

Original Research

The reliability of inferred archaic segments

Nancy Bird ^{*}, Erin Walker, Garrett Hellenthal ^{*}

Department of Genetics, Evolution and Environment, University College London, London, UK;
Email: erin.walker.22@ucl.ac.uk

^{*} **Correspondence:** Nancy Bird; Email: nancy.bird.18@ucl.ac.uk;
Garrett Hellenthal; Email: g.hellenthal@ucl.ac.uk

Supplementary Materials

The following supplementary materials are available on the website of this paper:

Figure S1. Cartoon depicting the segment merging scheme used in cp-archaic (Methods).
Figure S2. Performance of archaic segment detection methods under a Papuan-like simulated demography.
Figure S3. Summary of ‘true’ archaic segments from the three demographic simulations in Figures 1 and S1.
Figure S4. Population-level archaic segment coverage distributions of simulated segments under different demographies.
Figure S5. Manhattan plots from 10 independent simulations of chromosome 1 of the Gower demography.
Figure S6. Relationship between local recombination rate and frequency of overlap with simulated archaic desert regions. Figure S7. Uncertainty in recall and precision estimates.
Figure S8. Summary of inferred archaic segments, using the four methods in simulated (top) and real (bottom) data.
Figure S9. Summary of inferred archaic segments, using the four methods and the ‘published’ parameters, where segments are considered at a haplotype level (lighter colours, right) and a genotype-level (merged across haplotypes, darker colors, left).
Figure S10. Overlap of archaic SNP calls across methods under optimized and published filtering schemes at SNP, window, and population levels in the Skov simulation.
Figure S11. Accuracy of inferred archaic segments across recombination rate and length bins for different methods.
Figure S12. The performance (recall and precision) of cp-archaic when varying the prior probability of archaic ancestry on the Skov simulation (2% was used for all analyses on simulated individuals).
Figure S13. Impact of phasing strategy on precision and recall for phased-based archaic inference methods in the Skov simulation.
Figure S14. Comparing the performance (recall and precision) of cp-archaic with the ChromoPainter-based method used in Jacobs et al., (2019) for chromosomes 12-22 under the Skov demography (see Methods for a description of the approaches).
Figure S15. Overlap of archaic SNP calls across methods in 1000 Genomes CEU and CHB populations at SNP, window (1Mb), and population levels.
Figure S16. Manhattan plots of archaic coverage percentage in 98 CEU 1000 genomes individuals in 10kb windows for each of the four methods (rows).
Data S1. Top 10 best F1 scores for each of the four methods under different filtering schemes and the Skov demographic simulation.
Data S2. Top 10 best F1 scores for each of the four methods under different filtering schemes and the Gower demographic simulation.

Data S3. Top 10 best F1 scores for each of the four methods under different filtering schemes and the Denisovan demographic simulation.

Data S4. Top 5 best F1 scores for cp-archaic applied to chromosome 2 and 15 of masked data, under different filtering schemes and the Skov demographic simulation.

Text S1. Msprime code used to simulate each demography.

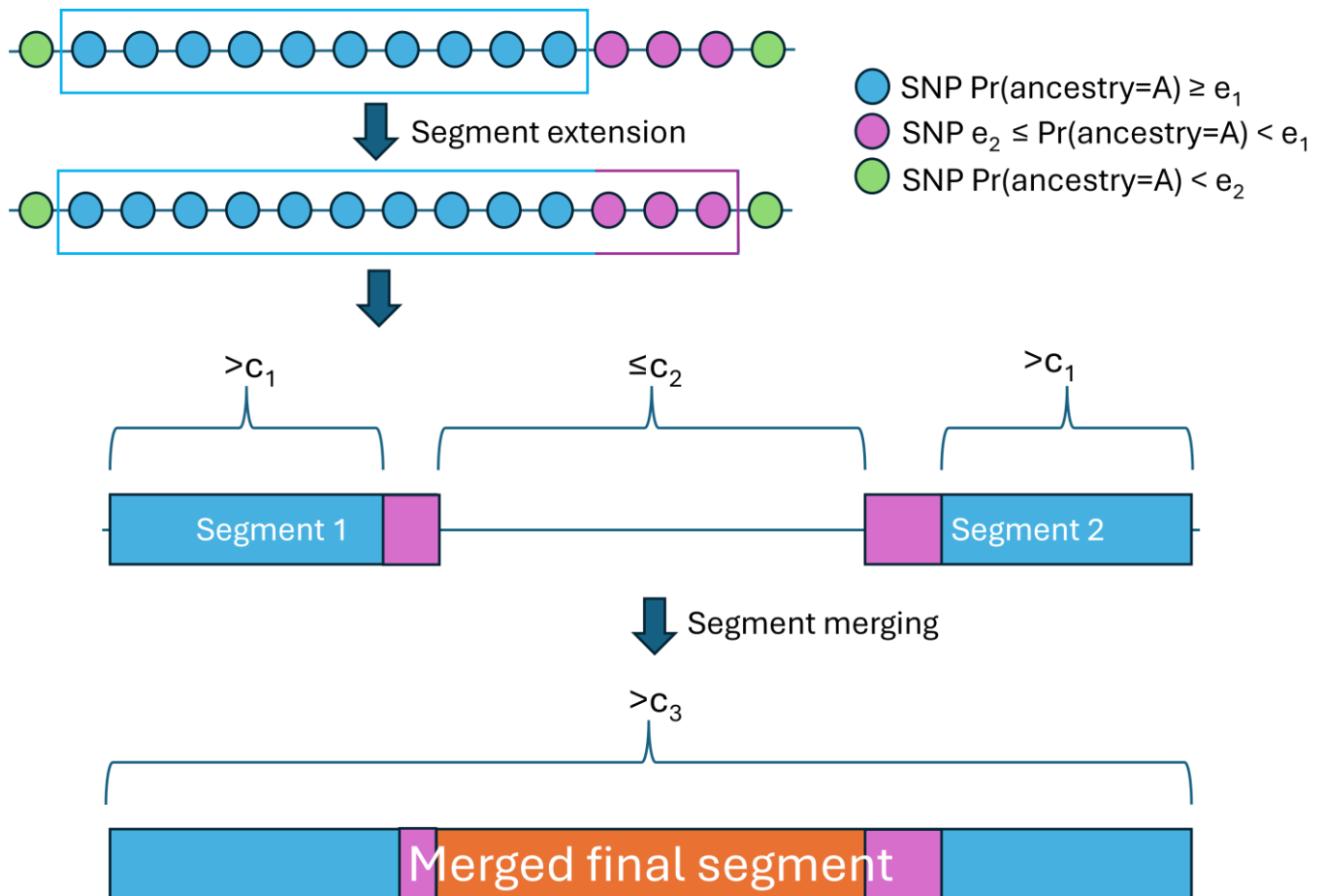


Figure S1. Cartoon depicting the segment merging scheme used in cp-archaic (Methods). We first identify segments that have ≥ 10 contiguous SNPs (circles) with $\Pr(\text{ancestry}=A)$, the probability of carrying archaic, $\geq e_1$ spanning a sequence length $> c_1$ (blue segments). These initial segments are then extended either side (purple segments) until the probability of carrying archaic falls below e_2 . Next, consecutive segments can be merged if the sequence length between, after extending, (orange segment) is $\leq c_2$. Finally, after the merging step, any segment with final sequence length $> c_3$ is retained in the final set of called archaic segments.

