# Machine learning methods for the interpretation of next generation sequencing data: applications to leukemia and demographic inference

Christopher B. Cole

Exeter College
University of Oxford

*A thesis submitted for the degree of*
*Doctor of Philosophy*

Michaelmas 2021

## Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Pellentesque sit amet nibh volutpat, scelerisque nibh a, vehicula neque. Integer placerat nulla massa, et vestibulum velit dignissim id. Ut eget nisi elementum, consectetur nibh in, condimentum velit. Quisque sodales dui ut tempus mattis. Duis malesuada arcu at ligula egestas egestas. Phasellus interdum odio at sapien fringilla scelerisque. Mauris sagittis eleifend sapien, sit amet laoreet felis mollis quis. Pellentesque dui ante, finibus eget blandit sit amet, tincidunt eu neque. Vivamus rutrum dapibus ligula, ut imperdiet lectus tincidunt ac. Pellentesque ac lorem sed diam egestas lobortis.

Suspendisse leo purus, efficitur mattis urna a, maximus molestie nisl. Aenean porta semper tortor a vestibulum. Suspendisse viverra facilisis lorem, non pretium erat lacinia a. Vestibulum tempus, quam vitae placerat porta, magna risus euismod purus, in viverra lorem dui at metus. Sed ac sollicitudin nunc. In maximus ipsum nunc, placerat maximus tortor gravida varius. Suspendisse pretium, lorem at porttitor rhoncus, nulla urna condimentum tortor, sed suscipit nisi metus ac risus.

Aenean sit amet enim quis lorem tristique commodo vitae ut lorem. Duis vel tincidunt lacus. Sed massa velit, lacinia sed posuere vitae, malesuada vel ante. Praesent a rhoncus leo. Etiam sed rutrum enim. Pellentesque lobortis elementum augue, at suscipit justo malesuada at. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Praesent rhoncus convallis ex. Etiam commodo nunc ex, non consequat diam consectetur ut. Pellentesque vitae est nec enim interdum dapibus. Donec dapibus purus ipsum, eget tincidunt ex gravida eget. Donec luctus nisi eu fringilla mollis. Donec eget lobortis diam.

Suspendisse finibus placerat dolor. Etiam ornare elementum ex ut vehicula. Donec accumsan mattis erat. Quisque cursus fringilla diam, eget placerat neque bibendum eu. Ut faucibus dui vitae dolor porta, at elementum ipsum semper. Sed ultrices dui non arcu pellentesque placerat. Etiam posuere malesuada turpis, nec malesuada tellus malesuada.

# Machine learning methods for the interpretation of next generation sequencing data: applications to leukemia and demographic inference

Christopher B. Cole

Exeter College

University of Oxford

A thesis submitted for the degree of

*Doctor of Philosophy*

Michaelmas 2021

# Acknowledgements

## Personal

This is where you thank your advisor, colleagues, and family and friends.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vestibulum feugiat et est at accumsan. Praesent sed elit mattis, congue mi sed, porta ipsum. In non ullamcorper lacus. Quisque volutpat tempus ligula ac ultricies. Nam sed erat feugiat, elementum dolor sed, elementum neque. Aliquam eu iaculis est, a sollicitudin augue. Cras id lorem vel purus posuere tempor. Proin tincidunt, sapien non dictum aliquam, ex odio ornare mauris, ultrices viverra nisi magna in lacus. Fusce aliquet molestie massa, ut fringilla purus rutrum consectetur. Nam non nunc tincidunt, rutrum dui sit amet, ornare nunc. Donec cursus tortor vel odio molestie dignissim. Vivamus id mi erat. Duis porttitor diam tempor rutrum porttitor. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed condimentum venenatis consectetur. Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Aenean sit amet lectus nec tellus viverra ultrices vitae commodo nunc. Mauris at maximus arcu. Aliquam varius congue orci et ultrices. In non ipsum vel est scelerisque efficitur in at augue. Nullam rhoncus orci velit. Duis ultricies accumsan feugiat. Etiam consectetur ornare velit et eleifend.

Suspendisse sed enim lacinia, pharetra neque ac, ultricies urna. Phasellus sit amet cursus purus. Quisque non odio libero. Etiam iaculis odio a ex volutpat, eget pulvinar augue mollis. Mauris nibh lorem, mollis quis semper quis, consequat nec metus. Etiam dolor mi, cursus a ipsum aliquam, eleifend venenatis ipsum. Maecenas tempus, nibh eget scelerisque feugiat, leo nibh lobortis diam, id laoreet purus dolor eu mauris. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Nulla eget tortor eu arcu sagittis euismod fermentum id neque. In sit amet justo ligula. Donec rutrum ex a aliquet egestas.

## Institutional

If you want to separate out your thanks for funding and institutional support, I don't think there's any rule against it. Of course, you could also just remove the subsections and do one big traditional acknowledgement section.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut luctus tempor ex at pretium. Sed varius, mauris at dapibus lobortis, elit purus tempor neque, facilisis sollicitudin felis nunc a urna. Morbi mattis ante non augue blandit pulvinar. Quisque nec euismod mauris. Nulla et tellus eu nibh auctor malesuada quis imperdiet quam. Sed eget tincidunt velit. Cras molestie sem ipsum, at faucibus quam mattis vel. Quisque vel placerat orci, id tempor urna. Vivamus mollis, neque in aliquam consequat, dui sem volutpat lorem, sit amet tempor ipsum felis eget ante. Integer lacinia nulla vitae felis vulputate, at tincidunt ligula maximus. Aenean venenatis dolor ante, euismod ultrices nibh mollis ac. Ut malesuada aliquam urna, ac interdum magna malesuada posuere.

# Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Pellentesque sit amet nibh volutpat, scelerisque nibh a, vehicula neque. Integer placerat nulla massa, et vestibulum velit dignissim id. Ut eget nisi elementum, consectetur nibh in, condimentum velit. Quisque sodales dui ut tempus mattis. Duis malesuada arcu at ligula egestas egestas. Phasellus interdum odio at sapien fringilla scelerisque. Mauris sagittis eleifend sapien, sit amet laoreet felis mollis quis. Pellentesque dui ante, finibus eget blandit sit amet, tincidunt eu neque. Vivamus rutrum dapibus ligula, ut imperdiet lectus tincidunt ac. Pellentesque ac lorem sed diam egestas lobortis.

Suspendisse leo purus, efficitur mattis urna a, maximus molestie nisl. Aenean porta semper tortor a vestibulum. Suspendisse viverra facilisis lorem, non pretium erat lacinia a. Vestibulum tempus, quam vitae placerat porta, magna risus euismod purus, in viverra lorem dui at metus. Sed ac sollicitudin nunc. In maximus ipsum nunc, placerat maximus tortor gravida varius. Suspendisse pretium, lorem at porttitor rhoncus, nulla urna condimentum tortor, sed suscipit nisi metus ac risus.

Aenean sit amet enim quis lorem tristique commodo vitae ut lorem. Duis vel tincidunt lacus. Sed massa velit, lacinia sed posuere vitae, malesuada vel ante. Praesent a rhoncus leo. Etiam sed rutrum enim. Pellentesque lobortis elementum augue, at suscipit justo malesuada at. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Praesent rhoncus convallis ex. Etiam commodo nunc ex, non consequat diam consectetur ut. Pellentesque vitae est nec enim interdum dapibus. Donec dapibus purus ipsum, eget tincidunt ex gravida eget. Donec luctus nisi eu fringilla mollis. Donec eget lobortis diam.

Suspendisse finibus placerat dolor. Etiam ornare elementum ex ut vehicula. Donec accumsan mattis erat. Quisque cursus fringilla diam, eget placerat neque bibendum eu. Ut faucibus dui vitae dolor porta, at elementum ipsum semper. Sed ultrices dui non arcu pellentesque placerat. Etiam posuere malesuada turpis, nec malesuada tellus malesuada.

# Contents

**Appendices**

# List of Figures

# List of Tables

# Glossary

**ARG** Ancestral Recombination Graph. 10

**CAHG** Central African Hunter Gatherer. 12

**IBD** Identity by Descent. 9

**IMF** Integrated Migration Fraction. 17

**ka BP** kiloanni (thousands of years) before present. 43

**mtDNA** Mitochondrial DNA. 9

$N_e$ Effective Population Size. 37, 38

**NGS** Next Generation Sequencing. 9, 13

**SAHG** South African Hunter Gatherer. 12

**SMCSMC** Sequential Monte Carlo for the Sequentially Markovian Coalescent. 10, 12, 16

**TMRCA** Time to Most Recent Common Ancestor. 9

*In the beginning there was nothing, which exploded.*

— Terry Pratchett, *Lords and Ladies*

# 1 Next Generation Sequencing

Introduction Outline:

- Genomic Sequencing Data
    1. Enabling advances in Sequencing technologies
    2. Applications of sequencing technologies
        (a) Population variation (whole genome sequencing)
            i. Technical error and variant calling
            ii. Issues with assembly in diverse populations
        (b) Functional genomics (ATAC-seq, DNAse-seq, ChIP-seq, etc.)
            i.
- Machine Learning and Statistical Inference for sequencing data
    1. Reasons for considering machine learning for these two problems.
- **Aim**: Use sequencing data and machine learning to find new ways to answer outstanding issues at the population and cellular scale.
    1. **Specific Aim 1**: Estimate a coloured ancestral recombination graph from WGS Data

   (a) Modelling ancestry with the coalescent

   (b) Traditional approaches to modelling demography

   (c) Sequential Monte Carlo

   (d) Advantages, including directional migration

2. **Specific Aim 2**: Predict Regulatory Regions of MLL-AF4 Leukemia from ATAC-seq

   (a) Traditional methods for genomic sequence annotation

   (b) Artificial Neural Networks

   (c) Adopting artificial intelligence models for sequence based learning

   (d) Advantages , including in silico mutagenesis

# 2

# A particle filter for demographic inference

### 2.0.1 The Ancestral Recombination Graph (ARG) and admixture inference in the ancient past

Should I take this out and put it in the previous chapter introducing the method? It's kind of a reminder here but not super relevant as this is the "application" chapter

The phylogenetic trees over a set of samples as they change along the genome through recombination, collectively referred to as the ancestral recombination graph (ARG) [1, 2], record all information about the samples' evolutionary history. This history itself is shaped by the population's demography, a statistical relationship that is quantified by the coalescent-with-recombination (CwR) model [1]. The ARG is a complex data structure which is only weakly constrained by the observed genetic polymorphisms, making inference of demography difficult. By making approximations to the CwR, for instance by making an independent-sites assumption, efficient parametric inference of demography becomes possible [3, 4]. Methods including `PSMC` [5], `diCal` [6] and `SMC++` [7] allow non-parametric inference of demography under a closer approximation to the CwR, but one that does not include gene flow between populations. MSMC [8] introduced the cross-coalescent rate and `MSMC-IM` described how to interpret this rate in the context of a isolation-migration model to estimate a migration rate between populations[9]. However, these methods are not well suited for estimating directional migration rates.

Here we extend SMCSMC (Sequential Monte Carlo inference of the Sequentially Markovian Coalescent, [10]) to allow inference of directional migration. SMCSMC is a Bayesian method that uses a particle filter to explicitly sample from the posterior distribution of ARGs over multiple diploid samples under the full CwR model. Since particle filters operate by simulating latent variables (here the ARG) under the statistical model of interest, it becomes possible to handle complex demographic scenarios. We exploit this by extending the CwR model to include time-varying directional migration rates in a two-island demographic model. We use the posterior sample of ARGs including migration events to update the parameters of the demographic model, using either expectation-maximization or a variational Bayes procedure, and iterate these steps until convergence. We apply SMCSMC to estimate directional migration rates in whole genome sequencing data from the Simons Genome Diversity Panel (SGDP) [11] and the Human Genome Diversity Panel (HGDP) [12] to investigate population structure around the OoA event.

### 2.0.2 Migration initialisation Values

We select a more comprehensive set of initiation pardameters and particle values and use them to analyze a Yoruban and French individual from SGDP (Fig **??**). The effect of the initial migration rate seems relatively consistent for low values (0.5 - 2.0),while an increasingly small migration peak is seen for higher initial magnitudes 4.0 - 10.0. Again, beginning with an initial rate of zero tends to lead to highly unstable estimates of effective population size and migration rates. For the remainder of the analyses in this article, we choose to use an initial rate of 1.0.

## Contents

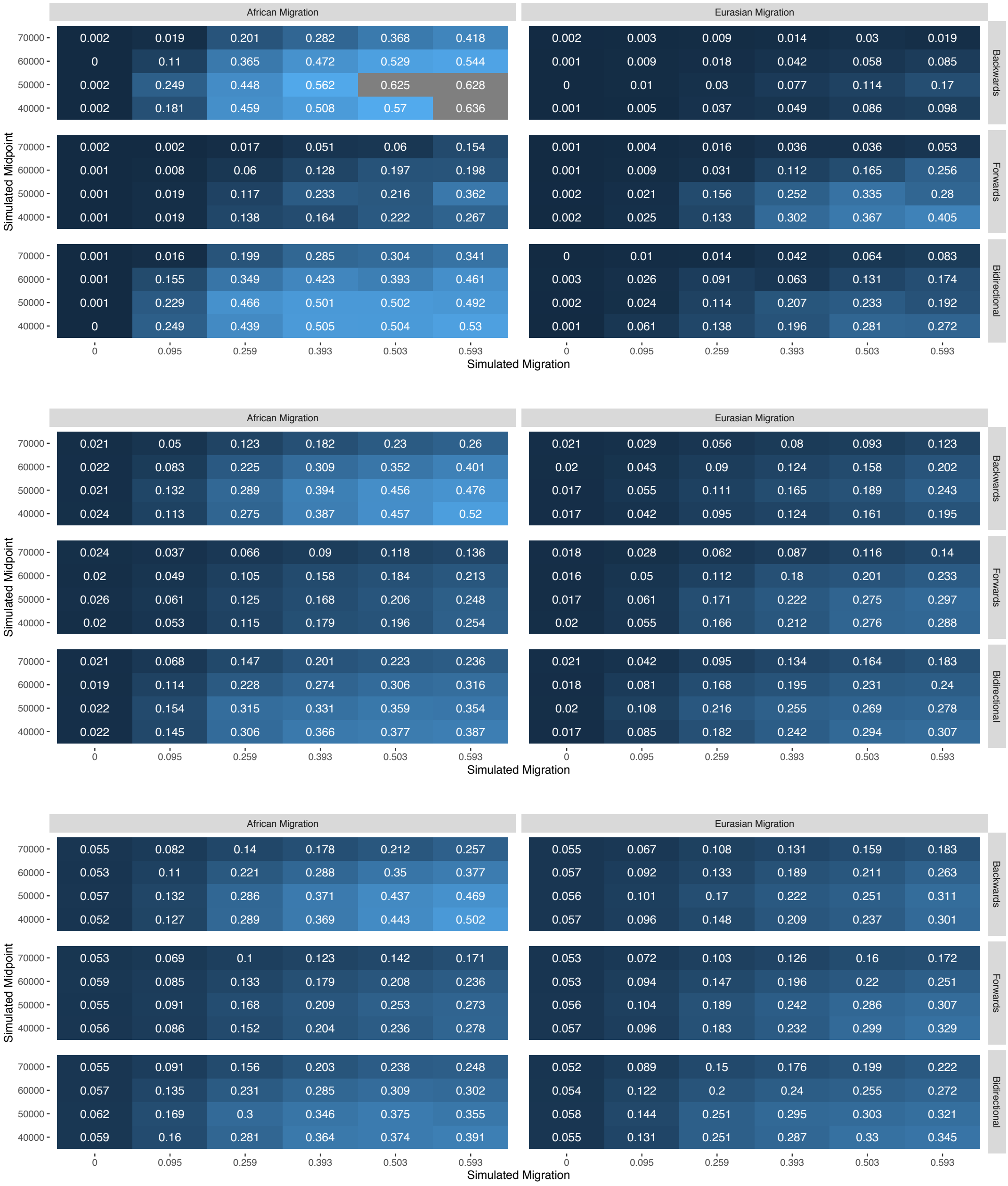


**Figure 2.1: Integrated migration fraction (IMF) in the last 100ky for three cases of simulated demography shown in 3.15, 3.16, and 3.17.** Simulations were performed as per Supplemental Section 3.2.7 with an additional parameter for the initial migration rate used to initialise the SMCSMC particle filter. The timing and IMF simulated are as per the aforementioned figures. From top to bottom, inference was initiated with 0, 1, and 5 $4N_0$ population replacement per generation in the specified direction (backwards, bidirectionally, and forwards).

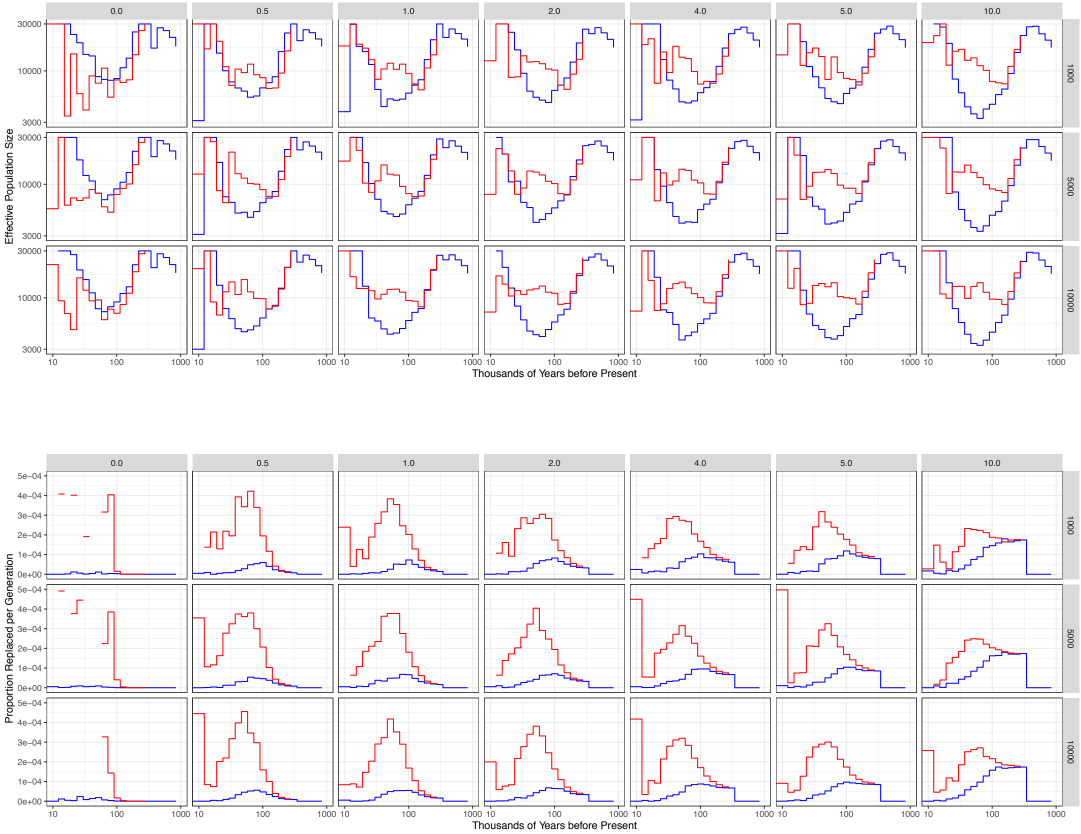


**Figure 2.2: The effect of initial migration parameter on demographic inference.** Effective population size and migration history of a Yoruban (`S_Yoruba-1`) and a French (`S_French-1`) individual from the Simons Genome Diversity panel were modelled with SMCSMC. The initial migration proportion, in units of 4 $N_0$ proportion of the population replaced per generation was varied along the X axis, while the number of particles is varied along the Y. 10 iterations of variational Bayes was used for parameter inference, while 5000 particles were used infer the ancestral recombination graph.

# 3
# Ancient Admixture into Africa from the Ancestors of non-Africans

Text and figures in this chapter have been largely adopted from:

Christopher B. Cole et al. “Ancient Admixture into Africa from the ancestors of non-Africans”. In: *bioRxiv* (2020).

Genetic diversity across human populations has been shaped by demographic history, making it possible to infer past demographic events from extant genomes. Next generation sequencing of diverse populations has allowed for the possibility of understanding foundational events in human evolution, however methodological barriers lie between these complex datasets and meaningful inference. In this chapter, I apply the extensions made to the SMCSMC framework to infer directional migration alongside population size in a poorly understood period of history, the Late Middle Paleolithic and the Out of Africa migration. I find evidence for substantial migration from the ancestors of present-day Eurasians into African groups between 40 and 70 thousand years ago, predating the divergence of Eastern and Western Eurasian lineages. In addition, I present suggestive evidence that this lineage diverged from the remainder of the migrants before their interbreeding event with Neanderthals. This migration event accounts for previously unexplained genetic diversity in African populations around this epoch, and supports the existence of novel population substructure in the Late Middle Paleolithic. Additionally, I demonstrate the importance of parameterizing the full migration matrix for the inference of other demographic parameters of interest. These results indicate that our species’ demographic history around the out-of-Africa event is more complex than previously appreciated.

## Contents

## 3.1 Introduction

The history of a population shapes its patterns of genetic diversity (cite). A nuanced understanding of the historical relationships between global populations has been hindered by both the availability of diverse Next Generation Sequencing (NGS) data and methodological hurdles. With the recent completion of several repositories of genomic variation across the globe, and the advances presented in the previous chapter allowing for tractable inference of directional migration rates, here I aim to investigate ancient contributors to extant genomic variation using machine learning on sequence data.

