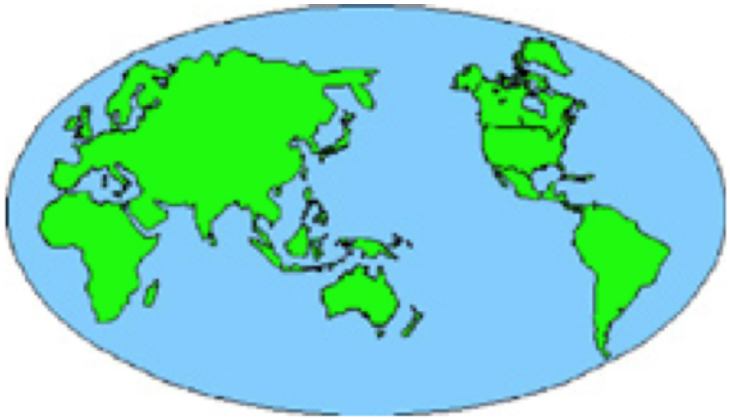
Computational Population Genomics

Rasmus Nielsen

Departments of Integrative Biology and Statistics, UC Berkeley

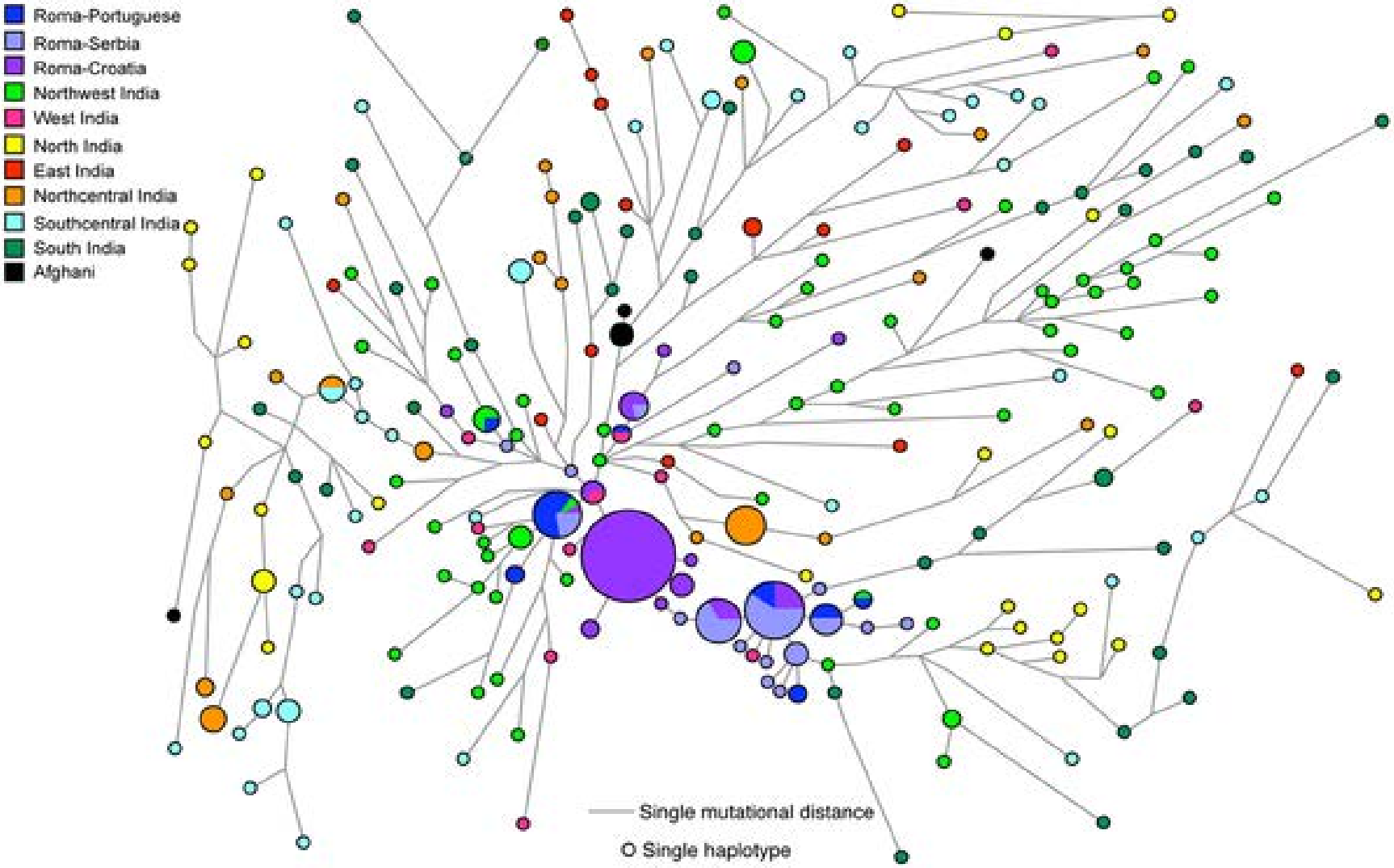
Department of Biology, University of Copenhagen



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

Data: DNA sequences

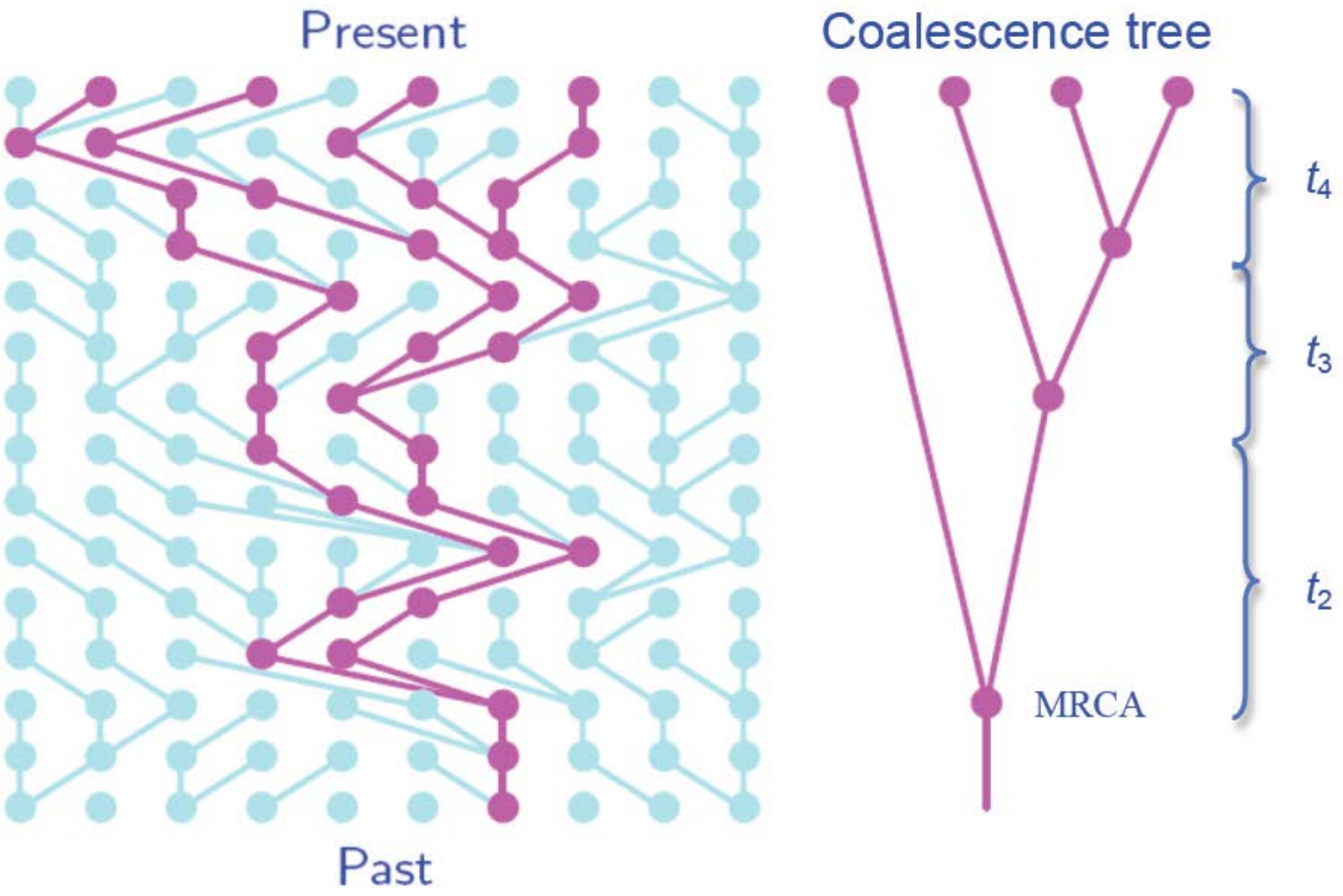
```
Eastern 1 CGAAAAGTATCCATCTCGCAGTGCTGAGCTAGACA
Eastern 2 CGAAAAGTATCCATCTCGCAGTGCTGAGCTAGACA
Eastern 3 CGAAAAGTATCCATCTCGCAGTGCTGAGCTAGACA
Eastern 4 CGAAAGGTATCCATCTCGCAGTGCTGAGTTAGACA
Eastern 5 CAAAAAGTATCCATCTCGCAGTGCTGAGTTAGACA
Eastern 6 CAAAAAGTATCCATCTCGCAGTGCTGAGTTAGACA
Eastern 7 CGAAAAGTATCCATCTCGCAGTGCTGAACTAGACA
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
```



Rai N, Chaubey G, Tamang R, Pathak AK, et al. (2012) The Phylogeography of Y-Chromosome Haplogroup H1a1a-M82 Reveals the Likely Indian Origin of the European Romani Populations. PLoS ONE 7(11): e48477. doi:10.1371/journal.pone.0048477
http://www.plosone.org/article/info:doi/10.1371/journal.pone.0048477

Dangers of naïve phyleography

- Trees are random.
- Many different trees can arise under the same population history.
- The same tree can be compatible with many different population histories.


Present
Past
Coalescence tree
t_4
t_3
t_2
MRCA

A sample of size two

In a Wright-Fisher model with $2N$ gene copies

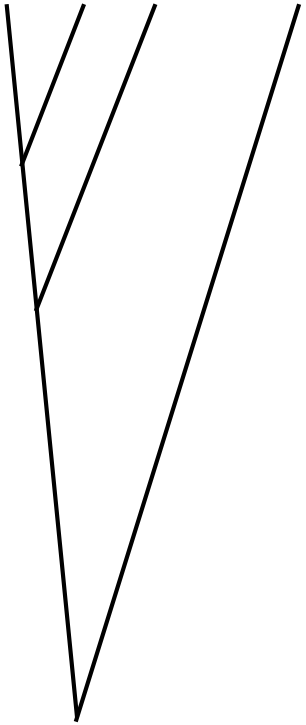
P(two gene copies do not have the same ancestor in m generations) = $\left(1-1/2N\right)^{m}$

Now scale by the population size and consider the limit of large population sizes. Set $m = 2Nt$ and let $N \to \infty$, then

$$\left(1-1/2N\right)^{2Nt} \to e^{-t}$$

The waiting time to a coalescence event is exponentially distributed with mean 1.

Coalescence trees



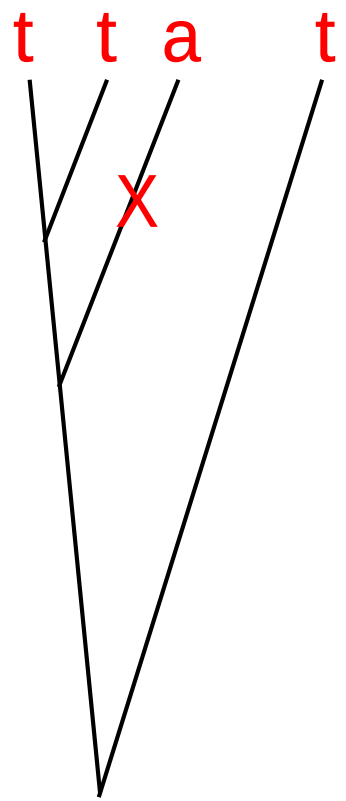
Kingman 1982 proved that the following genealogical process arises in the limit of large *N:*

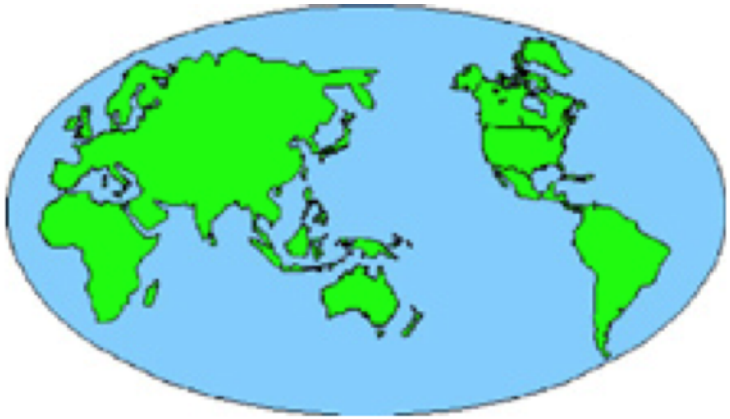
The ancestral process $\{A_n(t), t \geq 0\}$ for a sample of size n, with state space on $\{n, n-1,\ldots,1\}$ is a pure death process with transition rates

$$q_{ij} = \begin{cases} \binom{i}{2} \text{ if } j = i-1, j > 0 \\ 0 \text{ otherwise} \end{cases}$$

and absorbing state in $A_n(t) = 1$. Time (t) is here scaled in terms of the population size $2N$. Since all alleles gene copies are exchangeable, the probability that any particular two gene copies among the i gene copies coalesce is $= i(i-1)/2$.

