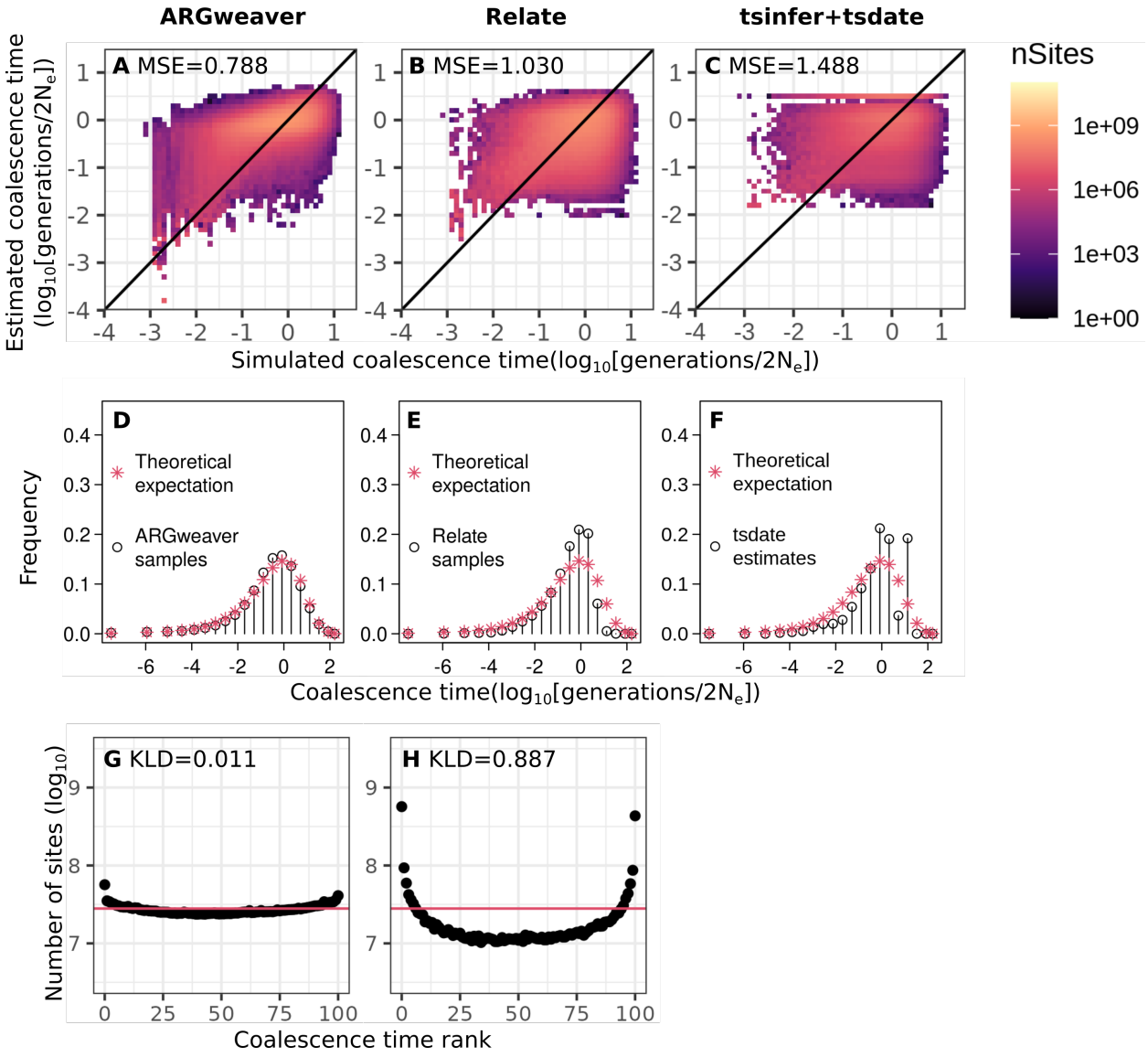

Figure A.5: Point estimates (A-C), distribution of coalescence times (D-F) and counts of ranks from simulation-based calibration (G,H) from ARGweaver (A,D,C), Relate (B,E,H) and tsinfer+tsdate (C,F). Simulations with **reduced mutation rate** ($\mu = 2 \times 10^{-9}$ and $\rho = 2 \times 10^{-8}$). Compared to simulations with mutation rate equal to recombination rate, mean square error (MSE) values are all larger (Figure 1.3), distributions of coalescence times deviate more from the theoretical expectation (Figure 1.4), and KLD is lower in ARGweaver but higher in Relate (Figure 1.5).

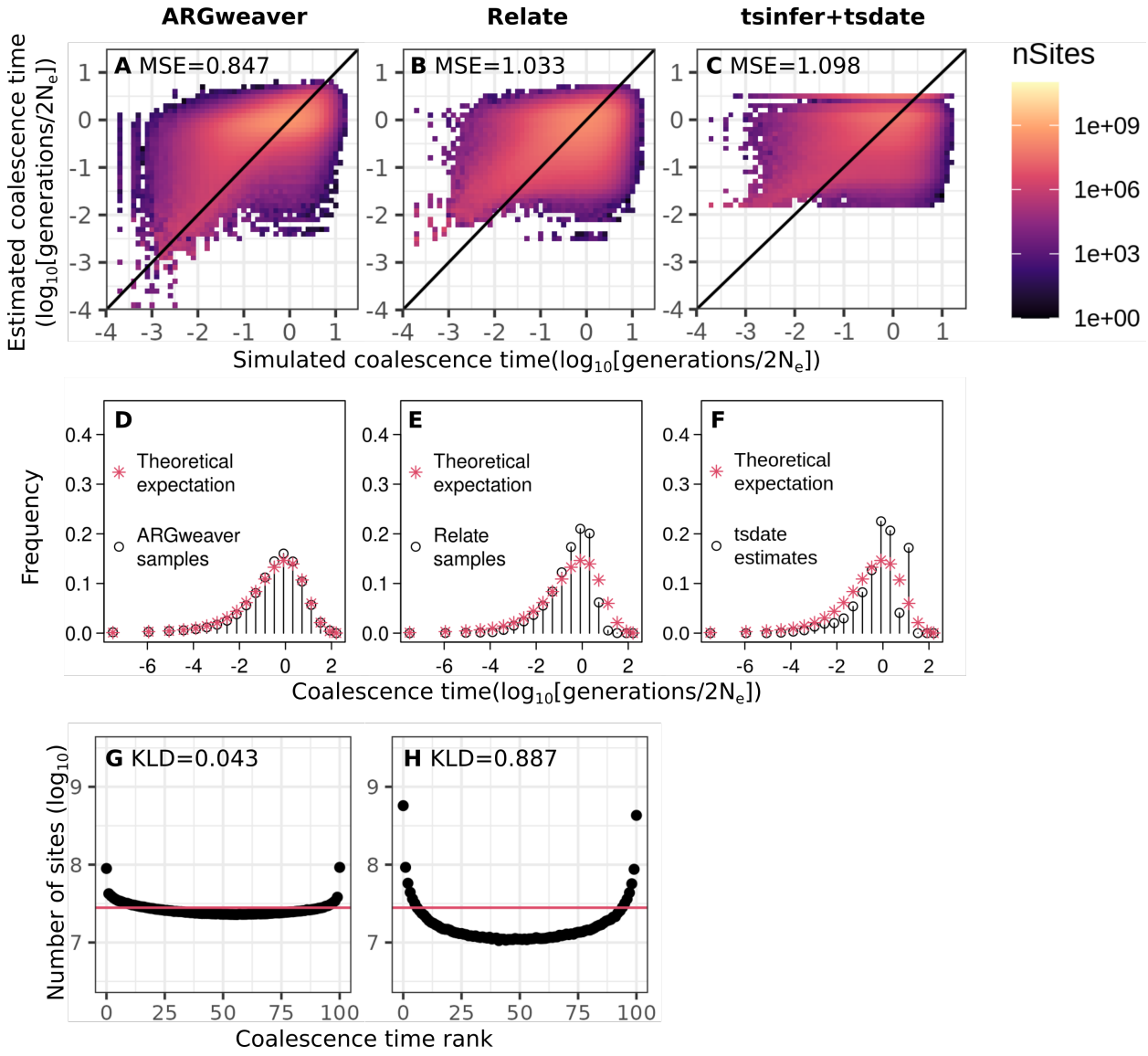

Figure A.6: Point estimates (A-C), distribution of coalescence times (D-F) and counts of ranks from simulation-based calibration (G,H) from ARGweaver (A,D,C), Relate (B,E,H) and tsinfer+tsdate (C,F). Simulations with **increased recombination rate** ($\mu = 2 \times 10^{-8}$ and $\rho = 2 \times 10^{-7}$). Compared to simulations with mutation rate equal to recombination rate, Mean square error (MSE) values are all larger (Figure 1.3), distributions of coalescence times deviate more from the theoretical expectation (Figure 1.4), and KLD is lower in ARGweaver, but higher in Relate (Figure 1.5).

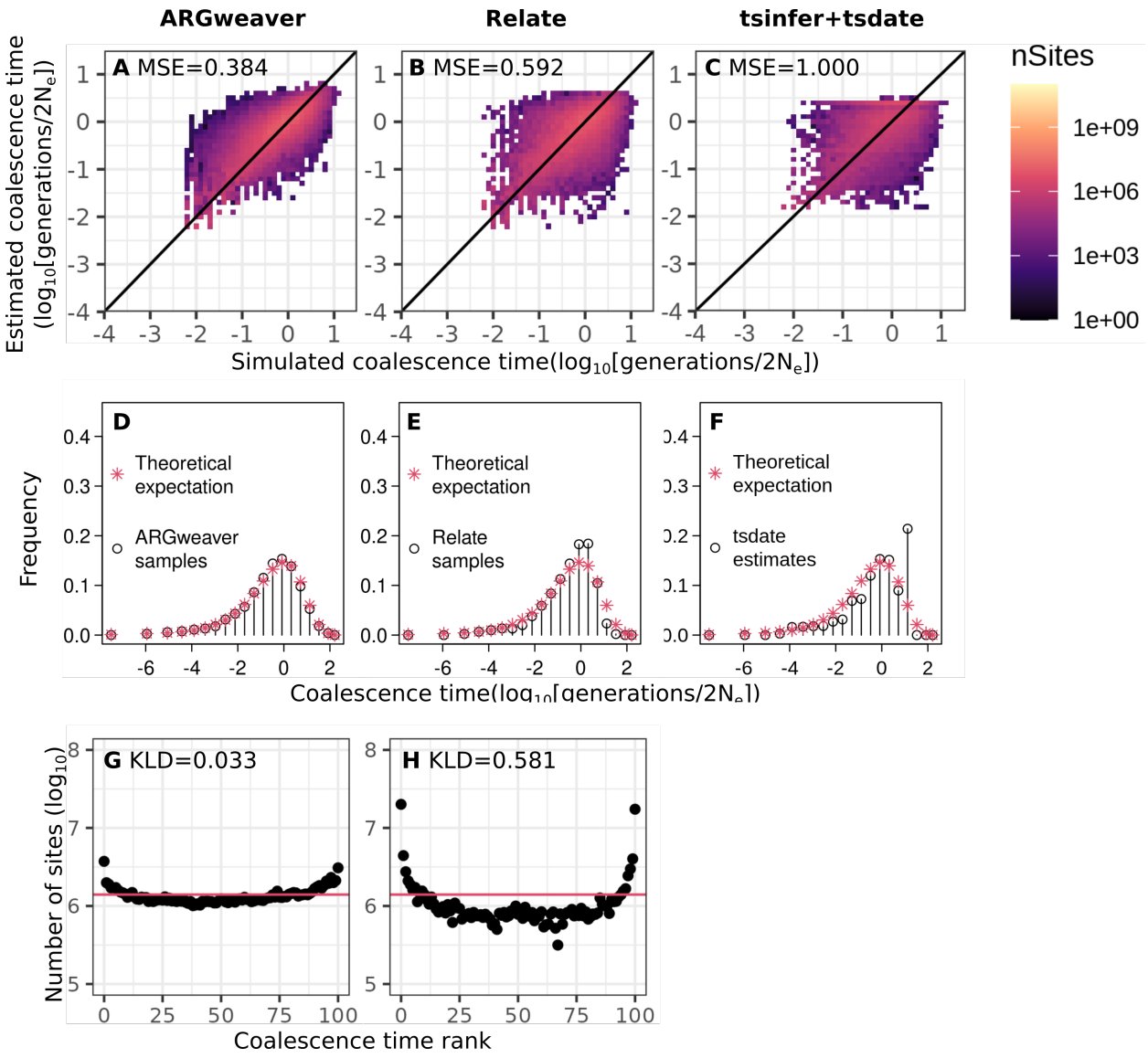

Figure A.7: Point estimates (A-C), distribution of coalescence times (D-F) and counts of ranks from simulation-based calibration (G,H) from ARGweaver (A,D,C), Relate (B,E,H) and tsinfer+tsdate (C,F). Simulations with sample size of 8 haplotypes, $\mu = \rho = 2 \times 10^{-8}$, and **input sequence length of 5Mb**.