Before the era of NGS, an understanding of diverse populations was mostly based on combinations of variants in Y chromosome and Mitochondrial DNA (mtDNA) expected to derive from a common ancestor [14]. As these pieces of the genome do not recombine, each comprise a single genealogical tree dating back to a common ancestor. This attractive property allowed geneticists in the late twentieth century to study the evolution of these two lineages in great depth, leading to an extensive catalog of variation within so-called haplogroups (see `https://isogg.org/tree/index.html`). As mtDNA is inherited maternally and Y chromosome DNA is interited paternally, they represent unique viewpoints into sexually dimorphic anthropological processes (for a review, see Kivisild [15]). Results from haplogroup analyses have motivated many important discoveries, however caution must be taken when interpreting the results of a single tree in light of the whole of human history [16]. For the general case of investigating population structure across the globe in light of human history, the nuclear genome must be focus of our study.

Historically, patterns in genomic diversity have been investigated through differences in allele frequencies between subpopulations. Metrics such as the fixation index are often use to summarize differences between subsets of the population, either in terms of their relative variances, their probability of Identity by Descent (IBD) or their Time to Most Recent Common Ancestor (TMRCA) [14]. Modern adaptation of Wright's F statistics include so-called drift statistics, which quantity the degree to which two populations share genetic drift [17]. These statistics can be built up to create intricate descriptions and statistical tests for shared drift, ($f_3$) admixture ($D$, $f_4$), number of introgression events (`qpWave`) and even entire global topologies of population differentiation with branch lengths (`qpAdm`) [18]. These approaches form a highly useful toolkit for studying easy to gather polymorphism data, but do not attempt to recapitulate the processes by which this variation is produced. In order to explore this much larger parameter space, machine learning is necessary.

A given set of individuals will share recombination and coalescence events along their genomes, represented by a series of trees encoded in a data structure known as the Ancestral Recombination Graph (ARG) [2]. Theoretically, all information about a sample's genetic history is encoded in its ARG, if it could be examined directly [19]. Recently, methods which use stochsatic inference techniques have been able to directly sample from the posterior distribution of marginal genomic trees and produce representations of probable ancestral recombination graphs [2, 19, 20]. Methods for doing this inference analytically are currently outside the scope of the field, so machine learning based approaches which learn probable representations from data represent a crucial element of ARG inference. As introduced in chapter 2, I extend the approach taken by SMCSMC to infer directional migration rates simultaneously with effective population size, enabling a unique viewpoint into the ancient past.

A particularly interesting time in human history is the end of the Middle Paleolithic approximately ∼60 ka BP, which saw the divergence of the most deeply sampled lineages of human genetic variation, introgression from multiple archaic sources, and the expansion of anatomically modern humans Out of Africa. As archaeological evidence and ancient DNA from this period are scarce, inference of demography from present-day genetic data is potentially very informative, though technically challenging. Here, we use the previously developed approach, SMCSMC, to infer population size and directional migration in a unique and dynamic period of ancient human development.

This section is structured as follows. Firstly, I give an introduction to pertinent historical and anthropological theories relating to this period of time so as to orient the reader. Secondly, I introduce competing approaches for the inference of ancestral recombination graphs and motivate the usage of SMCSMC for this application. Following this, I outline my contributions to this area of research, which involve the identification and characterization of a putative directional migration from the ancestors of modern day Eurasians to the ancestors of modern day Africans.

### 3.1.1 Out of Africa and the peopling of Eurasia

An abundance of archaeological and genetic evidence has shown that the continent of Africa is the historical source of all modern humans [21]. It contains more genetic sequence diversity than any other region of the world, so much so that the two haplotypes in a single African genome are less similar than any two haplotypes outside of the continent [11]. Evidence from climate science suggests that a combination of a gradual shift away from aridity in Northern Africa as well as short term dry-wet cycles may have motivated both global and local range expansions [22, 23]. The most pertinent of these range expansions is the migration Out of Africa and into Eurasia, an event which will form the basis of modern population structure around the globe. These migrants were probably diverged from sister populations within the continent for many tens of thousands of years before their eventual dispersal into the Levant and beyond [8, 12]. The population which formed this successful range expansion out of Africa experienced at least one, and potentially multiple, breeding events with Neanderthals [24]. Eventually, the group split into two distinct subpopulations around the time of the Ust'-Ishim individual, approximately 45 ka BP [25]. One of these populations went East, forming the basis of for East Asians and Aboriginal Australians, while the other went West, forming the initial Upper Paleolithic European hunter gatherers [26, 27]. These were not the only derivative groups from the original Out of Africa, as another earlier diverged population popularly known as "Basal Eurasians" are thought to have branched before the initial contact with Neanderthal populations and contribute to later European population structure [28]. From this point on, diversification occurred on a highly regional basis, beyond the scope of this brief review.

However, little is known about population structure within Africa prior to the expansion of agriculturalists and pastoral groups [29, 30]. Recent evidence from the handful of successfully sequenced ancient African genomes hint at large-scale population movements and admixture from multiple highly divergent, extinct populations, with complex affinities to current groups [31–33]. The majority of structure in the continent is derived from events in the Holocene, including the spread of Bantu languages from Western Central Africa both East and South, as well as admixture from pastoralists in the Near East and Western Eurasia [29]. Eastern Africans are the most closely related group to the ancestral Out of Africa migrants, though they show particularly high levels of ancestry related to neolithic populations from Iran and the Levant consistent with multiple waves of back-migration in the Holocene [26]. Evidence for recent admixture from Eurasian sources is well established, however the lack of ancient African DNA from the Pleistocene has confounded efforts to uncover interactions between the earliest inhabitants of the continent. While the migration event associated with establishing current global population structure has been confidently dated between 60-80 ka BP, Central African Hunter Gatherer (CAHG) and South African Hunter Gatherer (SAHG) such as the Mbuti and KhoeSan (without implying linguistic unity, defined as southern African hunter gatherers who speak non-Bantu languages which include a click consonant) may have diverged from other groups 200-250 ka BP [32, 34]. In the intervening millennia, fossils identified as AMH have been found in China about 80-120kya, Sumatra about 63-73 ka BP, and artifacts from Australia 65 ka BP [35–37]. Support for multiple migrations across Eurasia additionally comes from climate science, where four distinct periods of warming may have provided vegetated migration routes out of Africa (OoA) as early as 120 ka BP [23]. The extent of contributions to modern day populations from "ghost" populations is unknown, though controversially suggested in Australasia and South East Asia [11, 38–41] and Africa [19, 32, 42–45]. To a large degree, the fate of these anciently diverged populations and their contributions, if any, to modern day populations remains an open question. Using an extension to SMCSMC, I aim to use machine learning to investigate ancient contributions to modern day population structure within global NGS datasets.

## 3.2 Methods

### 3.2.1 A Particle Filter for Demographic Inference

Details of the Sequential Monte Carlo for the Sequentially Markovian Coalescent (SMCSMC) algorithm have been previously published [10] (see the URLs for an implementation). Additionally, an introduction to the methodology and my contribution to establishing power for migration related inference is outlined in Chapter 2: A particle filter for demographic inference. Briefly, SMCSMC builds an approximation of the posterior distribution of genealogical trees conditional on observed mutations along the genome using a particle filter, a method also known as sequential Monte Carlo sampling. It does so by simulating a number of sequences of genealogical trees (particles) under a fixed set of demographic parameters $\theta$ using the sequential coalescent sampler `SCRM` [46]. Simulated recombination events may change the local trees along the sequence. Particles are then weighted according to their conditional likelihood given observed polymorphisms. To avoid sample depletion, the set of particles is regularly resampled, which tends to remove and duplicate particles with low and high weight respectively. To further increase the efficiency of the procedure, the resampling procedure targets not the partial posterior distribution that includes polymorhpisms up to the current location, but also includes a "lookahead likelihood" term that approximates a particle's likelihood's dependence on subsequent polymorphisms, while ensuring that the estimate of the posterior tree distribution remains asymptotically exact. From a sample of trees from the posterior distribution, Variational Bayes (VB) or Stochastic Expectation Maximization (SEM) is used to update the estimates of demographic parameters $\hat{\theta}$. This is repeated over a given number of iterations, or until the estimate of $\theta$ has converged.

To add the ability to infer time-varying migration rates, we exploit the capabilities of SCRM to simulate ARGs under complex demographic scenarios, and collect

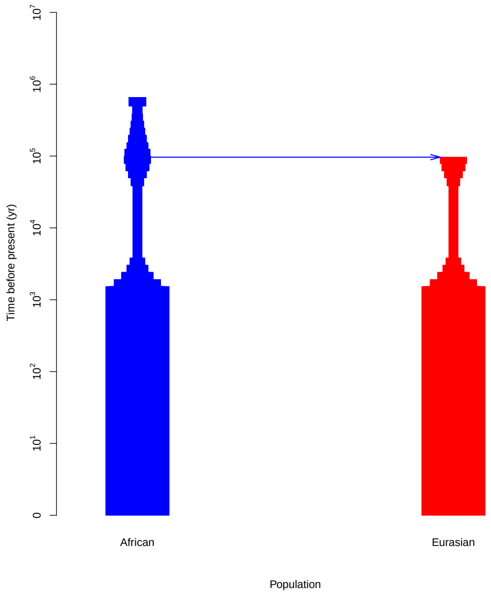


**Figure 3.1:** Demographic model used as initialisation for SMCSMC analysis visualised using PopDemog [47].

sufficient statistics (migration opportunity, and number, time and direction of simulated migration events) for each particle. This is described in chapter 2.

We use SMCSMC to infer effective population sizes and migration matrices in pairs of unrelated individuals from the phased release of the Simons Global Diversity Panel. We set a uniform recombination rate of $3 \times 10^{-9}$ and a neutral mutation rate of $1.25 \times 10^{-8}$, both in units of events per nucleotide per generation; previous results indicate that modeling recombination hotspots minimally affects results [5]. To reduce the number of iterations to convergence, we initialise the particle filter with an approximation of human demographic history (Figure 3.1). We seed the model with an initial constant symmetric migration rate of 0.0092 ($M_{i,j}$; proportion per generation of the sink population replaced by migrants from the source backwards in time). We arrive at this value through simulation in the previous chapter (subsection 2.0.2).

Unless otherwise noted, the directionality of migration is given forwards in time. That is, from population A in the past to population B in the future.

### 3.2.2 Multiply Sequential Markovian Coalescent

We use MSMC2 to estimate the effective population size of pairs of African and Eurasian individuals using default configurations and scripts provided in `msmc-tools` (see URLs) [8, 9]. We use a fixed recombination rate in line with our SMCSMC analysis and skip ambiguously phased sites. Twenty iterations are performed by default. We additionally compute the relative cross-coalescent rate to examine relative gene flow by transforming the coalescent rates generated by MSMC2 as indicated in the software documentation.

A brief example is given below, where I estimate coalescence rates within population 1 (Eurasians), population 2 (Africans), and between them given properly formatted input.

```
msmc2 -I 0,1 --fixedRecombination --skipAmbiguous -t 1 -o {output_prefix} {input_string}
msmc2 -I 2,3 --fixedRecombination --skipAmbiguous -t 1 -o {output_prefix} {input_string}
msmc2 -I 0-2,0-3,1-2,1-3 --fixedRecombination --skipAmbiguous -t 1 -o {output_prefix} {input_string}
python combineCrossCoal.py {input[2]} {input[0]} {input[1]} > {output[0]}
```

In order to convert between SMCSMC style input (seg files) and MSMC style input files, we use a custom script found here.

### 3.2.3 Inferring population size and migration rates in the Simons Genome Diversity Panel

This section describes analysis of the SGDP with both SMCSMC and MSMC. SMCSMC version 1.0.1 was installed from the conda package manager (also found at `https://github.com/luntergroup/smcsmc/releases/tag/v1.0.2`), MSMC2 version 2.1.2 was installed from Github (found at `https://github.com/stschiff/msmc2/releases/tag/v2.1.2`) and all analyses were performed on the Oxford Biomedical Research Computation cluster.

We download prephased sequencing data from `https://sharehost.hms.harvard.edu/genetics/reich_lab/sgdp/phased_data/` and mask for the strict accessibility mask from the 1000 genomes project. We additionally mask for any sites absent Chimpanzee ancestry due to a known issue with the phasing algorithm [9]. We perform this masking in vcftools. We use the SMCSMC python package function smcsmcv̇cf_to_seg to convert the sequence data from VCF to seg file format, a format very similar to MSMC format. We provide a script to convert from seg file format to MSMC file format as well. Unless otherwise noted, the names of individuals used in this paper are the first in their population (i.e. an individual named Yoruban is S_Yoruba1 in the SGDP nomenclature, full list in 3.1). We select two diploid individuals from each population in Africa and infer piecewise constant population size and directional migration rates. Specifically, we use the following options for SMCSMC:

```
smc2 -c -chunks 100 -no_infer_recomb -nsam 4 -I 2 2 2 -mu 1.25e-8
  -rho 3e-9 -calibrate_lag 1.0 -EM \${EM} -tmax 3.5 -alpha 0.0 \
  -apf 2 -N0 14312 -Np ${Np} -VB ${DEMOGRAPHIC_MODEL}
  -P 133 133016 31*1 -arg -o ${OUTPUT} -segs ${SEGS}
```

In order, we invoke the use of a QSUB cluster with -c and split our analysis into 100 chunk. We do not infer recombination sites along with the demographic model in order to reduce runtime. Four haploid samples, two from each population, are analysed with a fixed mutation rate of 1.25 $\times 10^{-8}$, a fixed recombination rate of 3 $\times 10^{-9}$, and accumulating events for one unit of survival time along the sequence. We use a given number of epochs for parameter units, and bound the upper limits of the trees at 3.5 times the effective population size (set to 14312). We use the look-ahead likelihood to guide the resampling process for a given number of particles Np and use variational Bayes in place of the default stochastic expectation maximization algorithm. Parameters are inferred over 31 equally spaced intervals from 133 to 133016 generations in the past, and the sampled posterior ARGs are reported. The choice of these parameters is discussed in depth in [48].

The demographic model used to seed inference is given on page 64. We visualise this demographic model in the POPdemog package in Figure 3.1 [47]. This demographic model has been designed to roughly mimic human population size history without overly biasing the results from inference.

Each SMCSMC analysis gives a final output file detailing migration and coalescent events, their rates, and their opportunities which denote the total opportunity for an event to occur during a particular epoch. Output files are trimmed to only visualise the final iteration of variational Bayes inference and assessed for convergence. Times and rates are interpreted differently than `scrm` output. Rates are in units of $4N_0$ per generation (defined here as 29 years as per [49]), while times are given in generations.

We implement the above in a `Snakemake` pipeline. Sample size and relative cross-coalescent rates are transformed as described in the documentation using the same parameter values for mutation rate and generation time used for SMCSMC analysis.

Migration during the last 100ky is integrated into a metric we call the Integrated Migration Fraction (IMF). This is related to the cummulative migration fraction (CMF) as introduced in MSMC-IM [9], except the quantity is integrated in a particular epoch. IMF is calculated as a function of time $F(t) = e^{-\int_{t=0}^{T} \rho(t)dt}$ given an upper bound $T$. A practically identical solution can be found from first prinicples. Consider $p$ proportion of the population are replaced every generation. Start with 0 individuals from the source $N_{source}$ population in the sink population $N_{sink}$, each generation replace $p$ proportion of the sink population with the source. We track the proportion of the population which are replaced by the source $P$.

$$\begin{aligned}
P_0 &= 0 \\
P_1 &= pN_{sink} \\
P_2 &= pN_{sink} + p(N_{sink} - pN_{sink}) \\
&= pN_{sink} + pN_{sink}(1-p) \\
P_3 &= pN_{sink} + pN_{sink}(1-p) + p((N_{sink} - pN_{sink}) - p(N_{sink} - pN_{sink})) \\
&= pN_{sink} + pN_{sink}(1-p) + p(N_{sink}(1-p) - pN_{sink}(1-p)) \\
&= pN_{sink} + pN_{sink}(1-p) + pN_{sink}(1-p)(1-p) \\
&\dots \\
P_n &= N_{sink}p(1-p)^n
\end{aligned}$$

In practice, both methods give essentially identical proportions for all considered questions.

### 3.2.4 Statistical Analysis of Migrated Segments

We run SMCSMC with the `-arg` flag to report the posterior estimate of the ancestral recombination graph. We use this to isolate segments of the African genome where predicted migration events occurred between 50 and 70kya and used these segments to calculate drift statistics. The isolation procedure is implemented in `smcsmc.find_segments`, and involves sequentially reconstructing marginal trees and keeping track of which contain migration events in a particular epoch. We isolate segments from the marginal trees of all SGDP comparisons.

### 3.2.5 Length Distribution of Isolated Segments

Under the Markovian model of the SMC', the length of admixed tracts $L$ is an exponential process with scale factor $2N(1-m)\left(1-e^{-T/2N}\right)$, with a proportion $m$ of the sink population being replaced with the source $T$ generations in the past and an effective population size of $N$ [50, 51]. This gives an approximate mean length $[(1-m)r(T-1)]^{-1}$ with recombination rate $r$ in units of Morgans, which is well approximated by $(rT)^{-1}(1-m)$ [52]; we use this approximation to derive expected distribution of fragment sizes. When analysing populations with SMCSMC, we fix the recombination rate at $3 \times 10^{-9}$ uniformly across the genome, in line with that used by tt MSMC in simulations [8]. This value is a conservative underestimate, accounting for the presence of recombination hotspots and SMCSMC's inability to deconvolve recombinations in these areas, effectively underestimating the true $r$. For estimates of ancestral tract lengths, we use the more universally accepted value of $1 \times 10^{-8}$, equivalent to a one percent chance of a cross-over per megabase and per generation [53].

### 3.2.6 Drift Statistics

Patterson's drift statistics were calculated with ADMIXTOOLS [17] and the admixr package [18] in R. We converted the above sequence data to Eigenstrat format with vcf2eigenstrat formerly distributed with admixr. We merged SGDP and archaic Eigenstrat datasets with convertf and mergeit implemented in ADMIXTOOLS.

Here, Yoruba-1 is used as a representative of Western African groups, and used for ascertaining putatively migrated segments. Yoruba-2 is used as a comparison individual from the same populations. In this way, we look for evidence above another individual in the same population of similarity to Eurasians.

### 3.2.7 Simulation procedure

Coalescent simulations were performed under the sequential coalescent with recombination model (SCRM) [46]. 1 gigabase (Gb) of sequence was simulated. In addition to branches in local genealogical trees, SCRM retains non-local branches in the ancestral recombination graph (ARG) within a user-specified sliding window. In the limit of a chromosome-sized windows SCRM is equivalent to the coalescent with recombination, while for a zero-length window it is equivalent to the sequentially Markovian coalescent (SMC') [4, 50]; we use a 100kb sliding window to approximate the CwR and improve accuracy over SMC' while retaining tractable inference.

We modelled migration as a 10ky pulse of constant migration rate resulting in an integrated migration fraction (IMF) of 0 to 0.593. The migration pulse was centered at various times between 40 and 70 kya. Due to the amount of compute required, we then used SMCSMC to infer the demographic parameters using a reduced set of 5000 particles and 5 iterations of the VB procedure. To aid convergence, we started inference at a reasonable approximation of human demographic history (Figure 3.2). We modelled $N_e$ and migration rates as piecewise continuous functions and set 32 exponentially spaced epochs from 133 to 133016 generations in the past. To convert evolutionary rates to years we set a generation time of 29 years [49]. For computational efficiency, individual genomes were split into 120 chunks and processed in parallel, with sufficient statistics collected and processed together in the VB steps.

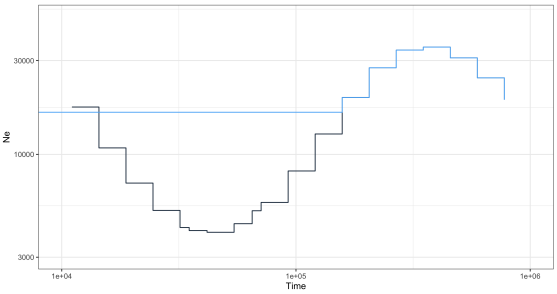


**Figure 3.2: Effective Population size model used for simulations.** Following the simulation procedure in Supplemental Section **??**, the population size model is plotted per epoch, with the effectively African population plotted in blue and the effectively non-African population in black.

Times are in units of $4N_0g$ while population sizes are in units of $N_0$. For $g = 29, N_0 = 14312$, the demographic model is as shown in Figure 3.2. Exact specifications are given in A

The demographic model which we have assumed for both population's effective sizes has been shown to recapitulate similar inference to real data (data not shown). The migration parameter must be initiated at a given magnitude; further back in time, the particle filter is less able to identify lineage's true populations, and the inference of migration rates becomes essential uniform. Thus, we see a "drop-off" effect, where in the ancient past, the inference remains at the initiation value, and as more certainty about different histories is obtained, the migration values recapitulate real information. Thus the choice of an appropriate parameter for the initial migration rate is a crucial step in SMCSMC analysis, and here we chose to arrive at this value through simulation.

We simulate back-migration scenarios of varying total migration proportions from 0 (no migration) up to 60% population replacement. For each simulation, we initiate the particle filter at either 0, 1, or 5 $4N_0$ proportion replaced per generation (which are the units used internally by `scrm` and `ms` for simulation). SMCSMC is then used to infer effective population size and migration histories in five iterations with 5000 particles. As a cautionary note, these simulations are almost certainly not fully converged, and are used as an indication of power. Their power, theoretically, approaches 1, as particle filters asymptotically exactly approach the true posterior distribution. However, these low resolution attempts are indicative of a "quick" overview of the abilities of the algorithm. With 600 cores available, each of the cases (forward, backward, or bidirectional) was able to run in approximately 20 hours.

Generally, beginning with a higher migration rate seems to recover a higher proportion of the simulated migration. However, as in the case of a 60% replacement simulated 40kya, beginning with 5 $4N_0$ rather than 1 $4N_0$ recovers similar proportions of backwards migration (0.502 vs 0.52) yet the higher migration rate finds 0.301 Eurasian migration rather than 0.195. The higher initial migration rates thus slightly reduce power (though, not in all cases, and for fully converged solutions, we would expect both proportions to be similar up to noise) while additionally finding an increased migration in the opposite direction. Beginning with a zero rate leads to highly unstable estimates of the migration rate and effective population size, and we exclude it from our analysis.

