Ancestral inference on coalescent histories

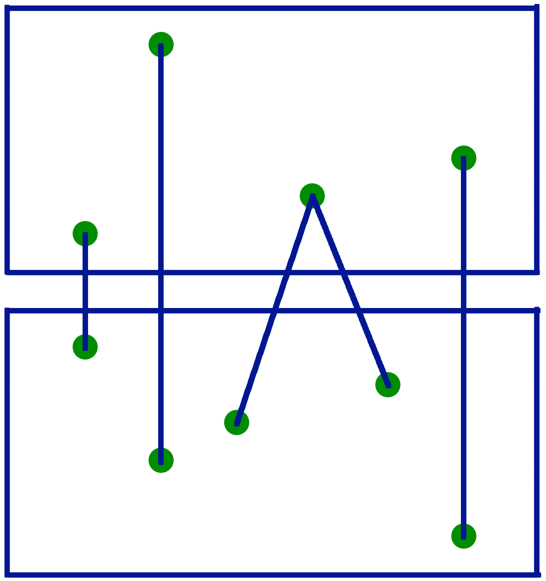
Bob Griffiths
University of Oxford

Collaborators: Melanie Bahlo, Ignazio Carbone, Graham Coop, Yvonne Griffiths, Rosalind Harding, Maria De Iorio, Paul Jenkins, Yun Song, Simon Tavaré

Wright-Fisher model

A population of M genes.
Generation $k+1$ is formed from generation k by choosing M genes at random with replacement.

Generation k



Generation $k+1$

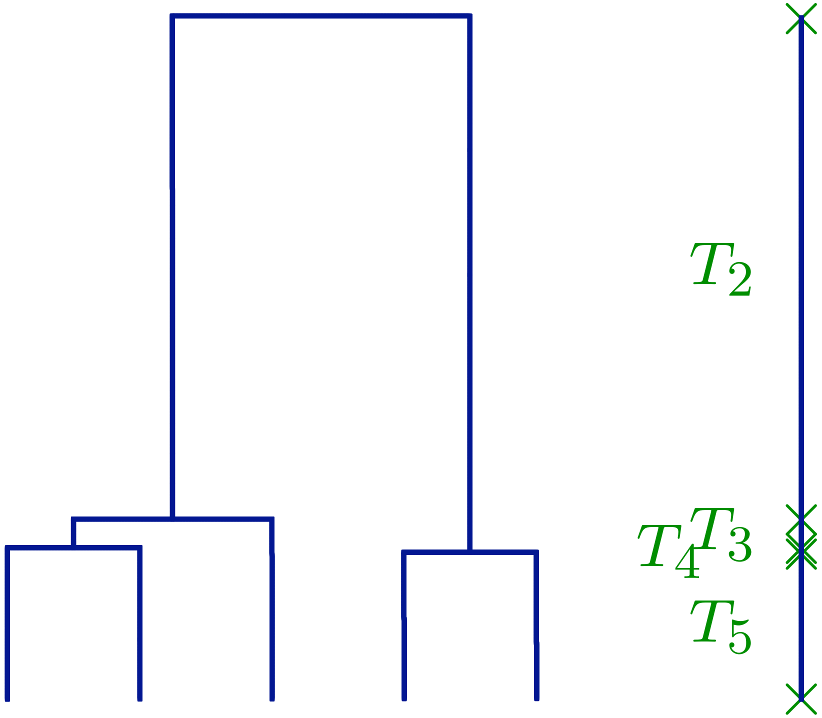
The Coalescent

Kingman, J. F. C. (1982). The coalescent.
Stochastic Processes and their Applications **13**, 235–248.

If time is measured in units of M generations, and $M \to \infty$, then the ancestral tree of n genes back in time in the Wright-Fisher model converges to a coalescent tree.