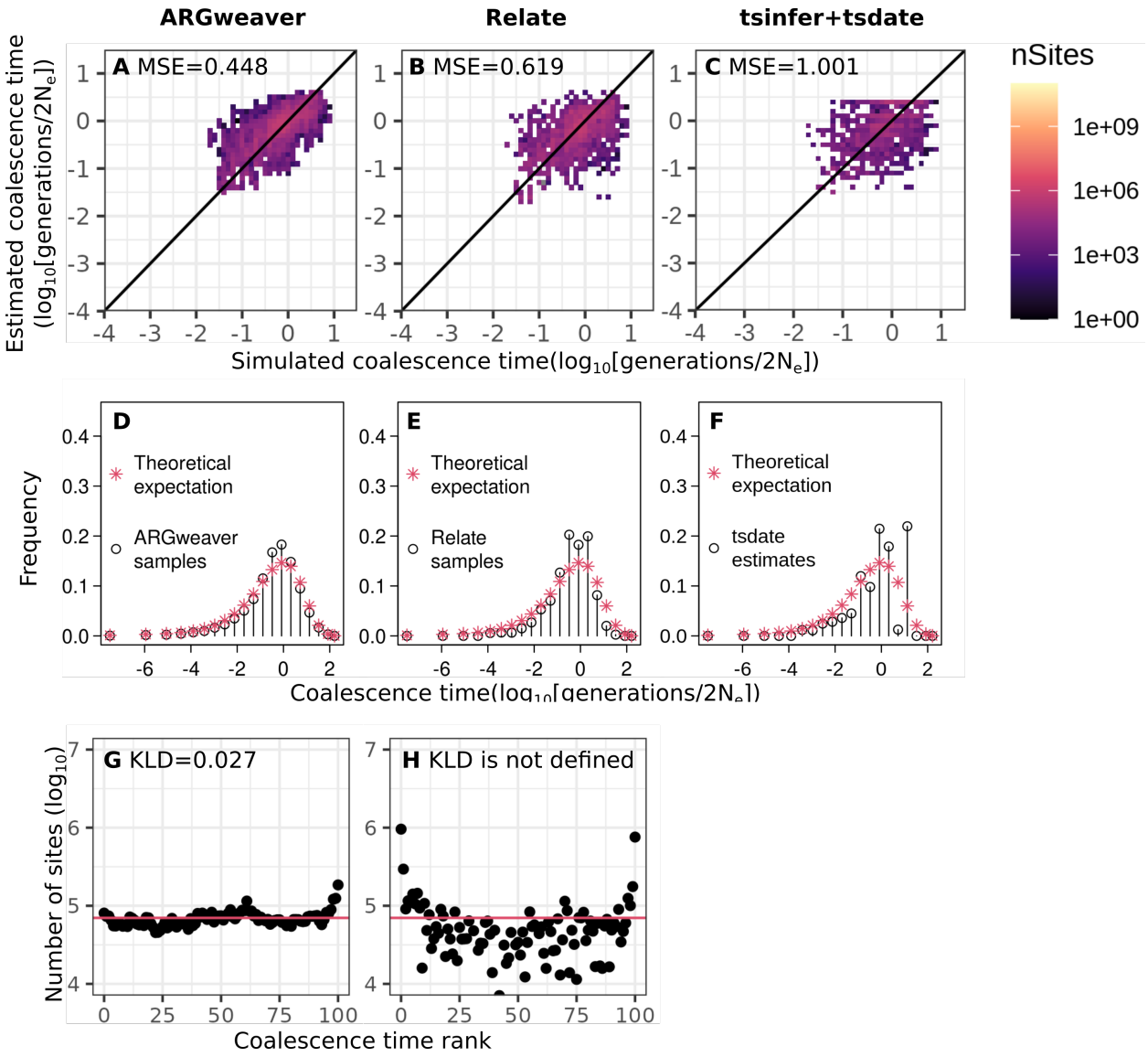

Figure A.8: Point estimates (A-C), distribution of coalescence times (D-F) and counts of ranks from simulation-based calibration (G,H) from ARGweaver (A,D,C), Relate (B,E,H) and tsinfer+tsdate (C,F). Simulations with sample size of 8 haplotypes, $\mu = \rho = 2 \times 10^{-8}$, and **input sequence length of 250kb**. In H, KLD is not defined because counts for one of the ranks is zero.

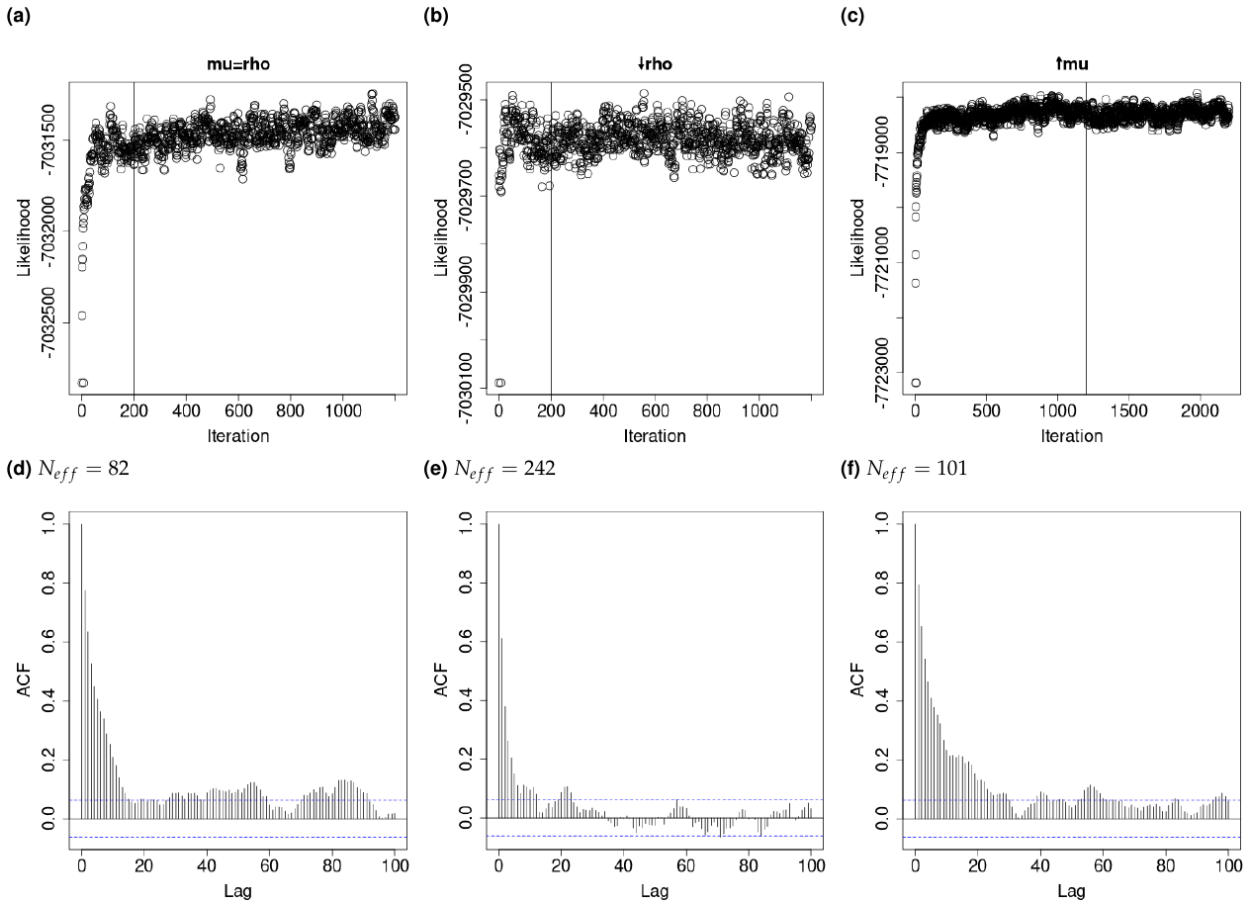

Figure A.9: ARGweaver likelihood traces (top) and autocorrelation between consecutive MCMC iterations (bottom, also showing effective sample sizes (N_{eff})) for the number of iterations used in the main text. Left column: simulations with 8 haplotypes, mutation rate equal to the recombination rate(2×10^{-8}). Potential scale reduction factor (PSRF) is 1.02, upper confidence interval (CI) is 1.05. Middle column: simulations with recombination rate decreased to 2×10^{-9}. PSRF is 1.04, upper CI is 1.11. For both of these simulated datasets we used a burn in of 200 iterations (indicated by vertical line) and ran them for 1200 iterations in total, sampling every 10th iteration. Right column: simulations with mutation rate increased to 2×10^{-7}. PSRF is 1.01, upper CI is 1.02. For this dataset we used a burn in of 1200 iterations (indicated by vertical line) and ran them for 2200 iterations in total, sampling every 10th iteration.

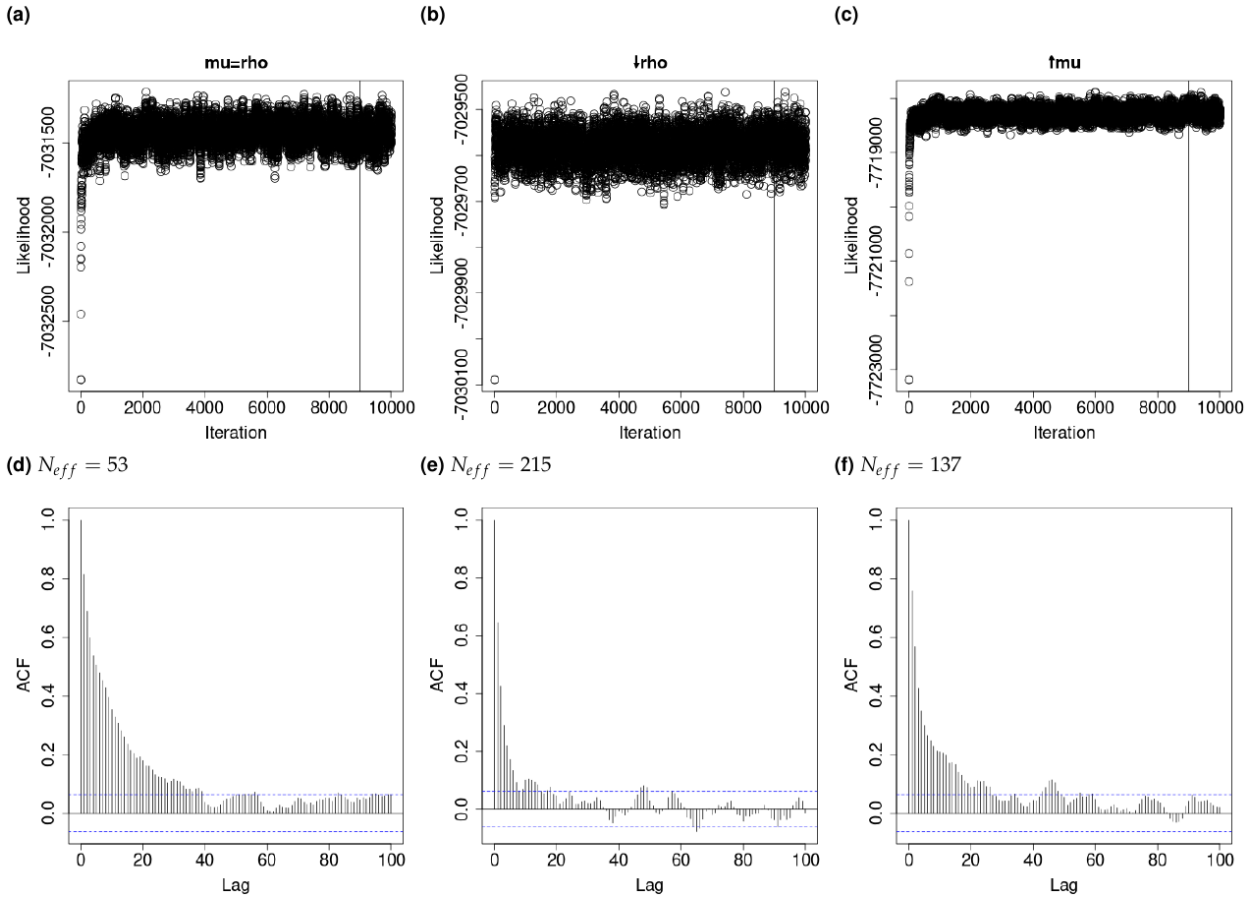

Figure A.10: Similar to Figure A.9, but running ARGweaver for 10 thousand iterations, with a burn in of 9 thousand applied before calculating effective sample sizes, to keep the same number of samples (1000). ARGweaver likelihood traces (A,B,C) and autocorrelation between consecutive MCMC iterations (D,E,F). Left column: simulations with 8 haplotypes, mutation rate equal to the recombination rate (2×10^{-8}). Middle column: simulations with recombination rate decreased to 2×10^{-9}. Right column: simulations with mutation rate increased to 2×10^{-7}.

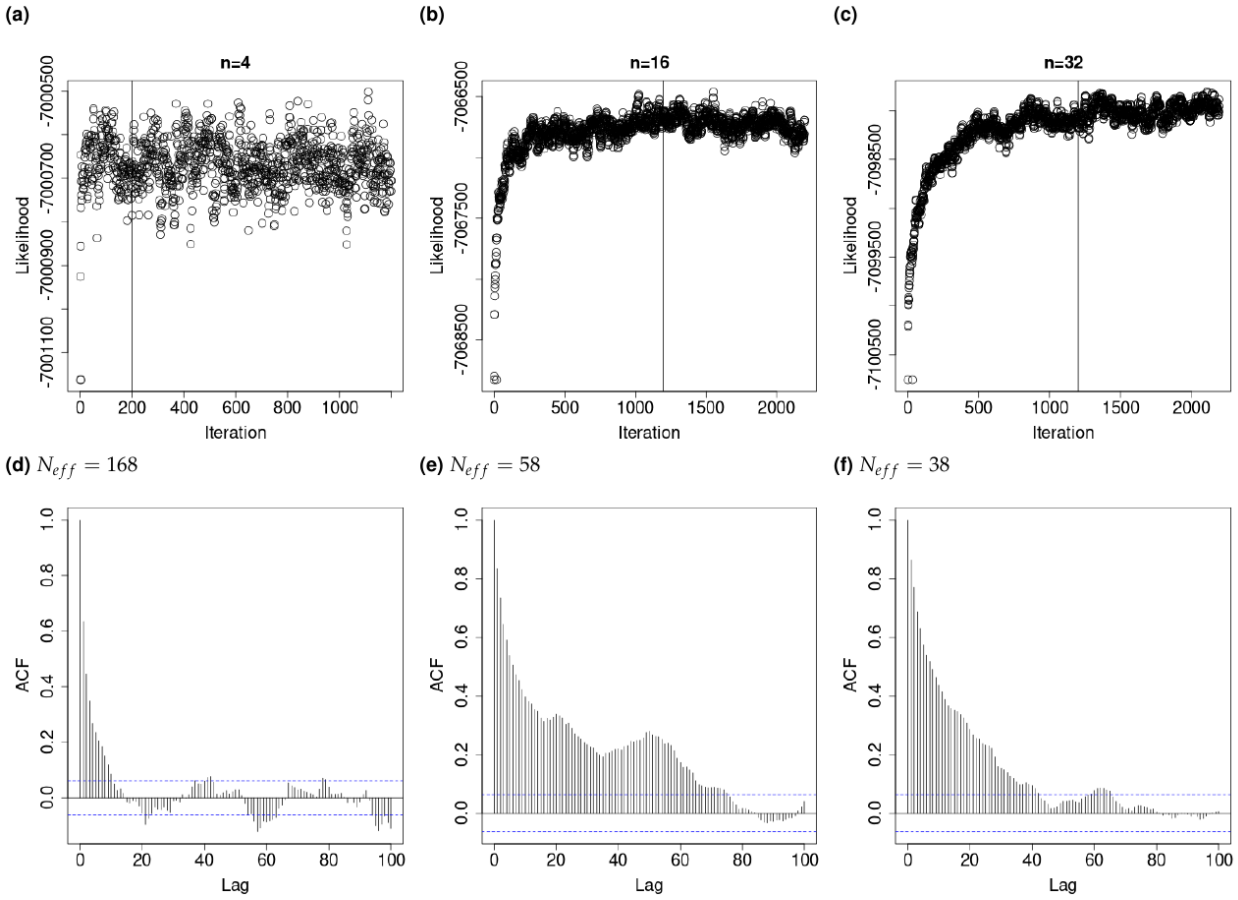

Figure A.11: ARGweaver likelihood traces (top) and autocorrelation between consecutive MCMC iterations (bottom). A,D: simulations with 4 haplotypes, mutation rate equal to recombination rate (2×10^{-8}). For this simulated dataset we used a burn in of 200 iterations (indicated by vertical line) and ran them for 1200 iterations in total, sampling every 10th iteration. B,E: simulations with 16 haplotypes. C,F: simulations with 32 haplotypes. For both of these datasets we used a burn in of 1200 iterations (indicated by vertical line) and ran them for 2200 iterations in total, sampling every 10th iteration.

