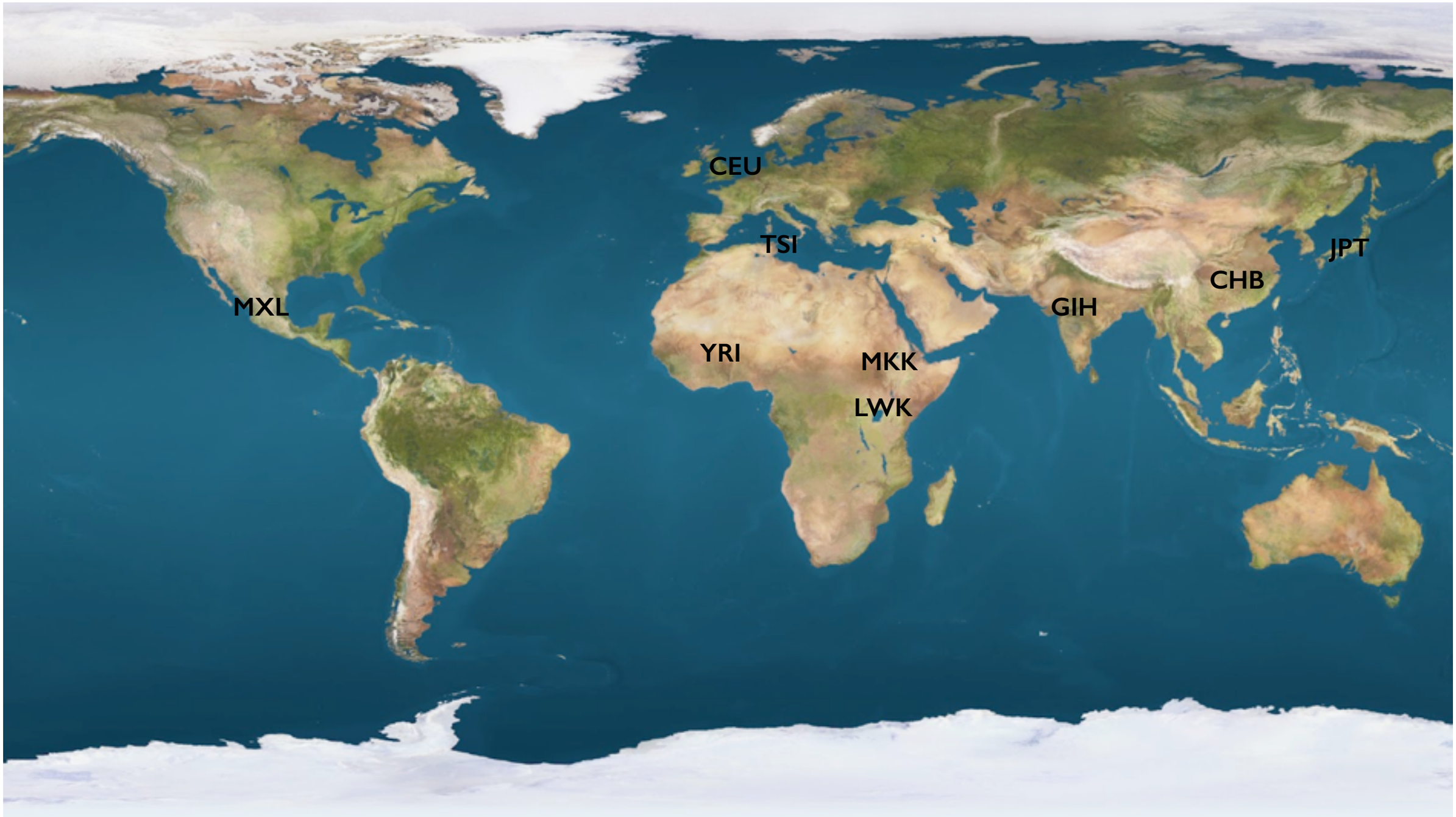
Inferring human population history from multiple genome sequences

Stephan Schiffels
Postdoctoral Fellow with Richard Durbin,
Wellcome Trust Sanger Institute, Cambridge, UK

EMBO Fellow

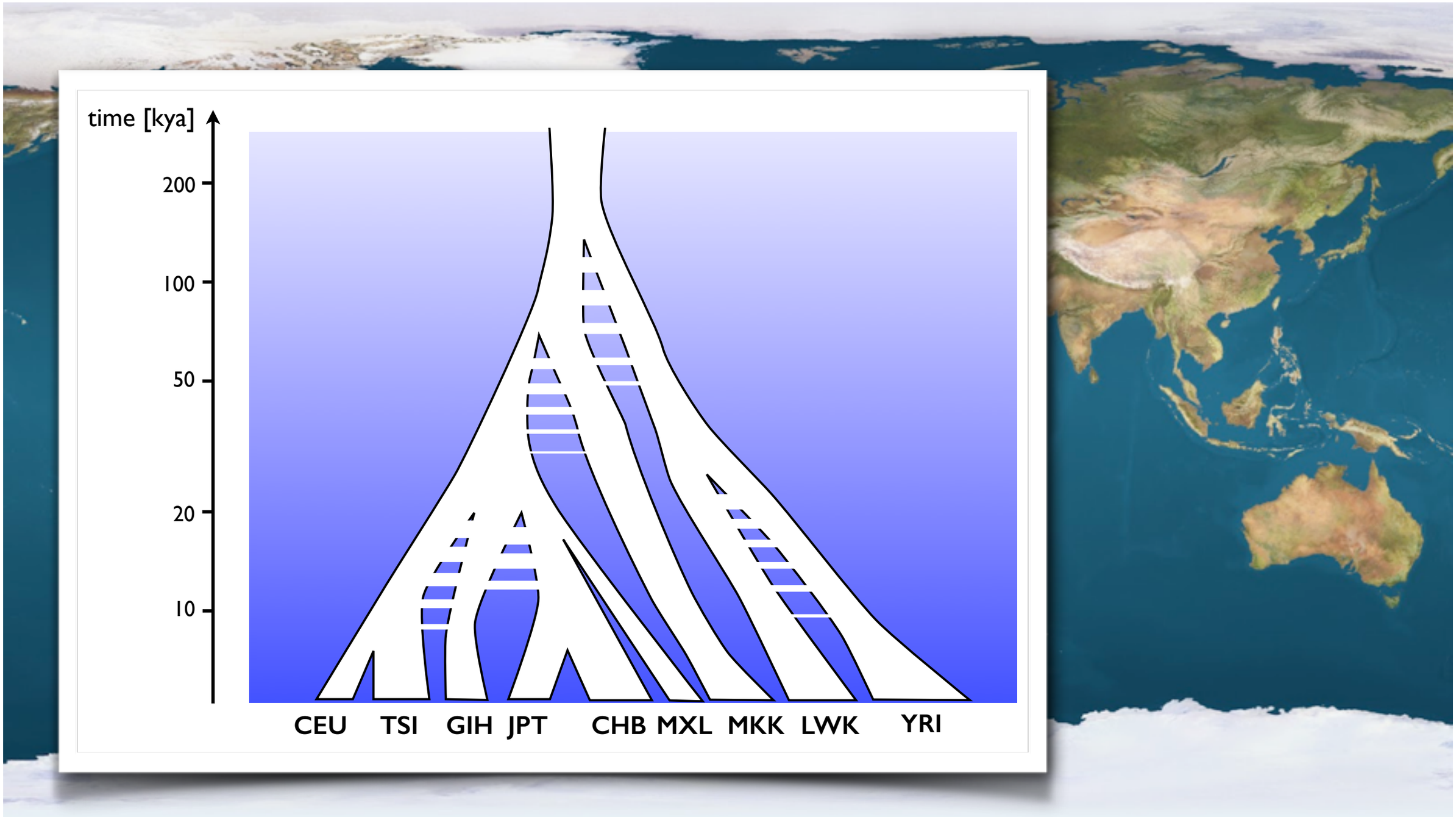
wellcome trust sanger institute

From genome sequences to human history



[Sequence data from Complete Genomics]

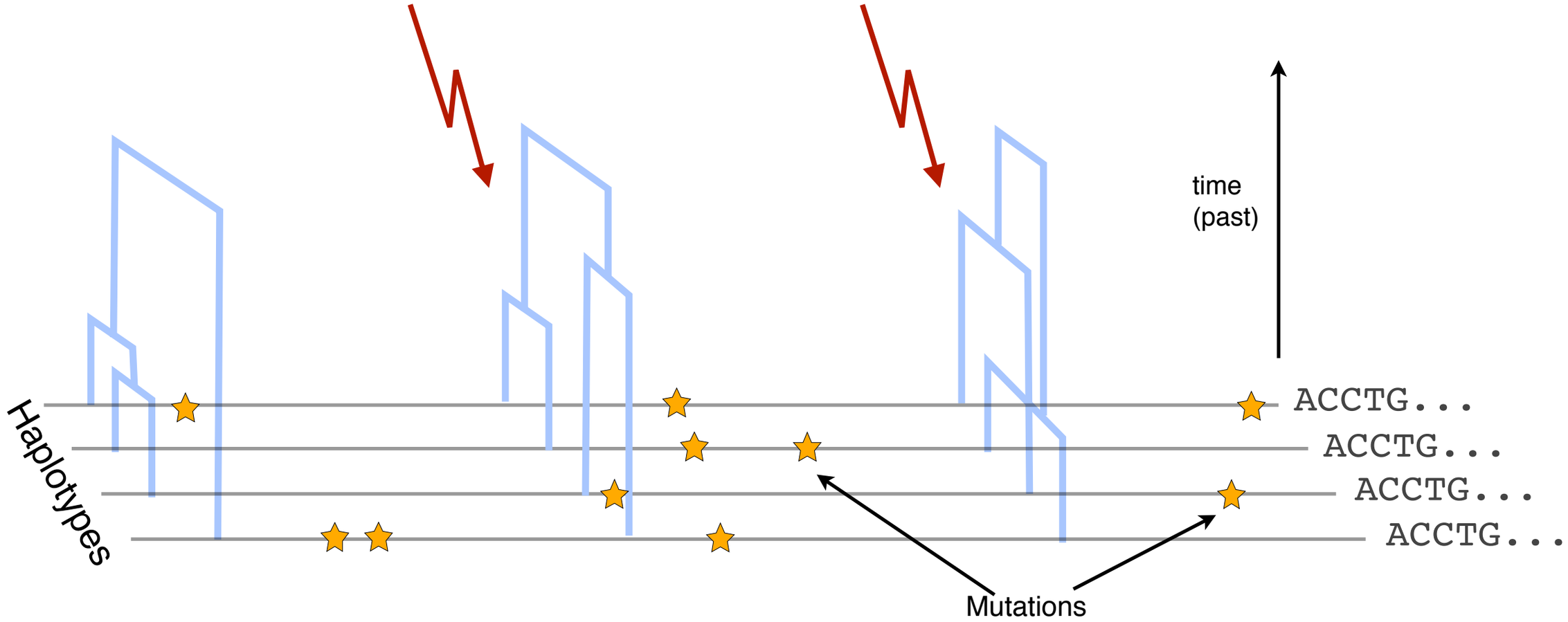
From genome sequences to human history



[Sequence data from Complete Genomics]

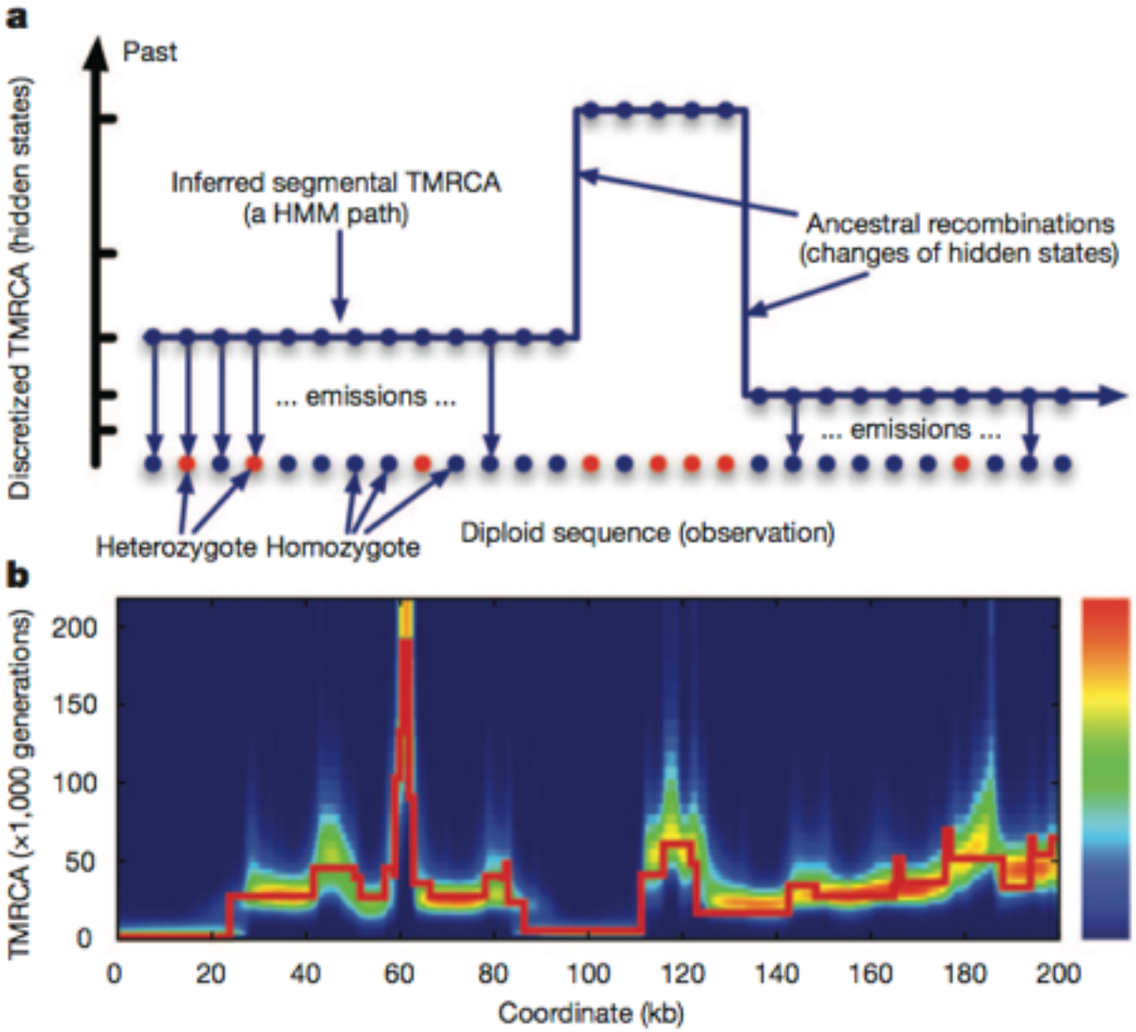
How are sequences related?

Ancestral recombinations change trees along the sequences



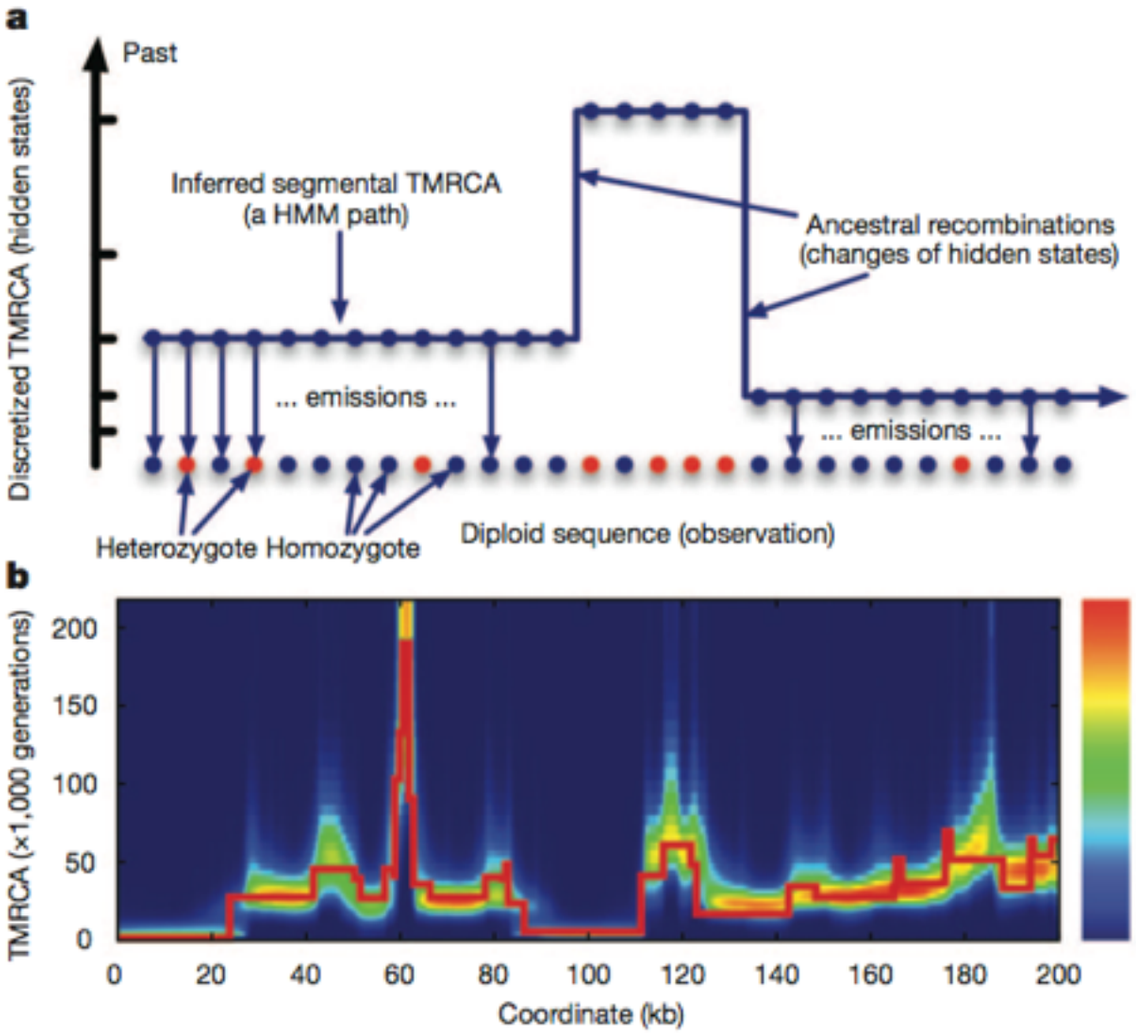
Problem: Estimate trees only from observed mutations