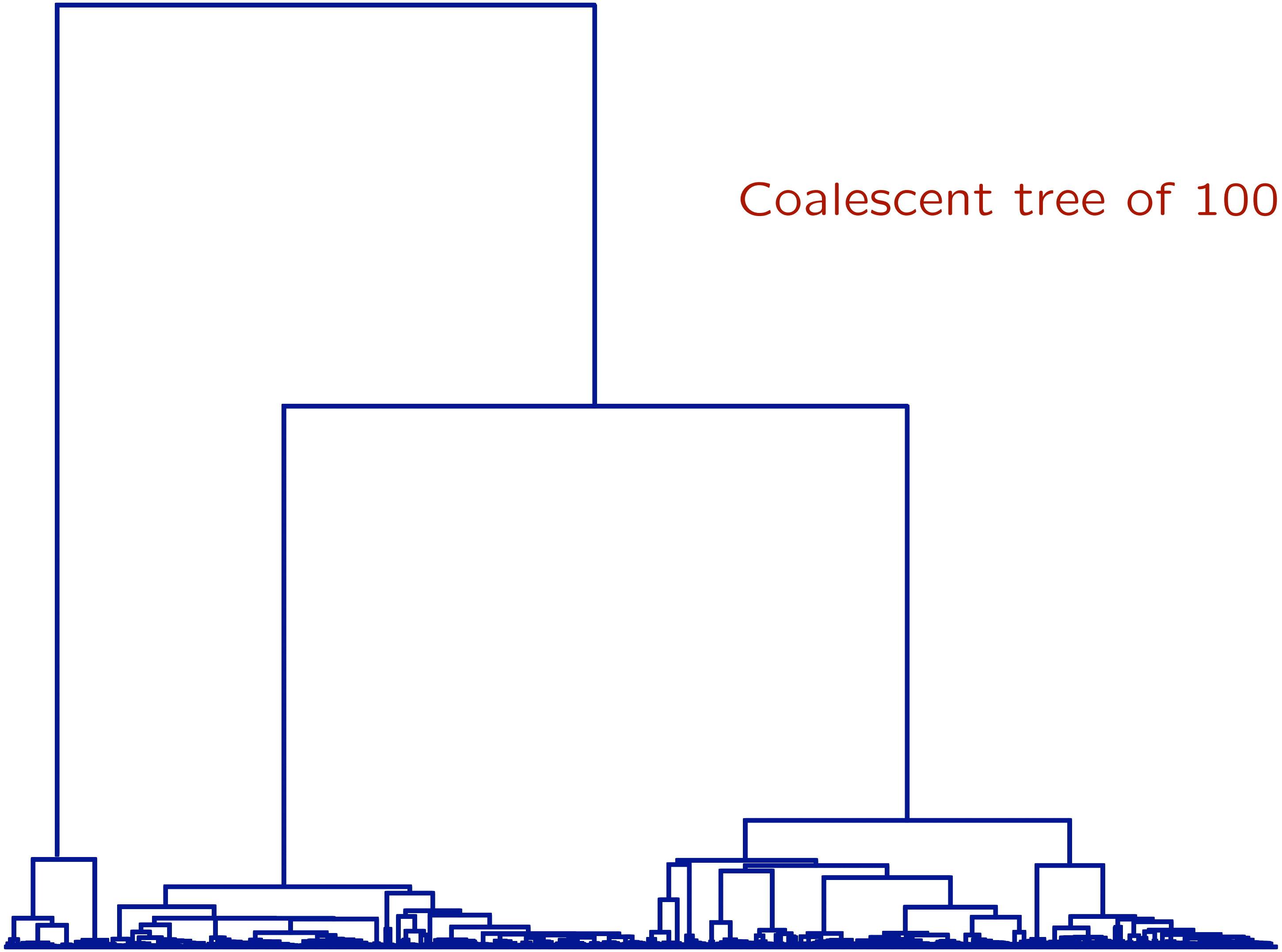
Two ancestral lineages coalesce when they have a common ancestor forming a vertex in the tree.

Coalescent tree



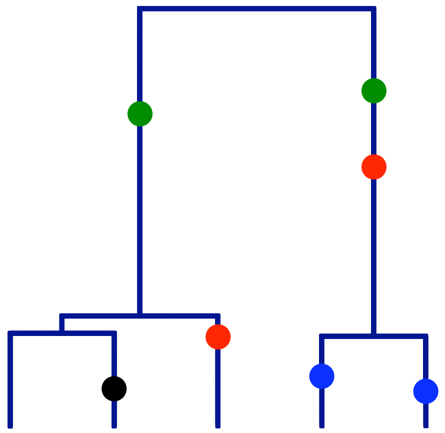
$\{T_j; j = n, \ldots, 2\}$ are independent exponential random variables with rates $\{\binom{j}{2}; j = n, \ldots, 2\}$.

Mean time to the most recent common ancestor is $2(1 - 1/n)$.