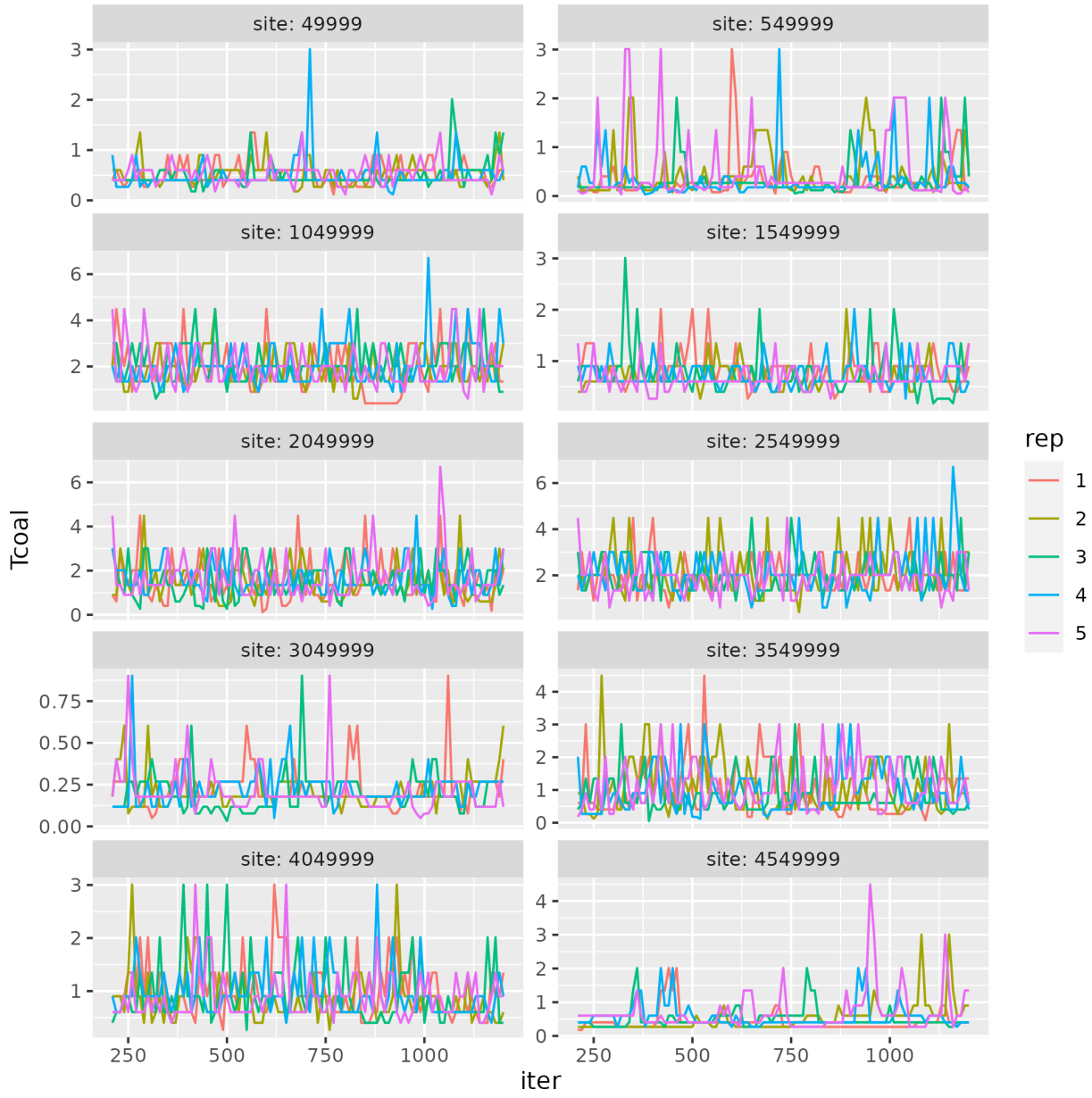

Figure A.12: Coalescence times for one pair of samples inferred by 5 independent runs of ARGweaver at 10 sites equally spaced sites along the 5Mb sequence. Simulations with 8 samples and mutation rate equal to recombination rate.

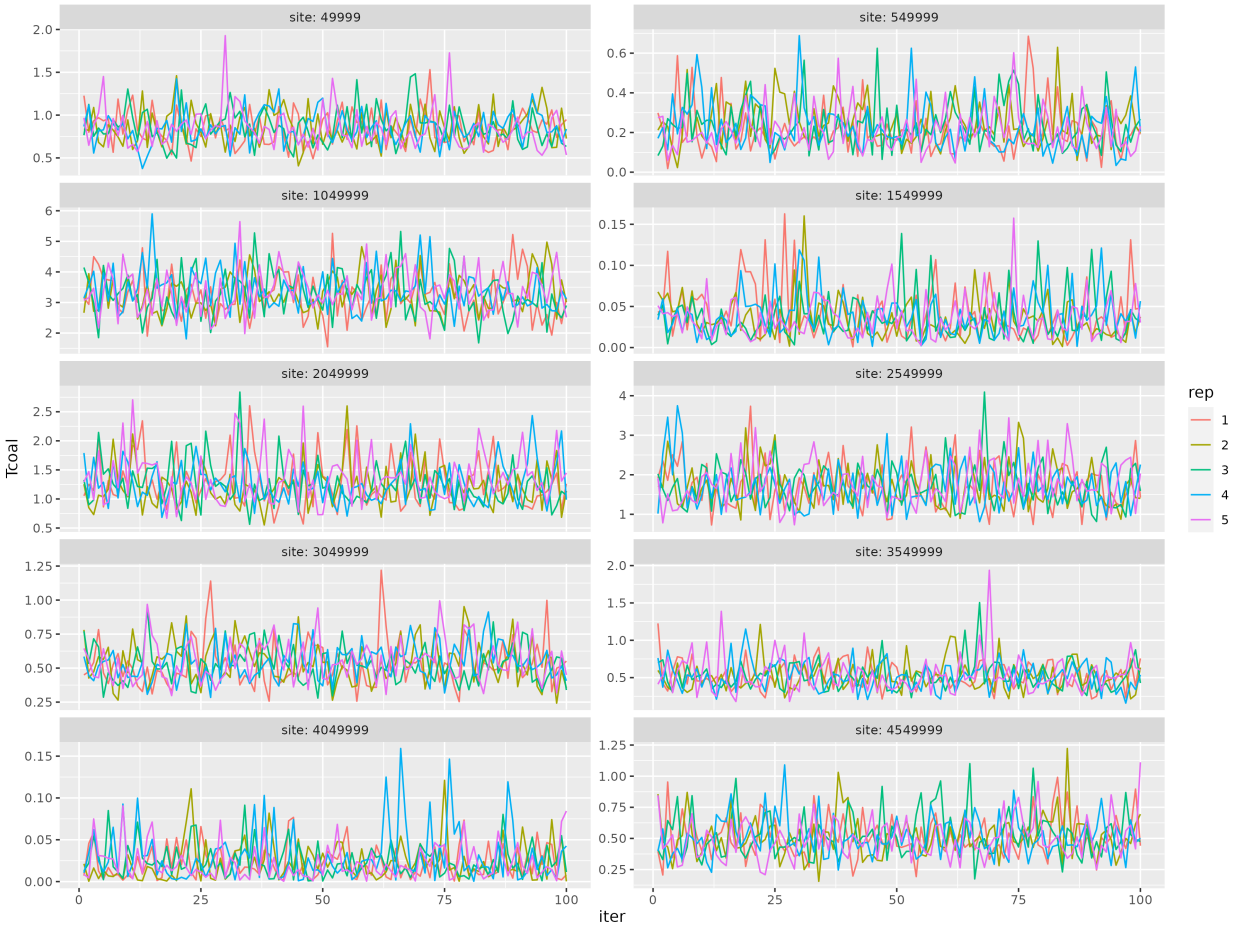

Figure A.13: Coalescence times for one pair of samples inferred by 5 independent runs of Relate at 10 sites equally spaced sites along the 5Mb sequence. Simulations with 8 samples and mutation rate equal to recombination rate.

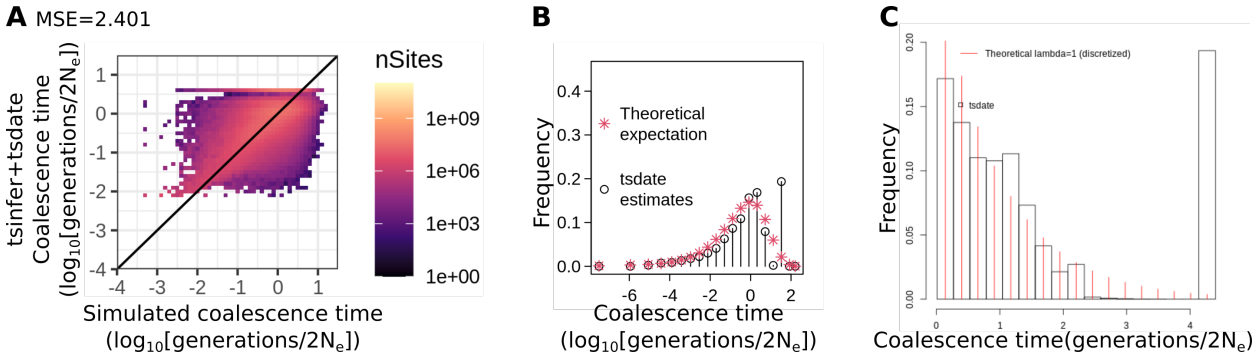

Figure A.14: Tsdate results with a **prior grid constructed with timepoints=100**. (A) Comparisons of estimated and simulated point estimates of pairwise coalescence times. (B) Comparisons of the distribution of coalescence times to the expected exponential distribution, using ARGweaver time discretization bins. (C) Same as B, but without imposing ARGweaver time discretization.

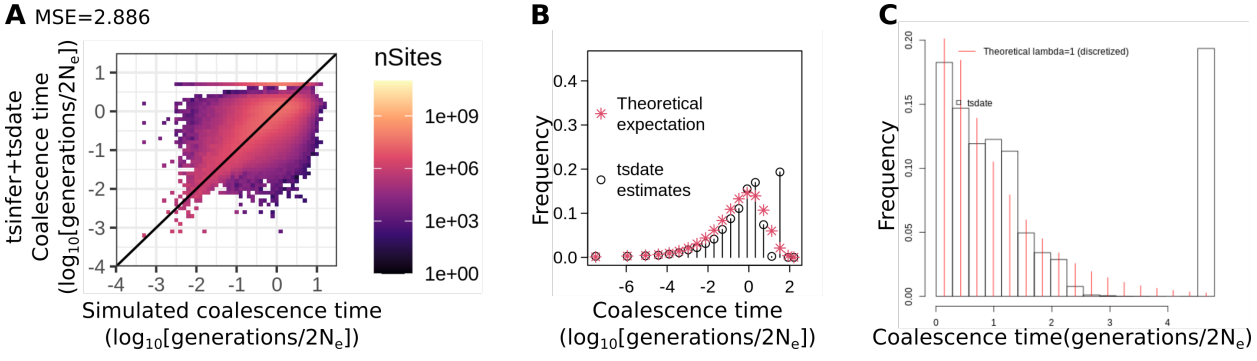

Figure A.15: Tsdate results with a **prior grid constructed with a maximum value of 12**. (A) Comparisons of estimated and simulated point estimates of pairwise coalescence times. (B) Comparisons of the distribution of coalescence times to the expected exponential distribution, using ARGweaver time discretization bins. (C) Same as B, but without imposing ARGweaver time discretization.

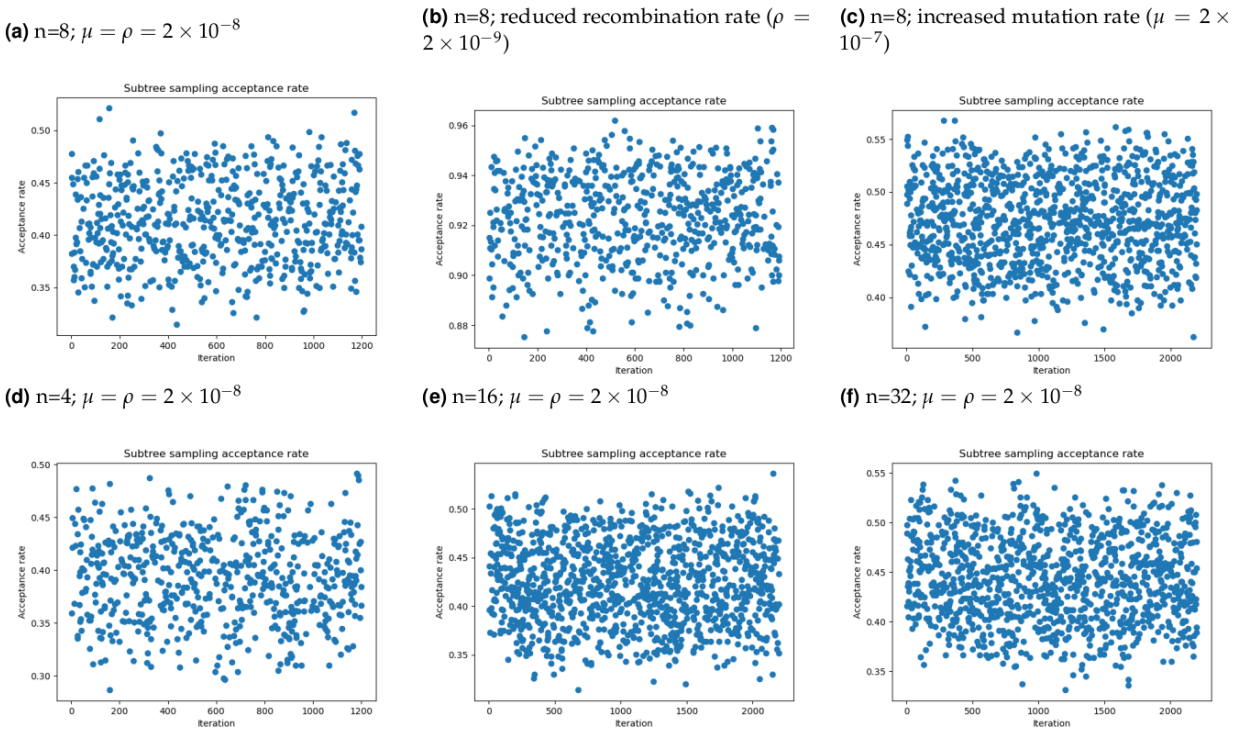

Figure A.16: Acceptance rate from ARGweaver subtree sampling steps in one 5Mb region of each simulation.