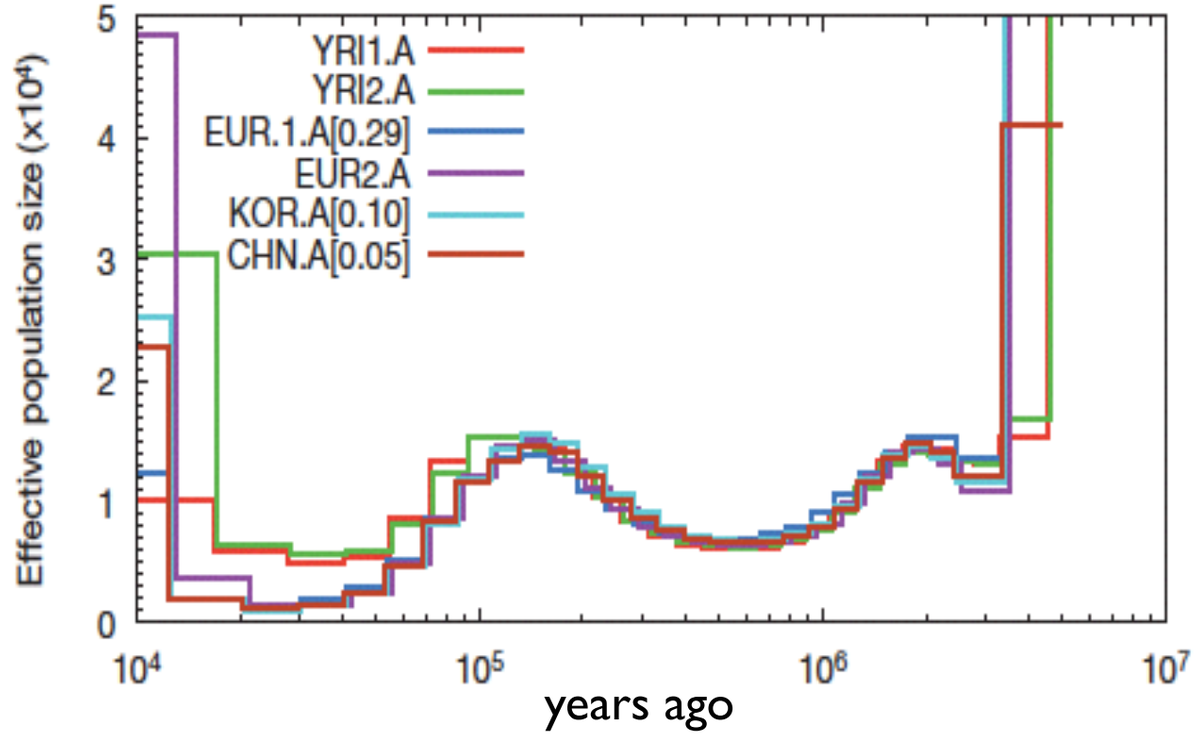
Previous work: PSMC



[Li and Durbin, 2011]

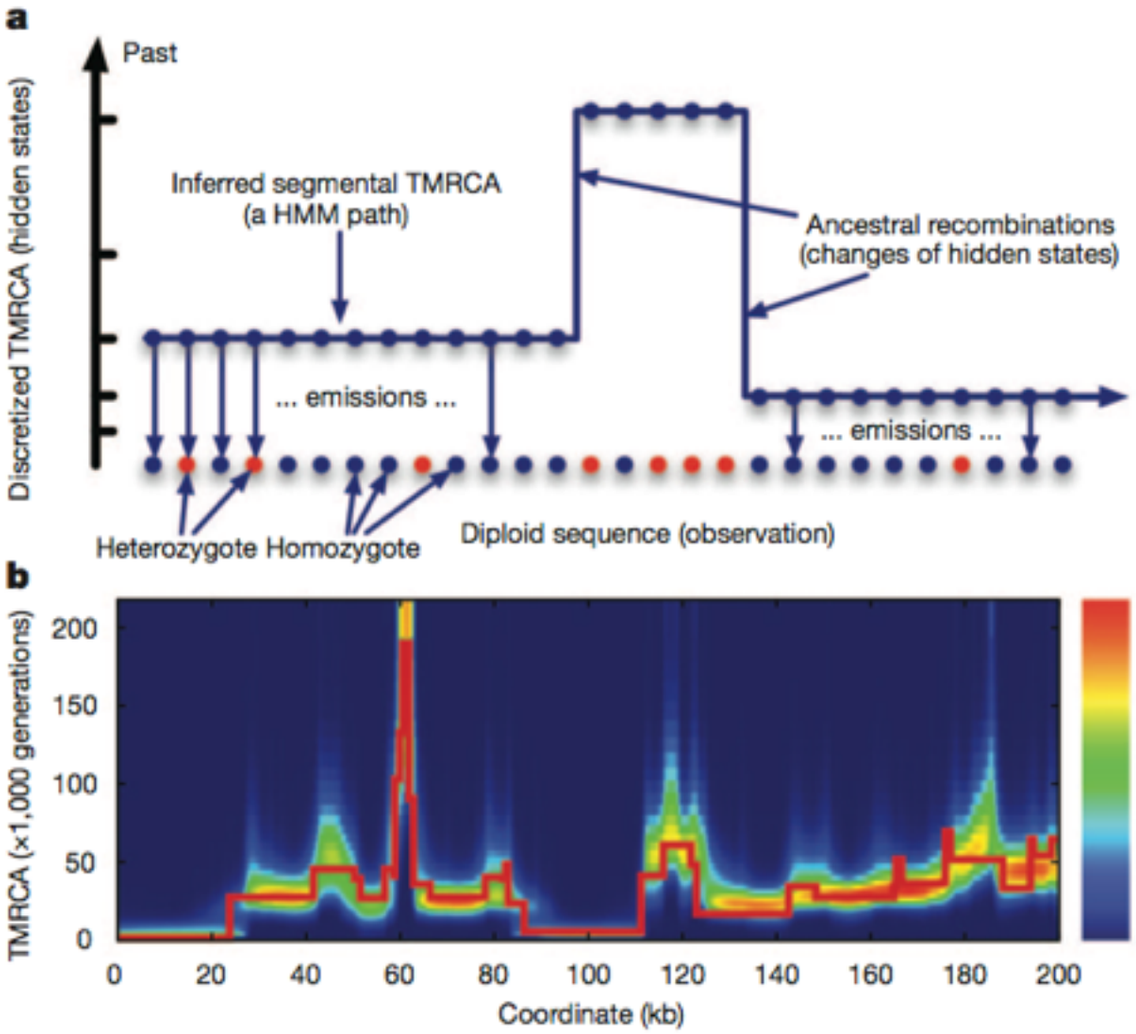
Previous work: PSMC

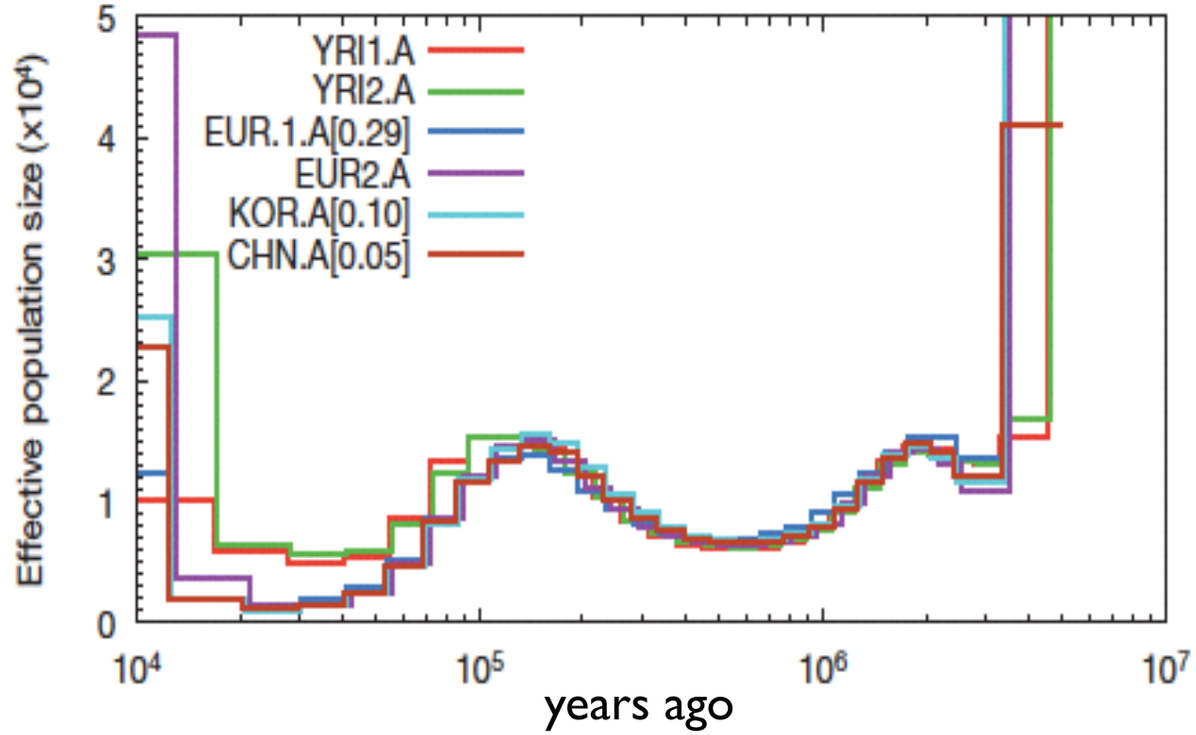




[Li and Durbin, 2011]

Previous work: PSMC

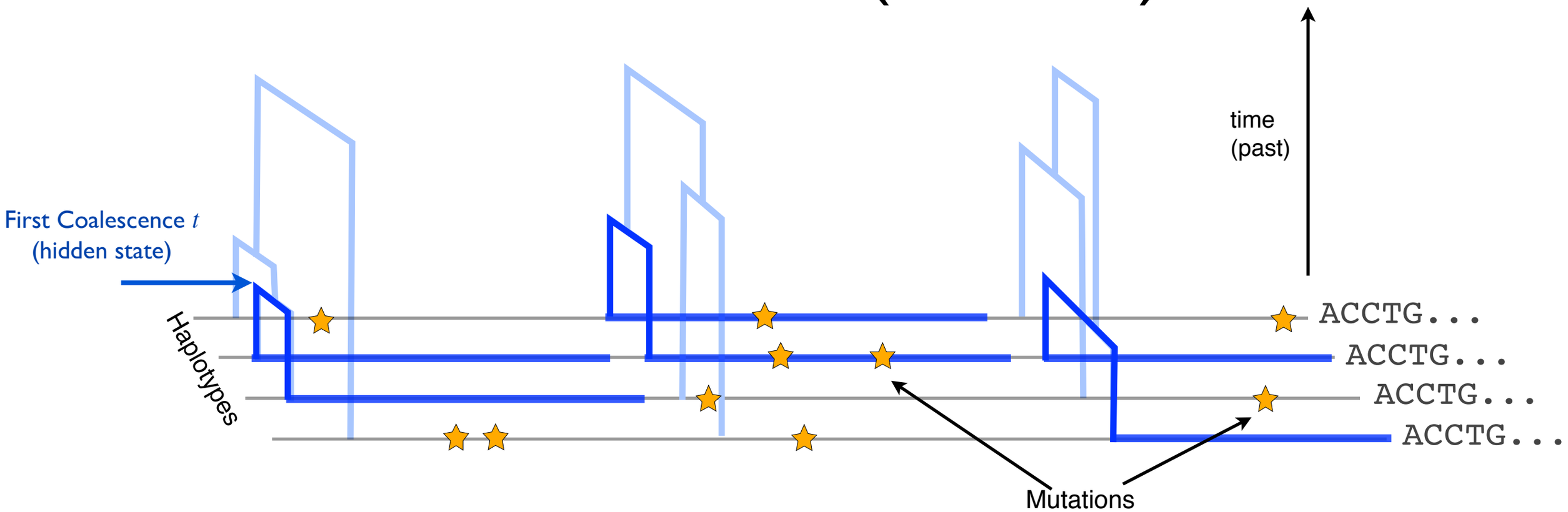




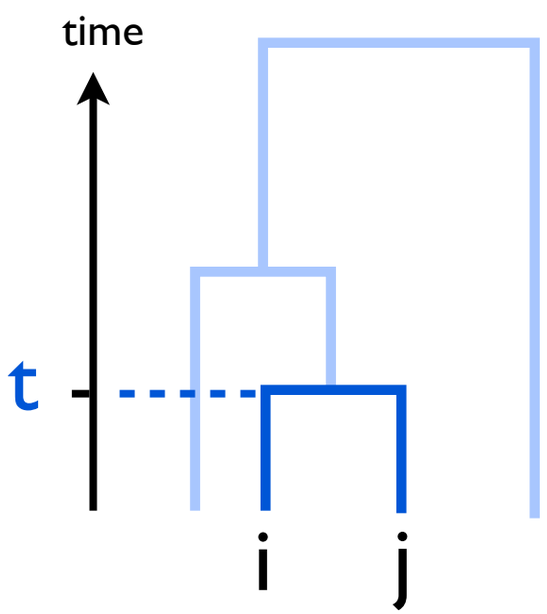
But: Inference from only two sequences limited to times beyond 20kya. Also: population splits difficult to model

[Li and Durbin, 2011]

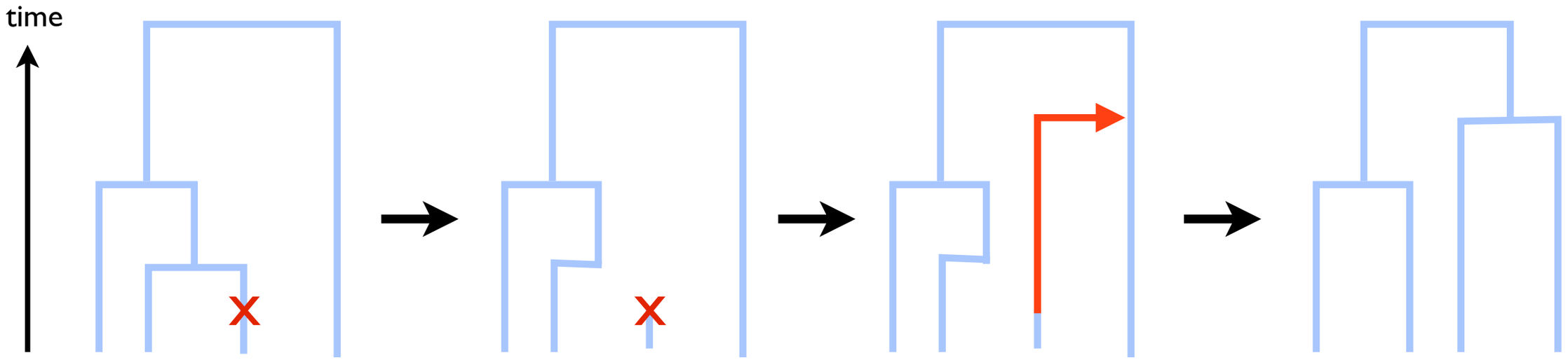
Multiple sequentially Markovian Coalescent (MSMC)



MSMC hidden state:
triple (t, i, j)



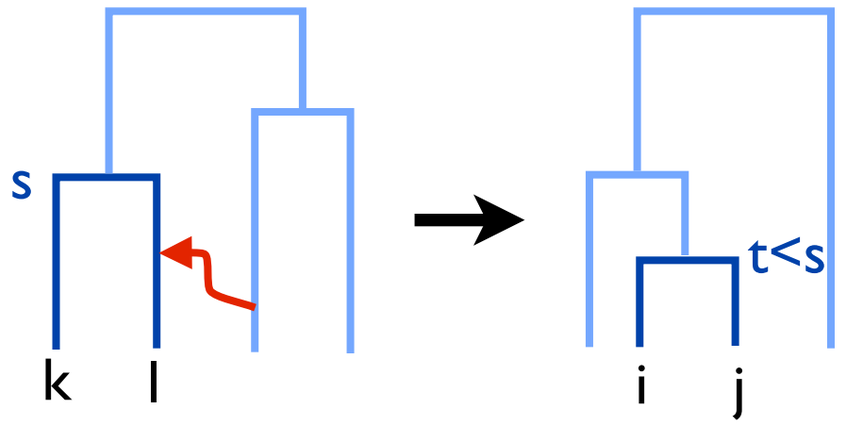
Effect of recombination on Genealogies



[*Sequentially Markovian Coalescent*, McVean and Cardin, 2005]