Mutation





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
```

Two population island model

The ancestral process $\{A(t), t \geq 0\}$ for a sample (n_1, n_2), is a Markov process with state space on $\{n_1 + n_2, n_1 + n_2 - 1, \ldots, 0\} \times \{n_1 + n_2, n_1 + n_2 - 1, \ldots, 0\} \backslash (0,0)$, with initial state (n_1, n_2), and transition rates

$$q\big((i,j) \rightarrow (i-1,j)\big) = \binom{i}{2} \text{ if } i \geq 2$$

$$q\big((i,j) \rightarrow (i,j-1)\big) = \binom{i}{2} r \text{ if } j \geq 2$$

$$q\big((i,j) \rightarrow (i-1,j+1)\big) = M_1 i \text{ if } i \geq 1$$

$$q\big((i,j) \rightarrow (i+1,j-1)\big) = M_2 j \text{ if } j \geq 1$$

and absorbing states in $A(t) = (0,1)$ and $A(t) = (1,0)$. Time (t) is here scaled in terms of the population size of population 1. Since all alleles gene copies are exchangeable, the probability that any particular two gene copies among the i gene copies coalesce is $= i(i - 1)/2$.

Inference

- Assume we have some DNA sequence data (X).
- We are interested in a possibly vector valued parameter Θ.
- We have $p(G \mid \Theta)$, the density (with respect to a multidimensional Lebesgue measure) of genealogies (coalescent trees), G.
- We can also easily calculate $p(X \mid G)$, using standard Markov chain theory and a dynamic programming algorithm.

This suggests the following representation:

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.

Importance Sampling

$$\int_{\psi} p(X|G)p(G|\Theta)dG = \int_{\psi} p(X|G)p(G|\Theta)\frac{h(g)}{h(g)}dG$$

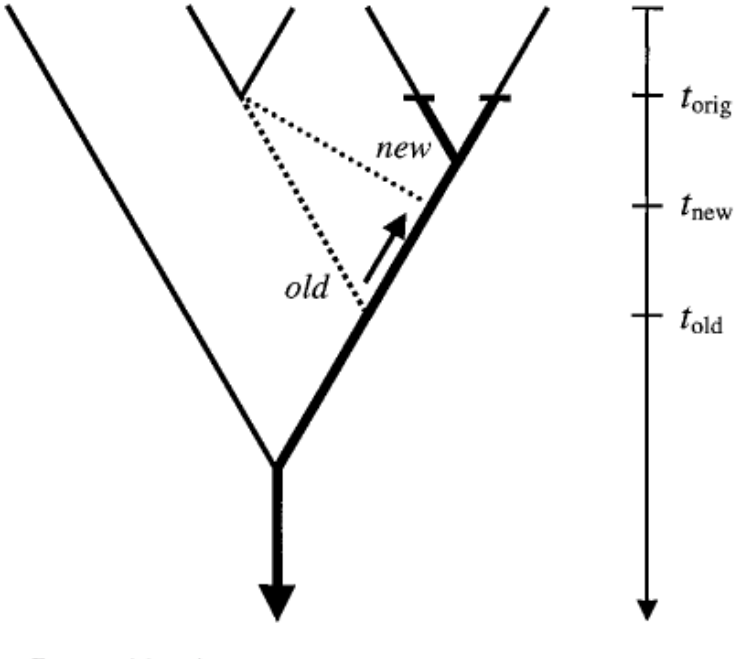
$$= E\left[\frac{p(X|G)p(G|\Theta)}{h(g)}\right]$$

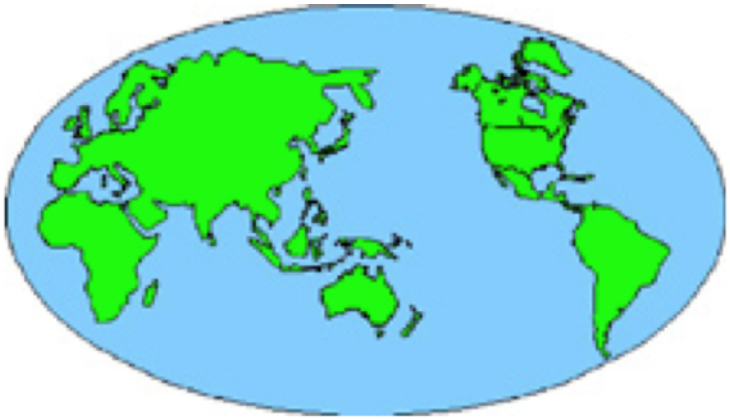
So

$$p(X|\Theta) \approx \frac{1}{k}\sum_{i=1}^{k}\frac{p(X|G_i)p(G_i|\Theta)}{h(g_i)}$$

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

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
```

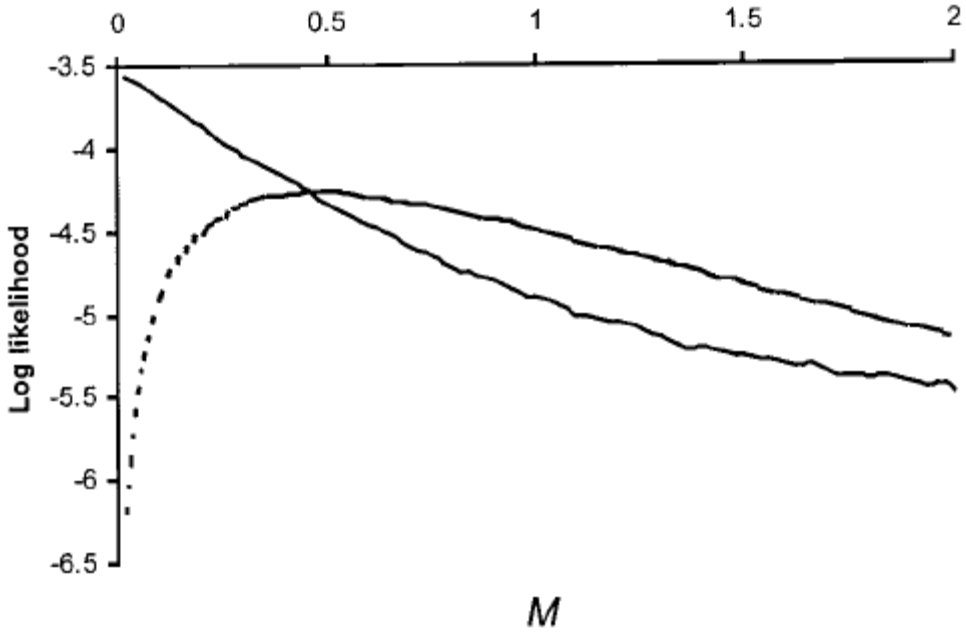
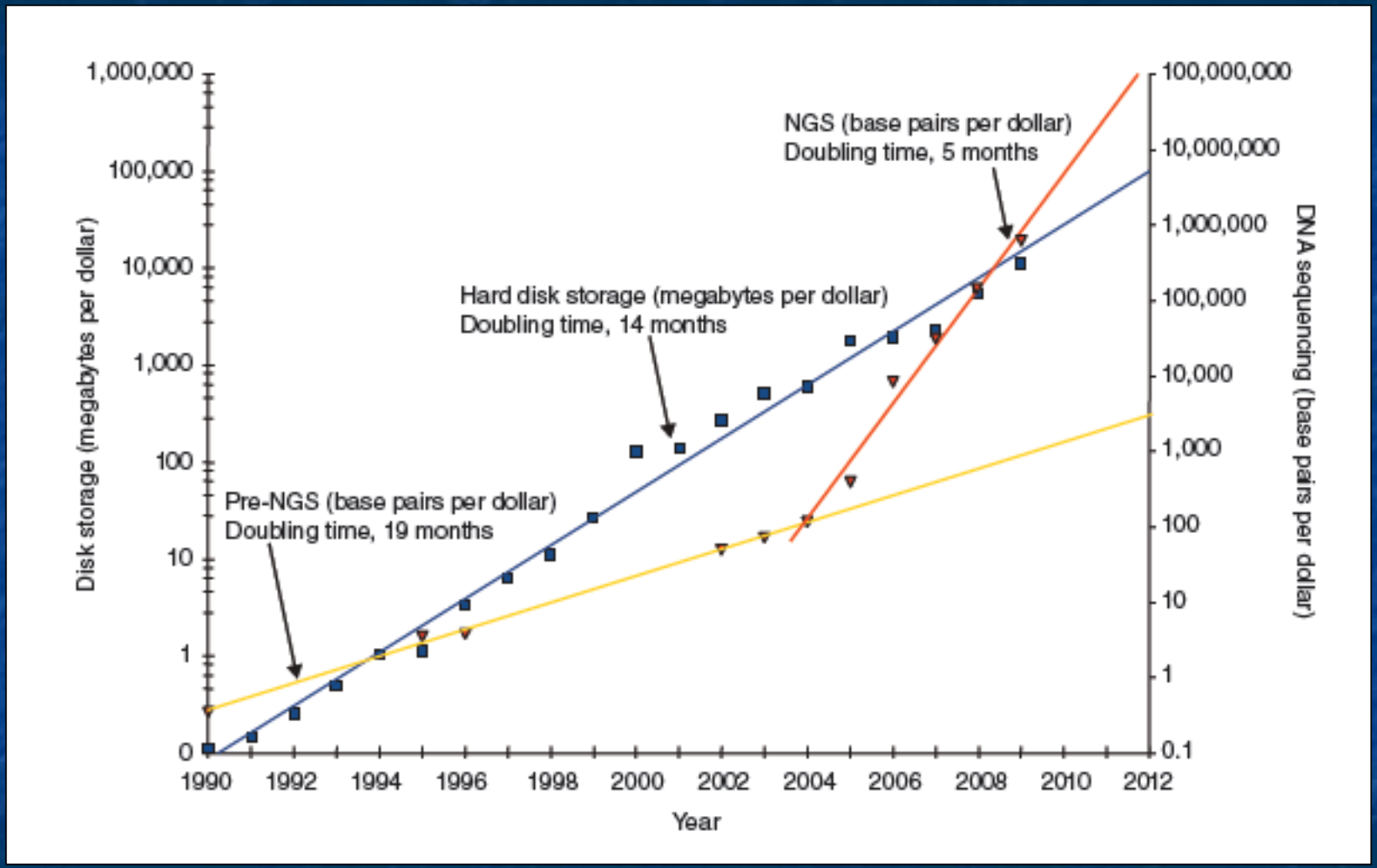


FIGURE 8.—The integrated likelihood surfaces for M_1 (dots) and M_2 (solid lines) estimated from the data by ORTI *et al.* 1994.


Genome Analyzer
illumina

Price of Sequencing

- 1990: 1 dollar per base.
- 2000: 0.01 dollars per base.
- 2012: 10^{-8} dollar per base.


1,000,000
100,000
10,000
1,000
100
10
1
0
Disk storage (megabytes per dollar)
100,000,000
10,000,000
1,000,000
100,000
10,000
1,000
100
10
1
0.1
DNA sequencing (base pairs per dollar)
NGS (base pairs per dollar)
Doubling time, 5 months
Hard disk storage (megabytes per dollar)
Doubling time, 14 months
Pre-NGS (base pairs per dollar)
Doubling time, 19 months
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
Year

Additional Challenges

(1) Each site in a genome may have it's own tree.

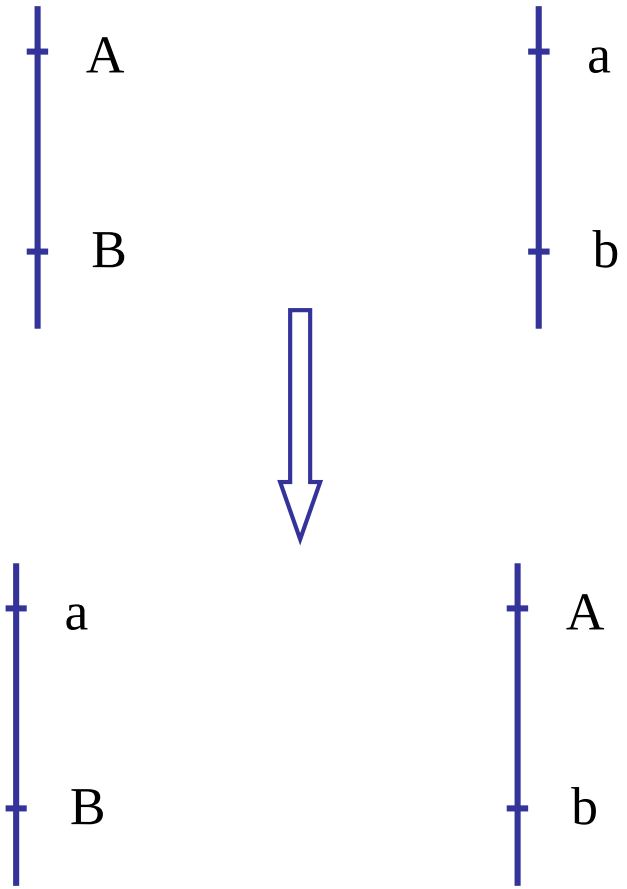
(2) Missing data and errors.

Additional Challenges

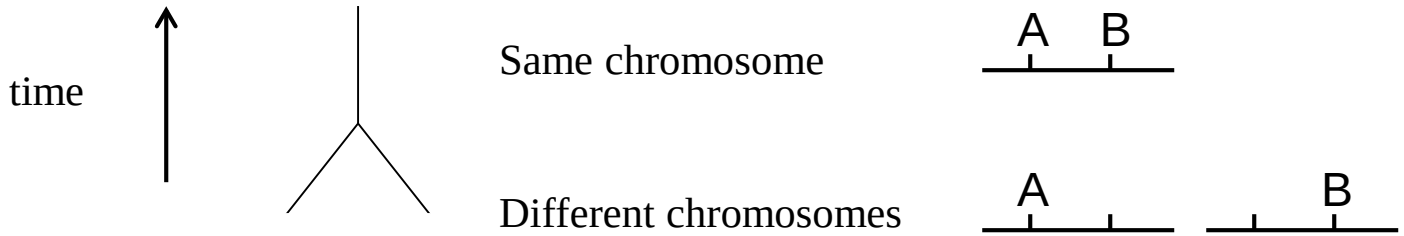
(1) Each site in a genome may have it's own tree.

(2) Missing data and errors.

