Genomic Evidence for Prehistoric Migrations to Africa and Population Structure in the Paleolithic

Abstract

Characterising human evolutionary history requires an understanding of genetic diversity in Africa, however demographic inference in the ancient past is difficult. Here, we develop a method called SMCSMC which is able to estimate the ancestral recombination graph. We use this method to find evidence for substantial migration from the ancestors of modern Eurasians into African groups between 50 and 70 thousand years ago (kya), accounting for previously unexplained genetic diversity in the ancestral African population approximately 100kya.

1 Introduction

Understanding global patterns of genetic variation requires a firm characterisation of their origins in Africa. Our knowledge of prehistory in Africa is limited, at least in part due to the complexity of inferring demographic models over long periods of time. Trees along the genome, collectively known as the ancestral recombination graph (ARG), represent a high-dimensional data structure whose inference is confounded by, to name a few, background selection, population substructure, and gene flow. Recent methods for inferring the ARG focus on estimating changes in population size and selective pressure, but do not deal with the effects of migration [1, 2]. In this article we introduce SMCSMC, a particle filter which is able to efficiently infer the ARG of up to four diploid samples and simultaneously estimate demographic parameters leading to the observed trees. We apply SMCSMC to the Simons Genome Diversity Panel (SGDP) and the Human Genome Diversity Panel (HGDP) and demonstrate complex interactions between modern populations in the ancient past.

2 Results

Directional Migration Explains Excess African Diversity 100kya Inference of African effective population size in the ancient past has consistently reported an increase immediately subsequent to the divergence of Eurasian populations [3, 4]. We use SMCSMC to infer the ARG in one individual from each African population in the SGDP along with three seperate partner Eurasian populations (Han Chinese, French European, and Papua New Guinean) and repeated the analysis three times to estimate error (all inference shown in S2). Our inference identifies significantly lower African effective population size than MSMC, with the populations decreasing together until a later split time (Figure 1). We then asked if these observations were due to systematic differences between the inference methods. To do this, we modelled the effective population size of one single African genome in SMCSMC. In this scenario, the N_e inference was comparable between the single-genome SMCSMC and MSMC, with a higher N_e and temporary increase around the population divergences (green in Figure 1). The differences between the two SMCSMC analyses are due to the incorporation of Eurasian migration. We asked if we could recreate this difference artificially, and used `scrm` to simulate one gigabase of sequence from demographic scenarios involving several levels of gene flow between Africa and Eurasia in either and both directions. We analysed the simulated data together, inferring N_e and migration together, and for each of the diploid genomes separately, inferring only N_e. We found that inference in a single genome where migration back to Africa was simulated significantly inflated N_e is a similar period to real data. This includes scenarios with bidirectional migration, which we cannot

conclusively rule out. Migration from Africans to Eurasians showed no inflation in the single genome analysis. We next asked if this difference in inference could be replicated in a different data set, and used MSMC and SMCSMC to infer effective population size in a physically phased subset of the Human Genome Diversity Panel. We find comparable results to the SGDP, with the incorporation of migration leading to a more recent split time and smaller African N_e. We do not expect that errors made during statistical phasing would have any systematic impact on the inferences made by either MSMC or SMCSMC, however this replication in a physically phased dataset suggests that our intuition is correct. From this we conclude that previous inferences of African N_e have been biased by undetected directional migration.

SMCSMC Infers Gene Flow from Eurasian Ancestors to Africans 50-70kya SMCSMC infers high rates of migration from Eurasians to Africans in the ancient past, consistently several times higher than gene flow from Africa to Eurasia (Figure 3). The peak of the migration is slightly different between the individual SGDP population analysed. However, in Niger-Kordofanian and Nilo-Saharan populations, the peak migration either occurs in the epoch ranging approximately 35-45kya, 45-55kya, or 55-70kya (integrated migration proportions in S5, all inference in S3). We asked if SMCSMC was well powered to detect migration in this time period, and used `scrm` to simulate a variety of demographic models involving migration in either direction and both. We integrate the recovered migration over the last 100ky, and find that SMCSMC has more power to detect recent migrations than more ancient ones (Figure 2). In addition, we see that in order to recover a migration of a similar magnitude of that which we see in real data, a migration in excess of 50% replacement must be simulated. SMCSMC is better powered to detect migration from Eurasia to Africa than the reverse, as both forwrd and bidirectional migration recovers less of the true signal than the backwards case (bottom two rows of 2). Again, we sought to understand the effect of phasing errors and replicated the analysis in the physically phased HGDP. We found comparable magnitudes of migration in both datasets.

Enriched Affinity to Eurasians in Portions of the Genome Predicted to Migrate We asked whether we could find evidence for enriched allele sharing with Eurasians in the portions of the genome inferred to have a migration event 50-70kya. We used the inferred ARG to select these segments, and filtered the Reich Human Origins genotype dataset into two sets (details in Supplementary Section S2). We isolate segments in one African, and make sets contains all markers available, and only those variants which fall within our predicted segments. Using a Yoruban as a representative of Western African populations, we find evidence for admixture with negative f_3(Comparison African, Eurasian, Test African) in the subset of markers falling in segments ($|Z| > 3$), but not in the whole panel ($|Z| < 3$). We also find that the individual is closer to Out of African groups than a comparable Yoruban ($D < 0, |Z| > 3$, Table S2). Reversing the direction of gene flow, we can see that OoA groups have contributed more to test individual than another Yoruban ($|Z| > 3$, Table S5). We then asked if, in general, Africans were more related to Eurasians than another representative of their population and computed D(Test African, Comparable African; Eurasian Population, Chimpanzee) for all African populations in the SGDP and all Eurasian populations in the Reich Human Origins variant dataset 4. We find that D statistics are systematically inflated in the identified segments.

Something about f3 statistics?

Less Gene Flow to Central and South African Hunter Gatherers Though there is substantial migration inferred in all African groups, the integrated proportion of the available KhoeSan (SAHG) groups is significantly lower than any other language family ($P < 0.05$ for all donor populations, two tailed T test, Table S1). We ask if this difference could be seen in MSMC, and calculate the relative cross-coalescent rate (xcoal) for eight donating Eurasian populations (right side of Figure 3). We see evidence for contrasting migration histories in the MSMC inferred xcoal as well in the SMCSMC inference (all xcoal rates in Figure S3, and for additional donating populations in S8). We find different xcoal curves for the KhoeSan and the Yoruban, indicating different histories of migration in the ancient past. KhoeSan groups are inferred to have been replaced at approximately half the rate of Nilo-Saharan and Niger-Kordofanian groups (Figure S6). The Mbuti and Biaka, both South African Hunter Gatherer populations, are inferred to have the lowest degree of population replacement outside of KhoeSan. The Biaka have a higher integrated migration proportion, potentially inherited from known extensive recent gene flow from Bantu speaking ancestors, than the Mbuti,

who had less interaction with Bantu speaking groups [5]. Migration in the KhoeSan and Mbuti amounts to between 25-35%, whereas the Biaka integrate to between 42-50% replacement.

Less Gene Flow from Papuans than Han Chinese or French Europeans We compare the inferred proportions of replacement between the three different Eurasian populations acting as donors and find that, on average, Papuans contribute approximately 5% less than do French or Han (Figure S6).

No Evidence for Excess Neanderthal Ancestry Previous studies have proposed that a backflow from Eurasia may have brought Neanderthal ancestry [6]. To evaluate these claims, we once again isolate the segments of the African genome with a migration event 50-70kya. Isolating segments from a representative Yoruban, we find no evidence for gene flow with a Vindija Neanderthal on the Mbuti baseline, or when compared to a different Yoruban ($|Z| < 3$, Table S6). 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 ($|Z| < 3$, Table S7). In general, we find no differences in affinity to Neanderthals or Denisovans between the variants which fall in segments and the whole genome (Figure 4).

3 Discussion V2

This is a summary paragraph, I'm basing the structure off of the Akey paper
We developed a novel approach for estimating the Ancestral Recombination Graph (ARG) from whole genome sequence data. SMCSMC provides a framework for estimating demographic parameters such as effective population size and directional migration, with the ability to be extended to any parameter which may be simulated along the sequence. We use this method to investigate ancient migration, and find evidence for a substantial backfrom from Eurasia in the Late Middle Pleistocene. We study population substructure during this era, and identify differential rates of replacement in Central and South African populations.

This paragraph talks about the SOURCE of the migration
Applying SMCSMC to whole genome sequencing datasets revealed unexpected patterns of migration in the ancient past. We found that a population related to modern day Eurasians replaced upwards of half of the ancestral African population. There is no difference between French Europeans and Han Chinese migration, therefore the replacing population must have split from the OoA migrants before the East/West Eurasian divergence. This implies a lowerbound on the timing of the migration of approximately 40kya. We also see some evidence that the migrating population was more like French and Chinese populations than like Papuans. The most likely interpretation of this observation involves the sequential divergence from the OoA lineage of the Papuans, our replacing population, and the East/West Eurasian split, in that order. Another involves the dilution of "migrant-like" genetic diversity in Oceanian populations due to established high levels of admixture with archaic humans. However, this explaination is less convincing, as incorporating high (<20%) levels of archaic admixture was not sufficient to resolve distinct cross-coalescence curves indicating a migration involving the ancestors of Eurasians and Yorubans, but not Papuans or San [7].

This paragraph is about the SINK of the migration
We found that populations in Africa did not share the same migration history. While the ancestors of Niger-Kordofanian and Nilo-Saharan populations experienced generally consistent replacement, the ancestors of Central and South African Hunter Gatherer populations did not. We additionally believe that the history of the Afroasiatic populations has been confounded by extensive recent admixture from Eurasian populations during the Holocene. In CAHG and SAHGs, the amount of estimated replacement is approximately half of the other groups, between 25% and 35%. The Biaka show a higher level of replacement, though a likely explanation is the extensive documented admixture from Western African groups not shared with the Mbuti [6, 8]. The lower levels of replacement imply at least partial diversification of these populations at the time of the migration, placing an upper bound on the timing. The date of genetic diversification of both the KhoeSan and CAHGs is contested, though another candidate upper bound comes from our simulations, which indicate that if the migration occured more than 70kya, SMC2 would have little power to detect it.

This paragraph is about the Ne artifact
This migration has biased previous inference of the African population size. We show that incorporating directional migration resolves unexplained African genetic diversity approximately 100kya. As N_e is traditionally estimated as the scaled inverse coalescent rate, inference can be biased in the presence of population substructure and migration [9, 4]. We show with real data and simulation that backflow would delay coalescences, causing an increase in estimated N_e. Incorporating directional migration allows us to recover a more recent split time between the two populations.

This paragraph is about ARCHAICS
It has been previously reported that Eurasian backflow may be responsible for putatively introgressed Neanderthal material in the African genome; however, we find no evidence for Neanderthal-like enrichment in putatively Eurasian-derived segments of the genome [6]. There are several potential explanations for this observation. Firstly, the replacing population may have diversified from the OoA group prior to admixture from Neanderthals. However, this calls into question the proposed timeline with the Papuans, who are thought to have experienced the same introgression as Eurasians. Secondly, talk to Gerton about the Neanderthal results here becuase they show something different...

Historical recap and conclusion Here we present evidence for complex population structure during and after the Out of Africa migration. Evidence has been mounting for multiple migrations into the Eurasian continent, possibly mediated by climatic drivers [10, 6, 11]. Recent Eurasian backflow during the Holocene has been well established [12, 13], however earlier migrations have been proposed due to the spatial distribution of certain haplogroups [14, 15, 16, 17, 18, 19, 20, 6]. At the same time, evidence has been mounting for extreme heterogeneity in the history of sub-Saharan Africans, with several unsampled population theorised to contribute at various points in the past [21]. Inference of the coalescent intensity function shows distinct Eurasian histories in sub-Saharan African groups, supporting both an early split between the groups and a substantial replacement of genetic material between 50-70kya [22]. In light of these recent studies, the observations in this paper echo a body of evidence for complex population structure and migration surrounding the Out of Africa event with a substantial replacement of the African population in the Late Middle Paleolithic.

4 Discussion

As a geographically complex region with the longest history of continuous AMH occupation, patterns of genetic variation within the African continent contain a wealth of information about evolutionary history. Here, we analyse individuals from two global datasets and study ancient directional migration. Our analysis suggests that the ancestral African population was structured, with at least two distinct groups recieving substantially different amounts of migrants from Eurasia between 50 and 70kya. The San peoples, who are known to have branched from the ancestral population prior to the split of Out of Africa populations, received a small amount of migration from Eurasian sources, while the remainder of populations show a high degree of migrations from an unsampled population closer to modern day Han Chinese and French individuals than to modern day Papuans.

Analysing the effective population size of the African population with MSMC shows a temporarily deflated coalescent rate in African populations immediately prior (as in, more ancient) to the inferred migration event. This same inference was made by the Pairwise Sequentially Markovian Coalescent (PSMC) model, where Li and Durbin theorised that this "bump" in the African N_e was not a historical event but rather a consequence of substructure within the ancestral African population. While population substructure may play a role, we expect that delaying coalescences within a population would not result in a directional increase in cross-population coalescences such as we observe with `smcsmc`. We are able to replicate this temporary N_e inflation in a single population analysis, though when we simultaneously fit parameters for directional migration over time, the artifact is resolved in all populations. This effect is not driven by errors in statistical phasing, as analysis in a physically phased subset of the HGDP behaves in the same manner. Several simulated demographic scenarios involving directional migration also support this conclusion. From this we conclude that migration in the history of African populations may cause the inverse instantaneous coalescent rate (IICR) to deviate from the true population size, a known result in structured populations [9].

