Detecting archaic introgression without archaic reference genomes

Laurits Skov[1*], Ruoyun Hui[2], Asger Hobolth[1], Aylwyn Scally[2], Mikkel Heide Schierup[1], Richard Durbin[2*]

1. Bioinformatics Research Centre, Aarhus University, 8000 Aarhus C., Denmark
2. Department of Genetics, University of Cambridge, Cambridge CB2 3EH United Kingdom

*Correspondence: lskov@cs.au.dk, rd@sanger.ac.uk

Abstract

Human populations out of Africa have experienced at least two bouts of introgression from archaic humans, Neandertal and Denisovans. In Papuans there is prior evidence of both these introgressions. Here we present a new approach to detect segments of individual genomes of archaic origin without using an archaic reference genome. The approach is based on the detection of genomic regions with a high SNV density of SNVs not seen in unadmixed populations. We show using simulations that this provides a powerful approach to identifying segments of archaic introgression with a small rate of false detection. Furthermore our approach is able to accurately infer admixture proportions and divergence time of human and archaic populations.

We apply the model to detect archaic introgression in 89 Papuans and show how the identified segments can be assigned to likely Neandertal or Denisovan origin. We report more Denisovan admixture than previous studies and directly find a shift in size distribution of fragments of Neandertal and Denisovan origin that is compatible with a difference in admixture time. Furthermore we identify small amounts of Denisova ancestry in West Eurasians, South East Asians and South Asians.

Introduction

Archaic introgression into modern humans occurred at least twice (Neandertals and Denisovans) (MEYER *et al.* 2012; PRUFER *et al.* 2014) and had a phenotypic effect on humans (HUERTA-SANCHEZ *et al.* 2014; DANNEMANN *et al.* 2017; RACIMO *et al.* 2017). A large part of Neandertal and Denisovan genetic material is still present in modern humans and we can learn about archaic populations from studying the effect of their genetic variants in humans.

To harness this information a number of methods have been developed to infer segments of archaic ancestry in an individual's genome.

Scanning along the genome, Hidden Markov Models (HMMs)(PRUFER *et al.* 2014; SEGUIN-ORLANDO *et al.* 2014) and Conditional Random Fields (CRF)(SANKARARAMAN *et al.* 2016) can identify haplotypes in non-Africans that are 1. closer to the archaic reference genomes, than to Africans and 2. are longer than expected by incomplete lineage sorting, and these are then identified as archaic introgressed segments. Another approach is to identify segments with more variants in higher linkage disequilibrium (LD) that are unique to non-Africans than expected given a certain demographic scenario (PLAGNOL and WALL 2006). The latest implementations of this method also use an archaic reference genome for refining the set of putative archaic haplotypes (VERNOT *et al.* 2016).

The use of archaic reference genomes for identification of introgressed fragments has drawbacks. First, since the Neandertal reference genomes are closer to the introgressing Neandertal (80,000-145,000 years divergence)(PRUFER *et al.* 2017) , than the introgressing Denisova is to the Denisova genome (276,000-403,000 years divergence)(PRUFER *et al.* 2014) detecting Denisovan ancestry will be harder. Second, the reliance on having reference genomes implies that the introgression maps generated by these methods need updates whenever more archaic reference genomes are sequenced (PRUFER *et al.* 2017). Finally, it may be hard to identify introgressing segments of unknown archaic origin if such exists, as in the case of the putative archaic introgression into Pygmies (HSIEH *et al.* 2016) and Andamanese islanders (MONDAL *et al.* 2016).

Here we present a new method for the identification of archaic introgression that does not require a reference genome or prior knowledge of demographic parameters, but uses density of variants

in individuals private to their population of origin. We demonstrate with Papuans how we can estimate demographic parameters relevant to introgression and infer more archaic material than previously. Furthermore, we can separate introgression events into Denisovan and Neandertal components that display different length distributions in accordance with different admixture times.

Method

Model

An archaic genomic segment introgressed into a population is expected to have a high density of variants not found in populations without the introgression. We use a Hidden Markov Model (HMM) to classify genomic segments into states with varying variant density. We focus on a scenario where introgression with a deeply divergent archaic population only happened into an ingroup and not the outgroup, see Figure 1, panel a.

We can then remove variants found in the outgroup in order to better distinguish the variant density in introgressed segments and non-introgressed segments, see Figure 1, panel a. These remaining variants which we denote private variants (because they are private with respect to the ingroup) can either have occurred on the branch starting from the split of the ingroup and outgroup or if they are introgressed, they could have occurred on the introgressing population's branch. Because the introgressing segments have had a longer time to accumulate variants, they have a higher probability of emitting private variants.

Thus, we define a HMM with two states. The hidden states are Ingroup and Archaic, and the probability for changing state in the Ingroup is *p* and the probability for changing state in the Archaic is *q*, see Figure 1, panel b. The probability of changing state can also be expressed in terms of a constant recombination rate between windows $r \cdot L$, the admixture time T_{admix} and admixture proportion a, Figure 1, panel b. We show how to derive it in Supplementary note 1.

The number of private variants observed in a window of length *L* (typically $L = 1000\ bp$) is Poisson distributed with a rate $\lambda_{Ingroup}$ and $\lambda_{Archaic}$, respectively where $\lambda_{Ingroup} = \mu \cdot L \cdot T_{Ingroup}$ and $\lambda_{Archaic} = \mu \cdot L \cdot T_{Archaic}$. μ is the mutation rate, $T_{ingroup}$ is the mean coalescence time for the ingroup and the outgroup and $T_{Archaic}$ is the mean coalescence time for the archaic population and the outgroup, see Figure 1, panel c.

We make a correction of the rates to take into account the number of missing bases in a window and the local mutation rate. For window i we have $\lambda^i_{Ingroup} = \mu_i \cdot L_i \cdot T_{Ingroup}$ and $\lambda^i_{Archaic} = \mu_i \cdot L_i \cdot T_{Archaic}$, where μ_i is the local mutation rate and L_i is the number of called bases in a widow.

The set of transition parameters p, q and the Poisson parameters $\lambda_{Ingroup}$, $\lambda_{archaic}$ that maximizes the likelihood given the observations are found using the Baum-Welch algorithm for an individual genome. These parameters are informative of the mean coalescence with the ingroup and archaic with the outgroup, the admixture time and the admixture proportion if we assume a known mutation rate μ and a known recombination rate between windows rL.

Once the set of optimal parameters are found they can be used to decode the genome, using posterior decoding. We call each window where the posterior probability of being in the archaic state is bigger than 0.5 as archaic. We group consecutive archaic windows together into archaic segments and calculate the mean posterior probability of being archaic for the entire segment.

Results

Testing the model with simulations

To investigate the ability of our model to identify archaic (Neandertal and Denisovan) admixture into Papuans we simulated un-phased whole autosome data with admixture with an archaic hominin 1,500 generations ago replacing 5% of the population – (a script with all demographic parameters are shown in Supporting information – Simulation script.py and a graphic representation of the demography is shown in Supporting figure 1). We simulated three scenarios to test the effects of missing data and varying recombination rate. The mutation rate were kept constant across the genome for all simulations.