### 3.2.8 Isolating Anciently Admixed Segments

We sampled genealogical trees with migration events from the posterior distribution estimated by the particle filter under the final, converged, demographic parameters. We scan along the sequence and identified marginal trees with migration events from the source (Eurasian) population to the sink (African) population (forward in time) within the desired time period along with the beginning and end position of that tree in the genome sequence. In this process, we ignore recombination events that alter a tree in such a way that the migration event is retained.

### 3.2.9 Sequence Data and Preparation

We downloaded whole genome sequence (WGS) data from the phased release of the Simons Genome Diversity Panel and converted it to `.seg` file format using scripts provided (See URLs). We apply two masks to the data. First, we mask the data with the strict accessibility mask provided by the 1000 genomes project (see

URLs). Second, we mask any sites absent chimpanzee ancestry, to address a known variant issue in the data that resulted in artificially long runs of homozygosity [9]. We develop a `Snakemake` [54] pipeline for efficiently analysing sequence data with both SMCSMC and MSMC2. We assume a mutation rate of $1.25 \times 10^{-8}$ and a recombination rate of $3 \times 10^{-9}$ (events per nucleotide per generation), in line with recent literature [55, 56]. The number of particles, and the number of VB iterations, are set per analyses, and are reported in figure captions. Unless otherwise noted, the names of individuals used in this paper are the first in their population (e.g. an individual named Yoruban is `S_Yoruba-1` in the SGDP nomenclature); a complete list of sample identifiers is provided in Supplemental Table 3.1.

### 3.2.10 Integrated Migration Fraction

The IMF, the total fraction of a particular population $A$ replaced during a particular time period from $T_0$ to $T_1$ generations in the past is found as follows. Let $\rho(t)$ be the instantaneous rate of migration out of $A$ per unit of time in the backward direction (i.e. into $A$ forwards in time), and $F(t)$ the fraction not migrated in the epoch $[T_0, t]$, then $\frac{d}{dt}F(t) = -\rho(t)F(t)$ with solution $F(t) = e^{-\int_{T_0}^{T_1} \rho(t)\mathrm{d}t}$, so that the IMF is given by $1 - F(T_1)$. The integral is calculated as a finite sum since $\rho$ is piecewise constant.

## 3.3 Results

### 3.3.1 Substantial Migration from Eurasian to African Ancestors

We use SMCSMC to analyse pairs of individuals from the SGDP and simultaneously infer migration rates and effective population sizes ($N_e$) under a two-island model with directional migration. Population sizes and migration rates are modeled as piece-wise constant across 32 exponentially spaced epochs from 133 to 133016 generations in the past, corresponding to 3.8 thousand to 3.8 million years ago (3.8kya–3.8Mya) using a generation time $g = 29$ years [49]. We find that the method infers high rates of migration from descendants of the OoA event ('non-Africans') to Africans, but not in the opposite direction, in the period 30–70kya corresponding to the Late Middle Paleolithic (3.4). In populations from the Niger-Kordofanian and Nilo-Saharan language groups, comprising the majority of the population on the African continent, the peak inferred migration rate from Eurasian populations (2.5–$3.0 \times 10^{-4}$ and 3.5–$4.0 \times 10^{-4}$, in units of proportion of the target (ancestral African) population replaced per generation) most frequently falls in the epochs spanning 35–45kya, while peak migration rates in the opposite direction are substantially lower (0.5-1.0 $\times$ $10^{-4}$) and occur earlier, in the epochs spanning 55–70kya (3.5). Populations in the Afroasiatic language group show evidence of large amounts of directional migration in the Holocene (3.10), which is consistent with previous findings of relatively recent European introgression into these populations [29, 57].

We track the overall peak of migration rate in different populations (Figure 3.5a,b). The most common backwards migration peak falls in the epoch between 35–45kya in the Nilo-Saharan and Niger-Kordofanian groups. Forwards migration has an earlier peak, in the epoch spanning 55–70kya. This result must be interpreted in light of the simulation results presented below.

We model the migration adjusted $N_e$ in Eurasian populations, averaged over African partners, and African populations averaged over Eurasian partners 3.13. The resulting curves largely represent our prior knowledge of world history, with an early divergence of Papuans consistent with the timing proposed in [38], and a second bottleneck of populations inhabiting North America such as the Karitiana and Pima. Because we do not explicitly infer population split times, and have no convenient metric like MSMC2, more fine-scale trends are difficult to identify. The African population size models show more discrepancy between populations, including an OoA-like bottleneck in Afroasiatic populations, and a large historical population size in hunter-gatherer groups such as proposed in [32].

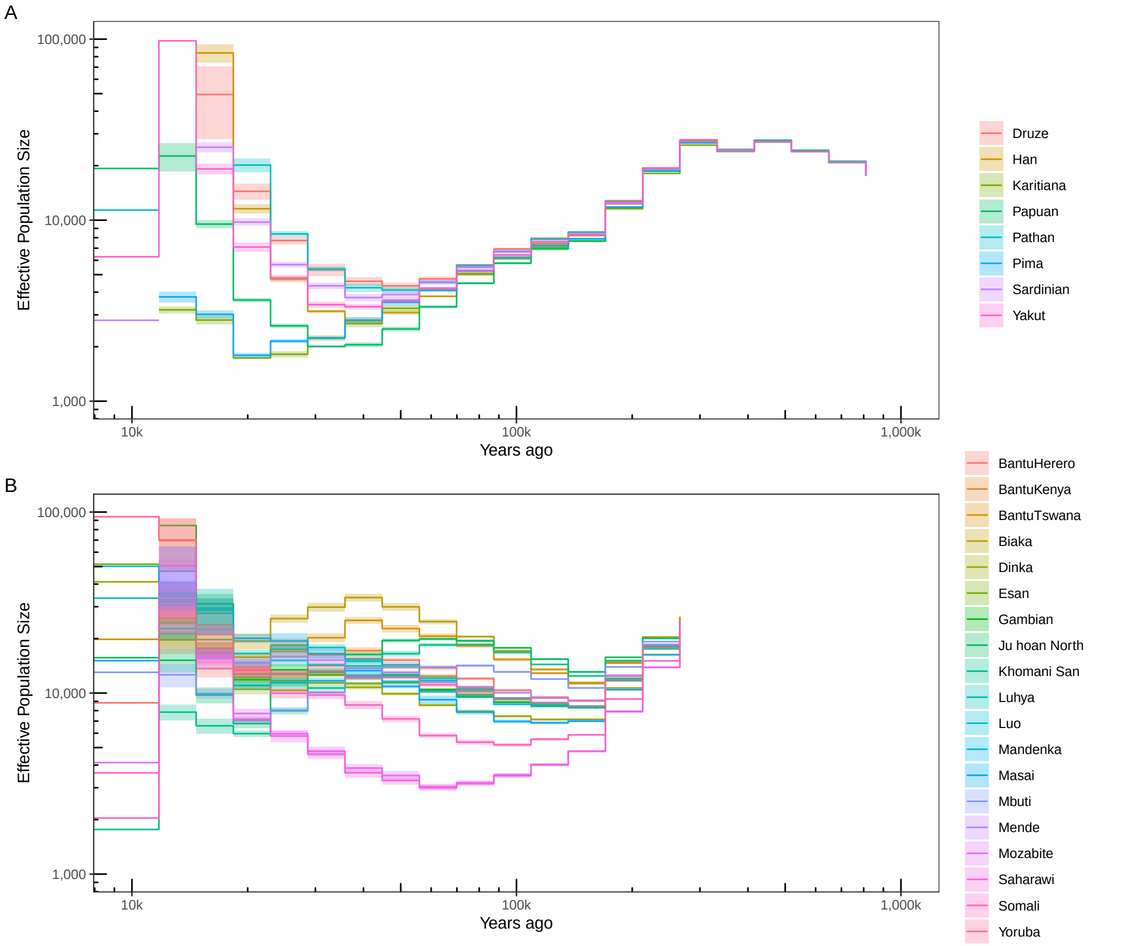


**Figure 3.3: Estimates of individual population sizes incorporating directional migration.** Using `SMCSMC` the effective population size of global populations in the Simons Genome Diversity Panel is inferred while simultaneously fitting directional migration estimates. Averages are plotted by epoch, with shaded regions denoting the standard deviation. **a.** Estimates of Eurasian population sizes when averaged over Eurasian donor populations. This analysis uses the eight Eurasian populations matched to HGDP populations averaged over the four matched African populations. B. Estimates of African population sizes when averaged over Eurasian recipient populations. This analysis uses the three donor Eurasian populations used for the majority of the analyses in the main text (French, Han, and Papuan) along with the given African populations. Before approximately 250kya, the populations share the same population size within the model, and are not plotted. 10,000 particles are used to approximate the ancestral recombination graph in the SMCSMC particle filter and 15 iterations are used to update demographic parameter values.

To assess the impact of errors introduced by statistical phasing, as is the case for the SGDP, we repeated the analyses above on a subset of physically phased individuals from the Human Genome Diversity Project (HGDP) [11]

### 3.3.2 Validation in a physically phased subset of the Human Genome Diversity Panel (HGDP)

Phased data is not essential for demographic inference using SMCSMC; however, the use of phase alongside the look-ahead likelihood allows for more efficient convergence. The Human Genome Diversity Project collected 929 genomes from a diverse collection of human populations [12]. 36 of these genomes, two each from nine Eurasian and four African populations, were physically phased by use of linked-read sequencing technologies. This resource allows us to validate our inference both in an independent dataset, and evaluate the effect of phasing errors on SMCSMC inference.

To analyse these data, the same `Snakemake` pipeline was used with minor adjustments in wildcard constrains to account for differences in sample names. 120 chunks of the genome were run in parallel for reasons of computational efficiency, while fixed recombination rate and mutation rates were held at the same values as the SGDP analysis, and an identical demographic model was used to initiate the analysis. Three replicates of the analysis were performed to assess the impact of stochastic sampling variation on inference. We infer both effective population size and directional migration in each of these 9x4 comparisons between Eurasian and African populations (Figure 3.11a). The resulting inference allows us to verify and validate many observations from the SGDP.

Firstly, we calculate the timing of the migration peak and its magnitude, and find the estimates largely in line with the SGDP inference (Figure 3.5c,d). For instance, inferred backwards migration in the Yoruban and Biaka populations peak at 40-50kya, while the Mbuti and San show earlier migration peaks around 50-60kya (Figure Figure 3.5c). The migration rate at the peak shows the same qualitative trends as the SGDP, with the peak in the Yoruban (approximately $2.5 \times 10^{-4}$) far exceeding the peaks in the Biaka, Mbuti, or San (between $0.1 - 0.175 \times 10^{-4}$)

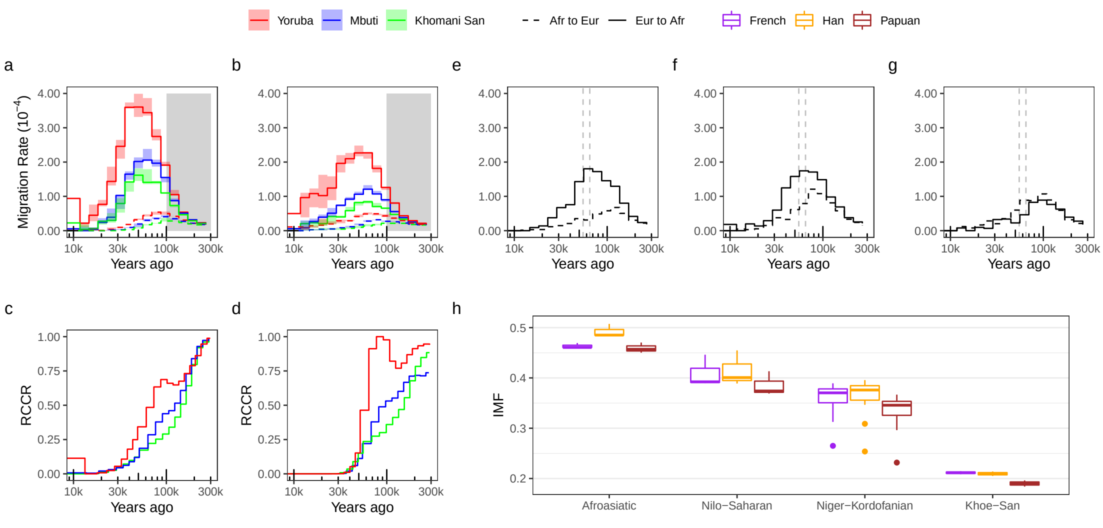


**Figure 3.4: Migration rate inference a**. Inferred migration between an African individual and a Han Chinese individual in the SGDP. Three replicates were performed, with the median estimate plotted and the range shaded. Solid lines show inferred migration from Eurasians to Africans (forward in time) while dotted lines show the reverse migration. The SMCSMC analysis used 10000 particles to estimate the posterior distribution of marginal trees, and 25 iterations of variational Bayesian inference to achieve converged parameter estimates. The shaded grey regions represents a time period where simulation shows SMCSMC has very little power to infer migration (details found in 3.2.7). **b**. The same analysis as in a. except using individuals from the physically phased subset of the HGDP, showing similar differences between populations but systematically lower migration overall. Three replications were performed to estimate error and the standard deviation is shaded. The same SMCSMC settings were used as in a. **c**. Relative cross-coalescence rate (RCCR) estimated by MSMC in three different populations in the SGDP, supporting gene flow between Eurasians and Yorubans not shared by Mbuti or Khoe-San. 40 iterations were used to achieve parameter convergence. **d**. The same analysis as in c. but performed on individuals in the physically phased subset of the HGDP, similarly supporting shared gene flow between the Yoruban and Eurasians not shared by Mbuti or Khoe-San. **e**, **f** and **g**. Inferred migration rates from from data simulated under a two-island model with, from left to right, a backward Eurasia-to-Africa, a bidirectional, and a forward migration pulse lasting 10 ka BP (dashed vertical lines) and replacing 40% of the recipient population(s) approximately 60kya. The migration rate from Africa to Eurasia is not well estimated by SMCSMC (3.15–3.17 and 3.2.7), but SMCSMC is well powered to infer migration from Eurasia to Africa in this period. **h**. Integrated total migration fraction (IMF) over the last 100 thousand years stratified by language phyla in the SGDP and comparison Eurasian population used to estimate migration. Afroasiatic (Mozabite, Saharawi, and Somali), Nilo-Saharan (Dinka, Luo, and Masai), Niger-Kordofanian (BantuHerero, BantuKenya, BantuTswana, Biaka, Esan, Gambian, Luhya, Mandenka, Mbuti, and Mende), and San (Khomani San and Ju hoan North) are grouped as in [57]. Similar levels of migration are inferred from French and Han Chinese to all language groups, with significantly less migration from Papuan groups ($p \leq 0.05$, two-tailed paired t-test, 3.3). Ourliers in the Niger-Kordofanian group are the Mbuti.

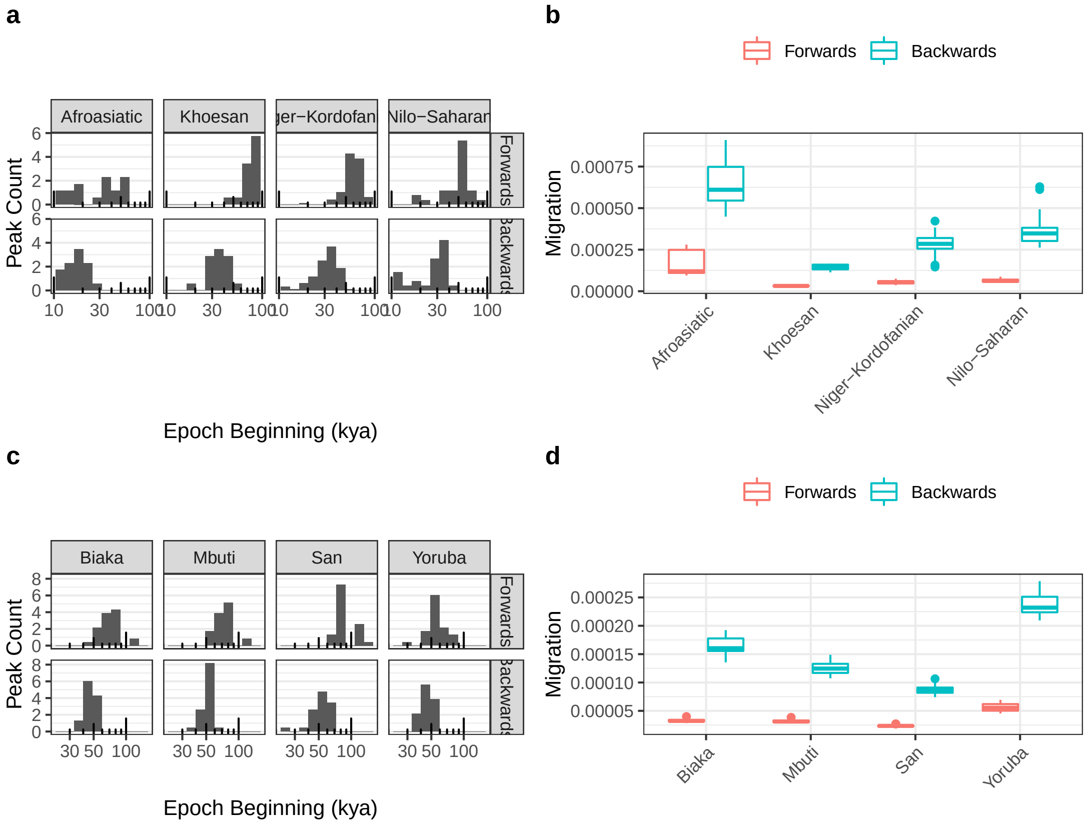


**Figure 3.5: Timing and average maximum rate of directional migration in HGDP and SGDP. a** Migration is inferred in evenly spaced epochs on the log scale from 3.8 thousand to 3.8 million years ago. For each population in the SGDP, we record the epoch with the highest inferred directional migration rate (the "peak" of migration) and plot this as a histogram. Backwards migration refers to migration from Eurasians to Africans, whilst forward represents the reverse. **b** In the epochs of highest migration identified in a., we record the inferred rate per population and plot these as a boxplot. Whiskers represent 1.5 times the interquartile range. The migration rate is given in proportion of the population replaced per generation. **c** and **d** represent the same analyses as in a. and b. calculated for the Human Genome Diversity Panel, rather than the SGDP.

(Figure Figure 3.5d). This replication in the HGDP confirms the presence of a large directional migration in the Late Middle Pleistocene, and demonstrates that statistical errors in phasing the SGDP are not large contributors to the qualitative trends observed.

We integrate migration between 40–70kya to obtain the inferred IMF for each of the comparisons in the HGDP (Figure 3.11b). Differences amung the African populations mirror those in the SGDP, though the proportions are uniformly smaller (main text). However, the migration rates backwards into Africa are

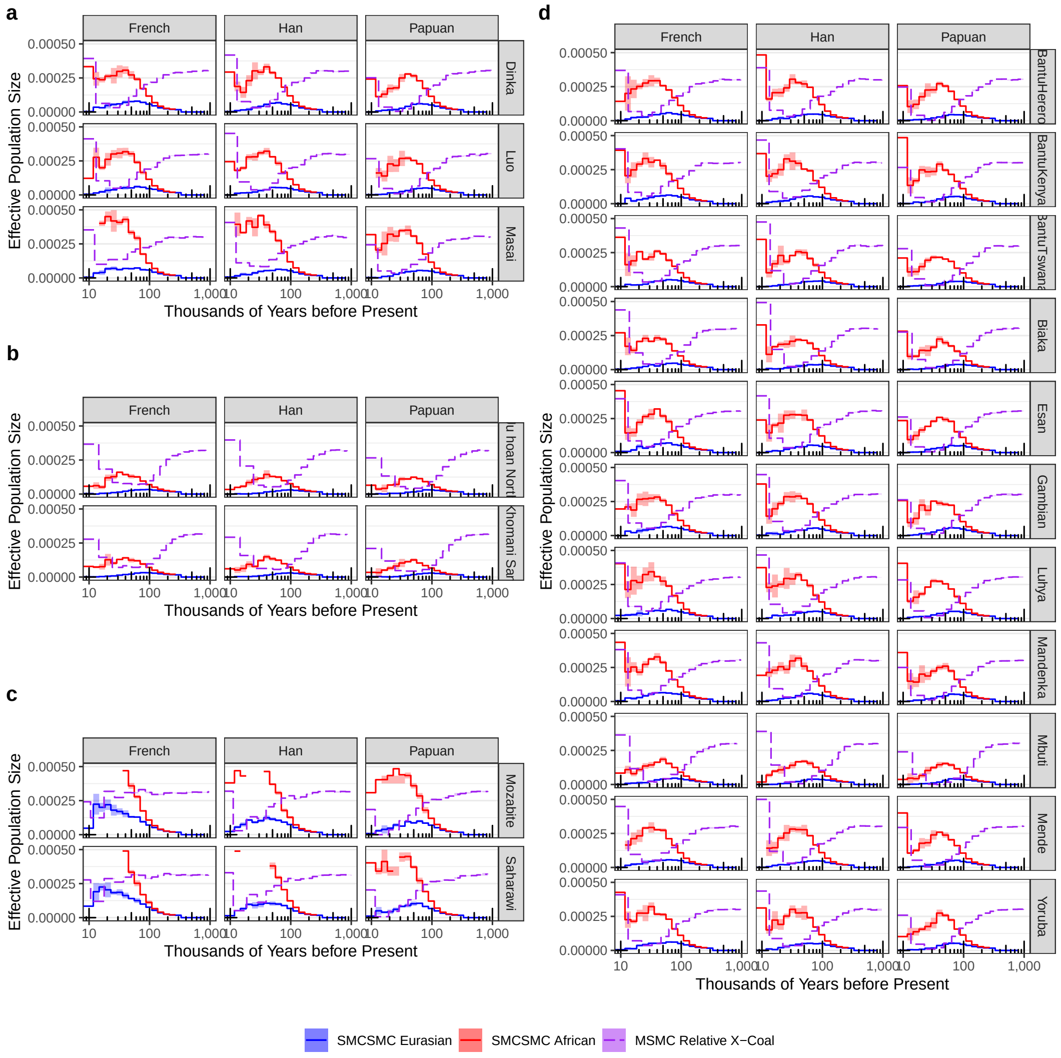


**Figure 3.6: Inference of directional migration in the Simons Genome Diversity Project.** The SMCSMC particle filter was used to infer directional migration rates in both directions from one of three Eurasian populations (French, Han, and Papuan) to one of 18 African populations. 5000 particles were used to approximate the ancestral recombination graph with 10 iterations of variational Bayesian inference to update demographic parameter values. Panels represent a. Nilo-Saharan, b. KhoeSan, c. Afroasiatic, and d. Niger-Kordofanian language families. Alongside the SMCSMC inference, we use MSMC2 to infer the relative cross coalescent rate (RCCR) with default settings and 20 iterations for convergence.

apparent in all comparisons, and the order of populations IMF remains the same. Differences between individual Eurasian donor populations are small, and with the exception of the Papuan, insignificant. A discussion of the Papuan comparisons appears in the subsequent section.

