Nuclear DNA and the History of Population Size

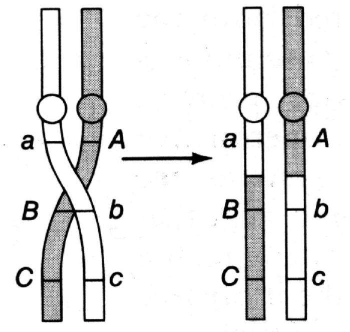
Alan R. Rogers

September 10, 2021

Advantages and disadvantages of the nuclear genome

- Huge amounts of data.
- Recombination complicates things.

Nuclear genes recombine



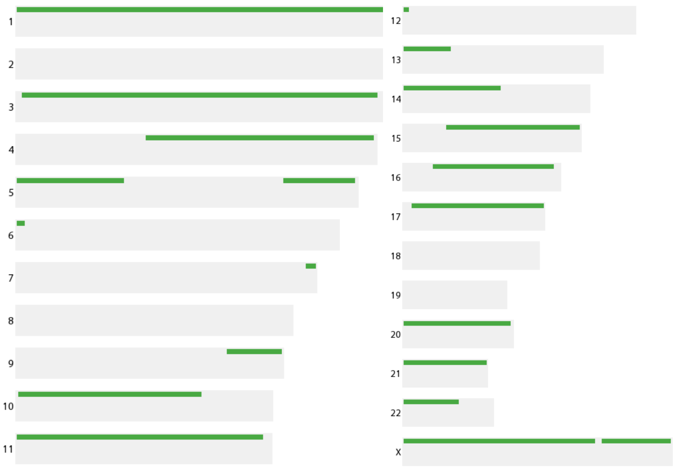
Useful data began to appear in about 2000.

Crossovers shuffle DNA

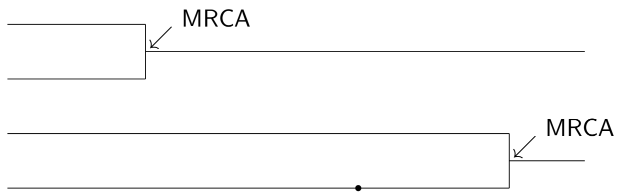
Gamete in gamete may differ from either parental chromosome; if so, it's a *recombinant* chromosome.

Each chromosome has many gene genealogies, which vary in length.

Chromosome sharing by my mother and daughter

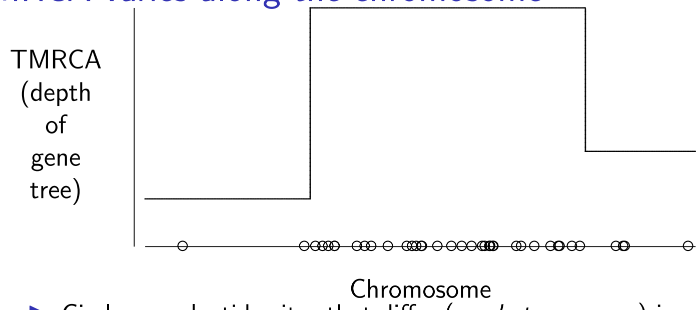


Gene trees at two loci



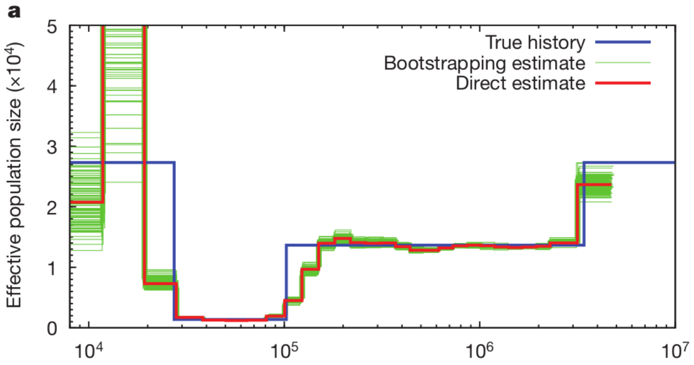
- TMRCA: time of the most recent common ancestor
- Gene trees vary in length across the genome.
- Mutation (•) is more likely on a deep gene tree.

