Learning history from the SFS of low- and high-coverage ancient DNA

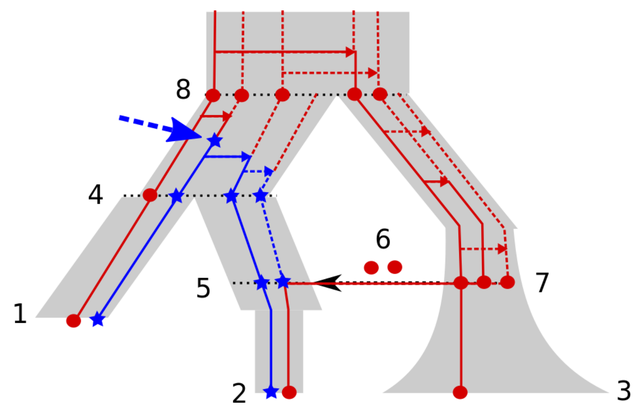
Jack Kamm, Richard Durbin

Human Genetics Programme, Wellcome Trust Sanger Institute
Department of Genetics, University of Cambridge

Joint site frequency spectrum (SFS)

- Distn. of SNP derived allele counts **x**
- Can be used to infer demographic history
- Computed under coalescent model

$$\mathbf{x} = (x_1, x_2, x_3) = (1, 1, 0)$$
$$\mathbf{n} = (n_1, n_2, n_3) = (2, 2, 1)$$
$$\text{SFS} = \mathbb{P}(\mathbf{x})$$



Ancient DNA

Added resolution and power

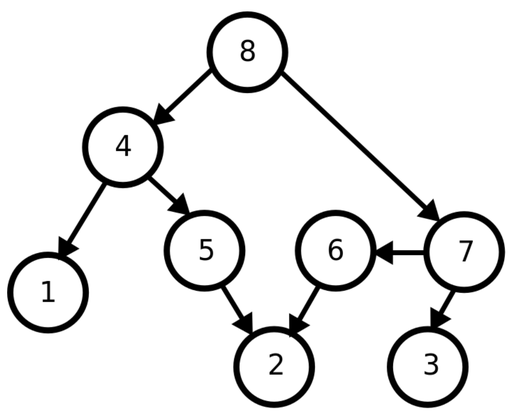
- See further into the past
- Can estimate dates without mutation rates
- Modern pops can be modeled as mixture of ancient pops

But errors, bias distort SFS

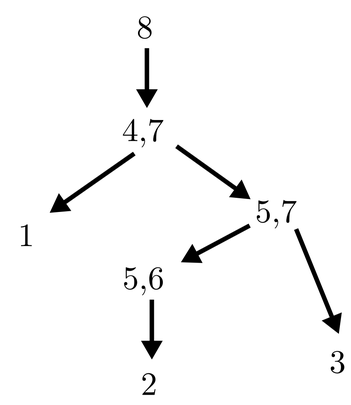
- Differences in coverage
  - Different rates of variant discovery
- Deamination