First, we simulated five individuals where every base in the genome is called equally well and there is a constant recombination rate of $1.2 \cdot 10^{-8}$ events per basepair per generation. We call this dataset the ideal data. Second, we simulated five individuals and removed all variants that are in repetitive regions (using the repeatmask track for the human reference genome hg19 (SMIT *et al.* 2013)) to test how the model performs with missing data. Third, we simulated five individuals with missing data and using a varying recombination rate (using HapMap phase II (INTERNATIONAL

HAPMAP *et al.* 2007)) to test the effect of missing data and recombination. We analyzed all genomes in bins of 1000 bp, and removed all variants found in 500 simulated Africans, 100 simulated Europeans and 100 simulated Asians. We combine two haplotypes to form genotype data for the simulated individuals. This will be more similar to situations where phased data is not available. For diploid data the conversion from p, q, $\lambda_{Ingroup}$ and $\lambda_{Archaic}$ to $T_{archaic}$, $T_{Ingroup}$, T_{admix} and a changes as is shown in Supplementary note 1.

We found the transition and emission parameters that optimized the likelihood, using the Baum-Welch algorithm and used them to get an estimate for the admixture time T_{admix}, the admixture proportion a and the mean coalescent times with the outgroup $T_{ingroup}$ and $T_{archaic}$ for the ingroup and archaic segments respectively, see Figure 2 panel b.

For all scenarios the coalescence time between the ingroup and outgroup ($T_{ingroup}$) the mean estimate is 2,625 generations ago(max = 2,647, min = 2,595), and the average simulated coalescent time with the outgroup is 3,109 generations ago. For the coalescence time between the outgroup and the archaic ($T_{archaic}$) the mean estimate is 37,345 generations ago (max = 37,832, min = 37,028) and the simulated values were 35,543 generations ago.

We find that the mean estimate of the admixture proportion a when using the transition matrix is between 4.62 % and 5.34 %.

We also find that with a posterior cutoff at 0.8 for segments (mean posterior probability of being archaic for all windows in segment), the amount of false positives can be reduced to around 50%, while still keeping 90% of the true segments, see Supporting figure 4.

An estimate of the false negative rate of the model is counting the amount of simulated archaic segments that have zero overlap with the putative archaic sequence, which is 11.1 Mb for ideal simulations, 32.2 Mb for simulations with missing data and 26.3 Mb for simulations with missing data and a varying recombination rate, see Figure 2 panel a. The model has less power to identify short segments as can be seen in Supporting figure 2.

If we estimate the false positive rate as the amount of putative archaic segments that have zero overlap with the simulated archaic segments we find 8.4 Mb for ideal simulations, 4.1 Mb for simulations with missing data and 9.0 Mb for simulations with missing data and a varying

recombination rate, see Figure 2 panel a. In total, we recover 243 Mb, 198 Mb and 184 Mb of archaic sequence for Ideal simulations, simulations with missing data and simulations with missing data and varying recombination rate respectively.

An example of how the simulated and putative archaic segments overlap is shown in Figure 2, panel c for the a 10 Mb window. A map of all simulated archaic segments and putative archaic segments can be seen in Supporting figure 3.

The mean estimate for the admixture time using transition matrix is around 1,704 generations ago when using the ideal data and 1,522 generations ago when adding missing data. When we vary the recombination rate across the genome the average estimate of the admixture time is 1,146 generations ago if we estimate it using the transition matrix. The underestimate of the admixture time might be due to fact that the model fail to identify around 80% of the short segments. This would make the average segment length longer and make the admixture time seem more recent.

Application to Papuan genomes

Having verified the validity of the model, we applied it to 14 Papuan individuals from Simons diversity project (MALLICK *et al.* 2016), 40 Papuans from (MALASPINAS *et al.* 2016) and an additional 35 Papuans (VERNOT *et al.* 2016). In addition to this, we also analyzed individuals from West Eurasia, East Asia and South East Asia.

For each individual we used two different sets of variants as outgroup. We estimate the background mutation rate in windows of 100 kb, using the variant density of all variants in the African populations from the 1000 genomes project.

Our model will not be able to distinguish Neandertal from Denisova segments in Papuans, because the Denisovans and Neandertals share a common ancestor before they do with humans and therefore the mean coalescence time with humans will be the same (PRUFER *et al.* 2014). This means that the Poisson parameters will be the same as they both depend on $T_{archaic}$. However, we should be able to enrich for Denisova and Neandertal segments by using different outgroups in our filtering step.

First, we used only variants found in Sub-Saharan populations as an outgroup. This should remove variation in the common ancestor of Sub-Saharan Africans and the Papuans, retaining archaic variants of Neandertal and Denisova origin as both are present in Papuans, but mainly absent in Africa (SANKARARAMAN *et al.* 2016; VERNOT *et al.* 2016). We also used this filter when analyzing Eurasian populations.

Second we remove variants found in all non Papuan populations, only retaining variants that are unique to Papuan populations. This should remove Neandertal variants that are shared with other non-African populations (PRUFER *et al.* 2014) and also to some extend remove variants of Denisova origin that are mainly found in Asians and Native Americans (SKOGLUND and JAKOBSSON 2011; QIN and STONEKING 2015). Thus removing all variants from 1000 genomes should enrich for Denisova segments while the segments that do not overlap when using the two different outgroups should be enriched for Neandertal segments.

We found the optimal set of transition and emission parameters for each Papuan individual and found them to be largely consistent across the different datasets, See Supporting figure 5. The parameters were converted into estimates of T_{admix}, a, $T_{ingroup}$ and $T_{archaic}$ using an average recombination rate of and mutation rate of $1.2 \cdot 10^{-8}$ events per basepair per generation and an average mutation rate of $1.25 \cdot 10^{-8}$ mutations per base pair per generation, see Figure 3, panel a and b.

We find that mean coalescence time between Papuans and non-Papuan individuals happened more recently (1,395 – 1,540 generations ago) than the mean coalescence time with Sub-Saharan Africans (1,953-2,293 generations ago) reflecting that Papuans are more closely related to other Non-Africans than to Africans. The mean coalescence time between Papuans and other non-Africans also provides an upper limit for Neandertal introgression because it happened into the common ancestor of these populations.

Using only Sub-Saharan individuals as an outgroup we find a mean coalescence time to between the archaic and outgroup to be between 29,404 and 33,944 generations ago. When using non-Papuans as an outgroup the estimate is between 25,268 and 30,352 generations ago. The lower estimate is likely due to the fact that some of the variants in the common ancestor of Denisovans and Neandertals have been removed.

We estimate the admixture proportion of archaic sequence to between 4.1-4.4 % and the admixture proportion that is unique to Papuans between 1.5-1.8 %. This means that around 2.6 % is shared with Non-Papuans, see Figure 3, panel a.