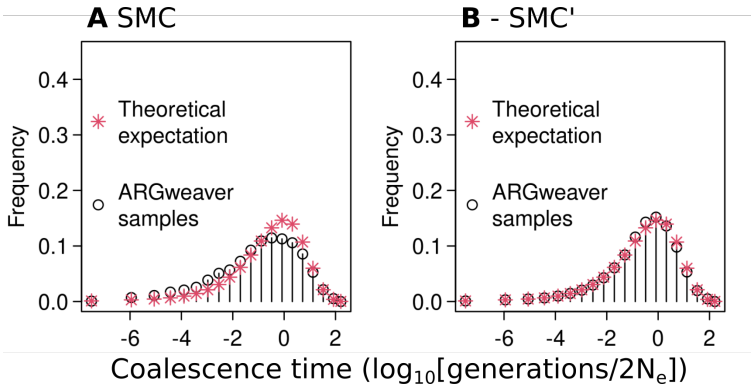

Figure A.17: Distribution of coalescence times in msprime simulations using the SMC (A) or SMC' model (B). ARGweaver inference is done using the same model used in the simulations.

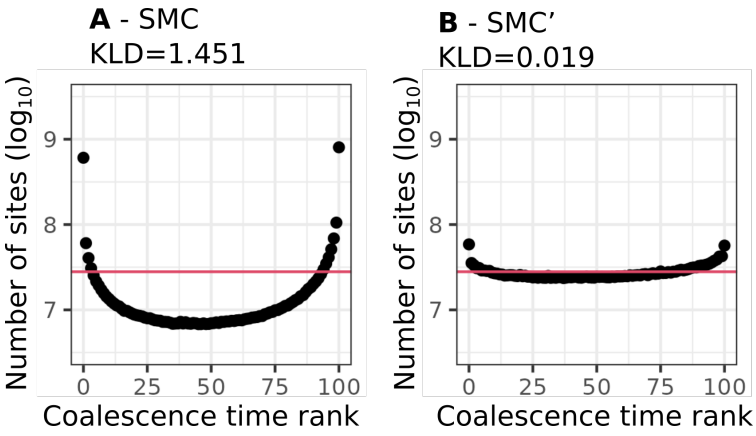

Figure A.18: Simulation-based calibration results in msprime simulations using the SMC (A) or SMC' model (B). ARGweaver inference is done using the same model used in the simulations.

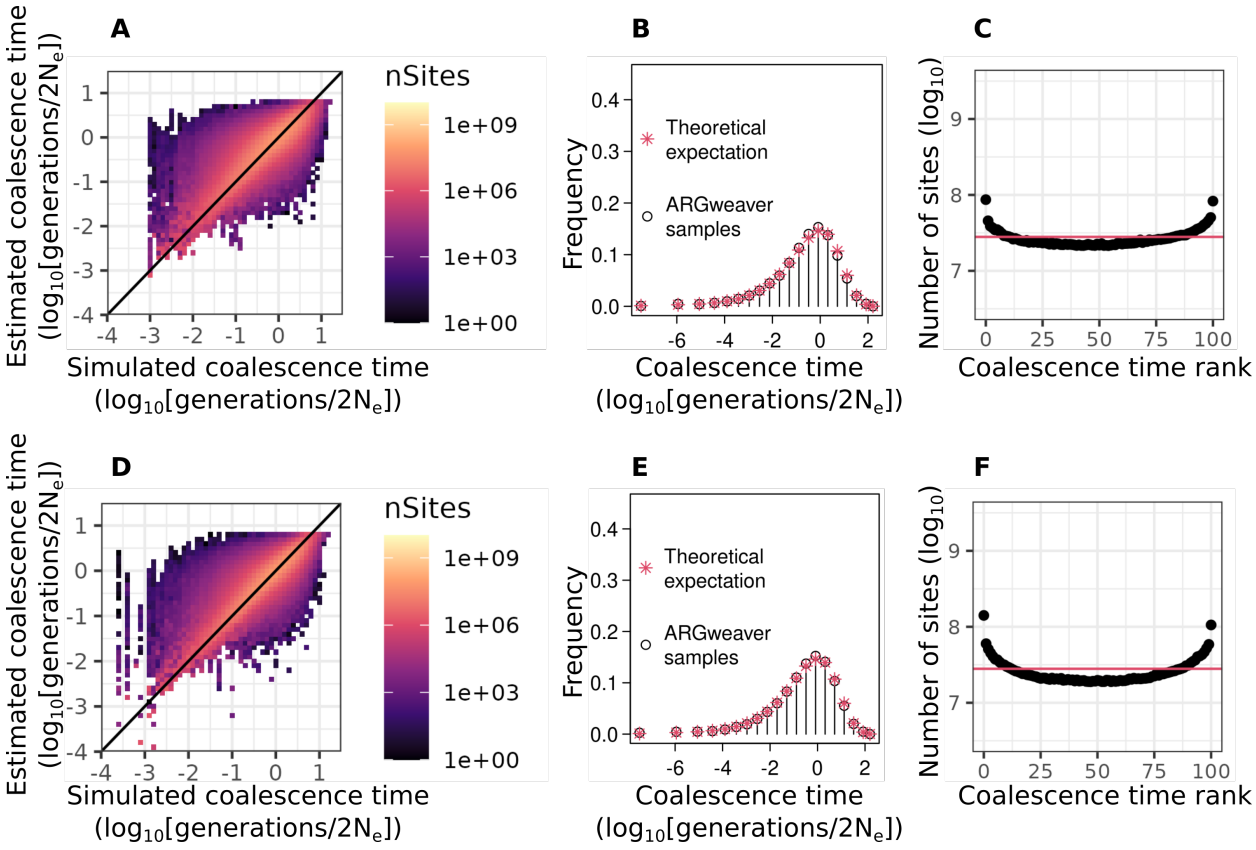

Figure A.19: Evaluation of ARGweaver point estimates (A,D), distribution of coalescence times (B,E) and posterior calibration (C,F) for simulations with mutation rate to recombination rate ratio of 2 (A-C, $\mu = 4 \times 10^{-8}$, $\rho = 2 \times 10^{-8}$) and mutation rate to recombination rate ratio of 4 (D-F, $\mu = 8 \times 10^{-8}$, $\rho = 2 \times 10^{-8}$)

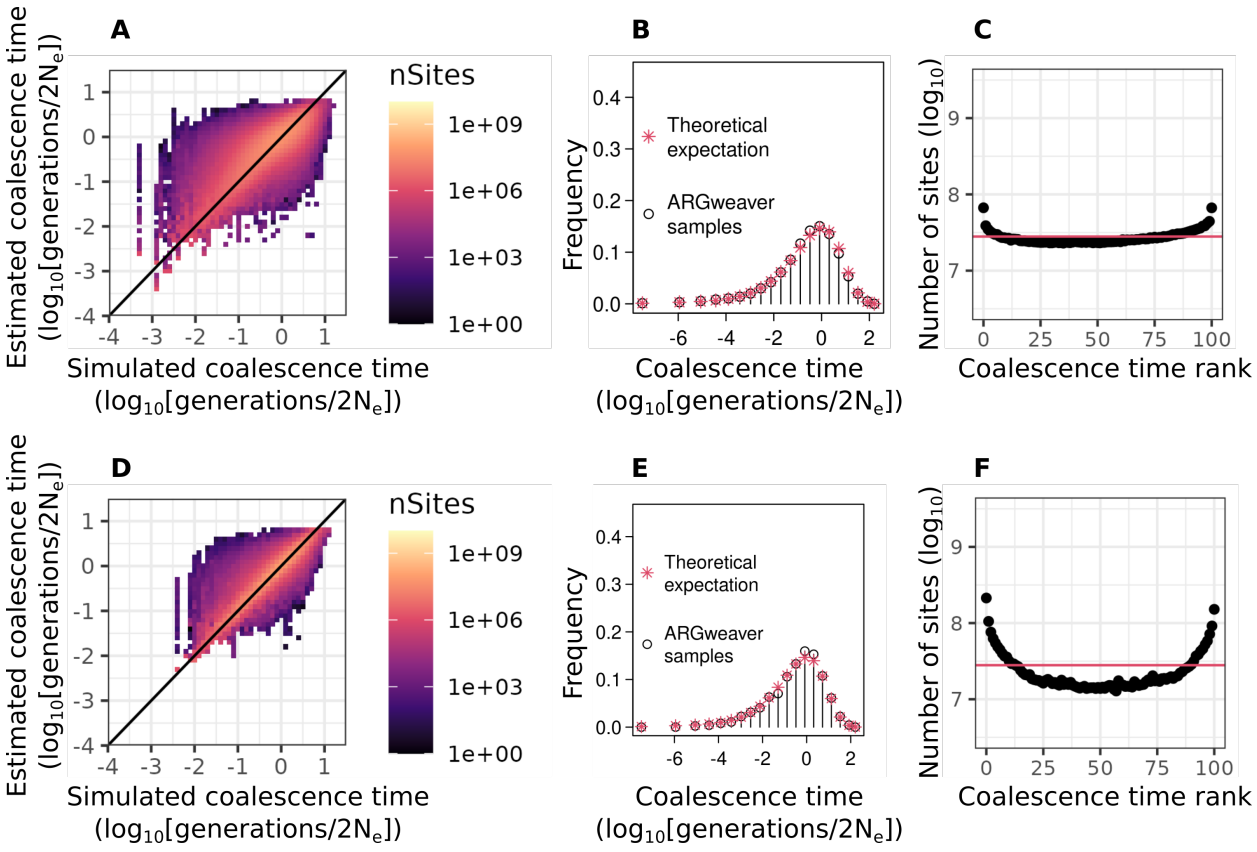

Figure A.20: Evaluation of ARGweaver point estimates (A,D), distribution of coalescence times (B,E) and posterior calibration (C,F) with simulations under the Jukes and Cantor mutational model. A-C: simulations with 8 haplotypes and $\mu = \rho = 2 \times 10^{-8}$. D-F: simulations with 8 haplotypes and $\mu = 2 \times 10^{-8}$ and $\rho = 2 \times 10^{-9}$

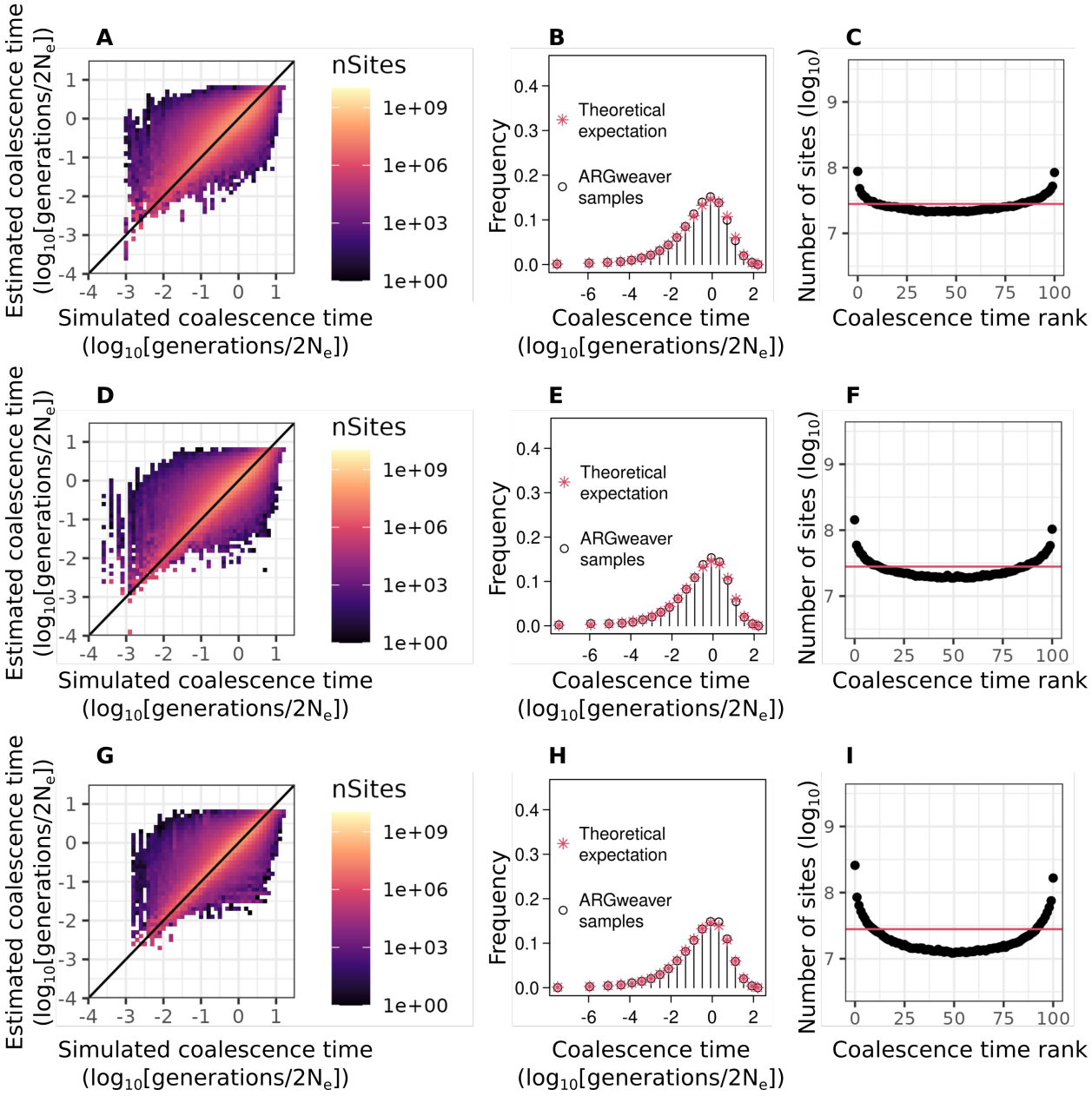

Figure A.21: Evaluation of ARGweaver point estimates (A,D,G), distribution of coalescence times (B,E,H) and posterior calibration (C,F,I) with simulations under the Jukes and Cantor mutational model. In all cases we simulated 8 haplotypes and used $\rho = 2 \times 10^{-8}$. A-C: $\mu = 4 \times 10^{-8}$. D-F: $\mu = 8 \times 10^{-8}$. G-I: $\mu = 2 \times 10^{-7}$

A.5 Supplementary tables

Table A.1: Potential scale reduction factor point estimates (PSRF), their upper confidence intervals (C.I.) and effective sample sizes (N_{eff}) for ARGweaver stats. μ: mutation rate, ρ: recombination rate.