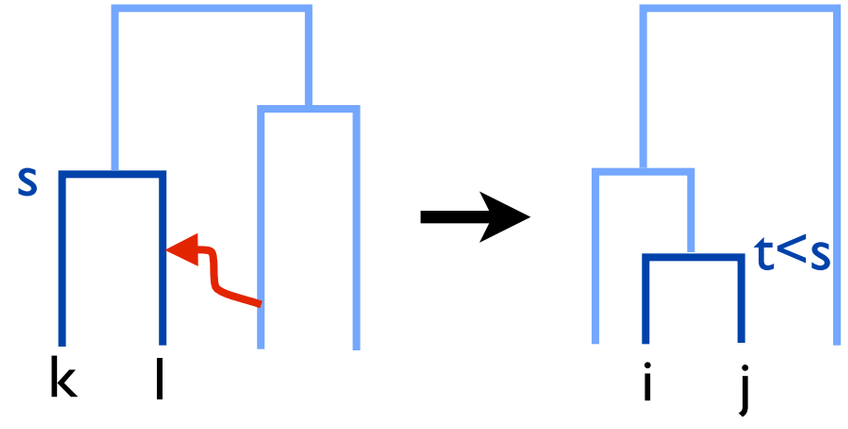
MSMC: state transitions

(s,k,l) → (t,i,j)
where t<s

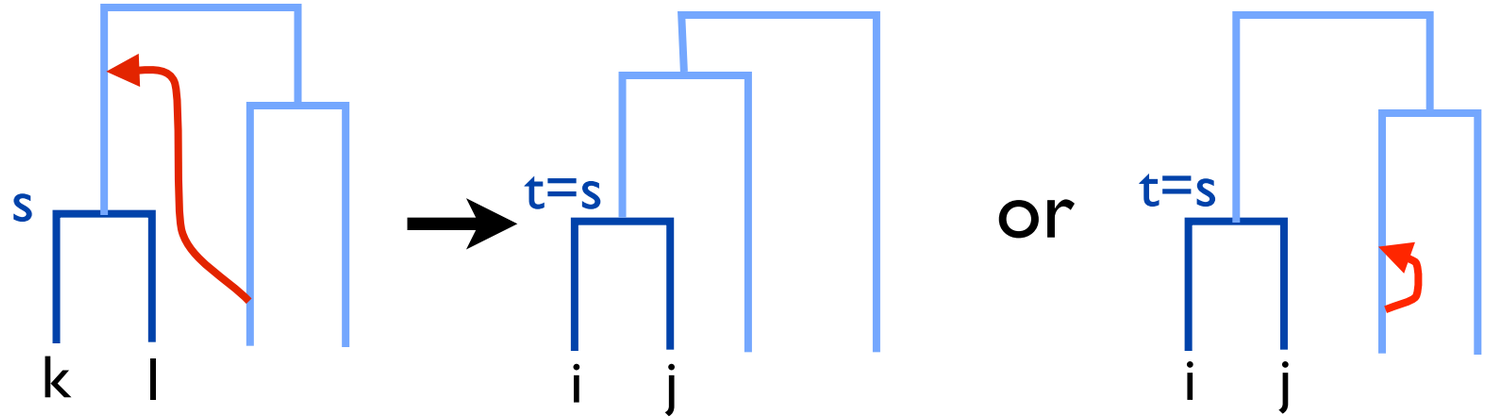


MSMC: state transitions

(s,k,l) → (t,i,j)
where t<s

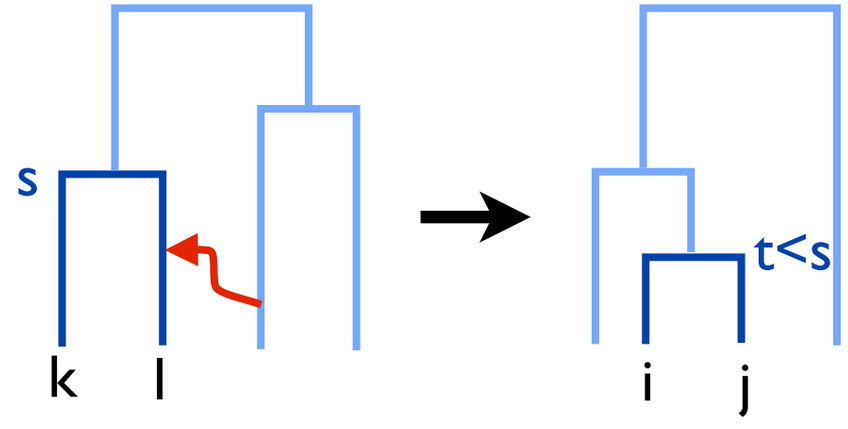


(s,k,l) → (t,i,j)
where t=s

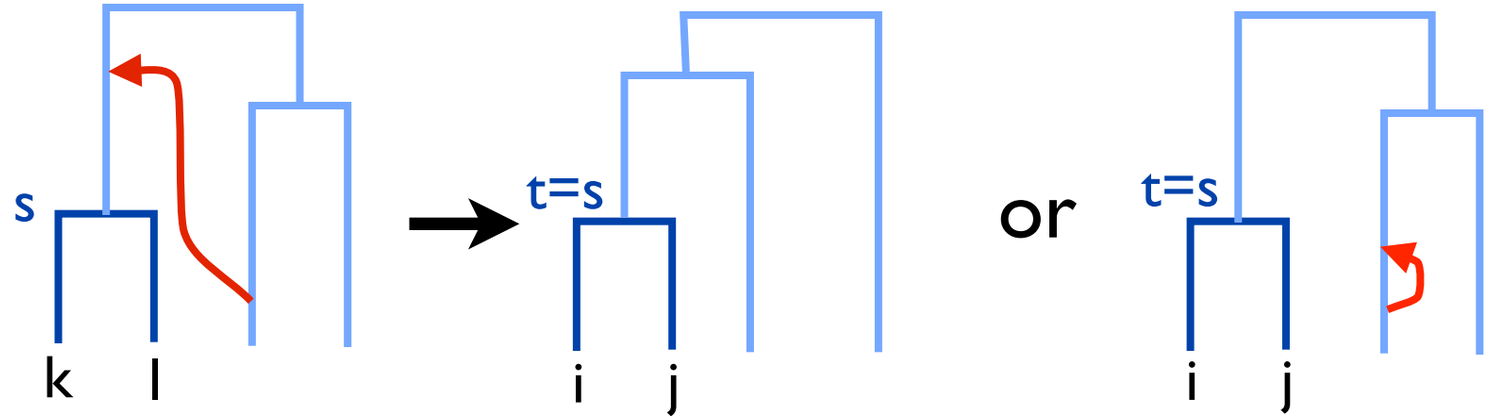


MSMC: state transitions

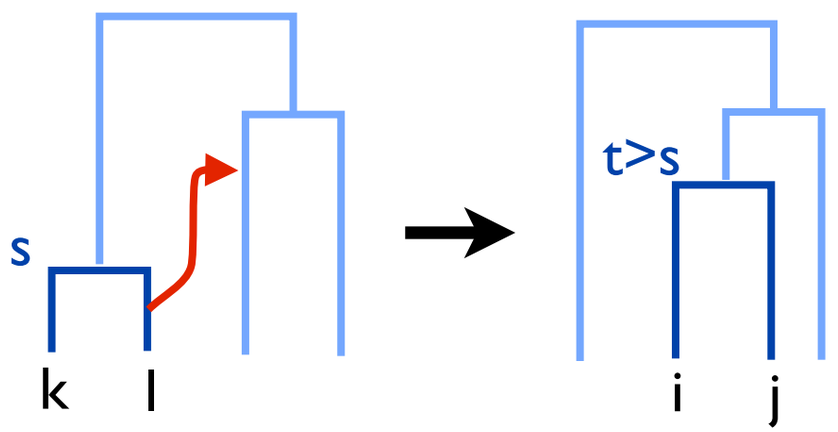
(s,k,l) → (t,i,j)
where t<s



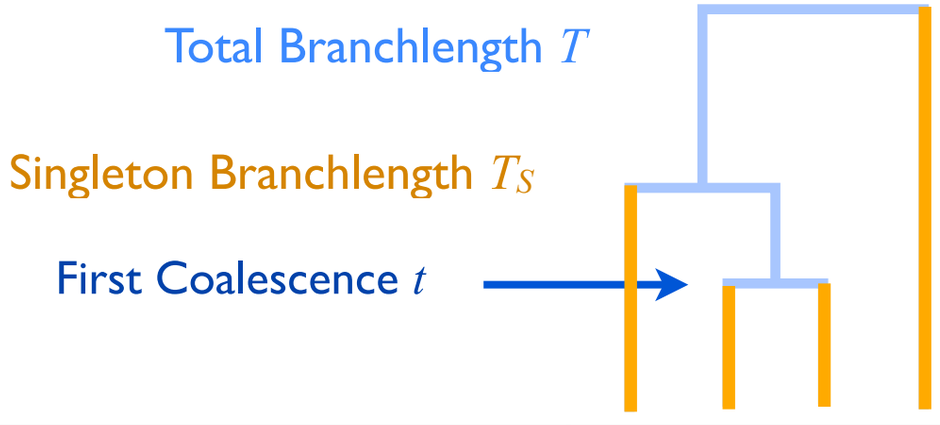
(s,k,l) → (t,i,j)
where t=s



(s,k,l) → (t,i,j)
where t>s



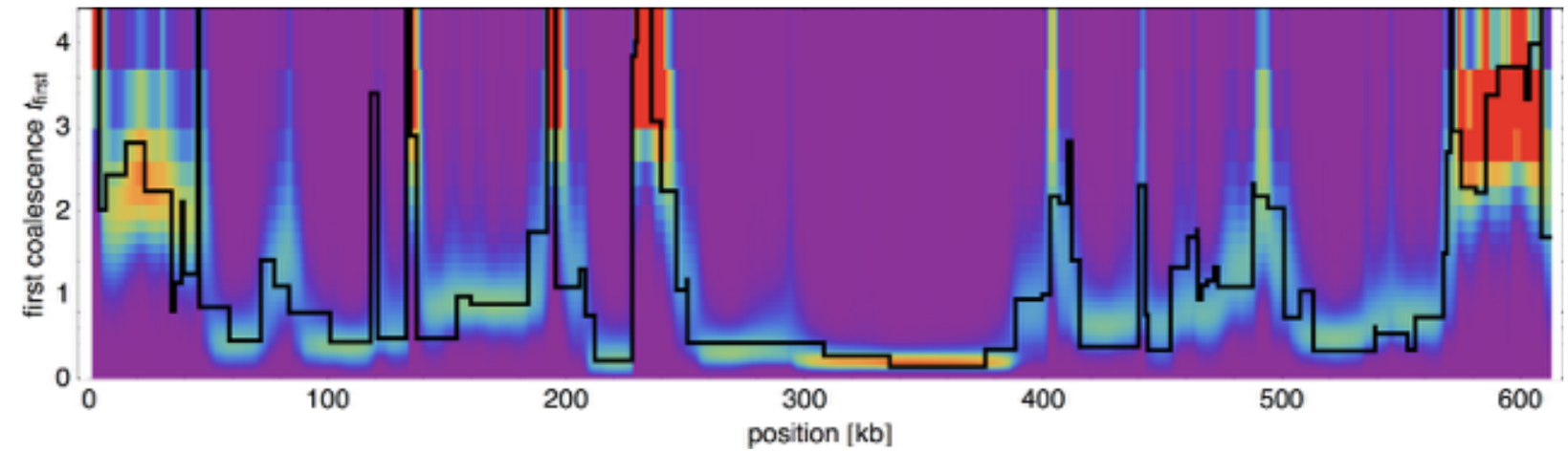
MSMC: mutation probability



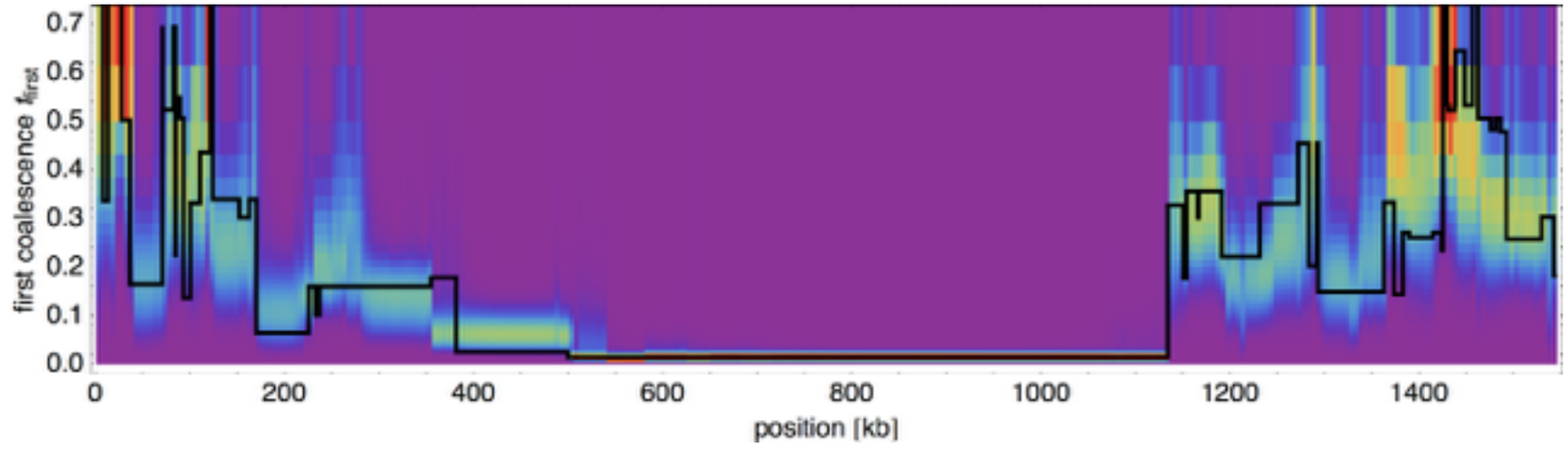
No mutation	$1 - \mu T$	A	A	A	A
Singleton within pair	$\mu\, t$	A	C	A	A
Singleton outside pair	μ	A	A	A	T
Double mutation	0	A	T	A	T
Higher Frequency	μ	A	T	T	A
Missing Data	1	A	A	-	-

Local Inference of first coalescence time

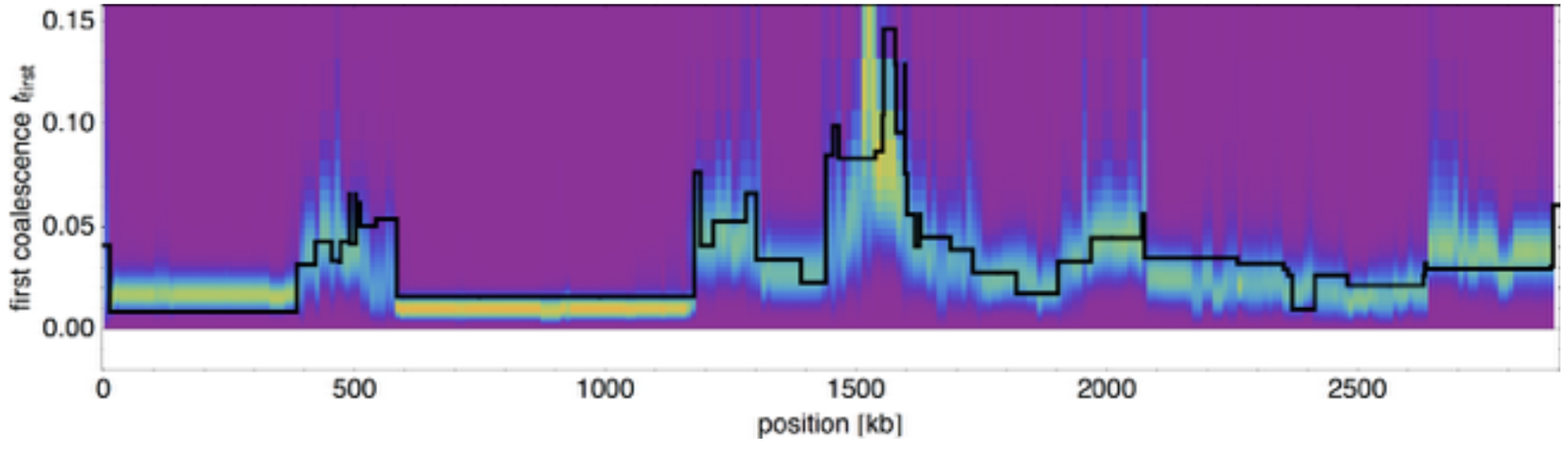
2 haplotypes,
similar to PSMC
[Li and Durbin, 2011]