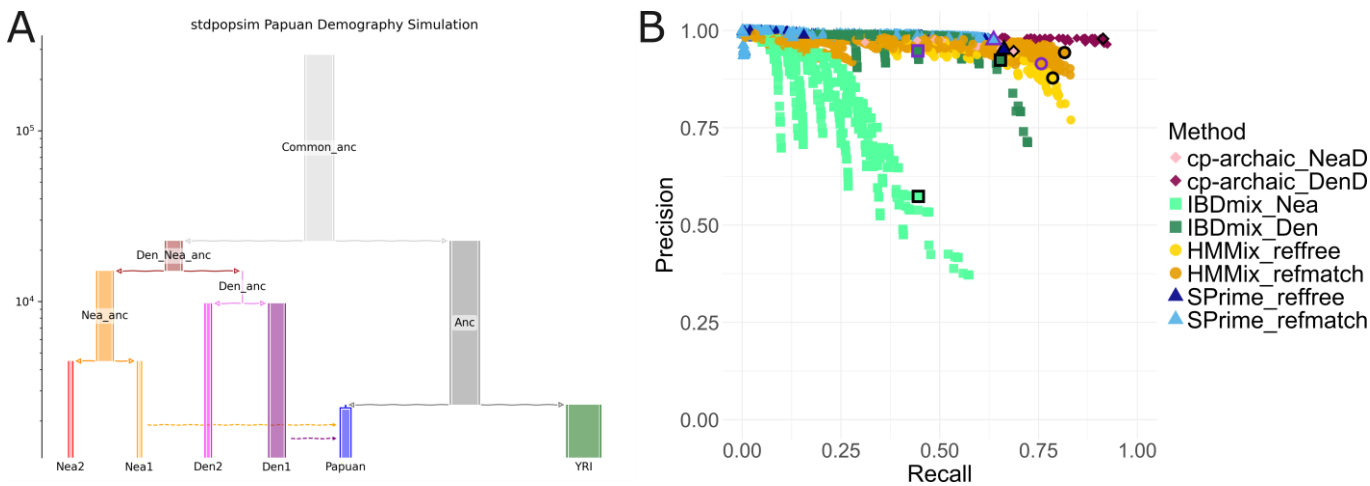


Figure S2. Performance of archaic segment detection methods under a Papuan-like simulated demography. (A) Simulated demography based on a Papuan population with parameters from stdpopsim run with msprime. Parameters available in text S1. (B) Recall and precision values for four different methods, cp-archaic, IBDmix, HMMix and SPrime run under a variety of different parameters and filters, for example minimum length, score, match rate to archaics. IBDmix was run using either a sampled individual from the Neanderthal 2 (the non-introgressing Neanderthal) population or a sampled individual from the Denisovan 2 (the non-introgressing Denisovan) population as a reference. cp-archaic was run with either three Neanderthals (two from Nea1, one from Nea1) as ‘donors’ and a non-introgressing Neanderthal (from Nea2) as the surrogate archaic (NeaD, see Methods), or two Neanderthals (from Nea 1) and the non-introgressing Denisovan (Den 2) as donors and a non-introgressing Neanderthal as the surrogate archaic (DenD). Black points indicate the best F1 score for the method. Thick purple borders indicate ‘recommended’ or commonly used parameters for published methods. These two values are the same for cp-archaic.

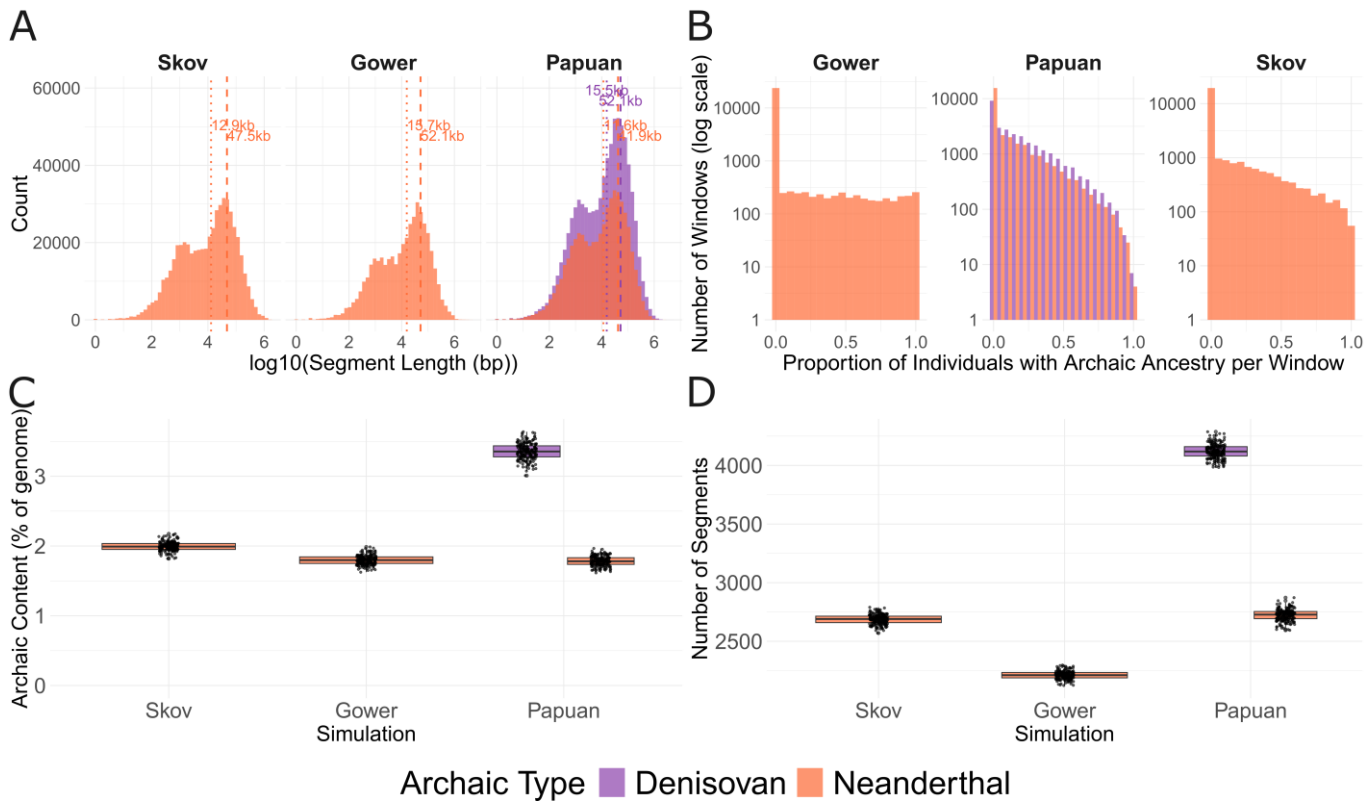
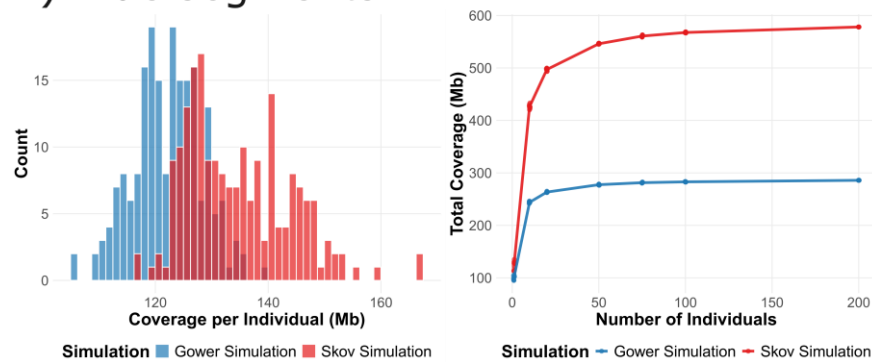


Figure S3. Summary of ‘true’ archaic segments from the three demographic simulations in Figures 1 and S1. (A) Histograms of the distribution of segment lengths in each simulation with the mean (dashed line) and median (dotted line) length labeled. (B) Histogram of the percentage of individuals (out of the 200 simulated) with any archaic segments overlapping 100kb windows across the genome. (C) Percentage archaic ancestry per individual in each of the simulations and D) number of archaic segments per individual in each of the simulations.