To compare the HGDP inference with the SGDP inference, we construct a set of the SGDP with the same donor populations as the HGDP.

### 3.3.3 Comparisons between the HGDP and a subset of the SGDP

Previously, inference in the SGDP has relied on three candidate Eurasian donor populations. However, the physically phased subset of the HGDP provides a higher resolution view into global migration patterns with nine Eurasian populations represented. In order to compare effectively between the inferences made in these two datasets, we find representatives from these nine Eurasian populations in the SGDP dataset and use them as donor populations to the same four African populations (Yoruban, San, Mbuti, and Biaka), effectively recreating the analysis done in the physically phased subset of the HGDP. We select the Khomani San as a representative of the San, and only use one of the Papuan populations in the HGDP to compare (Highlands, as opposed to Sepik), creating the same 8x4 analysis table for both data sets. We infer the effective population sizes and migration rates using both SMCSMC and MSMC, with analysis details effectively identical to the original comparisons listed above (Figure 3.7). We average over inferences to visually compare trends between the two datasets, in the same populations and compute the inferred IMF between 40–70kya (Figure 3.11).

In both the HGDP and the SGDP, MSMC estimates of African population size are higher than SMCSMC estimates in the ancient past (80 – 300kya) (Figure 3.11a). By modelling directional migration, we are able to account for excess genetic diversity in the ancestral African population in both datasets. Uncertainty in the estimates increases substantially nearer to the present, as would be expected with the SMCSMC method.

We summarise migration from 40–70kya in the HGDP similarly to the SGDP. The total inferred migration is lower in the HGDP than in the SGDP (Figure Figure 3.11b). We use this comparison setup to additionally test the differences between Papuan donors and the remainder of Eurasians. We construct a linear model predicting IMF based on an indicator variable of Papuan/not Papuan and the receptor donor population, and find that in both the SGDP and the HGDP, Papuans show approximately 2% less IMF than other donor populations (Table 3.5, Table 3.4). While this difference is small, it is highly significant. However, the demographic scenario causing this difference in inferred IMF is not obvious; it is possible that the Papuan group had begun to diverge from the donating population prior to the admixture event, or alternatively that differences in archaic admixture between Eurasian and Papuan groups make up the difference in affinity.

However, the qualitative patterns in inferred directional migration rates between populations are similar in both datasets (Figure 3.11c). In both datasets, the highest rates are found in the Yorubans, follow by the Biaka, then the Mbuti and San. The MSMC curves are interestingly dissimilar between the different data sets, with a much steeper ascent around the period of our inferred migration in the HGDP than the SGDP.

### 3.3.4 Simulation demonstrates power to infer large directional migration pulses

We asked whether SMCSMC has power to detect a large back-migration event in the Late Middle Paleolithic and distinguish it from other demographic scenarios. To answer this we used `SCRM` [46] to simulate a gigabase of sequence data under a two-island demographic model, with effective population sizes chosen to be comparable to typical African and Eurasian populations as inferred from real data. To this we added a 10ky pulse of forward, backward or bidirectional migration of varying strengths, with the midpoint of the migration pulse within the range 40 to 70kya. To quantify the inferred amount of migration we calculate the integrated migration fraction (IMF), defined as one minus the probability that a lineage in

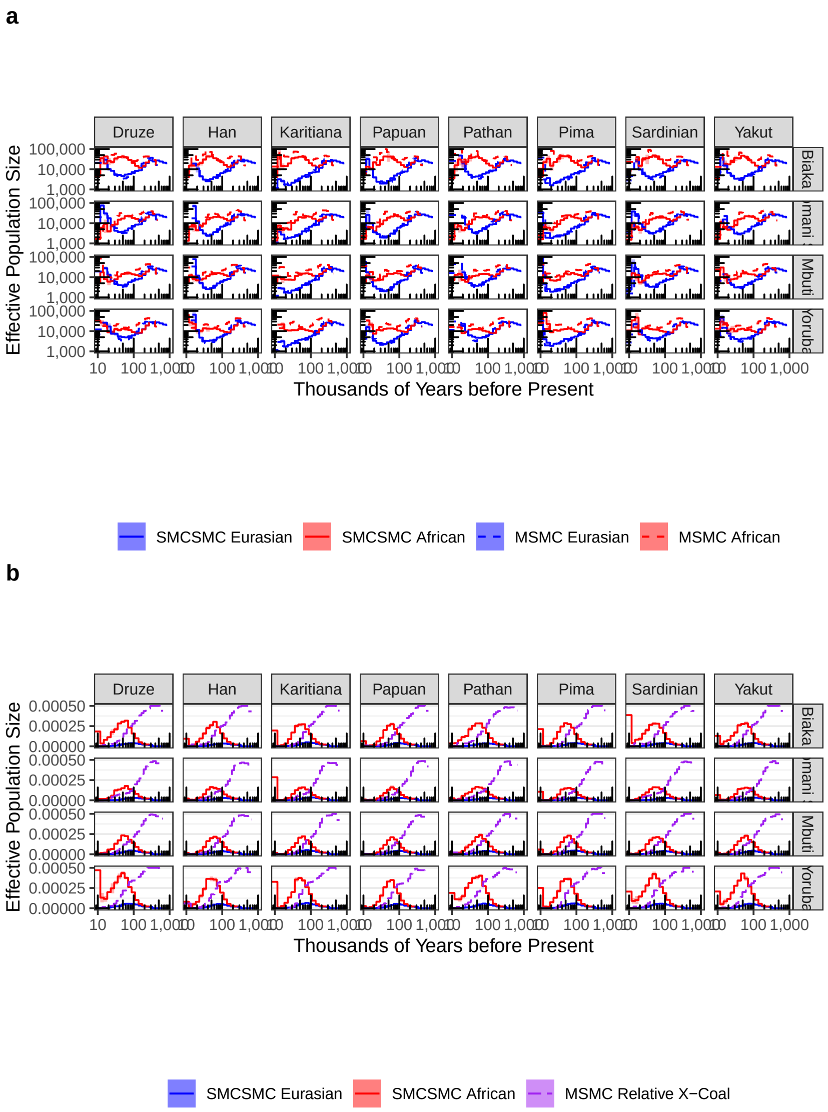


**Figure 3.7: Demographic inference in a matched subset of the Simons Genome Diversity Panel**. **a.** SMCSMC and MSMC2 inferred effective population size of several populations in the Simons Genome Diversity Panel. These samples were selected to match, as closely as possible, those in the physically phased subet of the Human Genome Diversity Project panel. b. Inferred migration using SMCSMC in the Simons Genome Diversity Panel along with the scaled relative cross-coalescent rate estimated by MSMC2. 10,000 particles were used to approximate the ancestral recombination graph in the SMCSMC particle filter and 25 iterations were used to update demographic parmaeters. 20 iterations were used for MSMC2.

the destination (e.g. African) population traced backwards in time remains in that population across a given epoch according to the migration model (see Methods). For the simulations, we chose the most recent 100kya as epoch, and used scenarios with IMFs ranging from 0 to 0.593. For each simulation we report the inferred IMF in both the forward and backward direction (Figure 3.8). We find that SMCSMC has good power to detect backward migration pulses up to 60kya (median ratio of inferred and true IMF, 0.91), while power drops off at 70kya (IMF ratio 0.46). In the pure backward migration case, some forward migration is falsely inferred, but this is always substantially less than the inferred backward migration (median ratio inferred forward to true backward IMF, 0.37; true migration peak $\leq$ 60kya). However, in the case of true forward migration as well as bidirectional migration, roughly equal mixtures of forward and backward migration are inferred (Figure 3.8). We conclude that in the epoch 40–70kya the forward and bidirectional scenarios are difficult to distinguish from each other, but both can be distinguished from backward migration, the only scenario resulting in substantially different inferred backward and forward migration.

To validate the existence of the migration pulse, though not its direction, we next analyzed the same data using MSMC, which is widely used to estimate gene flow in the ancient past by estimating the relative cross-coalescent rate (RCCR) between two populations [8, 57–59]. We use the updated implementation MSMC2 recommended by the authors and first published in [38]. Each of the SMCSMC analyses are repeated using MSMC2 to estimate effective population size and RCCR (Figure 3.10, Figure 3.11, Figure 3.9). Consistent with previous analyses conducted with MSMC2, our estimates show high RCCR in the Late Middle Pleistocene in both the SGDP and the HGDP (Figure 3.4c,d) [12, 57]. These observations confirm the existence of a substantial pulse of ancient gene flow between Eurasians (Han Chinese) and Africans.

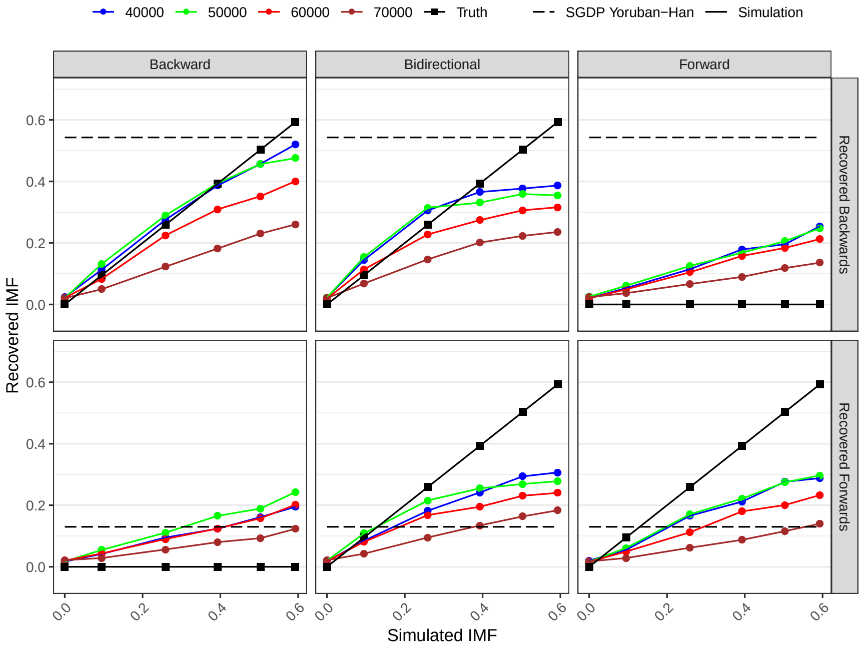


**Figure 3.8: Simulation study**. SCRM was used to simulate 1 gigabase sequence data for two diploid individuals under three different migration models. Migration was simulated backwards (from a Eurasian-like population to an African-like population), forward (the reverse), and symmetrically (equal migration in both directions). The amount of migration indicates the proportion of the sink population replaced by the source over a 10ky period centered at 40, 50, 60, or 70kya. The total IMF inferred by SMCSMC over the last 100ky is plotted and compared to the true simulated amount. For reference, the inferred IMF in either direction across 0-100kya for a Yoruban and Han individual is given in dashed lines. 5 iterations of variational Bayes and 5000 particles were used for inference. The effective population size model and additional details are given in Supplemental Section **??**.

### 3.3.5 Migration Pre-dates East-West Eurasian Divergence

To assess whether the inferred back-migration shows variation across the descendants of the OoA event, we repeated the analyses using three representative non-African groups in the SGDP: Han Chinese, French European, and Papuans. Since simulations show that SMCSMC has little power to detect migration predating 70kya, and to exclude Holocene migration, the epoch we use to calculate real-data IMFs comprise the period of peak inferred migration up to the period of diminishing power (30–70kya); we use this epoch for all subsequent analyses. Inferred IMFs are not significantly different between Han Chinese and European populations in non-Afroasiatic populations (p=0.14, two-tailed paired t-test; Figure 3.4h and Figure 3.12, Table 3.3), consistent with migration occurring before the European-

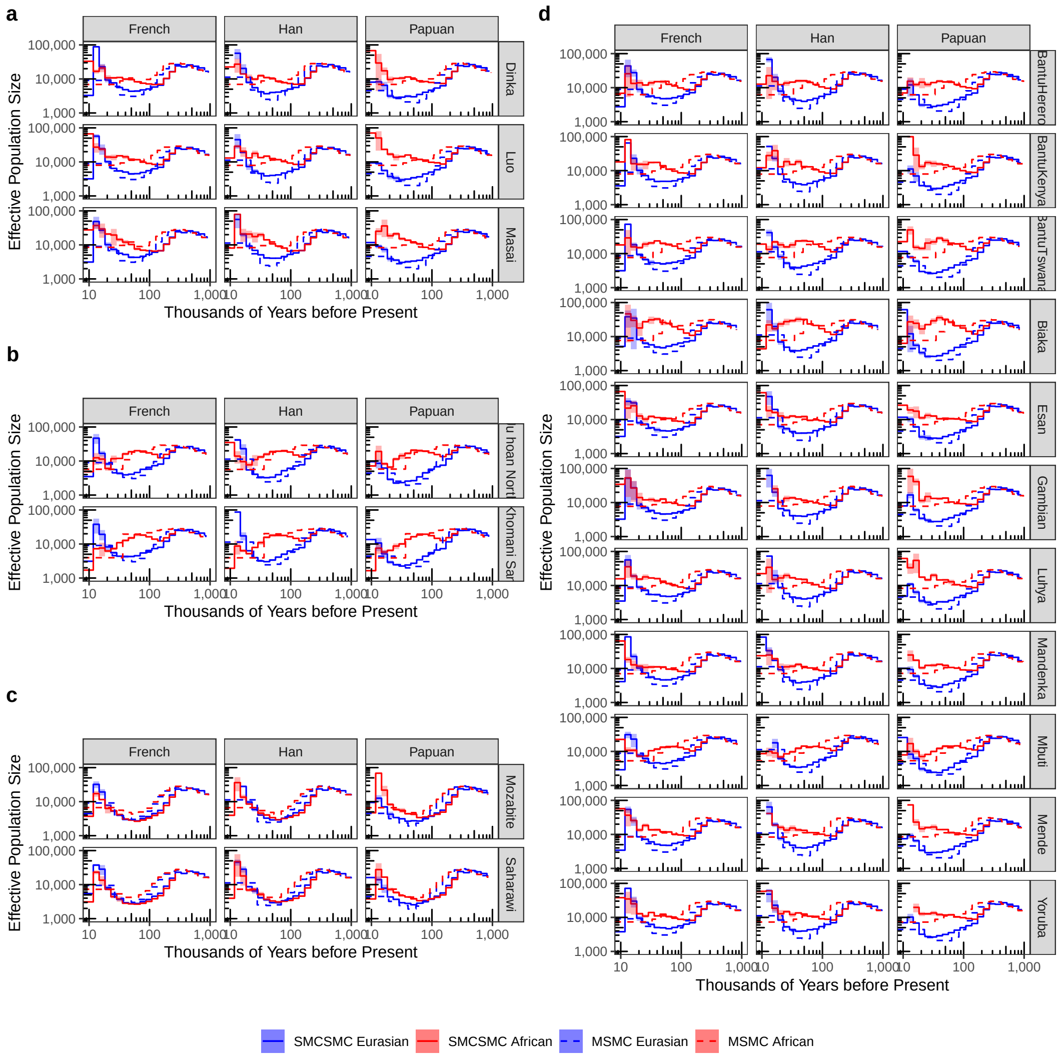


**Figure 3.9: Inference of historical effective population size in the Simons Genome Diversity Project.** The SMCSMC particle filter was used to infer directional migration rates and effective population size in both directions from one of three Eurasian populations (French, Han, and Papuan) to one of 18 African populations. 5000 particles were used to approximate the ancestral recombination graph with 10 iterations of variational Bayesian inference to update demographic parameter values. Panels represent a. Nilo-Saharan, b. KhoeSan, c. Afroasiatic, and d. Niger-Kordofanian language families. Alongside the SMCSMC inference, we use MSMC2 to infer the same values with default settings and 20 iterations for convergence.

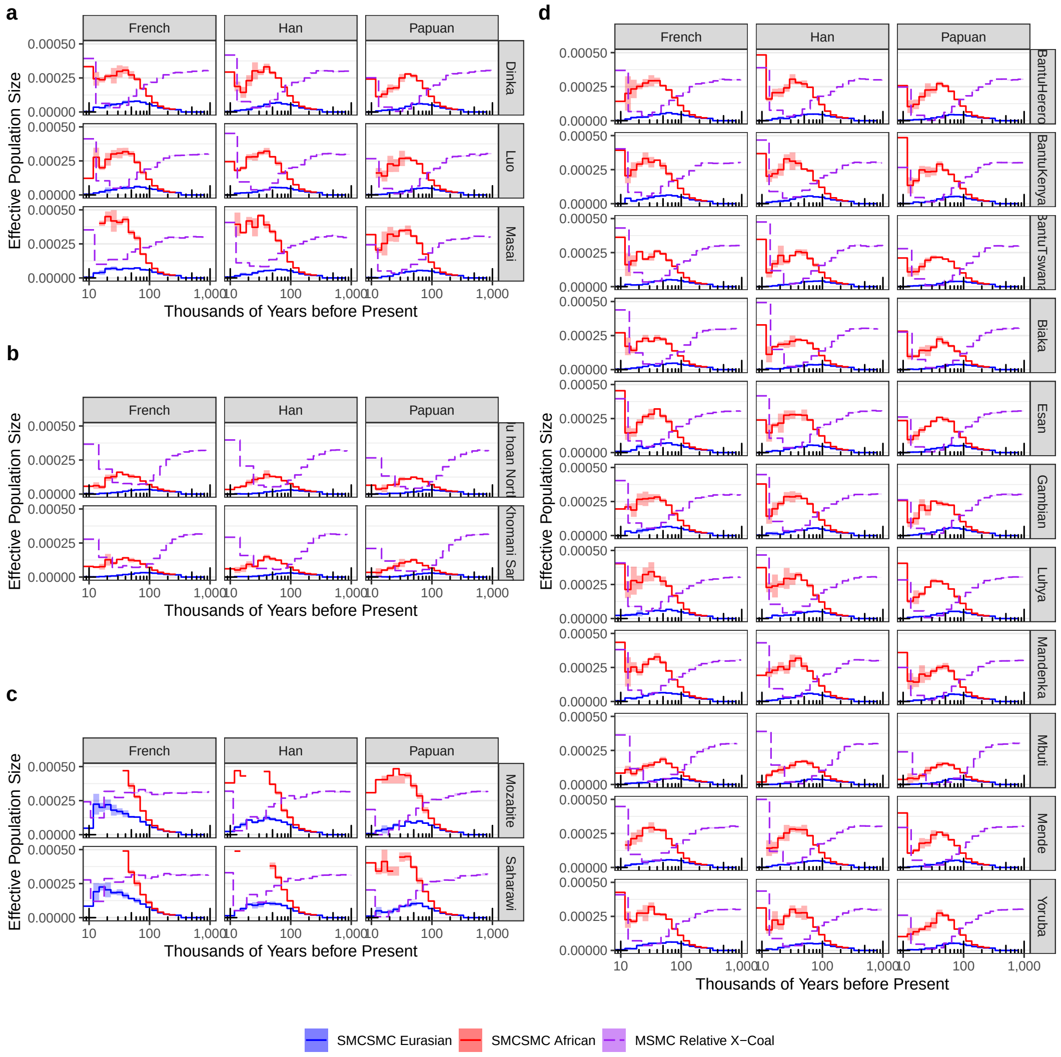
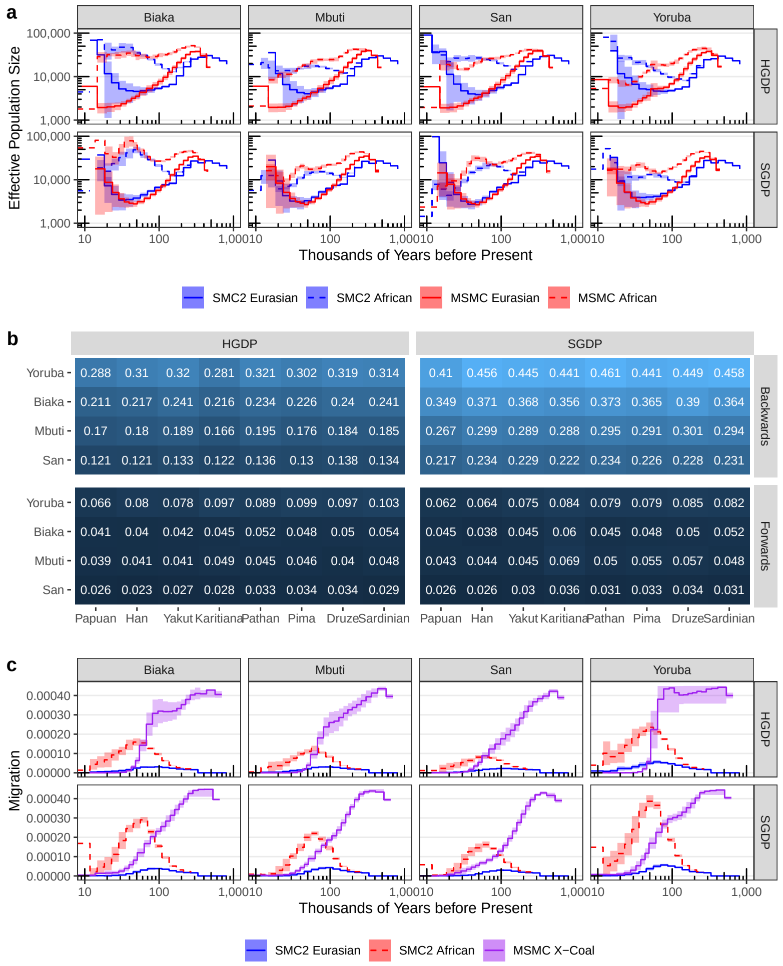
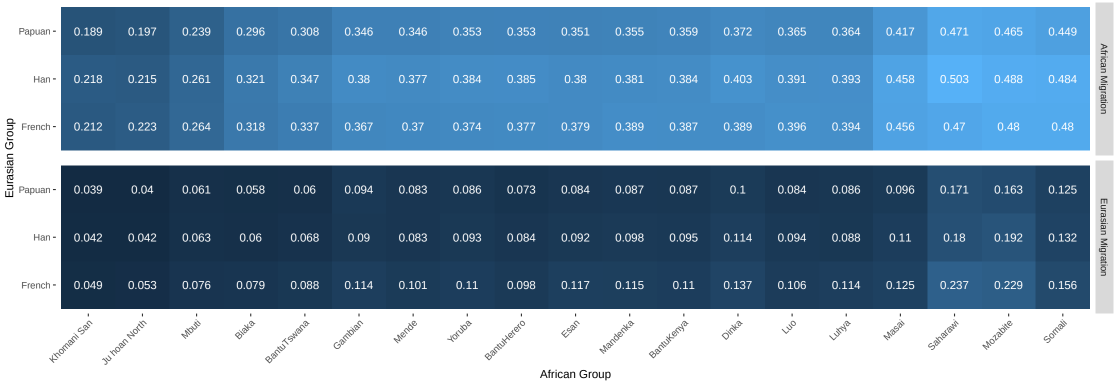