Recombination



Consider the ancestral history of a single chromosome backwards in time



A sample of size one

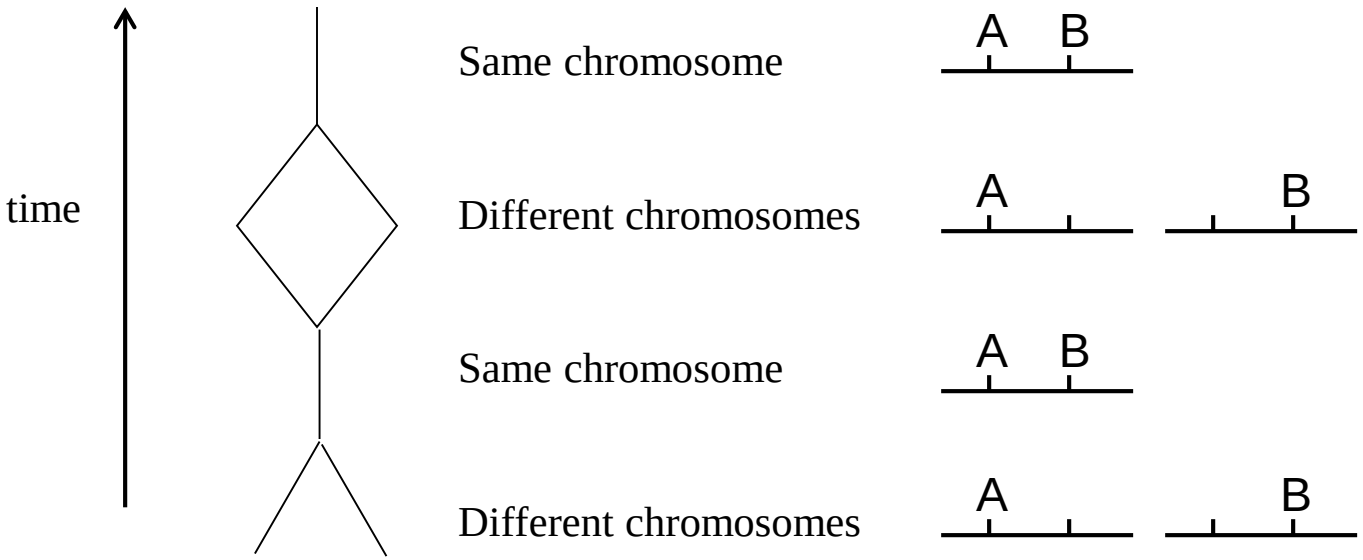
We will look at a model in which recombination events occur in each generation with probability r between two homologous sequences**.** Then

P(gene copy does not recombine in m generations) = $(1-r)^m$

We now look at the similar time scale as for the case of the coalescent process without recombination, i.e. we scale by the population size and consider the limit of large population sizes. Set $m = 2Nt$, $R = 2Nr$ and let $N \to \infty$, then

$$(1-R/2N)^{2Nt} \to e^{-Rt}$$

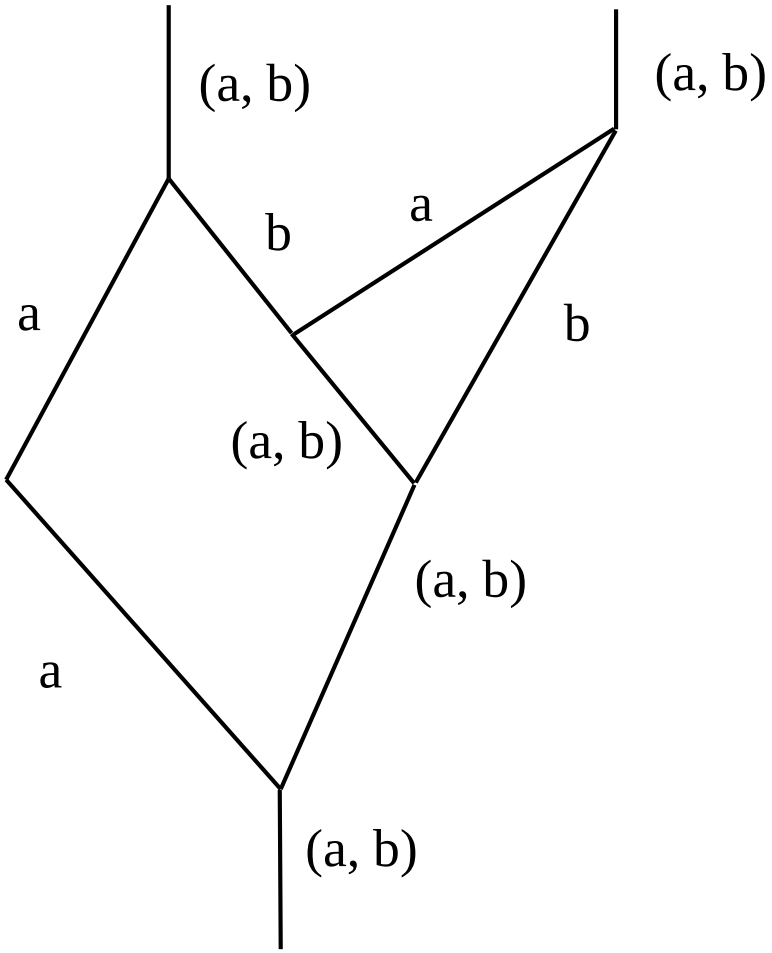
The waiting time to a recombination event is exponentially distributed with parameter $R = 2Nr$.



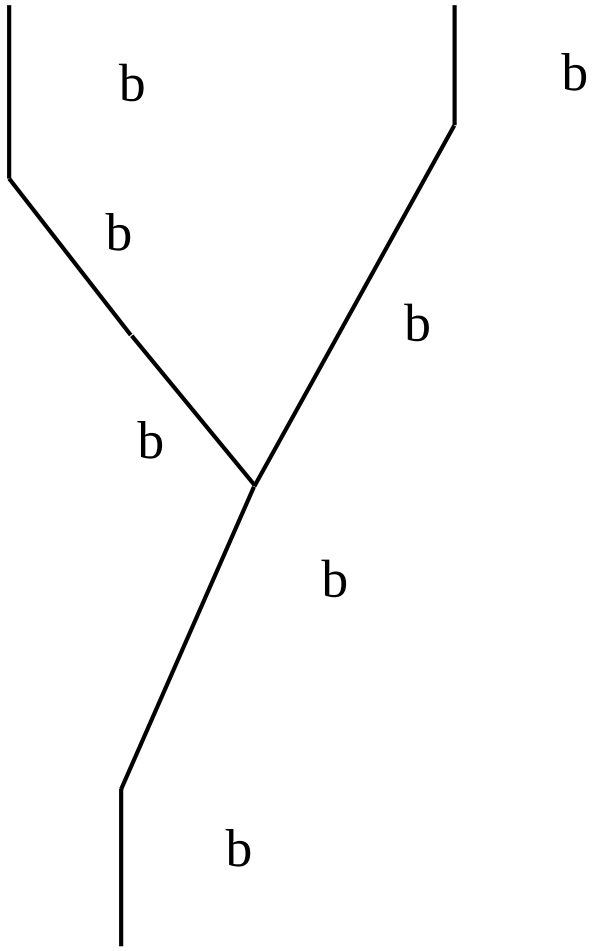
Mean waiting time to recombination: $R = 2Nr$

Mean waiting time to coalescence: $2N$

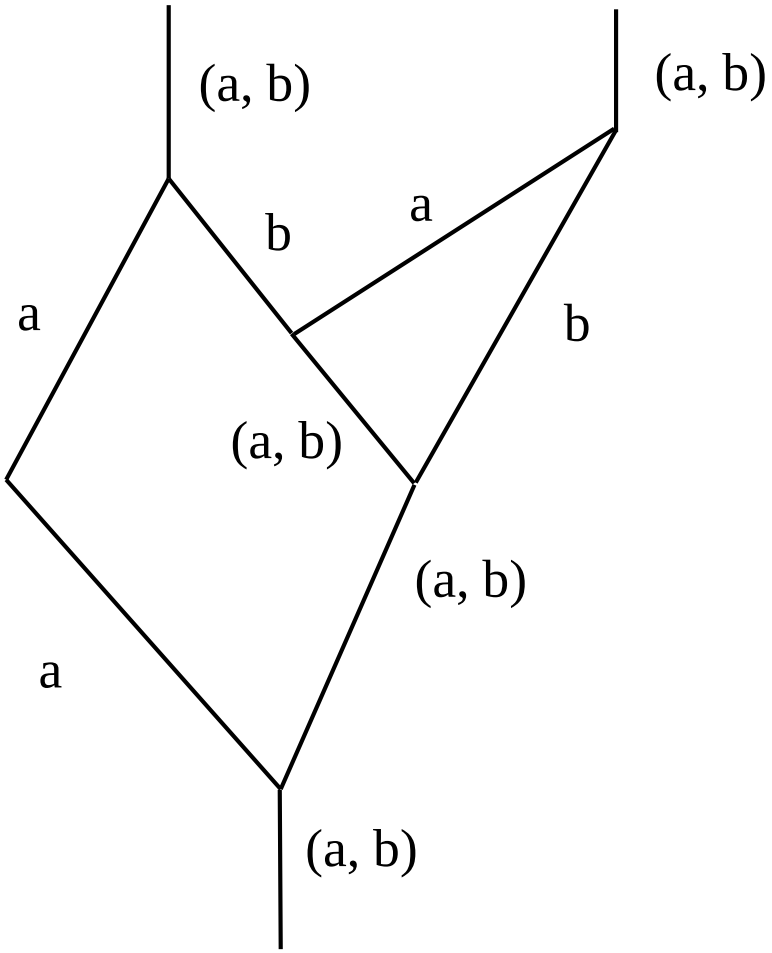
Ancestral recombination Graph for 2 loci and $n = 2$



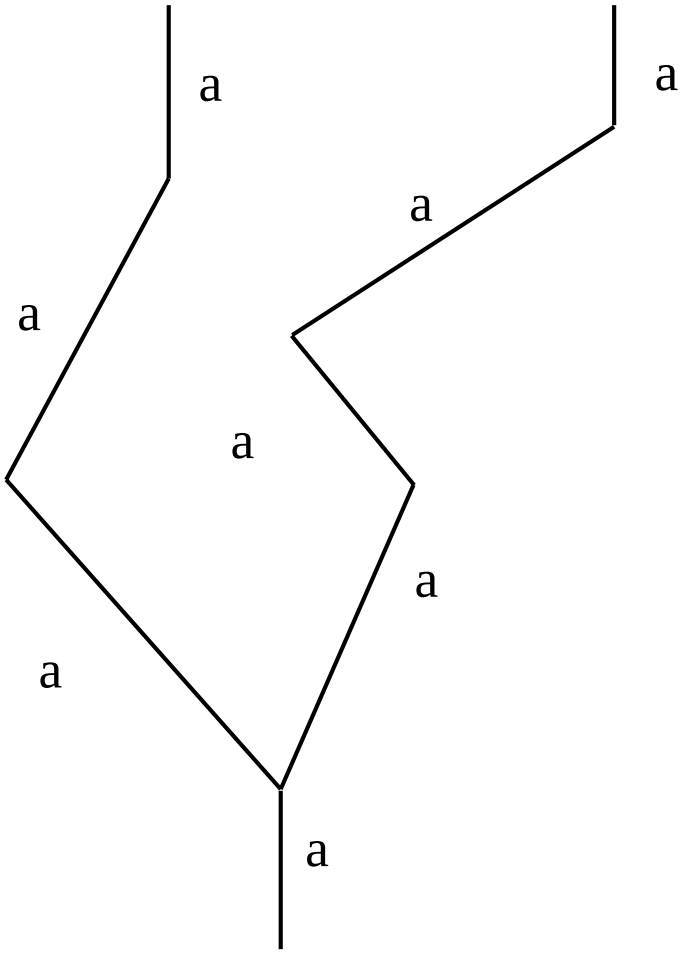
Ancestral recombination Graph for 2 loci and $n = 2$



Ancestral recombination Graph for 2 loci and $n = 2$



Ancestral recombination Graph for 2 loci and $n = 2$



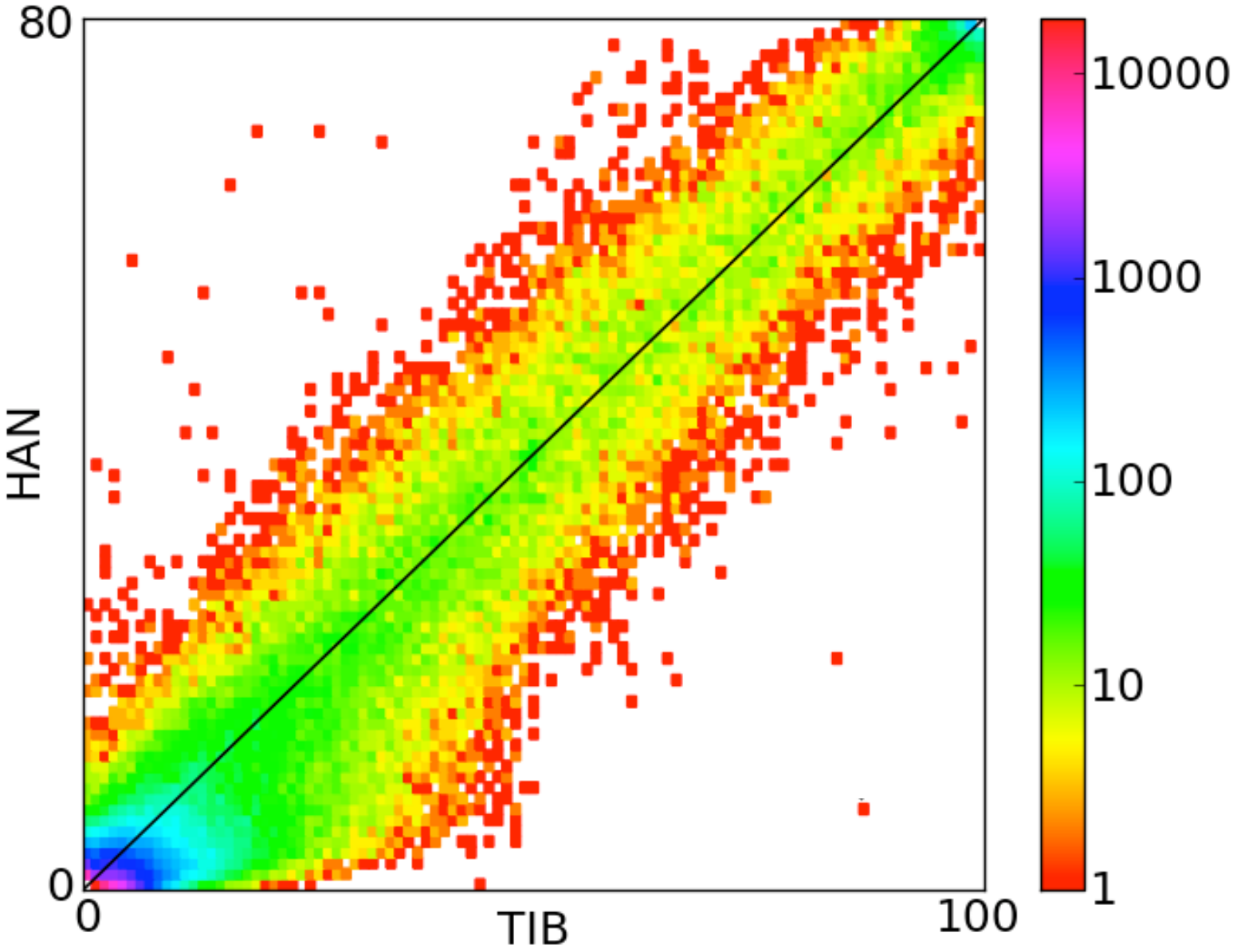
Each site in a genome may have it's own tree.

- Because of recombination, each position in the genome has it own tree – possibly shared with some nearby positions.
- A process on Ancestral Recombination Graphs (ARG) can describe the coalescence process with recombination.
- The size of increases of the ARG increases fast with the size of the region considered.
- Methods based on full likelihood are very computationally intensive, implementation heavy, and cannot be applied to large genomics regions.

So we cheat!

- Use statistics that are not sufficient.
- Use approximating models.
- Do both...
- Give up on parametric inferences.

Site Frequency Spectrum



Cheating 1: Composite likelihood

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

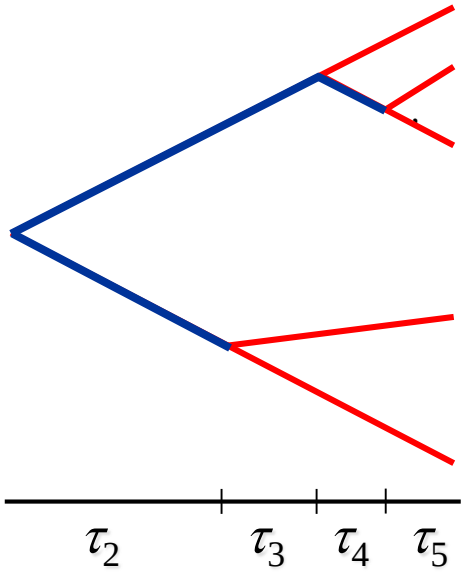
Sampling probability of a SNP with site pattern *j* (a binary vector)

Number of SNPs with pattern *j* in the data

SNPs within a gene are correlated. But estimator is consistent (Wiuf 2006).

