Statistical Inferences in Population Genetics

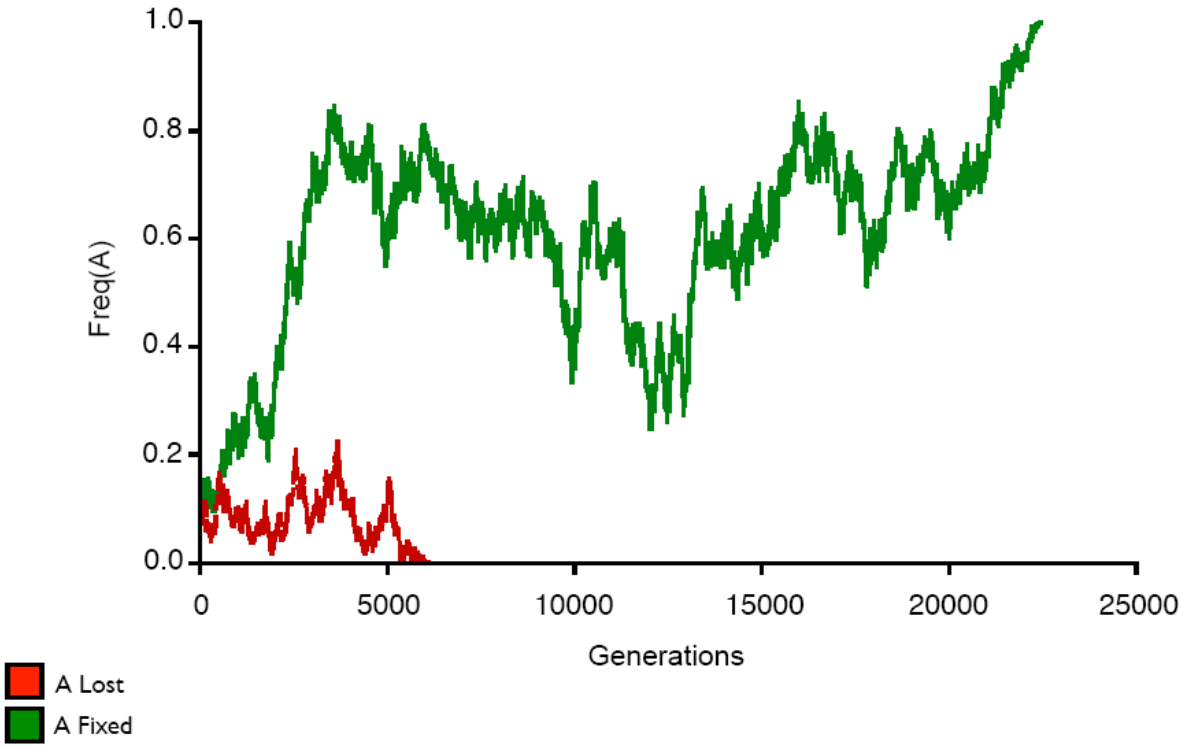
Rasmus Nielsen 6/7/2022

Inferences of

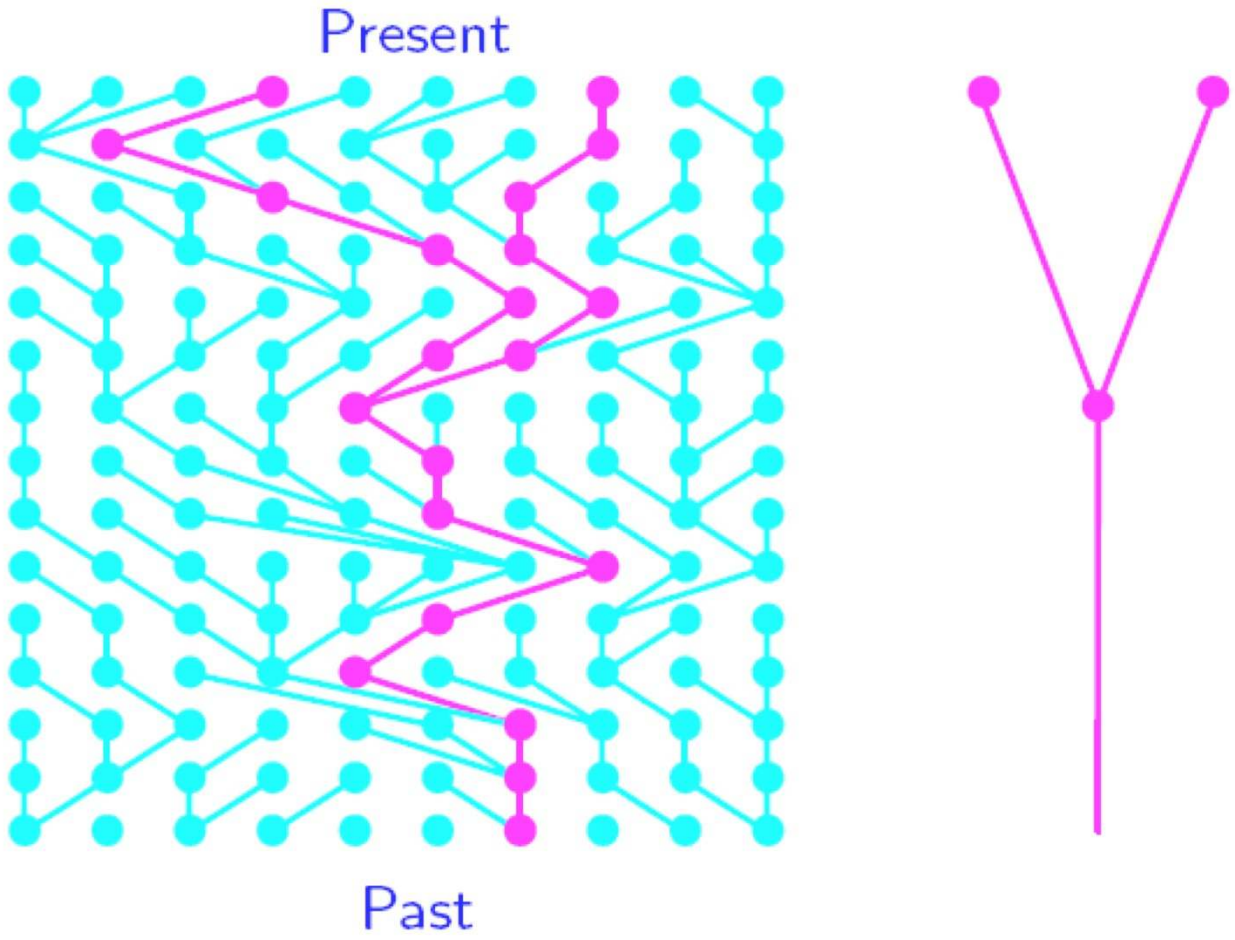
- Demography (migration rates, divergence times, changes in population size, admixture proportions)
- Selection (identifying loci under positive or negative selection, estimating selection coefficients, time of selection)

Genetic Drift

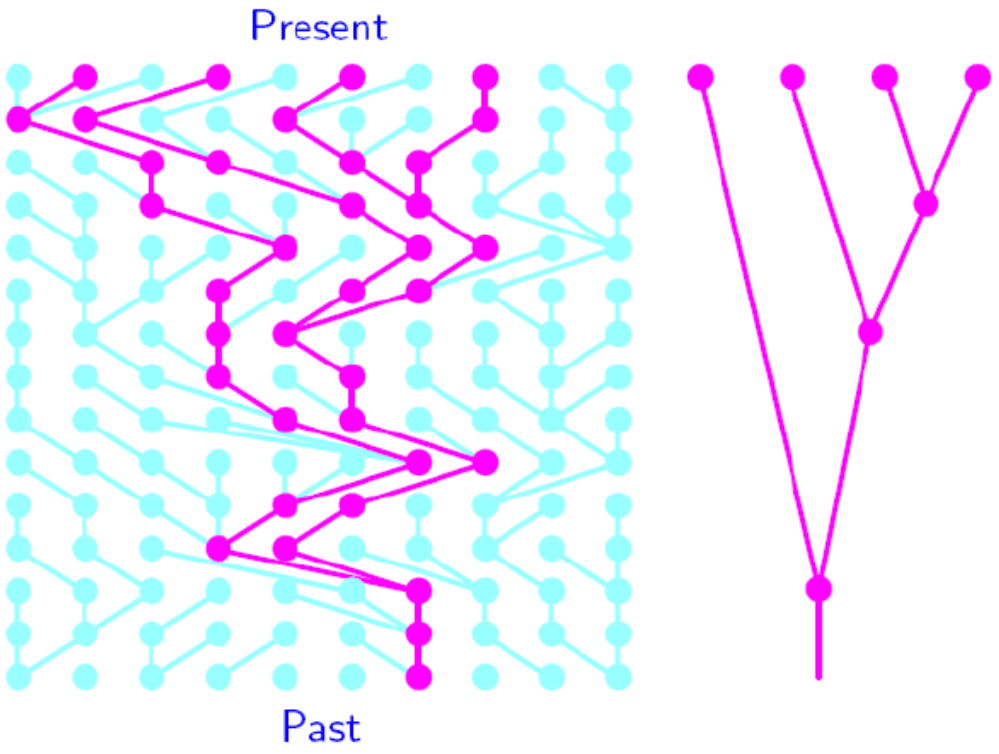
Change of allele frequencies through time.



Coalescence theory



The n-coalescent



The *n*-coalescent

For a sample of size n, the probability that all gene copies have distinct parental gene copies in the previous generation is

$$\left(1-\frac{1}{2N}\right)\left(1-\frac{2}{2N}\right)\times\ldots\times\left(1-\frac{n-1}{2N}\right)=1-\frac{\binom{n}{2}}{2N}+O\left(N^{-2}\right)$$

$P(n$ gene copies do not find a common ancestor in r generations) is then approximately

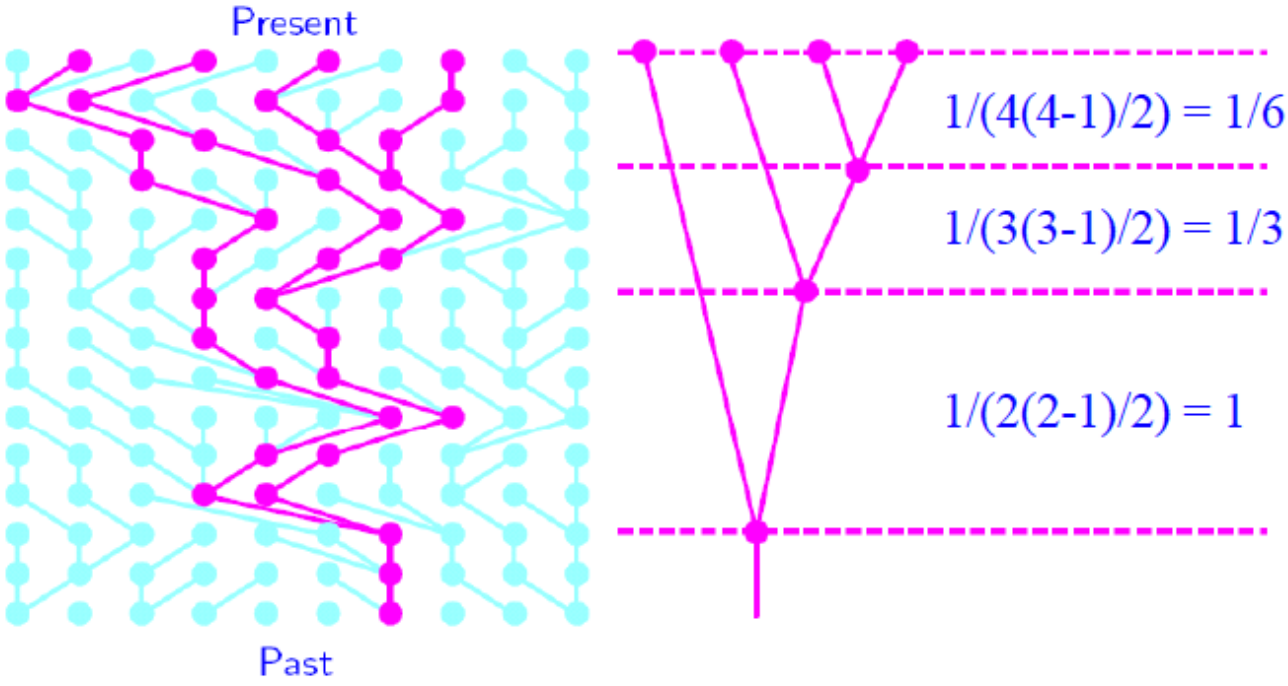
$$\left[1-\binom{n}{2}(2N)^{-1}+O\left(N^{-2}\right)\right]^{r}$$

Set $r = 2Nt$ and let $N \to \infty$

$$\left[1-\binom{n}{2}(2N)^{-1}+O\left(N^{-2}\right)\right]^{r} \approx e^{-\binom{n}{2}t}$$

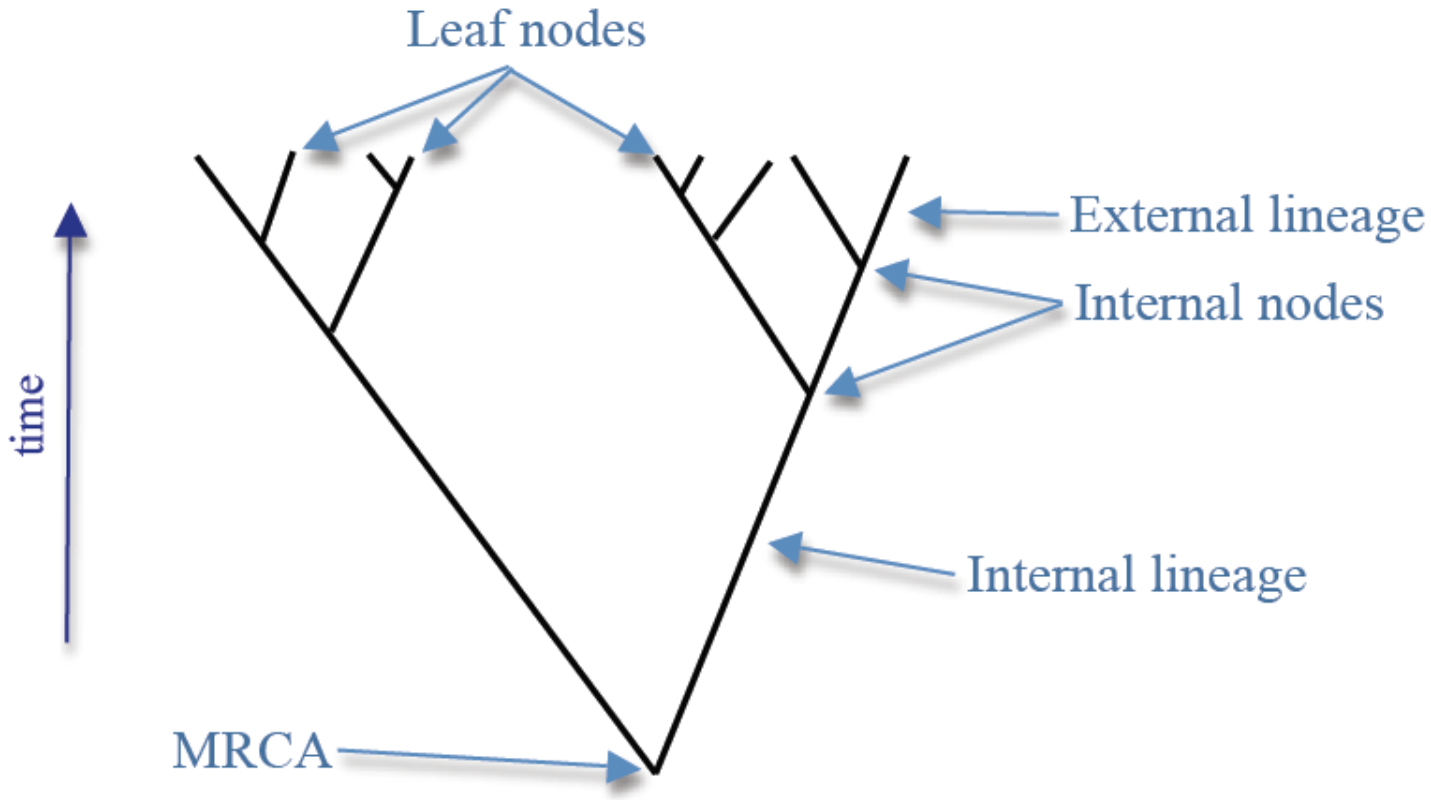
In general, the time in which there a j ancestors in the sample is exponentially distributed with mean

$$\binom{j}{2}^{-1}$$



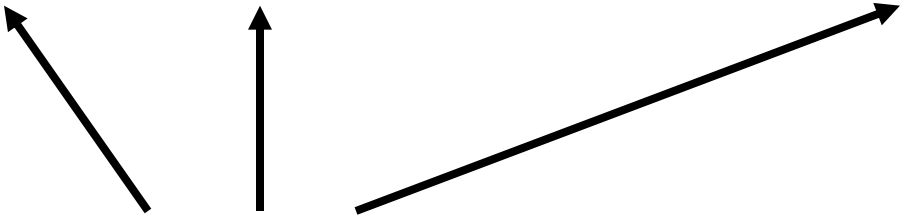
(1) Time is scaled in terms of $2N$. The smaller the population – the faster the coalescence process.

(2) The coalescence process is faster when there are many lineages. The further in the past you get in the tree – the slower the coalescence process. Branches deep in the tree will be longer


Leaf nodes
External lineage
Internal nodes
time
Internal lineage
MRCA

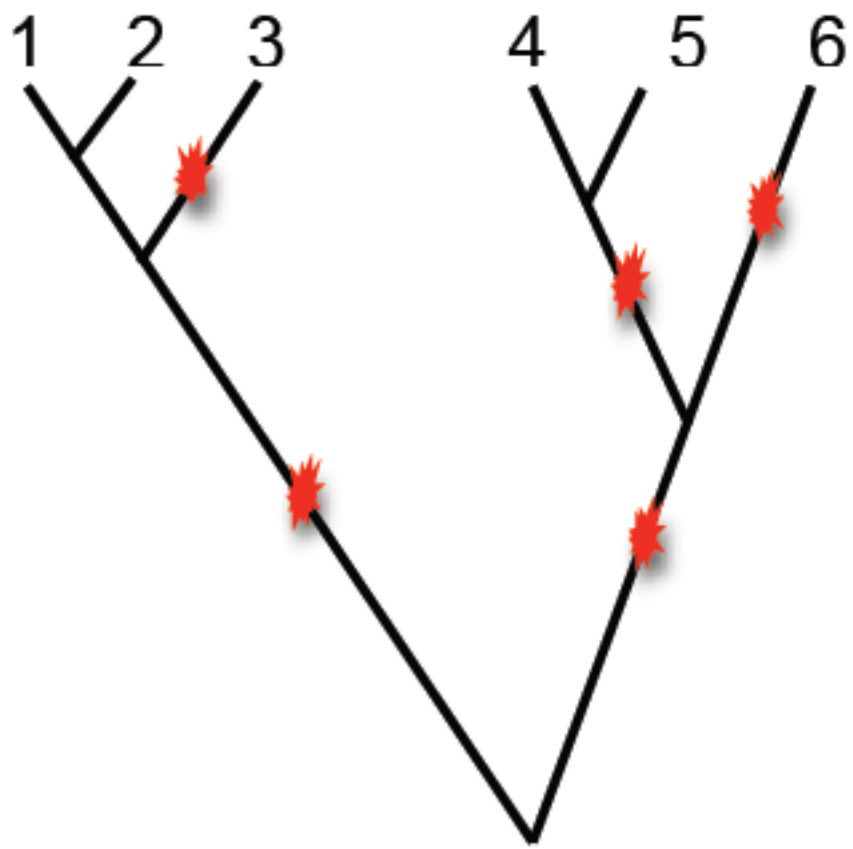
Sequence comparisons

```
Human1 CGC TAT TCC CCG ATT CAG AAT GCC CAG
Human2 CGC TAT TCC CAG ATT CAG AAT GCA CAG
Human3 CGC TAC TCC CCG ATT CAG AAT GCA CAG
Human4 CGC TAT TCC CAG ATT CAG AAT GCC CAG
Human5 CGC TAT TCC CCG ATT CAG AAT GCC CAG
```



Segregating sites with
Single Nucleotide Polymorphisms (SNPs)

Infinite Sites Model



Sequences	
1	01000
2	01000
3	11000
4	00110
5	00110
6	00011

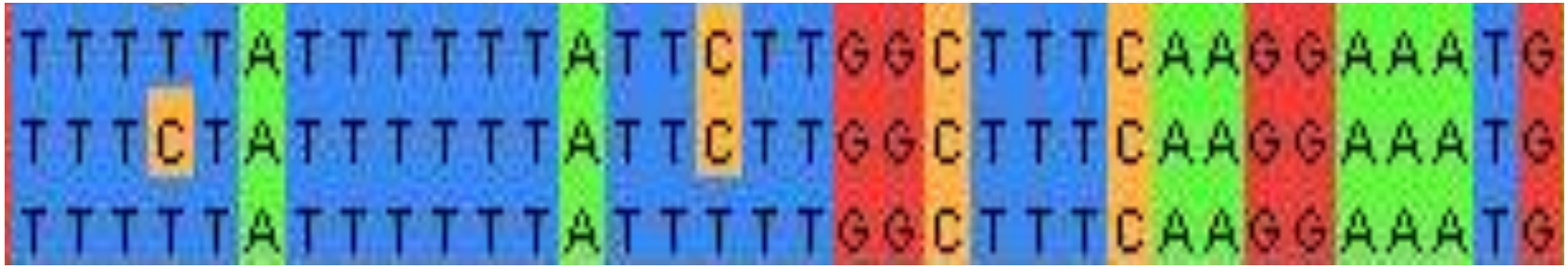
Assume there are $2N$ individuals (we model a haploid population with N individuals as a diploid population with $2N$ individuals).

Probability of having same parent in one generation = $1/2N$

Mean number of generations until they have the same parent (time to the most recent common ancestor, tMRCA) = $2N$

Mean number of mutations separating two individuals = $2N \times 2 \times \mu = 4N\mu = \theta$

Tajima's estimator of *4Nu* = θ

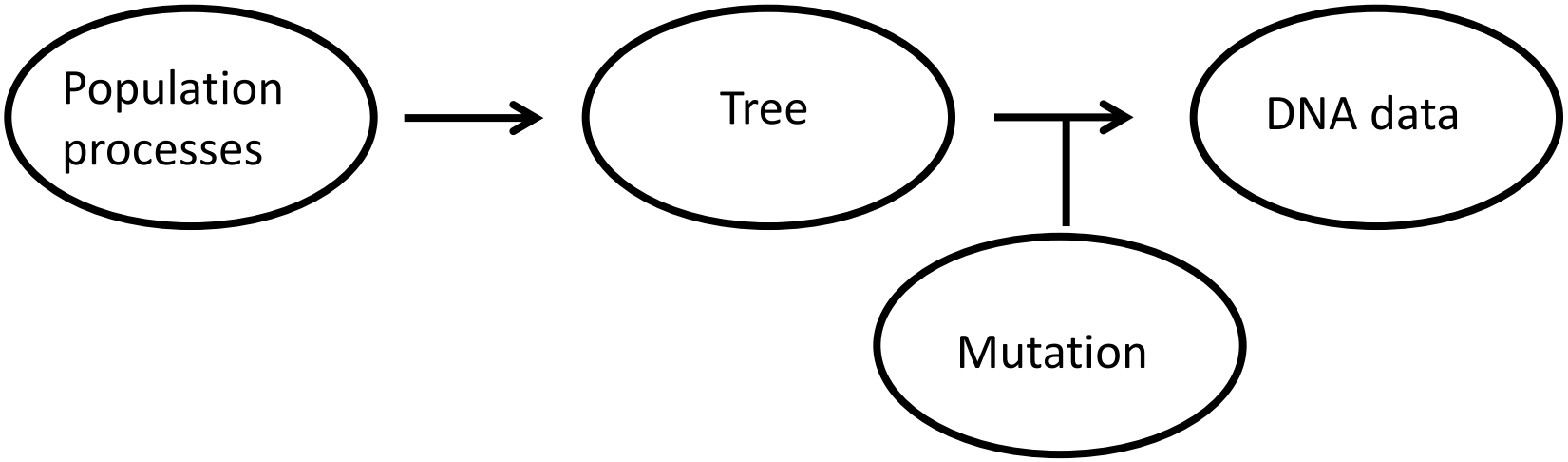


$$\pi = (1+1+2)/3 = 4/3$$

π (=$\hat{\theta}_T$) is an *estimator* of *4Nu* = θ

(In this case the total value for the entire region)

Coalescence Trees



Felsenstein's Equation

$$p(X \mid \Theta) = \int_{G \in \psi} p(X \mid G)\, p(G \mid \Theta) dG$$

- Can only be evaluated directly in very simple cases.
- Simulation based approaches are typically used for real data.

MCMC

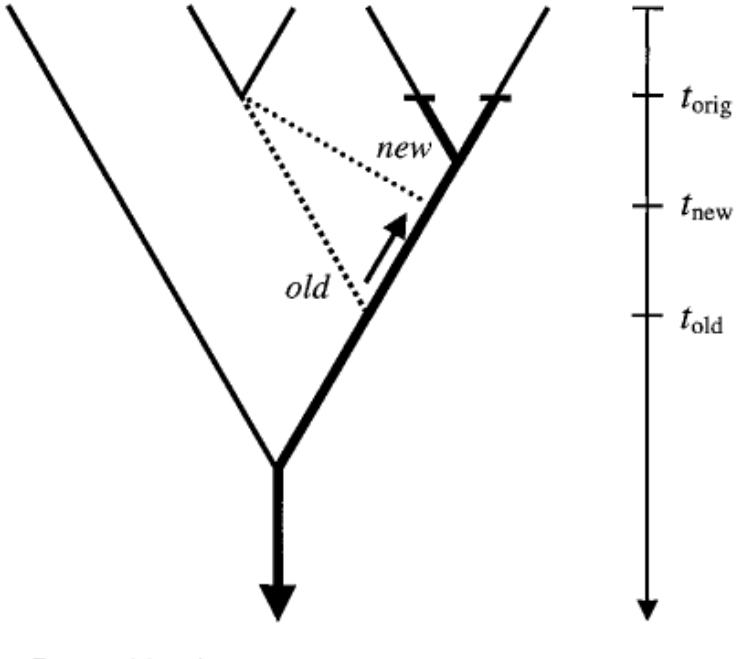
Set up a Markov chain on state space on all supported values of Θ and G and with stationary distribution $p(\Theta, G \mid X)$. Now since

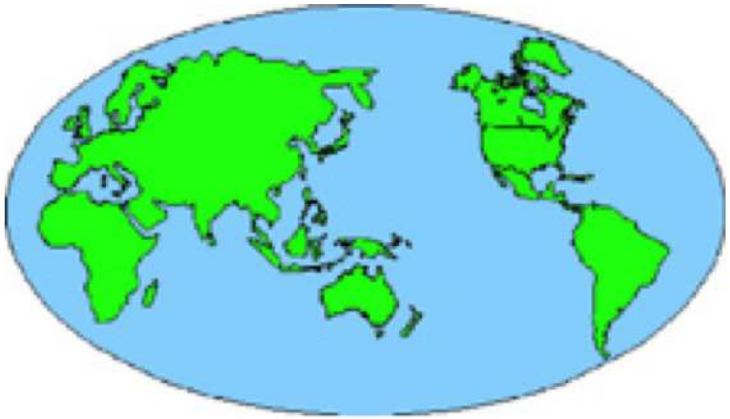
$$p(\Theta, G \mid X) \propto \Pr(X \mid G)\, p(G \mid \Theta)\, p(\Theta)$$

this can easily be done using Metropolis-Hastings sampling, i.e. updates to Θ and G are proposed from a proposal distribution $q(\Theta, G \rightarrow \Theta', G')$ and accepted with probability

$$\frac{P(X \mid G')P(G' \mid \Theta')q(\Theta', G' \rightarrow \Theta, G)}{P(X \mid G)P(G \mid \Theta)q(\Theta, G \rightarrow \Theta', G')}$$

MCMC algorithms

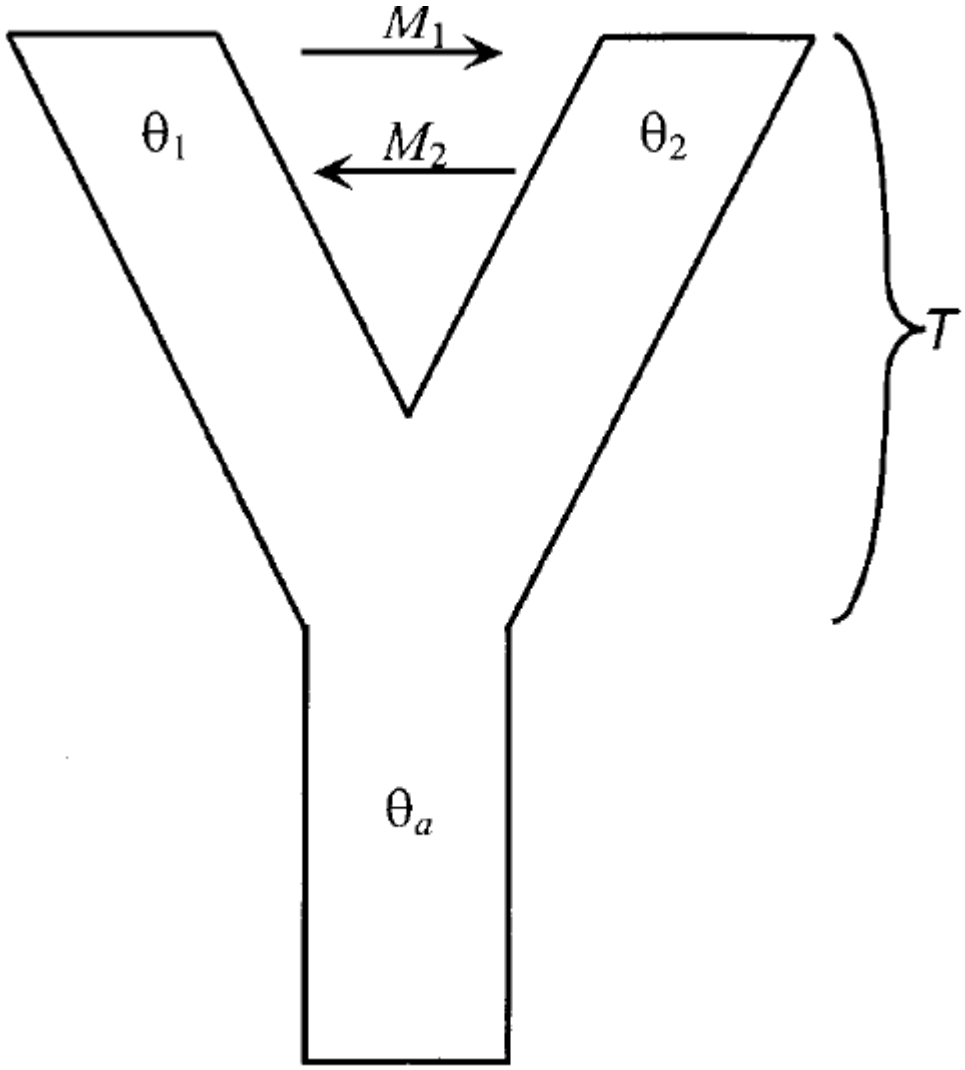




Demographic history of eastern pacific stickleback vs. western pacific stickleback.