4 haplotypes

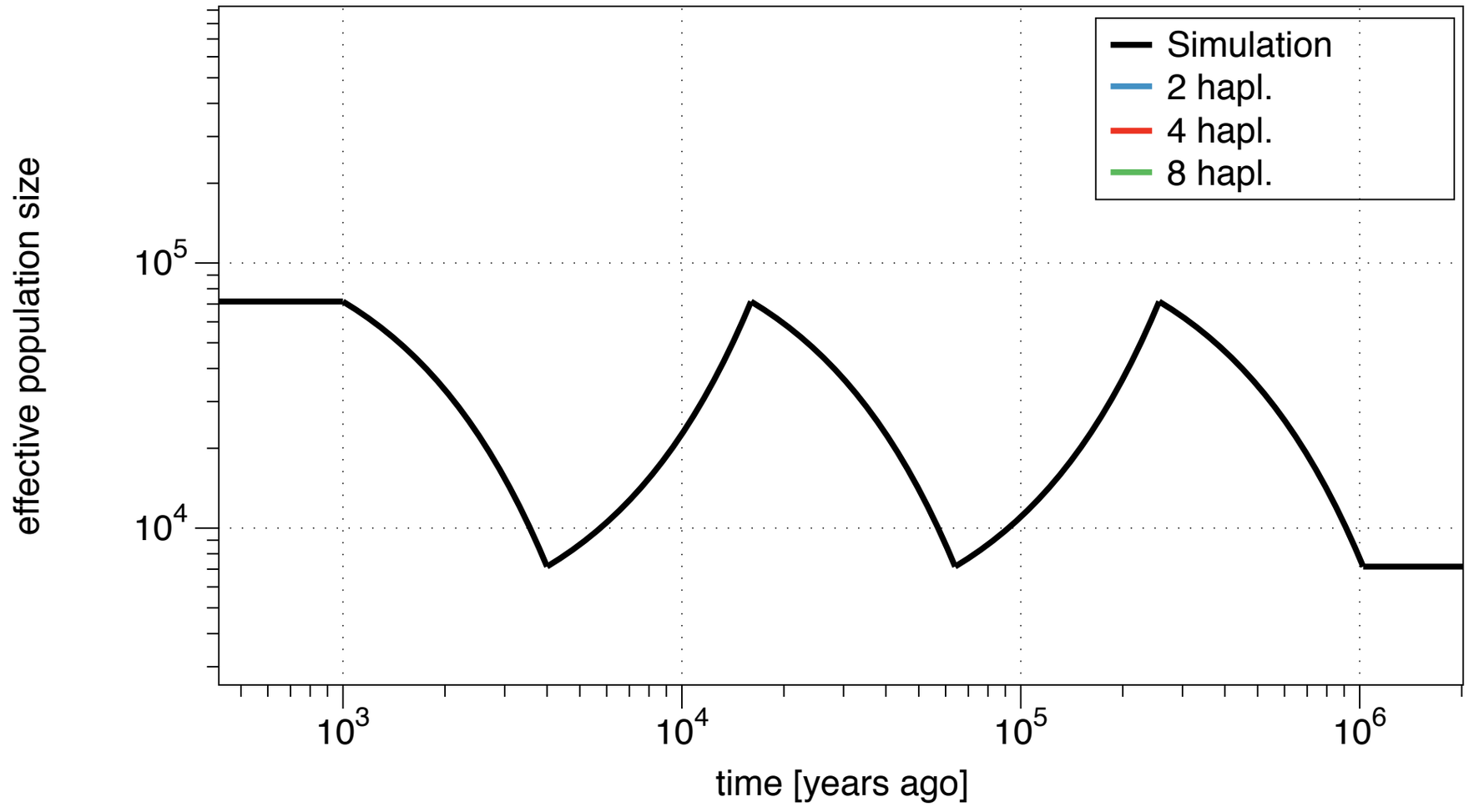


8 haplotypes



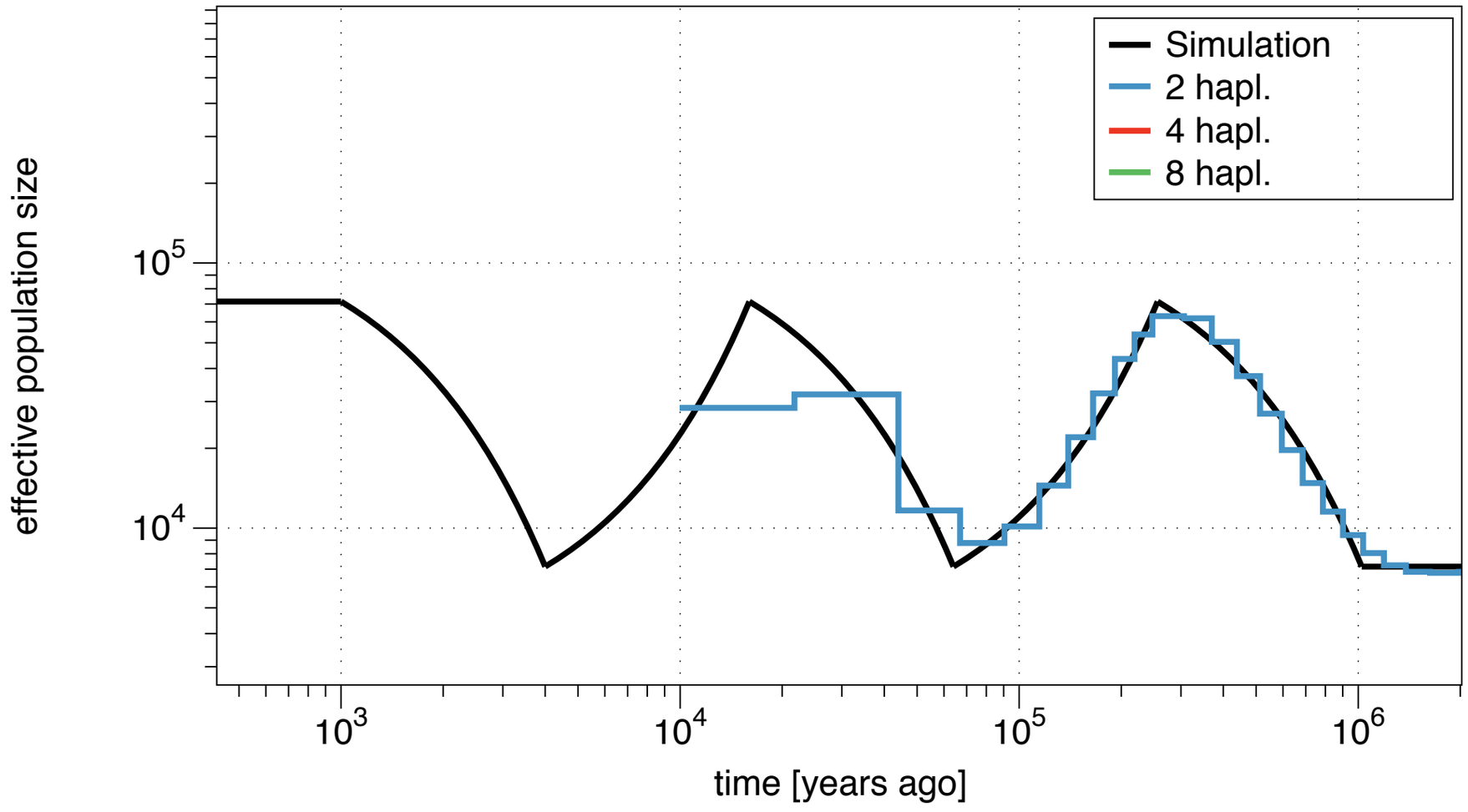
Test with simulations

Example: Exponentially growing and shrinking population size



Test with simulations

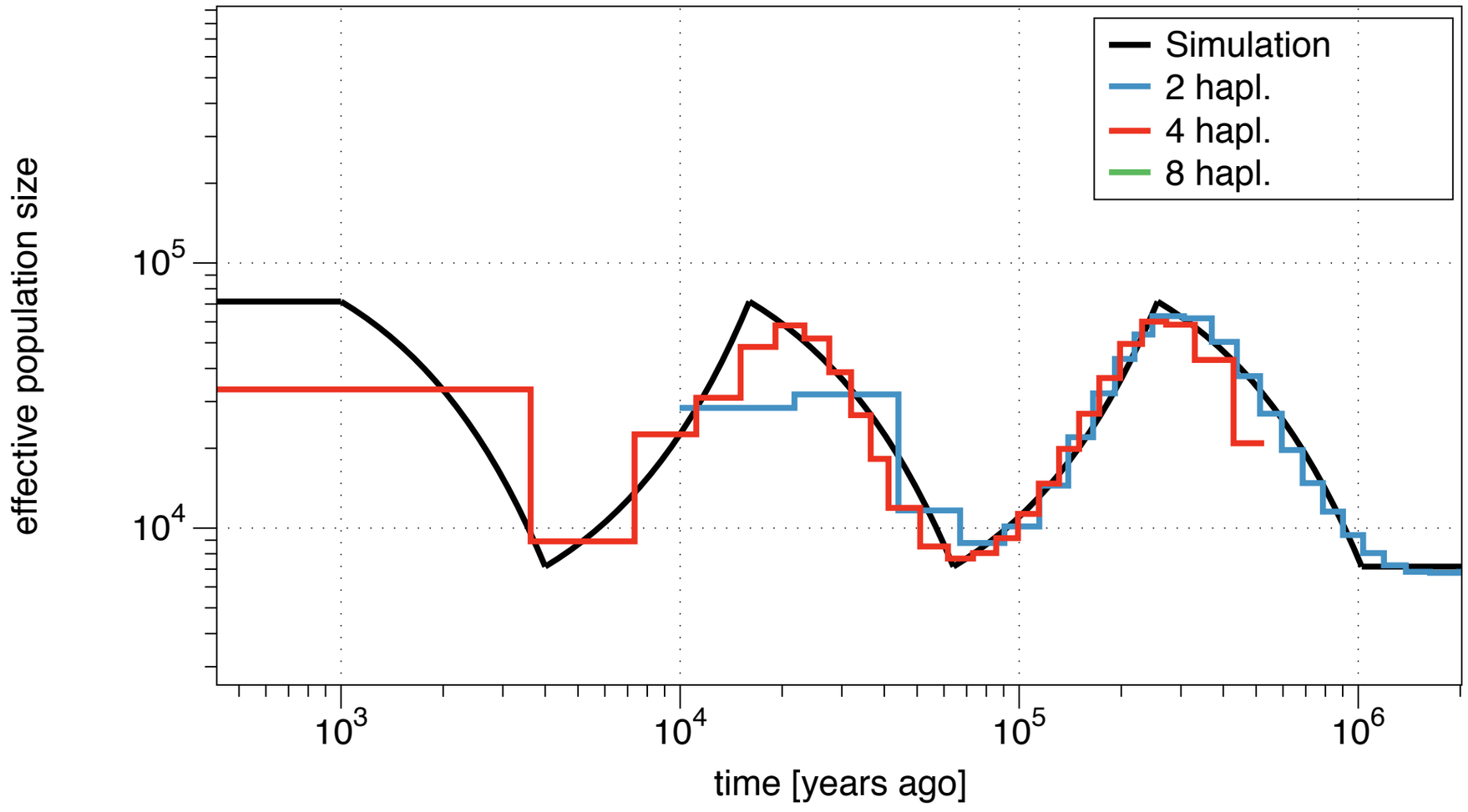
Example: Exponentially growing and shrinking population size



2 haplotypes: 40kya-3Mya

Test with simulations

Example: Exponentially growing and shrinking population size

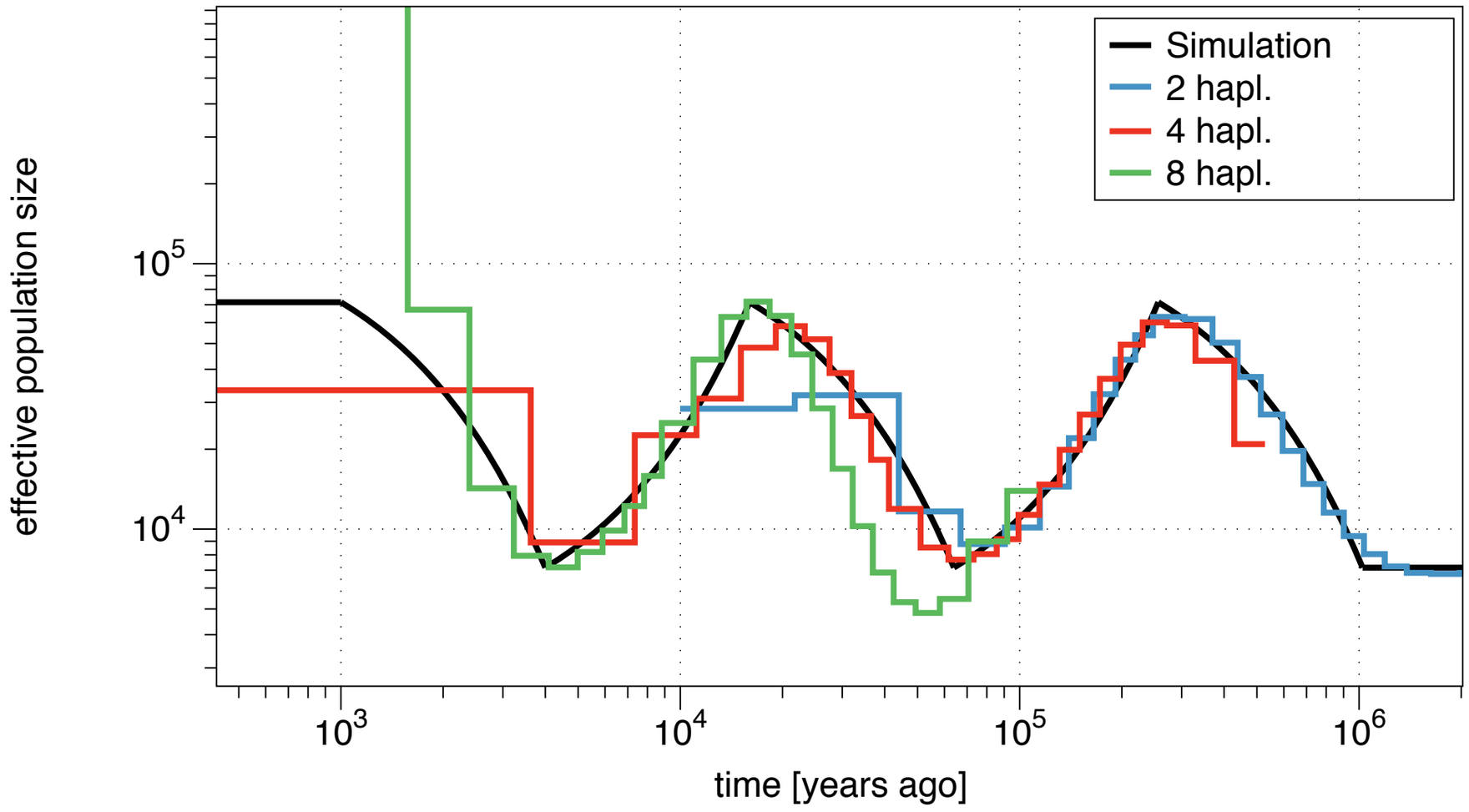


2 haplotypes: 40kya-3Mya

4 haplotypes: 8kya-300kya

Test with simulations

Example: Exponentially growing and shrinking population size

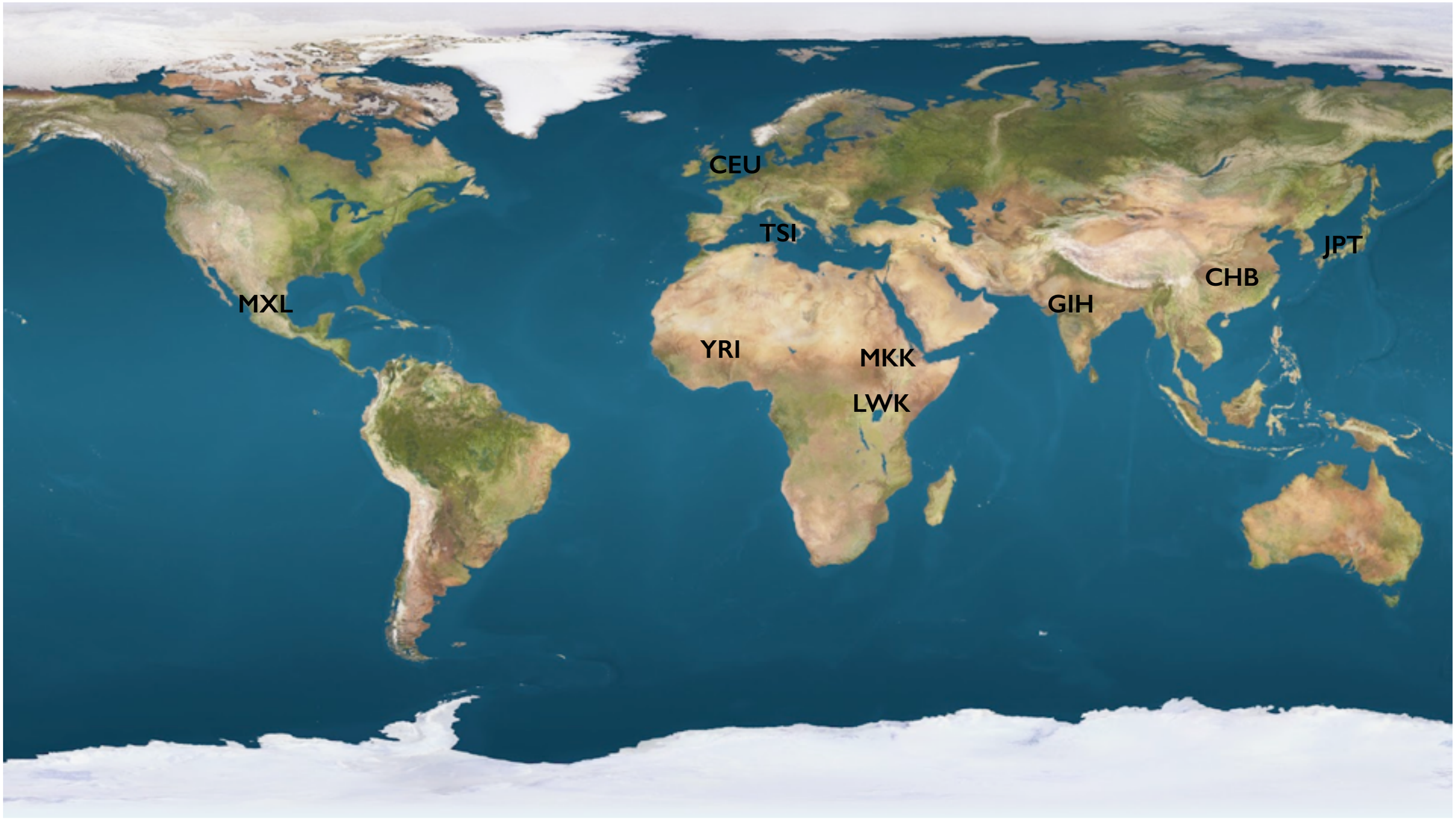


2 haplotypes: 40kya-3Mya

4 haplotypes: 8kya-300kya

8 haplotypes: 2kya-80kya

From genome sequences to human history

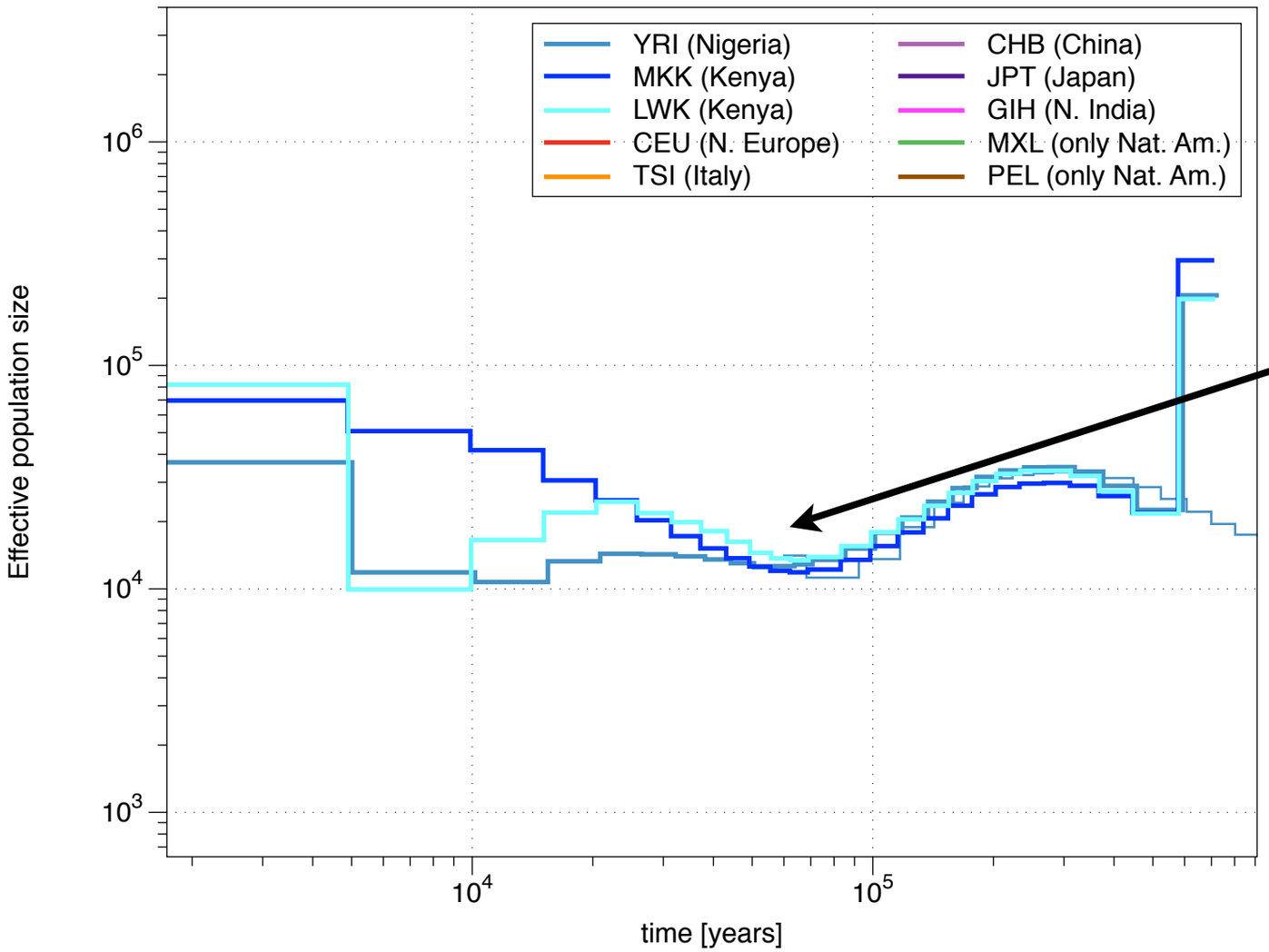


[Sequence data from Complete Genomics]

Inferring historical Population Sizes

real time scaling using
mutation rate per generation
$\mu = 1.25 \times 10^{-8}$ and a generation
time of 30 years

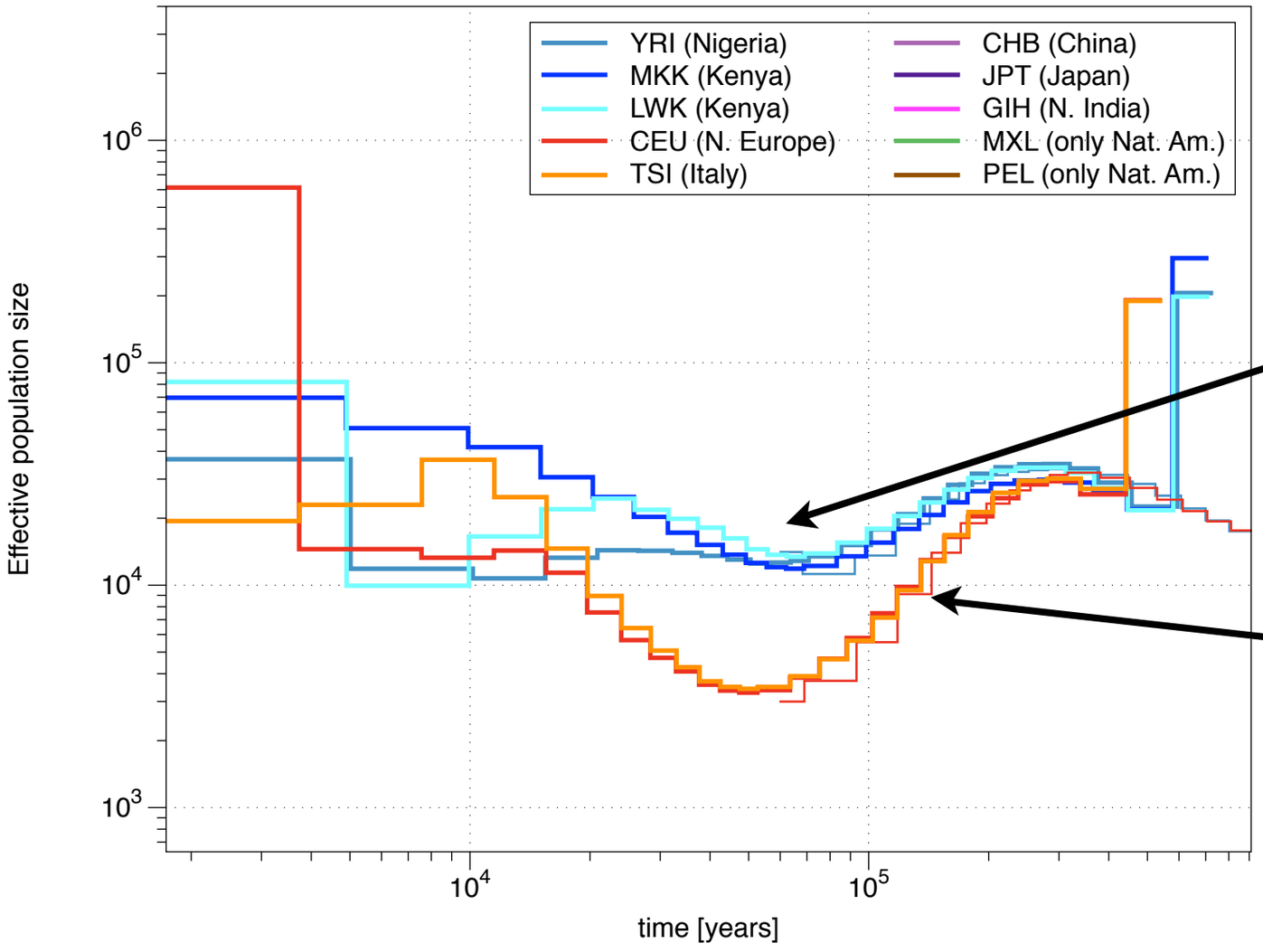
Inferring historical Population Sizes



real time scaling using mutation rate per generation $\mu = 1.25 \times 10^{-8}$ and a generation time of 30 years