Languages in Africa may be broadly classified, with the exception of the Austronesian group in Madagascar, into four families which broadly align with the phylogeny of the continent [23, 24]. Niger-Kordofanian and Nilo-Saharan groups show the highest levels of migration from this "ghost" population, between 50 and 60 percent replacement, while we strongly believe that inference in Afroasiatic populations such as the Mozabite may be confounded by established recent admixture in the Holocene [25]. The Khoesan, a representative of South African Hunter Gatherers (SAHG), show the lowest levels of inferred migration, amounting to approximately 30% replacement over the last 100ky. Though the specifics of population divergences are contested, recent evidence strongly supports an early divergence and distinct geographical distribution of SAHGs with limited interaction with the majority of other African groups [26, 27, 8, 28, 29]. The Mbuti, and to a lesser degree the Biaka, show the lowest non-Khoesan rates. Recent evidence places the divergence times of Central African Hunter Gatherers between 200-250kya; though yet to be resolved, this would be consistent with our results [21]. In this study, the Biaka were found to have a much higher Bantu-related component than the Mbuti, consistent with previous evidence of Bantu gene flow to Eastern CAHG, but not Western [30]. Interestingly, the next lowest migration magnitude is found in Bantu speakers in Botswana, known to have high levels of gene flow from the Khoesan [31]. Simulations show that `smcsmc` has weak power to detect the actual magnitude of ancient migrations, so we suggest that care is taken in the interpretation of these values. An admixture event into the common ancestor of non-Khoesan, non-CAHG groups with some historical gene flow would be a parsimonious explanation for the observed migration.

For several decades, evidence for an ancient directional migration "back to Africa" has been mounting. A recent pulse of admixture c.3kya has been well supported by ancient DNA [12, 13]. However, a much earlier migration during the Paleolithic period has been suggested to explain the spatial distribution of several haplogroups, such as E-M96, L3, M1, U6*, and DE*, proposed to have a Eurasian origin and subsequent back-flow variously dated to between 40 and 75kya [14, 15, 16, 17, 18, 19, 20]. A separate body of evidence supports a highly divergent lineage contributing to the ancestors of Western Africans [21, 32, 1]. Evidence also comes from studies on whole genome sequencing data, where a clean split between African and OoA groups is inconsistent with MSMC X-Coal curves [3], and supports gene flow between the ancestors of Niger-Kordofanian groups and Eurasians, but not Papuans [7]. We find lower levels of overall migration in the Papuans than Han Chinese and French, broadly supporting this assertion. In addition, we use the posterior ancestral recombination graph (ARG) generated by the `smcsmc` particle filter to isolate segments of the African genome with a migration event in this period. These segments do not show more drift with Neanderthals or Denisovans, consistent either with a scenario where the admixing population either has not yet seen the introgression event from Neanderthals, or where this material was donated but subsequently selected against. It is currently unclear whether the inferred migration events may be confounded by super-archaic introgression [33, 34] from populations not represented in the fossil record.

With the relative scarcity of ancient DNA in Africa, it is difficult to explain the historical situation leading to a directional migration. Because we know that the ghost population is more genetically similar to modern day Eurasians than modern day Africans, we can be fairly certain that it was not the result of a failed earlier migration OoA as has been previously suggested. Instead we suggest that the source may be related to a group splintering off of the main OoA population (known as a "ghost" population in [7]) after its divergence with the group who would eventually inhabit Sahul. Periodic changes in the climate created a vegetated migration path between Eurasia and Africa, providing the requisite periods of isolation and genetic differentiation followed by migration [10]. A more parsimonious explanation for the ghost population from [7] may be a retreat back to Africa during one of these green periods, causing an effective directional migration of Eurasian-like genetic material into the African gene pool.

Our results imply that a portion of sub-Saharan African ancestry derives from a group related to the main OoA migration. We identify directional migration in two large databases, extensively verify its plausibility through simulation, and improve the estimation of population divergence times by explaining artificially low intra-population coalescence rates. This work calls for a more nuanced investigation of gene flow in the ancient past, and reinforces the importance of fully parameterizing demographic inference models. Additional genome sequences from unsampled groups, both ancient and modern, may help to resolve the circumstances leading to a migration Back to Africa.

5 Methods

A Particle Filter for Demographic Inference Details of the Sequential Monte Carlo for the Sequentially Markovian Coalescent (`smcsmc`) algorithm have been previously published [35] (see the URLs for an implementation). Briefly, `smcsmc` builds an approximation of the posterior distribution of genealogical trees along the genome using sequential Monte Carlo, also known as a particle filter. It does so by simulating a number of sequences of genealogical trees (particles) under a fixed set of demographic parameters θ. 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 an approximate sample of trees from the posterior, Variational Bayes (VB) or Stochastic Expectation Maximization (SEM) is used to update the estimates of demographic parameters θ. This is repeated over a given number of iterations, or until the demographic parameters θ have converged.

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 3×10^{-9} and a neutral mutation rate of 1.25×10^{-8}, both in units of events per nucleotide per generation; previous results indicate that modeling recombination hotspots minimally affects results [4]. 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).

Coalescent Simulation Coalescent simulations were performed under the sequential coalescent with recombination model (SCRM) [36]. 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 CwR, while for a zero-length window it is equivalent to the sequentially Markovian coalescent (SMC') [37, 38]; we use a 100 kilobase (kb) sliding window to approximate the CwR and improve accuracy over SMC' while retaining tractable inference. We simulated a model of exponentially decreasing migration into Eurasia from Africa between 200 and 100 thousand years ago (kya).

We modelled migration back to Africa as an epoch with a constant migration rate, parameterized by the total proportion migrated (0.4, 0.55, and 0.7), the total duration of the migration event (10000, 30000, or 50000 years), and the midpoint of the migration (50kya, 60kya, or 70kya). The proportions simulated are unrealistically high, in order to obtain estimates of migration that resemble observations from real data. We then used `smcsmc` to infer the demographic parameters using 10000 particles and 30 iterations of the VB procedure. We started inference at a reasonable approximation of human demographic history to aid in convergence, and set 31 logarithmically spaced epochs from 133 to 133016 generations in the past. We use a generation time of 29 years throughout For computational reasons, individual genomes were split into 120 chunks and processed in parallel, with sufficient statistics collected and processed together in the VB steps.

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 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 event that alter a tree in such a way that the migration event is retained.

Expectation of tract length 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 [38, 39]. 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 $(rT)^{-1}$ [40]; we use this approximation to derive expected distribution of fragment sizes. When analysing populations with `smcsmc` , we fix the recombination rate at 3×10^{-9} uniformly across the genome, in line with that used by MSMC in simulations [41, Supp. section 7]. 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 1×10^{-8}, equivalent to a one percent chance of a cross-over per megabase and per generation [42].

Sequence Data and Preparation We download whole genome sequence (WGS) data from the phased release of the Simons Genome Diversity Panel and convert it to seg file format using a utility provided by the `smcsmc` software implementation (`smcsmc.vcf_to_seg`). We apply two masks to the data. Firstly, we mask the data with the strict accesibility mask provided by the 1000 genomes project (see URLs). Secondly, we mask any sites absent chimpanzee ancestry, due to a known issue in calling which resulted in artificially long runs of homozygosity. We develop a `snakemake` pipeline for efficiently analysing sequence data with both `smcsmc` and MSMC, available at the project's github page (see URLs). We assume a mutation rate of 1.25×10^{-8} and a recombination rate of 3×10^{-9}, in line with recent literature. Two parameters must be set on a run-by-run basis. As the inference portion of `smcsmc` uses a variational Bayesian approach, a number of maximum epochs must be set. Additionally, the number of particles to be used must be specified. 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_Yoruba-1` in the SGDP nomenclature).

Formal Statistics Patterson's formal statistics were calculated with ADMIXTOOLS [31] and the `admixr` package [43] 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`.

Integrating Migration Rates over Time We wish to find the total fraction of a particular population replaced during a particular time period. We track the probability that a particular individual has not migrated after time T (in generations). Let $\rho(t)$ be the instantaneous rate of migration per unit of time. In this formulation, $\frac{d}{dt}F(t) = -\rho(t)F(t)$, whose solution is $F(t) = e^{-\int_{t=0}^{T}\rho(t)dt}$; this gives a total migration probability in a range of epochs E of $1 - F(T) = 1 - e^{\sum_{i \in E} r_i}$.

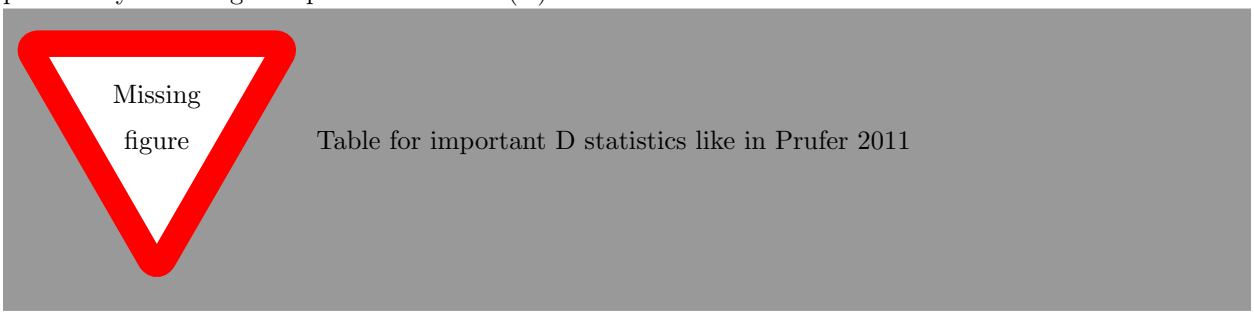


Table for important D statistics like in Prufer 2011

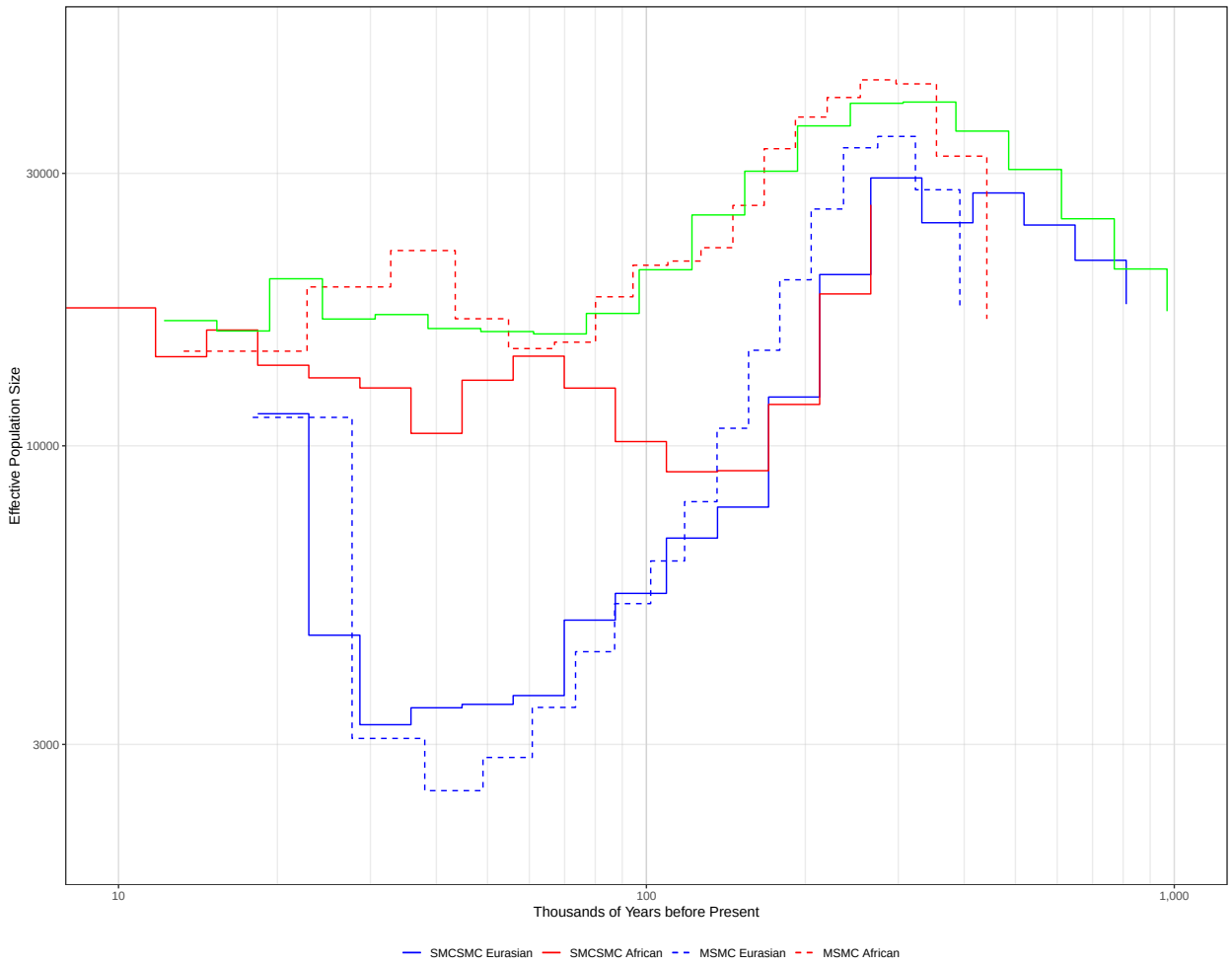


Figure 1: Inferred effective population size using `smcsmc` and MSMC from real and simulated whole genome sequence data. A. Inferred N_e of Han Chinese and Yoruban individuals from the SGDP in both MSMC and `smcsmc`. B. N_e of Han Chinese and Yoruban individuals from the physically phased subset of the HGDP. C, D, and E simulate demographic histories with a 33%, 42.5%, and 50% total replacement 60kya, over the span of 10ky, in `scrm` and reinfer the N_e with `smcsmc`. We replicate a PSMC-like analysis by analysing one diploid Yoruban with `smcsmc` (PSMC^2).

