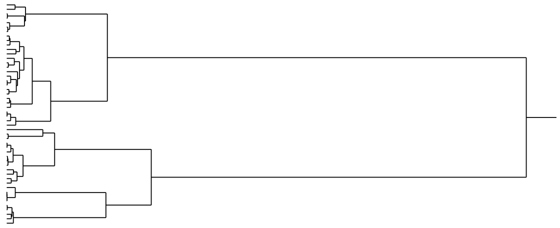
The History of Population Size from Whole Genomes

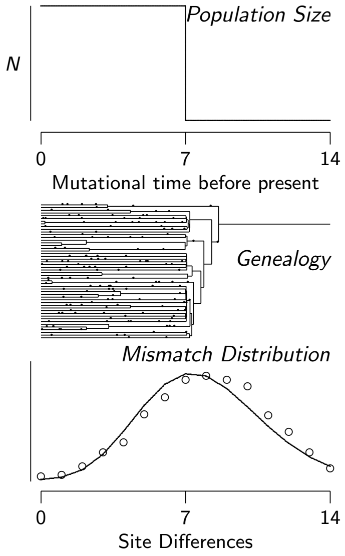
Alan R. Rogers

April 2, 2024

Simulated gene genealogy of a sample of size 50 from a population of constant size



- Short terminal branches; long basal ones.
- Large samples tell us about recent past.
- Not necessary for ancient past.



Effect of a population explosion

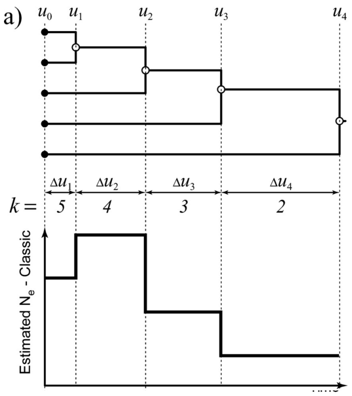
Middle: genealogy of 50 individuals; dots are mutations.

1 mutational diff per time unit

Bottom: ○ = simulated data, line = theory.

Wave peaks at population expansion.

Skyline Plot

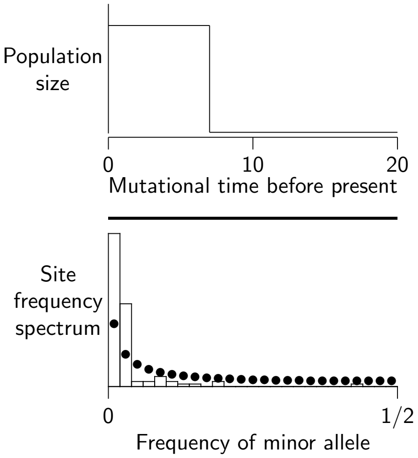


(Drummond et al 2005)

- Use mutations to estimate length of each interval.
- Long intervals imply large population size.
- Won't work with nuclear DNA: too few mutations per tree

Nuclear genome

- Huge amounts of data.
- Recombination makes previous methods unusable.



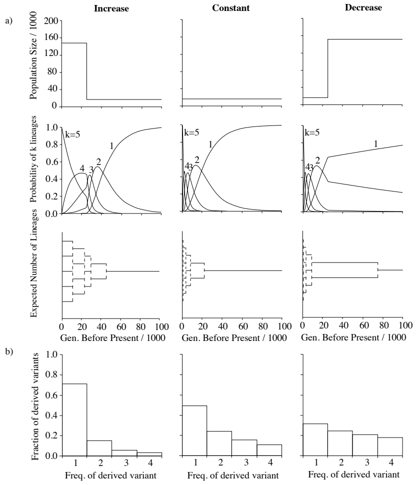
Site frequency spectrum

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

Population growth inflates the number of singletons. Bars show simulated spectrum; filled circles show expectation under constant population size.

To estimate history, we need a theory that works under complex population histories.

