File S1
Supporting Information

1 Simulation details

The following `ms` commands were used to simulate data under three population size change histories:

S1: `ms 10 1 -T -r 10000 1000000 -eN 0.05 0.1 -eN 0.2 0.5 -eN 0.5 1.25`
S2: `ms 10 1 -T -r 10000 1000000 -eN 0 10 -eN 0.05 0.1 -eN 0.2 0.5 -eN 0.5 1.25`

Note that `ms` times are in units of $4N_0$ generations, so we multiplied the raw times above by 2 to compare to PSMC and our method diCal. Mutation rates were not specified above, since the only `ms` output used was tree at each base (`-T` flag). Mutations were then added to the trees using a finite sites model, the mutation matrix in Table 1, and a mutation rate $\theta = 0.01 \times 1.443$. The factor of 1.443 accounts for the fact that this mutation matrix allows mutations that do not actually change the base (i.e., an A → A transition); see Chan et al. (2012) for further explanation. This mutation matrix was also used for the real data analysis.

The following style of command was used to run PSMC. We used 20 iterations as described in the PSMC paper (Li and Durbin, 2011), and the same pattern of parameters we used for diCal:

```
psmc -p 3+2+2+2+2+2+3 -t 7 -N 20 -r 1 -o output.psmc input.psmcfa
```

To run our method on simulated data, the following style of command was used:

```
java -Xmx25G -d64 diCal_EM -i input.fasta -p params.txt -n 9 -t 5 -a "3 2 2 2 2 2 3"
```

The parameter file includes the number of loci in each sequence, the number of alleles (4 in our case), an estimate of the mutation rate, mutation matrix, and recombination rate, and the discretization. The `-n` flag specifies the number of haplotypes to use in the trunk, so there are $n + 1$ total. The `-t` flag specifies the number of threads to use; memory requirements scale linearly with this parameter. If `-t 1` was specified in the case, then `-Xmx5G` could be used for the memory requirement. The `-a` flag specifies the pattern of parameters, in an analogous fashion to PSMC.

To run our method on real data, the following style of command was used:

```
java -Xmx20G -d64 diCal_EM -i input.fasta -p params.txt -n 9 -t 2 -r 1.25 -a "4 2 2 2 2 2 2"
```

Table 1: Mutation matrix for realistic human data. The rows represent the original base, and the columns represent the mutated base.

base	A	C	G	T
A	0.503	0.082	0.315	0.100
C	0.186	0.002	0.158	0.655
G	0.654	0.158	0	0.189
T	0.097	0.303	0.085	0.515

The `-r` flag specifies the $T_{\max}$ (analogous to the `-t` flag for PSMC), since for humans we know the approximate date range of interest. For the real data we used a longer sequence, so the memory requirements scale accordingly (linearly).

2 Comparison of diCal to PSMC

Although diCal and PSMC are both implementations of the sequentially Markov coalescent in a discrete-time framework, they have significant differences that must be considered when comparing results from the two programs. One difference is that PSMC scales all population sizes with respect to an inferred parameter $\theta_{\text{psmc}} = 4N_{\text{psmc}}\mu$. In contrast, diCal scales population sizes with respect to a fixed input $\theta_{\text{smcsd}} = 4N_{\text{smcsd}}\mu$. Neither θ is right or wrong, they are just scaled with respect to a different N_0. If we arbitrarily set $N_{\text{smcsd}} = 1$, then

$$N_{\text{psmc}} = \theta_{\text{psmc}}/\theta_{\text{smcsd}}$$

Thus when analyzing the results, we multiplied the PSMC sizes and times by N_{psmc}. We also multiplied the `ms` times by 2, since they are in units of $4N_0$ generations.

To compare the performance of the two programs fairly, we gave both PSMC and diCal the same amount of data. Specifically, we compared the performance of diCal with a n-sequence leave-one-out scheme to the performance of PSMC with the same n sequences, but paired up sequentially (i.e. sequence 1 with 2, sequence 3 with 4, etc).

References

Chan, A. H., Jenkins, P. A., and Song, Y. S. 2012. Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*. *PLoS Genet.*, **8,**(12) e1003090.

Li, H. and Durbin, R. 2011. Inference of human population history from individual whole-genome sequences. *Nature*, **10,** 1–5.