	$\mu = \rho = 2 \times 10^{-8}$			$\rho = 2 \times 10^{-9}$			$\mu = 2 \times 10^{-7}$		
	PSRF	C.I.	N_{eff}	PSRF	C.I.	N_{eff}	PSRF	C.I.	N_{eff}
prior	1.06	1.15	224	1.01	1.03	494	1.00	1.01	216
likelihood	1.02	1.05	294	1.04	1.11	964	1.01	1.02	499
joint	1.06	1.16	216	1.01	1.02	486	1.01	1.02	219
recombs	1.04	1.1	254	1.01	1.03	559	1.00	1.01	229
noncompats	1.02	1.04	406	1.01	1.03	1290	1.01	1.04	518
arglen	1.06	1.16	348	1.08	1.21	459	1.05	1.12	319
Multivariate	1.15			1.1			1.05		

Table A.2: Potential scale reduction factor (PSRF) mean, variance and range for each of 200 coalescence times in ARGweaver, the multivariate PSRF Plummer et al. (2006) and the number of coalescence times for each the effective sample size (N_{eff}) is smaller than 100. Unless otherwise noted, mutation rate (μ) and recombination rate (ρ) are 2×10^{-8} and sample sizes (n) are 8 haplotypes.

PSRF	$\mu = \rho$	$\rho = 2 \times 10^{-9}$	$\mu = 2 \times 10^{-7}$	n=4	n=16	n=32
Mean	1.055	1.069	1.211	1.028	1.242	244.152
Variance	0.005	0.010	1.053	0.001	0.233	11613699
Range	0.994 - 1.415	0.994 - 1.709	0.994 - 13.740	0.991 - 1.199	1.001 - 4.847	0.994 - 4.783×10^4
Multivariate	4.92	4.29	21.2	2.78	24.9	110560
Number of $N_{eff} < 100$ (out of 200)	4	16	14	0	32	45

Table A.3: Potential scale reduction factor (PSRF) mean, variance and range for each of 200 coalescence times in Relate, the multivariate PSRF Plummer et al. (2006) and the number of coalescence times for each the effective sample size (N_{eff}) is smaller than 100. Unless otherwise noted, mutation rate (μ) and recombination rate (ρ) are 2×10^{-8} and sample sizes (n) are 8 haplotypes.

PSRF	$\mu = \rho$	$\rho = 2 \times 10^{-9}$	$\mu = 2 \times 10^{-7}$	n=4	n=16	n=32
Mean	1.007	1.007	1.008	1.008	1.009	1.008
Variance	10^{-4}	8.7×10^{-5}	10^{-4}	10^{-4}	10^{-4}	10^{-4}
Range	0.991 - 1.076	0.993 - 1.051	0.992 - 1.051	0.991 - 1.102	0.992 - 1.061	0.993 - 1.049
Multivariate	2.24	2.12	2.57	2.24	3.31	2.49
Number of $N_{eff} < 100$ (out of 200)	0	0	0	0	0	0

Table A.4: Minimum and maximum acceptance rates of ARGweaver subtree sampling steps for each simulation.

Simulation	Acceptance rates Min	Max
n=8; $\mu = \rho = 2 \times 10^{-8}$	0.283	0.532
n=8; $\mu = 2 \times 10^{-8}$; $\rho = 2 \times 10^{-9}$	0.838	0.965
n=8; $\mu = 2 \times 10^{-7}$; $\rho = 2 \times 10^{-8}$	0.345	0.582
n=4; $\mu = \rho = 2 \times 10^{-8}$	0.262	0.511
n=16; $\mu = \rho = 2 \times 10^{-8}$	0.299	0.567
n=32; $\mu = \rho = 2 \times 10^{-8}$	0.307	0.568

Table A.5: Comparison of ARGweaver results with simulations under infinite sites mutational model and Jukes-Cantor finite sites mutational model, including simulations with values of mutation to recombination rate ratio in between the ones shown in the main text. * indicate results shown in the main text and presented here again for comparison.

μ/ρ	Point estimates (MSE) Infinite sites	Finite sites (JC)	Ranks (KLD) Infinite sites	Finite sites (JC)
$\frac{2\times10^{-8}}{2\times10^{-8}} = 1$	0.397*	0.396	0.027*	0.026
$\frac{4\times10^{-8}}{2\times10^{-8}} = 2$	0.285	0.285	0.049	0.053
$\frac{8\times10^{-8}}{2\times10^{-8}} = 4$	0.195	0.197	0.113	0.112
$\frac{2\times10^{-7}}{2\times10^{-8}} = 10$	0.117*	0.120	0.350*	0.353
$\frac{2\times10^{-8}}{2\times10^{-9}} = 10$	0.120*	0.119	0.286*	0.291

Appendix B

Appendix of Chapter 2

B.1 Supplementary Methods

Markov process describing state of lineages in a coalescence tree with four tips

Let us encode a lineage state at time t by (k, l, i), where $k, l \in \{0, 1, 2\}$, $k + l > 0$ are the number of descendants of this lineage that were sampled (at time $t = 0$) from populations 1 and 2 respectively, and $i \in \{1, 2\}$ is the population where the lineage is at time t. We do not consider equations corresponding to states with a single lineage, *i.e.* before the most recent common ancestor of the 4 samples ($\{2, 2, i\}, i = 1, 2$), because they do not contribute to variable sites.

Before split time there is one single ancestral population, so a similar approach holds but the last index i is not needed to encode a lineage. So, there are only 8 possible states of the Markov process, and an additional absorbing state $(2, 2)$ which does not contribute to SFS. We proceed with the derivation of the case of two ancestral populations, as it is a more complex one.

Every Markov state $\{(k_j, l_j, i_j)\}$ is a set of lineages (enumerated with the index j, $1 < j \leq 4$) with the condition $\sum_j k_j = \sum_j l_j = 2$. At the time of observation $t = 0$ the initial state is $\{(1, 0, 1), (1, 0, 1), (0, 1, 2), (0, 1, 2)\}$. Two lineages (k_1, l_1, i_1) and (k_2, l_2, i_2) can coalesce only if $i_1 = i_2$, and the resulting lineage is $(k_1 + k_2, l_1 + l_2, i_1)$.

Let us consider the state $L = \{(1, 1, 1), (1, 0, 1), (0, 1, 2)\}$, and write the equation for the derivative $P_L(t)$ which is the change in the probability of the Markov process being in state L at time t.

Transitions into state L are possible from the following four states

- $L_1 = \{(1, 0, 1), (1, 0, 1), (0, 1, 1), (0, 1, 2)\}$ through coalescence of any of two lineages $(1, 0, 1)$ and the lineage $(0, 1, 1)$ with the total rate of coalescence $2/N_{L1}$,

- $L_2 = \{(1,1,2),(1,0,1),(0,1,2)\}$ through migration of the lineage $(1,1,2)$ from population 2 into population 1 with the migration rate m_{21},
- $L_3 = \{(1,1,1),(1,0,2),(0,1,2)\}$ through migration of the lineage $(1,0,2)$ with the rate m_{21},
- $L_4 = \{(1,1,1),(1,0,1),(0,1,1)\}$ through migration of the lineage $(0,1,1)$ with the rate m_{12}.

Transitions from state L are possible into four states

- $L_5 = \{(2,1,1),(0,1,2)\}$ through coalescence of the lineages $(1,1,1)$ and $(1,0,1)$ with the coalescence rate $1/N_{L1}$,
- $L_2 = \{(1,1,2),(1,0,1),(0,1,2)\}$ through migration of the lineage $(1,1,1)$ from population 1 into population 2 with the migration rate m_{12},
- $L_3 = \{(1,1,1),(1,0,2),(0,1,2)\}$ through migration of the lineage $(1,0,1)$ with the rate m_{12},
- $L_4 = \{(1,1,1),(1,0,1),(0,1,1)\}$ through migration of the lineage $(0,1,2)$ with the rate m_{21}.

So, the corresponding equation is

$$P'_L(t) = -\left(\frac{1}{N_{L1}} + 2m_{12} + m_{21}\right) P_L(t) + \frac{2}{N_{L1}} P_{L1}(t) + m_{21}P_{L2}(t) + m_{21}P_{L3}(t) + m_{12}P_{L4}(t).$$

Mutation on a lineage (k,l,i) contributes to the $f_{k,l}$ entry of the SFS. More specifically, $f_{k,l}$ is proportional to the total probability (from $t = 0$ to infinity) of lineages $(k,l,1)$ and $(k,l,2)$.

Assume that in the matrix form the equation has the form

$$P'(t) = M(t)P(t),$$

where P is the vector of probabilities of states, and M is a transition matrix depending on coalescence and migration rates. In order to calculate SFS, we need to compute the corresponding integrals $\int_0^\infty P(t)dt$ of the time spent in each of the states. We assume that the local effective population sizes and migration rates are picewise constant, hence M is piecewise constant too. On each time interval $[t_0, t_1]$ the solution of the matrix equation is $P(t) = \exp(Mt)P(t_0)$, and the integral

$$\int_{t_0}^{t_1} P(t)dt = -M^{-1}(\exp(M(t_1 - t_0)) - E)P(t_0),$$

where exp is the matrix exponent and E is the identity matrix.

Simulations

Simulations were done with the software *ms* Hudson (2002) and using GNU parallel Tange (2011) to run replicates and explore parameter values. First, we simulated two types of populations: *Population 1* remained with constant intermediate size, while *population 2* underwent a bottleneck followed by expansion, similar to the scenario simulated in Figure 2 of Li and Durbin (2011). The exact ms command line for simulating size changes of populations 1 and 2 is the following: "4 100 -t 15000 -r 1920 30000000 -l -I 2 2 2 -n 1 1.5 -n 2 3.0 -en 0.025 2 0.2 -en 0.175 2 1.5 -ej 0.625 2 1 -eN 3 3".

We added continuous migration using the *-em* flag in ms, and pulse migration was included using *-es* to split the receiver population into a third population, followed by *-ej* to merge the third population into the donor population at the exact same time.

To simulate a single panmictic population with the same historical ordinary effective population size as population 2 (used for Figure 2.4C,E), we used the following ms command line: "2 10000 -T -t 15000 -r 1920 30000000 -l -eN 0.0 3.233 -eN 0.01 3.69 -eN 0.02 3.996 -eN 0.025 0.311 -eN 0.03 0.359 -eN 0.04 0.436 -eN 0.05 0.531 -eN 0.06 0.644 -eN 0.07 0.777 -eN 0.08 0.926 -eN 0.09 1.088 -eN 0.1 1.369 -eN 0.125 1.723 -eN 0.15 1.928 -eN 0.175 2.274 -eN 0.2 2.049 -eN 0.25 1.859 -eN 0.3 1.703 -eN 0.4 1.597 -eN 0.5 1.545 -eN 0.625 1.5 -eN 3.0 3.0".

We also simulated past population size changes that closely approximate those inferred by PSMC for human populations. An approximation of the **Han-French** effective population size trajectories, was simulated using the command: "4 1000 -t 1500 -r 192 3000000 -l -I 2 2 2 -n 1 6.3 -n 2 2.4 -ej x 2 1 -eN 0.0225 0.3 -eN 0.15 2.7 -eN 0.5 1.3 -eN 2.5 2.9 -eN 5 2.7", using the following values of the split times (x): 0.0225 (900 generations, the end of the bottleneck), 0.043 (1714 generations, the split time inferred by MiSTI, within the bottleneck), 0.15 (6000 generations, right before the bottleneck) and 0.575 (23000 generations, before the expansion that precedes the main bottleneck).

An approximation of the **San-Dinka** effective population size trajectories, was simulated using the command: "4 1000 -t 1500 -r 192 3000000 -l -I 2 2 2 -n 1 1 -n 2 1 -ej x 2 1 -en 0.05 1 2 -en 0.11 1 3 -en 0.15 2 2.7 -eN 0.5 1.3 -eN 2.5 2.9 -eN 5 2.7", using the following values of the split times (x): 0.05 (2000 generations), 0.09 (3728 generations, the split time inferred by MiSTI), 0.21 (8554 generations, the split time inferred by TT), 0.25 (10000 generations) and 0.5 (20000 generations).

An approximation of the **Dinka-Sardinian** effective population size trajectories, was simulated using the command: "4 1000 -t 1500 -r 192 3000000 -l -I 2 2 2 -n 1 1 -n 2 1 -ej x 2 1 -en 0.0225 2 0.3 -eN 0.15 2.7 -eN 0.5 1.3 -eN 2.5 2.9 -eN 5 2.7", using the following values of the split times (x): 0.0225 (900 generations, the end of the bottleneck), 0.064 (2553 generations, the split time inferred by TT), 0.099 (2963 generations, the split time inferred by MiSTI), 0.25 (10000 generations) and 0.5 (20000 generations).

Times in the ms command lines are given in ms units, *i.e.* generations/$(4 \times N_0)$, where $N_0 = 10000$.

Data processing

We applied MiSTI to datasets from human and puma populations. MiSTI takes as input PSMC results for one individual from each population. If estimation of migration rates is desired, a joint site frequency spectrum of both genomes is also required. The joint site frequency spectrum can be generated with ANGSD Korneliussen et al. (2014), and a Python program is provided with MiSTI to convert ANGSD 2D site frequency spectrum format to MiSTI input format. For both humans and pumas, we applied filters to keep only sites with mapping quality above 30 and coverage between one third and twice the average genomewide coverage. In all analyses of human data, we applied the 1000 Genomes strict accessibility genome mask, and a filter for positions where the ancestral state was conserved among three species of great apes (Chimpanzee, Gorilla and Orangutan). The accessibility mask file can be downloaded from `http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/accessible_genome_masks/20140520.strict_mask.autosomes.bed`, and the ancestral state data were downloaded from `https://zenodo.org/record/4441887`. We ran PSMC with parameters -N25 -t15 -r5 -p "4+25*2+4+6" for both species. For humans, we used a mutation rate of 1.25×10^{-8} per base pair per generation, and generation time of 29 years. For pumas, we used mutation rate of 5×10^{-9} per base pair per generation, and generation time of 5 years Saremi et al. (2019).