We do have such a theory



Row 1: 3 histories of population size

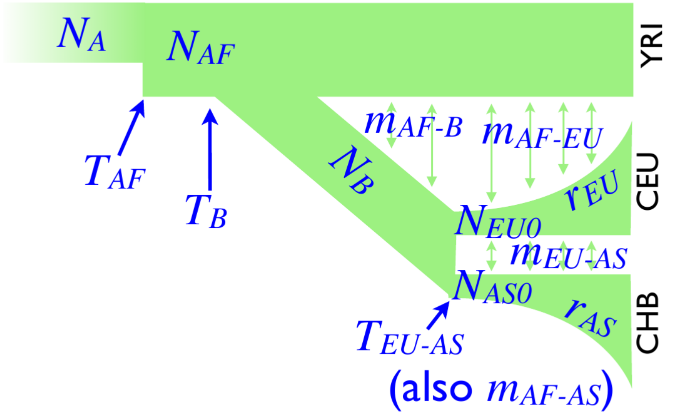
Row 2: Prob that there are *k* lineages at time *t* in a sample of size 5.

Row 3: Expected lengths each coalescent intervals.

Row 4: Expected spectrum

Griffiths & Tavaré (1998); Wooding & Rogers (2002)

∂a∂i: inferences from the site frequency spectrum

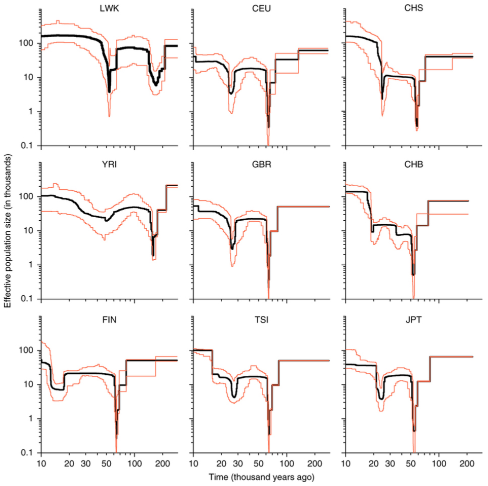


(Gutenkunst et al 2009)

Stairway plot (Liu & Fu 2015)

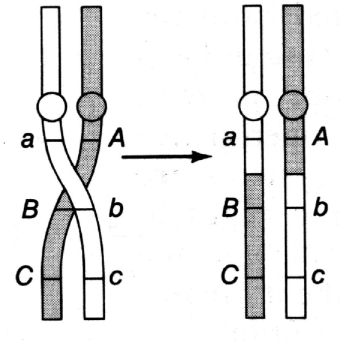
- uses site frequency spectrum
- no need for phased data
- can deal with samples of hundreds of individuals

Stairway Plot results



1000-Genomes data; (Liu & Fu 2015)

Also: recombination is our friend

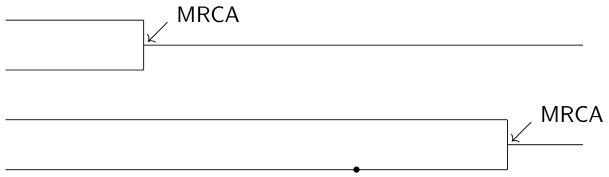


Useful data began to appear in about 2000.

Crossovers shuffle DNA

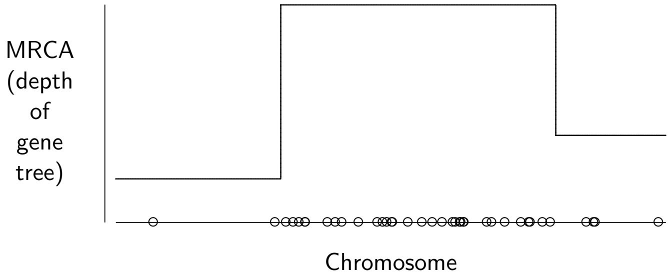
Each chromosome has many gene genealogies, which vary in length.

Two (hypothetical) loci in a single diploid genome



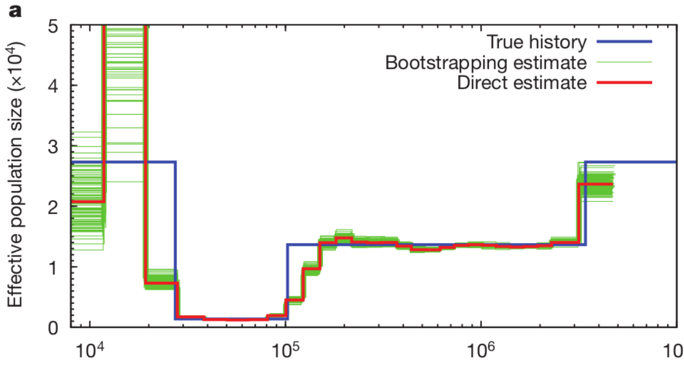
- MRCA: most recent common ancestor
- Gene trees vary in length across the genome.
- Mutation (•) is more likely on a deep gene tree.