Data: DNA sequences

```
Eastern 1 CGAAAAGTATCCATCTCGCAGTGCTGAGCTAGACA
Eastern 2 CGAAAAGTATCCATCTCGCAGTGCTGAGCTAGACA
Eastern 3 CGAAAAGTATCCATCTCGCAGTGCTGAGCTAGACA
Eastern 4 CGAAAGGTATCCATCTCGCAGTGCTGAGTTAGACA
Eastern 5 CAAAAAGTATCCATCTCGCAGTGCTGAGTTAGACA
Eastern 6 CAAAAAGTATCCATCTCGCAGTGCTGAGTTAGACA
Eastern 7 CGAAAAGTATCCATCTCGCAGTGCTGAACTAGACA
Eastern 8 CGAAAAGTATCCATCTCGCAGTGCTAAGCTAGACA
Eastern 9 TGAAAAGTATCCATCTCGCAGTGCTAAGCTAGACA
Western 1 CGAAAAGTATCCATCTCGCAGTGCTGAGCTAGACA
Western 2 CGCGAAGCACTTGCCCCATAGCGCTAAGCCGCGTT
Western 3 CGCGTAGCACTTGCCCCATAGCGCTAAGCCGCGTT
Western 4 CGCAAAGCGCTTGCCCCATAACGCTAAGCCGCGTT
Western 5 CGCAAAGCGCTTGCCCCATAACGCTAAGCCGCGTT
Western 6 CGCAAAGCACTTGCCCCATAACGCTAAGCCGCGTT
Western 7 CGCAAAGCACTTGCCCCATAACGCTAAGCCGCGTT
```


M_1
M_2
θ_1
θ_2
T
θ_a

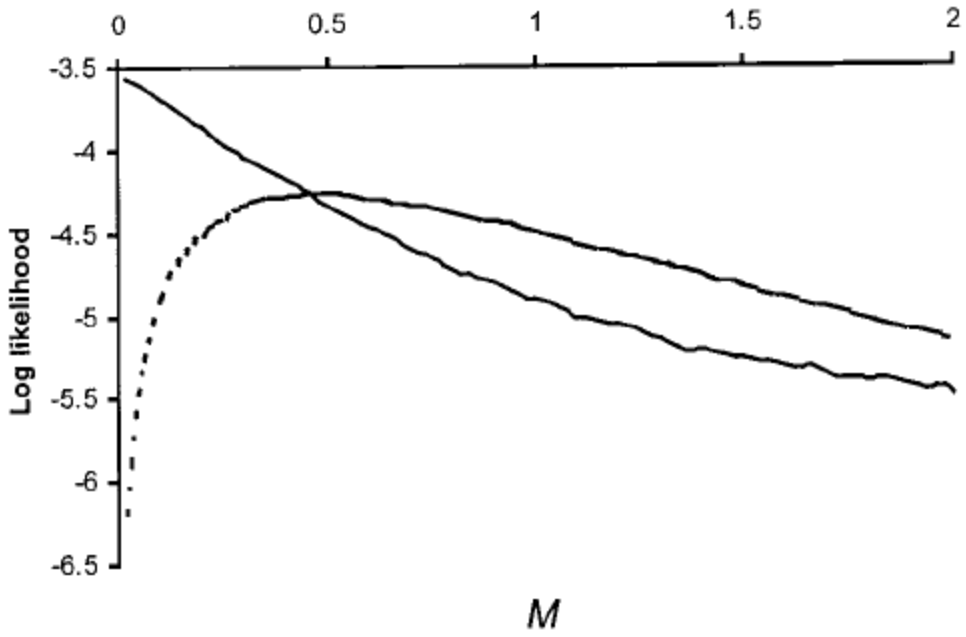
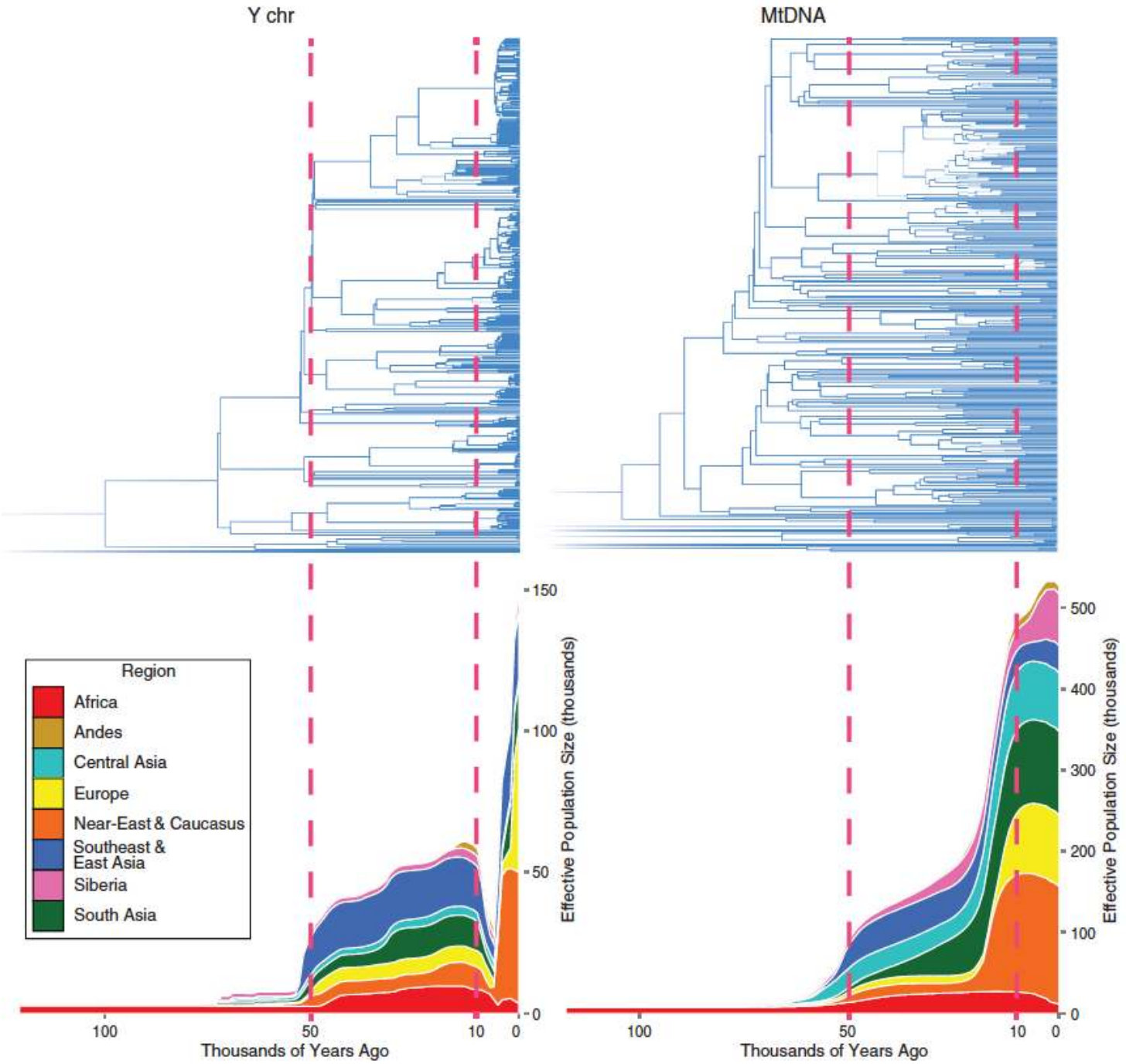
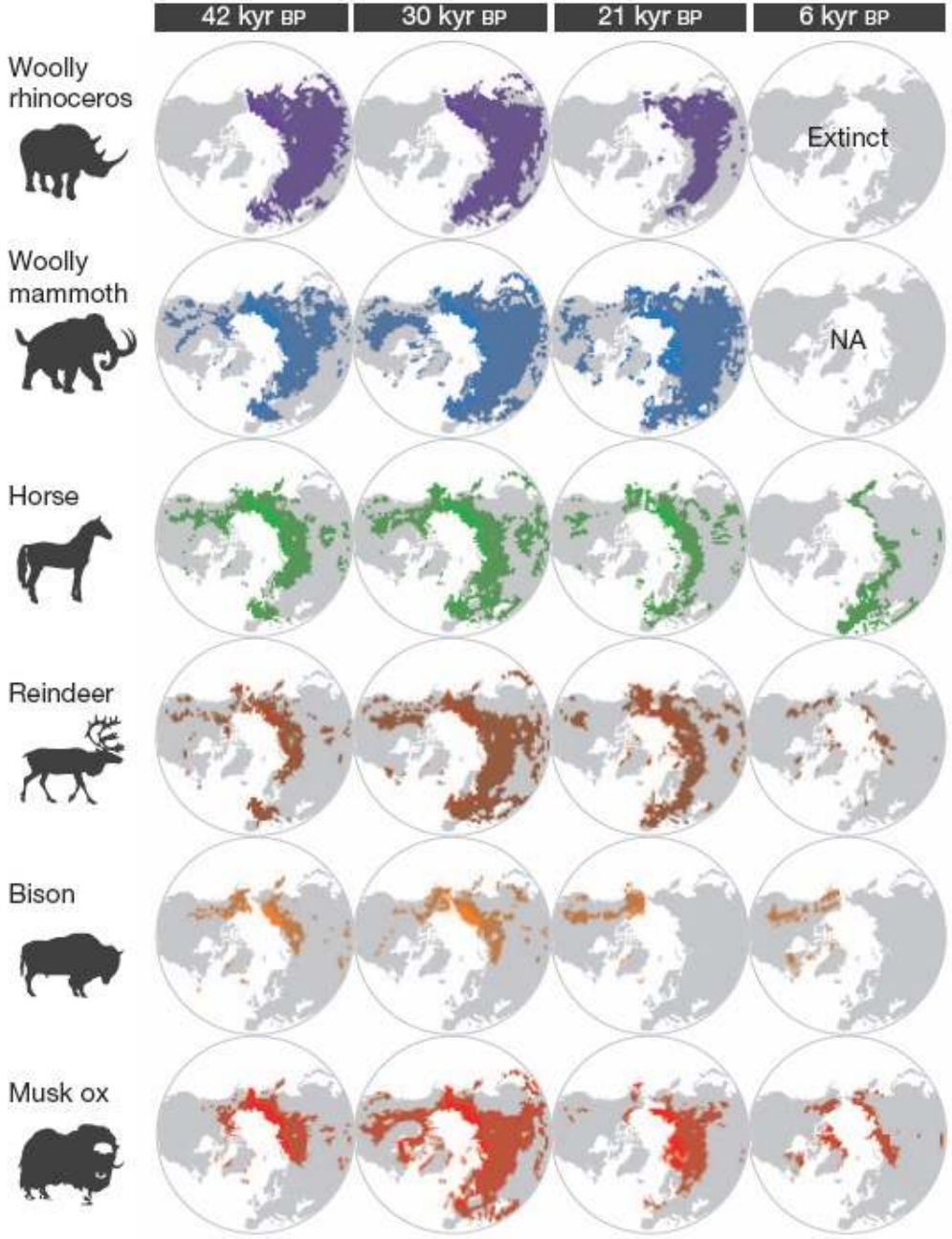
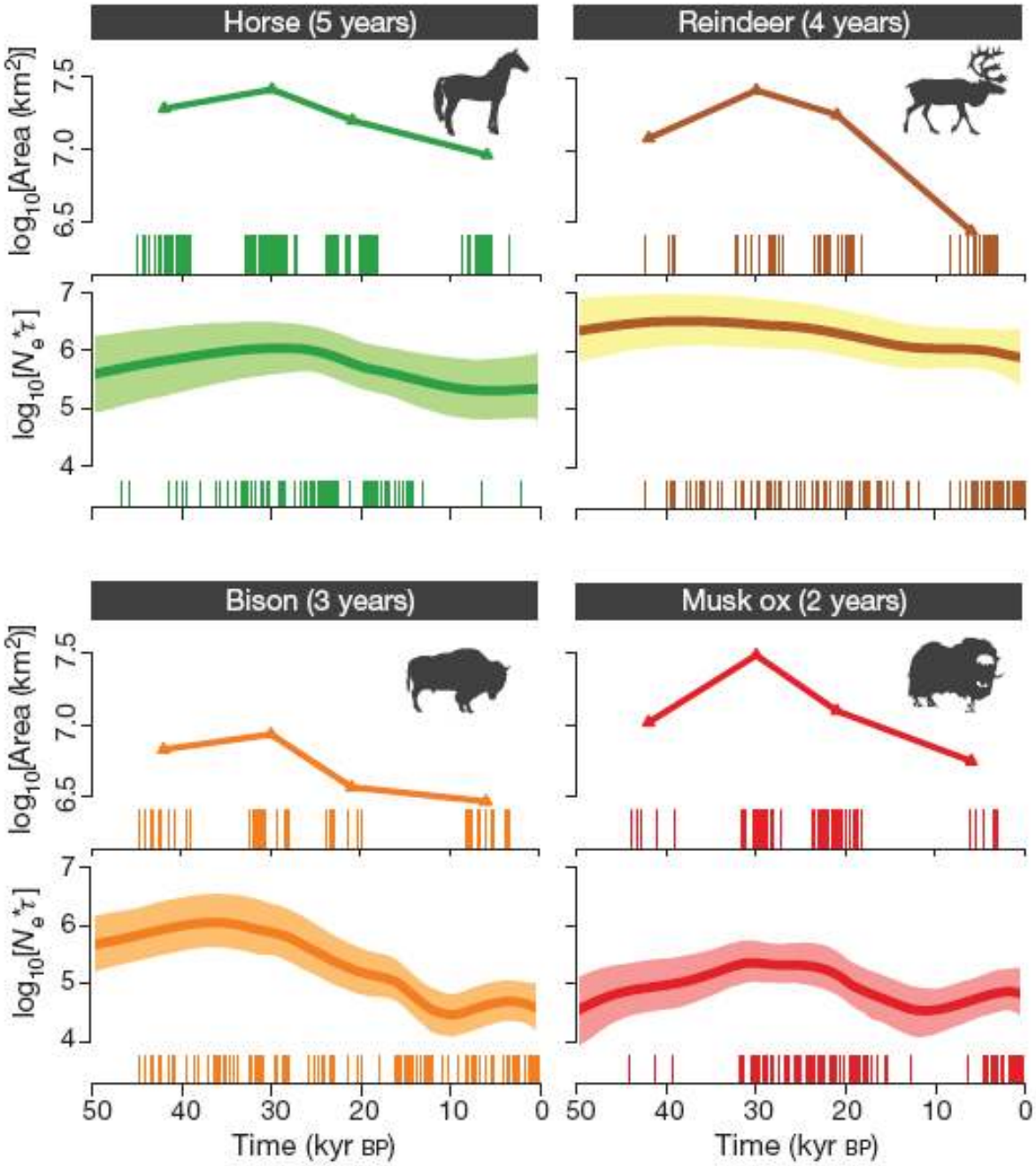


FIGURE 8.—The integrated likelihood surfaces for M_1 (dots) and M_2 (solid lines) estimated from the data by ORTI *et al.* 1994.


Y chr
MtDNA
Region
Africa
Andes
Central Asia
Europe
Near-East & Caucasus
Southeast & East Asia
Siberia
South Asia
Effective Population Size (thousands)
150
100
50
0
500
400
300
200
100
0
100
50
10
0
Thousands of Years Ago


42 kyr BP
30 kyr BP
21 kyr BP
6 kyr BP
Woolly rhinoceros
Extinct
Woolly mammoth
NA
Horse
Reindeer
Bison
Musk ox


Horse (5 years)
Reindeer (4 years)
Bison (3 years)
Musk ox (2 years)
log10[Area (km2)]
log10[Ne*τ]
7.5
7.0
6.5
7
6
5
4
50
40
30
20
10
0
Time (kyr BP)

Some MCMC approaches

- BEAST (Rambaut)
- Various version of IM (Hey)
- MIGRATE (Beerli)

Approximate Bayesian Computation

1. Draw $\theta_i \sim \pi(\theta)$.
2. Simulate $x_i \sim p(x \mid \theta_i)$.
3. Reject θ_i if $\rho(S(x_i), S(y)) > \epsilon$.

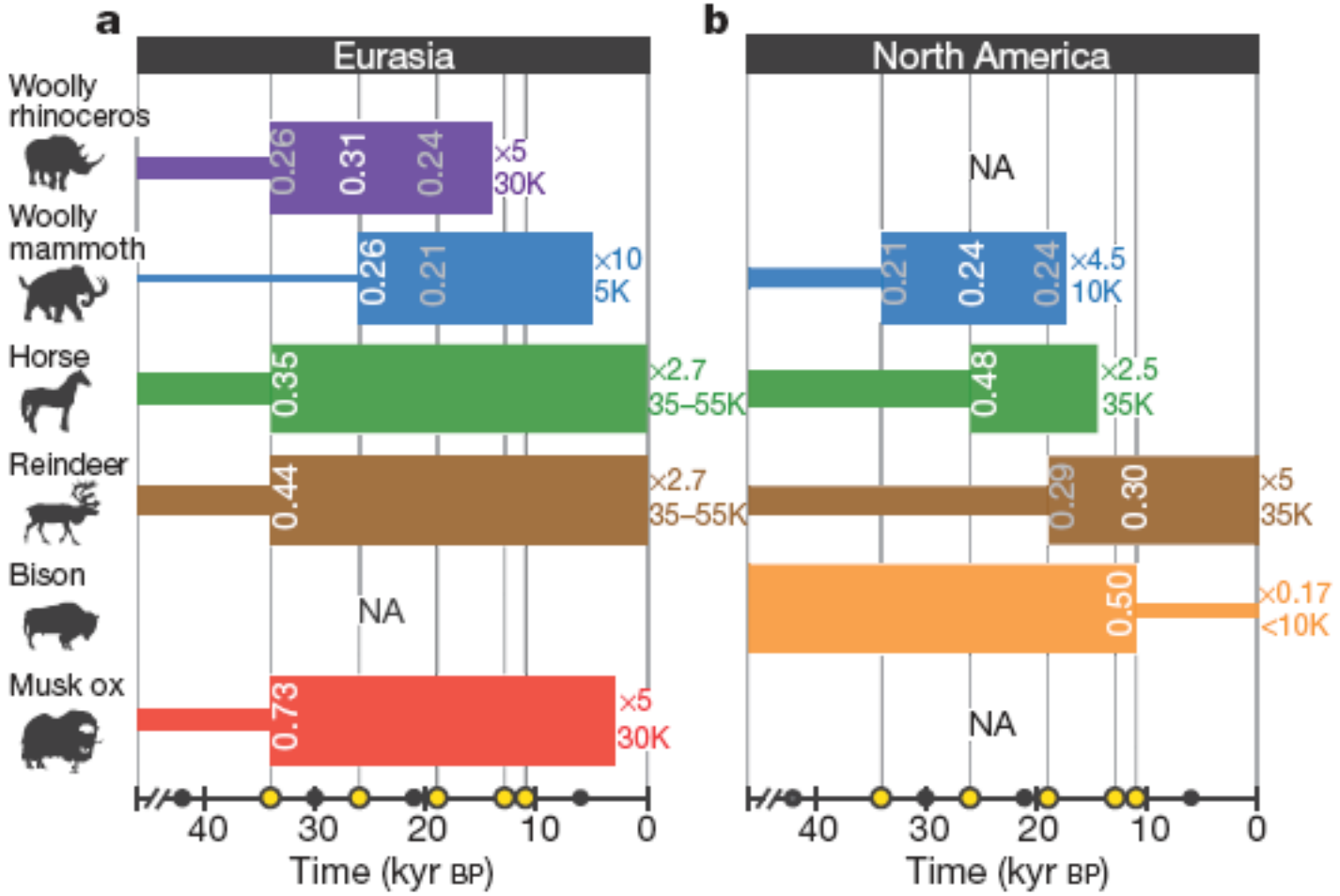
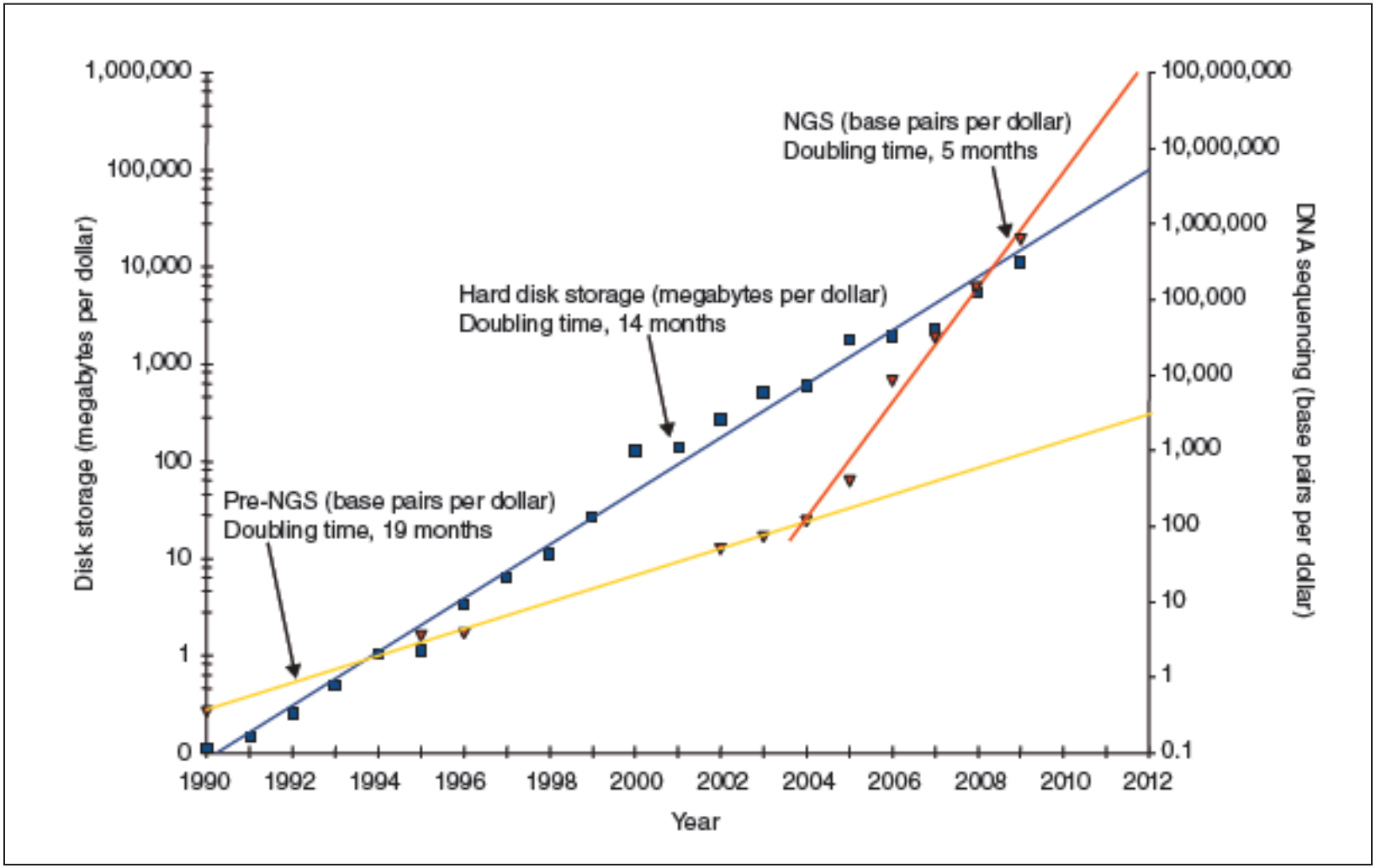
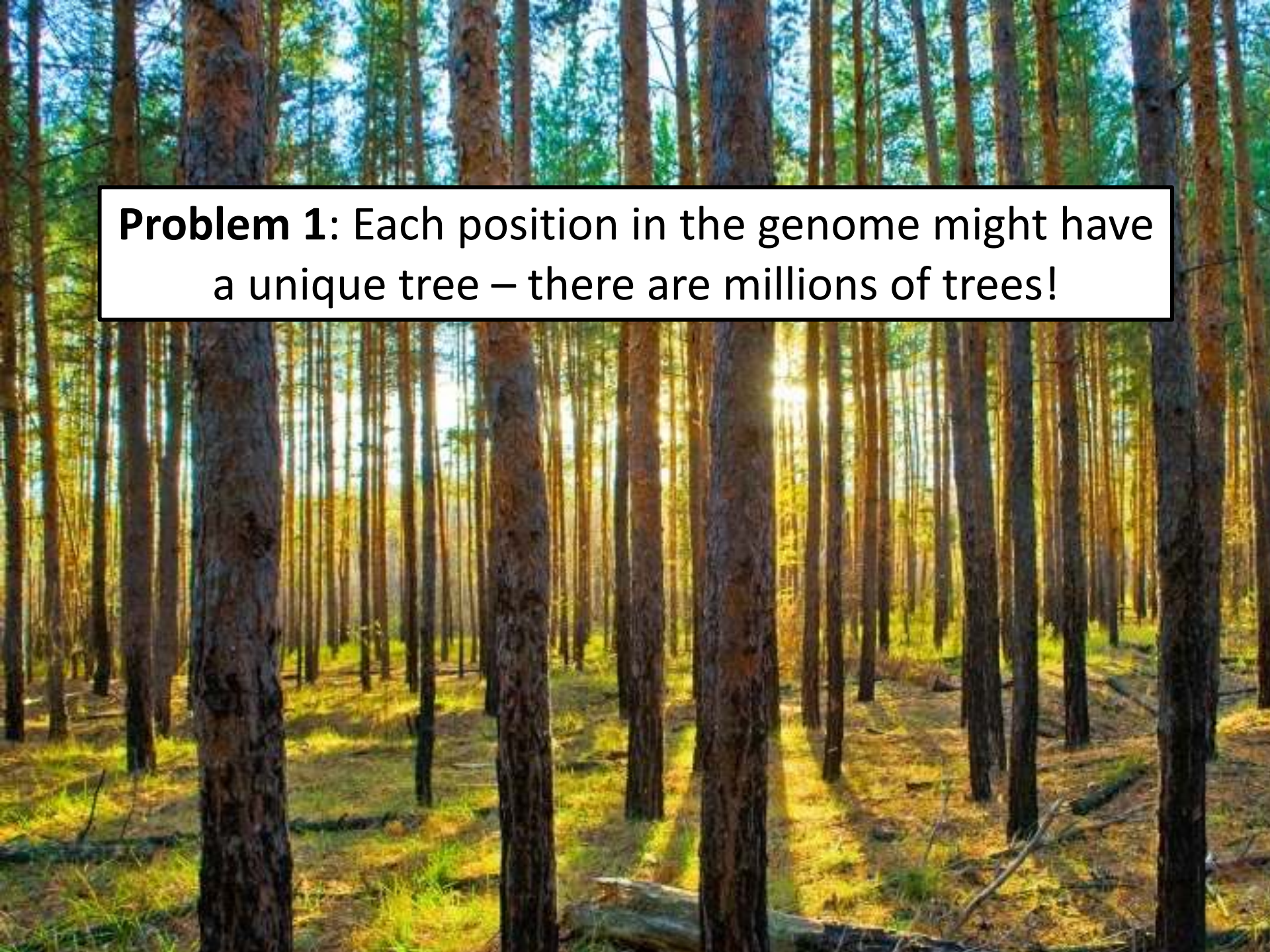
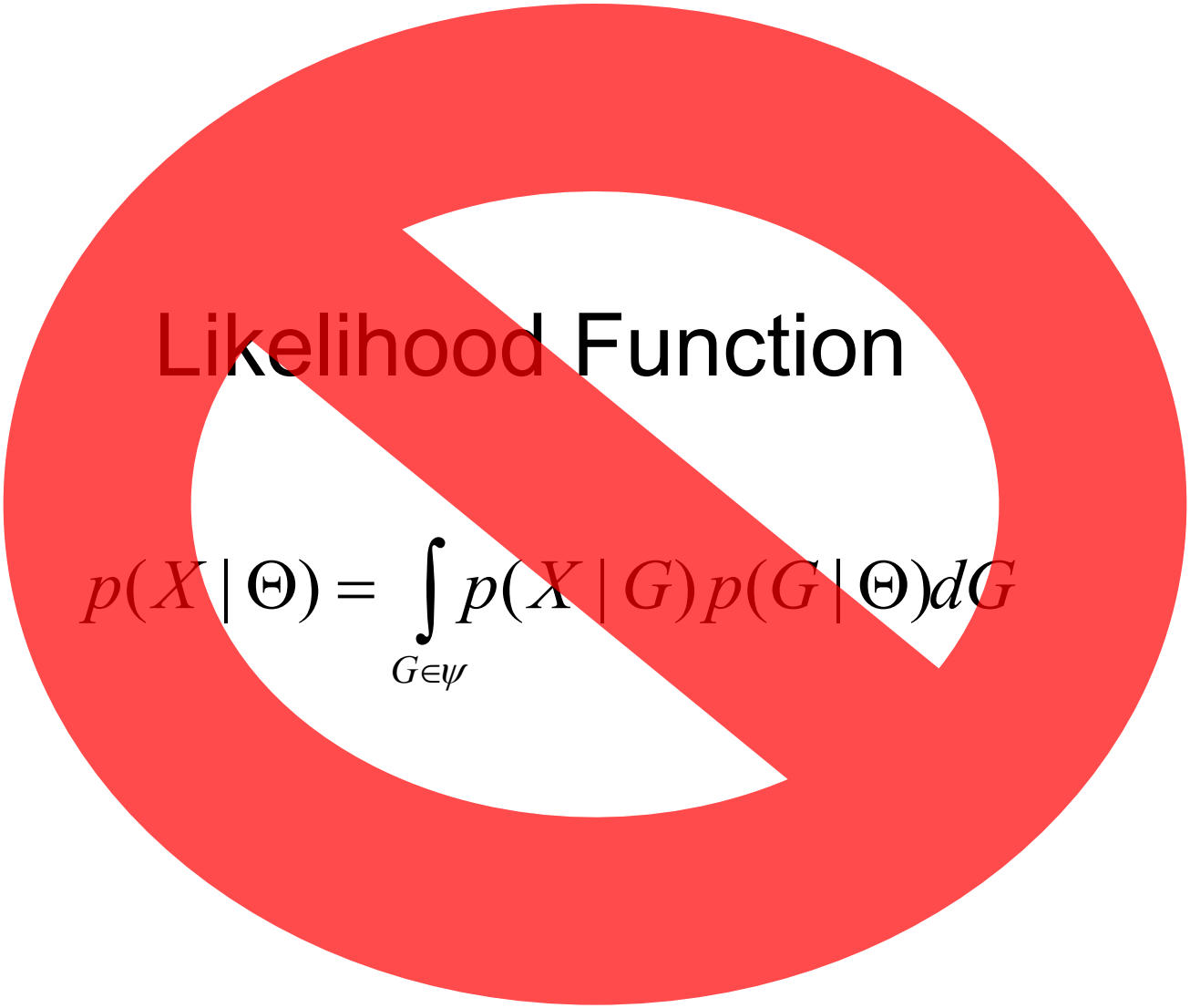


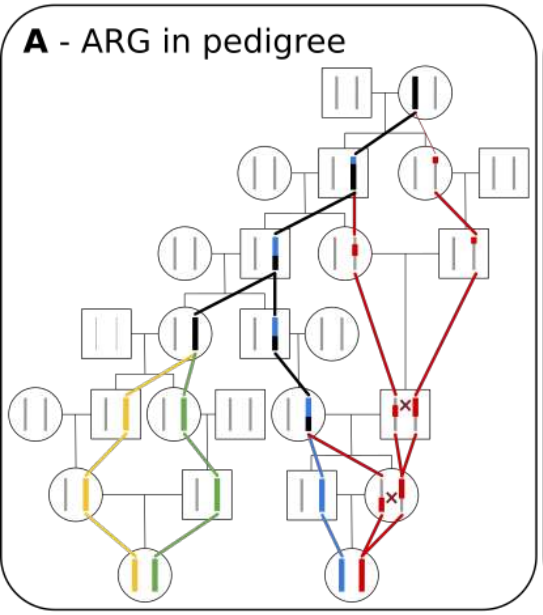
Figure 3 | Best-supported demographic models inferred by approximate Bayesian computation model-selection. a, Eurasia; b, North America. Grey

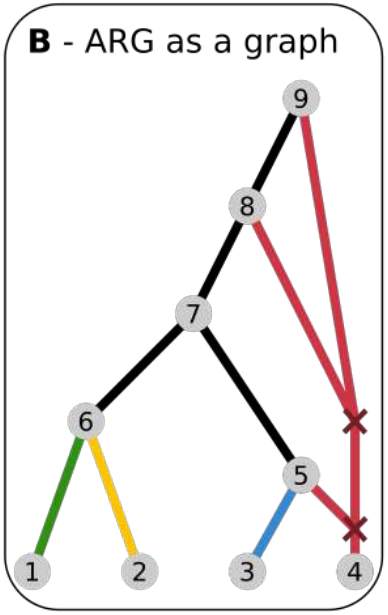
Then this happened…

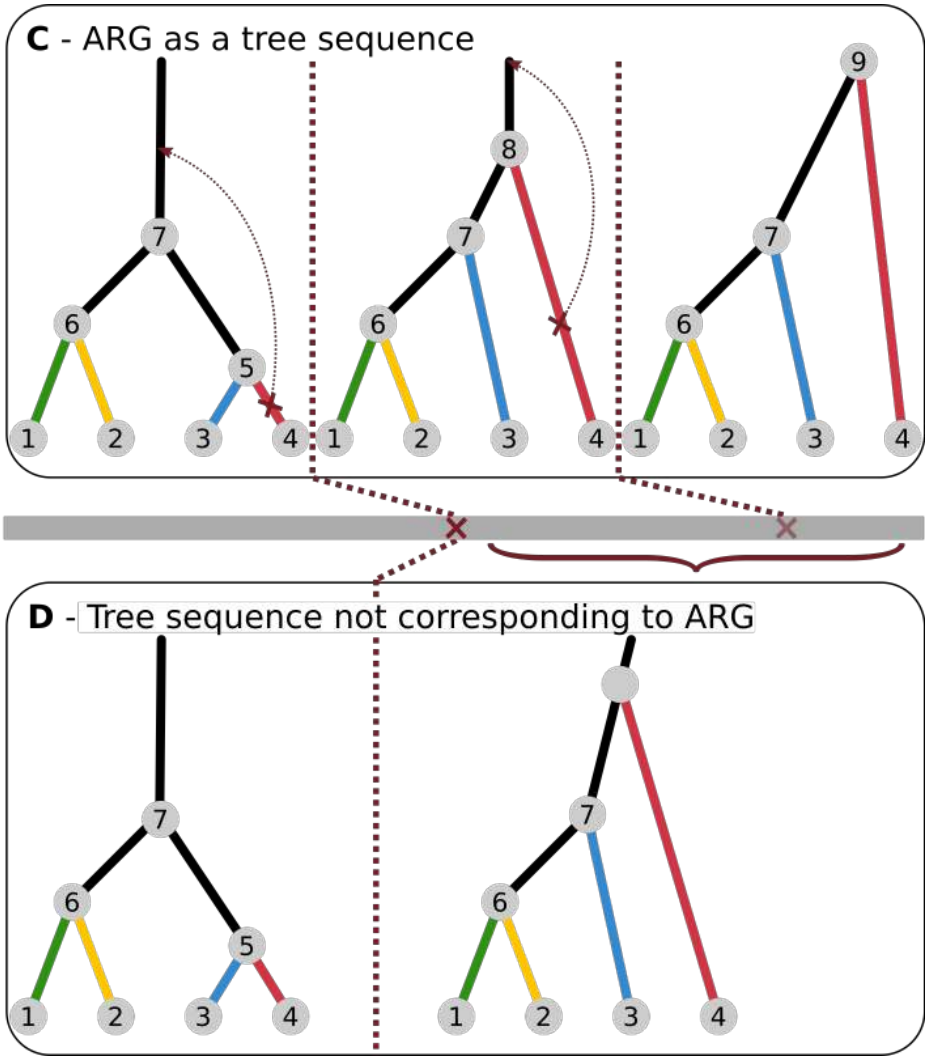



Problem 1: Each position in the genome might have a unique tree – there are millions of trees!