Using the transition parameters we estimate that both admixture events happened around 1000 generations ago which is likely an underestimate as it was for the simulated data with missing data and varying recombination rate. Neandertal admixture likely occurred some 2,000 generations ago after the out of Africa migration (FU *et al.* 2014; SANKARARAMAN *et al.* 2016) and Denisova admixture likely occurred before the peopling of Sahul which is thought to be 47-55 thousand years ago (1,620-1,896 generations ago assuming a generation time of 29 years) (CLARKSON *et al.* 2015; O'CONNELL and ALLEN 2015), if one assumes that Denisova ancestry in Asians, Native Americans and Papuans have the same origin.

We use a threshold of 0.8 posterior probability as in the case of the simulated data. By comparing to the Vindija (PRUFER *et al.* 2017) and Denisova (MEYER *et al.* 2012) genomes we find that this cutoff removes around 50% of the short segments that don't share variants with any archaic reference genome and could be deeply coalescing modern human haplotypes, see Supporting figure 8.

When we use a cutoff of 0.8 we find that 84 % of the segments unique to Papuans (80 % of the total sequence) shared more variants Denisova genome than Vindija and that 78 % the segments that are shared with other non-Africans (83 % of the total sequence) shared more variants with the Vindija Neandertal than the Denisova, See Figure 3, panel c. This means that a majority of the archaic sequence unique to Papuans likely comes from a population more closely related to Denisovans.

However, segments that are unique to Papuans are longer on average (94.2 kb) compared to those shared with other non-African populations (76.9 kb), See Figure 3, panel d. The difference in length distributions are not seen as clearly when using Sstar or CRF, see Supporting figure 6. Moreover, the length distribution of archaic segments that are not unique to Papuans are more similar to other non-African populations, see Supporting figure 7.

We compared our archaic segments to those previously reported using other methods (SANKARARAMAN *et al.* 2016; VERNOT *et al.* 2016). We find that 67% of the archaic sequence found

using CRF are also recovered using our method, and that 74% of the archaic sequence found using Sstar are also recovered using our method.

Comparing to the archaic reference genomes our method finds more Denisova in Papuans than it finds Neandertal unlike the CRF. It also finds a significant amount of additional Denisova segments in East and South East Asians, see Table 1.

Model	*Pop*	*Both*	*Denisova*	*None*	*Vindija*	*Total*
HMM	Papuan	4.40	77.00	11.39	71.44	164.23
	eastasia	1.48	5.69	9.96	61.37	78.49
	southasia	1.62	5.85	10.12	51.36	68.95
	westeurasia	1.47	2.39	10.14	43.95	57.94
Sstar	Papuan	26.5	43.11		49.21	118.82
	eastasia	-	0.00	-	65.02	65.02
	southasia	-	0.00	-	55.18	55.18
	westeurasia	-	0.00	-	51.23	51.23
CRF	Papuan	-	58.17	-	84.72	142.89
	eastasia	-	3.21	-	72.92	76.14
	southasia	-	2.79	-	61.36	64.15
	westeurasia	-	0.68	-	57.29	57.97

Table 1. Amount of sequence of different origins. The amount of sequence (in Mb) that is equally related to Denisova and Vindija, more closely related to Denisova, doesn't share any variation with either and is more closely related to Vindija are shown different populations and different methods.

Discussion

Since emission probabilities are very different between the human and archaic states in our model, we expect a low rate of false positive archaic inference and this is also what we see in simulations made to match the real data closely. However, since recombination rates are highly variable, a lot of very short archaic segments are expected and these have a very high false negative rate. Our inability to identify these cause us to underestimate of the admixture time. However the model does seem to find the correct size distribution for longer segments (> 50 kb), see Supporting figure 2. The mean coalescence times of modern and archaic humans is estimated well in simulations. In real data, the potential presence of super-archaic introgression as reported into the sequenced Denisovan (PRUFER *et al.* 2014) should cause us to overestimate this quantity, but this effect is expected to be small since there is little difference between Europeans and Papuans for this quantity. It remains to be seen whether there is also evidence for super-archaic admixture in the introgressing Denisova. One way to do this would be to just add an extra super-archaic hidden state to our model.

Our model reports more Denisova segments than approaches relying on the Denisovan reference. This is likely because our method does not rely on putative Denisova segments being more closely related to the Denisova genome than the Vindija Neandertal genome. Given that the introgressing and sequenced Denisova split shortly after the Neandertals split from Denisovans (PRUFER *et al.* 2014) many segments are expected to be equally close to the Vindija Neandertal and Denisova. It is also expected that a fraction of segments introgressed for Denisova are more closely related to Vindija and vice verse due to incomplete lineage sorting. It is therefore also reassuring that we do not find the same large excess of Neanderthal fragments in Papuans compared to Asian populations as has been reported previously, see Table 1.

We find no clear evidence for an introgression with a new archaic hominin in Papuans, but we do find segments that do not share variation with any of the sequenced archaic populations. These segments could likely represent variation in Neandertals and Denisovans that is not captured by the three high coverage archaic reference genomes. In the future it will be interesting to compare

these segments to other human populations that might also have archaic segments of unknown origin (HSIEH *et al.* 2016; MONDAL *et al.* 2016).

Our model works particularly well when it is possible to remove all the common variation between the ingroup and outgroup. As a larger number of individuals from different species are being sequenced, this method could be used as an alternative method for identifying introgression in other species, e.g. chimp bonobo (DE MANUEL *et al.* 2016), bears (LIU *et al.* 2014), elephants (PALKOPOULOU *et al.* 2018) or gibbons (CARBONE *et al.* 2014).

Materials and methods

Simulations

To simulate data we used Msprime (KELLEHER *et al.* 2016). We simulated 5 Papuans and as an outgroup we simulated 500 Africans, 100 Europeans and 100 Asians using demographic parameters from (MALASPINAS *et al.* 2016). We simulated data where we varied the recombination rate according to HapMap recombination maps (INTERNATIONAL HAPMAP *et al.* 2007) for 5 individuals and removed variants within non-callable regions and variants that were found in the simulated outgroup. We grouped all autosomes into bins of 1000 base pairs and counted the number of variants. For each 1000 bp windows we calculated the number of called bases using the repeat masked segments.

Train parameters and decode segments

We trained and decoded the segments using our HMM, which is available at:
https://github.com/LauritsSkov/Introgression-detection/

Data sets

We used 14 Papuans, 71 WestEurasians, 72 East Asians and 39 South Asians individuals from Simons diversity project (MALLICK *et al.* 2016), 40 Papuans form (MALASPINAS *et al.* 2016) and an additional 35 Papuans (VERNOT *et al.* 2016).