A) True segments



B) Inferred segments (Skov simulation)

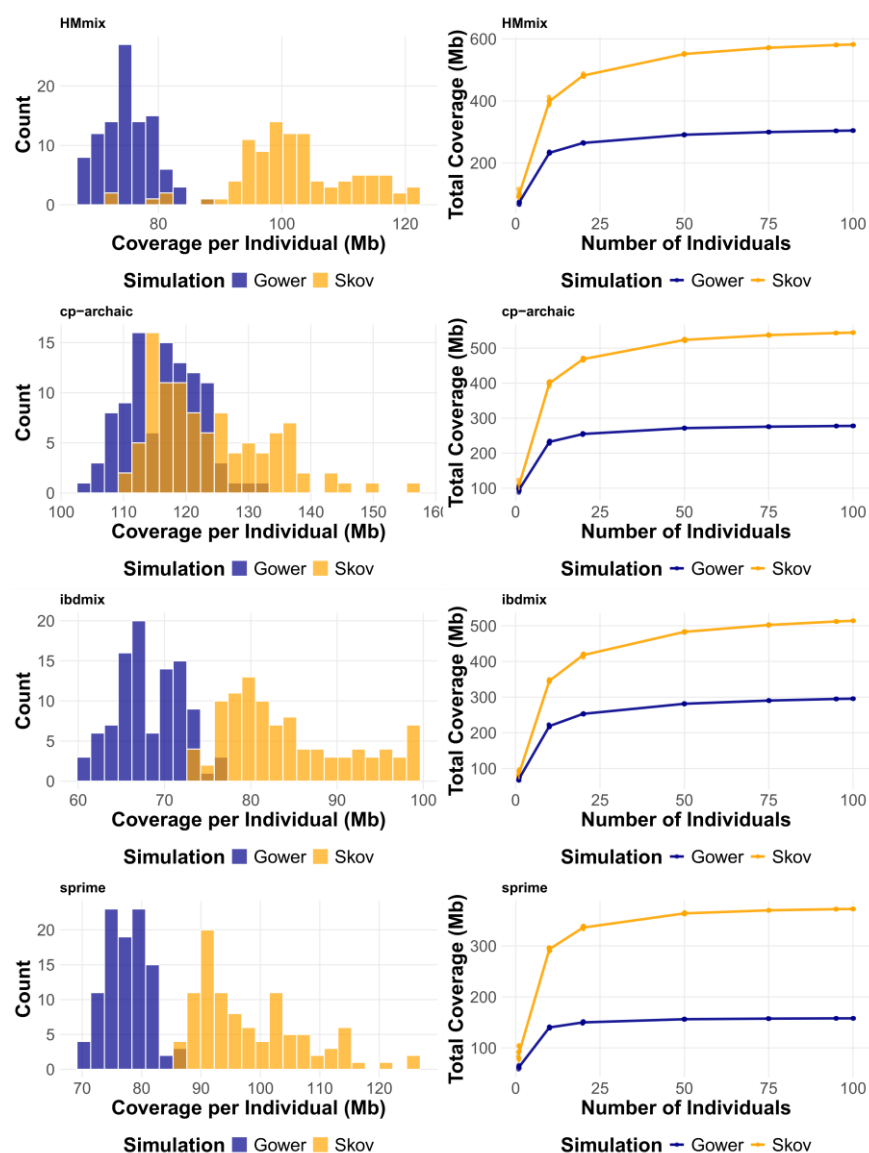


Figure S4. Population-level archaic segment coverage distributions of simulated segments under different demographies. showing the 'true' coverage of archaic ancestry in each of the Skov (red) and Gower (blue) demography simulations (A), and the inferred coverage of archaic ancestry by each of the four methods of the Skov (yellow) and Gower (blue) demography simulations and the 'best' filtering parameters (B). Left is a histogram of individual archaic coverage per individual, and right shows the total coverage when combining archaic segments from different numbers of individuals, mimicking the plots for 1000 Genome data in Figure 2.

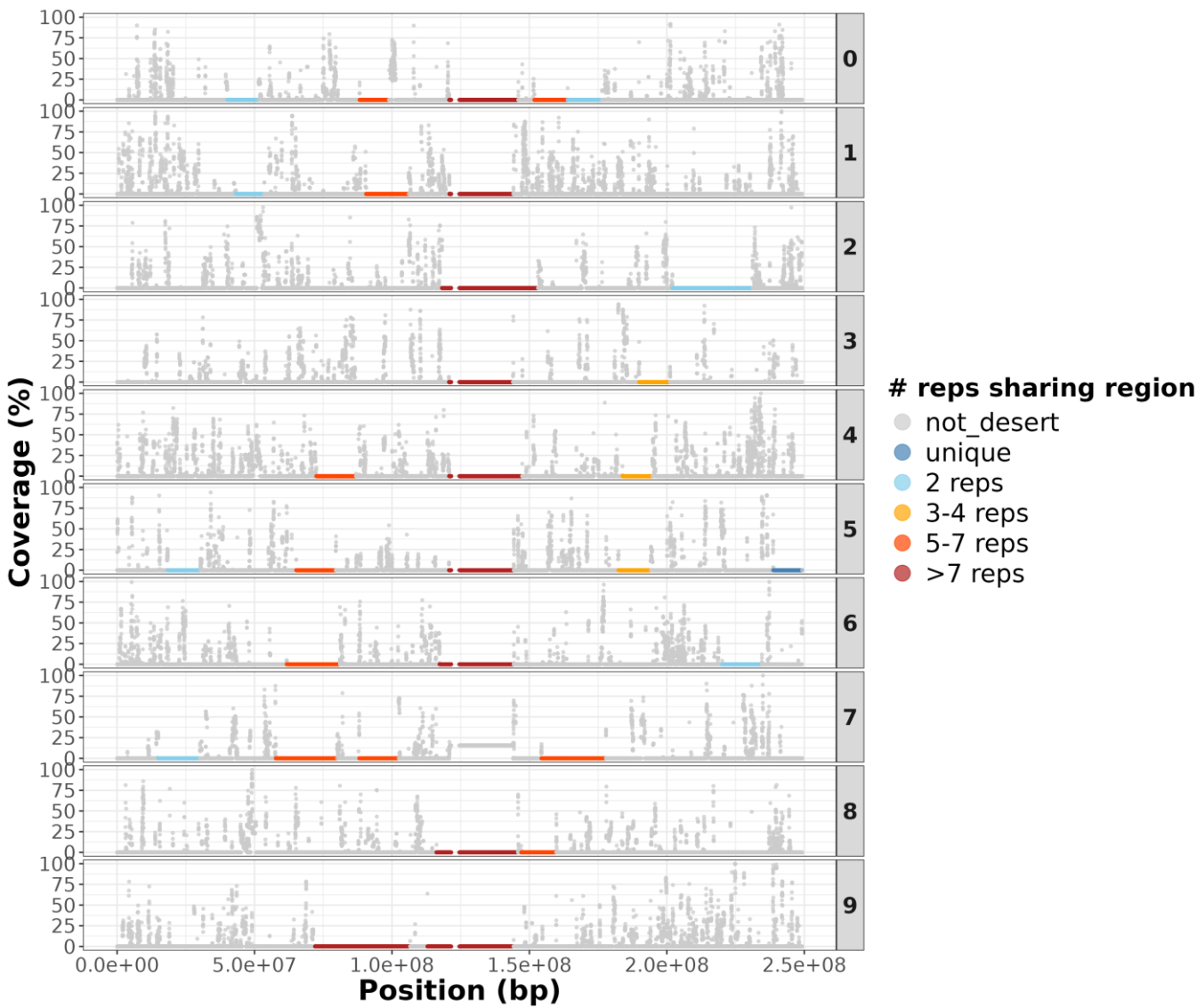


Figure S5. Manhattan plots from 10 independent simulations of chromosome 1 of the Gower demography. Desert regions (>10Mb with 0% archaic ancestry) are highlighted in colour, depending on the number of desert regions in other simulations that they overlap with.

