Learning with Admixture: Modeling, Optimization, and Applications in Population Genetics

Jade Yu Cheng

PhD Dissertation

Bioinformatics Research Centre
Department of Computer Science
Aarhus University
Denmark

Learning with Admixture: Modeling, Optimization, and Applications in Population Genetics

A Dissertation
Presented to the Faculty of Science and Technology
of Aarhus University
in Partial Fulfillment of the Requirements
for the PhD Degree

by
Jade Yu Cheng
July 31, 2016

Abstract

Population genetics is a branch of applied mathematics. It is a translation of scientific observations into mathematical models and their manipulations in order to produce quantitative predictions about evolution. Combining knowledge from genetics, statistics, and computer science, population geneticists strive to establish working solutions to extract information from massive volumes of biological data. The steep increase in the quantity and quality of genomic data during the past decades provides a unique opportunity but also calls for new and improved algorithms and software to cope with the big data era.

In this PhD dissertation, I present my work on methods and tools developed for two projects, Admixture CoalHMM and Ohana, both of which have been designed to study historical admixture and its influence on population evolution. In Admixture CoalHMM, I make use of full genomic sequences from a few individuals to perform demographic inference. In Ohana, I use site-independent genomic data from many individuals to analyze individual admixture, to infer population trees, and to identify selection signals.

The development of CoalHMM at the Bioinformatics Research Centre at Aarhus University dates back to 2007 [11]. CoalHMM is a hidden Markov model constructed on the foundation of coalescence theory with the key approximation that the distribution of local genealogies is Markovian along the sequence alignment. Through parametrized modeling, CoalHMM attempts to recover a full demography including population splits, effective population sizes, gene flow, etc. Since joining the CoalHMM development team in 2014, I have mainly contributed in two directions: 1) improving optimizations through heuristic-based evolutionary algorithms and 2) modeling of historical admixture events.

Ohana, meaning "family" in Hawaiian, is a novel project I started at the Center for Theoretical Evolutionary Genetics at the University of California Berkeley. Ohana provides a set of methods and tools for structure analysis, population tree inference, and selection study that fully takes advantage of structured genomic data. Ohana's admixture module is based on classical structure modeling [29] but uses new optimization subroutines through quadratic programming, which outperform the current state-of-the-art software in both speed and accuracy. Ohana presents a new method for phyloge-

netic tree inference using Gaussian approximation. With the estimated global ancestry and population relationships, Ohana provides a flexible selection signal detection process that considers any prior knowledge on the covariance structure, e.g population bottleneck or local adaptation.

Statistical modeling and numerical optimization form the foundation for both CoalHMM and Ohana. Optimization modeling has been the main theme throughout my PhD, and it will continue to shape my work for the years to come. The algorithms and software I developed to study historical admixture and population evolution fall into a larger family of machine learning, and their underlying techniques have a wide range of applications that go beyond just bioinformatics and population genetics.

Resumé

Populationsgenetik er en gren af anvendt matematik. Målet er at oversættelse videnskabelige observationer til matematiske modeller for at producere kvantitative forudsigelser om evolution. Ved at kombinere viden fra genetik, statistik og datalogi, forsøger genetikere at konstruere arbejdsmodeller der kan trække viden ud af massive mængder af biologiske data. Den kraftige stigning i mængden og kvaliteten af genomiske data i løbet af de seneste årtier er en enestående mulighed, men kræver også nye og forbedrede algoritmer og software til at håndtere denne æra af store datamænder.

I denne ph.d.-afhandling præsenterer jeg metoder og værktøjer udviklet i to projekter, Admixture CoalHMM og Ohana, som begge er designet til at studere historiske blanding af befolkninger og deres indflydelse på befolkningens evolution. I Admixture CoalHMM udnytter jeg komplette genomiske sekvenser fra et par af individer til at inferere demografiske parametre. I Ohana udnytter jeg uafhængig genomiske varianter fra mange individer til at analysere hvilken blanding af urbefolkninger hvert individ er fra, til at udlede urbefolkningernes historie, og til at finde signaler for selektion.

Udviklingen af CoalHMM på Center for Bioinformatik ved Aarhus Universitet begyndte i 2007. CoalHMM er en skjult Markov model baseret på coalescent teori, forsimplet ved at antage at tiden til coalescent langs et genom er Markov. Gennem parametriserede modeller bruges CoalHMMer til at forstå demografiske hændelser så som befolkningsopsplitning, effektiv befolkningstørelse, gen-udveksling, osv. Siden jeg startede i CoalHMM gruppen in 2014 har jeg bidraget i to retninger: 1) forbedring af parameter-optimeringer gennem heuristisk-baserede evolutionær algoritmer og 2) modellering af historiske genudvekslinger.

Ohana, som betyder "familie" på hawaiiansk, er et nyt projekt jeg startede på Center for Teoretisk Evolutionary Genetik ved University of California Berkeley. Ohana indeholder en række metoder og værktøjer til strukturanalyse, inferens af befolkningstræer, og identifikation af selektion, der fuldt ud udnytter strukturerede genomiske data. Ohanas admixturmodel er baseret på den klassiske struktur modellering men bruger nye optimering subrutiner gennem kvadratisk programmering, der udkonkurrerer state-of-the-art software i både hastighed og nøjagtighed. Ohana præsenterer en ny metode til fylogenetisk træ inferens ved hjælp Gaussisk approksimation. Med de estimerede

globale herkomst og befolkning relationer kan Ohana bruges som en fleksibel metode til at finde signaler om selektion der ikke kræver nogen forudgående viden om kovariansen struktur, f.eks flaskehalseffekten eller lokal tilpasning.

Statistisk modellering og numerisk optimering danner grundlaget for inferens i både CoalHMM og Ohana. Optimering af modeller har været det vigtigste tema i hele min ph.d., og det vil fortsætte med at forme mit arbejde for de kommende år. De algoritmer og software jeg udviklet for at studere historiske blanding af befolkinger og befolkningers evolution er en del af en familie af machine learning metoder, og deres underliggende teknikker har en bred vifte af applikationer, der går ud over blot bioinformatik og populationsgenetik.

Contents

Chapter 1

Introduction

Admixture, gene flow, and hybridization events play an important role in shaping evolutionary history. Population geneticists have long recognized the importance of quantifying ancestry admixing among populations, and its applications span many sub-fields such as conservation genetics, association study, and migration pattern research [29]. With the decrease in costs to sequence full genomes and the increase in accuracy of genotype technologies, it has become possible to infer admixture from large genomic datasets. The availability of large volumes of high quality data also calls for fast and accurate tools to perform admixture-related analysis.

Similar to other branches of population genetics, the study of admixture is a translation of scientific observations into mathematical models and its manipulations in order to produce quantitative predictions. We combine knowledge from genetics, applied mathematics, and computer science to establish working solutions of extracting meaningful information from a overwhelmingly large volume of biological data.

A mathematical model is an abstract model that describes portions of reality expressed in the language of mathematics. We call it a probabilistic model if we use the mathematics of probability theory to express all forms of uncertainty and noise associated with the model and to describe the data that one can observe from the system. We use this kind of modeling for parameter learning, information filtration, model prediction, etc. In our study, we formulate mathematical abstractions to describe admixture in the context of evolution.

Numerical optimization is an important method for finding values of variables that optimize an objective. The objective depends on certain characteristics of a system, and the process of identifying the objective, variables, and constraints for a given problem is called modeling. There is no universal optimization algorithm but rather a collection of algorithms, each of which being designed for a particular type of optimization problem [27]. In our study, we numerically optimize the likelihoods of parametrized mathematical models.

Learning the model parameters provides us a quantitative view of admixture in the evolutionary history.

In this dissertation, we develop novel methods and present new tools to study historical admixture and gene flow in the field of population genetics. In Chapter 2, we present Admixture CoalHMM, a project that uses full genomic sequences from a few modern-day individuals to perform demographic inference. In Chapter 3, we present Ohana, a project that uses site-independent genomic data from many individuals to perform structural analysis, to infer population trees, and to conduct selection study.

Admixture CoalHMM

CoalHMM is a hidden Markov model constructed on the foundation of coalescence theory with the key approximation that the distribution of local genealogies is Markovian along the sequence alignment.

In Chapter 2, we extend the state-of-the-art CoalHMM framework with heuristic-based optimizations, complex demographic model construction, and most importantly, the capability of modeling admixture events. We have validated this system through extensive simulation studies and have used it to investigate the evolutionary histories of a range of species: baboons, bears, equids, humans, and lynxes. A corresponding implementation of Admixture CoalHMM is available on GitHub:

https://github.com/jade-cheng/Jocx

This work has resulted in two papers (one in preparation) that developed the method. The first paper focused on the heuristic optimization algorithms, and the second paper focused on the modeling of the admixture events. This work has also resulted in three papers (two in preparation) that applied the method to biological data.

- Jade Yu Cheng and Thomas Mailund. 2016 "A coalescent hidden Markov model for inferring admixture relationships" (in preparation, Appendix A).
- Jade Yu Cheng and Thomas Mailund. 2015 "Ancestral population genomics using coalescence hidden Markov models and heuristic optimisation algorithms" Computational Biology and Chemistry, 57, pp.80-92. (Appendix B).
- Tianying Lan, Jade Yu Cheng, Aakrosh Ratan, Webb Miller, Stephan C. Schuster, Karyn Rode, Todd Atwood, Sean Farley, Dick Richard T. Shideler, Sandra L. Talbot, Thomas Mailund, Charlotte Lindqvist. 2016 "Genome-wide evidence for a hybrid origin of modern polar bears" bioRxiv, p.047498.

Ohana

Over the past two decades, population geneticists have begun using unsupervised learning methods to analyze population structure. While we have made great strides, the effort to improve the accuracy and speed never diminishes. In addition, a need has emerged to identify selection signals while fully taking advantage of the structured data.

In Chapter 3, we develop a suite of statistical methods to infer individual ancestries, to estimate population covariances, and to detect covariance outliers as selection signals. We have validated different stages of this system through simulation studies, and we have used the methods in several collaborative studies with great success. A corresponding implementation, along with installation instructions, documentation, and example workflows, is available on GitHub:

https://github.com/jade-cheng/ohana

This work has resulted in four papers, two that developed the methods (one in the planning stage) and two that applied the methods to biological data.

- Jade Yu Cheng, Thomas Mailund, and Rasmus Nielsen. 2016 "Ohana, a tool set for population genetic analyses of admixture components" Submitted to Bioinformatics (Appendix C)
- Georgios Athanasiadis, Jade Yu Cheng, Bjarni J. Vilhjálmsson, Frank Grønlund Jørgensen, Thomas D. Als, Stephanie Le Hellard, Thomas Espeseth, Patrick F. Sullivan, Christina M. Hultman, Peter Kjærgaard, Mikkel Heide Schierup, Thomas Mailund. 2016 "Nationwide genomic study in Denmark reveals remarkable population homogeneity" to appear in Genetics (Appendix E)
- Anna-Sapfo Malaspinas, Michael C Westaway, Craig Muller, Vitor C Sousa, Oscar Lao, Isabel Alves, Anders Bergström, Georgios Athanasiadis, Jade Y Cheng, . . . , Eske Willerslev. 2016 "The genomic history of Australia" to appear in Nature (Appendix D)

Overview

Throughout this dissertation, we study admixture through mathematical modeling and numerical optimization. In Chapter 2, we build a complex statistical model using two major concepts, continuous time Markov chains and hidden Markov models. Because of the non-analytical procedure of the likelihood computation, we investigate a range of black-box style optimizers with a special emphasis on heuristic-based evolutionary algorithms. In Chapter 3, we

obtain the analytical forms of Ohana's statistical models. Because of the extensive number of model parameters, we develop two sequential quadratic programming methods tailored for the problem. We also describe the Nelder-Mead simplex method for optimization tasks dealing with small-dimensioned parameter spaces.

In each of the following chapters, we will begin with an introduction to the field and its state-of-the-art research. Next, we will detail the mathematical models of the framework and discuss the optimization techniques created for these models. We will demonstrate simulation studies and present real genomic data analysis. Each chapter will end with remarks regarding the models, optimizations, and inference results.

Chapter 2

Admixture CoalHMM

2.1 Introduction

Coalescence theory describes a class of retrospective modeling in population genetics [9]. In coalescence theory, we represent the ancestries of current day genes as gene genealogies, similar to species genealogies. We assign probabilities to all possible genealogies that could have created the gene variations we see in present samples. The classical coalescence model is described as a continuous time Markov process running backward in time, during which events such as coalescence and recombination happen.

Considering coalescence events exclusively, gene genealogies are tree structures, but after involving recombination events, a lineage can also split into two backward in time. The outcome of both processes is a directed acyclic graph rather than a tree. We call this an ancestral recombination graph (ARG). We can consider an ARG as a collection of trees merged together, where each position on the genome takes a single tree as its local genealogy.

Armed with coalescence theory, we can model structured populations by controlling the possibility of coalescence because only samples residing in the same population can coalesce. We assign lineages to different populations. When the populations are isolated, no coalescence can occur. When migrations starts coalescence occurs, and it happens at a rate related to the rate of migration. We can also model population splits and admixture events by re-assigning lineages to reflect these changes.

The coalescence hidden Markov model is a hidden Markov model (HMM) constructed on the foundation of coalescence theory with the key approximation that the distribution of local genealogies is Markovian along the sequence alignment. Being Markovian means the genealogy tree at one alignment location depends on only the genealogy tree at the previous alignment location and none of the earlier locations. This allows us to investigate only two loci at a time, and the joint probability of two neighboring genealogies would form the transition probabilities for the HMM rather than exploring the entire ARG.

In the family of coalescence HMM based methods, PSMC developed by Li and Durbin [20] leads the way of popularity. PSMC stands for pairwise sequentially Markovian coalescent, and it focuses on modeling varying coalescent rates over time. As the name indicates, PSMC models a pair of samples, possibly two haplotypes of one individual. MSMC [31] is the successor of PSMC, and it stands for multiple sequential Markovian coalescent. Compared to PSMC, MSMC increases the inference power for recent history. MSMC also incorporates the capability to estimate population splits. CoalHMM described in [21] uses continuous Markov chain to explore all possible configurations of samples with respect to population structures. The modeling in CoalHMM is similar to PSMC and MSMC but more flexible and complex. The estimated quantities, hence the goals, are also different. PSMC focuses on recovering effective population sizes over time. CoalHMM attempts to recover the demography including population splits and all forms of gene flow.

Even with the approximation of being Markovian along the sequence, coalescence HMM modeling is still computational expensive and nearly impossible to scale when the number of samples increases. We can further restrict multifurcating trees because, probabilistically speaking, they occur very rarely and can be ignored. The complexity situation, however, does not change. To see this, we can simply consider the number of local genealogies of K samples as N_K. Increasing to $K+1$ we have $N_{K+1} = N_K \times (2K-2) + 1$. $2K-2$ is the number of edges in the binary tree with K leaves, $N_K \times (2K-2)$ is the number of ways to insert the new node on an existing edge, and the 1 is for the new node to form an outgroup of itself. This recursive relation defines an exponential growth. To complicate the matter, we also discretized time. This corresponds to forming different trees by varying branch lengths. To circumvent this obstacle but maintain the capability of analyzing multiple sequences, we use a composite likelihood approach over HMMs constructed from pairwise sequence alignments.

2.2 Mathematical models

Hidden Markov model

The observations of this HMM are the symbols on the sequence alignment. The possible outcomes, therefore, are being the same, different, or unknown due to missing data. The hidden states of this HMM are the different coalescent trees that could have caused such sequence alignment. Since we consider just two sequences, the coalescent trees, i.e. the hidden states, are just the time points when coalescence takes place between the two samples. The number of hidden states is therefore determined by the total number of time slices.

Assume we have k time slices with break points $\tau_0,\ \tau_1, \cdots,\ \tau_k$. HMM state i corresponds to the two samples coalescing in the time interval $[\tau_i,\ \tau_{i+1}]$, where $\tau_k = \infty$, implying the two samples eventually coalesce in one of the

given time slices. For example, in Figure 2.1 we have four HMM states, two in the migration period and two in the ancestral period. Since samples reside in two different populations, no coalescent events can occur during the isolation period.

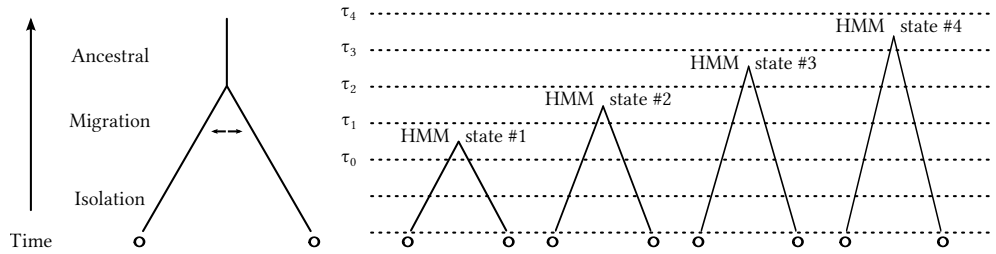


Figure 2.1: Example HMM hidden states for two samples in a demography that involves an isolation period, a migration period, and an ancestral period. During the migration and ancestral periods, coalescence is possible between the samples. Then number of hidden states is determined by the discretization of time for the periods when coalescence is possible.

Conditional on the hidden states, HMM emission probabilities are the probabilities of seeing a certain alignment column. Similar to other coalescence HMM methods, we apply Jukes-Cantor's substitution model [15], which is a simple mutation model in which the rate of substitution λ is constant.

Transition probabilities

CoalHMM uses the continuous time Markov chain (CTMC) to explicitly explore all possible nucleotide configurations given a population structure. From the CTMC, we calculate the probability of samples starting and ending with certain configurations. Concatenating the CTMCs for a sequence of time slices gives us the probability of a certain sequence of events. In particular, we are interested in all of the probabilities of two neighboring nucleotides reaching their common ancestors.

Continuous time Markov chain

To compute the joint probability of two neighboring genealogies, we explicitly evaluate all states of a two-loci coalescent process [10, 21, 32, 33]. Figure2.2 shows the state space for two samples in a single population, the single CTMC.

In our implementation, we represent each state as follows. The sets $\{1\}$ and $\{2\}$ denote a locus on sequence 1 and sequence 2 and that they have not found their common ancestor. The set $\{1, 2\}$ denotes an ancestral linkage of $\{1\}$ and $\{2\}$. The two neighboring nucleotides are represented as a pair of such states; for example $(\{1, 2\}, \{1\})$ denotes a lineage in which the left nucleotide has found a common ancestor, and the ancestor is linked to a neighboring

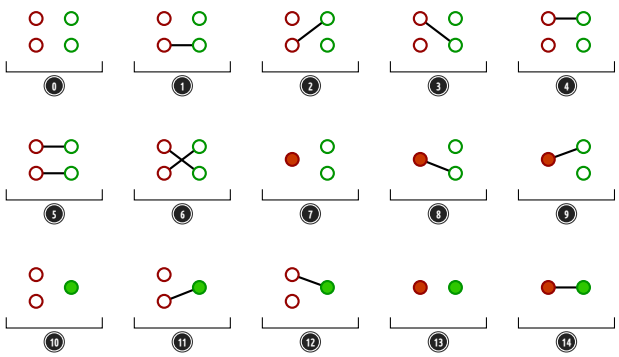


Figure 2.2: Graphical representation of the single CTMC, where two samples reside in a single population. The samples are free to coalesce and recombine.

nucleotide from just sequence 1, which has not yet found a common ancestor with sequence 2. Finally, to assign lineages to populations, we pair them with a population symbol, for example $(1, (\{1, 2\}, \{1\}))$ denotes a two-nucleotide lineage $(\{1, 2\}, \{1\})$ residing in population 1. A set of all possible configurations is shown in Figure 2.3, which forms the set representation of the single CTMC shown in Figure 2.2

```
0    {(0, ({1}, {})), (0, ({2}, {})), (0, ({}, {1})), (0, ({}, {2}))}
1    {(0, ({1}, {1})), (0, ({2}, {})), (0, ({}, {2}))}
2    {(0, ({1}, {2})), (0, ({2}, {})), (0, ({}, {1}))}
3    {(0, ({1}, {})), (0, ({2}, {1})), (0, ({}, {2}))}
4    {(0, ({1}, {})), (0, ({2}, {2})), (0, ({}, {1}))}
5    {(0, ({1}, {1})), (0, ({2}, {2}))}
6    {(0, ({1}, {2})), (0, ({2}, {1}))}
7    {(0, ({1, 2}, {})), (0, ({}, {1})), (0, ({}, {2}))}
8    {(0, ({1, 2}, {1})), (0, ({}, {2}))}
9    {(0, ({1, 2}, {2})), (0, ({}, {1}))}
10   {(0, ({1}, {})), (0, ({2}, {})), (0, ({}, {1, 2}))}
11   {(0, ({1}, {1, 2})), (0, ({2}, {}))}
12   {(0, ({1}, {})), (0, ({2}, {1, 2}))}
13   {(0, ({1, 2}, {})), (0, ({}, {1, 2}))}
14   {(0, ({1, 2}, {1, 2}))}
```

Figure 2.3: Set representation of the single CTMC. The corresponding graphical representation is shown in Figure 2.2.

The rates of state transitions in this CTMC directly follow the rates of different actions, i.e. coalescence, recombination, and migration. Merging two samples of one locus requires a coalescent event at this locus. Linking two samples at two loci requires a coalescent event somewhere else on the sequence. Disconnecting two samples at two loci requires a recombination event. Relocating a lineage requires a migration event. If one CTMC state is not reachable from another by any of these actions, the rate of transition would be zero. Table 2.1 shows the rate matrix for the single CTMC shown in Figure 2.2, where R is the recombination rate, and C is the coalescent rate.

Q	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	-	C	C	C	C	0	0	C	0	0	C	0	0	0	0
1	R	-	0	0	0	C	0	0	C	0	0	C	0	0	0
2	R	0	-	0	0	0	C	0	0	C	0	C	0	0	0
3	R	0	0	-	0	0	C	0	C	0	0	0	C	0	0
4	R	0	0	0	-	C	0	0	0	C	0	0	C	0	0
5	0	R	0	0	R	-	0	0	0	0	0	0	0	0	C
6	0	0	R	R	0	0	-	0	0	0	0	0	0	0	C
7	0	0	0	0	0	0	0	-	C	C	0	0	0	C	0
8	0	0	0	0	0	0	0	R	-	0	0	0	0	0	C
9	0	0	0	0	0	0	0	R	0	-	0	0	0	0	C
10	0	0	0	0	0	0	0	0	0	0	-	C	C	C	0
11	0	0	0	0	0	0	0	0	0	0	R	-	0	0	C
12	0	0	0	0	0	0	0	0	0	0	R	0	-	0	C
13	0	0	0	0	0	0	0	0	0	0	0	0	0	-	C
14	0	0	0	0	0	0	0	0	0	0	0	0	0	R	-

Table 2.1: Rate matrix for the single CTMC. R is the recombination rate, and C is the coalescent rate. The CTMC's graphical representation is shown in Figure 2.2, and the set representation is shown in Figure 2.3.

With the full CTMC state space and its rate matrix, according to the CTMC theory we can compute the probability of observing a certain state at a certain time by computing the matrix exponentiation. Figure 2.4 shows another CTMC state space involving migration, and Table 2.2 shows its rate matrix.

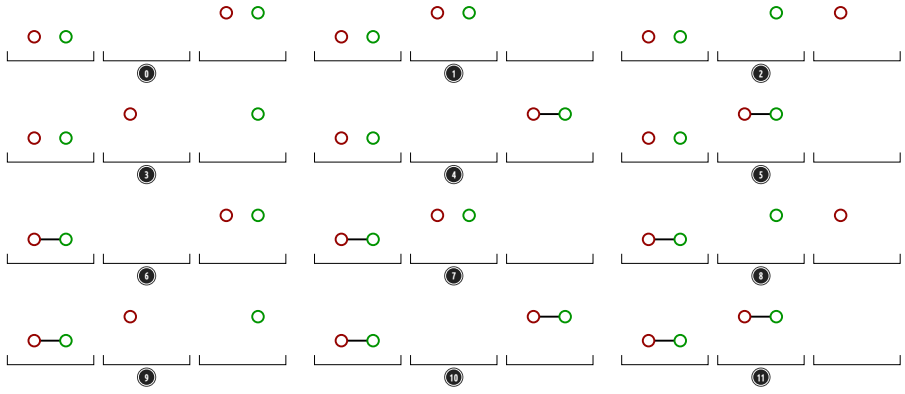


Figure 2.4: Graphical representation of another example CTMC. Two samples reside in a three populations. The second and third populations allow migrations between them.

Joint probability

We divide CTMC states into four categories, begin (B), left (L), right (R), and end (E). In a B type CTMC state, neither loci have coalesced, e.g. all states in the CTMC shown in Figure 2.4 are type B. In an L type CTMC state, the left and only the left locus has coalesced, e.g. state 7, 8, and 9 in the CTMC shown in Figure 2.2 are type L. In an R type CTMC state, only

Q	0	1	2	3	4	5	6	7	8	9	10	11
0	-	0	m32	m32	c3	0	c1	0	0	0	0	0
1	0	-	m23	m23	0	c2	0	c1	0	0	0	0
2	m23	m32	-	0	0	0	0	0	c1	0	0	0
3	m23	m32	0	-	0	0	0	0	0	c1	0	0
4	r	0	0	0	-	m32	0	0	0	0	c1	0
5	0	r	0	0	m23	-	0	0	0	0	0	c1
6	r	0	0	0	0	0	-	0	m32	m32	c3	0
7	0	r	0	0	0	0	0	-	m23	m23	0	c2
8	0	0	r	0	0	0	m23	m32	-	0	0	0
9	0	0	0	r	0	0	m23	m32	0	-	0	0
10	0	0	0	0	r	0	r	0	0	0	-	m32
11	0	0	0	0	0	r	0	r	0	0	m23	-

Table 2.2: Rate matrix of the CTMC shown in Figure 2.4. C1, C2, C3 are the coalescent rates in three populations, respectively; m23 and m32 are the migration rates going from population 2 to 3 and vice versa; finally R is the recombination rate.

the right locus has coalesced, e.g state 10, 11, and 12 in the CTMC shown in Figure 2.2 are type R. Finally, in an E type CTMC state, both loci have found their common ancestors, e.g state 13 and 14 in the CTMC shown in Figure 2.2 are type E.

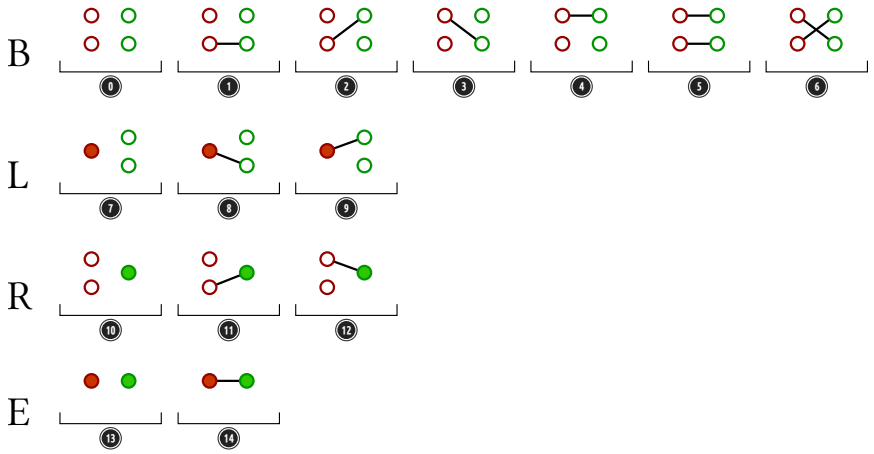


Figure 2.5: State categorization for the single CTMC. The graphic representation of this CTMC is shown in Figure 2.2. This CMC contains 7 begin states (B), 3 left states (L), 3 right states (R), and 2 end states (E)

Under the four category schema, we can split each CTMC's rate matrix and probability matrices into 16 sections. Shown in Figure 2.6 is an example of one such categorization of the rate matrix for the single CTMC. Some sections are entirely zeros. This is because coalescence is an irreversible process. Once two samples find their common ancestor, they stay merged. These sections are: L to B, R to B, E to B, L to R, R to L, E to L, or E to R. In other words, valid transitions consist of only B to B, B to L, B to R, B to E, L to E, R to E, and E to E.

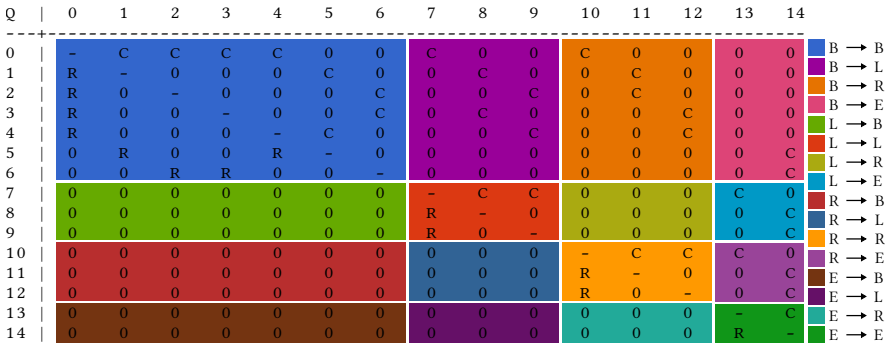


Figure 2.6: The categorized rate matrix for the single CTMC. This CTMC's graphical representation is shown in Figure 2.2 and the set representation is shown in Figure 2.3.

The transition probability matrix has size N by N, where N is the number of hidden states. With the CTMC and a specific path going through each time slice, we can compute the probability of this particular path. Combining all paths that have the same beginning, State #5 shown in Figure 2.5, and one of the E states, we can compute a joint probability. State #5 is a B state, and it depicts the modern-day condition of two loci on two sampled sequences. Any one of the E states would suffice as the ending condition because we are concerned with only the coalescence status of the two loci.

Let's call the joint probability matrix J, where J_{ij} is the probability of observing coalescence of the left nucleotide in time slice i and coalescence of the right nucleotide in time slice j. Let's call the HMM's transition matrix T. To get T_{ij}, we normalize row i in J,$T_{ij} = J_{ij} / \sum_m^N J_{mj}$. To calculate J_{ij}, we use the probability matrices produced by the sequence of CTMCs. Let's call the initial state the *zero state*. State #5 shown in Figure 2.5 is the zero state for the single CTMC. Shown below are the three possible ways to reach a type E state.

$$\begin{aligned}
&1.\ (B \to B)_{\text{zero or more}},\ B \to E \\
&2.\ (B \to B)_{\text{zero or more}},\ B \to L,\ (L \to L)_{\text{zero or more}},\ L \to E \\
&3.\ (B \to B)_{\text{zero or more}},\ B \to R,\ (R \to R)_{\text{zero or more}},\ R \to E
\end{aligned} \tag{2.1}$$

The joint probability matrix is symmetric because the chance of taking path #3 is the same as taking path #2. According to CTMC theory, we compute the probability matrix for each time slice $P = \exp(Q \cdot \Delta t)$, where Δt is the duration of the time slice. We need the probability matrices from the most recent time to the time slice when both loci find their common ancestor.

We consider a collection of event sequences by multiplying together the probability matrices. This corresponds to integrating all paths that belong to a certain type. For example, $(P)_{BB}$ represents the top-left slice of P, like

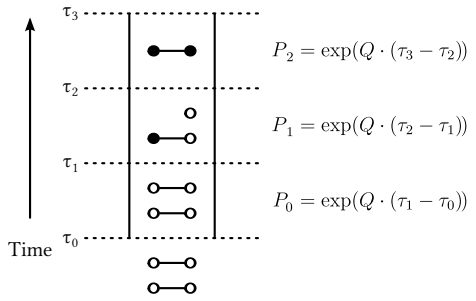


Figure 2.7: Example probability matrices over time slices. This diagram illustrates one particular path of two samples starting from a B state and ending in a E state. Specifically, the two loci are connected in the modern samples, i.e two adjacent loci on the alignment. Going backward in time, nothing changed in the first time slice; the left locus coalesced during the second time slice, and the right locus coalesced during the third time slice.

shown in Figure 2.6. The sum of all values in $(P)_{BB}$ is the probability of starting the time slice in one of the begin states and ending the time slice also in one of the begin states. The exact begin and end states are not important. Together with the three paths, we can now derive the analytical forms to calculate the joint probability matrix.

$$J_{ij} = \begin{cases} \sum_\alpha \sum_\beta M_{a\beta} & \text{when } i \leq j \\ J_{ji} & \text{when } i > j. \end{cases} \tag{2.2}$$

$$M_{ij} = \begin{cases} (P_0)_{0B} \times \cdots \times (P_{i-1})_{BB} \times (P_i)_{BL} \times (P_{i+1})_{LL} \times \cdots \times (P_{j-1})_{LL} \times (P_j)_{LE} & i < j \\ (P_0)_{0B} \times (P_1)_{BB} \times \cdots \times (P_{i-1})_{BB} \times (P_i)_{BE} & i = j \end{cases} \tag{2.3}$$

For example, the following is the joint probability of observing a sequence of events through three time slices in which the left locus coalesces in the second and the right locus coalesces in the third. One such path is illustrated in Figure 2.7

$$J_{23} = \sum_\alpha \sum_\beta \left((P_0)_{0B} \times (P_1)_{BL} \times (P_2)_{LE} \right)_{\alpha\beta}$$

Figure 2.8 shows two other possible paths. Explicitly listing all possible paths becomes impossible when we have more time slices and more types of CTMCs. The number of combinations increases exponentially.

Projection Matrix

Under different CTMCs, matrix dimensions or the mapping of B, L, R, and E would fail to match. Mostly likely both properties would fail. This forbids concatenating probability matrices through multiplication. When this happens,

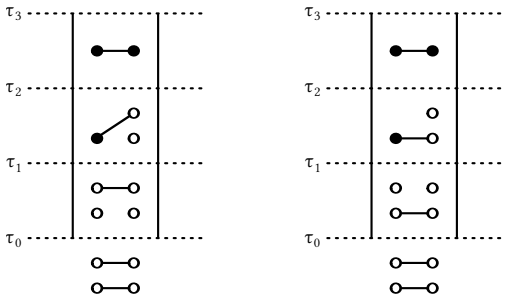


Figure 2.8: Another two paths covered by joint probability J_{23}. The path shown in Figure 2.7 is also covered by J_{23}.

we need projection matrices to move samples from one kind of CTMC state space to another. We obtain this state mapping by resetting population labels in the set representations of the CTMC states like the ones shown in Figure 2.3. For example, if CTMC state $(1, (\{1\}, \{1\}))\,, (2, (\{2\}, \{2\}))$ in a two-sample two-population CTMC were to be merged into a state in a single-population CTMC, we could relabel the population symbols, $(0, (\{1\}, \{1\}))\,, (0, (\{2\}, \{2\}))$. Extending to all state configurations, if we were to connect the aforementioned two kinds of CTMCs, we would have the following projection mapping.

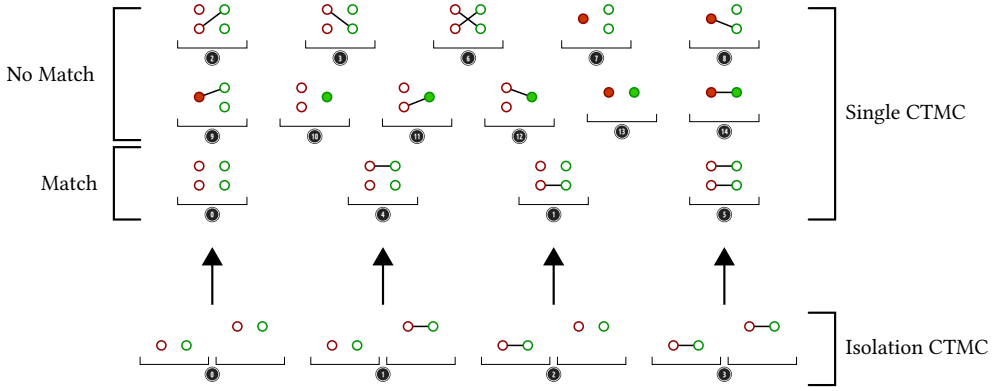


Figure 2.9: Graphical representation of the projection mapping from the isolation CTMC to the single CTMC for two samples. The single CTMC's graphical representation is shown in Figure 2.2, and its set representation is shown in Figure 2.3. The isolation CTMC involves two isolated populations and two samples each residing in one population.

To place the projection matrix in the context of calculating the joint probabilities, we would insert appropriate projection matrices into the chain of matrix multiplications. For example, in the situation illustrated in Figure 2.10, we have three time slices in which two belong to a two-population migration CTMC and one belongs to a single-population ancestral CTMC.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0

Table 2.3: Matrix representation of the projection mapping from the isolation CTMC to the single CTMC for two samples. The four isolation CTMC states map to four single CTMC states. We record 1 for each mapping and 0 for the rest.

We would like to compute $(P_0)_{0B} \times (P_1)_{BL} \times (P_2)_{BE}$, where state zero is the initial state reflecting the modern sample configuration of the two loci. But the first matrix multiplication is illegal because P_0 has a dimension of 4×4 but P_1 and P_2 have a dimension 15×15. We must instead do $(P_0)_{0B} \times (P_{\text{Isolation}\to\text{single}})_{BB} \times (P_1)_{BL} \times (P_2)_{BE}$, where $P_{\text{isolation}\to\text{single}}$, as shown in Figure 2.9, and Table 2.3 maps states between two CTMCs.

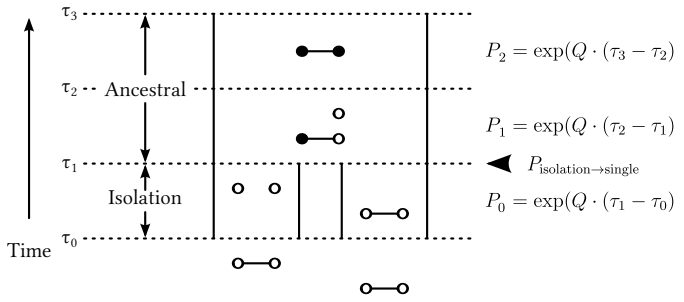


Figure 2.10: Example probability matrices. Joint probability J_{23} covers the illustrated path. To compute J_{23} requires probability matrices from three time slices and the CTMC projection matrix given in Table 2.3.

Admixture CoalHMM

With a foundation of CTMC and HMM constructions, we proceed to investigate a special kind of demographic that involves gene flow in the form of admixture events. We try to infer parameters that are associated with the admixture events. We call this modeling *admixture CoalHMM*.

Introduction to admixture

In previous CoalHMM models, gene flow happens as continuous migration. This does not, however, model cases in which gene flow occurs quickly among populations, possibly caused by certain environmental changes. We call the latter *admixture events*, and we model this kind of gene flow as single instantaneous occurrences rather than periods with migration rates. Admixture events

introduce changes to the proportions of samples in various of populations, like a reshuffling with controlled rates.

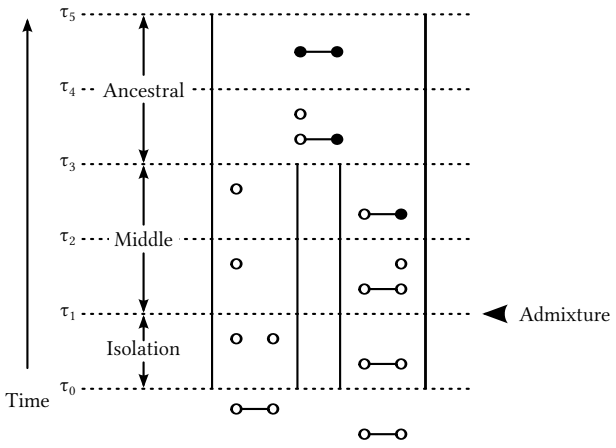


Figure 2.11: Admixture example. This path illustrates two samples residing in two isolated populations for the first time slice, experiencing an admixture event between the first and second time slices, and eventually entering a period when the two populations become a single ancestral population. The admixture event introduced lineage exchange between the populations and hence formed a different CTMC. In this case, the middle period CTMC is identical to a two-population migration CTMC.

Figure 2.11 shows an example of an admixture event occuring during a two-population isolation period. Without this admixture event, the middle epoch would also be an isolation epoch. With the gene flow introduced by the admixture event, the middle epoch behaves like a two-population migration epoch. During the admixture event, samples in the two isolated populations are shuffled. Any combination of togetherness is allowed. The resulting CTMC configuration is, therefore, identical to one that describes a two-population migration scenario.

Admixture projection

It comes down to the construction of admixture projections, which are inserted into the chain of matrix multiplications to accommodate the sample reordering during admixture events. We compute the admixture projection mapping by explicitly calculating the probabilities of starting from a CTMC state and ending at a CTMC state after going through an admix event. In other words, we fully describe the admixture effect as a collection of probabilities, p_{ij}, for moving a sample from population i to population j.

To do so, we first identify the number of pieces, i.e. connected components, in a source state. We then identify the location for each piece. In the example

illustrated in Figure 2.12, we have two populations and two connected pieces. In the source states, each population has one piece.

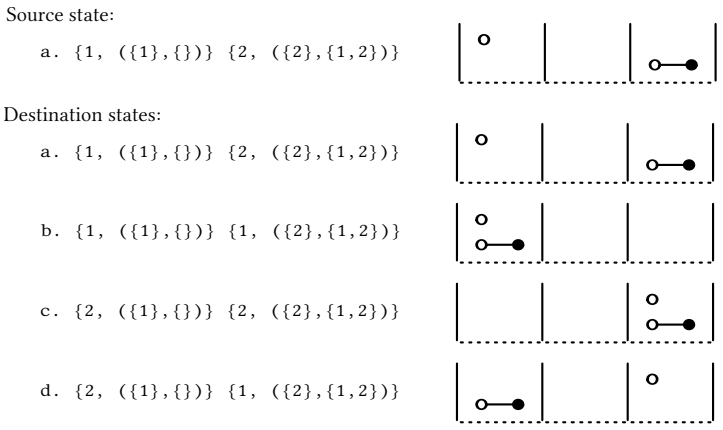


Figure 2.12: Admixture projection source and destination example. During an admixture event between two populations, each connected component has a choice of moving to the other population or staying in its original population. These movements change the state configuration.

During an admixture event, a connected component may either move or stay, so there are k^n outcomes, where k is the number populations involved in this demographic and n is the number of pieces in the source state. In Figure 2.12, we have $k = 2$ and $n = 2$, and therefore, $2^2 = 4$ possible outcomes, i.e. 4 possible destination states.

To calculate the probabilities of transiting to each one of the possible destination states, we use admixture proportions, which describe the strength of a historical admixture event. The proportions are part of model parameters. In our two-population case, we simply iterate the binary representation of values from 0 to 4, and assign a probability for each one accordingly. Table 2.4 shows an example of an admixture proportion of $p = 0.1$ for a sample going from population 1 to 2 and a proportion of $q = 0.2$ for a sample going from population 2 to 1.

binary index	pieces	destination	probability
$0_d = 00_b$	nobody moves	a	$(1-p)\cdot(1-q) = 0.72$
$1_d = 01_b$	left piece stays; right piece moves	b	$(1-p)\cdot q = 0.18$
$2_d = 10_b$	left piece moves; right piece stays	c	$p\cdot(1-q) = 0.08$
$3_d = 11_b$	both pieces move	d	$p\cdot q = 0.02$

To form the admixture projection matrix, we perform this calculation for all source states, which are CTMC states immediately prior to the occurrence of the admixture event. The sum of all probabilities from one source state is

..	... b ..... d ..... a ..... c ...			
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
a	0.18	0.02	0.72	0.08
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

Table 2.4: Example admixture projection calculation. This example follows what is shown in Figure 2.12. With specific admixture proportions, $p = 0.1$ and $q = 0.2$, we can compute the probability of arriving at each destination state. The proportions p and q are the probabilities of a sample moving from population 1 to 2 and from 2 to 1, respectively.

always one, $(1-p)\cdot(1-q) + (1-p)\cdot q + p\cdot(1-q) + p\cdot q = 1 + pq - p - q + q - pq + p - pq + pq = 1$. A source state inevitably turns into one of the destination states. From the calculations in Table 2.4, we can fill one row in the admixture projection matrix.

HMM construction

Depending upon the availability of the samples from extant populations, we can construct HMMs from pairwise alignments of different configurations. Figure 2.13 to 2.16 demonstrate four ways of constructing HMMs from two sequences under a three-population admixture demography.

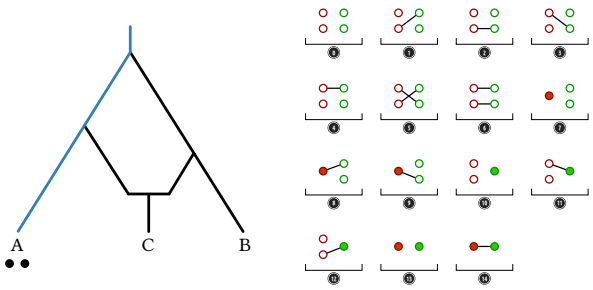


Figure 2.13: CTMC construction for HMM with two samples both from a source population. The model parameters are coalescent rate and recombination rate.

The HMM shown in Figure 2.13 requires one type of CTMC. In this CTMC, we have a single population. The two samples coalesce and recombine freely. This CTMC contains 15 states.

The HMM shown in Figure2.14 requires two types of CTMCs. The recent CTMC involves two isolated populations. During this time, samples located in the same population coalesce and recombine freely—but not across populations. This CTMC contains 4 states. The distant CTMC involves a single

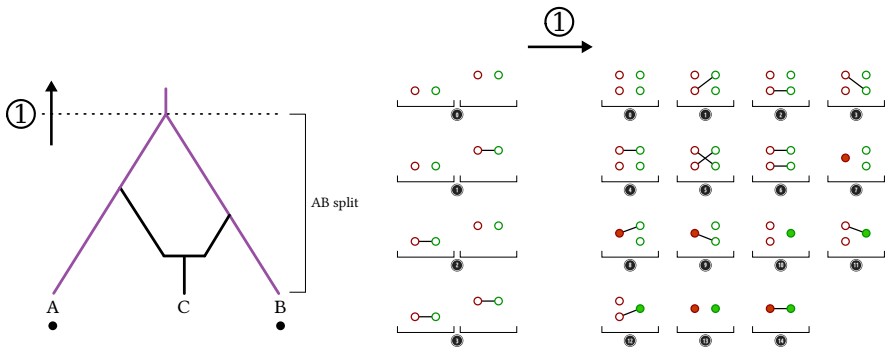


Figure 2.14: CTMC construction for HMM with two samples, one from each of the two source populations. The model parameters are AB split, coalescent rate, and recombination rate.

ancestral population. During this time, samples coalesce and recombine freely. This CTMC contains 15 states.

The HMM shown in Figure2.15 requires four types of CTMCs. The most recent CTMC is the same as the ones from Figure 2.13 and 2.14. The second most recent CTMC involves three populations. During the admixture event, a sample can move to one of the two populations, each with a certain probability. The CTMC state space is equivalent to allowing migration between the second and third populations while keeping the first population isolated. Samples in the second and third populations, therefore, coalesce, recombine, and migration freely—but not with samples in the first population. This CTMC contains 12 states. The second most distant CTMC involves two isolated ancestral populations. During this time, samples coalesce and recombine freely within a population. Since the first population is formed by merging the first two populations from the previous CTMC, the situation is equivalent to having an asymmetric migration, i.e. allowing for migration from the second population to the first but not the other way around. This CTMC contains 29 states. The most distant CTMC involves a single ancestral population, where all samples coalesce and recombine freely. This CTMC contains 15 states.

The HMM shown in Figure2.16 requires three types of CTMCs. The recent CTMC involves a single population, where all samples coalesce and recombine freely. This CTMC contains 15 states. In the middle CTMC, all samples are free to coalesce, recombine, and migrate freely. This CTMC contains 94 states. The distant CTMC involves a single ancestral population. During this time, two pairs of samples coalesce and recombine freely. This CTMC contains 15 states, the same as the recent CTMC.

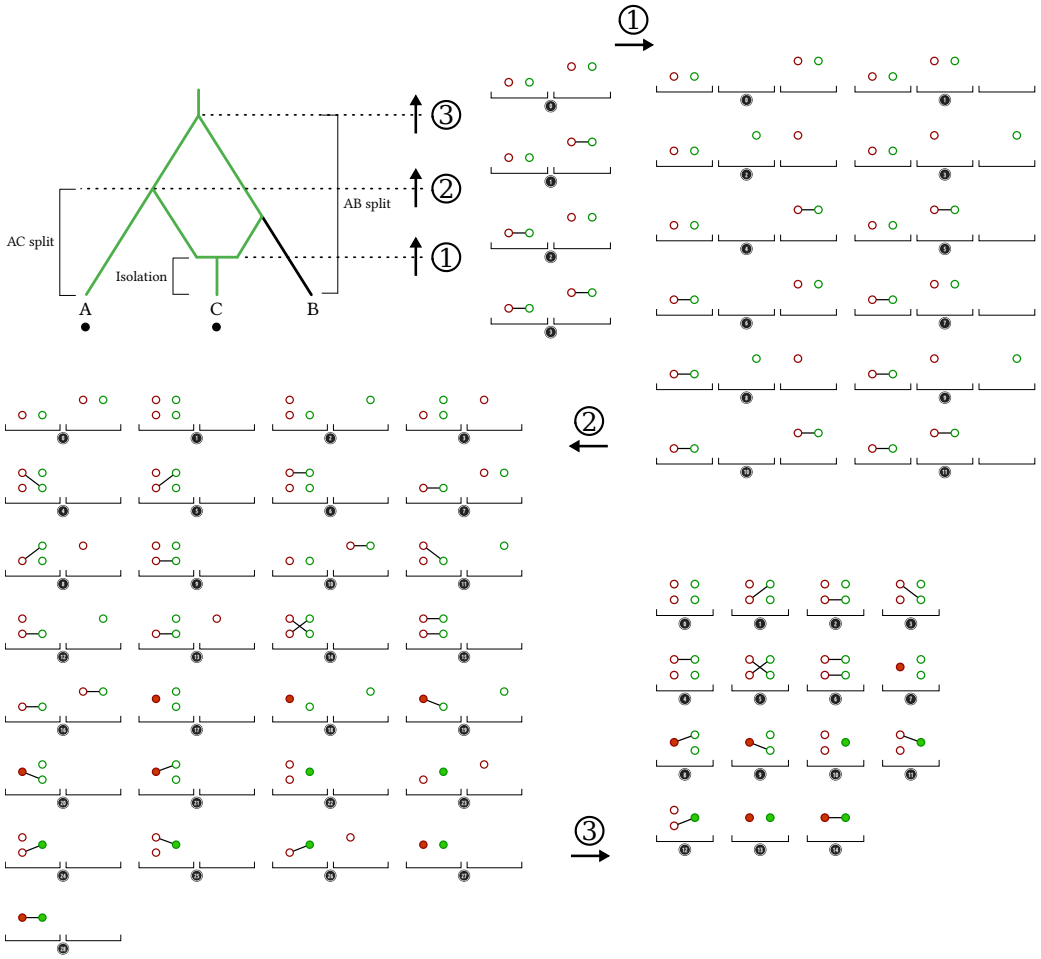


Figure 2.15: CTMC construction for HMM with two samples, one from the admixed population, the other from one of the two source populations. The model parameters are admixture time, AC split, AB split, coalescent rate, and recombination rate.

Composite likelihood

When we have a multiple sequence alignment, the idea of integrating paths belonging to a certain category would still work, but it becomes more tedious. Instead of a 4×4 dicing of the rate and probability matrices, the type of CTMC stats would increase exponentially as the number of the sequence increases. To take advantage of multiple sequence alignments while avoiding the computational complexity from exact integration, we apply a composite likelihood approach.

In the composite likelihood schema, we extract all pairwise alignments from a multiple sequence alignment and form a HMM for each pairwise alignment as

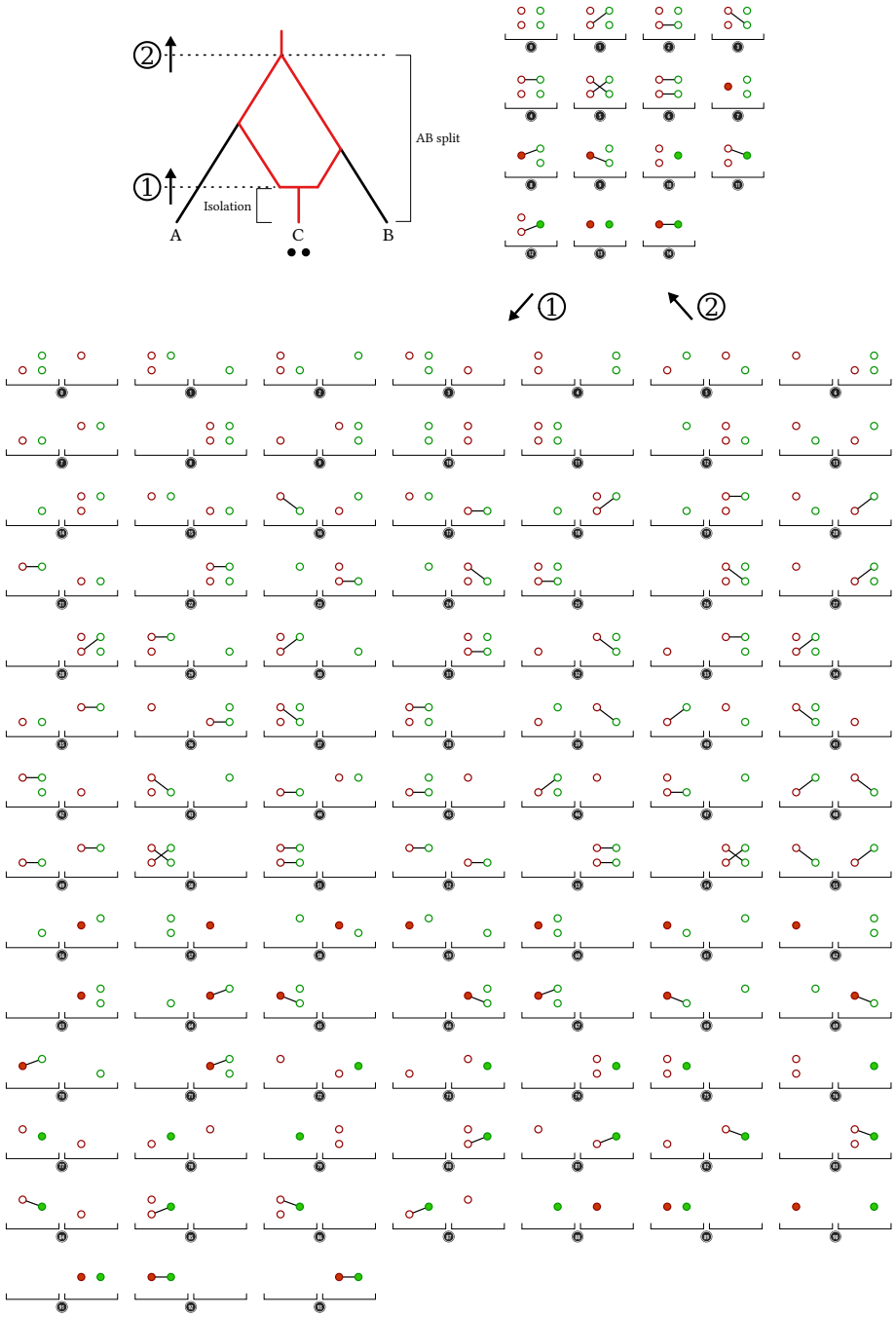


Figure 2.16: CTMC construction for HMM with two samples both from the admixed population. The model parameters are the admix time, AB split, coalescent rate, and recombination rate.

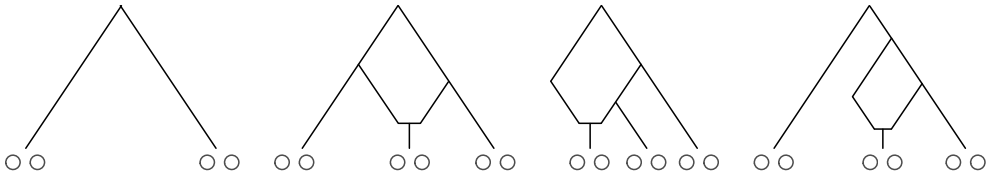


Figure 2.17: Example models that can be constructed using pairwise alignments and the composite likelihood approach.

described in previous sections. We infer parameters based on the summation of the likelihood values from all HMMs. Using the composite likelihood schema and the HMMs shown from Figure 2.13 to Figure 2.16, we can construct a range of admixture CoalHMM models. These models are useful under different conditions depending on the availability of data from extant populations.

Model \ Population	A	B	C	samples per population
#1	×	×	✓	• •
#2	✓	×	✓	• •
	×	✓	✓	• •
#3-1	✓	✓	✓	•
#3-2	✓	✓	✓	• •
#3-3	✓	✓	✓	★ •

Figure 2.18: A summary for models described from Figure 2.19 to Figure 2.23.

We use Model #1 when we have access to only the admixed population. In this case, we make inference from two sequences collected from population C and build a single HMM from them.

If we have data for the admixed population and one of the two source populations, we can use Model #2. We construct three HMMs: one using two sequences, both from the admixed population (the same as in Model #1); one using two sequences, both from the source population; and one using two sequences, one from each population.

If we have data from all three populations, we have more ways to explore the data. In Model #3-1, we use only one sample per population. For Model #3-2, we use two samples per population, and we construct one HMM for each of the six types of HMMs. For Model #3-3, we construct fifteen HMMs for all pairwise alignments of the six samples, two from each population.

2.3 Parameter inference

In the maximum likelihood estimate, we infer CoalHMM parameters by optimizing the log-likelihood value calculated from HMM's forward algorithm.

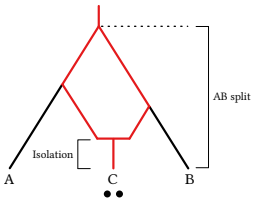


Figure 2.19: Model #1: Admixture model with only the admixed population.

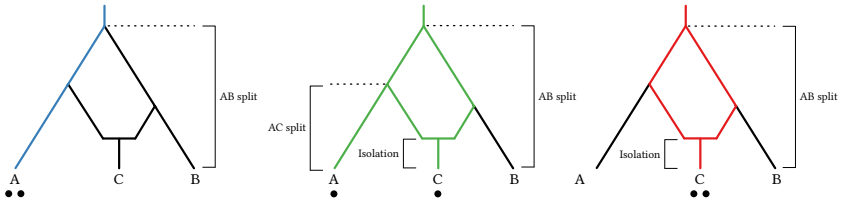


Figure 2.20: Model #2: Admixture model with the admixed population and one of the two source populations.

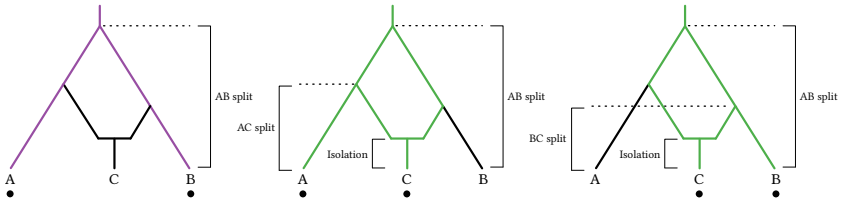


Figure 2.21: Model #3-1: Full admixture model with one sample from each population.

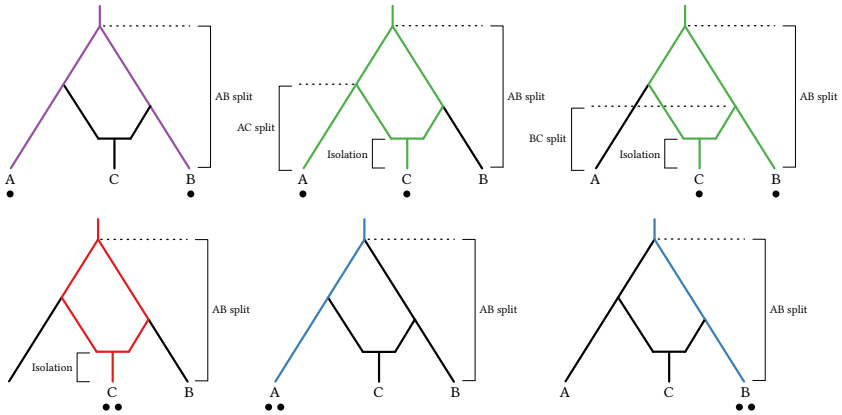


Figure 2.22: Model #3-2: Full admixture model with two sample from each population and six HMMs.

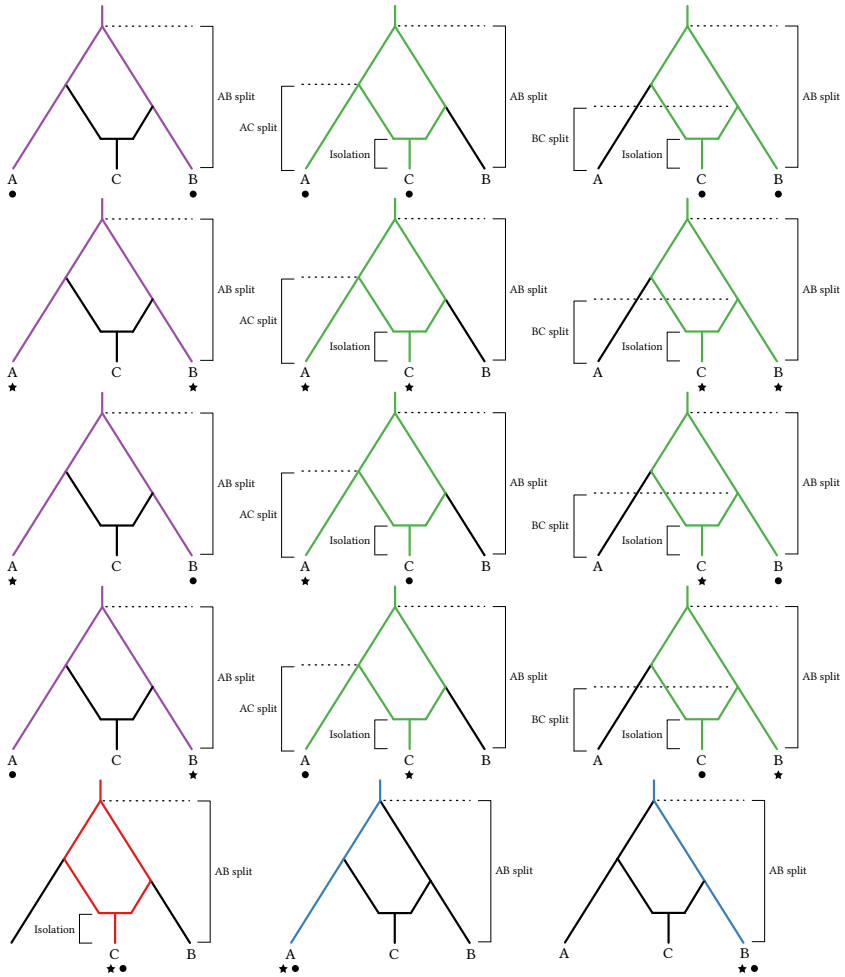


Figure 2.23: Model #3-3: Full admixture model with two sample from each population and fifteen HMMs.

We treat this processes as a black-box style optimization. We implement and compare two types of optimization processes: the deterministic simplex approach and nondeterministic evolutionary algorithms.

Nelder-Mead simplex method

John Nelder and Roger Mead introduced the Nelder-Mead simplex method in 1965 [25] as a technique to minimize an objective function in a many-dimensional space. The pseudo-code below summarizes this method.

This method uses several algorithm coefficients to determine the amount of effect of possible actions. They are the reflection coefficient ρ, the expansion coefficient χ, the contraction coefficient γ, and the shrinkage coefficient σ.

Algorithm 1 Nelder-Mead simplex method

repeat
  evaluate each point in the simplex using the objective function
  determine the point $p_{\min}$ with the lowest fitness
  reflect $p_{\min}$ through the centroid of the remaining vertices to p_r
  if the fitness at p_r is the highest in the simplex **then**
    expand p_r away from the centroid to p_e
    use p_e in place of $p_{\min}$
  else if the fitness at p_r is still the lowest **then**
    contract p_r toward the centroid to point p_c
    if the fitness at p_c is no longer the lowest **then**
      use p_c to replace $p_{\min}$
    else
      determine the point $p_{\max}$ with the highest fitness
      shrink all points in the simplex around $p_{\max}$
    end if
  else
    use p_r to replace $p_{\min}$
  end if
until termination condition is reached

Standard values recommended in [3] are $\rho = 1$, $\chi = 2$, $\gamma = 1/2$, and$\sigma = 1/2$.

Evolutionary algorithms

Evolutionary algorithms belong to a subfield of artificial intelligence in computer science. They are population-based heuristic optimization methods inspired by biological processes such as evolutionary reproduction and insects swarming. We investigate and implement two methods from this class, the genetic algorithm (GA) and the particle swarm optimization (PSO)

Genetic algorithm (GA)

John Holland first introduced GAs in the 1970s [13]. The idea is to encode each solution as a chromosome-like data structure and operate on them through actions analogous to genetic alterations, which usually involves selection, recombination, and mutation. For each type of alteration, people have developed different techniques.

Selection determines a subset of the current population to use when forming the next generation. The Roulette Wheel Selection algorithm by [12] and Stochastic Universal Sampling by [2] lead the way as two fitness proportion selection methods. Tournament Selection by [23] is ranking-based, and it selects the highest fitness value from a random subset of the population.

Recombination, also known as crossover, is the second stage of a GA. It combines the selected individual to breed the next generation. Similar to the biological process, crossover can happen with arbitrarily many junction points. The uniform crossover scheme by [34], however, does not have a close biological analogue. When a complex scheme fails to offer a gain in optimization, we resort to a simpler scheme, such as one-point crossover.

Mutation is the last stage. It introduces a single point of modification to the solution represented by the newly-generated population. To create a random change, naturally, we can use various processes for random sampling. Uniform mutation from [22] and Gaussian mutation from [5] are the most common choices. In practice, the selection of mutation greatly influences the optimization process.

Particle swarm optimization (PSO)

Eberhart and Kennedy first introduced PSO in 1995 [6] as an optimization technique relying on stochastic processes, similar to GAs. As its name implies, each individual solution mimics a particle in a swarm. Each particle holds a velocity and keeps track of the best positions it has experienced and best position the warm has experienced. The former encapsulates the social influence, i.e. a force pulling towards the swarm's best. The latter encapsulates the cognitive influence, i.e. a force pulling towards the particle's best. Both forces act on the velocity and drive the particle through a hyper parameter space.

Parameter space rescaling

To allow the optimization procedures to cover a wide range of values, we perform a log scaling of the parameter space prior to optimizing, and we scale the estimated quantities reversely before each objective function evaluation.

2.4 Simulation study

To evaluate our framework, we conduct an extensive amount of simulations. In this section we describe two main sets of simulation studies. In the first, we use the program **ms** [14] to generate ancestral recombination graphs under standard neutral evolutionary models with recombination, speciation, variable populations, migrations, etc. We then use the **seq-gen** [30] program to produce sequence samples of length 10 Mbp. Using the phylogenetic trees simulated by **ms** as input, **seq-gen** evolves the sequences along the phylogeny. In the second, we use the program **fastsimcoal2** [7, 8] for continuous-time sequential Markovian coalescent simulations. We simulate demographics involving splitting and fusing of populations, admixture events, changes in migration matrices, etc. From the simulated polymorphic sites of a pairwise sequence, we calculate the HMM observations.

CoalHMM model with isolation and migration

The isolation model is the simplest demographic model in our simulation study. The model contains three parameters: a coalescence rate, a recombination rate, and a split time, where the ancestral population is split into two isolated populations [21].

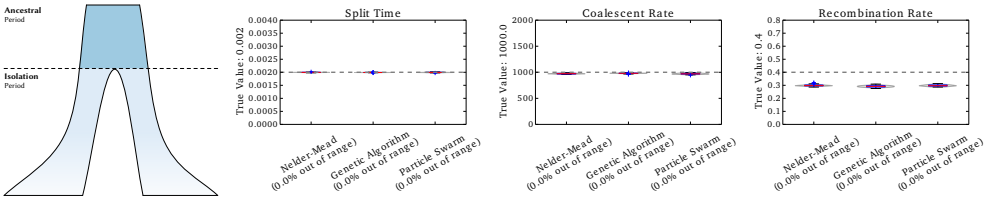


Figure 2.24: Isolation model. Parameter estimations are good.

The isolation with initial migration model is the second simplest model. This model contains five parameters: the time period during which the two populations are isolated, the time period during which migration persists, a shared coalescence rate for all populations, a migration rate for the migration epoch, and a recombination rate.

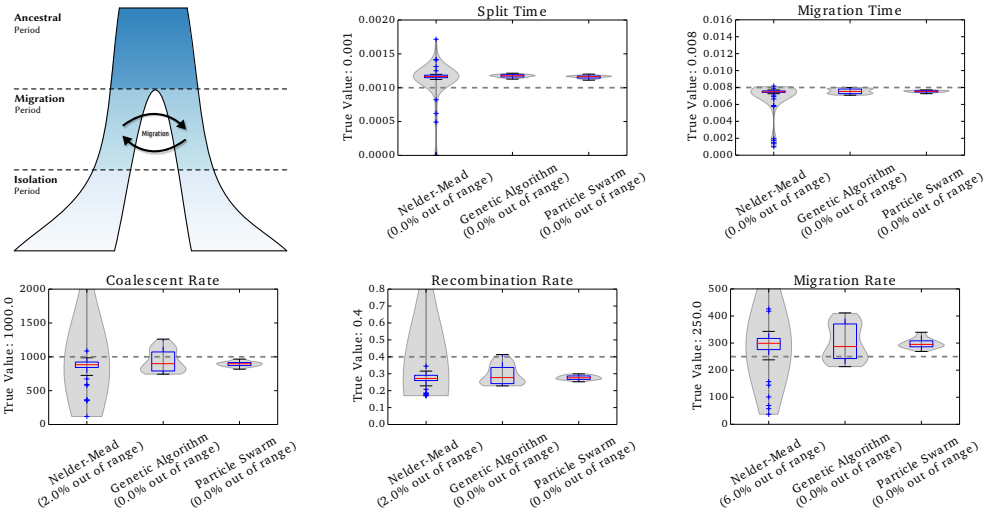


Figure 2.25: Isolation with migration model. Parameter estimations are good.

The isolation with initial migration epochs model extends from the isolation with initial migration model. This model allows for multiple epochs within the isolation period, the migration period, and the ancestral period. Both coalescence rates and migration rates can vary freely between epochs.

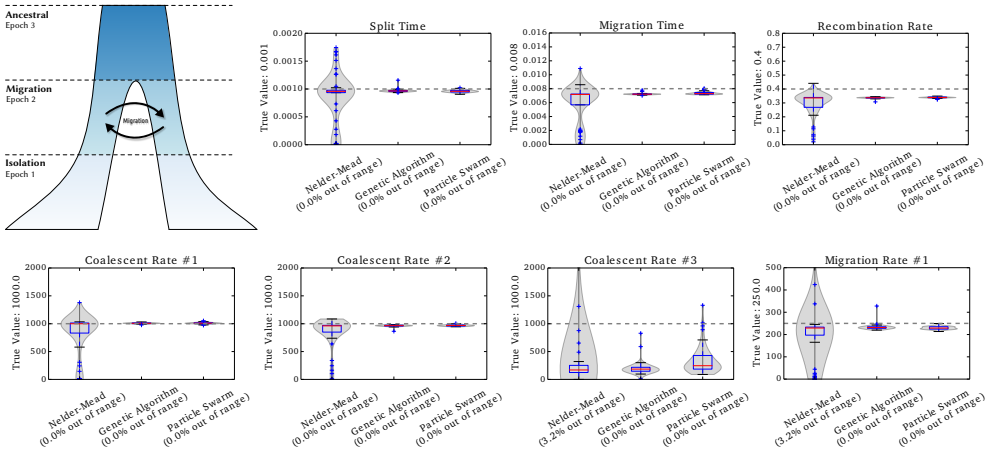


Figure 2.26: Isolation with migration 3-epochs model. Parameter estimations are good except for the coalescent rate far back in time.

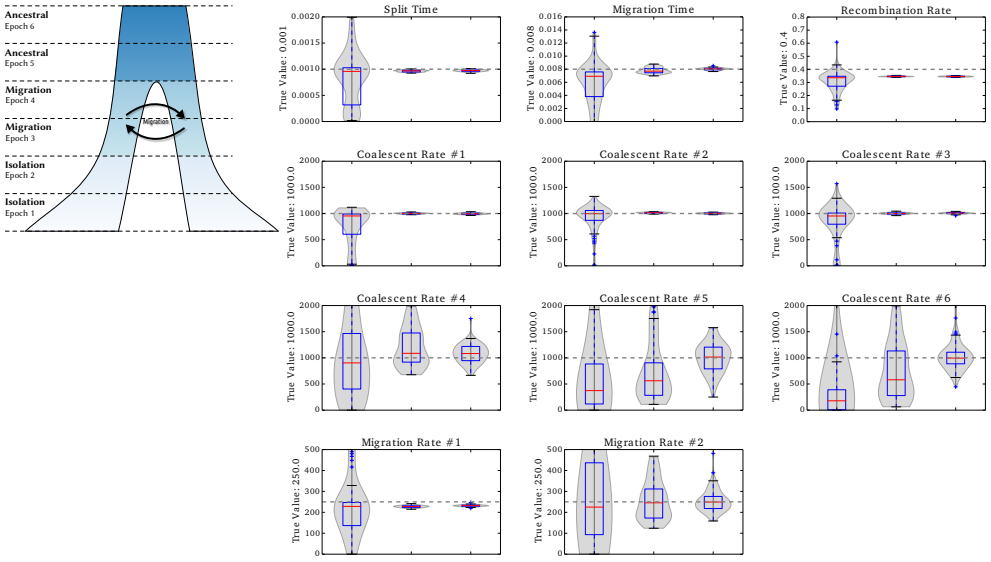


Figure 2.27: Isolation with migration 6-epochs model. Parameter estimations are good except for the coalescent rate far back in time.