Filtering variants in real data

We used two sets of outgroups. One is all Sub-Saharan Africans (populations: YRI, MSL, ESN) from 1000 Genomes Project (GENOMES PROJECT *et al.* 2015) and all Sub-Saharan African populations from Simons (MALLICK *et al.* 2016) (not Masai, Somali, Sharawi and Mozabite).
The other ougroup is all individuals from the 1000 Genomes (GENOMES PROJECT *et al.* 2015) as outgroup and all Non-Papuans from Simons diversity project.
For all human data sets, we also removed sites that fell within repeatmasked (SMIT *et al.* 2013) regions, and sites that were not in the Strict callability mask for the 1000 genomes Project.
Repeat mask regions
hgdownload.cse.ucsc.edu/goldenpath/hg19/bigZips/chromFaMasked.tar.gz

Strict callability mask for 1000 genomes:
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/accessible_genome_masks/StrictMask/

The background mutation rate was calculated using the variants density of all variants from populations YRI, LWK, GWD, MSL and ESN in windows of 100 Kb divided by the mean variant density of the whole genome.

Comparison to Sstar and Conditional Random Field

We called Neandertal and Denisova segments in the 14 Papuans and compared them to the segments called with CRF with more than 50 posterior probability (SANKARARAMAN *et al.* 2016) available at:
https://sriramlab.cass.idre.ucla.edu/public/sankararaman.curbio.2016/
The path to the haplotypes is:
summaries/2/denisova/oceania/summaries/haplotypes/CRHOM.thresh-50.length-0.00.haplotypes

We called Neandertal and Denisova segments in the 35 Papuans and compared them to the segments called with Sstar with more than 99 posterior probability (VERNOT *et al.* 2016) available at:
https://drive.google.com/drive/folders/0B9Pc7_zItMCVWUp6bWtXc2xJVkk
The path to the haplotypes is:
introgressed_haplotypes/LL.callsetPNG.mr_0.99.den_calls_by_hap.bed.merged.by_chr.bed

Calculation of admixture time using length distribution

We fitted a linear regression line to the log of the length distribution of putative archaic segments with a posterior probability greater than 0.5. The slope is an estimate of the mean of the exponential distribution and can be converted into an admixture time using $mean\ of\ exponential = \frac{1}{(1-m)r(t-1)}$ where where m is the admixture proportion, t is the admixture time and r is the recombination rate per base pair.

340 **Figure legends**

341 **Figure 1. Overview of the model.** Illustration on small test dataset. a) An archaic segment
342 introgresses into the ingroup population at time T_{admix} with admixture proportion a. The
343 segments in the ingroup has a mean coalescence time with a segment from the outgroup at time
344 $T_{Ingroup}$ and an archaic segment has a mean coalescence time with a segment from the outgroup
345 at time $T_{Archaic}$. Removing all variants found in the outgroup (light orange points) should remove
346 all the variants in the common ancestor of ingroup and outgroup, leaving only private variants that
347 either occurred on the ingroup branch (dark orange) or on the archaic branch (dark blue). This will
348 make the archaic segment have a higher variant density. The genome is then binned into windows
349 of *L* (here 1000 bp) and the number of private variants are counted in each window. These are the
350 observations and the hidden states are either Ingroup state or Archaic state. When decoding the
351 sequence the most likely path through the sequence is found. b) The transition matrix between
352 the archaic state and ingroup state. c) The emission probabilities are modelled as Poisson
353 distributions with means $\lambda_{Ingroup}$ and $\lambda_{Archaic}$. It is more likely to see more private variants in the
354 Archaic state than in the Ingroup state.

355

356 **Figure 2. Evaluation of the model on simulated data.** a) Average amount of sequence per
357 individual that come from segments that are classified as false archaic (zero percent overlap with
358 any true archaic segment), found < 50% (segment where there is less than 50 % overlap with true
359 archaic segments), found > 50 % (segments where more than 50 % overlap with true archaic
360 segments) and missed archaic which are segments that the model does not identify as archaic. The
361 bars are colored according to what simulation scenarios they belong to. All segments are used
362 here are required to have > 0.8 posterior probability. b) The estimation of the four parameters
363 T_{admix}, a, $T_{ingroup}$ and $T_{archaic}$ are shown for the different simulation scenarios. c) And example
364 of how simulated archaic segments and putative archaic segments overlap in a 10 Mb window.
365 The x-axis is the genomic coordinates in Mb and the y-axis is the different simulation scenarios.

366

Figure 3. Application of model to Papuan genomes. a) Relationship between modern and archaic humans with the outgroup branches (Sub-Saharan Africans) colored in red. The average coalescence times for ingroup and outgroup $T_{Ingroup}$ and archaic and outgroup $T_{Archaic}$ are shown. The admixture proportions a and admixture time T_{admix} are shown for segments that are shared with other non-African populations. b) The outgroup colored in red is now all non-Papuans, and the new demographic parameters are shown. c) The segments that are shared with other Non-Africans share more variation with the Vindija Neandertal than they do with the Altai Denisova. Segments that are unique to Papuan individuals share more variation with Altai Denisova than they do with the Vindija Neandertal. d) The length distribution (all archaic segments with a mean posterior probability > 0.5 are kept) for segments that are shared with other non-African populations is shorter than segments that are unique to Papuans.

Supporting Figure legends

Supporting figure 1 – Demographic parameters for simulation. The effective population sizes, split times and bottleneck population sizes are shown for the simulated populations.

Supporting figure 2 - Total segments and sequence called SIM. The first column show the total number of segments found and the second column show the total amount of sequence that these segments add up to. The rows are different simulation scenarios and the colors of the stacked bar plot show the amount/number of segments that are not found using posterior decoding, where less than half of the segment overlap with the true archaic segments or where more than half of the segment overlaps with the true archaic segment.

Supporting figure 3- True and inferred archaic for the whole genome for simulated data. The x-axis is the genomic coordinates and the y-axis is the simulated individual. We colored the segments according to if they are true archaic of inferred archaic by the model using a cutoff of 0.5 for the mean posterior probability of the segment.

Supporting figure 4 – Effect of adjusting cutoff for when to include a putative archaic segment. The rows are different simulation scenarios and the rows are different classifications of putative archaic segments. False is segments with zero overlap to the true archaic segments, found<50%

395 are archaic segments that overlap with less than 50% with the true archaic segments and
396 found>50% are segments that overlap with more than 50% with the true archaic segments. On
397 the x-axis is the mean posterior probability of an archaic segment and the y-axis is the amount of
398 sequence left when applying the filter.

399

400 **Supporting figure 5 - Parameter estimation of Papuans.** The different facets show the estimates
401 for the parameters t_admix, a, T_ingroup and T_archaic depending on which outgroup was used
402 (Sub-Saharan Africans or the whole world (non-Papuans). The bars are colored according to which
403 dataset they came from.

404 **Supporting figure 6 - Length distribution other methods.** The length distribution of all Denisova
405 and Neandertal segments found using conditional random field (CRF), the hidden Markov model
406 (HMM) and Sstar. For the HMM Neandertal are those segments that are shared with other non-
407 African populations and Denisova are those unique to Papuans.

408

409 **Supporting figure 7 - Length distribution of Asia, Europe and Papuans.** The length distribution of
410 segments unique to Papuans (Denisova) and segments shared with other non-African populations
411 (Neandertal) are shown for four different population groups.