Estimation

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



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

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

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

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

Data

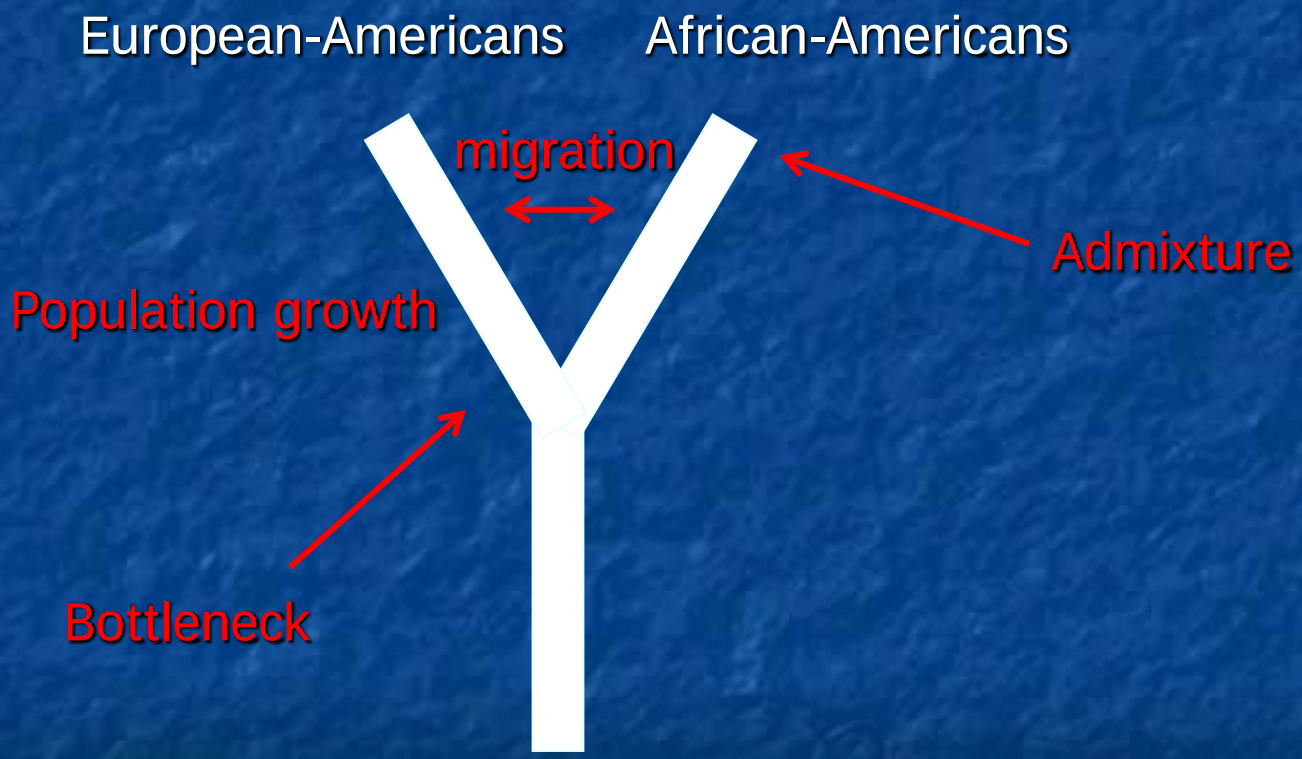
Directly sequenced polymorphism data from 20 European-Americans, 19 African-Americans and one chimpanzee from 9,316 protein coding genes (Bustamante et al. 2005).

Objectives

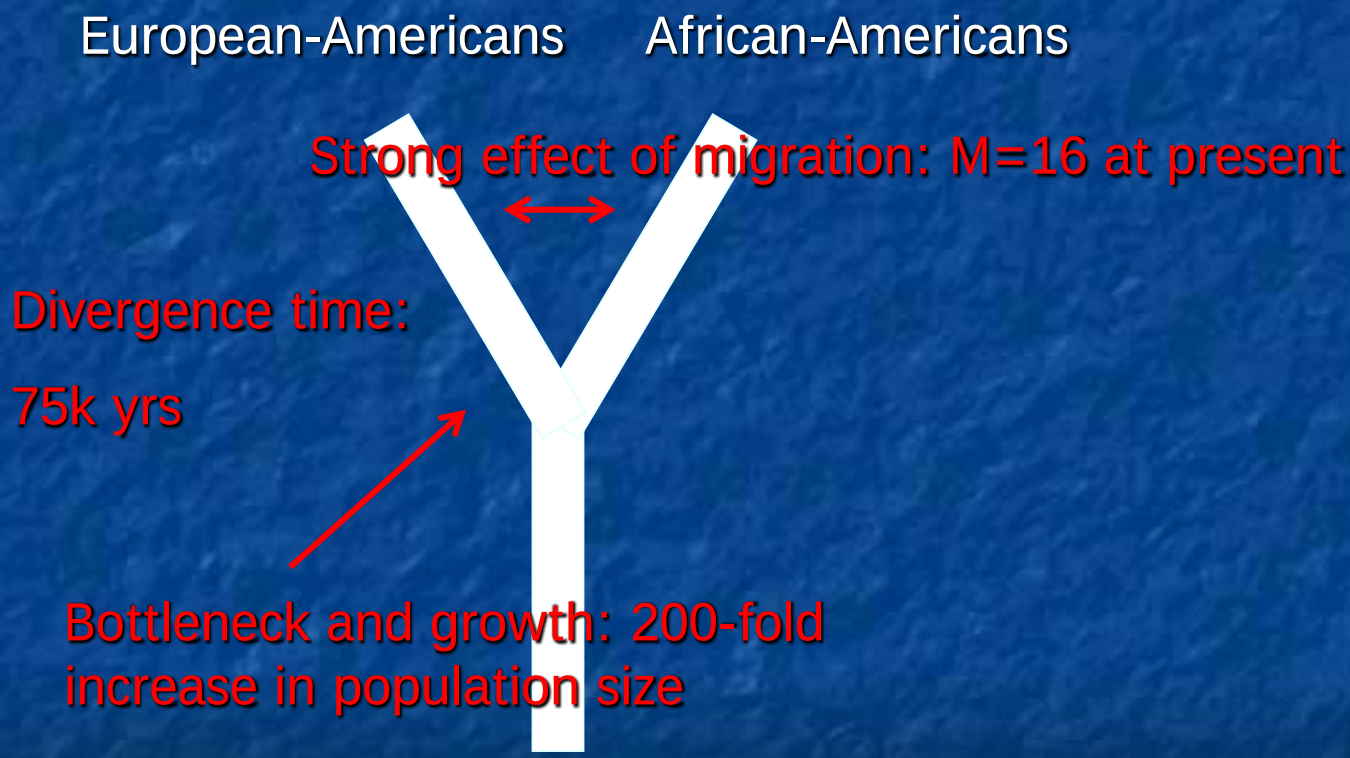
To detect natural selection in individual genes using the frequency spectrum.

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

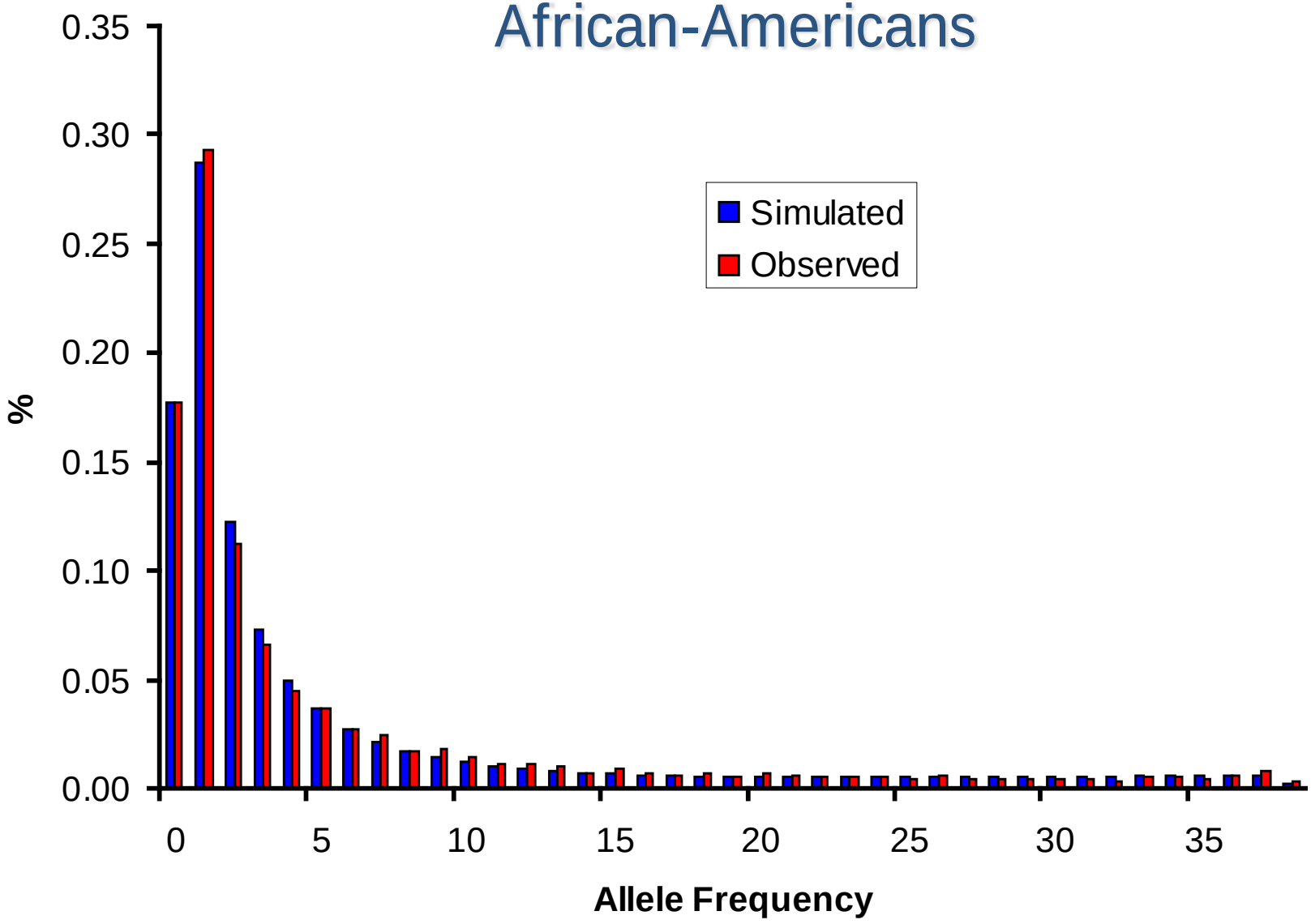
Demographic model



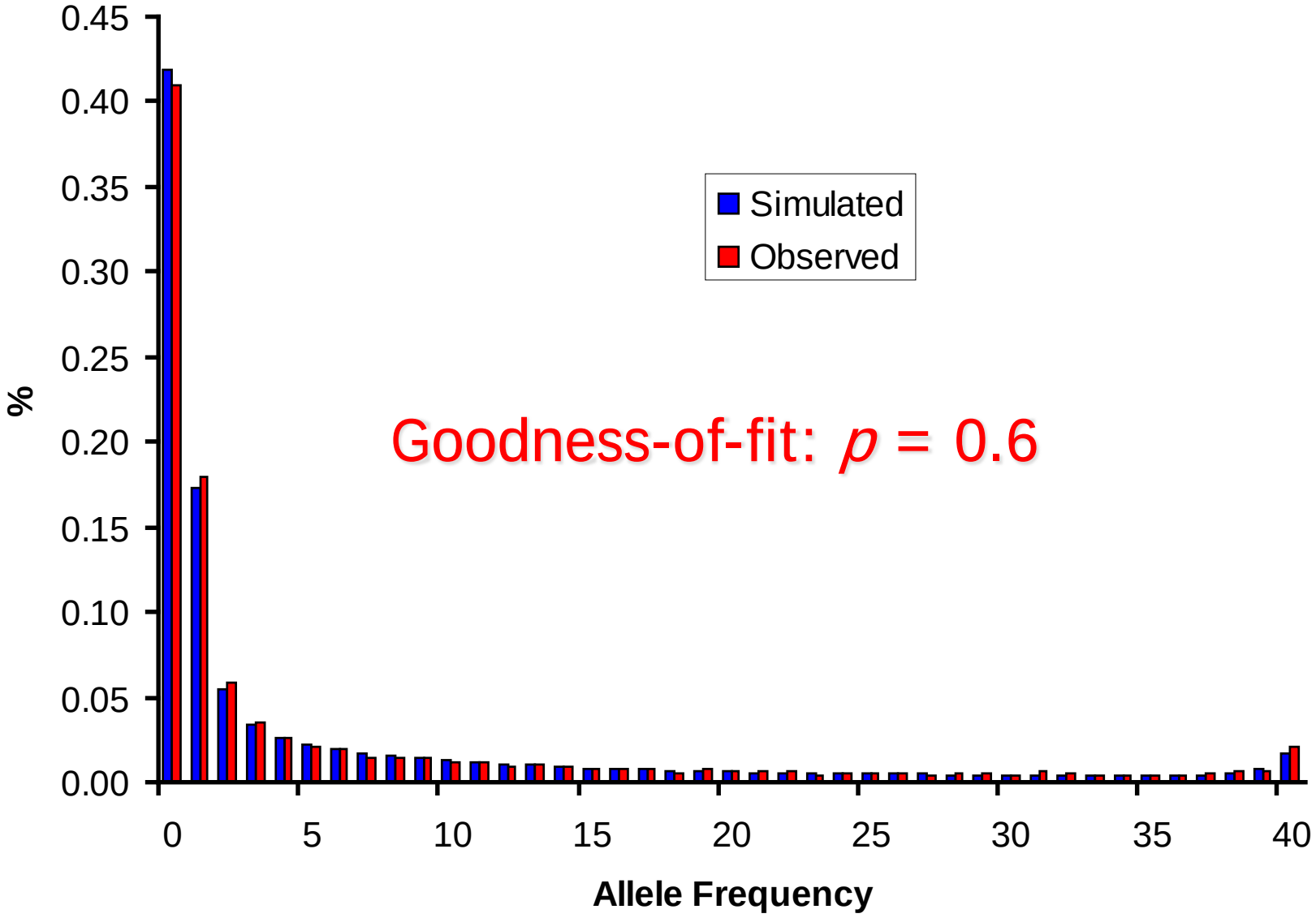
Estimates


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



Cheating 2: 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 $\rho(S(x_i), S(y)) > \epsilon$.