Mild bottleneck in African ancestors

Inferring historical Population Sizes

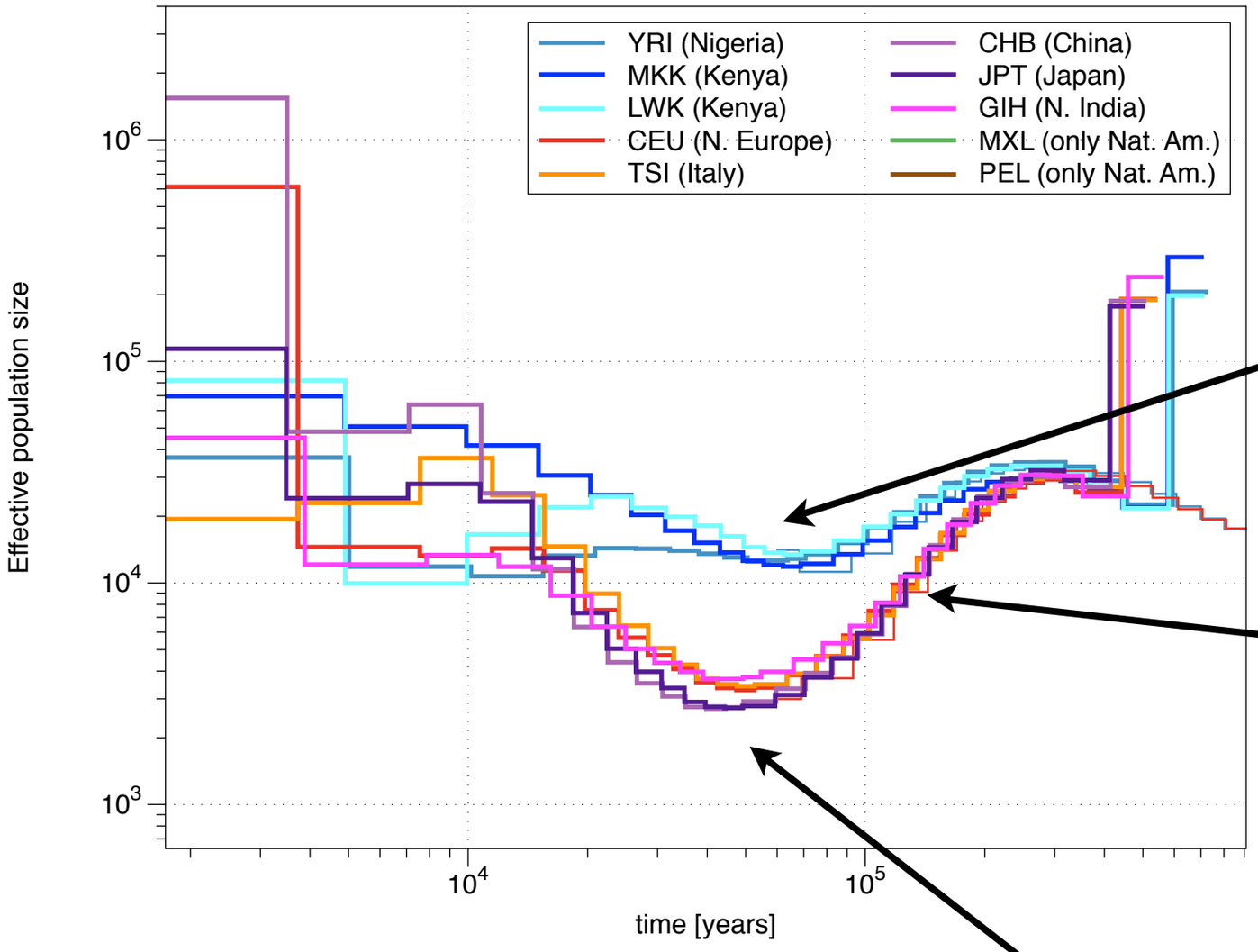


real time scaling using mutation rate per generation $\mu = 1.25 \times 10^{-8}$ and a generation time of 30 years

Mild bottleneck in African ancestors

identical history of non-African populations before 50kya

Inferring historical Population Sizes



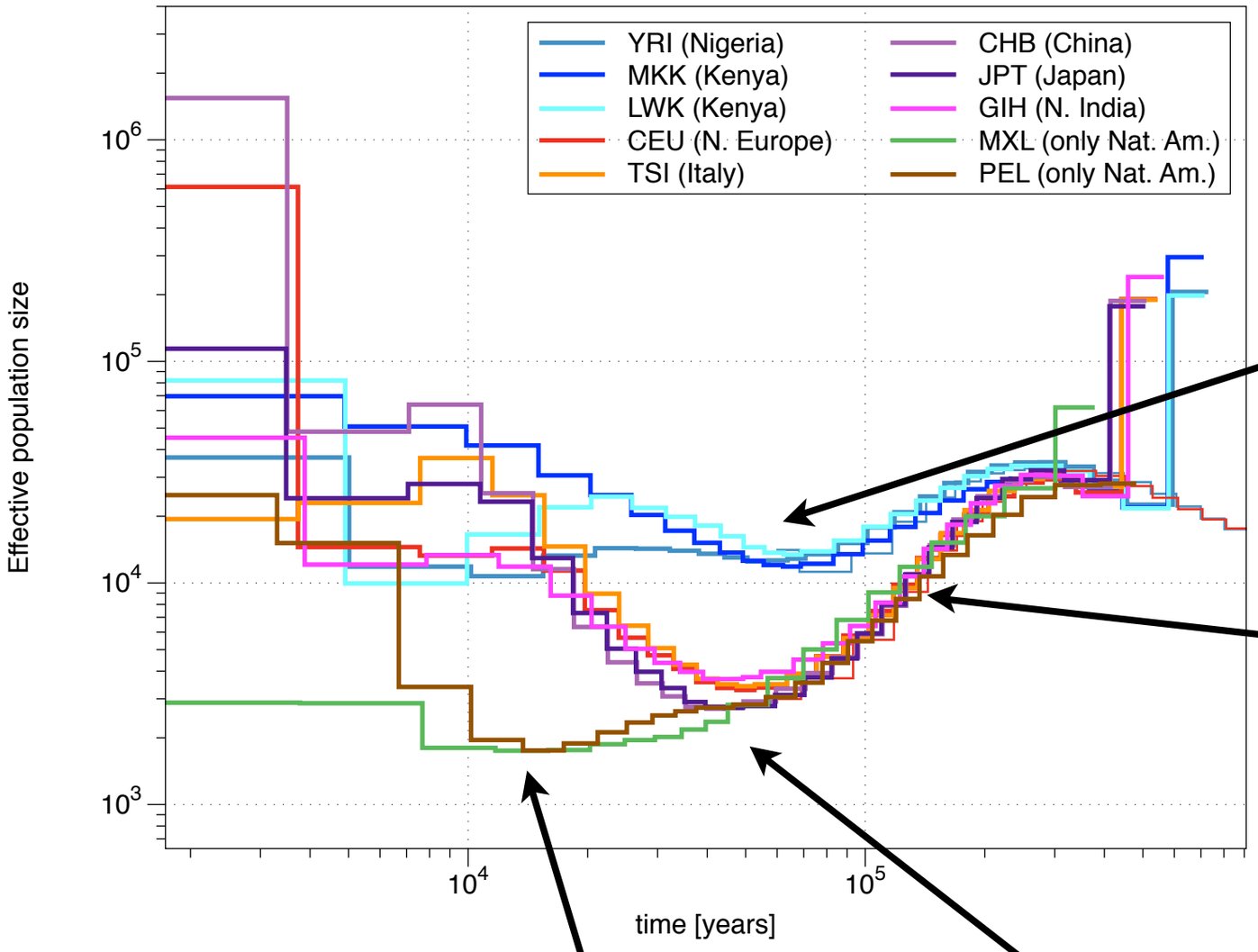
real time scaling using mutation rate per generation $\mu = 1.25 \times 10^{-8}$ and a generation time of 30 years

Mild bottleneck in African ancestors

identical history of non-African populations before 50kya

strong bottleneck in non-African populations 50-60kya, coincides with out-of-Africa dispersal

Inferring historical Population Sizes



real time scaling using mutation rate per generation $\mu = 1.25 \times 10^{-8}$ and a generation time of 30 years

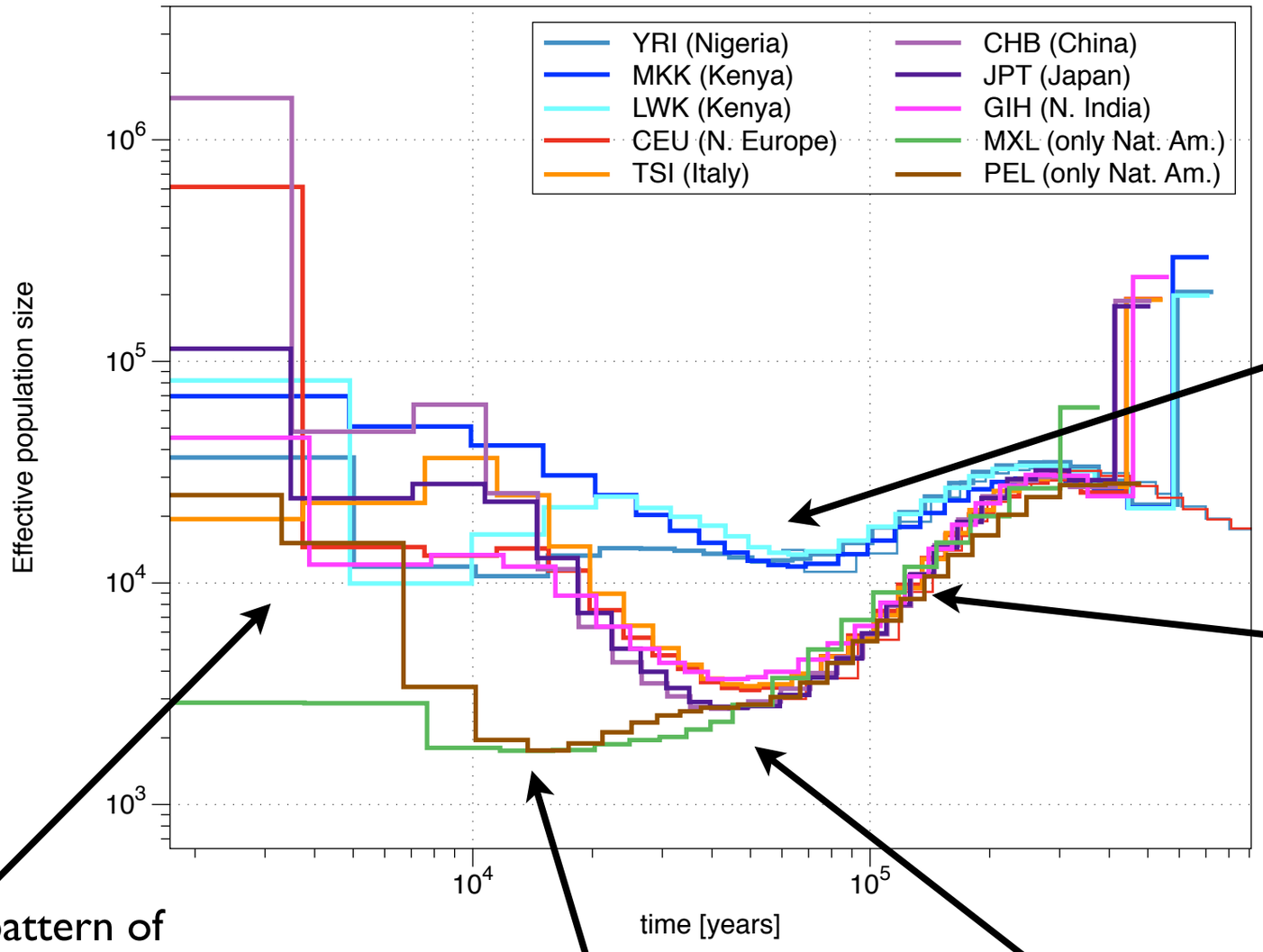
Mild bottleneck in African ancestors

identical history of non-African populations before 50kya

strong bottleneck in non-African populations 50-60kya, coincides with out-of-Africa dispersal

deep bottleneck in Native American populations around 15kya: Beginning of the Peopling of America via Alaskan Landbridge

Inferring historical Population Sizes



real time scaling using mutation rate per generation $\mu = 1.25 \times 10^{-8}$ and a generation time of 30 years

Mild bottleneck in African ancestors

identical history of non-African populations before 50kya

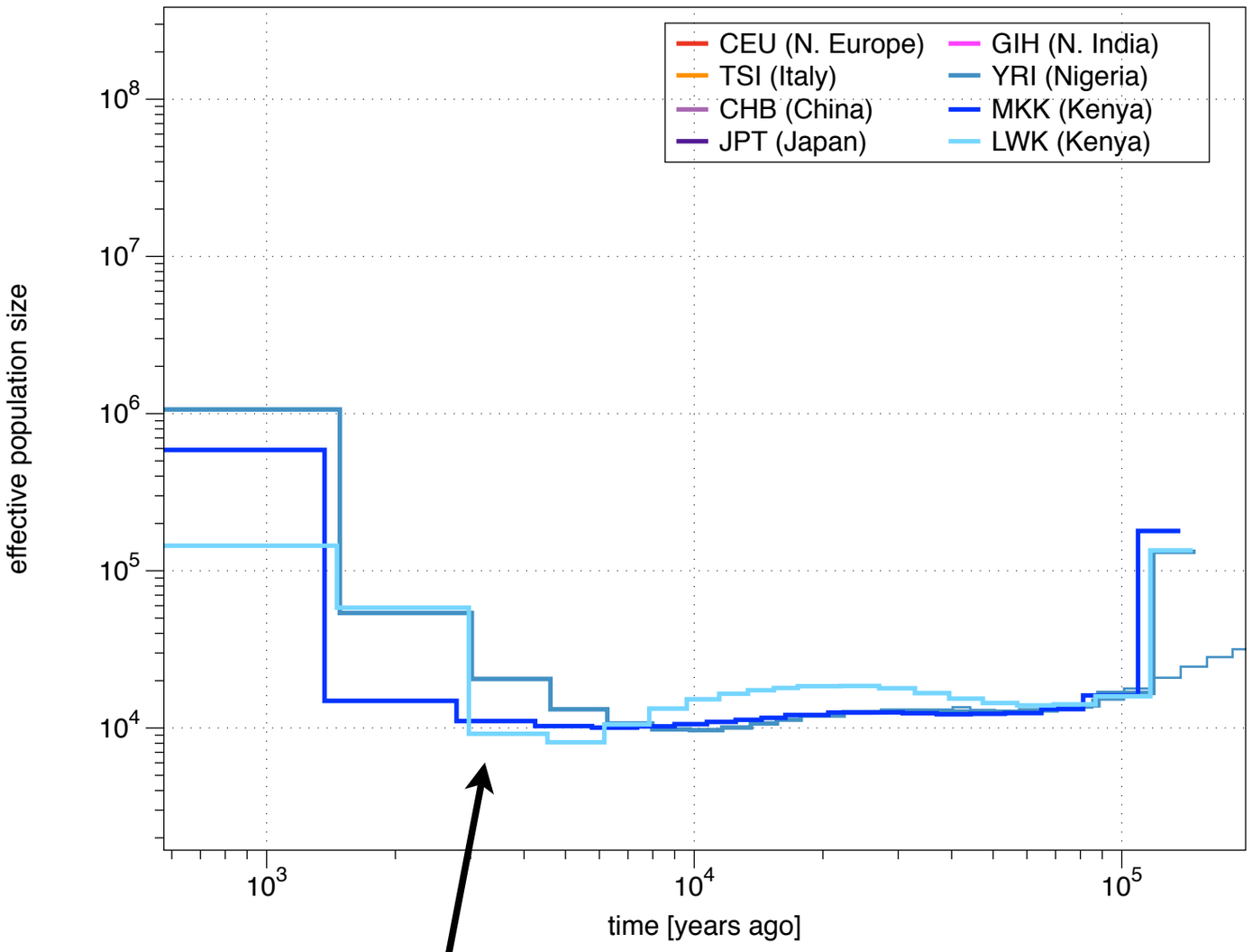
Complex pattern of neolithic expansions in the last 10,000 years within Europe, Asia and Africa

deep bottleneck in Native American populations around 15kya: Beginning of the Peopling of America via Alaskan Landbridge

strong bottleneck in non-African populations 50-60kya, coincides with out-of-Africa dispersal

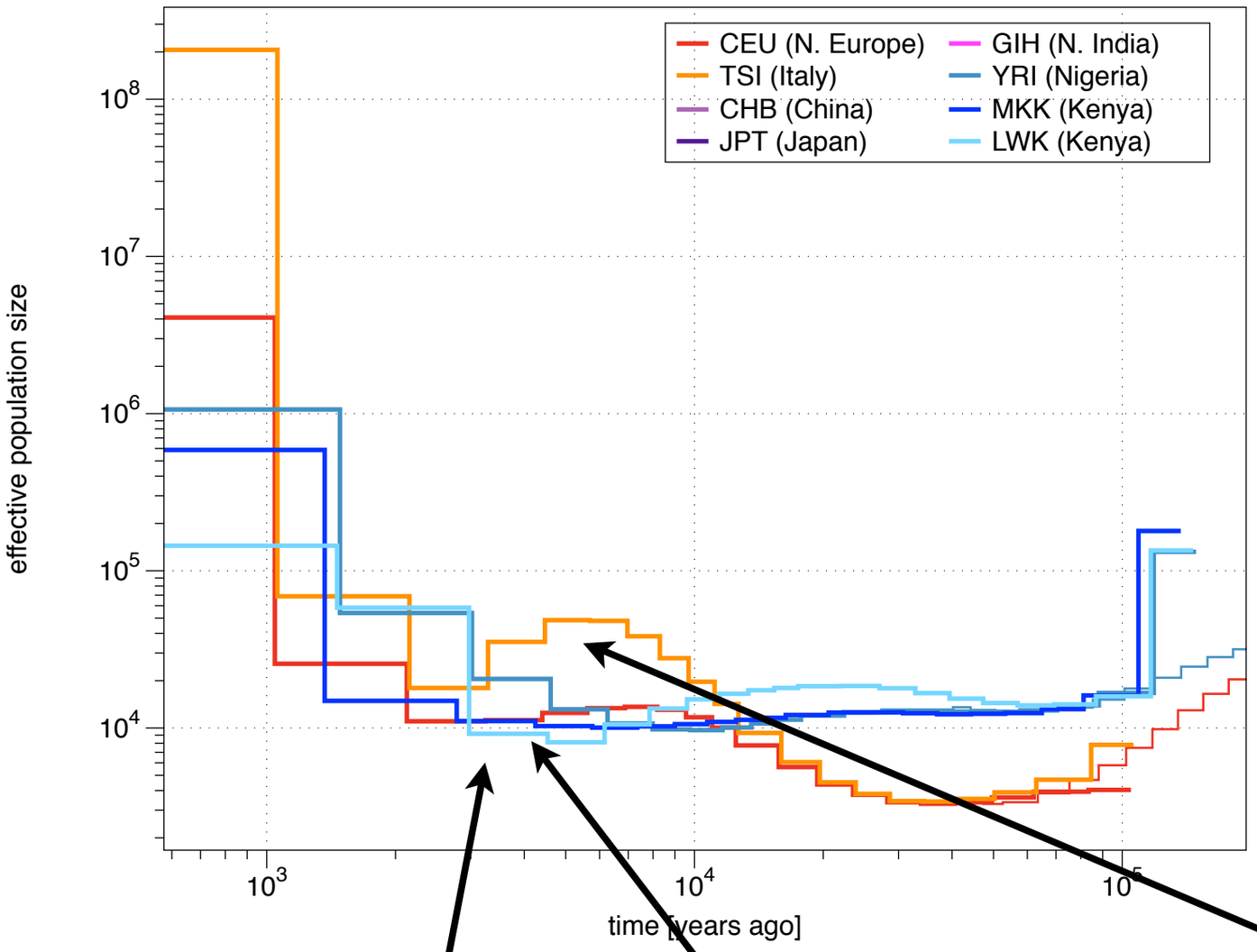
Population size inference from 8 haplotypes

Population size inference from 8 haplotypes



Bantu Expansion within Africa?