CoalHMM model with admixture

In this section, we investigate the performance for admixture CoalHMM modeling. First we show the estimation accuracy. To do this, we simulate sequence data under different demographic scenarios, and we apply a range of admixture models. Second, we show the effect when the admixture model is mis-specified. In other words, we show the admixture inference results when the sequences are simulated under demographic scenarios that are different

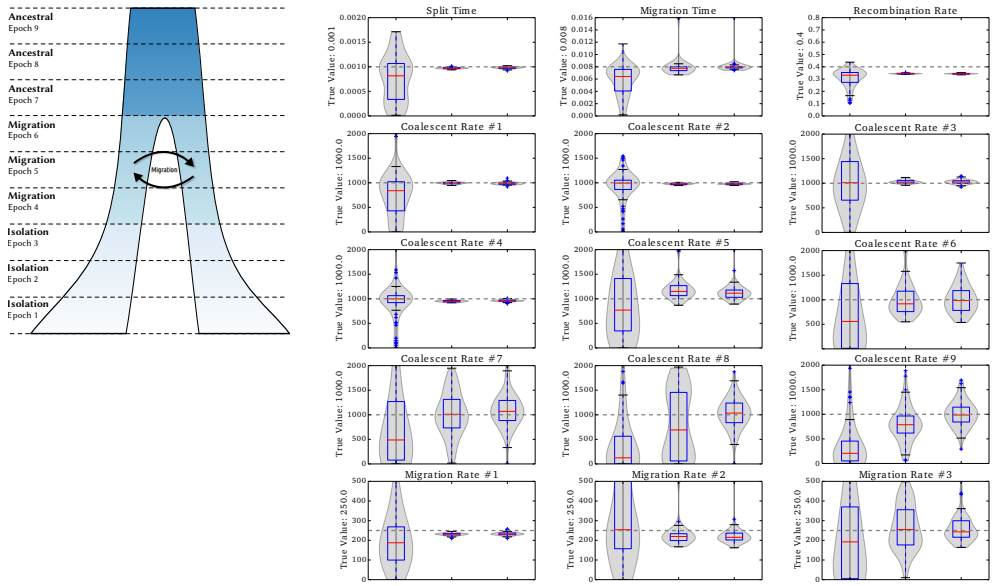


Figure 2.28: Isolation with migration 9-epochs model. Parameter estimations are good except for the coalescent rate far back in time.

from what is specified by the admixture CoalHMM model.

Admixture CoalHMM estimation accuracy

Figure 2.29 summarizes the estimation accuracy of the five admixture models outlined in Figure 2.18. Details of the model constructions using the composite likelihood approach are illustrated from Figure 2.19 to Figure 2.23. Figure 2.30 shows the recovery of varying admixture proportions. Figure 2.31 shows the recovery of varying coalescent rates.

Effect of wrongly specified demography

In parametrized modeling, the parametrizations of are not generally unique. This brings forward the question about model mis-specification, which happens when the model described in the method fails to capture reality. In this section, we show how the inference results are affected when the true demography is different from what is modeled in admixture CoalHMM.

Admixture CoalHMM models summarized in Figure 2.18 model a three-population admixing scenario, where two populations are related to the ancestors of the two source populations of an admixture event, and the third population is the direct descendant of the admixed population. Many alternative demographies exist for three populations. Figure 2.32 illustrates the alternative where one population is in fact an out-group. Figure 2.33 illustrates the alternative where continuous gene flow occurs after the admixture

event. Figure 2.34 illustrates the alternative where continuous gene flow occurs between two ancestor populations. The admixture CoalHMM models shown in Figure 2.18 fail to accurately describe any of these demographics.

From Figure 2.32 to Figure 2.34, we observe failed attempts when applying admixture CoalHMM on any of the mis-specified models. Some of them are affected more dramatically than others. The existence of an out-group significantly skews all admixture-related parameters, as shown in Figure 2.32. The effect of continuous gene flow affects the estimates, but at different levels. Ancient gene flow prior to the admixture event appears to have only a minor effect, as shown in Figure 2.34, but recent gene flow skews admixture-related estimates significantly, as shown in Figure 2.33.

2.5 Biological data analysis

We have applied admixture CoalHMM to several data projects, and we highlight two data studies in this section. The first is a study of the admixture history for the bear family including the brown bear, polar bear, ABC island bear, and black bear. We used admixture CoalHMM to analyze these bear species under various admixing scenarios. We mentioned this work in the second method paper, included as Appendix B, and the analysis results were also used to support the conclusions in [19]. We studied baboons in another data project. As members of the baboon consortium, we used the most probable admixture graph and analyzed several major species in the baboon family, cynocephalus, ursinus, kindae, hamadryas, and papio. We are currently composing this manuscript.

In both data projects, for each sample triplet, we analyzed the full autosomal genome and obtained the variance in the estimates from a blocked bootstrap with the genome split into 10Mb blocks and 100 repetitions. We also performed simulation tests of goodness-of-fit and de-biasing of estimates. To examine which parameters are likely to be biased, and by how much, we simulated data with parameters in a grid of time points around the estimated points and estimated the final parameters from the simulated data.

2.6 Concluding remarks

Three years ago, when I started my PhD in the field of population genetics, I joined the CoalHMM project team and proposed to expand the existing CoalHMM framework to study historical admixture. Prof. Thomas Mailund and I identified several major stages towards this goal. Firstly, I would improve the optimization module of the framework so that we could produce reliable estimates especially when the number of model parameters increases. Secondly, I should be able to build CoalHMM models in a modular fashion so

that different model constructions would be possible. Finally, I would model admixture events and estimate admixture-related model parameters.

The paper in Appendix A summarizes my work in the first two stages. I investigated, implemented, and compared several black-box style optimization techniques with the emphasis on heuristic-based evolutionary algorithms. This paper also presents a range of models demonstrating the capability of complex model construction. Finally, in this paper I present simulation studies to evaluate these new additions: the optimization module and model constructions.

The manuscript in Appendix B summarizes my work during the third stage. I implemented admixture CoalHMM to infer historical admixture events and constructed multiple admixing demographics. Admixture CoalHMM not only learns the admixture time but also the proportions of gene flow from different source populations. Also in this paper, I present a panel of simulation evaluations, and I demonstrate good inference accuracy under different demographics. I also show the effect of using admixture CoalHMM on wrongly modeled demographics. Together, I present admixture CoalHMM a as new tool to study historical admixture events.

I have applied admixture CoalHMM to several large collaborative projects. As a member of the baboon consortium, I studied the admixing relationship among several baboon species. I contributed in a bear gnomic study by analyzing several bear species' speciation times and admixing history. I also analyzed equids, lynx, and human genome data using admixture CoalHMM.

Future work

The future of CoalHMM lies in the automation of model construction. Together with the composite likelihood approach, I now have the means to construct CoalHMM models for any given demography involving population splits, continuous migration, and admixture events. In theory, I can estimate model parameters and recover the evolutionary history depicted by any given demography. In the current stage of CoalHMM's development, model construction is a case-by-base process. This framework would, however, be more powerful if model construction were automated. It will be a significant undertaking in the algorithm and software design. Besides that, the inference procedure will face optimization challenges and state-space complexity issues. The increased number of parameters calls for better optimization. Current optimization methods do not scale well with the number of parameters. In addition, the CTMC state-space grows exponentially with factors such as the number of populations and the presence of gene flow. CTMC state-space directly concerns the complexity of matrix exponentiation, and because of that, certain demographic epochs may become computational bottlenecks.

After automated model construction it comes model selection. With the power to explore a range of demographics of different population splits and gene flow, a natural question to ask is which demography is the best and

should be proposed as the evolution history of the species under investigation. Many model selection techniques exist, such as Akaike information criterion and Approximate Bayesian computation. It is a straightforward task to incorporate model selection into the framework. The challenge, however, is the sensitivity of the maximum likelihoods. In other words, the optimization power and run-time complexity, as described in the previous paragraph, may also pose potential difficulties to model selection.

I see the possibility of splitting the CoalHMM framework into two stages: the model construction plus complexity prediction stage and the parameter inference plus model selection stage. From the user's point of view, the workflow would be as follows. First, in a hypothesizing phase, the user would sketch several demographic scenarios. Second, in a modeling phase, the user would use CoalHMM to outline several modeling plans. This step would also include a complexity analysis that identifies parameter count for optimization strength and CTMC state spaces for possible computational bottlenecks. Third, in a inferencing phase, the user would execute the modeling plans and infer parameters that describe the hypothesized demographics. Finally, in a decision phase, the user would conduct model selection among the proposed hypotheses and identify the best demography.

The first phase stems from the researcher's understanding of the biological data. The second phase should be computationally inexpensive, and it should provide a guideline to refine the first stage. The third phase would be expensive in terms of time, CPU, and space. I would advise a high performance cluster service for this phase. Due to the limitations originating from optimization and CTMC state space, the third and fourth phases might not be feasible for complex demographies.

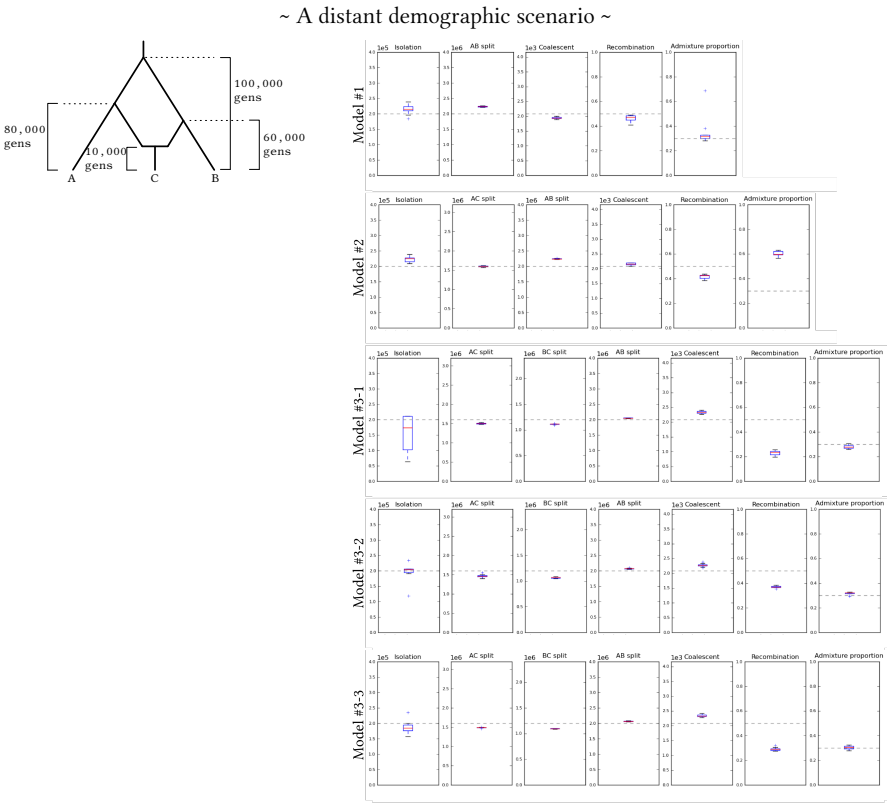


Figure 2.29: Parameter estimation from five admixture CoalHMM models under two demographic scenarios. One is distant back in time (top), and the other is more recent (bottom).

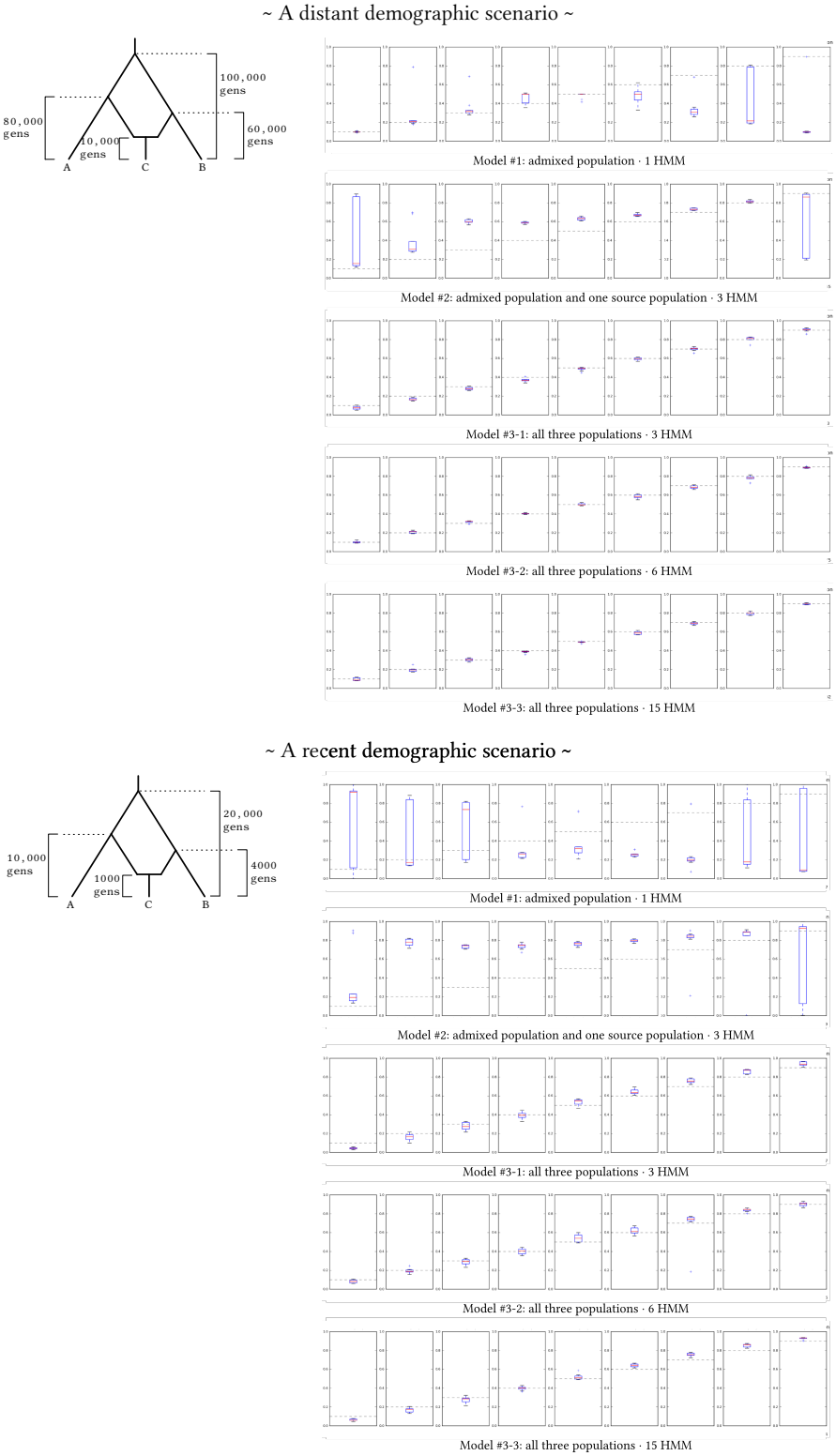


Figure 2.30: Admixture proportion estimations from five admixture CoalHMM models under two demographic scenarios. One is distant back in time (top), and the other is more recent (bottom).

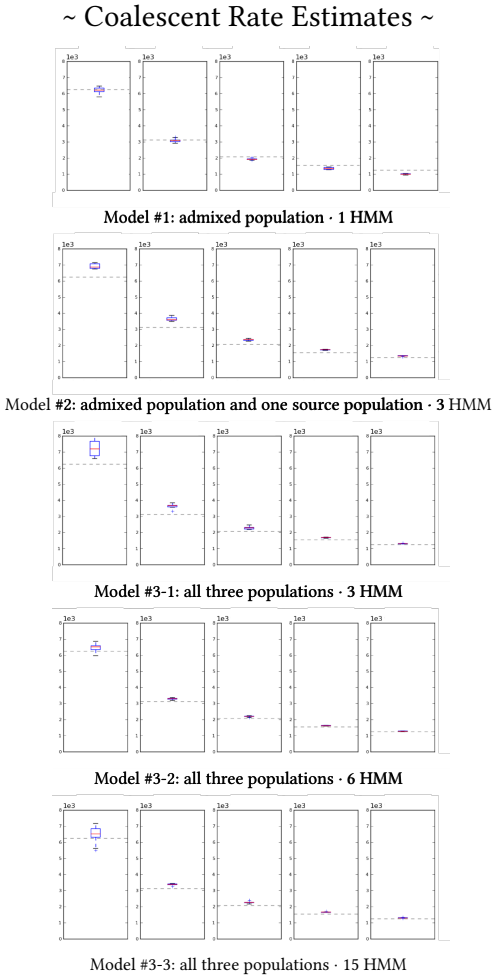


Figure 2.31: Coalescent rate estimations from five admixture CoalHMM models for a ranges of five different simulation values.

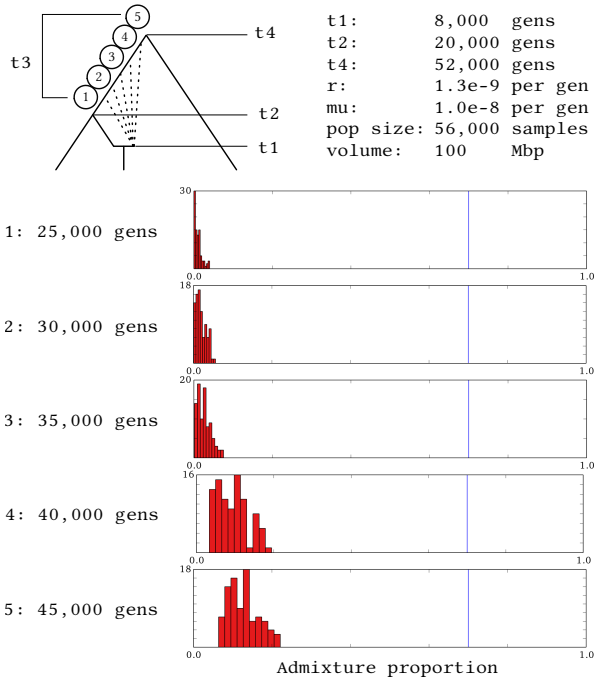


Figure 2.32: Mis-specified demography type1. In this scenario, an outgroup is modeled as a source population.

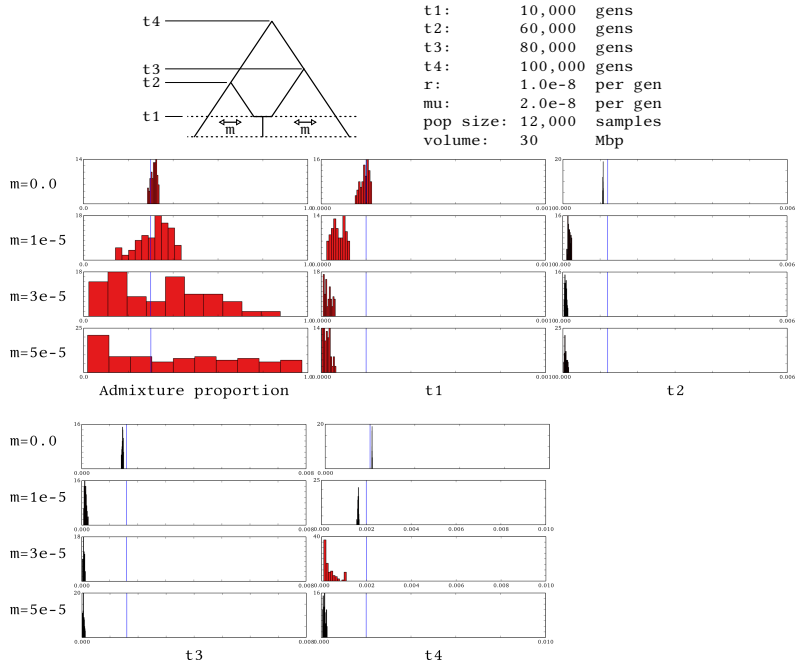


Figure 2.33: Mis-specified demography type2. In this scenario, the admixed population maintains a constant gene flow between the two source populations after the admixture event.

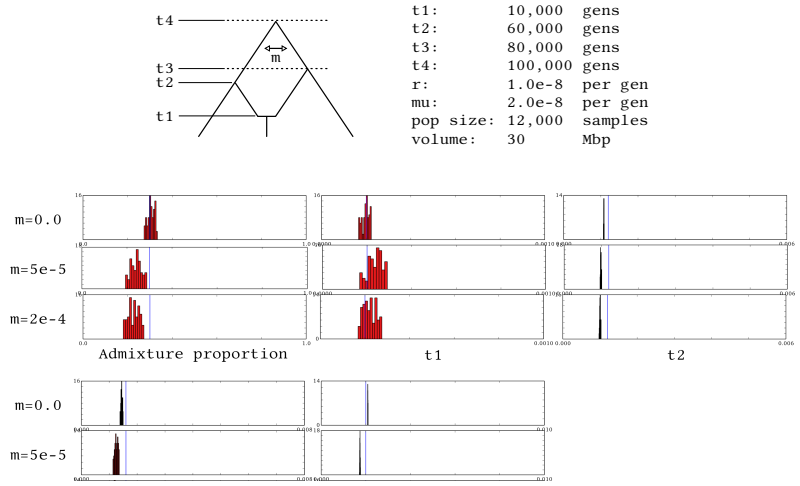


Figure 2.34: Mis-specified demography type3. In this scenario, the ancestors of the two source populations go through a period of time when there is continuous gene flow between them.

Chapter 3

Ohana

3.1 Introduction

The population genetics community has been using unsupervised learning models to study population structure and individual admixture for the past two decades [29]. This process assigns fractional memberships of a set of ancestry components to each individual. The natural next question to ask is regarding the evolutionary relationships among these ancestry components. Allele frequencies estimated during the structure analysis can provide us this information [28]. Furthermore, researchers have also begun using allele frequencies caused by population structure as evidence to detect ongoing positive selection [26]. Selection may increase the level of genetic differentiation among populations by acting on local adaptation-related mutations. Selection may also act on beneficial mutations that arise in specific geographical locations and cause a temporary increase in the level of population differentiation. Local positive selection, therefore, could be responsible for loci exhibiting genetic distances larger than the average genetic distance among the populations [17, 26].

We present Ohana, a set of tools that starts with structure analysis, proceeds to population tree inference, and finally conducts selection analysis while fully taking advantage of the structured genomic data. Ohana builds its mathematical models and optimization techniques on top of well-established methods. Ohana infers individual clustering from which we identify outliers for selection analyses.

The current release of Ohana contains five programs: **qpas**, **cpax**, **nemeco**, **selscan**, and **convert**. **qpas** and **cpax** perform structure inference using genotype observations or genotype likelihoods. To achieve this goal, **qpas** and **cpax** use different algorithms to solve sequential quadratic programming. **qpas** uses an adaptation of the active set algorithm, while **cpax** uses a variation of the complementarity pivoting algorithm. At the current stage of development, we recommend using **qpas** because it achieves better likelihoods

in benchmark tests. Program **nemeco** infers population variances and covariances assuming components are rooted at one of the populations. Program **selscan** selects covariance outliers as candidates for selection study. Program **convert** facilitates different stages of the analysis by providing file conversions and fast approximations through five submodules, **ped2dgm**, **bgl2lgm**, **cov2nwk**, **nwk2svg**, and **nwk2cov**.

The source code, installation instructions, high-level documentation, and example workflows are available on GitHub.

https://github.com/jade-cheng/ohana

Detailed doxygen documentation is available at the following URL.

http://jade-cheng.com/ohana/

The rest of this section will be divided into four subsections. In the first subsection, we will outline the probabilistic models used in the inference processes. In the second subsection, we will describe the different numerical optimization strategies and their mathematical derivations. In the third subsection, we will present simulation studies to evaluate Ohana. Finally in the last subsection, we will analyze real genomic data and present their interpretations. Except for the joint inference process, all modules are released on GitHub.

3.2 Mathematical models

We outline two mathematical models, first to infer global ancestry and second to infer population relations. We implement the first model in program **qpas** and **cpax**. We implement the second model in program **nemeco**.

When performing data analysis, we first infer admixture using **qpas** or **cpax** by supplying genetic data, either called genotypes or genotype likelihoods. We then infer population covariances using **nemeco** by supplying allele frequencies, a byproduct produced from the admixture inference.

Structure analysis

Structure analysis estimates overall genome-wide percentage contributions from different ancestries for each sample individual. The main outcome is a set of component labels and their percentages for each individual. The input of this process is the genomic data containing certain markers from certain individuals. This data can be genotype observations for high-coverage data or genotype likelihoods for low- or medium-coverage data. The first piece of output is percentage values indicating admixture proportions for each individual at each component. The second piece of output is allele frequencies for each component at each marker.

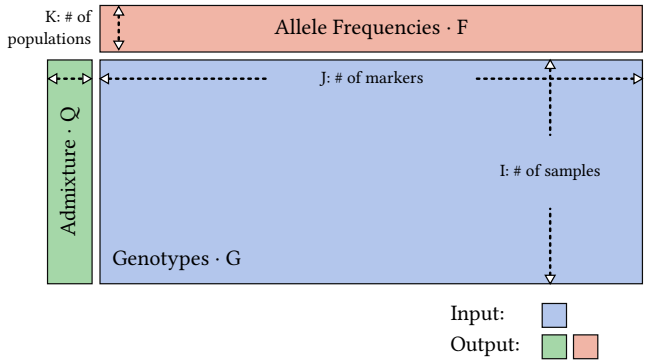


Figure 3.1: An illustration of the input and output for program **qpas** and **cpax**, the admixture inference module of Ohana. The genotype input, G matrix shown above, can take one of two forms: genotype observations or genotype likelihoods. High-coverage genomic data produce reliable called genotypes. This is, however, not always available, so Ohana also models genotype likelihoods and provides support for low- or medium-coverage data.

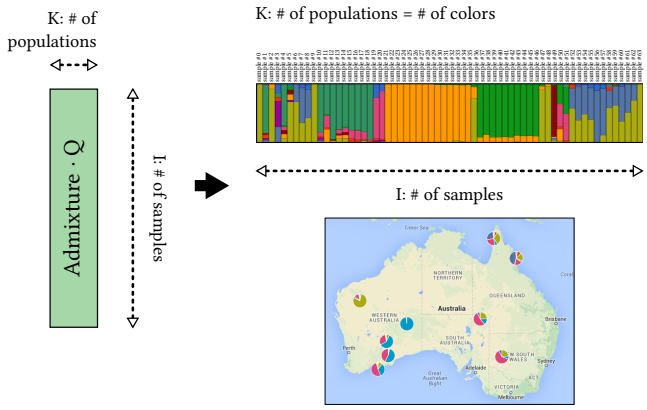


Figure 3.2: Bar charts provide a direct graphic representation of the admixture results. Together with geographic records of the participating samples, map diagrams provide a more informative view of the same results.

We apply a model-based inference approach. The structure analysis part of the statistical model $P_1(Q,\ F)$ is similar to what is used in existing software such as STRUCTURE [29], FRAPPE [35], ADMIXTURE [1], and SPA [36]. It denotes the probability of assigning admixture proportions Q to individuals with corresponding allele frequencies F. We can work with not only high quality genomic data in the form of genotype observations $P_1^O(Q,\ F)$ but also lower coverage data in the form of genotype likelihoods $P_1^L(Q,\ F)$.

Genotype observations

We formulate the analytical expression of the likelihood model as the following when we apply genotype observations as the input. We count the occurrences of major and minor alleles from a given dataset. We denote K as the number of populations, I as the number of individuals, and J as the number of polymorphic sites.

$$\ln\left[P_1^O\left(Q,\ F\right)\right] \quad = \quad \sum_i^I \sum_j^J \left\{ g_{ij} \cdot \ln\left[\sum_k^K q_{ik} \cdot f_{kj}\right] + (2 - g_{ij}) \cdot \ln\left[\sum_k^K q_{ik} \cdot \left(1 - f_{kj}\right)\right]\right\}.$$

Genotype likelihoods

We formulate the analytical expression of the likelihood model as the following when we apply genotype likelihoods as the input. Values g_{ij}^{AA}, g_{ij}^{Aa}, and g_{ij}^{aa} are the probabilities of observing the sequence data at the ith individual's jth marker, conditioned on genotype AA, Aa (or aA), and aa, respectively. Let us define the probability of having a major allele, conditioned on Q and F, at the ith individual's jth marker to be $A_{ij} = \sum_m^K q_{im} \cdot f_{mj}$ and the probability of having a minor allele, conditioned on Q and F, at this location to be $B_{ij} = \sum_m^K q_{im} \cdot (1 - f_{mj})$.

$$\begin{aligned}
P_1^L\left(Q,\ F\right) &= \sum_g \left[Pr\left(\text{read data} \mid g\right) \cdot Pr\left(g \mid Q,\ F\right)\right] \\
Pr\left(\text{read data}_{ij} \mid g_{ij}\right) &= \begin{cases} g_{ij}^{AA} & \text{for } AA \\ g_{ij}^{Aa} & \text{for } Aa \text{ or } aA \\ g_{ij}^{aa} & \text{for } aa \end{cases} \\
\ln\left[P_1^L\left(Q,\ F\right)\right] &= \sum_i^I \sum_j^J \ln\left(g_{ij}^{AA} \cdot A_{ij}^2 + g_{ij}^{aa} B_{ij}^2 + g_{ij}^{Aa} \cdot 2A_{ij} B_{ij}\right).
\end{aligned}$$

Population covariance analysis

Population covariance analysis estimates variances and covariances among population components. Given a covariance matrix, we can draw conclusions about genetic distances among these components. The inputs of this process are allele frequencies for each population component at each marker. If we assume these components follow a tree-like evolution, we can approximate the estimated covariance matrix into a tree structure, which provides a direct representation, like shown in Figure 3.3.

We model the joint distribution of allele frequencies across all ancestry components as a multivariate Gaussian similar to TreeMix [28]. $P_2\left(F\right)$ is the probability of having such ancestral allele frequencies given the populations. The variance of the multivariate normal distribution is a product of two factors

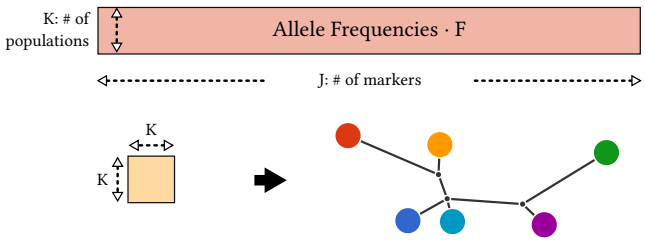


Figure 3.3: One way to visualize the covariance estimates is to approximate the covariance matrix into a phylogenetic tree structure. From the tree representation we learn the genetic relationships among population components.

[28]. The first term $\mu_j (1 - \mu_j)$ is site-specific. The second term, Ω, is constant across sites, and Ω captures population variances and covariances.

$$P(f_j \mid \Omega,\ \mu_j) \quad \sim \quad \mathcal{N}(\mu_j,\ \mu_j (1 - \mu_j)\, \Omega)\,.$$

The ovariance matrix Ω is symmetric and of size $K \times K$. f_j is a vector of size K containing the allele frequencies at site j. Here μ_j is the average allele frequency at site j. It is calculated by dividing the occurrences of the allele by the total alleles at site j, i.e. the sum of the major-major sites and half of the major-minor sites divided by the total number of sites.

$$\begin{aligned}
\ln[P_2(F)] &= \ln\left\{\prod_j^J \left[\frac{1}{\sqrt{|2\pi c_j \Omega|}} \exp\left(-\frac{1}{2}(f_j - \mu_j)^T (c_j\Omega)^{-1} (f_j - \mu_j)\right)\right]\right\} \\
&= -\frac{1}{2} \cdot \sum_j^J \left\{K \cdot \ln(2\pi c_j) + \ln[\det(\Omega)] + \frac{1}{c_j} \cdot (f_j - \mu_j)^T \Omega^{-1} (f_j - \mu_j)\right\} \\
\text{where } c_j &= \mu_j (1 - \mu_j)\,.
\end{aligned}$$

This system is under-determined. It happens because of the symmetry of the Gaussian distribution. One unrooted tree corresponds to multiple different covariance matrices, i.e. rooted trees. These covariance matrices all induce the same probability distribution on the allele frequencies. To address this, we root the tree in one of the observations. This corresponds to calculating the conditional probability of the data given the value observed in one of the populations, which can be arbitrarily chosen. We use the first population as the "root population". Allele frequencies at other loci are replaced by the difference f'_j of the original values f_j, and the corresponding frequency in the first population f_{j_0}. We obtain Ω', a symmetric matrix of size $(K-1) \times (K-1)$.

$$\begin{aligned}
\ln\left[P_2\left(F\right)\right] &= \ln\left\{\prod_j^J\left[\frac{1}{\sqrt{\left|2\pi c_j\Omega'\right|}}\exp\left(-\frac{1}{2}\cdot f_j'^T\cdot\left(c_j\Omega'\right)^{-1}\cdot f_j'\right)\right]\right\} \\
&= -\frac{1}{2}\cdot\sum_j^J\left\{\left(K-1\right)\cdot\ln\left(2\pi c_j\right)+\ln\left[\det\left(\Omega'\right)\right]+\frac{1}{c_j}\cdot f_j'^T\cdot\Omega'^{-1}\cdot f_j'\right\} \\
\text{where } c_j &= \mu_j\left(1-\mu_j\right) \\
f_j' &= f_j - f_{j_0}.
\end{aligned}$$

3.3 Parameter Inference

SQP for structure analysis

In general, we can approximate a function F with its second order Taylor expansion F_T. Each Newton update attempts to find the Δx such that the derivative of F_T with respect to Δx is zero.

$$F_T\left(x_n+\Delta x\right) = F\left(x_n\right)+F'\left(x_n\right)\Delta x+\frac{1}{2}F''\left(x_n\right)\Delta x^2$$

If we use quadratic programming to solve for each Newton's update step, this becomes a sequential quadratic programming (SQP) problem. The second order Taylor expansion forms the quadratic form, $\frac{1}{2}\bar{x}^TQ\bar{x}+c^T\bar{x}$, where c corresponds to $\left[F'\left(x_n\right)\right]^T$, Q corresponds to $F''\left(x_n\right)$, and $F\left(x_n\right)$ is a constant term that can be dropped. In each round of quadratic optimization, we solve for Δx that tells us the direction and amount to take for the next step. This tends to an optima on the objective surface. When we have a bounded problem, the overall problem is convex. In that case, the local optima we find through SQP is also the global optima.

In our case, we need to satisfy a sequence of constraints while searching for Δx. Specifically, $\forall\Delta q_{ik},\ q_{ik}+\Delta q_{ik}\in[0,\ 1]$, $\forall\Delta f_{kj},\ f_{kj}+\Delta f_{kj}\in[0,\ 1]$, and $\forall\Delta q_{ik},\ \sum_k^K\Delta q_{ik}=0$ because $\sum_k^K q_{ik}=1$. We have an equality- and inequality- constraint SQP problem.

In the implementation, we avoid dealing with individual values and iterating through matrix elements in a sequential fashion. We accomplish matrix operations at the highest level possible, i.e. vectors or matrices as a whole.

Derivatives for genotype observation model

We derive the first and second differentials for $\ln\left[P_1^O\left(Q,\ F\right)\right]$ with respect to values in Q.

$$\begin{aligned}
\frac{\partial\left(\ln\left[P_1^O\left(Q,\ F\right)\right]\right)}{\partial q_{ik}} &= \frac{\partial\left(\sum_i^I\sum_j^J\left\{g_{ij}\cdot\ln\left[\sum_k^K q_{ik}\cdot f_{kj}\right]+(2-g_{ij})\cdot\ln\left[\sum_k^K q_{ik}\cdot\left(1-f_{kj}\right)\right]\right\}\right)}{\partial q_{ik}} \\
&= \sum_j^J\left[\frac{g_{ij}\cdot f_{kj}}{\sum_m^K q_{im}\cdot f_{mj}}+\frac{(2-g_{ij})\cdot\left(1-f_{kj}\right)}{\sum_m^K q_{im}\cdot(1-f_{mj})}\right] \\
\frac{\partial^2\left(\ln\left[P_1^O\left(Q,\ F\right)\right]\right)}{\partial q_{ik}\partial q_{i'k'}} &= \begin{cases}-\sum_j^J\left\{\frac{g_{ij}\cdot f_{kj}\cdot f_{k'j}}{\left(\sum_m^K q_{im}\cdot f_{mj}\right)^2}+\frac{(2-g_{ij})\cdot(1-f_{kj})\cdot\left(1-f_{k'j}\right)}{\left[\sum_m^K q_{im}\cdot(1-f_{mj})\right]^2}\right\} & \text{if } i=i' \\ 0 & \text{if } i\neq i'.\end{cases}
\end{aligned}$$

We derive the first and second differentials for $\ln\left[P_1\left(Q,\ F\right)\right]$ with respect to values in F.

$$\begin{aligned}
\frac{\partial\left(\ln\left[P_1^O\left(Q,\ F\right)\right]\right)}{\partial f_{kj}} &= \frac{\partial\left(\sum_i^I\sum_j^J\left\{g_{ij}\cdot\ln\left[\sum_k^K q_{ik}\cdot f_{kj}\right]+(2-g_{ij})\cdot\ln\left[\sum_k^K q_{ik}\cdot\left(1-f_{kj}\right)\right]\right\}\right)}{\partial f_{kj}} \\
&= \sum_i^I\left[\frac{g_{ij}\cdot q_{ik}}{\sum_m^K q_{im}\cdot f_{mj}}-\frac{(2-g_{ij})\cdot q_{ik}}{\sum_m^K q_{im}\cdot(1-f_{mj})}\right] \\
\frac{\partial^2\left(\ln\left[P_1^O\left(Q,\ F\right)\right]\right)}{\partial f_{kj}\partial f_{k'j'}} &= \begin{cases}-\sum_j^J\left\{\frac{g_{ij}\cdot q_{ik}\cdot q_{ik'}}{\left(\sum_m^K q_{im}\cdot f_{mj}\right)^2}+\frac{(2-g_{ij})\cdot q_{ik}\cdot q_{ik'}}{\left[\sum_m^K q_{im}\cdot(1-f_{mj})\right]^2}\right\} & \text{if } j=j' \\ 0 & \text{if } j\neq j'.\end{cases}
\end{aligned}$$

In the implementation, for each row of Q^i, calculating the derivatives requires the corresponding vector G^i, A^i, B^i, and all of Fa and Fb, where $A=Q\cdot Fa$ and $B=Q\cdot Fb$.

$$\begin{aligned}
\frac{\partial\left(\ln\left[P_1^O\left(Q,\ F\right)\right]\right)}{\partial Q^i} &= \sum_j^J\left[\frac{G_j^i}{A_j^i}\cdot Fa^j+\frac{\left(2-G_j^i\right)}{B_j^i}\cdot Fb^j\right] \\
\frac{\partial\left(\ln\left[P_1^O\left(Q,\ F\right)\right]\right)}{\partial Q^i\partial Q^i} &= -\sum_j^J\left\{\frac{G_j^i}{\left(A_j^i\right)^2}\cdot\left[Fa^j\left(Fa^j\right)^T\right]+\frac{2-G_j^i}{\left(B_j^i\right)^2}\cdot\left[Fb^j\left(Fb^j\right)^T\right]\right\}.
\end{aligned}$$

For each column of F^j, calculating its derivatives requires the corresponding G^j, A^j, B^j, and all of Q.

$$\begin{aligned}
\frac{\partial\left(\ln\left[P_1^O\left(Q,\ F\right)\right]\right)}{\partial F^j} &= \sum_i^I\left[\left(\frac{G_i^j}{A_i^j}-\frac{2-G_i^j}{B_i^j}\right)\cdot Q^i\right] \\
\frac{\partial\left(\ln\left[P_1^O\left(Q,\ F\right)\right]\right)}{\partial F^j\partial F^j} &= -\sum_i^I\left\{\left[\frac{G_i^j}{\left(A_i^j\right)^2}+\frac{2-G_i^j}{\left(B_i^j\right)^2}\right]\cdot\left[\left(Q^i\right)^T Q^i\right]\right\}.
\end{aligned}$$

Derivatives for genotype likelihood model

We derive the first and second derivatives of $\ln\left[P_1^L\left(Q,\ F\right)\right]$ with respect to values in the Q matrix.

$$
\begin{aligned}
\frac{\partial\left(\ln\left[P_1^L\left(Q,\ F\right)\right]\right)}{\partial q_{ik}} &= \frac{\partial\left(\sum_i^I\sum_j^J \ln\left(g_{ij}^{AA}\cdot A_{ij}^2+g_{ij}^{aa}\cdot B_{ij}^2+g_{ij}^{Aa}\cdot 2A_{ij}B_{ij}\right)\right)}{dq_{ik}} \\
&= \sum_j^J\left[\frac{G_Q\left(i,j,k\right)}{F\left(i,j\right)}\right] \\
\frac{\partial^2\left(\ln\left[P_1^L\left(Q,\ F\right)\right]\right)}{\partial q_{ik}\partial q_{i'k'}} &= \begin{cases}\sum_j^J\left[\frac{F(i,j)\cdot H_Q\left(i,j,k,k'\right)-G_Q(i,j,k)\cdot G_Q\left(i,j,k'\right)}{F^2(i,j)}\right] & \text{if } i=i' \\ 0 & \text{if } i\neq i'\end{cases} \\
F\left(i,j\right) &= g_{ij}^{AA}\cdot A_{ij}^2+g_{ij}^{aa}\cdot B_{ij}^2+g_{ij}^{Aa}\cdot 2A_{ij}B_{ij} \\
G_Q\left(i,j,k\right) &= \frac{\partial F\left(i,j\right)}{\partial q_{ik}} \\
&= 2g_{ij}^{AA}\cdot f_{kj}\cdot A_{ij}+2g_{ij}^{aa}\cdot\left(1-f_{kj}\right)\cdot B_{ij}+ \\
&\quad 2g_{ij}^{Aa}\cdot\left[A_{ij}\cdot\left(1-f_{kj}\right)+B_{ij}\cdot f_{kj}\right] \\
H_Q\left(i,j,k,k'\right) &= \frac{\partial G\left(i,j,k\right)}{\partial q_{ik'}} \\
&= 2g_{ij}^{AA}\cdot f_{kj}\cdot f_{k'j}+2g_{ij}^{aa}\cdot\left(1-f_{kj}\right)\cdot\left(1-f_{k'j}\right)+ \\
&\quad 2g_{ij}^{Aa}\left[f_{k'j}\cdot\left(1-f_{kj}\right)+\left(1-f_{k'j}\right)\cdot f_{kj}\right].
\end{aligned}
$$

We derive the first and second derivatives of $\ln\left[P_1^L\left(Q,\ F\right)\right]$ with respect to values in the F matrix.

$$
\begin{aligned}
\frac{\partial\left(\ln\left[P_1^L\left(Q,\ F\right)\right]\right)}{\partial f_{kj}} &= \frac{d\left(\sum_i^I\sum_j^J \ln\left(g_{ij}^{AA}\cdot A^2+g_{ij}^{aa}\cdot B^2+g_{ij}^{Aa}\cdot AB\right)\right)}{df_{kj}} \\
&= \sum_i^I\left[\frac{G_F\left(i,j,k\right)}{F\left(i,j\right)}\right] \\
\frac{\partial^2\left(\ln\left[P_1^L\left(Q,\ F\right)\right]\right)}{df_{kj}df_{k'j'}} &= \begin{cases}\sum_i^I\left[\frac{F(i,j)\cdot H_F\left(i,j,k,k'\right)-G_F(i,j,k)\cdot G_F\left(i,j,k'\right)}{F^2(i,j)}\right] & \text{if } j=j' \\ 0 & \text{if } j\neq j'\end{cases} \\
F\left(i,j\right) &= g_{ij}^{AA}\cdot A_{ij}^2+g_{ij}^{aa}\cdot B_{ij}^2+g_{ij}^{Aa}\cdot 2A_{ij}B_{ij} \\
G_F\left(i,j,k\right) &= \frac{\partial F\left(i,j\right)}{\partial f_{kj}} \\
&= 2g_{ij}^{AA}\cdot q_{ik}\cdot A_{ij}-2g_{ij}^{aa}\cdot q_{ik}\cdot B_{ij}+ \\
&\quad 2g_{ij}^{Aa}\cdot\left(B_{ij}\cdot q_{ik}-A_{ij}\cdot q_{ik}\right) \\
H_F\left(i,j,k,k'\right) &= \frac{\partial G\left(i,j,k\right)}{\partial f_{k'j}} \\
&= 2g_{ij}^{AA}\cdot q_{ik}\cdot q_{ik'}+2g_{ij}^{aa}\cdot q_{ik}\cdot q_{ik'}-4g_{ij}^{Aa}\cdot q_{ik}\cdot q_{ik'}.
\end{aligned}
$$

In the implementation, for each row of Q^i, calculating the derivatives requires the corresponding vector G^{AAi}, G^{Aai}, G^{aai}, A^i, B^i, and all of Fa and Fb.

$$
\begin{aligned}
\frac{\partial\left(\ln\left[P_1^L\left(Q,\ F\right)\right]\right)}{\partial Q^i} &= \sum_j^J\left(\frac{\beta_j^i\cdot Fa^j+\gamma_j^i Fb^j}{\alpha_j^i}\right)\\
\frac{\partial^2\left(\ln\left[P_1^L\left(Q,\ F\right)\right]\right)}{\partial Q^i\partial Q^i} &= \sum_j^J\left\{\frac{2G_j^{AAi}\left[Fa^j\left(Fa^j\right)^T\right]+2G_j^{aai}\left[Fb^j\left(Fb^j\right)^T\right]}{\alpha_j^i}\right\}\\
&\quad \sum_j^J\left\{\frac{2G_j^{Aai}\left[Fa^j\left(Fb^j\right)^T\right]+2G_j^{Aai}\left[Fb^j\left(Fa^j\right)^T\right]}{\alpha_j^i}\right\}\\
&\quad -\sum_j^J\left\{\frac{\left(\beta_j^i\right)^2\left[Fa^j\left(Fa^j\right)^T\right]+\left(\gamma_j^i\right)^2\left[Fb^j\left(Fb^j\right)^T\right]}{\left(\alpha_j^i\right)^2}\right\}\\
&\quad -\sum_j^J\left\{\frac{\beta_j^i\gamma_j^i\left[Fa^j\left(Fb^j\right)^T\right]+\beta_j^i\gamma_j^i\left[Fb^j\left(Fa^j\right)^T\right]}{\left(\alpha_j^i\right)^2}\right\}\\
\alpha_j^i &= G_j^{AAi}\left(A_j^i\right)^2+G_j^{aai}\left(B_j^i\right)^2+2G_j^{Aai}\cdot A_j^iB_j^i\\
\beta_j^i &= 2G_j^{AAi}A_j^i+2G_j^{Aai}B_j^i\\
\gamma_j^i &= 2G_j^{aai}B_j^i+2G_j^{Aai}A_j^i.
\end{aligned}
$$

For each column of F^j, calculating its derivatives requires the corresponding G^j, A^j, B^j, and all of Q.

$$
\begin{aligned}
\frac{\partial\left(\ln\left[P_1^L\left(Q,\ F\right)\right]\right)}{\partial F^j} &= \sum_i^I\left(\frac{\zeta_j^i}{\alpha_j^i}Q^i\right)\\
\frac{\partial^2\left(\ln\left[P_1^L\left(Q,\ F\right)\right]\right)}{\partial F^j\partial F^j} &= \sum_i^I\left\{\frac{\left(2G_j^{AAi}+2G_j^{aai}-4G_j^{Aai}\right)}{\alpha_j^i}\left[\left(Q^i\right)^TQ^i\right]\right\}\\
&\quad -\sum_i^I\left\{\frac{\left(\zeta_j^i\right)^2}{\left(\alpha_j^i\right)^2}\left[\left(Q^i\right)^TQ^i\right]\right\}\\
\alpha_j^i &= G_j^{AAi}\left(A_j^i\right)^2+G_j^{aai}\left(B_j^i\right)^2+2G_j^{Aai}\cdot A_j^iB_j^i\\
\zeta_j^i &= 2G_j^{AAi}A_j^i-2G_j^{aai}B_j^i+2G_j^{Aai}B_j^i-2G_j^{Aai}A_j^i.
\end{aligned}
$$

Block structure

The log likelihood function $L\left(Q,\ F\right)$ is concave in Q for a fixed F and concave in F for a fixed Q. In each iteration of updating Q and F, we perform two quadratic optimizations, one for Q and one for F.

$$H_Q = \begin{bmatrix} * & * & 0 & 0 & 0 & 0 \\ * & * & 0 & 0 & 0 & 0 \\ 0 & 0 & * & * & 0 & 0 \\ 0 & 0 & * & * & 0 & 0 \\ 0 & 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & 0 & * & * \end{bmatrix} \qquad H_F = \begin{bmatrix} * & * & 0 & 0 & 0 & 0 & 0 & 0 \\ * & * & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & * & * & 0 & 0 & 0 & 0 \\ 0 & 0 & * & * & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & * & * & 0 & 0 \\ 0 & 0 & 0 & 0 & * & * & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & * & * \end{bmatrix}$$

$$D_Q = \begin{bmatrix} * & * & * & * & * & * \end{bmatrix} \qquad D_F = \begin{bmatrix} * & * & * & * & * & * & * & * \end{bmatrix}$$

For example, if we have $I = 3$, $J = 4$, and $K = 2$, for updating Q, we have a Hessian matrix H_Q of size 6×6 and a derivative vector D_Q of size 6; for updating F, we have a Hessian matrix H_F of size 8×8 and a derivative vector D_F of size 8. Based on quadratic calculations, they would take the form of the above matrices where asterisks represent non-zero entries.

In theory, with this information we can perform quadratic optimization to find the ΔQ and ΔF. In practice, however, we have very large I and even larger J. We would have very large and sparse H_Q and H_F. We avoid solving and storing matrix inversions by solving the system of linear equations. To solve the system of linear equations, we first observe the Hessian matrices. It is clear that rather than solving the full linear system, the problem can be split into a collection of smaller problems consisting of solving systems with Hessian matrices of $K \times K$ instead of $IK \times IK$ and $JK \times JK$.

$$H_{Q_i} = \begin{bmatrix} * & * \\ * & * \end{bmatrix} \qquad D_{Q_i} = \begin{bmatrix} * & * \end{bmatrix} \qquad \forall i \in \{0,\ 1,\ ,\cdots,\ I-1\}$$

$$H_{F_j} = \begin{bmatrix} * & * \\ * & * \end{bmatrix} \qquad D_{F_j} = \begin{bmatrix} * & * \end{bmatrix} \qquad \forall j \in \{0,\ 1,\ ,\cdots,\ J-1\}$$

Active set algorithm

To solve these inequality- and equality-constraint quadratic optimization problems, we can apply the Active Set Algorithm [24]. A constraint is called active when its equality is satisfied and inactive when its strict inequality is satisfied [27]. An equality constraint is always active. A rough outline of this algorithm is described below.

To find a feasible starting point and initialize the corresponding active set, we could use linear programming. With our problem, however, we can take a shortcut. We have box constraints, i.e. $a_i \leq \Delta_i \leq b_i$, where $a_i < 0$ and $b_i > 0$, $\forall i$. We can always provide $[a_0,\ 0,\ \cdots,\ 0]$ as the starting point and initialize the active set to contain one constraint, $\Delta_0 \leq -a_0$.

To solve the equality problem defined by the active set and compute the Lagrange multipliers of the active set, we use the Karush-Kuhn-Tucker (KKT) approach [16, 18]. It is a nonlinear programming generalization of the Lagrange multiplier method. It allows only equality constraints. The active set

Algorithm 2 Quadratic programming with the active set algorithm

```
Find a feasible starting point
Initialize the corresponding active set
loop
  Solve the equality problem defined by the active set
  Compute the Lagrange multipliers of the active set
  if the solved approximation is within the feasible region then
    if all Lagrange multipliers are negative then
      return the solved approximation
    else
      Remove the constraint with the largest Lagrange multiplier
    end if
  else
    Take the shortest step back into the feasible region
    Insert the corresponding constraint into the active set
  end if
end loop
```

algorithm operates based on solving for equality quadratic subproblems. The general form of the KKT procedure can be summarized in the following equation, where H is the Hessian, A is the coefficients of constraints defined in the active set, x is the solution vector to be tested, $\mathcal{L}$ is the Lagrange multipliers, D is the derivatives, and b is the right hand side of the active constraints.

$$\begin{array}{ll} \max_{\Delta} & \left\{\frac{1}{2}x^T Hx + Dx\right\} \\ \text{s.t.} & Ax = b \\ \text{solve} & \begin{bmatrix} H & A^T \\ A & 0 \end{bmatrix} \cdot \begin{bmatrix} x \\ \mathcal{L} \end{bmatrix} = \begin{bmatrix} -D \\ b \end{bmatrix}. \end{array}$$

In each iteration of the main loop, the active set algorithm tries to find a better solution by walking along the active constraints. It deviates from the bounds when the Lagrange multipliers signal a better solution toward the feasible region. The maximum iterations of the main loop is the number of inequality constraints. In the worst case, the algorithm walks along each inequality constraint once. We have $2K + 1$ constraints for updating Q_i and $2K$ constraints for updating F_j. Without block relaxation, we have $2IK + I$ constraints for updating Q and $2JK$ constraints for updating F. The runtime complexity for each update step, therefore, improves from $\Theta\left(I^2K^2 \cdot (I + 2IK) + J^2K^2 \cdot 2JK\right)$ to $\Theta\left(IK^2 \cdot (1 + 2K) + JK^2 \cdot 2K\right)$ taking advantage of the block structure.

To update Q_i, we maximize a quadratic form subject to one equality constraint and $2K$ inequality constraints.

$$
\begin{array}{ll}
\max_{\Delta_{Q_i}} & \left\{ \frac{1}{2} \Delta_{Q_i}^T H_{Q_i} \Delta_{Q_i} + D_{Q_i}^T \Delta_{Q_i} \right\} \\
\text{s.t.} & A\Delta_{Q_i} \leq a \\
& B\Delta_{Q_i} = b \\
\text{where} & A = \begin{bmatrix} -1 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & -1 \\ 1 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} \quad a = \begin{bmatrix} q_{i0} \\ \vdots \\ q_{ik} \\ 1 - q_{i0} \\ \vdots \\ 1 - q_{ik} \end{bmatrix} \\
& B = \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix} \quad b = [0]
\end{array}
$$

To update F_j, we maximize a quadratic form subject to just $2K$ inequality constraints.

$$
\begin{array}{ll}
\max_{\Delta_{F_j}} & \left\{ \frac{1}{2} \Delta_{F_j}^T H_{F_j} \Delta_{F_j} + D_{F_j}^T \Delta_{F_j} \right\} \\
\text{s.t.} & A\Delta_{F_j} \leq a \\
\text{where} & A = \begin{bmatrix} -1 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & -1 \\ 1 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} \quad a = \begin{bmatrix} f_{0j} \\ \vdots \\ f_{kj} \\ 1 - f_{0j} \\ \vdots \\ 1 - f_{kj} \end{bmatrix}
\end{array}
$$

A concrete example Here we present a concrete example of the active set algorithm applied to solve a plain quadratic minimization problem. We define the quadratic form following the convention of the active set algorithm, i.e. a minimization problem with less-than inequality constraints.

$$
\begin{array}{ll}
\min\limits_{x} & \left\{ x^2 + y^2 - 8x - 6y \right\} \\
\text{s.t.} & -x \leq 0 \\
& -y \leq 0 \\
& x + y \leq 5
\end{array}
$$

We can easily derive the Hessian matrix and the derivative vector. We can then compute the D vector and express this problem in its quadratic form, $\min\limits_{\bar{x}} \left\{ \frac{1}{2} \bar{x}^T H \bar{x} + D\bar{x} \right\}$.

$$
\begin{aligned}
\frac{\partial\left(x^2+y^2-8x-6y\right)}{\partial x} &= 2x-8 \\
\frac{\partial\left(x^2+y^2-8x-6y\right)}{\partial y} &= 2y-6 \\
\frac{\partial^2\left(x^2+y^2-8x-6y\right)}{\partial x \partial x} &= 2 \\
\frac{\partial^2\left(x^2+y^2-8x-6y\right)}{\partial y \partial y} &= 2 \\
\frac{\partial^2\left(x^2+y^2-8x-6y\right)}{\partial x \partial y} &= \frac{\partial^2\left(x^2+y^2-8x-6y\right)}{\partial y \partial x} = 0 \\
\Rightarrow H &= \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \\
D &= \begin{bmatrix} 2x-8 \\ 2y-6 \end{bmatrix} - H \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -8 \\ -6 \end{bmatrix}
\end{aligned}
$$

We can apply a linear programming step to find a good initial starting point, but since the problem is simple, we can just use the origin as the starting point. This would render the first two constraints active, so our initial active set contains the first two constraints.

$$
\begin{aligned}
\min_{\bar{x}} \quad & \left\{ \frac{1}{2} \bar{x}^T \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \bar{x} + \begin{bmatrix} -8 \\ -6 \end{bmatrix} \bar{x} \right\} \\
\text{s.t.} \quad & \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \bar{x} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\
\text{solve} \quad & \begin{bmatrix} 2 & 0 & -1 & 0 \\ 0 & 2 & 0 & -1 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ \mathcal{L}_1 \\ \mathcal{L}_2 \end{bmatrix} = \begin{bmatrix} 8 \\ 6 \\ 0 \\ 0 \end{bmatrix} \\
\Rightarrow \begin{bmatrix} x \\ y \\ \mathcal{L}_1 \\ \mathcal{L}_2 \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \\ -8 \\ -6 \end{bmatrix}.
\end{aligned}
$$

The first round of solving KKT does not move anywhere. This is, of course, the case because the intersect of the first two constraints is unique, the origin. But we receive negative Lagrangian which indicate better solutions towards the feasible region. Following the algorithm we remove the constraints with the smallest Lagrangian.

$$
\begin{aligned}
\text{solve} \quad & \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & -1 & 0 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ \mathcal{L}_1 \end{bmatrix} = \begin{bmatrix} 8 \\ 6 \\ 0 \end{bmatrix} \\
\Rightarrow \begin{bmatrix} x \\ y \\ \mathcal{L}_1 \end{bmatrix} &= \begin{bmatrix} 4 \\ 0 \\ -6 \end{bmatrix}.
\end{aligned}
$$

The second round of solving KKT takes us from $(0, 0)$ to $(4, 0)$. We still have a negative Lagrangian, so we remove the only constraint currently active.

$$\begin{aligned} &\text{solve} && \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 8 \\ 6 \end{bmatrix} \\ \Rightarrow &\begin{bmatrix} x \\ y \end{bmatrix} &&= \begin{bmatrix} 4 \\ 3 \end{bmatrix}. \end{aligned}$$

The third round of solving KKT takes us from $(4, 0)$ to $(4, 3)$, which is outside of the feasible region because $4 + 3 > 5$. Following the algorithm, we backtrack towards the previous valid point, $(4, 0)$, and add the constraint that is crossed during this backtrack.

$$\begin{aligned} &\text{solve} && \begin{bmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ \mathcal{L}_1 \end{bmatrix} = \begin{bmatrix} 8 \\ 6 \\ 5 \end{bmatrix} \\ \Rightarrow &\begin{bmatrix} x \\ y \\ \mathcal{L}_1 \end{bmatrix} &&= \begin{bmatrix} 3 \\ 2 \\ 2 \end{bmatrix}. \end{aligned}$$

The forth round of solving KKT takes us from $(4, 3)$ to $(3, 2)$, and the only Lagrangian is now positive. We therefore return $(3, 2)$ as the final solution.

Complementarity pivoting algorithm

To solve these inequality- and equality-constraint quadratic optimization problems, we can also use an adaptation of the complementarity pivoting algorithm designed for linear complementarity problems [24]. The complementarity pivoting algorithm is similar to the simplex algorithm used for linearly programming problems. The first goal is to convert our problem to a quadratic optimization problem in the following form.

$$\begin{aligned} \min_{x} \quad & \left\{ \frac{1}{2} x^T Q x + C^T x \right\} \\ \text{s.t.} \quad & Ax \geq b \\ & x \geq 0 \end{aligned}$$

Then we can map it to a linear complementarity problem $w = M \cdot z + q$ where $w_i \cdot z_i = 0, \ \forall i$, and we can implement Lemke's complementarity pivot algorithm to solve for z and hence for x.

$$\begin{aligned} M &= \begin{bmatrix} Q & -A^T \\ A & 0 \end{bmatrix} \\ q &= \begin{bmatrix} c \\ -b \end{bmatrix} \\ z &= \begin{bmatrix} x \\ \lambda \end{bmatrix} \end{aligned}$$

In our problem, we use the SQP approach by first approximating the objective functions, both the genotype observation version and the genotype

likelihood version, with their second order Taylor expansions. The quantity we optimize is Δx_Q or Δx_F for updating Q or F, respectively. These quantities record the changes in values for Q or F, so they need to satisfy a sequence of constraints during the search for Δx, $\forall \Delta q_{ik},\ q_{ik} + \Delta q_{ik} \in [0,\ 1]$, $\forall \Delta f_{kj},\ f_{kj} + \Delta f_{kj} \in [0,\ 1]$, and $\forall \Delta q_{ik},\ \sum_k^K \Delta q_{ik} = 0$ because $\sum_k^K q_{ik} = 1$.

As we can see, the conditions do not quite satisfy the requirements from Lemke's algorithm. Specifically, we have equality constraints for updating Q, and the values to solve can be negative. For equality constraints, we replace them with pairs of inequality constraints. For negativity, we perform a re-parameterizing based on the knowledge of our bounded parameter space. We let $\bar{\Delta} q = \Delta q + v$ and $\bar{\Delta} f = \Delta f + v$ for all values in Q and F. We proceed to solve for $\bar{\Delta} q$ and $\bar{\Delta} f$, and we then recover the original values with $\Delta q = \bar{\Delta} q - v$ and $\Delta f = \bar{\Delta} f - v$.

$$
\begin{aligned}
F_T\left(x_n + \bar{\Delta} x - v\right) &= F\left(x_n\right) + F'\left(x_n\right)\left(\bar{\Delta} x - v\right) + \frac{1}{2}\left(\bar{\Delta} x - v\right)^T F''\left(x_n\right)\left(\bar{\Delta} x - v\right) \\
&= F\left(x_n\right) - v \cdot F'\left(x_n\right) + F'\left(x_n\right)\bar{\Delta} x + \\
&\quad \frac{1}{2}\left[\left(\bar{\Delta} x\right)^T F''\left(x_n\right)\bar{\Delta} x - v\left(\bar{\Delta} x\right)^T F''\left(x_n\right) - vF''\left(x_n\right)\bar{\Delta} x + v^2 F''\left(x_n\right)\right] \\
&= F\left(x_n\right) - v \cdot F'\left(x_n\right) + \frac{1}{2}v^2 F''\left(x_n\right) + \\
&\quad \left[F'\left(x_n\right) - vF''\left(x_n\right)\right]\bar{\Delta} x + \frac{1}{2}\left(\bar{\Delta} x\right)^T F''\left(x_n\right)\bar{\Delta} x \\
\Rightarrow Q_{\bar{\Delta} x} &= Q_{\Delta x} \\
C_{\bar{\Delta} x} &= C_{\Delta x} - v \cdot Q_{\Delta x}
\end{aligned}
$$

If we select a $v \geq 1$, we would satisfy $\bar{\Delta} x \geq 0$ as required by the basic form for Lemke's algorithm. And we would insert equality constraints as two inequality constraints. Finally, we have a maximization problem rather than minimization. It clear that if we phrase our problem as the following, we simply need to negative all quantities, Q, C, A, and b, before Lemke's operations.

$$
\begin{aligned}
\max_x \quad & \left\{\frac{1}{2}x^T Q x + C^T x\right\} \\
\text{s.t.} \quad & Ax \leq b \\
& x \leq 0
\end{aligned}
$$

To update Q_i, we maximize a quadratic form subject to 1 equality constraint and $2K$ inequality constraints.

$$
\begin{array}{ll}
\max_{\Delta_{\bar{Q}_i}} & \left\{ \frac{1}{2} \Delta_{\bar{Q}_i}^T H_{\bar{Q}_i} \Delta_{\bar{Q}_i} + D_{\bar{Q}_i}^T \Delta_{\bar{Q}_i} \right\} \\
\text{s.t.} & A\Delta_{\bar{Q}_i} \leq a \\
& B\Delta_{\bar{Q}_i} = b
\end{array}
$$

$$
\text{where} \quad A' = \begin{bmatrix} -1 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & -1 \\ 1 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & 1 \\ 1 & \cdots & 1 \\ -1 & \cdots & -1 \end{bmatrix} \quad a' = \begin{bmatrix} q_{i0} - v \\ \vdots \\ q_{ik} - v \\ 1 + v - q_{i0} \\ \vdots \\ 1 + v - q_{ik} \\ k \cdot v \\ -k \cdot v \end{bmatrix}
$$

To update F_j, we maximize a quadratic form subject to just $2K$ inequality constraints.

$$
\begin{array}{ll}
\max_{\Delta_{\bar{F}_j}} & \left\{ \frac{1}{2} \Delta_{\bar{F}_j}^T H_{\bar{F}_j} \Delta_{\bar{F}_j} + D_{\bar{F}_j}^T \Delta_{\bar{F}_j} \right\} \\
\text{s.t.} & A\Delta_{\bar{F}_j} \leq a
\end{array}
$$

$$
\text{where} \quad A = \begin{bmatrix} -1 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & -1 \\ 1 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} \quad a = \begin{bmatrix} f_{0j} - v \\ \vdots \\ f_{kj} - v \\ 1 + v - f_{0j} \\ \vdots \\ 1 + v - f_{kj} \end{bmatrix}
$$

A concrete example Here we present a concrete example of the complementarity pivoting algorithm used to solve a plain quadratic minimization problem. We define the quadratic form following the convention of Lemke's algorithm, i.e. a minimization problem with greater-than inequality constraints.

$$
\begin{array}{ll}
\min\limits_{x} & \left\{ x^2 + y^2 - 8x - 6y \right\} \\
\text{s.t.} & x \geq 0 \\
& y \geq 0 \\
& -x - y \geq -5
\end{array}
$$

Like in the example for the active set algorithm, we first derive the Hessian matrix and the derivative vector. We then compute the C vector and express this problem in its linear complementarity form.

$$
\begin{array}{ll}
& \min\limits_{\bar{x}} \left\{ \frac{1}{2} \bar{x}^T H \bar{x} + D\bar{x} \right\} \\
& H = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \; C = \begin{bmatrix} -8 \\ -6 \end{bmatrix} \\
\Rightarrow \quad & M = \begin{bmatrix} 2 & 0 & 1 & 0 & -1 \\ 0 & 2 & 0 & 1 & -1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 & 0 \end{bmatrix}, \; q = \begin{bmatrix} -8 \\ -6 \\ 0 \\ 0 \\ 5 \end{bmatrix}
\end{array}
$$

We pad the column of z_0 with values -1 and initiate the pivoting at column z_0. We first select location (w_1, z_0) as the pivot because -8 is the most negative q. We perform Gaussian-Jordan elimination on pivot (w_1, z_0) to obtain the second tabula. We dropped w_1 from the base vector (BV), so we pivot on column z_1. Because 1 is the smallest ratio we pivot on (w_2, z_1). We continue this process until the present basis is a complementary feasible basis, and then the algorithm terminates.

BV	w1	w2	w3	w4	w5	z1	z2	z3	z4	z5	z0	q
w1	1.000	0.000	0.000	0.000	0.000	-2.000	-0.000	1.000	0.000	-1.000	-1.000	-8.000
w2	0.000	1.000	0.000	0.000	0.000	-0.000	-2.000	0.000	1.000	-1.000	-1.000	-6.000
w3	0.000	0.000	1.000	0.000	0.000	-1.000	-0.000	-0.000	-0.000	-0.000	-1.000	-0.000
w4	0.000	0.000	0.000	1.000	0.000	-0.000	-1.000	-0.000	-0.000	-0.000	-1.000	-0.000
w5	0.000	0.000	0.000	0.000	1.000	1.000	1.000	-0.000	-0.000	-0.000	-1.000	5.000

BV	w1	w2	w3	w4	w5	z1	z2	z3	z4	z5	z0	q
z0	-1.000	-0.000	-0.000	-0.000	-0.000	2.000	0.000	-1.000	-0.000	1.000	1.000	8.000
w2	-1.000	1.000	0.000	0.000	0.000	2.000	-2.000	-1.000	1.000	0.000	0.000	2.000
w3	-1.000	0.000	1.000	0.000	0.000	1.000	0.000	-1.000	-0.000	1.000	0.000	8.000
w4	-1.000	0.000	0.000	1.000	0.000	2.000	-1.000	-1.000	-0.000	1.000	0.000	8.000
w5	-1.000	0.000	0.000	0.000	1.000	3.000	1.000	-1.000	-0.000	1.000	0.000	13.000

BV	w1	w2	w3	w4	w5	z1	z2	z3	z4	z5	z0	q
z0	0.000	-1.000	-0.000	-0.000	-0.000	0.000	2.000	0.000	-1.000	1.000	1.000	6.000
z1	-0.500	0.500	0.000	0.000	0.000	1.000	-1.000	-0.500	0.500	0.000	0.000	1.000
w3	-0.500	-0.500	1.000	0.000	0.000	0.000	1.000	-0.500	-0.500	1.000	0.000	7.000
w4	0.000	-1.000	0.000	1.000	0.000	0.000	1.000	0.000	-1.000	1.000	0.000	6.000
w5	0.500	-1.500	0.000	0.000	1.000	0.000	4.000	0.500	-1.500	1.000	0.000	10.000

BV	w1	w2	w3	w4	w5	z1	z2	z3	z4	z5	z0	q
z0	-0.250	-0.250	-0.000	-0.000	-0.500	0.000	0.000	-0.250	-0.250	0.500	1.000	1.000
z1	-0.375	0.125	0.000	0.000	0.250	1.000	0.000	-0.375	0.125	0.250	0.000	3.500
w3	-0.625	-0.125	1.000	0.000	-0.250	0.000	0.000	-0.625	-0.125	0.750	0.000	4.500
w4	-0.125	-0.625	0.000	1.000	-0.250	0.000	0.000	-0.125	-0.625	0.750	0.000	3.500
z2	0.125	-0.375	0.000	0.000	0.250	0.000	1.000	0.125	-0.375	0.250	0.000	2.500

BV	w1	w2	w3	w4	w5	z1	z2	z3	z4	z5	z0	q
z5	-0.500	-0.500	-0.000	-0.000	-1.000	0.000	0.000	-0.500	-0.500	1.000	2.000	2.000
z1	-0.250	0.250	0.000	0.000	0.500	1.000	0.000	-0.250	0.250	0.000	-0.500	3.000
w3	-0.250	0.250	1.000	0.000	0.500	0.000	0.000	-0.250	0.250	0.000	-1.500	3.000
w4	0.250	-0.250	0.000	1.000	0.500	0.000	0.000	0.250	-0.250	0.000	-1.500	2.000
z2	0.250	-0.250	0.000	0.000	0.500	0.000	1.000	0.250	-0.250	0.000	-0.500	2.000

We then read the solution from the last tabular, $z_1 = x = 3$ and $z_2 = y = 2$. The rest of the z values are Lagrangian. If the process terminates due to a pivoting column with entirely non-positive values, then the pivoting algorithm fails to find the optima. This does not happen in our case, however, because our problem is well-shaped, while the complementarity pivoting algorithm attempts to solve linear complementarity problems in their general form, where M is not restricted to symmetric and positive semi-definite matrices.

Comparison between the Active set algorithm and the Complementarity pivoting algorithm

Ohana's **qpas** program implements the active set algorithm, and Ohana's **cpax** program implements the complementarity pivoting algorithm. We tailor both methods to solve the SQP problem defined by the classical structure model. Both methods take advantage of the block structure of Hessians, where

K	cpax lle	qpas lle	diff (cpax-qpas)
2	-1967734	-1967733	-1
3	-1956799	-1956785	-14
4	-1946379	-1946218	-161
5	-1935939	-1935775	-164
6	-1925641	-1925636	-5
7	-1915558	-1915552	-6
8	-1905394	-1905430	36
9	-1895298	-1895372	74
10	-1885465	-1885306	-160
11	-1875303	-1875503	200
12	-1865490	-1865492	2
13	-1855715	-1855502	-213
14	-1846133	-1845732	-401
15	-1836169	-1836315	147

Table 3.1: Highest log likelihoods achieved from Ohana's **cpax** and **qpas** programs over a range of K values. For each program, and each K, we execute 100 times using random seeds $0, 1, ..., 99$. Two thirds of the time, Ohana's **qpas** reports higher likelihoods. This dataset contains 118 Europeans of 17,507.

most of the off-diagonal values diminish, so both methods deal with sequences of small matrix operations rather than large matrix operations. Both methods perform calculations at the highest possible level, vector or matrix, rather than iterating each matrix element sequentially.

At this stage of Ohana's development, we recognize **qpas** as the better solver over capx. From our benchmark tests, **qpas** generally reached better likelihood values. Table 3.1 shows such an experiment. The time durations required for both programs to reach their plateau likelihoods are similar.

NM for population covariances

After obtaining population stratification from the structure model, the natural next question to ask is regarding the evolution history of these ancestry components. We achieve this goal by modeling the allele frequencies produced from the structure analysis using Gaussian approximation. Gaussian modeling corresponds to a Brownian motion approximation of genetic drift rather than, say, Wright-Fisher diffusion. We show in the simulations that Gaussian modeling is reasonably robust, but it underestimates branch lengths for long divergence scenarios. This limitation exits in all methods and tools that apply Gaussian approximation, such as TreeMix [28].

Eq 3.1 captures this likelihood model. We can evaluate a given valid covariance matrix into a likelihood value. We can, therefore, treat this as a black-box optimization problem and consider gradient-less optimization methods. In this category, we implemented and experimented with two. The first

is Particle Swarm Optimization (PSO). The second is Nelder-Mead (NM) optimization [25]. Between these two, we prefer the deterministic NM over the heuristic based PSO. NM has behaved well in simulations.

We use sample covariances as the initial starting point for the Nelder-Mead optimizer. The general form for computing sample covariances for a multivariate distribution is $S_c = \frac{1}{n} \cdot \sum_i^n (x_i - \bar{x}_i)(x_i - \bar{x}_i)^T$. We do not have a single mean value per population. Each site has its own mean value, and this value is the same across populations at that site.

$$S_c = \frac{1}{J} \cdot \sum_j^J (x_j - \bar{x}_j)(x_j - \bar{x}_j)^T$$

$$x_j = \begin{bmatrix} f_0 \\ \vdots \\ f_{K-1} \end{bmatrix}_j \quad \bar{x}_j = \begin{bmatrix} f_A \\ \vdots \\ f_A \end{bmatrix}_j \quad x_j - \bar{x}_j = \begin{bmatrix} f'_0 \\ \vdots \\ f'_{K-1} \end{bmatrix}_j .$$

3.4 Phylogenetic trees estimation

Ohana not only produces the covariance relationships among ancestry components, but it also estimates the most compatible phylogenetic tree structure to depict the evolutionary history. Covariance matrices and their most compatible trees have a one-to-one mapping. Of course, the evolutionary history of the ancestry components may not be a tree-like demography. In this case, we show in the simulations that the tree inference process produces the closest approximation.