For the analysis of human data, we downloaded a modern European genome (in bam format) from the CEPH/UTAH (CEU) population from the European Nucleotide Archive (ENA), accession number ERR194158. Neanderthal tracts specific for that individual were obtained from Steinrücken et al. (2018). The Neanderthal bam file was downloaded from `http://cdna.eva.mpg.de/neandertal/altai/AltaiNeandertal/bam/`. Bam files from one Han Chinese sample (HGDP00778), one French sample (HGDP00521), one San sample (HGDP01029) and one Dinka sample (DNK02) and one Sardinian sample (HGDP00665) were downloaded from `http://cdna.eva.mpg.de/denisova/BAM/human/`. When inferring migration from real data, we allowed it to start from the 4th PSMC interval and going until the split time, to avoid using the first PSMC intervals that have a lot of uncertainty.

We ran PSMC on the CEU genome before and after masking its tracts of Neanderthal ancestry, and we used MiSTI to correct the PSMC of the unmasked CEU genome, assuming 1.5% and 3% of pulse admixture from Neanderthals. The split time was set to 662 kya, considering that the average archaic-modern human split time inferred in Prüfer et al. (2014) is 570 thousand years, with a mutation rate of 5×10^{-10} per base pair per year, and adjusting for the mutation rate we use here (which translates to 4.3×10^{-10} per year). The pulse migration time was set to 60 kya, and the sample age was set to 50 kya, using MiSTI's *–sdate* parameter.

For the analysis of puma data, we obtained bam files and masks for runs of homozygosity (due to recent inbreeding) from the authors of Saremi et al. (2019). We focused on one sample from Florida (EVG21) that showed an inflated PSMC trajectory in Saremi et al. (2019), likely due to its known history of admixture with Central American pumas, and another sample from Florida that does not have Central American ancestry (CYP47). We

masked the runs of homozygosity from these genomes and we ran PSMC on them. We used MiSTI to correct the inferred effective population size trajectories for a plausible scenario of continuous migration and recent pulse admixture from Central America to Florida, based on the known history of this species. Saremi et al. (2019) inferred that the split time between the Florida pumas and Brazilian pumas was 300 thousand years ago. We have assumed a more recent split time of 200 thousand years between the Florida pumas and the Central American pumas that were ancestors of EVG21, which is the time when PSMC trajectories of CYP47 and EVG21 diverge. The resulting trajectory of the admixed individual, after correcting for its Florida ancestry component, is a putative effective population size trajectory for Central American pumas, that were not sampled.

Running MiSTI

To run MiSTI with parameter optimization, we recommend starting with inference of split times without migration. Once the best split time for this model (T^*) is found, the user can optimize the migration rates in each direction under split times equal to or larger than T^*. The user can run the migration rate optimization from different starting values, and we recommend using gnu-parallel Tange (2011) to provide the starting values to MiSTI.

Time discretization

MiSTI merges time points from two PSMC files, so that effective population size of both populations are constant on each time interval. This discretization is the principal time scale for MiSTI, and the default search of the split time is performed at the nodes of this discretization. If more precision is needed, one may use `-d N` key to split all the intervals in the search range into N equal parts.

Effective population size before split

Of course, when we choose a split time point, the estimated values for the effective population size before the split do not necessary coincide. So, we need to find a consensus effective population size from two estimates. The consensus effective population size before the split time is computed so that the expected number of coalescences for two haplotypes be the same as the sum of expected number of coalescence for the first and for the second genomes. More formally, let P_1 and P_2 be the probabilities of lineages sampled from populations 1 and 2 respectively not to coalesce before given time interval based on their corresponding distributions of effective population size. The probability not to coalesce within the time interval of length t and with estimated effective population sizes N_1 (from the first genome) and N_2 (from the second genome) is

$$P_{nc} = \frac{P_1 e^{-t/N_1} + P_2 e^{-t/N_2}}{P_1 + P_2}.$$

The consensus value for effective population size N at this time interval is

$$N = -\frac{t}{\log P_{nc}}. \tag{B.1}$$

B.2 Limitations of MiSTI: an analysis of likelihood surfaces

The demographic history of a recently admixed puma

In Florida, USA, there is a population of pumas (or mountain lions, *Puma concolor*) known as the Florida Panther. These pumas have a series of morphological signs of inbreeding depression. Interestingly, in the Everglades National Park (EVG), a population does not present these typical signs, likely because of an influx of genetic diversity brought by the introduction of individuals with Central American ancestry O'Brien et al. (1990). The PSMC trajectory from one of these individuals (EVG21) shows larger effective population sizes than other Florida panthers, as expected due to its admixed ancestry Saremi et al. (2019).

We have corrected the PSMC trajectory of this admixed individual with MiSTI, modeling it is a Central American puma that received admixture from a Florida Panther. We use a sample from the Big Cypress National Preserve (CYP47) as the unadmixed Florida Panther. This population is partially isolated from the EVG population and does not show inflated effective population size like EVG21.

Even though we do not have data from Central American pumas, by using MiSTI to correct the PSMC trajectory of EVG21 for the Florida admixture component (CYP47), we aimed to recover the effective population size trajectory of the unsampled Central American puma.

When we model a split time of 200ky between Central American and Florida panthers, and a single pulse admixture event very close to the present, we can fit a pulse of 0.30 from Florida to Central American. Under this model, we infer that the Central American population had more constant effective population size through time than other puma populations (Figure B.1).

The range of *Puma concolor* used to be connected from East to West of North America until a great population decline started in the 1800s, when these animals suffered extreme habitat loss and were hunted almost to extinction. Therefore, it is likely that there has been some degree of continuous migration between Florida Panthers and Central American pumas until historical times. Allowing for continuous migration, the model fits the data better (llh=-131360 instead of llh= -172712). This model includes continuous migration (migration rate from Central America to Florida of 1.0, and 0.65 in the other direction) and the same 0.30 recent pulse migration from Florida to Central America. Under this model, we also infer that the Central American puma population had relatively constant effective population size through time, and that both populations had smaller local effective population size than the ancestral effective population size inferred by PSMC (Figure B.2). We note that the original

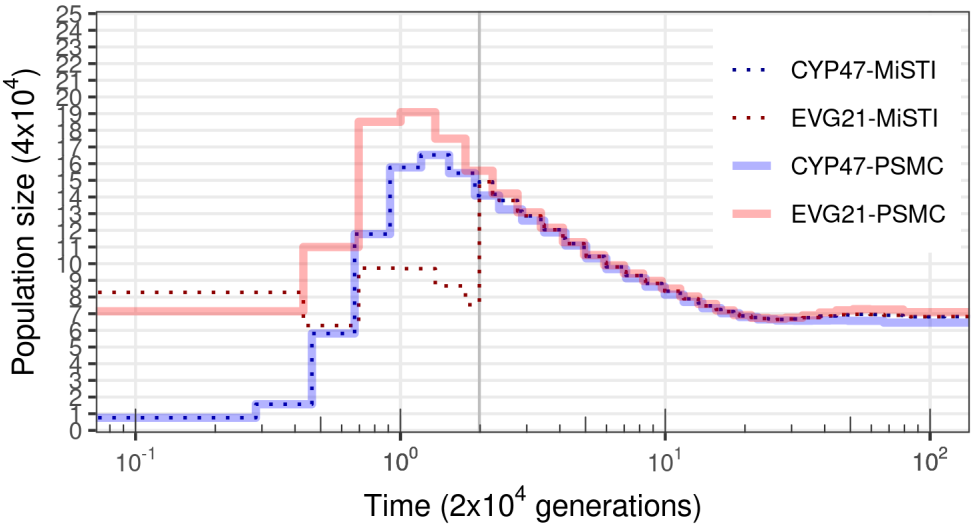

Figure B.1: MiSTI correction of the PSMC curve of the admixed puma sample EVG21 with a single pulse of 0.30 admixture from CYP47 at the most recent time interval. Split time (200 thousand years) is indicated by a vertical gray bar.

PSMC effective population sizes are very high (over 600 thousand individuals), and likely unrealistic for a population of large carnivores. This is likely a consequence of some degree of population structure and continuous migration across the range of pumas until recent times.

We note, however, that the composite likelihood surface for the split time and pulse of migration used as a model for the Pumas is not smooth (Figure B.3). The empty area of the composite likelihood surface indicates rates of continuous migration that are incompatible with the PSMC trajectories. For those values of migration rates, the corrected local effective population size would become negative, which is nonsensical. The fact that the highest composite likelihoods are at the border of the likelihood surface and next to these incompatible values of migration rates indicates that the data does not fit well with the MiSTI model of pulse and continuous migration we used. This uneven composite likelihood surface could also be due to some issue in the PSMC inference step. One possible source of problems is that there is large uncertainty in the PSMC inference at the most recent time intervals, which is when we model the pulse of migration in the Pumas.

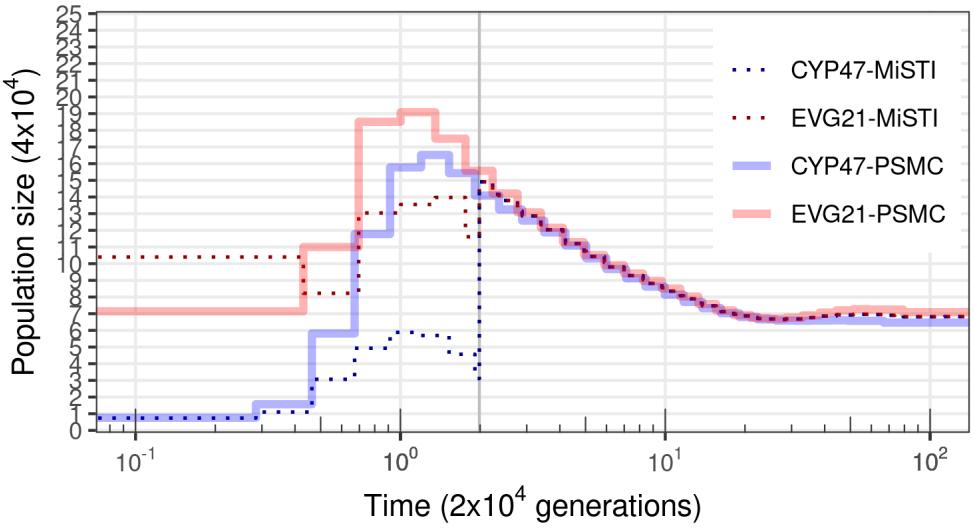

Figure B.2: MiSTI correction of the PSMC curve of the admixed puma sample EVG21 with continuous migration allowed since the split time, and a single pulse of 0.30 admixture from CYP47 at the most recent time interval. Split time (200 thousand years) is indicated by a vertical gray bar.

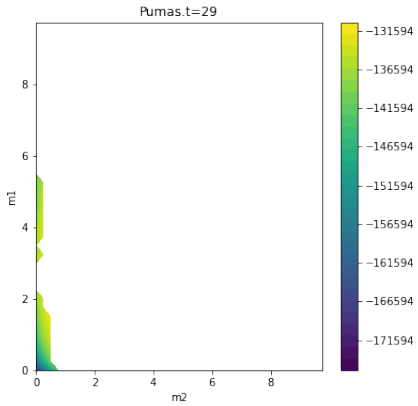

Figure B.3: Composite likelihood surface for the model of Florida panther split time 200 kya, pulse of migration at the most recent time interval, and a range of continuous migration rates.

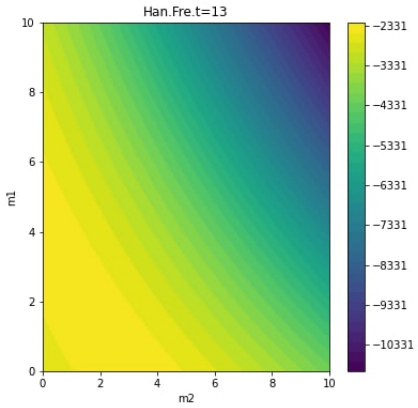

Figure B.4: Composite likelihood surface from MiSTI for the inferred split time between Han and French (1505 generations ago), for different values of migration rates.

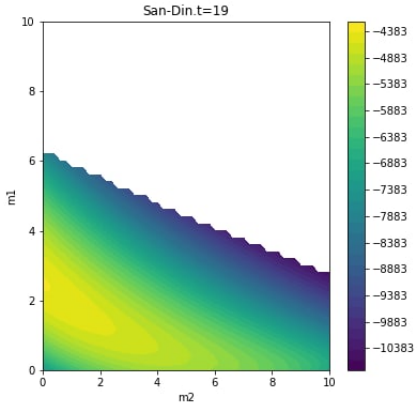

Figure B.5: Composite likelihood surface from MiSTI for the inferred split time between San and Dinka (3729 generations ago), for different values of migration rates.

Likelihood surfaces of inferred models of split and migration between pairs of human populations

Here we show the composite likelihood surfaces for a range of migration rates under the best inference of split time for pairs of human populations discussed in the main text. These composite likelihood surfaces differ from the puma case above in that their peak is not near the steep drop in composite likelihood.

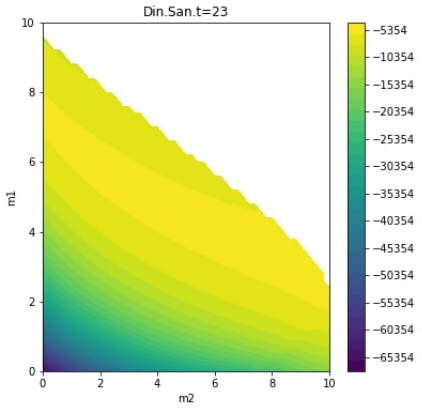

Figure B.6: Composite likelihood surface from MiSTI for the inferred split time between Dinka and Sardinian (3963 generations ago), for different values of migration rates.

B.3 Simulations of the San-Dinka split time with migration

Here we show simulations of the best inference of split time and migration rates for the San-Dinka pair. We did ten replicate simulations with the effective population size trajectories inferred from the data with PSMC, and with split time 3729 generations ago, and migration rate of 2.5 from Dinka to San (Table 2.2). We applied both the TT method and MiSTI to infer split times in each simulation. The TT method largely overestimated the split time in all cases, while MiSTI underestimated the split time when no migration is allowed in the model (left most column, Figure B.7). When migration in the direction simulated (from Dinka to San) is allowed, MiSTI estimates split times closer to the simulated values, and migration rates are also estimated in the correct direction (columns 2 and 4, Figure B.7). When migration is only allowed in the direction opposite to the simulated, it is largely overestimated, and the split time is underestimated (third column of Figure B.7).