`momi`: compute SFS using Moran model + Bayesian graph

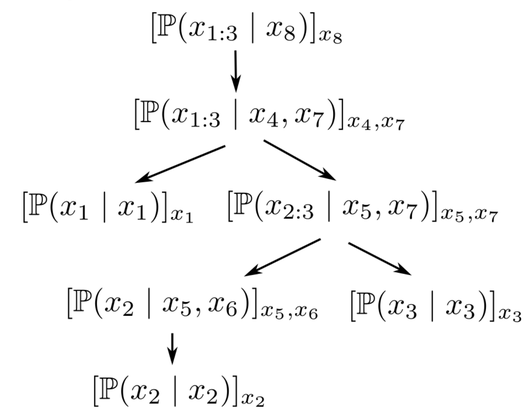
1. Represent demography as DAG



2. Convert DAG to tree



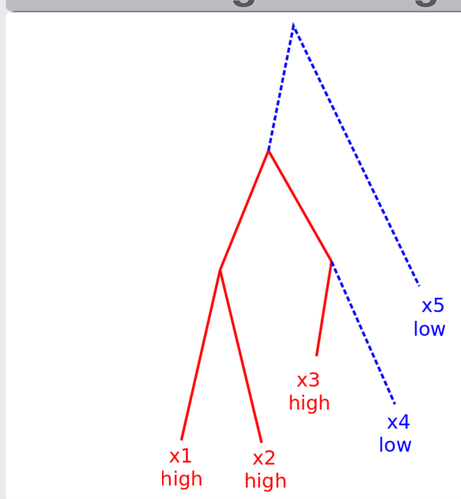
3. Compute likelihoods at each node



Likelihoods and other statistics

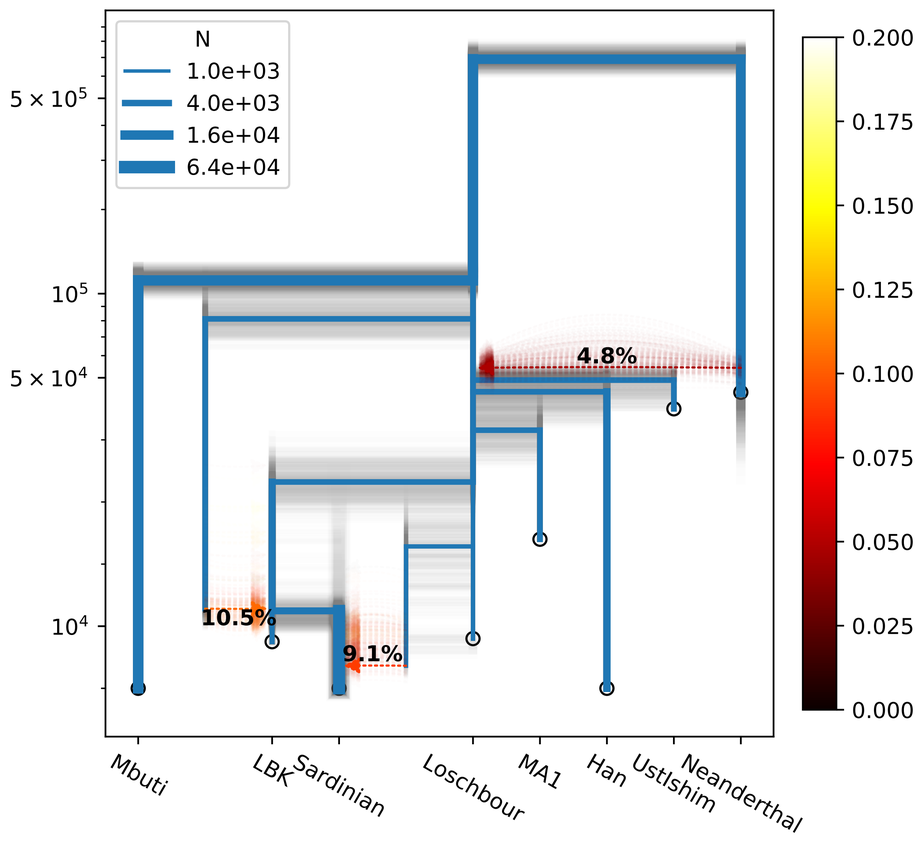
- `momi` algorithm is a DP on the tree
- To compute $\mathbb{P}(x_1, x_2, x_3)$:
  - set leafs to $\mathbf{e}_{x_1}, \mathbf{e}_{x_2}, \mathbf{e}_{x_3}$
  - Propagate likelihoods up tree
    - "Tree-peeling"
- Also efficiently compute:
  - Total branch length
  - TMRCA
  - f_2, f_3, f_4 statistics
  - F_{ST}
  - Tajima's D
  - $\mathbb{E}[f(X_1)g(X_2)h(X_3)]$

Correcting Coverage & Ascertainment Bias



- Ascertain in high-coverage
- Random allele call in low coverage
- Normalize by subtree branch length to compute conditional probability $\mathbb{P}(\mathbf{x} \mid x_1, x_2, x_3 \text{ polymorphic})$

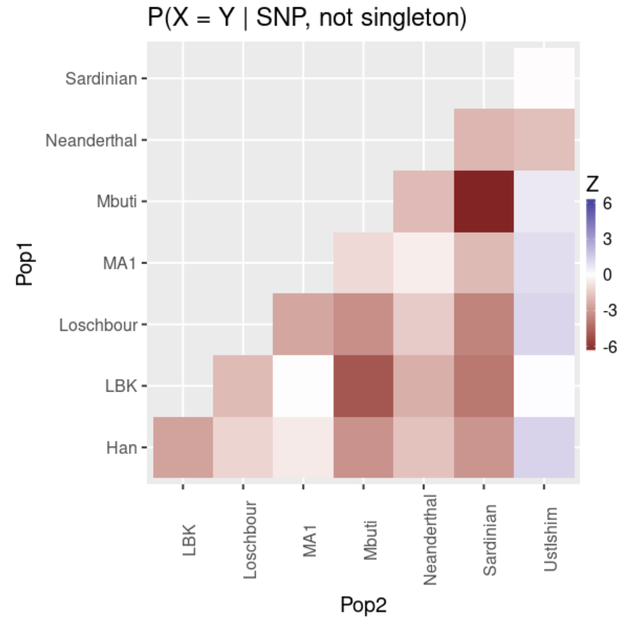
Estimating "Basal Eurasian" gene flow



- Lazaridis et al 2014, 2016: $f_4(HG, Farmer; East, Out) < 0$;
  - Posit "Basal Eurasian" component in Europe/Middle East
- We estimated parameters for 8 population model with SFS
- Used 300 nonparametric bootstraps for confidence intervals

Assessing model fit

- ABBA/BABA statistics
  - Like *qpGraph* (Patterson et al., 2012), but with recent mutations
  - Compare model expectation to observed ABBA/BABA
    - Basal Eurasian model fits reasonably well
    - $|Z| < 3.2$ ($p > .05$ after Bonferonni)
- Pairwise similarity
  - Comparing expectation to observed, Mbuti and Sardinian have excess pairwise similarity; UstIshim has excess dissimilarity



Mutation rate estimation

We used the expected within-population nucleotide diversity to estimate mutation rate. We estimated 1.11 to 1.26 $\times 10^{-8}$ depending on which populations we included in the model.

This research is supported by the Wellcome Trust (grant 206194) and the Human Frontiers Science Program (LT000402/2017).

{jk21,rd}@sanger.ac.uk