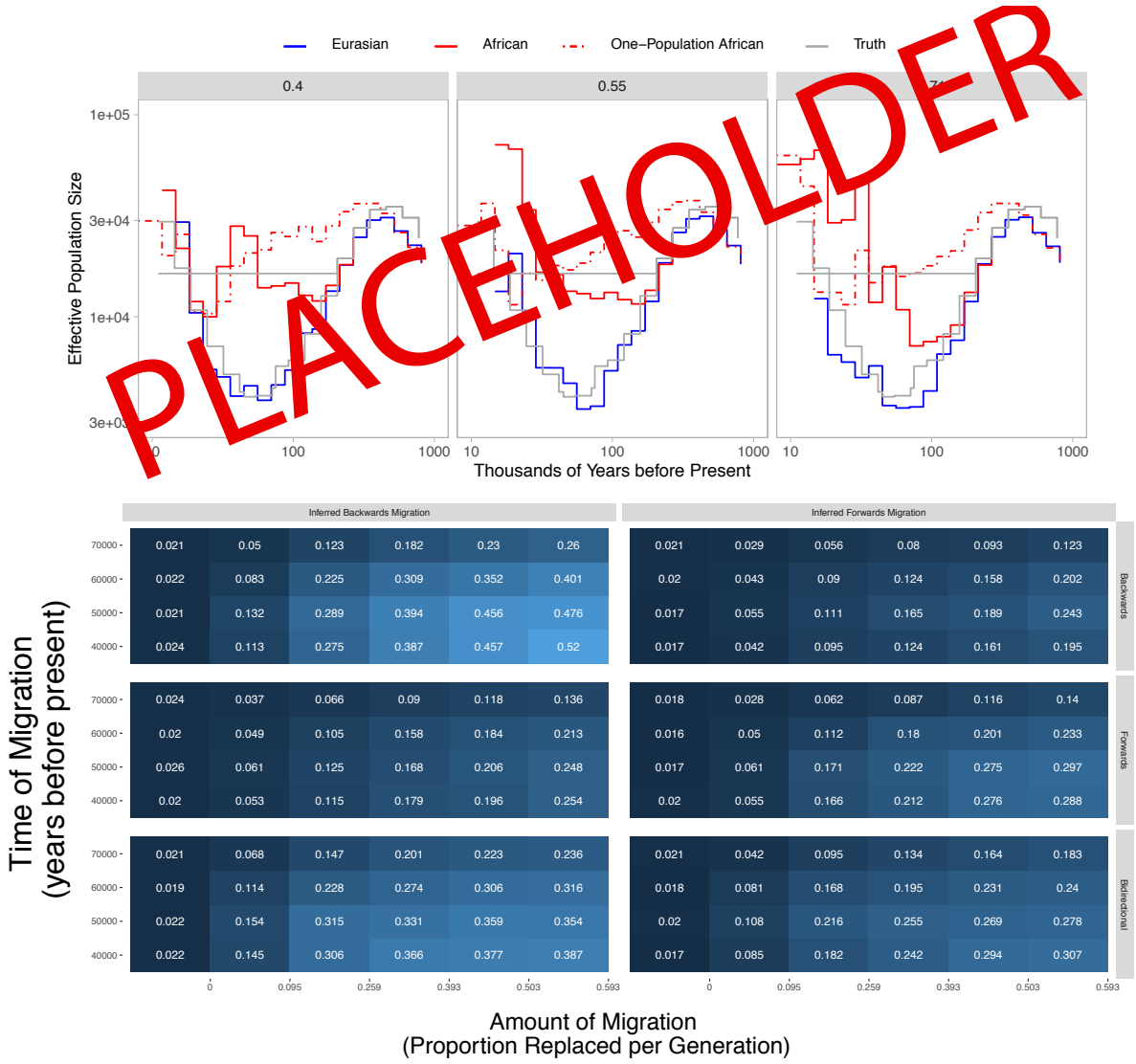


Figure 2: Simulation study concludes that `smcsmc` has power to detect a back-migration. Top: One gigabase of sequence was simulated by `scrm` under models with increasing migration from an "African" population to a "Eurasian" population. The magnitude of the population replaced by the migration is indicated in the grey ribbon. In Figure S we additionally vary the timing and duration of the migration, which for this figure are fixed at 60kya and 10ky respectively. Dotted line indicates a single diploid "African" genome which was analysed in a one-population model. Grey indicates the true population size simulated. Bottom: Under three different scenarios (migration "backwards" from Eurasia to Africa, migration "forwards" from Africa to Eurasia, and "bidirectional" migration, all forward in time), `scrm` was used to simulate one gigabase of sequence and 5 iterations of `smcsmc` with 5000 particles were used to infer migration. The axis of each heatmap shows increasing simulated proportion replaced over 100,000 while the Y shows the midpoint of the migration simulated. The full migration and population size inference over time along with a more substantial set of simulations is available in Supplemental Section S3.

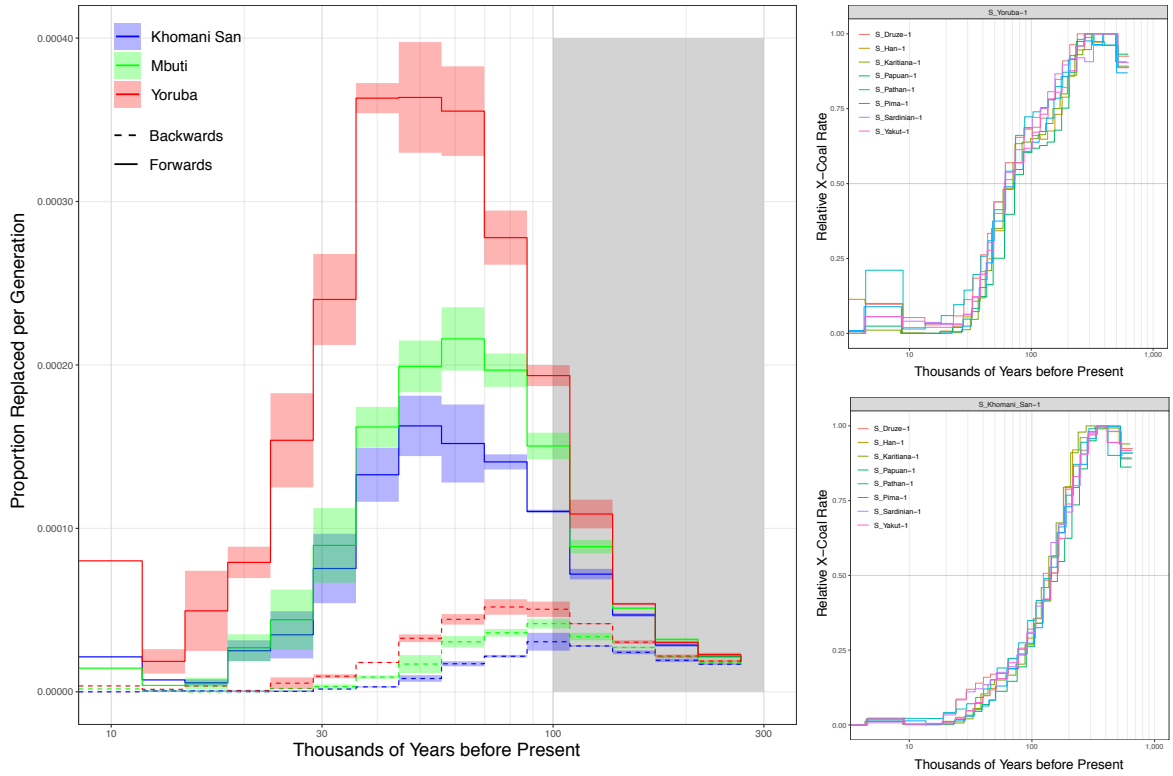


Figure 3: Migration estimates from the Simons Genome Diversity Panel. A. shows three replicates of `smcsmc` inferred migration to with a Han Chinese individual in three contrasting populations representing Niger-Kordofanian and San populations, with the Mbuti a unique intermediate. Analysis used 10000 particles and 25 iterations of variation Bayesian inference to converge. Trend line represents the mean of the replications, while the shaded regions denote the standard deviation. B. and C. show relative cross-coalescent (x-coal) curves for the Yoruban and San individuals modelled with a panel of Out of Africa OoA populations showing contrasting migration histories.

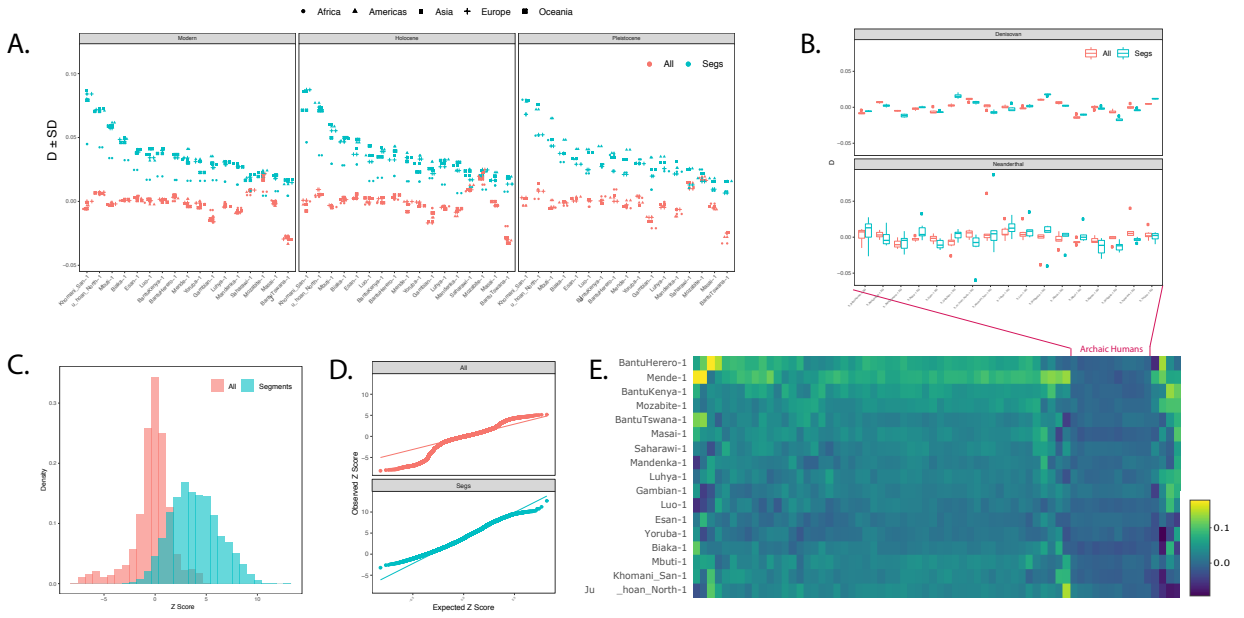


Figure 4: Analysis of $D(A, A_p, Y, Chimp)$ statistics for all analysed African populations A paired with a partner from the same population A_p and analysed for gene flow with a global population Y from the Reich Lab genotype database. A. compares average D statistic by continent in different African populations, calculated for both all available markers and for the portion of the genome whose tree contains a back-migration event. This is stratified by the age of the comparison sample, either Modern (Present - 1kya), Holocene (1kya - 11.7kya), or Pleistocene ($>$ 11.7kya). B. displays the distribution of D statistics for both sections, while C. gives a QQ plot of their Z statistics. D. shows a heatmap of D statistics for all of the Pleistocene samples, with E. being a blow-out of the Denisovan and Neanderthal statistics.