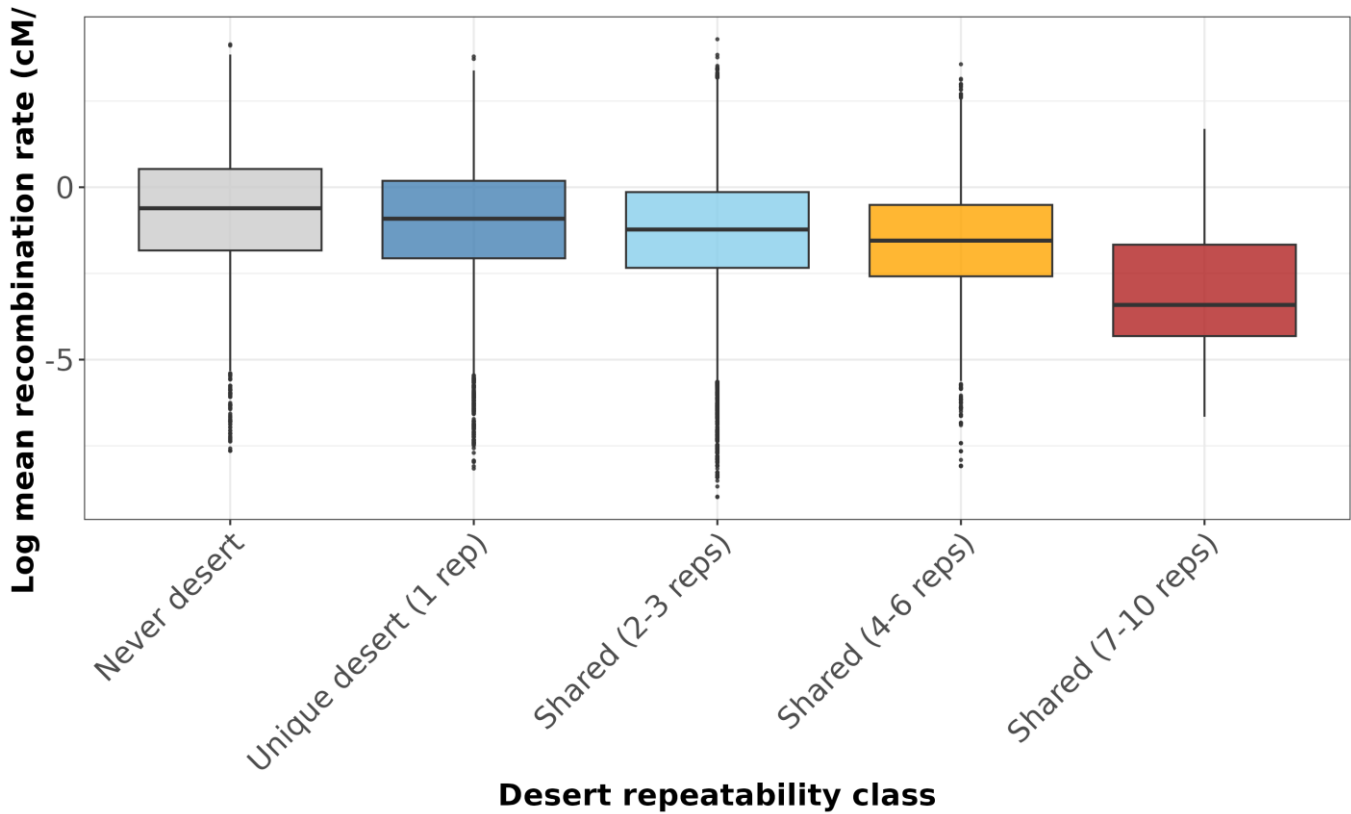


Figure S6. Relationship between local recombination rate and frequency of overlap with simulated archaic desert regions. . Log mean recombination rate of 10kb non-overlapping window regions plotted against how often the window overlaps an archaic desert region in each of the 10 independent simulation repeats from **Figure S5**.

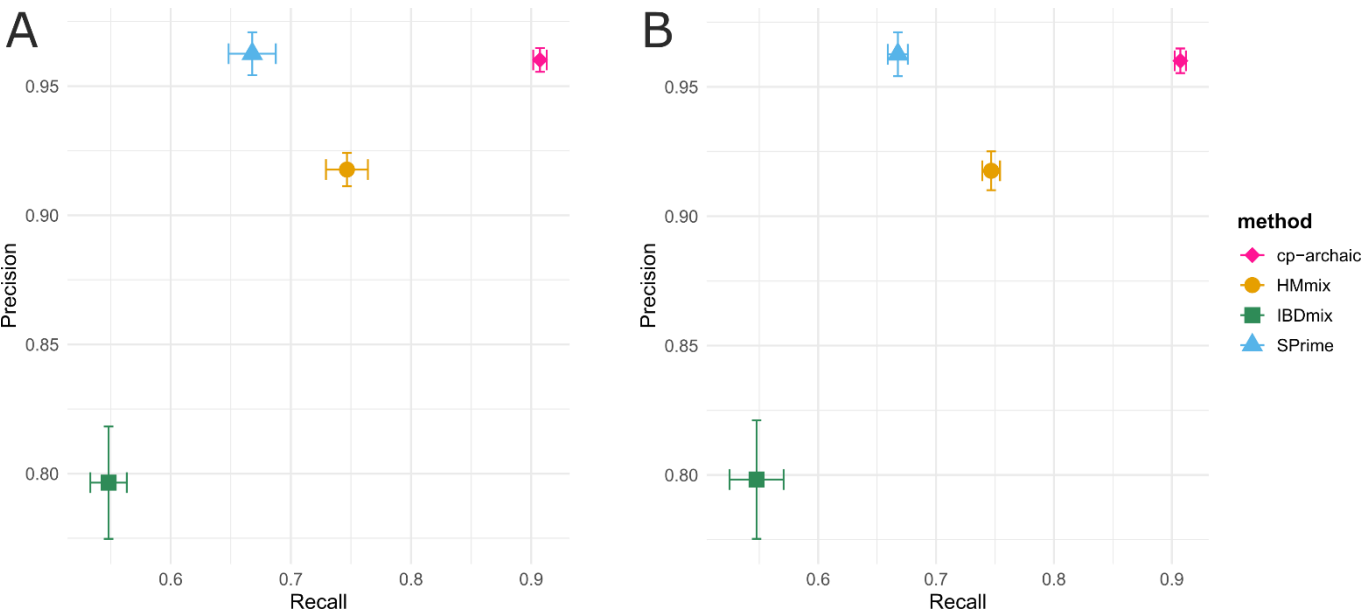


Figure S7. Uncertainty in recall and precision estimates. . Recall and precision confidence intervals calculated for the 'best' parameters for each method using (A) chromosome jackknifing [1] or (B) resampling 10 chromosomes for 100 repeats.