MRCA varies along the chromosome



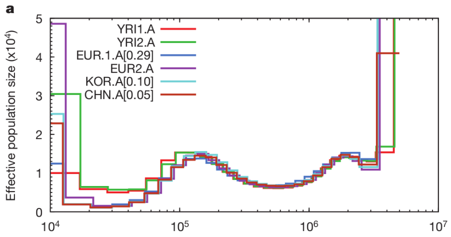
- Circles: nucleotide sites that differ (are *heterozygous*) in a single diploid sample.
- Heterozygous sites are denser where gene tree is deep.
- Population size → length of MRCA segments and genetic variation within segments.
- PSMC uses this pattern to estimate population history

PSMC is accurate from 20 ky to 3 my ago.



(Li and Durbin 2011)

PSMC estimates from autosomes

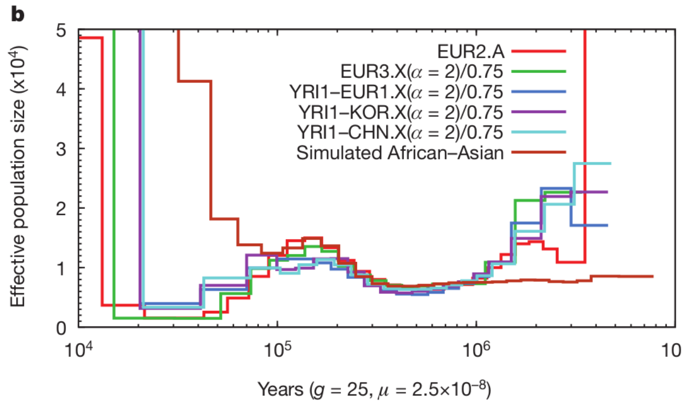


↑ 2 mya (origin of *Homo*); ↑ 200 kya (origin of modern humans); ↑ 20 kya (beginning of Holocene).

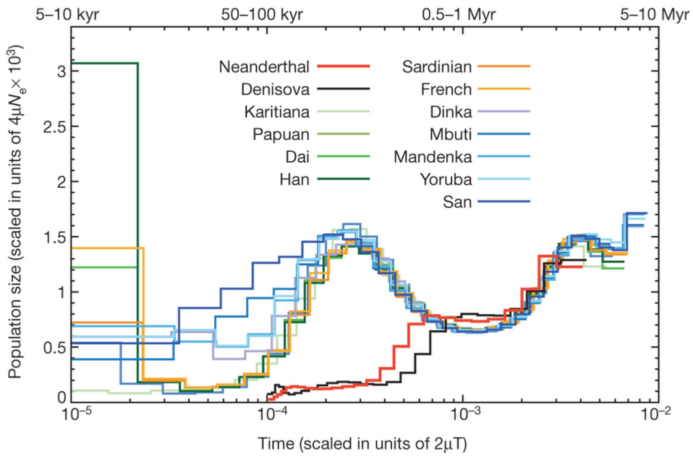
Eurasian/African split 150 kya.

African bottleneck short and shallow.

PSMC estimates from X chromosomes

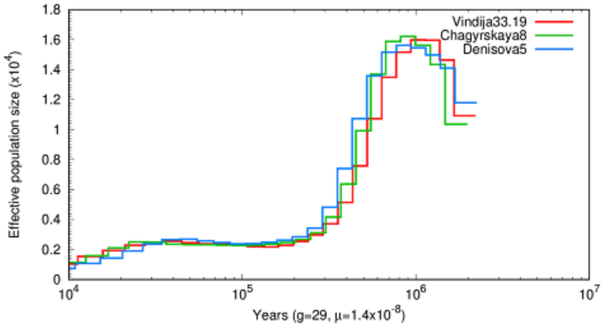


PSMC with Neanderthal as well as Denisova



(Prüfer et al 2014)

All 3 high-coverage archaic genomes



Horizontal axis shows time backwards from time of fossil's death. Curves of younger fossil (Vindija) is shifted to right.