Population size inference from 8 haplotypes

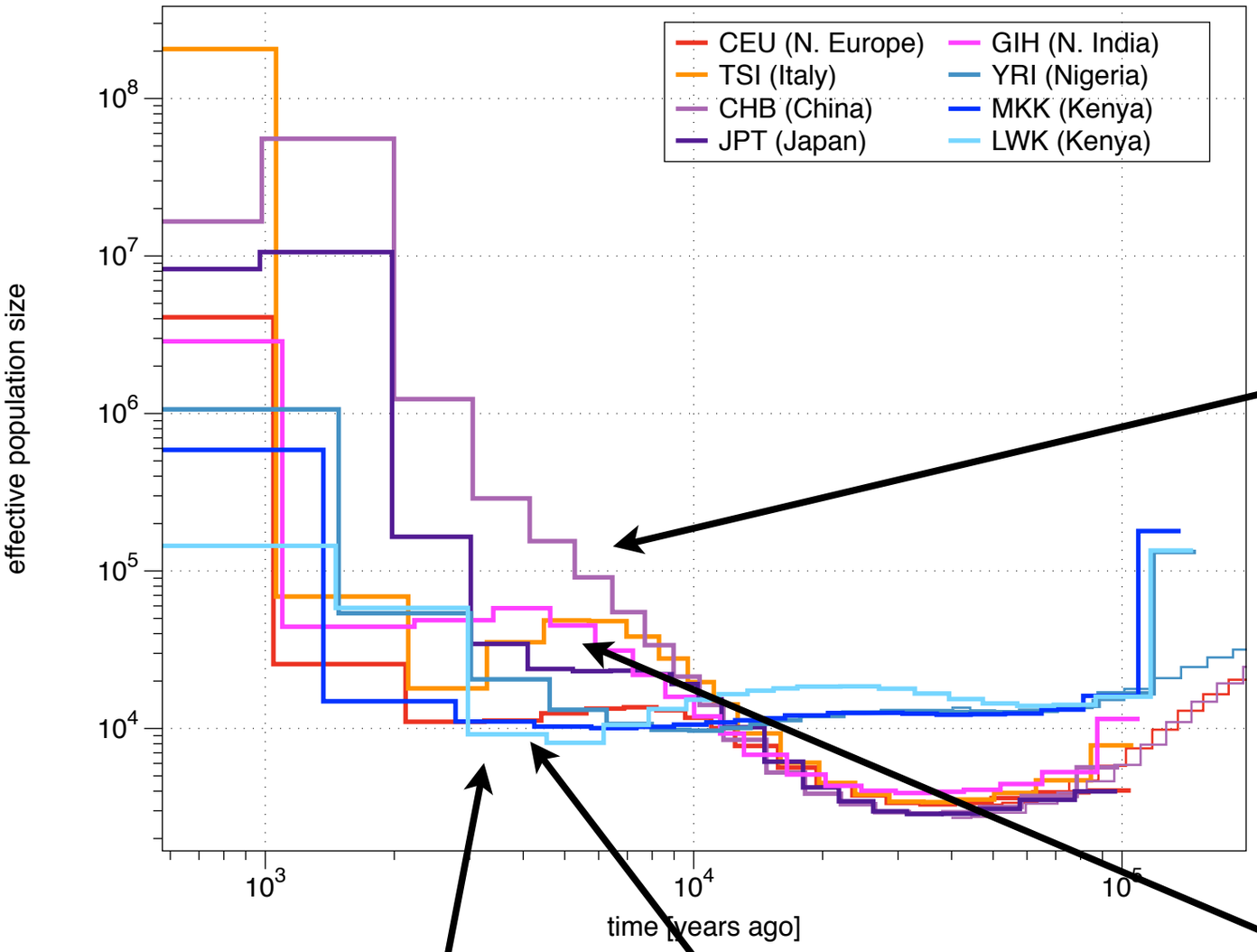


Bantu Expansion within Africa?

CEU: Constant population size until 2kya, then expansion

Structure and admixture in Southern Europe [see Botigué et al., 2013]

Population size inference from 8 haplotypes



Continuous expansion in Chinese ancestors

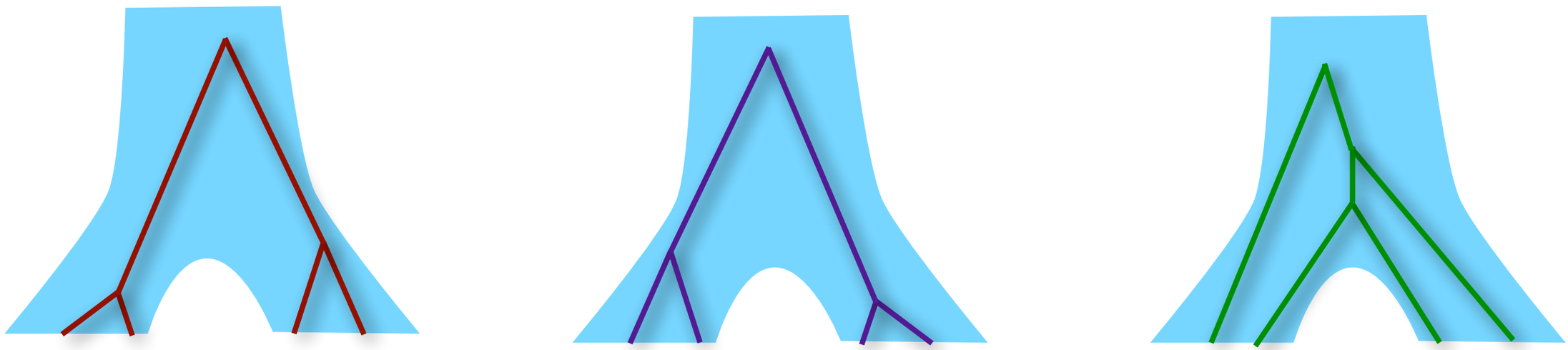
Structure and admixture in Southern Europe [see Botigué et al., 2013]

CEU: Constant population size until 2kya, then expansion

Bantu Expansion within Africa?

Divergence between populations

- Idea: Infer separate coalescence rates within and between populations:



First Coalescence within Population 1

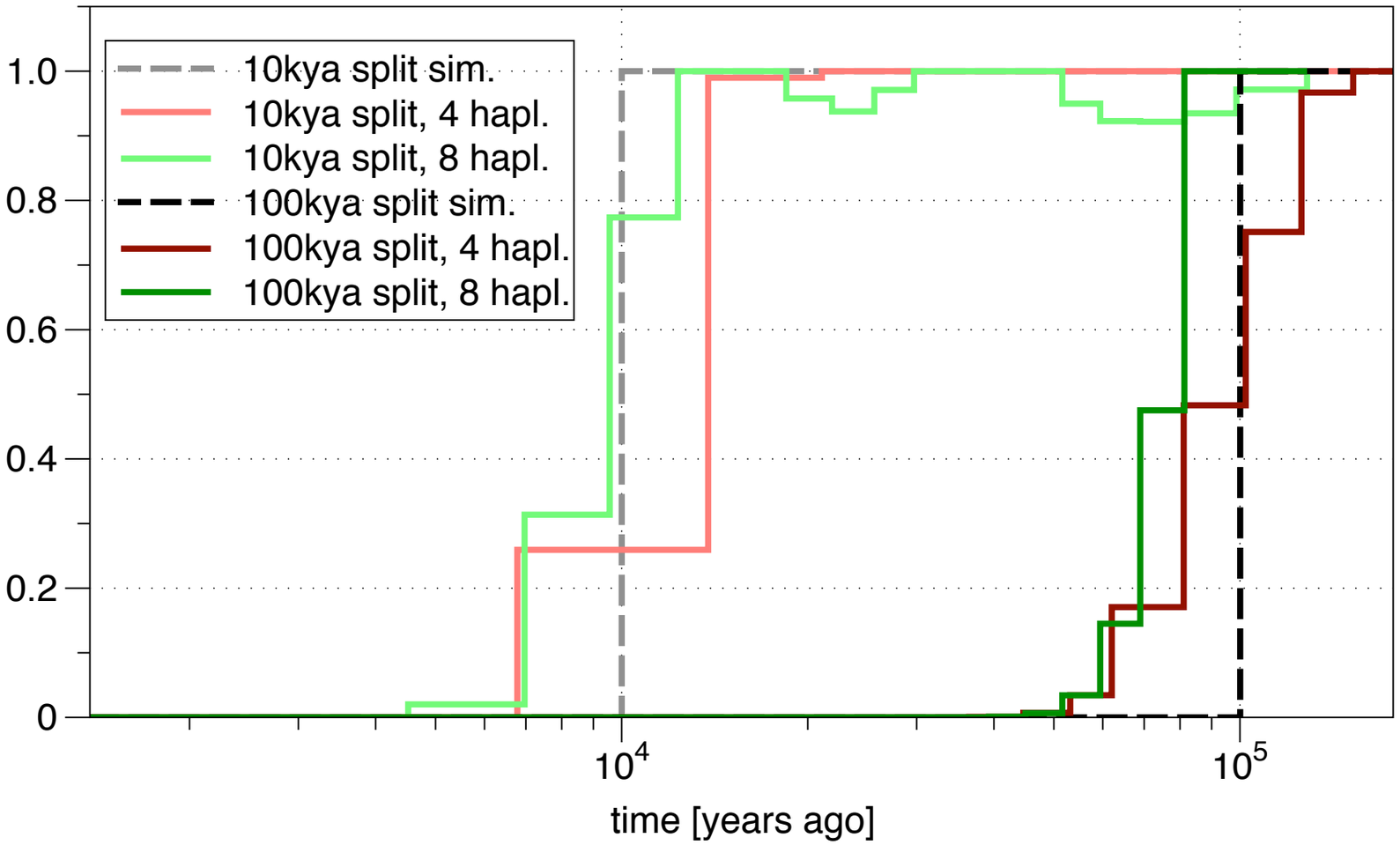
First Coalescence within Population 2

First Coalescence across both populations

- MSMC can infer separate coalescence rates within populations,
- Given rates within populations, $\lambda_{11}(t)$ and $\lambda_{22}(t)$, and across populations, $\lambda_{12}(t)$, compute relative gene flow as ratio

$$m(t) = \frac{\lambda_{12}(t)}{[\lambda_{11}(t) + \lambda_{22}(t)] / 2}$$

Testing gene flow inference with simulated split



☜ $m(t)=1$: perfectly mixed

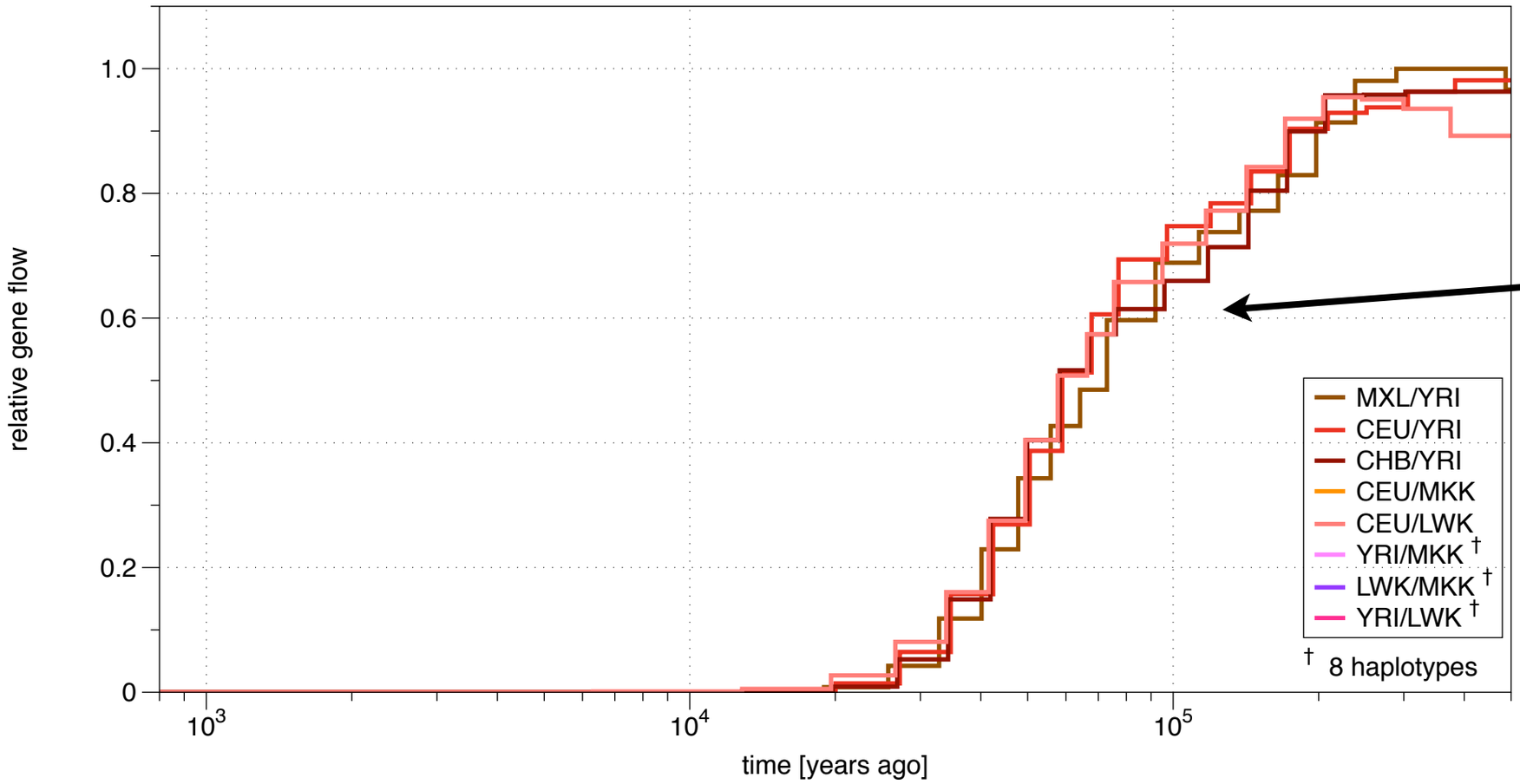
☜ $m(t)=0$: perfectly split

4 haplotypes: good for splits 50-200kya.

8 haplotypes: good for splits 5-50kya.