Coalescent tree of 1000

Coalescent tree with mutations



Coalescence of edges occurs at rate $\binom{k}{2}$ when k edges.

Mutations occur at a rate of $\theta/2$ on the edges of the coalescent tree in the coalescent time scale according to a Poisson process, given the edge lengths of the tree. Gene type space $E = \{1, 2, \ldots, d\}$. A type change occurs by mutation from a parent to an offspring according to a transition matrix P.

Population gene frequency distribution

Gene type space $E = \{1, 2, \ldots, d\}$. A type change occurs by mutation from a parent to an offspring according to a transition matrix P.

Population frequencies of types of genes $(X_i)_{i \in E}$ are distributed according to the stationary distribution in a diffusion process with state space $\{\mathbf{x} \in [0,1]^d : \sum_1^d x_i = 1\}$ and generator

$$\mathcal{L} = \frac{1}{2} \sum_{i,j \in E} x_i(\delta_{ij} - x_j) \frac{\partial^2}{\partial x_i \partial x_j} + \sum_{j \in E} \left(\sum_{i \in E} x_i r_{ij} \right) \frac{\partial}{\partial x_j}$$

where $R = \frac{\theta}{2}(P - I)$.

Sample distribution

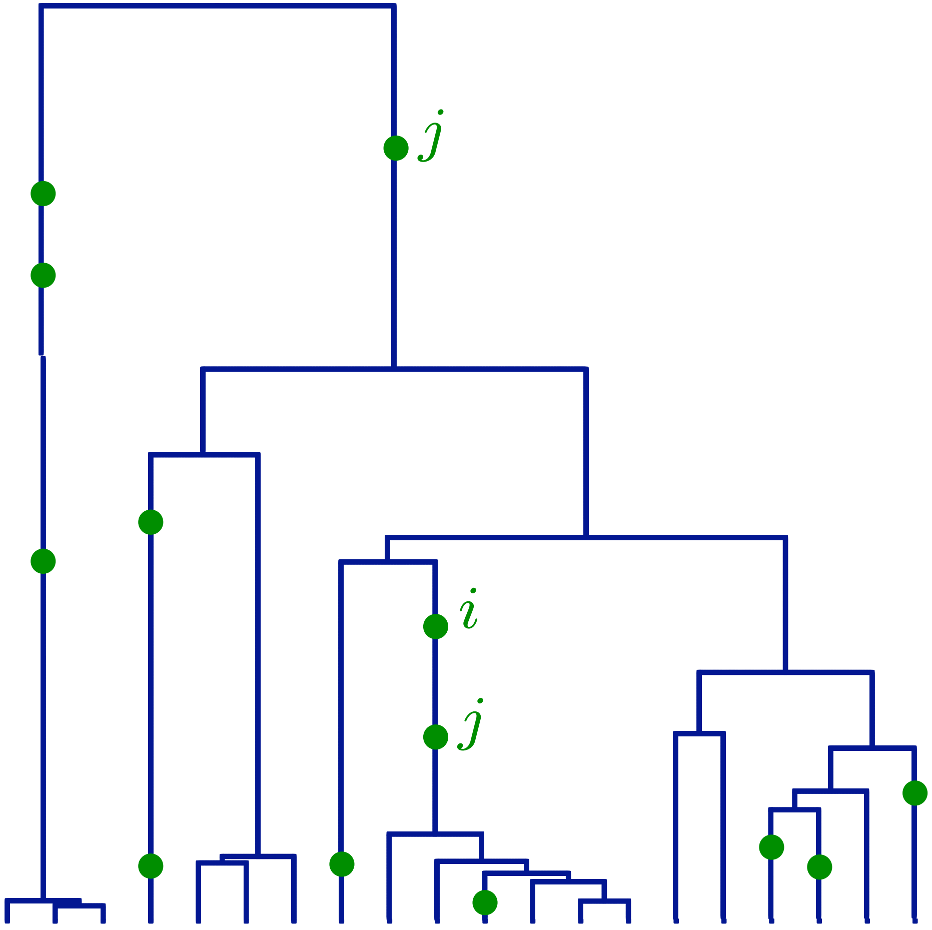
The sample distribution is a multinomial mixture

$$p(\mathbf{n}) = \frac{n!}{\prod_{i\in E} n_i!} E\left(\prod_{i\in E} X_i^{n_i}\right)$$

where E denotes expectation in the stationary distribution of the diffusion process. From the diffusion generator, or a coalescent tree

$$\begin{aligned} p(\mathbf{n}) \;&=\; \frac{\theta}{n+\theta-1} \sum_{i,j\in E, n_j>0} \frac{n_i+1-\delta_{ij}}{n} P_{ij} p(\mathbf{n}+\mathbf{e}_i-\mathbf{e}_j) \\ &+ \frac{n-1}{n+\theta-1} \sum_{j\in E, n_j>0} \frac{n_j-1}{n-1} p(\mathbf{n}-\mathbf{e}_j). \end{aligned}$$