Cheating 3: approximating models

Sequentially Markov Coalescent (SMC)

- Assumes $p(G_i \mid G_{i\text{-}1}, G_{i\text{-}2},\ldots) = p(G_i \mid G_{i\text{-}1})$.

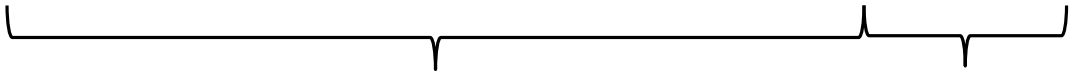
- Greatly simplifies calculations.

- At least two versions: SMC (McVean and Cardin) and SMC' (Marjoram and Wall).

Identity-By-State (IBS) tracts

CATGACGTGAGACCAGATATAGCAGATGG

CATGGCGTGAGACCAGATATACCAGTTGG



Tract 1 Tract 2

Identity-By-State (IBS) tracts

Joint probability that a randomly chosen site is the left endpoint of an L-base IBS region and that t is the TMRCA of the rightmost base pair of the region:

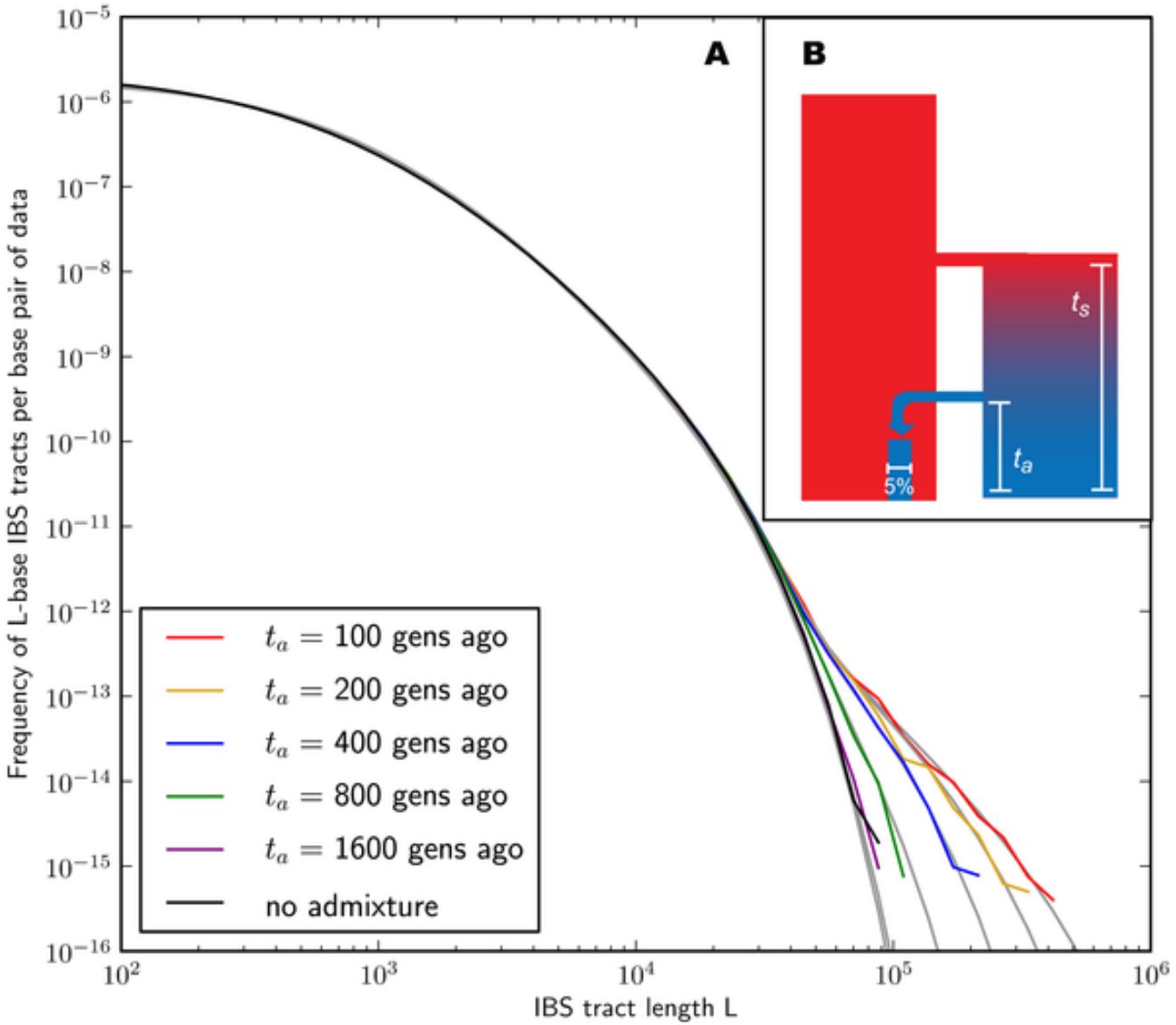
$$P(L,t) = e^{-t\theta} \int_{t_0=0}^{\infty} P(L-1,t_0) P(T_b = t \mid T_{b-1} = t_0) dt_0$$

T_b: TMRCA of site b
θ: $4N\mu$

Harris and Nielsen 2013. *PLoS Genetics* doi:10.1371

Kelley Harris

Identity-By-State (IBS) tracts



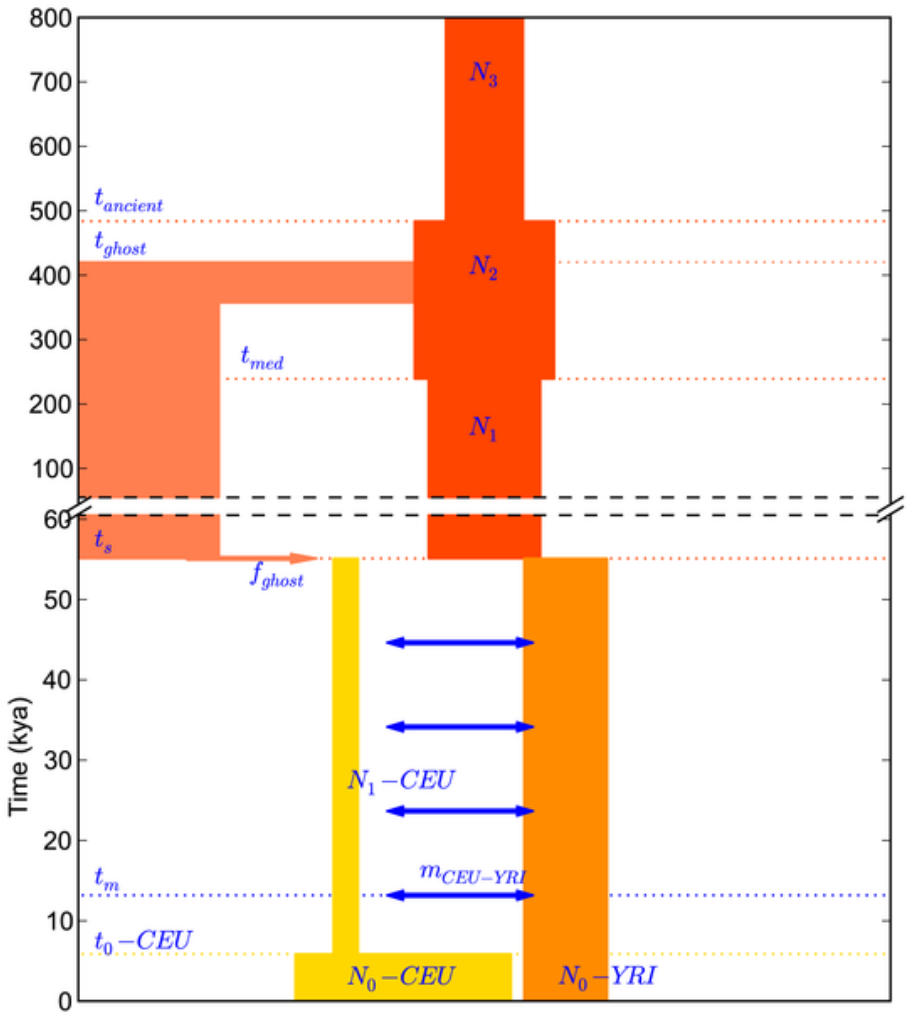
Identity-By-State (IBS) tracts

	τ_a (gens)	τ_s (gens)	f	N
True value:	400	2,000	0.05	10,000
Mean:	431	1,990	0.0505	9,806
Std dev:	51	41	0.00652	27
Bias:	31	−10	0.0005	−194
Mean squared error:	3280	1781	4.27×10^{-5}	3.84×10^{4}
True value:	200	2,000	0.05	10,000
Mean:	220	1,983	0.0499	10,003
Std dev:	28	39	0.00328	287
Bias:	20	−17	−0.0001	−3
Mean squared error:	1184	1810	1.08×10^{-5}	8.23×10^{4}

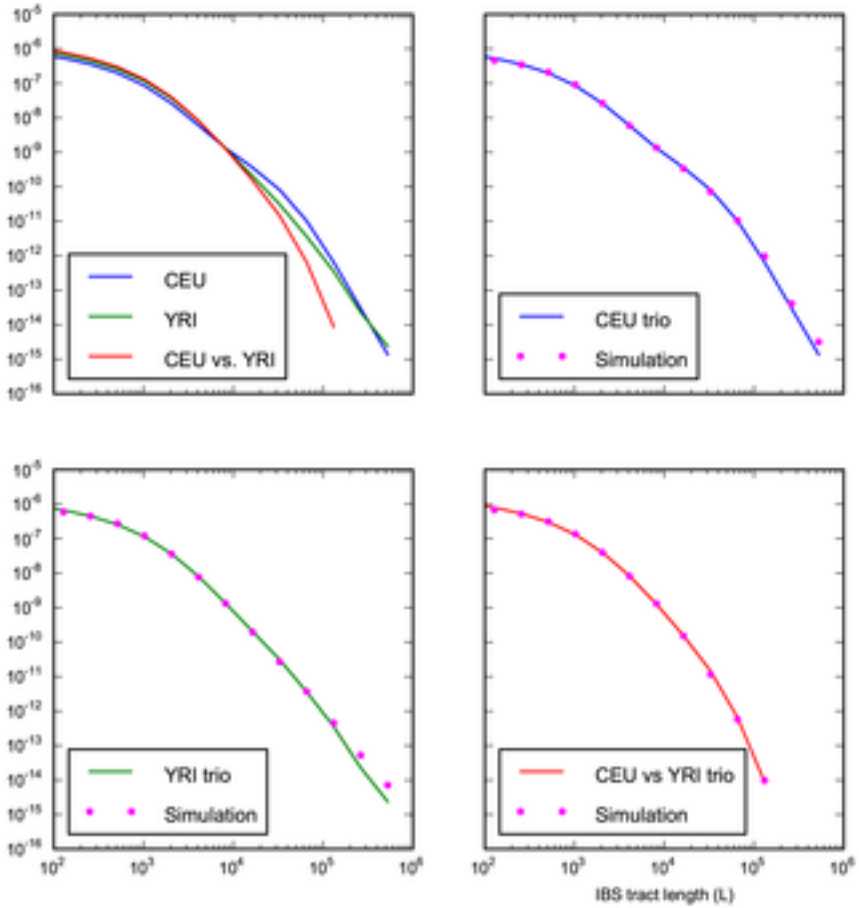
Using MS, we simulated 200 replicates of the admixture scenario depicted in Figure 2B. In 100 replicates, the gene flow occurred 400 generations ago, while in the other 100 replicates it occurred 200 generations ago. Our estimates of the four parameters τ_a, τ_s, f, N are consistently close to the true values, showing that we are able distinguish the two histories by numerically optimizing the likelihood function.

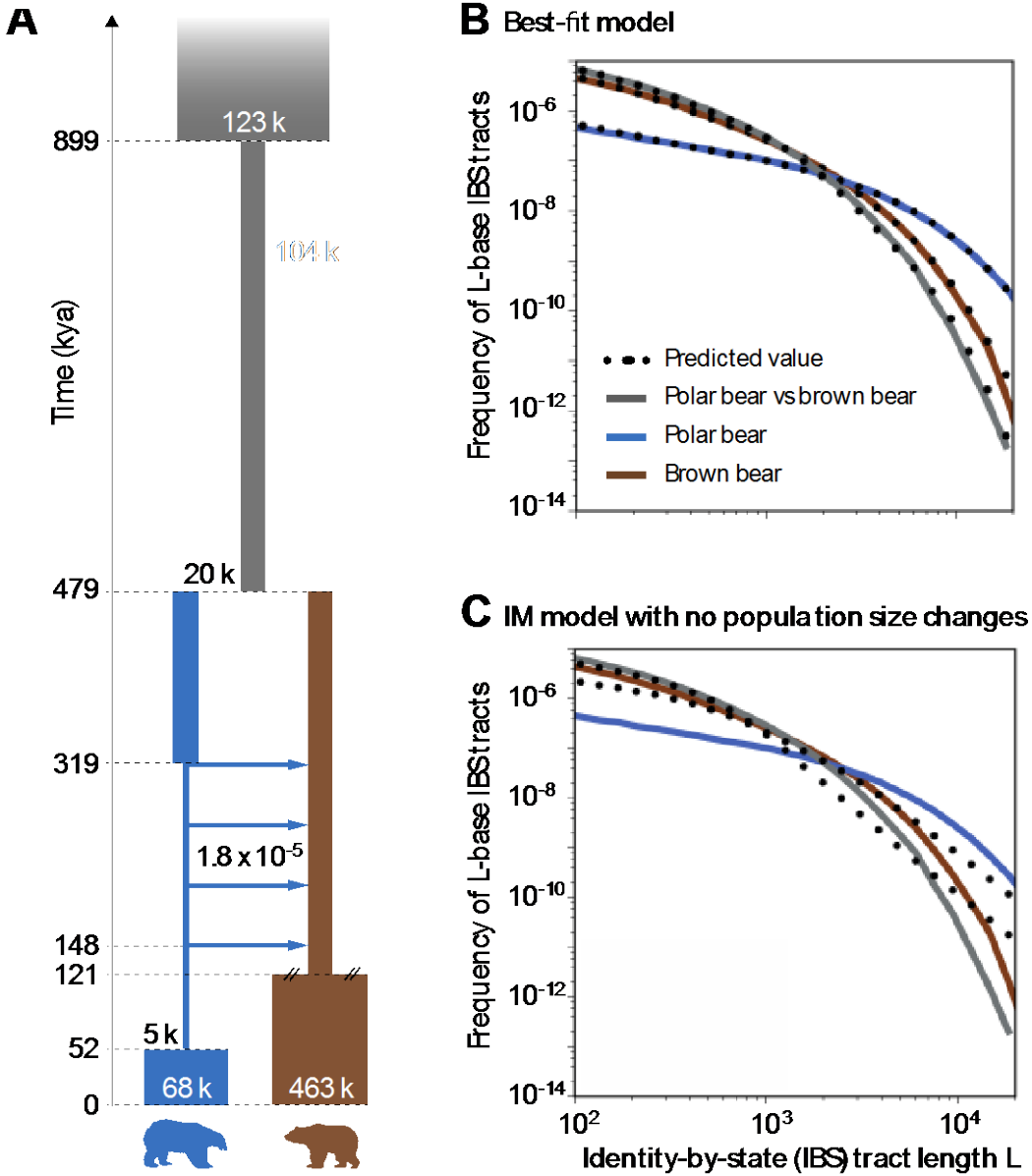
doi:10.1371/journal.pgen.1003521.t001

Identity-By-State (IBS) tracts



Identity-By-State (IBS) tracts





Liu *et al.* 2014. *Cell* 157: 785–794

Polar Bear Physiology

- Diet is extremely rich in fat
- Fasting triglyceride levels: 292 mg/dl
- Fasting cholesterol levels 381 mg/dl

Bear Genomics

Sequencing of 79 polar bear and 10 brown bear genomes (BGI).