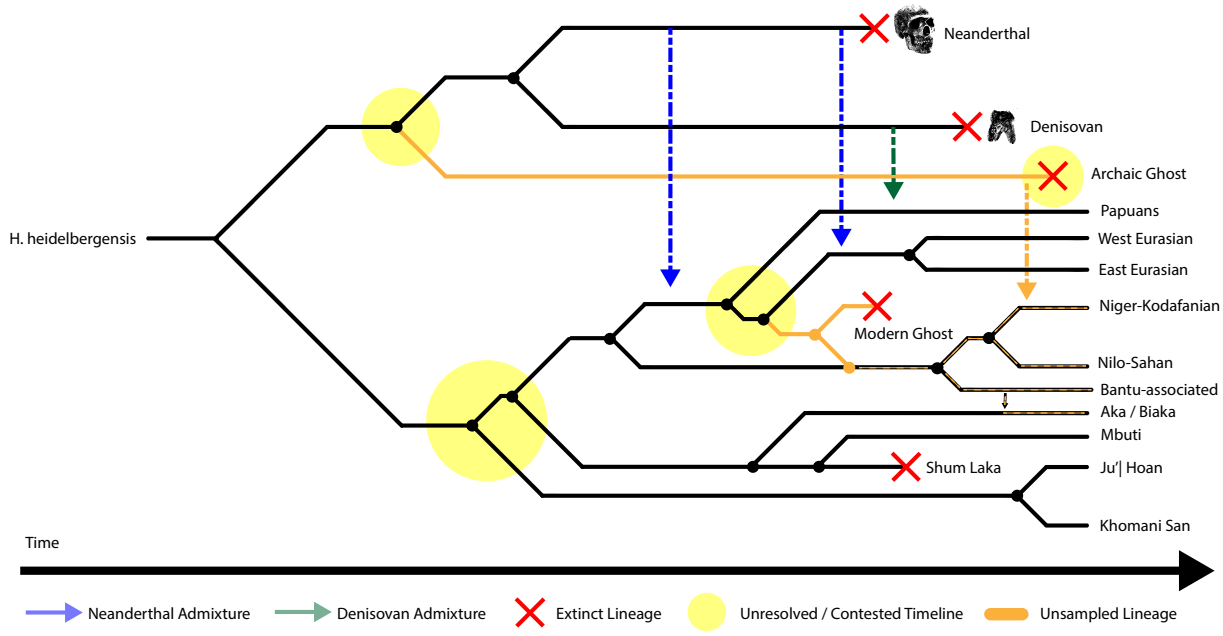


Figure 5: Proposed Demographic Model. Population splits coloured in yellow are contested. The existance of an archaic ghost lineage which has contributed to Western African populations has been broadly supported in the literature, but the time and order of divergences relative to Neanderthals and Denisovans remains an open question. Until recently, the San peoples were considered to be the most anciently diverged group in Africa, though recent evidence places Central African Hunter Gatherers on a similar timespan, with the addition of a modern ghost population. This existance of this population is additionally supported in the literature and in this article, though the order of divergence is contested. In this article we posit that the ghost diverged from the common ancestor of Eurasians after Papuans had diverged, similar to that suggested in [7].

References

[1] Leo Speidel et al. "A method for genome-wide genealogy estimation for thousands of samples". In: *Nature Genetics* 51.9 (2019), pp. 1321–1329. ISSN: 15461718. DOI: 10.1038/s41588-019-0484-x. URL: http://dx.doi.org/10.1038/s41588-019-0484-x.

[2] Jerome Kelleher et al. "Inferring whole-genome histories in large population datasets". In: *Nature Genetics* 51.9 (2019), pp. 1330–1338. ISSN: 15461718. DOI: 10.1038/s41588-019-0483-y. URL: http://dx.doi.org/10.1038/s41588-019-0483-y.

[3] 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. ISSN: 15461718. DOI: 10.1038/ng.3015. arXiv: 005348 [10.1101].

[4] Heng Li and Richard Durbin. "Inference of human population history from individual whole-genome sequences". In: *Nature* 475.7357 (2011), pp. 493–496. ISSN: 00280836. DOI: 10.1038/nature10231. arXiv: arXiv:1011.1669v3. URL: http://dx.doi.org/10.1038/nature10231.

[5] 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. ISSN: 10959203. DOI: 10.1126/science.aal1988.

[6] 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. ISSN: 0092-8674. DOI: 10.1016/j.cell.2020.01.012. URL: https://doi.org/10.1016/j.cell.2020.01.012.

[7] Anna Sapfo Malaspinas et al. "A genomic history of Aboriginal Australia". In: *Nature* 538.7624 (2016), pp. 207–214. ISSN: 14764687. DOI: 10.1038/nature18299. arXiv: NIHMS150003. URL: http://dx.doi.org/10.1038/nature18299.

[8] 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. ISSN: 07374038. DOI: 10.1093/molbev/msq294.

[9] 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. ISSN: 1365-2540. DOI: 10.1038/s41437-017-0005-6. URL: https://doi.org/10.1038/s41437-017-0005-6.

[10] Axel Timmermann and Tobias Friedrich. "Late Pleistocene climate drivers of early human migration". In: *Nature* 538.7623 (2016), pp. 92–95. ISSN: 14764687. DOI: 10.1038/nature19365. URL: http://dx.doi.org/10.1038/nature19365.

[11] Luca Pagani et al. "Genomic analyses inform on migration events during the peopling of Eurasia". In: *Nature* 538.7624 (2016), pp. 238–242. ISSN: 14764687. DOI: 10.1038/nature19792.

[12] 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. ISSN: 1176-9343. DOI: 10.4137/EBO.S33489. URL: http://www.ncbi.nlm.nih.gov/pubmed/27127403%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4844272.

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

[14] 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 (Aug. 1997), pp. 462–6. ISSN: 0002-9297. DOI: 10.1016/S0002-9297(07)64077-4. URL: http://www.ncbi.nlm.nih.gov/pubmed/9311756%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1715891.

[15] 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 (Apr. 1998), pp. 427–441. ISSN: 0737-4038. DOI: 10.1093/oxfordjournals.molbev.a025939. URL: https://academic.oup.com/mbe/article-lookup/doi/10.1093/oxfordjournals.molbev.a025939.

[16] 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. ISSN: 00029297. DOI: 10.1086/340257.

[17] A. Chandrasekar et al. "YAP insertion signature in South Asia". In: *Annals of Human Biology* 34.5 (Jan. 2007), pp. 582–586. ISSN: 0301-4460. DOI: 10.1080/03014460701556262. URL: http://www.tandfonline.com/doi/full/10.1080/03014460701556262.

[18] 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 (Dec. 2018), p. 98. ISSN: 1471-2148. DOI: 10.1186/s12862-018-1211-4. URL: https://bmcevolbiol.biomedcentral.com/articles/10.1186/s12862-018-1211-4.

[19] 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 (May 2016), p. 25501. URL: https://doi.org/10.1038/srep25501%20http://10.0.4.14/srep25501%20https://www.nature.com/articles/srep25501%7B%5C#%7Dsupplementary-information.

[20] 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. ISSN: 0016-6731. DOI: 10.1534/genetics.119.302368. URL: http://www.genetics.org/lookup/doi/10.1534/genetics.119.302368.

[21] Mark Lipson et al. "Ancient West African foragers in the context of African population history". In: November 2018 (2019). DOI: 10.1038/s41586-020-1929-1.

[22] Patrick K. Albers and Gil McVean. "Dating genomic variants and shared ancestry in population-scale sequencing data". In: *bioRxiv* (2019), p. 416610. DOI: 10.1101/416610. URL: https://www.biorxiv.org/content/10.1101/416610v2.

[23] Roger Blench. *Archaeology, Language, and the African Past*. Tech. rep. 2007.

[24] Shaohua Fan et al. "African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations". In: *Genome Biology* (2019). ISSN: 1474760X. DOI: 10.1186/s13059-019-1679-2.

[25] George Bj Busby et al. "Admixture into and within sub-Saharan Africa". In: *eLife* (2016). ISSN: 2050084X. DOI: 10.7554/eLife.15266.

[26] Laurent Excoffier et al. "Robust Demographic Inference from Genomic and SNP Data". In: *PLoS Genetics* (2013). ISSN: 15537390. DOI: 10.1371/journal.pgen.1003905.

[27] Doron M. Behar et al. "The Dawn of Human Matrilineal Diversity". In: *American Journal of Human Genetics* (2008). ISSN: 00029297. DOI: 10.1016/j.ajhg.2008.04.002.

[28] Wentao Shi et al. "A worldwide survey of human male demographic history based on Y-SNP and Y-STR data from the HGDP-CEPH populations". In: *Molecular Biology and Evolution* (2010). ISSN: 07374038. DOI: 10.1093/molbev/msp243.

[29] Carina M. Schlebusch et al. "Genomic Variation in Seven Khoe-San". In: 1187.October (2012), pp. 374–379. DOI: 10.1126/science.1227721.

[30] Lluís Quintana-Murci et al. "Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers". In: *Proceedings of the National Academy of Sciences of the United States of America* (2008). ISSN: 10916490. DOI: 10.1073/pnas.0711467105.

[31] Nick Patterson et al. "Ancient Admixture in Human History". In: 192.November (2012), pp. 1065–1093. DOI: 10.1534/genetics.112.145037.

[32] Pontus Skoglund et al. "Reconstructing Prehistoric African Population Structure". In: *Cell* 171.1 (2017), 59–71.e21. ISSN: 10974172. DOI: 10.1016/j.cell.2017.08.049.

[33] Arun Durvasula and Sriram Sankararaman. "Recovering signals of ghost archaic introgression in African populations". In: *bioRxiv* (2019), p. 285734. DOI: 10.1101/285734. URL: http://biorxiv.org/content/early/2019/02/28/285734.abstract.

[34] Joseph Lachance et al. "Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers". In: *Cell* 150.3 (2012), pp. 457–469. ISSN: 00928674. DOI: 10.1016/j.cell.2012.07.009.

[35] Donna Henderson, Sha (Joe) Zhu, and Gerton Lunter. "Demographic inference using particle filters for continuous Markov jump processes". In: *bioRxiv* (Aug. 2018), p. 382218. DOI: 10.1101/382218. URL: https://www.biorxiv.org/content/early/2018/08/01/382218.

[36] Paul R. Staab et al. "scrm: efficiently simulating long sequences using the approximated coalescent with recombination". In: *Bioinformatics* 31.10 (May 2015), pp. 1680–1682. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/btu861. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu861.

[37] Gilean A.T. McVean and Niall J. Cardin. "Approximating the coalescent with recombination". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* (2005). ISSN: 09628436. DOI: 10.1098/rstb.2005.1673.

[38] Paul Marjoram and Jeff D Wall. "Fast "coalescent" simulation". In: *BMC Genetics* 7.1 (Mar. 2006), p. 16. ISSN: 1471-2156. DOI: 10.1186/1471-2156-7-16. URL: https://doi.org/10.1186/1471-2156-7-16.

[39] Mason Liang and Rasmus Nielsen. "The Lengths of Admixture Tracts". In: *Genetics* 197.3 (2014), pp. 953–967. ISSN: 0016-6731. DOI: 10.1534/genetics.114.162362. URL: http://www.genetics.org/content/197/3/953.

[40] Fernando Racimo et al. "Evidence for archaic adaptive introgression in humans". In: *Nature Reviews Genetics* 16 (May 2015), p. 359. URL: http://dx.doi.org/10.1038/nrg3936%20http://10.0.4.14/nrg3936%20https://www.nature.com/articles/nrg3936%7B%5C#%7Dsupplementary-information.

[41] 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. ISSN: 15461718. DOI: 10.1038/ng.3015. arXiv: 005348 [10.1101]. URL: http://dx.doi.org/10.1038/ng.3015.

[42] Beth L Dumont and Bret A Payseur. "EVOLUTION OF THE GENOMIC RATE OF RECOMBINATION IN MAMMALS". In: *Evolution* 62.2 (Feb. 2008), pp. 276–294. ISSN: 0014-3820. DOI: 10.1111/j.1558-5646.2007.00278.x. URL: https://doi.org/10.1111/j.1558-5646.2007.00278.x.

[43] Martin Petr, Benjamin Vernot, and Janet Kelso. " admixr —R package for reproducible analyses using ADMIXTOOLS ". In: *Bioinformatics* January (2019), pp. 1–2. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz030.

S1 Details of Data Analysis

S1.1 Inferring population size and migration rates in the Simons Genome Diversity Panel

This section describes analysis of the Simons Genome Diversity Panel 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/v1.0.1`), MSMC version 1.1.0 was installed from Github (found at `https://github.com/stschiff/msmc/releases/tag/v1.1.0`) 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. We do this masking in `vcftools`. We use `smcsmc` 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_Yoruba-1` in the SGDP nomenclature). 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 mutation rate of 1.25×10^{-8}, a recombination rate of 3×10^{-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 lookahead 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.