To construct phylogenetic trees, we first transform covariance matrices to distance matrices. The distance between two populations is the sum of the variances of these two populations, less two multiplied by the covariance between them. We then use the neighbor-joining method to construct phylogenetic trees from distance matrices. Finally, we visualize phylogenetic trees in scalable graphics. This pipeline is illustrated below.

3.5 Selection study

Following the structure analysis and population tree inference, a natural extension of this framework is to detect SNPs that, according to the likelihood models, prefer to deviate from the globally estimated covariance structure. In other words, we detect SNPs that exhibit different speeds of genetic drift. This process is not a fixed procedure. Instead, each execution should follow a certain hypothesis and focus the selection strength to signals that reflect features such as local adaptation, population bottleneck, ancient versus modern data, etc.

Specifically, we scan for covariance outliers by applying a likelihood model to each locus, similar to the one used genome-wide but with certain scalar

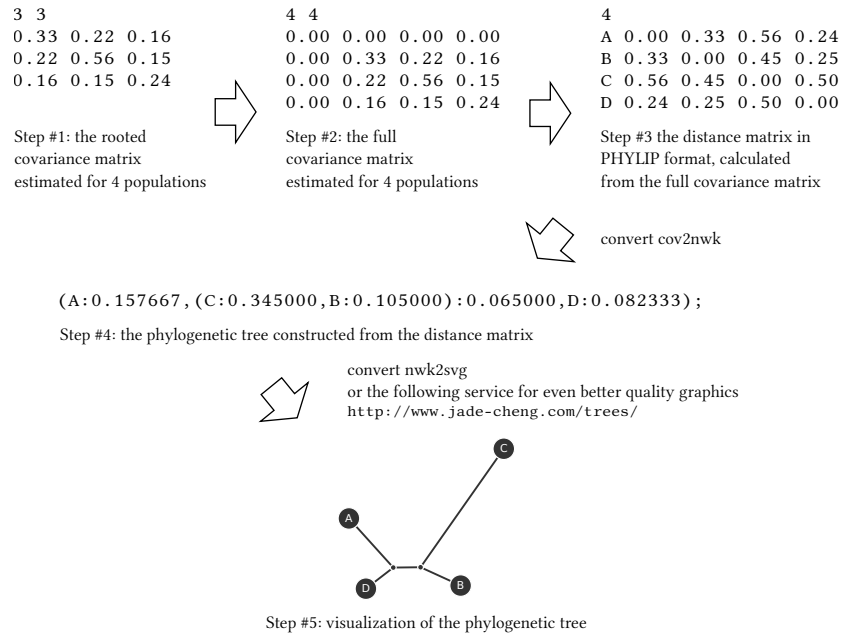


Figure 3.4: Pipline for phylogenetic tree construction. We start with the estimated variances and covariances. We compute the distance matrix and approximate the distance matrix into a tree structure. Finally, we visualize the tree using scalar graphics.

factor variations. This creates a nested likelihood model. Through a likelihood ratio test, it identifies loci in which the variance among populations is larger than expected from the genome-wide estimated covariance matrix. This method is inspired by a number of similar, recently-developed methods that use a Gaussian distribution as an approximation to model the distribution of allele frequencies among populations [4, 28].

To work with local adaptation hypotheses, we construct nested likelihood models with the scalar multiplier applied to only portions of the covariance structure rather than the entire covariance matrix.

We supply the selection algorithm with two covariance structures, one that is globally estimated and another that reflects the research hypothesis. The scanning process linearly interpolates between these two covariance structures and records the best intermediate state that produces the optimal local likelihood value for each marker location. Figure 3.5 demonstrates two examples where the selection analysis is localized at different portions of the covariance structure.

To summarize, we establish a framework that starts with population admixture analysis through unsupervised learning, estimates global covariance relations among ancestry components, and performs selection scanning with

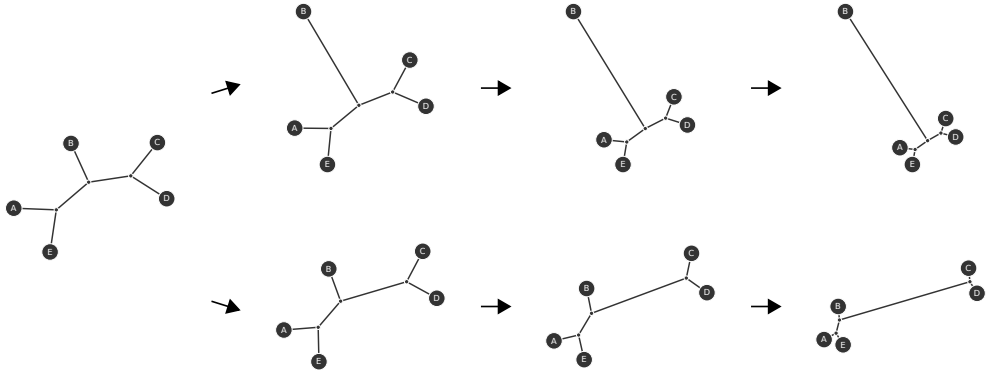


Figure 3.5: Selection analysis localized at different branches of the covariance tree. Top: selection analysis localized at the branch leading to component B. Bottom: selection analysis localized at the branch leading to component C and D. The left most covariance tree captures the global estimated covariance structure. The right most covariance trees capture two different local evolution hypotheses.

the capability of incorporating any given local evolutionary hypotheses. Algorithm 3.5 presents the high-level pseudo-code for this test in the case where we vary the covariance matrix by multiplying to it a scalar factor, $\alpha \times \Omega'$.

Algorithm 3 Selection scan to test for covariance outliers

```
Obtain the full genotype dataset G with N markers and M samples
Sample N' markers with respect to LD to form G'
Perform admixture analysis on G' of size M by N'
Produce admixture proportions Q' of size M by K
Produce allele frequencies F' of size K by N'
QPAS over ln(P1) using G while fixing Q'
to produce F of size K by N
for each marker in F do
  l_ratio ← 0
  for each α in a range of equal intervals starting from 1 do
    l_new ← ln(P2) calculated using α × Ω'
    l_old ← ln(P2) calculated using Ω'
    if 2 × (l_new − l_old) > l_ratio then
      l_ratio ← 2 × (l_new − l_old)
    end if
  end for
  emit l_ratio
end for
```

3.6 Joint inference for structure and covariances

The principle of joint estimation can be expanded to perform structure analysis with additional information or restraints. Ohana's sequential quadratic programming framework allows additional likelihood modules to be inferred jointly as long as the appendices are second-order differentiable and hold the same block structure, where most of the off-diagonal values in the Hessian diminish. Although this module is not currently released with Ohana, the idea of structure analysis with additional constraints is important and worth further exploration.

Here we show the mathematics for weighing the structure analysis using the inferred covariances. Over an iterative process, two parties of quantities should produce the best joint likelihood values.

Algorithm 4 Selection scan to test for covariance outliers

Estimate Q and F without weights from Ω
Estimate Ω using Q and F weighted by Ω
repeat
  Estimate Q' and F' weighted by Ω starting from Q and F
  $Q \leftarrow Q'$ and $F \leftarrow F'$
  Estimate Ω' using Q and F starting from Ω weighted by Ω
  $\Omega \leftarrow \Omega'$
until the joint likelihood value plateaus

In the joint inference process, we first estimate the initial pair of the Q and F starting with a random Q and F. We then start Nelder-Mead with the sample covariance matrix to estimate the initial Ω'. Then we estimate Q and F weighted by Ω'. The updated Q and F are in turn used to update Ω'. This process repeats until the joint likelihood value stops improving.

We derive the first differentials of $\ln\left[P_2\left(F\right)\right]$. Note we have $\left(\alpha \cdot A\right)^{-1} = \frac{1}{\alpha} \cdot A^{-1}$, $\det\left(\alpha \cdot A\right) = a^r \cdot \det\left(A\right)$, and $A = A^T \Rightarrow A^{-1} = \left(A^{-1}\right)^T$, where A is a square matrix of $r \times r$ and α is a scalar.

$$
\begin{aligned}
\frac{\partial\left(\ln\left[P_2\left(F\right)\right]\right)}{\partial f_{kj}} &= \frac{\partial\left(-\frac{1}{2}\cdot\sum_j^J\left\{\left(K-1\right)\cdot\ln\left(2\pi c_j\right)+\ln\left[\det\left(\Omega'\right)\right]+\frac{1}{c_j}\cdot f_j'^T\cdot\Omega'^{-1}\cdot f_j'\right\}\right)}{\partial f_{kj}} \\
&= \frac{\partial\left(-\frac{1}{2c_j}\cdot\sum_j^J\left(f_j'^T\cdot\Omega'^{-1}\cdot f_j'\right)\right)}{\partial f_{kj}} \\
&= -\frac{1}{2c_j}\cdot\frac{\partial\left(\sum_n^{K-1}\sum_m^{K-1}\left(f'_{mj}\cdot\Omega'^{-1}_{mn}\cdot f'_{nj}\right)\right)}{\partial f_{kj}} \\
&= -\frac{1}{2c_j}\cdot\frac{\partial\left(\sum_n^{K-1}\sum_m^{K-1}\left[\left(f_{(m+1)j}-f_{0j}\right)\cdot\Omega'^{-1}_{mn}\cdot\left(f_{(n+1)j}-f_{0j}\right)\right]\right)}{\partial f_{kj}}
\end{aligned}
$$

$$
\begin{aligned}
\text{when } k \neq 0: \quad &= -\frac{1}{2c_j} \cdot \sum_{n}^{K-1} \left[2 \cdot \Omega'^{-1}_{(k-1)n} \cdot \left(f_{(n+1)j} - f_{0j} \right) \right] \\
&= -\frac{1}{c_j} \cdot \sum_{n}^{K-1} \left[\Omega'^{-1}_{(k-1)n} \cdot \left(f_{(n+1)j} - f_{0j} \right) \right] \\
&= -\frac{1}{c_j} \cdot \sum_{n}^{K-1} \left(\Omega'^{-1}_{(k-1)n} \cdot f'_{nj} \right) \\
\text{when } k = 0: \quad &= -\frac{1}{2c_j} \cdot \sum_{n}^{K-1} \sum_{m}^{K-1} \left[-\Omega'^{-1}_{mn} \cdot \left(f_{(n+1)j} - f_{0j} + f_{(m+1)j} - f_{0j} \right) \right] \\
&= \frac{1}{2c_j} \cdot \sum_{n}^{K-1} \sum_{m}^{K-1} \left[\Omega'^{-1}_{mn} \cdot \left(f_{(n+1)j} + f_{(m+1)j} - 2f_{0j} \right) \right] \\
&= \frac{1}{2c_j} \cdot \sum_{n}^{K-1} \sum_{m}^{K-1} \left[\Omega'^{-1}_{mn} \cdot \left(f'_{mj} + f'_{nj} \right) \right] \\
\text{where } c_j \quad &= \mu_j \left(1 - \mu_j \right)
\end{aligned}
$$

We derive the second differentials of $\ln\left[P_2\left(F\right)\right]$.

$$
\begin{aligned}
\frac{\partial^2 \left(\ln\left[P_2\left(F\right)\right] \right)}{\partial f_{kj} \partial f_{k'j'}} \quad &= \quad \frac{\partial^2 \left(-\frac{1}{2} \cdot \sum_j^J \left\{ \left(K-1\right) \cdot \ln\left(2\pi c_j\right) + \ln\left[\det\left(\Omega'\right)\right] + \frac{1}{c_j} \cdot f_j'^T \cdot \Omega'^{-1} \cdot f'_j \right\} \right)}{\partial f_{kj} \partial f_{k'j}} \\
\text{when } k \neq 0 \text{ and } k' \neq 0: \quad &= \quad \frac{\partial \left(-\frac{1}{c_j} \cdot \sum_n^{K-1} \left[\Omega'^{-1}_{(k-1)n} \cdot \left(f_{(n+1)j} - f_{0j} \right) \right] \right)}{\partial f_{k'j}} \\
&= \quad -\frac{1}{c_j} \cdot \Omega'^{-1}_{(k-1)(k'-1)} \\
\text{when } k \neq 0 \text{ and } k' = 0: \quad &= \quad \frac{\partial \left(-\frac{1}{c_j} \cdot \sum_n^{K-1} \left[\Omega'^{-1}_{(k-1)n} \cdot \left(f_{(n+1)j} - f_{0j} \right) \right] \right)}{\partial f_{0j}} \\
&= \quad \frac{1}{c_j} \cdot \sum_{n}^{K-1} \left(\Omega'^{-1}_{(k-1)n} \right)
\end{aligned}
$$

$$
\begin{aligned}
\text{when } k = 0 \text{ and } k' \neq 0: \quad &= \quad \frac{\partial \left(\frac{1}{2c_j} \cdot \sum_n^{K-1} \sum_m^{K-1} \left[\Omega'^{-1}_{mn} \cdot \left(f_{(n+1)j} + f_{(m+1)j} - 2f_{0j} \right) \right] \right)}{\partial f_{k'j}} \\
&= \quad \frac{1}{c_j} \cdot \sum_{n}^{K-1} \left(\Omega'^{-1}_{(k'-1)n} \right) \\
\text{when } k = 0 \text{ and } k' = 0: \quad &= \quad \frac{\partial \left(\frac{1}{2c_j} \cdot \sum_n^{K-1} \sum_m^{K-1} \left[\Omega'^{-1}_{mn} \cdot \left(f_{(n+1)j} + f_{(m+1)j} - 2f_{0j} \right) \right] \right)}{\partial f_{0j}} \\
&= \quad -\frac{1}{c_j} \cdot \sum_{n}^{K-1} \sum_{m}^{K-1} \left(\Omega'^{-1}_{mn} \right) \\
\text{where } c_j \quad &= \quad \mu_j \left(1 - \mu_j \right).
\end{aligned}
$$

With the full likelihood function to update Q and F, we use the sum of the two parts. Since the second part of the likelihood function does not concern Q, the functional forms for Q's derivatives stay the same.

$$\begin{aligned}
\frac{\partial\left(\ln\left[P_1\left(Q,\ F\right)\right]+\ln\left[P_2\left(F\right)\right]\right)}{\partial q_{ik}} &= \frac{\partial\left(\ln\left[P_1\left(Q,\ F\right)\right]\right)}{\partial q_{ik}} \\
\frac{\partial^2\left(\ln\left[P_1\left(Q,\ F\right)\right]+\ln\left[P_2\left(F\right)\right]\right)}{\partial q_{ik}\partial q_{i'k'}} &= \frac{\partial^2\left(\ln\left[P_1\left(Q,\ F\right)\right]\right)}{\partial q_{ik}\partial q_{i'k'}}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial\left(\ln\left[P_1\left(Q,\ F\right)\right]+\ln\left[P_2\left(F\right)\right]\right)}{\partial f_{kj}} &= \frac{\partial\left(\ln\left[P_1\left(Q,\ F\right)\right]\right)}{\partial f_{kj}}+\frac{\partial\left(\ln\left[P_2\left(F\right)\right]\right)}{\partial f_{kj}} \\
\frac{\partial^2\left(\ln\left[P_1\left(Q,\ F\right)\right]+\ln\left[P_2\left(F\right)\right]\right)}{\partial f_{kj}\partial f_{k'j'}} &= \frac{\partial^2\left(\ln\left[P_1\left(Q,\ F\right)\right]\right)}{\partial f_{kj}\partial f_{k'j'}}+\frac{\partial^2\left(\ln\left[P_2\left(F\right)\right]\right)}{\partial f_{kj}\partial f_{k'j'}}.
\end{aligned}$$

3.7 Simulation studies

Data simulation

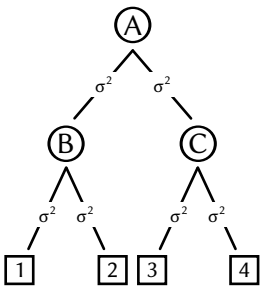
To evaluate the inference framework implemented in Ohana, we perform simulation studies. We apply two simulations schema, either simulating allele frequencies directly or simulating populations of nucleotide sequences according to a given demographic scenario. In both schema, we simulate admixture proportions directly, and we form genotype observations by calculating the probability of observing certain genotypes given the allele frequencies and the admixture proportions.

Simulate F directly

To directly simulate allele frequencies, we sample from given distributions with respect to a demography. We use Figure 3.6 as an example. A is the ancestral population. B and C are two intermediate ancestral populations. Populations 1, 2, 3, and 4 are the populations to be inferred using $K = 4$.

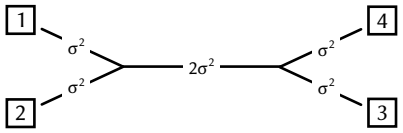
The allele frequency matrix F has size $K \times J$, where K is the number of populations and J is the number of SNPs. We can simulate a SNP by sampling f_A from a beta distribution. We can then sample f_B and f_C from a normal distribution with a mean of f_A. We can further sample f_1 and f_2 from a normal distribution with a mean of f_B, and we can sample f_3 and f_4 from a normal distribution with a mean of f_C. We set the allele frequency to the simulated value, clamped between zero and one.

$$\begin{aligned}
f_A &\sim \mathcal{B}\left(5,\ 5\right) \\
f_B,\ f_C &\sim \mathcal{N}\left(f_A,\ f_A\cdot\left(1-f_A\right)\cdot\sigma^2\right) \\
f_1,\ f_2 &\sim \mathcal{N}\left(f_B,\ f_B\cdot\left(1-f_B\right)\cdot\sigma^2\right) \\
f_3,\ f_4 &\sim \mathcal{N}\left(f_C,\ f_C\cdot\left(1-f_C\right)\cdot\sigma^2\right).
\end{aligned}$$



$$\Omega = \begin{bmatrix} & 1 & 2 & 3 & 4 \\ 1 & 2\sigma^2 & \sigma^2 & 0 & 0 \\ 2 & \sigma^2 & 2\sigma^2 & 0 & 0 \\ 3 & 0 & 0 & 2\sigma^2 & \sigma^2 \\ 4 & 0 & 0 & \sigma^2 & 2\sigma^2 \end{bmatrix}$$

Figure 3.6: Example direct allele frequency simulation. Left: population covariances to be simulated. Right: covariance matrix reflecting the tree structure on the left.



$$\Omega' = \begin{bmatrix} & 2 & 3 & 4 \\ 2 & 2\sigma^2 & \sigma^2 & \sigma^2 \\ 3 & \sigma^2 & 4\sigma^2 & 3\sigma^2 \\ 4 & \sigma^2 & 3\sigma^2 & 4\sigma^2 \end{bmatrix}$$

Figure 3.7: Example direct allele frequency simulation, rooted at population #1. Left: population covariances to be simulated. Right: covariance matrix reflecting the tree structure on the left.

Simulate F from sequences simulated from a demographic

In this schema, we perform coalescence simulation to obtain sequences under a given demography. From samples produced for each population, we calculate this population's allele frequency for each marker. We apply filtration to the simulated sequences as we do for genomic data collected from biological samples. We keep polymorphic sites with biallelic genotypes. We remove sites for which minor allele frequencies are less than 5% for any of the populations. We use the software `fastsimcoal2` for sequential Markov coalescent simulation of genomic data.

Simulate Q directly

We simulated admixture proportions directly. The admixture proportion matrix Q has size $I \times K$ where I is the number of individuals. We simulate for both un-admixed and admixed scenarios. For the un-admixed case, we simply assign portions of the samples to different populations. For the admixed case, we simulate Q_i independently from various symmetric Dirichlet distributions, $\mathrm{Dir}\,(\alpha,\ \alpha,\ \alpha)$ [1, 29]. The parameter α reflects the degree of admixture. When $\alpha < 1$, most individuals show little admixture, and when $\alpha > 1$, the opposite occurs.

Simulate genotype observations

We sample genotype for each individual at each marker using the simulated F and Q. The genotype matrix G has size $I \times J$. For each SNP location, the probabilities p^0, p^1, and p^2 for getting 0, 1, and 2 are calculated using the Q and F matrices described above. We first calculate the major allele frequency f_{ij} for each individual at each SNP location. We then calculate the probability of getting each genotype under Hardy-Weinberg Equilibrium.

$$\begin{aligned} f_{ij} &= \sum_{k}^{K} Q_{ik} \cdot F_{kj} \\ p_{ij}^0 &= f_{ij}^2 \\ p_{ij}^1 &= 2 \cdot f_{ij} \cdot (1 - f_{ij}) \\ p_{ij}^2 &= (1 - f_{ij})^2 . \end{aligned}$$

Effect of different divergence times

We designed a simulation study to investigate the effect of different divergence times. In the simulation study shown in Figure 3.8, we have three divergence scenarios: short, medium, and long. For each scenario, we simulated 120 individuals belonging to four groups, each with 30 individuals. The first three groups were un-admixed. The fourth group was a mixture of the first three under $\mathrm{Dir}\,(1.0,\ 1.0,\ 1.0)$. We simulated allele frequencies directly from distributions. We simulated 10,000 markers for each scenario. We simulated allele frequencies based on the covariance trees shown on the left in Figure 3.8. We estimated covariance trees for each scenario, shown on the right.

In the short divergence scenario, the simulation poorly recovered the admixture proportions and covariance tree. A small difference in allele frequencies across populations hinders the accurate estimation of allele frequencies and leads to a poor estimation for the covariance tree. In the joint inference process, a small difference would also cause a poor estimation of the admixture. This is a limitation for any inference system using similar statistical models such as STRUCTURE [29], FRAPPE [35], ADMIXTURE [1], and SPA [36].

In the long divergence scenario, the simulation poorly estimated the covariance tree but nicely recovered the admixture proportions. Because of the assumption in the Gaussian approximation, allele frequencies are bounded by zero and one, and the accuracy of modeling allele frequencies as a multivariate Gaussian decreases as the variances increase since more values land outside of the bounds. This is a limitation for any inference system using similar statistical models such as TreeMix [28].

In the medium-length divergence scenario, for suitable demographics, the simulation nicely recovered both the admixture proportions and the covariance

trees. The inferred covariance trees showed an accurate estimation of the relative positions of populations.

Effect of unknown number of components

To mimic real data analysis where the number of ancestry components is unknown, we designed two simulation studies, un-admixed and admixed.

In the simulation study shown in Figure 3.9, we simulated 120 individuals, un-admixed, in 6 populations, 20 individuals per populations. We simulated sequences of size 2,000,000 bp from which 12,961 markers survived the filtration. Simulation parameters are described in the figure. We then estimated using a range of K values.

In the simulation study shown in Figure 3.10, we simulated 140 individuals, un-admixed, in 7 groups, 20 individuals per group. The first six groups were un-admixed. The last group was a mixture of the first three. We simulated sequences of size 20,000,000 bp from which 125,787 markers survived the filtration. Simulation parameters are as described in the figure.

For simulations shown in Figure 3.9 and 3.10, we observe the progress of inferences when the number of assigned ancestry components increases. This mimics the process of real data analysis in which we apply a range of K values and evaluate the admixture and tree results without knowing the best component assignment. Both the un-admixed case shown in Figure 3.9 and the admixed case shown in Figure 3.10 demonstrate an ideal progression of the estimated quantities. Both the admixture and tree results fit with simulation. When the K value becomes too large ($K = 7$ in these simulations), we observe over-fitting where arbitrary individuals are assigned to the new component, which in turn forms a new branch on the phylogenetic tree that deviates from the simulated tree. The rest of the admixture and tree results still map to simulation properly.

Effect of joint inference

In the joint inference process, we attempt to find the best overall likelihood value $P_1 + P_2$. We first estimate the initial pair of Q and F, starting with random values. We then perform Nelder-Mead with the sample covariance matrix to estimate the initial rooted covariance matrix Ω'. After that we estimate Q and F, weighted by this matrix, and the updated Q and F are used in the next iteration to update Ω'.

The principle of weighted admixture inference is important because it provides the possibility of incorporating additional information, i.e. prior knowledge of the system. This is an important direction for future work.

The estimated quantities, however, move away from simulations. We observe a good progression in the likelihood values over the iterative process shown in Figure 3.11 and , and the admixture part of the likelihood P_1 reaches

the absolute optima after the first round of optimization, unweighted. But then it decreases steadily in later iterations, while the covariance part of the likelihood P_2 increases steadily. The sum $P_1 + P_2$ increases over iterations and eventually plateaus. We do not have a good explanation for this phenomenon.

Effect of unsampled population

Interpreting individual admixture is not always as straightforward as it appears. A simple two-component scenario can occur for many different reasons.

In this section, we present a simulation study demonstrating the effect of strange admixture caused by an unsampled population. We simulate four populations in which one receives gene flow from an unsampled population. When we infer admixture using all population, we see admixture results that fit simulation. When we drop all samples from the donor population, however, we occasionally observe strange admixture results. This is demonstrated in Figure 3.13.

The most common interpretation for the two-component sharing scenario is inward gene flow forming the minority component. Under different demographic conditions, however, it could be inward or outward gene flow, common ancestry but no gene flow, or the existence of unsampled populations like the ones shown in Figure 3.13

Effect of admixture-graph-like demography

After the inference of population covariances, Ohana provides a process to approximate the covariance values into a phylogenetic tree. This module gives us a fast visualization of the inferred quantities, but it does not always fit with the reality, which may or may not be tree-like.

In this section, we present a simulation study demonstrating the effect of a demography that is not tree-like. The simulated demography, shown in Figure 3.15, involves an admixture event, which forms the ancestor of one of the modern populations sampled for the analysis. For the comparison, we also present a simulation study that has a tree-like demography, shown in Figure 3.16.

We observe in Figure 3.16 an accurate recovery of the simulated demography. This is expected because this simulation follows a tree-like evolution. We observe from Figure 3.15 that the estimated covariance tree indeed does its best to visualize population relationships restricted to a tree-like evolution. The estimated population tree groups together populations that are close to each other with respect to split times and gene flow. For the portion of demography that does follow a tree-like evolution, the estimated population tree accurately recovers the simulated demography.

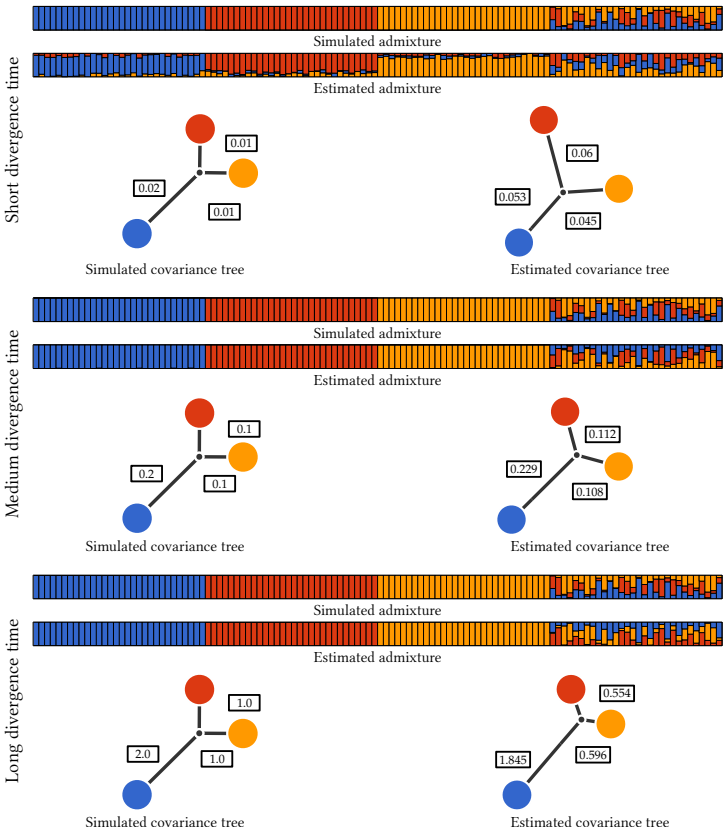


Figure 3.8: Simulation study for different divergence scenarios. The short divergence scenario (top) fails to accurately estimate individual admixtures. This leads to an inaccurate estimation of the population tree. The medium divergence scenario (middle) recoveries both the admixture and population tree nicely. The deep divergence scenario (bottom) produces accurate individual admixtures but poorly estimates the population tree.

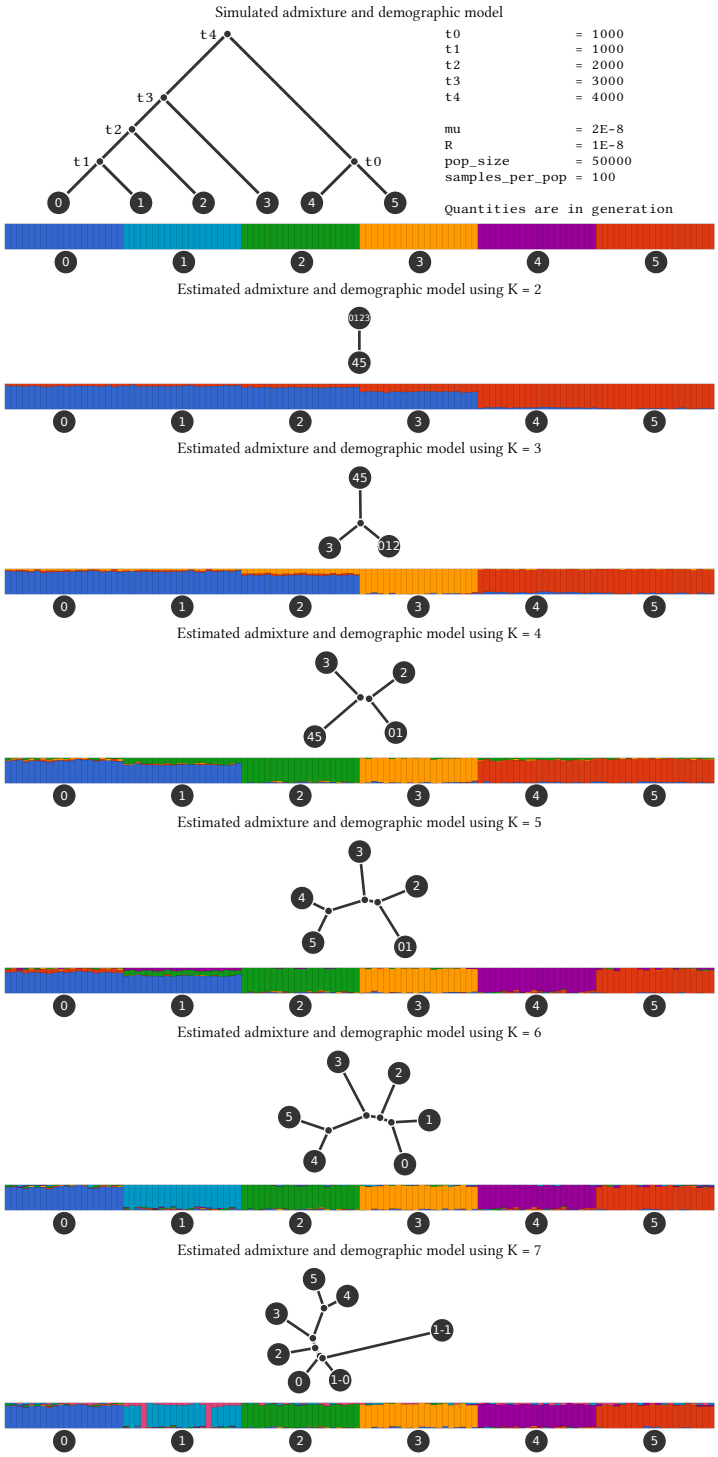


Figure 3.9: Simulation study for un-admixed case. We use coalescence simulation to produce sequences according to a given tree structure (top left). We estimate with a range of K values to mimic real data analysis. We observe good estimates for all quantities and a nice progression when the K value increases.

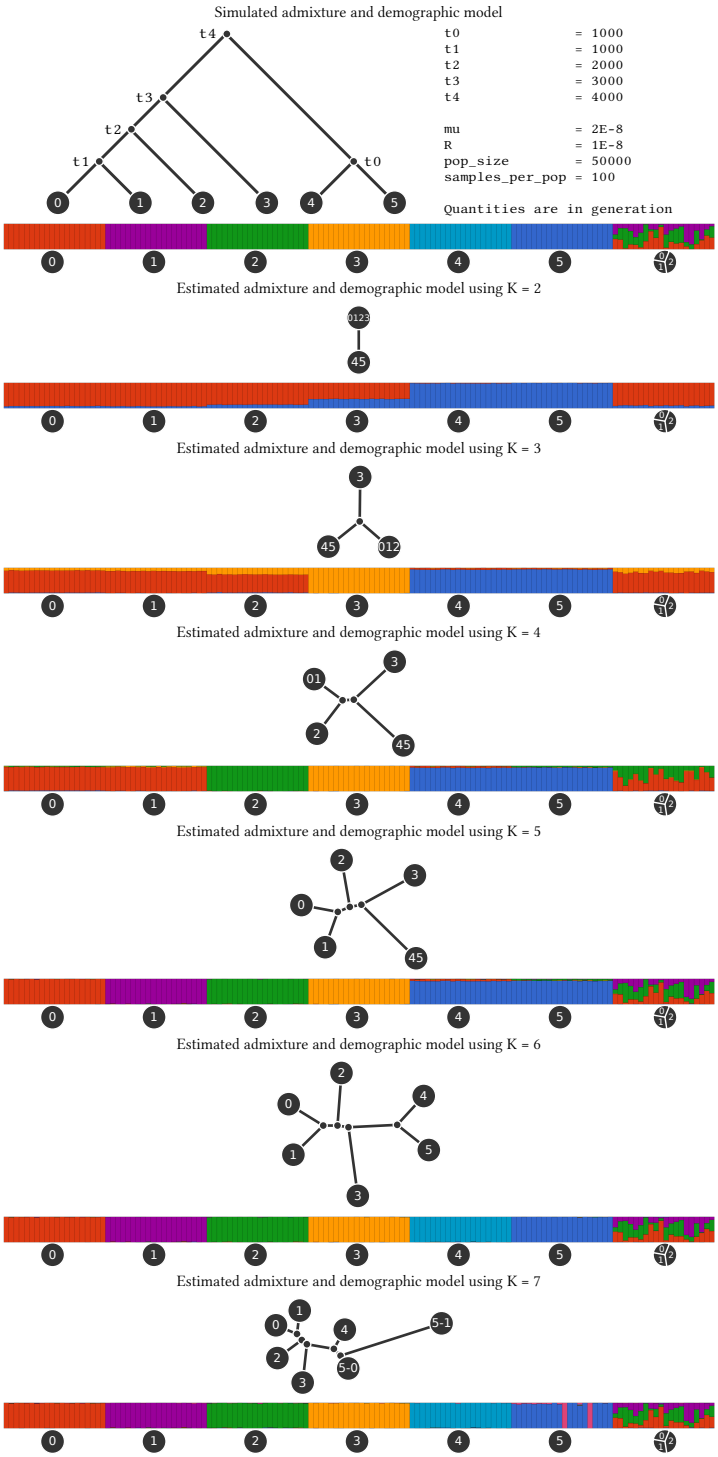


Figure 3.10: Simulation study for admixed case. This simulation experiment is similar to the one shown in Figure 3.9 except that we simulate one more group of samples and that these groups are admixed. All good results from the previous experiment shown Figure 3.9 persist after the addition of the admixed individuals and population.

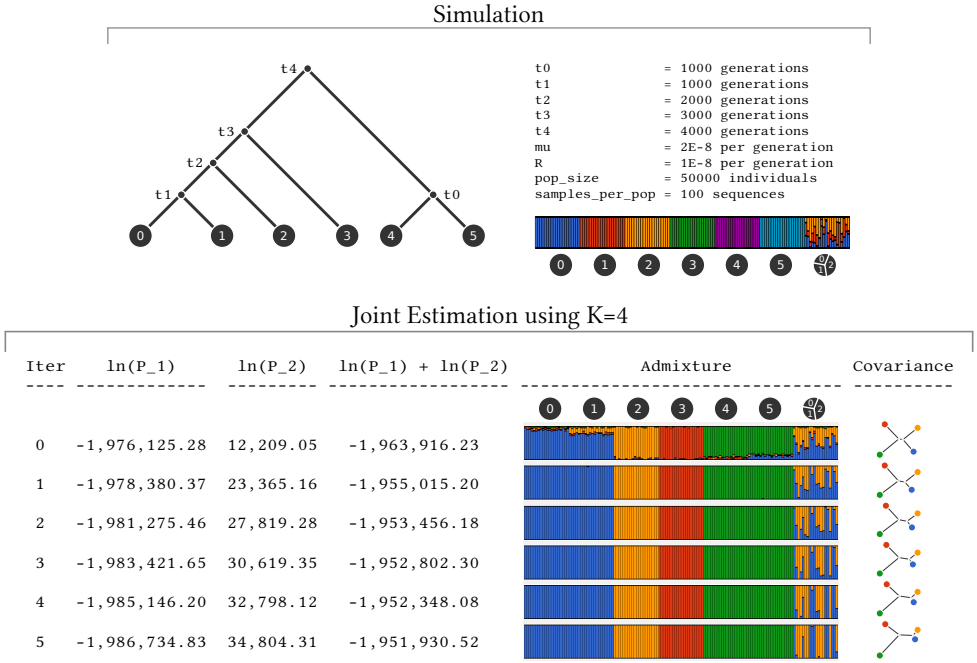


Figure 3.11: Effect of joint inference in simulation data, case #1. We use coalescence simulation to produce sequences according to a given tree structure (top left). We perform joint estimation over iterations of inferring admixture and covariances. We observe a good progression of P_1, P_2, and their sum, but over iterations the estimated quantities deviate from the simulation.

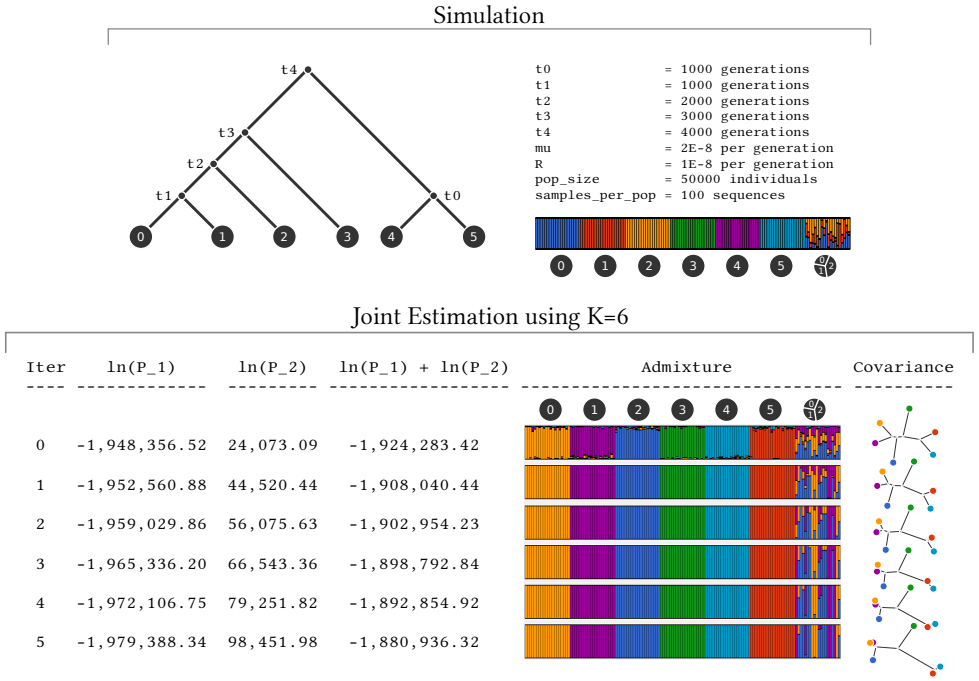


Figure 3.12: Effect of joint inference in simulation data, case #2. This simulation experiment is similar to the one shown in Figure 3.11 except we use a higher K value. The conclusions observed from Figure 3.11 still hold for this experiment.

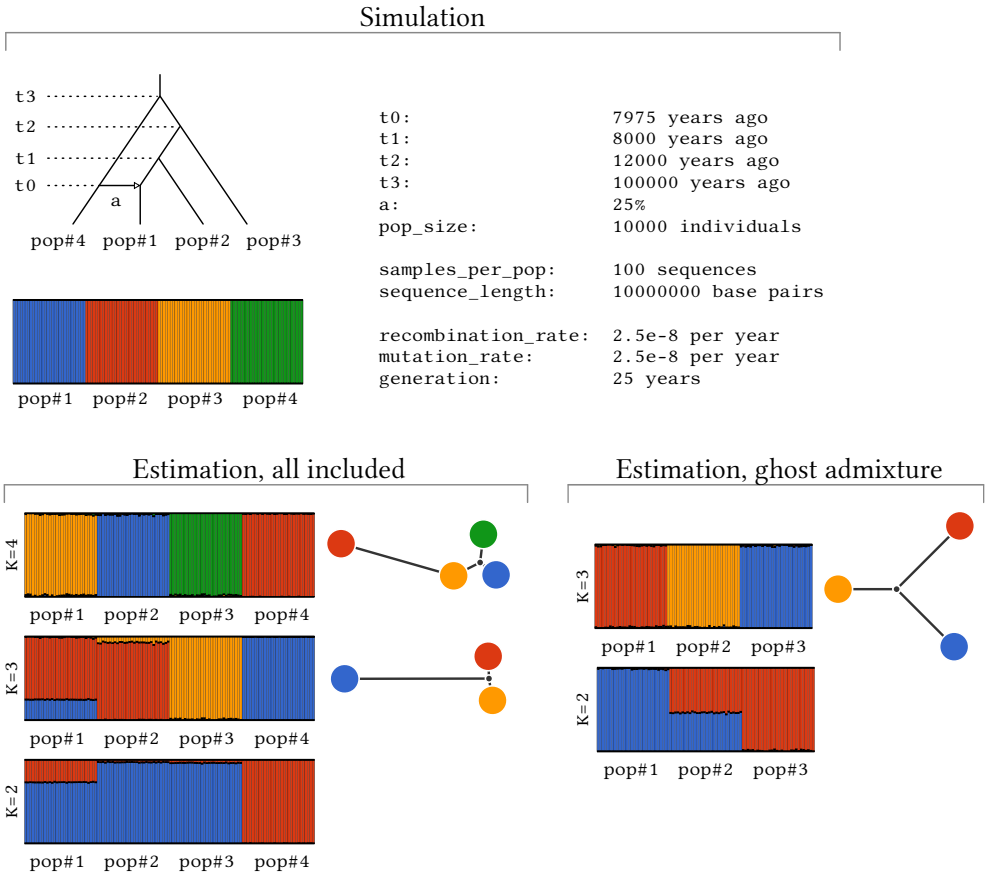


Figure 3.13: Unsampled source population causes abnormal admixture under certain simulation conditions. We simulate pop #1 to receive gene flow from pop #4. When we infer admixture using all population, we see proper admixing in pop #1. This is shown in $K = 3$ on the left. But when we remove pop #4 samples from the input data, we arrive at pop #2 as the admixed population. This is shown in $K = 2$ on the right.

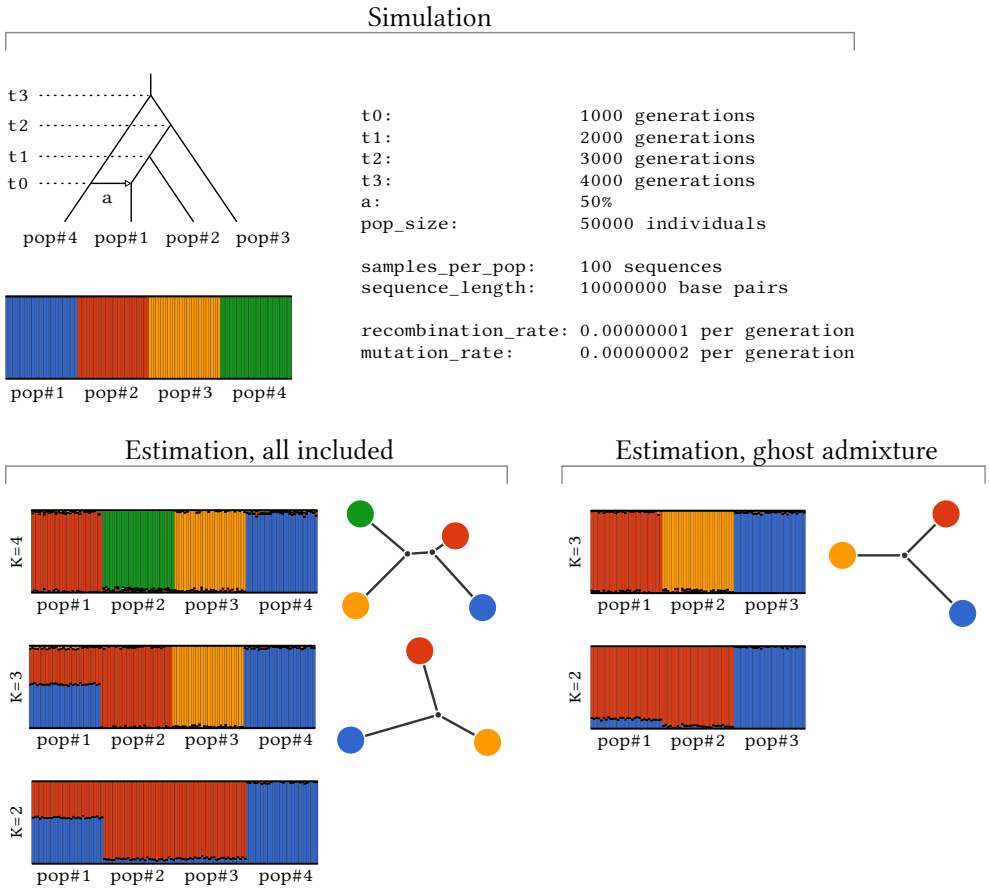


Figure 3.14: Unsampled source population, in most cases, does not cause strange admixture estimates. Except for the simulation parameters, this experiment is identical to the one in Figure 3.13. We do not observe the strange two-component sharing in pop #2.

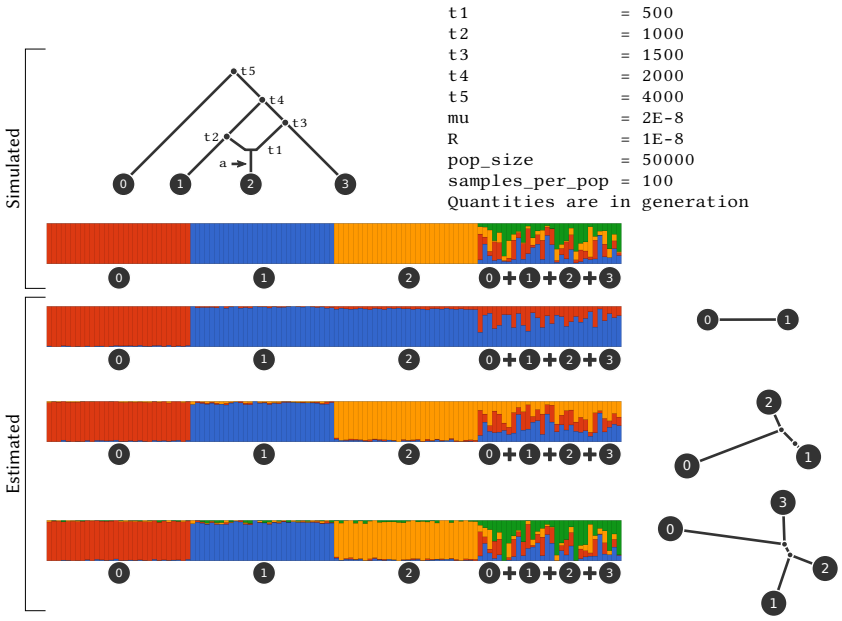


Figure 3.15: Simulation with a demography involving an admixture event. We observe good admixture results. The estimated tree structure provide the closest tree approximation of the simulated demography.

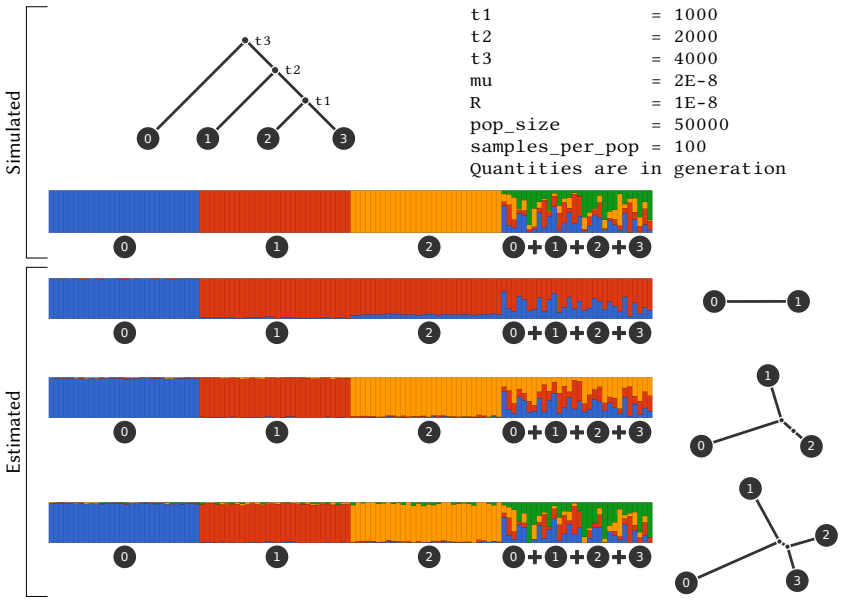


Figure 3.16: Simulation with a tree-like demography. We observe good inference results. The admixture and covariance relationship of all populations, including the un-sample population, are recovered.

3.8 Biological data analysis

Admixture and population tree

We have used Ohana in several collaborative projects to analyze real biological data. Since the initial software release, many other groups have also been using Ohana to analyze their own data. Due to the sensitivity of genomic data for any to-be-published work, in this section, we present a set of real data analysis that does not involve results that may be of interest in any data projects.

This dataset is a compilation of world population containing 80 individuals and 11,793 markers. For each K value, we collect 32 executions with random seeds from 0 to 31, and we report results from the execution that reached the best likelihood for this K. Figure 3.17 and 3.18 demonstrate the estimated admixture and population trees for a range of K values.

Selection study

The essential goal of Ohana is to perform selection study while fully taking advantage of structured genomic data. The selection model is fully developed but not as well tested through simulations. This is the immediate future task. We do, however, have some real data analysis.

We used a compilation of English, Han, and Yoruba from the 1000 Genomes project. The dataset contains 90 individuals, 30 per group, and 5,660,192 markers. We first sampled 161,645 markers to infer the admixture proportions. We then estimated allele frequencies for all markers. We scanned for these allele frequencies for covariance outliers using $\alpha \times \Omega$. We selected peak SNPs based on two criteria: this SNP has the highest likelihood ratio within 100,000 bp, and this SNP is not the only one within 100,000 bp that falls into the most extreme 1% of the likelihood ratios. Copied below are the top 50 highest peaks.

Figure 3.20 shows the top four genes within 1,000,000 regions. These are the top peaks, but multiple peaks within 1,000,000 regions are not shown. For example, peaks on SULT1C4 and GCC2 rank high, but they are within the 1,000,000 range of the peak on EDAR, which ranks even higher, so only the peak on EDAR is listed below.

chr	pos	rsid	English	Han	Yoruba	LLRatio	Gene
15	48426484	rs1426654	0.1167	0.4167	0.4000	14.0643	SLC24A5
5	33951693	rs16891982	0.1000	0.2000	0.1000	13.2122	SLC45A2
2	109513601	rs3827760	0.2667	0.3167	0.5667	10.6272	EDAR
17	4400392	rs11657785	0.5833	0.5000	0.3667	10.1500	x
2	108997262	rs4149433	0.4000	0.0333	0.1167	10.1104	SULT1C4
2	109118851	rs12618349	0.2667	0.0000	0.0000	9.6052	GCC2
2	109257152	rs1866188	0.1667	0.3500	0.1500	9.6052	LIMS1
15	28187772	rs1545397	0.0333	0.1000	0.3000	9.3464	OCA2
2	109362162	rs12620921	0.5333	0.5500	0.3667	9.1118	RANBP2
1	234332412	rs666263	0.4500	0.0167	0.1167	8.7796	SLC35F3
16	48375777	rs6500380	0.0000	0.0000	0.0333	8.7035	LONP2, MIR548AE2, MIR5095
10	78117929	rs2395375	0.3333	0.2333	0.2333	8.6358	C10orf11
2	26296089	rs34537429	0.2000	0.0667	0.5167	8.6302	RAB10
16	48258198	rs17822931	0.3000	0.7167	0.1000	8.3813	ABCC11
1	36037222	rs6425948	0.1167	0.1000	0.0500	8.3562	x
2	26113913	rs78404020	0.0000	0.0000	0.2833	8.2696	x
16	30602319	rs59385041	0.2000	0.1667	0.0167	8.2696	x
10	94855135	rs11187277	0.5000	0.0833	0.6833	8.1600	x
10	55613123	rs10763013	0.1333	0.4500	0.0000	8.1330	PCDH15
1	36158589	rs11264189	0.2000	0.0167	0.0000	8.1010	x
2	74761422	rs6707302	0.5500	0.4000	0.3167	8.1010	LOXL3
10	115148533	rs12262703	0.2667	0.4833	0.4667	8.0888	x
1	204859749	rs7541623	0.7000	0.1667	0.2333	8.0648	NFASC
12	80193361	rs2694658	0.1000	0.0000	0.1833	7.9769	PPP1R12A
1	35926150	rs1768560	0.2667	0.3333	0.5167	7.8710	KIAA0319L
12	101738184	rs6538985	0.1833	0.3500	0.2833	7.8698	UTP20
2	74864999	rs6739708	0.5333	0.4000	0.2667	7.7956	M1AP
10	119750413	rs7084970	0.1833	0.0000	0.0667	7.7956	x
20	2078995	rs2875718	0.0333	0.0000	0.2667	7.7769	x
15	28495956	rs12912427	0.1167	0.3000	0.0667	7.7576	HERC2
6	7058319	rs531077	0.1333	0.0667	0.3833	7.7549	x
2	242087712	rs7577489	0.2667	0.1500	0.1000	7.7457	PASK
1	168810025	rs10800388	0.1333	0.0333	0.2000	7.7028	x
1	1987803	rs2803309	0.9167	0.1000	0.1500	7.6561	x
2	74641624	rs2240444	0.3833	0.5500	0.1667	7.6200	C2orf81
3	64505376	rs11718026	0.5667	0.0833	0.1167	7.6200	ADAMTS9
20	2315543	rs6132532	0.3167	0.0000	0.7667	7.6066	TGM3
2	136407479	rs1446585	0.0500	0.1000	0.0330	7.6065	BNC2
15	28356859	rs1129038	0.3000	0.6000	0.0167	7.6065	HERC2
17	19174874	rs1467028	0.3167	0.4167	0.5167	7.6065	EPN2, EPN2-IT1
7	28065278	rs4722751	0.0500	0.0667	0.1667	7.5974	JAZF1
10	78889487	rs2616645	0.1667	0.1333	0.0833	7.5325	KCNMA1
2	13896241	rs12470874	0.2167	0.0000	0.0833	7.5227	x
12	80299468	rs10862022	0.1000	0.0000	0.2667	7.5005	PPP1R12A
10	119898665	rs853599	0.4000	0.3000	0.2000	7.4368	CASC2
1	35725203	rs11581846	0.0833	0.0000	0.1333	7.3991	x
1	1385211	rs1312568	0.8333	0.0167	0.1000	7.3947	ATAD3C
2	19075323	rs12472380	0.4500	0.0167	0.1167	7.3947	x

Table 3.2: The top 50 peaks of a selection scan of English, Han, and Yoruba. We identify peaks as the marker location that has the highest likelihood ratio within 100,000, and it is not the only marker within 100,000 that falls into the most extreme 1% of the likelihood ratios.

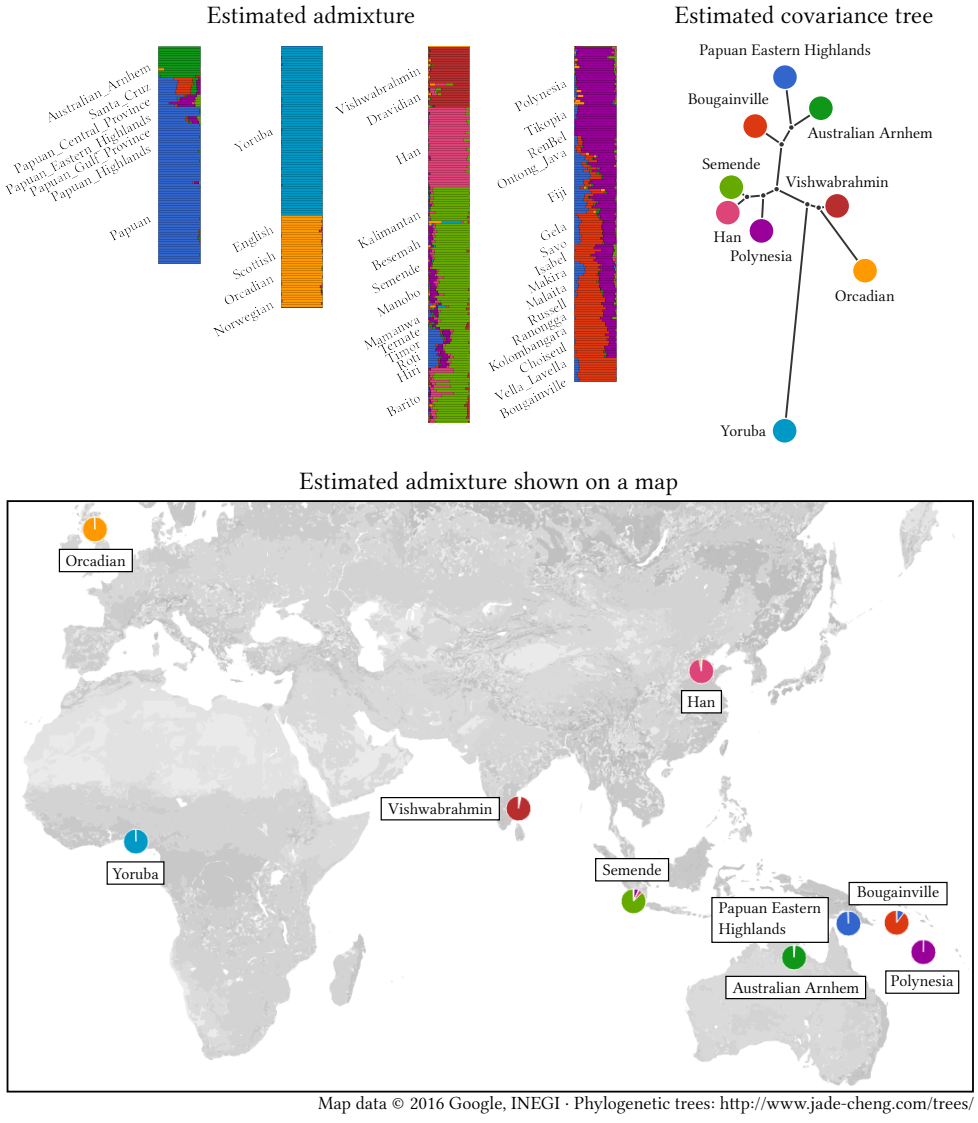


Figure 3.17: Admixture and population tree analysis of a world population containing 9 populations of 493 individuals and 11142 markers using $K = 9$.

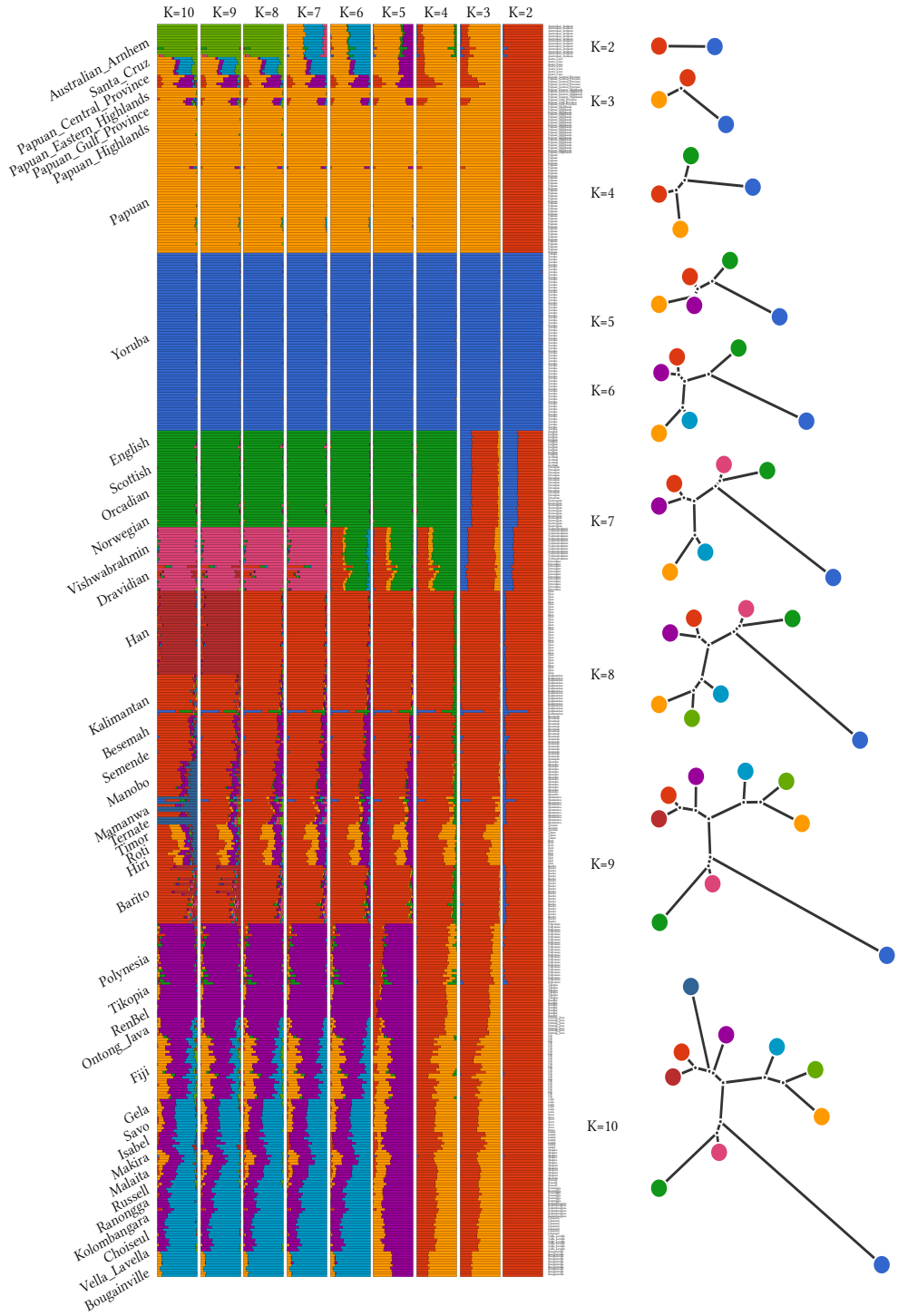


Figure 3.18: Admixture and covariance tree estimations for a range of K values. Here we see a nice progression of the estimated quantities. Each population component become isolated with respect to its relations with the rest of the populations.

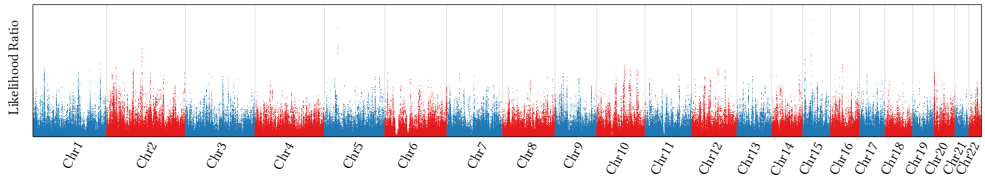


Figure 3.19: Summary of the selection scan for English, Han, and Yoruba. We obtain likelihood ratios from a selection scan using the scalar value multiplied to the entire covariance matrix. We plot the likelihood ratios against marker locations. Vast majority of the markers have a likelihood ratio of zero. The peak locations on this plot are the potential selection hot spots for further investigation.

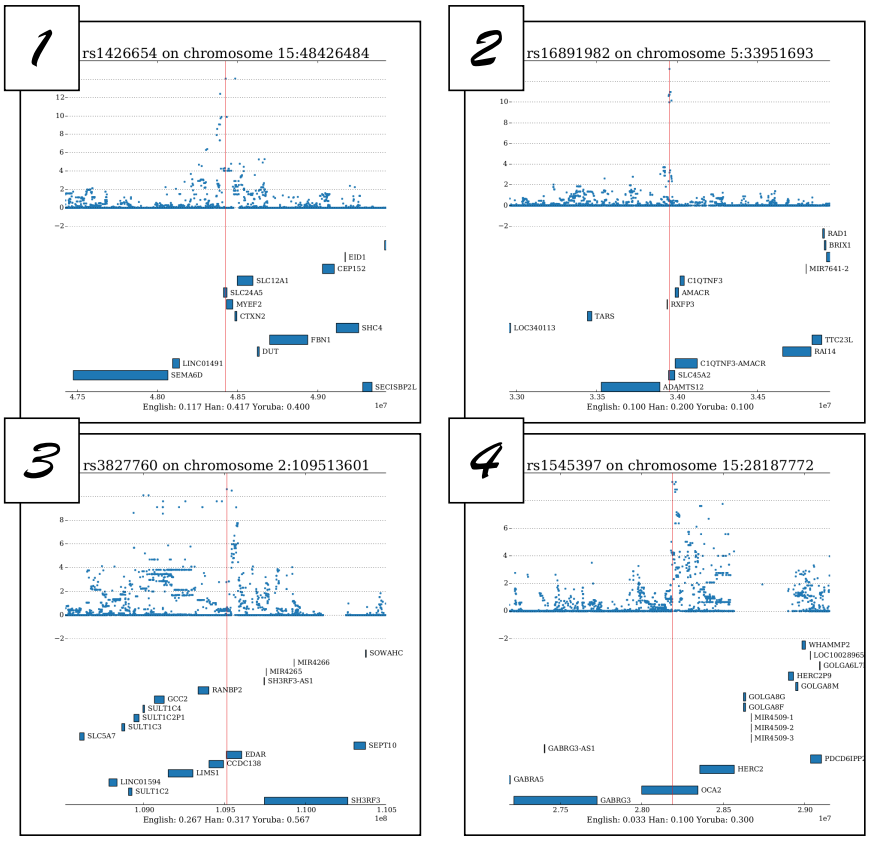


Figure 3.20: Example selection peaks and their annotations. They are the top four genes shown in a 1,000,000 regions. They are identified as the top four from likelihood ratios shown in Figure 3.19.

3.9 Software performance comparison

When genotype observations are used as the input, Ohana's **qpas** can be compared directed with previous software using the same modeling. This includes ADMIXTURE, STRUCTURE, and FRAPPE. In [1] we see the clear superiority of ADMIXTURE over other tools of the same class. In this section, we compare Ohana's **qpas** with ADMIXTURE. On average, Ohana outperforms ADMIXTURE for both accuracy and speed.

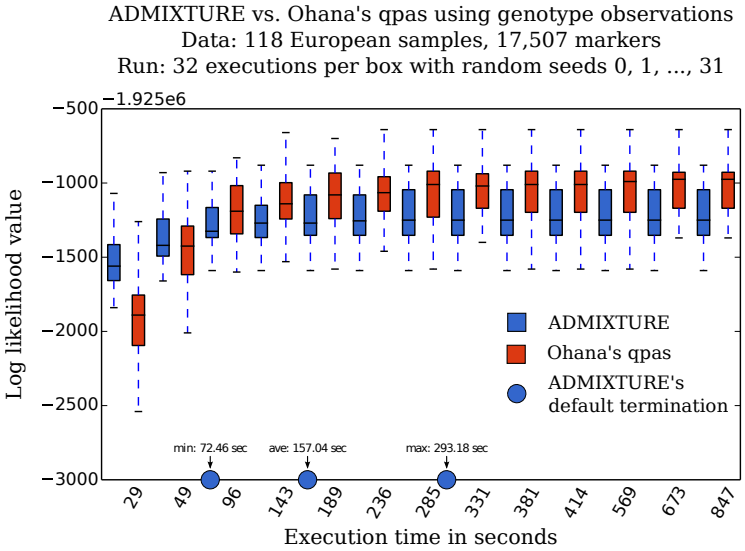


Figure 3.21: Performance comparison between Ohana's **qpas** and ADMIXTURE, currently the best software for admixture analysis. Ohana initiates with less optimal solutions, but it accelerates faster and outperforms ADMIXTURE in likelihood achievements.

	Dataset #1			Dataset #2		
K	Ohana	ADMIXTURE	Diff	Ohana	ADMIXTURE	Diff
2	-1967733	-1967733	0	-3835358	-3835365	7
3	-1956785	-1956799	14	-3799873	-3799887	14
4	-1946218	-1946244	26	-3788598	-3788607	10
5	-1935775	-1936025	250	-3777351	-3777361	11
6	-1925636	-1925877	241	-3766558	-3766540	-18
7	-1915552	-1915743	191	-3755851	-3755860	9
8	-1905430	-1905638	209	-3746227	-3745412	-815
9	-1895372	-1895879	507	-3735240	-3736079	839
10	-1885306	-1885466	160	-3725558	-3725624	66
11	-1875503	-1875853	350	-3715543	-3715157	-385
12	-1865492	-1865965	474	-3706069	-3707715	1646
13	-1855502	-1856262	760	-3697531	-3698519	987
14	-1845732	-1846490	758	-3688970	-3689124	154
15	-1836315	-1836775	460	-3681092	-3680829	-263
	Dataset #3			Dataset #4		
K	Ohana	ADMIXTURE	Diff	Ohana	ADMIXTURE	Diff
2	-1857263	-1857263	0	-288991	-288991	0
3	-1848450	-1848451	1	-279462	-279463	1
4	-1841198	-1841199	1	-275212	-275213	1
5	-1834377	-1834378	1	-271807	-271808	1
6	-1827829	-1827830	2	-268837	-268832	-5
7	-1821445	-1821458	13	-265907	-265923	17
8	-1815214	-1815214	0	-263052	-263096	44
9	-1809084	-1809101	18	-260268	-260440	172
10	-1802911	-1802906	-5	-257539	-257736	197
11	-1796763	-1796847	84	-254920	-254961	41
12	-1790671	-1790811	140	-252196	-252266	70
13	-1784688	-1784765	77	-249456	-249468	12
14	-1778599	-1778671	73	-246760	-246817	56
15	-1772555	-1772669	114	-244058	-244298	240

Table 3.3: Highest log likelihoods achieved from ADMIXTURE and Ohana's **qpas** over a range of K values. For each dataset, each program, and each K, we execute 100 times using random seeds $0, 1, ..., 99$. In 54 out of 60 cases, Ohana's **qpas** reports higher likelihoods. Dataset #1 contains 118 Europeans of 17,507. Dataset #2 is the benchmark dataset used in ADMIXTURE [1]. It contains 324 CEU, YRI, MEX, and ASW individuals and 13,928 markers. Dataset #3 is a compilation of 171 Han Chinese with 9,822 markers. Dataset #4 is a worldwide population of 60 individuals with 4,695 markers.

3.10 Concluding remarks

Project Ohana started during my time visiting prof. Rasmus Nielsen at the Center for Theoretical Evolutionary Genomics at the University of California Berkeley. I participated in a large, collaborative project researching the genetic history of Aborigine Australians. The admixture and population tree analysis I produced using Ohana fundamentally changed the nature of this project. The project concluded successfully, and a corresponding article was accepted and will appear in Nature. The results from Ohana's analysis appear in the main article. I also provided a brief outline of the framework, which appears in the Supplementary Information accompanying this Nature article, shown in Appendix D.

After the Australian project, I started to focus my effort on finalizing the methods and software development. The essential goal of this set of tools is to perform selection analysis fully taking advantage of structured genomic data. This naturally splits the framework into two parts, one for structure analysis and tree inference and the other for selection study. Prof. Rasmus Nielsen and I decided to write two manuscripts to describe the methods used in Ohana. The first one focused on the structure and population tree analysis, and the second one will focus on selection study.

The first method manuscript, shown in Appendix C, imparts the theory behind programs **qpas**, **cpax**, and **nemeco**. I evaluated the inference results with simulation study and real data analysis. I explored model limitations for Gaussian modeling of allele frequencies. I also performed software comparison to establish Ohana's admixture analysis as a faster and more accurate tool than the best tool currently available.

Besides the Australian project, I have been involved in several collaborative projects using Ohana. They are in different stages of development. Some are still in the data collection phase. Some have already reached the manuscript phase. A corresponding paper for one such project is included in Appendix E.

Many potentially prominent methods and applications could stem from the current framework implemented in Ohana. The modeling and optimization techniques used in Ohana are not restricted to population genetics. Mixture models are used in many subfields of machine learning, such as handwriting recognition, fuzzy image segmentation, and financial return models, to name a few. For the immediate future, we will continue to focus on population genetics.

Future work

The immediate next stage for project Ohana is the second method manuscript, which will focus on the selection module. For this work, I have implemented several selection models. These models leverage structured genomic data, and they vary in the strength of identifying certain types of selection signals,

e.g. focusing on local adaptation, differentiating ancient versus modern, and targeting each time slice over evolution. In addition, I have obtained some exciting results from real data analysis. What is most needed at this stage is additional simulation study, not only simulations under neutrality, but also simulations with selection signals. For the latter, I need to show that Ohana is able to identify selection signals, and then I will compare Ohana's selection power with other selection identification methods that require prior grouping, such as F statistics and population branch statistics.

The quadratic programing setup allows for additional likelihood modules to be inferred jointly, assuming their analytical forms satisfy certain conditions, such as being second-order differentiable and consistent with the block structure. This would provide ways to incorporate additional information or penalties while estimating the admixture and population trees. For example, it might be interesting to add weight parameters based on the geographic distances of the samples.

Another direction of extending the current framework could focus on the population relations and the tree-like assumption. Rather than estimating trees, population networks and admixture graphs would be more general and hence more informative. Exploring graph topologies or just tree topologies is a complex problem, but constraints could be added or a collection of possible topologies could be provided to assist the inference.

Appendix

CoalHMM method #1

This paper summarizes my work on CoalHMM's optimization and modular model construction. I investigated, implemented, and compared several black-box style optimization techniques with the emphasis on heuristic-based evolutionary algorithms. This paper also presents a range of models demonstrating the capability of complex model construction. Finally, in this paper I present simulation studies to evaluate these new additions.