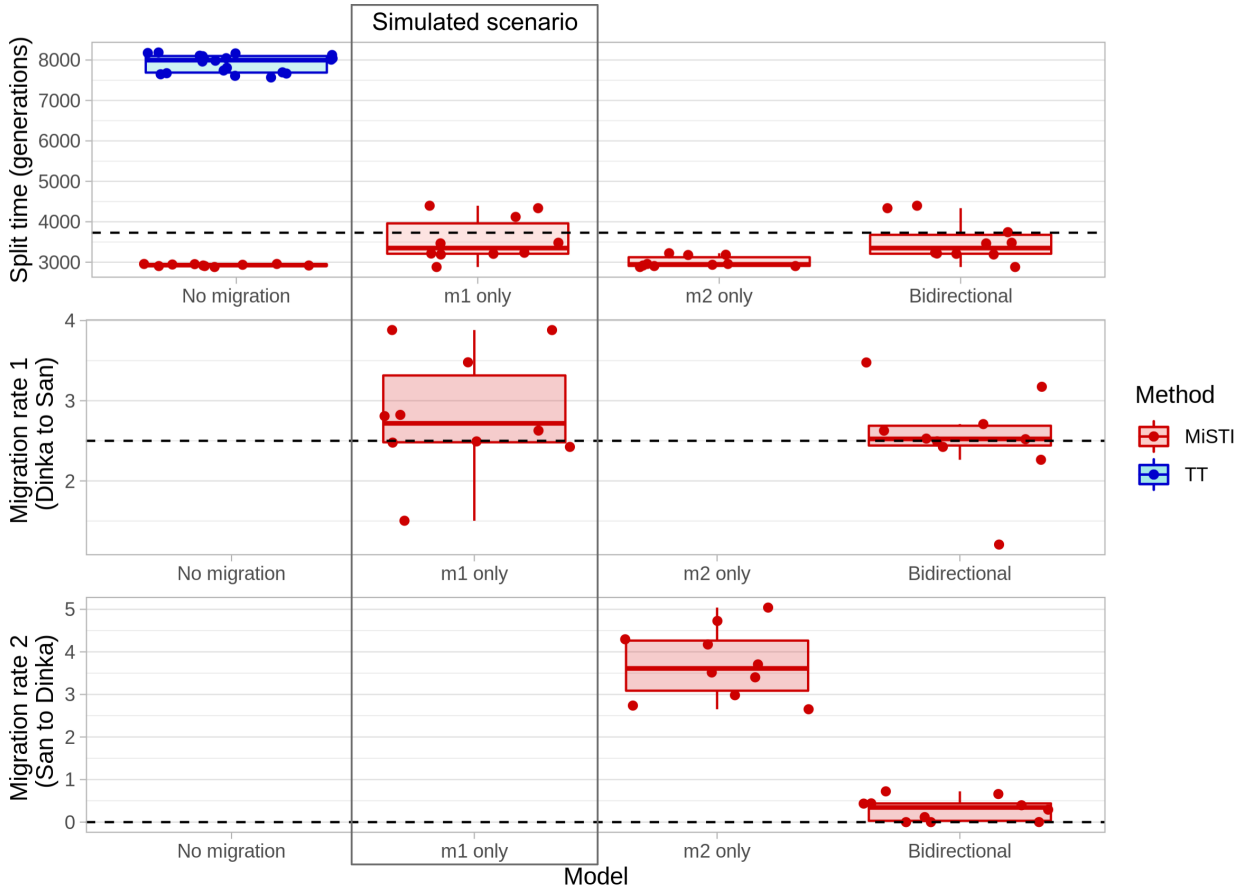

Figure B.7: Ten simulations of the split time and migration rates between San and Dinka inferred by MiSTI (split 3729 generations ago, m1=2.5 and m2=0, shown in the main text Table 2.2). In the top panel, we show split times inferred using the TT method and MiSTI. Middle and bottom panels shows values of inferred migration rates. In MiSTI, we inferred split times and migration rates for 4 models: no migration, m1 only, m2 only and bidirectional migration.

Appendix C

Appendix of Chapter 3

C.1 Supplementary figures

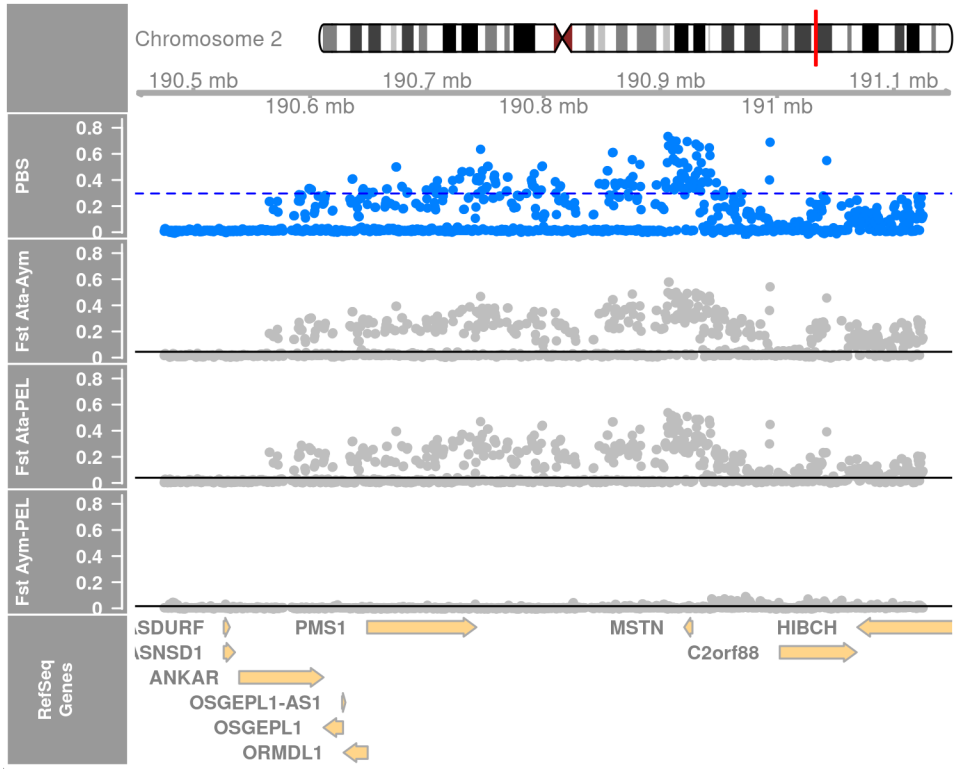

Figure C.1: PBS scan peak on chromosome 2 at 190Mb. Scan performed with windows of 1kb, slide of 500bp. Dashed blue line shows 0.1 percentile of PBS. Black lines show genomewide F_{ST} values for each population pair.

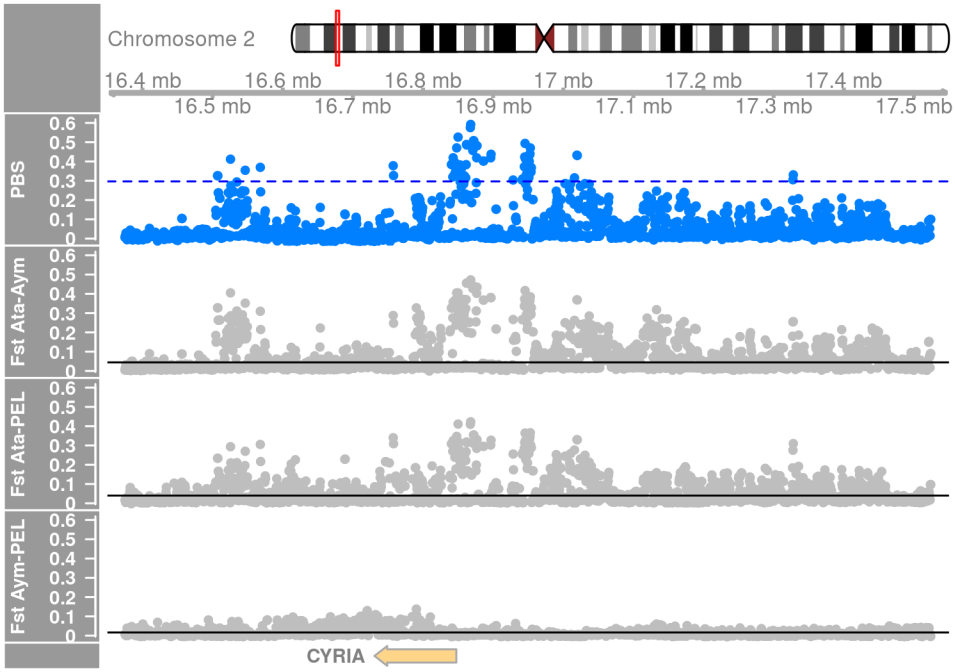

Figure C.2: PBS scan peak on chromosome 2 at 17Mb. Scan performed with windows of 1kb, slide of 500bp. Dashed blue line shows 0.1 percentile of PBS. Black lines show genomewide F_{ST} values for each population pair.

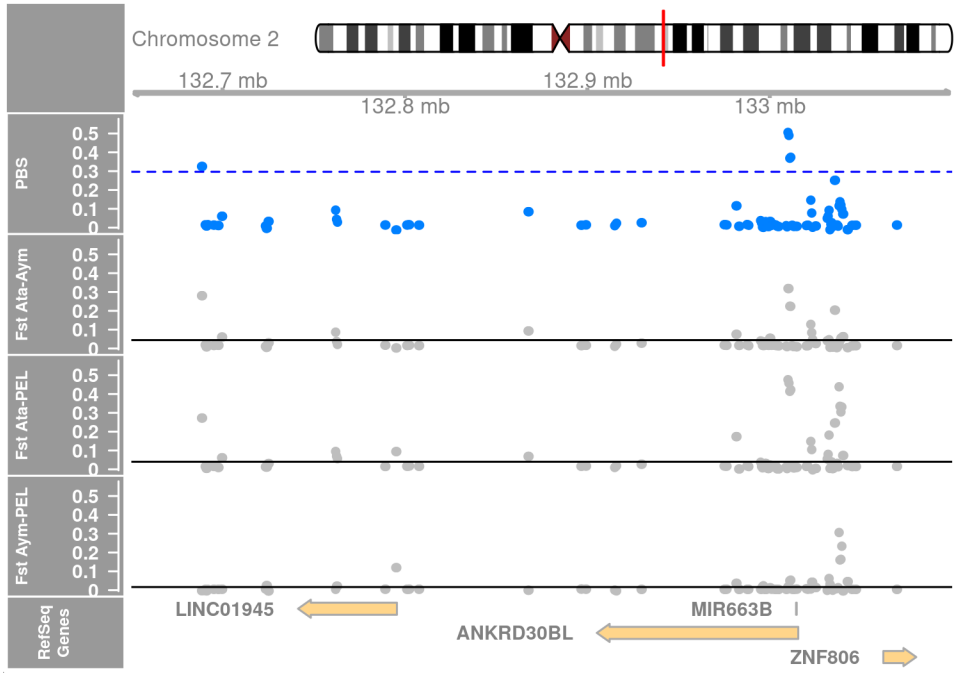

Figure C.3: PBS scan peak on chromosome 2 at 133Mb. Scan performed with windows of 1kb, slide of 500bp. Dashed blue line shows 0.1 percentile of PBS. Black lines show genomewide F_{ST} values for each population pair.

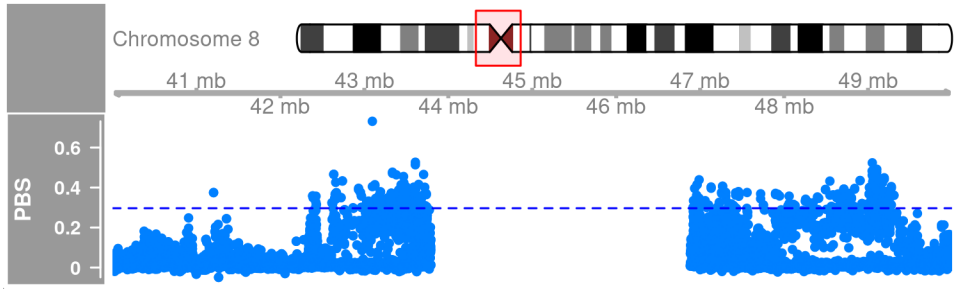

Figure C.4: PBS scan peak on chromosome 8. Peak of PBS is in a region of increased mutation next to the centromere. Dashed blue line shows 0.1 percentile of PBS.

C.2 Supplementary tables

Table C.1: Candidate genes in chromosome 10, at 105Mb. Gene summaries obtained from GeneCards (Safran et al., 2022).

Gene symbol	Gene name	Entrez Gene Summary
MFSD13A	Major Facilitator Superfamily Domain Containing 13A	Predicted to be integral component of membrane. [provided by Alliance of Genome Resources, Apr 2022]
ACTR1A	Actin Related Protein 1A	This gene encodes a 42.6 kD subunit of dynactin, a macromolecular complex consisting of 10-11 subunits ranging in size from 22 to 150 kD. Dynactin binds to both microtubules and cytoplasmic dynein. It is involved in a diverse array of cellular functions, including ER-to-Golgi transport, the centripetal movement of lysosomes and endosomes, spindle formation, chromosome movement, nuclear positioning, and axonogenesis. This subunit is present in 8-13 copies per dynactin molecule, and is the most abundant molecule in the dynactin complex. It is an actin-related protein, and is approximately 60% identical at the amino acid level to conventional actin. [provided by RefSeq, Jul 2008]
SUFU	SUFU Negative Regulator Of Hedgehog Signaling	The Hedgehog signaling pathway plays an important role in early human development. The pathway is a signaling cascade that plays a role in pattern formation and cellular proliferation during development. This gene encodes a negative regulator of the hedgehog signaling pathway. Defects in this gene are a cause of medulloblastoma. Alternative splicing results in multiple transcript variants.[provided by RefSeq, May 2010]

Table C.1 Continued from previous page.

Gene symbol	Gene name	Entrez Gene Summary
TRIM8	Tripartite Motif Containing 8	This gene encodes a member of the tripartite motif (TRIM) protein family. Based on similarities to other proteins, the encoded protein is suspected to be an E3 ubiquitin-protein ligase. Regulation of this gene may be altered in some cancers. Mutations resulting in a truncated protein product have been observed in early-onset epileptic encephalopathy (EOEE). [provided by RefSeq, Sep 2016]
ARL3	ADP Ribosylation Factor Like GTPase 3	ADP-ribosylation factor-like 3 is a member of the ADP-ribosylation factor family of GTP-binding proteins. ARL3 binds guanine nucleotides but lacks ADP-ribosylation factor activity. [provided by RefSeq, Jul 2008]
SFXN2	Sideroflexin 2	Predicted to enable serine transmembrane transporter activity. Involved in mitochondrial transmembrane transport. Located in mitochondrion. [provided by Alliance of Genome Resources, Apr 2022]
WBP1L	WW Domain Binding Protein 1 Like	Predicted to enable ubiquitin protein ligase binding activity. Predicted to act upstream of or within CXCL12-activated CXCR4 signaling pathway; hemopoiesis; and positive regulation of protein ubiquitination. Predicted to be integral component of membrane. [provided by Alliance of Genome Resources, Apr 2022]

Table C.1 Continued from previous page.

