The pairwise sequentially Markovian coalescent

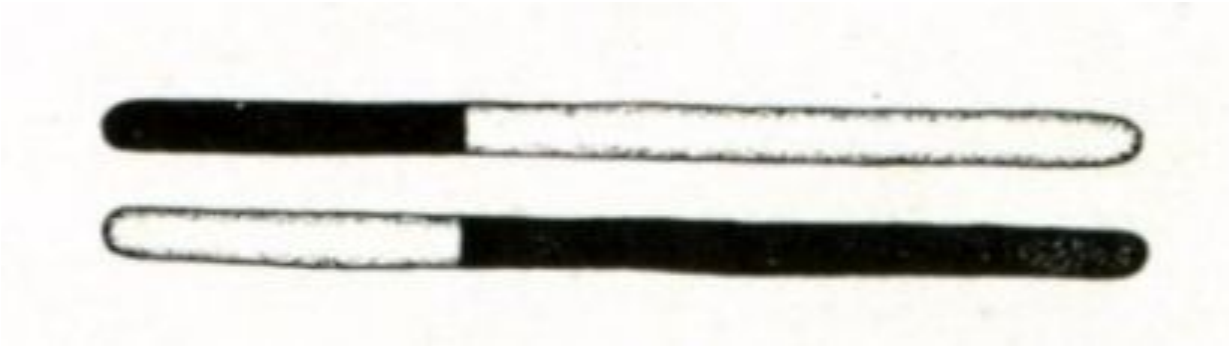
A seminar discussion on “Inference of human population history from individual whole-genome sequences” (Li & Durbin, 2011)

by: Hassan Nasif and Seth Temple

UNIVERSITY *of* WASHINGTON

Overview

- Infer bottlenecks and expansions in *deep time*
- Each haploid in a diploid organism is a history of recombination events
- The *local density of heterozygous sites* along a chromosome reflects these recombinations
- For each site, we can infer the time at which the two alleles of the diploid organism coalesce