(Mafessoni et al 2022)

What if the population is subdivided? (part 1)

Suppose there are K local groups, each of size N, which exchange migrants. We sample a single genome from some group. In the recent past, the hazard of a coalescent event is $h(t) = 1/2N$, and effective population size $(1/h(t))$ is $2N$.

Farther back in time, the ancestors of the 2 gene copies in our diploid sample may be in different groupss. At time t,

$$h(t) = \begin{cases} 1/2N & \text{if they're in the same group} \\ 0 & \text{if they're in different groups} \end{cases}$$

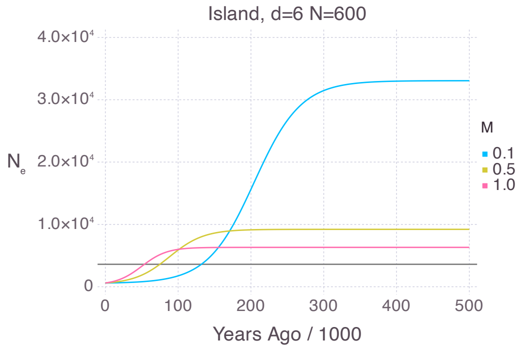
What if the population is subdivided? (part 2)

In the distant past, those loci that have still not coalesced tend to be those at which the two lineages haven't spent much time together in the same local group.

Consequently, the two lineages are *less likely* to be in the same group than are two drawn at random from the population as a whole.

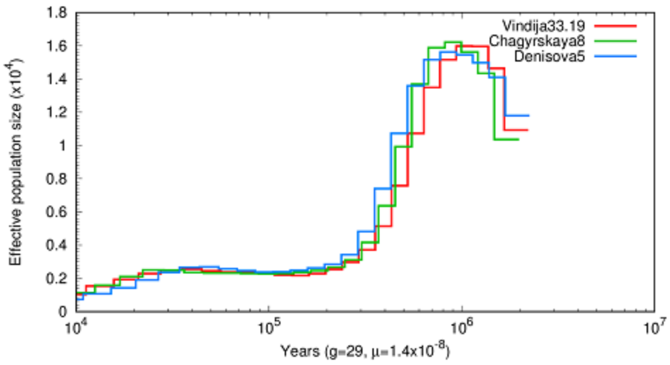
This means that in the distant past, N_e is greater than the actual population size.

PSMC gives misleading signal of decline in subdivided populations, even if population size is constant (Mazet et al 2014).



M is twice the number of immigrants per group per generation; horizontal black line is true population size.

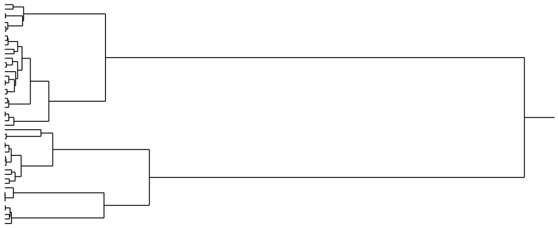
Population structure cannot explain the archaic decline



Decline due to population structure must be either faster or larger than that seen in archaic data.

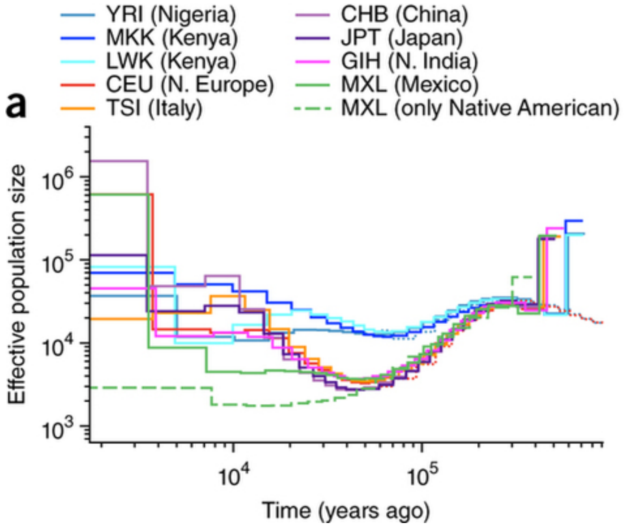
There must have been a change either in the number of individuals or in the rate or pattern of mobility among groups.