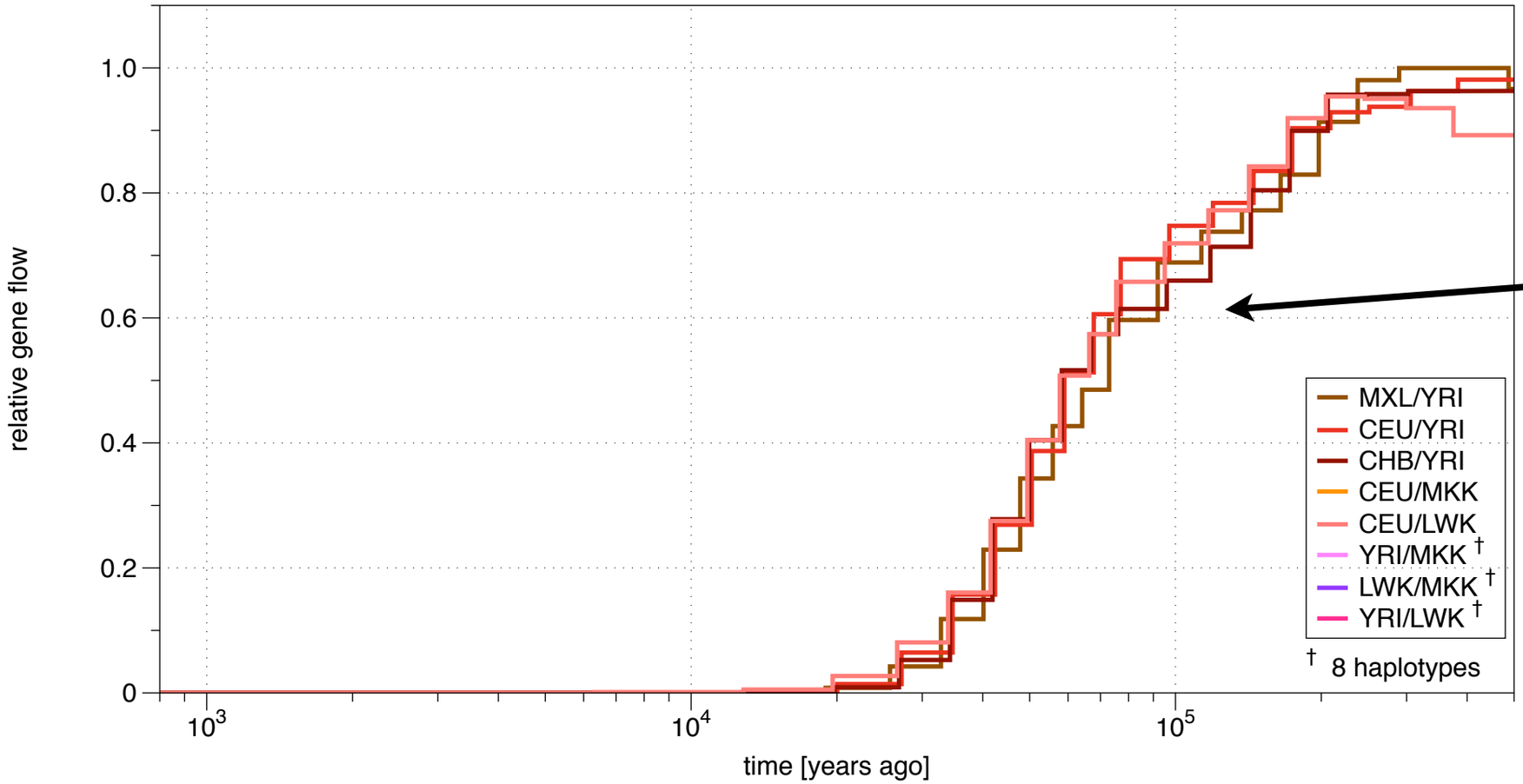
African Population separations

African Population separations

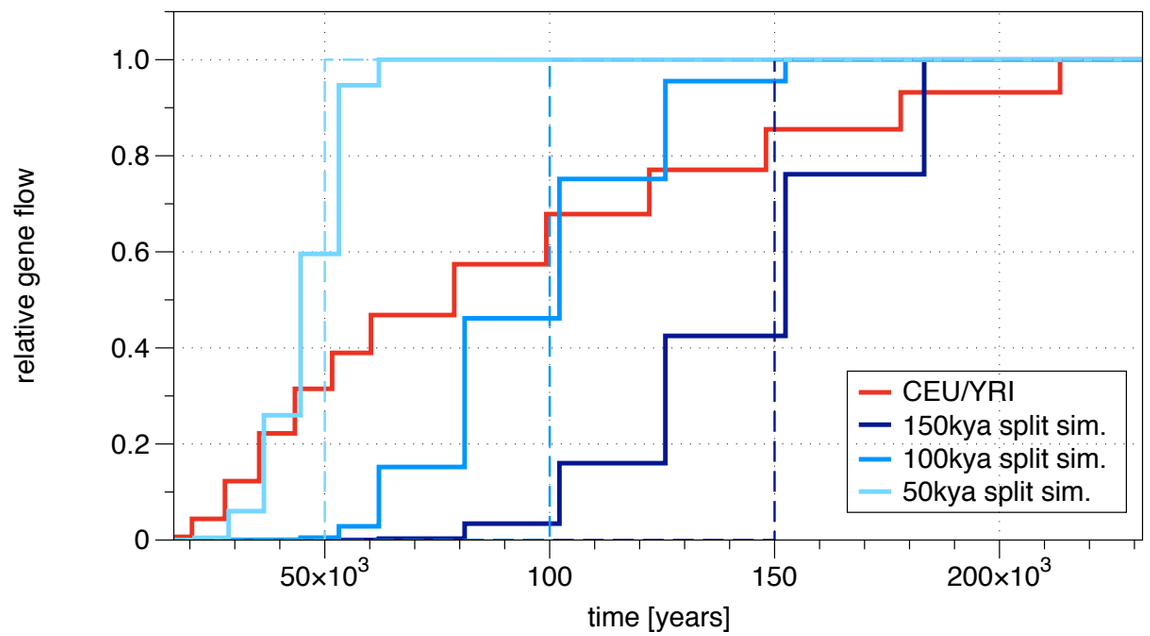


African(YRI)/Non-African separation: 150kya-40kya

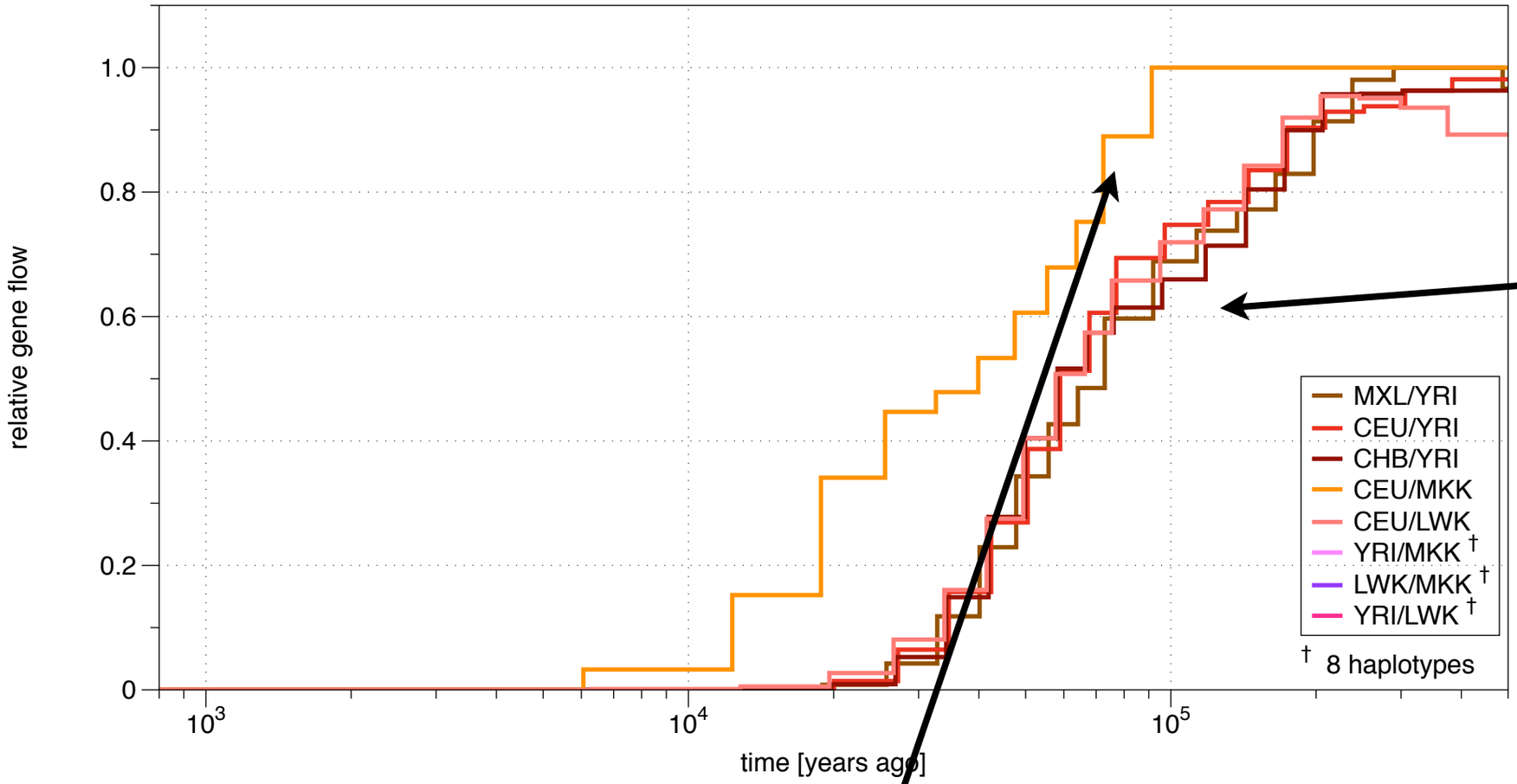
African Population separations



African(YRI)/Non-African separation: 150kya-40kya

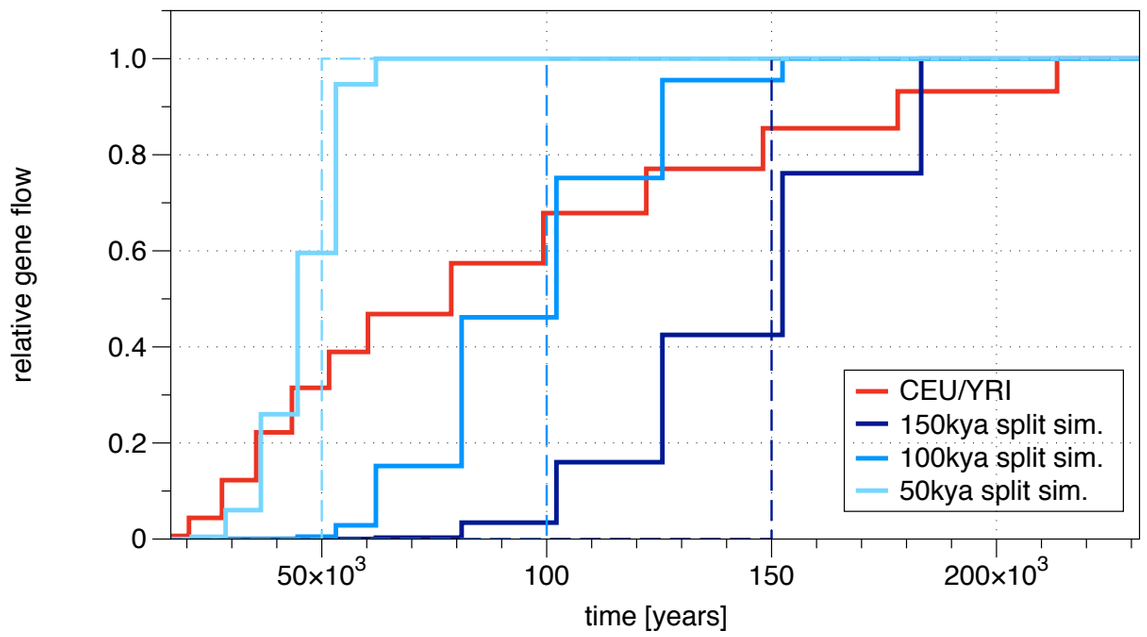


African Population separations

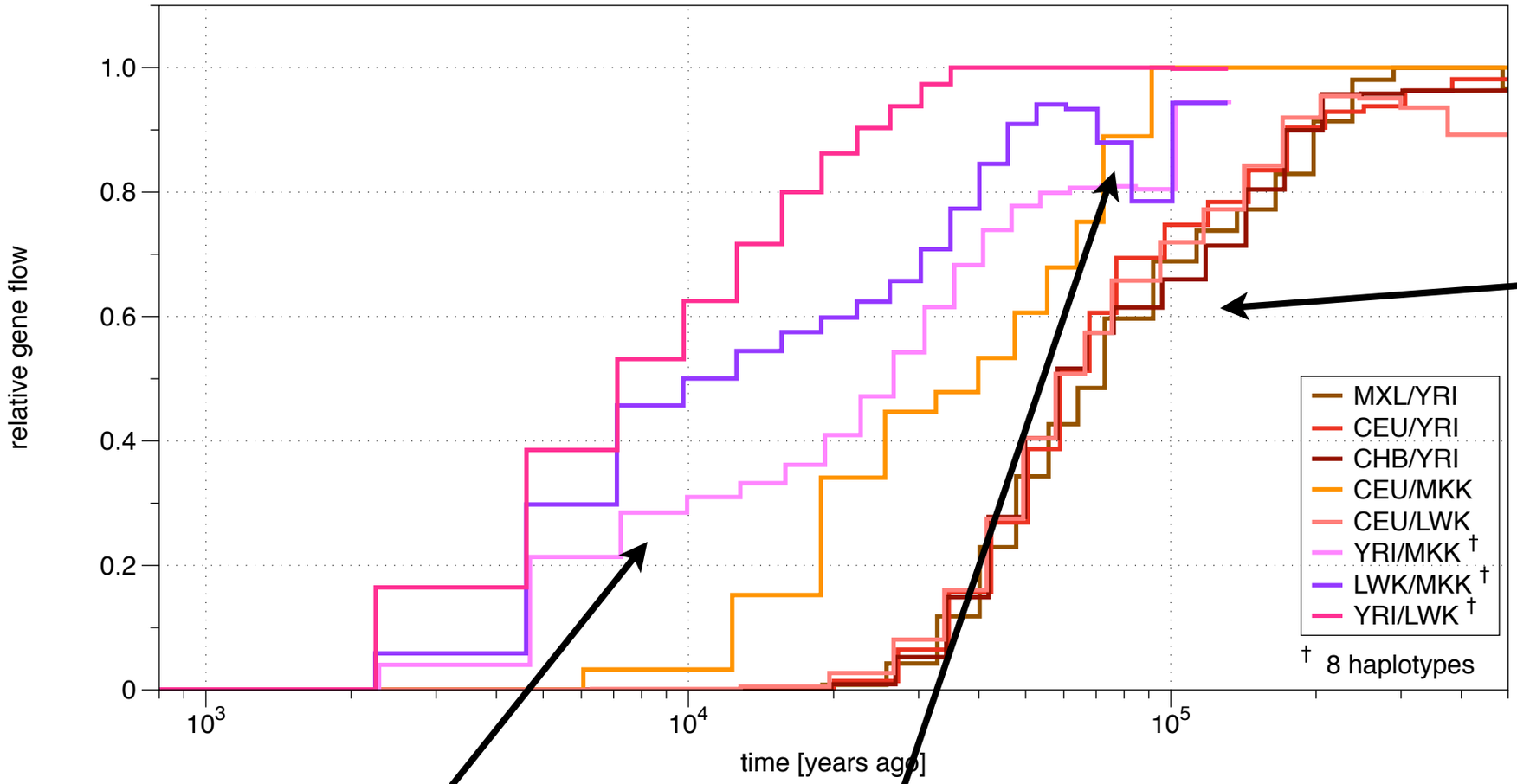


African(YRI)/Non-African separation: 150kya-40kya

Masaai remained mixing with European ancestors until ~80kya



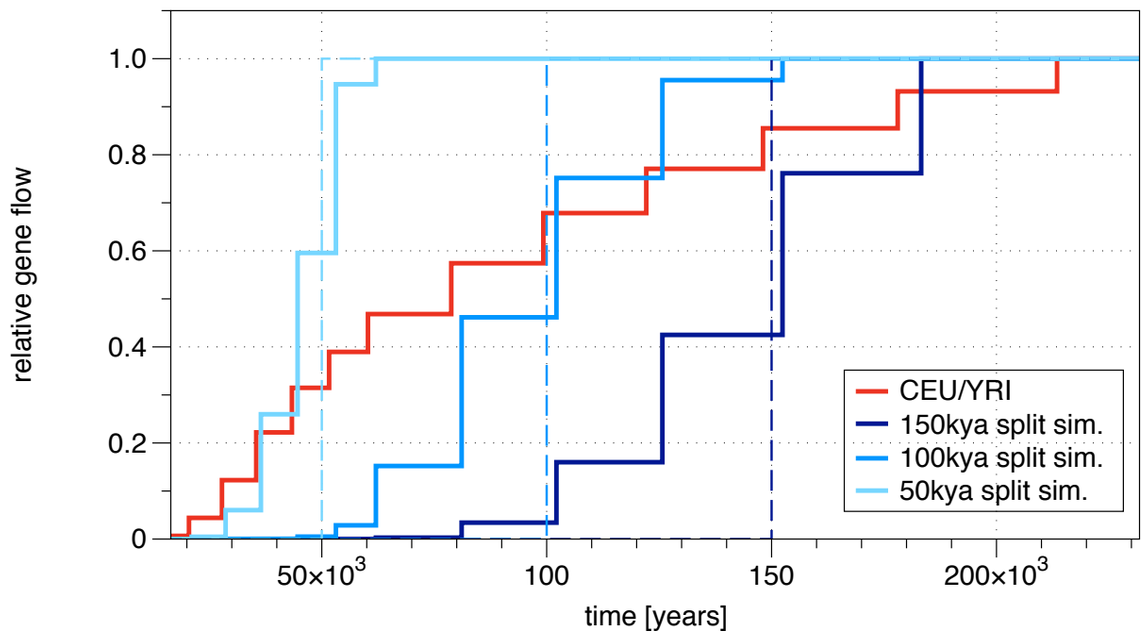
African Population separations



African(YRI)/Non-African separation: 150kya-40kya

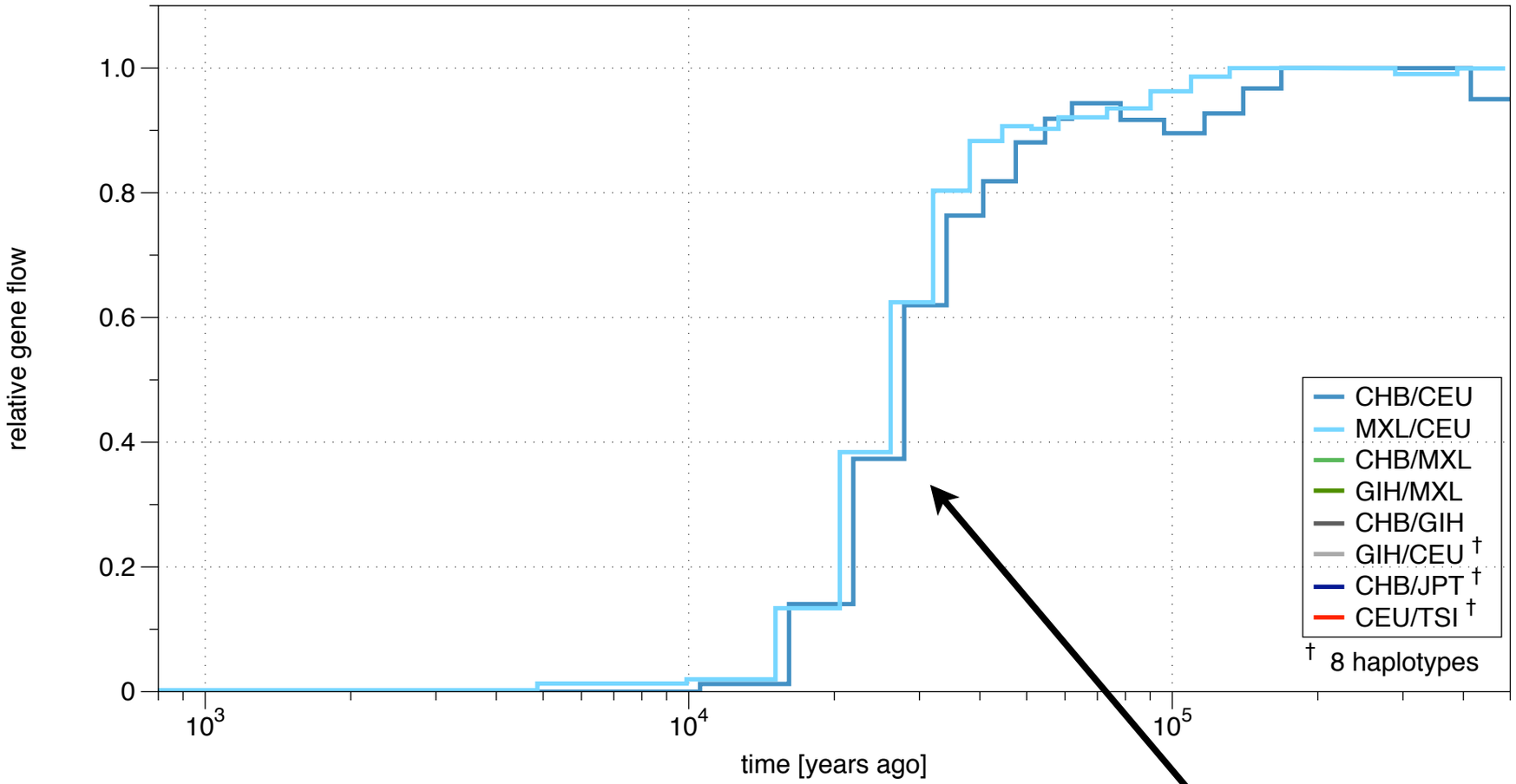
Complex splits within Africa: 5kya-30kya

Masaai remained mixing with European ancestors until ~80kya



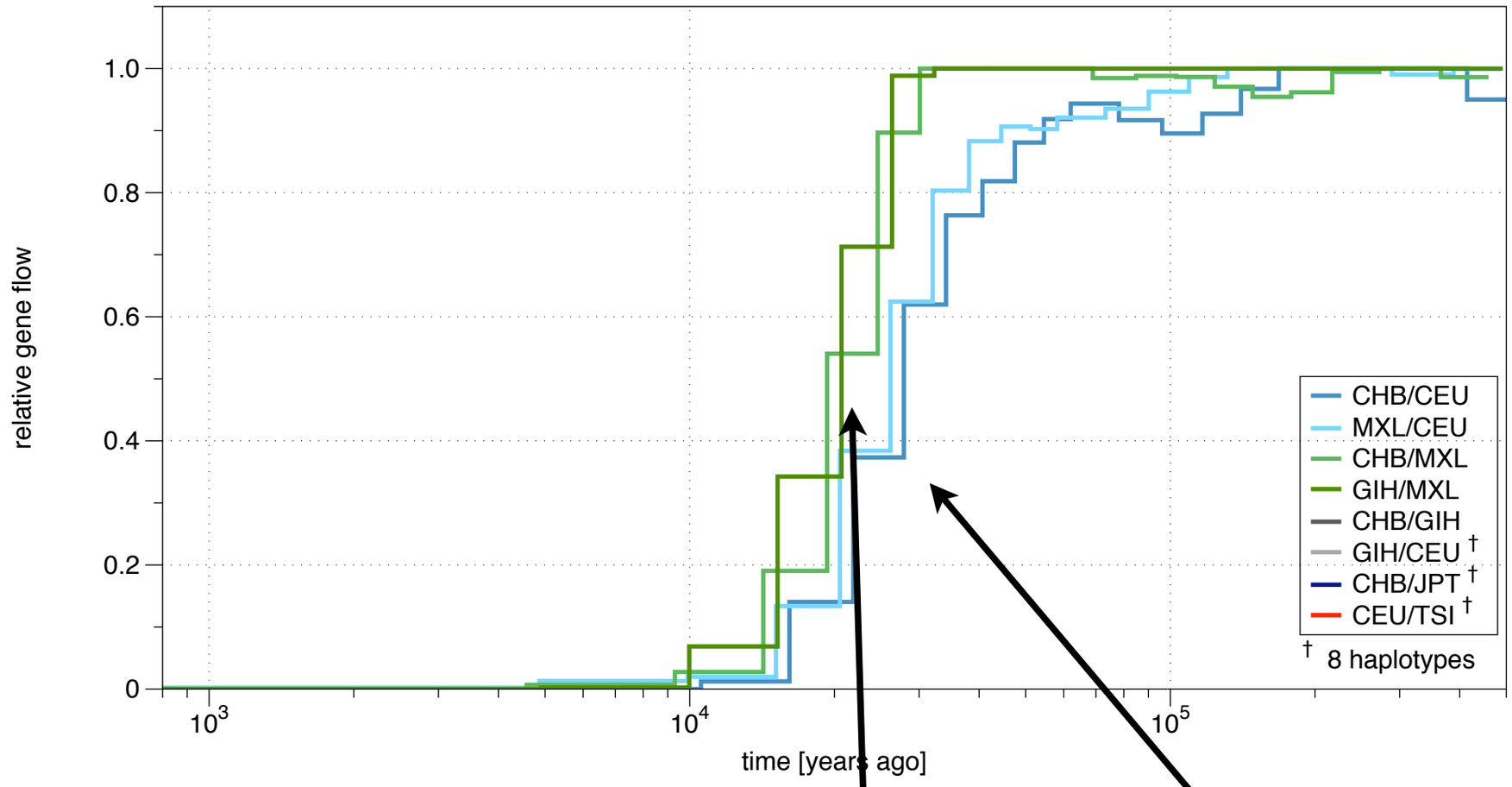
Non-African Population Separations

Non-African Population Separations



European/East-Asian split: 30kya, small older component?