**Figure 3.10: Inference of directional migration in the Simons Genome Diversity Project.** The SMCSMC particle filter was used to infer directional migration rates in both directions from one of three Eurasian populations (French, Han, and Papuan) to one of 18 African populations. 5000 particles were used to approximate the ancestral recombination graph with 10 iterations of variational Bayesian inference to update demographic parameter values. Panels represent a. Nilo-Saharan, b. KhoeSan, c. Afroasiatic, and d. Niger-Kordofanian language families. Alongside the SMCSMC inference, we use MSMC2 to infer the relative cross coalescent rate (RCCR) with default settings and 20 iterations for convergence.



| | HGDP | | | | | | | | SGDP | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Papuan | Han | Yakut | Karitiana | Pathan | Pima | Druze | Sardinian | Papuan | Han | Yakut | Karitiana | Pathan | Pima | Druze | Sardinian | |
| Yoruba | 0.288 | 0.31 | 0.32 | 0.281 | 0.321 | 0.302 | 0.319 | 0.314 | 0.41 | 0.456 | 0.445 | 0.441 | 0.461 | 0.441 | 0.449 | 0.458 | Backwards |
| Biaka | 0.211 | 0.217 | 0.241 | 0.216 | 0.234 | 0.226 | 0.24 | 0.241 | 0.349 | 0.371 | 0.368 | 0.356 | 0.373 | 0.365 | 0.39 | 0.364 | |
| Mbuti | 0.17 | 0.18 | 0.189 | 0.166 | 0.195 | 0.176 | 0.184 | 0.185 | 0.267 | 0.299 | 0.289 | 0.288 | 0.295 | 0.291 | 0.301 | 0.294 | |
| San | 0.121 | 0.121 | 0.133 | 0.122 | 0.136 | 0.13 | 0.138 | 0.134 | 0.217 | 0.234 | 0.229 | 0.222 | 0.234 | 0.226 | 0.228 | 0.231 | |
| Yoruba | 0.066 | 0.08 | 0.078 | 0.097 | 0.089 | 0.099 | 0.097 | 0.103 | 0.062 | 0.064 | 0.075 | 0.084 | 0.079 | 0.079 | 0.085 | 0.082 | Forwards |
| Biaka | 0.041 | 0.04 | 0.042 | 0.045 | 0.052 | 0.048 | 0.05 | 0.054 | 0.045 | 0.038 | 0.045 | 0.06 | 0.045 | 0.048 | 0.05 | 0.052 | |
| Mbuti | 0.039 | 0.041 | 0.041 | 0.049 | 0.045 | 0.046 | 0.04 | 0.048 | 0.043 | 0.044 | 0.045 | 0.069 | 0.05 | 0.055 | 0.057 | 0.048 | |
| San | 0.026 | 0.023 | 0.027 | 0.028 | 0.033 | 0.034 | 0.034 | 0.029 | 0.026 | 0.026 | 0.03 | 0.036 | 0.031 | 0.033 | 0.034 | 0.031 | |

**Figure 3.11: Inference of directional migration is comparable between data sets and phasing strategies.** We used SMCSMC to simultaneously infer directional migration rates and effective population size in the 36 genome physically phased subset of the Human Genome Diversity Panel. We match these 36 genomes with comparable individuals in the Simons Genome Diversity Panel (with the exception of one Papuan population, which has no comparable population in the SGDP) and perform an identical analysis. **a.** Average $N_e$ estimate across four populations in the physically phased subset of the HGDP and the subset of SGDP used to compare with HGDP inference. Inference of population size is averaged over eight Eurasian populations, with the bars representing standard deviation. For MSMC2, the time indexes were averaged to have consistent start and stop times for the steps. **b.** Inferred integrated migration fraction (IMF) from Africa to Eurasians (forwards) and from Eursaians to Africans (backwards) between 40 and 70 kya (see Methods). **c.** Directional migration inference in African populations averaged over Eurasian partners in the two data sets. Shaded regions denote standard deviations. For MSMC, the time indexes were averaged to have consistent start and stop times for plotting.



| | Khomani San | Ju hoan North | Mbuti | Biaka | BantuTswana | Gambian | Mende | Yoruba | BantuHerero | Esan | Mandenka | BantuKenya | Dinka | Luo | Luhya | Masai | Saharawi | Mozabite | Somali |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| African Migration | | | | | | | | | | | | | | | | | | | |
| Papuan | 0.189 | 0.197 | 0.239 | 0.296 | 0.308 | 0.346 | 0.346 | 0.353 | 0.353 | 0.351 | 0.355 | 0.359 | 0.372 | 0.365 | 0.364 | 0.417 | 0.471 | 0.465 | 0.449 |
| Han | 0.218 | 0.215 | 0.261 | 0.321 | 0.347 | 0.38 | 0.377 | 0.384 | 0.385 | 0.38 | 0.381 | 0.384 | 0.403 | 0.391 | 0.393 | 0.458 | 0.503 | 0.488 | 0.484 |
| French | 0.212 | 0.223 | 0.264 | 0.318 | 0.337 | 0.367 | 0.37 | 0.374 | 0.377 | 0.379 | 0.389 | 0.387 | 0.389 | 0.396 | 0.394 | 0.456 | 0.47 | 0.48 | 0.48 |
| Eurasian Migration | | | | | | | | | | | | | | | | | | | |
| Papuan | 0.039 | 0.04 | 0.061 | 0.058 | 0.06 | 0.094 | 0.083 | 0.086 | 0.073 | 0.084 | 0.087 | 0.087 | 0.1 | 0.084 | 0.086 | 0.096 | 0.171 | 0.163 | 0.125 |
| Han | 0.042 | 0.042 | 0.063 | 0.06 | 0.068 | 0.09 | 0.083 | 0.093 | 0.084 | 0.092 | 0.098 | 0.095 | 0.114 | 0.094 | 0.088 | 0.11 | 0.18 | 0.192 | 0.132 |
| French | 0.049 | 0.053 | 0.076 | 0.079 | 0.088 | 0.114 | 0.101 | 0.11 | 0.098 | 0.117 | 0.115 | 0.11 | 0.137 | 0.106 | 0.114 | 0.125 | 0.237 | 0.229 | 0.156 |

**Figure 3.12: Integrated migration fraction 40–70kya in `SMCSMC` analysed SGDP populations.** Directional migration was integrated by finding the cumulative probability of an individual migrating during the specified epoch. Directional migration backwards from Eurasia to Africa and forwards from Africa to Eurasia (both forward in time) are reported separately. Displayed is the average values from three technical replicates.

East Asian split approximately 40kya [60]. The contribution of this admixture event to extant African genetic variation is substantial; the estimated IMFs indicate that for individuals in the major African language groups, approximately a third of ancestral lineages trace their ancestry through the proto-Eurasian population (Niger-Kordofian group, $0.35 \pm 0.04$; Nilo-Saharan groups, $0.41 \pm 0.03$; Table 3.3). When we estimate these proportions using a Papuan sample to represent non-African descendants we find slightly but significantly smaller values compared to estimates using either the Han Chinese or European populations (mean difference of $0.029 \pm 0.002$, $p{=}9.2{\times}10^{-15}$, and $0.025 \pm 0.004$, $p{=}2.3 \times 10^{-10}$, paired t-tests, Table 3.3, Table 3.4). Similarly, in the HGDP, inferred migration in both Papuan groups (Sepik and Highlands) was $0.025 \pm 0.004$ ($p{=}1.4 \times 10^{-6}$) lower than French and Han (Table 3.5). We comment on this observation in the Discussion.

### 3.3.6 Directional Migration Explains Excess Inferred African Genetic Diversity 100kya

Previous studies looking at Effective Population Size ($N_e$) in human ancestral populations have consistently reported inflated inferences in African populations approximately 100kya, often hypothesized to be due to unaccounted-for population substructure within Africa [5, 8]. We use SMCSMC to analyze African individuals paired with an individual from one of three non-African populations (Han Chinese, French European, and Papuans) and infer $N_e$ for the African ancestral population under a two-island model with directional migration. Each analysis was repeated three times to assess the contribution of stochastic sampling to the inferences (Figure 3.14, Figure 3.9, per population $N_e$ in Supplemental Fig. Figure 3.13). SMCSMC infers substantially lower African $N_e$ than MSMC in the period 80kya–300kya. In addition, while MSMC inferences show convergence of African and Eurasian ancestral $N_e$ estimates only around 300kya, inferences from SMCSMC indicate convergence at 150kya (Fig. Figure 3.14a), closer to the hypothesized time of the diversification of the ancestral lineages prior to the main out-of-Africa migration episode [23, 38]. The same analysis on physically phased samples from HGDP show that these results are not driven by errors due to statistical phasing (Figure 3.11 and subsection 3.3.2). When we used SMCSMC to infer both African and European $N_e$ under a single-population model without migration, $N_e$ estimates were comparable to those from MSMC (Figure 3.14b), indicating that the SMCSMC inferences are not driven by methodological biases particular to SMCSMC.

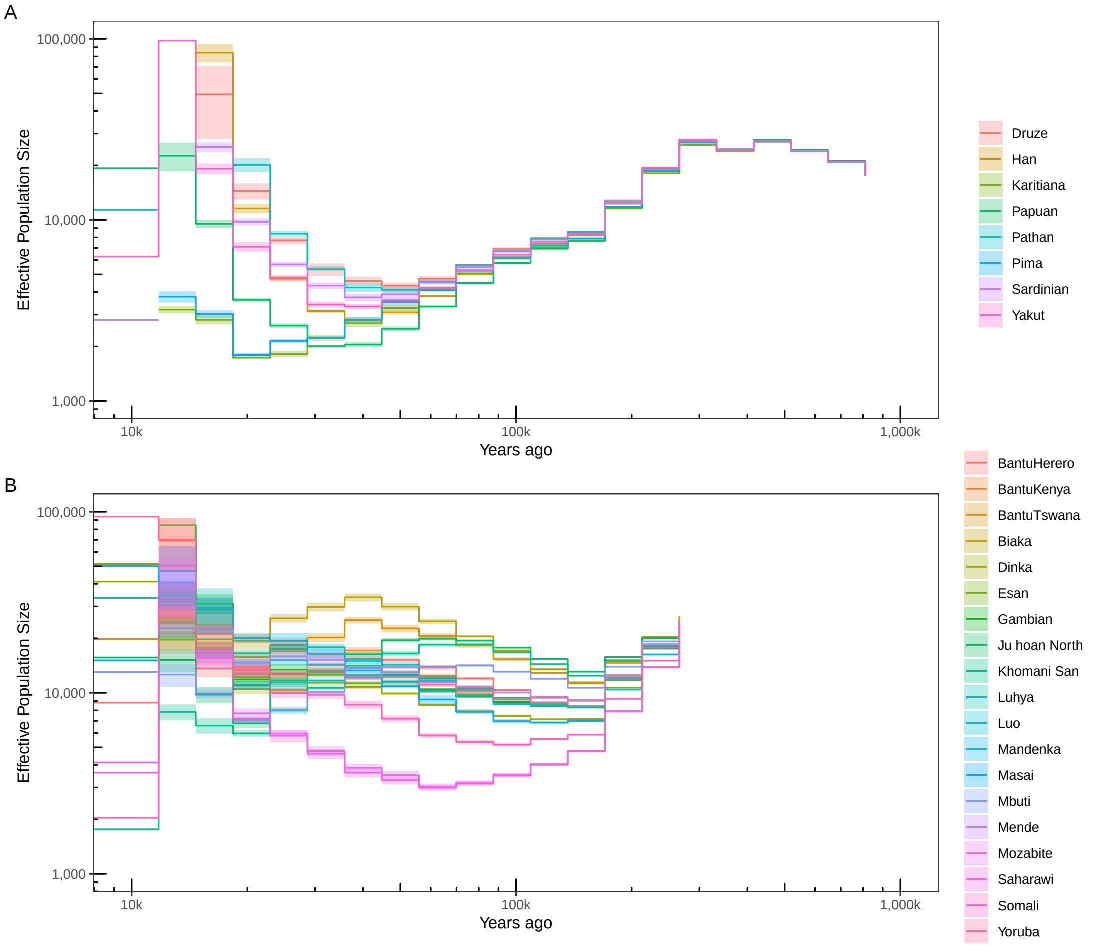


**Figure 3.13: Estimates of individual population sizes incorporating directional migration.** Using `SMCSMC` the effective population size of global populations in the Simons Genome Diversity Panel is inferred while simultaneously fitting directional migration estimates. Averages are plotted by epoch, with shaded regions denoting the standard deviation. **a.** Estimates of Eurasian population sizes when averaged over Eurasian donor populations. This analysis uses the eight Eurasian populations matched to HGDP populations averaged over the four matched African populations. B. Estimates of African population sizes when averaged over Eurasian recipient populations. This analysis uses the three donor Eurasian populations used for the majority of the analyses in the main text (French, Han, and Papuan) along with the given African populations. Before approximately 250kya, the populations share the same population size within the model, and are not plotted. 10,000 particles are used to approximate the ancestral recombination graph in the SMCSMC particle filter and 15 iterations are used to update demographic parameter values.

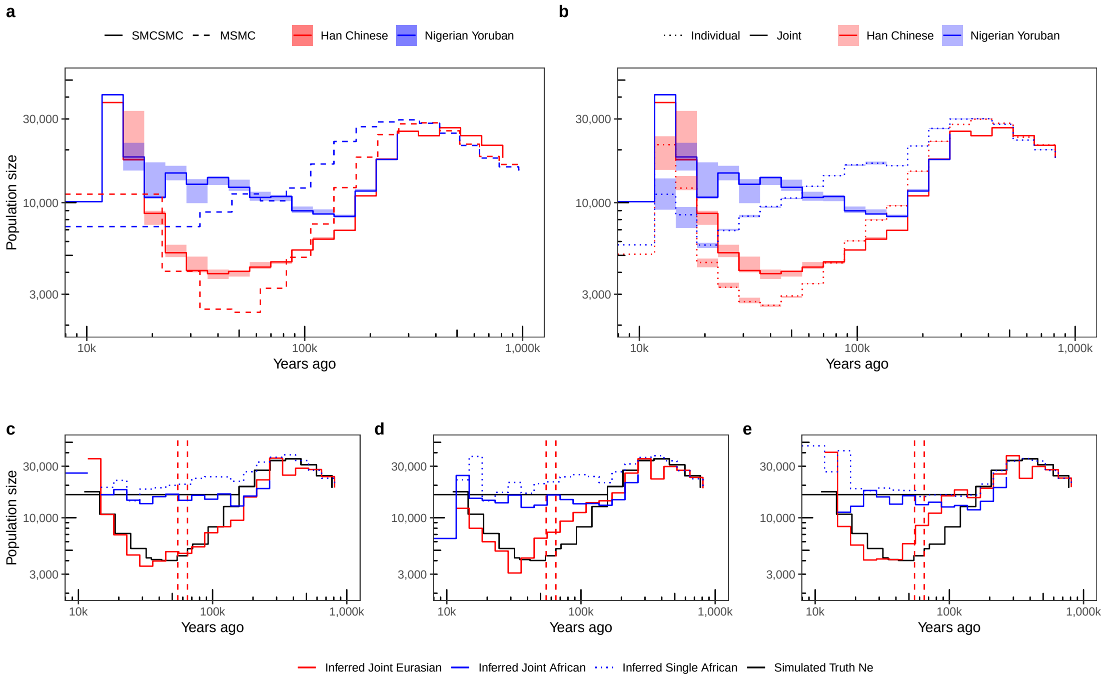


**Figure 3.14: Effective population size inference. a.** Analyzing a Nigerian Yoruban and a Han Chinese individual from the Simons Genome Diversity Panel jointly in a two-island model with directional migration using `SMCSMC` yields markedly lower $N_e$ estimates and a more recent apparent split time, than when the same data are analyzed using `MSMC` with a model that does not explicitly include migration. Analyses for `SMCSMC` repeated three times; range of the estimates shaded. **b.** When each individual is analysed separately, using a model not including migration, $N_e$ estimates from `SMCSMC` are similar to those of `MSMC2`. (Joint estimate from **a** included for comparison.) **c, d,** and **e.** Inferred Eurasian and African $N_e$ from data simulated under a two-island model with, from left to right, a backward Eurasia-to-Africa, a bidirectional, and a forward migration pulse lasting 10ky (dashed vertical lines; same data as for Fig. **??e-g**). Particularly for the backward migration case, inferred $N_e$ under a two-island model tracks the true values (black) well, while inferred $N_e$ under a single-population model are inflated around the split time. All `SMCSMC` analyses used 10000 particles and 25 variational Bayesian iterations; `MSMC` analyses used 40 iterations (Supplemental Section **??**).

To more directly support the interpretation that the lower African $N_e$ inferred by SMCSMC is due to appropriate modeling of directional migration, we again used coalescent simulation with `SCRM` to investigate various migration scenarios and their effects on inferred African $N_e$. Using the simulation framework as above, we examine $N_e$ estimates inferred under a two-island model with migration, and in addition $N_e$ separately inferred for each of the two simulated populations under a single-population model (Supplemental Section 3.2.7). Focusing on single-population inferences, we found that for simulated African populations that had received substantial migration from the simulated Eurasian population either through backward or bidirectional migration, inferred $N_e$ values indeed were substantially inflated compared to true values (Fig. Figure 3.14c,d), while this effect was not seen when forward (African-to-Eurasian) migration was simulated (Fig. Figure 3.14e). Similarly, single-population Eurasian $N_e$ estimates were inflated in the presence of forward and bidirectional migration, but not backward migration (Supplemental Figs. 3.15–3.17). In contrast, when using a model that includes migration, inferred African $N_e$ do not show inflation in any of the three scenarios (Fig. Figure 3.14c-e). We conclude that the inferences from SMCSMC and MSMC are compatible with substantial back-migration from ancestral Eurasians into Africans, but not substantial bidirectional or forward migration.