Liu *et al.* 2014. *Cell* 157: 785–794

Genes under lineage specific selection in Polar Bears

			Polymorphisms		Fixed mutations		
Gene	Coding seq length	Effective length	Polar bear	Brown bear	Polar bear	Brown bear	Rank of F_{ST}
TTN	100,203	99,416	571	803	74	8	0.994
APOB	13,290	13,264	50	162	27	1	0.976
OR5D13	945	871	4	38	12	0	0.906
FCGBP	7,410	5,216	42	83	20	1	0.941
XIRP1	5,538	3,848	23	52	17	0	0.970
COL5A3	5,166	4,402	46	92	18	1	0.900
LYST	11,400	11,172	50	83	22	2	0.980
ALPK3	4,848	3,007	28	47	19	0	0.989
VCL	3,438	3,106	24	56	12	0	0.906
SH3PXD2B	2,673	2,458	15	44	13	2	0.973
EHD3	1,608	1,230	12	38	9	0	0.981
IPO4	2,934	1,260	11	31	10	0	0.959
ARID5B	3,555	3,109	29	50	13	0	0.933
ABCC6	4,527	3,346	34	65	15	2	0.945
LAMC3	4,269	1,885	21	44	11	0	0.946
CUL7	4,914	2,701	28	42	14	0	0.926
C15orf55	3,414	3,001	17	53	9	1	0.974
POLR1A	5,154	4,499	37	58	15	1	0.916
AIM1	4,728	4,344	33	48	14	0	0.992
OR8B8	966	965	1	16	6	0	0.961

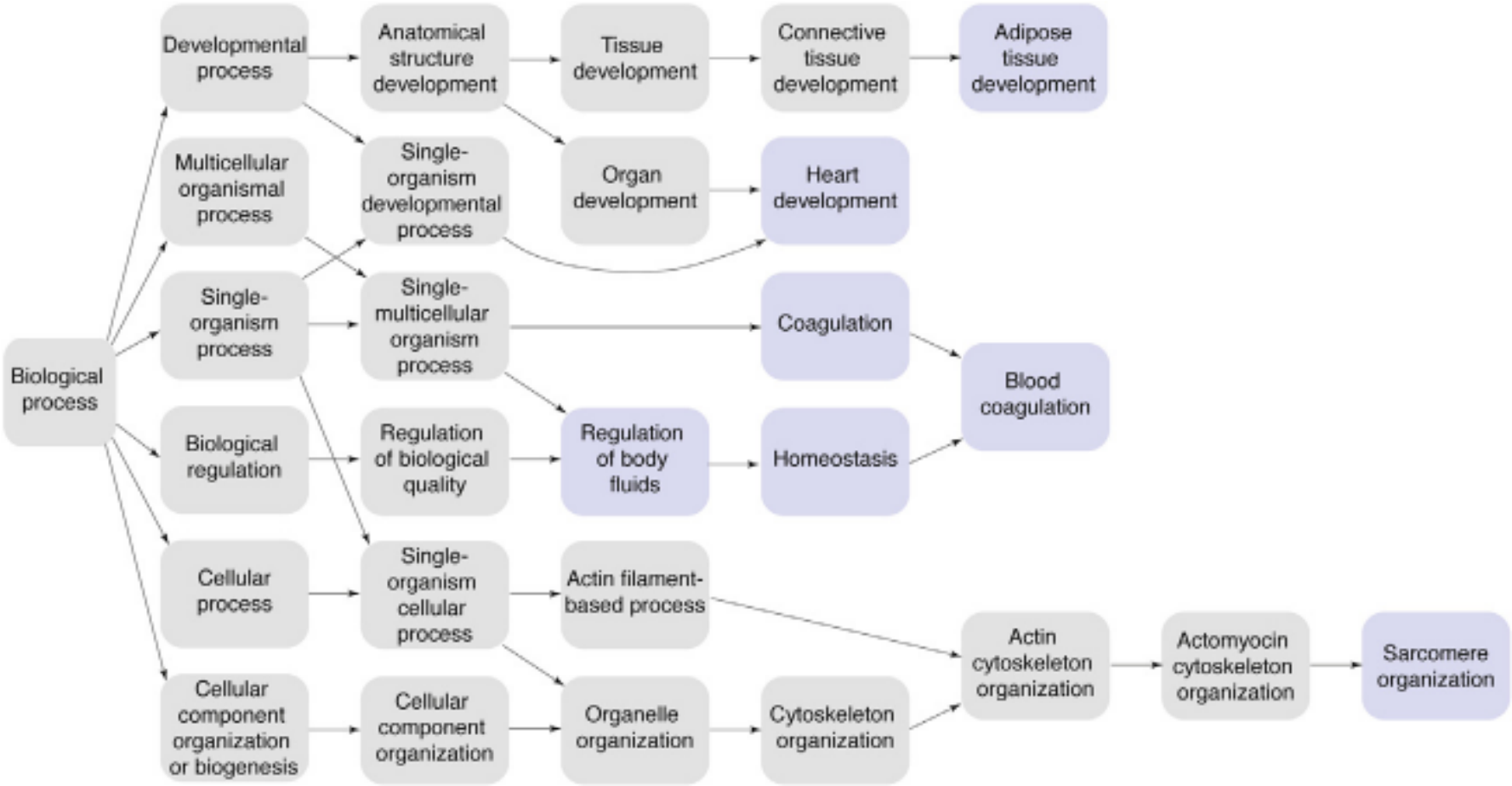
Liu *et al.* 2014. *Cell* 157: 785–794

Coat color

			Polymorphisms		Fixed mutations		
Gene	Coding seq length	Effective length	Polar bear	Brown bear	Polar bear	Brown bear	Rank of F_{ST}
TTN	100,203	99,416	571	803	74	8	0.994
APOB	13,290	13,264	50	162	27	1	0.976
OR5D13	945	871	4	38	12	0	0.906
FCGBP	7,410	5,216	42	83	20	1	0.941
XIRP1	5,538	3,848	23	52	17	0	0.970
COL5A3	5,166	4,402	46	92	18	1	0.900
LYST	11,400	11,172	50	83	22	2	0.980
ALPK3	4,848	3,007	28	47	19	0	0.989
VCL	3,438	3,106	24	56	12	0	0.906
SH3PXD2B	2,673	2,458	15	44	13	2	0.973
EHD3	1,608	1,230	12	38	9	0	0.981
IPO4	2,934	1,260	11	31	10	0	0.959
ARID5B	3,555	3,109	29	50	13	0	0.933
ABCC6	4,527	3,346	34	65	15	2	0.945
LAMC3	4,269	1,885	21	44	11	0	0.946
CUL7	4,914	2,701	28	42	14	0	0.926
C15orf55	3,414	3,001	17	53	9	1	0.974
POLR1A	5,154	4,499	37	58	15	1	0.916
AIM1	4,728	4,344	33	48	14	0	0.992
OR8B8	966	965	1	16	6	0	0.961

Liu *et al.* 2014. *Cell* 157: 785–794

GO analysis of selected genes



Liu *et al.* 2014. *Cell* 157: 785–794

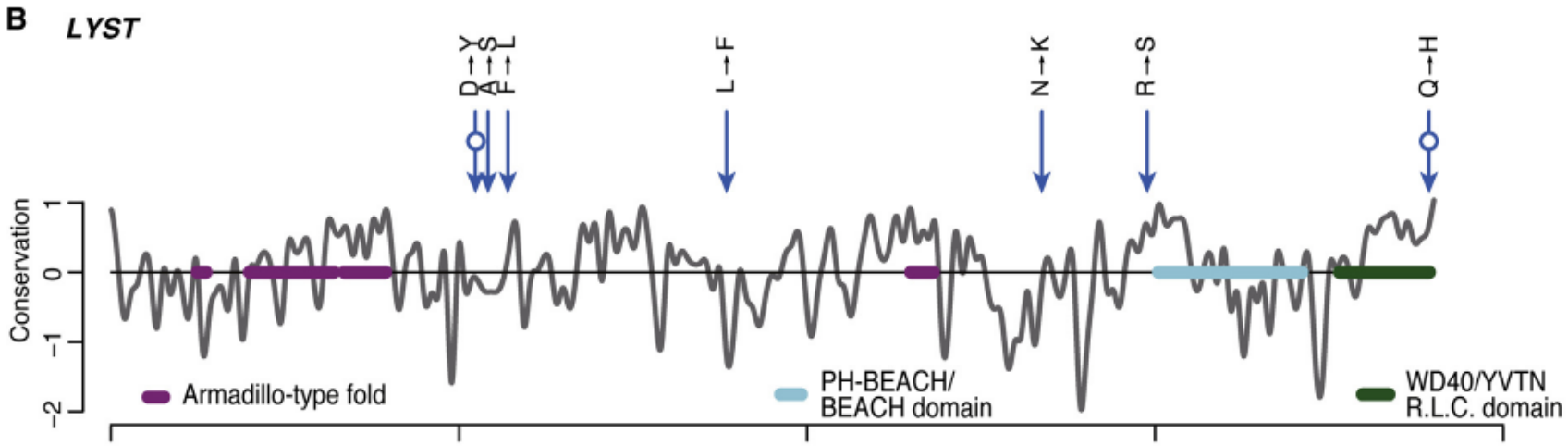
Related to cardiovascular function

			Polymorphisms		Fixed mutations		
Gene	Coding seq length	Effective length	Polar bear	Brown bear	Polar bear	Brown bear	Rank of F_{ST}
TTN	100,203	99,416	571	803	74	8	0.994
APOB	13,290	13,264	50	162	27	1	0.976
OR5D13	945	871	4	38	12	0	0.906
FCGBP	7,410	5,216	42	83	20	1	0.941
XIRP1	5,538	3,848	23	52	17	0	0.970
COL5A3	5,166	4,402	46	92	18	1	0.900
LYST	11,400	11,172	50	83	22	2	0.980
ALPK3	4,848	3,007	28	47	19	0	0.989
VCL	3,438	3,106	24	56	12	0	0.906
SH3PXD2B	2,673	2,458	15	44	13	2	0.973
EHD3	1,608	1,230	12	38	9	0	0.981
IPO4	2,934	1,260	11	31	10	0	0.959
ARID5B	3,555	3,109	29	50	13	0	0.933
ABCC6	4,527	3,346	34	65	15	2	0.945
LAMC3	4,269	1,885	21	44	11	0	0.946
CUL7	4,914	2,701	28	42	14	0	0.926
C15orf55	3,414	3,001	17	53	9	1	0.974
POLR1A	5,154	4,499	37	58	15	1	0.916
AIM1	4,728	4,344	33	48	14	0	0.992
OR8B8	966	965	1	16	6	0	0.961

Liu *et al.* 2014. *Cell* 157: 785–794

APOB Apolipoprotein B

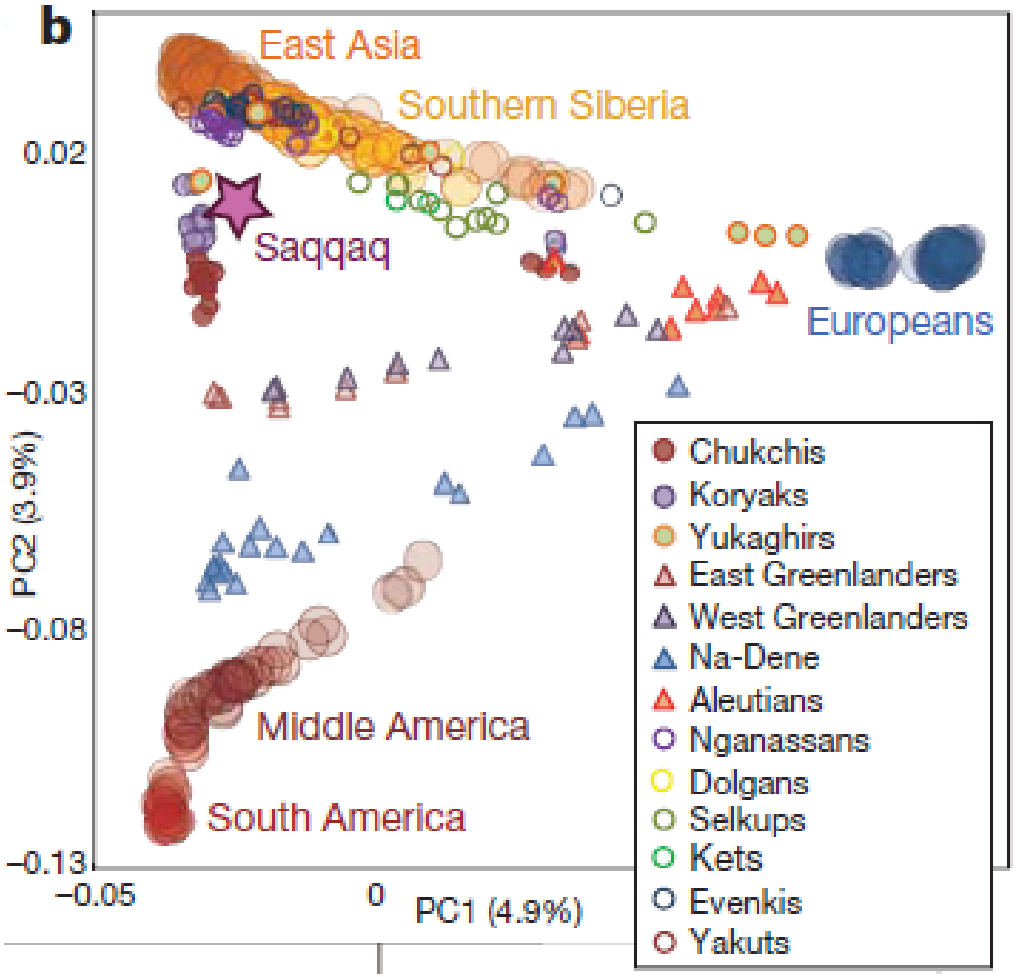
(LDL cholesterol component)