Likelihood Function
$p(X|\Theta) = \int_{G \in \psi} p(X|G)\,p(G|\Theta)\,dG$


A - ARG in pedigree

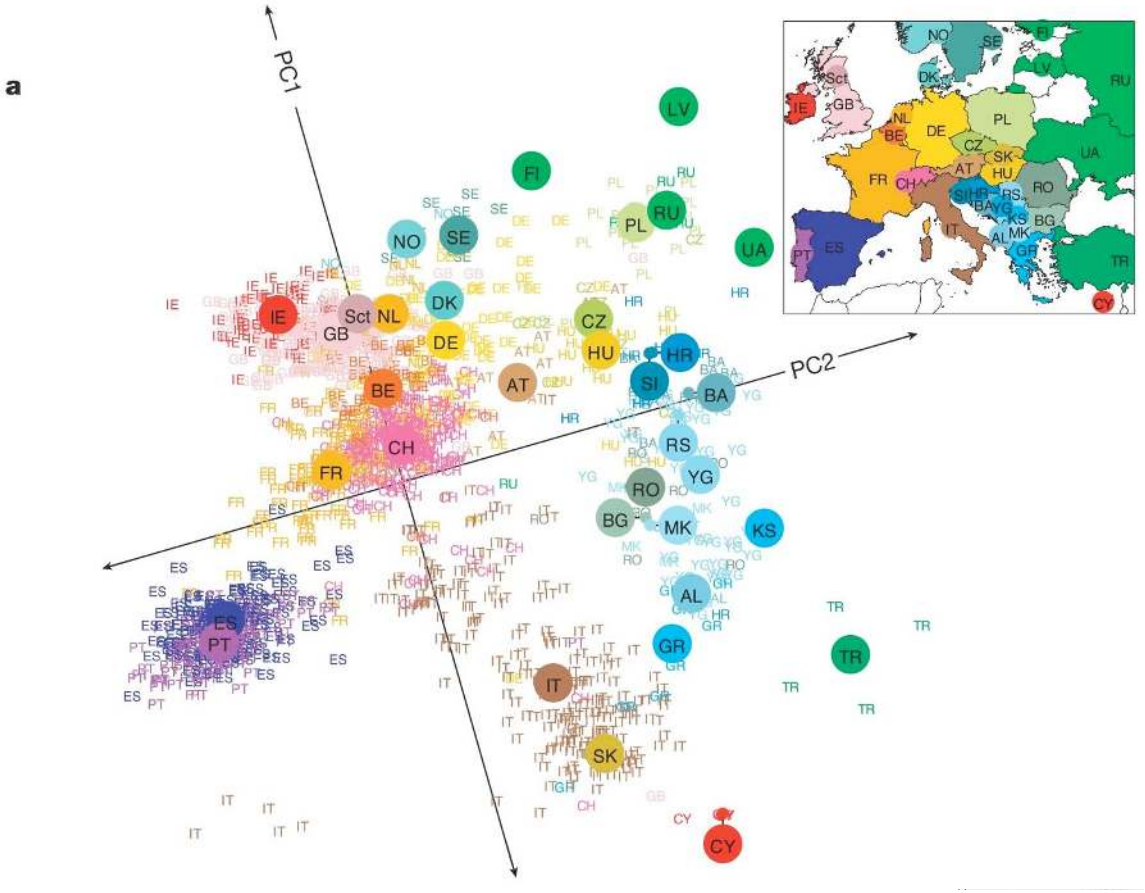

B - ARG as a graph
9
8
7
6
5
1
2
3
4


C - ARG as a tree sequence
D - Tree sequence not corresponding to ARG

Challenge

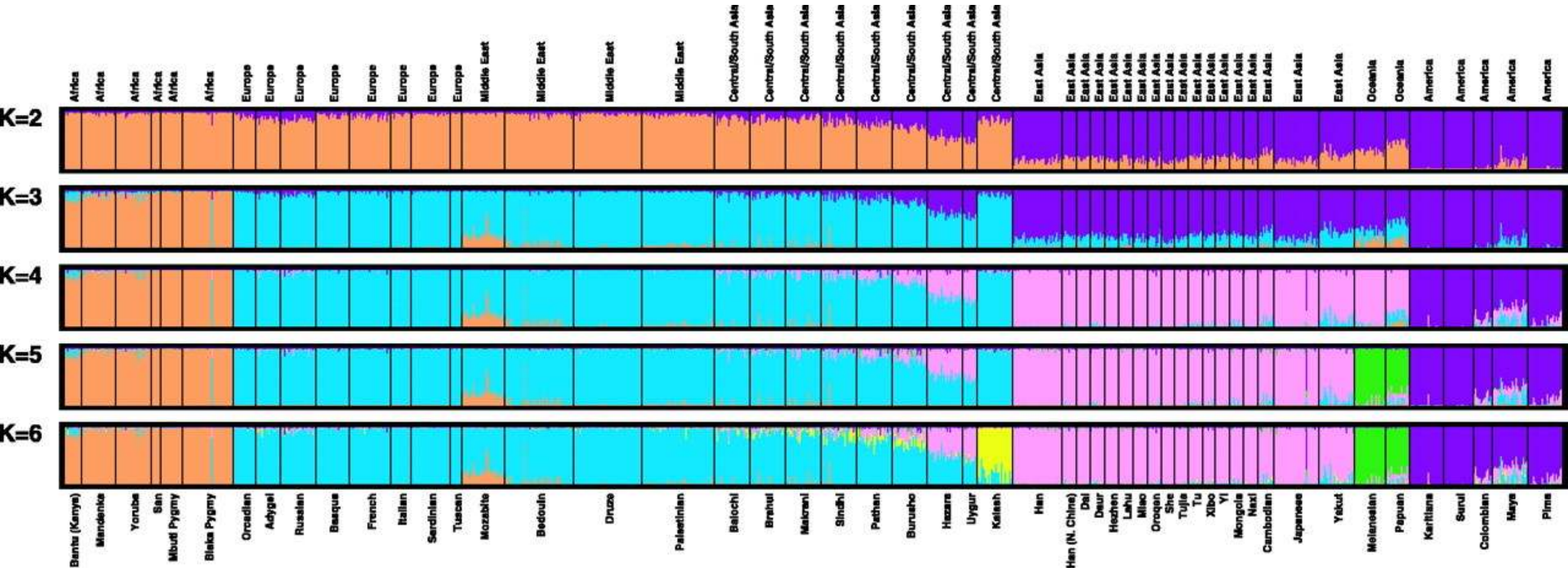
- Full likelihood inference considered intractable in samples of $n>2$ (but see PSMC).
- Instead, rely on summary statistics/various features extracted from the data.
- Examples: π, S, F_{ST}, Patterson's D
- More complex examples: The Site-Frequency-Spectrum, principle components, features extracted by a convolutional neural network.

Principle Component Analysis (PCA)

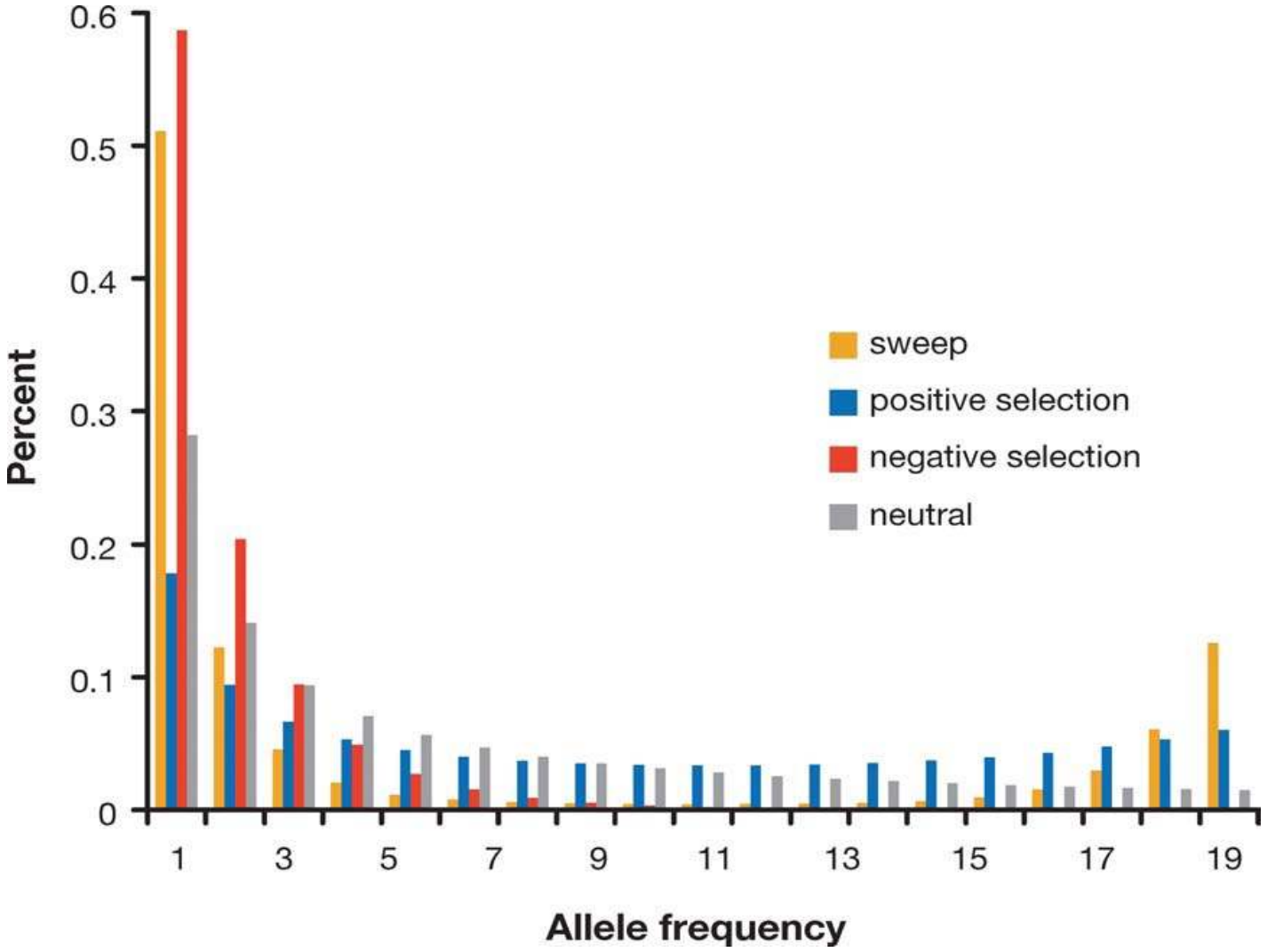


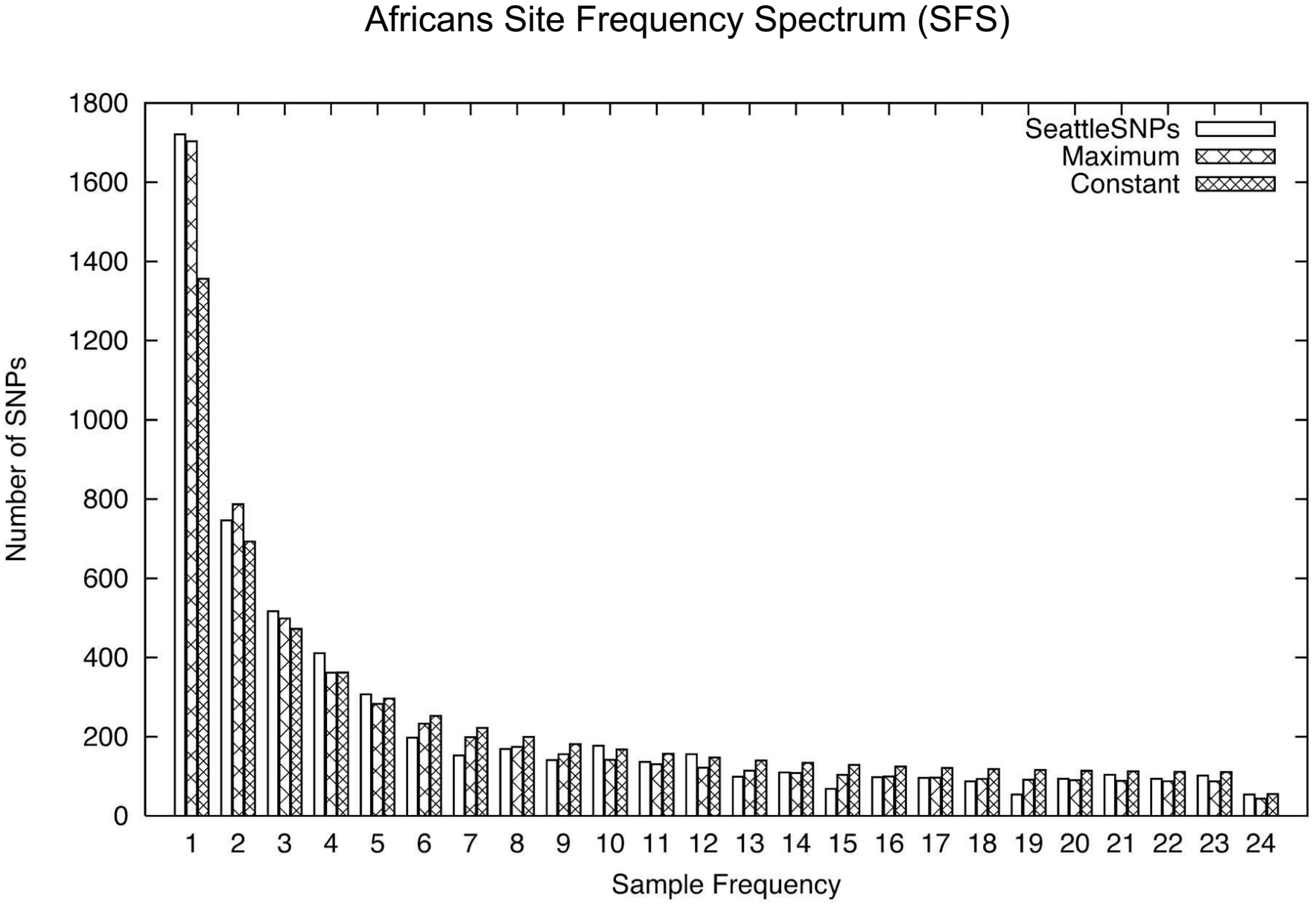
Structure analysis of 1056 individuals from 52 populations for 377 microsatelite loci.

Rosenberg et al. (2002)



Frequency spectrum




Africans Site Frequency Spectrum (SFS)
SeattleSNPs
Maximum
Constant
Number of SNPs
0
200
400
600
800
1000
1200
1400
1600
1800
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Sample Frequency

Composite likelihood function

$$L(\Theta) \equiv \prod_{j \in \Phi} \left(p_j(\Theta) \right)^{n_j}$$

Sampling probability of a SNP with site pattern *j* (a binary vector)

,

Number of SNPs with pattern *j* in the data

SNPs within a gene are correlated. But estimator is consistent. The estimate has the same properties as a real likelihood estimator except that it converges slightly slower because of the correlation (Wiuf 2006).

Marginal likelihood for a single site

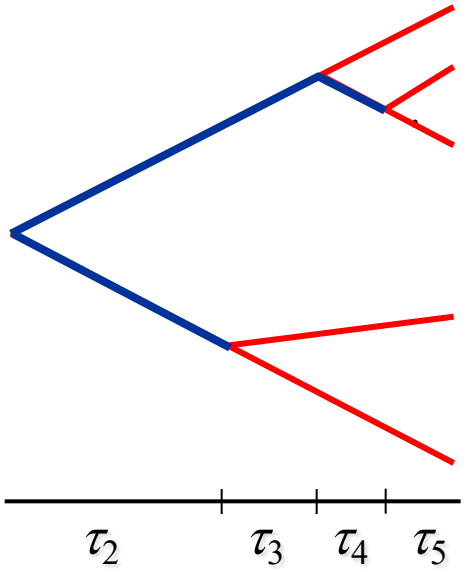
Nielsen (2000)

$$
\begin{aligned}
L(\Omega|X_i) &= \lim_{\theta \to 0} \Pr(X_i \,|\Omega, \theta, S_i > 0) \\
&= \lim_{\theta \to 0} \frac{\int (\theta/2)^{-1} \sum_{j:b_{ij} \in B_i} (1 - e^{-\theta T_{ij}/2})\, e^{-\theta (T_i - T_{ij})/2}\, dF(G|\Omega)}{\int (\theta/2)^{-1} (1 - e^{-\theta T_i/2})\, dF(G|\Omega)} \\
&= \frac{\int t_i dF(G|\Omega)}{\int T_i dF(G|\Omega)} = \frac{\mathrm{E}(t_i \mid \Omega)}{\mathrm{E}(T_i \,|\Omega)}.
\end{aligned}
$$

Estimation

(Nielsen 2000)

$$p_j(\Theta) = \frac{E[t]}{E[T]}$$



$$x = \{3, 2\}$$

$$T = 5\tau_5 + 4\tau_4 + 3\tau_3 + 2\tau_2.$$

$$t = \tau_4 + \tau_3 + 2\tau_2.$$

Sampling distribution can be calculated analytically or using simple simulation schemes.

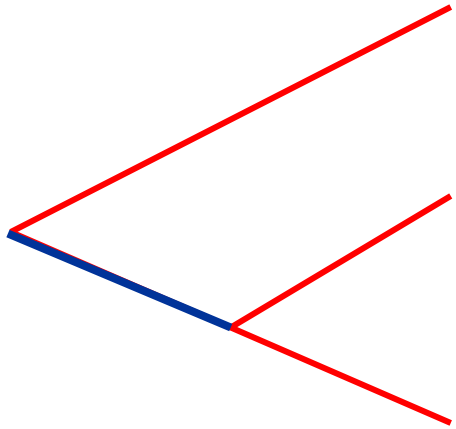
Question

Consider the polarized SFS. Under the standard (Kingman's coalescent) with n=3, what is the probability of the pattern {1, 2} meaning one ancestral and two derived alleles?

Hint: the expected time in the tree with i lineages is $2/(i(i-1))$

Question

Consider the polarized SFS. Under the standard (Kingman's coalescent) with *n*=3, what is the probability of the pattern {1, 2} meaning one ancestral and two derived alleles?



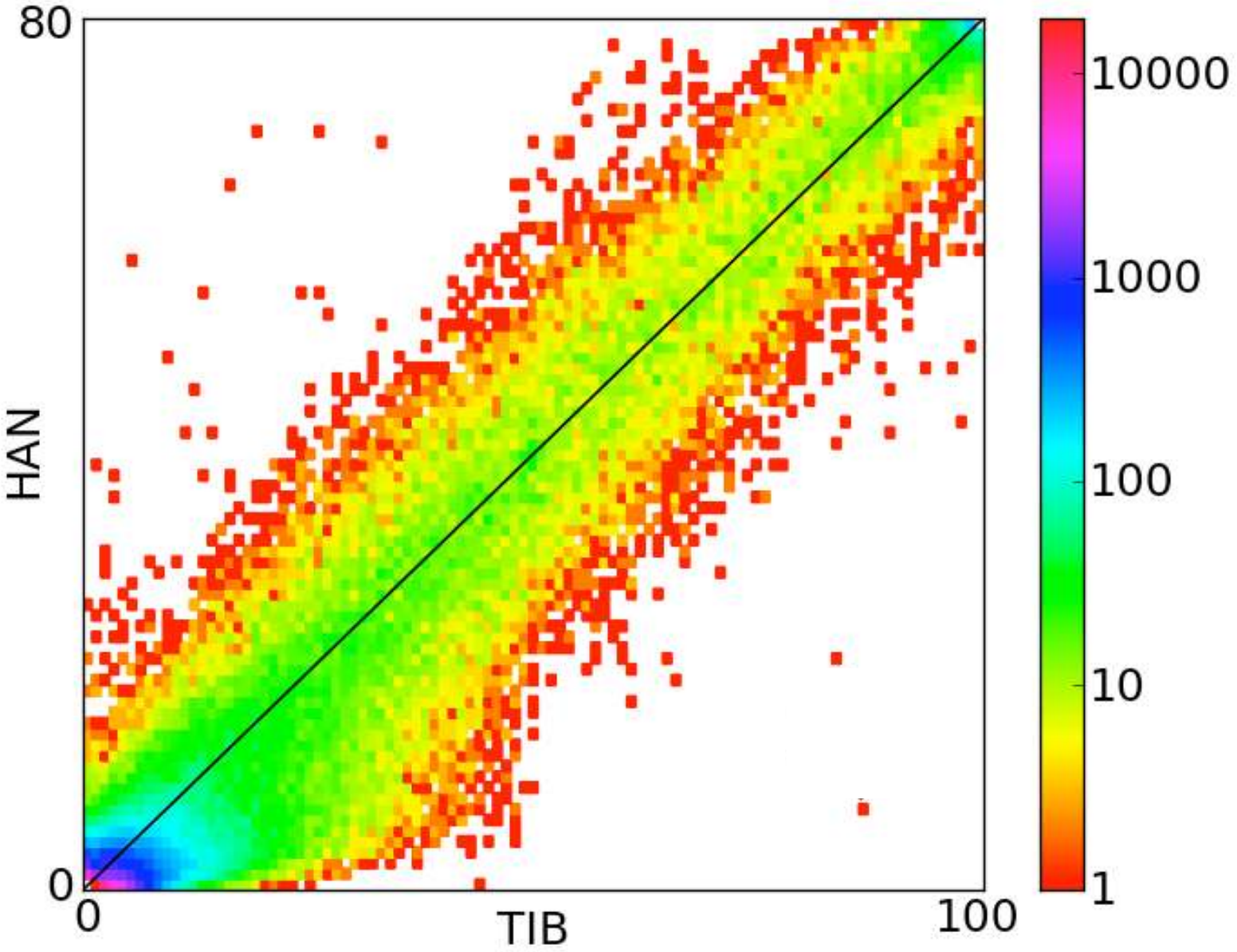
Expected total tree length:

$$3\times\left(\frac{1}{3}\right)+2\times1=3$$

Expected length of blue edge: 1

P(*x* ={1,2}) = 1/3

Site Frequency Spectrum



Data

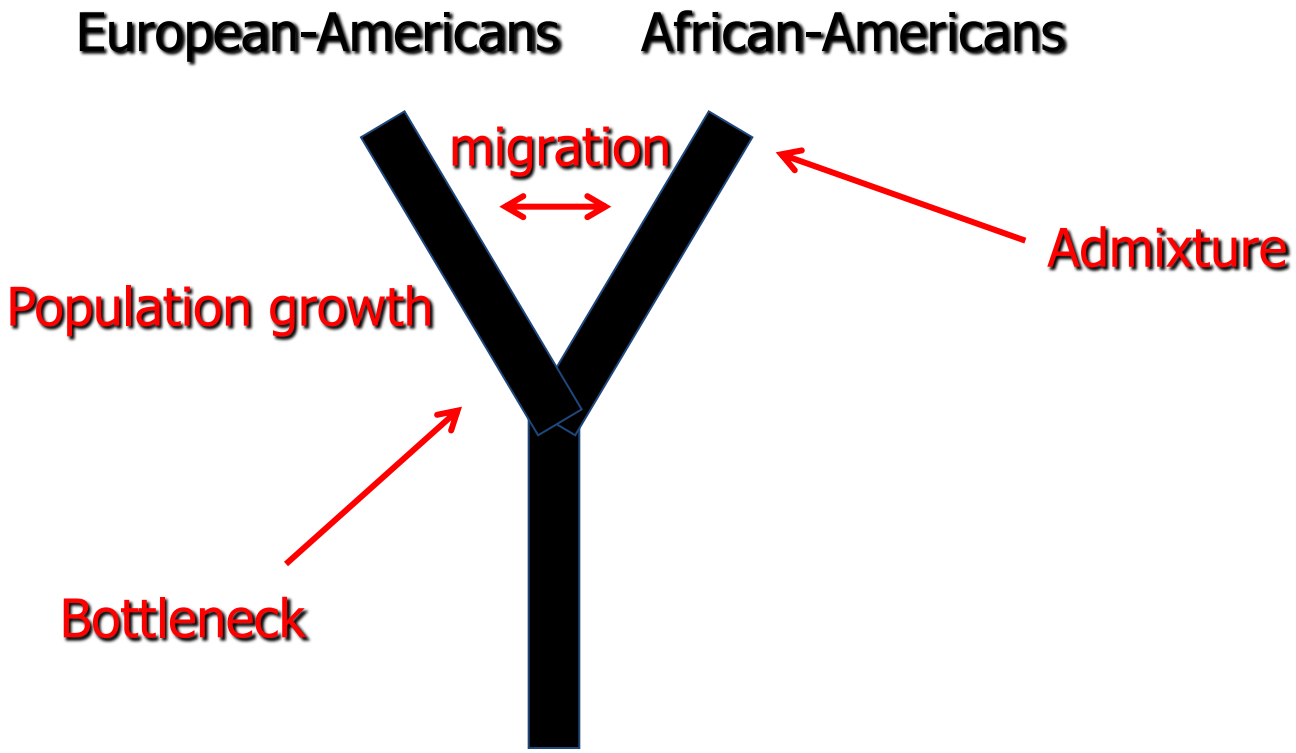
Directly sequenced polymorphism data from 20 European-Americans, 19 African-Americans and one chimpanzee from 9,316 protein coding genes (Bustamante et al. 2005).

Objectives

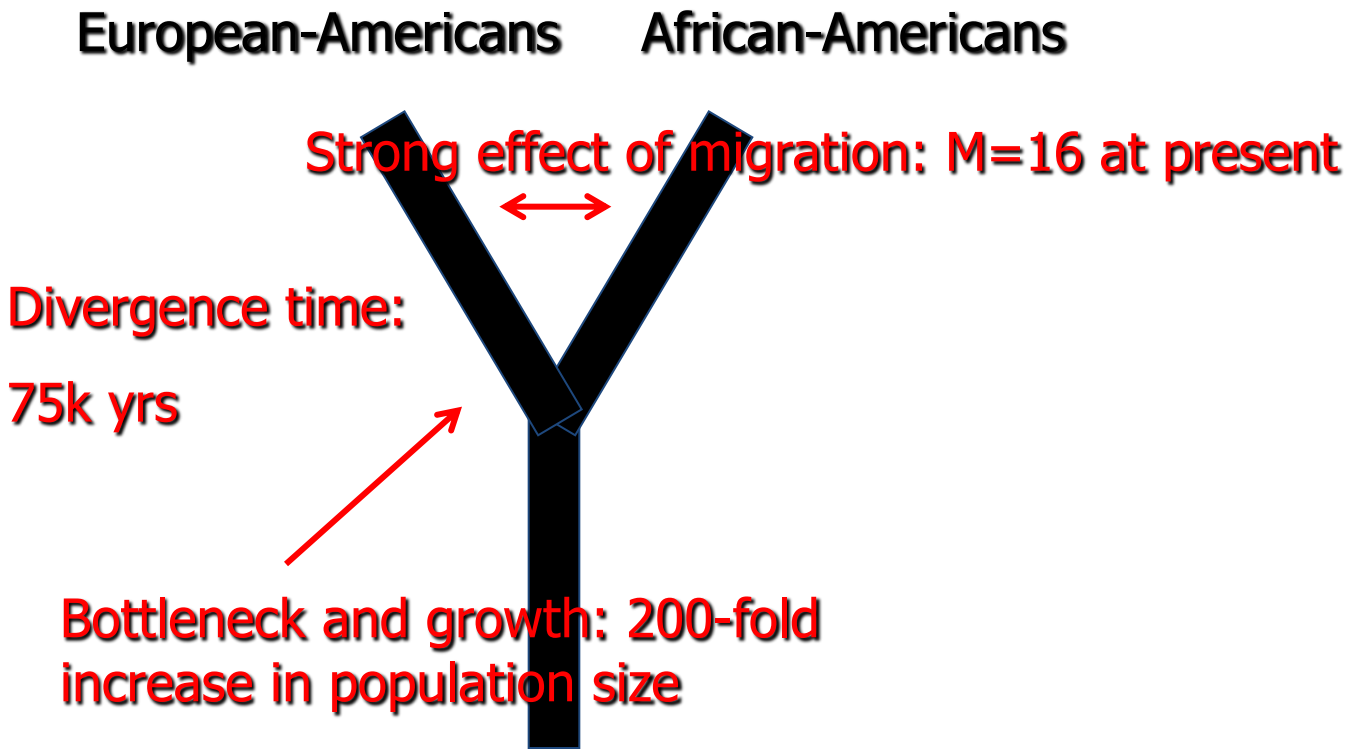
To detect natural selection in individual genes using the frequency spectrum.

To account for demography by estimating parameters of a demographic model.

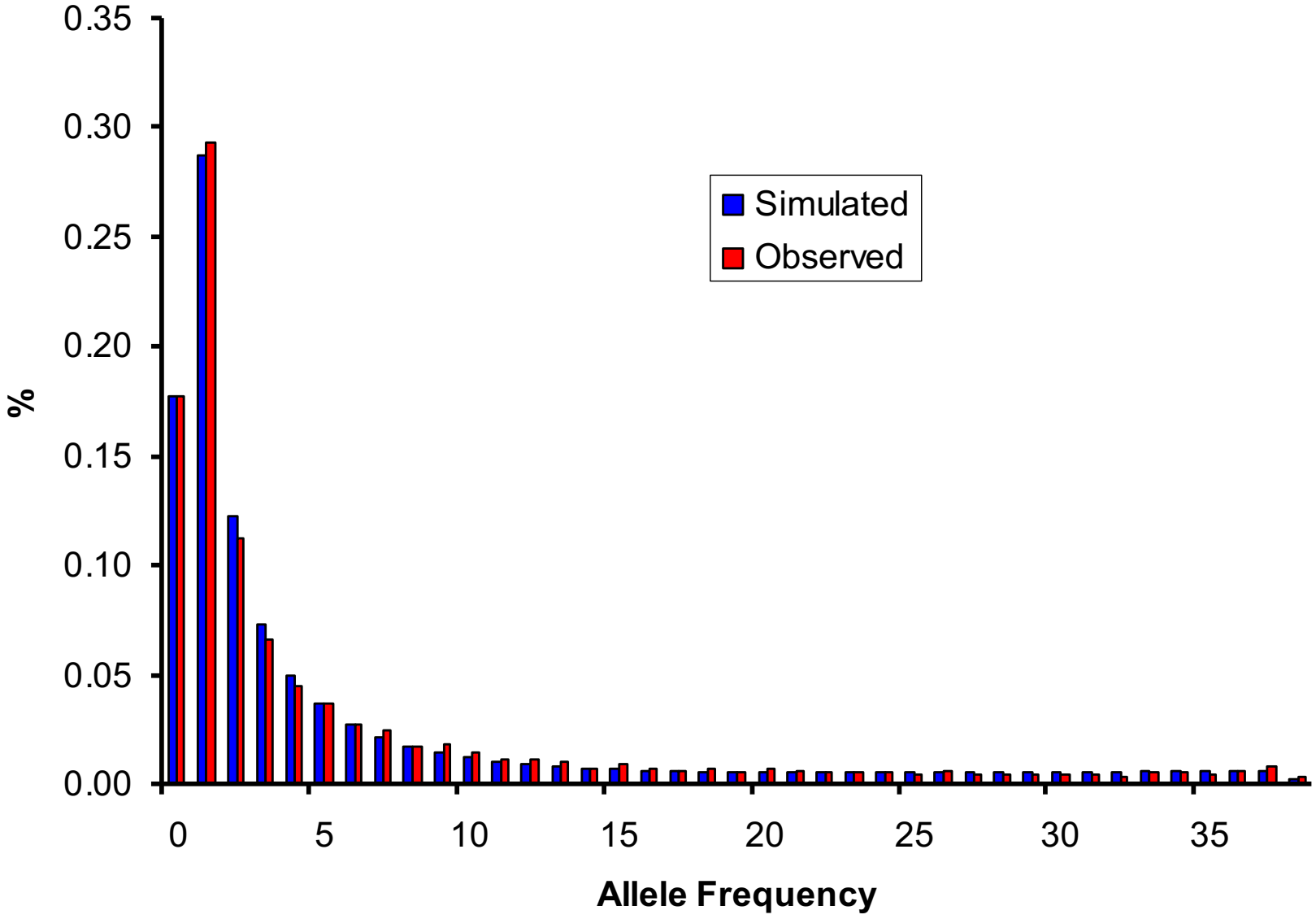
Demographic model



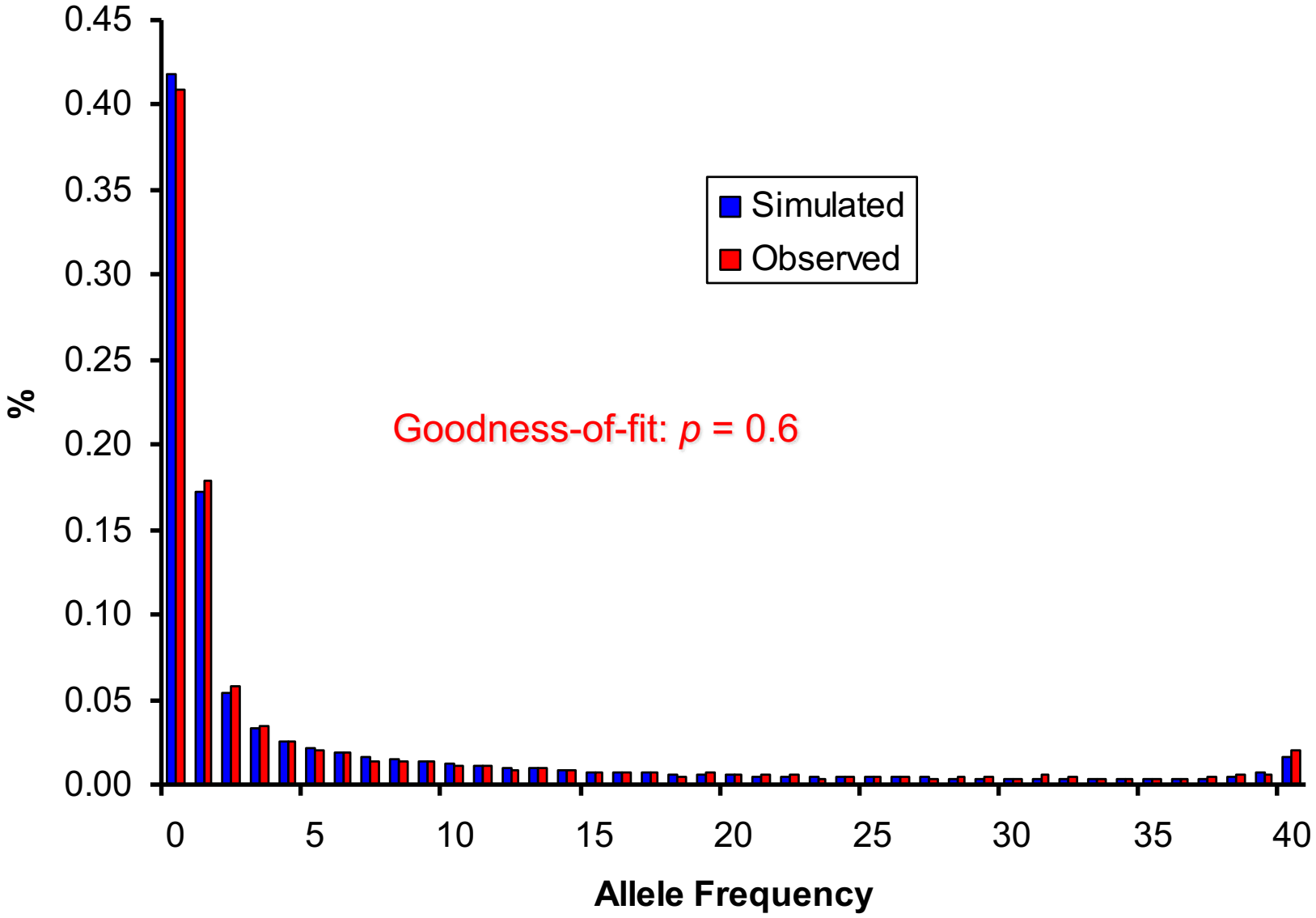
Estimates

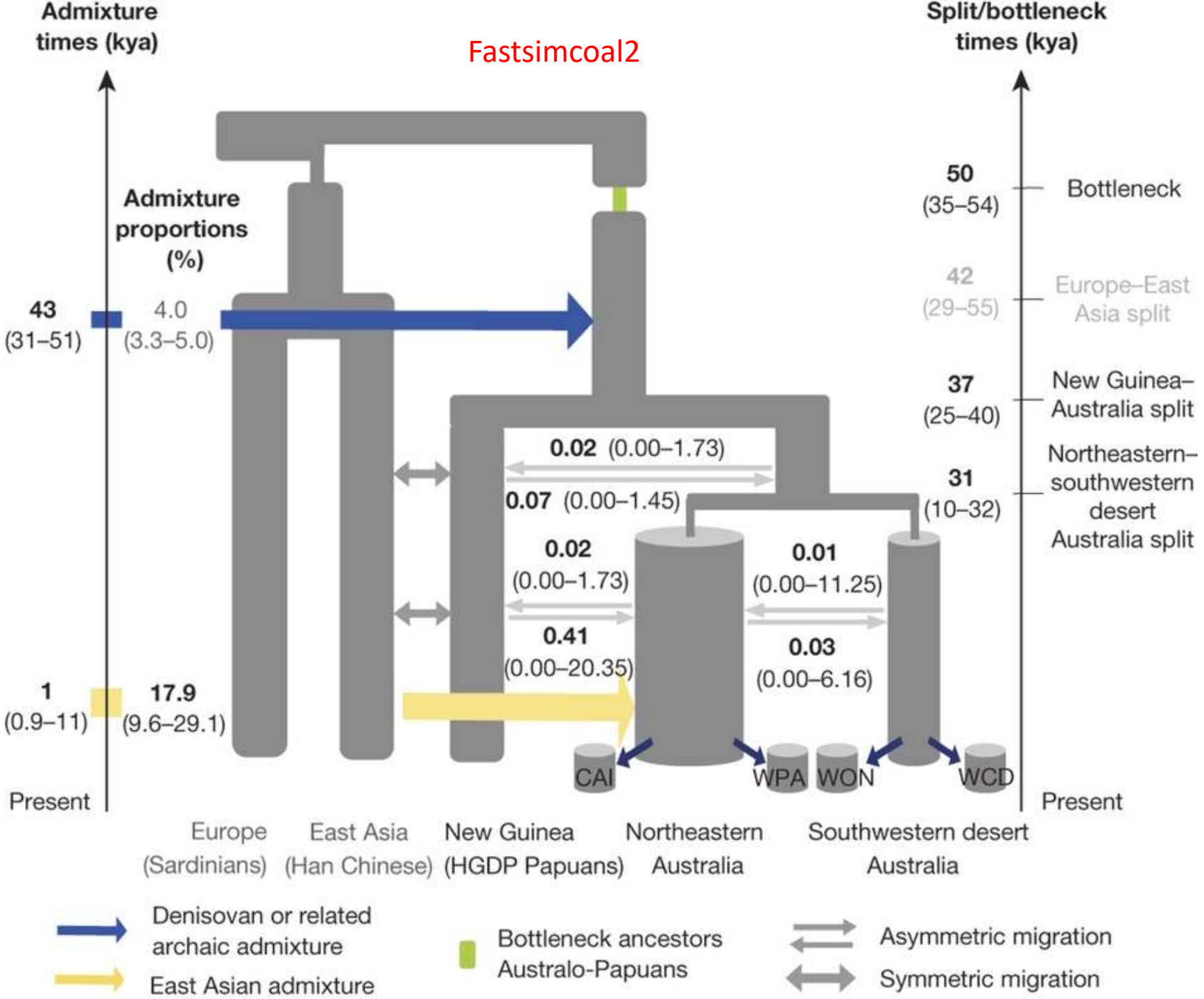

European-Americans
African-Americans
Strong effect of migration: M=16 at present
Divergence time:
75k yrs
Bottleneck and growth: 200-fold increase in population size