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

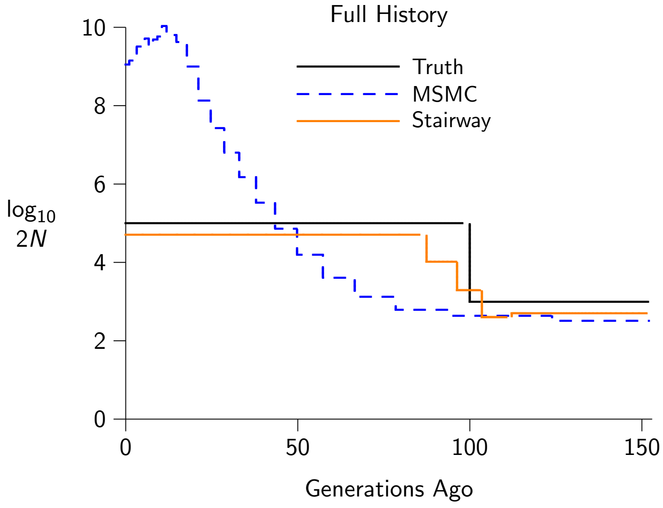


To estimate the *recent* history of population size, you need larger samples.

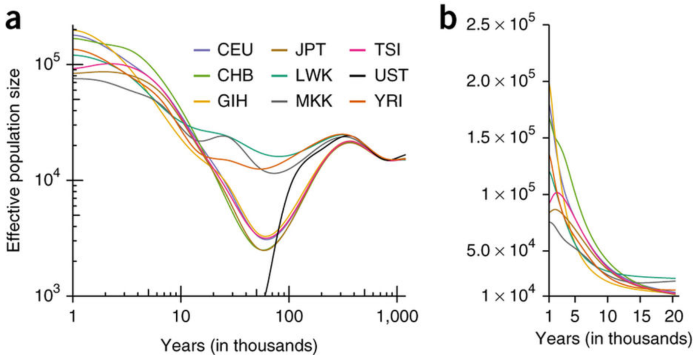
MSMC: using multiple genomes



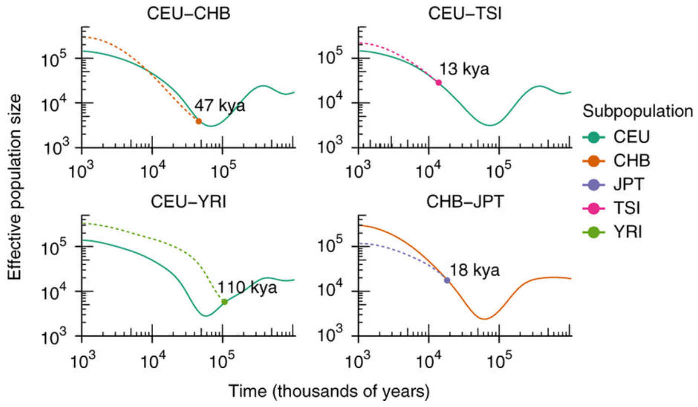
Very recent population growth is tough



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



Separation times (Terhorst, Kamm, & Song 2017)



Methods for studying the history of population size

- Large sample are useful for studying recent history; small samples are sufficient for ancient history.
- The site frequency spectrum can be used with large samples and is therefore useful in studying recent history. However, it ignores the information provided by LD.
- PSMC uses a single diploid genome. It is good for ancient history but not for recent history.
- MSMC uses several diploid genomes and has more power at recent time scales.
- SMC++ combines two sources of information: LD and the site frequency spectrum. It has power across a wide range of time scales.

Conclusions about human history

- History of population size affects depth of gene trees, genetic variation, and length of MRCA segments.
- We can use these facts to infer the history of population size.
- Human population has varied in size over past 3 my.
- Bottleneck during last ice age, ending 20 kya.
- African bottleneck was shorter and shallower.
- Eurasian/African split 150 kya.
- European/Asian split 20 kya.
- Effect of geographic population structure looks like population decline.