Back-migration V2

ccole

January 2020

Outline

1. Introduction
2. Methods
3. Current Results
   (a) Effective population sizes from SMC2 and MSMC, which finds differences in the earlier periods. This is similar to what was seen in PSMC.
   (b) To investigate this, we simulate things like PSMC, and find that we also can recover the incorrectly high population size. (What could be causing this high population size?)
   (c) Migration rate inference shows a peak of migration after this time period, acting as a possible explanation for the inflated population size. This trend breaks down mostly by language group (which are also validated evolutionary phyla).
   (d) Concerns about the phasing methodology made us validate this in HGDP. Language groups main trends line up well with results from SGPD.
   (e) Migration initialisation is important, talk about how different starting values can give different results. Include a brief supplemental section about seeding away from symmetric migration.
   (f) We can isolate the segments with a migration event in their history by using SMC2's estimated ARG. We can use these segments to look at the relationship between these individuals and all of the others in SGDP. We find that our statistics in our segments are higher, meaning that the segments are indeed more Eurasian.
   (g) Briefly talk about expectation of segment length, and how ours is consistent with a very large migration in that period, though we think this is unlikely, it does line up with the later evidence from the simulations. Caveats: not a single pulse (which makes it less than ideal, ideally would integrate similar to earlier paper (but this is very difficult to do analytically)), estimates not accurate with large values, selection almost certainly a factor.

(h) We simulate to find situations consistent with what we see, and find that we can recover basically the exact same thing with very large migration pulses. This is unrealistic, and means that we have some bias, at some level dependant upon the age of the migration, which is affecting inference. A large migration is consistent with the migration segments analysis, but is unlikely.

(i) **W**e use the segments we have to estimate relationships between the segments and a panel of ancient individuals, to try to get more clarity on the donating population. Ideally some estimate of consistency among populations, and pull in the average D statistic calculation. Not sure what the results will be here.

(j) We look at the relationship between the segments and Neanderthals, which are known to have contributed to OoA groups around this time period. Talk about the SMC2 analysis with Vindija, and Kay's analysis with the D statistics (which I will have to rerun).

(k) **W**e also use these D statistics to estimate a demographic model with ADMIXTOOLS, and show the resulting best-fit graph. Not sure what the results will be here.

4. Discussion

Things to (possibly) do.

1. ✗Run MSMC alongside the actual inference, get the MSMC N_e estimates on the same samples, overlay the plots. Have a figure which directly compares them on one particular case, then a supplemental that will have all of them with this overlay. Will probably want to include the x-coal rate somewhere in here. Put it on the migration plots probably.

2. ✓PSMC simulations have to be re-run with the correct migration parameter. (These are actually only single haploid inference, so they're fine. The simulations in general have to be rerun, but these ones are okay as is.)

3. ✓Simulations have to be run with the same magnitudes...? Did I do these already? Yes, I have all the magnitudes that I want, and I've run more midpoints with the different initiationalisation values for migration.

4. (Long) Run all groups in Africa with the correct migration initiation. Ideally against a few different groups (French, Han, Papuan, something else)

5. (Long) Replicate the SGDP runs with the HGDP.

6. After the whole thing is done, do the D statistics with all versus segments (ideally for all of the "ascertainment" schemes), and reproduce the plots from before.

7. The segment length estimation has to be redone, but there isn't all that much to do here.

8. Maybe redo a Vindija run, maybe not.

9. Kay's Neanderthal analysis has to be redone with the new segments. Hopefully this should be straightforward, as I'll just have a script that reruns his particular statistics. This will be a table, like the one in his Nddth paper, with the individdual statistics picked out. The rest can go in a Supplemental section.

10. Admixturegraph if its possible.

List of figures / tables:

Figures

1. a) SMC2 versus MSMC population size inference for a case or two b) PSMC simulation that shows the same thing.

2. a) Migration inference by language phyla (possibly superimposed all on top of each other), along with HGDP validation. b) Initialisation conditions. c) Segment lengths (by replicate, and HGDP/SGDP)

Supplementary Figures / Sections

1. All of the SMC2 versus MSMC N_e inference

2. All of the PSMC simulations.

3. Migration rate inference in SGDP and HGDP (with error bars)

4. Kay's statistics, and maybe add in an SMC2 run with the vindija. This is a low priority.

5. Probably a section for the "rest" of the D statistics, because I probably won't be able to use all of them. Potentially this can be a file though, instead of a table. The table doesn't really tell you very much.

1 Results

Using a French individual as a representative for Western Eurasians, we infer effective population size (N_e) and directional migration to African populations in the Simons Genome Diversity Panel with `smcsmc`. We simultaneously use MSMC with recommended parameters to infer effective population size on the same samples for comparison. Generally, both algorithms give comparable estimates of population size. However, around the divergence of the two lineages, MSMC shows an increase in African N_e relative to Eurasian N_e. This is consistent with a previous artefact identified and discussed in Li and Durbin 2012. `smcsmc`, on the other hand, shows both populations experiencing a similar bottleneck, and a much later effective divergence, more in line with generally accepted Out of Africa (OoA) timelines. We hypothesize that the difference between the two inferences is due to a migration from Eurasian populations to African populations directly after the period of population divergence. To test this hypothesis, we use `scrm` to simulate a variety of historical situations with and without migration in either or both directions. We use a skeleton of population history given in Supplemental Section S2 and mimic PSMC inference by analysing one haploid from each population. As expected, we find the hypothesized inflation in N_e, with a magnitude proportional to the amount of simulated migration. A full discussion of these simulation results may be found in supplemental section S3.