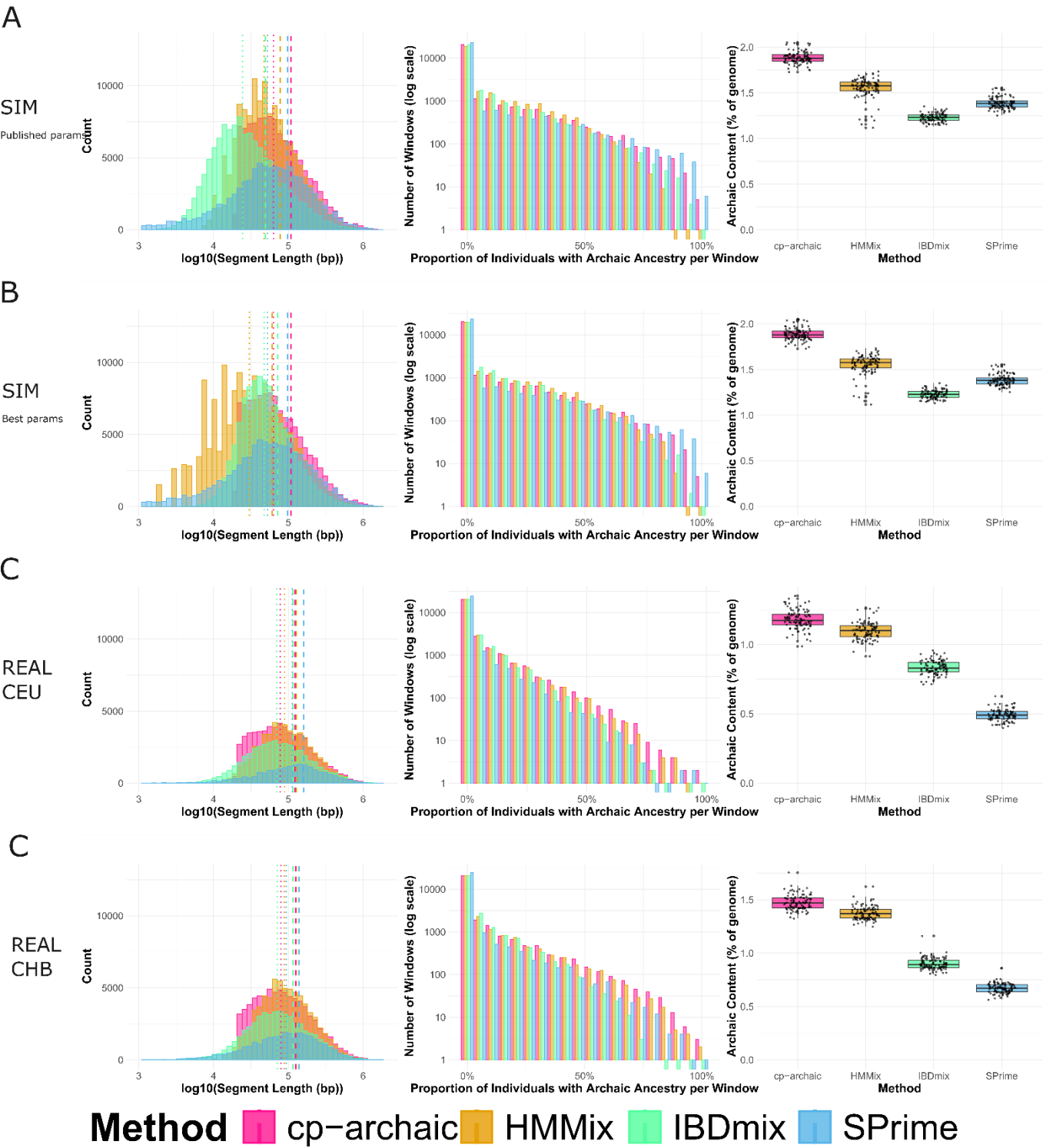


Figure S8. Summary of inferred archaic segments, using the four methods in simulated (top) and real (bottom) data. (A) and (B): segments inferred from the Skov demographic simulation in Figure 1A with (A) parameters used in published papers or with (B) the ‘best’ parameters (black outlined points in Figure 1B). (C) and (D): segments inferred in 1000 Genomes populations (C) CEU and (D) CHB. The first column plots are histograms of the distribution of segment lengths with the mean (dashed line) and median (dotted line) lengths highlighted. The second column shows histograms of the percentage of individuals with any archaic segments overlapping non-overlapping 100kb windows across the genome. A distribution skewed towards the left indicates that in most windows, few individuals have archaic ancestry. Though note that there are dozens of 100-kb windows where >80% of individuals are inferred by approaches to carry archaic ancestry, despite simulating under a neutral model (see also C and D showing similar high proportions for true segments). The final column shows boxplots of the percentage of archaic ancestry per individual in each of the simulated or real populations.

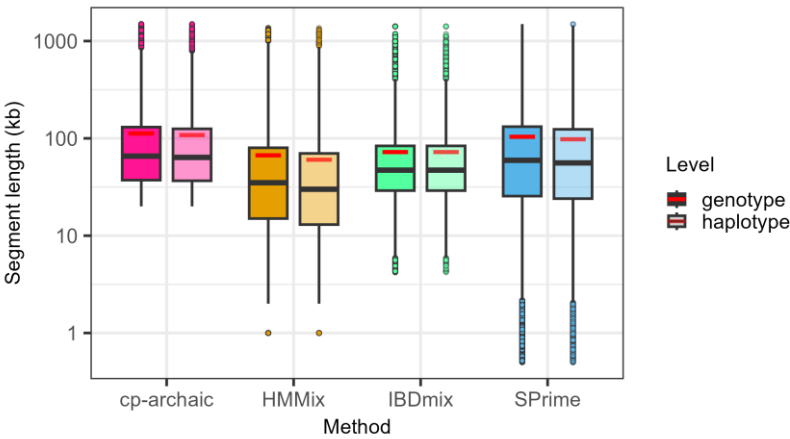


Figure S9. Summary of inferred archaic segments, using the four methods and the ‘published’ parameters, where segments are considered at a haplotype level (lighter colours, right) and a genotype-level (merged across haplotypes, darker colors, left). Boxplots of the distribution of segment lengths with the mean (red line) lengths highlighted.