African-Americans



European-Americans




Admixture times (kya)
Fastsimcoal2
Split/bottleneck times (kya)
Admixture proportions (%)
43 (31–51)
4.0 (3.3–5.0)
1 (0.9–11)
17.9 (9.6–29.1)
Present
50 (35–54)
Bottleneck
42 (29–55)
Europe–East Asia split
37 (25–40)
New Guinea–Australia split
31 (10–32)
Northeastern–southwestern desert Australia split
0.02 (0.00–1.73)
0.07 (0.00–1.45)
0.02 (0.00–1.73)
0.41 (0.00–20.35)
0.01 (0.00–11.25)
0.03 (0.00–6.16)
CAI
WPA
WON
WCD
Present
Europe (Sardinians)
East Asia (Han Chinese)
New Guinea (HGDP Papuans)
Northeastern Australia
Southwestern desert Australia
Denisovan or related archaic admixture
East Asian admixture
Bottleneck ancestors Australo-Papuans
Asymmetric migration
Symmetric migration

Alternative…

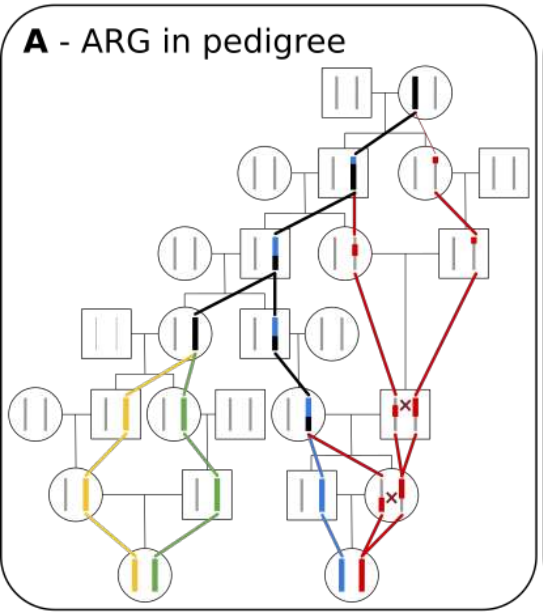
(e.g. Williamson et al. 2005)

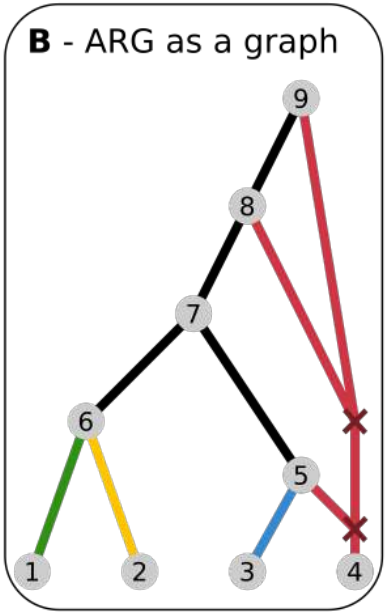
Use $\Pr(X \mid \Theta) = \int \binom{n}{x} p^x (1-p)^{n-x} f(p \mid \Theta) dp$

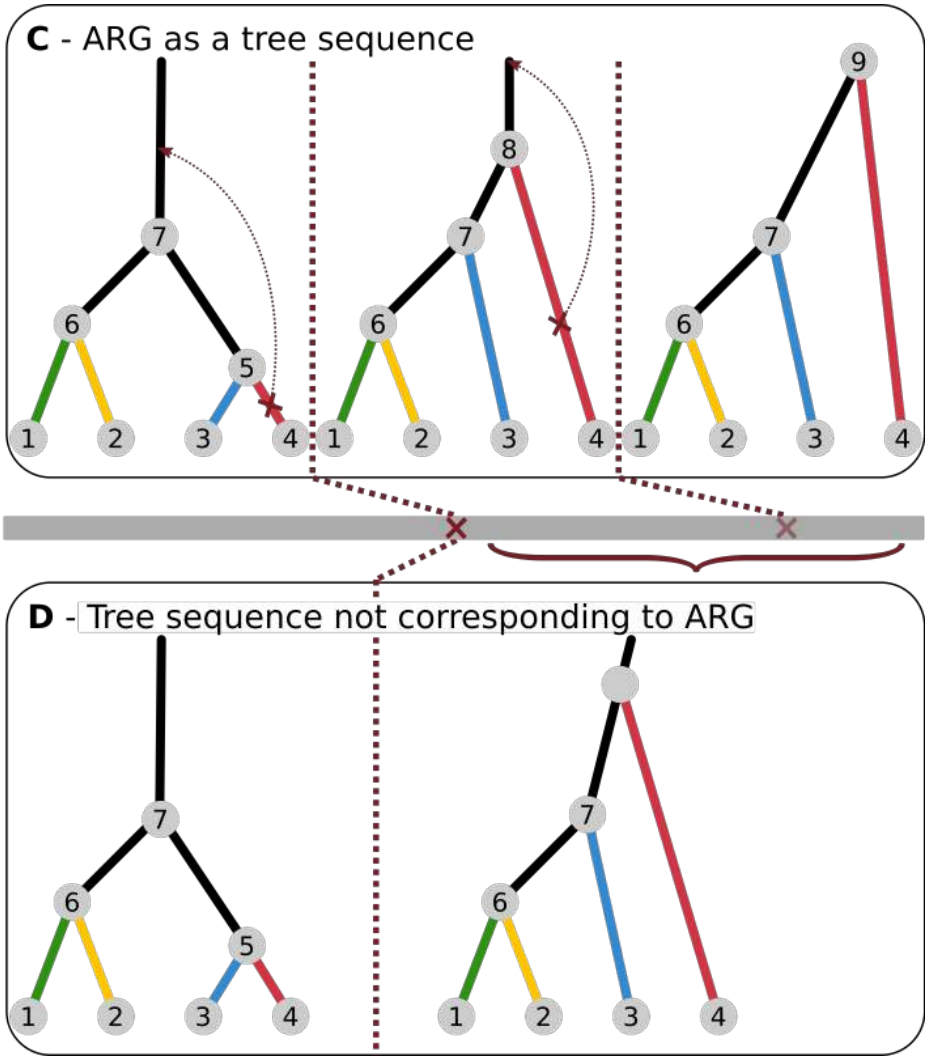
The density $f(p \mid \Theta)$ can be obtaiened by numerical solution of diffusion equations (e.g. Crank–Nicolson approximation).

SFS based infrences

- Simulation (e.g., Fastsimcoal2)
- Numerical solution of diffusion equation (e.g., dadi).


A - ARG in pedigree


B - ARG as a graph
9
8
7
6
5
1
2
3
4


C - ARG as a tree sequence
D - Tree sequence not corresponding to ARG

OPEN ACCESS Freely available online

PLOS | GENETICS

Genome-Wide Inference of Ancestral Recombination Graphs

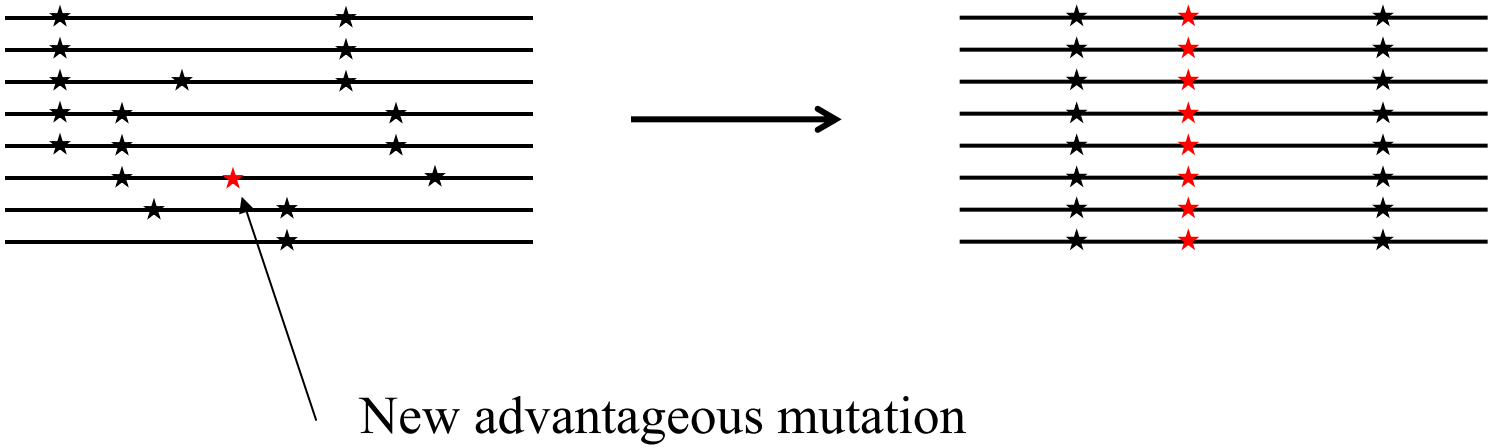
Matthew D. Rasmussen[1¤]*, Melissa J. Hubisz[1], Ilan Gronau[1], Adam Siepel[1,2]*

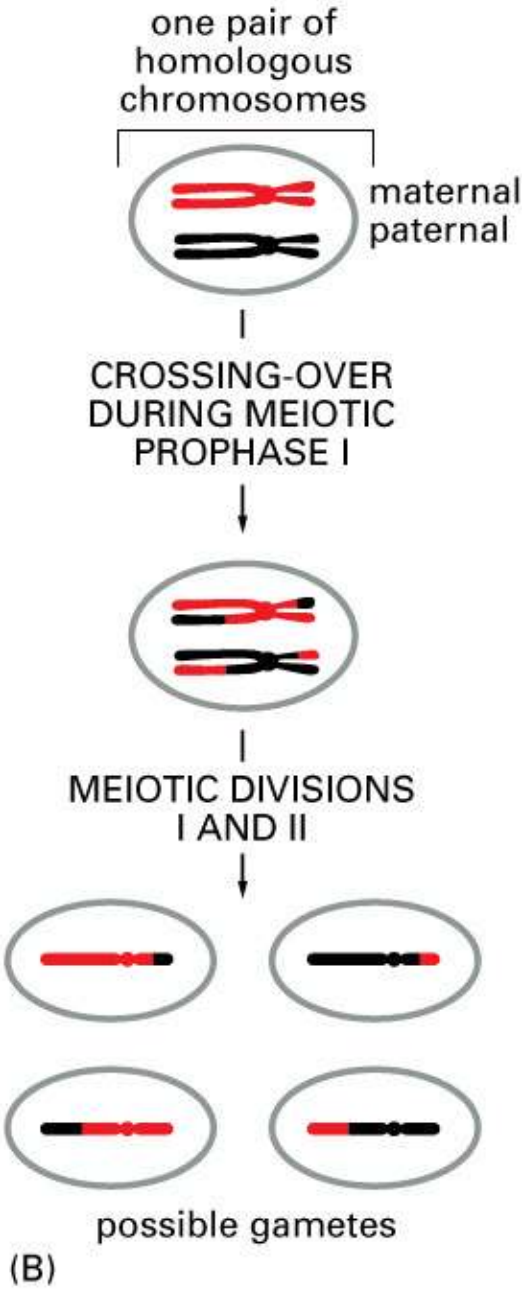
1 Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York, United States of America, **2** European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambs, United Kingdom

Abstract

The complex correlation structure of a collection of orthologous DNA sequences is uniquely captured by the "ancestral recombination graph" (ARG), a complete record of coalescence and recombination events in the history of the sample. However, existing methods for ARG inference are computationally intensive, highly approximate, or limited to small

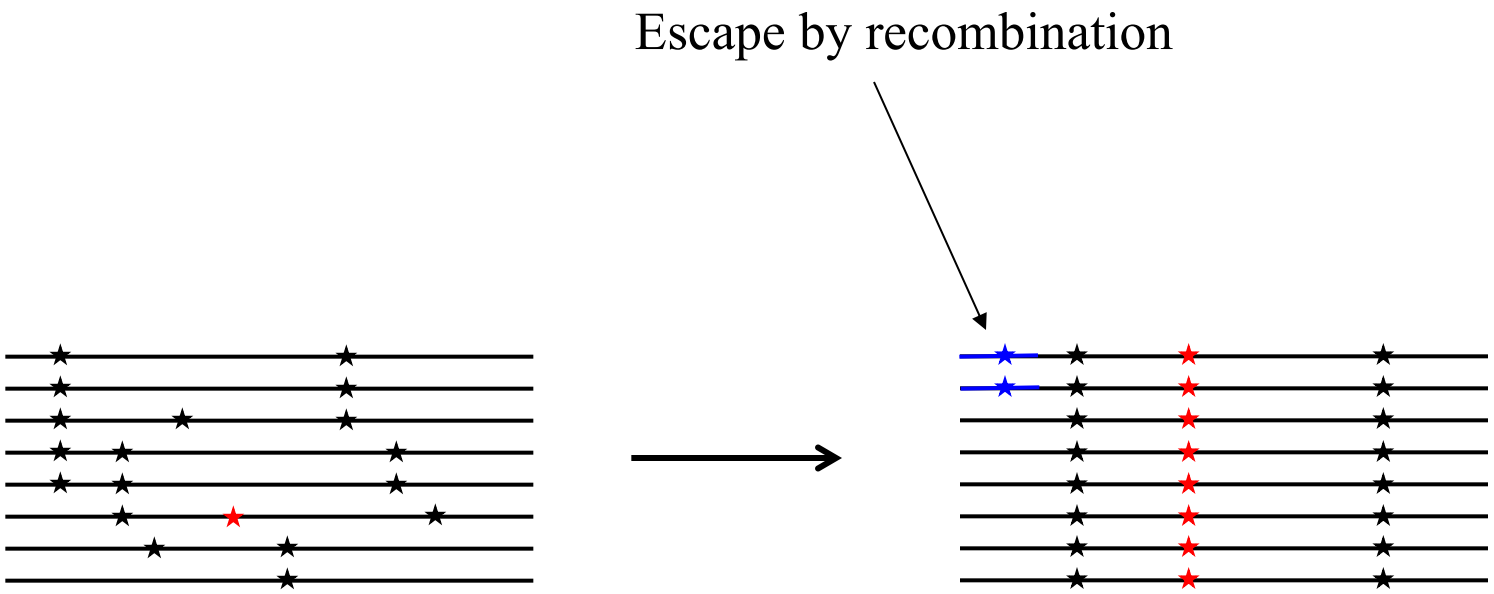
Selective Sweeps

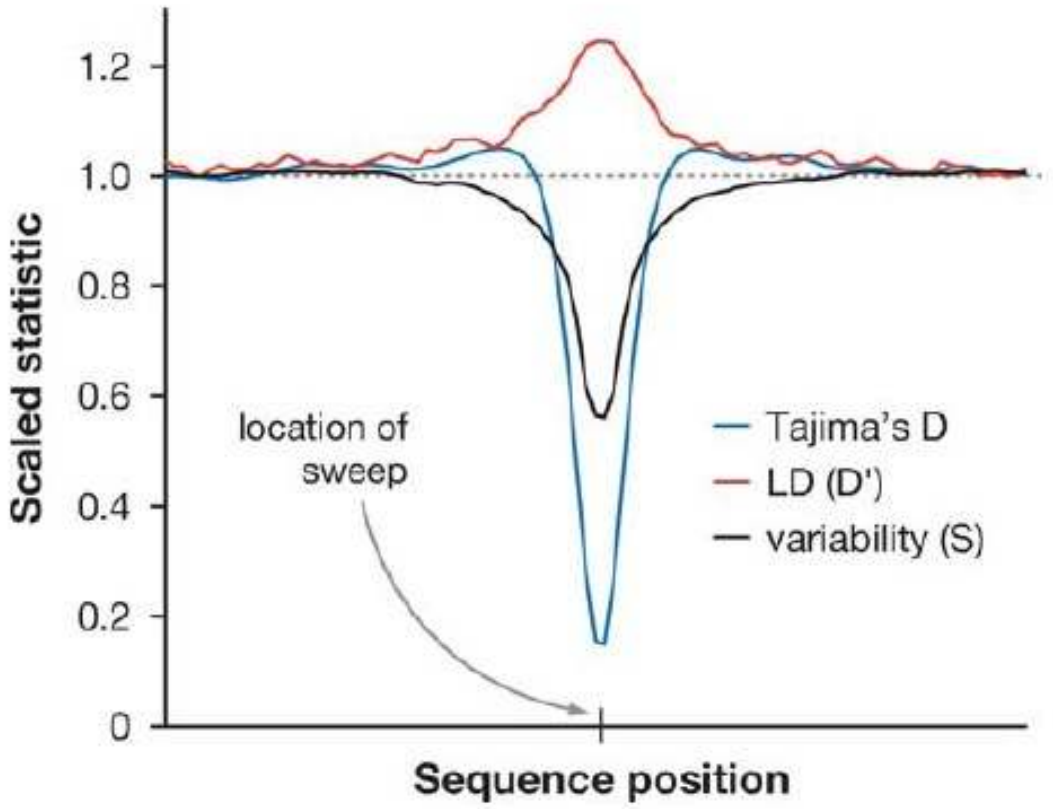



one pair of homologous chromosomes
maternal
paternal
CROSSING-OVER DURING MEIOTIC PROPHASE I
MEIOTIC DIVISIONS I AND II
possible gametes

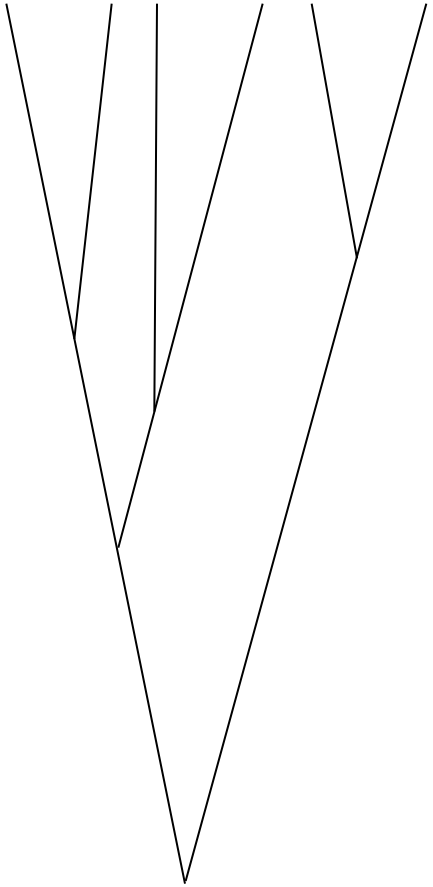
(B)

Selective Sweeps



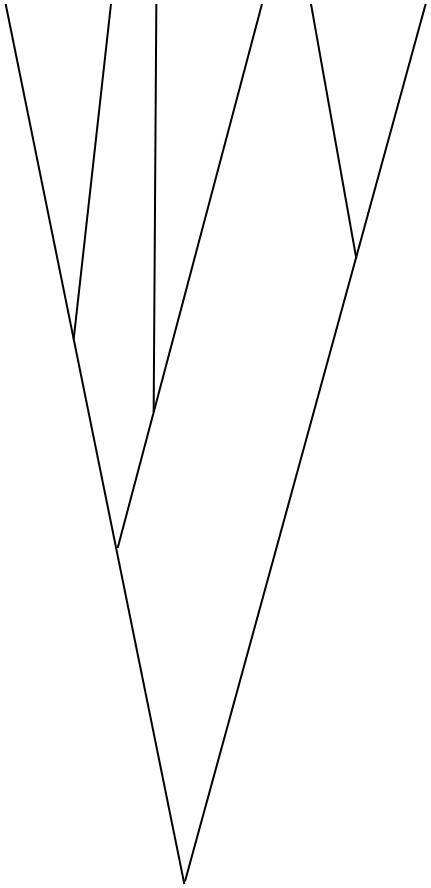

1.2
1.0
0.8
0.6
0.4
0.2
0
Scaled statistic
location of sweep
Tajima's D
LD (D')
variability (S)
Sequence position

Coalescence Tree

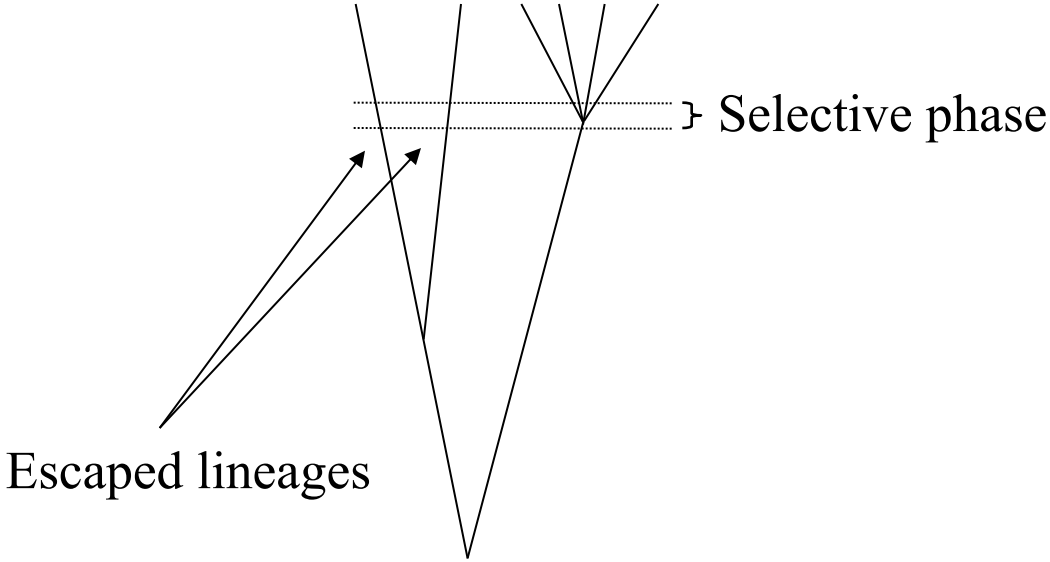


Coalescence Models

Neutral coalescence tree



Coalescence tree with sweep



Importance Sampling

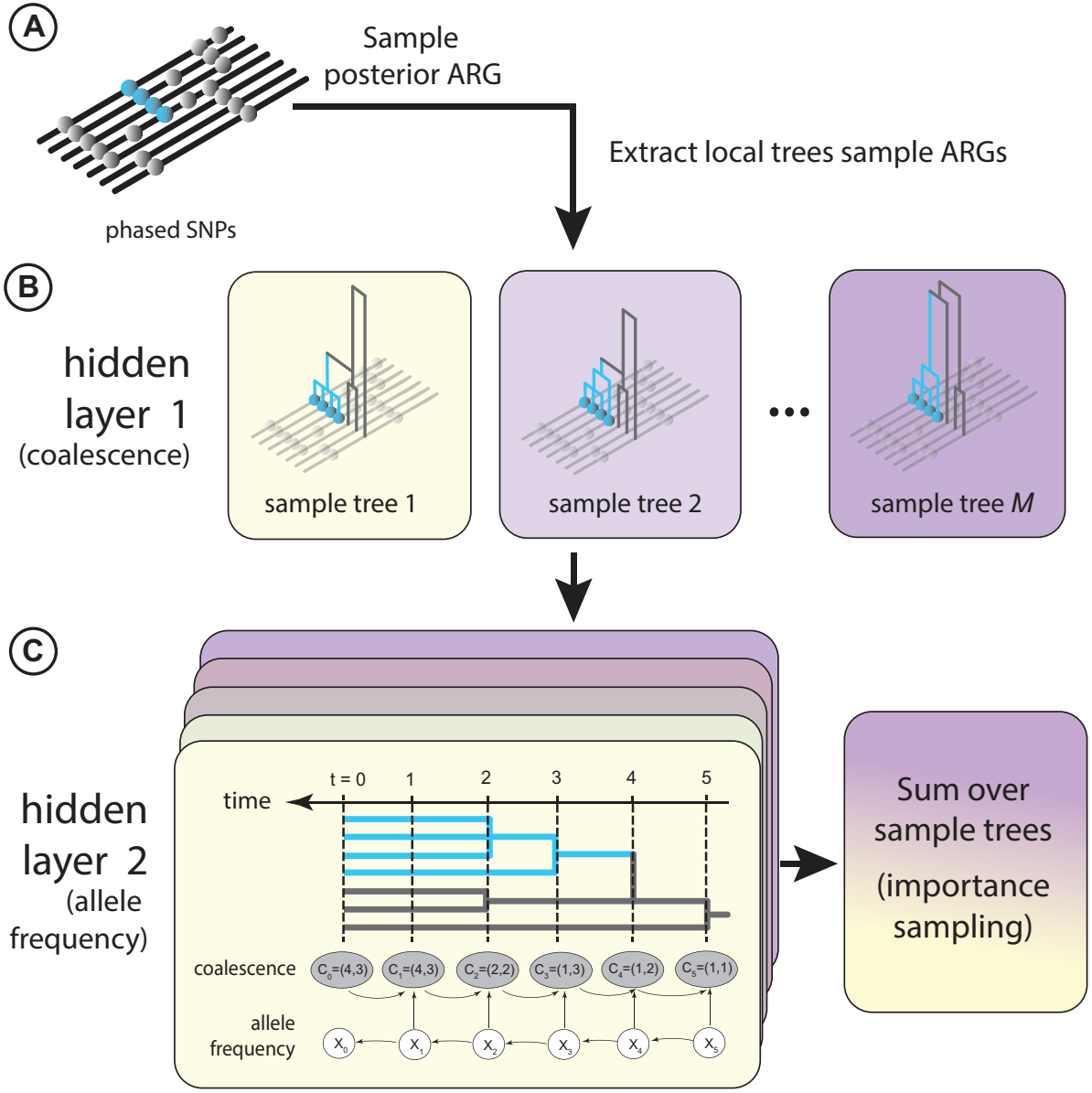
$$\int_{\psi} \Pr(X \mid G)\, p(G \mid \Theta) dG = \int_{\psi} \Pr(X \mid G)\, p(G \mid \Theta) \frac{h(G)}{h(G)} dG$$

$$= E\left[\frac{\Pr(X \mid G)\, p(G \mid \Theta)}{h(G)} \right]$$

So

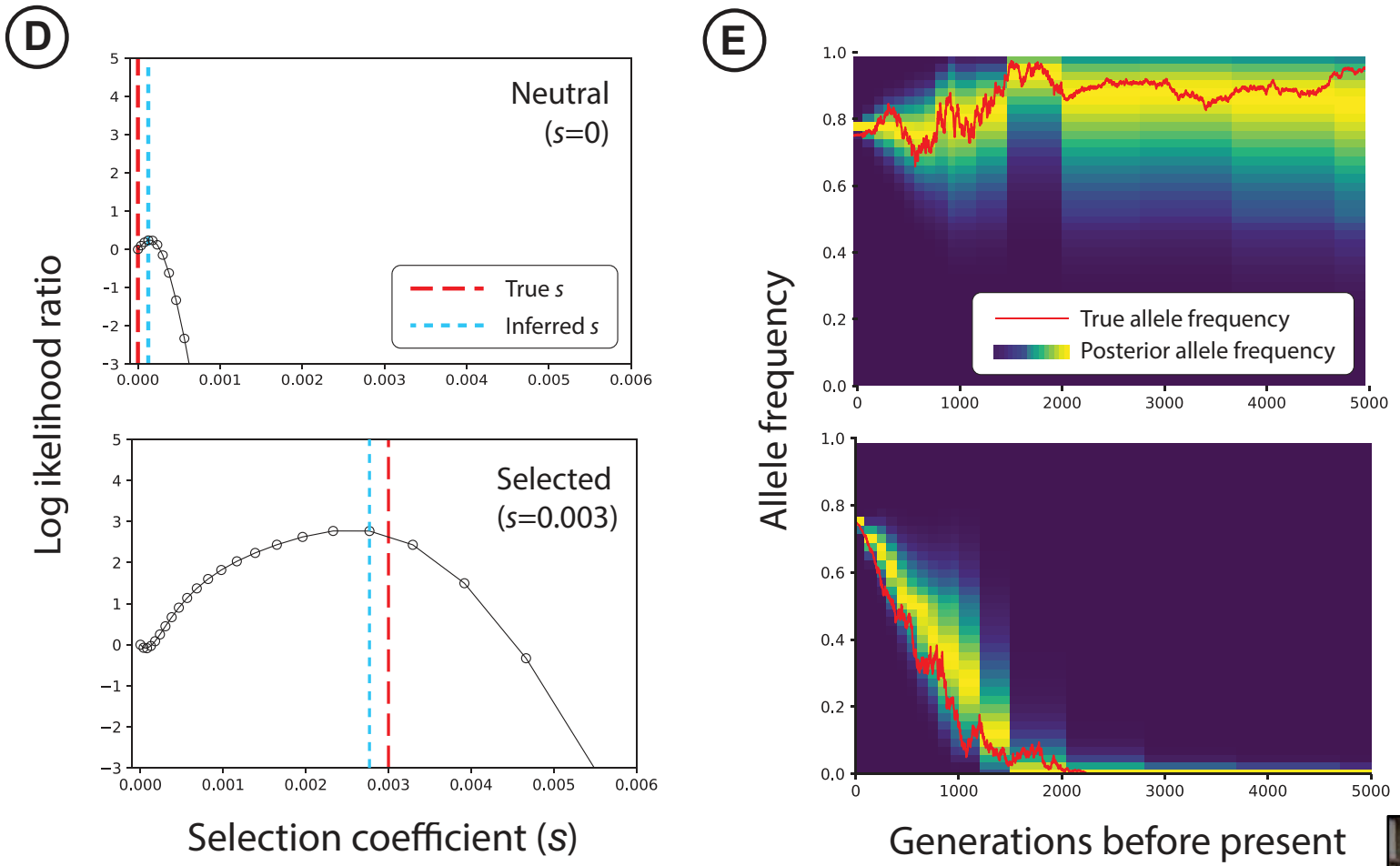
$$\Pr(X \mid \Theta) \approx \frac{1}{k} \sum_{i=1}^{k} \frac{\Pr(X \mid G_i)\, p(G_i \mid \Theta)}{h(G_i)}$$

where G_i, i=1,2,…k, has been simulated from $h(G)$.