Gene symbol	Gene name	Entrez Gene Summary
CYP17A1	Cytochrome P450 Family 17 Subfamily A Member 1	This gene encodes a member of the cytochrome P450 superfamily of enzymes. The cytochrome P450 proteins are monooxygenases which catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids. This protein localizes to the endoplasmic reticulum. It has both 17alpha-hydroxylase and 17,20-lyase activities and is a key enzyme in the steroidogenic pathway that produces progestins, mineralocorticoids, glucocorticoids, androgens, and estrogens. Mutations in this gene are associated with isolated steroid-17 alpha-hydroxylase deficiency, 17-alpha-hydroxylase/17,20-lyase deficiency, pseudohermaphroditism, and adrenal hyperplasia. [provided by RefSeq, Jul 2008]
BORCS7	BLOC-1 Related Complex Subunit 7	Part of BORC complex. [provided by Alliance of Genome Resources, Apr 2022]
AS3MT	Arsenite Methyltransferase	AS3MT catalyzes the transfer of a methyl group from S-adenosyl-L-methionine (AdoMet) to trivalent arsenical and may play a role in arsenic metabolism (Lin et al., 2002 [PubMed 11790780]).[supplied by OMIM, Mar 2008]
CNNM2	Cyclin And CBS Domain Divalent Metal Cation Transport Mediator 2	This gene encodes a member of the ancient conserved domain containing protein family. Members of this protein family contain a cyclin box motif and have structural similarity to the cyclins. The encoded protein may play an important role in magnesium homeostasis by mediating the epithelial transport and renal reabsorption of Mg2+. Mutations in this gene are associated with renal hypomagnesemia. Alternatively spliced transcript variants encoding multiple isoforms have been observed for this gene. [provided by RefSeq, Dec 2011]
NT5C2	5'-Nucleotidase, Cytosolic II	This gene encodes a hydrolase that serves as an important role in cellular purine metabolism by acting primarily on inosine 5'-monophosphate and other purine nucleotides. [provided by RefSeq, Oct 2011]

Table C.1 Continued from previous page.

Gene symbol	Gene name	Entrez Gene Summary
RPEL1	Ribulose-5-Phosphate-3-Epimerase Like 1	Predicted to enable metal ion binding activity and ribulose-phosphate 3-epimerase activity. Predicted to be involved in cellular carbohydrate metabolic process; pentose catabolic process; and pentose-phosphate shunt, non-oxidative branch. Predicted to be active in cytosol. [provided by Alliance of Genome Resources, Apr 2022]
INA	Internexin Neuronal Intermediate Filament Protein Alpha	Neurofilaments are type IV intermediate filament heteropolymers composed of light, medium, and heavy chains. Neurofilaments comprise the axoskeleton and they functionally maintain the neuronal caliber. They may also play a role in intracellular transport to axons and dendrites. This gene is a member of the intermediate filament family and is involved in the morphogenesis of neurons. [provided by RefSeq, Jun 2009]
PCGF6	Polycomb Group Ring Finger 6	The protein encoded by this gene contains a RING finger motif, which is most closely related to those of polycomb group (PcG) proteins RNF110/MEL-18 and BMI1. PcG proteins are known to form protein complexes and function as transcription repressors. This protein has been shown to interact with some PcG proteins and act as a transcription repressor. The activity of this protein is found to be regulated by cell cycle dependent phosphorylation. Alternatively spliced transcript variants encoding different isoforms have been identified. [provided by RefSeq, Jul 2008]

Table C.1 Continued from previous page.

Gene symbol	Gene name	Entrez Gene Summary
TAF5	TATA-Box Binding Protein Associated Factor 5	Initiation of transcription by RNA polymerase II requires the activities of more than 70 polypeptides. The protein that coordinates these activities is transcription factor IID (TFIID). This gene encodes an integral subunit of TFIID associated with all transcriptionally competent forms of that complex. This subunit interacts strongly with two TFIID subunits that show similarity to histones H3 and H4, and it may participate in forming a nucleosome-like core in the TFIID complex. Alternative splicing results in multiple transcript variants. [provided by RefSeq, Dec 2015]
ATP5MK, previoulsy ATP5MD	ATP Synthase Membrane Subunit K	Located in mitochondrion. Part of mitochondrial proton-transporting ATP synthase complex. Implicated in mitochondrial complex V (ATP synthase) deficiency nuclear type 6. [provided by Alliance of Genome Resources, Apr 2022]
MIR1307	MicroRNA 1307	microRNAs (miRNAs) are short (20-24 nt) non-coding RNAs that are involved in post-transcriptional regulation of gene expression in multicellular organisms by affecting both the stability and translation of mRNAs. miRNAs are transcribed by RNA polymerase II as part of capped and polyadenylated primary transcripts (pri-miRNAs) that can be either protein-coding or non-coding. [provided by RefSeq, Sep 2009]
PDCD11	Programmed Cell Death 11	PDCD11 is a NF-kappa-B (NFKB1; 164011)-binding protein that colocalizes with U3 RNA (MIM 180710) in the nucleolus and is required for rRNA maturation and generation of 18S rRNA.[supplied by OMIM, Oct 2008]
CALHM2	Calcium Homeostasis Modulator Family Member 2	Predicted to enable cation channel activity. Involved in positive regulation of apoptotic process. Predicted to be integral component of plasma membrane. [provided by Alliance of Genome Resources, Apr 2022]

Table C.2: Candidate genes in chromosome 1, at 155Mb. Gene summaries obtained from GeneCards (Safran et al., 2022).

Gene symbol	Gene name	Entrez Gene Summary
GBAP1	Glucosylceramidase Beta Pseudogene 1	GBAP1 (Glucosylceramidase Beta Pseudogene 1) is a Pseudogene.
GBA	Glucosylceramidase Beta	This gene encodes a lysosomal membrane protein that cleaves the beta-glucosidic linkage of glycosylceramide, an intermediate in glycolipid metabolism. Mutations in this gene cause Gaucher disease, a lysosomal storage disease characterized by an accumulation of glucocerebrosides. A related pseudogene is approximately 12 kb downstream of this gene on chromosome 1. [provided by RefSeq, Jan 2010]
ENTREP3, previously FAM189B	Endosomal Transmembrane Epsin Interactor 3	This gene is located near the gene for the lysosomal enzyme glucosylceramidase; a deficiency in this enzyme is associated with Gaucher disease. The encoded protein has been identified as a potential binding partner of a WW domain-containing protein which is involved in apoptosis and tumor suppression. [provided by RefSeq, Dec 2010]
SCAMP3	Secretory Carrier Membrane Protein 3	This gene encodes an integral membrane protein that belongs to the secretory carrier membrane protein family. The encoded protein functions as a carrier to the cell surface in post-golgi recycling pathways. This protein is also involved in protein trafficking in endosomal pathways. Two transcript variants encoding different isoforms have been found for this gene.[provided by RefSeq, May 2011]
CLK2	CDC Like Kinase 2	This gene encodes a dual specificity protein kinase that phosphorylates serine/threonine and tyrosine-containing substrates. Activity of this protein regulates serine- and arginine-rich (SR) proteins of the spliceosomal complex, thereby influencing alternative transcript splicing. [provided by RefSeq, Jun 2014]

Table C.2 Continued from previous page.

Gene symbol	Gene name	Entrez Gene Summary
HCN3	Hyperpolarization Activated Cyclic Nucleotide Gated Potassium Channel 3	This gene encodes a multi-pass membrane protein that functions as a voltage gated cation channel. The encoded protein is a member of a family of closely related cyclic adenosine monophosphate-binding channel proteins. Alternative splicing results in multiple transcript variants. [provided by RefSeq, Aug 2012]
PKLR	Pyruvate Kinase L/R	The protein encoded by this gene is a pyruvate kinase that catalyzes the transphosphorylation of phohspho-enolpyruvate into pyruvate and ATP, which is the rate-limiting step of glycolysis. Defects in this enzyme, due to gene mutations or genetic variations, are the common cause of chronic hereditary nonspherocytic hemolytic anemia (CNSHA or HNSHA). [provided by RefSeq, Jul 2008]
FDPS	Farnesyl Diphosphate Synthase	This gene encodes an enzyme that catalyzes the production of geranyl pyrophosphate and farnesyl pyrophosphate from isopentenyl pyrophosphate and dimethylallyl pyrophosphate. The resulting product, farnesyl pyrophosphate, is a key intermediate in cholesterol and sterol biosynthesis, a substrate for protein farnesylation and geranylgeranylation, and a ligand or agonist for certain hormone receptors and growth receptors. Drugs that inhibit this enzyme prevent the post-translational modifications of small GTPases and have been used to treat diseases related to bone resorption.[provided by RefSeq, Oct 2008]
RUSC1	RUN And SH3 Domain Containing 1	Predicted to enable actin binding activity. Involved in protein polyubiquitination. Located in cytosol. [provided by Alliance of Genome Resources, Apr 2022]
ASH1L	ASH1 Like Histone Lysine Methyltransferase	This gene encodes a member of the trithorax group of transcriptional activators. The protein contains four AT hooks, a SET domain, a PHD-finger motif, and a bromodomain. It is localized to many small speckles in the nucleus, and also to cell-cell tight junctions. [provided by RefSeq, Jul 2008]

Table C.2 Continued from previous page.

Gene symbol	Gene name	Entrez Gene Summary
MIR555	MicroRNA 555	microRNAs (miRNAs) are short (20-24 nt) non-coding RNAs that are involved in post-transcriptional regulation of gene expression in multicellular organisms by affecting both the stability and translation of mRNAs. miRNAs are transcribed by RNA polymerase II as part of capped and polyadenylated primary transcripts (pri-miRNAs) that can be either protein-coding or non-coding. [provided by RefSeq, Sep 2009]
POU5F1P4	POU Class 5 Homeobox 1 Pseudogene 4	POU5F1P4 (POU Class 5 Homeobox 1 Pseudogene 4) is a Pseudogene.
ASH1L-AS1	ASH1L Antisense RNA 1	ASH1L-AS1 (ASH1L Antisense RNA 1) is an RNA Gene, and is affiliated with the lncRNA class.
MSTO1	Misato Mitochondrial Distribution And Morphology Regulator 1	Involved in mitochondrion distribution. Located in cytosol and mitochondrial outer membrane. [provided by Alliance of Genome Resources, Apr 2022]
YY1AP1	YY1 Associated Protein 1	The encoded gene product presumably interacts with YY1 protein; however, its exact function is not known. [provided by RefSeq, Jul 2008]
DAP3	Death Associated Protein 3	Mammalian mitochondrial ribosomal proteins are encoded by nuclear genes and help in protein synthesis within the mitochondrion. Mitochondrial ribosomes (mitoribosomes) consist of a small 28S subunit and a large 39S subunit. This gene encodes a 28S subunit protein that also participates in apoptotic pathways which are initiated by tumor necrosis factor-alpha, Fas ligand, and gamma interferon. This protein potentially binds ATP/GTP and might be a functional partner of the mitoribosomal protein S27. [provided by RefSeq, Dec 2010]
MSTO2P	Misato Family Member 2, Pseudogene	MSTO2P (Misato Family Member 2, Pseudogene) is a Pseudogene. Diseases associated with MSTO2P include Gastric Cancer.

Table C.2 Continued from previous page.

Gene symbol	Gene name	Entrez Gene Summary
GON4L	Gon-4 Like	Predicted to enable transcription coregulator activity. Predicted to be involved in regulation of transcription, DNA-templated. Predicted to act upstream of or within B cell differentiation. Located in nuclear body. [provided by Alliance of Genome Resources, Apr 2022]
SYT11	Synaptotagmin 11	This gene is a member of the synaptotagmin gene family and encodes a protein similar to other family members that are known calcium sensors and mediate calcium-dependent regulation of membrane trafficking in synaptic transmission. The encoded protein is also a substrate for ubiquitin-E3-ligase parkin. [provided by RefSeq, Apr 2010]
RIT1	Ras Like Without CAAX 1	This gene encodes a member of a subfamily of Ras-related GTPases. The encoded protein is involved in regulating p38 MAPK-dependent signaling cascades related to cellular stress. This protein also cooperates with nerve growth factor to promote neuronal development and regeneration.[provided by RefSeq, Feb 2012]
KHDC4	KH Domain Containing 4, Pre-MRNA Splicing Factor	Enables RNA binding activity. Involved in mRNA splice site selection. Located in cytoplasm and nucleoplasm. Colocalizes with spliceosomal complex. [provided by Alliance of Genome Resources, Apr 2022]
SCARNA4	Small Cajal Body-Specific RNA 4	SCARNA4 (Small Cajal Body-Specific RNA 4) is an RNA Gene, and is affiliated with the scaRNA class.
RXFP4	Relaxin Family Peptide/INSL5 Receptor 4	GPR100 is a member of the rhodopsin family of G protein-coupled receptors (GPRs) (Fredriksson et al., 2003 [PubMed 14623098]).[supplied by OMIM, Mar 2008]
ARHGEF2	Rho/Rac Guanine Nucleotide Exchange Factor 2	Rho GTPases play a fundamental role in numerous cellular processes that are initiated by extracellular stimuli that work through G protein coupled receptors. The encoded protein may form complex with G proteins and stimulate rho-dependent signals. [provided by RefSeq, Jun 2009]

Table C.2 Continued from previous page.

Gene symbol	Gene name	Entrez Gene Summary
SSR2	Signal Sequence Receptor Subunit 2	The signal sequence receptor (SSR) is a glycosylated endoplasmic reticulum (ER) membrane receptor associated with protein translocation across the ER membrane. The SSR consists of 2 subunits, a 34-kD glycoprotein (alpha-SSR or SSR1) and a 22-kD glycoprotein (beta-SSR or SSR2). The human beta-signal sequence receptor gene (SSR2) maps to chromosome bands 1q21-q23. [provided by RefSeq, Jul 2008]
SCARNA26A	Small Cajal Body-Specific RNA 26A	SCARNA26A (Small Cajal Body-Specific RNA 26A) is an RNA Gene, and is affiliated with the scaRNA class.
SCARNA26B	Small Cajal Body-Specific RNA 26B	SCARNA26B (Small Cajal Body-Specific RNA 26B) is an RNA Gene, and is affiliated with the scaRNA class.
SNORA80E	Small Nucleolar RNA, H/ACA Box 80E	SNORA80E (Small Nucleolar RNA, H/ACA Box 80E) is an RNA Gene, and is affiliated with the snoRNA class.
MIR6738	MicroRNA 6738	microRNAs (miRNAs) are short (20-24 nt) non-coding RNAs that are involved in post-transcriptional regulation of gene expression in multicellular organisms by affecting both the stability and translation of mRNAs. miRNAs are transcribed by RNA polymerase II as part of capped and polyadenylated primary transcripts (pri-miRNAs) that can be either protein-coding or non-coding. [provided by RefSeq, Sep 2009]