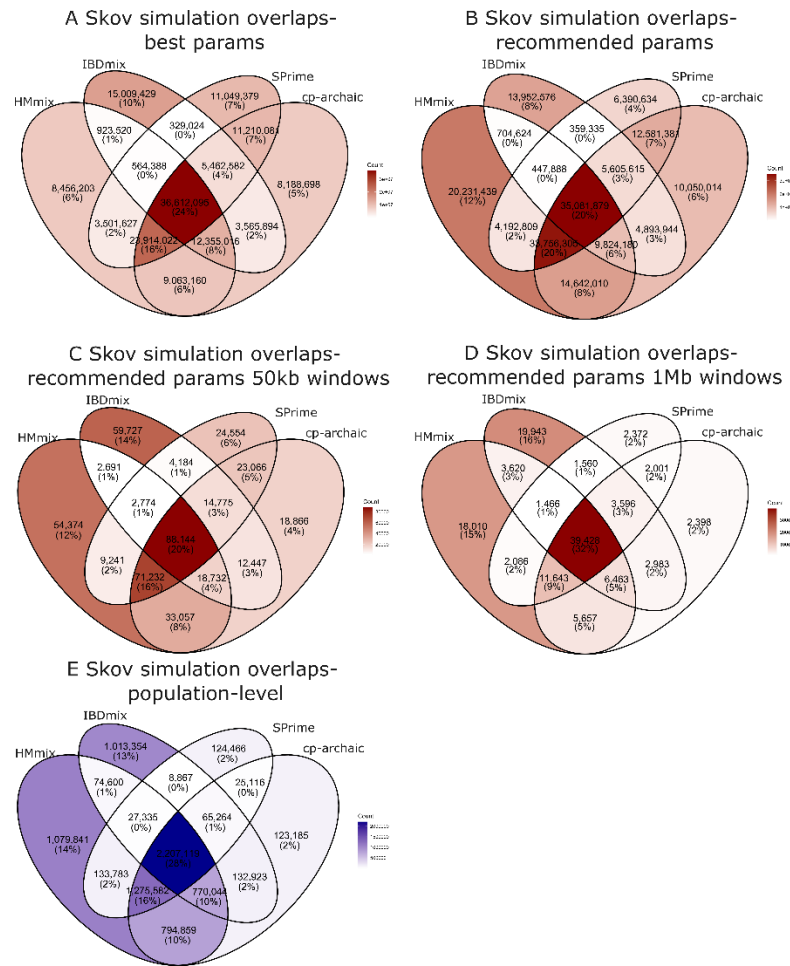


Figure S10. Overlap of archaic SNP calls across methods under optimized and published filtering schemes at SNP, window, and population levels in the Skov simulation. Venn diagram showing the SNP overlap of the four methods, given the SNP is called archaic in any of the methods for Skov demography simulation for 100 of the 200 simulated individuals, under (A) our best inferred filter conditions and (B) recommended previously used and published filters (see methods). (C) and (D) show percentage overlap under previously published filters for (C) 50kb windows and (D) 1Mb windows, where a window with any overlap with an archaic segment was counted as archaic. (E) shows population level SNP overlap, counting an overlap as a site called archaic by that method in any of the simulated individuals.

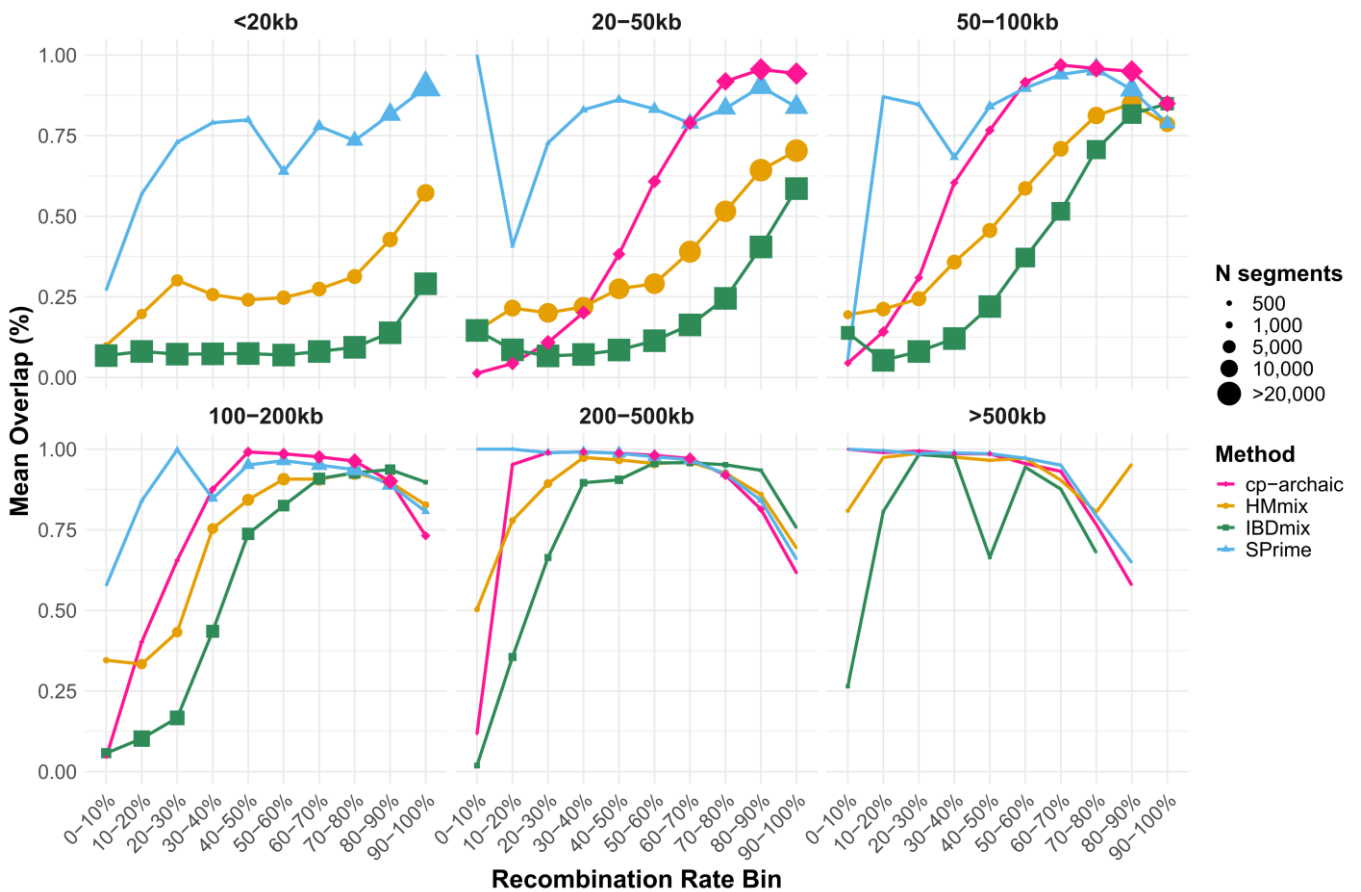


Figure S11. Accuracy of inferred archaic segments across recombination rate and length bins for different methods. . Plots showing mean percentage overlap with a true segment in each of the four methods (colours/shapes) for different recombination rate bins, where each plot shows segments from a different inferred segment length length bin. All inferred segments are shown, except for IBDmix where they are filtered for a LOD > 4, cp-archaic where the standard merging scheme leaves only segments >20kb, and HMmix where they are filtered for mean probability > 0.9 (following recommended filter parameters but ignoring length filters to show patterns). Size of point shows number of segments.

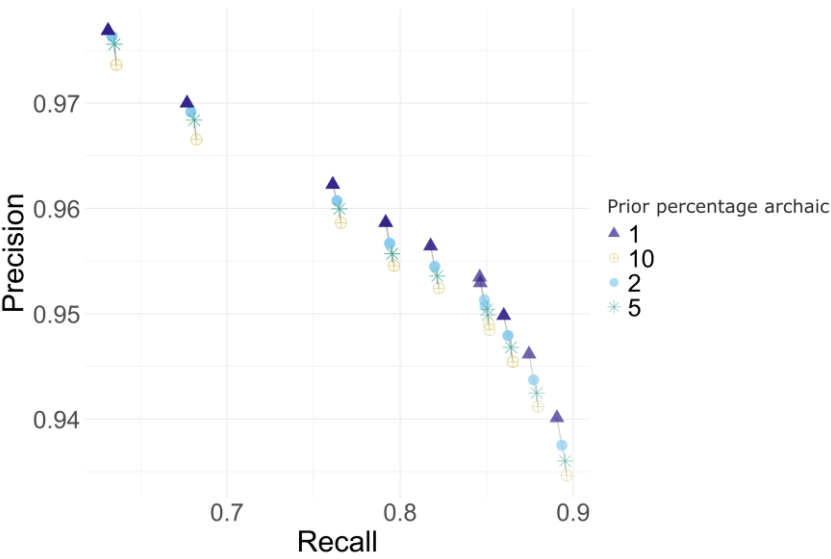


Figure S12. The performance (recall and precision) of cp-archaic when varying the prior probability of archaic ancestry on the Skov simulation (2% was used for all analyses on simulated individuals).