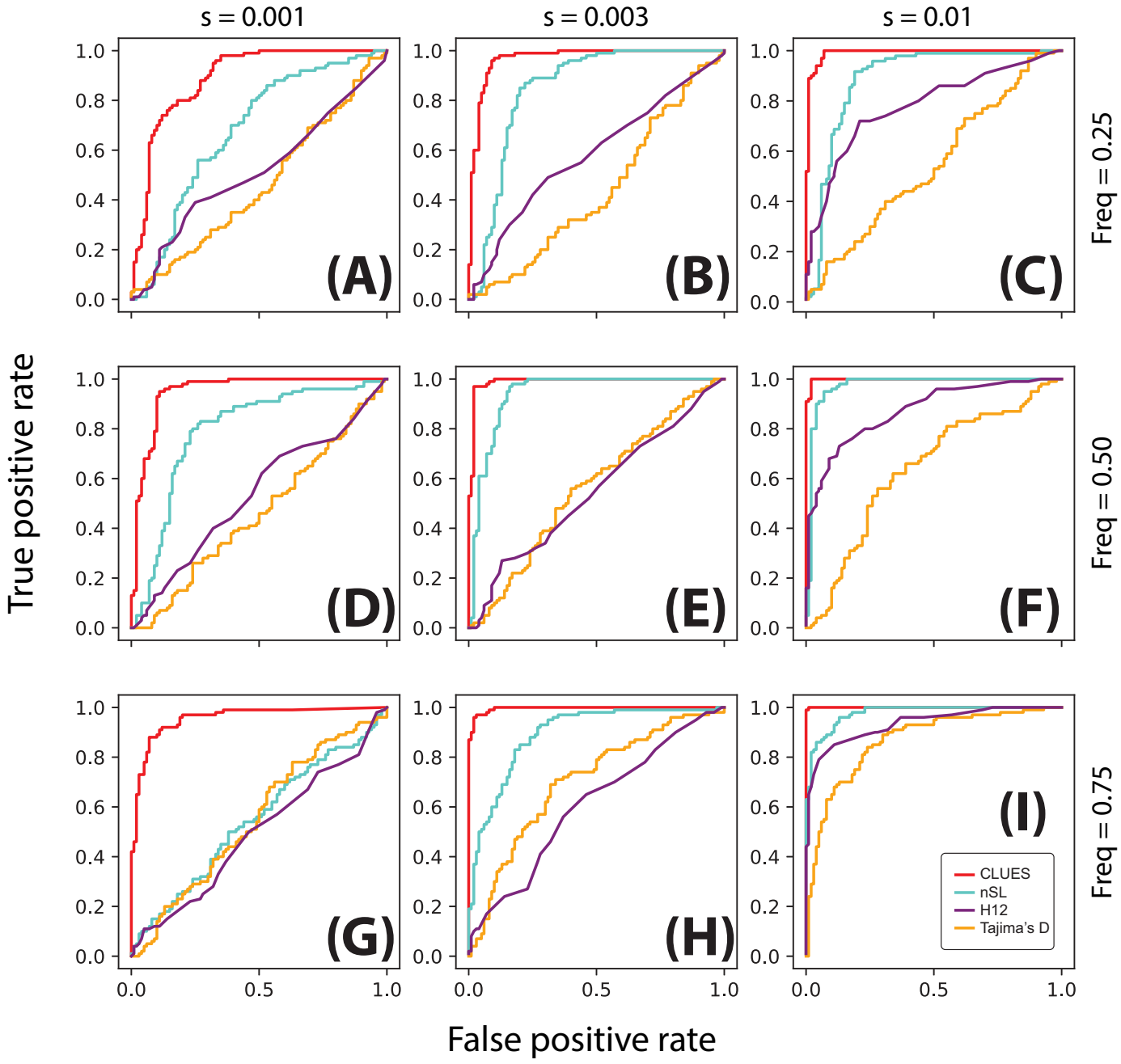


Aaron Stern

Stern et al. 2019. bioRxiv http://dx.doi.org/10.1101/592675.



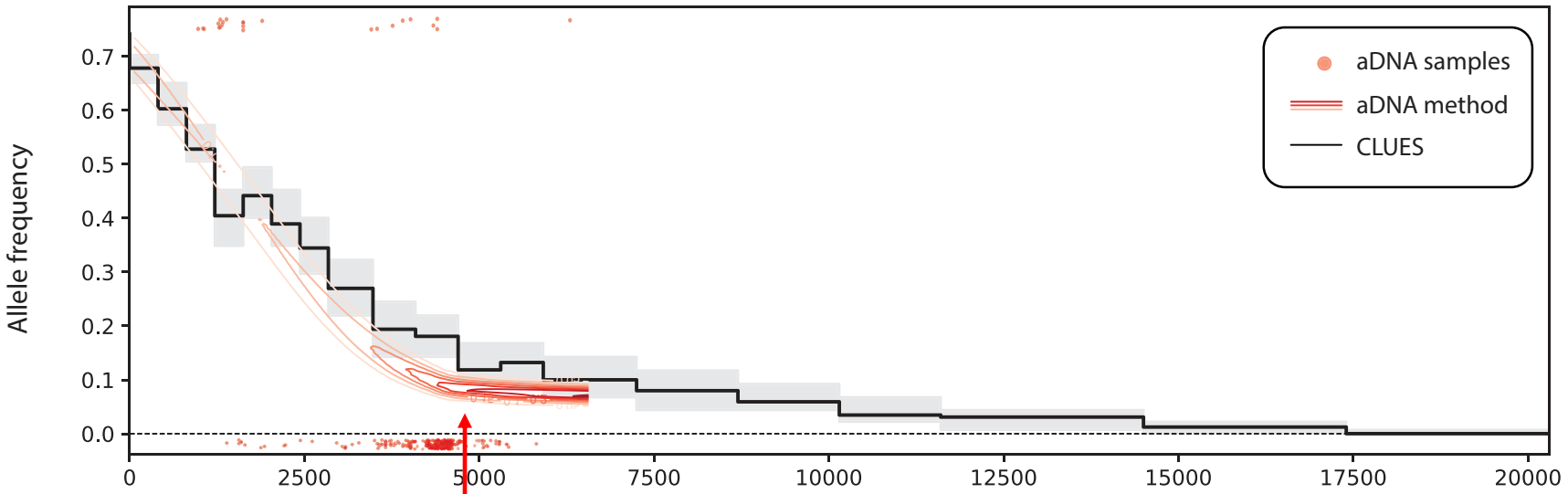
Stern et al. 2019. bioRxiv http://dx.doi.org/10.1101/592675.



Aaron Stern

Stern et al. 2019. bioRxiv http://dx.doi.org/10.1101/592675.

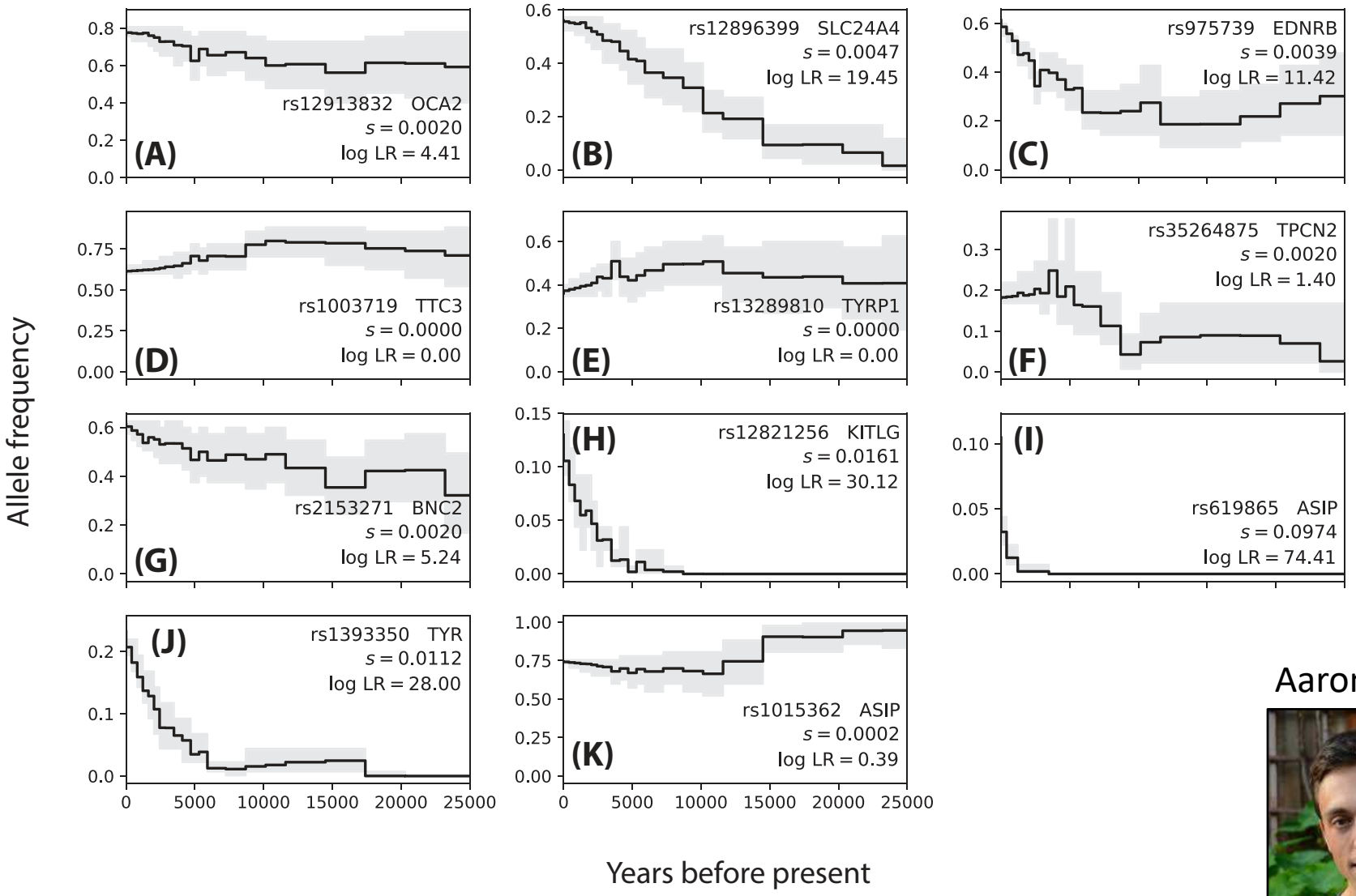
Lactase rs4988235



From Mathieson and Mathieson. 2018. *MBE* 35:2957–2970.

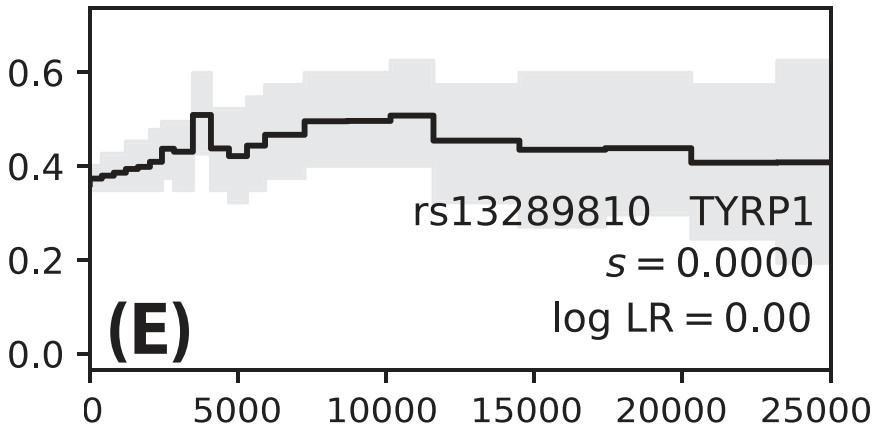
Aaron Stern

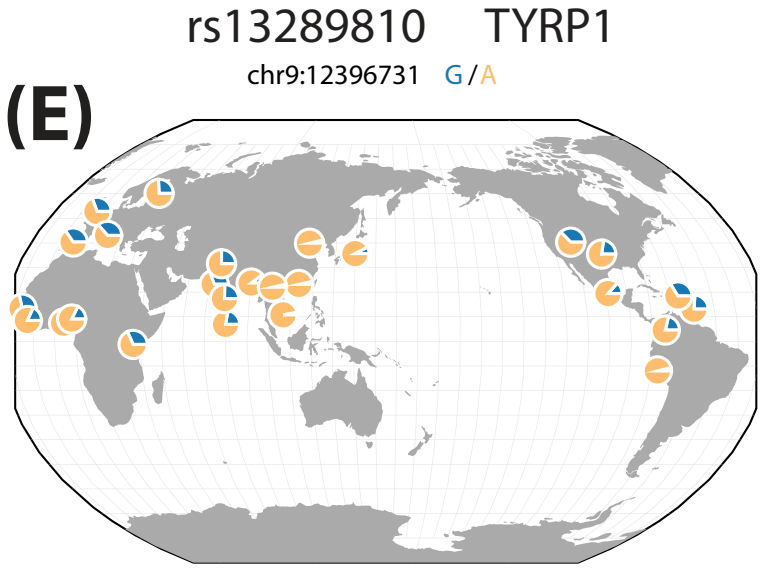
Stern et al. 2019. bioRxiv http://dx.doi.org/10.1101/592675.



Aaron Stern

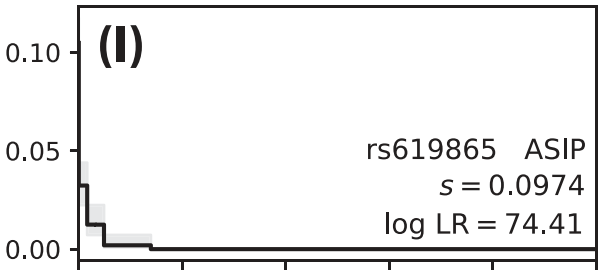
Stern et al. 2019. bioRxiv http://dx.doi.org/10.1101/592675.

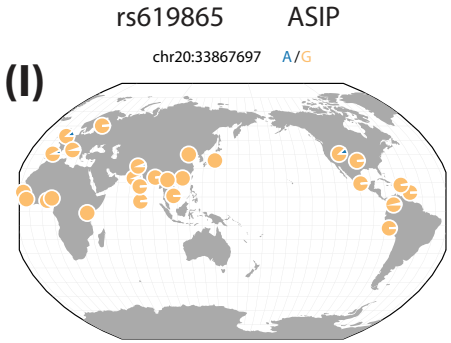


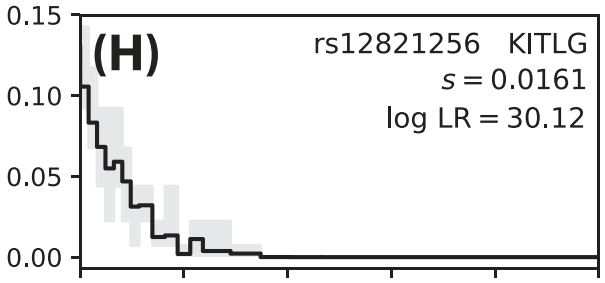


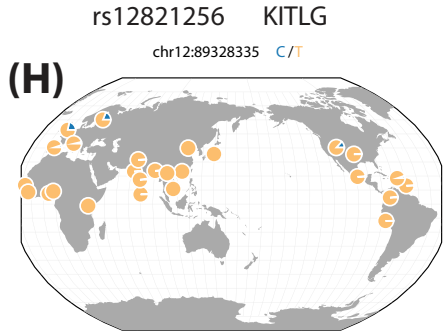
Aaron Stern

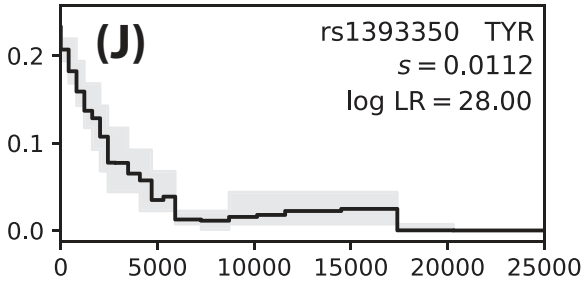
Stern et al. 2019. bioRxiv http://dx.doi.org/10.1101/592675.

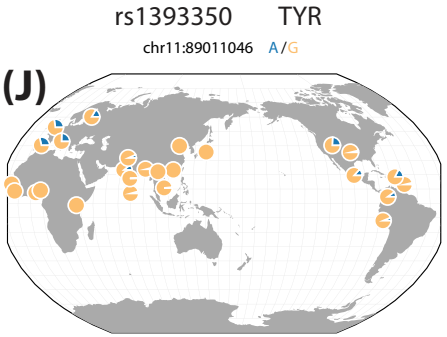

(I)
rs619865 ASIP
s = 0.0974
log LR = 74.41
0.10
0.05
0.00


rs619865 ASIP
chr20:33867697 A/G
(I)

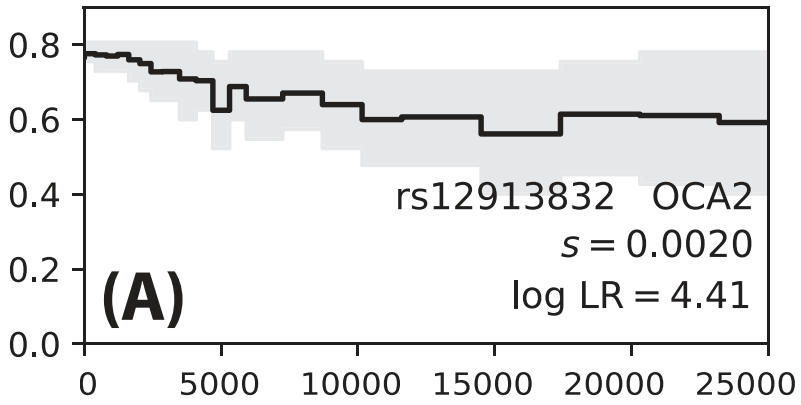

(H)
rs12821256 KITLG
s = 0.0161
log LR = 30.12
0.15
0.10
0.05
0.00

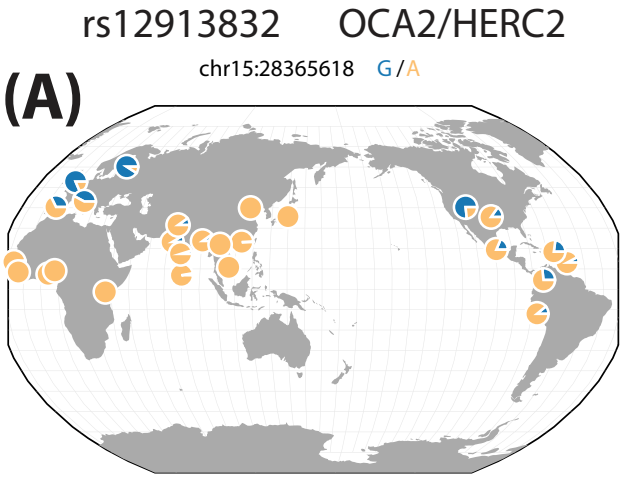

rs12821256 KITLG
chr12:89328335 C/T
(H)

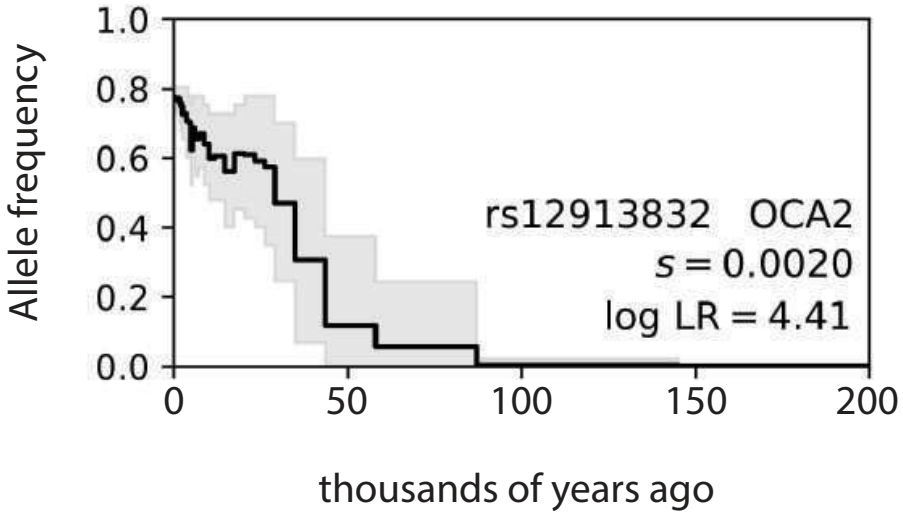

(J)
rs1393350 TYR
s = 0.0112
log LR = 28.00
0.2
0.1
0.0
0
5000
10000
15000
20000
25000


rs1393350 TYR
chr11:89011046 A/G
(J)

Aaron Stern



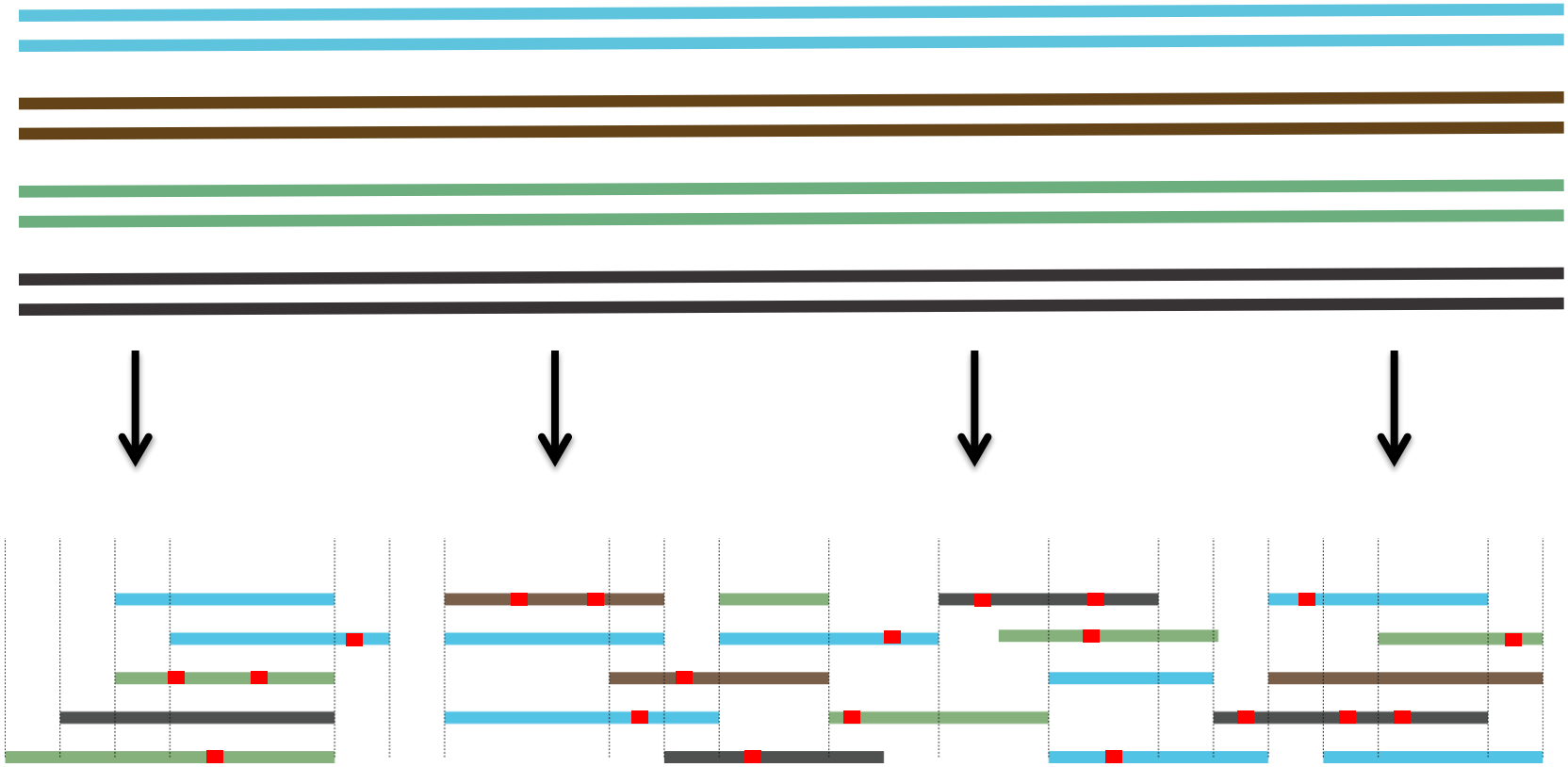




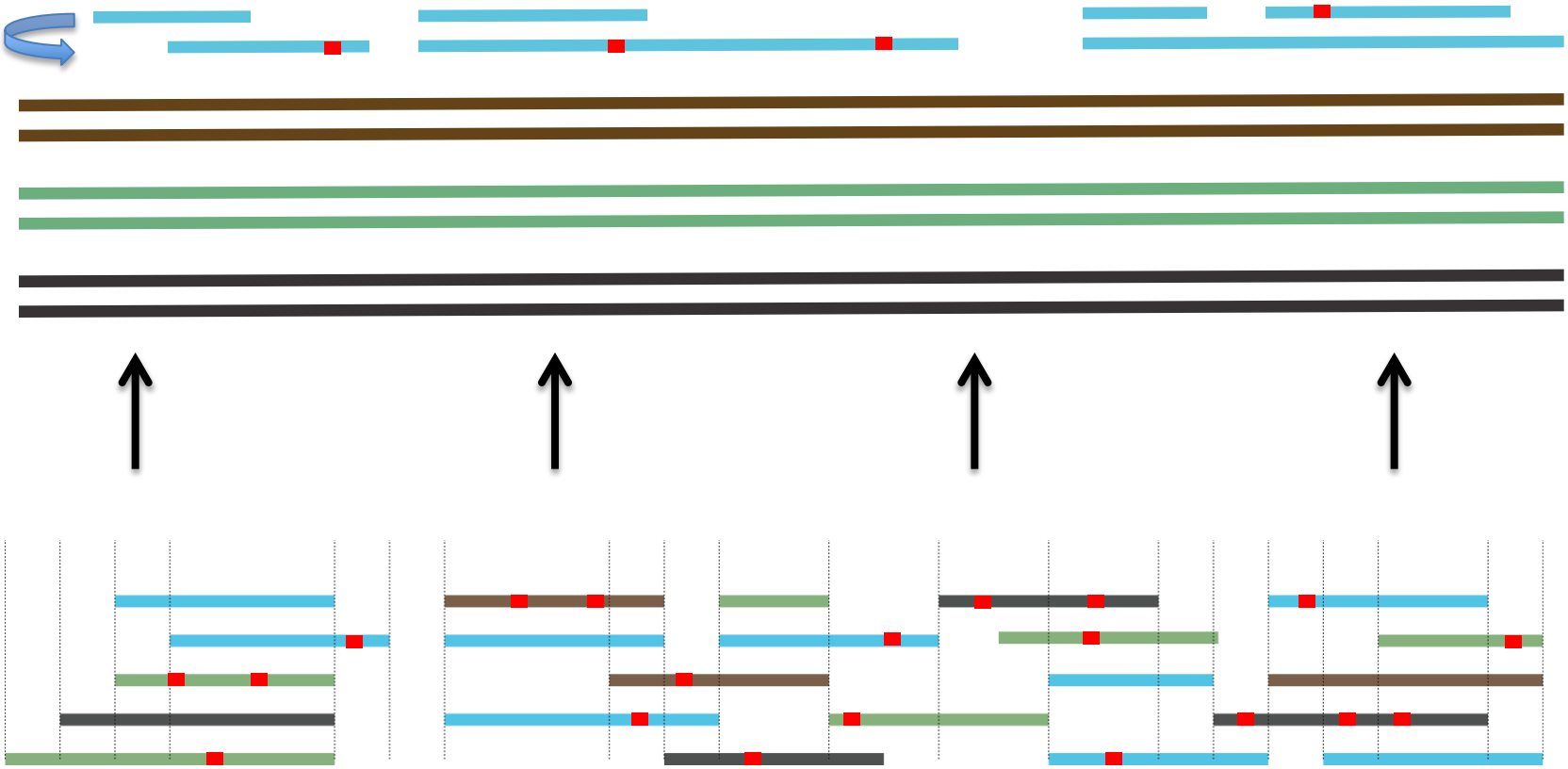
Aaron Stern

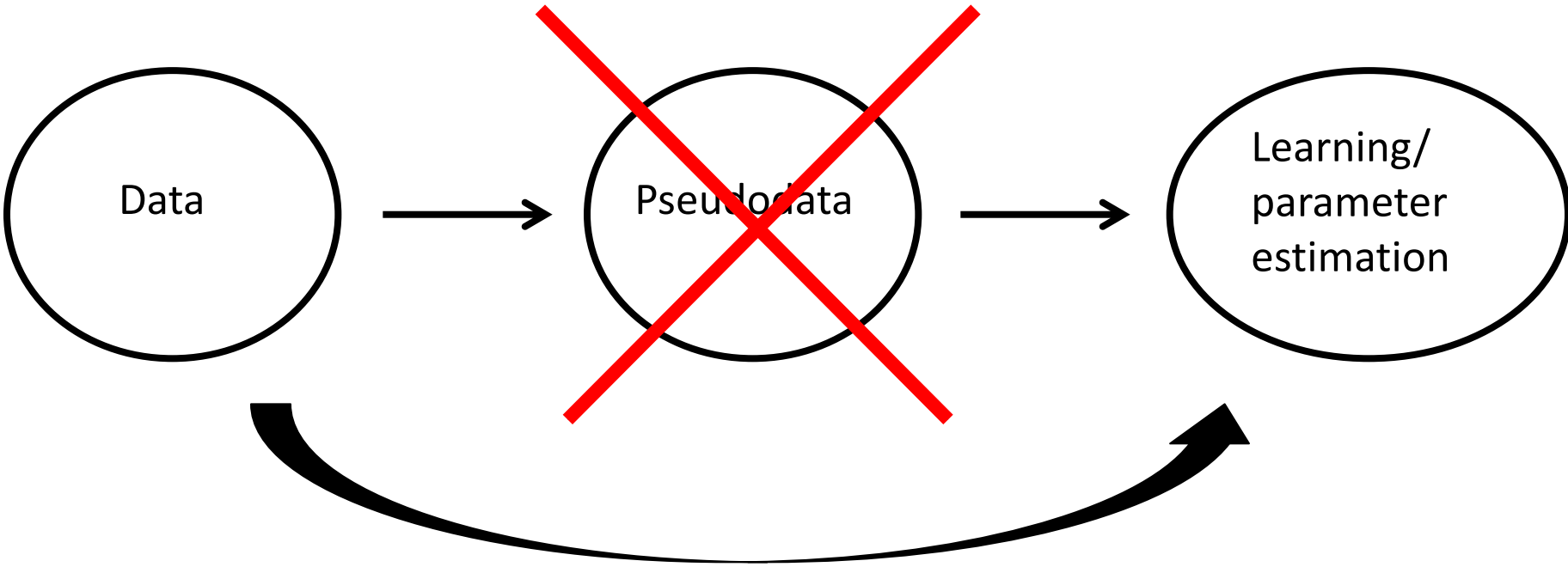
Stern et al. 2019. bioRxiv http://dx.doi.org/10.1101/592675.

Problem 2: DNA is chopped into short ‘reads’ with high error rates

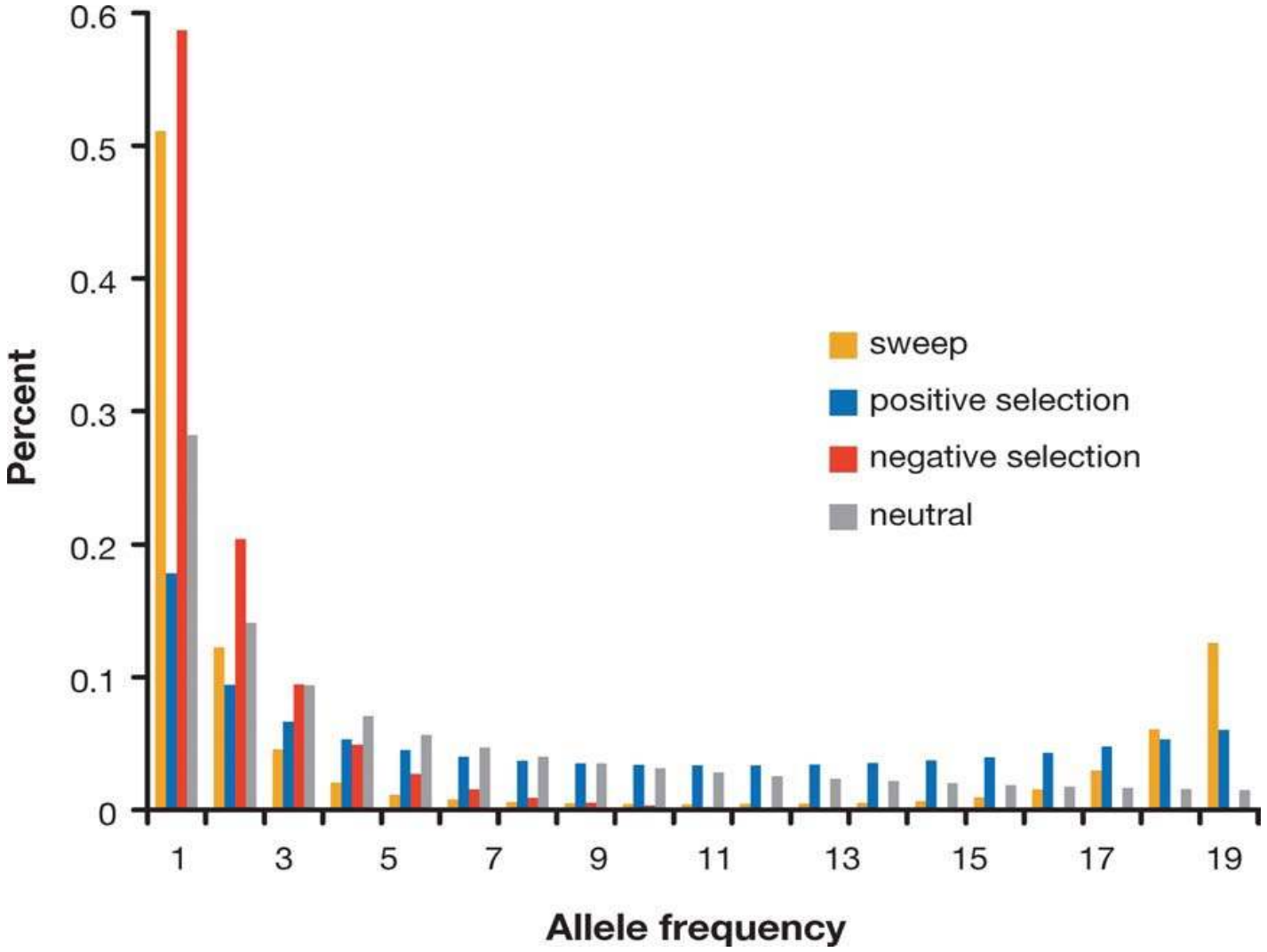


Reverse problem

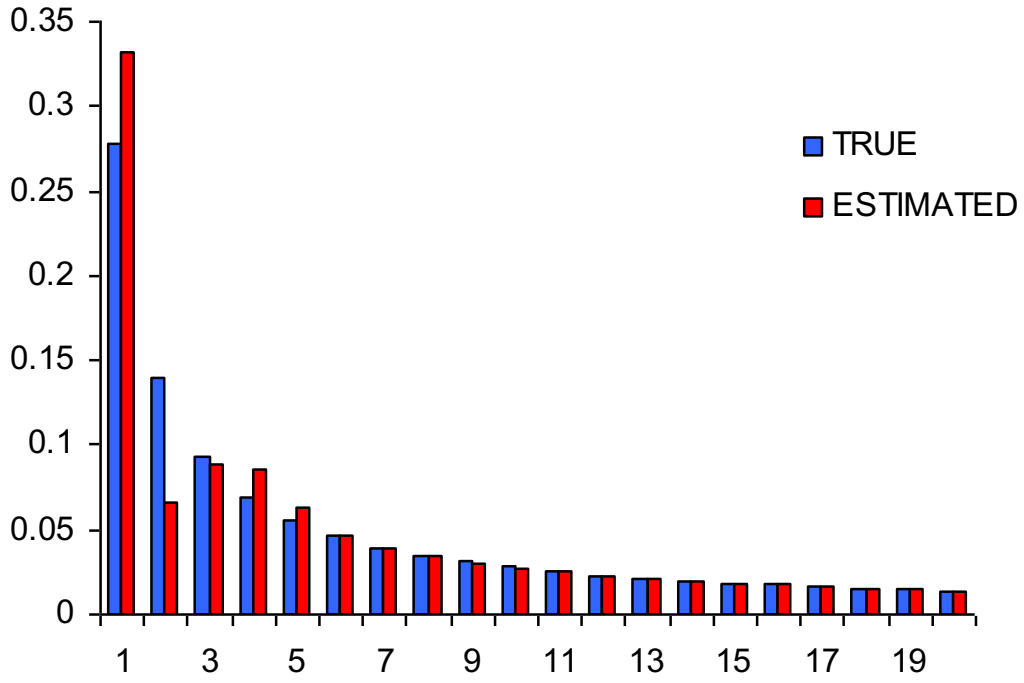



Data
Pseudodata
Learning/
parameter
estimation

Frequency spectrum



10X Next Gen. Sequencing data



Using a fixed cut-off for SNP calling can never produce unbiased allele frequency estimates for all values of *p*.