### 3.3.7 Less Gene Flow to Central and South African Hunter-Gatherers

We infer substantial Eurasian back-migration into all African groups, however the inferred IMFs for individuals from Khoe-San populations are significantly lower than for any other group (difference with Niger-Kordofians, $0.14 \pm 0.02$, $p = 4.4 \times 10^{-14}$; difference with Nilo-Saharans, $0.20 \pm 0.03$, $p = 6.9 \times 10^{-9}$, two-tailed t-test, Table 3.2). To further support this observation we used MSMC to estimate the relative cross-coalescent rate (RCCR) for several populations, and find evidence for gene flow between Yorubans and Eurasians that is not shared with the Khoe-San individuals in either the SGPD and the HGDP (Figure 3.4c,d). These results are consistent across Eurasian donor populations (Figure 3.7). The Khoe-San individuals are particular outliers, whose ancestors are inferred to have experienced approximately half the amount of admixture seen in Nilo-Saharan and Niger-Kordofanian groups (**??**). In addition, we find that the Mbuti and Biaka, both Central African hunter-gatherer populations, show levels of Eurasian gene flow that are intermediate between levels observed in the Khoe-San and Yorubans (3.4a,b, Table 3.3). This is mirrored by inferred IMFs for Central African Hunter Gatherers, which are significantly lower

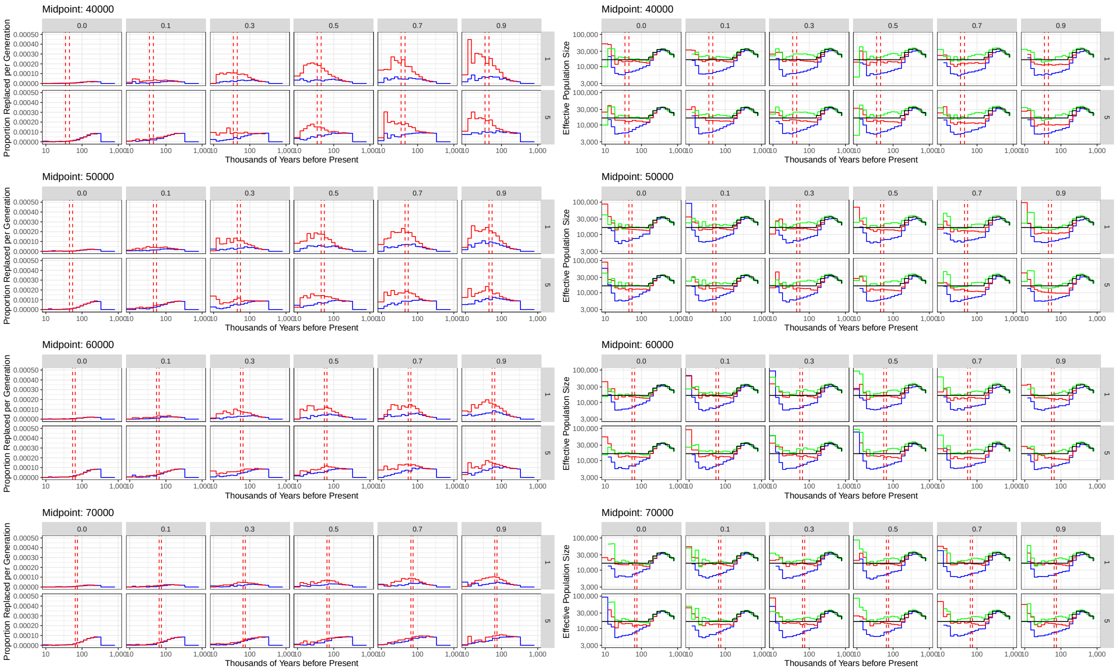


**Figure 3.15: Simulated migration backwards from effectively Eurasian to effectively African populations.** One gigabase of simulated sequence data was generated with SCRM for two diploid individuals from different populations using a 100kb sliding window approximation of the coalescent with recombination and a demographic model similar to Eurasians and Africans specified in Supplemental section 3.2.7. The timing, representing the midpoint of a 10ky migration episode, and the integrated migration fraction (IMF) were systematically varied. In this scenario, only backwards migration was simulated. The left panel displays the recovered directional migration, with red lines representing migration backwards from Eurasia and blue representing migration forwards to Eurasia. The timing of the migration episode is demarcated with dotted red lines. The right panel displays the inferred effective population size of both populations. Additionally, the effective population size of the African-like population was modelled separately, without simultaneously inferring migration. 5000 particles were used in the SMCSMC particle filter to approximate the ancestral recombination graph and 5 iterations of variational Bayesian inference were used to updated demographic parameter values.

than other Niger-Kordofanian groups (difference $-0.08 \pm 0.03$, $p = 1.2 \times 10^{-3}$, Table Table 3.2), possibly reflecting the proposed early split times of the Mbuti and Biaka from the remainder of ancestral African populations between 60 and 200kya [30, 32].

### 3.3.8 No Evidence for Excess Neanderthal Ancestry

Previous studies have proposed that a backflow from Eurasia may have brought Neanderthal ancestry into African populations [61]. To assess whether the proposed Late Middle Paleolithic back migration might have introduced Neanderthal material,

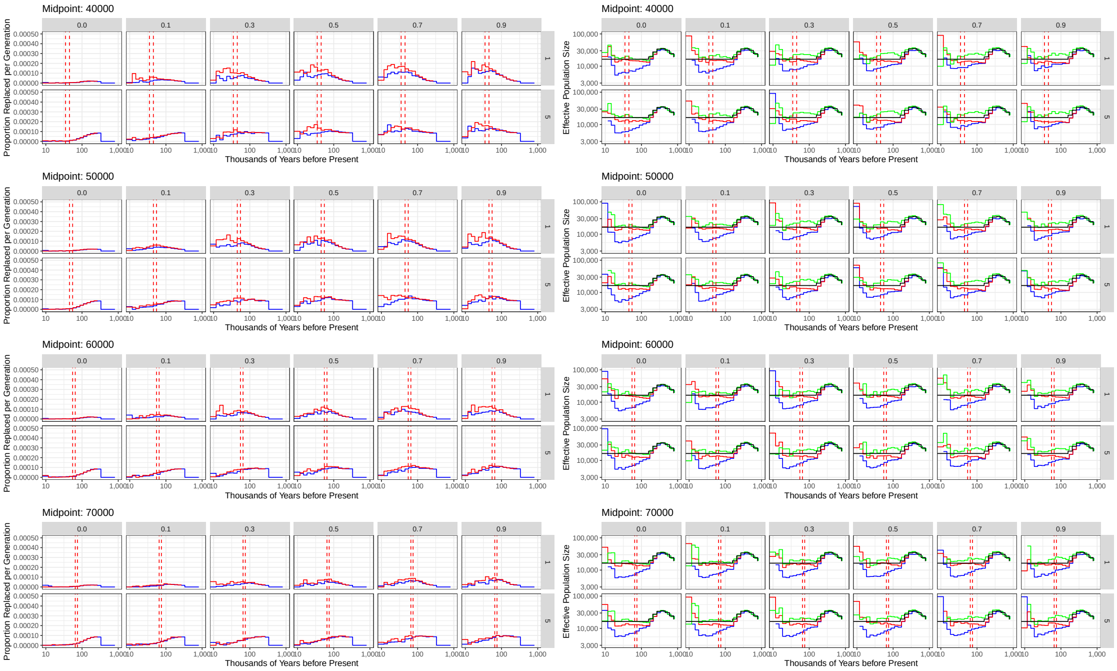


**Figure 3.16: Simulated bidirectional migration between effectively Eurasian and effectively African populations.** One gigabase of simulated sequence data was generated with SCRM for two diploid individuals from different populations using a 100kb sliding window approximation of the coalescent with recombination and a demographic model similar to Eurasians and Africans specified in Supplemental section **??**. The timing, representing the midpoint of a 10ky migration episode, and the integrated migration fraction (IMF) were systematically varied. In this scenario, migration between these two populations was simulated with equal rates. The left panel displays the recovered directional migration, with red lines representing migration backwards from Eurasia and blue representing migration forwards to Eurasia. The timing of the migration episode is demarcated with dotted red lines. The right panel displays the inferred effective population size of both populations. Additionally, the effective population size of the African-like population was modelled separately, without simultaneously inferring migration. 5000 particles were used in the SMCSMC particle filter to approximate the ancestral recombination graph and 5 iterations of variational Bayesian inference were used to updated demographic parameter values.

we analyzed a Yoruban and a French individual using SMCSMC to draw a sample from the posterior distribution of ARGs, isolated the marginal trees containing an inferred back-migration event in the epoch 30–70 kiloanni (thousands of years) before present (ka BP), and reported the inferred admixture tracts.

To assess whether the identified segments are plausible, we confirmed that their length distribution is consistent with IMF and timing of the migration inferred by SMCSMC. We take the isolated segments (Figure 3.18a) and compute the mean

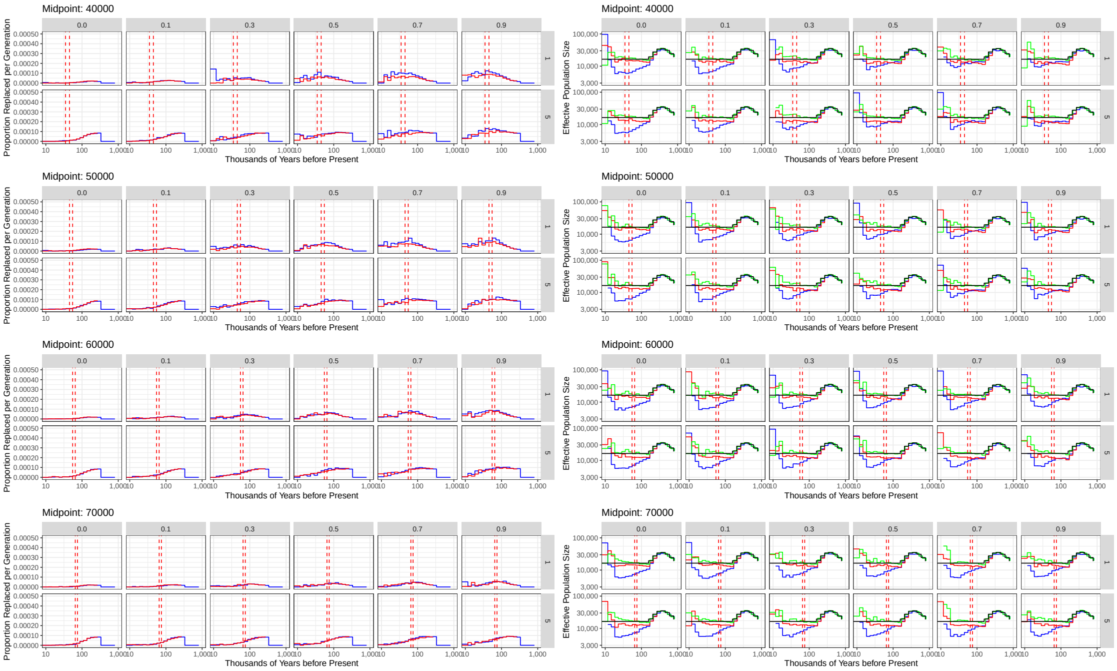


**Figure 3.17: Simulated migration forwards from effectively African to effectively Eurasian populations.** One gigabase of simulated sequence data was generated with SCRM for two diploid individuals from different populations using a 100kb sliding window approximation of the coalescent with recombination and a demographic model similar to Eurasians and Africans specified in Supplemental section **??**. The timing, representing the midpoint of a 10ky migration episode, and the integrated migration fraction (IMF) were systematically varied. In this scenario, only forwards migration was simulated. The left panel displays the recovered directional migration, with red lines representing migration backwards from Eurasia and blue representing migration forwards to Eurasia. The timing of the migration episode is demarcated with dotted red lines. The right panel displays the inferred effective population size of both populations. Additionally, the effective population size of the African-like population was modelled separately, without simultaneously inferring migration. 5000 particles were used in the SMCSMC particle filter to approximate the ancestral recombination graph and 5 iterations of variational Bayesian inference were used to updated demographic parameter values.

track length (Table 3.6). We use the approximation that the mean segment length should be approximately equal to $((1-m)r(T-1))^{-1}$ to determine that, if the migration happened in one pulse, our empirical distribution would suggest either a recent timing or a very large pulse (Figure 3.18b). However, we heavily caveat any interpretation of these data with the fact that they are explicitly generated under a model of a given migration proportion. The fact, therefore, that they are of a consistent length with a large migration is more evidence for the model producing internally-consistent tracts than any external validation of the results in this article.

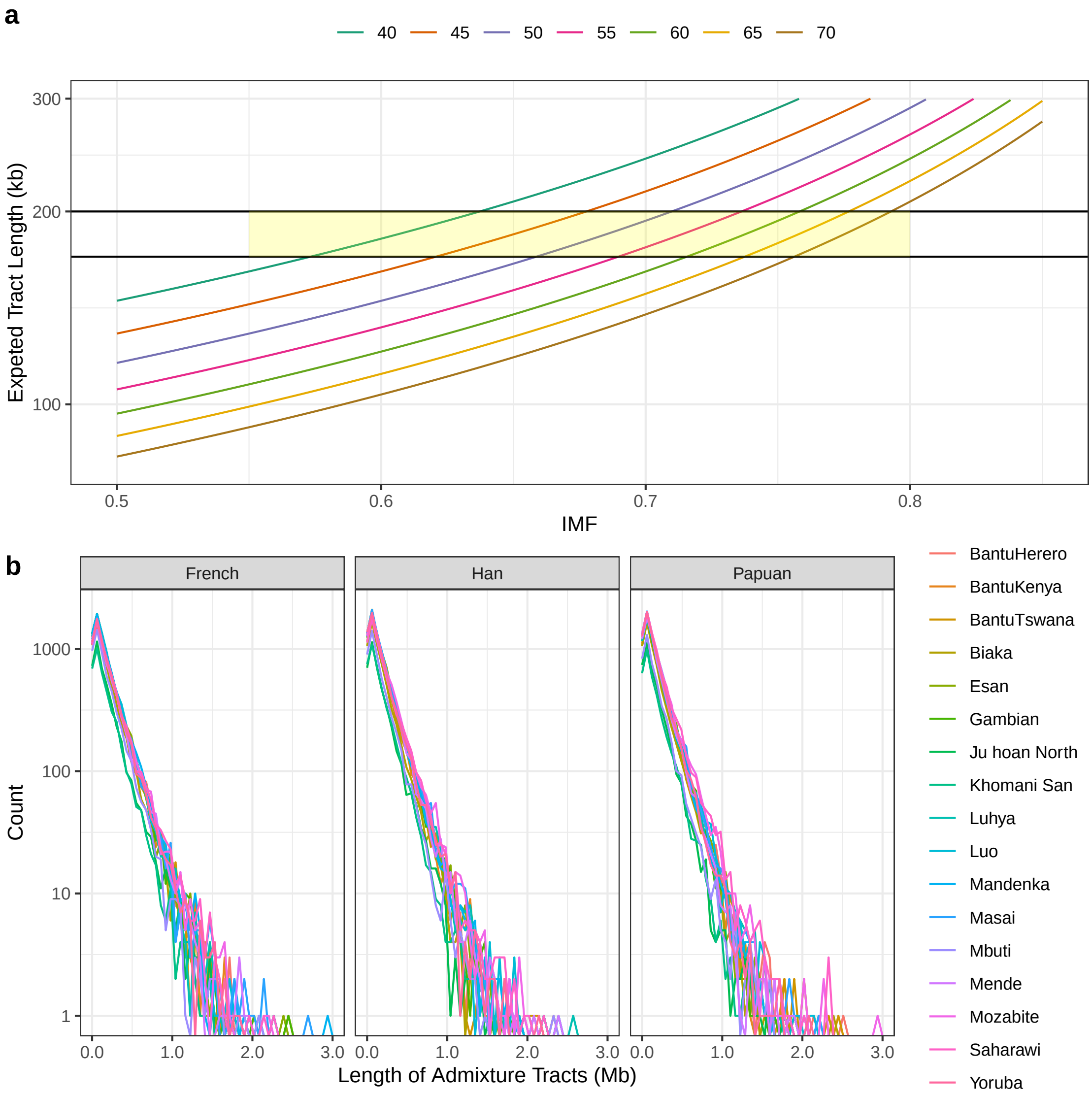


**Figure 3.18: Analysis of the length of putatively migrated segments. a.** Theoretical length distribution of admixture tract rate parameter under varying migration and admixture timing assuming that $L = ((1-m)r(T-1))^{-}1$, given $L$ length, $m$ symmetrical migration rate, $r$ recombination rate in events per nucleotide, and $T$ time in generations [51]. Shaded region denotes empirical range (with San at the bottom end, and Yoruban at the top end, Supplemental Table 3.6) of fragments observed in the Simons Genome Diversity Panel. **b.** Following the reconstruction of the ancestral recombination graph using different African and non-African individuals using the SMCSMC particle filter, we use a sample of the posterior distribution of marginal trees to reconstruct putatively migrated segments (see Methods). We plot the length distribution of admixture tracts between individuals in the SGDP using 50 bins. Length is given in megabases (Mb). Isolated from SMCSMC estimated ancestral recombination graphs. Migration rate is given in terms of proportion of the sink population replaced per generation.

The assumption that migration has occurred in a single wave is largely unrealistic. We used expectation maximisation to investigate if a mixture of exponential distributions explained the observed tract lengths better than a single distribution. We found that in some cases, two or three exponential distributions were better supposed by the data, however the differences in log likelihoods was negligible and the support for the different distributions was approximately inversely proportional to their number (data not shown). We found no strong support for multiple waves of migration from this analysis.

As expected, we found that these African segments with putative Eurasian ancestry tend to be more closely related to a Eurasian sample than another representative of the same African population (Table 3.7, Supplemental Fig. 3.19, 3.20) in a global dataset of modern and ancient individuals compiled by the Reich group (see URLs). Within these African segments that are likely enriched for material with Eurasian ancestry, we then used $D$ statistics [17] to identify enrichment for Neanderthal material compared to an African background.

We first use $f_3$ statistics to look for evidence of admixture between the African and various Eurasian groups. We calculate $f_3$(Yoruba-1, Eurasian group, Yoruba-2) for Papuans, French, Han Chinese, and the Vindija Neanderthal. We calculate this statistic in all available markers, and additionally for the segments isolated from the three Eurasians separately (Figure 3.19). These statistics show, firstly, that ascertaining in a particular group influences the shared drift with that group. This is exemplified by the non-significant shared drift with Papuans in French and Han ascertained segments. Secondly, these statistic show significant levels ($|Z| > 2$) of drift between the test individual and Eurasian populations, while also showing no increase in Neanderthal allele sharing ($f_3 = 0$). To find statistical evidence of admixture, we compute $f_3$(Yoruba-2, Eurasian, Yoruba-1) for the same Eurasian groups. We find statistical evidence for admixture in each of the groups examined, for all ascertainment schemes (Figure 3.20).

We use $D$ statistics to examine more nuanced scenarios. We find that the two Yorubans share more alleles than other groups in Africa (D(African group, Yoruba-

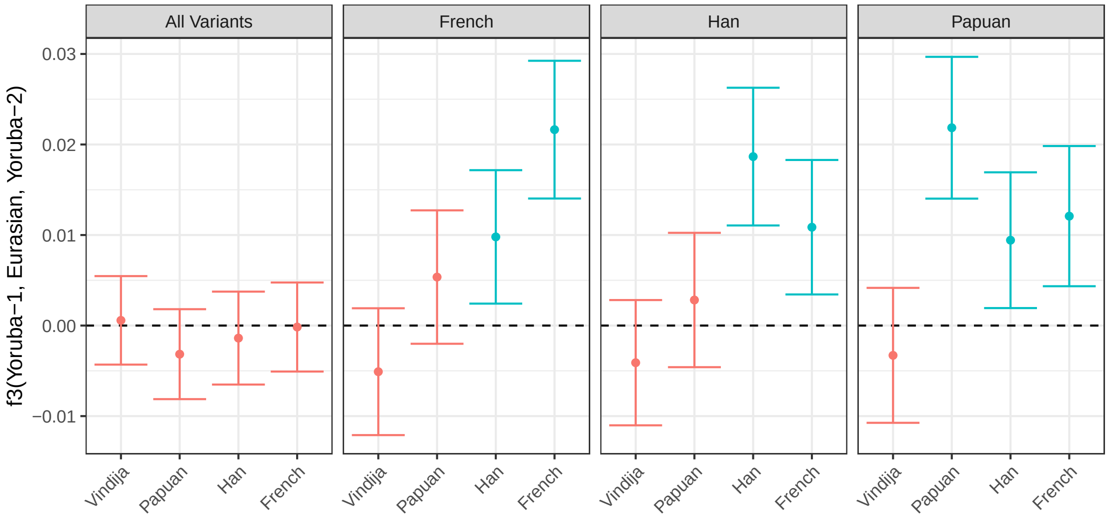


**Figure 3.19: $f_3$ statistics show evidence for shared drift with Eurasians.** Following the reconstruction of putatively migrated segments from the inferred ancestral recombination graph, we use the Reich Human Origins data set to investigate admixture using drift statistics using `ADMIXTOOLS` and `admixr` (see URLs). Tests are separated into all markers, and those ascertained through `SMCSMC` runs with the French, Han, and Papuans as a comparison Eurasian group. The $f_3$ statistics estimate is plotted, along with 95% C.I. computed via a block jackknife. Statistics which are significantly larger than zero are coloured blue, indicating shared drift.

1; Yoruba-2, Chimp) is significantly negative with $|Z| > 3$), but the individual of interest is closer to Out of Africa (OoA) groups such as the Han, French, and Papauns (D(OoA, Yoruba-2; Yoruba-1, Chimp) is significantly negative with $|Z| < 3$) than to its partner Yoruban (Table 3.7). This implies that SMCSMC has identified segments of the African Genome which are more closely related to OoA populations than to fellow Africans.

In summary, we find no evidence for gene flow with a Vindija Neanderthal on the Mbuti baseline, or when compared to a different Yoruban (Table 3.8, Table 3.9). We additionally find no evidence for increased affinity to the Vindija Neanderthal when compared to the Altai, as would be expected if the material were descended from admixing Eurasians (Table 3.10). However, we find that restricted to the identified segments, $D$ statistics have power to detect evidence for the known admixture from Vindija into a French individual (Figure 3.21), suggesting that lack of power does not explain the lack of evidence we find for Neanderthal admixture into Africans. In

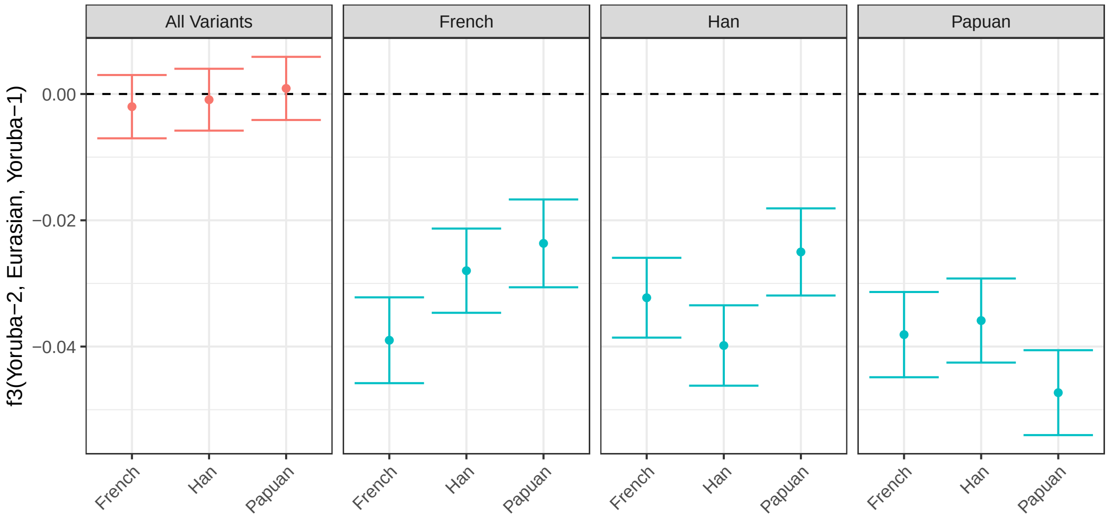


**Figure 3.20: $f_3$ statistics show evidence for Eurasian admixture**. Following the reconstruction of putatively migrated segments from the inferred ancestral recombination graph, we use the Reich Human Origins data set to investigate admixture using drift statistics using `ADMIXTOOLS` and `admixr` (see URLs). Tests are separated into all markers, and those ascertained through `SMCSMC` runs with the French, Han, and Papuans as a comparison Eurasian group. The $f_3$ statistics estimate is plotted, along with 95% C.I. computed via a block jackknife. Statistics which are significantly less than zero indicate statistical evidence for admixture, and are coloured blue.

addition, we find no differences in affinity to Neanderthals or Denisovans between the variants which fall in segments and the whole genome (Figure 3.21d). Taken together, this suggests that Eurasian-derived segments of the African genomes are not enriched with Neanderthal material.