Coalescent history connection



H_{-m}

$H_{\ell-1} = \mathbf{n}' - \mathbf{e}_j$
$H_\ell = \mathbf{n}'$

$H_{k-1} = \mathbf{n} - \mathbf{e}_j + \mathbf{e}_i$
$H_k = \mathbf{n}$

H_0

Coalescent history

Coalescent history $\{H_k, k = 0, -1, \ldots, -m\}$ is defined as the set of ancestral configurations at the imbedded events in the Markov process where coalescence, mutation or other events take place.

H_0 is the state at the current time.
H_{-m} the state when a singleton ancestor is reached.

Ancestor configurations $H_k = \mathbf{n}$ in the coalescent history have distribution $p(H_k) = p(\mathbf{n})$

Sequential Importance Sampling on Coalescent Histories

Griffiths and Tavaré (1994), Stephens and Donnelly (2000). Let H_j be the history configuration of gene types at step j back in the coalescent process of the sample, where at each step either a mutation or coalescence has occurred back in time. $\{H_j; j = 0, -1, \ldots, -m\}$ is the history process of the sample. A single MRCA is reached at $-m$.

$$p(H_j) = \sum p(H_j \mid H'_{j-1}) p(H'_{j-1})$$

with summation over possible configurations H'_{j-1}. Forward transition probabilities $p(H_j \mid H'_{j-1})$ are known. $p(H_j)$ and $\{p(H'_{j-1})\}$ are unknown.

Reverse IS transition probabilities $\hat{p}(H'_{j-1} \mid H_j)$

$$p(H_j) = \sum \frac{p(H_j \mid H'_{j-1})}{\hat{p}(H'_{j-1} \mid H_j)} \hat{p}(H'_{j-1} \mid H_j) p(H'_{j-1}).$$

The importance sampling representation is

$$\begin{aligned} p(H_0) &= E_{\hat{p}}\Big[\frac{p(H_0 \mid H_{-1})}{\hat{p}(H_{-1} \mid H_0)} \cdots \frac{p(H_{-m+1} \mid H_{-m})}{\hat{p}(H_{-m} \mid H_{-m+1})} p(H_{-m})\Big] \\ &= E_{\hat{p}}\Big[\frac{p(H_0 \mid H_{-1}) \ldots p(H_{-m+1} \mid H_{-m}) p(H_{-m})}{\hat{p}(H_{-m} \mid H_{-m+1}) \ldots \hat{p}(H_{-1} \mid H_0)}\Big] \\ &= E_{\hat{p}}\Big[\frac{p(\mathcal{H}_{\rightarrow})}{\hat{p}(\mathcal{H}_{\leftarrow})/p(H_0)}\Big]. \end{aligned}$$

Note that this is a general Markov chain construction.

Coalescent sampling distribution approximation

Let B_j be the event that a gene of type $j \in E$ is the first type to be involved in either a coalescent or mutation event back in time in a sample of n genes with type configuration $(n_j)_{j \in E}$.

Sample Approximation

$$\widehat{p}(B_j \mid \mathbf{n}) = \frac{n_j}{n}$$

Heuristic argument: (P_{ij}) has a stationary distribution (P_j)

$$p(B_j) = P_j$$

Sample result approximates the true probability.