Computational Biology and Chemistry 57 (2015) 80–92

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

Ancestral population genomics using coalescence hidden Markov models and heuristic optimisation algorithms

Jade Yu Cheng*, Thomas Mailund

Bioinformatics Research Centre, Aarhus University, C.F. Møllers Allé 8, 8000 Aarhus, Denmark

ARTICLE INFO

Article history:
Received 7 January 2015
Accepted 2 February 2015
Available online 5 March 2015

Keywords:
Sequential Markov coalescence
Coalescent hidden Markov models
Demographic inference
Numerical optimisation
Genetic algorithm
Particle swarm optimisation

ABSTRACT

With full genome data from several closely related species now readily available, we have the ultimate data for demographic inference. Exploiting these full genomes, however, requires models that can explicitly model recombination along alignments of full chromosomal length. Over the last decade a class of models, based on the sequential Markov coalescence model combined with hidden Markov models, has been developed and used to make inference in simple demographic scenarios. To move forward to more complex demographic modelling we need better and more automated ways of specifying these models and efficient optimisation algorithms for inferring the parameters in complex and often high-dimensional models.

In this paper we present a framework for building such coalescence hidden Markov models for pairwise alignments and present results for using heuristic optimisation algorithms for parameter estimation. We show that we can build more complex demographic models than our previous frameworks and that we obtain more accurate parameter estimates using heuristic optimisation algorithms than when using our previous gradient based approaches.

Our new framework provides a flexible way of constructing coalescence hidden Markov models almost automatically. While estimating parameters in more complex models is still challenging we show that using heuristic optimisation algorithms we still get a fairly good accuracy.

1. Background

Coalescence theory provide a very powerful framework for genetics modelling and inference, and the coalescence process with recombination underlies many important analysis tools. Drawing inference from sequences with recombination, however, often involves integrating over all possible ancestries, modelled as the so-called ancestral recombination graph (ARG), a process that rarely scales to more than a few, short sequences due to the complexity and state space size of the ARG. To alleviate this, the *sequential Markov coalescence* approximation assumes that statistical dependencies between local genealogies are Markov (McVean and Cardin, 2005; Marjoram and Wall, 2006; Chen et al., 2009; Hobolth and Jensen, 2014).

In recent years a number of inference tools have been developed based on combining the sequential Markov coalescence with hidden Markov models, constructing so-called *coalescence hidden Markov models* or CoalHMMs, that have been constructed for the inference of speciation times (Hobolth et al., 2007; Dutheil et al., 2009; Mailund et al., 2011), gene-flow patterns (Steinrücken et al., 2013; Mailund et al., 2012), changing population sizes (Li and Durbin, 2011; Sheehan et al., 2013; Schiffels and Durbin, 2014) or inference of recombination patters (Munch et al., 2014) and have been used in a number of whole genome analyses (Locke et al., 2011; Scally et al., 2012; Prado-Martinez et al., 2013; Prüfer et al., 2012; Miller et al., 2012). These models exploit that even a very small sample of full genomic sequences holds a wealth of information about the sample's ancestry: Loci sufficiently far apart in the genome can, because of recombination in the sample's history, be considered essentially independent samples from the underlying sample populations.

The crux of constructing a CoalHMM is describing the probability of transitioning from one local genealogy along a sequence alignment to the next in terms of the underlying population genetics parameters of interest. This is typically done either by considering the probability of changing to a new genealogy conditional on a current one (Hobolth and Jensen, 2014; Li and Durbin, 2011) or by considering the joint distribution of two neighbouring trees (Dutheil et al., 2009; Mailund et al., 2011). In either case it

* Corresponding author. Tel.: +45 87155572.
E-mail addresses: ycheng@birc.au.dk (J.Y. Cheng), mailund@birc.au.dk (T. Mailund).

http://dx.doi.org/10.1016/j.compbiolchem.2015.02.001

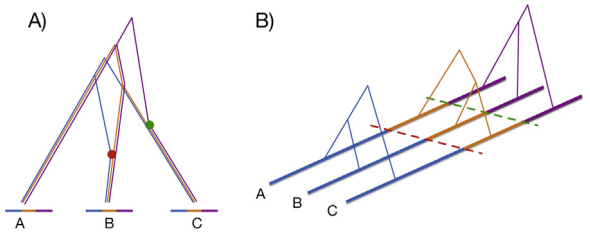


Fig. 1. (A) An ancestral recombination graph over three sequences, showing two recombinations and (B) the corresponding three local genealogies. The example shows the ancestry of three sequences in the case where they have experienced two recombination events, shown in red and green. These recombinations segments the sequences into three regions, shown in blue, orange and purple, each with different tree genealogies.

involves the explicit enumeration of all possible genealogies and a set of formulas for each possible transition. The formulas for transition probabilities, however, are very similar for transitions between similar genealogies and so constructing these formulas can be somewhat automated (Mailund et al., 2012).

Below we give a short introduction to the essentials of coalescence theory and coalescent hidden Markov models for inference of demographic parameters and in Section 3 we describe a new framework we have developed that makes it simple to construct so-called isolation-with-migration demographic models for analysis of pairwise alignments. This framework is similar to a more general framework for larger sample sizes (Mailund et al., 2012) but automates much of the model specification. The new framework is available under open source licence GPLv2 at https://github.com/mailund/IMCoalHMM.

1.1. Coalescence processes

Coalescence theory (Hein et al., 2005) describes the ancestry of a sample of present day genes and gives probabilities to all the possible genealogies that could have created the variation seen in the samples. The typical description of the coalescence model is as a continuous time Markov process running backwards in time, describing the various events that could have occurred in the past. An outcome of such a process is a tree-genealogy where inner nodes correspond to where two lineages find their most recent common ancestor. The time-depths of these nodes, and thus the branch lengths of the tree, are given by the rate of coalescence, a parameter that is determined by the size of the population the samples are taken from.

Extended with recombination, each lineage can also split into two. At a recombination event a lineage is split into a left and a right segment that then evolve back in time as two independent lineages. The outcome of this process is no longer a tree but a directed acyclic graph called the *ancestral recombination graph* or ARG (see Fig. 1A). While not a tree itself, the ARG represents a set of trees since at each position along the sample sequences a single tree describes the genealogy at that position (see Fig. 1B). At positions where a recombination has occurred the tree to the left and to the right of the recombination position can be different. The probability density over all possible ARGs thus also provides a joint probability for all the corresponding local tree-genealogies.

Structured populations can be modelled by assigning lineages to different populations, allow migration events to move lineages from one population to another, and only allow lineages to coalesce when within the same population. Population splits or admixing can be added simply by setting populations to be equal or randomly assigning lineages with one label to two or more new population labels.

Mutations on lineages can also be considered events that can occur as the process runs back in time, but typically mutations are put on the coalescence tree or ARG after it is simulated. There, the mutations can simply be put on the genealogy as a Poisson process or be put on inner nodes using a substitution model. The latter approach makes it possible to sum over all possible sequences at internal nodes using standard methods such as Felsenstein's peeling algorithm (Felsenstein, 1981) and this way obtain a joint probability distribution for the sequences at the leaves, i.e. the present day samples. This distribution depends only on the local tree-genealogies induced by the ARG since the possible nucleotides at any given position only depends on the tree for that given position.

If we denote by θ the relevant parameters for the coalescence process, e.g. coalescence rates, migration rates, recombination rates and mutation rates, we can let $f(\mathcal{G} \mid \theta)$ denote the probability density for the process producing the specific genealogy $\mathcal{G}$ and let $f(\mathcal{A} \mid \mathcal{G}, \theta)$ denote the probability that putting mutations on genealogy $\mathcal{G}$ produces the aligned samples $\mathcal{A}$. Typically the latter only depends on the mutation rate while the former is independent of the mutation rate but depends on rates (migration, recombination etc.) and time units (e.g. times where a population split apart or migration between two populations happend). These latter parameters can be expressed in time units of mutations, in essence setting $\mu = 1$, so we can simplify the two densities to just $f(\mathcal{G} \mid \theta)$ and $f(\mathcal{A} \mid \mathcal{G})$.

For demographic inference it is the parameters θ that are of interest rather than the actual underlying genealogy which is considered a nuisance parameter to be integrated out to get the likelihood

$$\mathrm{lhd}(\theta \mid \mathcal{A}) = \int f(\mathcal{A} \mid \mathcal{G}) f(\mathcal{G} \mid \theta) \;\; \mathrm{d}\mathcal{G}.$$

This integral over all possible genealogies is generally not efficiently computable and must either be approximated through sampling approaches or by approximating the coalescence process with a simpler model where the integral *can* be computed. The latter is the approach taken with coalescence hidden Markov models.

1.2. Coalescence hidden Markov models

The key approximation in CoalHMMs is assuming that the distribution of local genealogies along an alignment is Markov in the sense that when moving from one tree to another across a recombination point, the next tree depends only on the current tree and not any others. By approximating the distribution of local genealogies by a Markov chain the probability of the full genealogy reduces to specifying the joint probability of two neighbouring genealogies (which might be identical genealogies, e.g. if there is no recombination between them). Let ℓ denote the "left" genealogy and r the "right" genealogy and $J_\theta(\ell, r)$ their joint density. Then the "transition density" $T_\theta(r \mid \ell)$ is given simply by

$$T_\theta(r \mid \ell) = \frac{J_\theta(\ell, r)}{p_\theta(\ell)}.$$

where we define

$$p_\theta(\ell) = \int J_\theta(\ell, r') \mathrm{d}r'.$$

as the marginalisation over all possible right genealogies and thus the likelihood for just seeing the left genealogy.

If our data $\mathcal{A}$ consists of L nucleotides then the underlying genealogy $\mathcal{G}$ consists of L local trees $\mathcal{G} = \mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_L$ then

$$f(\mathcal{G} \mid \theta) = p_\theta(\mathcal{G}_1) \prod_{i=2}^{L} T_\theta(\mathcal{G}_i \mid \mathcal{G}_{i-1}).$$

The alignment probability given these local genealogies separates into probabilities for the individual nucleotides so if $\mathcal{A}_i$ denotes the i'th column in the alignment then

$$f(\mathcal{A} \mid \mathcal{G}) = \prod_{i=1}^{L} E(\mathcal{A}_i \mid \mathcal{G}_i).$$

where $E(\mathcal{A}_i \mid \mathcal{G}_i)$, the "emission probability", is the probability that the $\mathcal{A}_i$ column was produced by tree $\mathcal{G}_i$ and can be computed using the peeling algorithm.

In order to integrate over all genealogies we further approximate by discretising the possible time points where inner nodes can be found in the trees. We split the possible coalescence times into n intervals and place all events in the same interval at a single time point. This reduces the space of possible genealogies to a finite set that can be explicitly summed over, so

$$p_\theta(\ell) = \sum_{r'} J_\theta(\ell, r').$$

and

$$\int f(\mathcal{A} \mid \mathcal{G}) f(\mathcal{G} \mid \theta) \mathrm{d}\mathcal{G} = \sum_{\mathcal{G}_1, \ldots, \mathcal{G}_L} \left[p_\theta(\mathcal{G}_1) E(\mathcal{A}_1 \mid \mathcal{G}_1) \prod_{i=2}^{L} T_\theta(\mathcal{G}_i \mid G_{i-1}) E(\mathcal{A}_i \mid \mathcal{G}_i) \right].$$

This equation takes the form of a hidden Markov model (Rabiner, 1989) where the sequence $\mathcal{A}_1, \ldots, \mathcal{A}_L$ is the observable sequence and $\mathcal{G}_1, \ldots, \mathcal{G}_L$ the hidden Markov sequence. There is an exponential number of genealogies this way but by rearranging the sum and using dynamic programming in what is known as the Forward algorithm it can be computed in time $O(N^2 L)$ where L is the sequence length and N the number of possible genealogies. In the framework we describe in this paper we always consider pairwise alignments so a local genealogy consists simply of a coalescence time and with n time intervals there are n possible genealogies, and thus the likelihood of a demographic model can be computed in $O(n^2 L)$ running time using a CoalHMM, once $J_\theta(\ell, r)$ is specified.

In practise we can exploit repetitions in the alignment to reduce it further and in our framework we use the ZipHMM library (Sand et al., 2013) that lets us compute the likelihood of an entire genome alignment in a a few seconds to a few minutes depending on how finely we discretise time. For this library we simply need to specify the hidden Markov model using the transition matrix $T_\theta(r \mid \ell)$ which we compute using $J_\theta(\ell, r)$ and the emission matrix $E(\mathcal{A}_i \mid \mathcal{G}_i)$ which we compute using a Jukes-Cantor substitution model (Jukes and Cantor, 1969), where it is simply determined by the coalescence time of the $\mathcal{G}_i$ genealogy. The way our new framework makes it almost automatic to compute $J_\theta(\ell, r)$ is described in Section 2.

1.3. Parameter inference

Previous versions of our CoalHMM framework used the Nelder–Mead method (Nelder and Mead, 1965), or downhill simplex method, to estimate the parameter set for a CoalHMM by maximising the log-likelihood values calculated from the Forward algorithm. This optimisation method was developed by John Nelder and Roger Mead in 1965 as a technique to minimise an objective function in a many-dimensional space. In the context of CoalHMM, each dimension corresponds to a model parameter. CoalHMM infers parameters using maximum likelihood estimations, so the scores returned from its objective function simply correspond to the negated log-likelihood values.

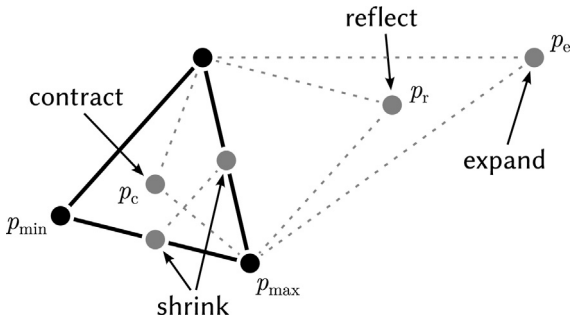


Fig. 2. An iteration of the Nelder–Mead method over two-dimensional space, showing point p_{min} reflected to point p_r, expanded to point p_e, or contracted to point p_c. If these test points do not improve the overall score of the simplex, then it shrinks around the point p_{max} with the highest score.

The Nelder–Mead method is an iterative process that continually refines a simplex, which is a polytope of $D+1$ vertices in D dimensions. During each iteration, the objective function is evaluated to determine a score at each point in the simplex (see Fig. 2). The point p_{min} with the lowest score is reflected through the centroid of the remaining vertices to point p_r. If the score at p_r is neither the highest nor the lowest score, then p_r is used in place of p_{min} to form the simplex for the next iteration. If the score at p_r is the highest score in the simplex, then this reflected point is expanded away from the centroid to p_e and used in place of p_{min} to form the next simplex. If the score at p_r is still the lowest score, then p_r is contracted toward the centroid to point p_c. If the score at p_c is no longer the lowest score, then it is used to replace p_{min} to form the next simplex. Otherwise, all points in the simplex shrink around the point p_{max} with the highest score. This process continues until the simplex collapses beyond a predetermined size, a maximum length of time expires, or a maximum number of iterations is reached.

The amount of effect these possible actions have on the simplex is controlled by supplying to the algorithm coefficients for reflection ρ, expansion χ, contraction γ, and shrinkage σ. Standard values are $\rho = 1$, $\chi = 2$, $\gamma = 1/2$, and $\sigma = 1/2$ (Baudin, 2009); but fine-tuning these coefficients has the potential to improve the performance of the algorithm.

1.3.1. Genetic algorithms

A Genetic Algorithm (GA) is a type of evolutionary algorithm. This optimisation technique gained popularity through the work of John Holland in the early 1970s (Holland, 1992). It operates by encoding potential solutions as simple chromosome-like data structures and then applying genetic alterations to those structures. Over many iterations, its population of chromosomes evolves toward better solutions, which it determines based on fitness values returned from an objective function. The algorithm typically terminates when the diversity of its population reaches a predetermined minimum, a maximum length of time expires, or a maximum number of iterations has completed.

GAs typically operate in three phases: Selection, Crossover, and Mutation (see Fig. 3). Selection determines a subset of a population what will breed the next generation of individuals, and a variety of selection schemes exist. In one scheme, Roulette Wheel Selection (RWS) (Goldberg, 1989), the algorithm selects individuals based on their relative fitness within the population; the probability p_i of selecting an individual i is given by $p_i = f_i / \sum_{j=1}^{N} f_j$, where f_i is the fitness of the individual and N is the population size. While RWS works by repeatedly sampling the population, a variation of RWS,

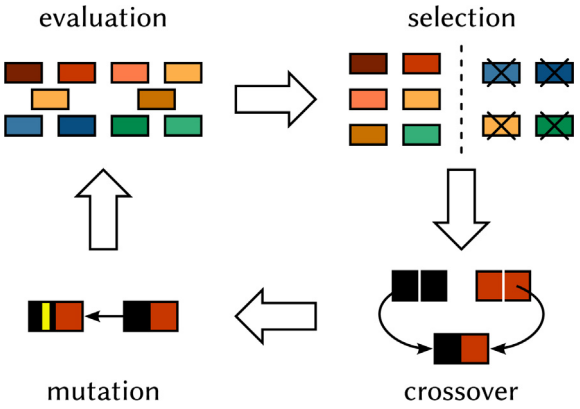


Fig. 3. In one iteration of the genetic algorithm's evolution, it operates in three stages: *Selection*, where it chooses a relatively fit subset of individuals for breeding; *Crossover*, where it recombines pairs of breeders to create a new population; and *Mutation*, where it potentially modifies portions of new chromosomes to help maintain the overall genetic diversity. Arrows in the diagram indicate transitions into the next genetic operation within one generation.

Stochastic Universal Sampling (SUS) (Baker, 1987), uses a single random value to sample all breeders by choosing them at evenly spaced intervals; this gives less fit individuals a greater chance to breed. RWS and SUS are both examples of fitness proportionate selection, but other selection schemes are based only on rank, and these are particularly beneficial when the lower and upper bounds of a fitness function are hard to determine. For example, in Tournament Selection (Miller et al., 1995), the algorithm selects an individual with the highest fitness value from a random subset of the population.

Crossover is a genetic operation used to combine pairs of individuals previously selected for breeding the following generation, and like Selection, several Crossover schemes exist. In One Point Crossover, the algorithm chooses a single point on both parents' chromosomes, and it forms the child by concatenating all data prior to that point from the first parent with all data after that point from the second parent. In Two Point Crossover, the algorithm instead chooses two points, which splits the parents' chromosomes into three regions; the algorithm then forms the child by concatenating the first region from the first parent, the second region from the second parent, and the third region from the first parent. While nature serves as the inspiration for One and Two Point Crossover, Uniform Crossover (Syswerda, 1989) has no such biological analogue. In Uniform Crossover, each position on the child's chromosome has equal opportunity to inherit its data from either parent.

Mutation is the third phase in many GAs. Every position on every chromosome has a certain probability to mutate, which helps the population maintain or even improve its genetic diversity. Several variants of this technique exist. In Uniform Mutation (Michalewicz, 1996), when a position mutates, the algorithm replaces its value with a new value, chosen at random, between a predetermined lower and upper bound. In another variant, Gaussian Mutation (Deb, 2001), when a position mutates, its current value increases or decreases based on a Gaussian random value.

1.3.2. Particle swarm optimisation

Particle Swarm Optimisation (PSO) is another type of heuristic based search algorithm. Eberhart and Kennedy first discovered and introduced this optimisation technique through simulation of a simplified social model in 1995 (Eberhart and Kennedy, 1995). Similar to GAs, PSOs are highly dependent on stochastic processes. Each individual in a PSO population maintains a position and a velocity as it flies through a hyperspace in which each dimension corresponds to one position in an encoded solution. Each individual contains a

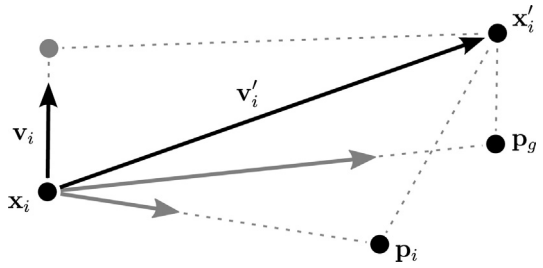


Fig. 4. Three vectors applied to a particle at position $\mathbf{x}_i$ in one iteration of a Particle Swarm Optimisation: a cognitive influence urges the particle toward its previous best $\mathbf{p}_i$, a social influence urges the particle toward the swarm's previous best $\mathbf{p}_g$, and its own velocity $\mathbf{v}_i$ provides inertia, allowing it to overshoot local minima and explore unknown regions of the problem domain.

current position, which evaluates to a fitness value. Each individual also maintains its personal best position $\mathbf{p}_i$ and tracks the global best position $\mathbf{p}_g$ of the swarm (see Fig. 4). The former encapsulates the cognitive influence, and the latter encapsulates the social influence. A PSO works as an iterative process. After each iteration, the algorithm adjusts the position of each individual based on its knowledge of $\mathbf{p}_i$ and $\mathbf{p}_g$. This adjustment is analogous to the crossover operation used by GAs. The inertia of an individual, however, allows it to overshoot local minima and explore unknown regions of the problem domain.

In PSO, we represent the position of the ith particle as $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,D})$ and its velocity as $\mathbf{v}_i = (v_{i,1}, v_{i,2}, \ldots, v_{i,D})$, where D is the number of dimensions in the parameter space. We represent the particle's previous position with its best fitness as $\mathbf{p}_i = (p_{i,1}, p_{i,2}, \ldots, p_{i,D})$. During each iteration, the algorithm adjusts the velocity $\mathbf{v}$ and position $\mathbf{x}$ according to the following equations:

$$\begin{aligned}\mathbf{v}'_{i,d} &\leftarrow \mathbf{v}_{i,d} \quad +\phi_p \cdot r_p \cdot (\mathbf{p}_{i,d} - \mathbf{x}_{i,d}) \\ &\qquad\quad +\phi_g \cdot r_g \cdot (\mathbf{p}_{g,d} - \mathbf{x}_{i,d}) \\ \mathbf{x}'_{i,d} &\leftarrow \mathbf{x}_{i,d} \quad +\mathbf{v}_{i,d}.\end{aligned}$$

where r_p and r_g are two random values between zero and one, and ϕ_p and ϕ_g are two positive constants representing cognitive and social influences. As Shi and Eberhart demonstrated (Shi and Eberhart, 1998), it can be beneficial to include a constant ω, which helps balance the global and local search forces. This term directly affects the inertia of the particle.

$$\begin{aligned}\mathbf{v}'_{i,d} &\leftarrow \omega \cdot \mathbf{v}_{i,d} \quad +\phi_p \cdot r_p \cdot (\mathbf{p}_{i,d} - \mathbf{x}_{i,d}) \\ &\qquad\quad +\phi_g \cdot r_g \cdot (\mathbf{p}_{g,d} - \mathbf{x}_{i,d}).\end{aligned}$$

2. Methods

We first describe how our framework supports constructing CoalHMMs for pairwise alignments and then the algorithms we have implemented for parameter estimation.

2.1. Framework for CoalHMMs for pairwise alignments

Our framework builds the joint probability distribution $J_\theta(\ell, r)$ by tracking all possible states of the coalescence process for two samples with two nucleotides similar to our previous work (Mailund et al., 2011, 2012, 2012). Demographic scenarios are specified by slicing the past into a number of "epochs" where each such has a fixed number of populations and a fixed number of constant rates with which events occur. Within each epoch we construct the state space of all possible configurations within the demographic model of the epoch and construct a continuous time Markov chain (CTMC). Finally we stack these CTMCs on top of each other to get

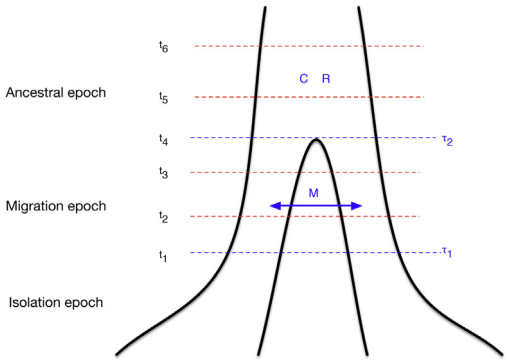


Fig. 5. The demographic IIM model. The model has three epochs and five parameters. An ancestral population epoch with one population and free coalescences, a migration epoch with two populations where lineages can only coalesce within the same population but can migrate between the populations, and an isolation epoch where the two populations are completely independent. The parameters are the time points where the system switches between the epochs, the coalescence and recombination rates (assumed to be the same in all populations) and a symmetric migration rate during the migration epoch. The time point $t_1, t_2, \ldots, t_6$ illustrates a possible discetisation of time into the intervals that becomes the states of the hidden Markov model.

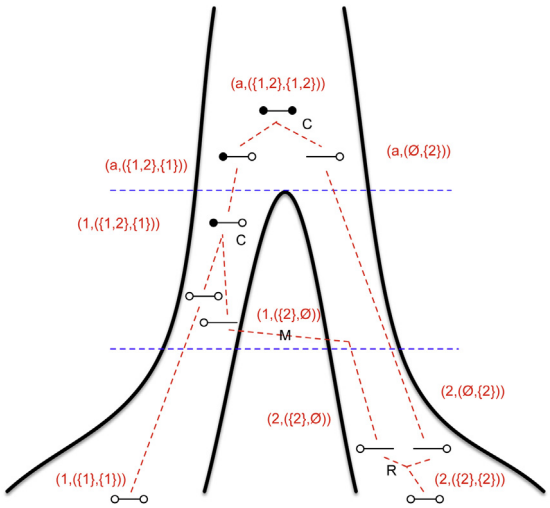


Fig. 6. An ancestral recombination graph in the IIM model with lineages in the notation of the transition system. The state at any particular point in time, corresponding to a horizontal line through the ARG, would be the number of lineages at that particular time. The initial state is {(1,({1},{1})), (2,({2},{2}))} that through a recombination transition (R) moves to {(1,({1},{1})), (2,({2},Ø)), (2,(Ø,{2}))}. The system now moves from its isolation epoch to its migration epoch and the next event is a migration event (M) that changes the state to {(1,({1},{1})), (1,({2},Ø)), (2,(Ø,{2}))} followed by a coalescence event (C) and the state {(1,({1,2},{1})), (2,(Ø,{2}))}. Now the system moves to the ancestral population epoch where this state is projected to the state {(*a*, ({1, 2}, {1})), (*a*,(Ø,{2}))} and the final event is a coalescence event changing the state to {(*a*, ({1, 2}, {1, 2}))}.

a coalescence process for the two samples for all the combined epochs and from this compute the joint probabilities of which intervals the left and right nucleotides will coalesce in.

As an example, consider the Isolation with Initial Migration (IIM) model from Mailund et al. (2012) and shown in Fig. 5. This model has three epochs. From the most recent to the most ancient these are (1) and epoch with complete isolation where lineages in the two populations can never coalesce, (2) an epoch with population structure where there are two distinct populations but where lineages can migration between them, and finally (3) an epoch with a single ancestral population.

The first epoch allows lineages to recombine and coalesce within each population but does not allow migrations. In this time period it is not possible for the two samples to find a common ancestor. In the migration period, lineages can cross from one population to another and coalesce into a common ancestor. In the final epoch the lineages can coalesce and find common ancestors freely. To build a CoalHMM for this demographic model it is necessary to build CTMCs for the three epochs, combine them to build a model for the entire demographic past and then use this model to specify the joint probability $J_\theta(\ell, r)$.

2.1.1. *Building continuous time Markov chains*

To track the possible histories within an epoch we explicitly construct the state space of the two-locus coalescence process; an approach taken in several earlier papers (Slatkin and Pollack, 2006; Simonsen and Churchill, 1997; Mailund et al., 2011; Hobolth and Jensen, 2014). Since explicitly enumerating all states and transitions is both tedious and error-prone we avoid this by letting the computer enumerate all states in a transition system. The states and transitions are defined as in our previous IIM paper (Mailund et al., 2012) but repeated below for completeness of this paper.

We represent lineages at a single nucleotide as sets. The sets $\{1\}$ and $\{2\}$ denote sequences 1 and 2 before they have found a common ancestor while $\{1, 2\}$ denote a lineage ancestral to both. We then model two neighbouring nucleotides as pairs of such states, so e.g. $(\{1, 2\}, \{1\})$ denote a lineage where the left nucleotide has found a common ancestor between sample 1 and 2 and is linked on the right to a nucleotide from the sequence 1, which has not found a common ancestor with sequence 2. To assign lineages to species, we pair them again, and let $[1, (l, r)]$ denote that lineage (l, r) is in population 1. A state in the CTMC corresponds to a set of such lineages assigned to species.

We define the following transitions of states:

$$\text{Coalescence: } \{[p_1, (l_1, r_1)]\} \cup \{[p_2, (l_2, r_2)]\} \cup \mathcal{S} \to \{[p_1, (l_1 \cup l_2, r_1 \cup r_2)]\} \cup \mathcal{S} \text{ if } p_1 = p_2$$

$$\text{Recombination: } \{[p, (l, r)]\} \cup \mathcal{S} \to \{[p, (l, \emptyset)]\} \cup \{[p, (\emptyset, r)]\} \cup \mathcal{S}$$

$$\text{Migration: } \{[p_1, (l, r)]\} \cup \mathcal{S} \to \{[p_2, (l, r)]\} \cup \mathcal{S} \text{ if } p_1 \neq p_2.$$

where $\mathcal{S}$ denotes the set of other lineages in the state.

When migration is not allowed in the epoch, as in the first epoch in the IIM model, we simply leave that transition out of the transition system when computing the state space. Fig. 6 shows and example of a run in this transition system specified for the IIM model.

As the initial state of the system, we use the state where sequence 1 is in population 1, sequence 2 is in population 2, and both sequences have their left and right nucleotides linked, $\{[1, (\{1\}, \{1\})], [2, (\{2\}, \{2\})]\}$, and we then compute a graph of all states reachable from this state through the transitions above, labelling each edge with the kind of transformation (coalescence, recombination or migration). From this state space we construct a rate matrix for the CTMC by first assigning a number to each state, and then setting rates (from our parameters θ) in the matrix in entries corresponding to edges in the graph. This is translated into an instantaneous rate matrix for the CTMC by setting all diagonal cells to minus the row sum. The result is a rate matrix for the CTMC, Q, such that $Q_{x,y}$ is the instantaneous rate of moving from state x to state y. From CTMC theory the probability of moving from state x to state y in time t is then given by $\left(e^{Qt}\right)_{x,y}$ where e^{Qt} is matrix exponentiation (Moler and Van Loan, 2003). For each time interval i in the CoalHMM we let Q^i denote the rate matrix of that interval. For intervals in the same epoch these will of course share the rate matrix but not necessarily the probability matrix for moving from one state to another when going through

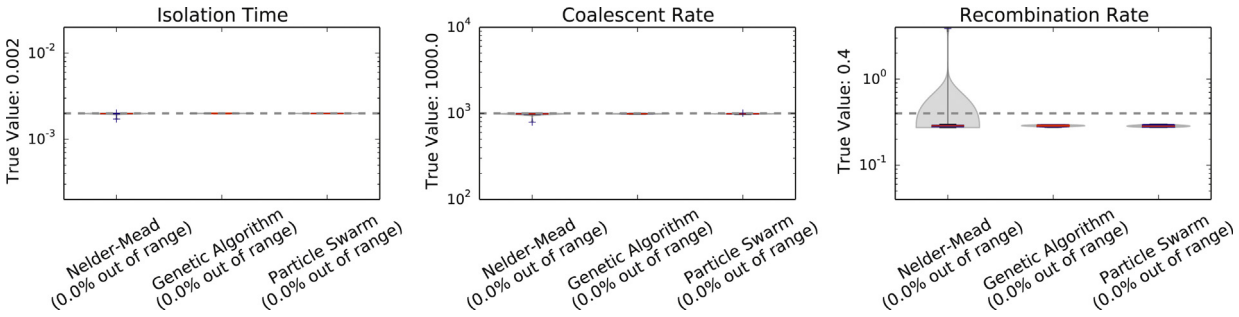


Fig. 7. Estimates for the isolation model from three optimisation algorithms. All three optimisers recover the simulated parameters, shown as dashed horizontal lines, reasonably well but with a higher variance for the Nelder–Mead optimiser. The estimates of the recombination rate are downwards biased, an effect we have previously observed and speculate is a consequence of the Markov assumption (Mailund et al., 2011).

the interval since the intervals do not necessarily have the same length.

2.1.2. Computing joint probabilities

To compute the $J_\theta(\ell, r)$ probabilities we use ideas from Mailund et al. (2011). Since coalescence times are discretised in time intervals we use $J_\theta(\ell = i, r = j)$ to mean that the left nucleotide coalesced in interval i and the right nucleotide in interval j. For this to be the case, and assuming interval i is earlier than interval j, neither left nor right nucleotide can have found a common ancestor between the two samples when entering interval i, the left but only the left must have when leaving interval i and this must remain the case until entering interval j, and when leaving interval j both left and right nucleotides must have found common ancestors for the two samples.

Regardless of the state space for the epoch CTMC we can always split the states into four non-overlapping (but possibly empty) sets: B: the "beginning states" where neither nucleotides have found common ancestors, L: the "left states" where the left nucleotides but not the right nucleotides have found a common ancestor, R: the "right states" where the right nucleotides but not the left nucleotides have found a common ancestor, and E: the "end states" where both nucleotides have found common ancestors. In terms of these sets we can reformulate the conditions for $J_\theta(\ell = i, r = j)$ as follows: when entering interval i we must be in a state in B but when leaving interval i we must be in a state in L and we must remain in L until we enter interval j and leave interval j in a state in E. It is straightforward to identify which of these sets each state belongs to and our framework does this automatically regardless of the epochs specification. We will use sub-scripts to indicate which interval and thus epoch the sets are associated with, so B_i, L_i, R_i and E_i are the sets for interval i.

Let $\mathcal{T}^i$ denote the probability transition matrix for changing states when going through interval i as computed from the matrix exponentiation of the rate matrix for the epoch of the interval $\mathcal{T}^i = e^{Q^i \Delta t_i}$ where Δt_i is the length of interval i. Since the state space in interval i and interval $i+1$ are not necessarily the same – if the intervals are from different epochs they might not be – we use the convention that the rows of $\mathcal{T}^i$ are indexed with the state space for interval i and the columns with the state space for interval $i+1$; the starting states for $\mathcal{T}^i$ are from the CTMC for interval i but the end states are from the CTMC for interval $i+1$. This makes it possible to always multiply together $\mathcal{T}$ matrices from adjacent intervals.

When two adjacent intervals are from the same epoch, then $\mathcal{T}^i$ is specified just from the matrix exponentiation, but when the interval i is from one epoch and $i+1$ from another, with a different state

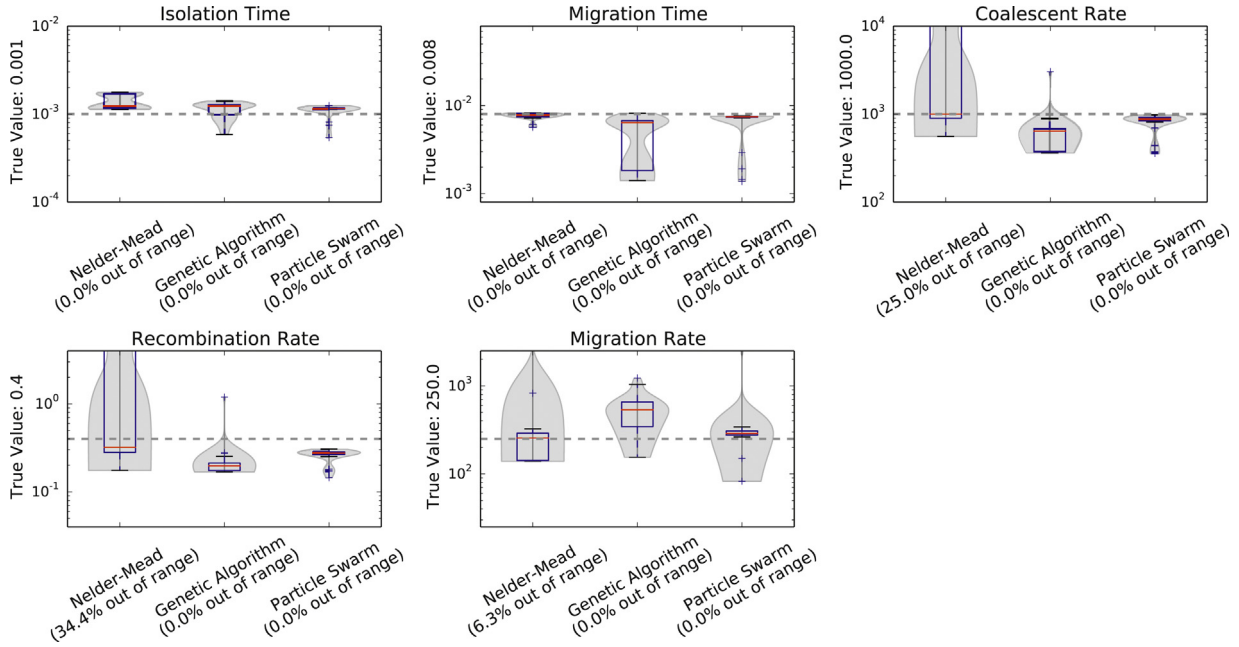


Fig. 8. Estimates for the isolation with initial migration model from all three optimisation algorithms. With more parameters to estimate, the variance in the estimates goes up as expected. The parameters are still reasonably well estimated for the two heuristic optimisers, especially for the particle swarm optimiser, but less so for the Nelder–Mead optimiser.

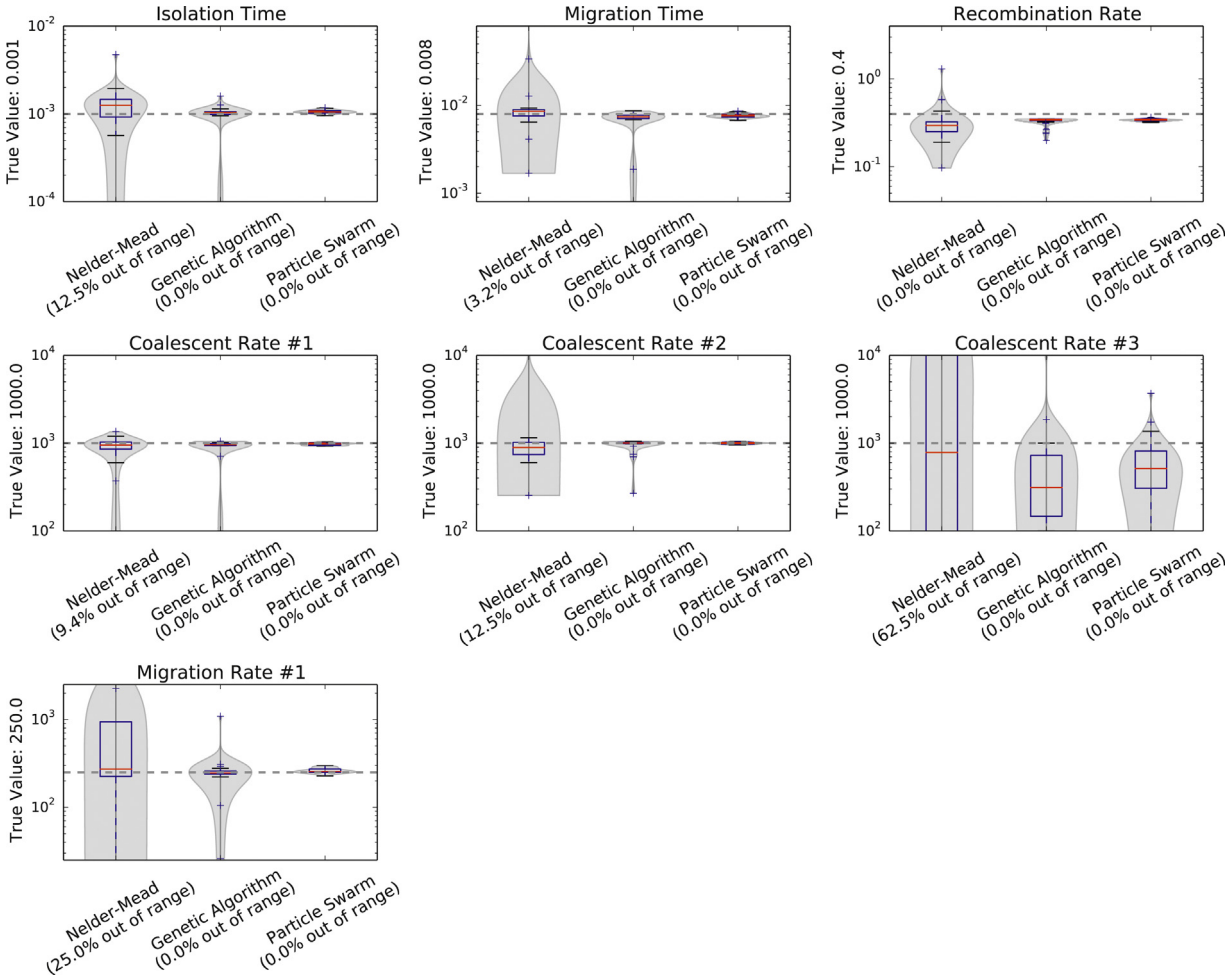


Fig. 9. Estimates for the isolation with initial migration model three epochs: One isolation epoch, one migration epoch and one ancestral epoch. This corresponds to the IIM model except that the coalescence rate is not assumed to be the same in all epochs. Again we see a failure for the Nelder–Mead to recover these parameters, and the last coalescence rate is not well estimated. The particle swarm optimiser performs the best among three optimisers.

space, a projection matrix is necessary. Such a matrix specifies how states in one CTMC correspond to states in another and by placing 1s in the relevant entries in a matrix P the $\mathcal{T}^i$ matrix is computed simply as $\mathcal{T}^i = e^{Q\Delta t_i} \cdot P$. In the case of the IIM model, moving from the isolation epoch to the migration epoch, lineages are mapped directly as $[p_i, (l, r)] \mapsto [p_i, (l, r)]$ since the lineages in the individual populations are the same; the state space is just larger when migration is allowed. For going from the migration epoch into the ancestral population both population p_1 and p_2 are simply mapped to the ancestral population p_A: $[p_i, (l, r)] \mapsto [p_A, (l, r)]$. We refer to the documentation in the framework for details on this and more complex projections.

Let $\mathcal{U}^i$ denote the transition matrix for going from time zero until the start point of interval i. This can be computed from the $\mathcal{U}^1$ matrix for getting from time zero to the first interval and $\mathcal{T}^i$ matrices for $j < i$:[1]

$$\mathcal{U}^i = \mathcal{U}^1 \prod_{j=1}^{i-1} \mathcal{T}^i.$$

If the first interval starts at time zero, $\mathcal{U}^1$ will just be the identity matrix. If it is not possible to coalesce for a certain time, as in the IIM model where the lineages are isolated until migration becomes possible, then the first interval starts later than time zero and $\mathcal{U}^1$ is used to address this. In the IIM $\mathcal{U}^1$ is computed by exponentiating the rate matrix from the isolation model multiplied with the isolation time.

Finally, let $\mathcal{B}^{i,j}$ for $i < j$ denote the probability matrix for going from the beginning of interval i to the end of interval j. This can be computed as

$$\mathcal{B}^{i,j} = \prod_{k=i}^{j} \mathcal{T}^k.$$

For computing $J_\theta(\ell = i, r = j)$ there are three cases: $i = j$, $i < j$ and $i > j$. All can be computed using the matrices defined above. Let ι denote the initial state for the coalescence system at time zero. For the IIM this would be the two lineages in separate populations. For $i = j$ we have

$$J_\theta(\ell = i, r = i) = \sum_{b \in B_i} \sum_{e \in E_{i+1}} \mathcal{U}^i_{\iota,b} \cdot \mathcal{T}^i_{b,e}.$$

with a special case for the last interval

$$J_\theta(\ell = n, r = n) = \sum_{b \in B_n} \mathcal{U}^n_{\iota,b}.$$

[1] In the actual implementation, intervals are indexed from zero and $\mathcal{U}^1$ is called $\mathcal{U}^0$ but we have chosen to index from 1 in the explanation of the algorithm here.

For $i<j$ we have

$$J_\theta(\ell = i, r = j) = \sum_{b\in E_i}\sum_{l\in L_{i+1}}\sum_{l'\in L_j}\sum_{e\in E_{j+1}} \mathcal{U}^i_{\iota,b}\cdot\mathcal{T}^i_{b,l}\cdot\mathcal{B}^{i+1,j-1}_{l,l'}\cdot\mathcal{T}^j_{l',e}.$$

with again a special case for the last interval

$$J_\theta(\ell = i, r = n) = \sum_{b\in E_i}\sum_{l\in L_{i+1}}\sum_{l'\in L_n} \mathcal{U}^i_{\iota,b}\cdot\mathcal{T}^i_{b,l}\cdot\mathcal{B}^{i+1,n-1}_{l,l'}.$$

Since the coalescence process is symmetric in left and right we can simply compute the cases for $j<i$ as $J_\theta(\ell = i, r = j) = J_\theta(\ell = j, r = i)$.

To specify a CoalHMM in our framework it is only necessary to specify the $\mathcal{T}^i$ and $\mathcal{U}^1$ matrices. Mostly this is a simple case of exponentiating rate matrices and specifying projections when moving between epochs.

2.2. Optimisation algorithms

We have enhanced our framework by incorporating two heuristic based optimisation algorithms. In both algorithms, the fitness of an individual solution is the negated log-likelihood values computed from the Forward algorithm from the CoalHMM.

2.2.1. Genetic algorithm optimiser

A GA optimiser in the CoalHMM framework initiates its first generation of individuals by uniformly selecting parameters within predetermined ranges. The GAs use population sizes of 100. Small

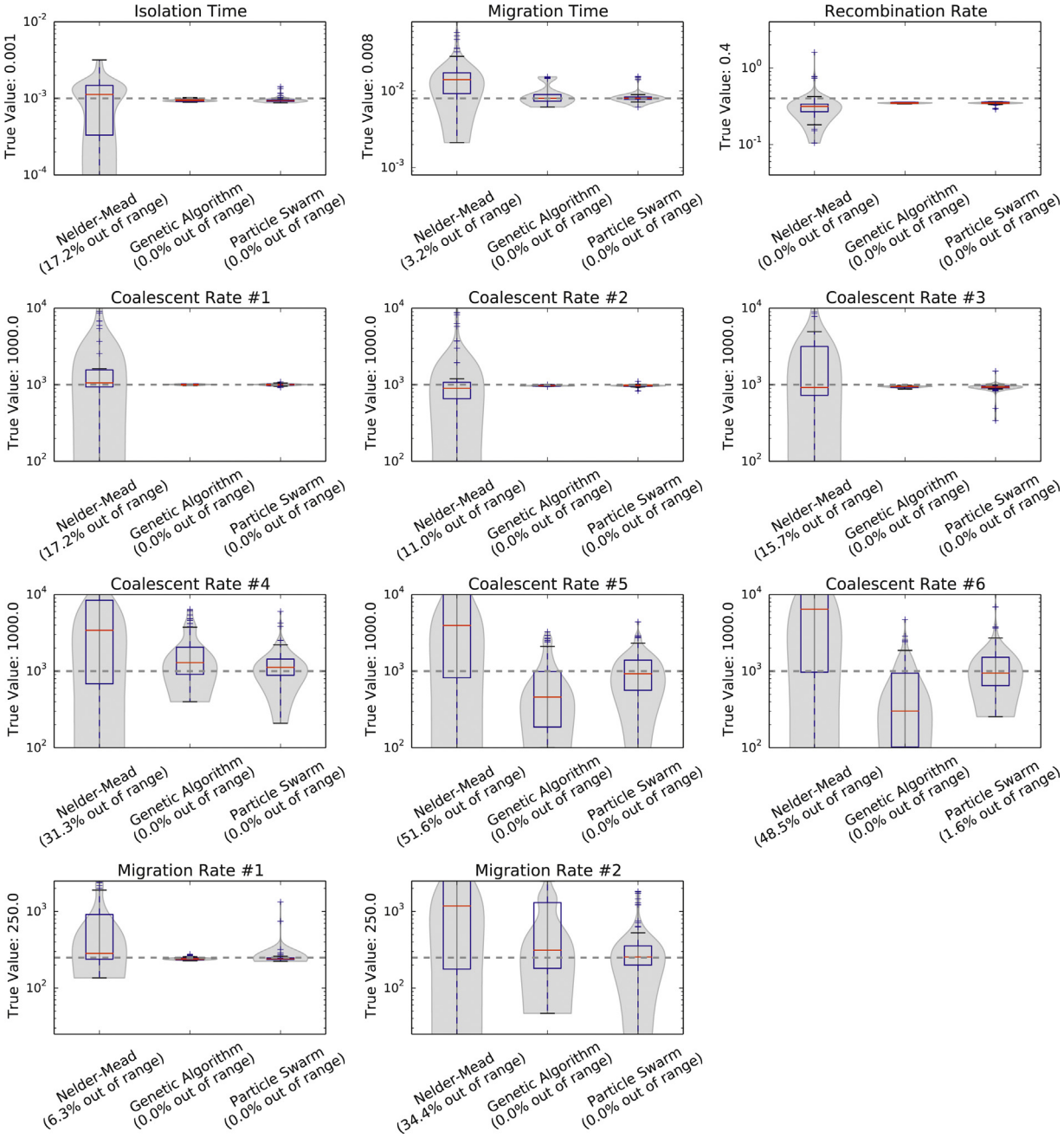


Fig. 10. Estimates for the isolation with initial migration model six epochs. Results are similar to the three epochs model. We see a failure to recover most of the parameters from the Nelder–Mead. We see some suboptimal results for the last two coalescence rates and second migration rate from the genetic algorithm. We see the best accuracy from the particle swarm.

populations lose genetic diversity quickly, while large populations result in better accuracy at the cost of increased running time. For our models, population sizes greater than 100 did not offer significant improvement. To form the breeding pool, we use Tournament Selection with a selection rate of 75% of the population size with tournament sizes of 10. We use a rank-based selection scheme because the lower and upper bounds of the fitness are unknown beforehand and differ from model to model; in order to use fitness proportionate selection, we would need an initial phase to estimate the fitness range. We then use One Point Crossover to

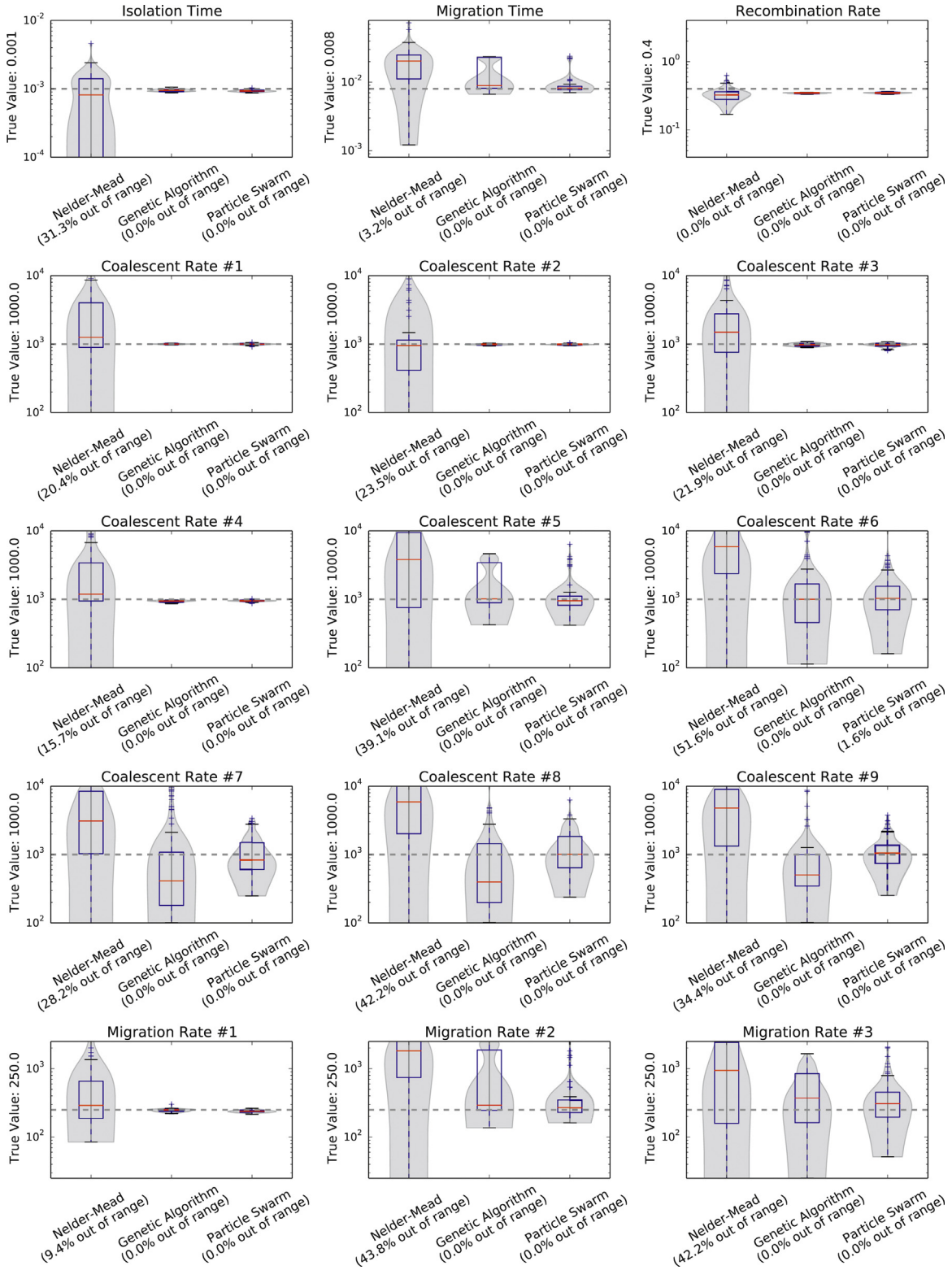


Fig. 11. Estimates for the isolation with initial migration model nine epochs. Results are similar to the three epochs and the six epochs models. We see a failure to recover most of the parameters from the Nelder–Mead. We see some suboptimal results for the late coalescence rates and migration rates from the genetic algorithm. We see the best accuracy from the particle swarm.

combine two breeders and generate individuals for the next generation. We chose this simple crossover scheme because other complex schemes failed to produce improved results. To help genetic diversity in a population, we apply point mutations at a rate of 15% and use Gaussian Mutation with $\mathcal{N}(\mu = 0, \sigma = 0.01)$. This relatively high point mutation rate is balanced by the relatively low σ; this configuration is suitable for our problem space, which consists of short chromosomes encoded with real numbers.

2.2.2. *Particle swarm optimiser*

Our framework also provides a PSO optimiser. Each model parameter corresponds to a dimension in the solution space. The

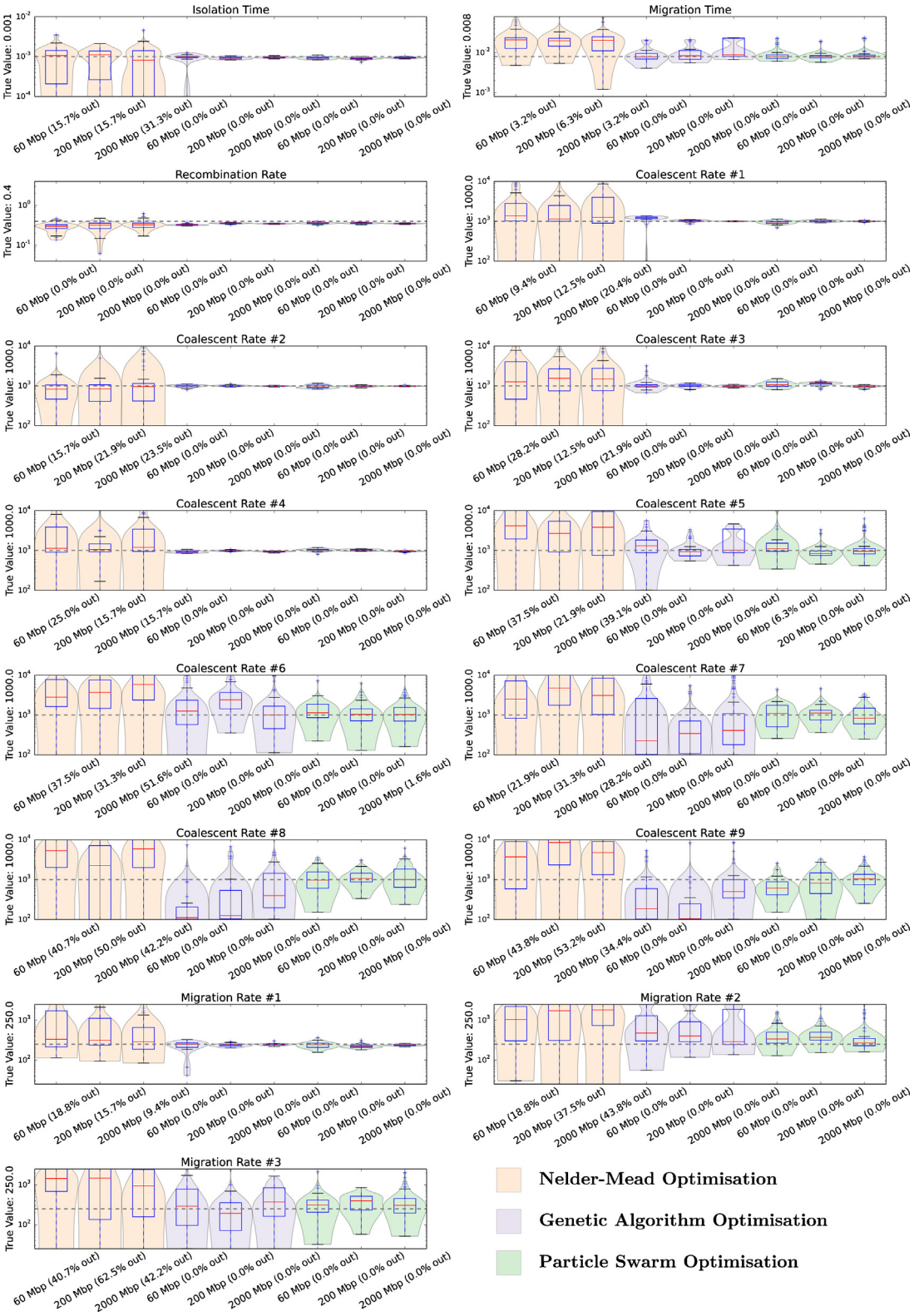


Fig. 12. Estimation accuracy with variable data sizes. For some of the parameters we see a reduction in the estimation variance with more data, but less than one would have hoped for a factor of more than three increase in alignment length.

optimiser initialises particle velocities from uniform random values within a range of 2% of the predetermined range for each parameter. During each iteration, we update the velocities of each particle using coefficients determined from trial and error. For the inertial coefficient, we use $\omega = 0.9$; i.e. a 10% decay in velocity if the particle is not affected by other forces. For the cognitive and social coefficients, we use $\phi_p = 0.3$ and $\phi_g = 0.1$, respectively. Larger values for ϕ had the tendency to accelerate the particles beyond acceptable ranges. Similar to our GA performance, we found population sizes greater than 100 did not significantly improve the performance, but they did dramatically increase the time required for the swarm to converge.

2.3. Simulated data

We use the program **ms** to generate ancestral recombination graphs under standard neutral evolutionary models with recombination, speciation, variable populations, migrations, etc. We then use the **seq-gen** program to produce sequence samples of length 10 Mbp. Using the phylogenetic trees simulated by **ms** as input, **seq-gen** evolves the sequences along the phylogeny.

3. Results and discussion

Below we illustrate how demographic inference can be done using our new CoalHMM framework by presenting a number of demographic models, from simple to more complex, and show how we can estimate parameters using our heuristic optimisation algorithms. All models are available as inference scripts in the framework.

3.1. Isolation model

The simplest model we will consider is the clean isolation model from Mailund et al. (2011). The model has three parameters: the split time where the ancestral population is split into two independent populations, a coalescence rate that is the same for the ancestral population and the two descedent populations, and a recombination rate.

Fig. 7 shows the estimation results for all three optimisers, operating on simulated sequences consisting of 1000 Mbp. The range on the y-axis corresponds to the range of possible values for the GA and PSO optimisers for each parameter. The Nelder–Mead optimiser is not limited to these ranges and the percentage of estimates that falls outside of the range is written below the x-axis. For this simplest model all three optimisers recover the simulated parameters, shown as dashed horizontal lines, reasonably well but with a higher variance for the Nelder–Mead optimiser. The estimates of the recombination rate are downwards biased, an effect we have previously observed and speculate is a consequence of the Markov assumption (see Mailund et al. (2011) Supplemental Text 1 and Fig. 4S in the same text).

3.2. Isolation with initial migration model

The next model we consider is the IIM model from Mailund et al. (2012) that we have used as an example in Section 2. This model has five parameters: The time period where the two populations are completely isolated, the time period where migration is ongoing, a shared coalescence rate for all populations, a migration rate for the migration epoch, and a recombination rate.

Fig. 8 shows the estimation results for this model for our three optimisers, operating on simulated sequences consisting of 1000 Mbp. With more parameters to estimate, the variance in the estimates goes up as expected. The parameters are still reasonably well estimated for the two heuristic optimisers, especially for the Particle Swarm optimiser, but less so for the Nelder–Mead optimiser. We still see a bias in the estimates of the recombination rate, but now also an slight upwards bias in the estimates of the split time (the time where gene flow finally ends). This was not obvious in our previous results (Mailund et al., 2012) because of the large variance in the optimiser we used there.

3.3. Multi-epochs isolation with initial migration models

For a more complex model we consider an extension of the IIM model not previously described. This model allows multiple epochs within the isolation period, the migration period and the ancestral population. Both coalescence rates and migration rates can vary freely between epochs. In our experiments we always have the same number of isolation, migration and ancestral epochs. The parameters are the end of gene flow (split time), the beginning of gene flow (migration time), one coalescence rate for each of the isolation, migration and ancestral epochs, a symmetric migration rate for each migration epoch and the recombination rate.

The first coalescence rate would be impossible to estimate with just a pairwise alignment of one sequence from each population since we observe no coalescence events there and so would have no hidden Markov model states in that epoch (Mailund et al., 2011). We solve this by constructing a composite likelihood from three different hidden Markov models: one where our pairwise alignment is from two samples from the first population, one where the alignment is from the second alignment and one with one sample from each population. These are all constructed with the same CTMCs and only differ in the initial state, ι, used for calculating the joint genealogy probability. We run all three models in parallel with the same parameters and add the log-likelihoods together to get a combined likelihood.

Fig. 9 shows results for a model with three epochs, operating on simulated sequences consisting of 2000 Mbp. This corresponds to the IIM model except that there are now three coalescence rates instead of one. Again we see a failure for the Nelder–Mead to recover these parameters. The the last coalescence rate is not as well estimated. The particle swarm optimiser performs the best among three optimisers.

Figs. 10 and 11 show results for models with six and nine epochs, respectively, operating on simulated sequences consisting of 2000 Mbp. Results are similar to the three epochs model. We see a failure to recover most of the parameters from the Nelder–Mead and some suboptimal results for the last coalescence rates and migration rates from the Genetic Algorithm. We see a better accuracy from the Particle Swarm. Even for the nine epochs model the Particle Swarm estimates reasonably well. The earlier migration rates are estimated better than last migration rate in both the six epochs model and the nine epochs model.

Fig. 12 shows the effect of increasing the data size from 60 Mbp to 2000 Gbp. For some of the parameters we see a reduction in the estimation variance with more data, but less than one would have hoped for a factor of more than three increase in alignment length.

4. Conclusions and future work

We have described a new framework for constructing coalescence hidden Markov models for demographic inference and showed that using heuristic optimisation algorithms we can accurately estimate parameters in a number of complex models. Using our framework it is relatively easy to construct CoalHMMs for even rather complex demographics, but a limiting factor is the accurate parameter estimation. We have shown that the Nelder-Mead algorithm we have previously used for estimation fails somewhat when the number of parameters increases and that the heuristic

optimisers do a better job. Good optimisation algorithms is still a topic for future work.

In this paper we have focused on maximum likelihood estimates of each parameter but not considered estimating error bars for the estimates. These can be computed using bootstrap or jackknife approaches but this comes at a cost in running time. Here, as well, future work is needed.

Being able to work with larger sample sizes than four could potentially improve the accuracy of parameter estimates as shown in the MSMC (Schiffels and Durbin, 2014) model compared to the PSMC model (Li and Durbin, 2011), and some of the approaches we take in our framework generalises to more samples. The construction of CTMCs for more samples is immediately possible as we have shown in previous work (Mailund et al., 2012), although this approach will only scale to a small number of samples due to the problem of dealing with very large state spaces for the CTMCs. Automatically combining CTMCs for such cases in a similar way to what we have presented here is more complex still and requires more work.

Despite these limitations we believe that our new framework will enable more complex models to be explored using the CoalHMM methodology and that the ideas underlying its design can be used for improved frameworks in the future.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TM implemented the CoalHMM framework. JC implemented the optimisation algorithms. Both authors designed the experiments and analysed the results. JC executed the experiments. Both authors drafted the manuscript.

Acknowledgements

This research was funded by the Danish Council of Independent Research Sapere Aude Grant 12-125062.

References

Baker, J.E.,1987. Reducing bias and inefficiency in the selection algorithm. In: Proceedings of the Second International Conference on Genetic Algorithms on Genetic Algorithms and Their Application. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, pp. 14–21 http://portal.acm.org/citation.cfm?id=42512.42515

Baudin, M., 2009. Nelder Mead User's Manual.

Chen, G.K., Marjoram, P., Wall, J.D., 2009. Fast and flexible simulation of dna sequence data. Genome Res. 19 (1), 136–142.

Deb, K., 2001. Multi-objective optimization using evolutionary algorithms.

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

Eberhart, R., Kennedy, J., 1995. A new optimizer using particle swarm theory. In: Proceedings of the Sixth International Symposium on Micro Machine and Human Science, 1995, pp. 39–43, http://dx.doi.org/10.1109/MHS.1995.494215.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17 (6), 368–376, http://dx.doi.org/10.1007/BF01734359.

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization and Machine Learning, 1st ed. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Hein, J., Schierup, M.H., Wiuf, C., 2005. Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory. Oxford University Press, USA.

Hobolth, A., Jensen, J.L., 2014. Markovian approximation to the finite loci coalescent with recombination along multiple sequences. Theor. Popul. Biol. 98, 48–58.

Hobolth, A., Christensen, O.F., Mailund, T., Schierup, M.H., 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS Genet. 3 (2), e7.

Holland, J.H., 1992. Genetic algorithms. Sci. Am. 267 (1), 66–72.

Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. In: Munro, M.N. (Ed.), Mammalian Protein metabolism, Vol. III. Academic Press, New York, pp. 21–132.

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

Locke, D.P., Hillier, L.W., Warren, W.C., Worley, K.C., Nazareth, L.V., Muzny, D.M., Yang, S.-P., Wang, Z., Chinwalla, A.T., Minx, P., Mitreva, M., Cook, L., Delehaunty, K.D., Fronick, C., Schmidt, H., Fulton, L.A., Fulton, R.S., Nelson, J.O., Magrini, V., Pohl, C., Graves, T.A., Markovic, C., Cree, A., Dinh, H.H., Hume, J., Kovar, C.L., Fowler, G.R., Lunter, G., Meader, S., Heger, A., Ponting, C.P., Marques-Bonet, T., Alkan, C., Chen, L., Cheng, Z., Kidd, J.M., Eichler, E.E., White, S., Searle, S., Vilella, A.J., Chen, Y., Flicek, P., Ma, J., Raney, B.J., Suh, B.B., Burhans, R., Herrero, J., Haussler, D., Faria, R., Fernando, O., Darré, F., Farré, D., Gazave, E., Oliva, M., Navarro, A., Roberto, R., Capozzi, O., Archidiacono, N., della Valle, G., Purgato, S., Rocchi, M., Konkel, M.K., Walker, J.A., Ullmer, B., Batzer, M.A., Smit, A.F.A., Hubley, R., Casola, C., Schrider, D.R., Hahn, M.W., Quesada, V., Puente, X.S., Ordo nez, G.R., López-Otín, C., Vinar, T., Brejova, B., Ratan, A., Harris, R.S., Miller, W., Kosiol, C., Lawson, H.A., Taliwal, V., Martins, A.L., Siepel, A., RoyChoudhury, A., Ma, X., Degenhardt, J., Bustamante, C.D., Gutenkunst, R.N., Mailund, T., Dutheil, J.Y., Hobolth, A., Schierup, M.H., Ryder, O.A., Yoshinaga, Y., de Jong, P.J., Weinstock, G.M., Rogers, J., Mardis, E.R., Gibbs, R.A., Wilson, R.K., 2011. Comparative and demographic analysis of orang-utan genomes. Nature 469 (7331), 529–533.

Mailund, T., Dutheil, J.Y., Hobolth, A., Lunter, G., Schierup, M.H., 2011. Estimating divergence time and ancestral effective population size of Bornean and Sumatran orangutan subspecies using a coalescent hidden Markov model. PLoS Genet. 7 (3), e1001319.

Mailund, T., Halager, A.E., Westergaard, M., Dutheil, J.Y., Munch, K., Andersen, L.N., Lunter, G., Prüfer, K., Scally, A., Hobolth, A., Schierup, M.H., 2012. A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. PLoS Genet. 8 (12), e1003125.

Mailund, T., Halager, A., Westergaard, M., 2012. Using colored Petri nets to construct coalescent hidden Markov models: Automatic translation from demographic specifications to efficient inference methods. In: Haddad, S., Pomello, L. (Eds.), Application and Theory of Petri Nets. Bioinformatics Research Center, Aarhus University/Springer, Denmark/Berlin, Heidelberg, pp. 32–50.

Marjoram, P., Wall, J.D., 2006. Fast "coalescent" simulation. BMC Genet. 7, 16.

McVean, G., Cardin, N.J., 2005. Approximating the coalescent with recombination, Philosophical transactions of the Royal Society of London. Series B. Biol. Sci. 360 (1459), 1387–1393.

Michalewicz, Z., 1996. Genetic algorithms -f data structures = evolution programs.

Miller, B.L., Miller, B.L., Goldberg, D.E., Goldberg, D.E., 1995. Genetic algorithms, tournament selection, and the effects of noise. Complex Syst. 9, 193–212.

Miller, W., Schuster, S.C., Welch, A.J., Ratan, A., Bedoya-Reina, O.C., Zhao, F., Kim, H.L., Burhans, R.C., Drautz, D.I., Wittekindt, N.E., Tomsho, L.P., Ibarra-Laclette, E., Herrera-Estrella, L., Peacock, E., Farley, S., Sage, G.K., Rode, K., Obbard, M., Montiel, R., Bachmann, L., Ingolfsson, Ó., Aars, J., Mailund, T., Wiig, Ø., Talbot, S.L., Lindqvist, C., 2012. Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. PNAS 109 (36), E2382–E2390.

Moler, C., Van Loan, C., 2003. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM Rev. 45 (1), 3–49.

Munch, K., Mailund, T., Dutheil, J.Y., Schierup, M.H., 2014. A fine-scale recombination map of the human-chimpanzee ancestor reveals faster change in humans than in chimpanzees and a strong impact of GC-biased gene conversion. Genome Res. 24 (3), 467–474, doi:10.1101/gr.158469.113.

Nelder, J.A., Mead, R., 1965. A simplex method for function minimization. Comput. J. 7 (4), 308–313, http://dx.doi.org/10.1093/comjnl/7.4.308 http://comjnl.oxfordjournals.org/content/7/4/308.abstract

Prüfer, K., Munch, K., Hellmann, I., Akagi, K., Miller, J.R., Walenz, B., Koren, S., Sutton, G., Kodira, C., Winer, R., Knight, J.R., Mullikin, J.C., Meader, S.J., Ponting, C.P., Lunter, G., Higashino, S., Hobolth, A., Dutheil, J., Karakoç, E., Alkan, C., Sajjadian, S., Catacchio, C.R., Ventura, M., Marquès-Bonet, T., Eichler, E.E., André, C., Atencia, R., Mugisha, L., Junhold, J., Patterson, N., Siebauer, M., Good, J.M., Fischer, A., Ptak, S.E., Lachmann, M., Symer, D.E., Mailund, T., Schierup, M.H., Andrés, A.M., Kelso, J., Pääbo, S., 2012. The bonobo genome compared with the chimpanzee and human genomes. Nature 486 (7404), 527–531.

Prado-Martinez, J., Sudmant, P.H., Kidd, H., Li, Jeffrey Mand, Kelley, J.L., Lorente-Galdos, B., Veeramah, K.R., Woerner, A.E., O'Connor, T.D., Santpere, G., Cagan, A., Theunert, C., Casals, F., Laayouni, H., Munch, K., Hobolth, A., Halager, A.E., Malig, M., Hernandez-Rodriguez, J., Hernando-Herraez, I., Prüfer, K., Pybus, M., Johnstone, L., Lachmann, M., Alkan, C., Twigg, D., Petit, N., Baker, C., Hormozdiari, F., Fernández-Callejo, M., Dabad, M., Wilson, M.L., Stevison, L., Camprubí, C., Carvalho, T., Ruiz-Herrera, A., Vives, L., Melé, M., Abello, T., Kondova, I., Bontrop, R.E., Pusey, A., Lankester, F., Kiyang, J.A., Bergl, R.A., Lonsdorf, E., Myers, S., Ventura, M., Gagneux, P., Comas, D., Siegismund, H., Blanc, J., Agueda-Calpena, L., Gut, M., Fulton, L., Tishkoff, S.A., Mullikin, J.C., Wilson, R.K., Gut, I.G., Gonder, M.K., Ryder, O.A., Hahn, B.H., Navarro, A., Akey, J.M., Bertranpetit, J., Reich, D., Mailund, T., Schierup, M.H., Hvilsom, C., Andrés, A.M., Wall, J.D., Bustamante, C.D., Hammer, M.F., Eichler, E.E., Marquès-Bonet, T., 2013. Great ape genetic diversity and population history. Nature 499 (7459), 471–475.

Rabiner, L., 1989. A tutorial on hidden markov models and selected applications in speech recognition. Proc. IEEE, 257–286, http://dx.doi.org/10.1109/5.18626.

Sand, A., Kristiansen, M., Pedersen, C.N., Mailund, T., 2013. Ziphmmlib: a highly optimised hmm library exploiting repetitions in the input to speed up the forward algorithm. BMC Bioinform. 14, 339, http://dx.doi.org/10.1186/1471-2105-14-339.