In the populations with this trend, `smcsmc` infers directional migration from Eurasian populations to African ones. Specifically, we select a representative from each African population in the Simons Genome Diversity and model migration to a French, Han, and Papuan individual in different analyses. In all cases, we initialise the inference with a symmetrical migration equal to 1.00 $4N_0$ proportion replaced per generation (henceforth, we assume an $N_0 = 14312$ **CITE?**), a choice we justify through simulation in Supplemental Simulation S4. A comparable magnitude of migration is found in Niger-Kordofanian and Nilo-Saharan populations, while analysis with Afroasiatic populations shows a sustained history of bidirectional migration consistent with the literature. San groups show a lower degree of migration, which is seen to a lower degree in Mbuti populations.

S1 Analysis of Simons Genome Diversity Panel

We download variant call format (VCF) 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. All of the following analyses were run three times.

1. Whole analysis
2. HGDP
3. HGDP subset in 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 S1). 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.

S2 Simulation procedure

S3 Population size simulations

S4 Choice of initiation parameters

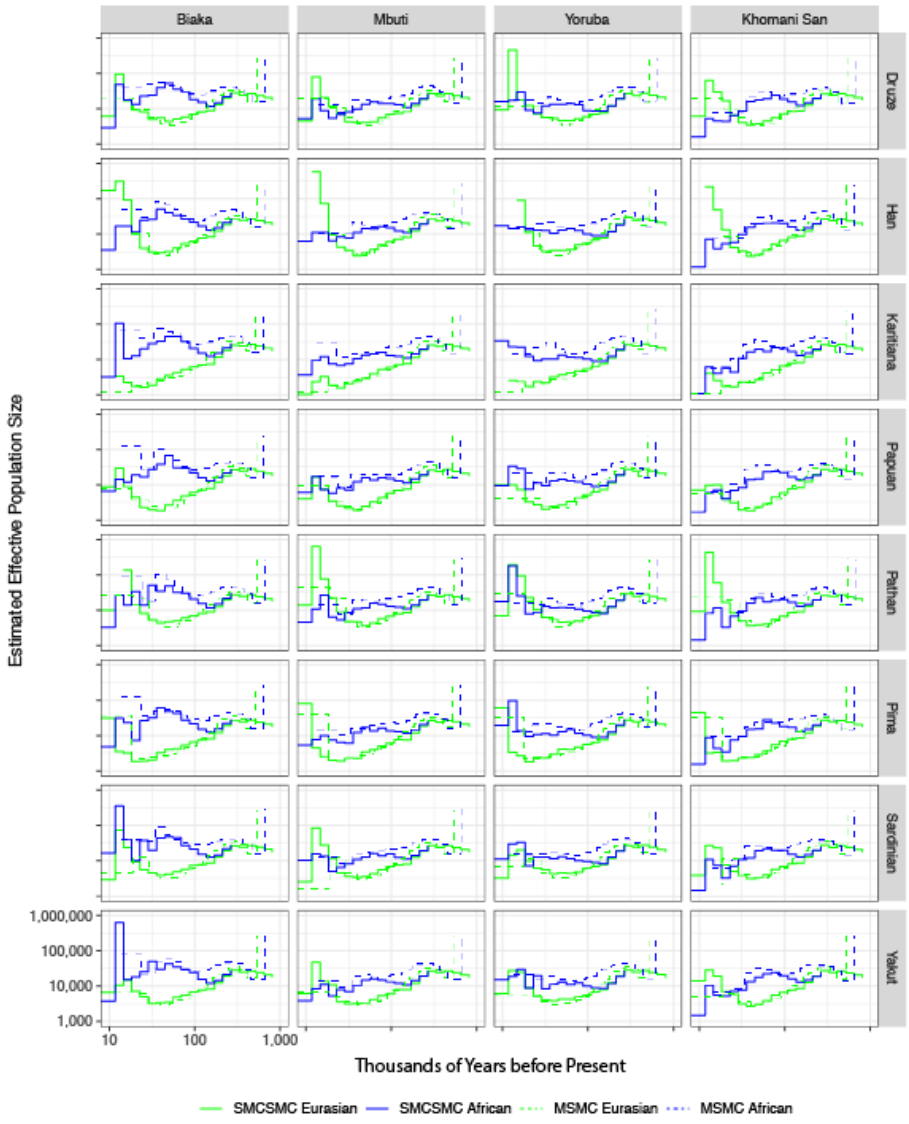

Figure S1: `smcsmc` and MSMC inferred effective population size of several populations in the Simons Genome Diversity Panel. These samples were selected to match, as closely as possible, those in the physically phased subet of the Human Genome Diversity Project panel. 10,000 particles and 25 iterations were used for `smcsmc` and 40 iterations for MSMC.