De Iorio and Griffiths (2004)

Equivalent Diffusion process generator approximation

$$\mathcal{L} = \frac{1}{2}\sum_{i,j\in E} x_i(\delta_{ij}-x_j)\frac{\partial^2}{\partial x_i \partial x_j} + \sum_{j\in E}\left(\sum_{i\in E} x_i r_{ij}\right)\frac{\partial}{\partial x_j} = \sum_{j\in E} L_j \frac{\partial}{\partial x_j}$$

In the stationary distribution of gene frequencies

$$E\Big(\mathcal{L}\binom{n}{\mathbf{n}}\prod_{i\in E} X_i^{n_i}\Big) = 0$$

Approximation leading to $\widehat{p}(\mathbf{n})$, $\widehat{p}(H_{k-1} \mid H_k)$:

$$E\Big(L_j \frac{\partial}{\partial X_j}\binom{n}{\mathbf{n}}\prod_{i\in E} X_i^{n_i}\Big) = 0, \text{ for each } j \in E$$

Reverse chain transition probabilities

Bayes' rule:

$$p(H_{k-1} \mid H_k) = p(H_k \mid H_{k-1})\frac{p(H_{k-1})}{p(H_k)}$$

Define $\pi(i \mid \mathbf{n})$ as the probability that an additional type chosen from the population is of type i, given a sample configuration of $\mathbf{n}$.

Reverse chain transition probabilities can be expressed in terms of π:

$$p(\mathbf{n}) = \frac{n}{n_j}\pi(j \mid \mathbf{n} - \mathbf{e}_j)p(\mathbf{n} - \mathbf{e}_j)$$

Importance sampling reverse proposal distrbutions

$$\mathbf{n} \to \mathbf{n} - \mathbf{e}_j \qquad \frac{n_j(n_j-1)}{n(n+\theta-1)}\frac{1}{\widehat{\pi}(j|\mathbf{n}-\mathbf{e}_j)}$$

$$\mathbf{n} \to \mathbf{n} + \mathbf{e}_i - \mathbf{e}_j \qquad \frac{\theta P_{ij}}{n+\theta-1}\frac{n_j}{n}\frac{\widehat{\pi}(i|\mathbf{n}-\mathbf{e}_j)}{\widehat{\pi}(j|\mathbf{n}-\mathbf{e}_j)}$$

Importance sampling weights

$$\mathbf{n} \to \mathbf{n} - \mathbf{e}_j \qquad \frac{n\widehat{\pi}(j|\mathbf{n}-\mathbf{e}_j)}{n_j}$$

$$\mathbf{n} \to \mathbf{n} + \mathbf{e}_i - \mathbf{e}_j \qquad \frac{n_i+1-\delta_{ij}}{n_j}\frac{\widehat{\pi}(j|\mathbf{n}-\mathbf{e}_j)}{\widehat{\pi}(i|\mathbf{n}-\mathbf{e}_j)}$$

Proposal probability × weight = coefficient in recursion

$$\begin{aligned} p(\mathbf{n}) &= \frac{\theta}{n+\theta-1} \sum_{i,j\in E, n_j>0} \frac{n_i+1-\delta_{ij}}{n} P_{ij} p(\mathbf{n}+\mathbf{e}_i-\mathbf{e}_j) \\ &\quad + \frac{n-1}{n+\theta-1} \sum_{j\in E, n_j>0} \frac{n_j-1}{n-1} p(\mathbf{n}-\mathbf{e}_j). \end{aligned}$$

Stephens and Donnelly's approximation

Approximate π by $\widehat{\pi}$, defined by

$$\widehat{\pi}(j \mid \mathbf{n}) = \sum_{i \in E} \frac{n_i}{n} \sum_{\ell=0}^{\infty} \Big(\frac{\theta}{n+\theta}\Big)^{\ell} \frac{n}{n+\theta} P_{ij}^{\ell}$$

De Iorio and Griffiths: Approximate the generator describing the population frequencies giving a general way to obtain $\widehat{\pi}(j \mid \mathbf{n})$ in more complex systems.

Coalescent approximation equivalent to the generator equation

Let B_j be the event that a gene of type $j \in E$ is the first to be involved in either a coalescent or mutation event back in time, and $\mathbf{Y}$ a random vector describing the configuration of types so that $P(\mathbf{Y} = \mathbf{n}) = p(\mathbf{n})$.

$$\begin{aligned} & P(\{\mathbf{Y} = \mathbf{n}\} \cap B_j) \\ & \quad = p(\mathbf{n}) P(B_j \mid \mathbf{Y} = \mathbf{n}) \\ & \quad = \frac{n-1}{n+\theta-1} \frac{n_j - 1}{n-1} p(\mathbf{n} - \mathbf{e}_j) \\ & \qquad + \frac{\theta}{n+\theta-1} \sum_{i \in E} \frac{n_i + 1 - \delta_{ij}}{n} P_{ij} p(\mathbf{n} + \mathbf{e}_i - \mathbf{e}_j) \end{aligned}$$

Approximate $P(B_j \mid \mathbf{Y} = \mathbf{n})$ by $\widehat{P}(B_j \mid \mathbf{Y} = \mathbf{n}) = n_j / n$.