Scally, A., Dutheil, J.Y., Hillier, L.W., Jordan, G.E., Goodhead, I., Herrero, J., Hobolth, A., Lappalainen, T., Mailund, T., Marques-Bonet, T., McCarthy, S., Montgomery, S.H., Schwalie, P.C., Tang, Y.A., Ward, M.C., Xue, Y., Yngvadóttir, B., Alkan, C., Andersen, L.N., Ayub, Q., Ball, E.V., Beal, K., Bradley, B.J., Chen, Y., Clee, C.M., Fitzgerald,

S., Graves, T.A., Gu, Y., Heath, P., Heger, A., Karakoc, E., Kolb-Kokocinski, A., Laird, G.K., Lunter, G., Meader, S., Mort, M., Mullikin, J.C., Munch, K., O'Connor, T.D., Phillips, A.D., Prado-Martinez, J., Rogers, A.S., Sajjadian, S., Schmidt, D., Shaw, K., Simpson, J.T., Stenson, P.D., Turner, D.J., Vigilant, L., Vilella, A.J., Whitener, W., Zhu, B., Cooper, D.N., de Jong, P., Dermitzakis, E.T., Eichler, E.E., Flicek, P., Goldman, N., Mundy, N.I., Ning, Z., Odom, D.T., Ponting, C.P., Quail, M.A., Ryder, O.A., Searle, S.M., Warren, W.C., Wilson, R.K., Schierup, M.H., Rogers, J., Tyler-Smith, C., Durbin, R., 2012. Insights into hominid evolution from the gorilla genome sequence. Nature 483 (7388), 169–175.

Schiffels, S., Durbin, R., 2014. Inferring human population size and separation history from multiple genome sequences. Nat. Genet. 46 (8), 919–925.

Sheehan, S., Harris, K., Song, Y.S., 2013. Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. Genetics 194 (3), 647–662.

Shi, Y., Eberhart, R., 1998. A modified particle swarm optimizer. In: IEEE World Congress on Computational Intelligence. The 1998 IEEE International Conference on Evolutionary Computation Proceedings, pp. 69–73, http://dx.doi.org/10.1109/ICEC.1998.699146.

Simonsen, K., Churchill, G., 1997. A Markov chain model of coalescence with recombination. Theor. Popul. Biol. 52 (1), 43–59.

Slatkin, M., Pollack, J.L., 2006. The concordance of gene trees and species trees at two linked loci. Genetics 172 (3), 1979–1984.

Steinrücken, M., Paul, J.S., Song, Y.S., 2013. A sequentially Markov conditional sampling distribution for structured populations with migration and recombination. Theor. Popul. Biol. 87, 51–61.

Syswerda, G., 1989. Uniform crossover in genetic algorithms. In: Schaffer, J.D. (Ed.), Proceedings of the Third International Conference on Genetic Algorithms. Morgan Kaufmann, pp. 2–9.

CoalHMM method #2

This manuscript summarizes my work on CoalHMM's admixture modeling. I implemented admixture CoalHMM to infer historical admixture events, and I constructed multiple admixing demographics. Admixture CoalHMM not only learns the admixture time but also the proportions of gene flow. Also in this paper, I present a range of simulation evaluations, and I demonstrate good inference accuracy under different demographics. I also show the effect of admixture CoalHMM used on wrongly modeled demographics. Together, I present admixture CoalHMM as a new tool to study historical admixture events.

A coalescent hidden Markov model for inferring admixture relationships

Jade Cheng[1], Thomas Mailund[1,*]
1 Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark
∗ **E-mail: mailund@birc.au.dk**

Introduction

Admixture and hybridisation events form new populations or species by mixing two or more source populations. Recent studies have shown these events to be common in various species, including bears [1–4], equids [5, 6], and hominins [7–11]. To understand the genetic relationship between populations and closely related species, gene flow and admixture events cannot be ignored, yet the tools we have for exploring and validating relationships with admixture and gene-flow are still limited.

Background

A wealth of methods exists for inferring phylogenetic tree relationships, and in recent years several approaches have been developed for testing the presence of gene flow, inferring the rate of gene flow between populations or inferring the admixture proportions between source populations and admixed populations, and dating admixture events. For a recent review we refer to Sousa & Hey [12].

These methods vary in the data they use and the parameters they estimate. A number of methods use the drift between populations—the random changes in allele frequencies that occur over time as a population evolves—to infer the relationship between populations and to detect when simple tree relationships cannot explain the data and gene-flow must have been present. The D statistics and the f_3 and f_4 statistics, originally developed to detect archaic admixture in human populations [13, 14] exploit correlations between allele frequencies in three or four populations to detect deviations from tree relationships. From five populations, the ratio of two f_4 statistics can infer the admixture proportions of an admixed population—when the topology of the population fits a specific pattern.

The Treemix method [15] uses correlations in allele frequencies to fit data to trees extended with gene-flow edges, and the $\partial a \partial i$ method [16] uses joint allele frequency patterns to infer parameters of a specified demographic scenario and test goodness-of-fit of the scenario. Similarly, the qpGraph method [14] fits drift parameters in an admixture graph to test the goodness-of-fit for a proposed relationship between populations.

Drift-based methods measures divergence in terms of changes in allele frequencies, a time measure that depends on the effective population size as well as time. Effective population sizes will vary between independent branches in an admixture graph, and time is thus not moving at the same speed in different parts of the graph. Because of this they cannot directly be used to infer the timing of events such as population splits and admixture events. For dating admixture events, linkage disequilibrium (LD) can be used. Admixture between two diverged populations introduces LD in the admixed population and this LD breaks down over time. By comparing the LD in an admixed population with that of its source populations it is possible to infer the number of generations since the admixture event, a property exploited by the Rolloff [14] and ALDER [17] methods.

The patterns of mutations between genomes from two diverged populations also holds information about the demographics that lead to the two populations, and the identity-by-state method of Harris & Nielsen [18] uses this to fit the distribution of distances between variable sites to a demographic model. Since this approach is based on substitutions between two genomes, the time between events can be worked out, assuming a molecular clock, and with a calibrated molecular clock the events can be dated.

Also assuming a molecular clock, the number of mutations between two haplotypes that have not undergone recombination is proportional to the time since their divergence. From a large number of

haplotypes, the coalescence density between two or more populations can be worked out, and since this density depends on the demographics, the demographics can be inferred. The so-called isolation-with-migration model [19,20] exploits this by integrating over genealogies using the Markov-chain Monte Carlo method. When only small sample sizes are considered—and the large number of haplotypes are obtained by using haplotypes spread over entire genomes—the coalescence densities can be analytically derived and fitted to the data [21–23]. More sophisticated sampling methods based on this, such as GPhoCS [24], exploits this to infer rates of gene-flow within a multi-population topology.

To exploit whole-genome data recombination must be explicitly modelled. The computational complexity of modelling genealogies with recombination prohibits this unless approximations are used. A framework based on assuming a Markov relationship between genealogies along a sequence alignment [25, 26] has enabled a new class of inference methods: coalescence hidden Markov models or CoalHMMs [27–33]. Once such methods, the MSMC [34]—while not explicitly modelling gene flow—reveals gene flow patterns from comparing the rate of coalescences within and between populations over time. The diCal method [30] combines a CoalHMM with a sequential sampling approach to infer gene-flow between diverging populations, and in earlier work we have developed a CoalHMM for an isolation-with-initial-migration model [35] to infer the presence of gene flow and estimate the extend in time this gene-flow took place after an initial population split.

In this paper we develop a coalescence hidden Markov model for inferring parameters for admixture events. By tracing lineages in an admixed population and its source populations back in time, and estimating the coalescence times of those lineages, we can infer the split time between the source populations, the time of admixture, and the admixture proportions. We validate the method using simulated sequences and apply the method to a number of polar bear and brown bear genomes to infer the complex population history of these bear species.

Results

Modelling admixture as a coalescent hidden Markov model

Samples from a species undergoing recombination are genealogically connected through an ancestral recombination graph capturing where—along the genome and back in time—the samples experienced recombination events and where they found their most recent common ancestors. This graph naturally defines a set of local genealogies: for each point along the genome the samples are connected in a tree genealogy; between recombination points the tree genealogy remains the same and at recombination points the trees before and after the recombination point are potentially different.

Coalescent hidden Markov models approximate the process in which local genealogies change along the genome. The genealogies are unobserved, but can be inferred from a sequence alignment of the samples, and are assumed to change along the genome as a Markov process. To use standard hidden Markov model inference algorithms the state space of this process must be finite, so time is discretised to get a finite set of possible tree genealogies. The size of the state space, however, is still at least as large as the number of possible tree topologies, which grows super-exponentially in the number of samples, so explicitly modelling all tree topologies is only feasible with small sample sizes. The smallest meaningful sample size is of course pairs of samples, where the local genealogy is equivalent to the local time to the most recent ancestor, T_{MRCA}. The T_{MRCA} of a pair of samples varies along the genome, and both the distribution of values it takes and patterns of changes along the genome are determined by the demographic history of the the samples.

We have constructed a coalescent hidden Markov model that specifies the changes of the T_{MRCA} along a genome between a pair of samples as a function of population divergence times and admixture proportions between populations. The model considers three populations: one admixed population and two populations related to the populations that created the admixed population. For any pair of samples,

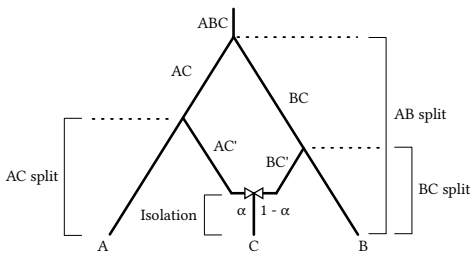


Figure 1. Admixture graph. The graph shows the relationship between three extant populations, A, B, and C. Population C originated as a mixture of two populations, AC' and BC', related to ancestral populations of A and B—AC and BC, respectively—that in the past split from a common ancestral population ABC.

either taken from the same population or from two different populations, we use a continuous time Markov model to compute how the T_{MRCA} will vary along the genome. Given samples from each population, each pair will be informative about a subset of the parameters in the admixture graph and we combine all pairs of samples in a composite likelihood function to estimate parameters.

Consider Fig. 1. The figure shows the relationship between three extant populations, A, B, and C, where C is admixed between two populations related to populations A and B. Lineages from the three populations, traced back in time, goes through different ancestral populations. Lineages from A will go through A then AC—the population ancestral to both A and C—and then ABC—the population ancestral to all populations. Similarly, lineages from population B will go through B, then AC, and then ABC. Lineages from population C, however, can take two different paths. They will first go through C and at the admixture time they will either go left, with probability α, and go through AC' and then AC until finally going to ABC, or they will turn right, with probability $1-\alpha$, and go through BC', then AB, and finally ABC.

Two lineages from population A can coalesce in any of populations A, AC, and ABC, at a rate that is inversely proportional to the effective population size in those populations. Similarly for two lineages from population B that can coalesce in populations B, BC, and ABC. Two lineages from population C can coalesce within populations C, AC', AC, and ABC with probability α^2; within populations C, BC', BC and ABC with probability $(1-\alpha)^2$, or within only C and ABC with probability $2\alpha(1-\alpha)$. For one lineage from population A and one lineage from population C these can coalesce within populations AC and ABC with probability α and within only population ABC with probability $(1-\alpha)$, and similarly for one lineage from population B and one from population C these can coalesce within populations AC and ABC with probability $(1-\alpha)$ or only within population ABC with probability α.

We model how the T_{MRCA} changes along the genome by specifying a continuous time Markov model that tracks two samples back in time through the admixture graph. For each sample we keep track of two neighbouring nucleotides and allow the left and right nucleotide of a sample to recombine apart, splitting a lineage into two, or coalesce, merging two lineages into one. When lineages from two samples coalesce one or two of the nucleotides will find their most recent common ancestor. By summing over all possible ancestries of the two samples in this way we obtain a joint probability of when the left and right nucleotides of our samples find their most recent common ancestor—the joint probability of the left and right T_{MRCA}—from which the transition probabilities of the coalescent hidden Markov models can be obtained.

Depending on which populations the two samples are from we get different transition matrices for changes in T_{MRCA}. We thus get different hidden Markov models for each choice of samples. All are, however, determined by the same set of parameters. By multiplying the likelihood we get for each hidden

Markov model for each choice of pairs of samples we get a composite likelihood we can use to estimate the parameters of the admixture graph.

The model parameters are: the time of the admixture, τ_{admix}, the divergence time between population A and C, τ_{AC}, the divergence time between population B and C, τ_{BC}, the divergence time between population A and B, τ_{ABC}, the admixture proportion, α, and the recombination and coalescence rates, R and C, that we for simplicity assume are constant across the admixture graph and along the sequence.

A composite likelihood combination of pairwise hidden Markov models

We apply the composite likelihood approach to deal with more than two samples. The more samples we have from each population the more HMMs we can construct and incorporate into the admixture model. We refer to Supplemental Text S1 for details on how each HMM is constructed and which are combined for each model.

Table 1 summarises the admixture models we use in the simulation study. The models differ depending on the samples we have available. We use Model #1 when we have access to only the admixed population. In this case we make inference from two chromosomes within population C and have a single hidden Markov model.

When we have genetic data for the admixed population and one of the source populations, we can apply Model #2. Here we assume we have two chromosomes from each population and we construct two hidden Markov models running on two samples from the same population – the first is the same as Model #1 and the second is similar except the two samples are from population A rather than C – and a third model running one one genome from either population.

When we have data from all three populations, we would use the full model. With data from all three populations we consider three different ways of exploiting this data: For Model #3.1, we have only one sample per population. In this case we can construct three HMMs for pairwise alignments of the three samples, one from each population. For Model #3.2, we have two samples per population, and we construct one HMM for each of the six types of HMMs. For Model #3.3 we construct fifteen HMMs for all pairwise alignments of six samples, two from each population.

Estimation accuracy

We simulated data using the program fastsimcoal2 [36, 37]. We first simulated two different scenarios, differing in the time since the population divergence and admixture events: An "Ancient" scenario ($\tau_{ABC} = 0.002$, $\tau_{AC} = 0.0016$, $\tau_{BC} = 0.0012$, $\tau_{\text{admix}} = 0.0002$) and a "Recent" ($\tau_{ABC} = 0.0004$, $\tau_{AC} = 0.0002$, $\tau_{BC} = 0.00008$, $\tau_{\text{admix}} = 0.00002$), both with coalescence rate $C = 3125$ and recombination rate $R = 0.5$. For both scenarios we varied the admixture proportions α from 10% to 90% in steps of 10%.

The estimated parameters for the two scenarios are shown in Fig. 2 where (A) show the estimates of the timing of events and the two rates and (B) shows the estimates of admixture proportions. The solid gray lines show the simulated parameter values and for the admixture proportions the solid lines shows the simulated α values while the dashed lines show $1 - \alpha$. Since Model #1 cannot estimate τ_{AC} and τ_{BC} and Model #2 cannot estimate τ_{BC}, estimates are naturally not shown for these parameter for those two models.

We generally recover the time parameters well, although τ_{AC} and τ_{BC} are slightly underestimated and τ_{ABC} is slightly overestimated for Model #1 and Model #2. We recover the coalescence rate well except for Model #1 in the Recent setup and we underestimate the recombination rate R. The latter is a general problem with our coalescent hidden Markov model approach also seen in earlier models [35,38]. Model #3.1, which has fewer hidden Markov models in its composite likelihood and has a larger variance in its estimates than the other models, especially when estimating the admixture time.

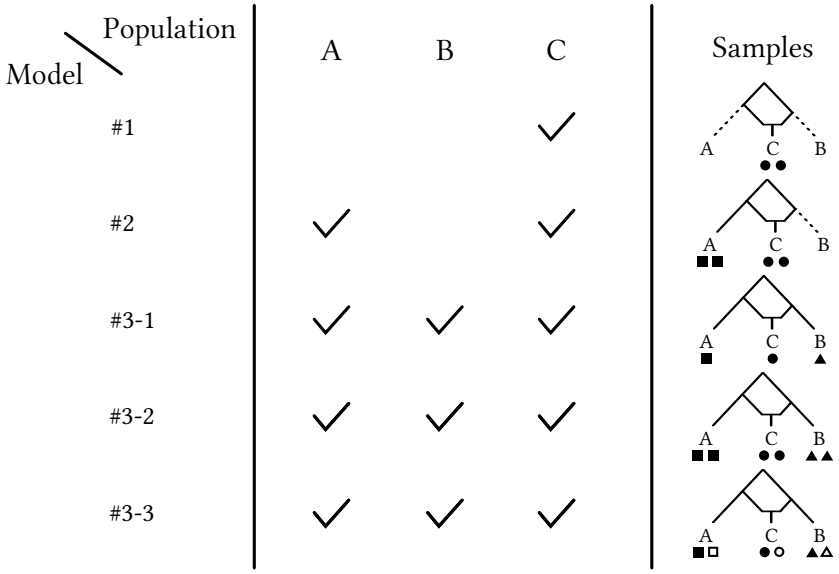


Table 1. Admixture models. The table gives an overview of the admixture models used in our simulation study. These models differ depending on the availability of data. We use Model #1 when we have access to only C. When we have genetic data for C and one of the source populations, we can apply Model #2. When we have data from A, B and C, we would use the full model. In this study, we experiment with three configurations for the full model. For Model #3.1, we have only one sample per population. In this case we can construct three HMMs for pairwise alignments of the three samples, one from each population. For Model #3.2, we have two samples per population, and we construct one HMM for each of the six types of HMMs. For Model #3.3 we construct fifteen HMMs for all pairwise alignments of six samples, two from each population (here the filled and hollow versions of the same shape denote that we treat samples from the same population as distinct when forming pairwise combinations, see Supplemental Text S1).

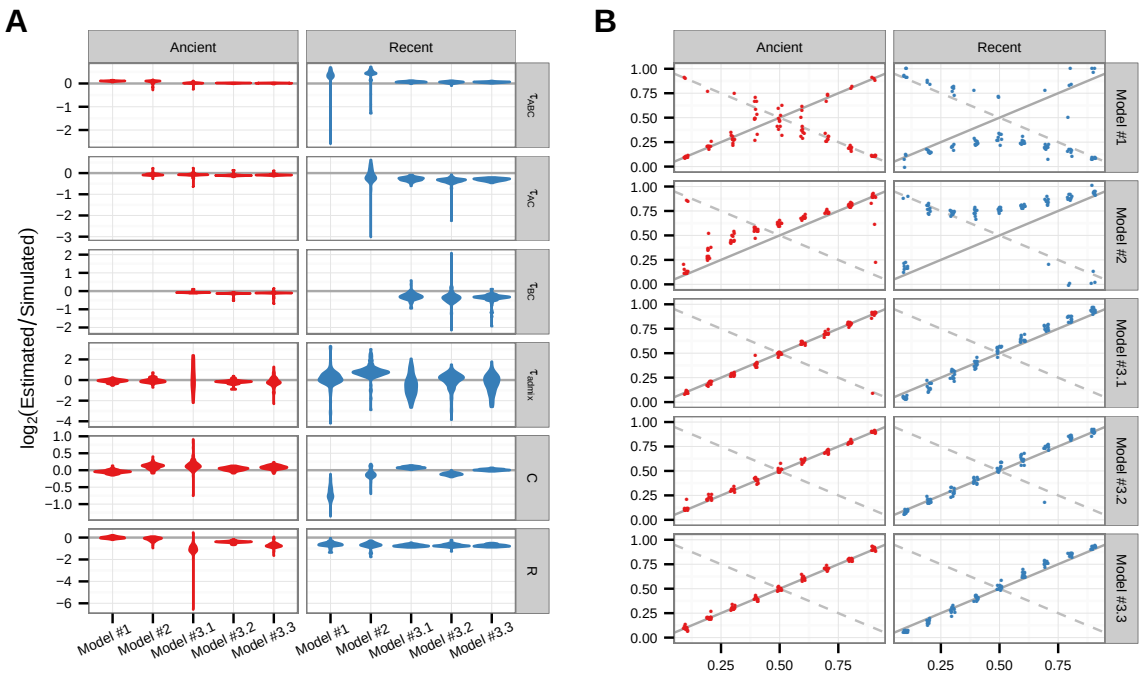


Figure 2. Parameter estimation accuracy. (A) Accuracy of parameter estimation for time parameters, the coalescence rate and the recombination rate. (B) Accuracy of estimation of admixture proportions.

We generally also recover the admixture proportions well, except for Model #1 and Model #2 where the model is just as likely to estimate $1-\alpha$ as we are to estimate α. This is because the likelihood for these models, where we only have samples from the admixed population (Model #1) or the admixed population and one of the source populations (Model #2) is symmetric in this parameter. The models with samples from all three populations do not have this problem and recover the admixture parameter very well.

Impact of varying the effective population size

The coalescence rate parameter captures the effective population size in the model since the effective population size of a population is inversely proportional to the rate at which lineages find a common ancestor, i.e. the coalescence rate. To examine the impact of the effective population size on parameter estimating we simulated data with different coalescence rates while keeping the other parameters fixed. Results are shown in Fig. 3.

The estimated parameters for the same two scenarios are shown in Fig. 3 where (A) show the estimates of the timing of events and the two rates and (B) shows the estimates of admixture proportions. Generally, we see a better estimation accuracy in the ancient scenario than in the recent scenario. This is the case for all timing parameters, coalescent rate, recombination rate, and admixture proportions. We also see the situation where α and $1-\alpha$ are both estimated. These observations are the same as the results shown in Fig 2.

With increasing coalescent rates, corresponding to decreasing effective population sizes, we do not see much effect on the estimation accuracy for admixture proportions shown in (B). Overall, we recover the admixture proportions just as well with the exception of Model #1 and Model #2 in the recent scenario, where high coalescent rates seem to help the inference, but we would not rely on these two models for admixture proportions because of the aformentioned issue that these two models do not distinguish gene flow directions.

Effect of split times relative to the admixture time

Fig. 4 shows the effect of varying the split time between the admixed population and one of the two source populations. When this split is close to the admixture event, we observe good estimates. We observe poor estimates in admixture-related parameters when this split is far back in time.

Admixture CoalHMM model fails to recover admixture proportions when the split is distant. This is caused by the big difference in time between the two key events, the admixture event and the split event related to the admixed population. When this happens the extant source population becomes much different from its ancestral population at the time of the split. The extant source population, therefore, does a poor job reflecting what truly happened during the admixture event. The population that is directly responsible for the admixing is a distant and ancient sbling of the extant source population.

Effect of modeling outgroup as a source population

Fig. 5 shows the effect of applying the admixture model with two source populations in a demographic scenario where one of the source populations to be inferred is actually an outgroup, and the admixture event happens between two ancestral siblings of one source population instead of involving two extant source populations. When the split times of the source population and its ancestral siblings are close to the admixture event, we observe an estimate of zero for the admixture proportion from one direction and one from the other direction. We also observe an overall inaccurate estimate of the admixture-related parameters because this is a model misspecification case where the model used here fails to capture the true demography.

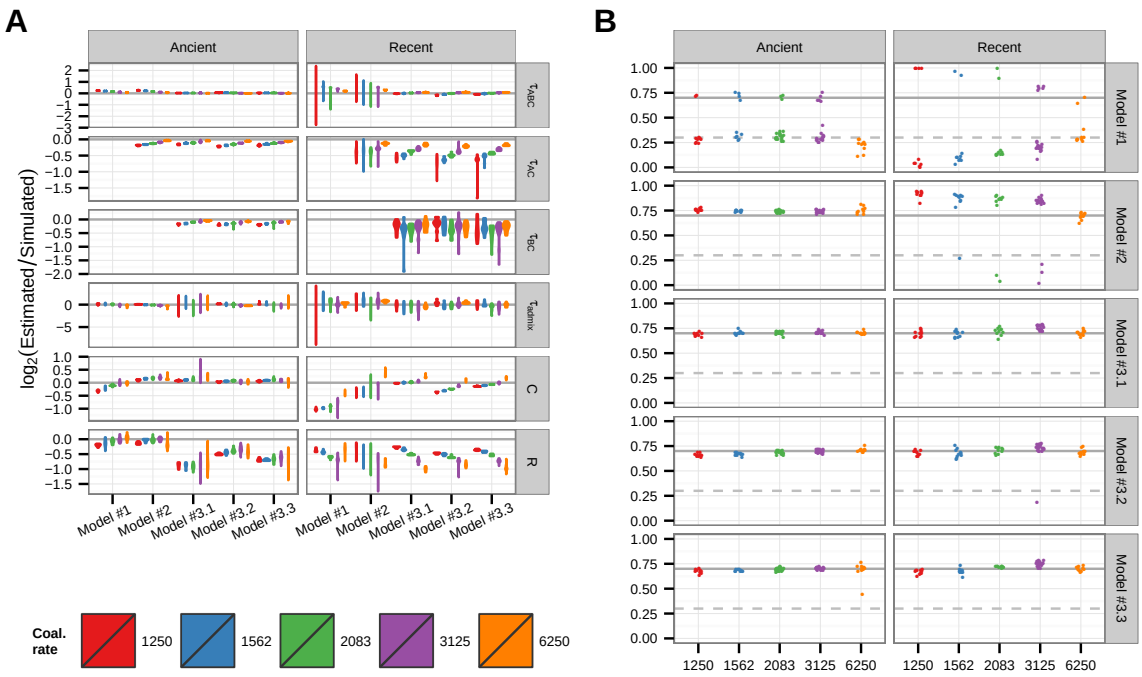


Figure 3. Varying the coalescence rate (effective population size). (A) Accuracy of parameter estimation for time parameters, the coalescence rate and the recombination rate. (B) Accuracy of estimating the admixture rates.

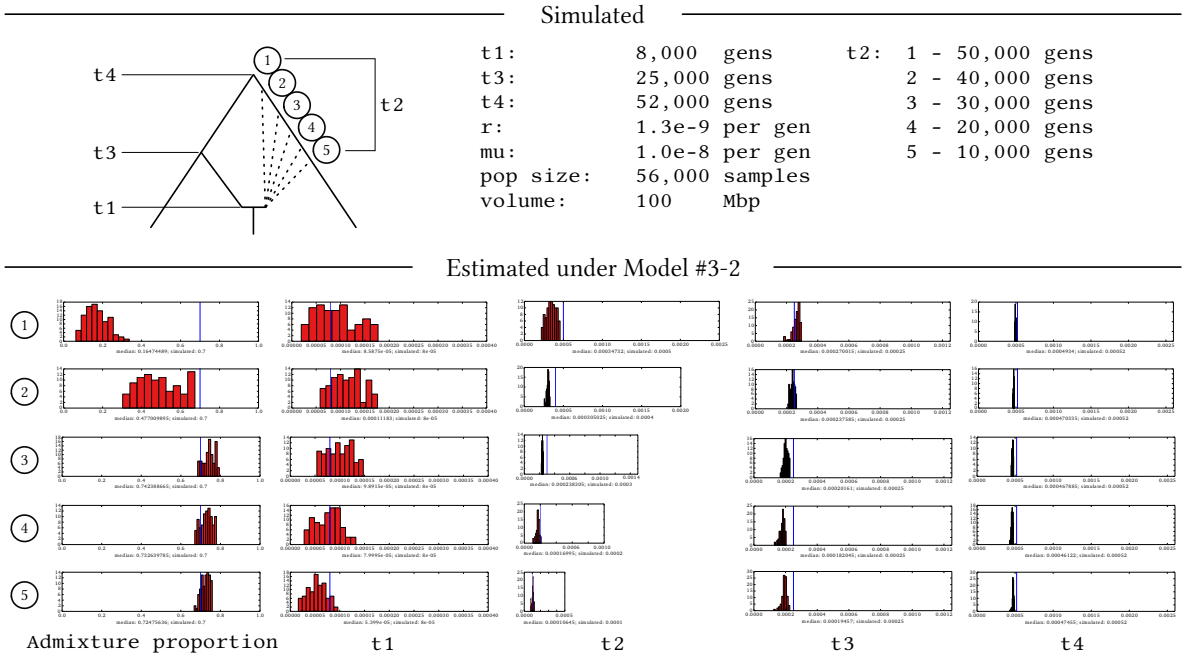


Figure 4. Effect of split times relative to the admixture time. Accuracy of parameter estimation with a range of split times, from early back in time to recent, close to the admixture event. We observe good estimates when the split is close to the admixture event. We observe poor estimates in admixture-related parameters when this split is far back in time. The first column of estimates for admixture proportions demonstrates this situation with clarity.

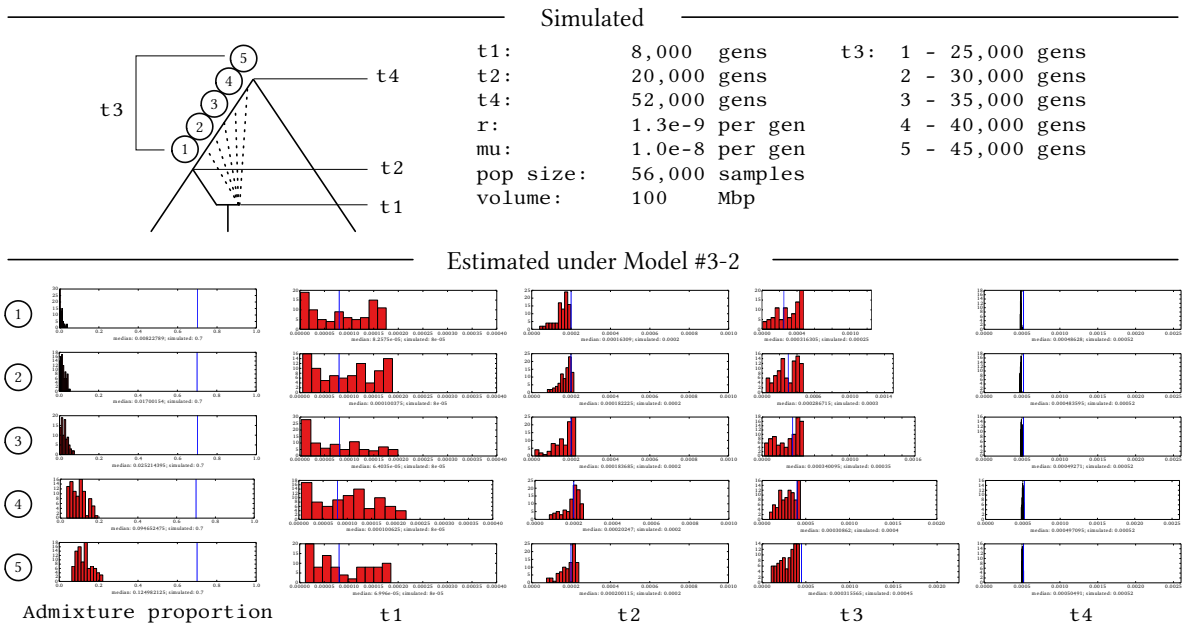


Figure 5. Effect of outgroup modeled as a source population. Accuracy of parameter estimation with a range of split times when an outgroup is mistakenly modeled as a source population. We observe a failure in recovering parameters related to the admixture event due to the wrong modeling, but the estimates accurately reflect reality by attributing gene flow to just one source population.

Admixture CoalHMM attributed all gene flow to one source population. This is an accurate reflection of the demography because the outgroup contributes zero percent in the admixture event while the one true source population contributes all of the gene flow.

Effect of continuous gene flow

In this section, we demonstrate the effect of applying the admixture model with two source populations on three types of demographic scenarios where continuous gene flow is involved among some of the extant or ancestral populations. The admixture model does not accurately model any of the these demographics, but we observe the effect when we apply a misspecified admixture model.

Effect of continuous gene flow in a three population isolation and migration scenario Fig. 6 shows the effect of applying the admixture model with two source populations in a scenario where the true demographic does not contain an admixture event. Instead, it is a three population isolation and migration model where we allow migration between the outgroup and one of the two closely related populations. When removing migration from the demography, the admixture model records false gene flow from the closely related population. With an increasing migration rate, the admixture model records an increasing admixture proportion from the outgroup.

Admixture CoalHMM attributes all gene flow to the closes-related population when migration is low. This agrees with the demographics in the best way if admixing is the only form of gene flow in the model. With an increasing migration between the middle population with the outgroup, the opposite best reflects the demographics.

Effect of recent gene flow between the admixed population and the extant source populations Fig. 7 shows the effect of applying the admixture model with two source populations in an admixture scenario where constant gene flows exist between the admixed population and the two source populations. When the rate of gene flow is zero, the admixture model accurately report the demographic parameters. When the rate of gene flow increases, we see a steady decline in the estimation accuracy for all parameters. The split times between the admixed population with the source populations are affected most significantly. Continuous gene flow dramatically reduces time estimates, which lead to populations with extremely recent splits.

Admixture CoalHMM does not model continuous gene flow, hence it attributes recent gene flow entirely to the admixture event leading to a extremely recent estimate for the admixture time. The recent migration also blurs the directions and quantities of gene flows from both directions leading to the failure in estimating admixture proportions. Finally, when migration is not incorporated into the model, the system compensates by producing recent splits and closely related extant populations.

Effect of distant gene flow between the two ancestral source populations Fig. 8 shows the effect of applying the admixture model with two source populations in an admixture scenario where constant gene flow exist between the two root ancestral populations after the root split but before the splits related to the admixed population. We observe a decline in the estimation accuracy for admixture-related parameters. Overall, this type of continuous gene flow has the smallest effect compared with the cases described in Fig. 6 and Fig. 7.

Admixture CoalHMM considers accumulated coalescence from its modern samples backward in time. Admixture-relevant parameters are affected most directly when the model fails to capture the demographics from the admixture event to current time. This is what we observed in Fig. 7. The simulated ancient gene flow in this experiment, however, only influences the admixture-relevant parameters indirectly and less significantly. As for the under estimation for the root splits, this is due to the ancient gene flow and the same compensation mechanism as discussed in the previous section.

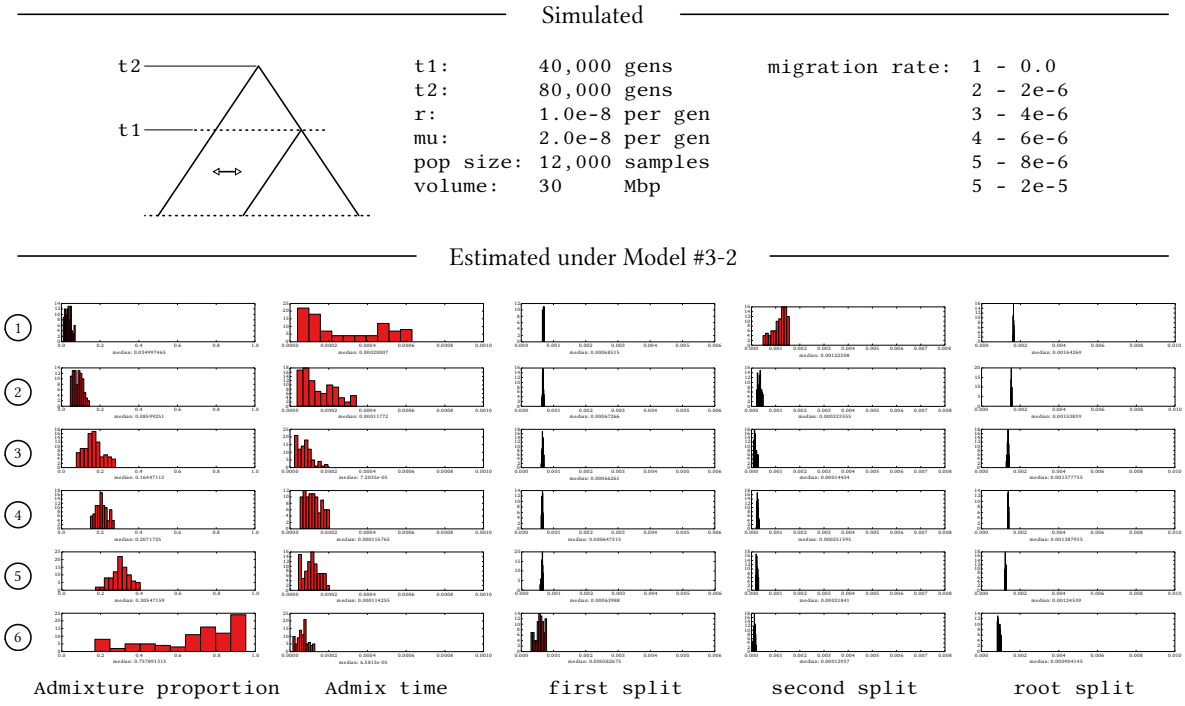


Figure 6. Effect of continuous gene flow, type 1. Accuracy of parameter estimation when applying the admixture model with two source populations in a scenario where the true demography is a three population isolation and migration model allowing migration between the outgroup and one of the two closely related populations. The admixture model records false gene flow from the closely related population, and this direction of gene flow dominates when migration rate is low. The admixture model records an increasing admixture proportion from the outgroup when we increase the migration rate in the simulated data.

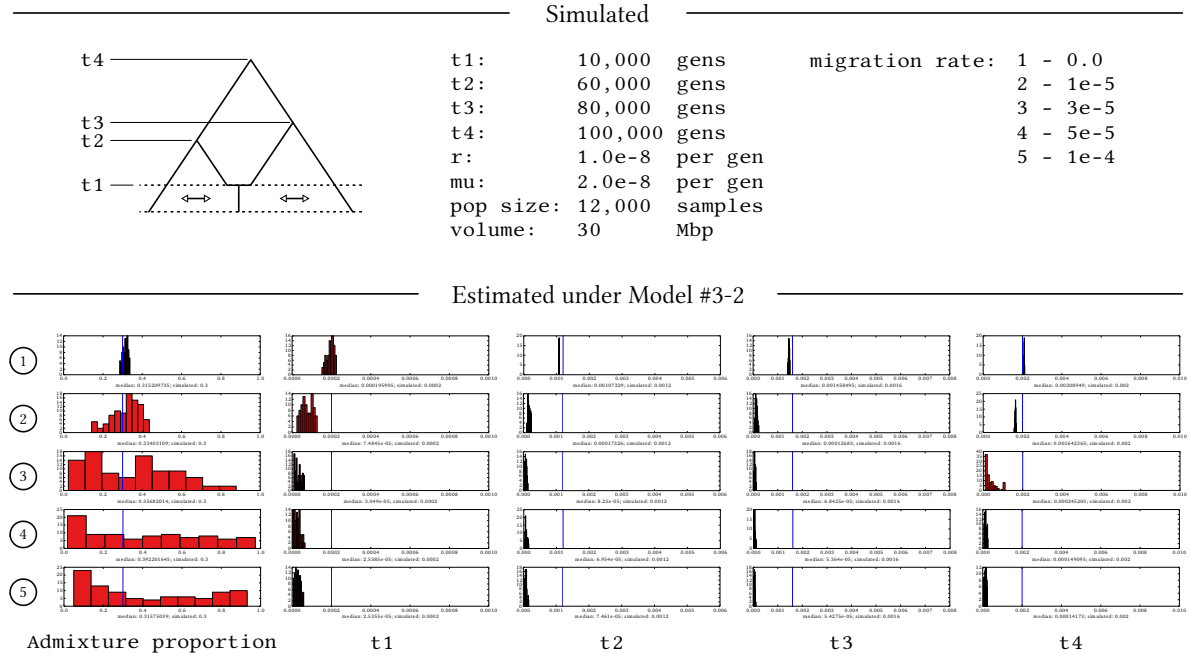


Figure 7. Effect of continuous gene flow, type 2. Accuracy of parameter estimation when a continuous gene flow exist between extant source populations and the admixed population. We observe accurate estimates when the rate of migration is zero, and we observe a failure in recovering any parameters when the migration rate increases. Specifically, all time estimates become extremely short due to the continuous gene flow in the simulated data.

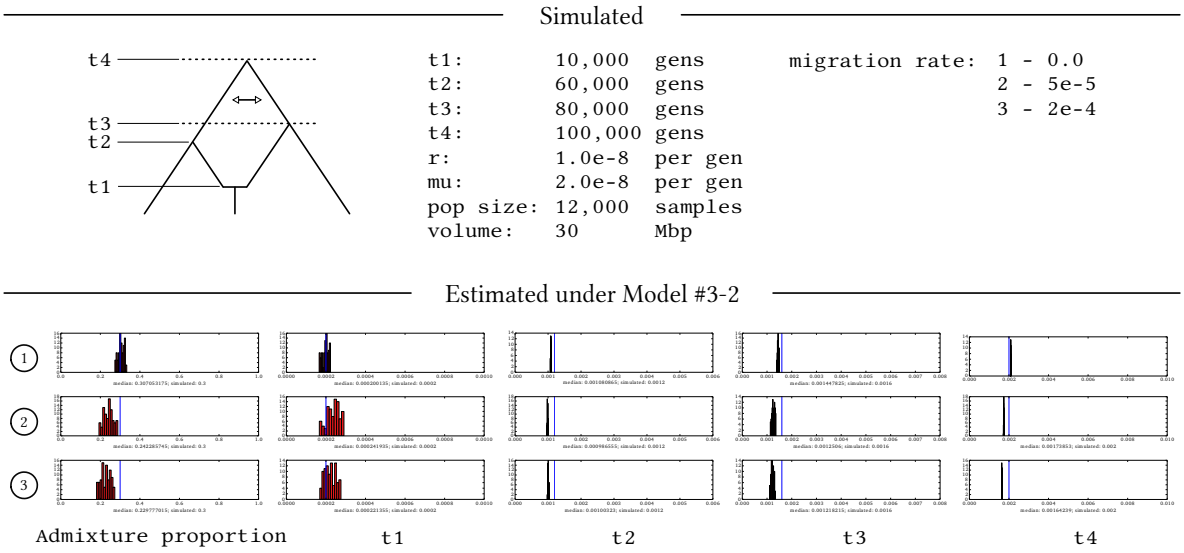


Figure 8. Effect of continuous gene flow, type 3. Accuracy of parameter estimation when continuous gene flow exists between the two root ancestral populations. This corresponds to the time after the root split but before the splits related to the admixed population. We observe a decline in estimation accuracy, but overall, compared with the cases shown in Fig. 6 and Fig. 7, this type of continuous gene flow has the smallest effect on admixture related inference.

Real data analysis

We applied the admixture CoalHMM framework of model construction and estimation to real genomic data. We obtained full genomes for various bear samples. Fig. 9 shows estimates from models that consider PB as an admixed population that received gene flow from a sibling population of BB and ABC as well as gene flow from a population ancestral to both BB and ABC. We consider five bear trios, PB4-BB020-BLK, PB4-ABC1-BLK, PB4-ABC2-BLK, PB4-BB020-ABC1, and PB4-BB020-ABC2.

Discussion

We have developed a coalescent hidden Markov model that enables us to estimate demographic parameters in scenarios where one population is the descendant from an admixture event, and we may or may not have samples from all extant populations. Through simulations we have shown that we recover most parameters well, with a noticeable exception being the recombination rate which we substantially underestimate, an effect we have also seen in our previous coalescence hidden Markov models [35,38]. Through simulations, we have also shown the effect of several common model mis-specifications, all of which fail to recover the mis-specified model parameters in expected manners.

Materials and Methods

Coalescence hidden Markov models infer demographic parameters from sequence alignments by modelling how the ancestry of the sequences vary along the alignment and how the sequences evolved over those ancestries. Conditional on the genealogies, the probability of the observed sequences can be computed

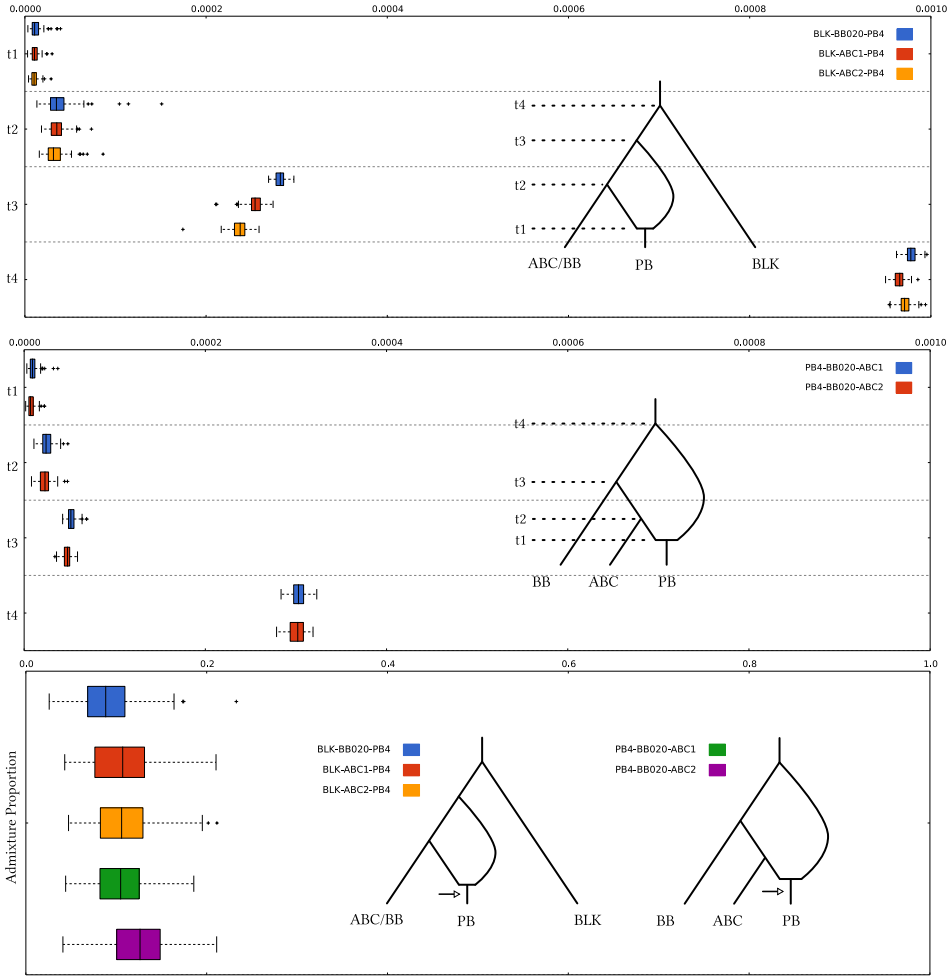


Figure 9. Bear admixture analysis and model comparison We perform 100 bootstrap executions for each bear trio. The two models consider PB as an admixed population which received gene flow from a sibling population of BB and ABC as well as gene flow from a population ancestral to both BB and ABC.

using standard algorithms [27, 35, 38, 39]. The crux of constructing a CoalHMM is thus specifying the probability of moving from one genealogy to the next as we scan along an alignment.

We take the approach, developed in earlier papers [32, 35, 38] of specifying the joint probability of two neighbouring genealogies using a continuous time Markov Chain (CTMC), similar to Simonsen & Churchill [40] and Slatkin & Pollack [41]. Let $J_\Theta(\ell, r)$ denote the joint probability of seeing a "left" genealogy ℓ and "right" genealogy r, given demographic parameters Θ, and let $P_\Theta(\ell)$ denote the probability of seeing the genealogy ℓ, then the transition probability for the hidden Markov model $T_\Theta(r \mid \ell)$ is given by $T_\Theta(r \mid \ell) = J_\Theta(\ell, r) \;/\; P_\Theta(\ell)$. We discretise time to obtain a finite set of possible genealogies and can then compute $P_\Theta(\ell)$ as $P_\Theta(\ell) = \sum_r J_\Theta(\ell, r)$, so for the full specification of the hidden Markov model we only need to specify $J_\Theta(\ell, r)$. This probability is computed by explicitly considering all possible histories that produces genealogies ℓ and r.

Tracing the ancestry of lineages using continuous time Markov chains

To model the ancestry of genealogies ℓ and r we construct a CTMC that tracks all ancestral states that our samples can go through. We start our system with two samples, either from the same population or from two different populations, and trace their lineages back in time. Initially, our samples consist of two nucleotides sitting on the same genome—the left and right nucleotide is linked—and back in time these nucleotides can be separated through recombination events and re-linked through coalescence events. At the admixture event, the lineages in the admixed population can jump to the source populations, independently, and when two or more lineages are in the same population they can coalesce into common ancestors. The time at which common ancestors are found defines the genealogies ℓ and r while all other events are integrated out using the CTMC framework.

We construct a CTMC to compute probabilities for all ancestries similar to our earlier work [42]: We specify a transition system that tracks all the possible events that lineages can undergo and assign rates to these events. The states of this transition system consists of a number of ancestral lineages, each lineage specified as a triplet, (p, ℓ, r), where p denotes the population the lineage is in, ℓ the set of samples that the lineage is ancestral to at the given time at the left nucleotide, and r the set of samples the lineage is ancestral to at the given time in the right nucleotide.

At any given time, a lineage (p, ℓ, r) can undergo recombination, at rate R, and split into two lineages: $(p, \ell, \emptyset)$ and $(p, \emptyset, r)$. Two lineages, in the same population, p, can coalesce, merging (p, ℓ_1, r_1) and (p, ℓ_2, r_2) into a single ancestral lineage $(p, \ell_1 \cup \ell_2, r_1 \cup r_2)$, at rate C_p (coalescence rates depend on the population since we allow the effective population size to vary between populations). We explicitly enumerate all possible states—the number of different ancestral lineages possible—and construct a rate matrix, Q, for the CTMC with off-diagonal values given by the recombination and coalescence rates and the diagonal values given by $Q_{i,i} = -\sum_{j \neq i} Q_{i,j}$. From this rate matrix, the transition probabilities of being in state y at time t, given we were in state x at time s is given by $\left[\exp\left(Q(t-s)\right)\right]_{x,y}$ where $\exp\left(Q(t-s)\right)$ is the matrix exponentiation [43].

Different time periods will have different state spaces: At the time before the admixture event, samples from each population will be confined to their initial population while after the admixture event lineages from the admixed population will have moved to one of the source populations and different states—reflecting that different coalescence events are now possible—will be reachable. We model this by having different Q matrices at different time periods and having projection matrices, η, when moving from one time period to the next. The responsibility of the η matrices is to map lineages in the admixed population to the source populations, with probabilities given by the admixture proportions, and to map lineages from the separate source populations into the ancestral population at the time of the initial split between source populations.

If we move from one state space, given by $Q^{(1)}$, at time τ, to another state space, given by $Q^{(2)}$, then the probability of going from state x at time $s < \tau$ to state $y > \tau$ will be given by the transition probability from time s to τ, then the projection matrix, η, and then the transition probability for going

from time τ to time t: $\left[\exp\left(Q^{(1)}(\tau - s)\right) \times \eta \times \exp\left(Q^{(2)}(t - \tau)\right)\right]_{x,y}$. When more than two time periods are involved, several projection matrices must be used, but the transition probabilities are computed by combining several such matrix multiplications.

The mapping from one state space to another can be done by mapping individual lineages. In the simplest case, there is a one-to-one mapping between lineages before the change in state space to after the change in state space: $\lambda : (p, \ell, r) \mapsto (p', \ell', r')$. This is the case when two populations, p_1 and p_2, merges into an ancestral population p_A back in time, where all lineages (p_i, ℓ, r) then maps to (p_A, ℓ, r) for $i = 1, 2$. Such a lineage-mapping induces a state-mapping σ_λ that maps states x to $\sigma_\lambda(x) = \{\, \lambda(l) \mid l \in x \,\}$. This defines the η mapping by

$$\eta_{x,y} = \begin{cases} 1 & \text{if } y = \sigma_\lambda(x) \\ 0 & \text{otherwise} \end{cases} .$$

When the change in state space is caused by an admixture event we still map individual lineages, but in this case each lineage can map to one or more different lineages with different probabilities. When population C is admixed from populations A and B with admixture proportions α and $\beta = 1-\alpha$ a lineage (C, ℓ, r) maps to lineage (A, ℓ, r) with probability α and to lineage (B, ℓ, r) with probability β. In this case, the lineage map gives us a set for each lineage $\lambda : (C, \ell, r) \mapsto \{\, [\alpha, (A, \ell, r)]\,,\ [\beta, (B, \ell, r)] \,\}$. For the elements in the image of λ let $\mathfrak{p}$ denote the probability, $\mathfrak{p}([\alpha, (p, \ell, r)]) = \alpha$ and let $\mathfrak{l}$ denote the lineage, $\mathfrak{l}([\alpha, (p, \ell, r)]) = (p, \ell, r)$. For a state, x, we have a set of potential states with corresponding probabilities: $X = \{\, z \,|\, z \in \lambda(l), l \in x \,\}$. The state corresponding to X is $\{\, \mathfrak{l}(z) \,|\, z \in y \,\}$ and the probability of moving to this state is $\prod_{z \in y} \mathfrak{p}(z)$, so

$$\eta_{x,y} = \prod_{z \in X} \mathfrak{p}(z)$$

where $y = \{\, \mathfrak{l}(z) \,|\, z \in X \,\}$ and $X = \{\, z \,|\, z \in \lambda(l), l \in x \,\}$.

Computing joint probabilities from two-locus CTMCs

The CTMC framework described above lets us assign probabilities for being in any given ancestral state at any given time, but for constructing our CoalHMM we must project this to the joint probability of seeing two genealogies ℓ and r.

To use the hidden Markov model framework we need a finite set of possible genealogies, so the first step is to discretise time. We take a simple approach and place breakpoints between intervals uniformly when a demographic period has both a start point and and end point, while for the last period—where we have a single ancestral population—we use an exponential distribution as in Mailund *et al.* [38].

Given breakpoints $0 = \tau_0, \tau_1, \ldots, \tau_N = \infty$ and corresponding state transition probability matrices $P^{(i)} = \exp\left(Q^{(i)}(\tau_{i+1} - \tau_i)\right) \times \eta^{(i)}$—where $\eta^{(i)}$ is the identity matrix unless $Q^{(i)}$ and $Q^{(i+1)}$ has different state spaces—the probability of being in state x at τ_i and in state y at time τ_j, for $j > i$, is $\left(P^{(i)} \times \cdots \times P^{(j-1)}\right)_{x,y}$. Or, letting $P^{[i:j]} = P^{(i)} \times \cdots \times P^{(j-1)}$: $P^{[i:j]}{}_{x,y}$.

For a sample of size two, the joint probability of genealogies ℓ and r are given by the coalescence time on the left and on the right, which means the probability that the left nucleotides coalesce in time interval i and the right nucleotides coalesce in time interval j: $J_\Theta(\ell = i, r = j)$. We denote by $B^{(i)}$ the states in the state space for the time period $[\tau_i : \tau_{i+1}]$ where neither left nor right nucleotides have found a common ancestor (the pair of ancestral configurations $(\{\{1\}, \{2\}\}, \{\{1\}, \{2\}\})$). Similarly, we let $L^{(i)}$ denote the states where the left nucleotides, but not the right, have found a common ancestor—$(\{\{1, 2\}\}, \{\{1\}, \{2\}\})$— $R^{(i)}$ denote the states where the right nucleotides, but not the left, have found a common ancestor—$(\{\{1\}, \{2\}\}, \{\{1, 2\}\})$—and $E^{(i)}$ the states where both left and right nucleotides have found a common ancestor—$(\{\{1, 2\}\}, \{\{1, 2\}\})$—see Mailund *et al.* [38]. Then

$$J_\Theta(\ell = i, r = i) = \sum_{b \in B^{(i-1)}} \sum_{e \in E^{(i)}} P^{[0:i]}_{\iota,b} \times P^{[i:i+1]}_{b,e}$$

where ι denotes the initial state—the state where both samples have left and right nucleotides linked. The probability of $J_\Theta(\ell = i, r = j)$ for $j > i$ is given by

$$J_\Theta(\ell = i, r = j) = \sum_{b \in B^{(i-1)}} \sum_{l_1 \in L^{(i+1)}} \sum_{l_2 \in L^{(j-1)}} \sum_{e \in E^{(i)}} P_{\iota,b}^{[0:i]} \times P_{b,l_1}^{[i:i+1]} \times P_{l_1,l_2}^{[i+1:j-1]} \times P_{b,e}^{[i:i+1]}$$

and by symmetry $J_\Theta(\ell = j, r = i) = J_\Theta(\ell = i, r = j)$. (Special cases, where some of the intervals are empty (e.g. $i + 1 = j - 1$) are handled by having the relevant matrices (e.g. $P^{[i+1:j-1]}$) being the identity matrix).

Composing demographic scenarios

A general admixture scenario was shown in Figure 1 that shows population C as admixed from two populations related to populations A and B. The admixture proportions are α from the population related to A and $\beta = 1 - \alpha$ from the population related to B. Time τ_{admix} is when the admixture happened and times τ_{AC} and τ_{BC} are when the source populations find shared ancestral populations with A and B respectively—the order of τ_{AC} and τ_{BC} can change depending on which split happened first—and τ_{admix} is the time where A and B find a shared ancestral population.

To compute the joint probability of left and right genealogies we compose CTMCs based on this graph. From time zero until τ_{admix} we model three populations independently. At time τ_{admix} lineages from population C splits to two different and independent populations with probabilities α and β through an η matrix. Between τ_{admix} and τ_{AC} there are four independent populations. At time τ_{AC} lineages from population A and the one source population merge into an ancestral population and at time τ_{BC} lineages in B and the other source population maps into the second ancestral population and finally at time τ_{ABC} all lineages map into a shared ancestral population.

There are five different CTMCs in use to model this scenario: one for the time period $[0, \tau_{\text{admix}})$, one for $[\tau_{\text{admix}}, \tau_{AC})$, one for $[\tau_{AC}, \tau_{BC})$, one for $[\tau_{BC}, \tau_{ABC})$, and a final one for $[\tau_{ABC}, \infty)$. The first η mapping matrix involves admixture probabilities and the rest simply maps lineages one-to-one.

Cases where we do not have samples from populations A or B are modelled simply by removing those branches in the graph, and the corresponding states in the CTMCs. When the A or B populations are the actual source populations, we can remove the branches with isolated populations after the admixture event and map lineages directly into the ancestral lineages of those populations, simplifying the CTMCs.

Building a hidden Markov model from joint probabilities

A hidden Markov model is fully specified through two matrices and one vector. The *transition matrix*, T, captures the probabilities of moving from one hidden state to the next as we move along a sequence; the *emission matrix*, E, captures the probability of seeing an observed state conditional on a hidden state; and the *initial probability vector*, π, captures the probability of starting the hidden Markov model in a given state [44].

For coalescent hidden Markov models, the hidden states are underlying genealogies and the observable states are alignment columns. The emission probabilities, the probability of seeing a given alignment column given a given underlying genealogy, can be computed using standard algorithms. The transition matrix is given by the transition probabilities computed from the joint probabilities, as described above, $T_{\ell,r} = T_\Theta(r \mid \ell) = J_\Theta(\ell, r)/P_\Theta(\ell)$ where $P_\Theta(\ell) = \sum_r J_\Theta(\ell, r)$, and the initial probability vector is simply $\pi_\ell = P_\Theta(\ell)$.

Parameter estimation

Given the hidden Markov model parameters, T, E, and π, the likelihood of seeing a sequence of observed states—the sequence alignment in the case of coalescence hidden Markov models—can be computed by

summing over all possible sequences of hidden states. This sum can be efficiently computed using dynamic programming via the so-called FORWARD algorithm. In its basic form, the FORWARD algorithm sums over all states for each position along the sequence, but repetitions in the sequence can be exploited to speed up this computation further, reusing computations when the sequence repeats. In our implementation we use the ZIPHMM algorithm [45] that previous experiments have shown is gives us a speedup in computing the likelihood of one or two orders of magnitude when analysing full genome alignments.

We apply a particle swarm optimiser, PSO, to infer the parameters. In PSO, we represent the position of the ith particle as $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,D})$ and its velocity as $\mathbf{v}_i = (v_{i,1}, v_{i,2}, \ldots, v_{i,D})$, where D is the number of dimensions in the parameter space. We represent the particle's previous position with its best fitness as $\mathbf{p}_i = (p_{i,1}, p_{i,2}, \ldots, p_{i,D})$. During each iteration, the algorithm adjusts the velocity $\mathbf{v}$ and position $\mathbf{x}$ according to the following equations, where r_p and r_g are two random values between zero and one, and ϕ_p and ϕ_g are two positive constants representing cognitive and social influences.

$$\mathbf{v}'_{i,d} \leftarrow \omega \cdot \mathbf{v}_{i,d} + \phi_p \cdot r_p \cdot (\mathbf{p}_{i,d} - \mathbf{x}_{i,d}) + \phi_g \cdot r_g \cdot (\mathbf{p}_{g,d} - \mathbf{x}_{i,d}).$$

$$\mathbf{x}'_{i,d} \leftarrow \mathbf{x}_{i,d} + \mathbf{v}_{i,d}.$$

Each model parameter corresponds to a dimension in the solution space. The optimiser initialises particle velocities from uniform random values within a range of 2% of the predetermined range for each parameter. During each iteration, we update the velocities of each particle using coefficients determined from trial and error. For the inertial coefficient, we use $\omega = 0.9$; i.e. a 90% decay in velocity if the particle is not affected by other forces. For the cognitive and social coefficients, we use $\phi_p = 0.3$ and $\phi_g = 0.1$, respectively. Larger values for ϕ had the tendency to accelerate the particles beyond acceptable ranges. We found population sizes greater than 100 did not significantly improve the performance, but they did dramatically increase the time required for the swarm to converge.

Simulation setup

We use the program fastsimcoal2 [36,37] for continuous-time sequential Markovian coalescent simulations. Fastsimcoal2 handles complex evolutionary scenarios. We simulate demographics involving splitting and fusing of populations, admixture events, changes in migration matrices, etc. From the simulated polymorphic sites of a pairwise sequence, we calculate the HMM observations, which are 0, 1, and 2, for being the same, different, and missing.

References

1. Miller W, Schuster SC, Welch AJ, Ratan A, Bedoya-Reina OC, et al. (2012) Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. Proceedings Of The National Academy Of Sciences Of The United States Of America 109: E2382–90.

2. Cahill JA, Green RE, Fulton TL, Stiller M, Jay F, et al. (2013) Genomic evidence for island population conversion resolves conflicting theories of polar bear evolution. PLoS Genet 9: e1003345–e1003345.

3. Liu S, Lorenzen ED, Fumagalli M, Li B, Harris K, et al. (2014) Population Genomics Reveal Recent Speciation and Rapid Evolutionary Adaptation in Polar Bears. Cell 157: 785–794.

4. Cahill JA, Stirling I, Kistler L, Salamzade R, Ersmark E, et al. (2014) Genomic evidence of geographically widespread effect of gene flow from polar bears into brown bears. Molecular Ecology : n/a–n/a.

5. Jónsson H, Schubert M, Seguin-Orlando A, Ginolhac A, Petersen L, et al. (2014) Speciation with gene flow in equids despite extensive chromosomal plasticity. Proceedings Of The National Academy Of Sciences Of The United States Of America : 201412627.

6. Schubert M, Jónsson H, Chang D, Der Sarkissian C, Ermini L, et al. (2014) Prehistoric genomes reveal the genetic foundation and cost of horse domestication. PNAS : 201416991.

7. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, et al. (2010) A draft sequence of the Neandertal genome. Science 328: 710–722.

8. Reich D, Green RE, Kircher M, Krause J, Patterson N, et al. (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. 468: 1053–1060.

9. Hammer MF, Woerner AE, Mendez FL, Watkins JC, Wall JD (2011) Genetic evidence for archaic admixture in Africa. Proceedings Of The National Academy Of Sciences Of The United States Of America 108: 15123–15128.

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

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

12. Sousa V, Hey J (2013) Understanding the origin of species with genome-scale data: modelling gene flow. Nat Rev Genet .

13. Durand EY, Patterson N, Reich D, Slatkin M (2011) Testing for ancient admixture between closely related populations. Molecular Biology and Evolution 28: 2239–2252.

14. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, et al. (2012) Ancient admixture in human history. Genetics 192: 1065–1093.

15. Pickrell J, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genetics 8: e1002967.

16. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genetics 5: e1000695.

17. Loh PR, Lipson M, Patterson N, Moorjani P, Pickrell JK, et al. (2013) Inferring admixture histories of human populations using linkage disequilibrium. Genetics 193: 1233–1254.

18. Harris K, Nielsen R (2012) Inferring demographic history from a spectrum of shared haplotype lengths. PLoS Genetics 91: 809–822.

19. Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics 158: 885.

20. Hey J (2010) Isolation with migration models for more than two populations. Molecular Biology and Evolution 27: 905–920.

21. Wang Y, Hey J (2010) Estimating divergence parameters with small samples from a large number of loci. Genetics 184: 363–379.

22. Hobolth A, Andersen LN, Mailund T (2011) On computing the coalescence time density in an isolation-with-migration model with few samples. Genetics 187: 1241–1243.

23. Andersen LN, Mailund T, Hobolth A (2014) Efficient computation in the IM model. Journal of mathematical biology 68: 1423–1451.

24. Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A (2011) Bayesian inference of ancient human demography from individual genome sequences. Nature Genetics 43: 1031–1034.

25. McVean G, Cardin NJ (2005) Approximating the coalescent with recombination. Philosophical transactions of the Royal Society of London Series B, Biological sciences 360: 1387–1393.

26. Marjoram P, Wall JD (2006) Fast "coalescent" simulation. BMC genetics 7: 16.

27. Hobolth A, Mailund T, Schierup MH, Hobolth A, Christensen OF, et al. (2007) Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS Genetics 3: e7.

28. Hobolth A, Dutheil JY, Hawks J, Schierup MH, Mailund T (2011) Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. Genome Research 21: 349–356.

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

30. Steinrücken M, Paul JS, Song YS (2013) A sequentially Markov conditional sampling distribution for structured populations with migration and recombination. Theor Popul Biol 87: 51–61.

31. Sheehan S, Harris K, Song YS (2013) Estimating variable effective population sizes from multiple genomes: a sequentially markov conditional sampling distribution approach. Genetics 194: 647–662.

32. Dutheil JY, Ganapathy G, Hobolth A, Mailund T, Uyenoyama MK, et al. (2009) Ancestral population genomics: the coalescent hidden Markov model approach. Genetics 183: 259–274.

33. Hobolth A, Jensen JL (2014) Markovian approximation to the finite loci coalescent with recombination along multiple sequences. Theor Popul Biol .

34. Schiffels S, Durbin R (2014) Inferring human population size and separation history from multiple genome sequences. Nature Genetics .

35. Mailund T, Halager AE, Westergaard M, Dutheil JY, Munch K, et al. (2012) A New Isolation with Migration Model along Complete Genomes Infers Very Different Divergence Processes among Closely Related Great Ape Species. PLoS Genetics 8: e1003125–e1003125.

36. Excoffier L, Foll M (2011) A continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. Bioinformatics 27: 1332-1334.

37. Excoffier DIHSE L, Foll M (2013) Robust demographic inference from genomic and SNP data. PLOS Genetics 9: e1003905.

38. Mailund T, Dutheil JY, Hobolth A, Lunter G, Schierup MH (2011) Estimating divergence time and ancestral effective population size of Bornean and Sumatran orangutan subspecies using a coalescent hidden Markov model. PLoS Genetics 7: e1001319.

39. Dutheil JY, Hobolth A (2011) Ancestral population genomics. Methods in molecular biology (Clifton, NJ) 856: 293–313.

40. Simonsen K, Churchill G (1997) A Markov Chain Model of Coalescence with Recombination. Theor Popul Biol 52: 43–59.

41. Slatkin M, Pollack JL (2006) The concordance of gene trees and species trees at two linked loci. Genetics 172: 1979–1984.

42. Mailund T, Halager A, Westergaard M (2012) Using Colored Petri Nets to Construct Coalescent Hidden Markov Models: Automatic Translation from Demographic Specifications to Efficient Inference Methods. In: Haddad S, Pomello L, editors, Application and Theory of Petri Nets. Bioinformatics Research Center, Aarhus University, Denmark, Berlin, Heidelberg: Springer Berlin / Heidelberg, pp. 32–50.

43. Moler C, Van Loan C (2003) Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later. SIAM Review 45: 3–49.

44. Durbin R, Eddy SR, Krogh A, Mitchison G (2005) Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids. Cambridge Univ Pr, cambridge univ pr edition.

45. Sand A, Kristiansen M, Pedersen CNS, Mailund T (2013) zipHMMlib: a highly optimised HMM library exploiting repetitions in the input to speed up the forward algorithm. BMC Bioinformatics 14: 339.

Ohana's admixture and population tree

This paper unfolds the theory behind programs **qpas**, **cpax** and **nemeco** implemented in Ohana. I show Ohana's strong inference power by showing inference results with simulated and real data. This paper explores model limitations. I also present some software comparisons to show that Ohana's admixture analysis is faster and more accurate than the current state-of-the-art tool in admixture analysis. Ohana introduces a new method to infer population trees, and it should be of use to other researchers as an additional component to their structure-style analysis.

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

Ohana, a tool set for population genetic analyses of admixture components

Jade Yu Cheng [1,2,3]* Thomas Mailund [1] and Rasmus Nielsen [2,3]

[1]Bioinformatics Research Centre, Aarhus University, Aarhus 8000 Denmark.
[2]Departments of Integrative Biology and Statistics, University of California, Berkeley, Berkeley, CA 94720, USA.
[3]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Oster Voldgade 5-7, Copenhagen 1350 Denmark.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Structure methods are highly used population genetic methods for classifying individuals in a sample fractionally into discrete ancestry components. **Contribution:** We introduce a new optimization algorithm of the classical Structure model in a maximum likelihood framework. Using analyses of real data we show that the new optimization algorithm finds higher likelihood values than the state-of-the-art method in the same computational time. We also present a new method for estimating population trees from ancestry components using a Gaussian approximation. Using coalescence simulations modeling populations evolving in a tree-like fashion, we explore the adequacy of the Structure model and the Gaussian assumption for identifying ancestry components correctly and for inferring the correct tree. In most cases, ancestry components are inferred correctly, although sample sizes and times since admixture can influence the inferences. Similarly, the popular Gaussian approximation tends to perform poorly when branch lengths are long, although the tree topology is correctly inferred in all scenarios explored. The new methods are implemented together with appropriate visualization tools in the computer package Ohana. **Availability:** Ohana is publicly available at *https://github.com/jade-cheng/ohana*. Besides its source code and installation instructions, we also provide example workflows in the project wiki site. **Contact:** jade.cheng@birc.au.dk

1 Introduction

To quantify population structure, researchers often use methods based on the Structure model (Pritchard *et al*., 2000). The basic assumption in this model is that individuals belong to a set of K discrete groups, each with unique allele frequencies and obeying Hardy-Weinberg Equilibrium, although the latter assumption can be relaxed (Gao *et al*., 2007). Furthermore, individuals are allowed to have fractional memberships of each group. The groups are often termed 'ancestry components' and are sometimes interpreted to represent ancestral populations. This interpretation may be correct in some scenarios, for example when analyzing balanced samples of recently admixed individuals from otherwise highly divergent groups. However, if basic model assumptions are violated, for example if populations truly are not discrete units, the interpretation is more unclear. Nonetheless, inferences under the Structure model have proven highly popular for quantifying population genetic variation and for exploring the basic structure and divisions of genetic diversity in a sample.

STRUCTURE (Pritchard *et al*., 2000), FRAPPE (Tang *et al*., 2005), and ADMIXTURE (Alexander *et al*., 2009) are arguably the three most commonly used programs that apply the Structure model. STRUCTURE uses a Bayesian approach and relies on a Markov Chain Monte Carlo (MCMC) algorithm to sample jointly the posterior distribution of allele frequencies and fractional group memberships. FRAPPE uses a maximum likelihood approach and optimizes the likelihood for both allele frequencies and fractional group memberships using an expectation-maximization (EM) algorithm. ADMIXTURE uses the same model and statistical framework as FRAPPE but uses a faster optimization algorithm. ADMIXTURE executes a two-stage process, first taking a few fast EM steps and then executing a sequential quadratic programming (QP) algorithm. ADMIXTURE uses a pivoting algorithm to solve each QP

problem and applies a quasi-Newton acceleration to each iteration. This acceleration does not respect parameter bounds. ADMIXTURE projects an illegal update to the nearest feasible point, and the acceleration step contributes only when it results in a better likelihood; otherwise the original QP update is used.

The interpretation of parameter estimates under the Structure model is somewhat contentious (Royal *et al*., 2010; Weiss and Long, 2009). It is not clear exactly what the groups, or ancestry components, represent, but in the most simple interpretation we can think of them as estimates of some idealized ancestral populations. If a researcher has inferred the existence of K ancestral populations and knows the fractional memberships of each individual in these populations, a next question would be to explore their evolutionary history. The estimated allele frequencies can provide information about this.

The first approaches for using allele frequencies to estimate population histories dates back to the seminal work by Edwards and Cavalli-Sforza (Cavalli-Sforza *et al*., 1964, 1967). They used Gaussian models for the joint distribution of allele frequencies of multiple populations to estimate genetic distances and to infer population trees. The use of Gaussian models to approximate genetic drift has recently had a resurgence after the availability of large Single Nucleotide Polymorphism (SNP) data sets. It is used in numerous methods and studies, including tests of local adaptation (e.g., (Coop *et al*., 2010; Gunther *et al*., 2013)) and the popular TREEMIX program developed by Pickrell *et al*. (2012). The basic idea in these methods is that you can define the joint allele frequencies among populations in terms of a Gaussian distribution with a covariance matrix dictated by a tree (or admixture graph). Under the Gaussian model, a tree corresponds to exactly one unique covariance matrix, and each covariance matrix corresponds to at most one tree. Furthermore, the likelihood function can be calculated very fast numerically without any need for pruning. The assumption of a Gaussian model for the allele frequencies corresponds to an assumption of a Brownian motion process to model genetic drift instead of, say, a Wright-Fisher diffusion. For small time intervals, the Brownian motion process can provide a close approximation to the Wright-Fisher diffusion. However, for longer time intervals, especially when the allele frequency is close to either of the boundaries (0 and 1), the Brownian motion model is clearly not a very accurate approximation to the Wright-Fisher diffusion. Nonetheless, the Gaussian models provide useful frameworks for inferences because of the distinct computational advantages.

A natural extension of the structure inference framework is to use similar models on the inferred ancestry groups to explore their evolutionary histories. A primary objective of this paper is to provide a computational tool for doing just this and to examine the performance of the Gaussian model in this context.

We present 'Ohana', a tool suite for inferring global ancestry, population covariances, and constructing population trees using Gaussian models. Ohana uses a maximum likelihood framework similar to ADMIXTURE, but it implements an optimization algorithm based on an Active Set (Murty *et al*., 1988) method to solve the QP problem that, as we will show in the results section, tends to find higher maximum likelihood values than ADMIXTURE in similar computational time. In addition, using the model of NGSADMIX (Skotte *et al*., 2013), it can work on genotype likelihoods from low coverage Next Generation Sequencing (NGS) data instead of called genotypes. It includes an optimization algorithm for estimating the best covariance matrix compatible with a tree, thereby estimating a tree, and simple algorithms and visualization tools for the obtaining a tree from the covariance matrix.

We evaluate the performance of the method on real and simulated data, and we also presents results on the limitations of the popular Gaussian model. We show, perhaps unsurprisingly, that the assumption of a Gaussian model in some cases can lead to severely biased branch lengths of population trees that have evolved under a Wright-Fisher diffusion process. This is a limitation of the approach implemented in Ohana and in other approaches that use Brownian motion models to approximate the Wright-Fisher diffusion.