412

413 **Supporting figure 8 - Closeness to archaic humans with cutoff.** The probability of a segment
414 sharing any variants with the archaic reference genomes as a function of the average posterior
415 probability of a segment.

References

Carbone, L., R. A. Harris, S. Gnerre, K. R. Veeramah, B. Lorente-Galdos *et al.*, 2014 Gibbon genome and the fast karyotype evolution of small apes. Nature 513: 195-201.

Clarkson, C., M. Smith, B. Marwick, R. Fullagar, L. A. Wallis *et al.*, 2015 The archaeology, chronology and stratigraphy of Madjedbebe (Malakunanja II): A site in northern Australia with early occupation. J Hum Evol 83: 46-64.

Dannemann, M., K. Prufer and J. Kelso, 2017 Functional implications of Neandertal introgression in modern humans. Genome Biol 18: 61.

de Manuel, M., M. Kuhlwilm, P. Frandsen, V. C. Sousa, T. Desai *et al.*, 2016 Chimpanzee genomic diversity reveals ancient admixture with bonobos. Science 354: 477-481.

Fu, Q., H. Li, P. Moorjani, F. Jay, S. M. Slepchenko *et al.*, 2014 Genome sequence of a 45,000-year-old modern human from western Siberia. Nature 514: 445-449.

Genomes Project, C., A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison *et al.*, 2015 A global reference for human genetic variation. Nature 526: 68-74.

Hsieh, P., A. E. Woerner, J. D. Wall, J. Lachance, S. A. Tishkoff *et al.*, 2016 Model-based analyses of whole-genome data reveal a complex evolutionary history involving archaic introgression in Central African Pygmies. Genome Res 26: 291-300.

Huerta-Sanchez, E., X. Jin, Asan, Z. Bianba, B. M. Peter *et al.*, 2014 Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature 512: 194-197.

International HapMap, C., K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds *et al.*, 2007 A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861.

Kelleher, J., A. M. Etheridge and G. McVean, 2016 Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. PLoS Comput Biol 12: e1004842.

Liu, S. P., E. D. Lorenzen, M. Fumagalli, B. Li, K. Harris *et al.*, 2014 Population Genomics Reveal Recent Speciation and Rapid Evolutionary Adaptation in Polar Bears. Cell 157: 785-794.

Malaspinas, A. S., M. C. Westaway, C. Muller, V. C. Sousa, O. Lao *et al.*, 2016 A genomic history of Aboriginal Australia. Nature 538: 207-214.

Mallick, S., H. Li, M. Lipson, I. Mathieson, M. Gymrek *et al.*, 2016 The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 538: 201-206.

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

Mondal, M., F. Casals, T. Xu, G. M. Dall'Olio, M. Pybus *et al.*, 2016 Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. Nat Genet 48: 1066-1070.

O'Connell, J. F., and J. Allen, 2015 The process, biotic impact, and global implications of the human colonization of Sahul about 47,000 years ago. Journal of Archaeological Science 56: 73-84.

Palkopoulou, E., M. Lipson, S. Mallick, S. Nielsen, N. Rohland *et al.*, 2018 A comprehensive genomic history of extinct and living elephants. Proc Natl Acad Sci U S A 115: E2566-E2574.

Plagnol, V., and J. D. Wall, 2006 Possible ancestral structure in human populations. PLoS Genet 2: e105.

Prufer, K., C. de Filippo, S. Grote, F. Mafessoni, P. Korlevic *et al.*, 2017 A high-coverage Neandertal genome from Vindija Cave in Croatia. Science.

Prufer, K., F. Racimo, N. Patterson, F. Jay, S. Sankararaman *et al.*, 2014 The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505: 43-49.

Qin, P., and M. Stoneking, 2015 Denisovan Ancestry in East Eurasian and Native American Populations. Mol Biol Evol 32: 2665-2674.

Racimo, F., D. Gokhman, M. Fumagalli, A. Ko, T. Hansen *et al.*, 2017 Archaic Adaptive Introgression in TBX15/WARS2. Mol Biol Evol 34: 509-524.

Sankararaman, S., S. Mallick, N. Patterson and D. Reich, 2016 The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. Curr Biol 26: 1241-1247.

Seguin-Orlando, A., T. S. Korneliussen, M. Sikora, A. S. Malaspinas, A. Manica *et al.*, 2014 Paleogenomics. Genomic structure in Europeans dating back at least 36,200 years. Science 346: 1113-1118.

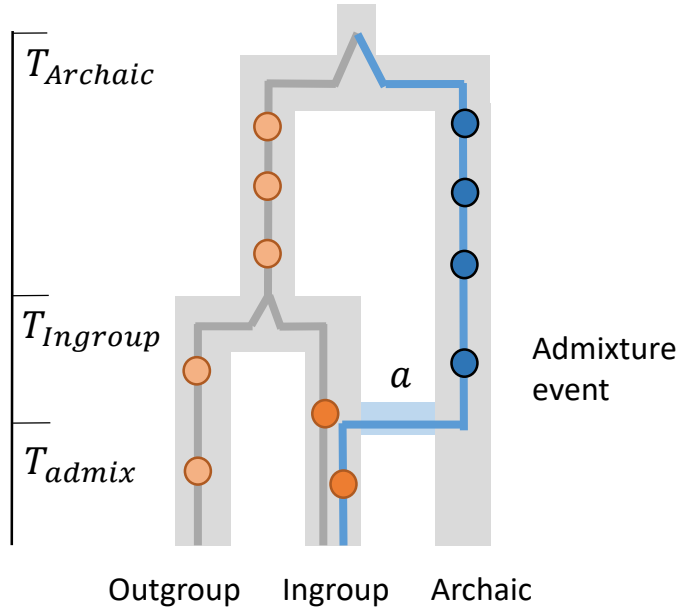
Skoglund, P., and M. Jakobsson, 2011 Archaic human ancestry in East Asia. Proc Natl Acad Sci U S A 108: 18301-18306.

Smit, A. F. A., R. Hubley and P. Green, 2013 RepeatMasker Open 4.0. RepeatMasker Open 4.0.