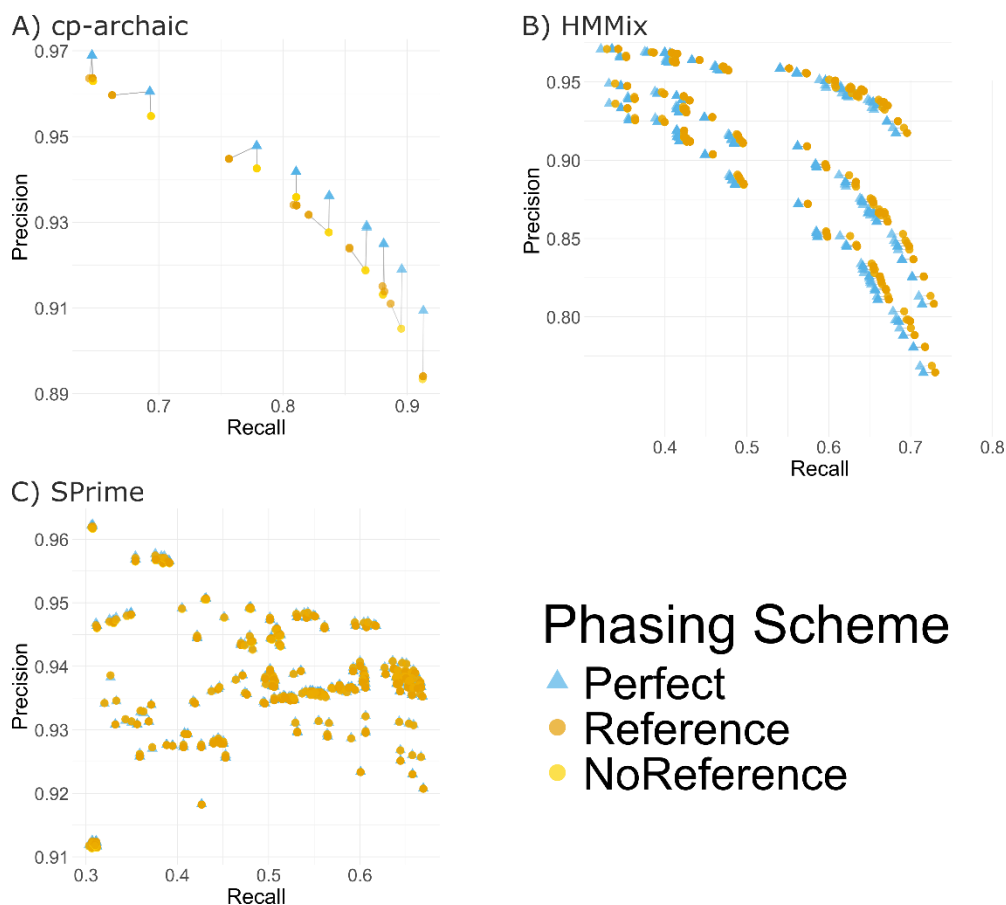


Figure S13. Impact of phasing strategy on precision and recall for phased-based archaic inference methods in the Skov simulation. The impact of phasing errors on the three methods that utilise phased data, cp-archaic (A), HMMix (B) and SPrime (C) on the Skov simulation. The simulated individuals (including the archaics) were phased either using a reference (comprising other target and outgroup simulated individuals, orange circle) or completely reference free (yellow circle, see Methods) and precision and recall were re-calculated.

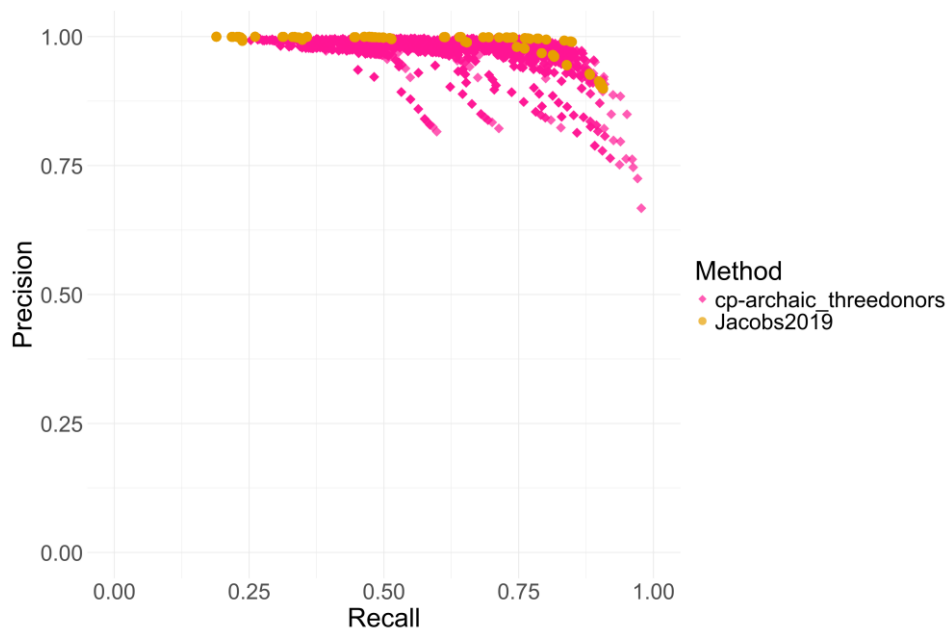


Figure S14. Comparing the performance (recall and precision) of cp-archaic with the ChromoPainter-based method used in Jacobs et al., (2019) [2] for chromosomes 12-22 under the Skov demography (see Methods for a description of the approaches).