2 Methods

Ohana's **qpas** program infers admixture using genotype observations stored in the ped format from Plink (Purcell *et al*., 2007) or genotype likelihoods in the bgl format from beagle (Browning *et al*., 2007). Ohana's **nemeco** program infers population covariances, and Ohana's **convert** program facilitates different stages of the analysis by providing file conversions and fast approximations. The source code, installation instructions, and example workflows are available on GitHub at *https://github.com/jade-cheng/ohana*.

2.1 Statistical Models

The likelihood model using genotype observations is given by

$$\ln\left[P_1^O\left(Q,\ F\right)\right] = \sum_i^I \sum_j^J \left\{ g_{ij} \cdot \ln\left[\sum_k^K q_{ik} \cdot f_{kj}\right] + (2 - g_{ij}) \cdot \ln\left[\sum_k^K q_{ik} \cdot (1 - f_{kj})\right]\right\}.$$

where K is the number of ancestry components, I is the number of individuals, and J is the number of polymorphic sites. This is the same as the model used in STRUCTURE (Pritchard *et al*., 2000), FRAPPE (Tang *et al*., 2005), ADMIXTURE (Alexander *et al*., 2009), and SPA (Yang *et al*., 2012).

Using the model in NGSADMIX (Skotte *et al*., 2013), **qpas** can also work on genotype likelihoods. In that case the likelihood model is given by

$$\ln\left[P_1^L\left(Q,\ F\right)\right] = \sum_i^I \sum_j^J \ln\left(g_{ij}^{AA} \cdot A_{ij}^2 + g_{ij}^{aa} B_{ij}^2 + g_{ij}^{Aa} \cdot 2A_{ij}B_{ij}\right).$$

$$A_{ij} = \sum_k^K q_{ik} \cdot f_{kj}$$

$$B_{ij} = \sum_k^K q_{ik} \cdot (1 - f_{kj})$$

where g_{ij}^{AA}, g_{ij}^{Aa}, and g_{ij}^{aa} are the probabilities of observing the sequence data at the ith individual's jth marker, conditioned on genotypes AA, Aa (or aA), and aa, respectively. This representation assumes markers with two alleles, although it could easily be generalized to multiple alleles. The advantage of working on genotype likelihoods instead of called genotypes is that genotype likelihoods incorporate the uncertainty regarding genotype calls inherent in much NGS data, and this makes it more applicable to low- or medium-coverage data (see e.g., (Skotte *et al*., 2013)).

To infer population histories, Ohana models the joint distribution of allele frequencies across all ancestry components as a multivariate Gaussian similar to TREEMIX (Pickrell *et al*., 2012) and Bayenv (Gunther *et al*., 2013). The covariance matrix Ω of dimension $K \times K$ is assumed to be constant among all sites, and the process has a mean μ_j at site j. The joint distribution of allele frequencies is then given by

$$P\left(f_j \mid \Omega,\ \mu_j\right) \sim \mathcal{N}\left(\mu_j,\ \mu_j\left(1 - \mu_j\right)\Omega\right).$$

This system is under-determined (see e.g., (Felsenstein, 2004) chapter 23), i.e. multiple covariance matrices induce the same probability distribution on the allele frequencies. Similar to Felsenstein's restricted maximum likelihood approach (Felsenstein, 1981), we therefore root the tree in one of the observations corresponding to conditioning on the allele frequencies in one of the populations when calculating the joint distribution

of allele frequencies in the other populations. We emphasize that the rooting is arbitrary but that it does not imply any assumptions of this population actually being ancestral (for time reversible models). We then obtain a new covariance matrix Ω', which has size $(K-1)\times(K-1)$ and a joint density of the form

$$\ln\left[P_2\left(F\right)\right] = \ln\left\{\prod_j^J\left[\frac{1}{\sqrt{\left|2\pi c_j\Omega'\right|}}\exp\left(-\frac{1}{2}\cdot f_j'^T\cdot\left(c_j\Omega'\right)^{-1}\cdot f_j'\right)\right]\right\}$$
$$= -\frac{1}{2}\cdot\sum_j^J\left\{(K-1)\cdot\ln\left(2\pi c_j\right)+\ln\left[\det\left(\Omega'\right)\right]+\frac{1}{c_j}\cdot f_j'^T\cdot\Omega'^{-1}\cdot f_j'\right\}$$

where $c_j=\mu_j\left(1-\mu_j\right)$

$f_j'=f_j-f_{j_0}$.

2.2 Parameter Inference

2.2.1 Inference for individual ancestries

To estimate Q and F, we use Newton's approach. In general, we can approximate a function $F\left(x\right)$ with its second order Taylor expansion. We proceed to minimize this second-order approximation by solving Δx. In our problem, ΔQ and ΔF are constrained by $\forall\Delta q_{ik},\ q_{ik}+\Delta q_{ik}\in[0,\,1]$, $\forall\Delta f_{kj},\ f_{kj}+\Delta f_{kj}\in[0,\,1]$, and $\sum_k^K\Delta q_{ik}=0$ because $\sum_k^K q_{ik}=1$. The analytical forms of the differential for $\ln\left[P_1^O\left(Q,\,F\right)\right]$ are presented below.

$$\frac{\partial\left(\ln P_1^O\right)}{\partial q_{ik}} = \sum_j^J\left[\frac{g_{ij}\cdot f_{kj}}{\sum_m^K q_{im}\cdot f_{mj}}+\frac{\left(2-g_{ij}\right)\cdot\left(1-f_{kj}\right)}{\sum_m^K q_{im}\cdot\left(1-f_{mj}\right)}\right]$$
$$\frac{\partial^2\left(\ln P_1^O\right)}{\partial q_{ik}\partial q_{i'k'}} = \begin{cases}-\sum_j^J\left\{\frac{g_{ij}\cdot f_{kj}\cdot f_{k'j}}{\left(\sum_m^K q_{im}\cdot f_{mj}\right)^2}+\frac{\left(2-g_{ij}\right)\cdot\left(1-f_{kj}\right)\cdot\left(1-f_{k'j}\right)}{\left[\sum_m^K q_{im}\cdot\left(1-f_{mj}\right)\right]^2}\right\} & \text{if } i=i'\\ 0 & \text{if } i\neq i'\end{cases}$$
$$\frac{\partial\left(\ln P_1^O\right)}{\partial f_{kj}} = \sum_i^I\left[\frac{g_{ij}\cdot q_{ik}}{\sum_m^K q_{im}\cdot f_{mj}}-\frac{\left(2-g_{ij}\right)\cdot q_{ik}}{\sum_m^K q_{im}\cdot\left(1-f_{mj}\right)}\right]$$
$$\frac{\partial^2\left(\ln P_1^O\right)}{\partial f_{kj}\partial f_{k'j'}} = \begin{cases}-\sum_i^I\left\{\frac{g_{ij}\cdot q_{ik}\cdot q_{ik'}}{\left(\sum_m^K q_{im}\cdot f_{mj}\right)^2}+\frac{\left(2-g_{ij}\right)\cdot q_{ik}\cdot q_{ik'}}{\left[\sum_m^K q_{im}\cdot\left(1-f_{mj}\right)\right]^2}\right\} & \text{if } j=j'\\ 0 & \text{if } j\neq j'\end{cases}.$$

The analytical forms of the differential for $\ln\left[P_1^L\left(Q,F\right)\right]$ can also be found below. For both $\ln\left[P_1^O\left(Q,F\right)\right]$ and $\ln\left[P_1^L\left(Q,\,F\right)\right]$, most off-diagonal values of the Hessians diminish. Leveraging this block structure, we convert the problem from manipulating huge matrices into manipulating sequences of small matrices of size K.

$$\frac{\partial\left(\ln P_1^L\right)}{\partial q_{ik}} = \sum_j^J\left[\frac{G_Q\left(i,j,k\right)}{F\left(i,j\right)}\right]$$
$$\frac{\partial^2\left(\ln P_1^L\right)}{\partial q_{ik}\partial q_{i'k'}} = \begin{cases}\sum_j^J\left[\frac{F(i,j)\cdot H_Q\left(i,j,k,k'\right)-G_Q(i,j,k)\cdot G_Q\left(i,j,k'\right)}{F^2(i,j)}\right] & \text{if } i=i'\\ 0 & \text{if } i\neq i'\end{cases}$$
$$F\left(i,j\right) = g_{ij}^{AA}\cdot A_{ij}^2+g_{ij}^{aa}\cdot B_{ij}^2+g_{ij}^{Aa}\cdot 2A_{ij}B_{ij}$$
$$G_Q\left(i,j,k\right) = \frac{\partial F\left(i,j\right)}{\partial q_{ik}}$$
$$= 2g_{ij}^{AA}\cdot f_{kj}\cdot A_{ij}+2g_{ij}^{aa}\cdot\left(1-f_{kj}\right)\cdot B_{ij}+$$
$$2g_{ij}^{Aa}\cdot\left[A_{ij}\cdot\left(1-f_{kj}\right)+B_{ij}\cdot f_{kj}\right]$$
$$H_Q\left(i,j,k,k'\right) = \frac{\partial G\left(i,j,k\right)}{\partial q_{ik'}}$$
$$= 2g_{ij}^{AA}\cdot f_{kj}\cdot f_{k'j}+2g_{ij}^{aa}\cdot\left(1-f_{kj}\right)\cdot\left(1-f_{k'j}\right)+$$
$$2g_{ij}^{Aa}\left[f_{k'j}\cdot\left(1-f_{kj}\right)+\left(1-f_{k'j}\right)\cdot f_{kj}\right].$$

$$\frac{\partial\left(\ln P_1^L\right)}{\partial f_{kj}} = \sum_i^I\left[\frac{G_F\left(i,j,k\right)}{F\left(i,j\right)}\right]$$
$$\frac{\partial^2\left(\ln P_1^L\right)}{\partial f_{kj}\partial f_{k'j'}} = \begin{cases}\sum_i^I\left[\frac{F(i,j)\cdot H_F\left(i,j,k,k'\right)-G_F(i,j,k)\cdot G_F\left(i,j,k'\right)}{F^2(i,j)}\right] & \text{if } j=j'\\ 0 & \text{if } j\neq j'\end{cases}$$
$$F\left(i,j\right) = g_{ij}^{AA}\cdot A_{ij}^2+g_{ij}^{aa}\cdot B_{ij}^2+g_{ij}^{Aa}\cdot 2A_{ij}B_{ij}$$
$$G_F\left(i,j,k\right) = \frac{\partial F\left(i,j\right)}{\partial f_{kj}}$$
$$= 2g_{ij}1^{AA}\cdot q_{ik}\cdot A_{ij}-2g_{ij}^{aa}\cdot q_{ik}\cdot B_{ij}+$$
$$2g_{ij}^{Aa}\cdot\left(B_{ij}\cdot q_{ik}-A_{ij}\cdot q_{ik}\right)$$
$$H_F\left(i,j,k,k'\right) = \frac{\partial G\left(i,j,k\right)}{\partial f_{k'j}}$$
$$= 2g_{ij}^{AA}\cdot q_{ik}\cdot q_{ik'}+2g_{ij}^{aa}\cdot q_{ik}\cdot q_{ik'}-4g_{ij}^{Aa}\cdot q_{ik}\cdot q_{ik'}.$$

To solve these inequality- and equality-constrained quadratic optimization problems, we use an adaptation of the Active Set Algorithm (Murty *et al*., 1988). To solve the equality problem defined by the active set and to compute the Lagrange multipliers of the active set, we use the Karush-Kuhn-Tucker (KKT) approach (Karush, 1939; Kuhn & Tucker, 1951). In each iteration, the algorithm searches for a better solution by considering the active constraints as equality constraints. It deviates from the bounds when the Lagrange multipliers signal a better solution toward the feasible region. The **qpas** program from Ohana performs this analysis. High-level pseudo-code of this algorithm appears in Algorithm 1 of the Supplementary Information (SI).

The maximum number of iterations performed by Ohana's **qpas** to update Q_i or F_j is the number of constraints. In the worst case, the algorithm considers each constraint once. We have $2K+1$ constraints for updating Q_i and $2K$ constraints for updating F_j. Solving systems of linear equations used in KKT is at most $\Theta\left(K^3\right)$. The runtime complexity for each update of Q and F, therefore, becomes $\Theta\left(IK^3\cdot(2K+1)+JK^3\cdot 2K\right) = \Theta\left(K^4\left(I+J\right)\right)$, taking advantage of the block structure.

2.2.2 Inference for population covariances

To optimize the likelihood model defined in the last equation of section 2.1, we use a black-box style of optimizer, the Nelder-Mead (NM) simplex method (Nelder & Mead *et al*., 1965). We use sample covariances, $S_c=\frac{1}{n}\cdot\sum_i^n\left(x_i-\bar{x}_i\right)\left(x_i-\bar{x}_i\right)^T$, as the initial starting point for the NM optimizer, and we use Cholesky decomposition (Cholesky, 1910) to determine the positive semi-definiteness and to compute matrix inverses and determinants. The **nemeco** program in Ohana performs this analysis. High-level pseudo-code of this algorithm appears in SI Algorithm 2.

2.3 Estimation of phylogenetic trees

With the estimated covariance matrix in hand, we can construct a phylogenetic tree. We use the Neighbor-Joining (NJ) method for this, taking advantage of the NJ theorem (Saitou and Masatoshi, 1987), which states that when a distance matrix is compatible with a phylogentic tree, this tree will be accurately reconstructed by the NJ method. To do so, we first transform the covariance matrix to a distance matrix by observing the distance between two populations is given by $\text{Dist}\left(p_1,p_2\right)=\text{Var}\left(p_1\right)+\text{Var}\left(p_2\right)-2\times\text{Cov}\left(p_1,p_2\right)$.

Notice that there is a one-to-one correspondence between the covariance matrix and distances. These distances are then fed to the NJ algorithm. Ohana's **convert** program performs all of these steps and in addition, provides an option to render the tree as SVG.

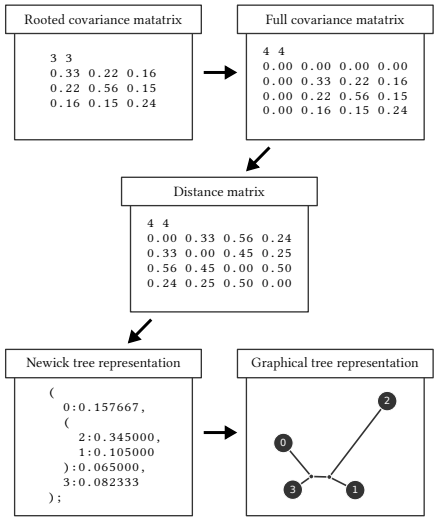


Fig. 1. Phylogenetic tree construction pipeline. Ohana's nemeco program estimates a rooted covariance matrix, where the root is arbitrarily chosen. Ohana's convert program with cov2nwk option then recovers the full covariance matrix, computes the distance matrix, and approximates the distance matrix as a tree structure using the NJ algorithm. Finally, Ohana's convert program with nwk2svg option renders the Newick tree in SVG format. For better control of the graphics, we recommend using our web service: http://www.jade-cheng.com/graphs/

2.4 Simulated data

We used the software **fastsimcoal2** (Excoffier *et al*., 2013) to produce genetic data using the Sequential Markov Coalescence (SMC) model (McVean and Niall, 2005; Marjoram and Simon, 2006). We simulated populations of nucleotide sequences according to a given demographic scenario. For each ancestry component, we simulated 100 sequences of size 20,000,000 bp under an identical population size of 50,000 for all components. We simulated demographic topologies with certain branch lengths by controlling population splits and effective population sizes.

We simulated admixture proportions for un-admixed and admixed scenarios. For un-admixed cases, we simply assigned a fraction of the sample to each population. For admixed cases, we simulated Q_i independently from Dirichlet distributions Dir $(\alpha,\ \alpha,\ \alpha)$, similarly to the simulations used in (Pritchard *et al*., 2000) and (Alexander *et al*., 2009).

Finally, we also simulated genotype observations by first calculating the major allele frequency f_{ij} for each individual at each marker location and then sampling genotypes under the assumption of Hardy-Weinberg Equilibrium, i.e. $p_{ij}^{AA} = f_{ij}^2$, $p_{ij}^{Aa} = 2 \cdot f_{ij} \cdot (1 - f_{ij})$, $p_{ij}^{aa} = (1 - f_{ij})^2$, where $f_{ij} = \sum_k^K Q_{ik} \cdot F_{kj}$, and p^{AA}, p^{Aa}, and p^{aa} are the probabilities of observing major-major, major-minor, or minor-minor genotypes for the locus.

2.5 Real data

We used four data sets for the software comparison with ADMIXTURE shown in Figure 2 and Table 1:

- Dataset #1, a compilation of Europeans containing 17,507 markers and 118 individuals; this data was obtained from the POPRES (Nelson *et al*., 2008), ALS (Laaksovirta *et al*., 2010), Swedish Schizophrenia (Ripke *et al*., 2013), and NCNG (Espeseth *et al*., 2012) projects. It is a subset of data compiled for a study of Danish genetics
- Dataset #2, a compilation of HapMap (HapMap *et al*., 2005) CEU, YRI, MEX, and ASW individuals containing 13,928 markers and 324 individuals. This is the benchmark dataset used in the original ADMIXTURE paper (Alexander *et al*., 2009)
- Dataset #3, a compilation of Han Chinese samples from the HapMap project (HapMap *et al*., 2005) containing 9,822 markers and 171 individuals.
- Dataset #4, a compilation of HapMap (HapMap *et al*., 2005) world population of 4,695 markers 60 individuals of 10 North European, 10 Japanese, 10 Guaharati, 10 Luhya, 10 Maasai Kinyawa, and 10 Tuscan.

For the admixture and covariance data analysis shown in Figure 5, we used a combination of world-wide samples containing 127,855 markers and 80 individuals from the HGDP project. We pruned for minor allele frequencies and Linkage Disequilibrium (LD) with Plink (Purcell *et al*., 2007) using the options *–indep 50 5 2 –geno 0.0 –maf 0.05*.

3 Results

3.1 Computational speed

ADMIXTURE has previously been shown to have the most efficient optimization algorithm among the previously published methods (Alexander *et al*., 2009). We therefore compare the optimization algorithm in Ohana to the algorithms implemented in ADMIXTURE. For a fair comparison, we show the distribution of likelihood values for the two methods, obtained after a fixed amount of computational time, for multiple different runs of Ohana and ADMIXTURE (Figure 2 and Table 1). We verify that the likelihood values are comparable between the two programs by calculating likelihood values for the same parameter values for both programs. We use four different real data sets described in the Methods section and explore a range of different values of K. For a very short amount of computational time, ADMIXTURE tends to find higher likelihood values. ADMIXTURE may possibly use better initial values for the optimization. However, after a relative short amount of time, the **qpas** algorithm in Ohana tends to find higher likelihood values than ADMIXTURE for the same computational time.

3.2 Estimation of admixture fraction and tree on simulated data

We simulated data on a tree using coalescence simulations as described in the Methods section and estimated for different values of K (Figure 3). This mimics the procedure often used in real data analyses in which multiple values of K are explored and presented without knowing the true value of K, although this value can be estimated using a variety of methods (Alexander *et al*., 2011; Scheet and Matthew, 2006; Wold, 1978).

The plots show good correspondence between the true and the estimated values, for both admixture proportions and demography. Furthermore, the changes in tree topology as K changes reflect the hierarchical structure of the tree. For example, at $K = 4$ the internal branch reflects the split between populations (0, 1, 2) and (3, 4, 5).

3.3 Model limitations

There are at least three reasons why tree estimation using a Gaussian model based on estimated allele frequencies may face challenges. First, the allele frequencies are treated as observed data, but they are truly estimates. This has the potential for introducing a variety of biases. Second, the use of a Brownian motion model to approximate genetic drift is inaccurate near the boundaries and for long divergence times, likely leading to underestimates of the lengths of long branches. Third, due to differences in sample sizes

	Dataset #1			Dataset #2			Dataset #3			Dataset #4		
K	Ohana	ADMIXTURE	Diff	Ohana	ADMIXTURE	Diff	Ohana	ADMIXTURE	Diff	Ohana	ADMIXTURE	Diff
2	-1967733	-1967733	0	-3835358	-3835365	7	-1857263	-1857263	0	-288991	-288991	0
3	-1956785	-1956799	14	-3799873	-3799887	14	-1848450	-1848451	1	-279462	-279463	1
4	-1946218	-1946244	26	-3788598	-3788607	10	-1841198	-1841199	1	-275212	-275213	1
5	-1935775	-1936025	250	-3777351	-3777361	11	-1834377	-1834378	1	-271807	-271808	1
6	-1925636	-1925877	241	-3766558	-3766540	-18	-1827829	-1827830	2	-268837	-268832	-5
7	-1915552	-1915743	191	-3755851	-3755860	9	-1821445	-1821458	13	-265907	-265923	17
8	-1905430	-1905638	209	-3746227	-3745412	-815	-1815214	-1815214	0	-263052	-263096	44
9	-1895372	-1895879	507	-3735240	-3736079	839	-1809084	-1809101	18	-260268	-260440	172
10	-1885306	-1885466	160	-3725558	-3725624	66	-1802911	-1802906	-5	-257539	-257736	197
11	-1875503	-1875853	350	-3715543	-3715157	-385	-1796763	-1796847	84	-254920	-254961	41
12	-1865492	-1865965	474	-3706069	-3707715	1646	-1790671	-1790811	140	-252196	-252266	70
13	-1855502	-1856262	760	-3697531	-3698519	987	-1784688	-1784765	77	-249456	-249468	12
14	-1845732	-1846490	758	-3688970	-3689124	154	-1778599	-1778671	73	-246760	-246817	56
15	-1836315	-1836775	460	-3681092	-3680829	-263	-1772555	-1772669	114	-244058	-244298	240

Table 1. A table of the highest log likelihoods achieved from ADMIXTURE and the qpas program in Ohana for a range K values. For each data set, each program, and each value of K, we executed 100 times using random seeds $0, 1, \ldots, 99$ and chose the highest value found in any run. This mimics the procedure often used for real data analysis. In the vast majority of cases, the qpas program in Ohana found significantly higher likelihood values than ADMIXTURE. Dataset #1 is a compilation of Europeans containing 17,507 markers and 118 individuals. Dataset #2 is the benchmark dataset used in ADMIXTURE (Alexander *et al*., 2009) containing 324 CEU, YRI, MEX, and ASW individuals and 13,928 markers. Dataset #3 is a compilation of 171 Han Chinese samples and 9,822 markers. Dataset #4 is a worldwide population of 60 individuals and 4,695 markers.

for different populations, the Structure model may not identify groups that correspond to natural units of a tree, even when the populations truly have evolved in a tree-like fashion.

We explore some of these issues in the following simulation study (Figure 4) by simulating trees with different divergence times: short, medium, and long. For very short divergence times (Figure 4-a), the covariance matrix was estimated poorly because of the small differences in allele frequencies across populations. This in turn leads to reduced accuracy in the estimation of the tree. While the topology is recovered correctly, the lengths of the external branches are overestimated. This likely happens because the Structure model tends to maximize allele frequency differences for finite sample sizes, i.e. the estimated difference in allele frequencies between pairs of populations tends to be larger than the true difference. This is an issue that can be mitigated with larger sample sizes and tends to be a problem only when branch lengths are very small. Nonetheless, it will likely affect many real data analyses.

In the long divergence scenario, Figure 4-c, another problem arises. For such long branches, the Brownian motion model is a poor approximation to genetic drift, and the mapping between the two transition probability functions (i.e. Wright-Fisher diffusion versus Brownian motion) is such that divergence times tend to be underestimated when they are long. The consequence is that the branch lengths of the tree are underestimated. We verify that this is the source of the bias by also simulating data under a Gaussian model directly and showing that under this model there is no significant bias for long branch lengths. This is described in SI Section 1. We note that the poor approximation of the Brownian motion model to the Wright-Fisher diffusion for long divergence times is a limitation for any inference system using similar statistical models such as TREEMIX (Pickrell *et al*., 2012) and Bayenv (Gunther *et al*., 2013), and it might be worthwhile in future work to explore the consequence of this effect for those methods as well.

In the medium-length divergence scenario (Figure 4-b), neither of the two previously mention sources of bias affect the inferences, and the estimates of the branch lengths are therefore quite close to the true values. In all three divergence scenarios, the tree topologies were always estimated accurately.

3.4 Other simulation scenarios

We also evaluated the performance of the method under several other simulation scenarios, and the results are presented in SI Section 2 to 5. A few noteworthy observations include: (1) In more than one simulation scenario with ancient admixture, the population was not inferred to be admixed but received a unique admixture component, SI Section 2 Figure 4 and Section 3 Figure 5. The probability of inferring admixture likely depends on the amount of drift since admixture. In the context of much human data showing evidence of ancient admixture, it might be worthwhile in future studies to explore how much drift after admixture is required to erase the signal of admixture. (2) When K is smaller than the true number of ancestry components, populations with few individuals represented in the sample tend to be (wrongly) inferred as admixed, SI Section 5 Figure 7. There is a clear dependence on sample size in inferences of admixture components in the Structure model. Similarly, the outgroup tends to be identified as the first admixture component that splits from the rest of the individuals, only when the outgroup is well-represented in the sample in terms of the number of individuals.

3.5 Real data analysis

To illustrate the method, we apply it to the panel of global human data described in the Methods section (Figure 5), using a range of K values. The topologies of the trees largely mimic what is already known about human ancestry (e.g., (Reich *et al*., 2012)), i.e. using a root in Africa, Asians and Native Americans cluster together, the European and middle Eastern groups cluster together, etc. In addition to Yorubans having a long branch because this group is an outgroup to the rest, we also notice a relatively long branch leading to Native Americans, reflecting the increased drift

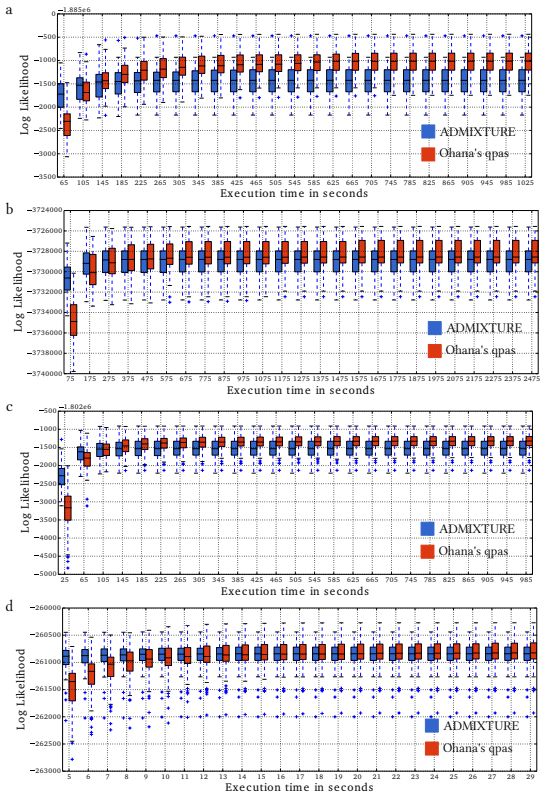


Fig. 2. Comparison of computational speed and efficacy of ADMIXTURE and the qpas program in Ohana. The plots show the change in the distribution of log likelihood values, produced from the two programs over time. For each data set, each program was executed 100 times using random seeds $(0, 1, \ldots, 99)$ and $K = 9$. (a, b, c, d) are four different data sets, same as in Table 1.

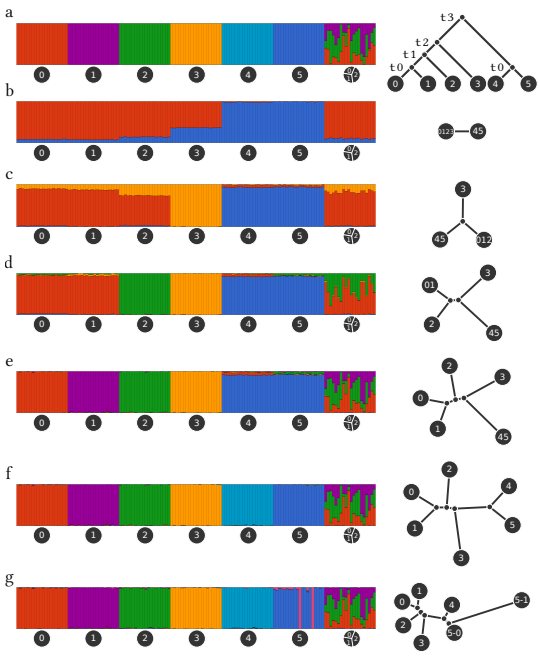


Fig. 3. An evaluation of the tree inference procedure in Ohana using coalescence simulations. We simulated 140 individuals in 7 groups, 20 individuals per group. The first 6 groups were un-admixed. The last group was an equal mixture of the first 3 groups. (a) Simulated admixture (left) and simulated demography (right). (b, c, d, e, f, g) Estimated admixture (left) and estimated demography (right) for $K = 2, 3, 4, 5, 6, 7$, respectively. For each of the 6 populations, we simulated 100 sequences of size 20,000,000 bp using fastsimcoal2 (Excoffier *et al.*, 2013). We used a mutation rate of 2×10^{-8} per generation, a recombination rate of 10^{-8} per generation, and a population size of 50,000. The time parameters were 1000, 2000, 3000, and 4000 generations for t_0, t_1, t_2, and t_3, respectively. A total of 125,787 markers survived filtration for being polymorphic, di-allelic, and with minor allele frequency greater than 5%. We then estimated admixture fractions and population trees using values of K ranging from 2 to 7.

in this group due to the bottleneck into the Americas and possibly small population sizes thereafter.

4 Discussion

In this paper, we introduced a new implementation of the Structure model in a maximum likelihood framework. We compared the new optimization algorithm to the one implemented in the hitherto fastest program, ADMIXTURE. The **qpas** program in our software, Ohana, generally outperformed ADMIXTURE by obtaining estimates with higher likelihood values in similar computational time.

In addition, we presented a new approach for estimating trees for ancestry components. Using coalescence simulations, we showed that when the trees are interpreted as reflecting true population trees, external branch lengths tend to be overestimated for small divergence times. However, for long divergence times, the use of a Gaussian model and its inaccuracy in approximating genetic drift cause branch length estimates to be downward biased. Nonetheless, the estimates of tree topology appear reasonably robust. The tree estimation and visualization tool should be of use to other researchers as an additional possible component of a Structure model analysis of the data. The tree is a visualization of the covariance structure of the admixture components, and it may as such be useful even if a strict interpretation of a evolutionary tree may not be warranted. There might be several reason why such an interpretation may not be appropriate, most of all because the true nature of the evolution of the ancestry components may not be well-described by a tree. Ancestry components are constructions that may or may not reflect true ancestral populations.

Acknowledgements

This work is funded by the Danish Council of Independent Research Sapere Aude grant 12-125062; *Conflict of Interest:* none declared.

References

Alexander, David H., John Novembre, and Kenneth Lange. "Fast model-based estimation of ancestry in unrelated individuals." Genome research 19, no. 9 (2009): 1655-1664.

Alexander, David H., and Kenneth Lange. "Enhancements to the ADMIXTURE algorithm for individual ancestry estimation." BMC

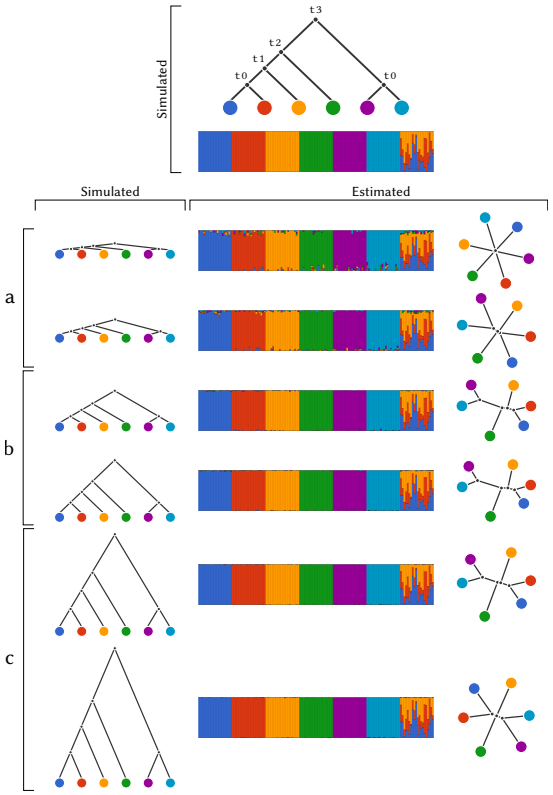


Fig. 4. A simulation study for different divergence times. We simulated 140 individuals in 7 groups, 20 individuals per group. The first 6 groups were un-admixed. The last group was an equal mixture of the first 3 groups. We illustrate the simulated demography on the top. We simulated 6 divergence scenarios, 2 short shown in (a), 2 medium shown in (b), and 2 long shown in (c). From the shortest to the longest divergence scenario (top to bottom), the split times (t_0, t_1, t_2, t_3) in generation were: (10, 20, 30, 40), (100, 200, 300, 400), (1000, 2000, 3000, 4000), (1500, 3000, 4500, 6000), (10000, 20000, 30000, 40000), (20000, 40000, 60000, 80000).

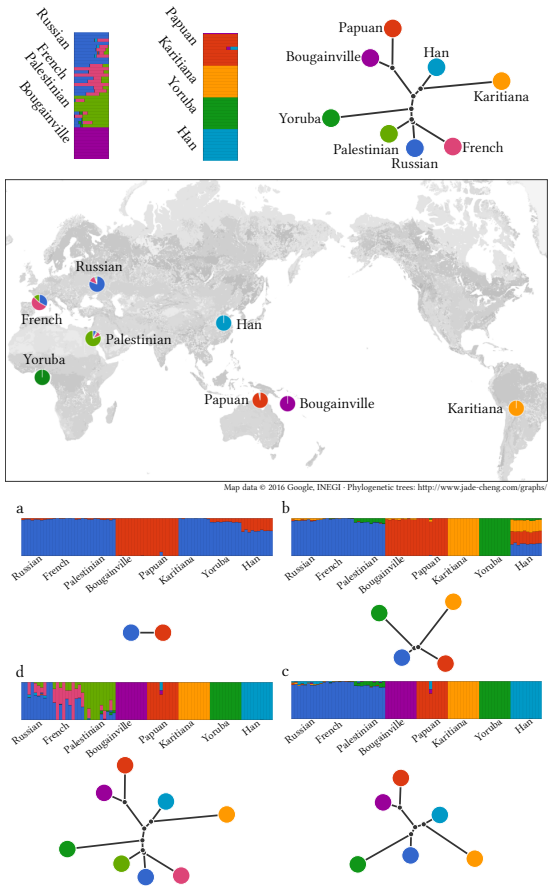


Fig. 5. Analysis of human global data. We used a data set compiled from the HGDP project containing 80 individuals from 8 populations, 10 per population. We filtered markers using Plink (Purcell *et al.*, 2007) with options –indep 50 5 2 –geno 0.0 –maf 0.05. A total of 125,787 markers survived the filtration and were used for the analysis. For each K value, we dispatched 32 executions with random seeds from 0 to 31. We report only results from the execution that reached the best likelihood for each K. The plots show individual admixture proportions and population trees for several different values of K. The map combines the admixture results and geographical records of the HGDP samples. Each slice of each pie chart shows the sum of one component estimated in samples collected at that region. (a, b, c, and d) show the admixture and tree estimates for $K = 2, 4, 6, 8$, respectively.

bioinformatics 12, no. 1 (2011): 1.

Browning, Sharon R., and Brian L. Browning. "Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering." The American Journal of Human Genetics 81, no. 5 (2007): 1084-1097.

Cavalli-Sforza, Luigi Luca, I. Barrai, and A. W. F. Edwards. "Analysis of human evolution under random genetic drift." In Cold Spring Harbor symposia on quantitative biology, vol. 29, pp. 9-20. Cold Spring Harbor Laboratory Press, 1964.

Cavalli-Sforza, Luigi Luca, I. Barrai, and A. W. F. Edwards. "Phylonative American population historygenetic analysis. Models and estimation procedures." American journal of human genetics 19.3 Pt 1 (1967): 233.

Cholesky, André-Louis. "Sur la résolution numérique des systèmes d'équations linéaires." Bulletin de la Sabix. Société des amis de la Bibliothèque et de l'Histoire de l'École polytechnique 39 (2005): 81-95. Harvard

Coop, Graham, David Witonsky, Anna Di Rienzo, and Jonathan K. Pritchard. "Using environmental correlations to identify loci underlying local adaptation." Genetics 185, no. 4 (2010): 1411-1423.

Espeseth, Thomas, Andrea Christoforou, Astri J. Lundervold, Vidar M. Steen, Stephanie Le Hellard, and Ivar Reinvang. "Imaging and cognitive genetics: the Norwegian Cognitive NeuroGenetics sample." Twin Research and Human Genetics 15, no. 03 (2012): 442-452.

Excoffier, Laurent, Isabelle Dupanloup, Emilia Huerta-Sanchez, Vitor C. Sousa, and Matthieu Foll. "Robust demographic inference from genomic and SNP data." PLoS Genet 9, no. 10 (2013): e1003905.

Felsenstein, Joseph. "Evolutionary trees from DNA sequences: a maximum likelihood approach." Journal of molecular evolution 17, no. 6 (1981): 368-376.

Felsenstein, Joseph, and Joseph Felenstein. Inferring phylogenies. Vol. 2. Sunderland: Sinauer Associates, 2004.

Gao, Hong, Scott Williamson, and Carlos D. Bustamante. "A Markov chain Monte Carlo approach for joint inference of population structure

and inbreeding rates from multilocus genotype data." Genetics 176, no. 3 (2007): 1635-1651.

Gunther, Torsten, and Graham Coop. "Robust identification of local adaptation from allele frequencies." Genetics 195, no. 1 (2013): 205-220.

International HapMap Consortium. "A haplotype map of the human genome." Nature 437, no. 7063 (2005): 1299-1320.

Karush, William. "Minima of functions of several variables with inequalities as side constraints." PhD diss., Master's thesis, Dept. of Mathematics, Univ. of Chicago, 1939.

Kayser, Manfred, Silke Brauer, and Mark Stoneking. "A genome scan to detect candidate regions influenced by local natural selection in human populations." Molecular Biology and Evolution 20, no. 6 (2003): 893-900.

Kuhn, HW-Tucker. "AW (1951) Nonlinear programming." In 2nd Berkeley Symposium. Berkeley, University of California Press. 1951.

Laaksovirta, Hannu, Terhi Peuralinna, Jennifer C. Schymick, Sonja W. Scholz, Shaoi-Lin Lai, Liisa Myllykangas, Raimo Sulkava et al. "Chromosome 9p21 in amyotrophic lateral sclerosis in Finland: a genome-wide association study." The Lancet Neurology 9, no. 10 (2010): 978-985.

Marjoram, Paul, and Simon Tavaré. "Modern computational approaches for analysing molecular genetic variation data." Nature Reviews Genetics 7, no. 10 (2006): 759-770.

McVean, Gilean AT, and Niall J. Cardin. "Approximating the coalescent with recombination." Philosophical Transactions of the Royal Society of London B: Biological Sciences 360, no. 1459 (2005): 1387-1393.

Murty, Katta G., and Feng-Tien Yu. Linear complementarity, linear and nonlinear programming. Berlin: Heldermann, 1988.

Nelder, John A., and Roger Mead. "A simplex method for function minimization." The computer journal 7, no. 4 (1965): 308-313.

Nelson, Matthew R., Katarzyna Bryc, Karen S. King, Amit Indap, Adam R. Boyko, John Novembre, Linda P. Briley et al. "The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research." The American Journal of Human Genetics 83, no. 3 (2008): 347-358.

Nicholson, George, Albert V. Smith, Frosti Jonsson, Omar Gustafsson, Kari Stefansson, and Peter Donnelly. "Assessing population differentiation and isolation from single-nucleotide polymorphism data." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64, no. 4 (2002): 695-715.

Nielsen, Rasmus, Ines Hellmann, Melissa Hubisz, Carlos Bustamante, and Andrew G. Clark. "Recent and ongoing selection in the human genome." Nature Reviews Genetics 8, no. 11 (2007): 857-868.

Nocedal, Jorge, and Stephen Wright. Numerical optimization. Springer Science & Business Media, 2006.

Pickrell, Joseph K., and Jonathan K. Pritchard. "Inference of population splits and mixtures from genome-wide allele frequency data." PLoS Genet 8, no. 11 (2012): e1002967.

Pritchard, Jonathan K., Matthew Stephens, and Peter Donnelly. "Inference of population structure using multilocus genotype data." Genetics 155, no. 2 (2000): 945-959.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller et al. "PLINK: a tool set for whole-genome association and population-based linkage analyses." The American Journal of Human Genetics 81, no. 3 (2007): 559-575.

Reich, David, Nick Patterson, Desmond Campbell, Arti Tandon, Stephane Mazieres, Nicolas Ray, Maria V. Parra et al. "Reconstructing native American population history." Nature 488, no. 7411 (2012): 370-374.

Ripke, Stephan, Colm O'Dushlaine, Kimberly Chambert, Jennifer L. Moran, Anna K. Kahler, Susanne Akterin, Sarah E. Bergen et al. "Genome-wide association analysis identifies 13 new risk loci for schizophrenia." Nature genetics 45, no. 10 (2013): 1150-1159.

Royal, Charmaine D., John Novembre, Stephanie M. Fullerton, David B. Goldstein, Jeffrey C. Long, Michael J. Bamshad, and Andrew G. Clark. "Inferring genetic ancestry: opportunities, challenges, and implications." The American Journal of Human Genetics 86, no. 5 (2010): 661-673.

Saitou, Naruya, and Masatoshi Nei. "The neighbor-joining method: a new method for reconstructing phylogenetic trees." Molecular biology and evolution 4, no. 4 (1987): 406-425.

Scheet, Paul, and Matthew Stephens. "A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase." The American Journal of Human Genetics 78, no. 4 (2006): 629-644.

Skoglund, Pontus, Swapan Mallick, Maria Catira Bortolini, Niru Chennagiri, Tabita Hunemeier, Maria Luiza Petzl-Erler, Francisco Mauro Salzano, Nick Patterson, and David Reich. "Genetic evidence for two founding populations of the Americas." Nature 525, no. 7567 (2015): 104-108.

Skotte, Line, Thorfinn Sand Korneliussen, and Anders Albrechtsen. "Estimating individual admixture proportions from next generation sequencing data." Genetics 195, no. 3 (2013): 693-702.

Tang, Hua, Jie Peng, Pei Wang, and Neil J. Risch. "Estimation of individual admixture: analytical and study design considerations." Genetic epidemiology 28, no. 4 (2005): 289-301.

Weiss, Kenneth M., and Jeffrey C. Long. "Non-Darwinian estimation: My ancestors, my genes' ancestors." Genome Research 19, no. 5 (2009): 703-710.

Wold, Svante. "Cross-validatory estimation of the number of components in factor and principal components models." Technometrics 20, no. 4 (1978): 397-405.

Yang, Wen-Yun, John Novembre, Eleazar Eskin, and Eran Halperin. "A model-based approach for analysis of spatial structure in genetic data." Nature genetics 44, no. 6 (2012): 725-731.

Ohana's application on Aborigine Australians

I participated in a collaborative project researching the genetic history of Aborigine Australians. Admixture and population tree analysis that I produced using Ohana fundamentally changed the nature of this very large collaborative project. This work has concluded with success. The article was accepted and will be published in Nature. Ohana's analysis results appear in the main article. I also provided a brief outline of the methods and additional analysis results. They appear in the Supplementary Information accompanying this Nature article.

ARTICLE

doi:10.1038/nature18299

A genomic history of Aboriginal Australia

Anna-Sapfo Malaspinas[1,2,3]*, Michael C. Westaway[4]*, Craig Muller[1]*, Vitor C. Sousa[2,3]*, Oscar Lao[5,6]*, Isabel Alves[2,3,7]*, Anders Bergström[8]*, Georgios Athanasiadis[9], Jade Y. Cheng[9,10], Jacob E. Crawford[10,11], Tim H. Heupink[4], Enrico Macholdt[12], Stephan Peischl[3,13], Simon Rasmussen[14], Stephan Schiffels[15], Sankar Subramanian[4], Joanne L. Wright[4], Anders Albrechtsen[16], Chiara Barbieri[12,17], Isabelle Dupanloup[2,3], Anders Eriksson[18,19], Ashot Margaryan[1], Ida Moltke[16], Irina Pugach[12], Thorfinn S. Korneliussen[1], Ivan P. Levkivskyi[20], J. Víctor Moreno-Mayar[1], Shengyu Ni[12], Fernando Racimo[10], Martin Sikora[1], Yali Xue[8], Farhang A. Aghakhanian[21], Nicolas Brucato[22], Søren Brunak[23], Paula F. Campos[1,24], Warren Clark[25], Sturla Ellingvåg[26], Gudjugudju Fourmile[27], Pascale Gerbault[28,29], Darren Injie[30], George Koki[31], Matthew Leavesley[32], Betty Logan[33], Aubrey Lynch[34], Elizabeth A. Matisoo-Smith[35], Peter J. McAllister[36], Alexander J. Mentzer[37], Mait Metspalu[38], Andrea B. Migliano[29], Les Murgha[39], Maude E. Phipps[21], William Pomat[31], Doc Reynolds[40], Francois-Xavier Ricaut[22], Peter Siba[31], Mark G. Thomas[28], Thomas Wales[41], Colleen Ma'run Wall[42], Stephen J. Oppenheimer[43], Chris Tyler-Smith[8], Richard Durbin[8], Joe Dortch[44], Andrea Manica[18], Mikkel H. Schierup[9], Robert A. Foley[1,45], Marta Mirazón Lahr[1,45], Claire Bowern[46], Jeffrey D. Wall[47], Thomas Mailund[9], Mark Stoneking[12], Rasmus Nielsen[1,48], Manjinder S. Sandhu[8], Laurent Excoffier[2,3], David M. Lambert[4] & Eske Willerslev[1,8,18]

The population history of Aboriginal Australians remains largely uncharacterized. Here we generate high-coverage genomes for 83 Aboriginal Australians (speakers of Pama–Nyungan languages) and 25 Papuans from the New Guinea Highlands. We find that Papuan and Aboriginal Australian ancestors diversified 25–40 thousand years ago (kya), suggesting pre-Holocene population structure in the ancient continent of Sahul (Australia, New Guinea and Tasmania). However, all of the studied Aboriginal Australians descend from a single founding population that differentiated ~10–32 kya. We infer a population expansion in northeast Australia during the Holocene epoch (past 10 kya) associated with limited gene flow from this region to the rest of Australia, consistent with the spread of the Pama–Nyungan languages. We estimate that Aboriginal Australians and Papuans diverged from Eurasians 51–72 kya, following a single out-of-Africa dispersal, and subsequently admixed with archaic populations. Finally, we report evidence of selection in Aboriginal Australians potentially associated with living in the desert.

During most of the last 100,000 years, Australia, Tasmania and New Guinea formed a single continent, Sahul, which was separated from Sunda (the continental landmass including mainland and western island Southeast Asia) by a series of deep oceanic troughs never exposed by changes in sea level. Colonization of Sahul is thought to have required at least 8–10 sea crossings between islands, potentially constraining the occupation of Australia and New Guinea by earlier hominins[1]. Recent assessments suggest that Sahul was settled by 47.5–55 kya[2,3] (Fig. 1). These dates overlap with those for the earliest evidence for modern humans in Sunda[4].

The distinctiveness of the Australian archaeological and fossil record has led to the suggestion that the ancestors of Aboriginal Australians and Papuans ('Australo-Papuans' hereafter) left the African continent earlier than the ancestors of present-day Eurasians[5]. Although some

[1]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5–7, 1350 Copenhagen, Denmark. [2]Institute of Ecology and Evolution, University of Bern, Baltzerstrasse 6, 3012 Bern, Switzerland. [3]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland. [4]Research Centre for Human Evolution, Environmental Futures Research Institute, Griffith University, Nathan, Queensland 4111, Australia. [5]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri i Reixac 4, 08028 Barcelona, Spain. [6]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain. [7]Population and Conservation Genetics Group, Instituto Gulbenkian de Ciência, 2780-156 Oeiras, Portugal. [8]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [9]Bioinformatics Research Centre, Aarhus University, 8000 Aarhus, Denmark. [10]Department of Integrative Biology, University of California, Berkeley, California 94720, USA. [11]Verily Life Sciences, 2425 Garcia Ave, Mountain View, California 94043, USA. [12]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany. [13]Interfaculty Bioinformatics Unit University of Bern, Baltzerstrasse 6, CH-3012 Bern, Switzerland. [14]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kemitorvet, Building 208, 2800 Kongens Lyngby, Denmark. [15]Department for Archaeogenetics, Max Planck Institute for the Science of Human History, Kahlaische Straße 10, D-07745 Jena, Germany. [16]The Bioinformatics Centre, Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen, Denmark. [17]Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Kahlaische Straße 10, D-07745 Jena, Germany. [18]Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK. [19]Integrative Systems Biology Laboratory, Division of Biological and Environmental Sciences & Engineering, King Abdullah University of Science and Technology, 23955-6900 Thuwal, Saudi Arabia. [20]Institute for Theoretical Physics, ETH Zürich, Wolfgang-Pauli-Str. 27, 8093 Zürich, Switzerland. [21]Jeffrey Cheah School of Medicine & Health Sciences, Monash University Malaysia, Jalan Lagoon Selatan, Sunway City, 46150 Selangor, Malaysia. [22]Evolutionary Medicine Group, Laboratoire d'Anthropologie Moléculaire et Imagerie de Synthèse, UMR 5288, Centre National de la Recherche Scientifique, Université de Toulouse 3, 31073 Toulouse, France. [23]Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Blegdamsvej 3B, 2200 Copenhagen N, Denmark. [24]CIMAR/CIIMAR, Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Rua dos Bragas 289, 4050-123 Porto, Portugal. [25]National Parks and Wildlife, Sturt Highway, Buronga, New South Wales 2739, Australia. [26]Explico Foundation, Håvågen 10, 6900 Florø, Norway. [27]Giriwandi, Gimuy Yidinji Country, Queensland 4868, Australia. [28]Research Department of Genetics, Evolution and Environment, University College London, Gower Street, London WC1E 6BT, UK. [29]UCL Department of Anthropology, 14 Taviton Street, London WC1H 0BW, UK. [30]Yinhawangka elder, Perth, Western Australia 6062, Australia. [31]Papua New Guinea Institute of Medical Research, PO Box 60, Goroka, Papua New Guinea. [32]Archaeology, School of Humanities & Social Sciences, University PO Box 320, University of Papua New Guinea & College of Arts, Society & Education, James Cook University, Cairns, Queensland 4811, Australia. [33]Ngadju elder, Coolgardie, Western Australia 6429, Australia. [34]Wongatha elder, Kurrawang, Western Australia 6430, Australia. [35]Department of Anatomy, University of Otago, Dunedin 9054, New Zealand. [36]2209 Springbrook Road, Springbrook, Queensland 4213, Australia. [37]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK. [38]Estonian Biocentre, Riia 23b, Tartu 51010, Estonia. [39]86 Workshop Road, Yarrabah, Queensland 4871, Australia. [40]Esperance Nyungar elder, Esperance, Western Australia 6450, Australia. [41]Atakani Street, Napranum, Queensland 4874, Australia. [42]2 Wynnum North Road, Wynnum, Queensland 4178, Australia. [43]School of Anthropology and Museum Ethnography, Oxford University, Oxford OX2 6PE, UK. [44]Centre for Rock Art Research and Management, M257, University of Western Australia, Perth, Western Australia 6009, Australia. [45]Leverhulme Centre for Human Evolutionary Studies, Department of Archaeology and Anthropology, University of Cambridge, Fitzwilliam Street, Cambridge CB2 1QH, UK. [46]Department of Linguistics, Yale University, 370 Temple Street, New Haven, Connecticut 06520, USA. [47]Institute for Human Genetics, University of California, San Francisco, California 94143, USA. [48]Departments of Integrative Biology and Statistics, University of California, Berkeley, California 94720, USA.
*These authors contributed equally to this work.

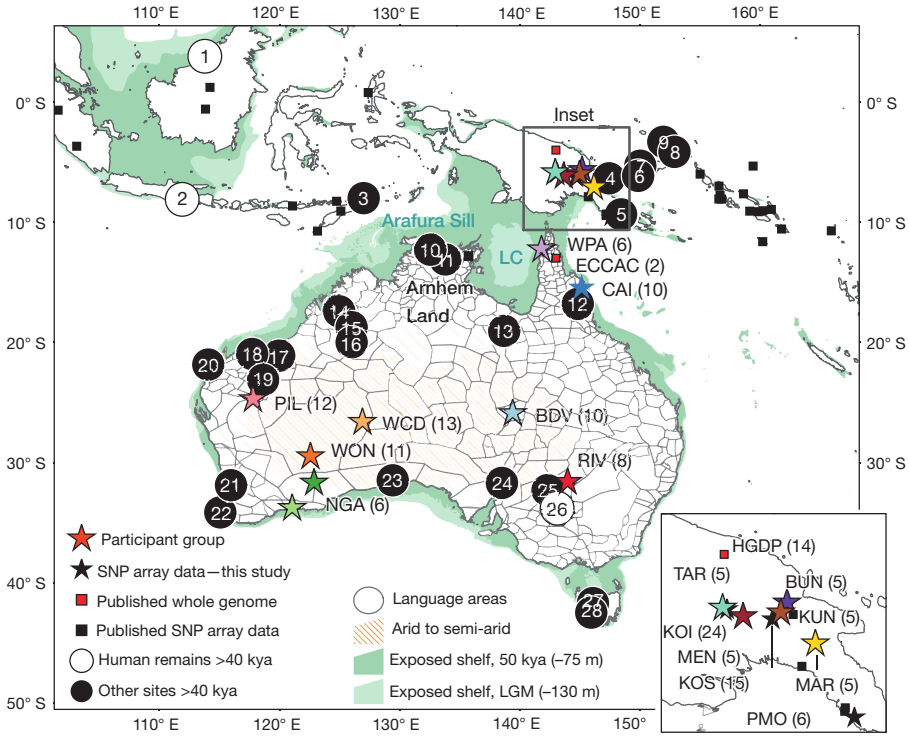


Figure 1 | Aboriginal Australian and Papuan samples used in this study, as well as archaeological sites and human remains dated to ~40 kya or older in southern Sunda and Sahul. The stars indicate the centroid location for each sampling group (sample size in parentheses). Publicly available genetic data (see Supplementary Information section S04) used as a reference panel in this study are shown as squares. Sites with dated human remains are shown as white circles and the archaeological sites as black circles. The associated dates can be found in Supplementary Information section S03. Grey boundaries correspond to territories defined by the language groups provided by the Australian Institute of Aboriginal and Torres Strait Islander Studies[45]. Sampled Aboriginal Australians self-identify primarily as: Yidindji and Gungandji from the Cairns region (CAI, $n = 10$, see also Supplementary Information section S02); Yupangati and Thanakwithi from northwest Cape York (WPA, $n = 6$), Wangkangurru and Yarluyandi from the Birdsville region (BDV, $n = 10$, 9 sequenced at high depth), Barkindji from southeast (RIV, $n = 8$); Pilbara area Yinhawangka and Banjima (PIL, $n = 12$), Ngaanyatjarra from western central desert (WCD, 13), Wongatha from Western Australia's northern Goldfields (WON, $n = 11$), Ngadju from Western Australia's southern Goldfields (NGA, 6); and Nyungar from southwest Australia (ENY, 8). Papuans include samples from the locations Bundi (BUN, $n = 5$), Kundiawa (KUN, $n = 5$), Mendi (MEN, $n = 5$), Marawaka (MAR, $n = 5$) and Tari (TAR, $n = 5$). We generated SNP array data (black stars) for 45 Papuan samples including 24 Koinambe (KOI) and 15 Kosipe (KOS)—described previously[46]—and 6 individuals with Highland ancestry sampled in Port Moresby (PMO). Lake Carpentaria (LC), which covered a significant portion of the land bridge between Australia and New Guinea 11.5–40 kya and thus potentially acted as a barrier to gene flow, is also indicated. Map data were sourced from the Australian Government, http://www.naturalearthdata.com/ and our research.

genetic studies support such multiple dispersals from Africa[6], others favour only one out-of-Africa (OoA) event, with one or two independent founding waves into Asia, of which the earlier contributed to Australo-Papuan ancestry[7,8]. In addition, recent genomic studies have shown that both Aboriginal Australian[8] and Papuan[9] ancestors admixed with Neanderthal and Denisovan archaic hominins after leaving Africa.

Increased desertification of Australia[10] during the last glacial maximum (LGM) 19–26.5 kya impacted the number and density of human populations[11]. In this context, unique morphological and physiological adaptations have been identified in Aboriginal Australians living in desert areas today[12]. In particular, desert groups were hypothesized to withstand sub-zero night temperatures without showing the increase in metabolic rates observed in Europeans under the same conditions.

At the time of European contact, Aboriginal Australians spoke over 250 distinct languages, two-thirds of which belong to the Pama–Nyungan family and cover 90% of the Australian mainland[13]. The place of origin of this language family and the effect of its extensive diffusion on its internal phylogenetic structure have been debated[14], but the pronounced similarity among Pama–Nyungan languages, together with shared socio-cultural patterns, have been interpreted as resulting from a mid-Holocene expansion[15]. Other changes in the mid-late Holocene (~4 kya) include the proliferation of backed blades and the introduction of the dingo[16]. It has been suggested that Pama–Nyungan languages, dingoes and backed blades all reflect the same recent migration into Australia[17]. Although an external origin for backed blades has been rejected, dingoes were certainly introduced, most likely via island Southeast Asia[16]. A recent genetic study found evidence of Indian gene flow into Australia at the approximate time of these Holocene changes[18], suggesting a possible association, while substantial admixture with Asians and Europeans is well documented in historical times[19].

To date, only three Aboriginal Australian whole genome sequences have been described—one deriving from a historical tuft of hair from Australia's Western Desert[8] and two others from cell lines with limited provenance information[20]. In this study, we report the first extensive investigation of Aboriginal Australian genomic diversity by analysing the high-coverage genomes of 83 Pama–Nyungan-speaking Aboriginal Australians and 25 Highland Papuans.

Data set

We collected saliva samples for DNA sequencing in collaboration with Aboriginal Australian communities and individuals in Australia (Supplementary Information section S01). We sequenced genomes at high-depth (average of 60×, range 20–100×) from 83 Aboriginal Australian individuals widely distributed geographically and linguistically (see Fig. 1 and associated legend for the location and label for each group as well as Extended Data Table 1, Supplementary Information

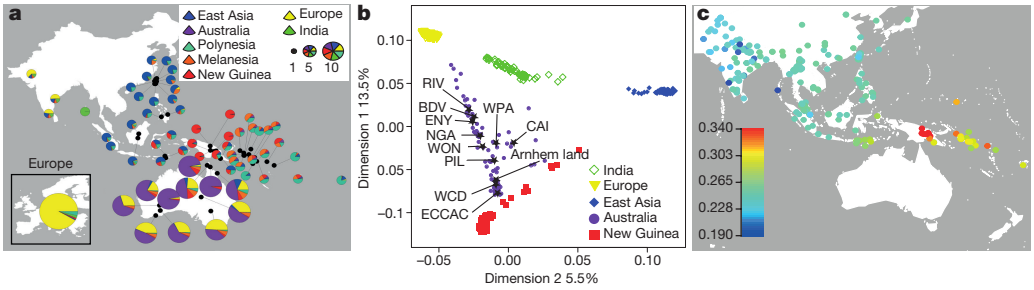


Figure 2 | Genetic ancestry of Aboriginal Australians in a worldwide context. a, Classical Multidimensional scaling (MDS) plot of first two dimensions based on an identity-by-state (IBS) distance matrix (based on 54,971 SNPs) between individuals from this study and worldwide populations, including publicly available data[9,18,26,47]. The first two dimensions explain 19% of the variance in the IBS distance matrix. Individuals are colour-coded according to sampling location, grouped into Australia (Arnhem Land, ECCAC, BDV, CAI, ENY, NGA, PIL, RIV, WCD, WON, WPA); East Asia (Cambodian, Dai, Han, Japanese, Naxi); Europe (English, French, Sardinian, Scottish, Spanish); India (Vishwabrahmin, Dravidian, Punjabi, Guaharati); and New Guinea (HGDP-Papuan, Central Province, Eastern Highlands, Gulf Province, Highlands, PMO, KOI, KOS, BUN, KUN, MEN, TAR, MAR). Stars indicate the centroid for each Aboriginal Australian group. Aboriginal Australians from this study as well as from previous studies are closest to Papuans and also show signals of admixture with Eurasians (see Supplementary Information section S05 for details). **b**, Estimation of genomic ancestry proportions for the best number of ancestral components ($K = 7$) based on Aboriginal Australian and Papuan whole genome sequence and SNP array data from this study (see Fig. 1), and publicly available SNP array data[9,18,26,47] (Supplementary Information section S05). Each ancestry component has been labelled according to the geographic region showing the corresponding highest frequency. The area of each pie chart is proportional to the sample size (as depicted in the legend). The genomes of Aboriginal Australian populations are mostly a mixture of European and Aboriginal Australian ancestry components. Northern Aboriginal Australian groups (Arnhem Land, CAI, ECCAC, PIL and WPA) are also assigned to components mainly present in East Asian populations, while northeastern Aboriginal Australian groups (CAI and WPA) also show components mainly present in New Guinean populations. A background of 5% 'Melanesian' component is observed in all the Aboriginal Australian populations; however, this component is widely spread over the geographic area shown in this figure, being present from Taiwan to India. We detected on average 1.5% 'Indian' component and 1.4% 'Polynesian' component across the Aboriginal Australian samples, but we attribute these residual ancestry components to statistical noise as they are present in other Southeast Asian populations and are not supported by other analyses (Supplementary Information section S05). **c**, A heat map displaying outgroup f_3 statistics of the form f_3(Mbuti; WCD02, *X*), quantifying genetic drift shared between the putatively unadmixed individual WCD02 chosen to represent the Aboriginal Australian population, and various populations throughout the broader region for which either array genotypes or whole-genome sequencing data were publicly available or generated in this study. We used 760,116 SNPs for which WCD02 had non-missing array genotypes that overlapped with any other data sets. Standard errors as estimated from block jack-knife resampling across the genome were in the range 0.00213–0.00713.

sections S02–S04 for more information). Additionally, we sequenced 25 Highland Papuan genomes (38–53×; Supplementary Information sections S03, S04) from individuals representative of five linguistic groups, and generated genotype data for 45 additional Papuans living or originating in the Highlands (Fig. 1). These data sets were combined with previously published genomes and SNP array genotype data, including Aboriginal Australian data from Arnhem Land, and from a human diversity cell line panel from the European Collection of Cell Cultures[20] (ECCAC, Fig. 1, Supplementary Information section S04).

We explored the extent of admixture in the Aboriginal Australian autosomal gene pool by estimating ancestry proportions with an approach based on sparse non-negative matrix factorization (sNMF)[21]. We found that the genomic diversity of Aboriginal Australian populations is best modelled as a mixture of four main genetic ancestries that can be assigned to four geographic regions based on their relative frequencies: Europe, East Asia, New Guinea and Australia (Fig. 2a, Extended Data Fig. 1, Supplementary Information section S05). The degree of admixture varies among groups (Supplementary Information section S05), with the Ngaanyatjarra speakers from central Australia (WCD) having a significantly higher 'Aboriginal Australian component' (median value = 0.95) in their genomes than the other groups sampled (median value = 0.64; Mann–Whitney rank sum test, one tail P value = 3.55×10^{-7}). The East Asian and New Guinean components are mostly present in northeastern Aboriginal Australian populations, while the European component is widely distributed across groups (Fig. 2a, Extended Data Fig. 1, Supplementary Information section S05). In most of the subsequent analyses, we either selected specific samples or groups according to their level of Aboriginal Australian ancestry, or masked the data for the non-Aboriginal Australian ancestry genomic components (Supplementary Information section S06).

Colonization of Sahul

The origin of Aboriginal Australians is a source of much debate, as is the nature of the relationships among Aboriginal Australians, and between Aboriginal Australians and Papuans. Using f_3 statistics[22], estimates of genomic ancestry proportions and classical multidimensional scaling (MDS) analyses, we find that Aboriginal Australians and Papuans are closer to each other than to any other present-day worldwide population considered in our study (Fig. 2b, c, Supplementary Information section S05). This is consistent with Aboriginal Australians and Papuans originating from a common ancestral population which initially colonized Sahul. Moreover, outgroup f_3 statistics do not reveal any significant differences between Papuan populations (Highland Papuan groups sampled in this study and the Human Genome Diversity Project (HGDP-Papuans)) in their genetic affinities to Aboriginal Australians (Extended Data Fig. 2a), suggesting that Papuan populations diverged from one another after or at the same time as the divergence between Aboriginal Australians and Papuans.

To investigate the number of founding waves into Australia, we contrasted alternative models of settlement history through a composite likelihood method that compares the observed joint site frequency spectrum (SFS) to that predicted under specific demographic models[23] (Fig. 3, Supplementary Information section S07). We compared HGDP-Papuans to four Aboriginal Australian population samples with low levels of European admixture (Extended Data Fig. 1) from both northeastern (CAI and WPA) and southwestern desert (WON and WCD) Australia. We compared one- and two-wave models, where each Australian region was either colonized independently, or by descendants of a single Australian founding population after its divergence from Papuans. The one-wave model provides a better fit to the observed SFS, suggesting that the ancestors of the sampled Aboriginal Australians diverged from a single ancestral population. This conclusion is also supported by MDS analyses (Fig. 2b), as well as by estimation of ancestry proportion where all Aboriginal Australians form a cluster distinct from the Papuan populations (Extended Data Fig. 1, Supplementary Information section S05). Additionally, it is supported by outgroup f_3 analyses, where all Aboriginal Australians are largely equidistant from Papuans when adjusting for recent admixture (Extended Data Fig. 2b).

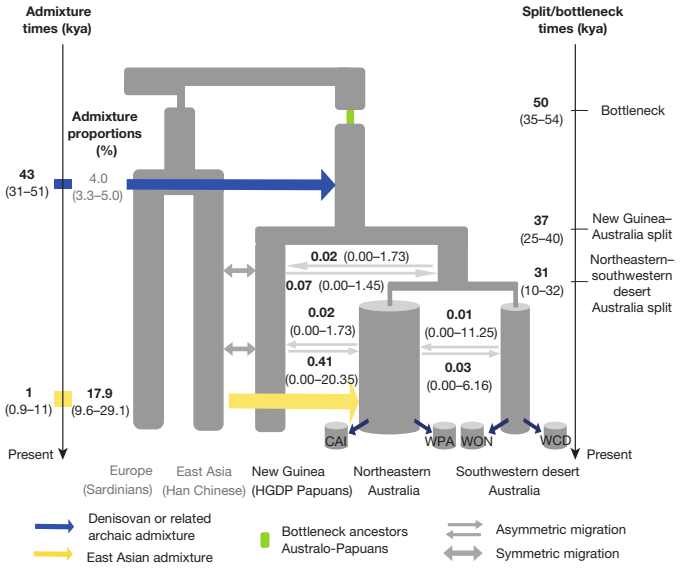


Figure 3 | Settlement of Australia. Best supported demographic model of the colonization of Australia and New Guinea. The demographic history of Aboriginal Australian populations was modelled by considering that sampled individuals are from sub-populations ('islands') that are part of two larger regions ('continents'), which geographically match the northeast and the southwestern desert regions of Australia. Maximum likelihood parameter estimates were obtained from the joint SFS of Han Chinese, HGDP-Papuans, CAI, WPA, WON and WCD. The 95% CI, obtained by non-parametric block bootstrap, are shown within parentheses. Estimated migration rates scaled by the effective population size ($2Nm$) are shown above/below the corresponding arrows. Only Aboriginal Australian individuals with low European ancestry were included in this analysis. In this model, we estimated parameters specific to the settlement of Australia and New Guinea (numerical values shown in black); keeping all the other demographic parameters set to the point estimates shown in Fig. 4 (numerical value shown in grey here). Only admixture events involving proportions >0.5% are shown. The inferred parameters were scaled using a mutation rate of 1.25×10^{-8} per generation per site[41] and a generation time of 29 years corresponding to the average hunter–gatherer generation interval for males and females[42]. See Supplementary Information section S07 for further details.

Thus, our results, based on 83 Pama–Nyungan speakers, do not support earlier claims of multiple ancestral migrations into Australia giving rise to contemporary Aboriginal Australian diversity[24].

The SFS analysis indicates that there was a bottleneck in the ancestral Australo-Papuan population ~50 kya (95% confidence intervals (CI) 35–54 kya, Supplementary Information section S07), which overlaps with archaeological evidence for the earliest occupation of both Sunda and Sahul 47.5–55 kya[2,3]. We further infer that the ancestors of Pama–Nyungan speakers and Highland Papuans diverged ~37 kya (95% CI 25–40 kya, Fig. 3, Supplementary Information section S07), which is in close agreement with results of multiple sequentially Markovian coalescent (MSMC) analyses (Extended Data Fig. 2c, Supplementary Information section S08), a method estimating cross coalescence rates between pairs of populations based on individuals' haplotypes[25]. This result is also in agreement with previous estimates, for example, based on SNP array data[18].

Archaic admixture

We characterized the number, timing and intensity of archaic gene flow events using three complementary approaches: SFS-based (Supplementary Information section S07), a goodness-of-fit analysis combining D-statistics (Supplementary Information section S09), and a method that infers putatively derived archaic 'haplotypes' (Supplementary Information section S10). Aboriginal Australians and Papuan genomes show an excess of putative Denisovan introgressed sites (Extended Data Fig. 3a, Supplementary Information section S11), as well as substantially more putative Denisovan-derived haplotypes (PDHs) than other non-Africans (Extended Data Fig. 3b, Supplementary Information section S10). The number and total length of those putative haplotypes vary considerably across samples. However, the estimated number of PDHs correlates almost perfectly ($r^2 = 0.96$) with the estimated proportion of Australo-Papuan ancestry in each individual (Extended Data Fig. 3c). We found no significant difference in the distribution of the number of PDHs or the average length of PDHs between putatively unadmixed Aboriginal Australians and Papuans (Mann–Whitney U-test, $P > 0.05$). Moreover, the genetic differentiation between WCD and Papuans was also similar for both autosomal SNPs and PDHs with F_{ST} values around 0.12. Taken together, these analyses provide evidence for Denisovan admixture predating the population split between Aboriginal Australians and Papuans (see also ref. 26) and widespread recent Eurasian admixture in Aboriginal Australians (Fig. 2a, b, Supplementary Information section S05). By constraining Denisovan admixture to have occurred before the Aboriginal Australian-Papuan divergence, the SFS-based approach results in an admixture estimate of ~4.0% (95% CI 3.3–5.0%, Fig. 4, Supplementary Information section S07), similar to that obtained by D-statistics (~5%, Supplementary Information section S09). The SFS analyses further suggest that Denisovan/Australo-Papuan admixture took place ~44 kya (95% CI 31–50 kya, Supplementary Information section S07).

The SFS analysis also provides evidence for a primary Neanderthal admixture event (~2.3%, 95% CI 1.1–3.5%, Fig. 4, Supplementary Information section S07) taking place in the ancestral population of all non-Africans ~60 kya (95% CI 55–84 kya, Fig. 4, Supplementary Information section S07). Although we cannot estimate absolute dates of archaic admixture from the lengths of PDHs and putative Neanderthal-derived haplotypes (PNHs) in our samples, we can obtain a relative date. We found that, for putatively unadmixed Aboriginal Australians and HGDP-Papuans, the average PNH and PDH lengths are 33.8 kb and 37.4 kb, respectively (Extended Data Fig. 3b). These are significantly different from each other ($P = 9.65 \times 10^{-6}$ using a conservative sign test), and suggest that the time since Neanderthal admixture was about 11% greater than the time since Denisovan admixture, roughly in line with our SFS-based estimates for the Denisovan pulse (31–50 kya) versus the primary pulse of Neanderthal admixture (55–84 kya). The SFS analysis also indicates that the main Neanderthal pulse was followed by a further 1.1% (95% CI 0.2–2.7%, Fig. 4, Supplementary Information section S07) pulse of Neanderthal gene flow into the ancestors of Eurasians. Finally, using our SFS- and haplotype-based approaches, we explored additional models involving complex structure among the archaic populations. We found suggestive evidence that the archaic contribution could be more complex than a model involving the discrete Denisovan and Neanderthal admixture pulses shown in Fig. 4[8,9] (Supplementary Information sections S07, S10).

Out of Africa

To investigate the relationship of Australo-Papuan ancestors with other world populations, we computed D-statistics[22] of the form ((H1 = Aboriginal Australian, H2 = Eurasian), H3 = African) and ((H1 = Aboriginal Australian, H2 = Eurasian), H3 = Ust'-Ishim). Several of these were significantly positive (Supplementary Information section S09), suggesting that Africans and Ust'-Ishim—a ~45 kya modern human from Asia[27]—are both closer to Eurasians than to Aboriginal Australians. These findings are in agreement with a model of Eurasians and Australo-Papuan ancestors dispersing from Africa in two independent waves. However, when correcting for a moderate amount of Denisovan admixture, Aboriginal Australians and Eurasians become equally close to Ust'-Ishim, as expected in a single OoA scenario (Supplementary Information section S09). Similarly, the D-statistics for ((H1 = Aboriginal Australian, H2 = Eurasian),

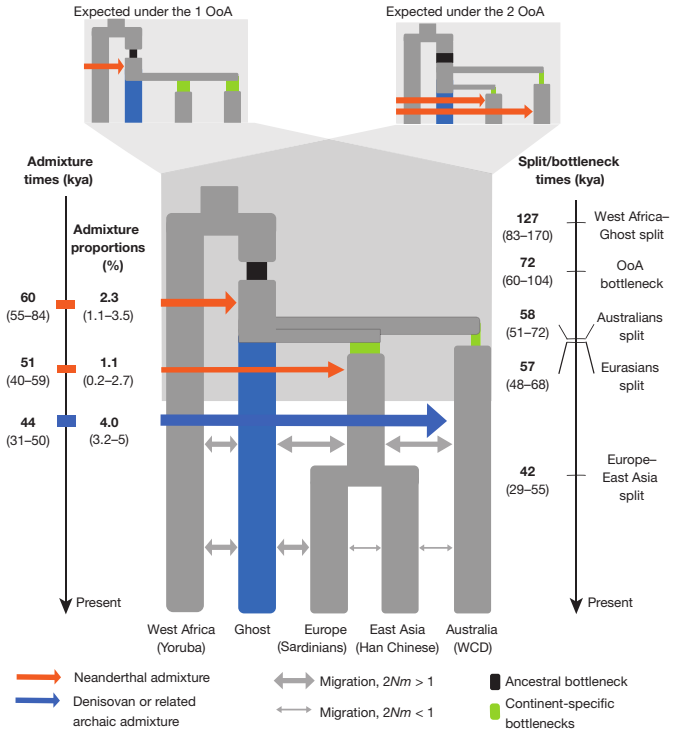


Figure 4 | Out of Africa. We used a likelihood-based approach to investigate whether the joint SFS supports the one-wave (1 OoA) or two-wave (2 OoA) scenarios. The maximum likelihood estimates are indicative of which scenario is best supported. As shown on the top left inset, under the 1 OoA scenario we expect (i) the presence of an ancestral bottleneck (in black); (ii) a relatively large Neanderthal admixture pulse shared by the ancestors of all non-Africans; and (iii) overlapping divergence times of the ancestors of Aboriginal Australians and Eurasians. In contrast, the top right inset shows parameters expected under a 2 OoA scenario: (i) a limited/absent ancestral bottleneck (in black) in the ancestors of all non-Africans; (ii) no shared Neanderthal admixture in the ancestors of all non-Africans; (iii) distinct divergence times for Aboriginal Australians and Eurasians. The main population tree shows the best fitting topology, which supports the 1 OoA scenario, and maximum likelihood estimates (MLEs) for the divergence and admixture times and the admixture proportions (with 95% CI obtained by non-parametric block bootstrap shown within square brackets). We assume that the OoA event is associated with the ancestral bottleneck. The 'Ghost' population represents an unsampled population related to Yoruba that is the source of the out-of-Africa event(s). Our results suggest that these two African populations split significantly earlier (~125 kya) than the estimated time of dispersals into Eurasia. Note that under a 1 OoA scenario, this ghost population becomes, after the ancestral bottleneck, the ancestral population of all non-Africans that admixed with Neanderthals. Arrow thicknesses are proportional to the intensity of gene flow and the admixture proportions, and only admixture events involving proportions >0.5% are displayed. The inferred parameters were scaled as for Fig. 3. See Supplementary Information section S07 for further details.