TMRCA varies along the chromosome



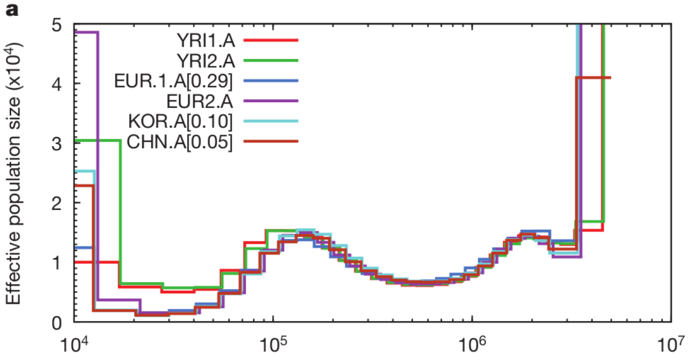
- Circles: nucleotide sites that differ (are *heterozygous*) in a single diploid sample.
- Heterozygous sites are denser where gene tree is deep.
- Population size → length of MRCA segments and genetic variation within segments.
- The "PSMC" method (Li & Durbin 2011) uses this pattern to estimate population history.

PSMC is accurate from 30 ky to 3 my ago.



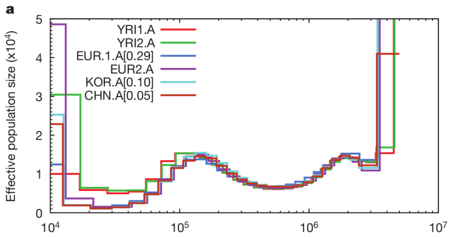
(Li and Durbin 2011)

PSMC estimates from autosomes



(Li and Durbin 2011)

PSMC estimates from autosomes

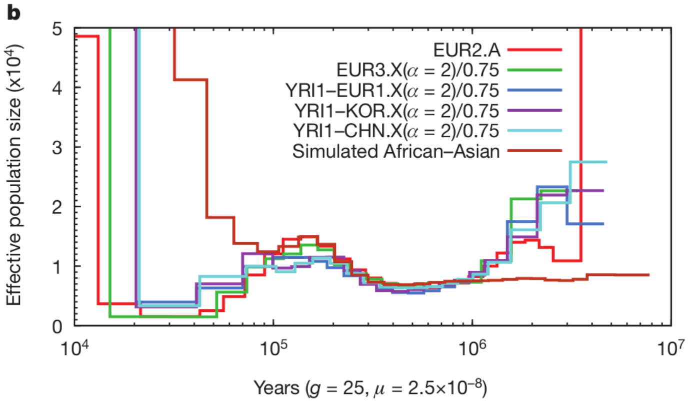


↑ 2 mya (origin of *Homo*); ↑ 200 kya (origin of modern humans); ↑ 20 kya (beginning of Holocene).

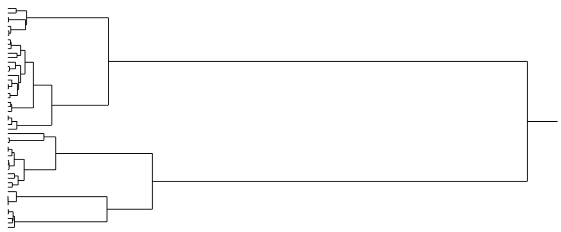
Eurasian/African split 150 kya.

African bottleneck short and shallow.

PSMC estimates from X chromosomes

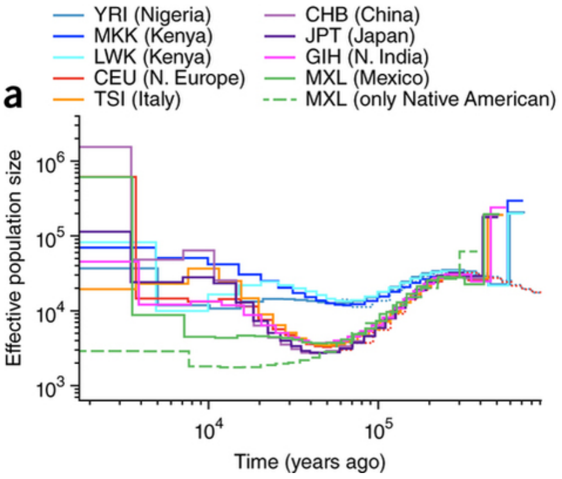


Once again: simulated gene genealogy of a sample of size 50 from a population of constant size