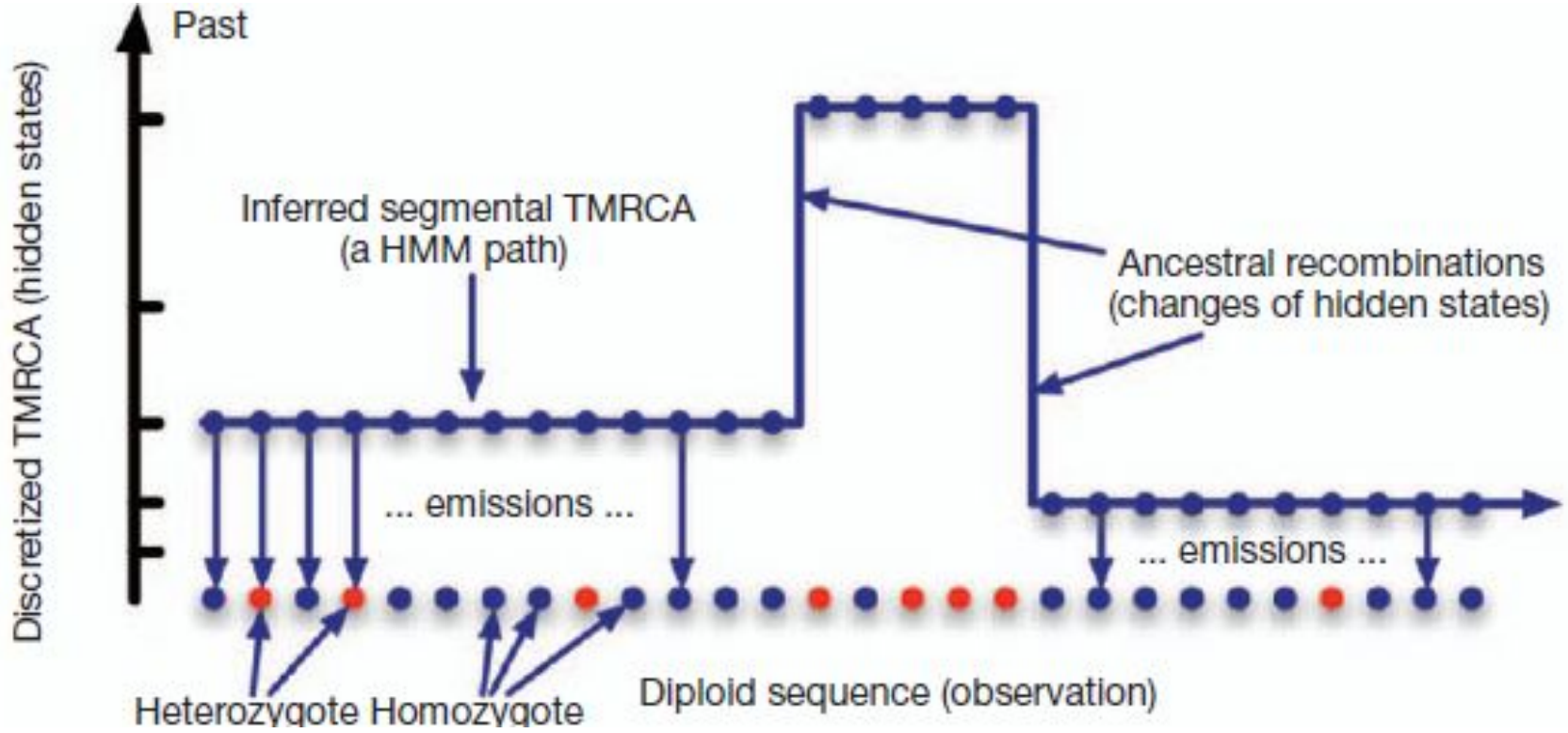


Methods

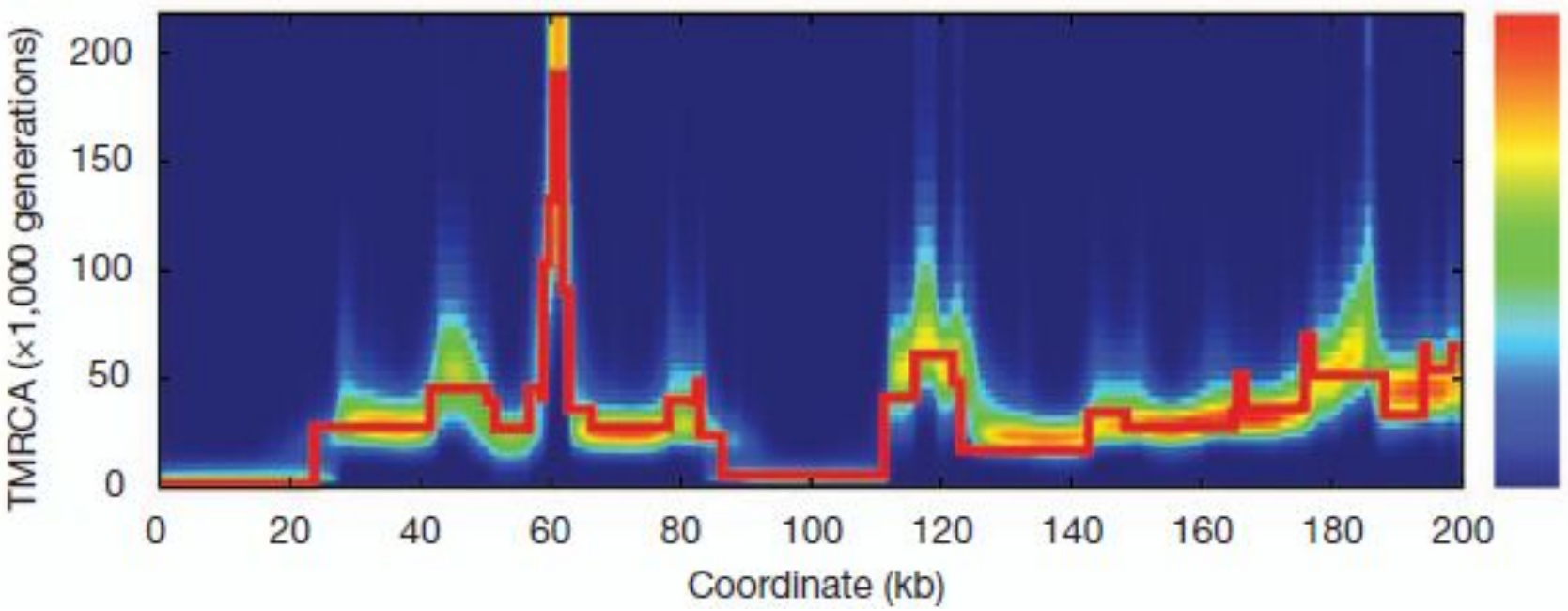
- Align a diploid sequence to a reference genome
- Tag each pair of alleles as 0 (homozygote), 1 (heterozygote), or . (missing)
- Set up a hidden Markov model where the observed states are as above and the hidden states are discretized TMRCAs
- Run the Baum-Welch EM algorithm 20 times