H3 = African) became much smaller after correcting for Denisovan admixture. Additionally, a goodness-of-fit approach combining D-statistics across worldwide populations indicates stronger support for two waves OoA, but when taking Denisovan admixture into account, a one-wave scenario fits the observed D-statistics equally well (Extended Data Fig. 4a, b, Supplementary Information section S09).

To investigate the timing and number of OoA events giving rise to present-day Australo-Papuans and Eurasians further we used the observed SFS in a model-based composite likelihood framework. When considering only modern human genomes, we find evidence for two waves OoA, with a dispersal of Australo-Papuans ~14,000 years before Eurasians (Supplementary Information section S07). However, when explicitly taking into account Neanderthal and Denisovan introgression into modern humans[9,20], the SFS analysis supports a single origin for the OoA populations marked by a bottleneck ~72 kya (95% CI 60–104 kya, Fig. 4, Supplementary Information section S07). This scenario is reinforced by the observation that the ancestors of Australo-Papuans and Eurasians share a 2.3% (95% CI 1.1–3.5%) Neanderthal admixture pulse. Furthermore, modern humans have both a linkage disequilibrium decay rate and a number of predicted deleterious homozygous mutations (recessive genetic load) that correlate with distance from Africa (Supplementary Information sections S05, S11, Extended Data Fig. 5), again consistent with a single African origin.

The model best supported from the SFS analysis also suggests an early divergence of Australo-Papuans from the ancestors of all non-Africans, in agreement with two colonization waves across Asia[8,9,18]. Under our best model, Australo-Papuans began to diverge from Eurasians ~58 kya (95% CI 51–72 kya, Fig. 4, Supplementary Information section S07), whereas Europeans and East Asians diverged from each other ~42 kya (95% CI 29–55 kya, Fig. 4, Supplementary Information section S07), in agreement with previous estimates[7,18,28]. We find evidence for high levels of gene flow between the ancestors of Eurasians and Australo-Papuans, suggesting that, after the fragmentation of the OoA population ('Ghost' in Fig. 4) 57–58 kya, the groups remained in close geographical proximity for some time before Australo-Papuan ancestors dispersed eastwards. Furthermore, we infer multiple gene flow events between sub-Saharan Africans and Western Eurasians after ~42 kya, in agreement with previous findings of gene flow between African and non-African populations[28].

MSMC analyses suggest that the Yoruba/Australo-Papuans and the Yoruba/Eurasians cross-coalescence rates are distinct, implying that the Yoruba and Eurasian gene trees across the genome have, on average, more recent common ancestors (Extended Data Fig. 4c, Supplementary Information section S08). We show through simulations that these differences cannot be explained by typical amounts of archaic admixture (<20%, Extended Data Fig. 4d). Moreover, the expected difference in phasing quality among genomes is not sufficient to explain this pattern fully (Supplementary Information section S08). While a similar separation in cross coalescence rate curves is obtained when comparing Eurasians and Australo-Papuans with Dinka, we find that, when comparing Australo-Papuans and Eurasians with San, the cross coalescence curves overlap (Extended Data Fig. 4c). We also find that the inferred changes in effective population size through time of Aboriginal Australians, Papuans, and East Asians are very similar until around 50 kya, including a deep bottleneck around 60 kya (Extended Data Fig. 6). Taken together, these MSMC results are consistent with a split of both Australo-Papuans and Eurasians from a single African ancestral population, combined with gene flow between the ancestors of Yoruba or Dinka (but not San) and the ancestors of Eurasians that is not shared with Australo-Papuans. These results are qualitatively in line with the SFS-based analyses (see for example, Fig. 4). While our results do not exclude the possibility of an earlier OoA expansion, they do indicate that any such event left little or no trace in the genomes of modern Australo-Papuans.

Genetic structure of Aboriginal Australians

Uniparental haplogroup diversity in this data set (Extended Data Table 1, Supplementary Information section S12) is consistent with previous studies of mitochondrial DNA (mtDNA) and Y chromosome variation in Australia and Oceania[29], including the presence of typically European, Southeast and East Asian lineages[30]. The combined results provide important insights into the social structure of Aboriginal Australian societies. Aboriginal Australians exhibit greater between-group variation for mtDNA (16.8%) than for the Y chromosome (11.3%), in contrast to the pattern for most human populations[31]. This result suggests higher levels of male- than female-mediated migration, and may reflect the complex marriage and post-marital residence patterns among Pama–Nyungan Australian groups[32]. As expected (Supplementary Information section S02), the inferred European

ancestry for the Y chromosome is much greater than that for mtDNA (31.8% versus 2.4%), reflecting male-biased European gene flow into Aboriginal Australian groups during the colonial era.

On an autosomal level, we find that genetic relationships within Australia reflect geography, with a significant correlation ($r_{GEN,GEO} = 0.77$, $P < 0.0005$, Extended Data Fig. 7b) between the first two dimensions of an MDS analysis on masked genomes and geographical location (Supplementary Information section S13). Populations from the centre of the continent occupy genetically intermediate positions (Extended Data Fig. 7a, b). A similar result is observed with an F_{ST}-based tree for the masked genomic data (Extended Data Fig. 7c, Supplementary Information section S05) as well as in analyses of genetic affinity based on f_3 statistics (Extended Data Fig. 2a), suggesting a population division between northeastern and southwestern groups. This structure is further supported by SFS analyses showing that populations from southwestern desert and northeastern regions diverged as early as ~31 kya (95% CI 10–32 kya), followed by limited gene flow (estimated scaled migration rate ($2Nm$) < 0.01, 95% CI 0.00 < Nm < 11.25). An analysis of the major routes of gene flow within the continent supports a model in which the Australian interior acted as a barrier to migration. Using a model inspired by principles of electrical engineering where gene flow is represented as a current flowing through the Australian continent and using observed F_{ST} values as a proxy for resistance, we infer that gene flow occurred preferentially along the coasts of Australia (Extended Data Fig. 7e–g, Supplementary Information section S13). These findings are consistent with a model of expansion followed by population fragmentation when the extreme aridity in the interior of Australia formed barriers to population movements during the LGM[33].

We used MSMC on autosomal data and mtDNA Bayesian skyline plots[34] (BSP) to estimate changes in effective population size within Australia. The MSMC analyses provide evidence of a population expansion starting ~10 kya in the northeast, while both MSMC and BSP indicate a bottleneck in the southwestern desert populations taking place during the past ~10 kya (Extended Data Fig. 6, Supplementary Information sections S08, S12). This is consistent with archaeological evidence for a population expansion associated with significant changes in socio-economic and subsistence strategies in Holocene Australia[35].

European admixture almost certainly had not occurred before the late 18th century, but earlier East Asian and/or New Guinean gene flow into Australia could have taken place. We characterized the mode and tempo of gene flow into Aboriginal Australians using three different approaches (Supplementary Information sections S06, S07, S14). We used approximate Bayesian computation (ABC) to compare the observed mean and variance in the proportion of European, East Asian and Papuan admixture among Aboriginal Australian individuals, to that computed from simulated data sets under various models of gene flow. We estimated European and East Asian admixture to have occurred on the order of ten generations ago (Supplementary Information section S14), consistent with historical and ethnographic records. Consistent with this, a local ancestry approach suggests that European and East Asian admixture is more recent than Papuan admixture (Extended Data Fig. 4a, Supplementary Information section S06). In addition, both ABC and SFS analyses indicate that the best-fitting model for the Aboriginal Australian-Papuan data is one of continuous but modest gene flow, mostly unidirectional from Papuans to Aboriginal Australians, and geographically restricted to northeast Aboriginal Australians ($2Nm = 0.41$, 95% CI 0.00–20.35, Fig. 3, Supplementary Information section S07).

To investigate gene flow from New Guinea further, we conducted analyses on the Papuan ancestry tracts obtained from the local ancestry analysis. We inferred local ancestry as the result of admixture between four components: European, East Asian, Papuan and Aboriginal Australian (Supplementary Information section S06). The Papuan tract length distribution shows a clear geographic pattern (Extended Data Fig. 8); we find a significant correlation of Papuan tract length variance with distance from WCD to other Aboriginal Australian groups ($r = 0.64$, $P < 0.0001$). The prevalence of short ancestry tracts of Papuan origin, compared to longer tracts of East Asian and European origin, suggests that a large fraction of the Papuan gene flow is much older than that from Europe and Asia, consistent with the ABC analysis (Supplementary Information section S14). We also investigated possible South Asian (Indian related) gene flow into Aboriginal Australians, as reported recently[18]. However, we found no evidence of a component that can be uniquely assigned to Indian populations in the Aboriginal Australian gene pool using either admixture analyses or f_3 and D-statistics (Supplementary Information section S05), even when including the original Aboriginal Australian genotype data from Arnhem Land. The different size and nature of the comparative data sets may account for this discrepancy.

Pama–Nyungan languages and genetic structure

To investigate whether linguistic relationships reflect genetic relationships among Aboriginal Australian populations, we inferred a Bayesian phylogenetic tree for the 28 different Pama–Nyungan languages represented in this sample[13] (Extended Data Table 1, Supplementary Information section S15). The resulting linguistic and F_{ST}-based genetic trees (Extended Data Fig. 7c, d) share several well-supported partitions. For example, both trees indicate that the northeastern (CAI and WPA) and southwestern groups (ENY, NGA, WCD and WON) form two distinct clusters, while PIL, BDV and RIV are intermediate. A distance matrix between pairs of languages, computed from the language-based tree, is significantly correlated with geographic distances ($r_{GEO,LAN} = 0.83$, Mantel test two-tail P value on 9,999 permutations = 0.0001, Supplementary Information section S13). This suggests that differentiation among Pama–Nyungan languages in Australia follows geographic patterns, as observed in other language families elsewhere in the world[36]. Furthermore, we find a correlation between linguistics and genetics ($r_{GEN,LAN} = 0.43$, Mantel test $P < 0.0005$, Supplementary Information section S13) that remains significant when controlling for geography ($r_{GEN,LAN.GEO} = 0.26$, partial Mantel test $P < 0.0005$, Supplementary Information section S13). This is consistent with language differentiation after populations lost (genetic) contact with one another. The correlation between the linguistic and genetic trees is all the more notable given the difference in time scales: the Pama–Nyungan family is generally accepted to have diversified within the last 6,000 years[37], while the genetic estimates are two to five times that age. The linguistic tree thus cannot simply reflect initial population dispersals, but rather reflects a genetic structure that has a complex history, with initial differentiation 10–32 kya, localized population expansions (northeast) and bottlenecks (southwest) ~10 kya, and subsequent limited gene flow from the northeast to the southwest. The latter may be the genetic signature that tracks the divergence of the Pama–Nyungan language family.

Selection in Aboriginal Australians

To identify selection signatures specific to Aboriginal Australians, we used two different methods based on the identification of SNPs with high allele frequency differences between Aboriginal Australians and other groups, similar to the population-branch statistics[38] (PBS, Supplementary Information section S16). First, we scanned the Aboriginal Australian genomes for loci with unusually large changes in allele frequency since divergence from Papuans, taking recent admixture with Europeans and Asians into account ('global scan'). Second, we identified genomic regions showing high differentiation associated with different ecological regions within Australia ('local scan', Supplementary Information section S16). Among the top ranked peaks (Extended Data Table 2) we found genes associated with the thyroid system (*NETO1*, seventh peak in the global scan, and *KCNJ2*, first peak in the local scan) and serum urate levels (eighth peak in the global scan). Thyroid hormone levels are associated with Aboriginal-Australian-specific adaptations to desert cold[39] and elevated serum

urate levels with dehydration[40]. These genes are therefore candidates for potential adaptation to life in the desert. However, further studies are needed to associate putative selected genetic variants with specific phenotypic adaptations in Aboriginal Australians.

Discussion

Australia has one of the longest histories of continuous human occupation outside Africa, raising questions of origins, relatedness to other populations, differentiation and adaptation. Our large-scale genomic data and analyses provide some answers but also raise new questions. We find that Aboriginal Australians and Eurasians share genomic signatures of an OoA dispersal—a common African ancestor, a bottleneck and a primary pulse of Neanderthal admixture. However, Aboriginal Australian population history diverged from that of other Eurasians shortly after the OoA event, and included private admixture with another archaic hominin.

Our genetic-based time estimates are relative, and to obtain absolute dates we relied on two rescaling parameters: the human mutation rate and generation time (assumed to be 1.25×10^{-8} per generation per site and 29 years, respectively, based on recent estimates[41,42]). Although the absolute estimates we report would need to be revised if these parameters were to change, the current values can be the starting point of future research and should be contextualized.

We find a relatively old divergence between the ancestors of Pama–Nyungan speakers and Highland Papuans, only ~10% younger than the European–East Asian split time. With the assumed rescaling parameters this corresponds to ~37 kya (95% CI 25–40 kya) implying that the divergence between sampled Papuans and Aboriginal Australians is older than the disappearance of the land bridge between New Guinea and Australia about 7–14.5 kya, and thus suggests ancient genetic structure in Sahul. Such structure may be related to palaeo-environmental changes leading up to the LGM. Sedimentary studies show that the large Lake Carpentaria (500 × 250 km, Fig. 1) formed ~40 kya, when sea levels fell below the 53-m-deep Arafura Sill[43]. Although Australia and New Guinea remained connected until the early Holocene, the flooding of the Carpentaria basin and its increasing salinity[43] may have thus promoted population isolation.

Our results imply that Aboriginal Australian groups are the descendants of the ancestral population that first colonized Australia[44]. They also indicate that the population that diverged from Papuans ~37 kya was ancestral to all Aboriginal Australian groups sampled in this study; yet, archaeological evidence shows that by 40–45 kya, humans were widespread within Australia (Fig. 1). Three non-exclusive scenarios could account for this observation: (1) the Aboriginal Australian ancestral population was widespread before the divergence from Papuans, maintaining gene flow across the continent; (2) it was deeply structured, and only one group survived to give rise to modern Aboriginal Australians; and (3) other groups survived, but the descendants are not represented in our sample. Additional genomes, especially from Tasmania and the non-Pama–Nyungan regions of the Northern Territory and Kimberley, as well as ancient genomes predating European contact in Australia and other expansions across Southeast Asia[17], may help resolve these questions in the future.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 7 August 2015; accepted 4 May 2016.
Published online xx xx 2016.

1. Davidson, I. The colonization of Australia and its adjacent islands and the evolution of modern cognition. *Curr. Anthropol.* **51,** S177–S189 (2010).
2. Clarkson, C. *et al.* The archaeology, chronology and stratigraphy of Madjedbebe (Malakunanja II): A site in northern Australia with early occupation. *J. Hum. Evol.* **83,** 46–64 (2015).
3. O'Connell, J. F. & Allen, J. The process, biotic impact, and global implications of the human colonization of Sahul about 47,000 years ago. *J. Archaeol. Sci.* **56,** 73–84 (2015).
4. Barker, G.*et al.* The 'human revolution' in lowland tropical Southeast Asia: the antiquity and behaviour of anatomically modern humans at Niah Cave (Sarawak, Borneo). *J. Hum. Evol.* **52,** 243–261 (2007).
5. Lahr, M. M. & Foley, R. Multiple dispersals and modern human origins. *Evol. Anthropol. Issues News Rev.* **3,** 48–60 (1994).
6. Reyes-Centeno, H. *et al.* Genomic and cranial phenotype data support multiple modern human dispersals from Africa and a southern route into Asia. *Proc. Natl Acad. Sci. USA* **111,** 7248–7253 (2014).
7. Wollstein, A. *et al.* Demographic history of Oceania inferred from genome-wide data. *Curr. Biol.* **20,** 1983–1992 (2010).
8. Rasmussen, M. *et al.* An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science* **334,** 94–98 (2011).
9. Reich, D. *et al.* Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am. J. Hum. Genet.* **89,** 516–528 (2011).
10. Reeves, J. M. *et al.* Climate variability over the last 35,000 years recorded in marine and terrestrial archives in the Australian region: an OZ-INTIMATE compilation. *Quat. Sci. Rev.* **74,** 21–34 (2013).
11. Hiscock, P. & Wallis, L. A. in *Desert Peoples* (eds Veth, P., Smith, M. & Hiscock, P.) 34–57 (Blackwell Publishing Ltd, 2005).
12. Birdsell, J. B. *Microevolutionary Patterns in Aboriginal Australia: A Gradient Analysis of Clines.* (Oxford University Press, 1993).
13. Bowern, C. & Atkinson, Q. Computational phylogenetics and the internal structure of Pama–Nyungan. *Language* **88,** 817–845 (2012).
14. Dixon, R. M. W. *Australian Languages: Their Nature and Development.* (Cambridge University Press, 2002).
15. Evans, N. & McConvell, P. The enigma of Pama–Nyungan expansion in Australia. *Archaeol. Lang.* **II,** 174–191 (1997).
16. Hiscock, P. *Archaeology of ancient Australia.* (Routledge, 2008).
17. Bellwood, P. *First Migrants: Ancient Migration in Global Perspective.* (Wiley-Blackwell, 2013).
18. Pugach, I., Delfin, F., Gunnarsdóttir, E., Kayser, M. & Stoneking, M. Genome-wide data substantiate Holocene gene flow from India to Australia. *Proc. Natl Acad. Sci. USA* **110,** 1803–1808 (2013).
19. Ellinghaus, K. Absorbing the 'Aboriginal problem': controlling interracial marriage in Australia in the late 19th and early 20th centuries. *Aborig. Hist.* **27,** 183–207 (2003).
20. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505,** 43–49 (2014).
21. Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G. & François, O. Fast and efficient estimation of individual ancestry coefficients. *Genetics* **196,** 973–983 (2014).
22. Patterson, N. J. *et al.* Ancient admixture in human history. *Genetics* **192,** 1065–1093 (2012).
23. Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C. & Foll, M. Robust demographic inference from genomic and SNP data. *PLoS Genet.* **9,** e1003905 (2013).
24. Thorne, A. G. in *The Origin of the Australians* (eds Kirk, R. L. & Thorne, A. G.) 81–95 (Canberra: Australian Institute of Aboriginal Studies, 1976).
25. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46,** 919–925 (2014).
26. Qin, P. & Stoneking, M. Denisovan Ancestry in East Eurasian and Native American Populations. *Mol. Biol. Evol.* **32,** 2665–2674 (2015).
27. Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514,** 445–449 (2014).
28. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5,** e1000695 (2009).
29. Bergström, A. *et al.* Deep roots for Aboriginal Australian Y chromosomes. *Curr. Biol.* **26,** 809–813 (2016).
30. Hudjashov, G. *et al.* Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis. *Proc. Natl Acad. Sci. USA* **104,** 8726–8730 (2007).
31. Lippold, S. *et al.* Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences. *Investig. Genet.* **5,** 13 (2014).
32. Radcliffe-Brown, A. R. The social organization of Australian tribes. *Oceania* **1,** 34–63 (1930).
33. Veth, P. Islands in the interior: a model for the colonization of Australia's arid zone. *Archaeol. Ocean.* **24,** 81–92 (1989).
34. Drummond, A. J., Rambaut, A., Shapiro, B. & Pybus, O. G. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **22,** 1185–1192 (2005).
35. Lourandos, H. & David, B. in *Bridging Wallace's Line: the Environmental and Cultural History and Dynamics of the SE Asian-Australasian Region* (eds Kershaw, A. P., David, B., Tapper, N., Penny, D. & Brown, J.) (97–118).
36. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes* (Princeton University Press, 1996).
37. Evans, N. & Jones, R. in *Archaeology and linguistics: Aboriginal Australia in global perspective* (Oxford University Press Australia, 1997).
38. Yi, X. *et al.* Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329,** 75–78 (2010).
39. Qi, X., Chan, W. L., Read, R. J., Zhou, A. & Carrell, R. W. Temperature-responsive release of thyroxine and its environmental adaptation in Australians. *Proc. Biol. Sci.* **281,** 20132747 (2014).

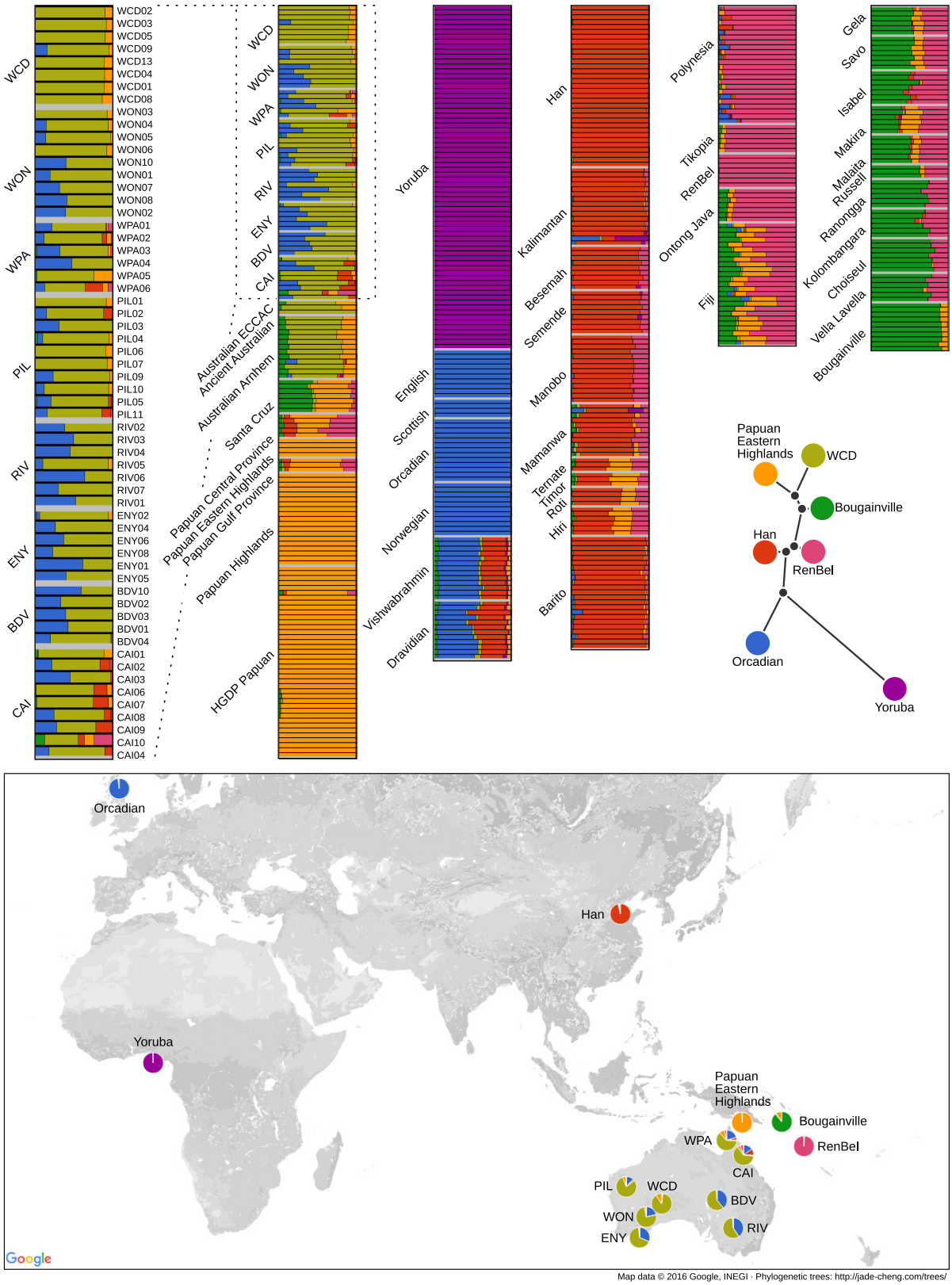
40. Tin, A. *et al.* Genome-wide association study for serum urate concentrations and gout among African Americans identifies genomic risk loci and a novel URAT1 loss-of-function allele. *Hum. Mol. Genet.* **20,** 4056–4068 (2011).
41. Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nat. Rev. Genet.* **13,** 745–753 (2012).
42. Fenner, J. N. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am. J. Phys. Anthropol.* **128,** 415–423 (2005).
43. Holt, S. *Palaeoenvironments of the Gulf of Carpentaria from the Last Glacial Maximum to the Present, as Determined by Foraminiferal Assemblages.* PhD thesis, Univ. Wollongong (2005).
44. Heupink, T. H. *et al.* Ancient mtDNA sequences from the First Australians revisited. *Proc. Natl Acad. Sci. USA* **113,** 6892–6897 (2016).
45. Horton, D. *The Encyclopedia of Aboriginal Australia.* (Australian Institute of Aboriginal and torres Strait Islander Studies, 1994).
46. Migliano, A. B. *et al.* Evolution of the pygmy phenotype: evidence of positive selection from genome-wide scans in African, Asian, and Melanesian pygmies. *Hum. Biol.* **85,** 251–284 (2013).
47. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513,** 409–413 (2014).
48. Wall, J. D. *et al.* Higher levels of Neanderthal ancestry in East Asians than in Europeans. *Genetics* **194,** 199–209 (2013).
49. Vernot, B. & Akey, J. M. Resurrecting surviving Neandertal lineages from modern human genomes. *Science* **343,** 1017–1021 (2014).
50. Fu, Q. *et al.* An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524,** 216–219 (2015).
51. Wang, C. *et al.* Comparing spatial maps of human population-genetic variation using Procrustes analysis. *Stat. Appl. Genet. Mol. Biol.* **9,** 13 (2010).
52. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7,** 214 (2007).

Supplementary Information is available in the online version of the paper.

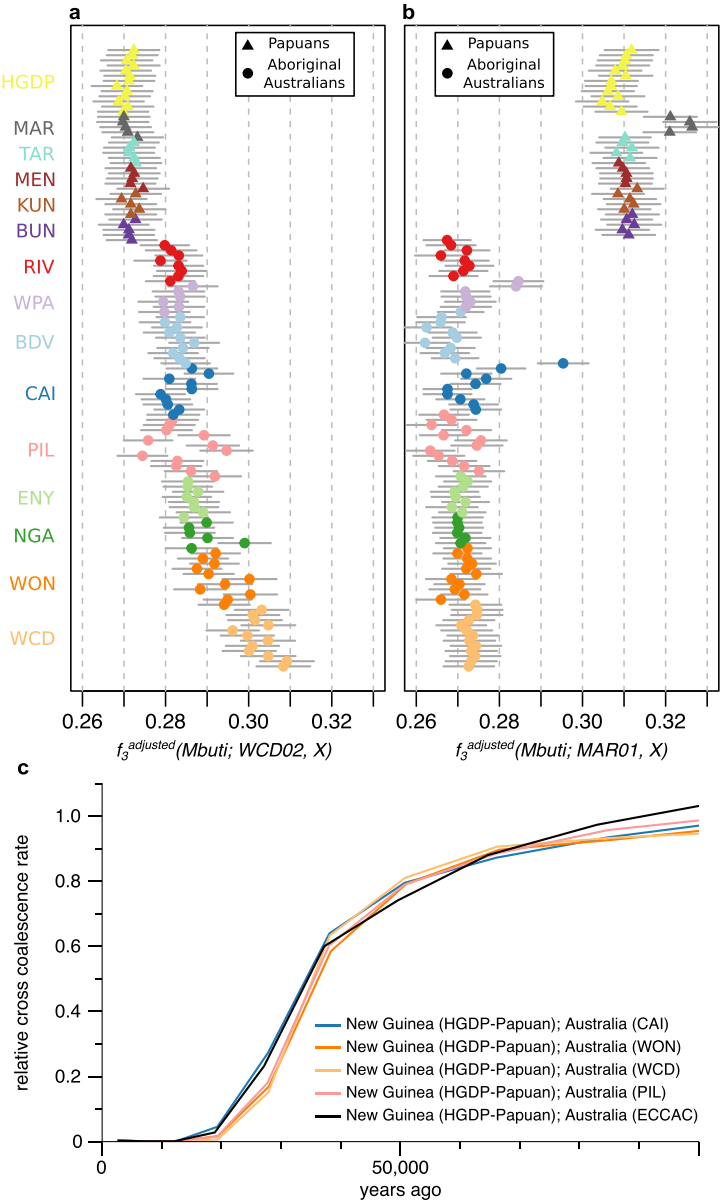
Acknowledgements We thank all sample donors for contributing to this study. We thank Macrogen (http://www.macrogen.com/) for sequencing of the Aboriginal Australian genomes, M. Rasmussen, C. Der Sarkissian, M. Allentoft, D. Cooper, R. Gray, S. Greenhill, A. Seguin-Orlando, T. Carstensen, M. Przeworski, J. D. Jensen and L. Orlando for helpful discussions. We thank E. Thorsby for sample collection and contributing the DNA extract for the P2077 genome, I. Lissimore for support with data storage and distribution. We thank T. Parks, K. Auckland, K. Robson, A. V. Hill, J. B. Clegg, D. Higgs, D. J. Weatherall and M. Alpers for assistance in sample collection and discussion. L.E., V.C.S., I.A., I.D. and S.P. are grateful to the High Performance Computation platform of the University of Berne for providing access to the UBELIX cluster. This work was supported by the Danish Research Foundation, the Lunbeck Lundbeck Foundation, and the KU2016 grant. A.-S.M. was supported by an ambizione grant with reference PZ00P3_154717 from the Swiss National Science Foundation (SNSF). M.C.W. was supported by the Australian Research Council (ARC) Discovery grants DP110102635 and DP140101405 and by a Linkage grant LP140100387. V.C.S., I.D. and S.P. were supported by SNSF grants to L.E. with references 31003A-143393 and CRSII3_141940. O.L. was supported by a Ramón y Cajal grant from the Spanish Ministerio de Economia y Competitividad (MINECO) with reference RYC-2013-14797 and by a BFU2015-68759-P (MINECO/FEDER) grant. I.A. was supported by a grant with reference SFRH/BD/73150/2010 from the Portuguese Foundation for Science and Technology (FCT). A.B., S.Sc., Y.X., C.T.-S. and R.D. were supported by a Wellcome Trust grant with reference WT098051. E.M., C.Ba., I.P., S.N. and M.S. acknowledge the Max Planck Society. S.Su. was supported by an ARC Discovery grant with reference DP140101405. J.L.W. was supported by a PhD scholarship from Griffith University. A.A. acknowledges the Villum foundation. I.M. was supported by a grant from the Danish Council for Independent Research with reference DFF–4090-00244. J.V.M.-M. acknowledges the Consejo Nacional de Ciencia y Tecnología (Mexico) for funding. N.B. and F.-X.R. were supported by the French Ministry of Foreign and European Affairs and French ANR with the grant ANR14-CE31-0013-01. S.B. was supported by a Novo Nordisk Foundation grant with reference NNF14CC0001. P.G. was supported by a Leverhulme Programme grant number RP2011-R-045 to A.B.M. at UCL Department of Anthropology and M.G.T. at UCL Department of Genetics, Evolution and Environment. A.J.M. was supported by a Wellcome Trust grant with reference 106289/Z/14/Z. M.M. acknowledges the EU European Regional Development Fund through the Centre of Excellence in Genomics to Estonian Biocentre; Estonian Institutional Research grant IUT24-1. A.B.M. was supported by a Leverhulme Programme grant number RP2011-R-045. M.G.T. was supported by a Wellcome Trust Senior Investigator Award with grant number 100719/Z/12/Z. S.J.O. was supported by a Wellcome Trust Core Award Grant Number 090532/Z/09/Z. A.Man. was supported by an ERC Consolidator Grant 647787 'LocalAdaptation'. M.E.P. would like to acknowledge the cardio-metabolic research cluster at Jeffrey Cheah School of Medicine & Health Sciences, Monash University Malaysia and Ministry of Science, Technology & Innovation, Malaysia for research grant 100-RM1/BIOTEK 16/6/2B. M.H.S. was supported by a grant from the Danish Independence Research Council with reference FNU 12-125062. R.A.F. was supported by the Leverhulme Trust. M.M.L. is supported by an ERC Advanced Grant 295907 'In-Africa'. C.Bo. was supported by USA National Science Foundation (NSF) grants BCS-0844550 and BCS-1423711, awarded to C.Bo. and Yale University. T.M. was supported by a grant from the Danish Independence Research Council with reference FNU 1323-00749. M.S.S. was supported by a Wellcome Trust grant with reference WT098051. L.E. was supported by Swiss NSF grant number 31003A-143393, D.M.L. was supported by ARC Discovery Grants DP110102635 and DP140101405 and Linkage grant LP140100387. E.W. is grateful to St John's College in Cambridge for help and support.

Author Contributions G.A., J.Y.C., J.E.C., T.H.H., E.M., S.P., S.R., S.Sc., S.Su. and J.L.W. contributed equally and are listed alphabetically in the author list; A.A., C.Ba., I.D., A.E., A.Mar., I.M. and I.P. contributed equally and are listed alphabetically in the author list; T.S.K., I.P.L., J.V.M.-M., S.N., F.R., M.Si. and Y.X. contributed equally and are listed alphabetically in the author list. E.W. and D.M.L. initially conceived and headed the project. L.E. led the genetic load and the SFS-based demographic analyses. M.S.S. headed the research at the Wellcome Trust Sanger Institute. A.-S.M. planned and coordinated the genetic analyses and the sequencing of the Aboriginal Australian genomes. C.M., J.L.W., T.H.H., P.F.C., W.C., G.F., D.I., B.L., A.L., P.J.M., L.M., D.R., T.W., C.W., J.D. and M.C.W. collaborated with local groups to collect Aboriginal Australian samples. N.B., P.G., G.K., M.L., A.J.M., A.B.M., W.P., F.-X.R., P.S., M.G.T. and S.J.O. collaborated with local groups to collect Papuan samples. S.E. collaborated with local groups to collect the Rapanui sample. A.Mar. extracted DNA for the Aboriginal Australian genomes. M.S.S., A.B. and C.T.-S. coordinated the design and sequencing of the Papuan genomes. O.L., V.C.S., I.A., A.-S.M., A.B., G.A., J.Y.C., J.E.C., T.H.H., E.M., S.P., S.R., S.Sc., S.Su., J.L.W., A.A., C.Ba., I.D., A.E., A.Man., I.M., I.P., T.S.K., I.P.L., J.V.M.-M., S.N., F.R., M.Si., F.A., S.B., L.E., J.D.W. and T.M. analysed genetic data. C.Bo. collected and analysed linguistic data. L.E., E.W., D.M.L., Y.X., M.E.P., C.T.-S., R.D., M.S.S., A.Man., M.H.S., T.M., M.St. and R.N. supervised genetic analyses. M.C.W., C.M., W.C., G.F., D.I., B.L., A.L., P.J.M., L.M., D.R., T.W., C.W., E.A.M.-S., M.M., M.E.P., S.J.O., J.D., A.B.M., R.A.F. and M.M.L. provided archaeological, anthropological and historical context. A.-S.M., V.C.S., O.L., I.A., A.B., M.M.L., R.N., L.E., D.M.L. and E.W. wrote the manuscript with critical input from G.A., T.H.H., E.M., S.Sc., S.Su., J.L.W., C.Ba., A.E., I.P., E.A.M.-S., M.S.S., S.J.O., C.T.-S., R.D., M.G.T., J.D., A.Man., M.H.S., R.A.F., C.Bo., J.D.W., T.M., M.St. and all other coauthors. A.-S.M., V.C.S., O.L., I.A. and A.B. revised and compiled the Supplementary Information.

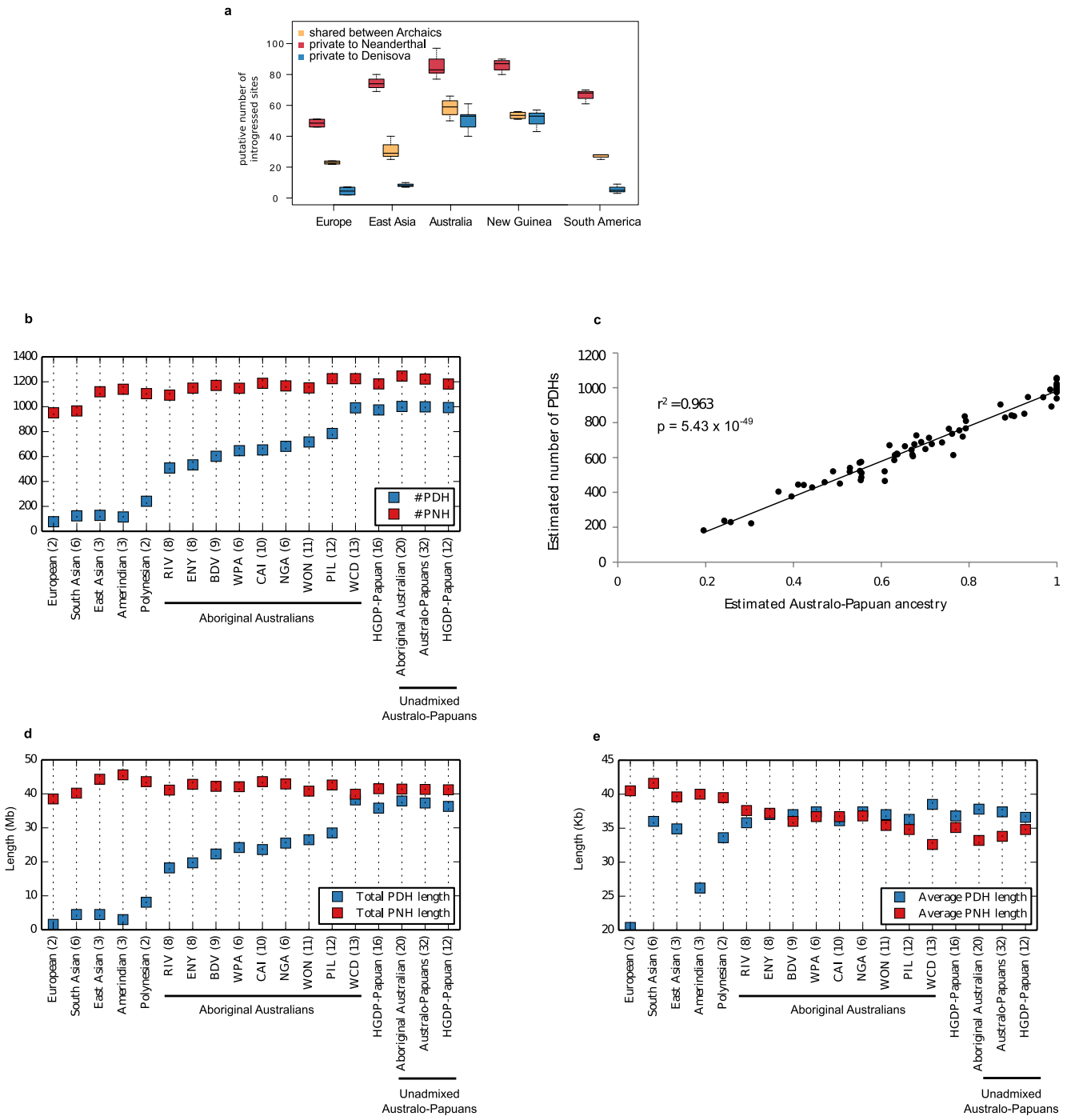
Author Information The Aboriginal Australian and Papuan whole genome sequenced data generated in this study have been deposited at the European Genome-phenome Archive (EGA, http://www.ebi.ac.uk/ega/), which is hosted by the EBI, under the accession numbers EGAS00001001766 and EGAS00001001247, respectively. The Papuan SNP array data generated in this study can be found under http://geogenetics.ku.dk/latest-news/alle_nyheder/2016/data. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.W. (ewillerslev@snm.ku.dk), D.M.L. (d.lambert@griffith.edu.au), L.E. (laurent.excoffier@iee.unibe.ch) and M.S.S. (ms23@sanger.ac.uk).



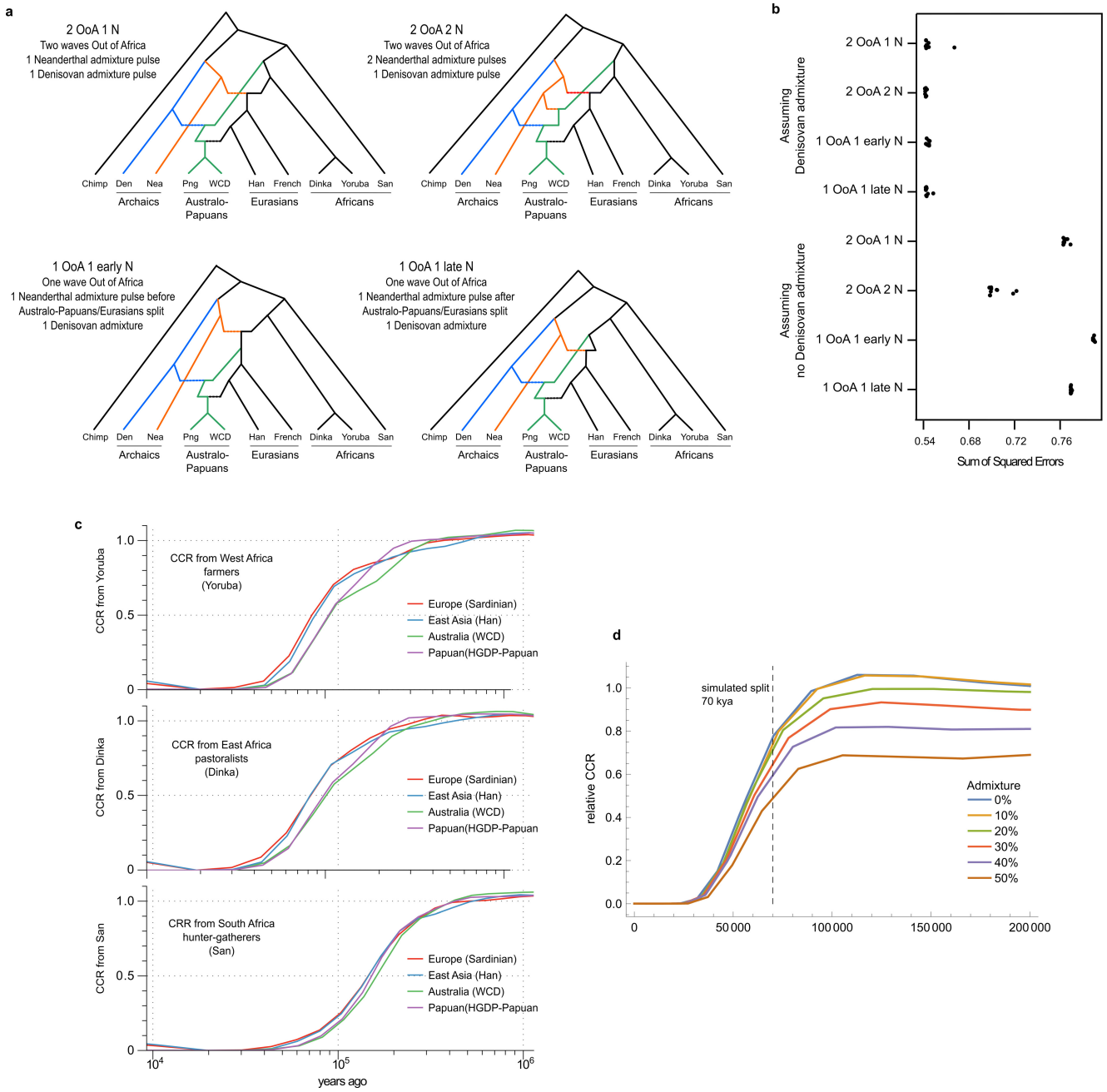
Extended Data Figure 1 | Per individual admixture proportions of $K = 7$ ancestral components including Aboriginal Australians, New Guineans, Europeans, Africans, Melanesians and Polynesians. The genome of each individual is depicted as a bar and is coloured according to the estimated genome-wide proportions of ancestry components. An unrooted tree showing the relationships between the identified ancestral components is also estimated by our method. Each ancestry has been labelled with the name of the population (see also map) showing the highest fraction of that ancestral component. The cross-validation error is minimized for this value of K for fivefold cross-validation (Supplementary Information section S05). The rooted tree supports the shared genetic origin of Aboriginal Australians, Papuans and Bougainvilleans. Note that only individuals with more than 50% of Aboriginal Australian ancestry in their genomes as defined in SOM06 were included in the analyses. Map data ©2016 Google, INEGI. Trees constructed with http://jade-cheng.com/trees/.



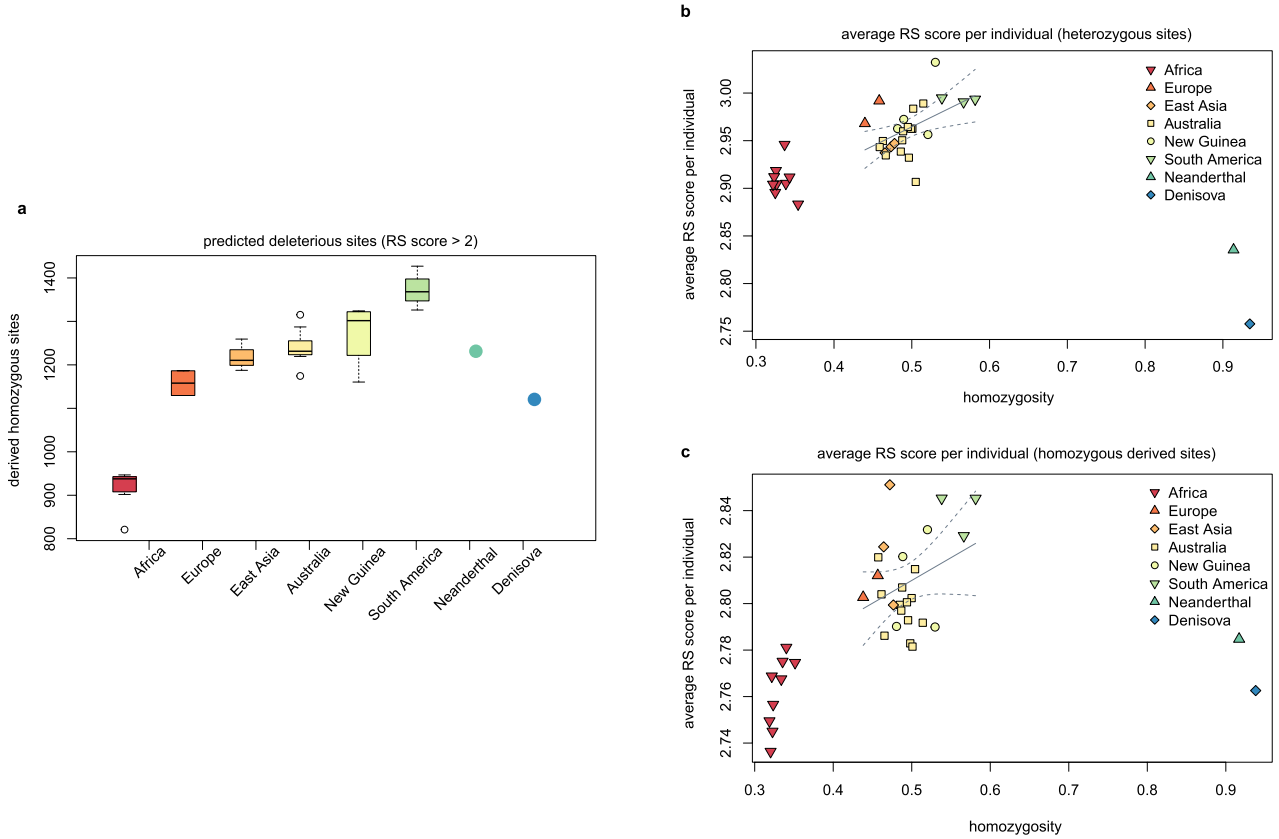
Extended Data Figure 2 | Genetic relationships of Aboriginal Australians and Papuans. a, Genetic affinities between a Western Central Desert (WCD02) genome and Aboriginal Australians and Papuans. Outgroup f_3 statistics between WCD02 and all other Aboriginal Australians and Highland Papuan individuals that were whole-genome sequenced for this study, using all genotypes called from the sequencing data. Because the widespread recent admixture in Aboriginal Australians has large confounding effects on the f_3 statistics, the values were adjusted using the slope coefficient from a simple linear regression model fitted to the relationship between f_3 and the fraction of non-indigenous (that is, not Aboriginal Australian nor Papuan) ancestry in each individual genome. The adjusted f_3 statistics display a genetic gradient that separates western and eastern Aboriginal Australian populations. However, we find no differences between Papuan population samples in their level of Aboriginal Australian affinity (Kruskal–Wallis test, $P = 0.083$). Horizontal lines correspond to ±1 standard error. **b**, Genetic affinities between a Papuan highlander genome and Aboriginal Australians and Papuans. The Papuan highlander sample MAR01 from the Marawaka area was arbitrarily chosen as a reference point for this analysis. f_3 values were adjusted for recent admixture as in **a**. All Aboriginal Australian groups display a similar level of Highland Papuan affinity (with the exception of three outlier individuals from the north-eastern WPA and CAI populations: WPA06, WPA05 and CAI10, the latter two of which are known to have at least one parent with origins in Papua New Guinea or the Torres Strait Islands). While some differences between groups are actually statistically significant (Kruskal–Wallis test, $P = 0.0002$, after removing the three outliers), which could be consistent with, for example, low levels of Papuan gene flow into some Aboriginal Australian groups (see Supplementary Information sections S06 and S07), we caution that some of these differences are probably due to imperfect adjustment for Eurasian admixture (the adjusted f_3 is highest in the WCD population, which has the least Eurasian admixture). Horizontal lines correspond to ±1 standard error. **c**, MSMC analyses. Linear interpolation through the midpoints of the time intervals of the relative cross coalescence rate estimates from MSMC[25] using pairs of individuals including one HGDP-Papuan and one other individual as indicated. We used CAI01, PIL06, WCD01, WON03 and an ECCAC sample for this analysis (see Supplementary Information section S08 for details). The MSMC results were scaled using a mutation rate of 1.25×10^{-8} per generation per site as suggested in ref. 41 and a generation time of 29 years corresponding to the average hunter–gatherer generation interval for males and females[42].



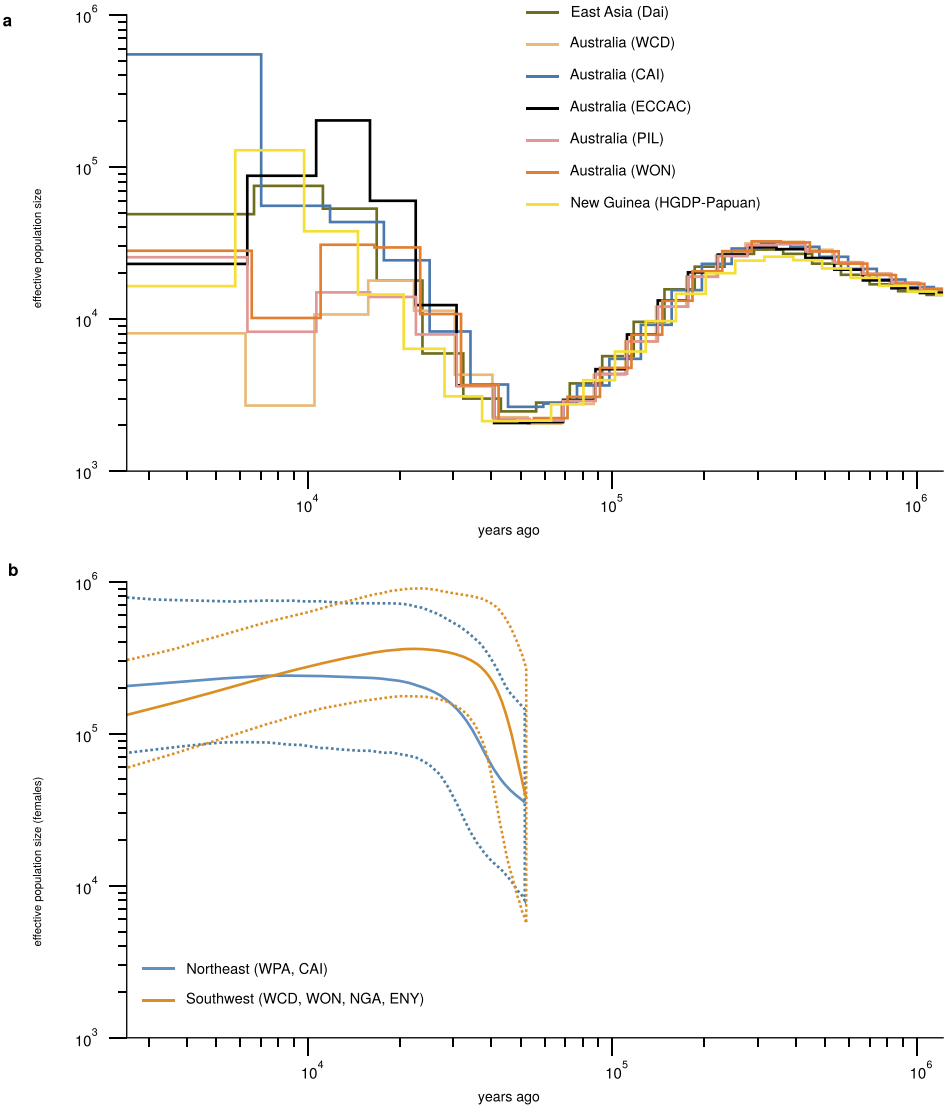
Extended Data Figure 3 | Introgressed archaic sites and putative Denisovan and Neanderthal haplotypes. a, Distribution of per individual number of putative introgressed sites from archaic humans. The number of Neanderthal-specific introgressed sites increases from Europe to Australia, and then decreases in Amerindians, which is consistent with recurrent Neanderthal (or Neanderthal-related archaic) gene flow during the expansion into Eurasia. Our results are thus indicative of several pulses of Neanderthal gene flow into modern humans, as inferred previously[48–50]. We note, however, that the apparent high levels in Neanderthal-specific introgressed sites in Australo-Papuans can be explained by the expected number of misclassified Neanderthal introgressed sites resulting from the shared ancestry with Denisovans (see Supplementary Information section S10 for details). **b**–**e**, Putative Denisovan (PDH) and Neanderthal haplotypes (PNH). The putative haplotypes correspond to clusters (four or more SNPs spanning at least 4 kb) of heterozygous or homozygous genotypes in complete linkage disequilibrium ('diplotypes') that are potentially the result of Neanderthal or Denisovan admixture. Those diplotypes are homozygous ancestral in 10 Africans, homozygous derived in the Denisovan for the PDH (respectively Neanderthal for the PNH), homozygous ancestral in the Neanderthal for the PDH (respectively Denisovan for the PNH), and with the derived allele segregating in all other contemporary non-African humans (see Supplementary Information section S11 for details). We report the average number of the PDHs and PNHs (**b**), the correlation between the estimated amount of Australo-Papuan ancestry (see Fig. 2b, Extended Data Fig. 1, Supplementary Information section S05) and the number of identified PDHs for each Australian sample (**c**), the sum of the lengths (**d**) and the average length (**e**) of the PDHs and PNHs per individual for worldwide populations included in our reference panel (see Supplementary Information section S03).



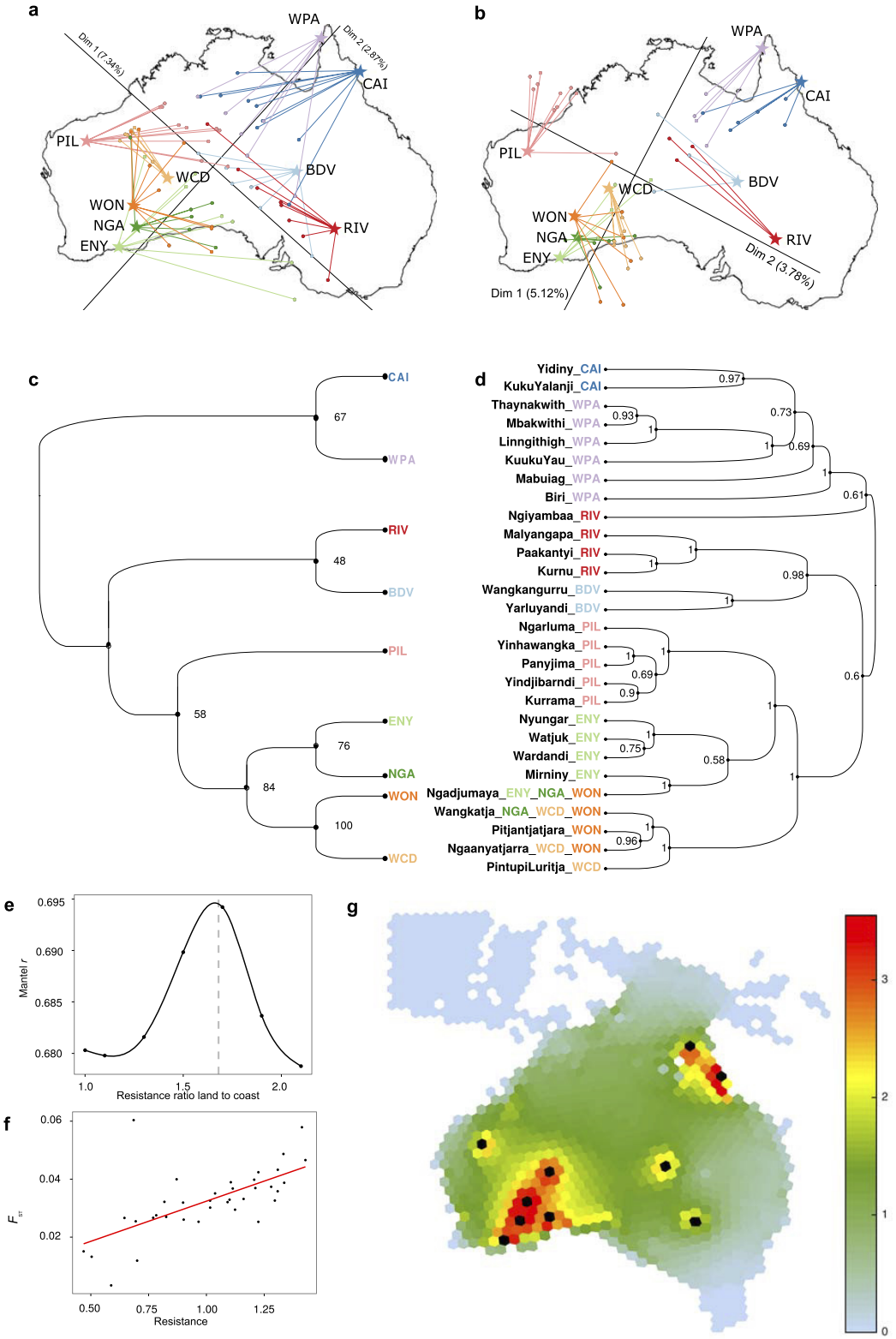
Extended Data Figure 4 | Out of Africa: admixture graphs based on D-statistics and MSMC analyses. a, Admixture graphs representing some of the topologies considered for the two waves and one wave Out of Africa models assuming Denisovan admixture. All topologies are identical except for the coloured lineages representing Australo-Papuans (green), Neanderthal (Nea, orange) and Denisovan (Den, blue). The graphs differ in (1) the number of OoA events, and (2) the number of Neanderthal admixture pulses. Png, HGDP-Papuan. **b**, Sum of squared errors between the observed D-statistics and the expectations for each quartet in the graph involving the chimpanzee as an outgroup for each of the admixture graphs shown in **a** and the corresponding four without Denisovan admixture. Each point is the result of the optimization procedure with different starting points. See Supplementary Information section S09 for details. **c**, MSMC analyses. Relative cross coalescence rate (CCR) estimates from MSMC[25] for pairs of individuals including one African sample (Yoruba, Dinka and San) and one other sample from Eurasia, as indicated in the legend. **d**, Simulation study to assess the effect of archaic admixture on the CCR rates. Relative CCR estimated for data simulated under a simple two-population divergence model where one of the populations admixed at different rates with an archaic population. See Supplementary Information section S08 for details.



Extended Data Figure 5 | Inferred deleterious mutations. a, Box plot of the number of derived homozygous sites per individual for worldwide populations that are predicted to be deleterious. Deleteriousness of SNPs was inferred using genomic evolutionary rate profiling (GERP) rejected substitution scores. Derived alleles with a rejected substitution score larger than 2 were considered to be deleterious, see Supplementary Information section S11. **b**, **c**, Average rejected substitution score per individual calculated across heterozygous sites (**b**), and derived homozygous sites (**c**). Each coloured symbol corresponds to estimates from a single individual. Homozygosity is calculated as the number of derived homozygous sites divided by the number of sites at which an individual carries at least one copy of the derived allele. Solid lines show the linear regression of homozygosity against average rejected substitution score per individual for non-African modern humans. Dashed lines indicate the 95% confidence interval for the linear regression. See Supplementary Information S11 for details.

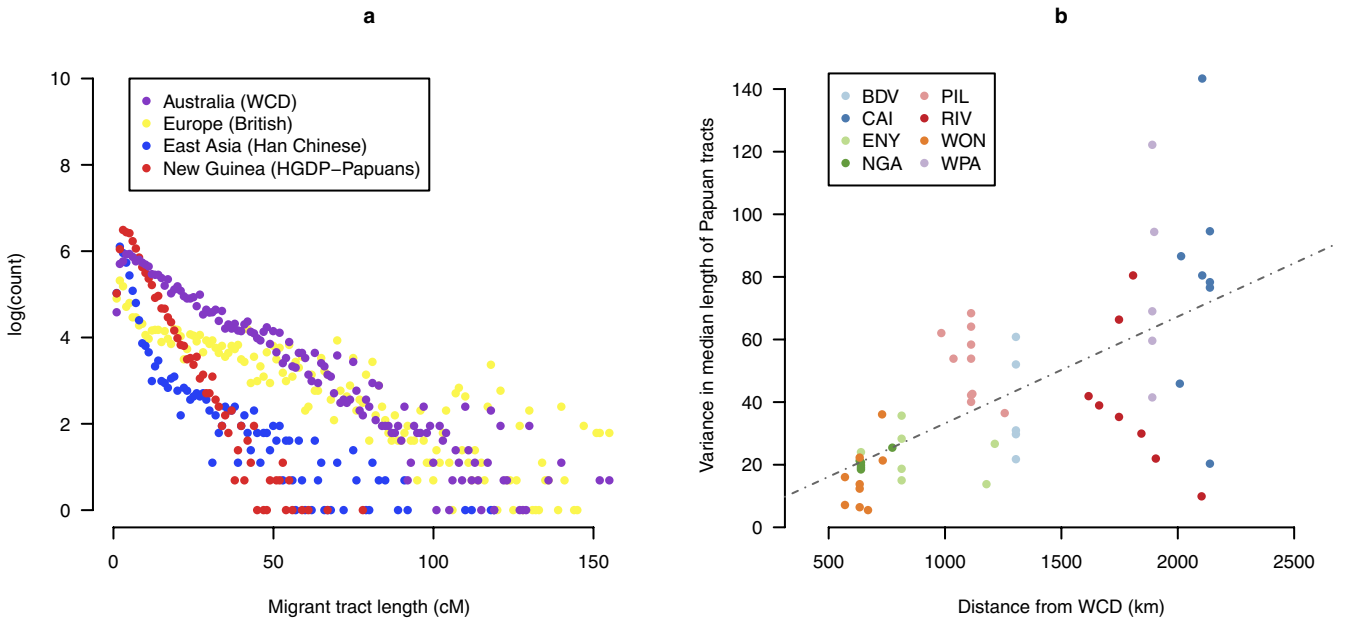


Extended Data Figure 6 | Effective population size changes over time. **a**, MSMC analyses. Population size estimates from MSMC for pairs of individuals from several populations within and outside of Australia. For each run, we used two individuals from each population, that is, four haplotypes in each run. MSMC results were scaled as in Fig. 3. **b**, Bayesian skyline plots (BSP) calculated from the mtDNA genome sequences, showing the effective population size estimates over time when considering either groups from northeastern Australia (CAI, WPA) or groups from southwestern Australia (ENY, NGA, WCD, WON). Solid lines are the estimates, dashed lines are the corresponding 95% credible intervals (see Supplementary Information section S12).



Extended Data Figure 7 | See next page for caption.

Extended Data Figure 7 | Genetics mirrors geography and languages. **a**, **b**, Procrustes analyses of the first two dimensions of a classical multidimensional scaling (MDS) analysis of the Aboriginal Australian genome sequences (autosomes). We considered two cases: an analysis including all variants (**a**), or only the variants remaining after genomic regions of putative recent European and East Asian origin are 'masked' (**b**, Supplementary Information section S06). Both MDS plots have been rotated towards the best overlap with geographic sampling locations as defined by Procrustes analysis[51]. In each plot, the arrows indicate the error of the MDS coordinates towards the assigned population sampling geographic coordinates. We find that the genetic relationships within Australia mirrors geography, with a significant correlation for both cases, that is, $r_{GEN,GEO} = 0.59$, $P < 0.0005$ for all variants and even higher ($r_{GEN,GEO} = 0.77$, $P < 0.0005$) for the masked data. We find using the bearing correlogram approach that the main axis of genetic differentiation in the masked Aboriginal Australian genomes is at angle = 65° compared to the equator, that is, in the southwest to northeast direction (Supplementary Information section S13). **c**, **d**, Correspondence between genetics and linguistics. Unrooted neighbour-joining F_{ST}-based genetic tree (cladogram). Weir and Cockerham F_{ST} distance was computed between the Aboriginal Australian populations after masking the Eurasian tracts. Statistical robustness of each branch was estimated by means of a bootstrap analysis (1,000 replicates, Supplementary Information section S05). **d**, Bayesian phylogenetic tree for the 28 different Pama–Nyungan languages represented in this sample (from ref. 13, see Supplementary Information section S15). Posterior probabilities are also indicated. Note that one language group can be shared by different Aboriginal Australian groups. The linguistic tree was built with BEAST[52]. **e**–**g**, Gene flow across the continent. **e**, Mantel non-parametric *r* (estimating the goodness of fit between genetic differentiation and connectivity) versus ratios of resistance of inland to coastal nodes, showing a peak at 1.7. **f**, Best fit of pairwise population genetic differentiation, F_{ST} (computed between the nine Aboriginal Australian groups after masking Eurasian tracts (Supplementary Information section S06)), versus pairwise connectivity based on the environment (estimated as resistance) when moving inland is 1.7 times harder than moving along coastal nodes. **g**, Gene flow across the Australian landscape, quantified as the cumulative current for pairwise connections among Aboriginal Australian groups (black circles), with larger current (warmer colours) representing greater gene flow.



Extended Data Figure 8 | European, East Asian and Papuan genomic tracts in Aboriginal Australians. a, Distribution of the tracts assigned to Aboriginal Australian (WCD), Papuan, East Asian or European ancestry for 58 unrelated non-WCD Aboriginal Australian samples. Most of the shorter tracts were of Papuan origin, suggesting that a large fraction of the Papuan gene flow is much older than that from Europe and East Asia, consistent with a Papuan influence spreading slowly from northeastern to southwestern Australia by ancient migration. **b**, Corresponding scatter plot with fitted line of per-individual variance in Papuan tract length versus geographic distance from WCD, the latter calculated using the great-circle distance formula for pairs of individual GPS coordinates. Papuan tract distribution showed a strong and significant correlation with distance from WCD ($r = 0.64$; $P < 1 \times 10^{-5}$), with 'younger tracts' (that is, with a larger variance) closer to New Guinea and 'older tracts' (that is, with a smaller variance) closer to WCD. This is also consistent with continuous Papuan gene flow spreading from the northeast.

Extended Data Table 1 | Whole genome sequence depth of coverage, haplogroup and language assignments for the Aboriginal Australian samples

indiv.	DoC*	mtDNA haplotype†	Ychr haplotype‡	Pama-Nyungan language§
BDV01	78	S2	-	Yarluyandi Wangkangurru
BDV02	75	S1a	R1b1a2a1a2c1g2a1a2	Yarluyandi Wangkangurru
BDV03	-	-	-	Yarluyandi Wangkangurru
BDV04	70	O1a	-	Yarluyandi Wangkangurru
BDV05	72	S1a	O1a	Yarluyandi Wangkangurru
BDV06	70	S1a	-	Yarluyandi Wangkangurru
BDV07	70	O1a	-	Yarluyandi Wangkangurru
BDV08	70	S1a	R1b1a2a1a2c1g2a1	Yarluyandi Wangkangurru
BDV09	74	S1a	-	Yarluyandi Wangkangurru
BDV10	72	S1a	I1a2a1d	Yarluyandi Wangkangurru
CAI01	84	P	K2b	Yidiny
CAI02	74	M42	K2b	Yidiny
CAI03	77	M42a	-	Yidiny
CAI04	71	P	-	Yidiny KukuYalanji
CAI05	80	P	O2a1a	Yidiny
CAI06	78	P	C1b	Yidiny
CAI07	71	N13	K2b	KukuYalanji
CAI08	70	P	K2b	Yidiny
CAI09	79	P	R1b1a2a1a2b1	Yidiny
CAI10	73	E1a2	K2b	-
ENY01	69	H1e1a3	R1b1a2a1a2b1c1	Nyungar
ENY02	79	R12	-	Ngadjumaya
ENY03	83	O	-	Mirniny
ENY04	83	M42	-	Nyungar
ENY05	78	S2	-	Ngadjumaya
ENY06	70	M42	-	Wardandi
ENY07	73	S2	E1b1b1b2a	Watjuk
ENY08	71	P4b1	C1b	Nyungar Ngadjumaya
NGA01	74	O1	-	Ngadjumaya
NGA02	52	O1a	-	Ngadjumaya
NGA03	73	O	-	Ngadjumaya
NGA04	75	O	R1b1a2a1a1b1a1a	Wangkatja
NGA05	56	R12	-	Ngadjumaya
NGA06	63	S1a	-	Wangkatja
PIL01	58	R	C1b	Yinhawangka
PIL02	61	M42	C1b	Yinhawangka
PIL03	56	M42	-	Yinhawangka
PIL04	64	M42	-	Yinhawangka
PIL05	68	M42	C1b	Yinhawangka
PIL06	59	O1	K2b	Panyjima
PIL07	63	O	-	Panyjima
PIL08	72	M42	C1b	Yindjibarndi Kurrama
PIL09	58	S5	R1b1a2a1a2e1	Kurrama
PIL10	61	R	-	Yinhawangka
PIL11	57	P3b	C1b	Kurrama
PIL12	63	P3b	C1b	Yindjibarndi
RIV01	73	M42a	-	Ngiyambaa
RIV02	62	P4b1	-	Paakantyi
RIV03	69	M42a	-	Paakantyi
RIV04	62	P4b1	I2a1a2a1a	Kurnu
RIV05	72	P4b1	-	Paakantyi
RIV06	66	H1bs	J2a1b	Ngiyambaa
RIV07	70	P4b1	R1b1a2a1a2c1c	Paakantyi
RIV08	66	P4b1	-	Paakantyi Malyangapa
WCD01	62	R12	K2b	Ngaanyatjarra
WCD02	59	S1a	C1b	Ngaanyatjarra
WCD03	61	R12	K2b	Wangkatja
WCD04	52	P3b	K2b	Ngaanyatjarra
WCD05	60	O1	C1b	Ngaanyatjarra
WCD06	58	O1a	C1b	Ngaanyatjarra
WCD07	61	M42	-	Ngaanyatjarra
WCD08	64	M42	-	Ngaanyatjarra
WCD09	59	R	J2a1b	Ngaanyatjarra
WCD10	63	M42	-	Ngaanyatjarra
WCD11	57	M42	K2b	Ngaanyatjarra
WCD12	59	M42	C1b	Ngaanyatjarra PintupiLuritja
WCD13	67	M14	C1b	Ngaanyatjarra
WON01	71	O	I1a2a1a3a	Wangkatja
WON02	101	O1a	-	Wangkatja
WON03	65	O1a	-	Wangkatja
WON04	58	R	-	Ngaanyatjarra
WON05	56	O1a	I2a2a1a2a2	Wangkatja
WON06	60	R12	-	Wangkatja
WON07	57	O	-	Ngadjumaya
WON08	52	O	-	Wangkatja
WON09	20	O	E1b1b1a1b1a4	Wangkatja
WON10	50	O1	R1b1a2a1a2a	Wangkatja
WON11	58	R12	-	Pitjantjatjara
WPA01	51	P5	-	Thaynakwith Linngithigh
WPA02	50	P	C1b	Mpakwithi Kaanju
WPA03	51	M42a	K2b	Thaynakwith Biri
WPA04	52	P5	-	Thaynakwith KukuYau
WPA05	56	M42	NA	Mabuiag Thaynakwith
WPA06	53	P5	O1a	Mpakwithi

*The depth of coverage (DoC) is the average number of reads covering every position in the genome (hg19) after duplicate removal (see Supplementary Information section S03).
†The average depth of coverage on the mitochondrial genome (mtDNA) is 3,484 ± 1,515 (mean ± s.d.) and haplogroups were called with haplogrep (http://haplogrep.uibk.ac.at/) and haplofind (https://haplofind.unibo.it/), see Supplementary Information section Supplementary Information section S12 for details and references.
‡The average depth of coverage on the Y chromosome (Ychr) is 28.88 ± 4.5 (mean ± s.d.). Haplogroup assignment was performed with an in-house script that matched our SNPs with the classification provided in ISOGG version 10.08, see Supplementary Information section Supplementary Information section S12 for details and references.
§Language group with which the speaker self-identifies, or to which they were assigned. Where more than one language is given, speakers either identified with more than one group, or they could not be assigned to a single group with certainty.