Linear equation system for $\widehat{\pi}$

$$\widehat{\pi}(j \mid \mathbf{n} - \mathbf{e}_j) = \frac{n_j}{n} \frac{\widehat{p}(\mathbf{n})}{\widehat{p}(\mathbf{n} - \mathbf{e}_j)}$$

From the coalescent approximation

$$\widehat{\pi}(j \mid \mathbf{n} - \mathbf{e}_j) = \frac{n_j - 1}{n + \theta - 1} + \sum_{i \in E} \frac{\theta}{n + \theta - 1} P_{ij} \widehat{\pi}(i \mid \mathbf{n} - \mathbf{e}_j)$$

Solution of the system is

$$\widehat{\pi}(j \mid \mathbf{n}) = \sum_{i \in E} \frac{n_i}{n} \sum_{\ell=0}^{\infty} \Big(\frac{\theta}{n + \theta}\Big)^{\ell} \frac{n}{n + \theta} P_{ij}^{\ell}$$

The parent independent mutation model has a mutation matrix P with rows $(p_1, \ldots, p_d)$, and

$$\pi(j \mid \mathbf{n}) = \widehat{\pi}(j \mid \mathbf{n}) = \frac{n_j}{n+\theta} + \frac{\theta}{n+\theta} p_j .$$

The stepwise mutation model in its simplest form has a mutation matrix $P_{ij} = 1/2$ if $|i-j| = 1$, and

$$\widehat{\pi}(j \mid \mathbf{n}) = \sum_{i \in E} \frac{n_i}{n} Q_{ij} \, ,$$

where

$$Q_{ij} = \frac{1-\rho}{\sqrt{1-\rho^2}} \cdot \left[\frac{\rho}{1+\sqrt{1-\rho^2}} \right]^{|j-i|} ,$$

with $\rho = \theta/(n+\theta)$. The jump $Z = j - i$ is distributed as a two-sided geometric for $Z > 0$ with an additional atom at 0

A model with migration

Subpopulations labelled by $\Gamma = \{1, \ldots, g\}$.
Relative subpopulation sizes are $(q_\alpha)_{\alpha \in \Gamma}$.

Types of genes $E = \{1, \ldots d\}$.
Mutation rate θ. Mutation transition matrix P.

Backward migration rates $M = (m_{\alpha\beta})_{\alpha\beta \in \Gamma}$.

Forward migration rates are $\widetilde{m}_{\beta\alpha} = q_\alpha m_{\alpha\beta} q_\beta^{-1}$

$X_{\alpha j}$ is the relative frequency of gene type j
in subpopulation α.
$\sum_{j \in E} X_{\alpha j} = 1$ for each $\alpha \in \Gamma$.

Diffusion process migration model

Infinitesimal means and covariances

$$E\Big(X_{\alpha j}(t+dt) \mid \{X_{\alpha i}(t)=x_{\alpha i}\}\Big)$$

$$= \Big(1-\Big[\frac{\theta}{2}+\frac{\widetilde{m}_\alpha}{2}\Big]dt\Big)x_{\alpha j} + \sum_{i\in E}\frac{\theta}{2}x_{\alpha i}P_{ij}dt$$

$$+ \sum_{\beta\neq\alpha} x_{\beta j}\frac{1}{2}\widetilde{m}_{\beta\alpha}q_\beta q_\alpha^{-1}dt$$

$$\text{Cov}\Big(X_{\alpha i}(t+dt), X_{\alpha j}(t+dt) \mid \{X_{\alpha i}(t)=x_{\alpha i}\}\Big)$$

$$= \Big(x_{\alpha i}\delta_{ij} - x_{\alpha i}x_{\alpha j}\Big)q_\alpha^{-1}dt$$