- Bootstrap sequences to measure variability
- Use WF-based backward simulator to show proof of concept on simulated data

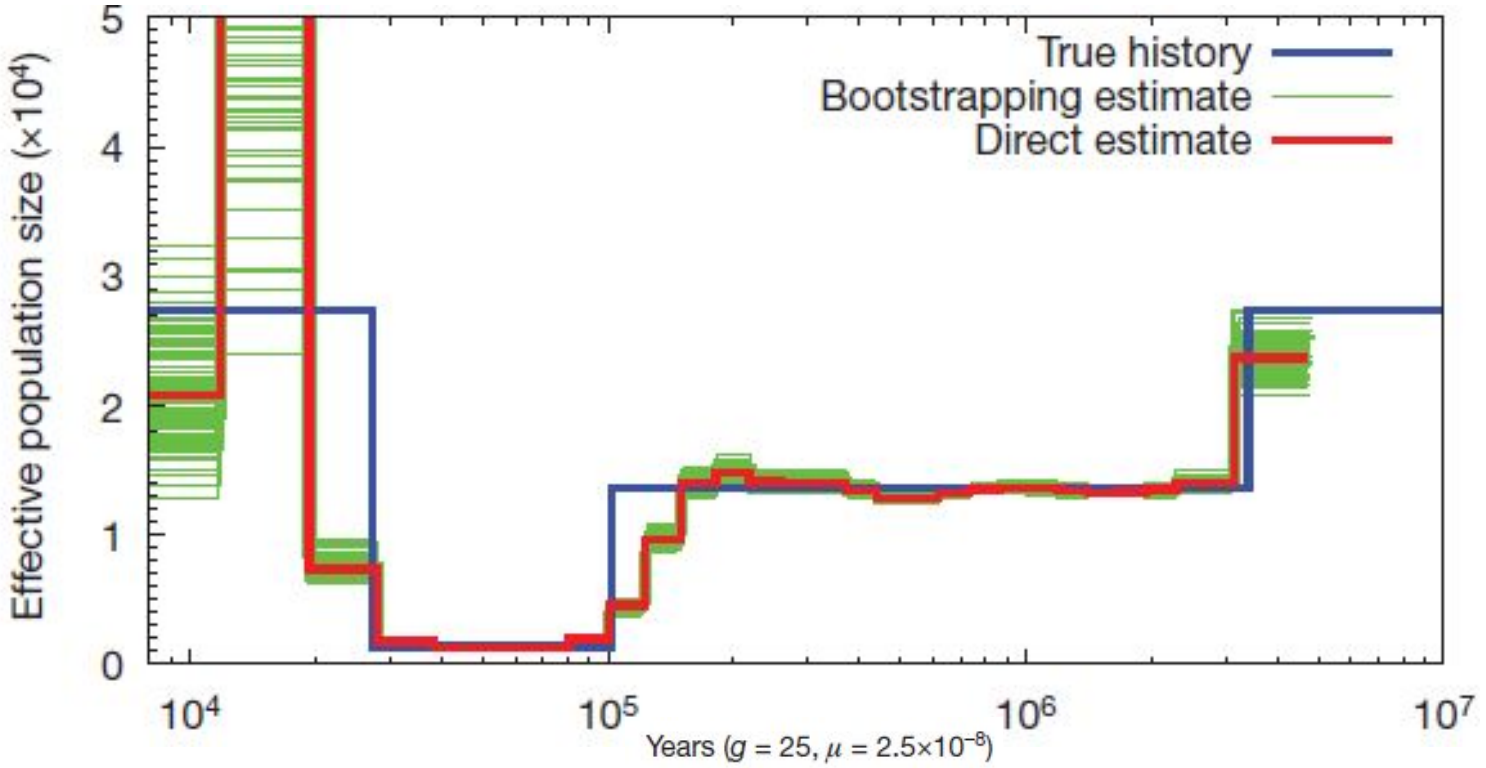
The hidden Markov model



The hidden Markov model

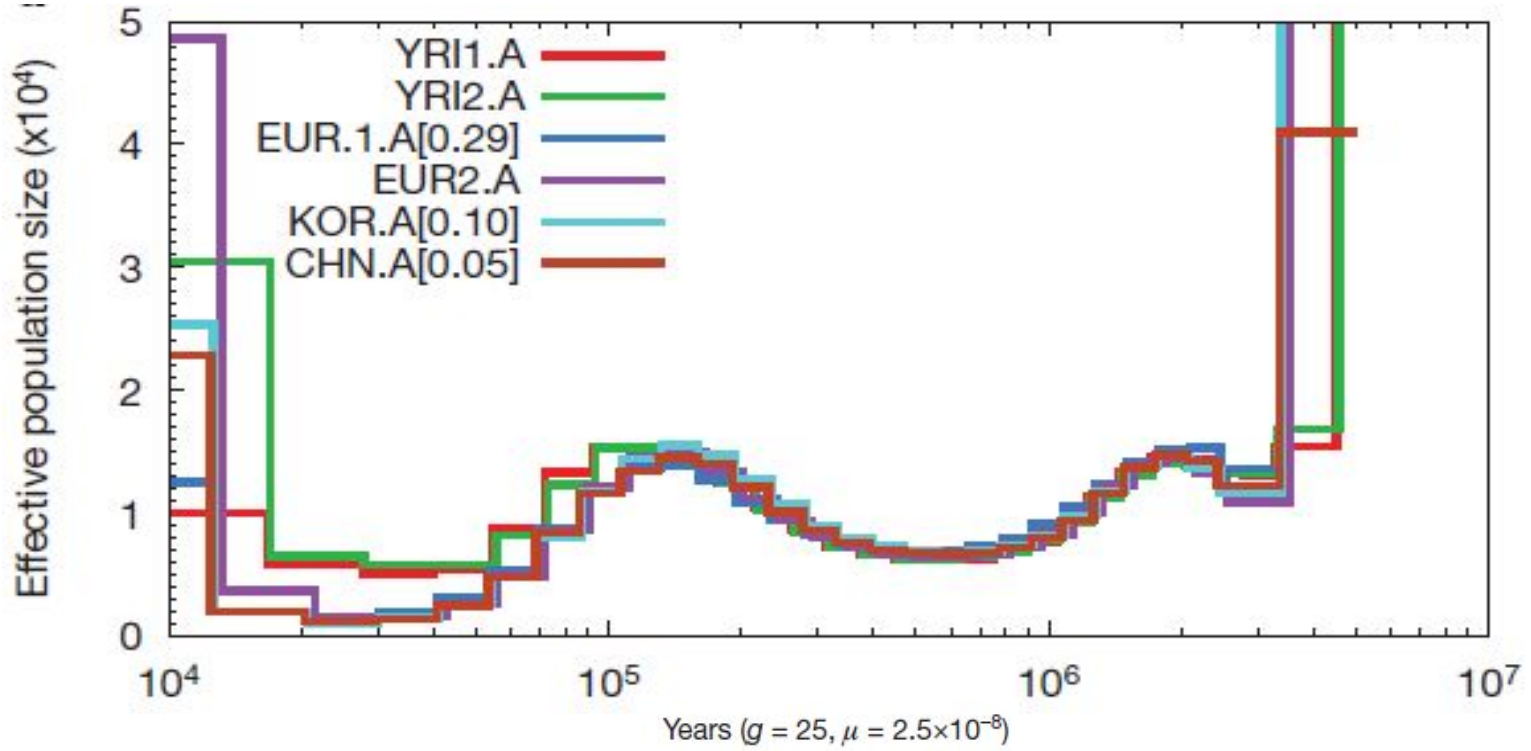


Results



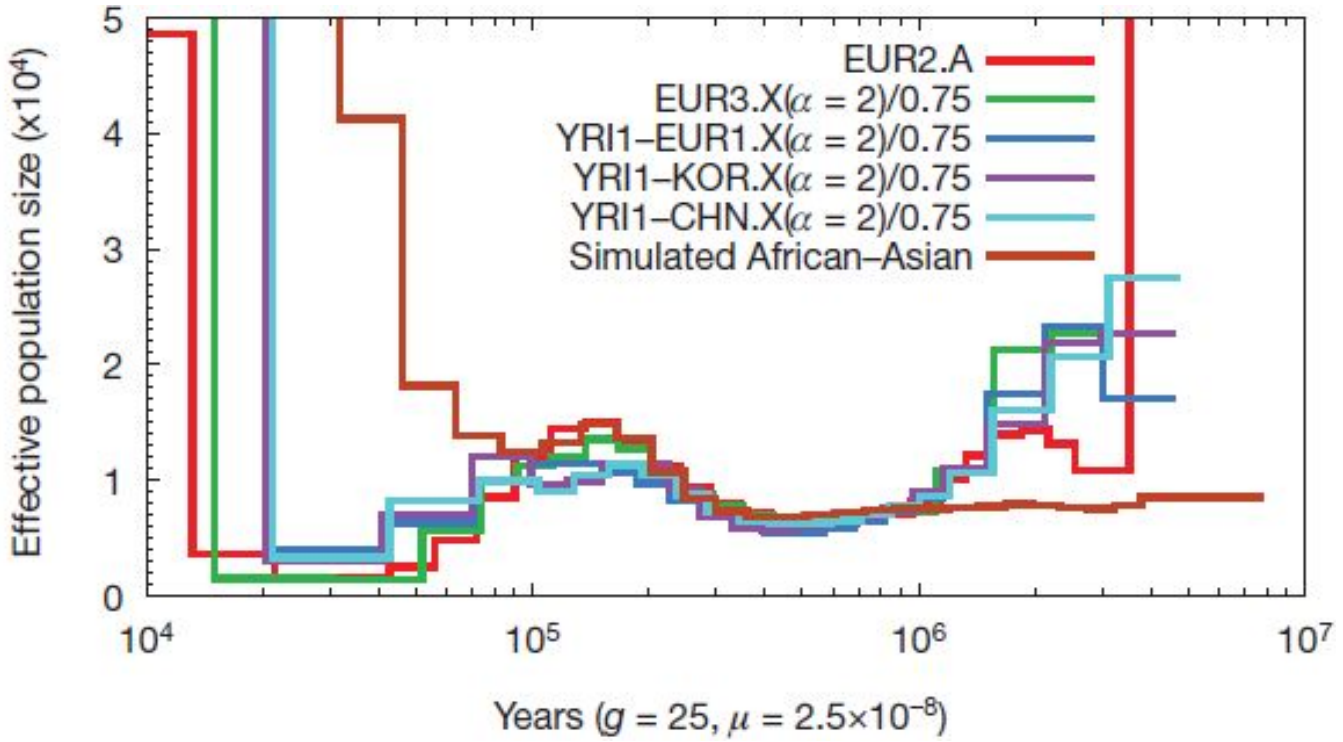
- Poor and highly variable estimates for recent generations
- Sharp bottlenecks are smoothed out

Results



- Two dips in Ne for all samples; recent population growth
- Ne for African samples is higher in recent 10-100 kya

Results



- Combine X chromosomes, and run same analysis
- Confirms West African and non-African populations largely remained as one population until 60-80 kya, but suggests continued genetic exchange between continental populations even after separation

Limitations

- Inferences sensitive to violations of assumptions
  - Coalescent assumes neutral evolution
  - Must define the generation times