Maximum Likelihood Estimation

- Assume genotype likelihoods, $p(X_i^{(v)} \mid G_i^{(v)})$, can be calculated in site v, individual d, for all v and d .
- Parameterize the frequency spectrum as $\mathbf{P} = (p_0, p_1, \ldots, p_{2k})$.
- Likelihood function:

$$L(\mathbf{P}) = \prod_v \left(\sum_{j=0}^{2k} p_j \left[\sum_{G_1^{(v)}} \cdots \sum_{G_k^{(v)}} c(j, G^{(v)}) \prod_{d=0}^{k} p(X_d^{(v)} \mid G_d^{(v)}) \right] \right)$$

$$c(j, G^{(v)}) = \begin{cases} \binom{2k}{j}^{-1} 2^{\sum_{d=1}^{k} I\left(G_d^{(v)}=1\right)} & \text{if } \sum_{d=1}^{k} G_d^{(v)} \neq j \\ 0 \text{ else} \end{cases}$$

$$G^{(v)} = (G_1^{(v)}, G_2^{(v)}, .., G_k^{(v)}), \quad G_d^{(v)} \in \{0,1,2\}$$

Nielsen *et al*. 2012. PloS One 10.1371

Dynamic Programming Algorithm

Initialization:

Set $h_0 = p(X_d \mid G_d = 0), h_1 = 2p(X_d \mid G_d = 1), h_2 = p(X_d \mid G_d = 2)$, and $h_j = 0$ for $j = 3,4,\ldots,2k$.

Recursion

For $d = 2, 3,\ldots, k$:

For $j = 2d, 2d\text{-}1,\ldots,2$:

$$\text{Set } h_j = p(X_d \mid G_d = 2)h_{j-2} + 2p(X_d \mid G_d = 1)h_{j-1} + p(X_d \mid G_d = 0)h_j$$

$$\text{Set } h_{j-1} = p(X_d \mid G_d = 0)h_{j-1} + 2p(X_d \mid G_d = 1)h_{j-2}$$

$$\text{Set } h_{j-2} = p(X_d \mid G_d = 0)h_{j-2}$$

Termination

$$\text{Set } h_{j,} = h_j \binom{2k}{j}^{-1} \text{ for } j=0,1,2,\ldots,2k.$$

Nielsen *et al*. 2012. PloS One 10.1371

Maximum Likelihood Estimation

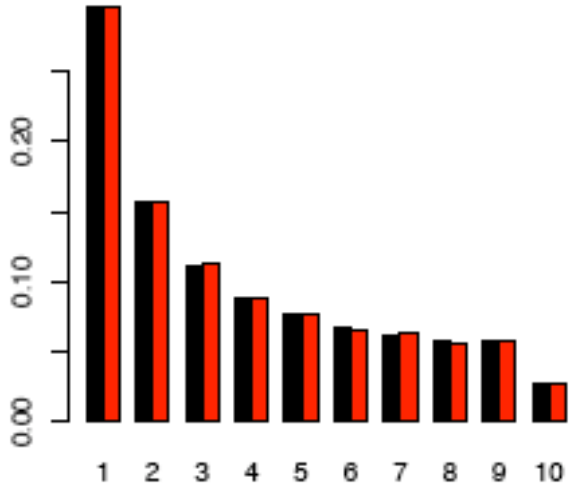
- The likelihood function can then be expressed as

$$L(\mathbf{P}) = \prod_{v} \left(\sum_{j=0}^{2k} p_j h_j^{(v)} \right)$$

- Derivatives can be calculated analytically.
- Algorithm which is quadratic in the number of individuals and linear in the number of sites.
- Optimization to eight decimals precision takes < 1 minute for 1 GB and 60 individuals on one desktop computer.

Nielsen *et al*. 2012. PloS One 10.1371

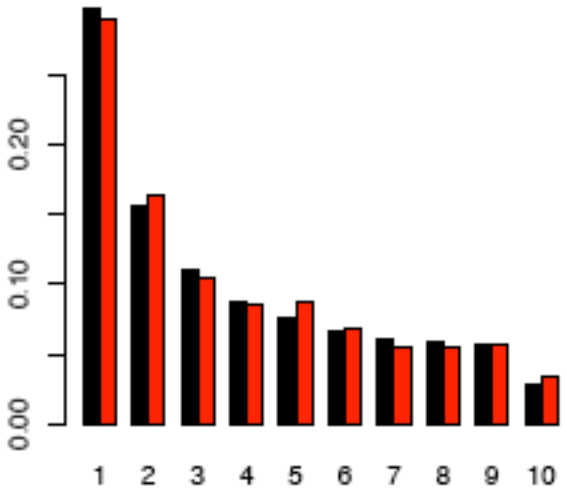
5X



The distribution of true and estimated folded SFS in a sample from 50 MB 10 diploid individuals, where 1% of all SNPs are variable in the population and follow a distribution of allele frequencies, *p*, proportional to 1/*p*. An error rate of 0.5% is assumed. The mean sequencing depths is 5X.

Nielsen *et al*. 2012. PloS One 10.1371

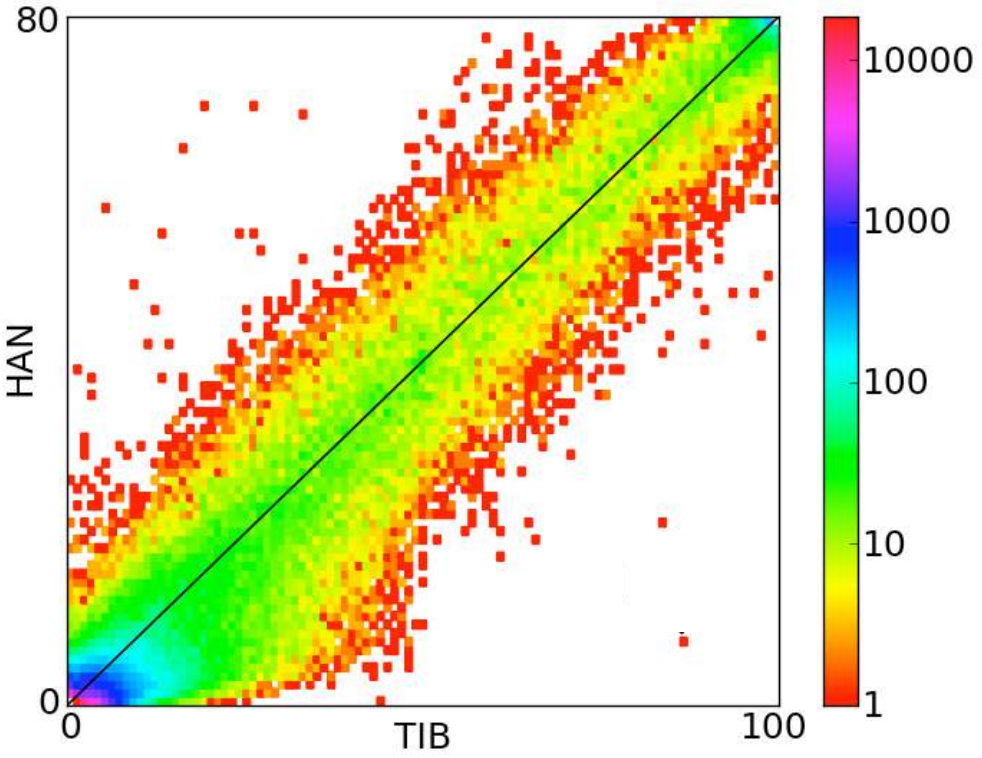
1X



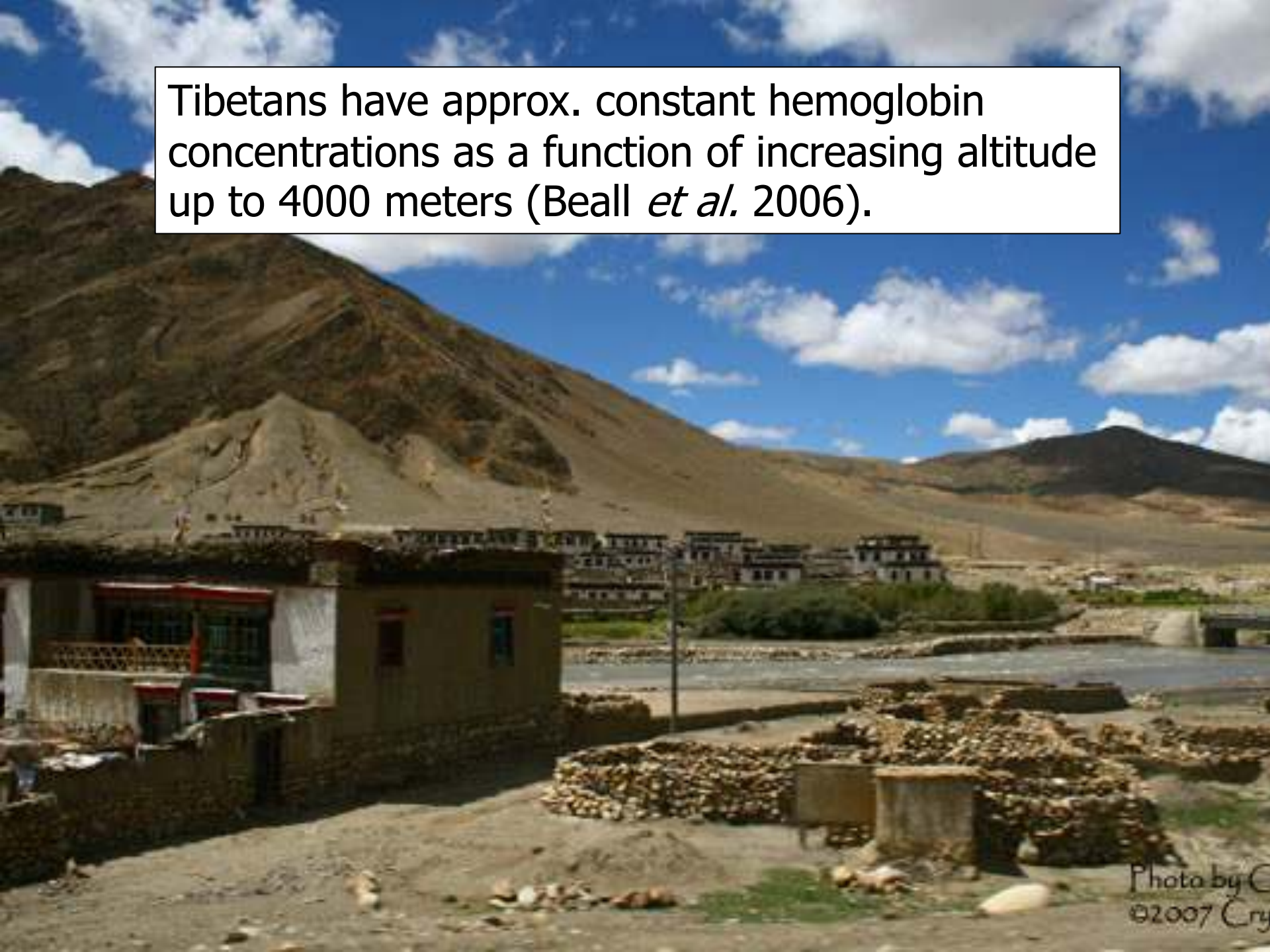
The distribution of true and estimated folded SFS in a sample from 50 MB 10 diploid individuals, where 1% of all SNPs are variable in the population and follow a distribution of allele frequencies, *p*, proportional to 1/*p*. An error rate of 0.5% is assumed. The mean sequencing depths is 1X

Nielsen *et al*. 2012. PloS One 10.1371

50 Tibetan Exomes

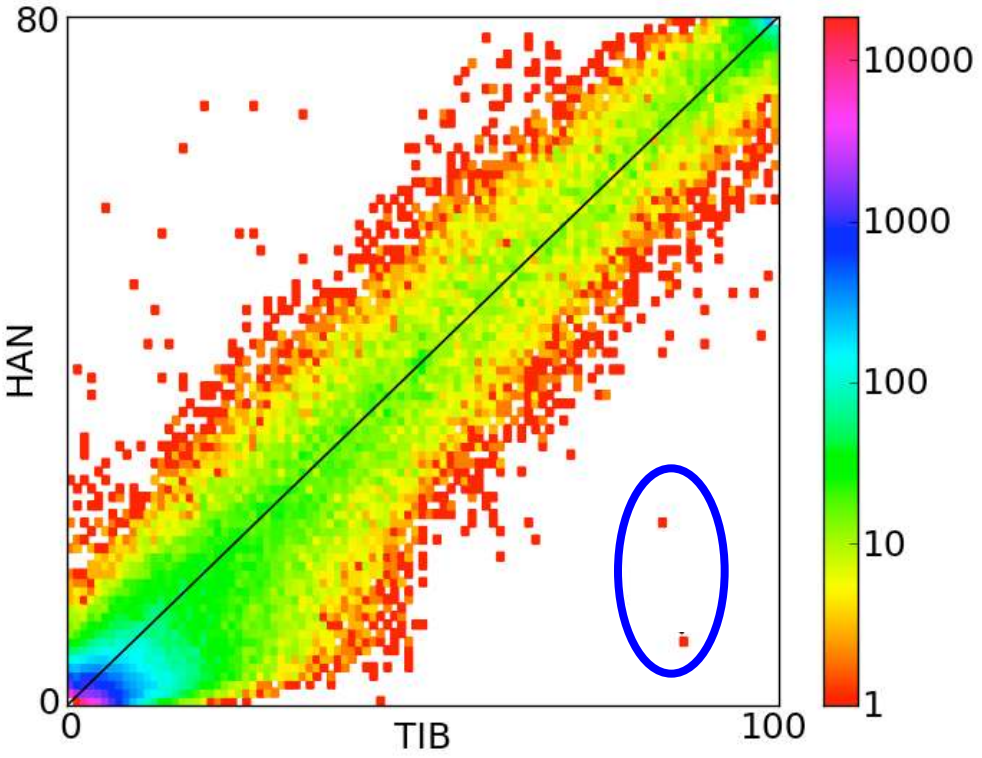


Yi, X. *et al.* 2010. *Science* 329: 75-78.



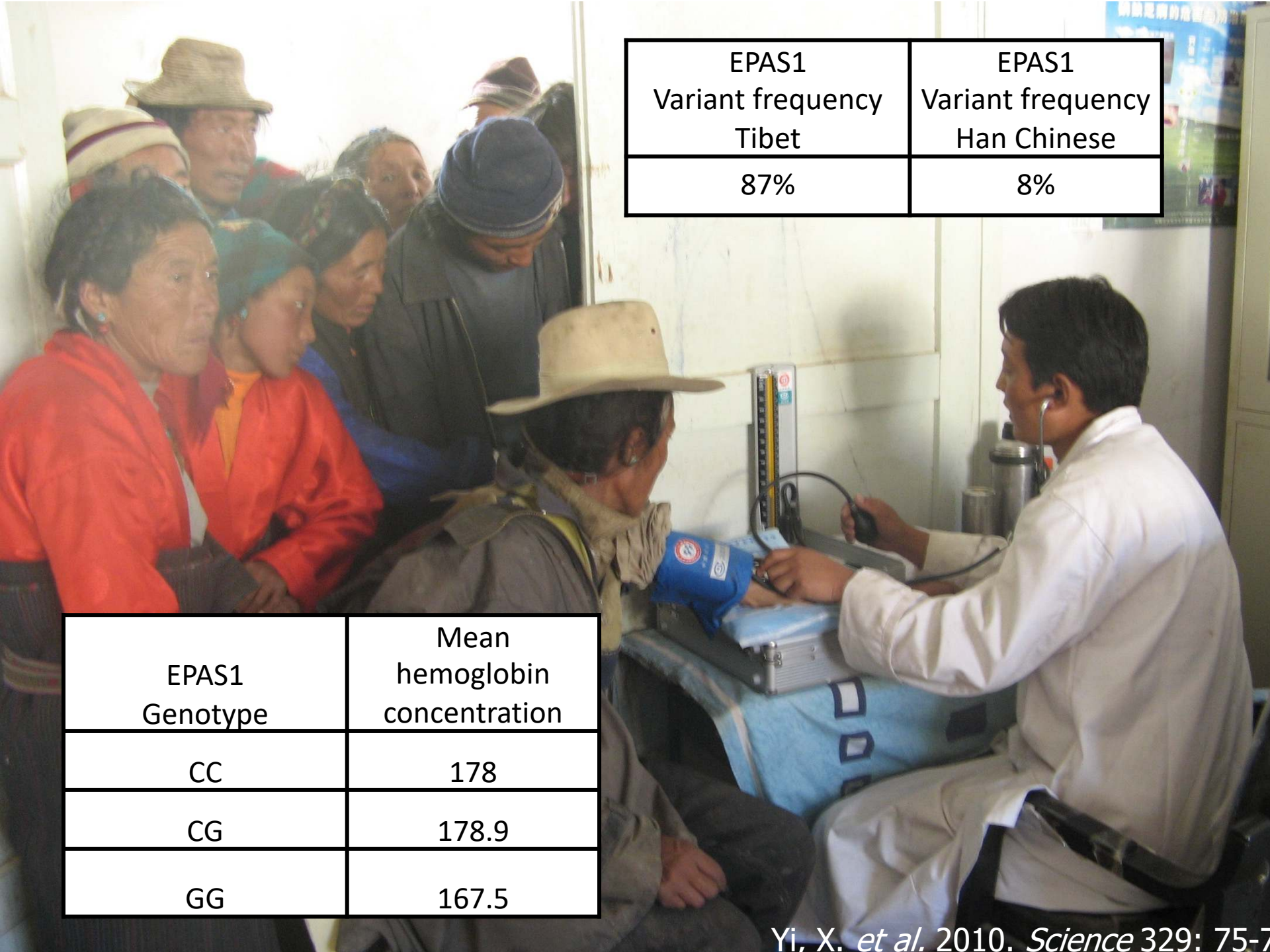
Tibetans have approx. constant hemoglobin concentrations as a function of increasing altitude up to 4000 meters (Beall *et al.* 2006).

50 Tibetan Exomes



EPAS1

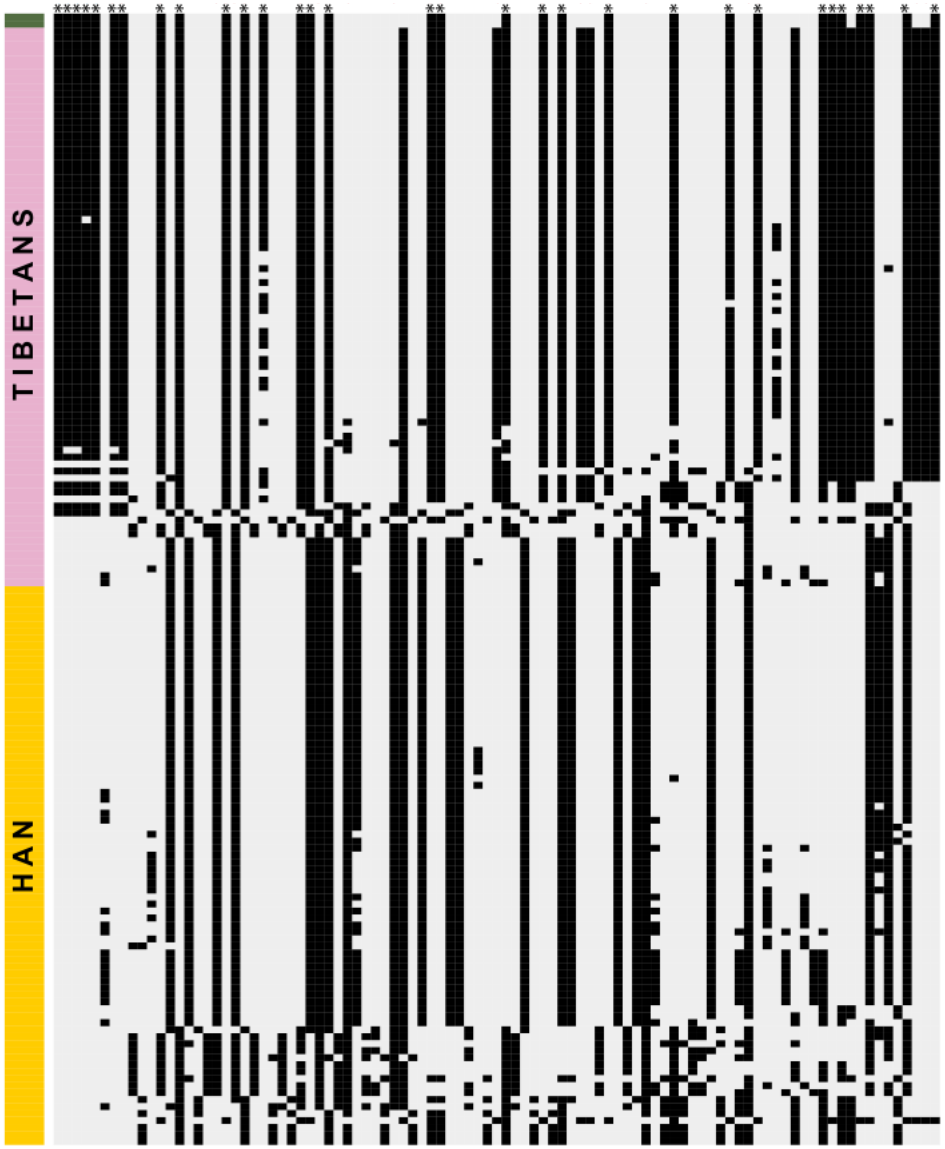
Yi, X. *et al.* 2010. *Science* 329: 75-78.



EPAS1 Variant frequency Tibet	EPAS1 Variant frequency Han Chinese
87%	8%

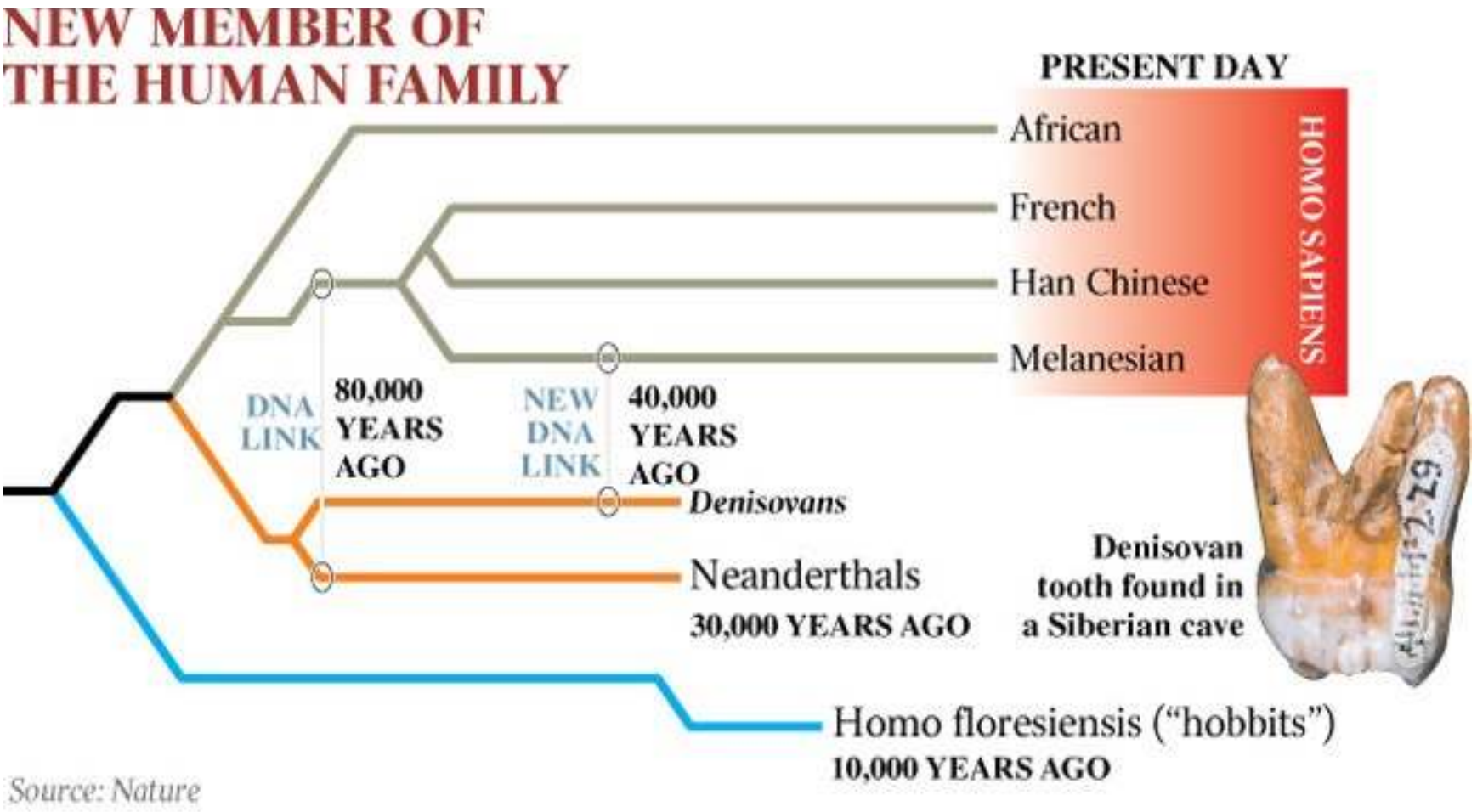
EPAS1 Genotype	Mean hemoglobin concentration
CC	178
CG	178.9
GG	167.5

Yi, X. *et al.* 2010. *Science* 329: 75-7

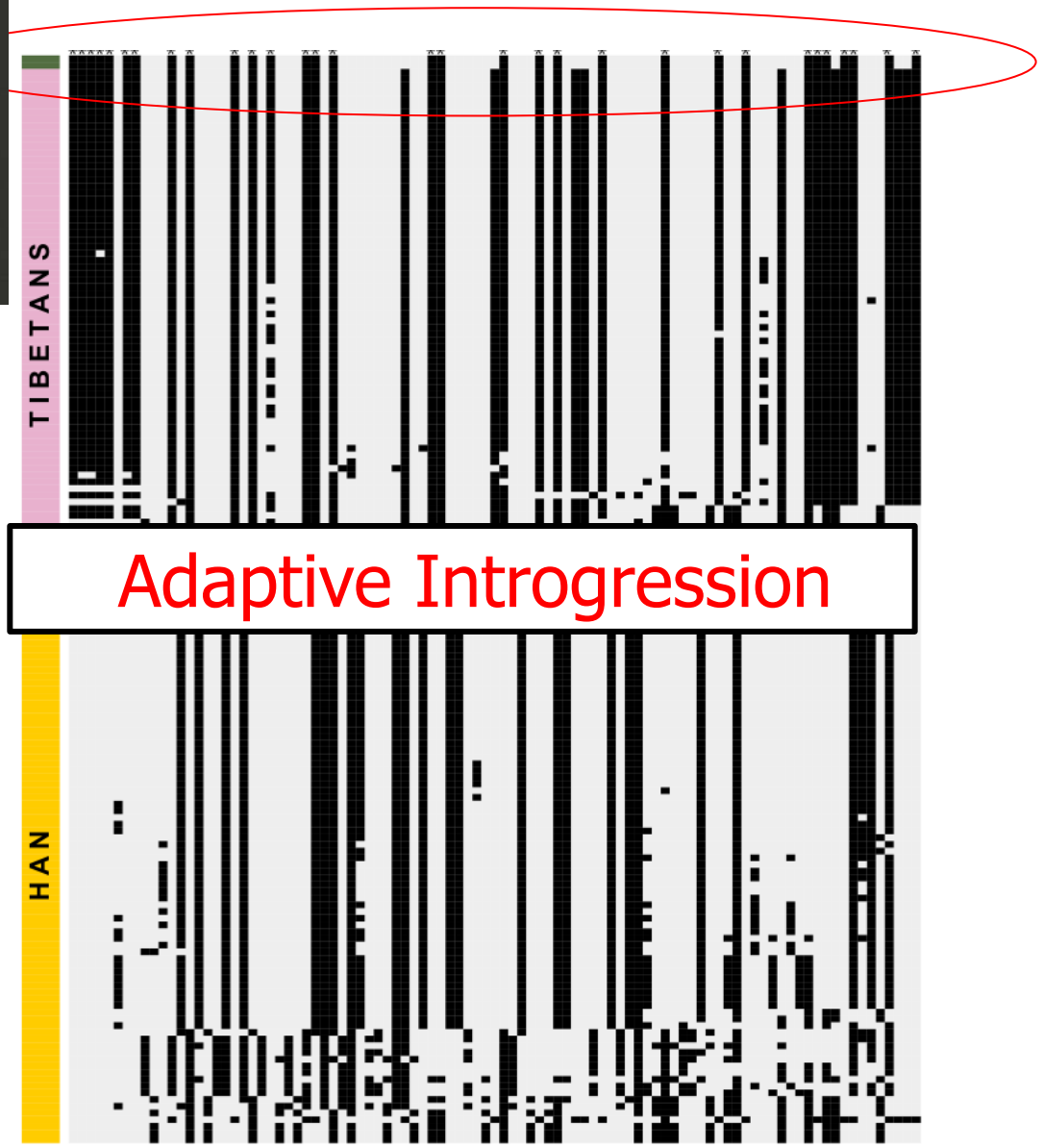


Huerta Sanchez *et al.* 2014. *Nature* 512: 194–197

Denisovans

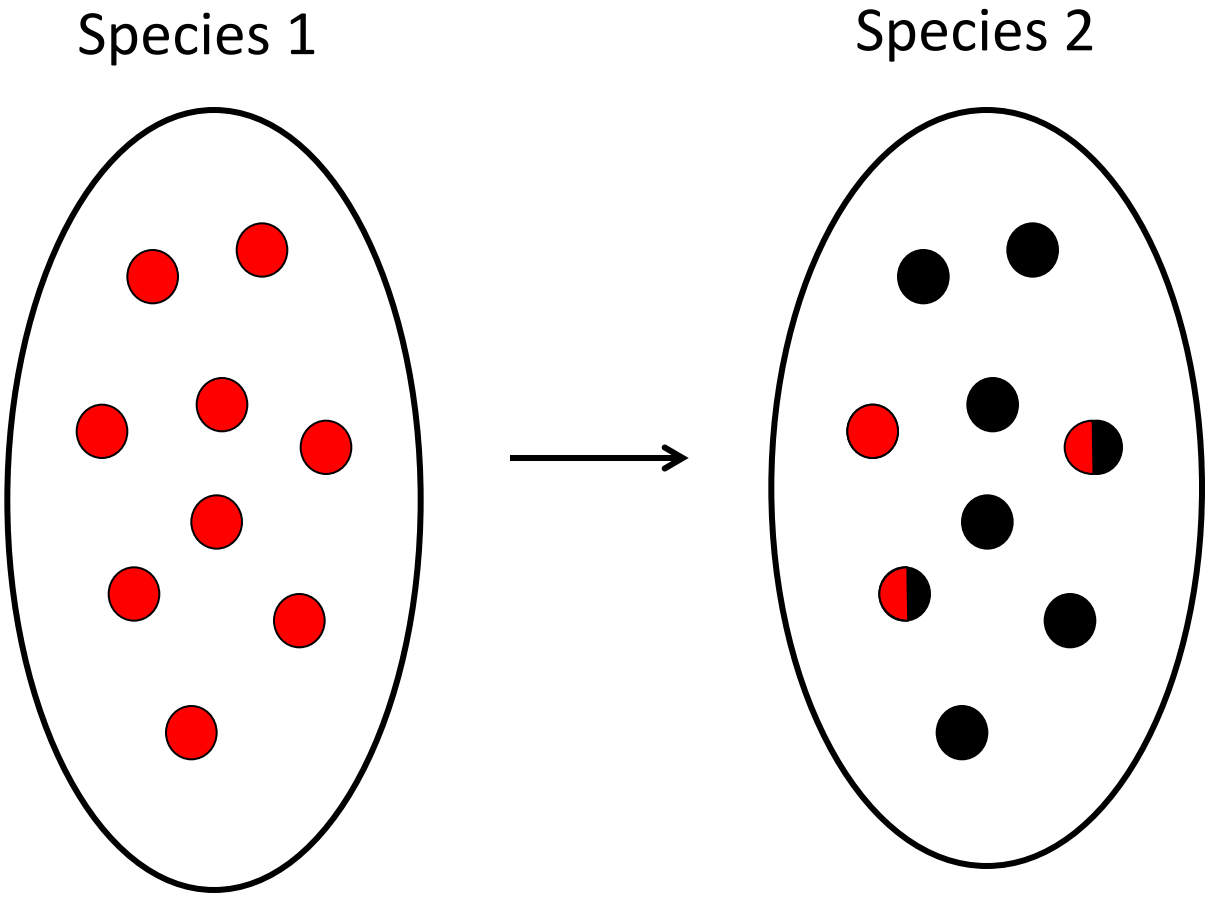

NEW MEMBER OF THE HUMAN FAMILY
PRESENT DAY
African
French
Han Chinese
Melanesian
HOMO SAPIENS
DNA LINK
80,000 YEARS AGO
NEW DNA LINK
40,000 YEARS AGO
Denisovans
Neanderthals
30,000 YEARS AGO
Denisovan tooth found in a Siberian cave
Homo floresiensis ("hobbits")
10,000 YEARS AGO
Source: Nature

Denisovan!