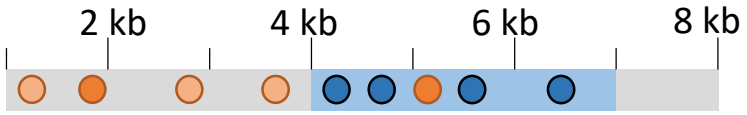
Vernot, B., S. Tucci, J. Kelso, J. G. Schraiber, A. B. Wolf *et al.*, 2016 Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. Science 352: 235-239.

a

Overview of the model



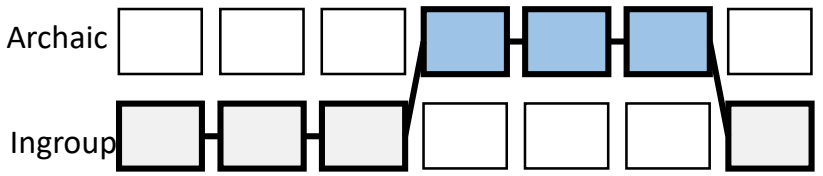
Observed variants



Remove variants found in outgroup

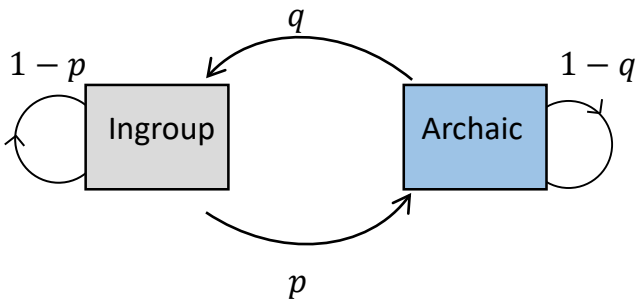
Count variants in window

Decode sequence



b

Transition probabilities

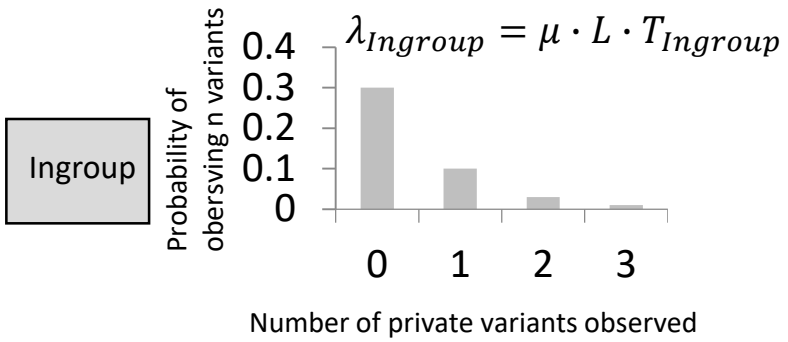


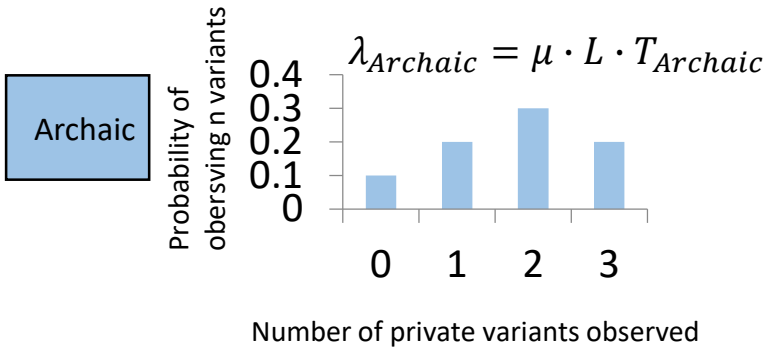
$$q \approx T_{admix} \cdot r \cdot L \cdot (1-a)$$