  - Must define scaled mutation, recombination rates, and constant population sizes
- No explicit hypothesis testing framework
  - Hard to say why Ne changes
- Extending to data with multiple individuals may be possible, but much more complex
  - Each HMM path would represent a recombination graph

Technical Discussion

- How do they go from time to most recent common ancestor to effective population size?
- Are there theoretical motivations for the proposed hidden Markov model?
- How do we interpret the observations for the pseudo-diploid studies? What additional considerations are required when working with X chromosomes?
- How sensitive is the method to the parameterization of time into discrete generations?

Limitations Discussion

- How do we feel about inferring a population history from a single diploid sample? For instance, displayed estimates differ between two YRI individuals.

- How much confidence do we have in PSMC given that proof of concept is from simulated data?
  - Genetic data produced by simulation fails to resemble empirical data (Mather et al., 2019).

- How does this method perform for recent history?

- How computationally demanding is this method?

Contextual Discussion

- How have their results been received?

- Are there comparable methods to this approach in recent research? How do these methods improve upon PSMC?
  - "Information in genetic data about effective population size comes from historical mutation events and also from historical recombination events. Approaches based on the ancestral recombination graph (ARG), such as the pairwise sequentially Markovian coalescent method, make use of both sources of information. However, because of computational constraints, they are limited to analysis of a small number of individuals, which restricts their ability to make inferences about the very recent past." (Browning and Browning, 2015)