## 3.4 Discussion

We have developed an approach for estimating demographic parameters and ARGs from whole genome sequence data, which can handle inference in complex demographic models, and implemented this in the software program SMCSMC [10]. We used SMCSMC to investigate ancient migration rates and population substructure, and found evidence for a substantial admixture from ancestors of present-day Eurasian populations into African populations in the Late Middle Paleolithic.

Our analysis suggests that a population ancestral to present-day Eurasians contributed as much as a third of the genetic material in many modern African

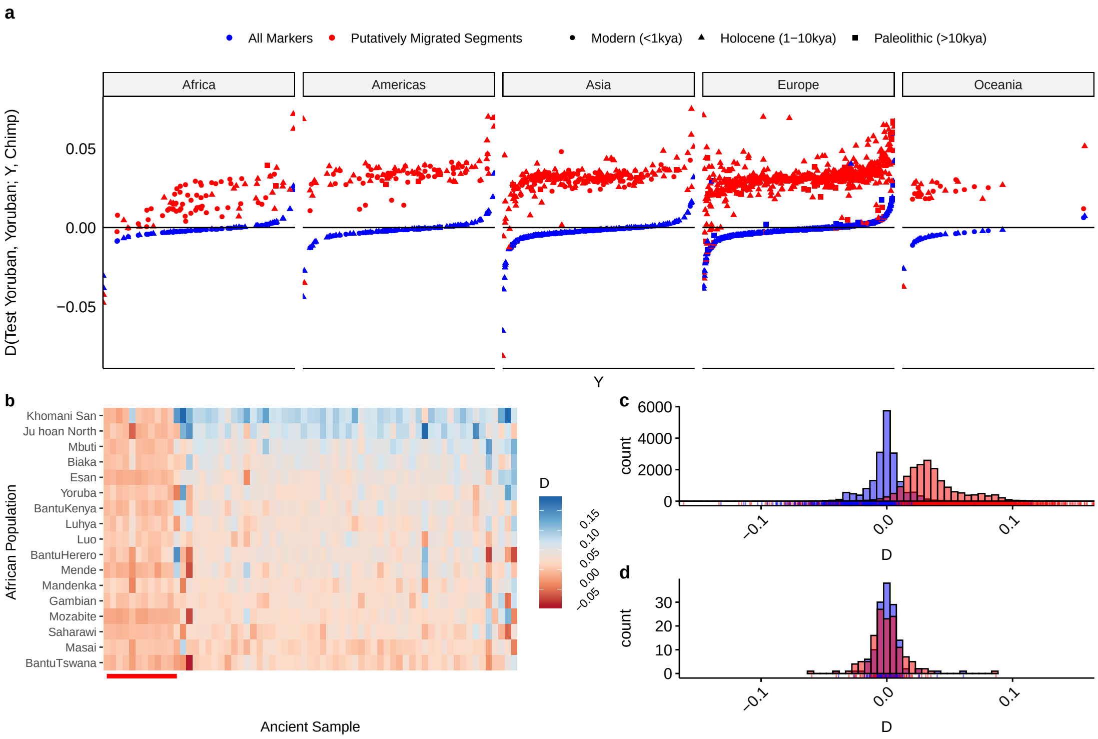


**Figure 3.21: African introgressed segments are more similar to Eurasians but show no Neanderthal or Denisovan enrichment. a.** $D$(Test Yoruban, Comparison Yoruban; Test population, Chimpanzee) calculated for all populations in the Reich Human Origins dataset (see URLs). $D$ statistics in the putatively migrated segments are higher across the board in 3589 ancient, 6472 present day individuals. **b.** The same $D$ statistic but computed for all African populations and individuals sampled from the Paleolithic. Neanderthal and Denisovan samples (marked with red bar) show low affinity to a Yoruban in putatively migrated segments. **c.** Histogram of $D$ statistics computed in a. showing clear inflation of statistics calculated in segments (red) versus all markers (blue). **d.** Subset of individuals from a. involving Neanderthal ($n = 6$), Denisovan ($n = 1$), and a unique mixture individual ($n = 1$) with statistics calculated in segments (red) and all markers (blue) for all $n = 17$ African individuals indicating no difference in this population.

populations. We find no difference in inferred admixture proportions when using French Europeans or Han Chinese as extant representatives of the donor population, indicating that the admixing population must have split from the out-of-Africa population before the East/West Eurasian divergence, implying a lower bound on the timing of the admixture of approximately 40kya [60]. It appears that our results suggest that the migrating population was more similar to present-day French and Chinese populations than to Papuans. However, up to 5% of the genomes of some present-day Papuans have been suggested to derive from archaic introgressions

[62], and these contributions will have reduced the inferred levels of admixture into Africans when using Papuans as a representative of the Eurasian ancestors. The alternative explanation, of an earlier divergence of Papuans and Eurasian ancestors, is possible but contested; in light of documented Eurasian admixture into Oceania, the effects of this early isolation are likely to be small relative to the large confounding effects of Denisovan admixture [38, 63].

The proposed period of admixture has biased previous inferences of the African population sizes. We show that including directional migration into the model resolves previously unexplained high inferred $N_e$ in the period 80 to 300kya. It is well known that effective population size estimates are biased in the presence of population substructure and migration [5, 64]. We use simulations to show that the proposed admixture event indeed causes an increase in estimated $N_e$ in analyses that do not explicitly model migration. Correctly modeling of directional migration recovers the correct $N_e$, and allows us to infer a more recent split time between the two populations than indicated by previous analyses, although we did not attempt to formally estimate this time of divergence.

We found that not all populations in Africa have been equally affected by the proposed migration event. While the ancestors of Niger-Kordofanian and Nilo-Saharan populations show evidence of similar levels of Eurasian admixture, the ancestors of Central African and South African hunter-gatherer populations show markedly lower levels. The date of genetic diversification of both the Central Hunter Gatherers and Khoe-San (SAHG) is contested [32], but a date of 100kya has been proposed [65], providing a putative upper bound on the main admixture event. Our simulations indicate that SMCSMC has little power to detect the impact of migration events occurring more than 70kya, providing an additional upper bound on the time of the migration episode, or the fraction of it that left a sufficiently distinct imprint on extant genetic material.

Compared to the remainder of the Niger-Kordofanians and Nilo-Saharans on the one hand, and the SAHG populations on the other, the Mbuti and Biaka show intermediate levels of admixture. Of these populations, the Biaka show slightly higher levels of admixture than the Mbuti, which is likely due to the well-documented admixture from Western African groups not shared with the Mbuti [66]. The lower levels of admixture in Mbuti and Biaka compared to Niger-Kordofian and Nilo-Saharan populations imply at least partial diversification of the former at the time of the migration, placing an upper bound on the timing. However, dating the diversification of these groups is difficult. Recent estimates using $f$ statistics place the split concurrent with the San in a large-scale early expansion 200-250kya [32], while older data consistently report an earlier split time between 50 and 90 kya [67]. Further clarity on the early structure and diversification of hunter-gatherer populations are necessary to interpret their interactions with Eurasian migrants. The Afroasiatic populations on the other hand show high levels of admixture, which also appears to be of much more recent origin, and it appears likely that this is the result of extensive admixture from Eurasian populations during the Holocene [29, 57].

It has previously been suggested that Eurasian back-migration may be responsible for Neanderthal material in Africans [61]; however, we find no evidence for enrichment of Neanderthal-like material in putatively Eurasian-derived genomic segments in Africans, indicating that Neaderthal introgression into Eurasians occurred after the African introgression event we study here, or that further population structure in the Eurasian ancestral population precluded substantial transmission of Neanderthal material into Africa.

Our findings are consistent with several other published observations. Migration rate estimates using `MSMC-IM` revealed high levels of admixture at times comparable to our results [9]. The coalescent intensity function additionally shows similar histories between sub-Saharan African and Eurasian groups with high coalescent intensity in epochs consistent with our inference and those of `MSMC-IM`, supporting both an early split between the groups and a substantial replacement of genetic material more recently than $\sim$ 100kya [68]. Evidence has been mounting for multiple migrations into the Eurasian continent, possibly mediated by climatic drivers [23, 39]. Eurasian backflow during the Holocene has been well established [21, 69], but earlier migrations have also been proposed before based on observations of the spatial distribution of Y chromosome and mitochondrial haplogroups [70–76]. At the same time, evidence has been mounting for extreme heterogeneity in the history of sub-Saharan Africans, with several unsampled population theorised to have contributed at various points in the past [19, 32, 42]. In light of these recent studies, the observations in this paper add to a growing body of evidence for complex population structure and migration surrounding the Out of Africa event leading to a substantial replacement of the African population in the Late Middle Paleolithic.

| Name | ID | Source |
|---|---|---|
| French | S_French-1 | SGDP |
| Han | S_Han-1 | SGDP |
| Papuan | S_Papuan-1 | SGDP |
| BantuHerero | S_BantuHerero-1 | SGDP |
| BantuKenya | S_BantuKenya-1 | SGDP |
| BantuTswana | S_BantuTswana-1 | SGDP |
| Biaka | S_Biaka-1 | SGDP |
| Dinka | B_Dinka-3 | SGDP |
| Esan | S_Esan-1 | SGDP |
| Gambian | S_Gambian-1 | SGDP |
| Ju hoan North | S_Ju_hoan_North-1 | SGDP |
| Khomani San | S_Khomani_San-1 | SGDP |
| Luhya | S_Luhya-1 | SGDP |
| Luo | S_Luo-1 | SGDP |
| Mandenka | S_Mandenka-1 | SGDP |
| Masai | S_Masai-1 | SGDP |
| Mbuti | S_Mbuti-1 | SGDP |
| Mende | S_Mende-1 | SGDP |
| Mozabite | S_Mozabite-1 | SGDP |
| Saharawi | S_Saharawi-1 | SGDP |
| Somali | S_Somali-1 | SGDP |
| Yoruba | S_Yoruba-1 | SGDP |
| Druze | HGDP00562 | HGDP |
| Han | HGDP00774 | HGDP |
| Karitiana | HGDP01013 | HGDP |
| PapuanHighlands | HGDP00549 | HGDP |
| PapuanSepik | HGDP00542 | HGDP |
| Pathan | HGDP00224 | HGDP |
| Pima | HGDP01043 | HGDP |
| Sardinian | HGDP00670 | HGDP |
| Yakut | HGDP00946 | HGDP |
| Yoruba | HGDP00930 | HGDP |
| San | HGDP01029 | HGDP |
| Mbuti | HGDP00450 | HGDP |
| Biaka | HGDP00460 | HGDP |
| Vindija | Vindija.DG | Pruefer et al. 2017 |
| Altai | Altai_published.DG | Pruefer et al. 2013 |
| Denisovan | Deniosva_published.DG | Myers et al 2012 |

**Table 3.1:** Sample IDs of the individuals used in this article and relevant resources.

| Language Family | Comparison Family | Mean Difference (95% CI) | Adjusted P |
|---|---|---:|---:|
| Niger-Kordofanian | Nilo-Saharan | -0.053 (-0.081–0.025) | 1.11e-03 |
| Niger-Kordofanian | Khoesan | 0.143 (0.125-0.161) | 4.39e-14 |
| Niger-Kordofanian | Afroasiatic | -0.125 (-0.142–0.107) | 1.67e-15 |
| Nilo-Saharan | Niger-Kordofanian | 0.053 (0.025-0.081) | 1.11e-03 |
| Nilo-Saharan | Khoesan | 0.196 (0.169-0.223) | 6.91e-09 |
| Nilo-Saharan | Afroasiatic | -0.072 (-0.098–0.045) | 1.23e-04 |
| Khoesan | Niger-Kordofanian | -0.143 (-0.161–0.125) | 4.39e-14 |
| Khoesan | Nilo-Saharan | -0.196 (-0.223–0.169) | 6.91e-09 |
| Khoesan | Afroasiatic | -0.268 (-0.284–0.252) | 3.62e-13 |
| Afroasiatic | Niger-Kordofanian | 0.125 (0.107-0.142) | 1.67e-15 |
| Afroasiatic | Nilo-Saharan | 0.072 (0.045-0.098) | 1.23e-04 |
| Afroasiatic | Khoesan | 0.268 (0.252-0.284) | 3.62e-13 |
| CAHG | Niger-Kordofanian | -0.084 (-0.119–0.049) | 1.18e-03 |

**Table 3.2: Differences in African IMF in the SGDP.** The integrated migration fraction (IMF) in the epoch 30–70kya is calculated as per the Methods section for all comparisons in the Simons Genome Diversity Project (SGDP), and a two tailed *t*-test is used to statistically test for differences between the inferred migration in African language groups. Averaged over three technical replicates to account for the influence of stochastic sampling variation. P values corrected for multiple testing using the Bonferroni method. Abbreviations: CAHG, Central African Hunter-Gatherers (include Mbuti and Biaka); CI, Confidence Interval.

| Language Family | Partner Population | Mean AFR IMF (SD) | Mean EUR IMF (SD) |
|---|---|---|---|
| Afroasiatic | All | 0.477(0.015) | 0.176(0.038) |
| Afroasiatic | French | 0.477(0.006) | 0.207(0.044) |
| Afroasiatic | Han | 0.492(0.01) | 0.168(0.032) |
| Afroasiatic | Papuan | 0.462(0.011) | 0.153(0.024) |
| Khoesan | All | 0.209(0.013) | 0.044(0.006) |
| Khoesan | French | 0.217(0.007) | 0.051(0.003) |
| Khoesan | Han | 0.217(0.002) | 0.042(0) |
| Khoesan | Papuan | 0.193(0.006) | 0.04(0) |
| Niger-Kordofanian | All | 0.352(0.04) | 0.088(0.017) |
| Niger-Kordofanian | French | 0.36(0.039) | 0.102(0.015) |
| Niger-Kordofanian | Han | 0.363(0.04) | 0.083(0.013) |
| Niger-Kordofanian | Papuan | 0.334(0.038) | 0.078(0.013) |
| Nilo-Saharan | All | 0.405(0.033) | 0.107(0.016) |
| Nilo-Saharan | French | 0.414(0.037) | 0.123(0.016) |
| Nilo-Saharan | Han | 0.417(0.036) | 0.106(0.01) |
| Nilo-Saharan | Papuan | 0.385(0.029) | 0.093(0.008) |

**Table 3.3: Integrated Migration Fraction (IMF) in either direction averaged over African language groups.** SMCSMC was used as in Supplemental Section **??** to infer directional migration and effective population size between populations in the Simons Genome Diversity Project. The total migration between each African and non-African population was integrated in the epoch 30–70kya (see Methods) and averaged over language family. Abbreviations: Integrated Migration Fraction (IMF), Stanard Deviation (SD), EUR (Eurasian), AFR (African)

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 0.3807 | 0.0036 | 105.42 | 1.79e-47 |
| Papuan | -0.0271 | 0.0017 | -15.60 | 7.44e-18 |
| BantuKenya | 0.0051 | 0.0050 | 1.02 | 3.16e-01 |
| BantuTswana | -0.0410 | 0.0050 | -8.13 | 9.49e-10 |
| Biaka | -0.0598 | 0.0050 | -11.87 | 3.53e-14 |
| Dinka | 0.0160 | 0.0050 | 3.17 | 3.05e-03 |
| Esan | -0.0016 | 0.0050 | -0.32 | 7.51e-01 |
| Gambian | -0.0073 | 0.0050 | -1.44 | 1.58e-01 |
| Ju hoan North | -0.1603 | 0.0050 | -31.79 | 1.82e-28 |
| Khomani San | -0.1653 | 0.0050 | -32.80 | 5.96e-29 |
| Luhya | 0.0118 | 0.0050 | 2.34 | 2.46e-02 |
| Luo | 0.0123 | 0.0050 | 2.45 | 1.94e-02 |
| Mandenka | 0.0032 | 0.0050 | 0.63 | 5.34e-01 |
| Masai | 0.0719 | 0.0050 | 14.25 | 1.32e-16 |
| Mbuti | -0.1171 | 0.0050 | -23.22 | 1.17e-23 |
| Mende | -0.0071 | 0.0050 | -1.42 | 1.65e-01 |
| Mozabite | 0.1060 | 0.0050 | 21.02 | 3.63e-22 |
| Saharawi | 0.1097 | 0.0050 | 21.75 | 1.12e-22 |
| Somali | 0.0995 | 0.0050 | 19.73 | 3.12e-21 |
| Yoruba | -0.0013 | 0.0050 | -0.26 | 7.97e-01 |

**Table 3.4: Linear model predicting integrated migration fraction in the SGDP.** The integrated migration fraction (IMF) in the epoch 30–70 kya is obtained as per the Methods section in the Simons Genome Diversity Project. A binary variable representing Papuan / not Papuan Eurasian donor and categorical variable representing African population were used to predict the IMF in a simple linear model. When adjusted for the different African populations, Papuans contribute less IMF than do other Eurasian partners (French and Han).

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 0.3392 | 0.0037 | 92.92 | 1.72e-39 |
| Papuan | -0.0252 | 0.0042 | -5.95 | 1.43e-06 |
| San | -0.1942 | 0.0050 | -38.94 | 6.85e-28 |
| Mbuti | -0.1410 | 0.0050 | -28.27 | 1.07e-23 |
| Biaka | -0.0883 | 0.0050 | -17.69 | 9.33e-18 |

**Table 3.5: Linear model predicting integrated migration fraction in the HGDP.** The integrated migration fraction (IMF) in the epoch 30–70 kya is obtained as per the Methods section in the physically phased subset of the Human Genome Diversity Project. A binary variable representing Papuan / not Papuan Eurasian donor and categorical variable representing African population were used to predict the IMF in a simple linear model. When adjusted for the different African populations, Papuans contribute less IMF than do other Eurasian partners (French and Han).

| African Population | Mean (SD) | Total (Mb) | Mean (SD) | Total (Mb) | Mean (SD) | Total (Mb) |
|---|---|---|---|---|---|---|
| Mbuti | 172.204 (189.255) | 939.371 | 165.822 (182.549) | 850.335 | 163.352 (181.607) | 756.320 |
| Biaka | 172.332 (184.183) | 1057.083 | 169.296 (185.457) | 1052.004 | 170.149 (185.853) | 1044.036 |
| Khomani San | 173.06 (188.492) | 671.125 | 169.005 (183.871) | 706.777 | 166.33 (175.899) | 597.125 |
| Ju hoan North | 175.415 (195.979) | 747.618 | 171.909 (187.835) | 711.186 | 161.146 (175.476) | 656.024 |
| Luhya | 177.303 (193.164) | 1370.729 | 181.025 (197.94) | 1403.123 | 175.821 (192.259) | 1211.762 |
| Esan | 177.483 (191.148) | 1364.132 | 178.429 (190.819) | 1426.180 | 170.451 (183.886) | 1270.204 |
| Gambian | 177.804 (195.646) | 1333.529 | 177.099 (191.542) | 1304.155 | 173.892 (185.544) | 1213.071 |
| Luo | 178.653 (192.447) | 1368.481 | 175.618 (192.833) | 1361.917 | 180.635 (201.159) | 1348.798 |
| BantuHerero | 179.691 (194.451) | 1354.869 | 176.473 (188.855) | 1327.779 | 178.544 (198.496) | 1310.511 |
| BantuTswana | 180.309 (195.176) | 1150.551 | 174.826 (188.153) | 1159.796 | 172.542 (192.6) | 1096.674 |
| Mende | 182.087 (200.591) | 1256.580 | 177.24 (197.23) | 1288.533 | 169.901 (182.944) | 1191.515 |
| Yoruba | 184.255 (199.949) | 1283.151 | 176.276 (193.69) | 1361.031 | 169.867 (181.948) | 1275.701 |
| BantuKenya | 184.419 (200.395) | 1265.297 | 173.485 (191.175) | 1269.739 | 179.415 (189.784) | 1214.104 |
| Mandenka | 185.35 (206.972) | 1411.440 | 176.908 (194.337) | 1328.754 | 172.03 (182.984) | 1265.625 |
| Masai | 186.769 (209.675) | 1353.890 | 182.366 (198.06) | 1483.361 | 181.22 (199.255) | 1438.160 |
| Mozabite | 194.378 (214.029) | 1318.273 | 193.469 (211.945) | 1557.040 | 188.867 (210.994) | 1532.842 |
| Saharawi | 197.668 (218.662) | 1383.280 | 196.156 (217.233) | 1406.245 | 195.045 (217.263) | 1585.907 |

**Table 3.6:** Summary of the length distribution for putatively migrated segments in different African individuals. Means and standard deviations are given in kilobases (kb) while the total length of all segments is given in megabases (Mb).

| Statistic | D | Z |
|---|---|---|
| D(KhomaniSan-1, Yoruba-1, Yoruba-2, Chimp) | -0.181 | -28.403 |
| D(Mbuti-1, Yoruba-1, Yoruba-2, Chimp) | -0.135 | -19.554 |
| D(Papuan-1, Yoruba-1, Yoruba-2, Chimp) | -0.026 | -3.422 |
| D(French-1, Yoruba-1, Yoruba-2, Chimp) | -0.006 | -0.866 |
| D(Han-1, Yoruba-1, Yoruba-2, Chimp) | 0.001 | 0.072 |
| D(KhomaniSan-1, Yoruba-2, Yoruba-1, Chimp) | -0.187 | -28.109 |
| D(Mbuti-1, Yoruba-2, Yoruba-1, Chimp) | -0.130 | -19.323 |
| D(Papuan-1, Yoruba-2, Yoruba-1, Chimp) | -0.008 | -1.003 |
| D(French-1, Yoruba-2, Yoruba-1, Chimp) | 0.030 | 4.355 |
| D(Han-1, Yoruba-2, Yoruba-1, Chimp) | 0.056 | 8.037 |

**Table 3.7:** Putatively migrated segments of a Yoruban are closer to Out of Africa groups than a comparable Yoruban.

| Statistic | D | Z |
|---|---|---|
| D(Mbuti-1, Yoruba-1, Vindija, Chimp) | -0.001 | -0.141 |
| D(Mbuti-1, Yoruba-2, Vindija, Chimp) | -0.003 | -0.306 |

**Table 3.8:** No difference in allele sharing with Vindija Neanderthal over Mbuti baseline.

| Statistic | D | Z |
|---|---|---|
| D(Yoruba-2, Yoruba-1, Vindija, Chimp) | 0.000 | 0.012 |

**Table 3.9:** No difference in allele sharing with Vindija Neanderthal.

| Statistic | D | Z |
|---|---|---|
| D(Vindija, Altai, Yoruba-1, Chimp) | 0.024 | 1.095 |
| D(Vindija, Altai, Yoruba-2, Chimp) | 0.034 | 1.526 |
| D(Vindija, Altai, Mbuti-1, Chimp) | 0.002 | 0.103 |
| D(Vindija, Altai, KhomaniSan-1, Chimp) | 0.023 | 1.008 |

**Table 3.10:** No increased affinity to Vindija Neanderthal over Altai, as would be expected if the source of any Neanderthal ancestry was Eurasian.