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
```

We visualise this demographic model in the POPdemog package in Figure S1.

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 epoch 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, while times are given in generations.

We implement the above in an open source `Snakemake` pipeline at `SNAKEM` which also implements a default analysis of MSMC with forty iterations to converge. Sample size and relative cross-coalescent rates are transformed as described in the documentation using the same parameter values for mutation rate

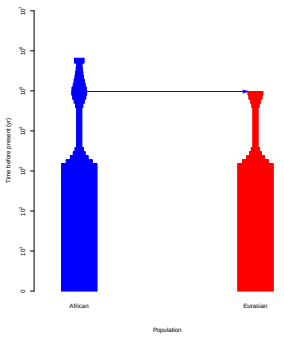


Figure S1: Demographic model used as a seed for SMC2 analysis

and generation time used for `smcsmc` analysis. Effective population size and migration estimates for the populations analysed in the SGDP are given in Figures S2 and S3. MSMC appears to consistently find a higher African N_e in the ancient past until the average estimates across populations stabalises approximately 100kya (Figure S4). We expand on a possible rational behind this effect in the main text of this article.

Migration during the last 100ky is integrated to observe overall trends (Figure S5). We use two methods to integrate migration, the first presented in the main text given by $F(t) = e^{-\int_{t=0}^{T} \rho(t)dt}$. Alternatively, 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. Inferred migration varies across language groups (Figure S6. Afroasiatic groups show high migration from Han and French populations, with a lower proportion deriving from Papuans. Niger-Kordofanian and Nilo-Saharan groups show an intermediate magnitude, between 50 and 60 percent replacement, though also significantly (P ¡ 0.05, two-tailed paired T test) closer to French and Han sources than Papuans. The Khoesan show the lowest migration, consistent with their early diversification from the remainder of African groups and the relative lack of gene-flow from Western African populations [21]. This is contrasted with the Mbuti and Biaka, Central African Hunter Gatherer populations who have historically recieved substantial amounts of gene flow from Western African sources. Both of these populations show the lowest migration in their language group (Table S1).

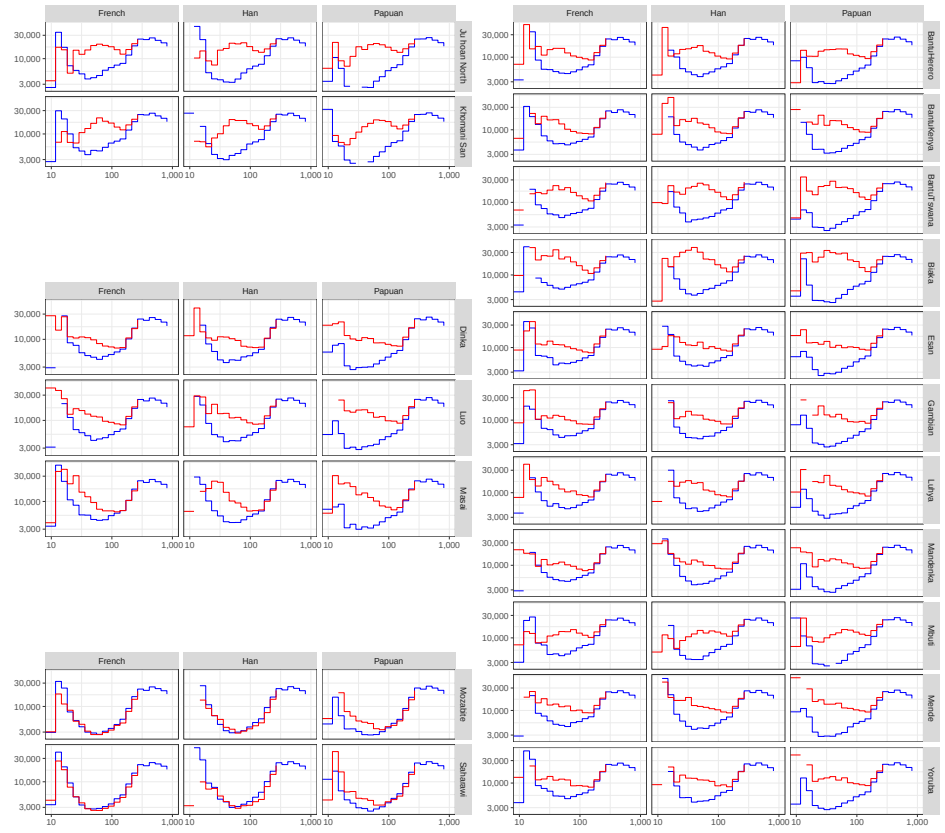


Figure S2: Estimated effective population size in different African and Eurasian groups with `smcsmc`. Left: From top to bottom, inferred N_e in Khoesan, Nilo-Saharan, and Afroasiatic populations. Right: Inferred N_e in Niger-Kordofanian populations. 5000 particles and 10 variational Bayes interations were used to achieve convergence.

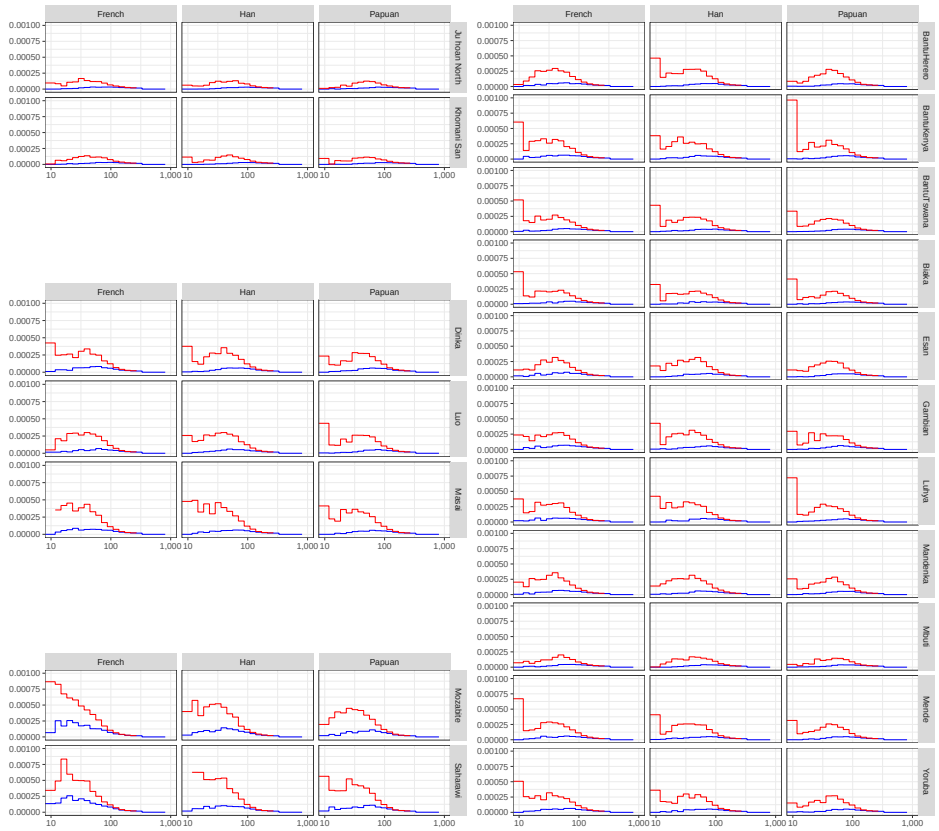


Figure S3: Estimated directional migration between African and Eurasian groups in the SGDP with `smcsmc`. Left: From top to bottom, inferred migration in Khoesan, Nilo-Saharan, and Afroasiatic populations. Right: Inferred migration in Niger-Kordofanian populations. 5000 particles and 10 variational Bayes interations were used to achieve convergence.

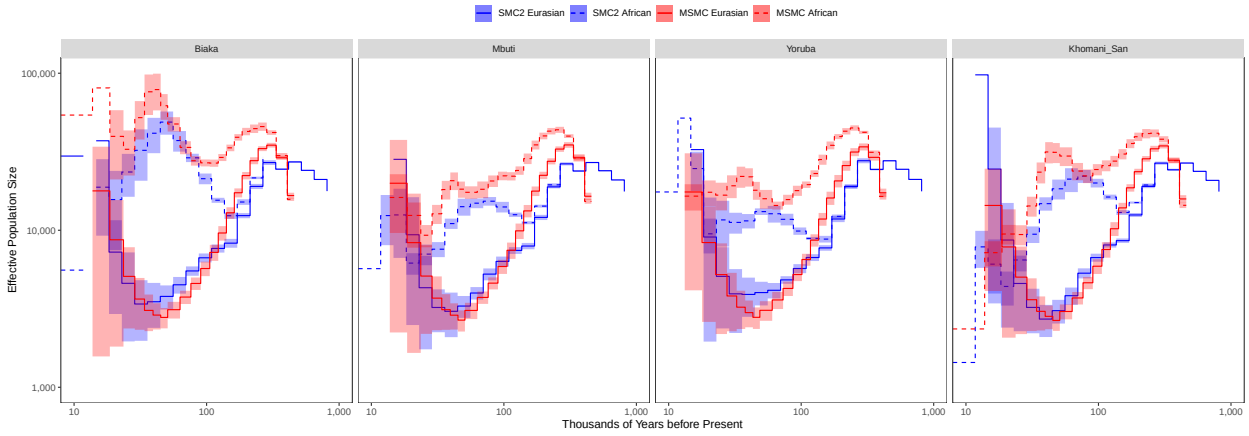


Figure S4: Average N_e estimate across four populations in the subset of SGDP used to compare with HGDP inference. Inference of population size is averaged over eight Eurasian populations, with the bars representing standard deviation. For MSMC, the time indexes were averaged to have consistent start and stop times for the steps.

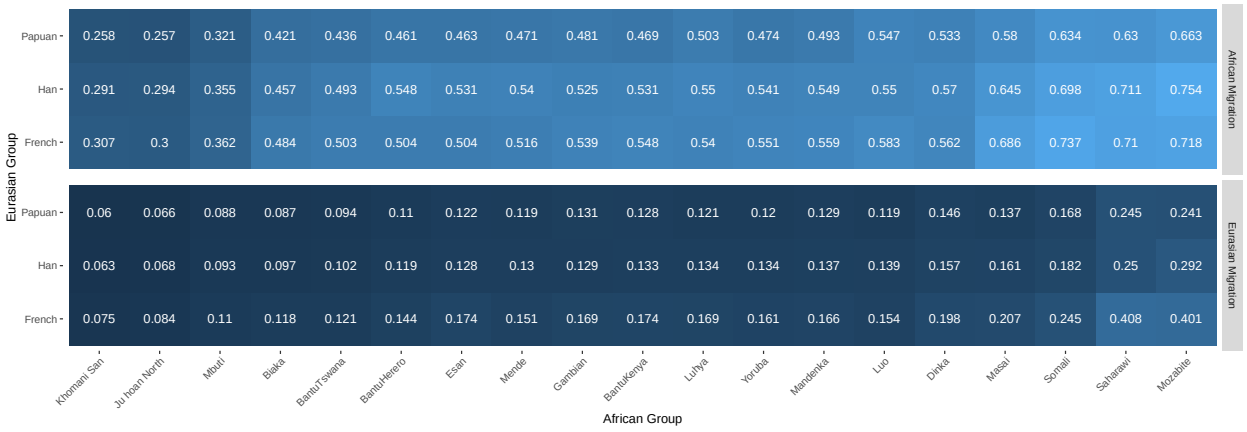


Figure S5: Integrated migration proportion in `smcsmc` analysed SGDP populations.

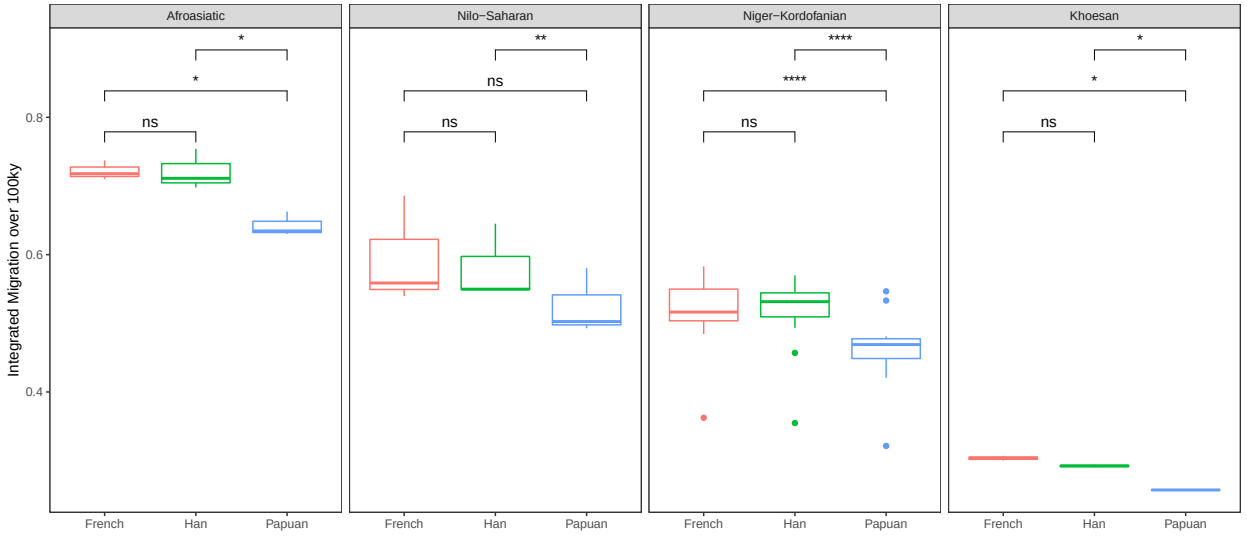


Figure S6: Integrated migration proportion over the last 100 thousand years (ky) between language families by comparison population. Papuans contributed significantly less to African populations across all populations in a two tailed paired T test. ns = Not Significant, * = $P < 0.05$, ** = $P < 0.01$, *** = $P < 0.001$, **** = $P < 0.0001$.

African	Eurasian	$M_{E,A}$ (SD)	$M_{A,E}$ (SD)
Afroasiatic	French	0.722(0.014)	0.351(0.092)
Afroasiatic	Han	0.721(0.029)	0.241(0.055)
Afroasiatic	Papuan	0.642(0.018)	0.218(0.043)
Khoesan	French	0.304(0.005)	0.079(0.006)
Khoesan	Han	0.292(0.002)	0.065(0.004)
Khoesan	Papuan	0.257(0.001)	0.063(0.004)
Niger-Kordofanian	French	0.514(0.059)	0.151(0.024)
Niger-Kordofanian	Han	0.513(0.061)	0.121(0.016)
Niger-Kordofanian	Papuan	0.462(0.059)	0.114(0.016)
Nilo-Saharan	French	0.595(0.079)	0.186(0.028)
Nilo-Saharan	Han	0.581(0.055)	0.152(0.012)
Nilo-Saharan	Papuan	0.525(0.048)	0.134(0.014)

Table S1: Average plus or minus standard deviation integrated directional migration from Eurasian to African populations in the last 100 thousand years (ky)

S1.2 Validation in a physically phased subset of the Human Genome Diversity Panel (HGDP)