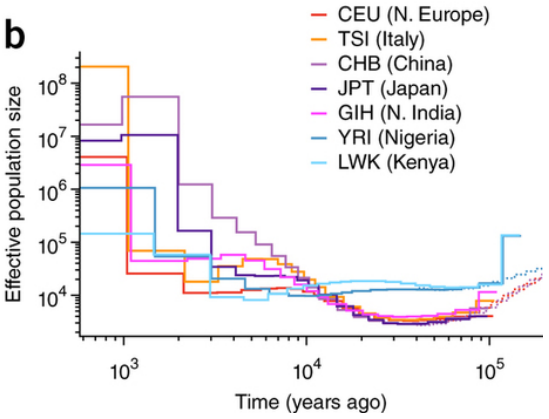


To estimate the *recent* history of population size, you need larger samples.

MSMC: using multiple genomes

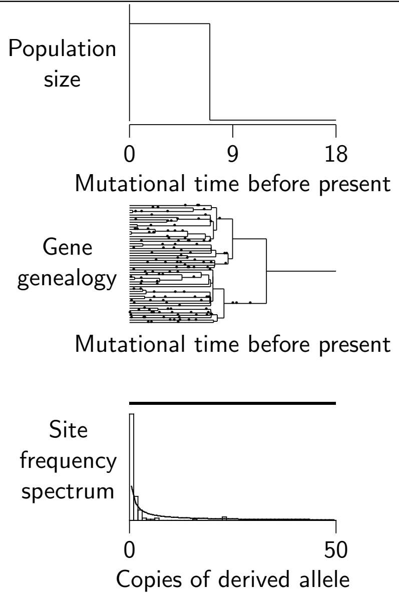


MSMC: using multiple genomes



MSMC: limitations

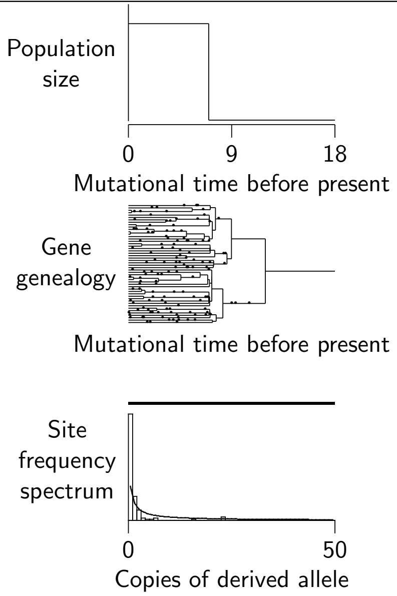
- can only deal with 4 diploid genomes
- data must be "phased": we need to know which nucleotides lie together along individual chromosomes
- phasing errors cause bias, especially during the last 10,000 years.



Site frequency spectrum

*i*th bar: number of sites at which derived allele is present in *i* copies.

Population growth or selection: an excess of rare derived alleles.



Site frequency spectrum

In nuclear DNA, there are millions of trees, most with no mutations, a few with one mutation.

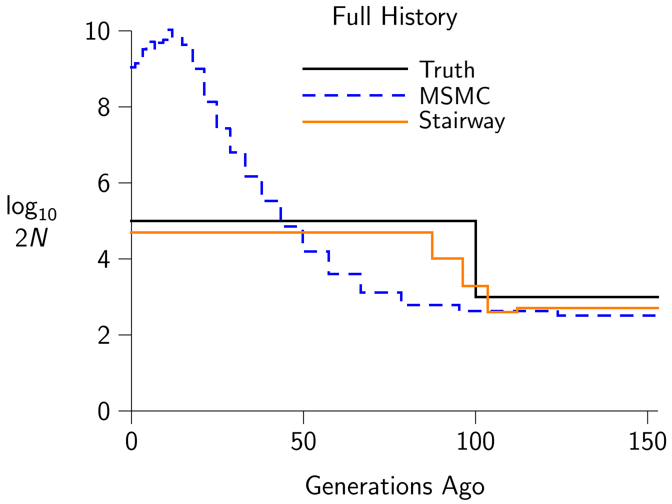
It's still true that most mutations are singletons if the population has grown.