# 4
# Topic Modelling and Latent Dirichlet Allocation for Unsupervised Clustering of Bulk ATAC-seq Experiments

**Contents**

## 4.1 Erythropoesis as a well-characterised model system

Because of the relative sparsity of well characterised MLL-r specific enhancers, we choose to first model a differentiation system which has been well characterised. Erythropoesis is the process by which hematopoetic stem cells first differentiate into progenitor populations before committing to the myeloid lineage and eventually undergoing enucleation and terminal differentition into mature erythrocytes. This system has been extensively studied (cite Ludwig and Corces and the protein one) and a catelog of differentially active transcription factors and associated enhancer elements can be readily derived from the literature. We seek to validate the proposed topic modelling approach on this model system by systematically building up a model of erythropoesis, including sequentially various detractor lineages such as a related differentiation to B cell lymphopoesis, asking whether topics are able to reproducibly recapitulate known dynamics of lineage committment and terminal differentiation.

1. Talk about other tools for performing this task and how this is not actually the same question. We need to be able to differentiate between arbitrary number of cells and cell states, not limited to just lookign at systems where the full differentiaiton hierarchy is available.

2. Talk about the methods for fidning these differentially accessible elements between the different stages.

### 4.1.1 Data preperation and peak calling

1. Where was the data downloaded from? How was it processed?

2. Peak calling. MaCS3 versus lance-o-tron. What kind of diagnostic features can we see here?

3. Number of peaks.

4. Peak length distribution

A

## 4.2 Results

- LDA is shown to be a viable method for single cell ATAC. Does it also work on bulk experiments?
- LDA is able to find reproducible topic loadings that differentiate pseudobulked clusters.
    - (Pseudobulk cluster topic loadings resemble single cell topic loadings)

  - **Question**: Have a view on what the best data for this question would have been * (put this into the discussion)

- LDA versus differential OCR ***(think more about this)***
- Topics are reproducible across stochastic replicates, but variously so.
  - Dimensionality reduction with and without the not-reproducible topics
  - Are there any systematic differences between the reproducible topics and the not reproducible ones (i.e. differences in strengths of loadings, distribution across the cells, localisation in a cluster, etc.)?
- Hyper parameter optimisation strategy
  - Dask application for bayesian optimisation of hyper parameters to maximise the LL while integrating over the number of topics selected
- Methods for isolating regions from region-topic annotations
  - Compare motifs identified through thresholding, gamma distribution threshold, top N approach, etc. Which produces the most reliable set of motifs for the various topics?
- Applicability to a biological system: Erythropoesis
  - Hyper parameter search using the Dask application
  - Topic loadings are stage specific.
  - Motifs represent biologically relevant transcription factors that are active at each stage and recapitulate the dynamics of erythropoesis
  - Major topics are consistent across replications and the stages always seperate based on topic loadings
- LDA recapitulates larger scale trends in chromatin accessibility:
  - Topic distribution in the ENCODE project recapitulates large scale trends in accessibility.
  - Dig into a couple of the interesting topics and try to figure out what they are doing here.

- Conclusion: LDA represents an interesting and useful method for Unsupervised clustering of different cell types from bulk ATAC-seq

### 4.2.1 Chapter 5: Topic Modelling Prioritises OCRs in MLL-AF4 Recombinant Cells

- **Describe what the appropriate validation should be here. From these results, I would like to see X, Y, Z experimental approaches.**

**Main question**: *What distinguishes MLL-AF4 from healthy cells in terms of open chromatin?*

- LoT results in cleaner peak calls for topic modelling in MLL-AF4 cells
- RS4;11 and SEM can be merged to form a cohesive MLL-AF4 group
  - Differential accessibility analysis between the two different cell types. (What is different, and why is this not important.)
- Differential peak analysis between MLL-AF4 and BCP (B cell precursors) to provide a baseline expectations of the regions we may see
-

# Appendices

# A
# Demographic Models

Generally, these models can be implemented in either `scrm` or `ms` through they have been written with the former in mind.

## A.1 Seed model for SMCSMC Inference

We seed the particle filter with a demographic model of population size and uniform symmetric migration rate, given by the following `scrm` command:

```
-ej 0.2324 2 1 -eM 0 1 -eN 0.0 6 -eN 0.0037 4.4 -eN 0.0046 3 -eN 0.0058 2 -eN 0.0073 1.4
-eN 0.0092 0.85 -eN 0.093 1.2 -eN 0.12 1.7 -eN 0.15 2.2 -eN 0.19 2.5 -eN 0.24 2.4
-eN 0.30 2.0 -eN 0.37 1.7 -eN 0.47 1.4 -eN 0.59 1.2 -eN 0.74 1.0 -eN 0.93 0.91 -eN 1.2 1.6
```

## A.2 Migration Simulations

The following models were used for population sizes:

### A.2.1 African Population Size

```
  -en 0.00000000 1 36.9124479 -en 0.00229999 1 14.8978177 -en 0.00299994 1 7.04453213
  -en 0.00391291 1 3.68961222 -en 0.00510371 1 2.06587476 -en 0.00665692 1 1.21617010
  -en 0.00868280 1 0.75362392 -en 0.01132521 1 0.49927968 -en 0.01477178 1 0.36258332
  -en 0.01926724 1 0.29687253 -en 0.02108190 1 0.28637149 -en 0.02513079 1 0.28071694
  -en 0.03277878 1 0.31028768 -en 0.03915210 1 0.36107482 -en 0.04275426 1 0.39815181
  -en 0.05576555 1 0.57528787 -en 0.07273654 1 0.88701054 -en 0.09487226 1 1.36014053
  -en 0.12374449 1 1.92573639 -en 0.16140334 1 2.36832894 -en 0.21052280 1 2.45284038
  -en 0.27459066 1 2.16222564 -en 0.35815613 1 1.71146032 -en 0.46715286 1 1.32388966
  -en 0.60932028 1 1.09778746 -en 0.79475315 1 1.04669123 -en 1.03661833 1 1.16969768
  -en 1.35208972 1 1.45788656 -en 1.76356769 1 1.80077313 -en 2.30026970 1 1.89942369
```

### A.2.2 Eurasian Population Size

```
-en 0.00000000 2 1.14422216 -en 0.00229999 2 1.14422216 -en 0.00299994 2 1.14422216
-en 0.00391291 2 1.14422216 -en 0.00510371 2 1.14422216 -en 0.00665692 2 1.14422216
-en 0.00868280 2 1.14422216 -en 0.01132521 2 1.14422216 -en 0.01477178 2 1.14422216
-en 0.01926724 2 1.14422216 -en 0.02108190 2 1.14422216 -en 0.02513079 2 1.14422216
-en 0.03277878 2 1.14422216 -en 0.03915210 2 1.14422216 -en 0.04275426 2 1.14422216
-en 0.05576555 2 1.14422216 -en 0.07273654 2 1.14422216 -en 0.09487226 2 1.36014053
-en 0.12374449 2 1.92573639 -en 0.16140334 2 2.36832894 -en 0.21052280 2 2.45284038
-en 0.27459066 2 2.16222564 -en 0.35815613 2 1.71146032 -en 0.46715286 2 1.32388966
-en 0.60932028 2 1.09778746 -en 0.79475315 2 1.04669123 -en 1.03661833 2 1.16969768
-en 1.35208972 2 1.45788656 -en 1.76356769 2 1.80077313 -en 2.30026970 2 1.89942369
```

# References

[1] Robert C. Griffiths and Paul Marjoram. “An Ancestral Recombination Graph”. In: 1997.

[2] Matthew D. Rasmussen et al. “Genome-Wide Inference of Ancestral Recombination Graphs”. In: *PLoS Genetics* (2014).

[3] Laurent Excoffier et al. “Robust Demographic Inference from Genomic and SNP Data”. In: *PLoS Genetics* (2013).

[4] Gilean A.T. McVean and Niall J. Cardin. “Approximating the coalescent with recombination”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* (2005).

[5] Heng Li and Richard Durbin. “Inference of human population history from individual whole-genome sequences”. In: *Nature* 475.7357 (2011), pp. 493–496. arXiv: `arXiv:1011.1669v3`.

[6] Matthias Steinrücken, John A. Kamm, and Yun S. Song. “Inference of complex population histories using whole-genome sequences from multiple populations”. In: *bioRxiv* (2015), p. 026591.

[7] Jonathan Terhorst and Yun S. Song. “Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum”. In: *Proceedings of the National Academy of Sciences* 112.25 (2015), pp. 7677–7682.

[8] Stephan Schiffels and Richard Durbin. “Inferring human population size and separation history from multiple genome sequences”. In: *Nature Genetics* 46.8 (2014), pp. 919–925. arXiv: `005348 [10.1101]`.

[9] Ke Wang et al. “Tracking human population structure through time from whole genome sequences”. In: *bioRxiv* (2019), pp. 1–21.

[10] Donna Henderson, Sha (Joe) Zhu, and Gerton Lunter. “Demographic inference using particle filters for continuous Markov jump processes”. In: *bioRxiv* (2018), p. 382218.

[11] Swapan Mallick et al. “The Simons Genome Diversity Project: 300 genomes from 142 diverse populations”. In: *Nature* 538.7624 (2016), pp. 201–206.

[12] Anders Bergström et al. “Insights into human genetic variation and population history from 929 diverse genomes”. In: *bioRxiv* (2019), p. 674986.

[13] Christopher B. Cole et al. “Ancient Admixture into Africa from the ancestors of non-Africans”. In: *bioRxiv* (2020).

[14] Kent E. Holsinger and Bruce S. Weir. *Genetics in geographically structured populations: Defining, estimating and interpreting FST*. 2009.

[15] Toomas Kivisild. *Maternal ancestry and population history from whole mitochondrial genomes.* 2015.

[16] Daniel Rubinoff and Brenden S. Holland. “Between two extremes: mitochondrial DNA is neither the panacea nor the nemesis of phylogenetic and taxonomic inference.” In: *Systematic biology* 54.6 (2005).

[17] Nick Patterson et al. “Ancient Admixture in Human History”. In: 192.November (2012), pp. 1065–1093.

[18] Martin Petr, Benjamin Vernot, and Janet Kelso. “admixr — R package for reproducible analyses using ADMIXTOOLS”. In: *Bioinformatics* January (2019), pp. 1–2.

[19] Leo Speidel et al. “A method for genome-wide genealogy estimation for thousands of samples”. In: *Nature Genetics* 51.9 (2019), pp. 1321–1329.

[20] Jerome Kelleher et al. “Inferring whole-genome histories in large population datasets”. In: *Nature Genetics* 51.9 (2019), pp. 1330–1338.

[21] Saioa López, Lucy van Dorp, and Garrett Hellenthal. “Human Dispersal Out of Africa: A Lasting Debate.” In: *Evolutionary bioinformatics online* 11.Suppl 2 (2015), pp. 57–68.

[22] Frank Schaebitz et al. “Hydroclimate changes in eastern Africa over the past 200,000 years may have influenced early human dispersal”. In: *Communications Earth & Environment* 2.1 (2021), pp. 1–10. URL: `http://dx.doi.org/10.1038/s43247-021-00195-7`.

[23] Axel Timmermann and Tobias Friedrich. “Late Pleistocene climate drivers of early human migration”. In: *Nature* 538.7623 (2016), pp. 92–95.

[24] Sriram Sankararaman et al. “The Date of Interbreeding between Neandertals and Modern Humans”. In: *PLoS Genetics* (2012).

[25] Qiaomei Fu et al. “Genome sequence of a 45,000-year-old modern human from western Siberia”. In: *Nature* 514.7253 (2014), pp. 445–449.

[26] Pontus Skoglund et al. “Reconstructing Prehistoric African Population Structure”. In: *Cell* 171.1 (2017), 59–71.e21.

[27] Mark Lipson, David Reich, and Jeffrey P. Townsend. “A working model of the deep relationships of diverse modern human genetic lineages outside of Africa”. In: *Molecular Biology and Evolution* 34.4 (2017), pp. 889–902.

[28] Iosif Lazaridis et al. “Ancient human genomes suggest three ancestral populations for present-day Europeans”. In: *Nature* 513.7518 (2014), pp. 409–413. arXiv: `1312.6639`. URL: `http://dx.doi.org/10.1038/nature13673`.

[29] George Bj Busby et al. “Admixture into and within sub-Saharan Africa”. In: *eLife* (2016).

[30] Etienne Patin et al. “Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America”. In: *Science* 356.6337 (2017), pp. 543–546.

[31] Pontus Skoglund and Iain Mathieson. “Ancient Genomics of Modern Humans: The First Decade”. In: *Annual Review of Genomics and Human Genetics* 19.1 (2018), pp. 381–404.

[32] Mark Lipson et al. “Ancient West African foragers in the context of African population history”. In: November 2018 (2019).

[33] M. Gallego Llorente et al. “Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent”. In: *Science* 350.6262 (2015), pp. 820–822. arXiv: `arXiv:1011.1669v3`.

[34] Carina M. Schlebusch et al. “Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago”. In: *Science* (2017).

[35] Chris Clarkson et al. “Human occupation of northern Australia by 65,000 years ago”. In: *Nature* 547.7663 (2017), pp. 306–310.

[36] Wu Liu et al. “The earliest unequivocally modern humans in southern China”. In: *Nature* 526.7575 (2015), pp. 696–699.

[37] K E Westaway et al. “An early modern human presence in Sumatra 73,000-63,000 years ago.” In: *Nature* 548.7667 (2017), pp. 322–325.

[38] Anna Sapfo Malaspinas et al. “A genomic history of Aboriginal Australia”. In: *Nature* 538.7624 (2016), pp. 207–214. arXiv: `NIHMS150003`.

[39] Luca Pagani et al. “Genomic analyses inform on migration events during the peopling of Eurasia”. In: *Nature* 538.7624 (2016), pp. 238–242.

[40] Morten Rasmussen et al. “An aboriginal Australian genome reveals separate human dispersals into Asia”. In: *Science* (2011).

[41] Pontus Skoglund et al. “Genetic evidence for two founding populations of the Americas”. In: *Nature* 525 (2015), p. 104.

[42] Arun Durvasula and Sriram Sankararaman. “Recovering signals of ghost archaic introgression in African populations”. In: *bioRxiv* (2019), p. 285734.

[43] M. F. Hammer et al. “Genetic evidence for archaic admixture in Africa”. In: *Proceedings of the National Academy of Sciences* 108.37 (2011), pp. 15123–15128.

[44] Vincent Plagnol and Jeffrey D Wall. “Possible Ancestral Structure in Human Populations”. In: 2.7 (2006).

[45] Aaron P. Ragsdale and Simon Gravel. “Models of archaic admixture and recent history from two-locus statistics”. In: *PLoS genetics* (2019).

[46] Paul R. Staab et al. “SCRM: efficiently simulating long sequences using the approximated coalescent with recombination”. In: *Bioinformatics* 31.10 (2015), pp. 1680–1682.

[47] Ying Zhou et al. “POPdemog: visualizing population demographic history from simulation scripts”. In: *Bioinformatics (Oxford, England)* (2018).

[48] Donna Henderson et al. “Demographic inference from multiple whole genomes using a particle filter for continuous Markov jump processes”. In: *PLOS ONE* 16.3 (2021), pp. 1–24. URL: `https://doi.org/10.1371/journal.pone.0247647`.

[49] Jack N. Fenner. “Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies”. In: *American Journal of Physical Anthropology* 128.2 (2005), pp. 415–423.

[50] Paul Marjoram and Jeff D Wall. “Fast “coalescent” simulation”. In: *BMC Genetics* 7.1 (2006), p. 16.

[51] Mason Liang and Rasmus Nielsen. “The Lengths of Admixture Tracts”. In: *Genetics* 197.3 (2014), pp. 953–967.

[52] Fernando Racimo et al. "Evidence for archaic adaptive introgression in humans". In: *Nature Reviews Genetics* 16 (2015), p. 359.

[53] Beth L Dumont and Bret A Payseur. "Evolution of the genomic rate of recombination in mammals". In: *Evolution* 62.2 (2008), pp. 276–294.

[54] Johannes Köster and Sven Rahmann. "Snakemake-a scalable bioinformatics workflow engine". In: *Bioinformatics* (2012).

[55] Aylwyn Scally and Richard Durbin. "Revising the human mutation rate: Implications for understand ing human evolution". In: *Nature Reviews Genetics* 13.10 (2012), pp. 745–753.

[56] Stephan Schiffels and Richard Durbin. "Inferring human population size and separation history from multiple genome sequences". In: *Nature Genetics* 46.8 (2014), pp. 919–925. arXiv: `005348 [10.1101]`.

[57] Shaohua Fan et al. "African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations". In: *Genome Biology* (2019).

[58] Luca Pagani et al. "Tracing the Route of Modern Humans out of Africa by Using 225 Human Genome Sequences from Ethiopians and Egyptians". In: *American Journal of Human Genetics* 96.6 (2015), pp. 986–991.

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

[60] Iain Mathieson and Gil McVean. "Demography and the Age of Rare Variants". In: *PLoS Genetics* 10.8 (2014). arXiv: `1401.4181`.

[61] Lu Chen et al. "Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals Article Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals". In: *Cell* (2020), pp. 1–11.

[62] Sriram Sankararaman et al. "The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans". In: *Current Biology* 26.9 (2016), pp. 1241–1247. URL: `http://dx.doi.org/10.1016/j.cub.2016.03.037`.

[63] Rasmus Nielsen et al. *Tracing the peopling of the world through genomics.* 2017.

[64] Lounès Chikhi et al. "The IICR (inverse instantaneous coalescence rate) as a summary of genomic diversity: insights into demographic inference and model choice". In: *Heredity* 120.1 (2018), pp. 13–24.

[65] Carina M. Schlebusch et al. "Genomic Variation in Seven Khoe-San". In: 1187.October (2012), pp. 374–379.

[66] Chiara Batini et al. "Insights into the demographic history of African pygmies from complete mitochondrial genomes". In: *Molecular Biology and Evolution* 28.2 (2011), pp. 1099–1110.

[67] Etienne Patin and Lluis Quintana-Murci. "The demographic and adaptive history of central African hunter-gatherers and farmers". In: *Current Opinion in Genetics and Development* 53.August (2018), pp. 90–97.

[68] Patrick K. Albers and Gil McVean. "Dating genomic variants and shared ancestry in population-scale sequencing data". In: *bioRxiv* (2019), p. 416610.

[69] M Gallego Llorente and A Manica. “Ancient Ethiopian genome reveals extensive Eurasian admixture in Eastern Africa”. In: *Science* 350.October (2015), pp. 820–825.

[70] T K Altheide and M F Hammer. “Evidence for a possible Asian origin of YAP + Y chromosomes.” In: *American journal of human genetics* 61.2 (1997), pp. 462–6.

[71] M. F. Hammer et al. “Out of Africa and back again: nested cladistic analysis of human Y chromosome variation”. In: *Molecular Biology and Evolution* 15.4 (1998), pp. 427–441.

[72] Fulvio Cruciani et al. “A Back Migration from Asia to Sub-Saharan Africa Is Supported by High-Resolution Analysis of Human Y-Chromosome Haplotypes”. In: *The American Journal of Human Genetics* 70.5 (2002), pp. 1197–1214.

[73] A. Chandrasekar et al. “YAP insertion signature in South Asia”. In: *Annals of Human Biology* 34.5 (2007), pp. 582–586.

[74] Vicente M. Cabrera et al. “Carriers of mitochondrial DNA macrohaplogroup L3 basal lineages migrated back to Africa from Asia around 70,000 years ago”. In: *BMC Evolutionary Biology* 18.1 (2018), p. 98.

[75] M Hervella et al. “The mitogenome of a 35,000-year-old Homo sapiens from Europe supports a Palaeolithic back-migration to Africa”. In: *Scientific Reports* 6 (2016), p. 25501.

[76] Marc Haber et al. “A Rare Deep-Rooting D0 African Y-Chromosomal Haplogroup and Its Implications for the Expansion of Modern Humans out of Africa”. In: *Genetics* (2019), genetics.302368.2019.