For the `smcsmc` algorithm, the use of phased data is not necessary but does help convergence. It additionally helps the lookahead-likelihood to better guide the resampling procedure. Therefore, we do not expect errors made during statistical phasing to significantly impact the inferred parameters. However, to test this, we replicate the above analysis in a physically phased subset of the HGDP downlaoded from `ftp://ngs.sanger.ac.uk/production/hgdp/hgdp_wgs.20190516/`. The same `snakemake` pipeline is used as in the analysis of the SGDP data. Data is additionally masked for the filters provided. We plot the effective population size and migration rates in Figures BLAH and BLAH.

All of these (with the exception of one) are complete, but do not have MSMC runs, so there's no use in making all the figures twice.

S1.3 Patterns in population size and migration fully replicate in an equivalent subset of the SGDP

To directly compare these results to those obtained in the SGDP, we select the closest matching samples to those in the physically phased HGDP dataset and analyse these with MSMC and `smcmsc` using 10k particles and 25 iterations to achieve convergence (Figure S7). The effective sample size around the OoA migration is similarly inflated in MSMC analyses, while the estimation of the Eurasian population size remains largely consistent.

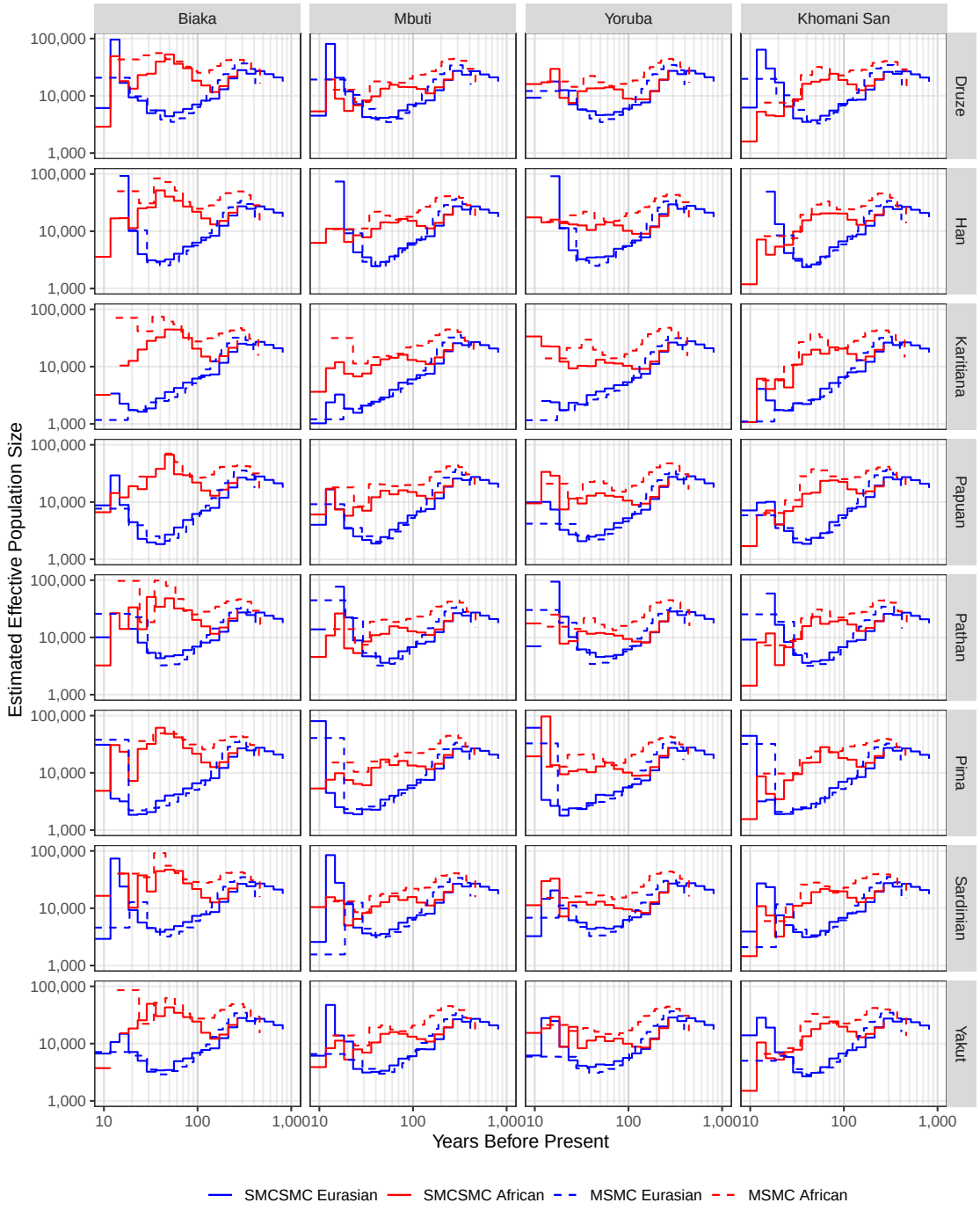


Figure S7: `smcsmc` and MSMC inferred effective population size of several populations in the Simons Genome Diversity Panel. These samples were selected to match, as closely as possible, those in the physically phased subet of the Human Genome Diversity Project panel. 10,000 particles and 25 iterations were used for `smcsmc` and 40 iterations for MSMC.

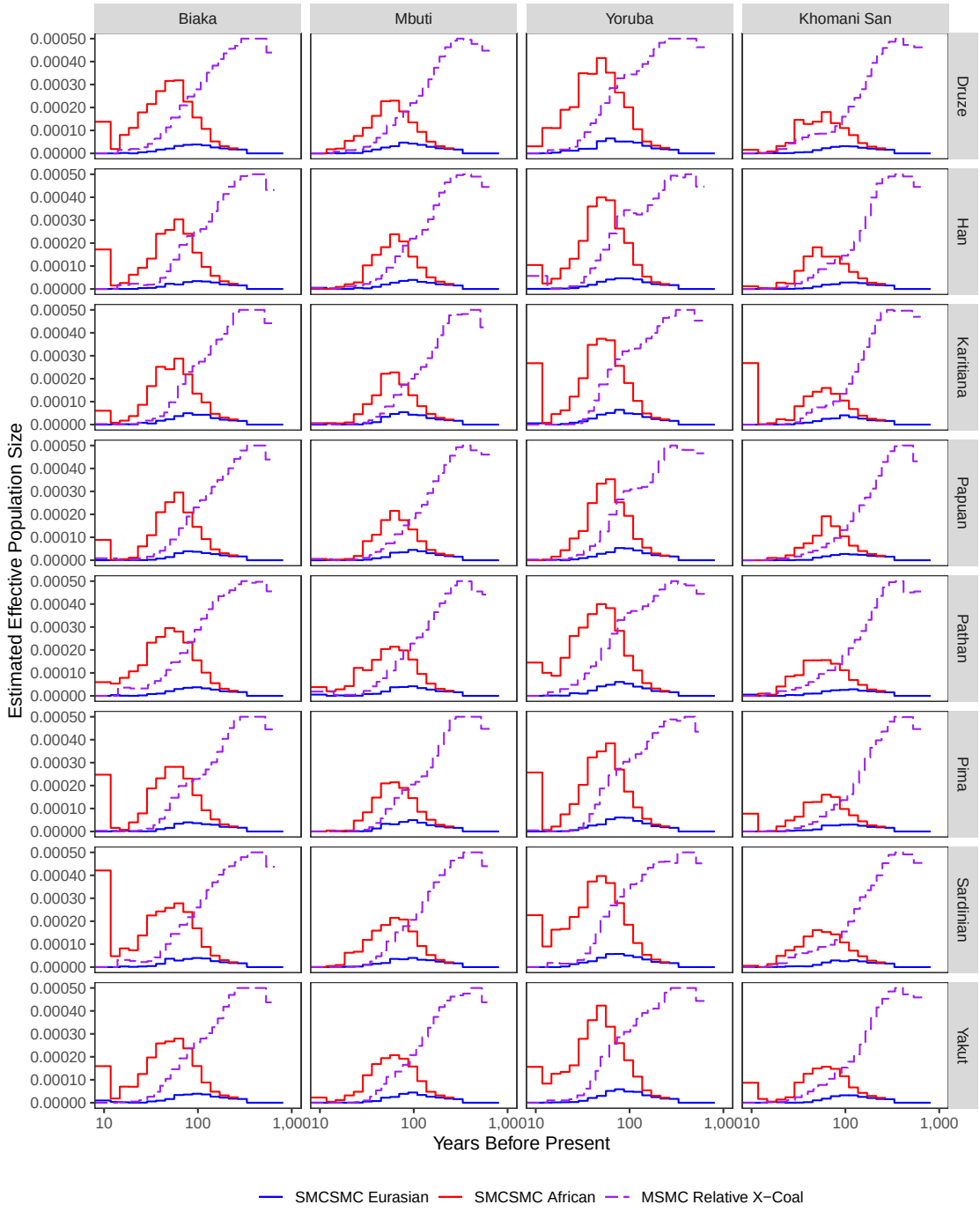


Figure S8: Inferred migration using `smcsmc` in the Simons Genome Diversity Panel along with the scaled relative cross-coalescent rate estimated by MSMC. Samples were chosen to match, as closely as possible, those in the physically phased subset of the Human Genome Diversity Project panel. 10,000 particles and 25 iterations were used, 40 in the case for MSMC.

S2 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. Here, Yoruba-1 is used as a representative of Western African groups, and used for ascertaining putatively migrated segments. The two Yorubans share more alleles than other groups in Africa (D(African group, Yoruba-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 S2). This implies that `smcsmc` has identified segments of the African Genome which are more closely related to OoA populations than to fellow Africans.

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 S2: Putatively migrated segments of a Yoruban are closer to Out of Africa groups than a comparable Yoruban.

Statistic	D	Z
D(KhomaniSan-1, Yoruba-1, French-1, Chimp)	-0.173	-25.685
D(KhomaniSan-1, Yoruba-1, Han-1, Chimp)	-0.208	-29.150
D(KhomaniSan-1, Yoruba-1, Papuan-1, Chimp)	-0.174	-24.085
D(KhomaniSan-1, Yoruba-2, French-1, Chimp)	-0.143	-19.523
D(KhomaniSan-1, Yoruba-2, Han-1, Chimp)	-0.161	-22.327
D(KhomaniSan-1, Yoruba-2, Papuan-1, Chimp)	-0.160	-21.258
D(Mbuti-1, Yoruba-1, French-1, Chimp)	-0.136	-19.835
D(Mbuti-1, Yoruba-1, Han-1, Chimp)	-0.167	-23.875
D(Mbuti-1, Yoruba-1, Papuan-1, Chimp)	-0.125	-17.088
D(Mbuti-1, Yoruba-2, French-1, Chimp)	-0.103	-14.514
D(Mbuti-1, Yoruba-2, Han-1, Chimp)	-0.116	-16.935
D(Mbuti-1, Yoruba-2, Papuan-1, Chimp)	-0.109	-14.459

Table S3: Both Yorubans share more alleles with OoA populations than San or Mbuti. The individual used to ascertain segments shares more alleles than a comparable individual.

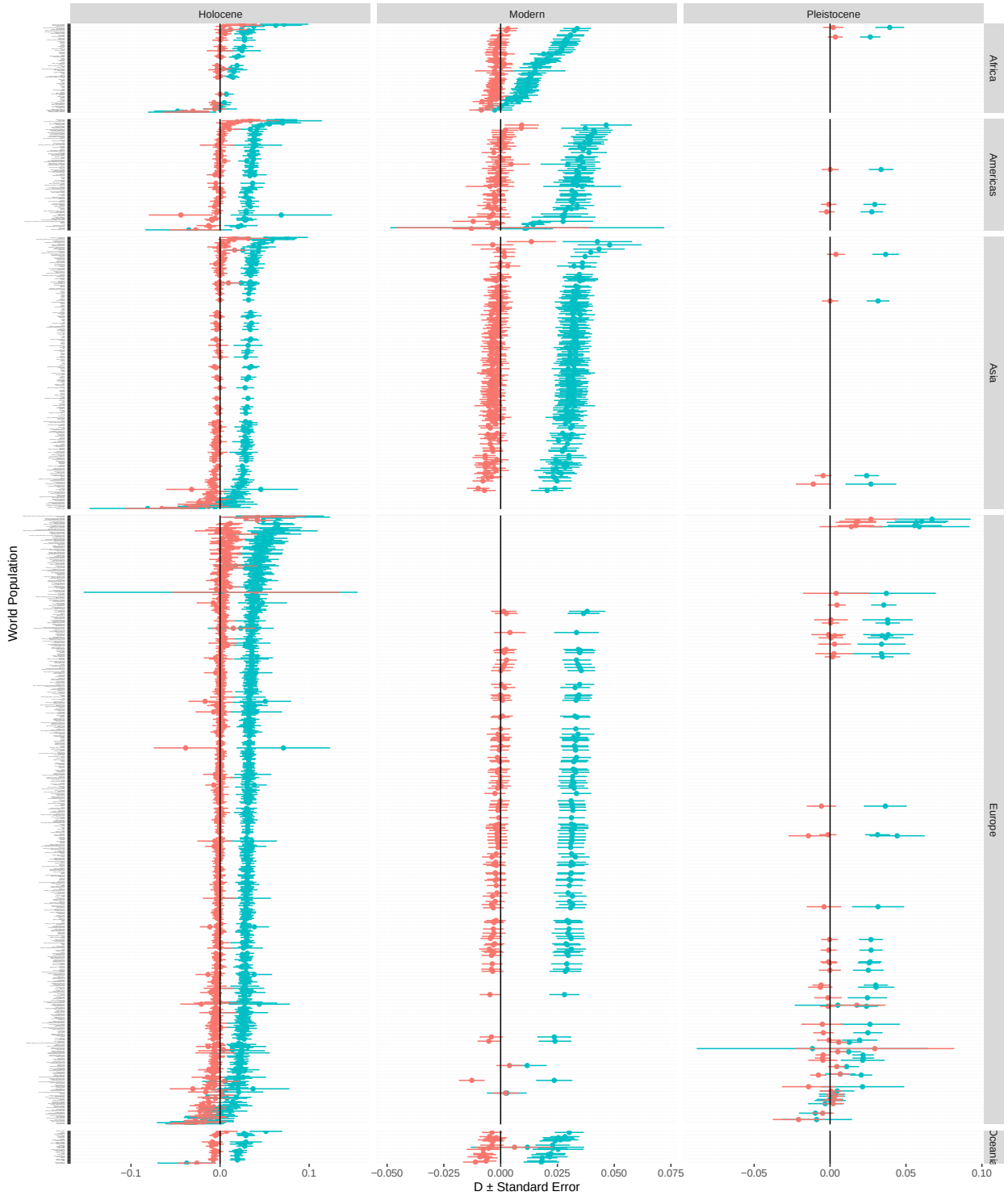


Figure S9: D statistics of the form (X, Chimp; Yoruba-1, Yoruba-2) for all global populations in the Human Origins dataset.

Statistic	D	Z
D(Yoruba-2, Yoruba-1, French-1, Chimp)	-0.036	-5.122
D(Yoruba-2, Yoruba-1, Han-1, Chimp)	-0.056	-7.888
D(Yoruba-2, Yoruba-1, Papuan-1, Chimp)	-0.018	-2.483

Table S4: The Yoruban used to ascertain segment is more closely related to OoA groups than a comparable Yoruban.

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

Table S5: No difference in allele sharing with Vindija Neanderthal.

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 S6: No difference in allele sharing with Vindija Neanderthal over Mbuti baseline.

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 S7: No increased affinity to Vindija Neanderthal over Altai, as would be expected if the source of any Neanderthal ancestry was Eurasian.

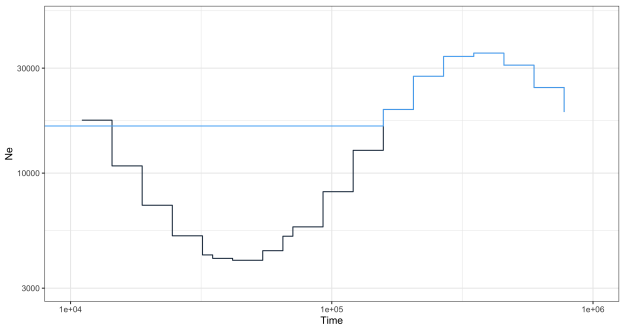


Figure S10: Population size model used for simulations.

S3 Simulation procedure

The ability of `smcsmc` to recover a back-migration signal is evaluated through simulation. One gigabase of sequence was simulated in `scrm`, and subsequently re-inferred by `smcsmc`. Migration is parameterised by three factors, magnitude, midpoint, and duration. A scenario is simulated where the midpoint is the center of a block of a given duration which has uniform migration which integrates to a given total proportion replacement over the period. We use the following demographic model for population size throughout all simulations:

The following commands can be used in either `ms` or `SCRM` to specify demographic models. The African population size is given by