$$p \approx T_{admix} \cdot r \cdot L \cdot a$$

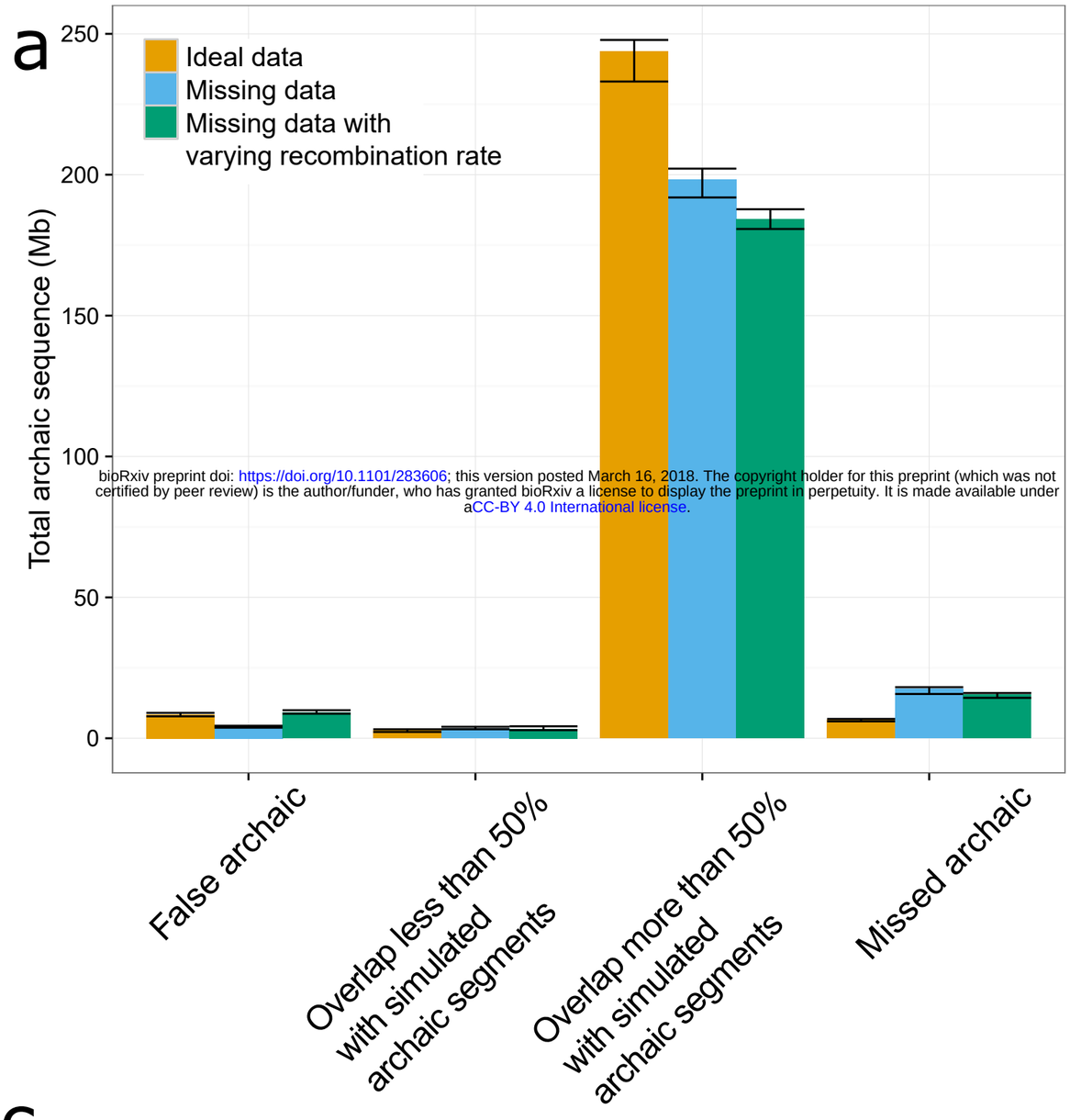
c

Emission probabilities

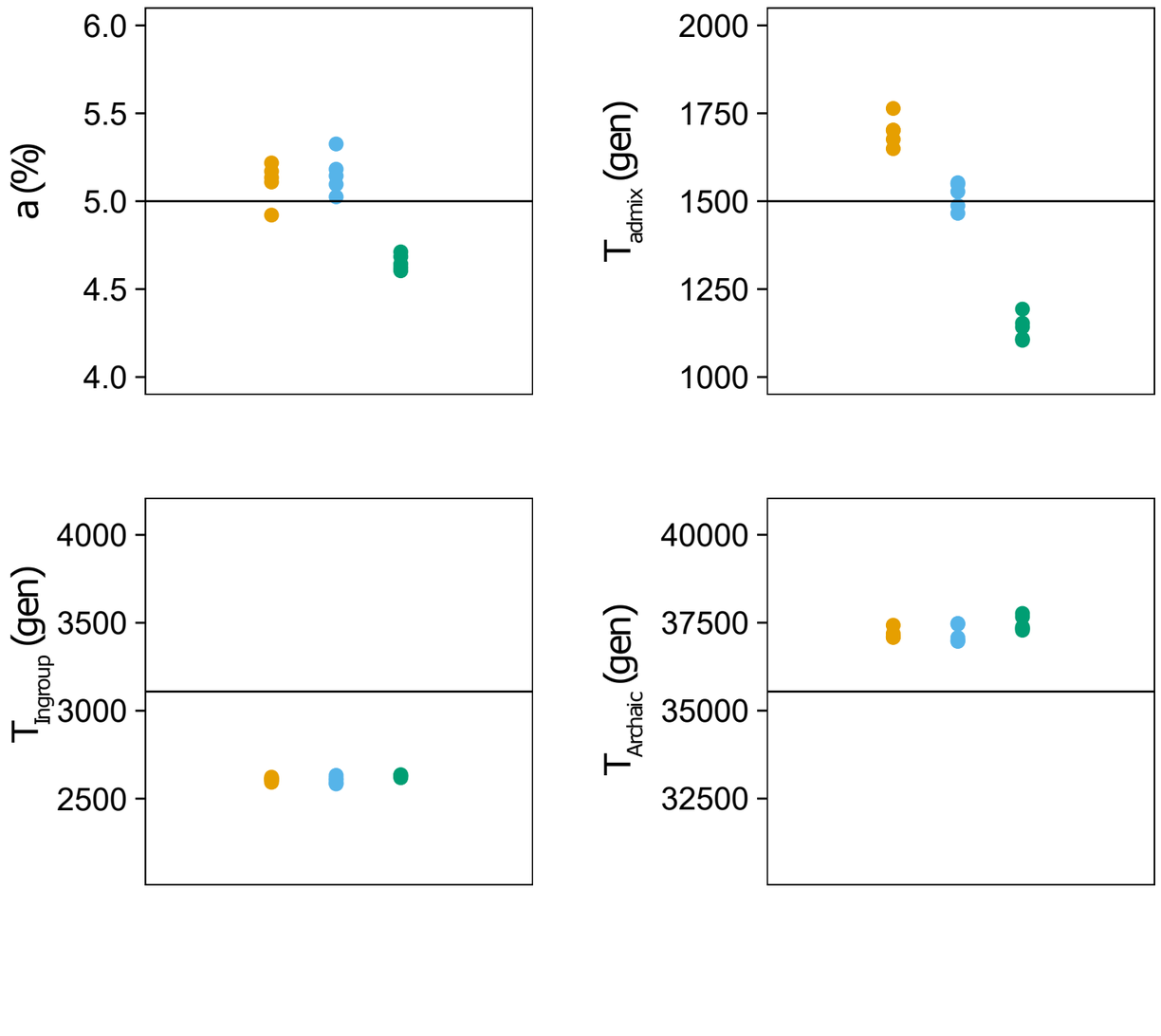




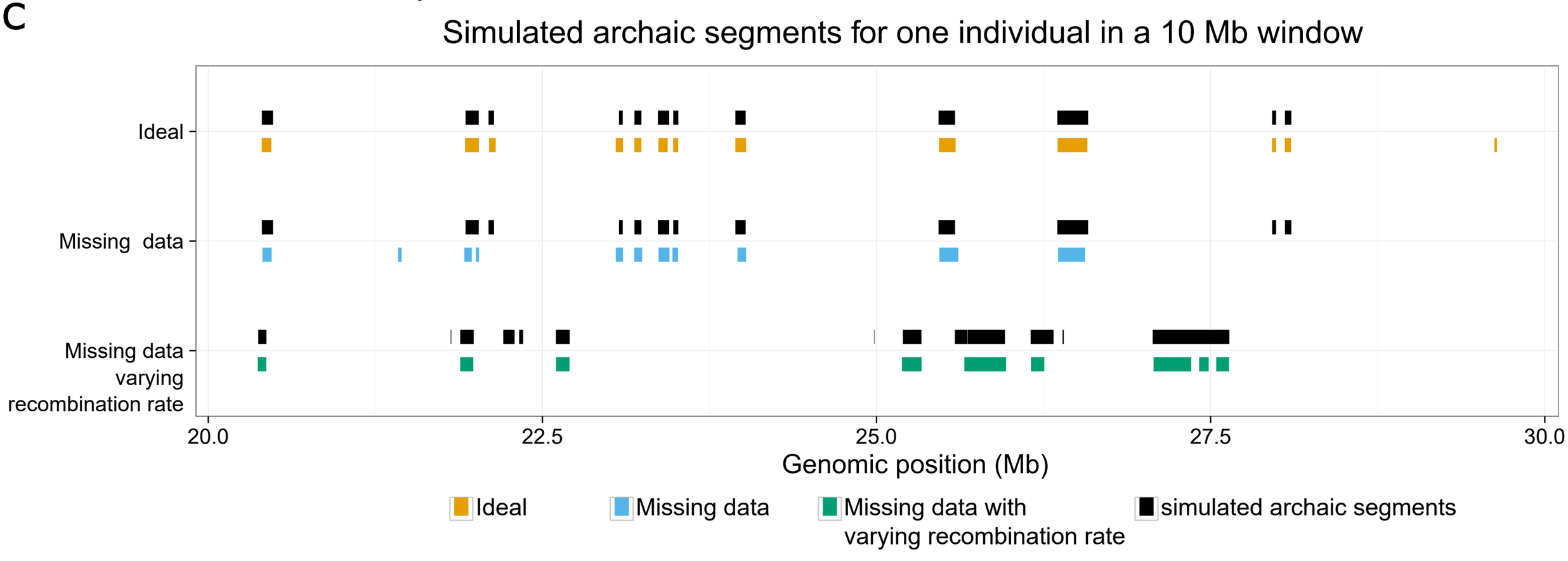