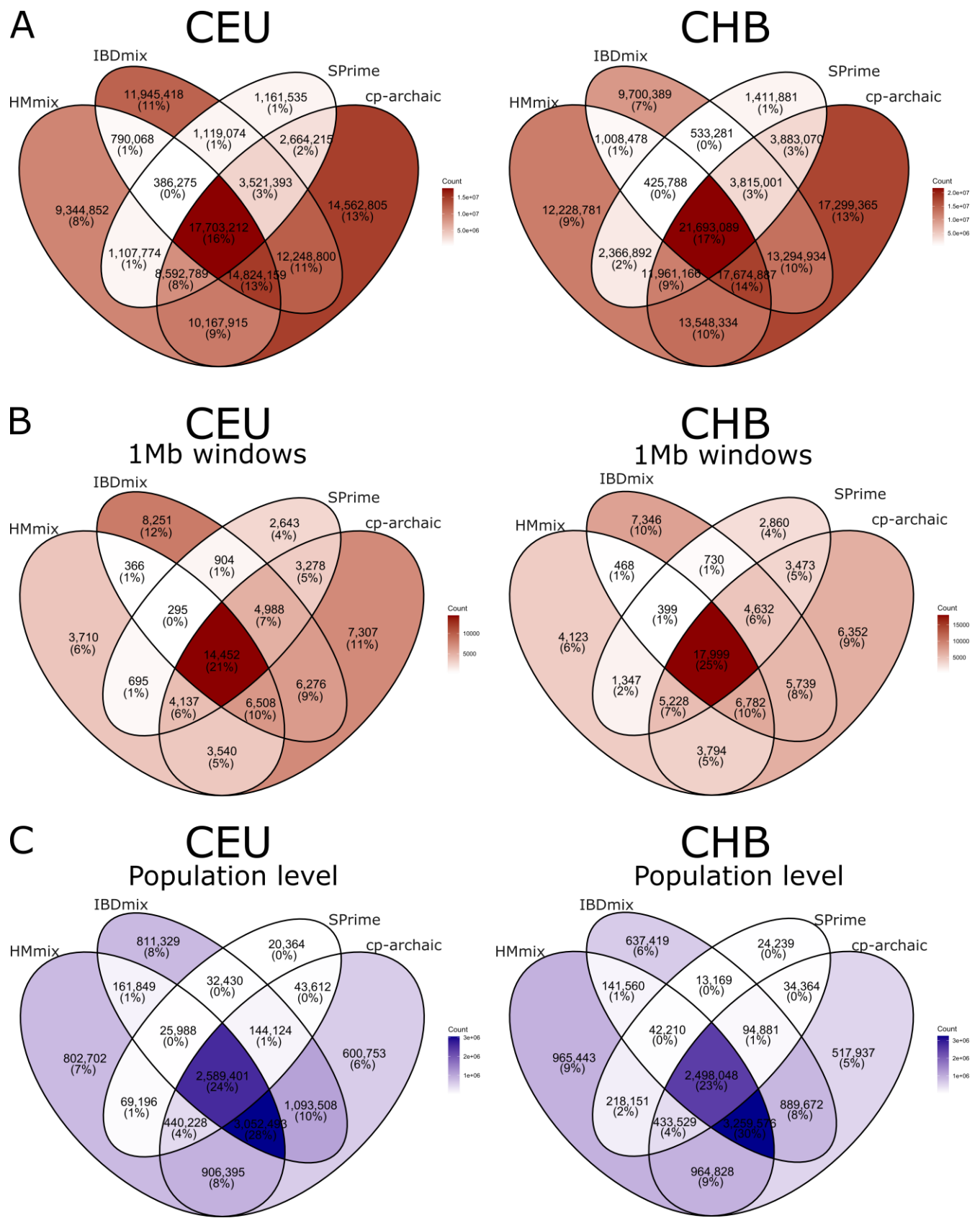


Figure S15. Overlap of archaic SNP calls across methods in 1000 Genomes CEU and CHB populations at SNP, window (1Mb), and population levels. (A) Venn diagrams showing the SNP overlap of the four methods, given the SNP is called archaic in any of the methods for 1000 genomes populations CEU and CHB. (B) show percentage overlap for 1Mb windows, where a window with any overlap with an archaic segment was counted as archaic. (C) shows population level SNP overlap, counting an overlap as a site called archaic by that method in any of the simulated individuals.

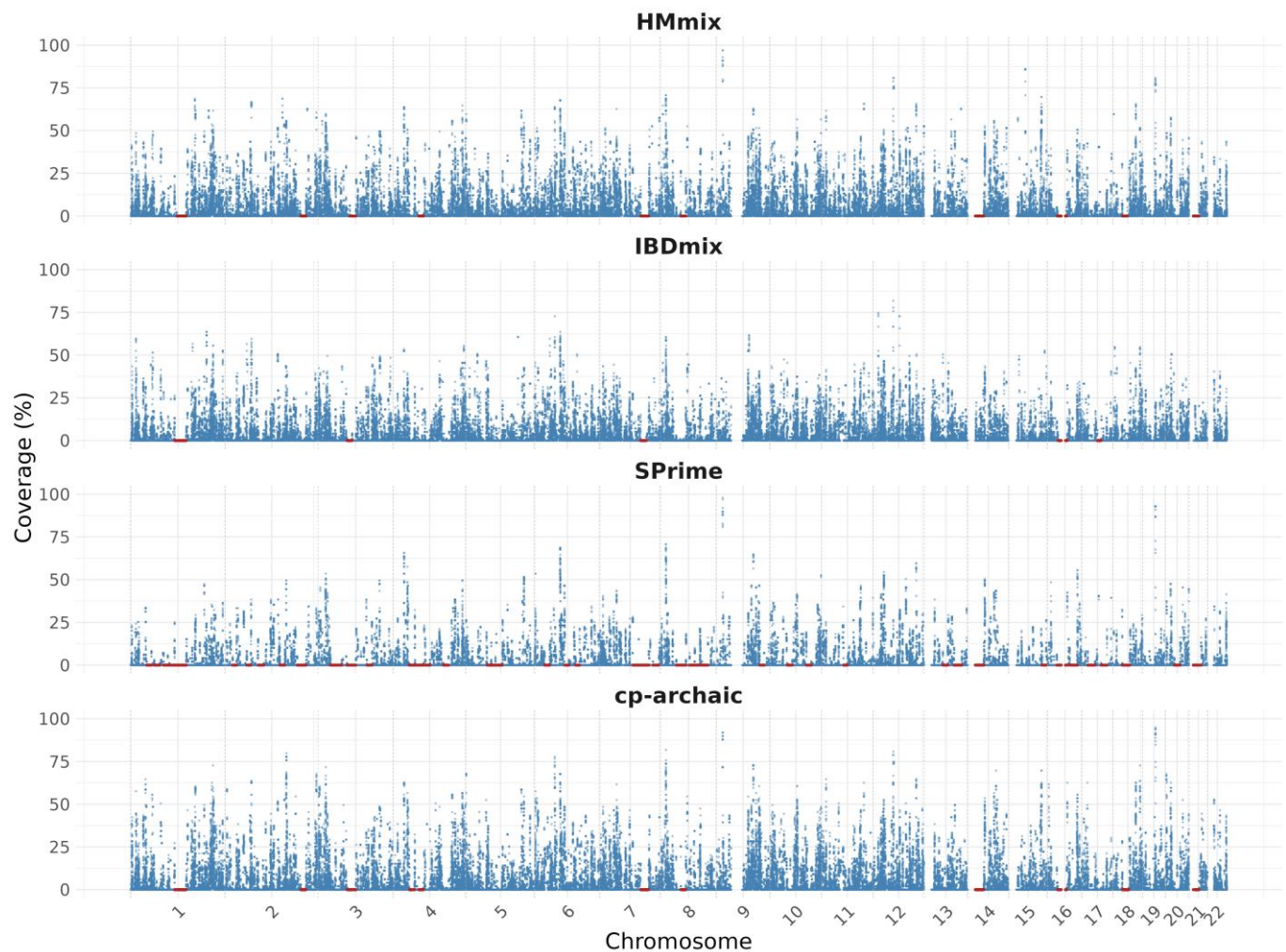


Figure S16. Manhattan plots of archaic coverage percentage in 98 CEU 1000 genomes individuals in 10kb windows for each of the four methods (rows). Archaic deserts (region $\geq 10\text{Mb}$ with $<0.1\%$ archaic ancestry) are highlighted in red, although these often overlap with centromeric regions. cp-archaic infers 11 deserts that do not overlap with the centromeres, IBDmix 5 HMmix 10 and SPrime 45.

References

1. Busing FMTA, Meijer E, Leeden R van der. Delete-m Jackknife for Unequal m. Stat Comput. 1999;9(1):3-8. [DOI](#)
2. Jacobs GS, Hudjashov G, Saag L, Kusuma P, Darusallam CC, Lawson DJ, et al. Multiple Deeply Divergent Denisovan Ancestries in Papuans. Cell. 2019;177(4):1010-1021.e32. [DOI](#)