Extended Data Table 2 | Selection scan in Aboriginal Australians

Focal Pop	Nearby Gene*	Position†	rsID	Dist ‡	*PBSn*§	F_{12} ¶	F_{13}	F_{23}	Function of gene product#
All	*TMEM86B*	55,833,076	rs734517	92,444	0.78	0.93	0.99	0.06	Catalyzes the degradation of lysoplasmalogen. Modulates cell membrane proteins.
All	*LRRC52*	165,621,695	rs4147601	88,510	0.74	0.96	0.91	0.01	Modulates voltage of potassium ion channels. Expressed in testis.
All	*MACROD2*	15,209,684	rs175279	901	0.70	0.92	0.89	-0.01	Involved in deacetylase activity. Possibly (but not conclusively) causative of Kabuki syndrome.
All	*JRKL*	96,747,146	rs72959058	507,105	0.74	0.99	0.87	0.15	Homologue to "jerky" gene in mouse.
All	*SPATA20*	48,631,324	rs73338243	287	0.70	0.96	0.85	0.09	Spermatid protein.
All	*NAA60*	3,537,933	rs73503305	970	0.71	0.91	0.91	-0.02	Histone acetyltransferase required for nucleosome assembly and chromosome segregation during anaphase. Human-specific imprinted gene.
All	*CBLN2*	70,019,066	rs12455116	184,848	0.69	0.92	0.87	0.00	*CBLN2*: cerebellum-specific protein involved in various signaling pathways. Possibly associated with pulmonary arterial hypertension.
	NETO1			390,482					*NETO1*: brain-specific transmembrane protein involved in the regulation of neuronal circuitry. Associated with thyroid function.
All	*SLC2A12*	134,391,056	rs4896021	17,267	0.76	0.96	0.95	-0.01	Catalyzes sugar absorption. Involved in the pathogenesis of diabetes. Associated with serum urate levels.
All	*LOC101927657*	127,358,509	rs145200081	16,731	0.65	0.94	0.80	0.13	Unknown (ncRNA).
All	*LOC102724612*	64,466,486	rs113341339	78,446	0.73	0.91	0.95	0.00	Unknown (ncRNA).
NE	*ZBTB20*	114,530,679	rs9289004	10,658	0.55	0.65	0.82	0.07	Transcriptional repressor associated with Primrose syndrome.
NE	*ANXA10*	168,646,016	rs2176513	367,671	0.49	0.61	0.61	-0.01	Calcium-dependent phospholipid-binding annexin.
NE	*TRPC3*	122,905,041	rs4502701	32,132	0.50	0.59	0.64	-0.01	Non-selective cation channel, associated with spinocerebellar ataxia.
NE	*HS3ST1*	11,634,592	rs7665516	204,055	0.45	0.45	0.71	0.07	Regulates rate of generation of anticoagulant heparan sulfate proteoglycan.
NE	*MIR548C*	65,027,511	rs2620721	11,126	0.50	0.55	0.73	0.03	Unknown (microRNA).
NE	*STARD13*	33,799,901	rs7318080	19,714	0.49	0.54	0.83	0.20	Involved in cell proliferation and fibroblast morphology.
NE	*AKAP11*	42,931,386	rs7319267	33,983	0.53	0.56	0.85	0.13	Directs protein kinase A activity and is involved in cAMP messenger signaling.
NE	*AGMO*	15,212,231	rs35557899	27,711	0.47	0.51	0.68	0.01	Catalyzes the cleavage of O-alkyl bonds of ether lipids.
NE	*RUNX1T1*	92,925,296	rs11776341	41,898	0.45	0.56	0.54	0.00	Involved in transcriptional repression. A translocation involving this gene is associated with acute myeloid leukemia.
NE	*FHAD1*	15,680,451	rs2473358	971	0.45	0.60	0.52	0.00	Unknown.
SW	*KCNJ2*	68,190,552	rs35167900	14,369	0.57	0.61	0.93	0.22	Potassium channel, associated with familial atrial fibrillation and periodic paralysis.
SW	*TACC2*	123,754,065	rs10159998	5,062	0.50	0.60	0.67	0.00	Belongs to a family of proteins that interact with the centrosome and microtubules, and that are implicated in cancer.
SW	*LOC101928708*	87,228,164	rs4843556	17,556	0.58	0.65	0.86	0.07	Unknown (ncRNA).
SW	*C16orf82*	27,187,689	rs72782349	107,202	0.51	0.60	0.69	0.02	Unknown.
SW	*LOC100507391*	194,520,805	rs56379930	17,908	0.55	0.66	0.75	-0.01	Unknown (ncRNA).
SW	*HAUS4*	23,416,252	rs2008951	127	0.49	0.50	0.83	0.16	A component of a microtubule-binding complex that plays a role in the generation of microtubules in the mitotic spindle.
SW	*KNG1*	186,438,819	rs5029990	815	0.51	0.56	0.72	0.01	During the inflammatory response, it is involved in vasodilation, coagulation, enhanced capillary permeability and pain induction.
SW	*MYDGF*	4,657,016	rs66891175	540	0.55	0.61	0.88	0.16	Unknown.
SW	*MSMP*	35,757,075	rs1951432	2,801	0.48	0.47	0.88	0.27	May be involved in the tumorigenesis of prostate cancer.
SW	*VAV2*	136,756,316	rs2519771	29,762	0.47	0.51	0.73	0.07	Member of an oncogene family. Involved in T-cell receptor signaling.

Top 10 peaks of differentiation from genome scans of all Aboriginal Australians combined (All) and two Aboriginal Australians subgroups living in different ecological regions in Australia.
*RefSeq protein coding gene with exon boundary near to windowed-*PBSn1* peak.
†Genomic position (hg19) of SNP with highest value of *PBSn1* within 200 Mb of the top window.
‡Distance between SNP and the nearest exon boundary of nearest gene.
§*PBSn1* statistic for top SNP.
¶F_{ST} statistics at top SNP for each comparison within the *PBSn1* calculation.
#Please see Supplementary Information section S16 for references.
*RefSeq protein coding gene with exon boundary near to windowed-*PBSn1* peak.

Ohana's application on Danish genetic history

I participated in a study of the Danish population's genetic history, in which approximately 800 students from 36 Danish high schools took part. In the early stage of this project, I used a range of web technology to retrieve and visualize data and analysis results stored at 23andme, a California-based personal genomics and biotechnology company. It involved several programming languages and various APIs: Python, C#, PHP, JavaScript, HTML and CSS, OAuth 2.0, RESTful services, Google's Maps API, Google's Visualization API, etc. In the later stage of this project, I helped to analyze the data using Ohana's admixture and population tree modules.

Nationwide genomic study in Denmark reveals remarkable population homogeneity

Georgios Athanasiadis[1,2]*, Jade Cheng[1], Bjarni J. Vilhjálmsson[1,3], Frank G. Jørgensen[4], Thomas D. Als[5,6,7], Stephanie Le Hellard[8,9], Thomas Espeseth[10,11], Patrick F. Sullivan[12,13,14], Christina M. Hultman[12], Peter C. Kjærgaard[2,15,17], Mikkel H. Schierup[1,2,16], Thomas Mailund[1,2]

[1]Bioinformatics Research Centre, Aarhus University, Aarhus, 8000, Denmark

[2]Centre for Biocultural History, Aarhus University, Aarhus, 8000, Denmark

[3]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA

[4]Tørring Gymnasium, Tørring, 7160, Denmark

[5]Department of Biomedicine, Aarhus University, Aarhus, 8000, Denmark

[6]The Initiative for Integrative Psychiatric Research (iPSYCH), Aarhus University, Aarhus, 8000, Denmark

[7]Center for Integrative Sequencing, iSEQ, Aarhus University, Aarhus, 8000, Denmark

[8]Dr E. Martens Research Group of Biological Psychiatry, Center for Medical Genetics and Molecular Medicine, Haukeland University Hospital, Bergen, 5021, Norway

[9]NORMENT, KG Jebsen Centre for Psychosis Research, Department of Clinical Science, University of Bergen, Bergen, 5021, Norway

[10]NORMENT, KG Jebsen Centre for Psychosis Research, Department of Psychology, University of Oslo, Oslo, 0424, Norway

[11]NORMENT, KG Jebsen Centre for Psychosis Research, Division of Mental Health and Addiction, Oslo University Hospital, Oslo, 0317, Norway

[12]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, 17177, Sweden

[13]Department of Genetics, University of North Carolina, Chapel Hill, NC 27599-7264, USA

[14]Department of Psychiatry, University of North Carolina, Chapel Hill, NC 27599-7264, USA

[15]Department of Culture and Society, Aarhus University, Aarhus, 8000, Denmark

[16]Department of Bioscience, Aarhus University, Aarhus, 8000, Denmark

[17]Present address: The Natural History Museum of Denmark, University of Copenhagen, Copenhagen, 1471, Denmark

*E-mail: athanasiadis@birc.au.dk

Abstract

Denmark has played a substantial role in the history of Northern Europe. Through a nationwide scientific outreach initiative, we collected and used genetic and anthropometrical data from ~800 high school students to elucidate the genetic makeup of the Danish population and to assess polygenic risk predictions of phenotypic traits in adolescents. We observed remarkable homogeneity across different geographic regions (e.g. average F_{ST} = 0.0002), although we could still detect weak signals of genetic structure (e.g. median distance between genomic relatives < 100 km). Denmark received substantial admixture contributions primarily from neighboring countries with overall influence of decreasing weight from Britain, Sweden, Norway, Germany and France. A Polish admixture signal was detected in Zealand coinciding with historical evidence of Wend settlements in the south of Denmark. We also observed considerably diverse demographic histories among Scandinavian countries, with Denmark having the smallest effective population size compared to Norway and Sweden. Finally, we found that polygenic prediction of self-reported adolescent height in the population was remarkably accurate (R^2 = 0.639±0.015). The high homogeneity of the Danish population could render population structure a lesser concern for the upcoming large-scale gene-mapping studies in the country.

Author summary

Denmark's genetic history has never been studied in detail. In this work, we analyze genetic and anthropometrical data from ~800 Danish students as part of an outreach activity promoting genomic literacy in secondary education. DNA analysis revealed remarkable homogeneity of the Danish population after discounting contributions from recent immigration. This homogeneity was reflected in PCA and AMOVA, but

also in more sophisticated LD-based methods for estimating admixture. Notwithstanding Denmark's homogeneity, we observed a clear signal of Polish admixture in the East of the country, coinciding with historical Polish settlements in the region before the Middle Ages. In addition, Denmark has a substantially smaller effective population size compared to Sweden and Norway, possibly reflecting further lack of strong population structure. None of these three Scandinavian countries seems to have suffered a depression due to the Black Death in the Middle Ages. Finally, we used the students' genetic data to predict their adult height after training a novel predication algorithm on public summary statistics from large GWAS. We validated our prediction using the students' self-reported height and found that we could predict height with a remarkable ~64% accuracy.

Introduction

In recent years, there has been an explosion of human genetic studies, which – aided by a variety of technological and methodological advancements – have contributed substantially to the characterization of patterns of intercontinental [1], intracontinental [2,3] and subcontinental [4] genetic variation; the reconstruction of population history in regions with poor/nonexistent historical records [5,6]; the study of local and global patterns of admixture in multiethnic societies [7]; and the study of admixture with other hominin species and its use in elucidating human dispersals [8,9].

The increased power of high-throughput genotype data and sophisticated computational methods, together with the diminishing cost of the former [10], has also boosted the emergence of single-country genomic projects in Europe [11–15] and the release of valuable data and results to the public [16]. Even though there is great

diversity in the objectives, data types and sample characteristics of these projects, all of them address to some extent the genetic structure of the region under study.

In this work, we extend the collection of single-country genetic studies in Europe by adding a new project from Denmark – a country whose area is comparable to that of the Netherlands but with a three times smaller population. Unlike previous genomic projects involving Denmark [17,18], ours was conceived from the beginning as a scientific outreach initiative with benefits for the general public and our research objectives. We invited ~800 high school students from across Denmark to participate in outreach activities whose primary goal was promoting genomic literacy in secondary education (G. Athanasiadis, personal communication). Most participants donated a DNA sample, which we used to explore the extent to which recent and more distant historical events left their mark on the genetic makeup of the Danish population.

With this work, we ultimately report back to our DNA donors the invaluable genetic insights we gained from analyzing the data. Our results showed remarkable homogeneity across different geographic regions in Denmark, though we were still able to detect weak signals of genetic structure. Denmark received substantial admixture contributions primarily from neighboring countries with overall influence of decreasing weight from Britain, Sweden and Norway. We also found evidence of considerably diverse demographic histories among the Scandinavian countries, as reflected in their historical effective population size. Finally, we found that height can be predicted with remarkable accuracy in the Danish population.

Results

Principal component analysis

To put Danes in a European genetic context, we first ran PCA on 3,858 samples from across Europe. In particular, we extended a previously published PCA within Europe [2] to include sizeable samples from Denmark, Norway and Sweden, as these were underrepresented in the original study. In this analysis, Denmark was represented by 407 individuals who had all four of their grandparents born in the country. Our Danish samples clustered in a geographically meaningful manner along the first two principal axes, partially overlapping with Norwegians and Swedes, and showing close genetic proximity to samples from Great Britain, the Netherlands, Germany and Poland (Figure 1A).

We then looked for fine-scale genetic structure within Denmark by focusing the analysis on 131 samples who had all four of their grandparents born in just one of the following six regions: Capital Region; Zealand; Funen; South, Central and North Jutland (Figure 1B). These six groups roughly correspond to the five administrative regions of Denmark (we further split the region of South Denmark into South Jutland and Funen). After repeating PCA for Denmark, Sweden, Norway and Germany alone (N = 1,168), we observed no geographically meaningful clustering of the 131 samples (Figure 1B). This lack of strong genetic structure was also supported by the low average F_{ST} value between the six regions (F_{ST} = 0.0002), as estimated by PLINK [19] using 459,425 autosomal SNPs. To check whether structure was simply too weak to be visually detected, we calculated for each Danish sample the average geographic coordinates of their grandparents' place of birth and regressed the resulting values on PC1 and PC2 eigenvectors. We repeated the procedure by gradually rotating the map clockwise and found a weak yet significant correlation between PC1 and latitude along a northwest-southeast axis at 32° ($r \approx 0.24$; $p < 0.001$; Figure 1C).

Chromosome painting, population clustering and admixture proportions

To further explore historical genetic interactions between Denmark and neighboring countries, we ran CHROMOPAINTER [20], fineSTRUCTURE [20] and GLOBETROTTER [21] to a subset of the studied populations. We used 2,745 individuals from 13 European countries (Norway, Sweden, Finland, the Netherlands, Germany, Poland, Austria, Hungary, France, Belgium, Great Britain, Spain and Portugal) and the 131 individuals from the six geographic regions in Denmark (Figure 1B) to calculate European admixture proportions in each of the six Danish groups. We used these six groups because we were unable to observe any alternative clustering with fineSTRUCTURE: the program clustered all Danish samples in a single large group (data not shown), reminiscent of the single Danish group observed in a previous study [15]. This observation also reflects the weak genetic structure in the Danish population.

After running CHROMOPAINTER and fineSTRUCTURE on the European donor samples, these were organized in eight major clusters: Norwegian, Swedish, Finnish, British, French, German, Polish and Iberian (Figure 2A). There was always one predominant country in the makeup of each cluster, except for the Iberian cluster, in which Spain and Portugal were present at almost equal proportions, and the German cluster, in which samples from Austria, Hungary and the Netherlands were also present at large numbers. We treated these clusters as ancestry components and used GLOBETROTTER to define their admixture contribution to each of the six Danish regions.

Figure 2B shows that the Swedish, Norwegian and British clusters made the most substantial contribution to the ancestry profiles of all six Danish groups, jointly accounting for 85.09-94.38% of the total admixture. The Scandinavian component (Sweden and Norway) surprisingly accounted for less than half of the total admixture

(range: 43.45-47.54%), on a par with the British component (40.14-47.93%), which peaked in South Jutland. Interestingly, Sweden's contribution (28.73-30.52%) was almost twice as large as Norway's (14.34-18.45%). This difference could be explained by the reduced landscape connectivity between Norway and Denmark and the increased connectivity between Sweden and Denmark, affecting gene flow correspondingly. It is also striking that the German cluster had little genetic influence on Denmark (2.69-7.28%), despite the proximity and historically fluid borders between the two countries. The French component was present in all Danish regions (3.36-7.36%) except for South Jutland. It is also worth noting that there was a small yet considerable contribution from the Polish component to Zealand (6.33%). Finally, there was no detectable contribution from the Finnish or the Iberian component to any of the six Danish groups.

Ancestry component analysis

We also estimated individual admixture proportions in the same imputed set of 12 European countries (without Finland) and six Danish regions by applying a new model-based method (R. Nielsen, personal communication). For this purpose, we assumed that each of the European samples was the result of admixture between K = 4 ancestral populations (i.e. ancestry components). The analysis returned an admixture pattern that was remarkably consistent with geography (i.e. North-South and East-West clines; Figure 3). In particular, we observed (i) a Southern European component, which was predominant in the Iberian peninsula but was also found at large proportions in France and Belgium; (ii) an Eastern European component, which was predominant in Poland but was also found at large proportions in neighboring countries (and notably in Scandinavia and East Denmark); (iii) a Nordic component, which was at higher proportions in Scandinavia yet far from being predominant; and

(iv) a Northwestern European component, at considerably higher proportions in Scandinavia and the Netherlands but also present at considerable proportions in most European countries (except for Iberia). Within Denmark, we observed that Zealand (and East Denmark in general) had a substantially larger membership to the Eastern European component, matching GLOBETROTTER's signal of Polish admixture observed in Zealand (Figure 2B).

Relatedness and identity by descent

We followed up the evidence for weak genetic structure in Denmark by exploring the degree and the geographic distribution of relatedness among Danish samples. Using KING [22], we found that the vast majority of the Danish students were at best distantly related (4^{th} degree or more distant), with only four pairs of individuals showing 2^{nd} or 3^{rd} degree relationships (Figure S1). Because of the inherent uncertainty in distinguishing between different degrees of distantly related individuals, we did not attempt to stratify the samples into more fine-grained categories by kinship coefficient.

After establishing that most participants in our sample were either unrelated or distantly related (Figure S1), we examined results from BEAGLE Refined IBD [23]. Using total genomic length of IBD tracts as a proxy for relatedness, we traced each individual's closest genomic relative within Denmark without explicitly defining the degree of relationship. Figure 4A shows the distribution of the geographic distance of each of 399 individuals from their closest genomic relative and from a randomly chosen sample. The distribution of the distance from the closest relative showed an enrichment for very short distances, i.e. less than ~50 km, as well as a median value of 99.3 km – significantly closer than expected by chance at 131.4 km (Mann-

Whitney test, U = 60,997; $p = 5.52 \times 10^{-9}$). This points at a weak yet highly significant overall tendency for participants to live close to their genomic relatives.

To gain more insight into the latter observation, we grouped our Danish samples into ranked bins by the amount of total genomic IBD shared with their closest genomic relative (the higher the rank the closer the relationship). We then calculated the median geographic distance of each participant to their closest relative within each bin and regressed this value against bin rank (Figure 4B) to observe a weak yet significant negative correlation ($r \approx -0.35$; $p < 0.01$). This observation points out that geographic distance tends to be significantly shorter between individuals who share more genomic IBD.

Historical effective population size

Historical N_e showed remarkable disparity between the three Scandinavian countries (Figure 5). In particular, N_e in Denmark, Norway and Sweden showed a dramatic ~273-fold, ~262-fold and ~995-fold increase over the past 150 generations (~4,500±300 years), following the general upward trend in Europe [24]. For most of this millennia-long period (until approximately the 10th Century), N_e in both Denmark and Sweden increased slowly in an almost indistinguishable manner. During the same period, Norway's N_e presented a less steep increase and was consistently smaller than in the other two countries, possibly due to earlier geographic isolation. During the High and Late Middle Ages, N_e in Denmark remained stable, suffering an almost imperceptible decline as a consequence of the otherwise devastating Black Death. On the contrary, N_e in Norway and Sweden were not affected by the Black Death and showed almost parallel increase rates, even though Norway's N_e was consistently smaller than Sweden's. Finally, from the 15th Century on, N_e in Denmark started to

rise again at a moderate rate, whereas N_e in Norway and Sweden rose at an even higher rate, resulting in Norway's N_e eventually surpassing Denmark's. Interestingly, even though Denmark and Norway currently have almost identical census population sizes (5.614 and 5.084 million, respectively), Norway's N_e is 1.76 times larger than Denmark's. It is also worth noting that the upward trend of Denmark's N_e was not impeded by other important epidemics in the recent history of the country, such as cholera, the Spanish flu or, more recently, polio (Figure 5).

Polygenic prediction of height and BMI

Polygenic risk prediction in our Danish sample was far more accurate for height than for BMI (Figure 6). In the case of height, we observed maximum accuracy when assuming infinitesimal genetic architecture and adjusting for age, sex and ten ancestry PCs (R^2 = 0.251±0.031). When age, sex and the ten PCs were included in the model, the prediction rose substantially to 0.639±0.015 ($p = 6.57\times10^{-71}$), a fact also reflected in the strong and significant correlation between real and predicted height (Figure 6A). In contrast, although the maximum accuracy of BMI prediction was also observed when we adjusted for age, sex and ten PCs, this was overall much poorer than for height (R^2 = 0.106±0.037). In this case, when age, sex and ten PCs were included in the model, the improvement of accuracy was not as notable as for height (R^2 = 0.195±0.034; $p = 5.49\times10^{-18}$), implying that BMI is only marginally affected by age and sex (Figure 6B).

Discussion

The most striking observation in this study is the high genetic homogeneity of the Danish population, possibly reflecting uninterrupted gene flow facilitated by the extended network of sea-based commerce and travelling [25], as well as the lack of

major geographical barriers in Denmark. It is important to point out that this observation should be appreciated in a historical context, as it did not entail genetic contributions from recent immigration: modern Danish society accommodates different ethnic and cultural groups [26], and this was also reflected in our sample, where ~4% of the participants were born in Afghanistan, China, Ethiopia, Finland, Germany, Greenland, Iraq, Jordan, Korea, Kosovo, the Netherlands or Zambia. This number went up when we looked at grandparental origin, where 14.6% of the participants had at least one grandparent born outside Denmark.

The homogeneity of the Danish population was evident in the relative lack of population structure in the classical PCA (Figure 1B), as well as in the extremely low average F_{ST} value. To put this observation in context, a previous study was able to detect subtle population structure along a north-south axis in the Netherlands [14], a country of almost identical size to Denmark's, and although we were able to detect significant correlation between PC1 and latitude (Figure 1C), this was admittedly weak. Similarly, a recent study of population structure in Great Britain [15] found that the average pairwise F_{ST} estimates between 30 geographic regions was 0.0007 – 3.5 times higher than the value we report here (0.0002).

In general, the study of admixture within the European continent is confounded by a well-grounded isolation-by-distance mechanism [2,3], as well as an increased historical complexity that renders most admixture models unrealistically simple. Denmark is no exception to these caveats: the Danish population has strong historical bonds with other Scandinavian countries, but also with Western and Eastern Europe via invasions, conquests and alliances that led to settlements as far as Britain, Estonia, the Faroe Islands, Iceland, Greenland and even Canada [25]. Even though it is tempting to explain the admixture proportions seen in Figures 2 and 3 as the result of

historical admixing events, a more judicious approach is to interpret such proportions as “mixture profiles” and use them for comparisons between the studied regions.

Bearing this in mind, we see that the mixture profiles of all six Danish groups comprise two major ancestry components, one predominant in Scandinavia and the other predominant in Northwestern Europe (Figure 3). In the GLOBETROTTER analysis, these two components were identified as the admixture contributions from Sweden/Norway on one hand, and Britain on the other (Figure 2B). In regard to the British contribution, however, this is actually more likely to reflect admixture in the opposite direction, i.e. from Denmark towards Britain, as shown recently [15].

It is worth noting that even though the mixture profiles of the six regions in Denmark were overall quite similar, they differed in their membership to the Eastern European component. This particular component was larger in East Denmark (Figure 3) and was independently observed in the GLOBETROTTER analysis as a minor contribution from the Polish component to Zealand (Figure 2B). It is tempting to interpret this signal as consequence of historical Wend settlements in the broader area around Lolland in the southernmost part of Denmark [25]. However, because similar mixture profiles were also observed in Sweden and Norway (Figure 3), it is hard to make a distinction between isolation by distance and actual admixture.

A weak signal of population structure was also observed when we studied the geographic distribution of IBD-based relatedness. The median distance to the closest genetic relative (99.3 km) was significantly smaller than to a randomly chosen sample (131.4 km), but it represents a minor effect (median difference ≈ 30 km; Figure 4A). In our regression analysis, we observed that this distance tended to be significantly smaller for pairs of individuals that were more closely related, yet the correlation was

overall modest ($r \approx -0.35$; Figure 4B). These observations point out that genetic structure does exist in Denmark, even though it is very weak.

It is also striking that Denmark, Sweden and Norway – three Scandinavian countries with a common historical, geopolitical, cultural and linguistic heritage – seem to differ considerably in their demographic history, as reflected in their historical N_e trajectories (Figure 5). Indeed, current Sweden-to-Norway census population size ratio is 1.89 – considerably close to their current N_e ratio (2.24) – implying that the two populations have had similar reproductive variance and population structure. On the contrary Denmark's N_e seems to have increased at a consistently lower rate after the Middle Ages. This could be reflecting weaker reproductive dynamics in the Danish population or the actual lack of population structure compared to Sweden and Norway.

Finally, we found that self-reported adolescent height could be predicted with remarkable accuracy using essentially nothing but information derived directly (genotypes) or indirectly (sex and ancestry) from DNA available at birth. When we combined SNP data with age, sex and PCA information, our prediction could explain more than half of the total phenotypic variance (63.9%). The remaining unexplained variance corresponded to a standard deviation of 5.43 cm. This means that, with 95% confidence, we are at most ~10.65 cm off in our prediction of adolescent height (Figure 6A). It is worth noting that adding age to the model did not yield significant improvements to the prediction accuracy of height (data not shown). This means that even though adolescent height may not be a perfect reflection of adult height, this was still a reliable measurement for the validation of the prediction. Similarly, even though adolescent height was self-reported and therefore potentially subject to inaccuracies, its strong correlation with adult height, as observed in the highly

significant genetic prediction accuracy, suggests that the students provided a reliable report of their personal data (Figure 6A). Finally, regarding the poorer prediction of BMI, it is worth noting that data training was carried out in adult individuals, whereas prediction was validated in adolescents. Improvement of the prediction as subjects advance in age is not an ungrounded possibility.

In conclusion, our analysis showed that, by applying the simple criterion of participants having all four of their grandparents born in Denmark, we obtained a largely homogeneous sample with extremely low F_{ST} values among different geographic regions. This remarkably high homogeneity has the potential of rendering population structure in large-scale Danish gene-mapping studies such as iPSYCH (http://ipsych.au.dk/) a lesser concern, regardless of statistical advancements such as genomic control, PCA [28] and mixed models [29]. We also found a remarkably disparate demographic history in Demark compared to other Scandinavian countries – a fact that can also be ascribed to the high homogeneity of the population – and that height in adolescents could be predicted with considerable accuracy using cutting-edge methods [30]. Lastly but not least importantly, this study stands as an example of how large-scale public engagement projects can generate mutual benefits for both the general public and the scientific community through the promotion of scientific knowledge.

Materials and Methods

Sample description

New data:- We recruited samples under the *Where Are You From?* project, a large-scale scientific outreach initiative by Aarhus University involving ~800 students from 36 high schools from across Denmark (G. Athanasiadis, personal communication).

We asked participants to provide a saliva sample for DNA analysis and to answer an online questionnaire about family origin (their own, their parents' and their grandparents place of birth) and basic anthropometrical data (e.g. self-reported height and weight). The institutional review board of Aarhus University approved the study. Because no health-related questions were asked, there was no requirement for additional approval by the University's medical ethics committee. Informed consent was obtained either from participants themselves (age > 18 years) or their parents (age < 18 years). We used the 23andMe (Mountain View, CA, USA) DNA analysis service for the genotyping of 723 participants. 23andMe uses a custom HumanOmniExpress-24 BeadChip from Illumina (San Diego, CA, USA). After excluding duplicated single nucleotide polymorphisms (SNPs), applying a per-locus missingness threshold of 2% with PLINK v1.9 [31] and removing SNPs ambiguously mapped to the forward DNA strand, 517,403 unique autosomal SNPs were available for analysis.

Additional data:- To put our study in a broader European context, we included four additional datasets: (i) the POpulation REference Sample (POPRES) [32]; (ii) Amyotrophic Lateral Sclerosis (ALS) Finland [33]; (iii) the Swedish Schizophrenia Study [34]; and (iv) the Norwegian Cognitive NeuroGenetics (NCNG) sample [35]. POPRES included 2,863 Europeans typed with the Affymetrix (Santa Clara, CA, USA) 500K chip. We used SMARTPCA from the EIGENSOFT v5.0.1 package [36] to identify and remove 25 extreme outliers from the sample. We then applied a per-locus and a per-individual missingness threshold of 2% with PLINK and removed SNPs ambiguously mapped to the forward strand to create a 2,833 individuals × 227,899 SNPs dataset. ALS Finland included 401 ALS cases and 495 controls from across Finland typed with three Illumina platforms (HumanCNV370v1, HumanCNV370-Quadv3_C and Human1M-Duov3_B). After removing all cases and

one extreme outlier in the control sample, as well as SNPs with genotype missingness > 2% and those ambiguously mapped to the forward strand, we ended up with a 494 individuals × 314,526 SNPs dataset. Data from the Swedish Schizophrenia Study initially included 3,736 controls genotyped using three platforms (Affymetrix 5.0, Affymetrix 6.0 and Illumina OmniExpress). After filtering for 2% *per-locus* missingness and removing SNPs ambiguously mapped to the forward strand, we used PLINK to run principal component analysis (PCA) jointly with the data from Finland and identified a large proportion of Swedish samples clustering with Finish samples (data not shown). We consequently sampled randomly 500 individuals to match approximately Denmark's sample size and excluded those Swedish subjects clustering with the Finns from the resulting sample, ending up with a 381 individuals × 577,252 SNPs dataset. The NCNG sample included 670 homogeneous controls from Norway typed with the Illumina Human 610-Quad Beadchip. We additionally filtered the data for 2% *per-locus* missingness and removed SNPs ambiguously mapped to the forward strand. Finally, we sampled randomly 300 individuals to match approximately Denmark's sample size, leading to a final dataset of 300 individuals × 537,306 SNPs.

Imputation

Because SNP intersection between the five datasets was small, we carried out genotype imputation within each dataset separately before combining them. As mentioned above, we first changed DNA strand orientation of several SNPs in all five datasets to create a uniform forward orientation. In particular, we first used SNPFLIP (https://github.com/endrebak/snp-flip) to detect reverse/ambiguous SNPs, which we then flipped/removed with PLINK. Table S1 shows the number of SNPs that were flipped/removed from each dataset. We finally used liftOver from the UCSC Genome Browser to "lift" genome coordinates from the NCBI36/hg18 (March 2006) to the

GRCh37/hg19 (February 2009) assembly. This task was necessary only for POPRES, ALS Finland and NCNG.

After QC checks, orientation to forward DNA strand and liftOver were accomplished, we used SHAPEIT v2.720 [37] to produce "prephased" haplotypes for each dataset. We then used these haplotypes together with the latest 1000 Genomes Phase 3 reference panel (b37, October 2014) for the separate imputation of the five datasets with IMPUTE2 v.2.3.1 [38]. We carried out the imputation on 5 Mbp-long chromosome segments excluding centromeres in all chromosomes, as well as acrocentric regions in chromosomes 13, 14, 15, 21 and 22. Finally, we concatenated the imputed data into separate chromosomes and filtered them for "info" ≥ 0.975 with QCTOOL (http://www.well.ox.ac.uk/~gav/qctool/#overview).

Principal component analysis

We ran PCA with PLINK to examine population structure in our Danish sample. PCA was run on two different data combinations: (i) POPRES, *Where Are You From?*, NCNG and the Swedish Schizophrenia Study; and (ii) *Where Are You From?*, NCNG, the Swedish Schizophrenia Study and Germany from POPRES, repeating the analysis for both real and imputed genotypes. To avoid undesirable clustering due to extensive linkage disequilibrium (LD), we thinned the imputed genotypes with PLINK (using a window and step size of 1,500 and 150 SNPs, respectively, and r^2 threshold = 0.80) and also removed SNPs from known high-LD genomic regions (e.g. MHC on chromosome 6).

Chromosome painting, population clustering and admixture proportions

We used a set of recently developed LD-based methods (CHROMOPAINTER [20], fineSTRUCTURE [20] and GLOBETROTTER [21]) to explore fine-grain population structure and admixture in Denmark. These methods require a set of phased SNP data from "donor" (i.e. the available European samples) and "recipient" populations (i.e. the Danish sample). In brief, the methods detect extended multi-marker haplotypes across the genome, which are organized in pairwise vectors of similarity counts (CHROMOPAINTER). These vectors are then used by an MCMC algorithm (fineSTRUCTURE) to hierarchically cluster individuals into groups that are often geographically, linguistically and/or historically meaningful [15]. As a final step, admixture proportions are estimated through a multiple linear regression on the average proportion of DNA that each recipient copies from each of the donor groups (GLOBETROTTER).

After jointly phasing 489,209 imputed autosomal SNPs across all five datasets with IMPUTE2, we ran CHROMOPAINTER using default options [15] on three different datasets: (i) Denmark alone; (ii) Europe without Denmark; and (iii) Europe and Denmark together. We previously ran each analysis ten times on a sample subset (~10% of the total number samples and only for chromosomes 4, 10, 15, and 22) to estimate the switch and global emission rates used by CHROMOPAINTER's Hidden Markov Model algorithm. Once similarity vectors were defined in the three datasets, we used fineSTRUCTURE to explore clustering in recipient (dataset i) and donor (dataset ii) populations. For the GLOBETROTTER analysis, we merged the similarity matrices from datasets ii and iii and ran the program with default options described in more detail elsewhere [15].

Ancestry component analysis

We used a new model-based method (R. Nielsen, personal communication) to estimate individual admixture proportions in 13 European countries – mostly Western and Northern – including Denmark. In brief, the method uses the same statistical model as does STRUCTURE [39], FRAPPE [40], and ADMIXTURE [41] for admixture, and Newton's method for optimization. After running the algorithm, we reported per-country admixture proportions by averaging out individual proportions within each country.

Relatedness and identity by descent

To explore relatedness in our Danish sample, we used two different methods: KING [22] for the initial calculation of pairwise kinship coefficients and BEAGLE Refined IBD [23] for the inference of DNA segments that were identical by descent (IBD) between pairs of individuals. We analyzed 406 individuals for whom we had complete information that all four of their grandparents were born in Denmark. We excluded from the analysis one individual from pairs of twins and siblings (preferentially the one with highest *per-locus* genotype missingness).

Historical effective population size

We also estimated the historical effective population size (N_e) of the Danish, Swedish and Norwegian populations through the combination of two IBD-based methods. In particular, we first applied IBDseq [42] to three datasets separately (407 Danes who had all four of their grandparents born in Denmark × 514,136 autosomal SNPs; 381 Swedes who did not show resemblance with samples from Finland × 577253 autosomal SNPs; and 300 Norwegians × 537,305 autosomal SNPs) in order to produce sets of pairwise IBD tracts. We then used IBDNe [24] to estimate N_e from the distribution of the inferred IBD tracts over the past 150 generations for each of the

three Scandinavian populations. To maximize power, we used the original rather than the imputed SNP data for this analysis.

Polygenic prediction of height and BMI

Apart from the characterization of population structure and demographic history in Denmark, an additional focus of this work has been the quantitative study of basic anthropometrical traits. To this end, we used the self-reported height and weight data from ~600 students of diverse ethnic backgrounds to perform polygenic risk prediction of height and body mass index (BMI) with LDpred [30], a summary statistic-based algorithm that models LD to improve the prediction. As training data for the model, we used public summary statistics from large genome-wide association studies of adult height [43] and BMI [44]. We first removed from our data SNPs with minor allele frequency (MAF) < 0.01, as well as SNPs with a MAF different from the one reported in the training data by a factor of 0.15. We then assessed SNP effect under different fractions of causal variants: p = {1, 0.5, 0.2, 0.1, 0.05}, whereby p = 1 corresponds to the infinitesimal model [45]. As an LD reference, we used a subset of 407 unrelated individuals (kinship coefficient < 0.05) who had all four of their grandparents born in Denmark. Finally, we validated LDpred's prediction of height in 578 individuals, as well as the prediction of BMI in 572 individuals. We calculated the prediction R^2 (i) by adjusting for age, sex and the first ten PCs, and (ii) by actually including age, sex and the first ten PCs in the model.

Acknowledgements

The authors would like to thank the high school students and their teachers for participating in the *Where Are You From?* project, as well as Dagmar Hedvig Fog Bjerre and Pia Johansson Nielsen for logistic support with organizing and shipping

the test kits to the participating schools. Special thanks to Prof. Jes Fabricius Møller for helping us gain useful insights into the history and demography of the Danish population. This project and related research were supported from the National Lottery Funds (*Danske Spil*), The Danish Ministry of Education and the Centre for Biocultural History, Aarhus University. B.J.V. was supported by a grant from the Danish Council for Independent Research (DFF-1325-0014). K.M.H. was supported by grants from the Swedish Research Council (K2014-62X-21445-05-3 and K2012-63X-21445-03-2). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science. 2008;319: 1100–1104. doi:10.1126/science.1153717

2. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. Nature. 2008;456: 98–101. doi:10.1038/nature07331

3. Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, et al. Correlation between genetic and geographic structure in Europe. Curr Biol CB. 2008;18: 1241–1248. doi:10.1016/j.cub.2008.07.049

4. Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. Nature. 2009;461: 489–494. doi:10.1038/nature08365

5. Moreno-Mayar JV, Rasmussen S, Seguin-Orlando A, Rasmussen M, Liang M, Flåm ST, et al. Genome-wide ancestry patterns in Rapanui suggest pre-European admixture with Native Americans. Curr Biol CB. 2014;24: 2518–2525. doi:10.1016/j.cub.2014.09.057

6. Moreno-Estrada A, Gignoux CR, Fernández-López JC, Zakharia F, Sikora M, Contreras AV, et al. Human genetics. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. Science. 2014;344: 1280–1285. doi:10.1126/science.1251688

7. Bryc K, Durand EY, Macpherson JM, Reich D, Mountain JL. The genetic ancestry of African Americans, Latinos, and European Americans across the United States. Am J Hum Genet. 2015;96: 37–53. doi:10.1016/j.ajhg.2014.11.010

8. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. Science. 2010;328: 710–722. doi:10.1126/science.1188021

9. Reich D, Patterson N, Kircher M, Delfin F, Nandineni MR, Pugach I, et al. Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. Am J Hum Genet. 2011;89: 516–528. doi:10.1016/j.ajhg.2011.09.005

10. Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) [Internet]. [cited 26 Nov 2015]. Available: http://www.genome.gov/sequencingcosts/

11. Jakkula E, Rehnström K, Varilo T, Pietiläinen OPH, Paunio T, Pedersen NL, et al. The genome-wide patterns of variation expose significant substructure in a founder population. Am J Hum Genet. 2008;83: 787–794. doi:10.1016/j.ajhg.2008.11.005

12. Price AL, Helgason A, Palsson S, Stefansson H, St Clair D, Andreassen OA, et al. The impact of divergence time on the nature of population structure: an example from Iceland. PLoS Genet. 2009;5: e1000505. doi:10.1371/journal.pgen.1000505

13. Karakachoff M, Duforet-Frebourg N, Simonet F, Le Scouarnec S, Pellen N, Lecointe S, et al. Fine-scale human genetic structure in Western France. Eur J Hum Genet EJHG. 2015;23: 831–836. doi:10.1038/ejhg.2014.175

14. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet. 2014;46: 818–825. doi:10.1038/ng.3021

15. Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T, et al. The fine-scale genetic structure of the British population. Nature. 2015;519: 309–314. doi:10.1038/nature14230

16. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014;42: D1001–1006. doi:10.1093/nar/gkt1229

17. Larsen MH, Albrechtsen A, Thørner LW, Werge T, Hansen T, Gether U, et al. Genome-Wide Association Study of Genetic Variants in LPS-Stimulated IL-6, IL-8, IL-10, IL-1ra and TNF-α Cytokine Response in a Danish Cohort. PloS One. 2013;8: e66262. doi:10.1371/journal.pone.0066262

18. Bae HT, Sebastiani P, Sun JX, Andersen SL, Daw EW, Terracciano A, et al. Genome-wide association study of personality traits in the long life family study. Front Genet. 2013;4: 65. doi:10.3389/fgene.2013.00065

19. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015;4. doi:10.1186/s13742-015-0047-8

20. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genet. 2012;8: e1002453. doi:10.1371/journal.pgen.1002453

21. Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, et al. A genetic atlas of human admixture history. Science. 2014;343: 747–751. doi:10.1126/science.1243518

22. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. Robust relationship inference in genome-wide association studies. Bioinforma Oxf Engl. 2010;26: 2867–2873. doi:10.1093/bioinformatics/btq559

23. Browning BL, Browning SR. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics. 2013;194: 459–471. doi:10.1534/genetics.113.150029

24. Browning SR, Browning BL. Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. Am J Hum Genet. 2015;97: 404–418. doi:10.1016/j.ajhg.2015.07.012

25. Derry TK. History of Scandinavia: Norway, Sweden, Denmark, Finland, and Iceland. University of Minnesota Press; 2000.

26. Jensen P, Pedersen PJ. To Stay or Not to Stay? Out-Migration of Immigrants from Denmark. Int Migr. 2007;45: 87–113. doi:10.1111/j.1468-2435.2007.00428.x

27. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999;55: 997–1004.

28. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38: 904–909. doi:10.1038/ng1847

29. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 2006;38: 203–208. doi:10.1038/ng1702

30. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. Am J Hum Genet. 2015;97: 576–592. doi:10.1016/j.ajhg.2015.09.001

31. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015;4: 7. doi:10.1186/s13742-015-0047-8

32. Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J, et al. The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am J Hum Genet. 2008;83: 347–358. doi:10.1016/j.ajhg.2008.08.005

33. Laaksovirta H, Peuralinna T, Schymick JC, Scholz SW, Lai S-L, Myllykangas L, et al. Chromosome 9p21 in amyotrophic lateral sclerosis in Finland: a genome-wide association study. Lancet Neurol. 2010;9: 978–985. doi:10.1016/S1474-4422(10)70184-8

34. Ripke S, O'Dushlaine C, Chambert K, Moran JL, Kähler AK, Akterin S, et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat Genet. 2013;45: 1150–1159. doi:10.1038/ng.2742

35. Espeseth T, Christoforou A, Lundervold AJ, Steen VM, Le Hellard S, Reinvang I. Imaging and cognitive genetics: the Norwegian Cognitive NeuroGenetics sample. Twin Res Hum Genet Off J Int Soc Twin Stud. 2012;15: 442–452. doi:10.1017/thg.2012.8

36. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2: e190. doi:10.1371/journal.pgen.0020190

37. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. Nat Methods. 2012;9: 179–181. doi:10.1038/nmeth.1785

38. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet. 2012;44: 955–959. doi:10.1038/ng.2354

39. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155: 945–959.

40. Tang H, Peng J, Wang P, Risch NJ. Estimation of individual admixture: analytical and study design considerations. Genet Epidemiol. 2005;28: 289–301. doi:10.1002/gepi.20064

41. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19: 1655–1664. doi:10.1101/gr.094052.109

42. Browning BL, Browning SR. Detecting identity by descent and estimating genotype error rates in sequence data. Am J Hum Genet. 2013;93: 840–851. doi:10.1016/j.ajhg.2013.09.014

43. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet. 2014;46: 1173–1186. doi:10.1038/ng.3097

44. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015;518: 197–206. doi:10.1038/nature14177

45. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era--concepts and misconceptions. Nat Rev Genet. 2008;9: 255–266. doi:10.1038/nrg2322

Figure titles and legends

Figure 1: (A) PCA of 105,672 imputed SNPs after merging four datasets: *Where Are You From?*, POPRES, NCNG and the Swedish Schizophrenia study without outliers clustering with Finland (total N = 3,858). Per-country box plots (median and interquartile range) of PC values were added to facilitate interpretation. Whiskers represent data within 1.5 times the interquartile range. IE: Ireland; ES: Spain; PT: Portugal; GB: Great Britain; FR: France; BE: Belgium; CH: Switzerland; NL: the Netherlands; DK: Denmark; DE: Germany; NO: Norway; SE: Sweden; AT: Austria; IT: Italy; PL: Poland; HU: Hungary; CZ: Czech Republic; HR: Croatia; RO: Romania; YU: Yugoslavia; GR: Greece. (B) PCA of 105,672 imputed SNPs from Denmark, Sweden, Norway and Germany with emphasis on the six geographic regions of Denmark (total N = 1,168). No clear genetic-geographic relationship was observed. Ca: Capital; Ze: Zealand; Fu: Funen; SJ: South Jutland; CJ: Central Jutland; NJ: North Jutland. (C) Correlation of PC1 and PC2 with average grandparent place-of-birth latitude along a 360° clockwise rotation. Maximum correlation was observed for PC1 at 32° ($r \approx 0.24$; $p < 0.001$).

Figure 2: (A) fineSTRUCTURE grouping of the 2,745 European donor samples into eight clusters roughly corresponding to well-defined geographic locations. FIN: Finnish; NOR: Norwegian; SWE: Swedish; POL: Polish; GER: German; BRI: British; FRA: French; IBE: Iberian. Color legend at the bottom of the map shows different donor countries. FI: Finland. (B) GLOBETROTTER admixture proportions of each of the eight European clusters in the six geographic regions of Denmark. Neither FIN nor IBE made substantial contributions to the mixture profiles of Denmark.

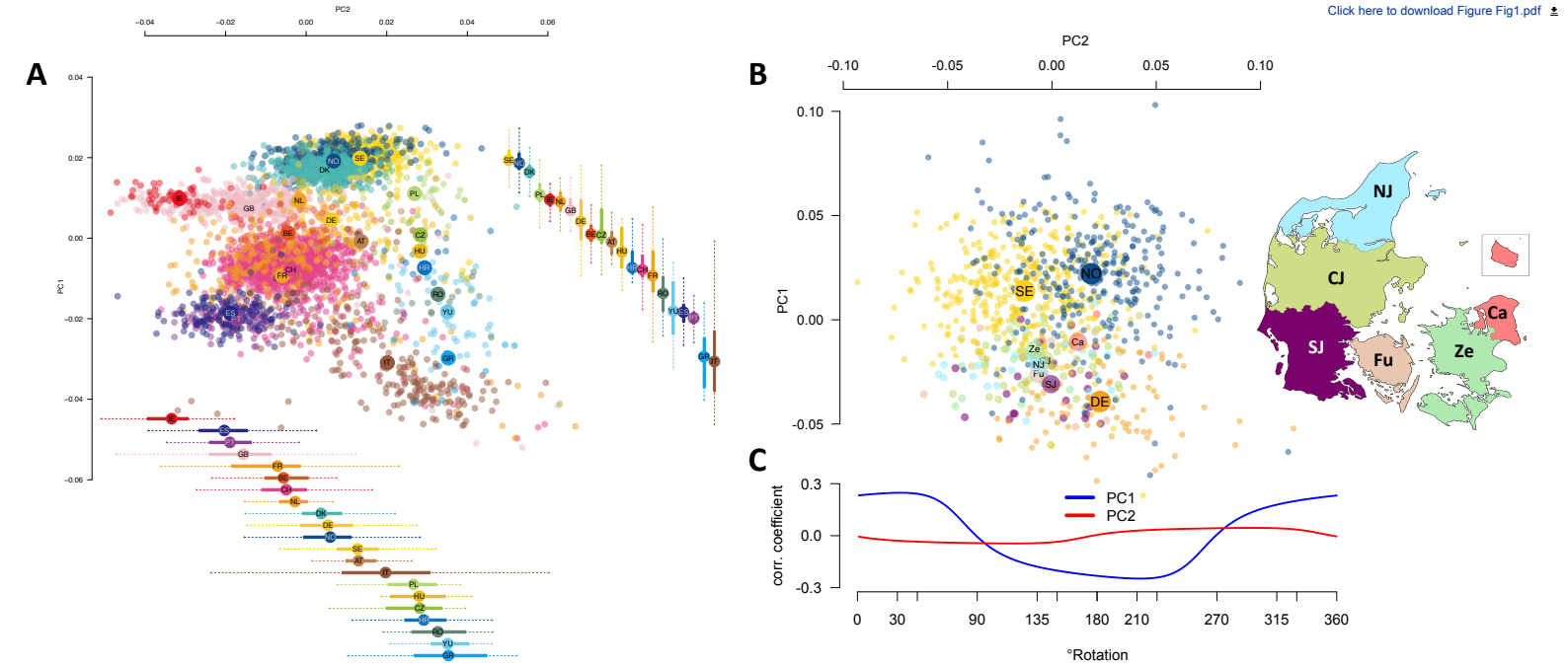
Figure 3: Ancestral component analysis of 13 European countries (including six well-defined geographic regions in Denmark shown in the inset) assuming K = 4 ancestral populations. Bar plot at the bottom shows per-individual membership to each of the four ancestral components, whereas pie charts on the map resume per-country (or per-region for Denmark) admixture proportions. Based on their preponderance in different parts of Europe, we interpret the four components as (i) Southern European (blue); (ii) Eastern European (yellow); (iii) Nordic (green); and (iv) Northwestern European (blue). Note that Sealand (including the Capital Region) and Funen have higher proportion of Eastern European ancestry, in accord with Figure 2B.

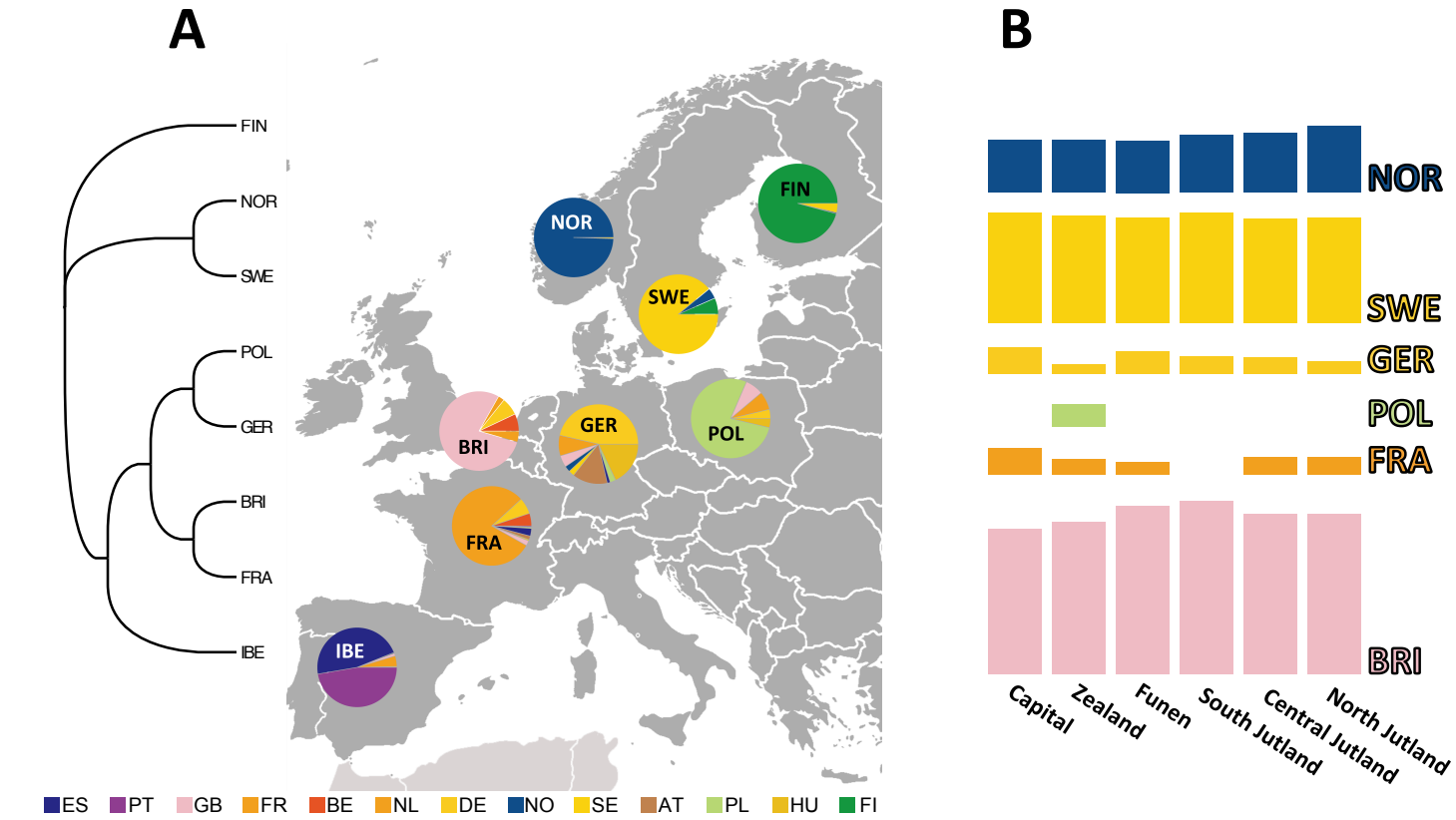
Figure 4: (A) Distribution of geographic distance of each participant's place of birth (N = 399) to that of their closest genomic relative (pink), and to that of a randomly chosen sample (green). Genomic relatedness was defined on the grounds of total genomic IBD. Arrows point at median values of the two distributions ($median_{rltv}$ = 99.3 Km; $median_{rand}$ = 131.4 Km). (B) Plot of rank of genomic relatedness vs. median geographic distance of each participant to their closest genomic relative. We created 57 equally-sized bins of individuals increasingly related to their closest genomic relative (seven individuals per bin). Alternative bin sizes also produced significantly negative correlations (data not shown).

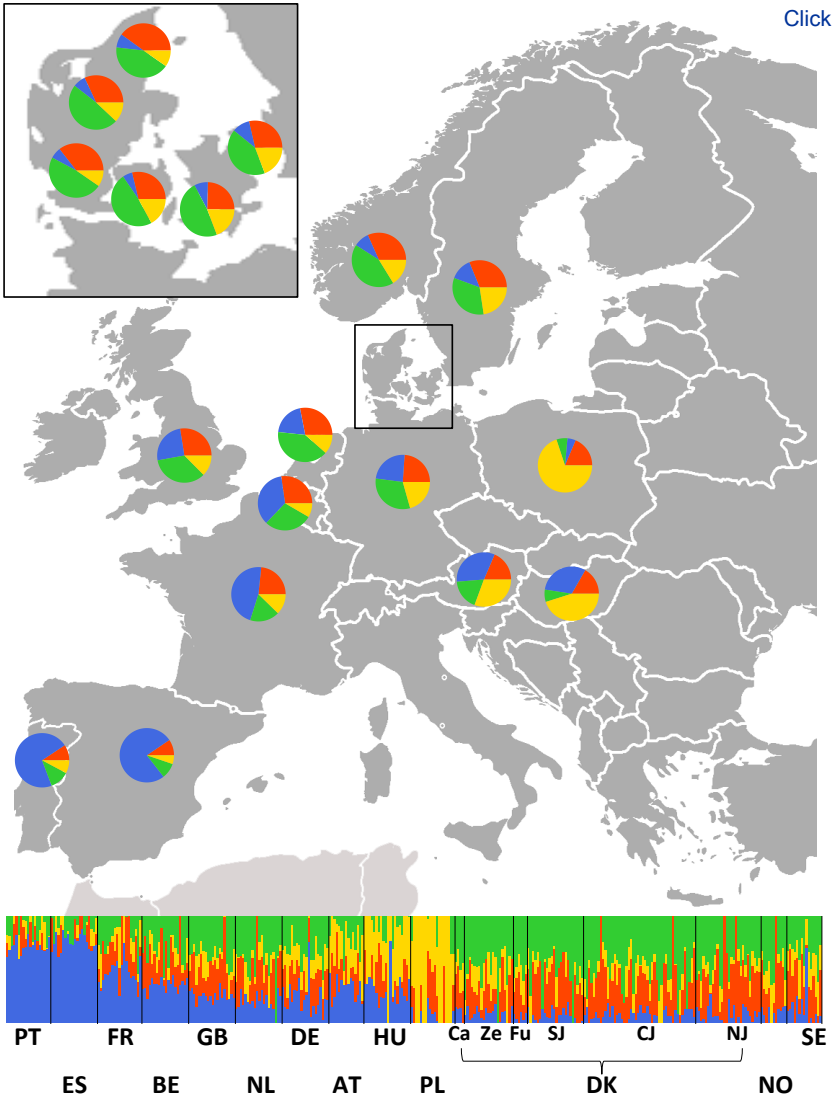
Figure 5: Change in effective population size (N_e) of the Danish, Swedish and Norwegian population over the past 150 generations. Shaded areas represent the upper and lower bounds of the 95% confidence intervals, after bootstrapping. Uncertainty in generation length is represented by year intervals on the x-axis, assuming that each generation lasts 30±2 years. Black segments represent major epidemics from the recent history of the Danish population and are plotted taking into account generation uncertainty. For more clarity, the inset shows the same graph in log-scale.

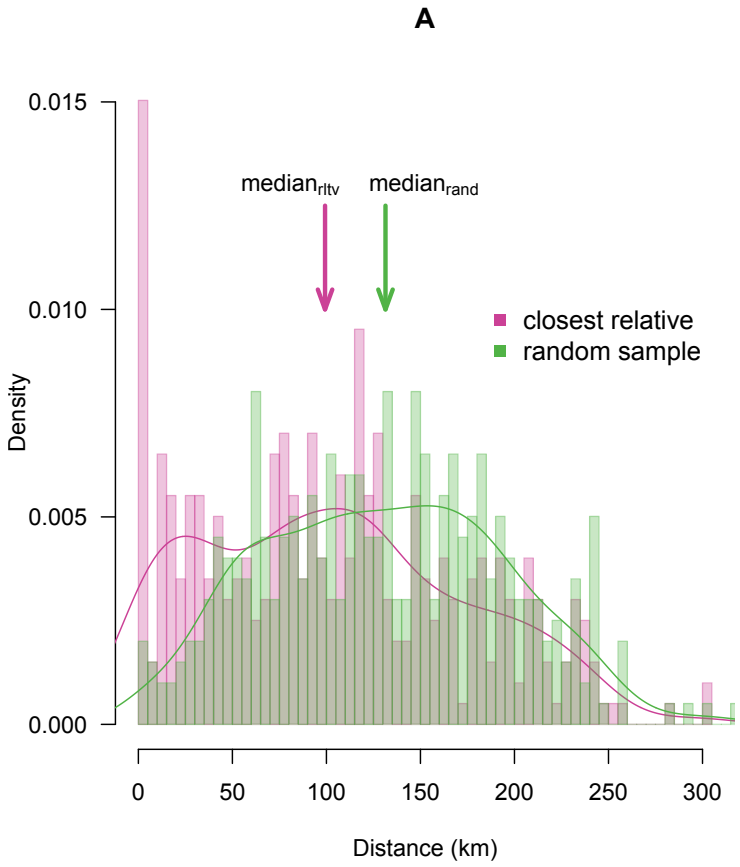
Figure 6: Prediction accuracy of LDpred for height (A) and BMI (B) when age, sex and PCA information were included in the model.

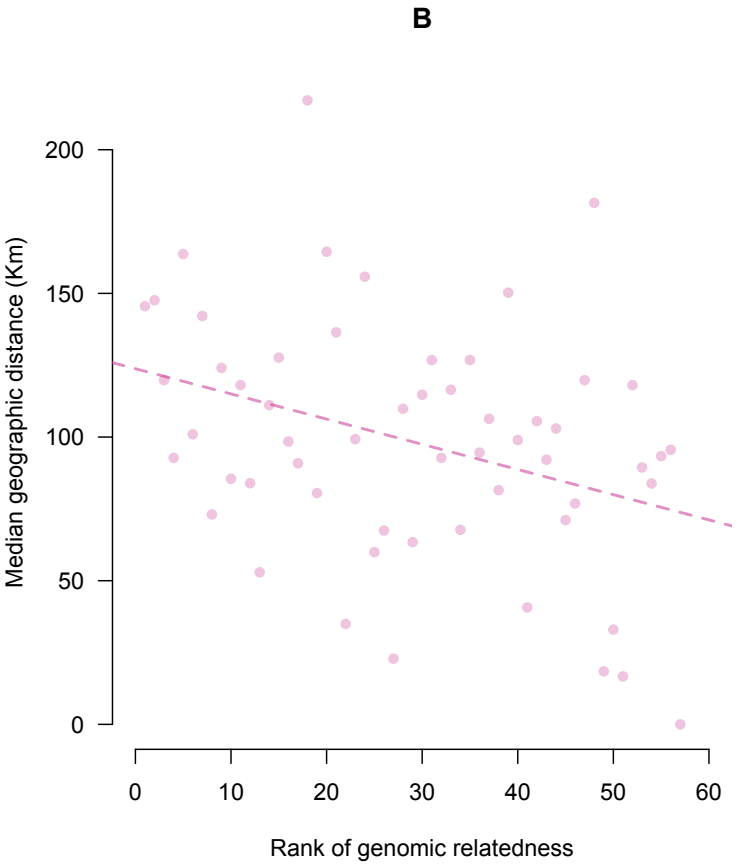
Figure S1: Proportion of zero IBS sharing vs. kinship coefficient for 82,215 pairs of individuals in a sample of 406 high school students who had all four of their grandparents born in Denmark. The vast majority of pairs were either unrelated (negative kinship coefficient in the grey-shaded part of the plot) or too distantly related (kinship $\in (0, 0.0221)$, in the red-shaded part of the plot).

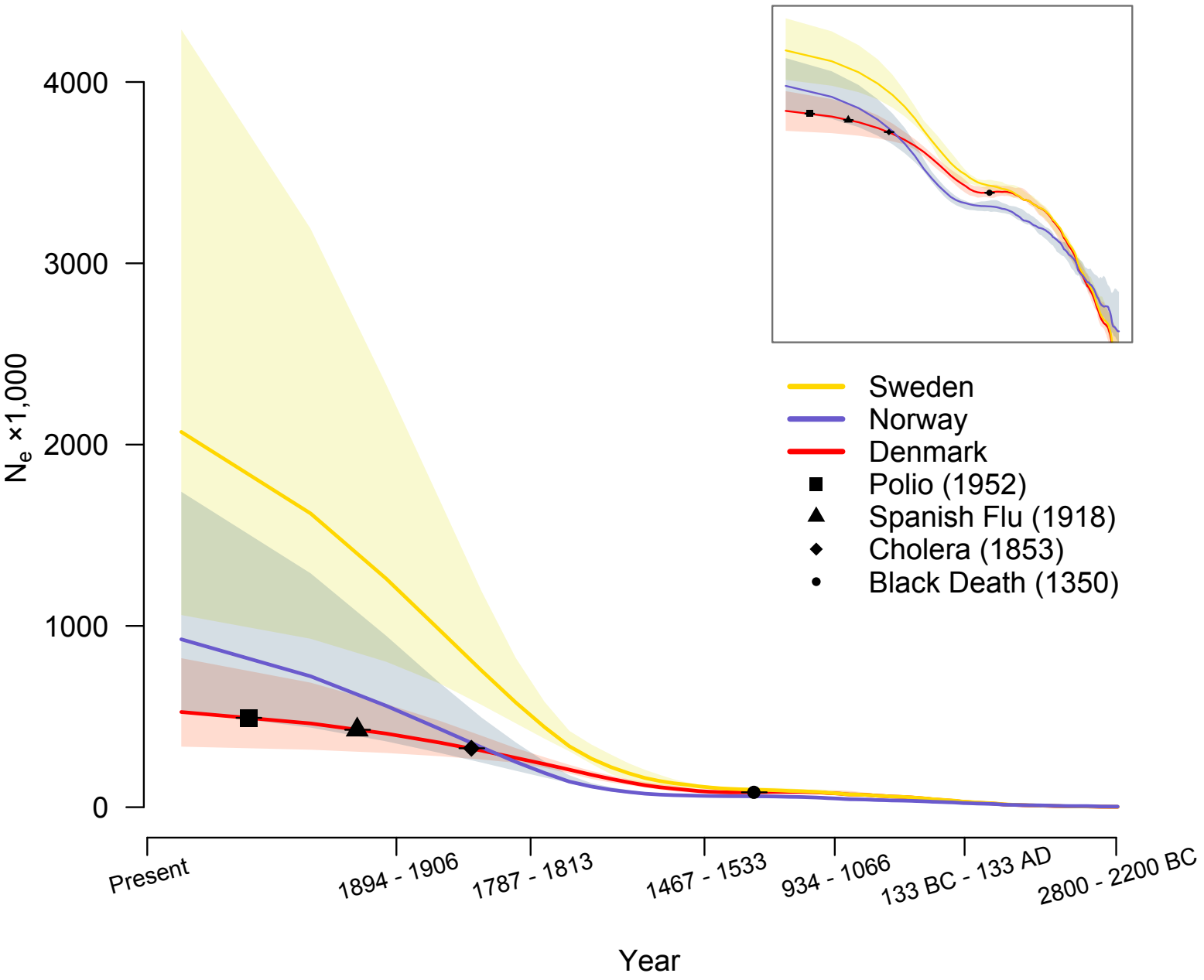


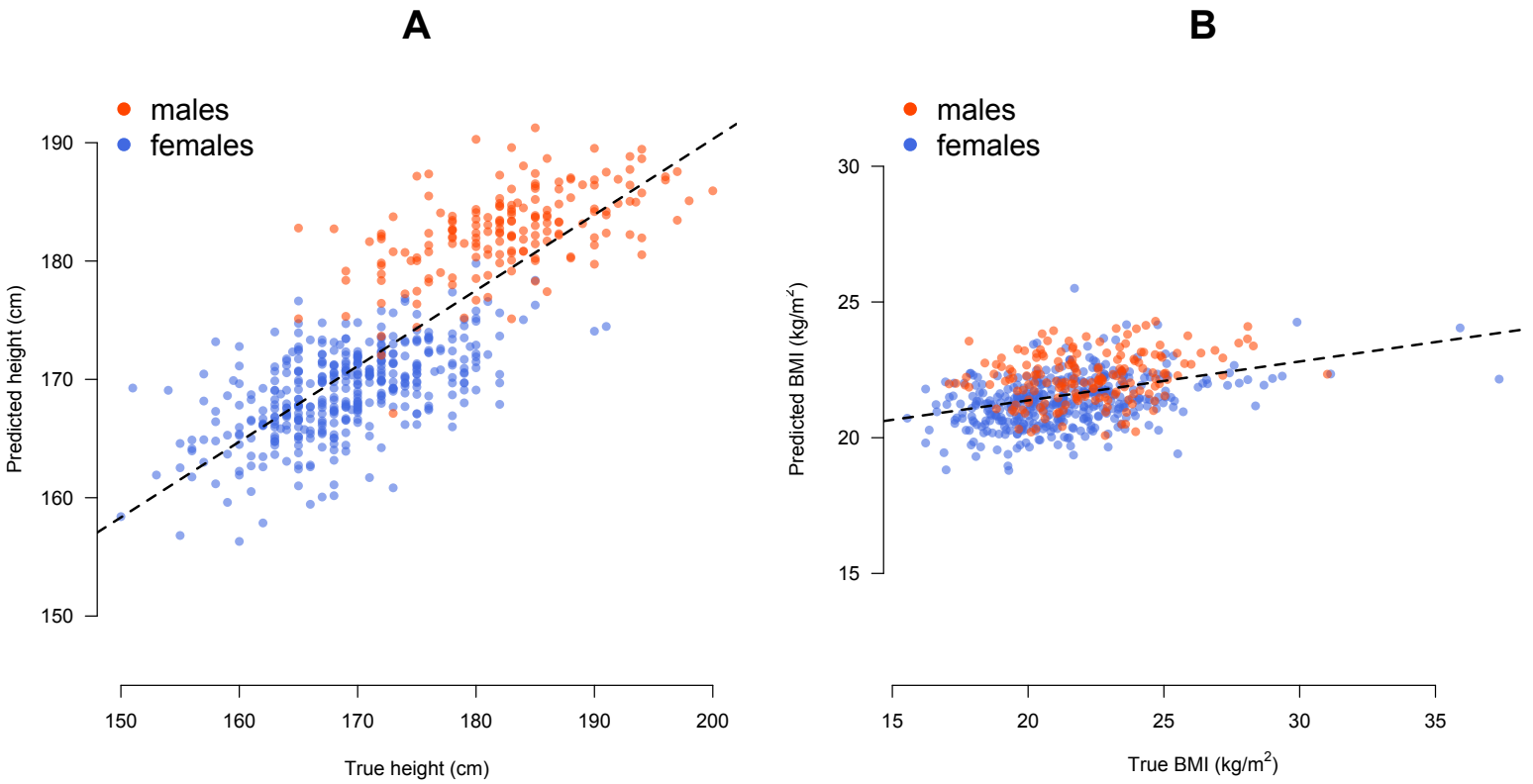












Bibliography

[1] David H Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9):1655–1664, 2009.

[2] James E Baker. Reducing bias and inefficiency in the selection algorithm. In *Proceedings of the second international conference on genetic algorithms*, pages 14–21, 1987.

[3] Michael Baudin. Nelder mead user's manual, 2009.

[4] Graham Coop, David Witonsky, Anna Di Rienzo, and Jonathan K Pritchard. Using environmental correlations to identify loci underlying local adaptation. *Genetics*, 185(4):1411–1423, 2010.

[5] Kalyanmoy Deb. *Multi-objective optimization using evolutionary algorithms*, volume 16. John Wiley & Sons, 2001.

[6] Russ C Eberhart, James Kennedy, et al. A new optimizer using particle swarm theory. In *Proceedings of the sixth international symposium on micro machine and human science*, volume 1, pages 39–43. New York, NY, 1995.

[7] Laurent Excoffier and Matthieu Foll. Fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics*, 27(9):1332–1334, 2011.

[8] Laurent Excoffier, Isabelle Dupanloup, Emilia Huerta-Sánchez, Vitor C Sousa, and Matthieu Foll. Robust demographic inference from genomic and snp data. *PLoS Genet*, 9(10):e1003905, 2013.

[9] Jotun Hein, Mikkel Schierup, and Carsten Wiuf. *Gene genealogies, variation and evolution: a primer in coalescent theory*. Oxford University Press, USA, 2004.

[10] Asger Hobolth and Jens Ledet Jensen. Markovian approximation to the finite loci coalescent with recombination along multiple sequences. *Theoretical population biology*, 98:48–58, 2014.

[11] Asger Hobolth, Ole F Christensen, Thomas Mailund, and Mikkel H Schierup. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden markov model. *PLoS Genet*, 3(2):e7, 2007.

[12] JH Holland. Genetic algorithms in search, optimization and machine learning, 1989.

[13] John Holland. Genetic algorithms. 1992.

[14] Richard R Hudson. ms a program for generating samples under neutral models. 2004.

[15] Thomas H Jukes and Charles R Cantor. Evolution of protein molecules. *Mammalian protein metabolism*, 3(21):132, 1969.

[16] William Karush. *Minima of functions of several variables with inequalities as side constraints*. PhD thesis, Master's thesis, Dept. of Mathematics, Univ. of Chicago, 1939.

[17] Manfred Kayser, Silke Brauer, and Mark Stoneking. A genome scan to detect candidate regions influenced by local natural selection in human populations. *Molecular Biology and Evolution*, 20(6):893–900, 2003.

[18] HW Kuhn and AW Tucker. Proceedings of 2nd berkeley symposium, 1951.

[19] Tianying Lan, Jade Cheng, Aakrosh Ratan, Webb Miller, Stephan Schuster, Sean Farley, Richard Shideler, Thomas Mailund, and Charlotte Lindqvist. Genome-wide evidence for a hybrid origin of modern polar bears. *bioRxiv*, page 047498, 2016.

[20] Heng Li and Richard Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496, 2011.

[21] Thomas Mailund, Julien Y Dutheil, Asger Hobolth, Gerton Lunter, and Mikkel H Schierup. Estimating divergence time and ancestral effective population size of bornean and sumatran orangutan subspecies using a coalescent hidden markov model. *PLoS Genet*, 7(3):e1001319, 2011.

[22] Zbigniew Michalewicz, Genetic Algorithms, and Data Structures. Evolution programs, 1996.

[23] Brad L Miller and David E Goldberg. Genetic algorithms, tournament selection, and the effects of noise. *Complex Systems*, 9(3):193–212, 1995.

[24] Katta G Murty and Feng-Tien Yu. *Linear complementarity, linear and nonlinear programming*. Citeseer, 1988.

[25] John A Nelder and Roger Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

[26] Rasmus Nielsen, Ines Hellmann, Melissa Hubisz, Carlos Bustamante, and Andrew G Clark. Recent and ongoing selection in the human genome. *Nature Reviews Genetics*, 8(11):857–868, 2007.

[27] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[28] Joseph K Pickrell and Jonathan K Pritchard. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet*, 8(11):e1002967, 2012.

[29] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2): 945–959, 2000.

[30] Andrew Rambaut and Nicholas C Grass. Seq-gen: an application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. *Computer applications in the biosciences: CABIOS*, 13(3):235–238, 1997.

[31] Stephan Schiffels and Richard Durbin. Inferring human population size and separation history from multiple genome sequences. *Nature genetics*, 46(8):919–925, 2014.

[32] Katy L Simonsen and Gary A Churchill. A markov chain model of coalescence with recombination. *Theoretical population biology*, 52(1):43–59, 1997.

[33] Montgomery Slatkin and Joshua L Pollack. The concordance of gene trees and species trees at two linked loci. *Genetics*, 172(3):1979–1984, 2006.

[34] Gilbert Syswerda. Uniform crossover in genetic algorithms. 1989.

[35] Hua Tang, Jie Peng, Pei Wang, and Neil J Risch. Estimation of individual admixture: analytical and study design considerations. *Genetic epidemiology*, 28(4):289–301, 2005.

[36] Wen-Yun Yang, John Novembre, Eleazar Eskin, and Eran Halperin. A model-based approach for analysis of spatial structure in genetic data. *Nature genetics*, 44(6):725–731, 2012.