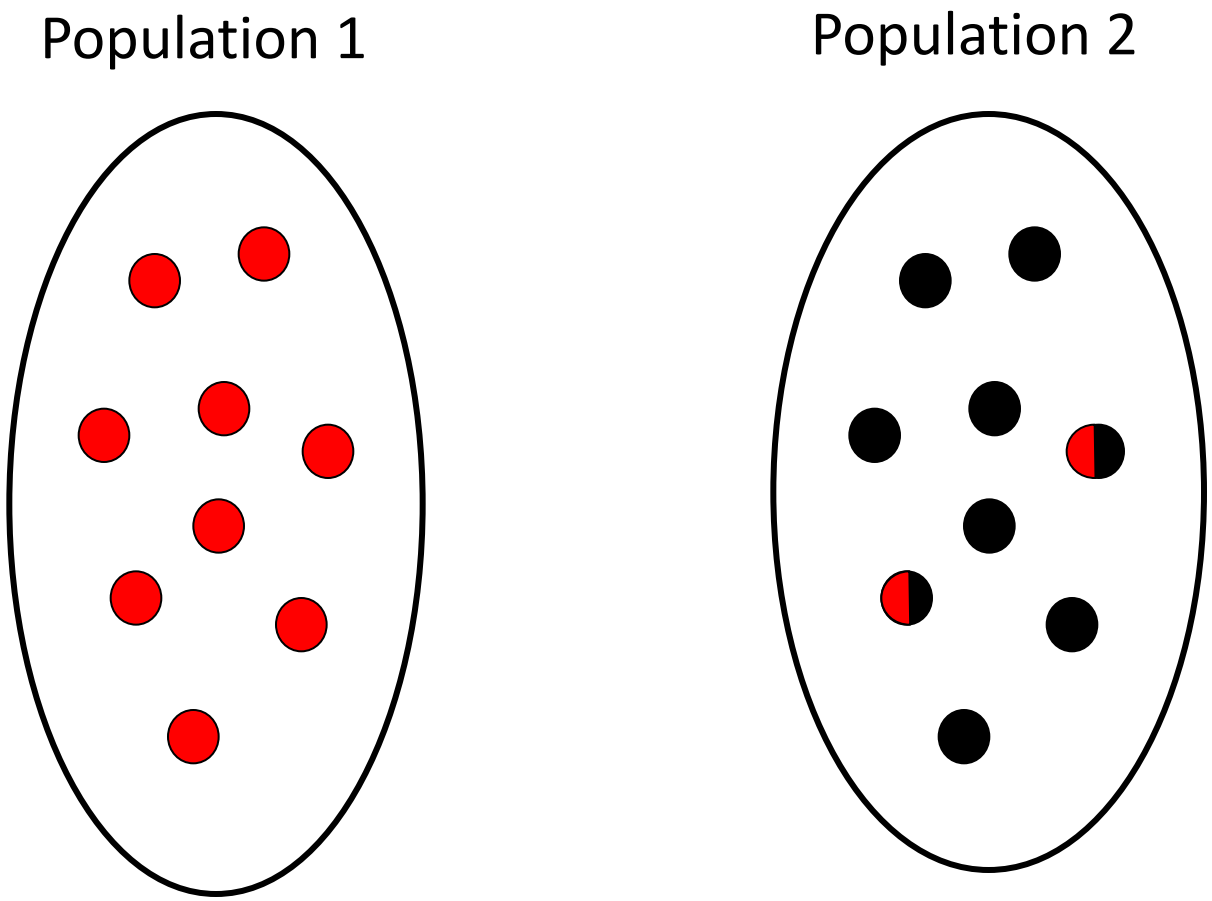


Huerta Sanchez *et al.* 2014. *Nature* 512: 194–197

Introgression



Admixture



Assignment and Admixture

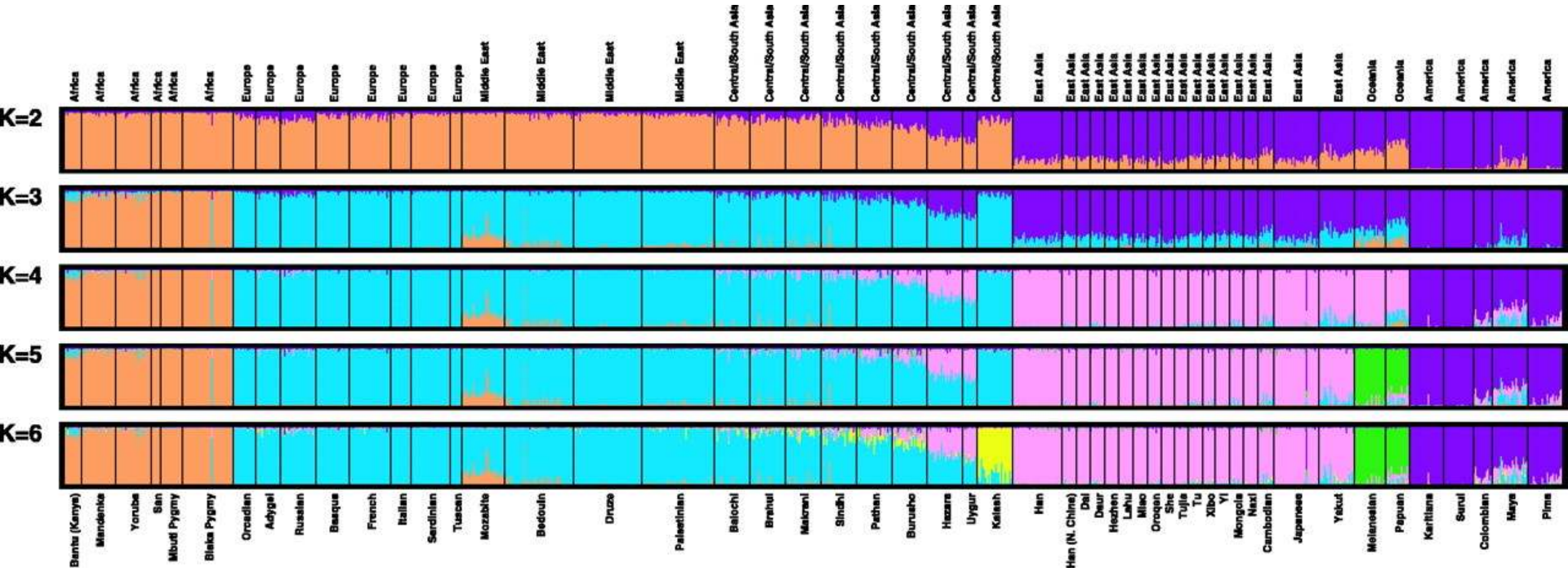
- Multi loci data?

-To which population does an individual belong?

- How many groupings should we choose?

Structure analysis of 1056 individuals from 52 populations for 377 microsatelite loci.

Rosenberg et al. (2002)



Transition to the Neolithic

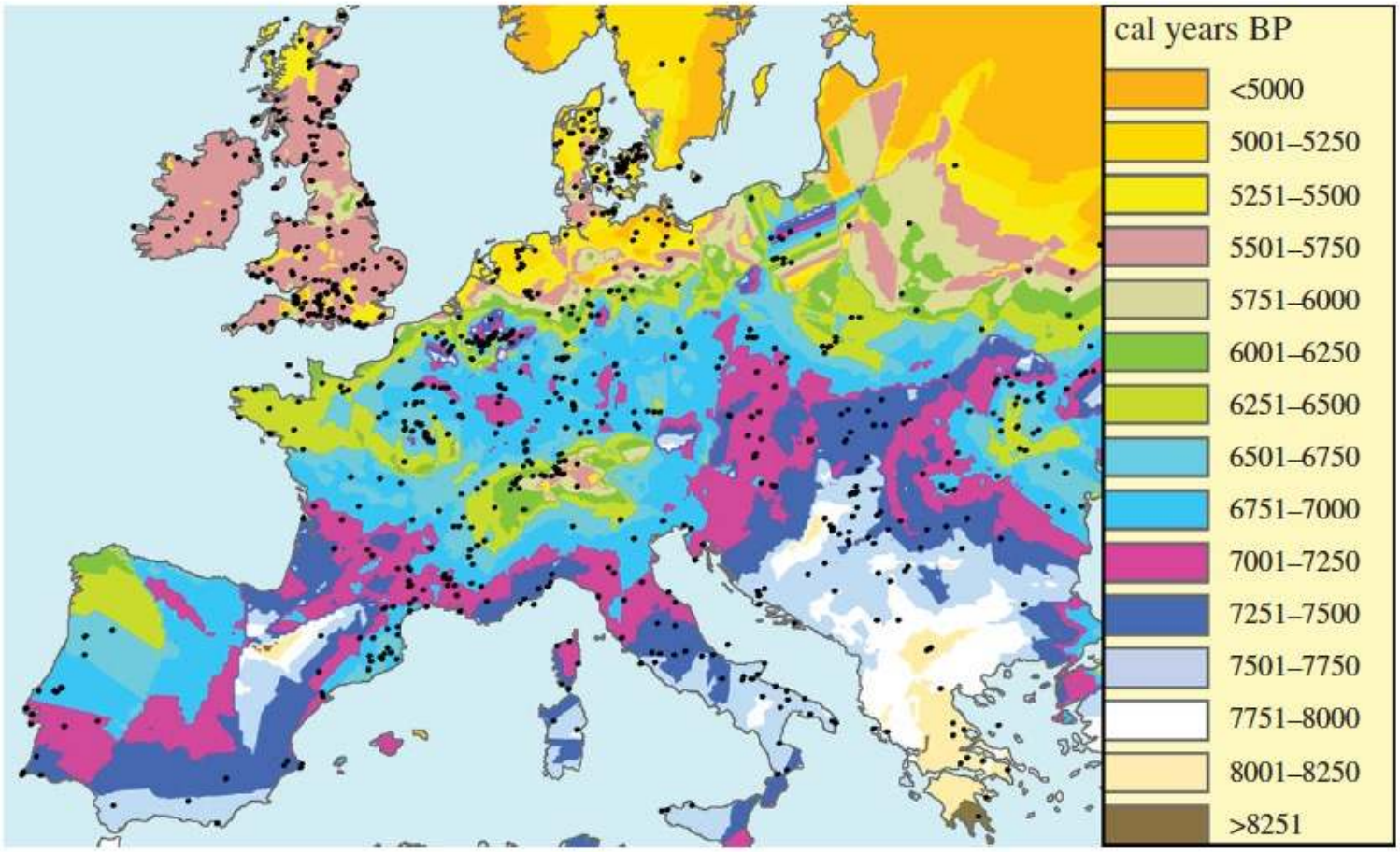
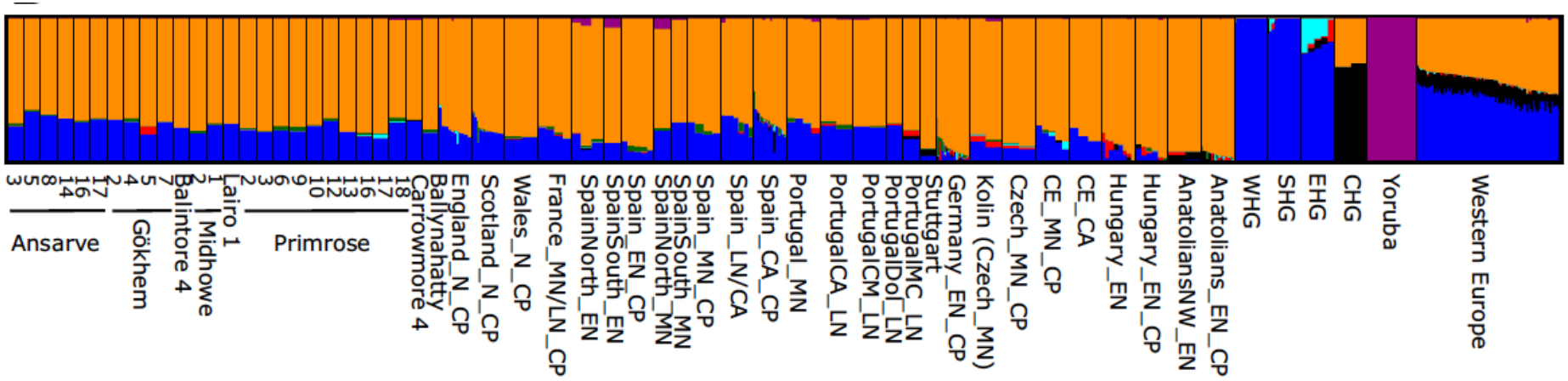
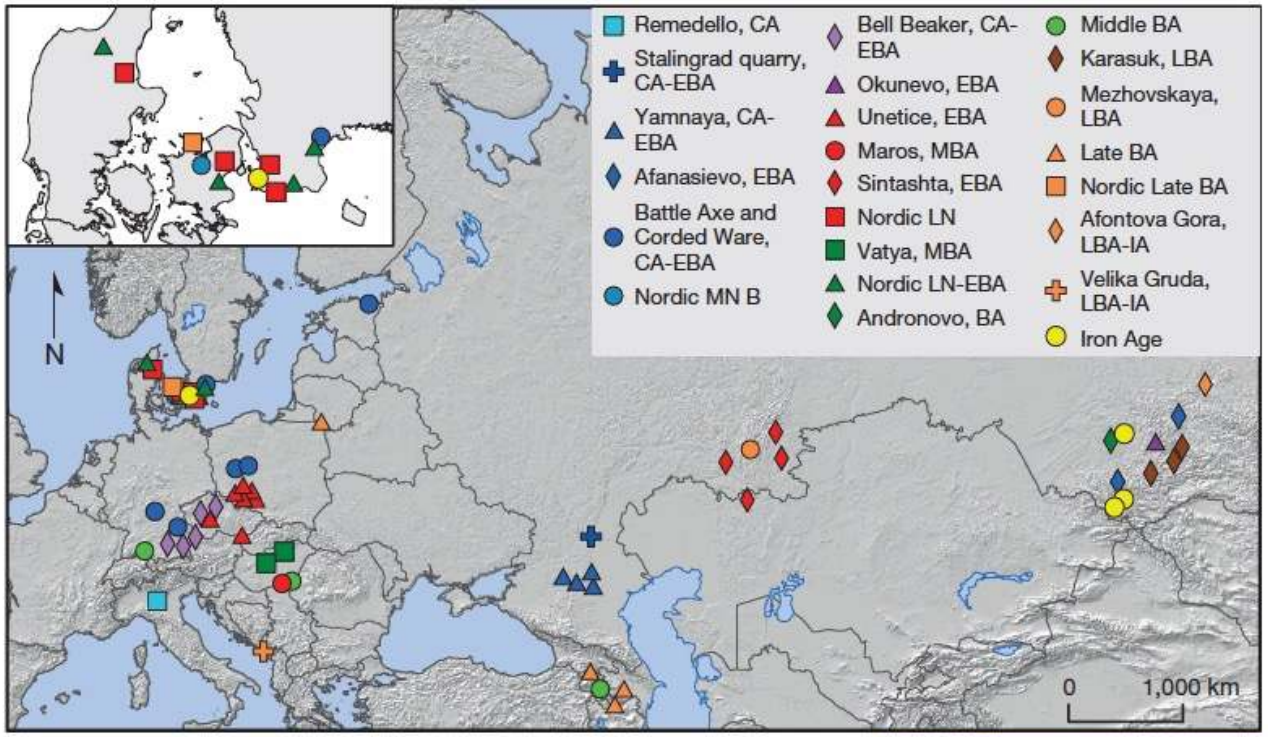
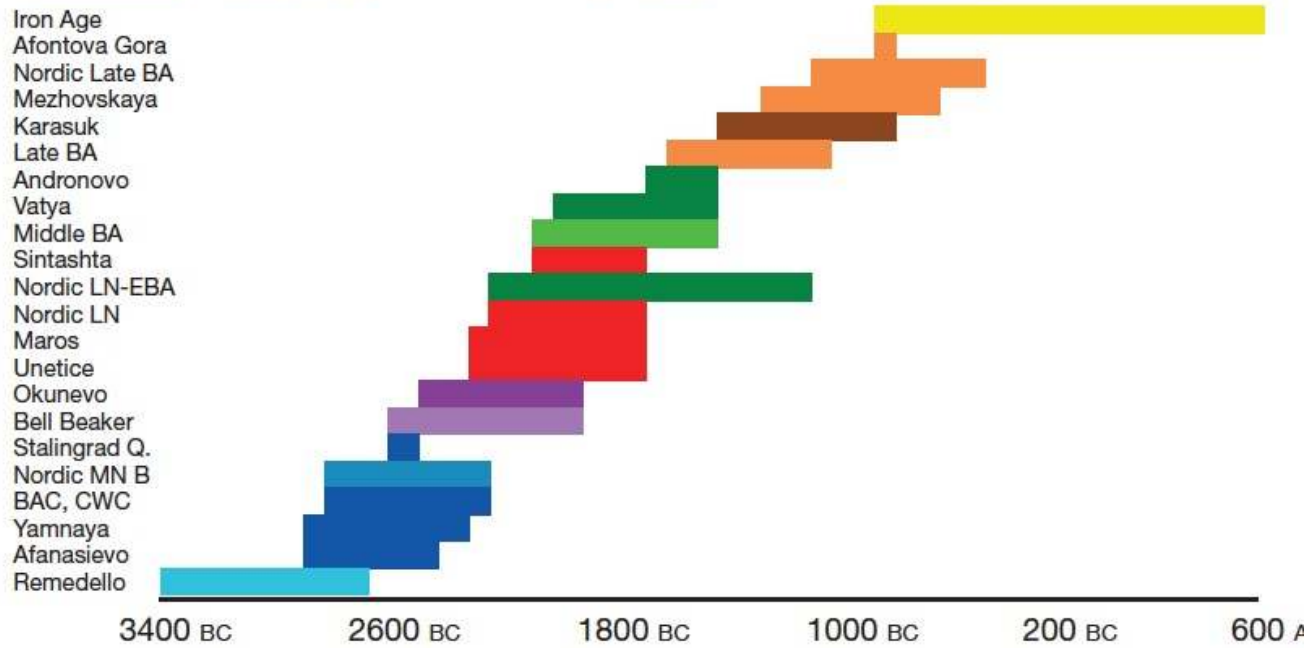


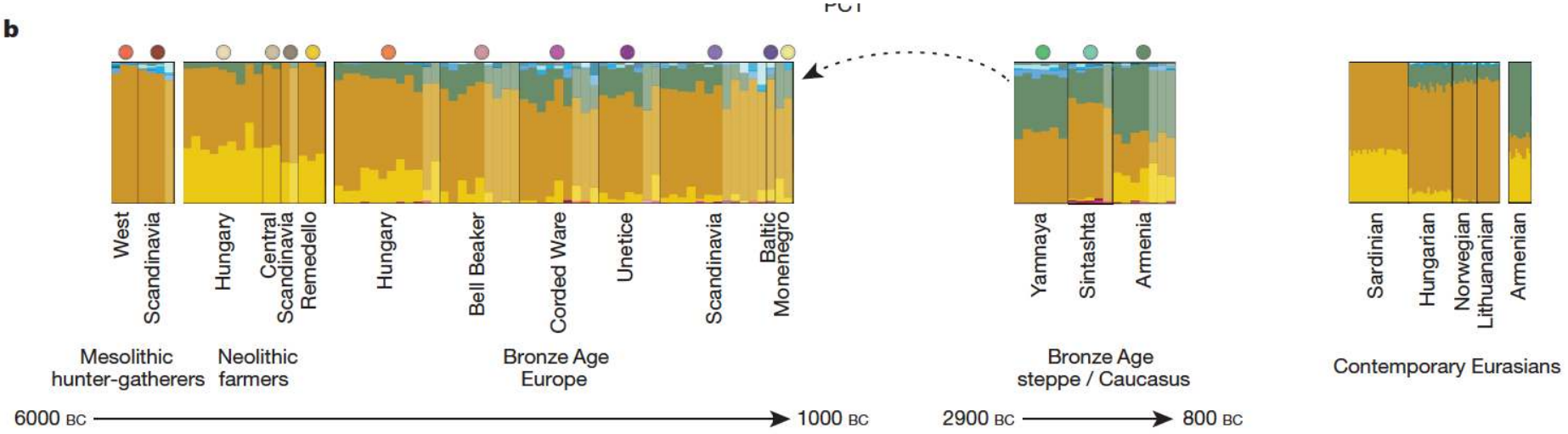
Figure 1. Isochrone map. The spread of the Neolithic transition, obtained by interpolating the dates in calibrated years before present (BP) of 918 Early Neolithic sites (circles) in Europe and the Near East (see the electronic supplementary material for details on the dataset and interpolation). Map created with ArcGIS 10.

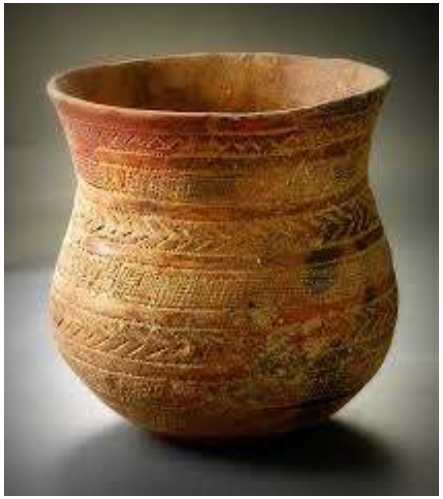
Transition to the Neolithic

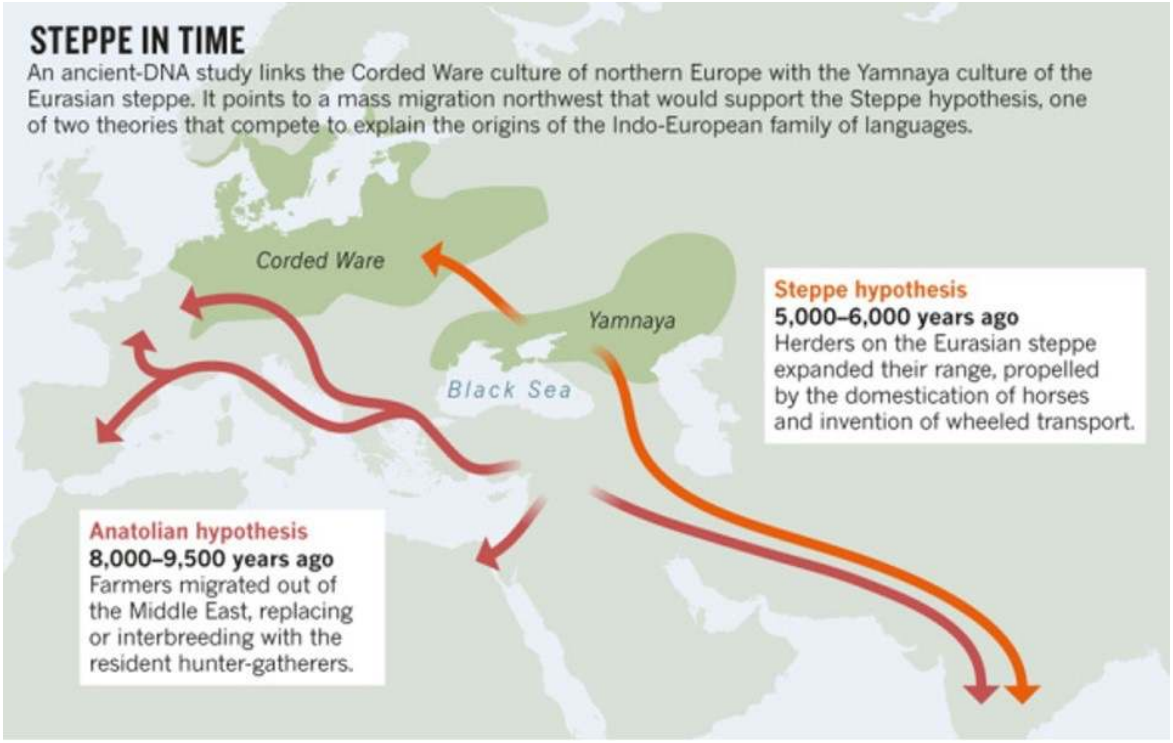



Remedello, CA
Stalingrad quarry, CA-EBA
Yamnaya, CA-EBA
Afanasievo, EBA
Battle Axe and Corded Ware, CA-EBA
Nordic MN B
Bell Beaker, CA-EBA
Okunevo, EBA
Unetice, EBA
Maros, MBA
Sintashta, EBA
Nordic LN
Vatya, MBA
Nordic LN-EBA
Andronovo, BA
Middle BA
Karasuk, LBA
Mezhovskaya, LBA
Late BA
Nordic Late BA
Afontova Gora, LBA-IA
Velika Gruda, LBA-IA
Iron Age
N
0
1,000 km

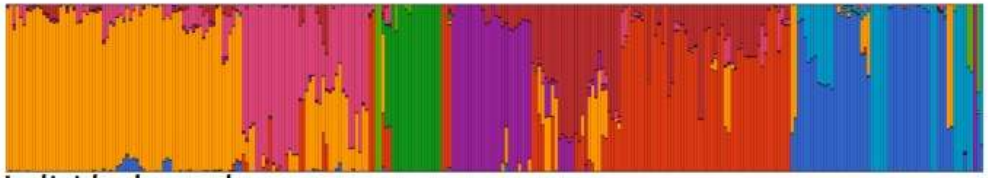

Iron Age
Afontova Gora
Nordic Late BA
Mezhovskaya
Karasuk
Late BA
Andronovo
Vatya
Middle BA
Sintashta
Nordic LN-EBA
Nordic LN
Maros
Unetice
Okunevo
Bell Beaker
Stalingrad Q.
Nordic MN B
BAC, CWC
Yamnaya
Afanasievo
Remedello
3400 BC
2600 BC
1800 BC
1000 BC
200 BC
600 A


b
PC1
West
Scandinavia
Hungary
Central Scandinavia
Remedello
Hungary
Bell Beaker
Corded Ware
Unetice
Scandinavia
Baltic
Monenegro
Yamnaya
Sintashta
Armenia
Sardinian
Hungarian
Norwegian
Lithuananian
Armenian
Mesolithic hunter-gatherers
Neolithic farmers
Bronze Age Europe
Bronze Age steppe / Caucasus
Contemporary Eurasians
6000 BC
1000 BC
2900 BC
800 BC

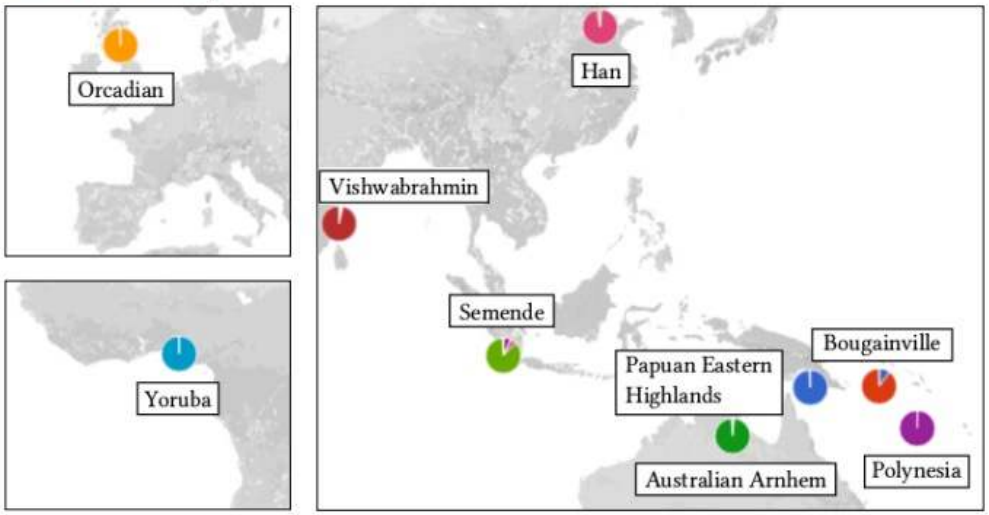




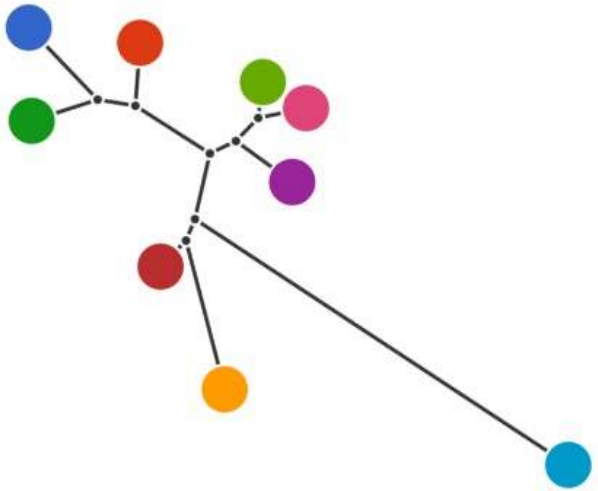
From Haak *et al.* (2015)



Individual samples ...