```
-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
```

While the European population size is given by

```
-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
```

Times are in units of $4gN_0$ while population sizes are in units of N_0. For $g = 29, N_0 = 14312$, the demographic model is as shown in Supplemental Figure S10.

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.

We select a more comprehensive set of initiation parameters and particle values and use them to analyze a Yoruban and French individual from SGDP (Fig S15). 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.

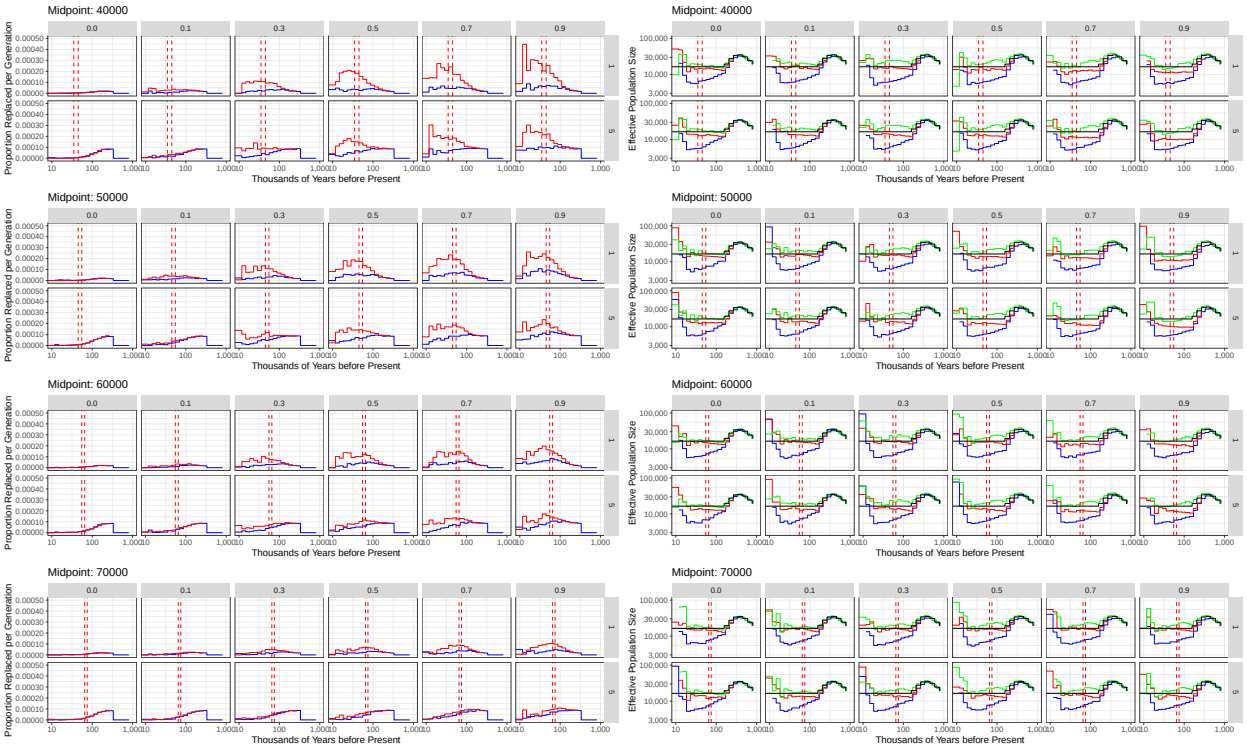


Figure S11: Backwards simulation

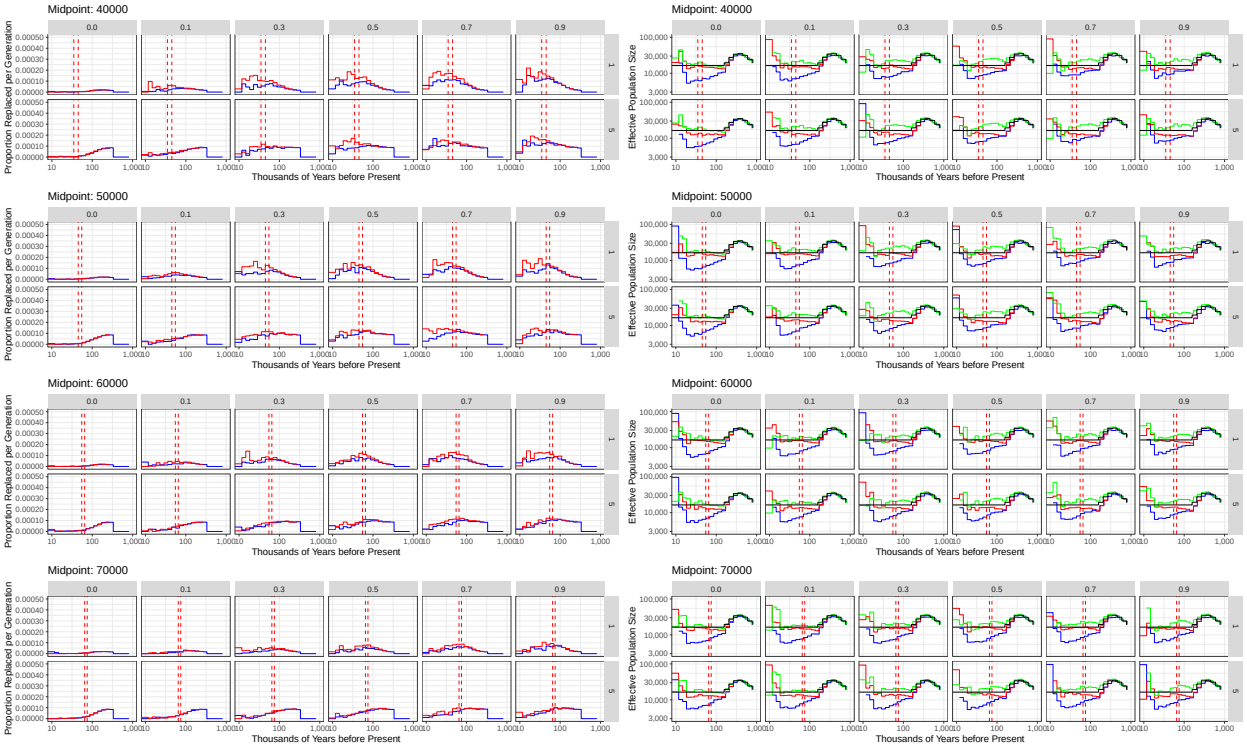


Figure S12: Bidirectional simulation

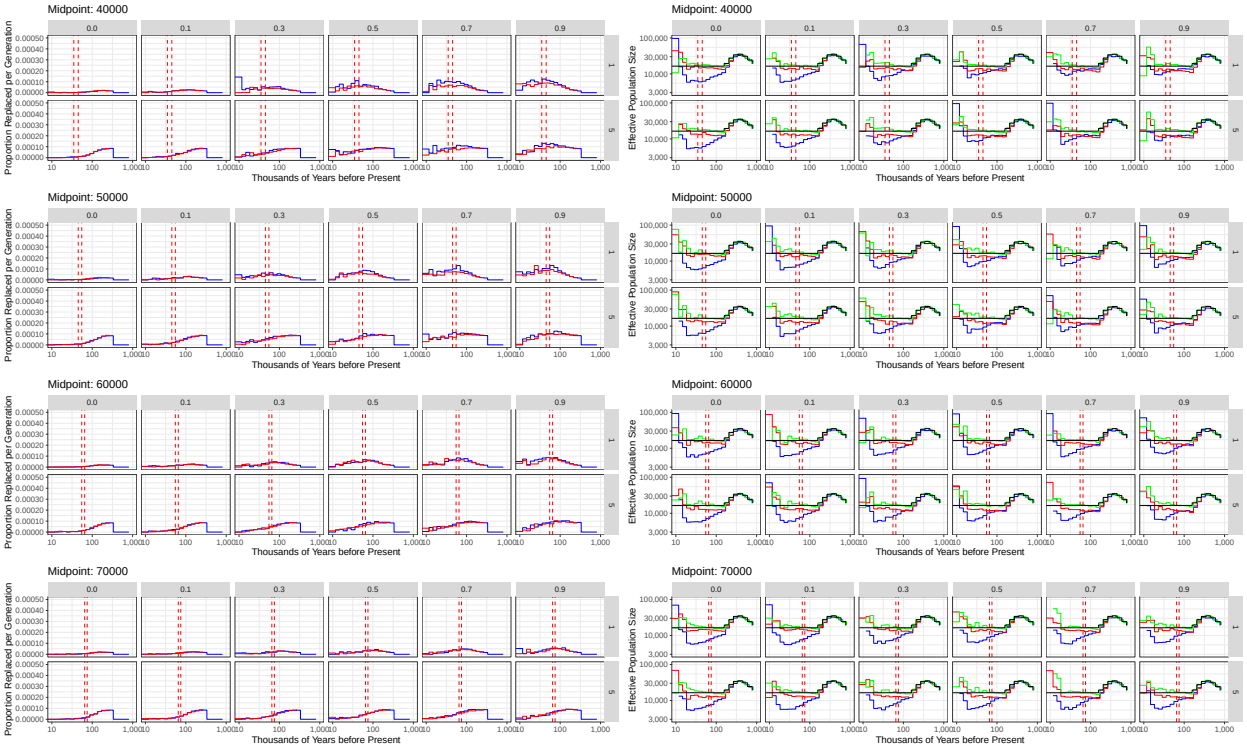
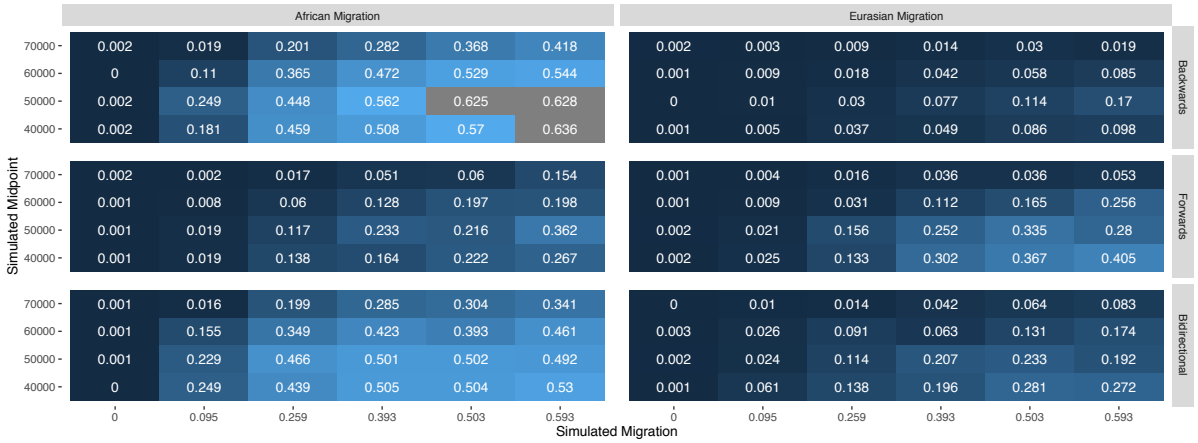
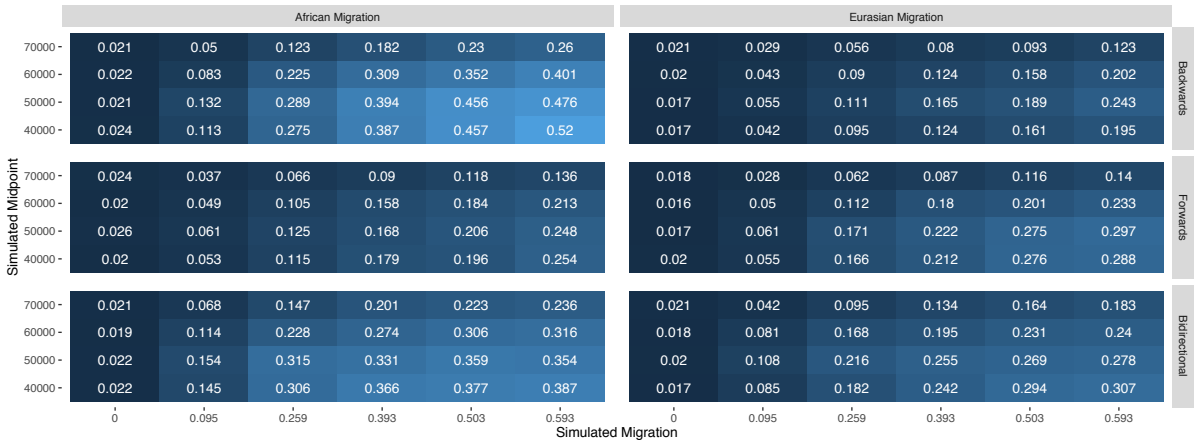


Figure S13: Forward simulation





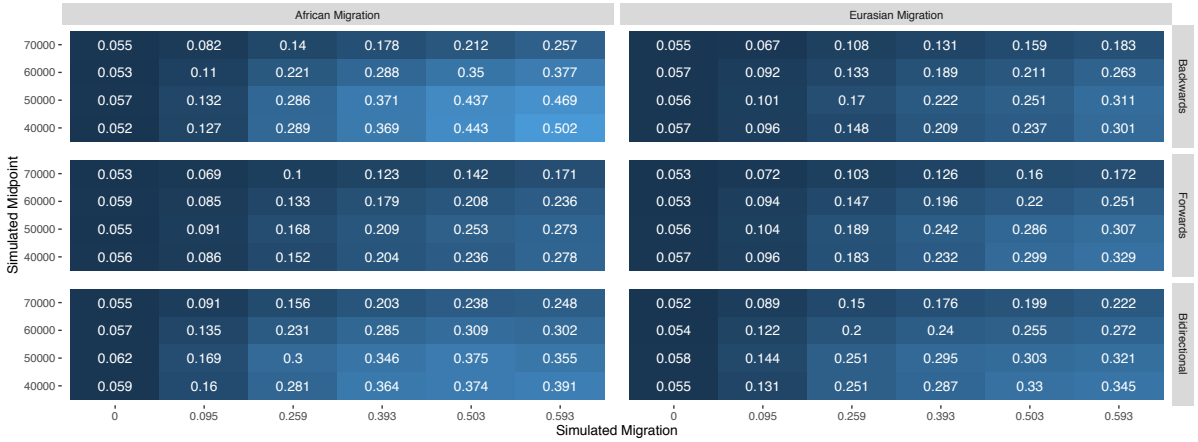


Figure S14: Area under the migration curve for three cases of simulated demography shown in Figures S11, S12, and S13. From top to bottom, inference was initiated with 0, 1, and 5 $4N_0$ population replacement per generation.

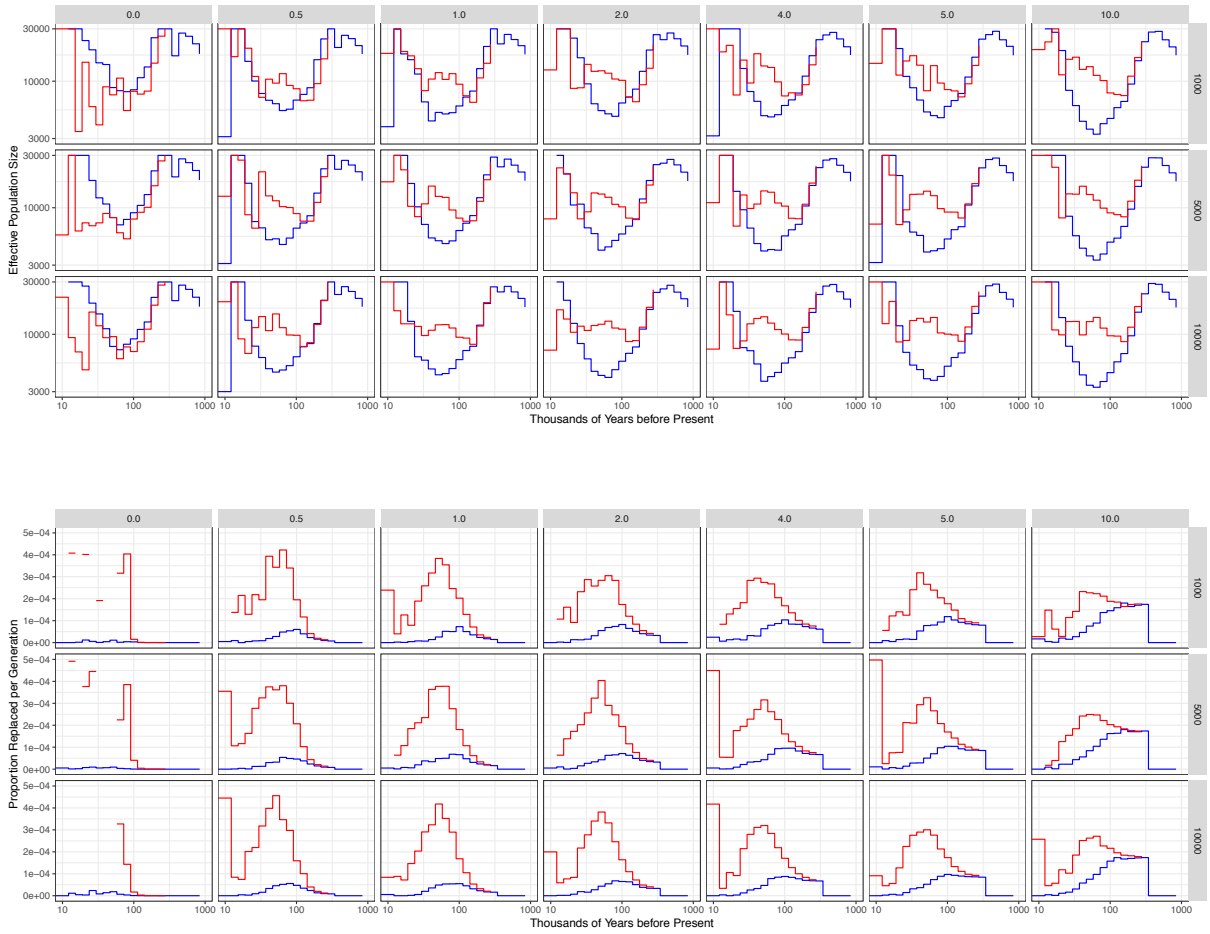


Figure S15: Effective population size and migration history of a Yoruban (`S_Yoruba-1`) and a French (`S_French-1`) individual from the Simons Genome Diversity panel. The initial migration proportion was varied along the X axis, while the number of particles is varied along the Y. 10 iterations of variational Bayes was used for parameter inference, while 5000 particles were used to sample from the posterior distribution of trees.

S4 Average D Statistics Among Populations

Here we consider the case of instantaneous admixture explored by Durand et al 2011 when a representative sample from the admixing population is available. Here, the expectation of D is reduced to

$$\begin{aligned} E[D(P_2, P_1, P0, O)] &= \frac{\Pr(\text{ABBA}) - \Pr(\text{BABA})}{\Pr(\text{ABBA}) + \Pr(\text{BABA})} \\ &= \frac{3f[t_{P3} - t_{GF}]}{3f[t_{P3} - t_{GF}] + 4N(1-f)(1-\frac{1}{2N})^{t_{P3}-t_{P2}} + 4Nf(1-\frac{1}{2N})^{t_{P3}-t_{GF}}} \end{aligned}$$

where t_{P_i} is the expected time to coalescence of any two lineages in P_i, f is the proportion of gene flow while t_{GF} is its timing, and N is the population size. We consider the case where D statistics $D(A_i, A_{pi}, Y, C)$ are summed and divided by their count. In this case, both A_i and A_{pi} are drawn from the same population, which implies that t_{P_3} is constant (with