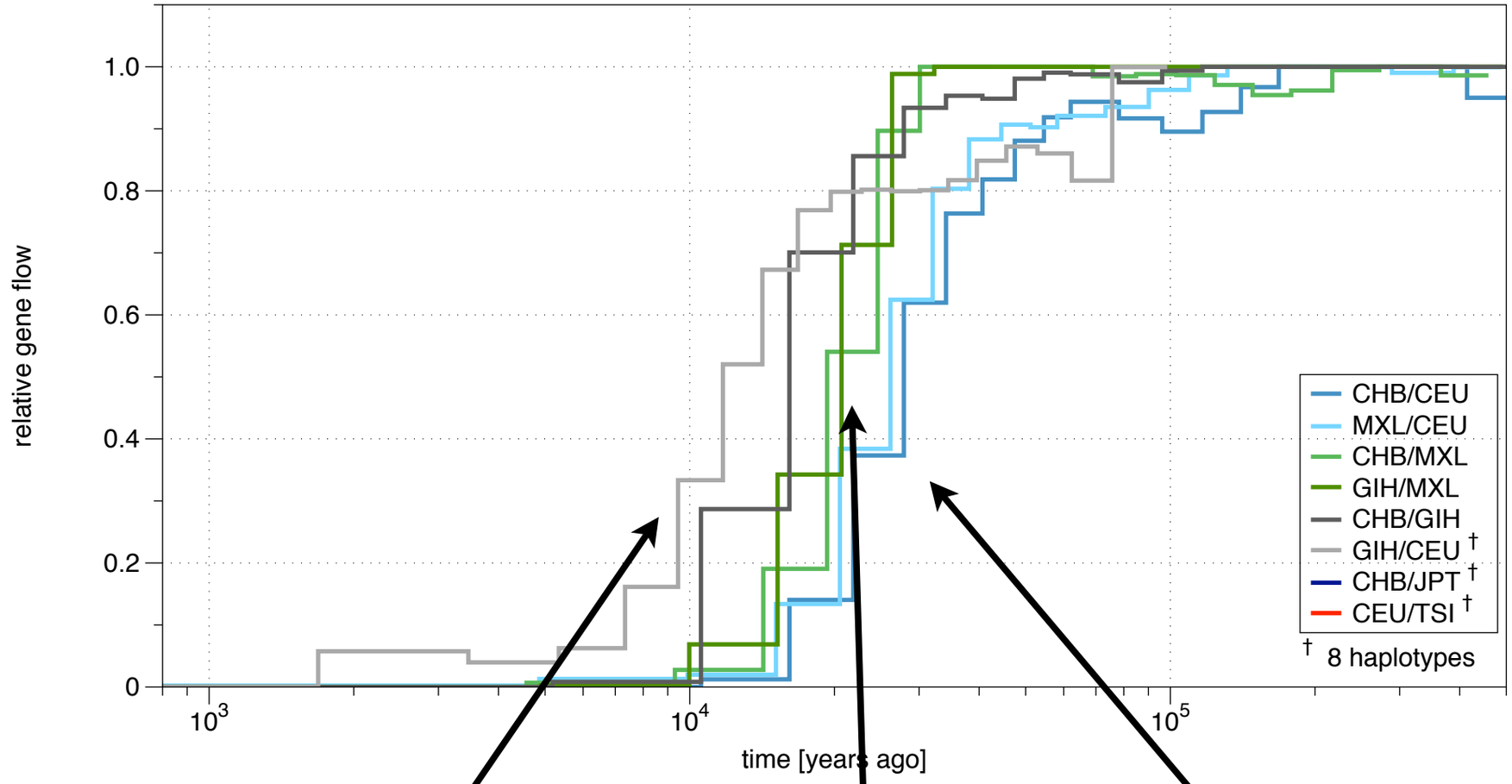
Non-African Population Separations



Asian/American
split: 20kya

European/East-Asian
split: 30kya, small older component?

Non-African Population Separations

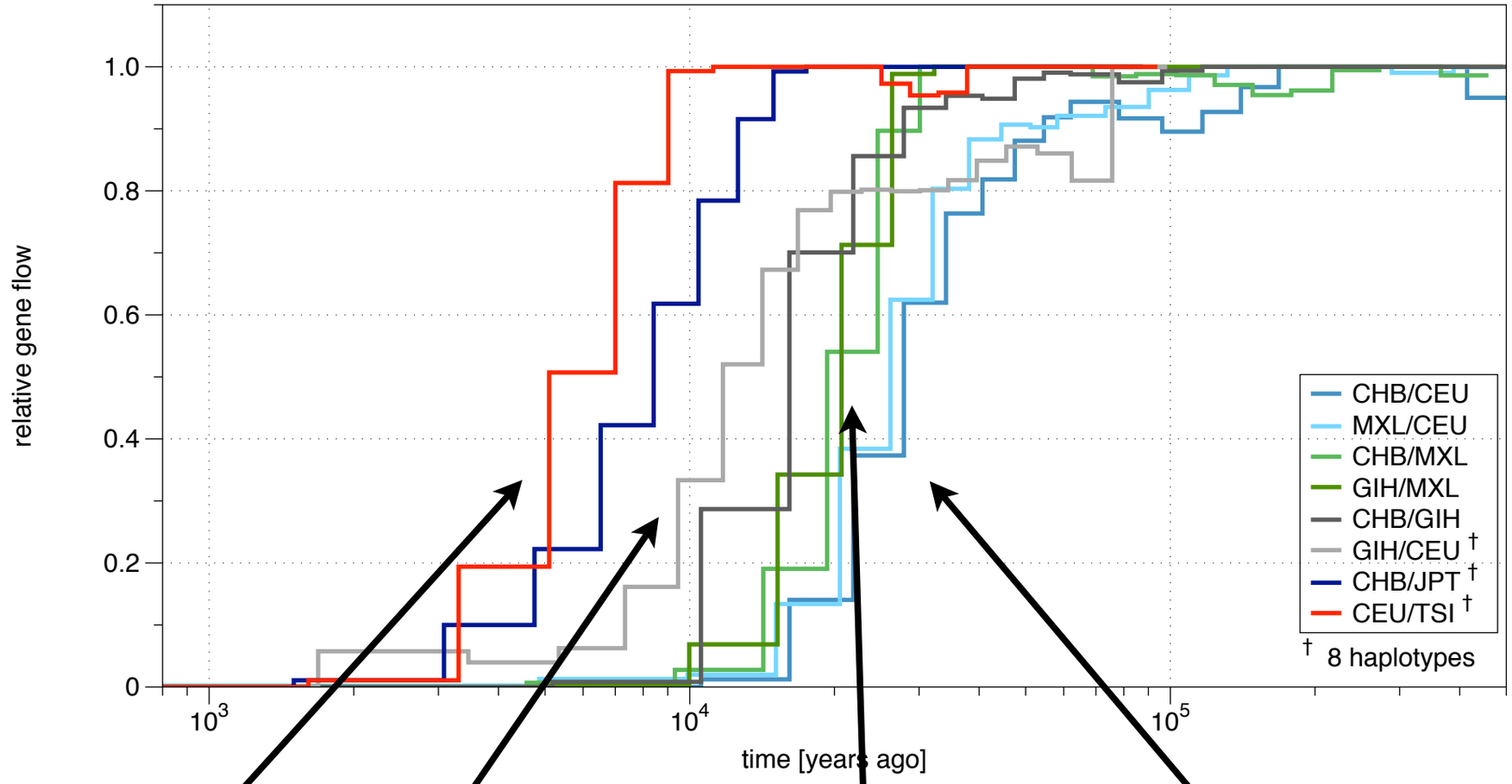


Complex Indian separations between 8kya and 20kya

Asian/American split: 20kya

European/East-Asian split: 30kya, small older component?

Non-African Population Separations



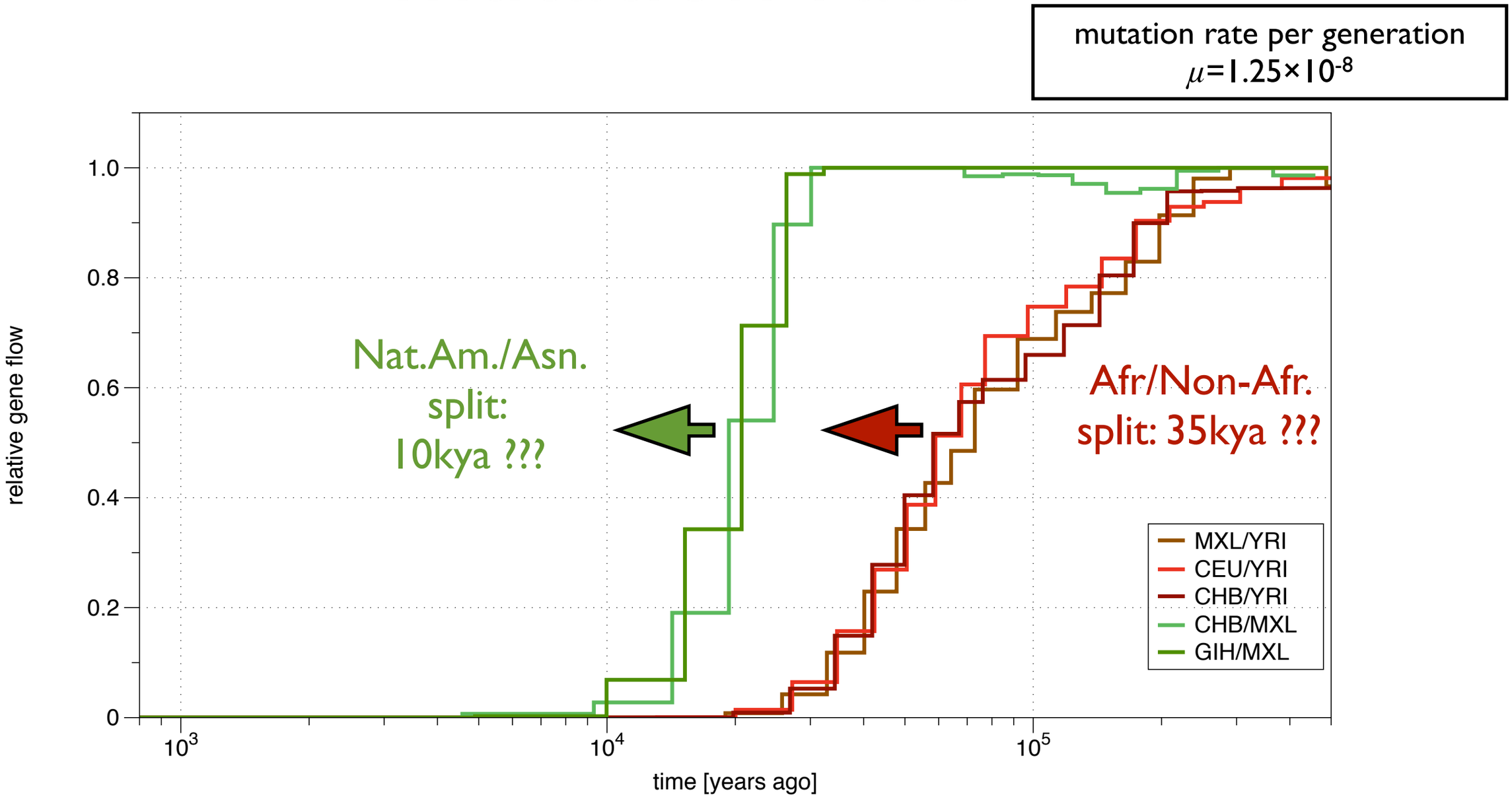
Splits within Europe and within East-Asia

Complex Indian separations between 8kya and 20kya

Asian/American split: 20kya

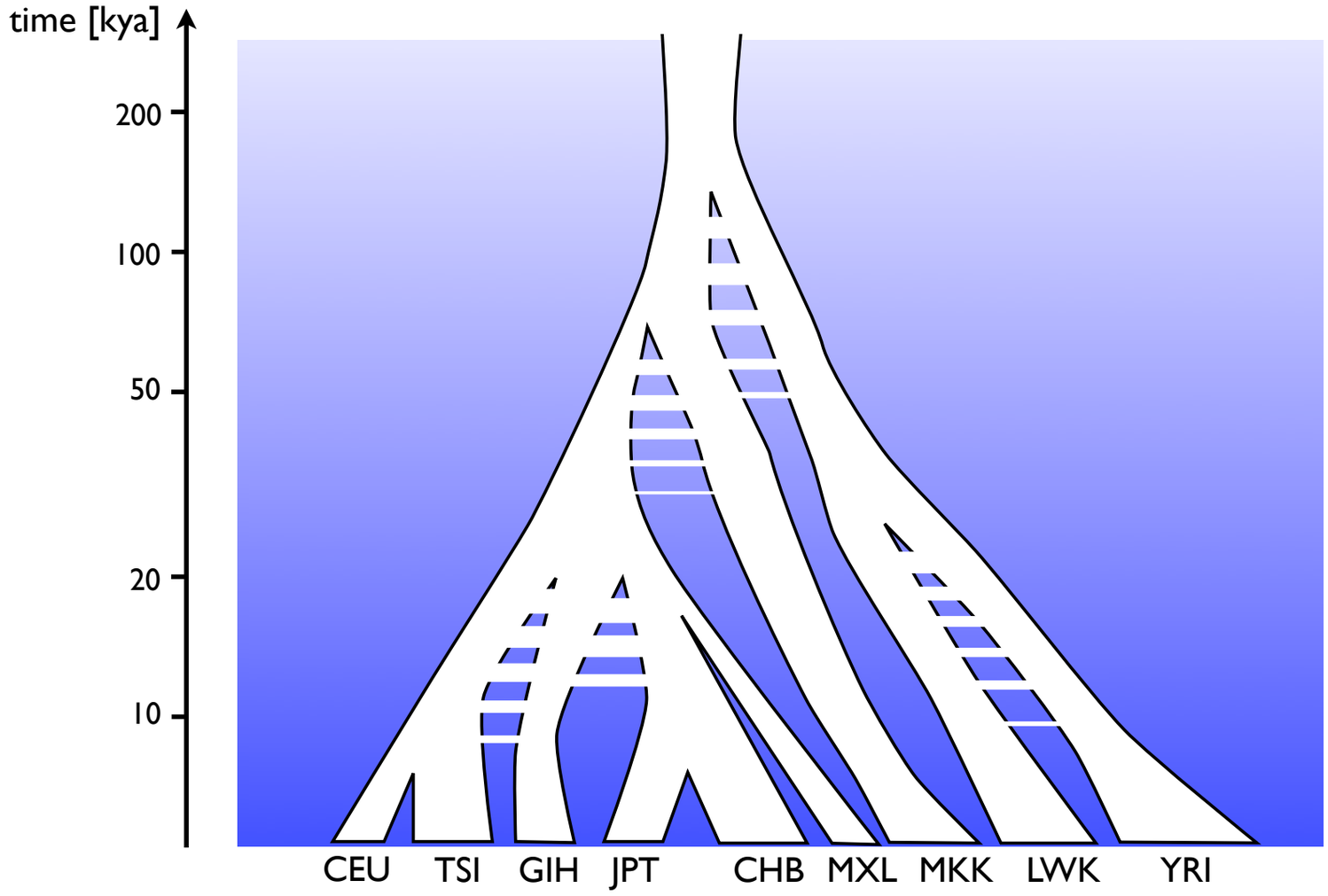
European/East-Asian split: 30kya, small older component?

Implications of higher human mutation rate



A higher mutation rate of 2.5×10^{-8} would push these splits towards too recent times

MSMC Summary on separation history



[Schiffels and Durbin, *Nature Genetics*, in press]

Acknowledgements

- Richard Durbin

- Aylwyn Scally

- Jeff Kidd, Simon Gravel, Eimear Kenny, Carlos Bustamante for MXL and PEL admixture tracts

- Article in press: Schiffels and Durbin, Nature Genetics
  Preprint available on biorxiv.

- Software available on https://github.com/stschiff/msmc

wellcome trust sanger institute

EMBO Fellow

Transition probability

$$q(i,\, j,\, t \mid k,\, l,\, s) = \delta(t-s)\,\delta_{i,k}\,\delta_{j,l}\,q_1(t) + q_2(t \mid s)$$

Probability to remain in state (i,j,t)

Probability to change time (and pair) of first coalescence

$$q_1(t) = e^{-M r t} + \left(1 - e^{-M r t}\right)\frac{1}{t}\,\frac{1}{M}\int_0^t \left(1 + (M-3)\exp\left(-M\int_u^t \lambda(v)\,dv\right)\right)du$$

$$q_2(t \mid s) = \left(1 - e^{-M r s}\right)\frac{1}{s}\,\frac{1}{M}\,2\,\lambda(t)\begin{cases} \int_0^t \exp\left(-M\int_u^t \lambda(v)\,dv\right)du & \text{if } t < s \\ \exp\left(-\binom{M}{2}\int_s^t \lambda(v)\,dv\right)\int_0^s \exp\left(-M\int_u^s \lambda(v)\,dv\right)du & \text{if } t > s. \end{cases}$$

depends on:

- coalescence rates λ(t)
- recombination rate r
- Number of sequences M