The spectrum is useful with nuclear as well as mitochondrial DNA.

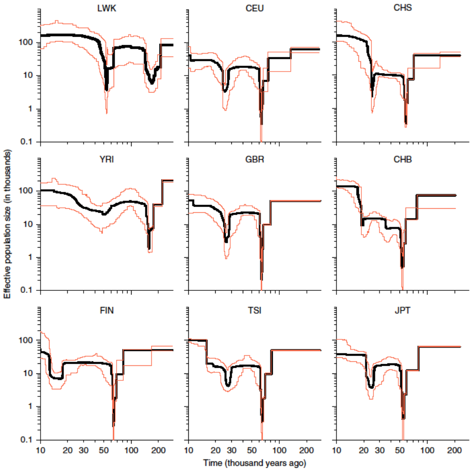
Stairway plot (Liu & Fu 2015)

- uses site frequency spectrum
- no need for phased data
- can deal with samples of hundreds of individuals

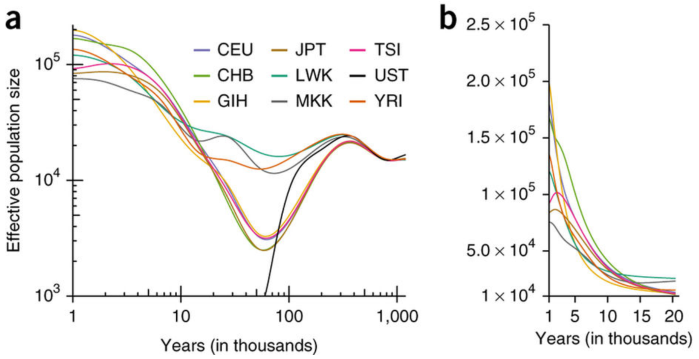
Very recent population growth is tough



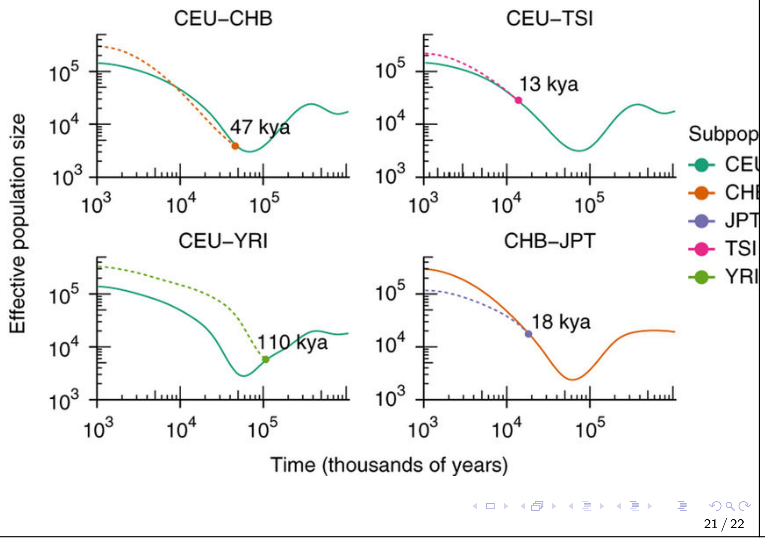
Stairway plot results (Liu & Fu 2015)



SMC++: combines PSMC and spectrum (Terhorst, Kamm, & Song 2017)



Separation times (Terhorst, Kamm, & Song 2017)



Summary

- Human population has varied in size over past 3 my.
- Bottleneck 60 kya, around the time Eurasians left Africa.
- Bottleneck during last ice age, 20 kya.
- African bottleneck was shorter and shallower.
- Eurasian/African split 110 kya.
- European/Asian split 50 kya.