Liu *et al.* 2014. *Cell* 157: 785–794

Cheating 4: Back to non-parametrics

Saqqaq Genome



Structure and Admixture Analyses

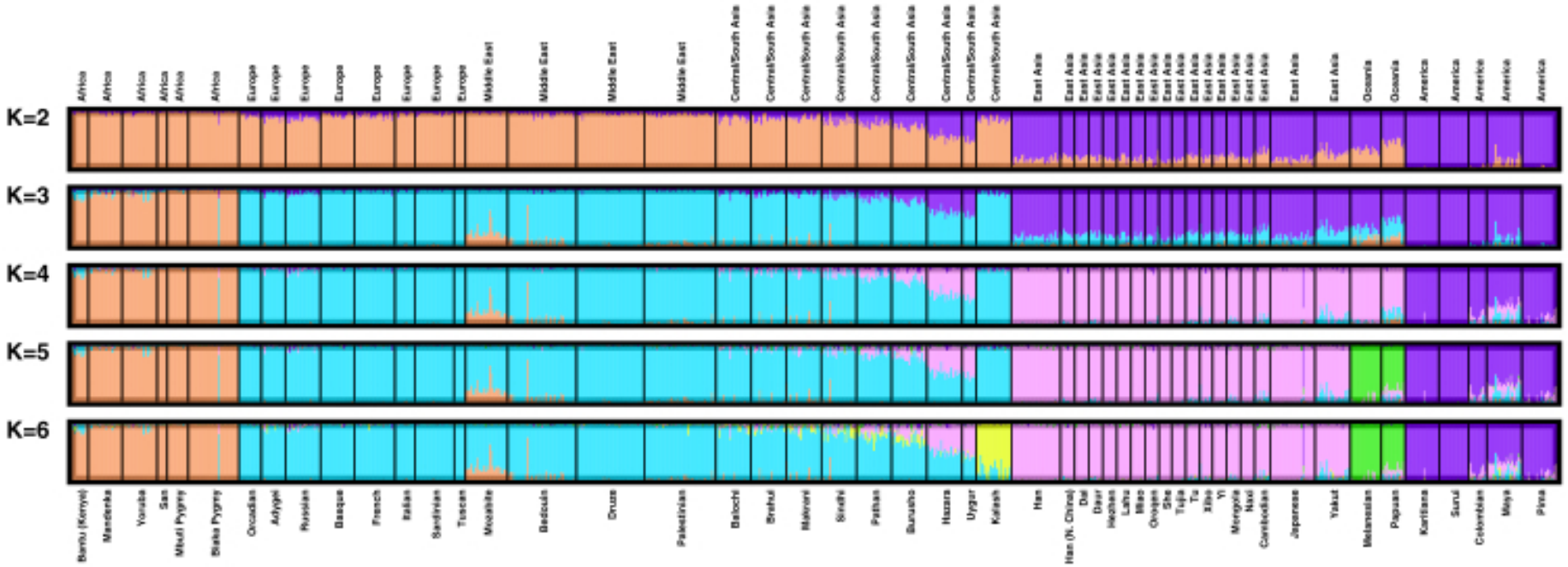
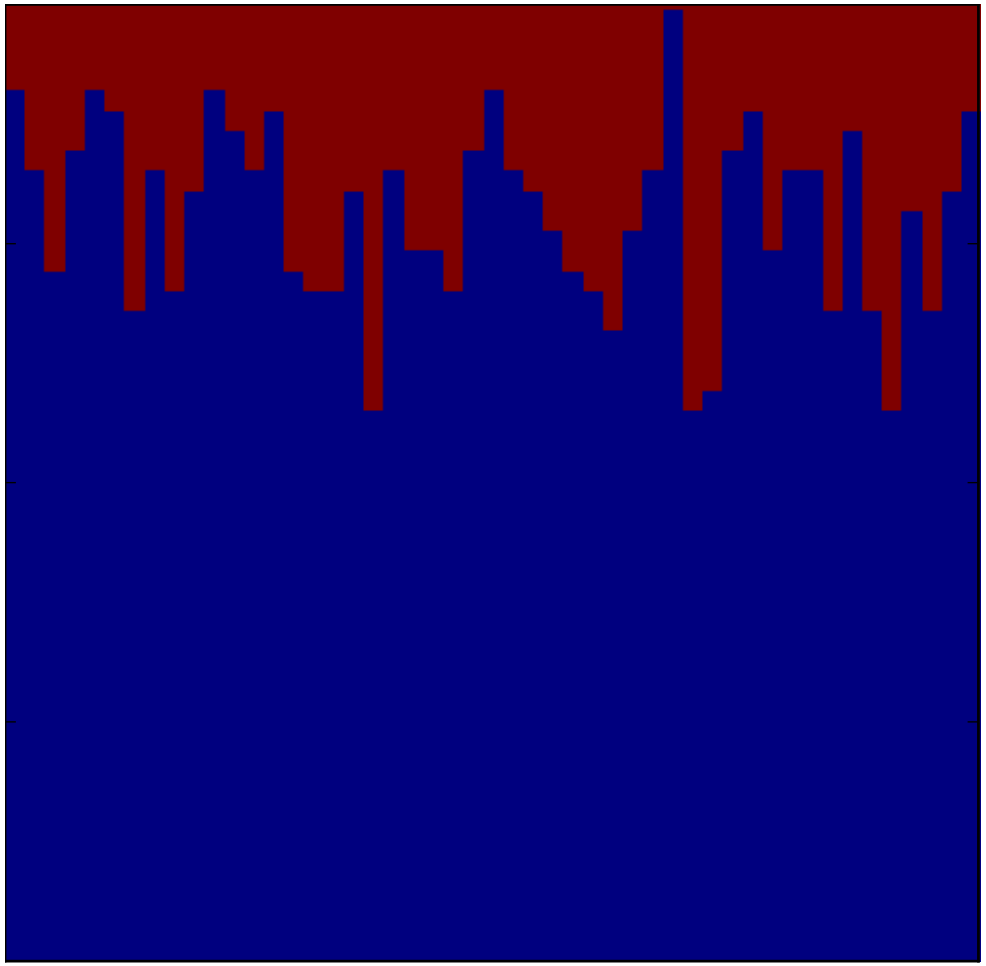


Fig. 1. Estimated population structure. Each individual is represented by a thin vertical line, which is partitioned into *K* colored segments that represent the individual's estimated membership fractions in *K* clusters. Black lines separate individuals of different populations. Populations are labeled below the figure, with their regional affiliations above it. Ten *structure* runs at each *K* produced nearly identical individual membership coefficients, having pairwise similarity coefficients above 0.97, with the exceptions of comparisons involving four runs at $K = 3$ that separated East Asia instead of Eurasia, and one run at $K = 6$ that separated Karitiana instead of Kalash. The figure shown for a given *K* is based on the highest probability run at that *K*.

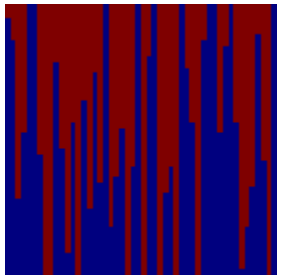
Admixture components



Mason Liang

Population Genetics of Admixture

- How do admixture components change through time?
- What is the relationship between a particular distribution of admixture components and particular population histories?



$$H = \frac{1}{L} \int_0^L A(\ell) d\ell$$

$$A(\ell) = \begin{cases} 0 \ : \ \ell \text{ descended from first source population} \\ 1 \ : \ \ell \text{ descended from second source population} \end{cases}$$

$$k_1 \equiv \frac{1}{n} \sum_{i=1}^{n} H_{i(g)}$$

$$\begin{aligned}
\mathbb{E}(k_1) &= \mathbb{E}(H_{1(g)}) \\
&= \frac{1}{L} \int_0^L \mathbb{P}\{A_{1(g)}(\ell) = 1\} d\ell \\
&= \mathbb{P}\{A_{1(g)}(0) = 1\}.
\end{aligned}$$

$$k_2 \equiv \frac{1}{n-1} \sum_{i=1}^{n} \left(H_{i(g)} - k_1 \right)^2$$

$$\begin{aligned} \mathbb{E}(k_2) &= \frac{1}{n-1} \sum_{i=1}^{n} \mathbb{E}(H_{i,g}^2) - \frac{1}{n(n-1)} \sum_{i,j=1}^{n} \mathbb{E}(H_{i,g} H_{j,g}) \\ &= \mathbb{E}(H_{1,g}^2) - \mathbb{E}(H_{1,g} H_{2,g}). \end{aligned}$$

$$\begin{aligned}
\mathbb{E}(H_{1(g)}^2) &= \frac{1}{L^2}\mathbb{E}\left(\int_0^L A_{1(g)}(\ell)d\ell \int_0^L A_{1(g)}(\ell)d\ell\right) \\
&= \frac{1}{L^2}\int_0^L \int_0^L \mathbb{E}\left(A_{1(g)}(\ell)A_{1(g)}(\ell')\right) d\ell d\ell' \\
&= \frac{1}{L^2}\int_0^L \int_0^L \mathbb{P}\left\{A_{1(g)}(\ell) = 1, A_{1(g)}(\ell') = 1\right\} d\ell d\ell'.
\end{aligned}$$

Similarly,

$$\mathbb{E}(H_{1(g)}H_{2(g)}) = \frac{1}{L^2}\int_0^L \int_0^L \mathbb{P}\left\{A_{1(g)}(\ell) = 1, A_{2(g)}(\ell') = 1\right\} d\ell d\ell'$$

In matrix form:

$$\mathbf{v}_{2(g)} = \begin{pmatrix} \mathbb{P}\{A_{1,g}(\ell) = 1, A_{1,g}(\ell') = 1\} \\ \mathbb{P}\{A_{1,g}(\ell) = 1, A_{2,g}(\ell') = 1\} \end{pmatrix}$$

$$\mathbb{E}(k_2) = \frac{1}{L^2} \int_0^L \int_0^L \begin{pmatrix} 1 & -1 \end{pmatrix} \mathbf{v}_{2(g)} d\ell d\ell'$$

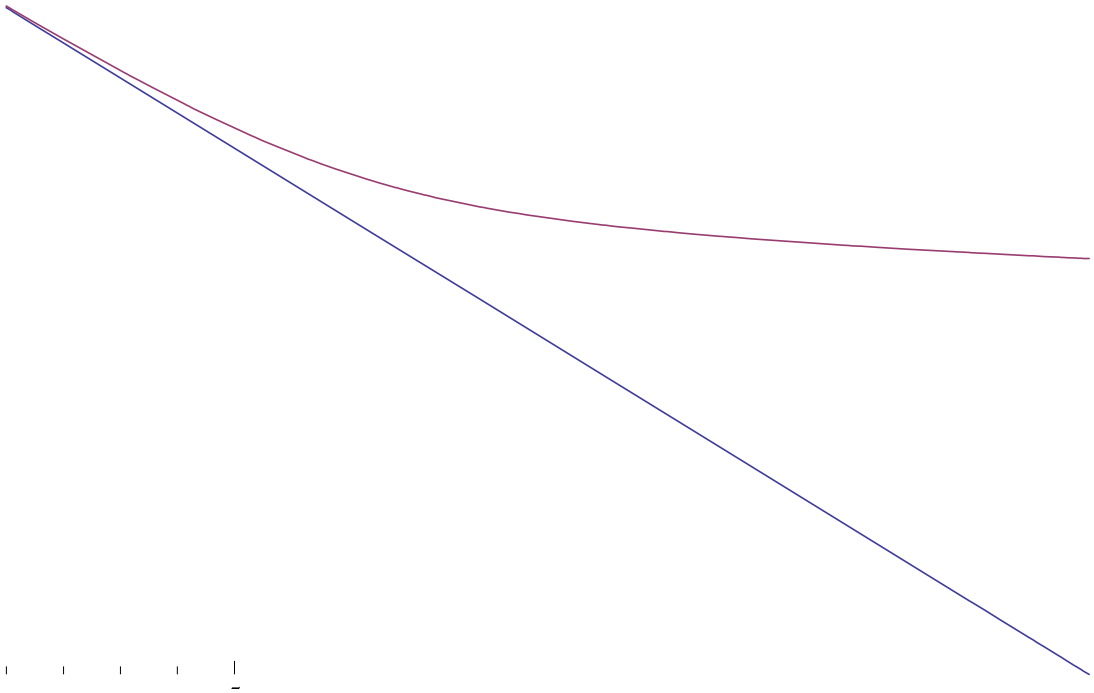
A single admixture event

$$\mathbb{E}(k_2) = \frac{1}{L^2}\int_0^L\int_0^L \begin{pmatrix} 1 & -1 \end{pmatrix} \cdot (\mathbf{L}_2\mathbf{U}_2)^g \, \mathbf{v}_{2(0)} d\ell d\ell'$$
$$= \frac{1}{L^2}\left(1-\frac{1}{N}\right)^g (s_{1,0} - s_{1,0}^2)\int_0^L\int_0^L [\ell\ell']^g d\ell d\ell'.$$

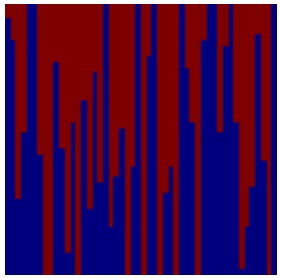
$$\mathbf{U}_2 = \begin{pmatrix} [\ell\ell'] & [\ell|\ell'] \\ 0 & 1 \end{pmatrix}$$

$$\mathbf{L}_2 = \begin{pmatrix} 1 & 0 \\ 1/N & 1-1/N \end{pmatrix}$$

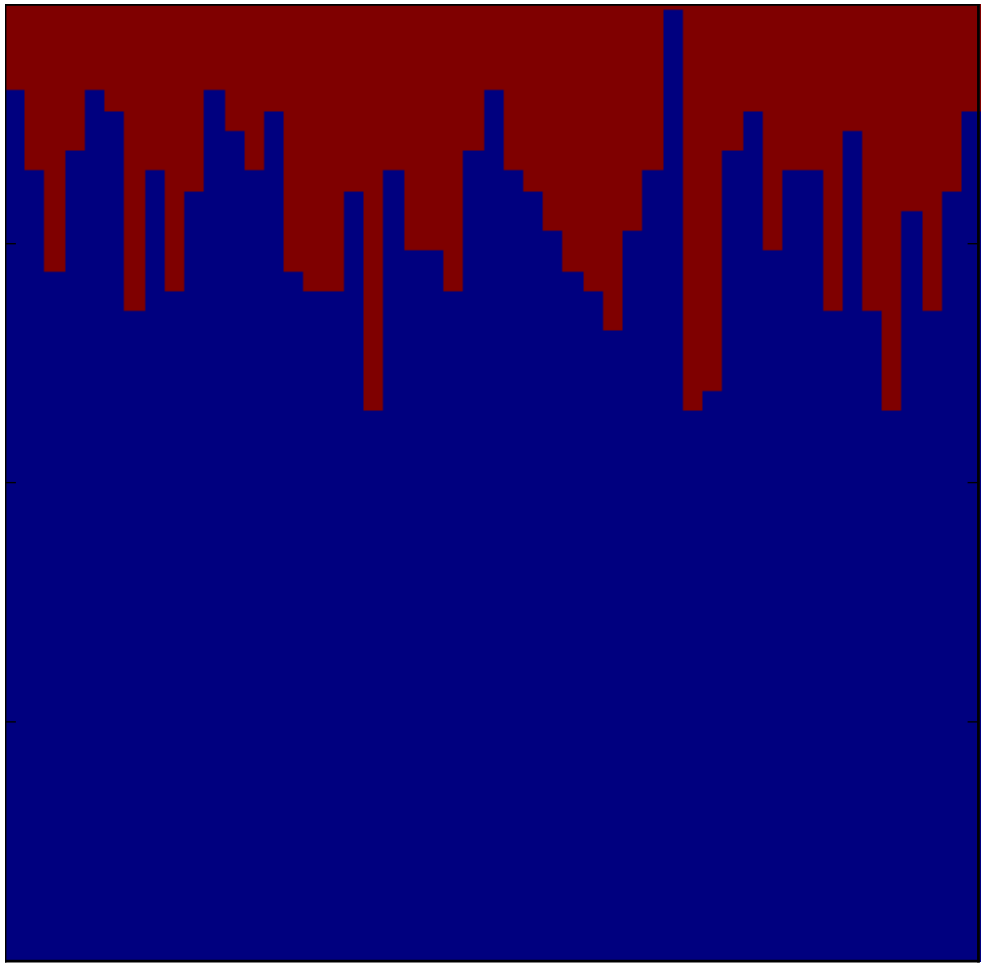
$$[\ell\ell'] = \frac{1+\exp(-2|\ell-\ell'|)}{2}$$