the exception of the San, who are excluded from these calculations). Since we assume that t_{P_3}, f, t_{GF} are held constant, we straightforwardly rearrange the definition of the D statistic to show that the expectation of averaging over the log of D values for n populations gives the expectation of a D statistic with the average t_{P2}, or time to divergence of lineages within the population.

$$\begin{aligned} \mathbb{E}[\frac{1}{n}\sum_i^n \ln D(P2, P1, P0, O)] &= \frac{1}{n}\sum_i^n \mathbb{E}[\ln D(P2, P1, P0, O)] \\ &= \frac{1}{n}\sum_i^n \ln \frac{3f[t_{P3} - t_{GF}]}{3f[t_{P3} - t_{GF}] + 4N(1-f)(1-\frac{1}{2N})^{t_{P3}-t_{P2}} + 4Nf(1-\frac{1}{2N})^{t_{P3}-t_{GF}}} \\ &= \frac{1}{n}\sum_i^n \ln\left(3f[t_{P3} - t_{GF}\right) - \ln(3f[t_{P3} - t_{GF}] + 4N(1-f)(1-\frac{1}{2N})^{t_{P3}-t_{P2}} \\ &+ 4Nf(1-\frac{1}{2N})^{t_{P3}-t_{GF}}) \\ &= \frac{1}{n}\sum_i^n \ln\left(3f[t_{P3} - t_{GF}\right) - \ln\left(3f[t_{P3} - t_{GF}]\right) + \ln\left(4N(1-f)(1-\frac{1}{2N})^{t_{P3}-t_{P2}}\right) \\ &- \ln\left(4Nf(1-\frac{1}{2N})^{t_{P3}-t_{GF}}\right) \\ &= \frac{1}{n}\left(n\left(\ln\left(3f[t_{P3} - t_{GF}\right) + \ln\left(3f[t_{P3} - t_{GF}]\right) + + \ln\left(4Nf(1-\frac{1}{2N})^{t_{P3}-t_{GF}}\right)\right)\right) \\ &+ \sum_i^n \ln\left(4N(1-f)(1-\frac{1}{2N})^{t_{P3}-t_{P2}}\right) \\ &= \ldots \sum_i^n (t_{P3} - t_{P2}) \ln\left(4N(1-f)(1-\frac{1}{2N})\right) \\ &= \ldots n \ln\left(4N(1-f)(1-\frac{1}{2N})\right)\sum_i^n (t_{P3} - t_{P2}) \\ &= \ldots n \ln\left(4N(1-f)(1-\frac{1}{2N})\right)\left(\sum_i^n t_{P3} - \sum_i^n t_{P2}\right) \\ &= \ldots n t_{P3} \ln\left(4N(1-f)(1-\frac{1}{2N})\right)\left(-\sum_i^n t_{P2}\right) \\ &\text{multiply through the } \frac{1}{n}\ldots \\ &= \ln\left(3f[t_{P3} - t_{GF}\right) + \ln\left(3f[t_{P3} - t_{GF}]\right) + \ln\left(4Nf(1-\frac{1}{2N})^{t_{P3}-t_{GF}}\right) \\ &+ t_{P3} \ln\left(4N(1-f)(1-\frac{1}{2N})\right)\left(-\frac{1}{n}\sum_i^n t_{P2}\right) \\ &= \ln\left(3f[t_{P3} - t_{GF}\right) + \ln\left(3f[t_{P3} - t_{GF}]\right) + \ln\left(4Nf(1-\frac{1}{2N})^{t_{P3}-t_{GF}}\right) \\ &+ \ln\left(4N(1-f)(1-\frac{1}{2N})^{t_{P3}-\frac{1}{n}\sum_i^n t_{P2}}\right) \\ &= \ln \frac{3f[t_{P3} - t_{GF}]}{3f[t_{P3} - t_{GF}] + 4N(1-f)(1-\frac{1}{2N})^{t_{P3}-\frac{1}{n}\sum_i^n t_{P2}} + 4Nf(1-\frac{1}{2N})^{t_{P3}-t_{GF}}} \\ &= \ln D(\bar{P2}, \bar{P1}, P0, O) \end{aligned}$$

Therefore, the expectation of average D statistics where both P_1 and P_2 are drawn from the same population produces the expectation of the D statistic with the average within-population coalescent rate.