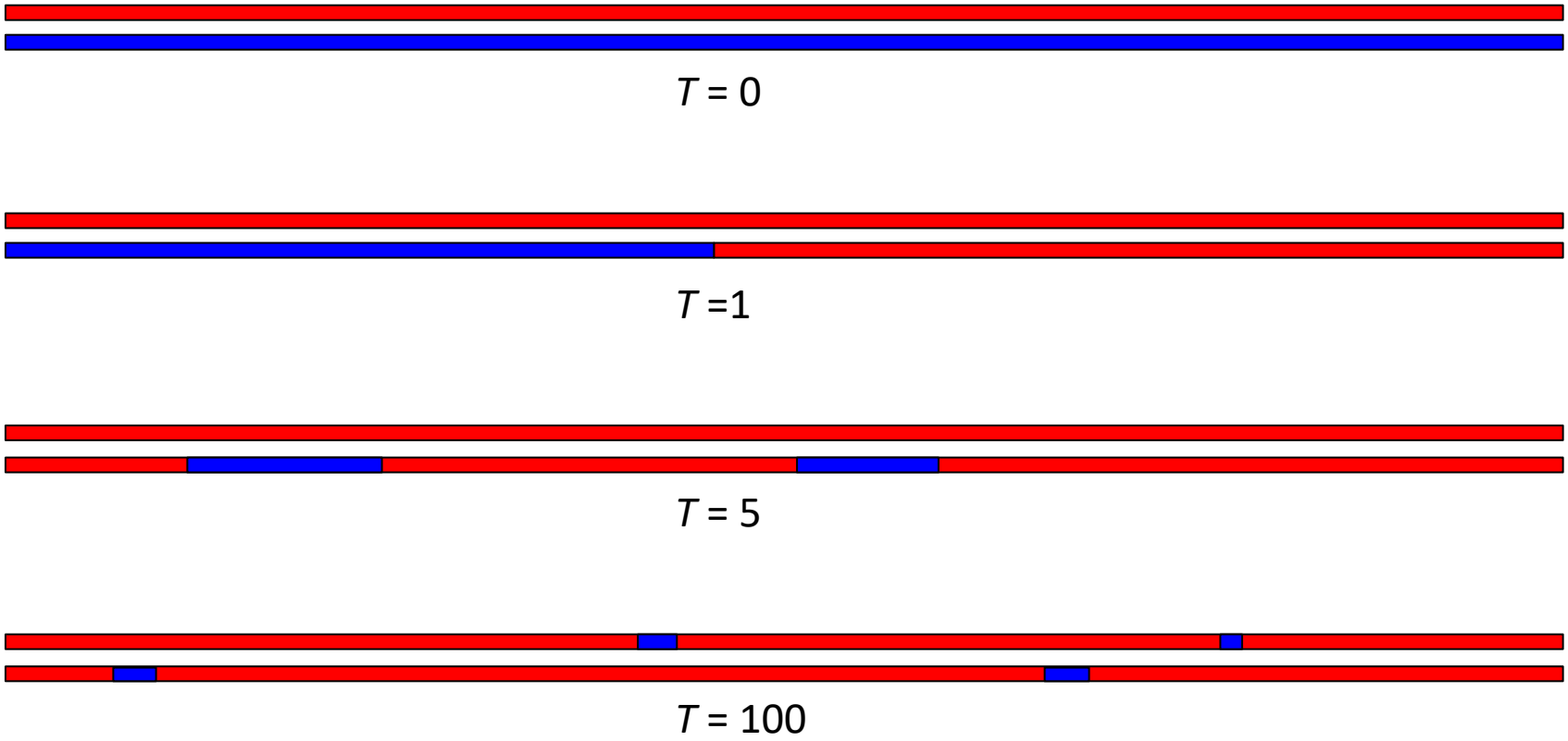


Map data © 2016 Google, INEGI · Phylogenetic trees: *http://www.jade-cheng.com/graphs/*



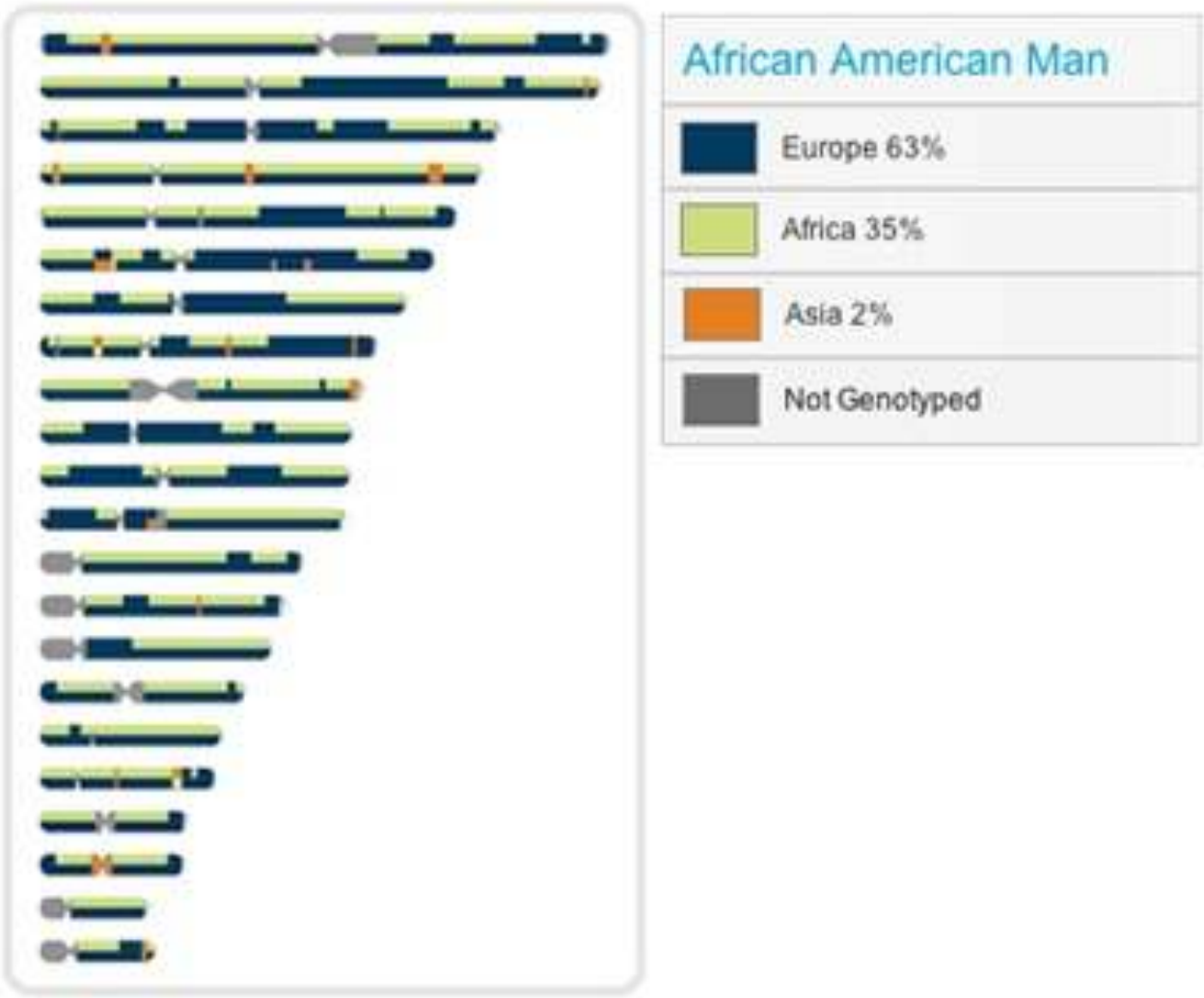
Jade Cheng

Admixture tracts

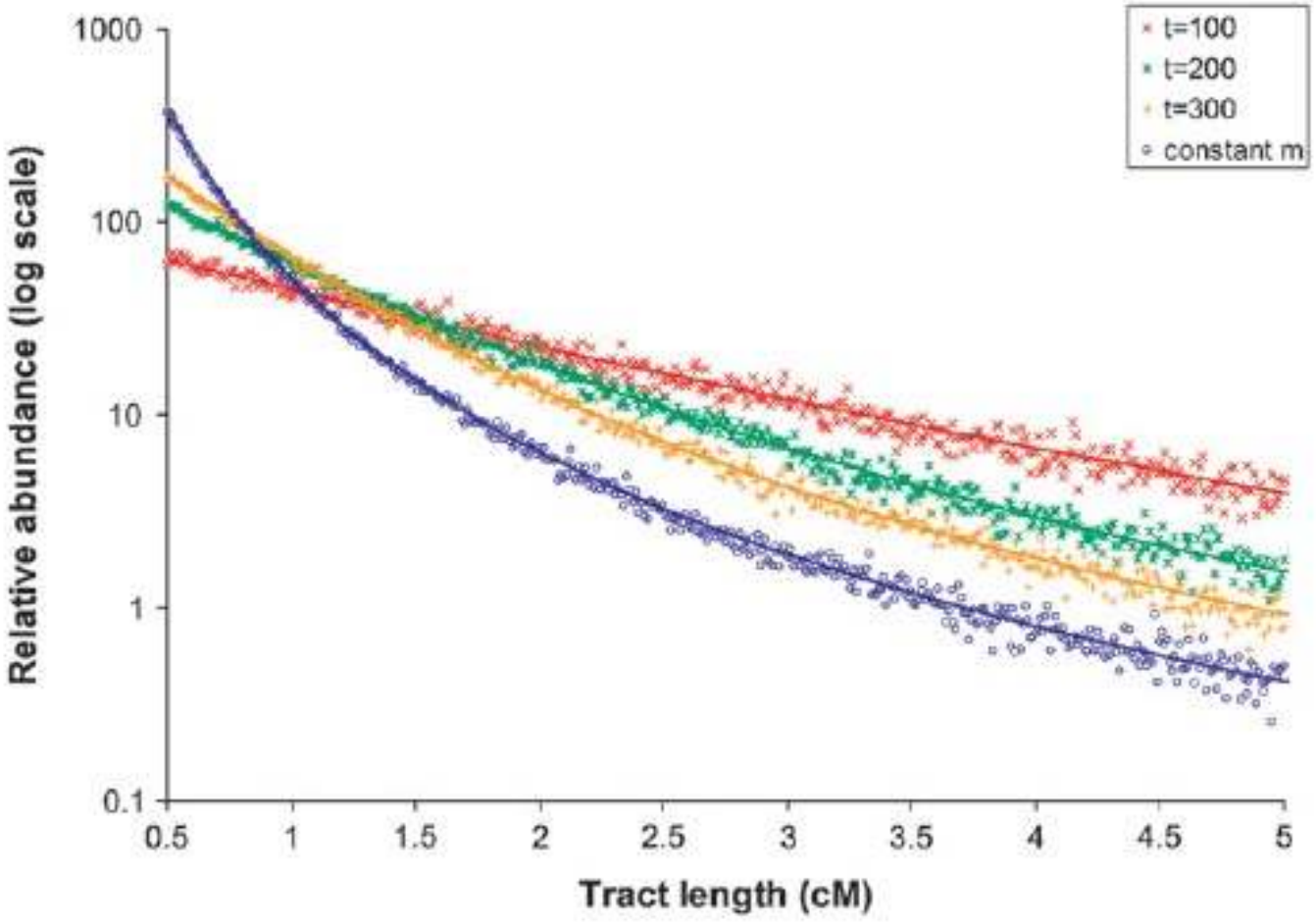


Recombination Clock!

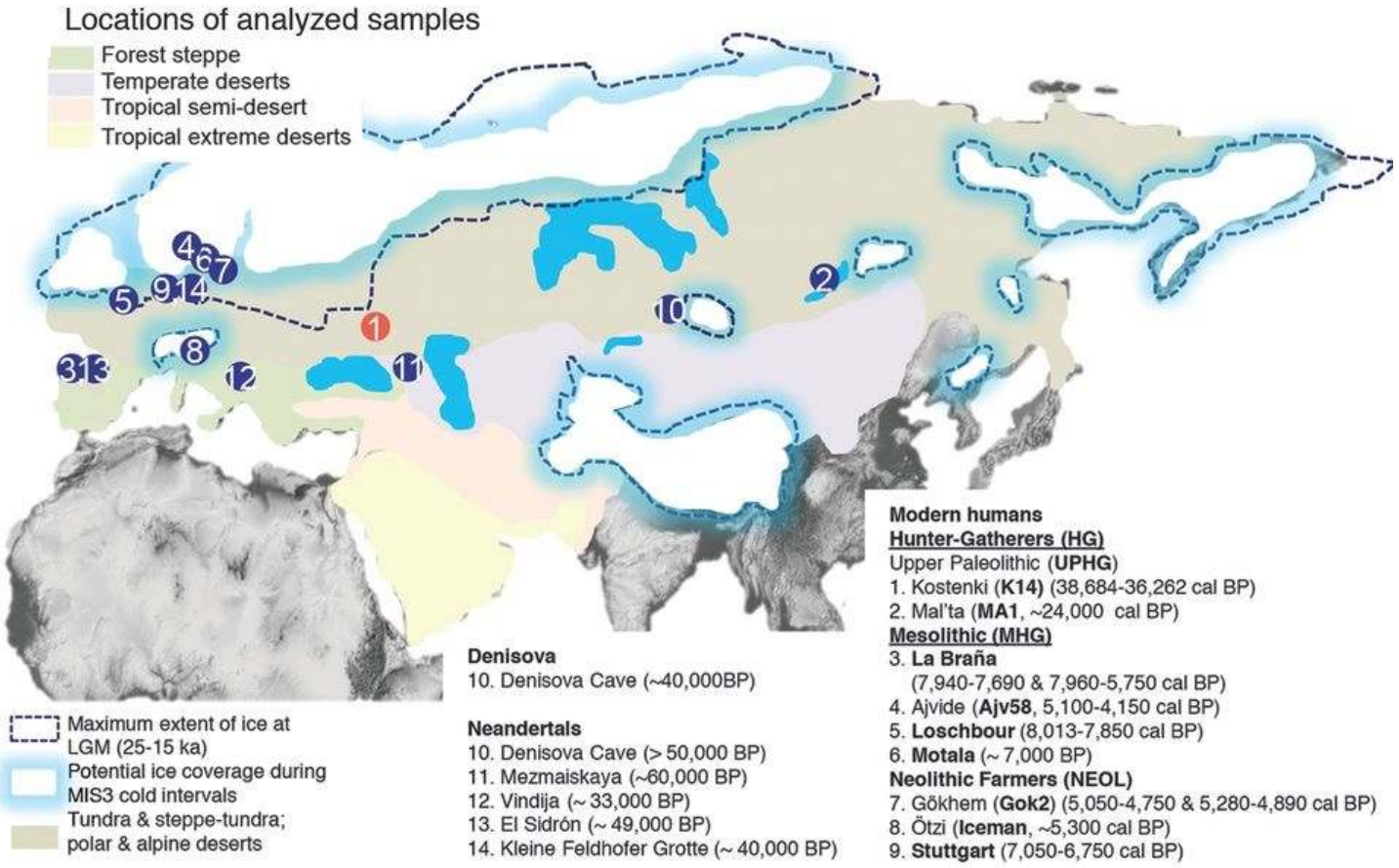
Ancestry painting (23andme.com)



Predictions assuming independence

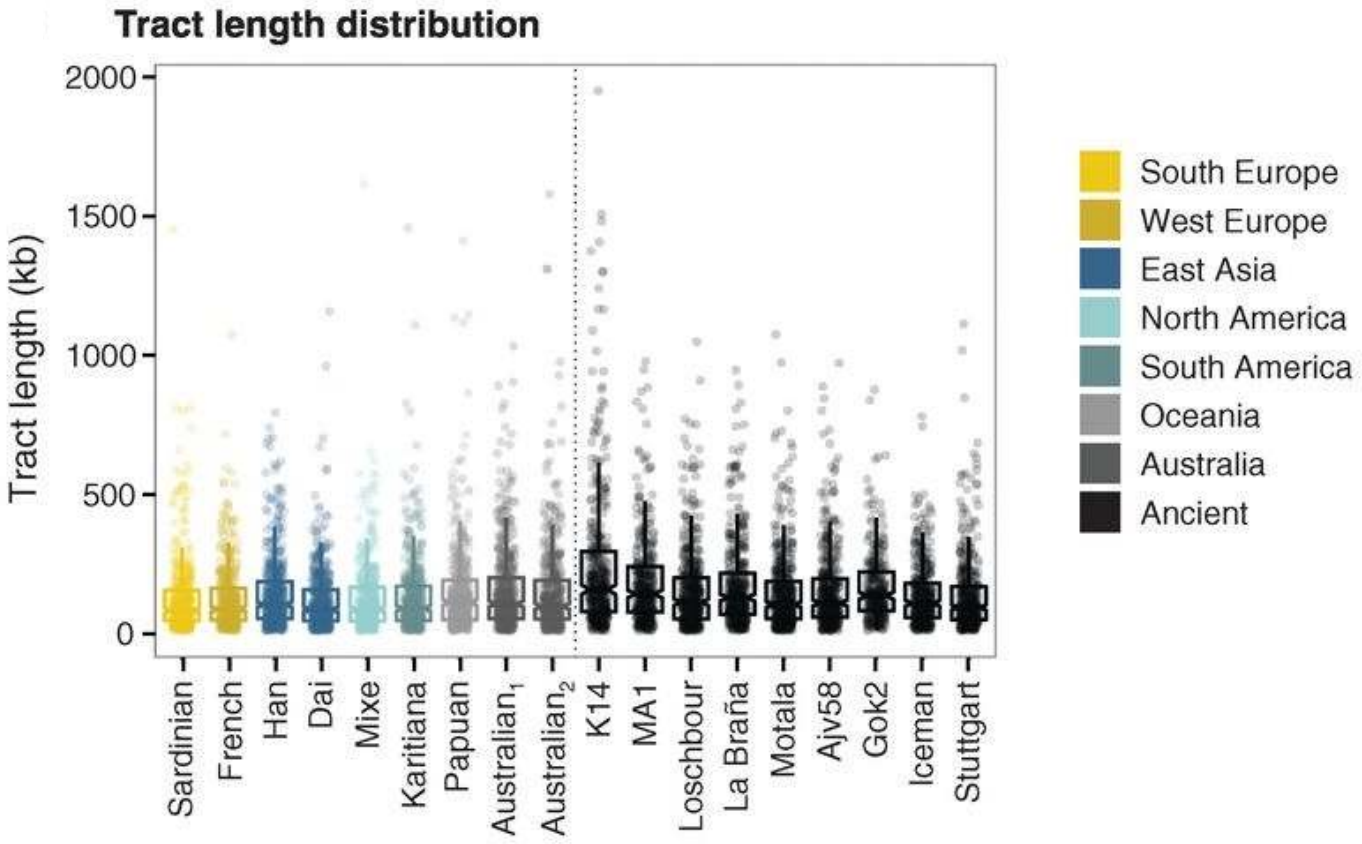


Kostenki



Andaine Seguin-Orlando *et al.* 2014. Science: 346: 1113-1118

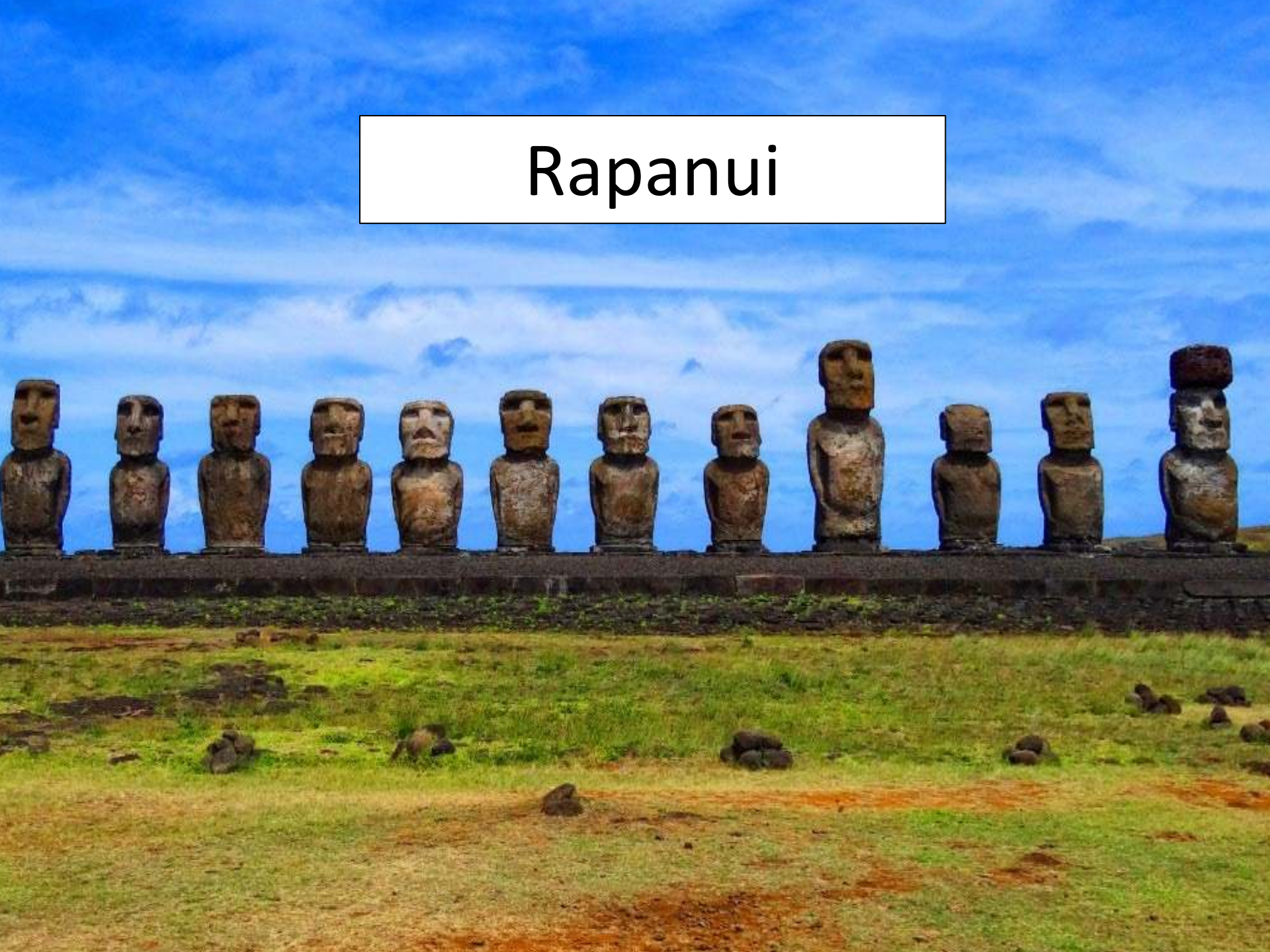
Kostenki



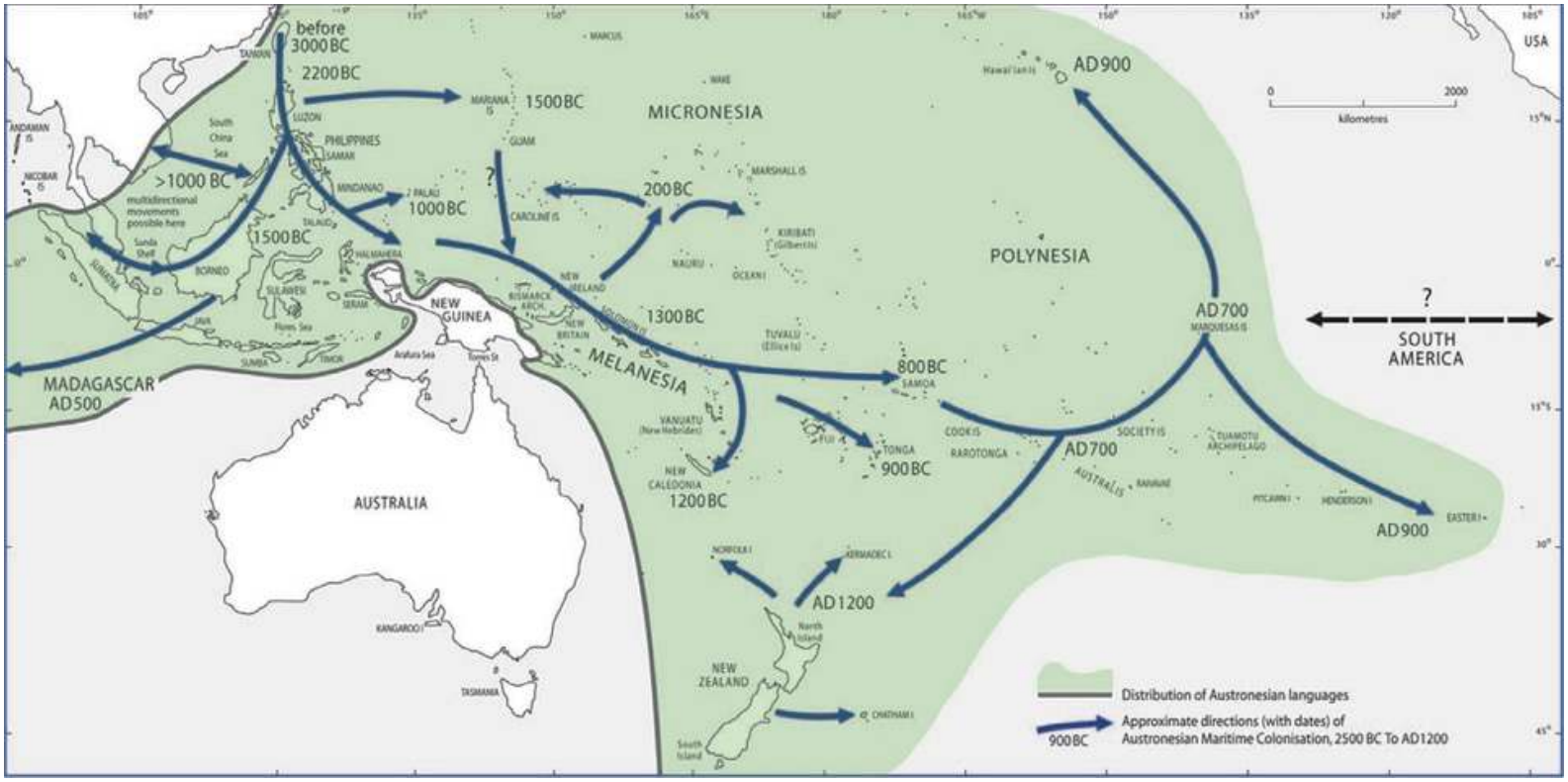
Age ~54,000 years

Andaine Seguin-Orlando *et al.* 2014. Science: 346: 1113-1118

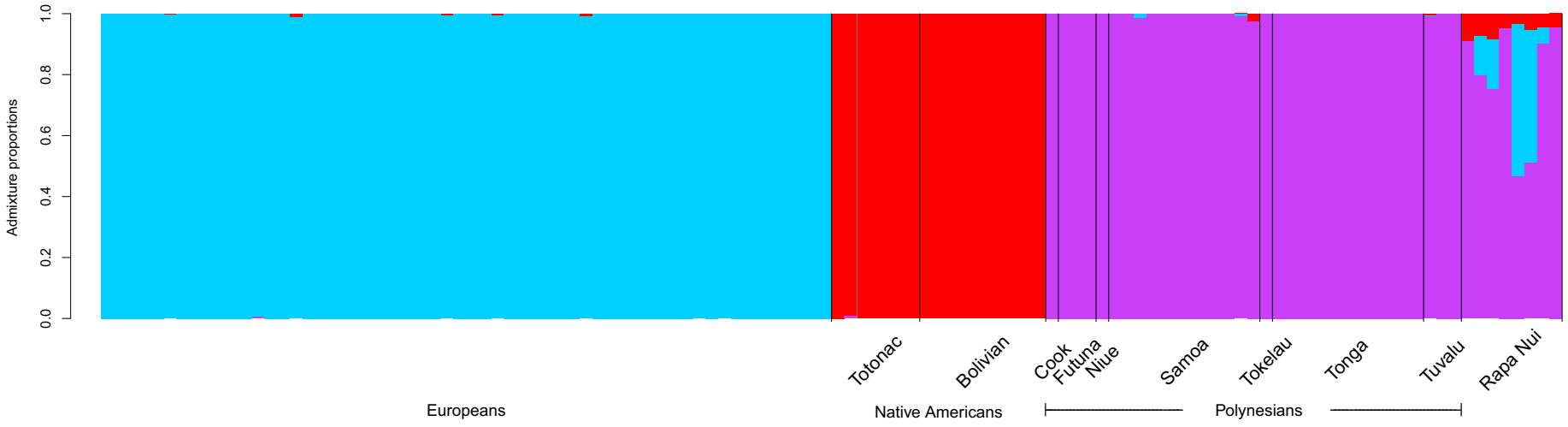
Rapanui



Polynesian Expansion

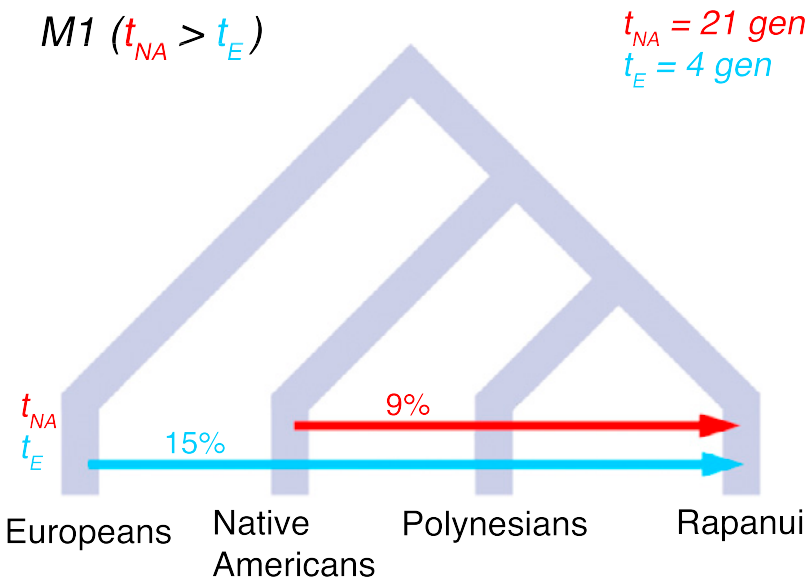


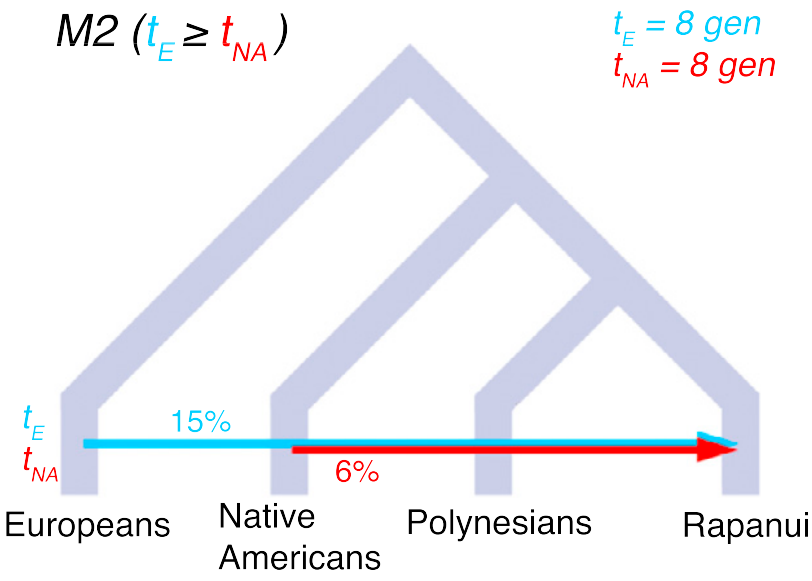
Rapanui



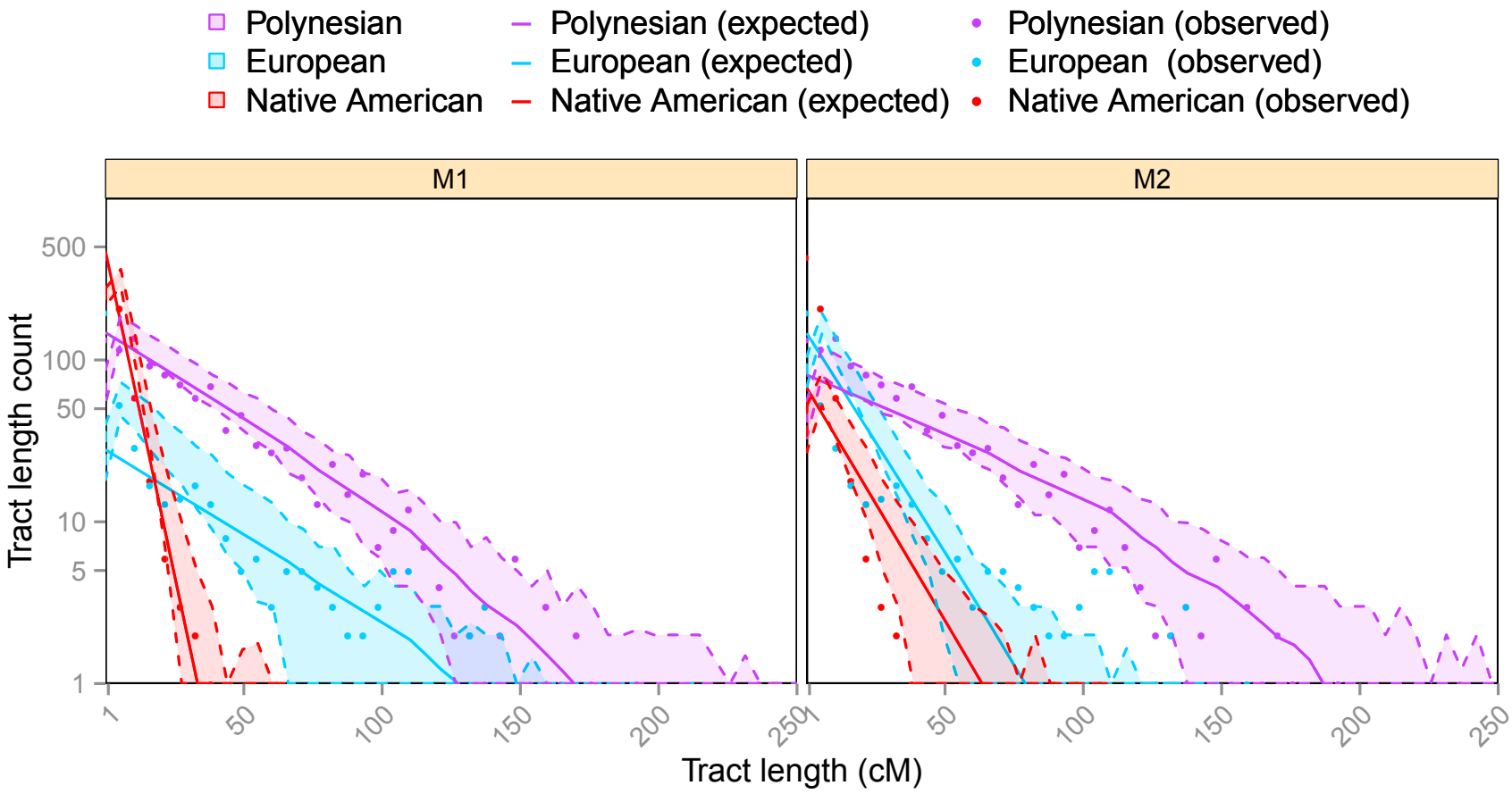
Moreno-Mayar et al. 2014. Current Biology 24: 2518–2525.

Rapanui



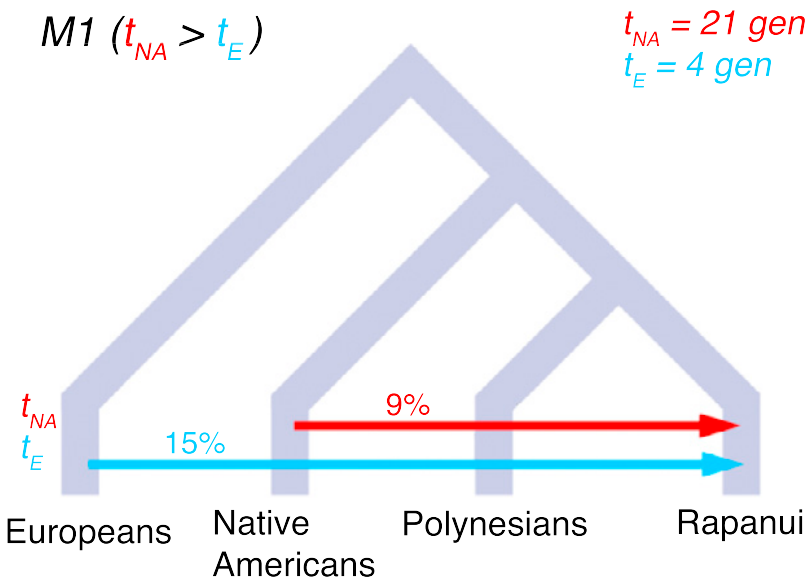


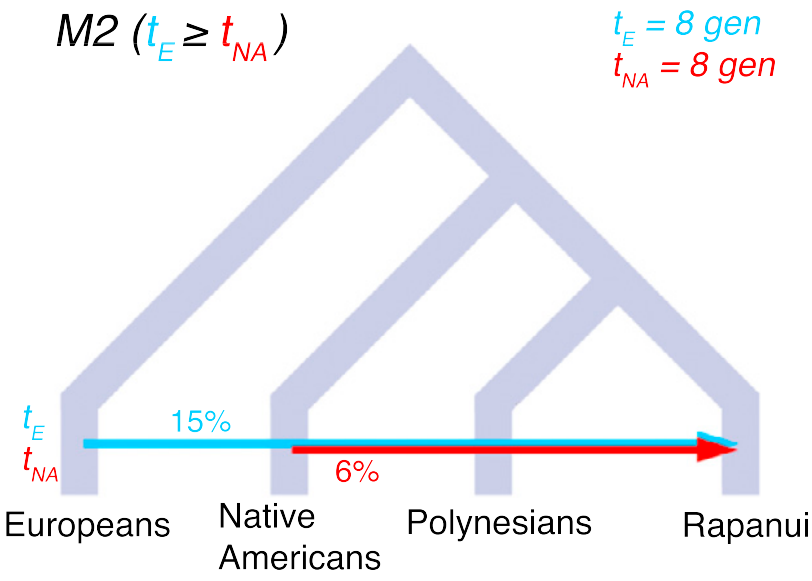
Moreno-Mayar et al. 2014. Current Biology 24: 2518–2525.



Moreno-Mayar et al. 2014. Current Biology 24: 2518–2525.

Rapanui





Moreno-Mayar et al. 2014. Current Biology 24: 2518–2525.