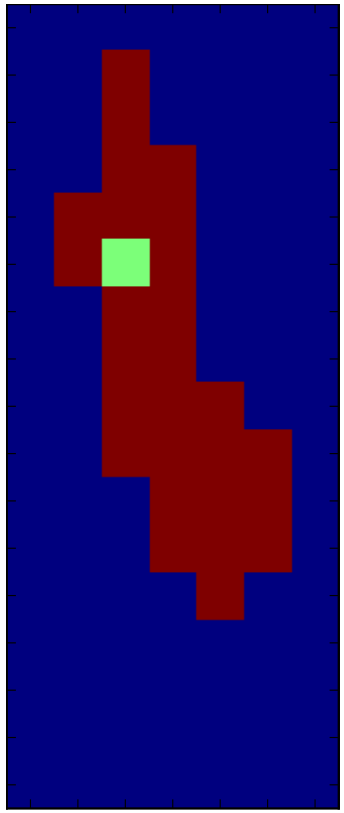


2011] and equation 1, plotted on a logarithmic scale. The variance we predict is always larger, but the two a very similar when g is small.



1000 Genomes Data





s colored green.

Conclusions (admixture components)

- We have analytical formulas for the moments of the admixture proportion distribution for multiple models.
- These results can be used as an additional method for estimating parameters of the admixture distribution.

Mason Liang

2. Admixture tracts

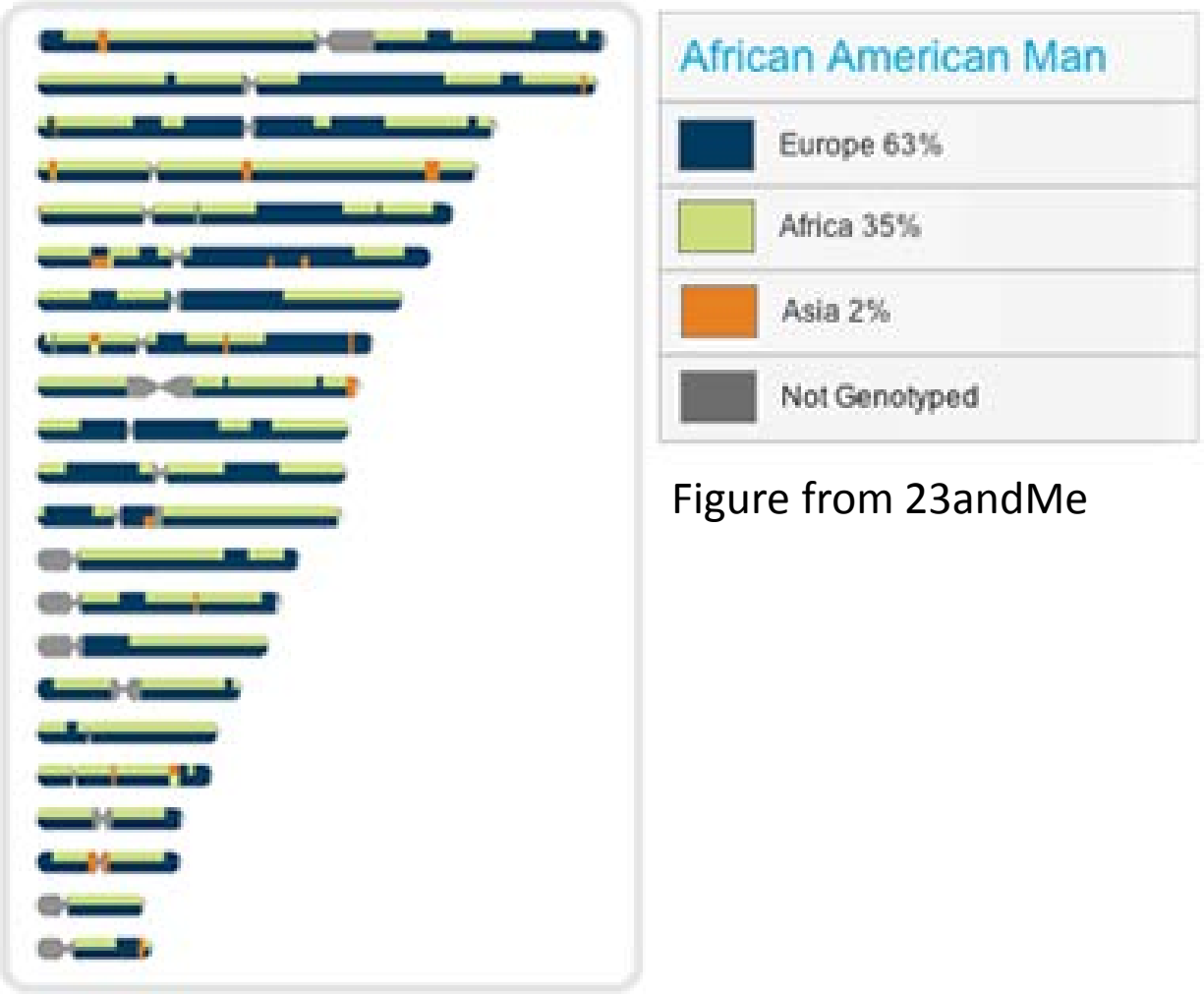
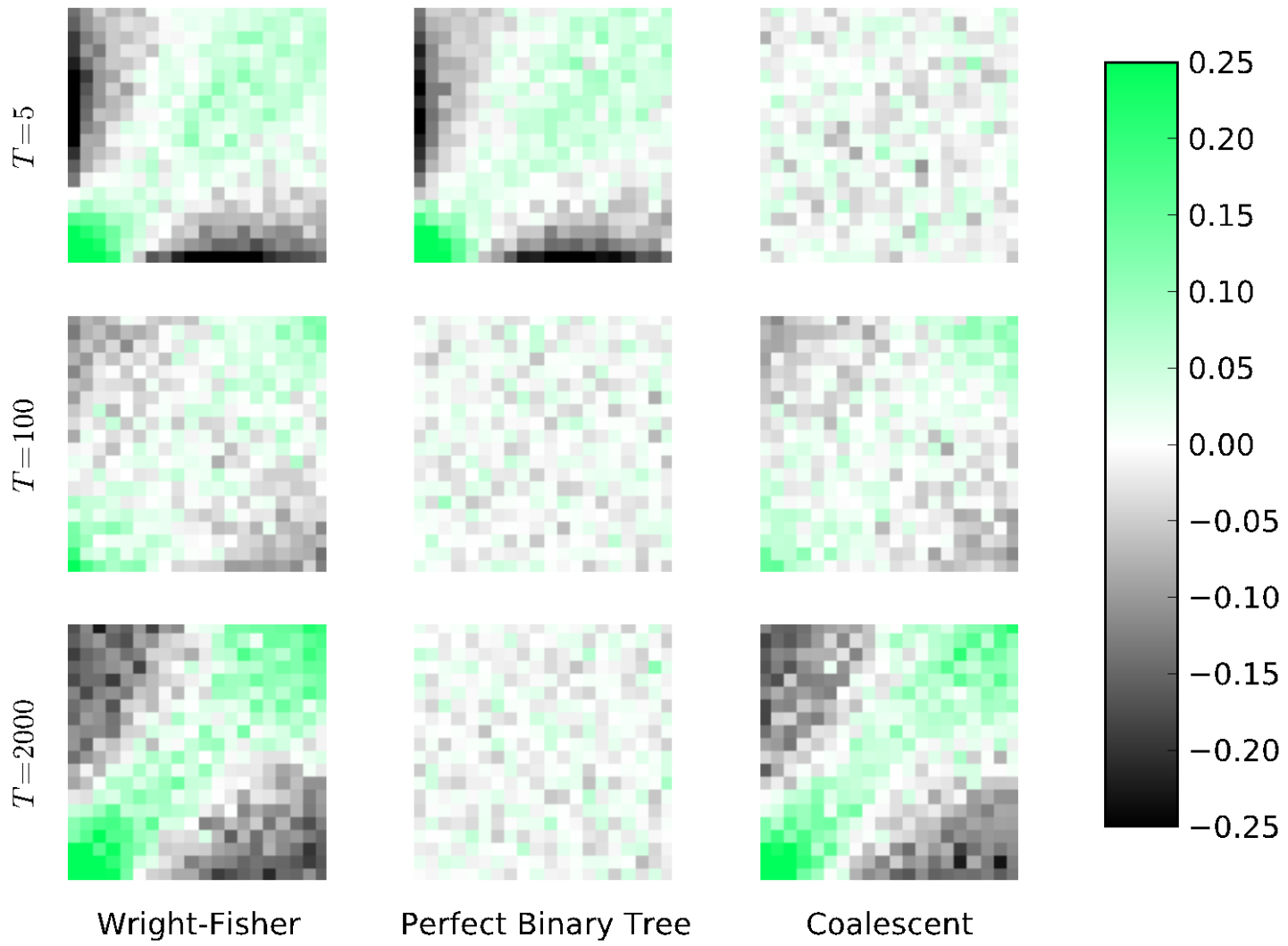
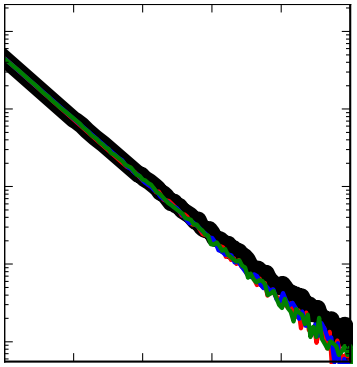


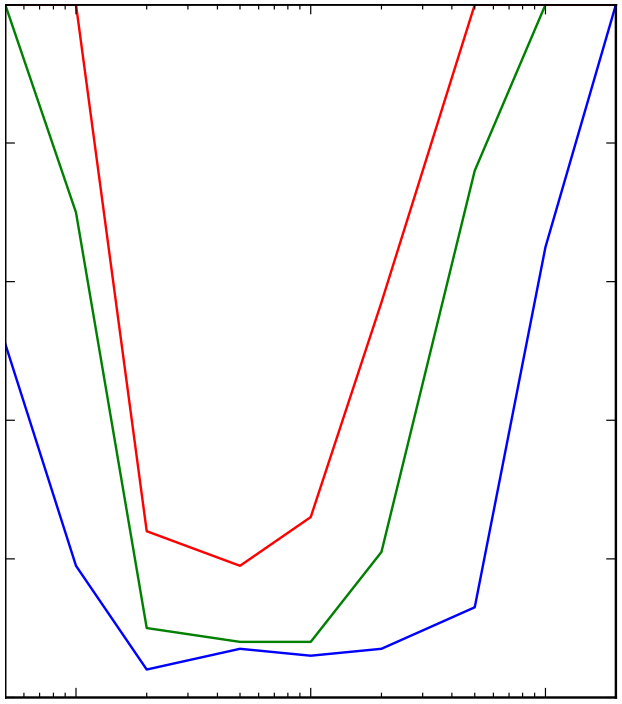
Figure from 23andMe

Mason Liang


$T=5$
$T=100$
$T=2000$
Wright-Fisher
Perfect Binary Tree
Coalescent
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
−0.25



Black: WF
Red: Coalescent
Blue: SMC
Green: SMC’



el. The red, green, and blue lines correspond to $m = 0.5$, 0.3, and 0.1. The left plot is for a likelihood-ratio test with $\alpha = 0.05$ and the right plot is with $\alpha = 0.002$.

Conclusions (tract lengths)

- Tracts lengths are clustered spatially.
- Tracts lengths are not exponentially distributed.
- Inferences of demographic parameters from tracts will be biased if not taking these two factors into account.

Mason Liang

Additional Challenges

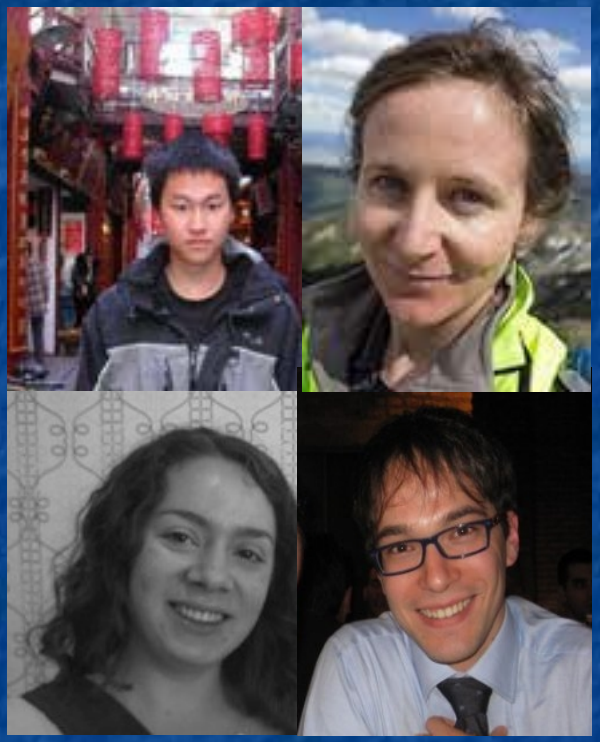
(1) Each site in a genome may have it's own tree.

(2) Missing data and errors.

Acknowledgements

Polar Bears:

Shiping Liu, Eline D. Lorenzen, **Matteo Fumagalli,** Bo Li, **Kelley Harris**, Zijun Xiong,Long Zhou, Thorfinn Sand Korneliussen, Mehmet Somel, Courtney Babbitt, Greg Wray, Jianwen Li, Weiming He, Zhuo Wang, Wenjing Fu, Xueyan Xiang, Claire C. Morgan, Aoife Doherty, Mary J. O'Connell, James O. McInerney, Erik W. Born, Love Dale , Rune Dietz, Ludovic Orlando, Christian Sonne, Guojie Zhang, Eske Willerslev, Jun Wang.



Admixture tracts.

Mason Liang

IBS tracts.

Kelley Harris