a



b



c

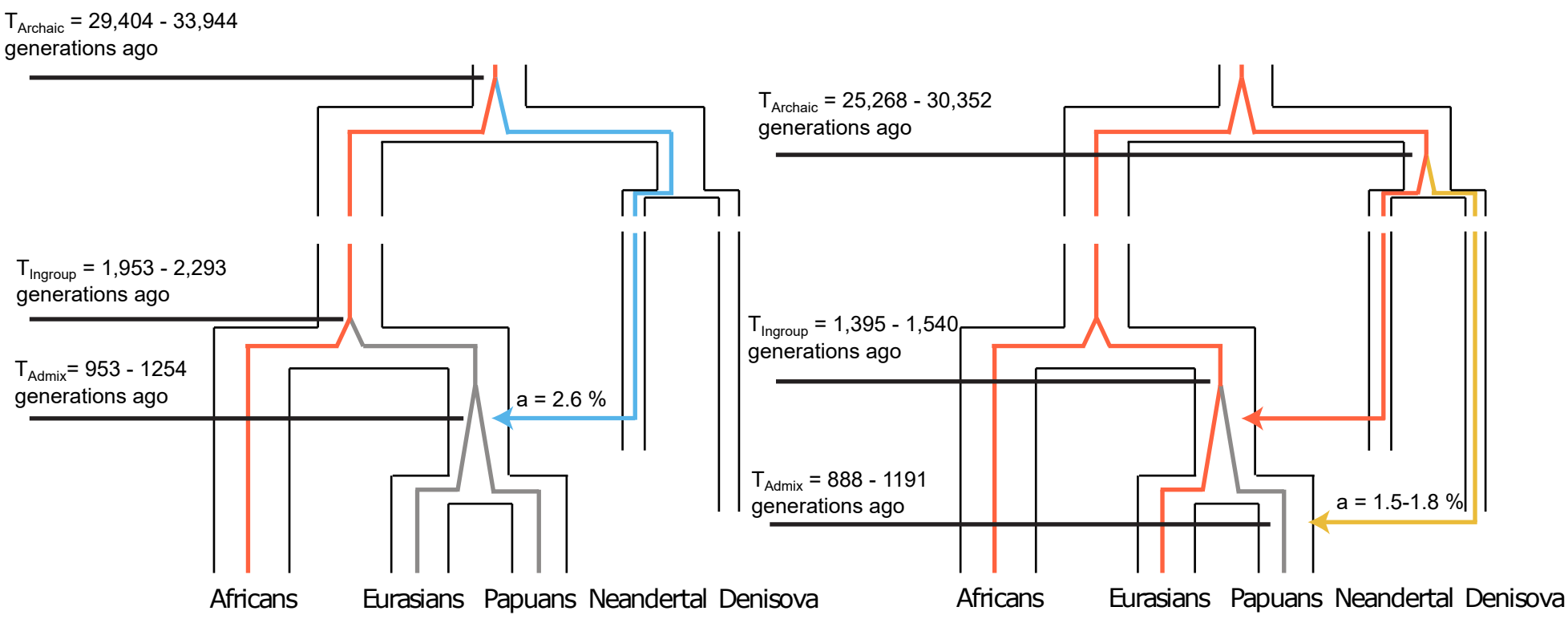


a

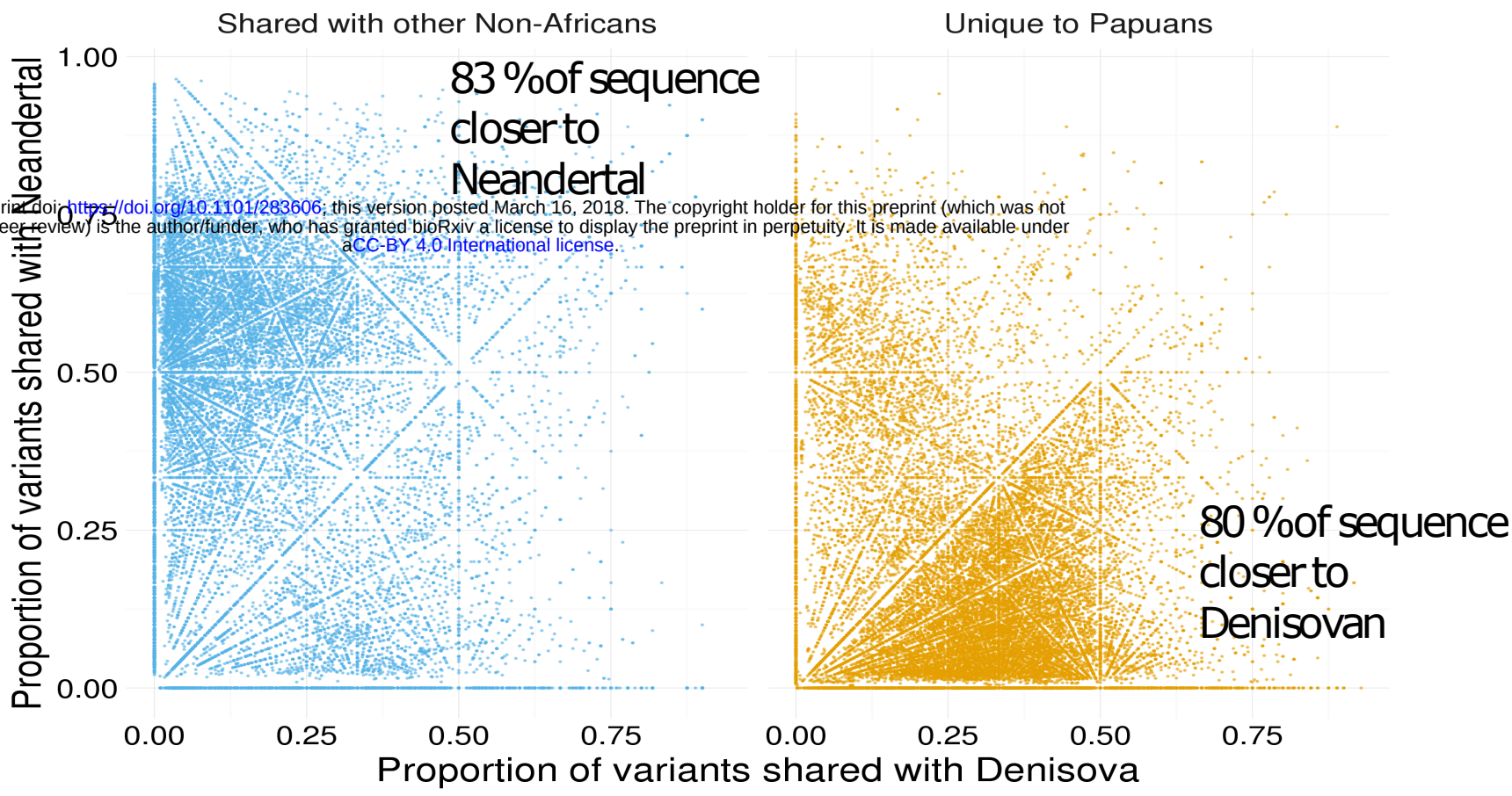
Using Africans as outgroup

b

Using Non-Papuans as outgroup



c



d