Reverse transition probabilities $H_k \longrightarrow H_{k-1}$

H_{k-1}	Proposal distribution
$\mathbf{n} - \mathbf{e}_{\alpha j}$	$\dfrac{n_{\alpha j}(n_{\alpha j}-1)q_{\alpha}^{-1}}{\widehat{\pi}(j \mid \alpha, \mathbf{n}-\mathbf{e}_{\alpha j})D(\mathbf{n})}$
$\mathbf{n} - \mathbf{e}_{\alpha j} + \mathbf{e}_{\alpha i}$	$\dfrac{n_{\alpha j}\theta P_{ij}\widehat{\pi}(i \mid \alpha, \mathbf{n}-\mathbf{e}_{\alpha j})}{\widehat{\pi}(j \mid \alpha, \mathbf{n}-\mathbf{e}_{\alpha j})D(\mathbf{n})}$
$\mathbf{n} - \mathbf{e}_{\alpha j} + \mathbf{e}_{\beta j}$	$\dfrac{n_{\alpha j}m_{\alpha\beta}\widehat{\pi}(j \mid \beta, \mathbf{n}-\mathbf{e}_{\alpha j})}{\widehat{\pi}(j \mid \alpha, \mathbf{n}-\mathbf{e}_{\alpha j})D(\mathbf{n})}$

where $D(\mathbf{n})$ is a scale constant.

Interpretation of the distribution $\widehat{\pi}(j \mid \alpha, \mathbf{n})$
from a walk in subpopulations

$\widehat{\pi}(j \mid \alpha, \mathbf{n})$ is the probability of choosing a type j gene from a fixed subpopulation α.

$M^{\circ} = (m_{\alpha\beta}/m_{\alpha})$ is a transition probability matrix constructed from the migration rate matrix M.

Denote

$$\phi_{\alpha} = \frac{m_{\alpha}}{n_{\alpha}q_{\alpha}^{-1} + m_{\alpha}} \quad \rho_{\alpha} = \frac{\theta}{n_{\alpha}q_{\alpha}^{-1} + m_{\alpha} + \theta}$$

and the transition probability matrix

$$P_{\alpha} = (1 - \rho_{\alpha})(I - \rho_{\alpha}P)^{-1}$$

Choosing a gene with type distribution $\widehat{\pi}(j \mid \alpha, \mathbf{n})$

Subpopulation walk

Choose a sequence of subpopulations starting with $\alpha_0 = \alpha$ and stopping at step τ in subpopulation α_τ, $\alpha_0, \alpha_1, \dots, \alpha_\tau$, with probability

$$\phi_{\alpha_0}\phi_{\alpha_1}\cdots\phi_{\alpha_{\tau-1}}(1-\phi_\tau)\cdot m^\circ_{\alpha_0\alpha_1}m^\circ_{\alpha_1\alpha_2}\cdots m^\circ_{\alpha_{\tau-1}\alpha_\tau}$$

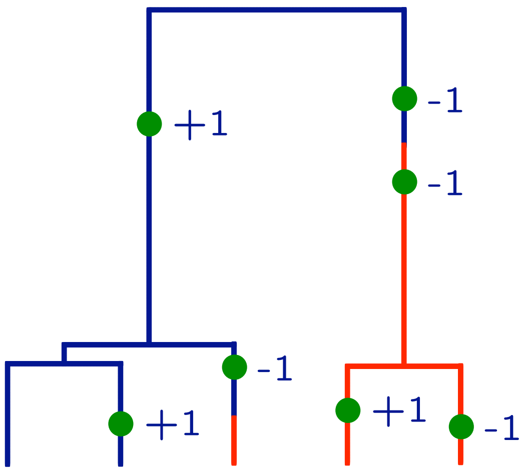
ϕ_α can be interpreted as the probability of moving from subpopulation α.

Mutations

Next choose a type at random from subpopulation α_τ, so that the probability of choosing a gene of type i is $n_{\alpha_\tau i}/n_{\alpha_\tau}$. Mutate back along the migration path to α_0, so that a sample path probability of mutations which start with type i and end with a type $i_0 = j$ gene is

$$\frac{n_{\alpha_\tau i_{\alpha_\tau}}}{n_{\alpha_\tau}} P_{\alpha_\tau;i_\tau i_{\tau-1}} \cdots P_{\alpha_1;i_2 i_1} P_{\alpha_0;i_1 i_0}$$

Microsatellite model



Mutations occur at a rate of $\theta/2$. The charge of each mutation is $+1, -1$ with probability 0.5, 0.5. Types of genes at the leaves of the tree are the sum of the ±1 charges along the path to the root. Gene type space $E = \{\ldots, -1, 0, 1, \ldots\}$.

Lineages may migrate between subpopulations.

Human Microsatellite data

Marshfield data sets. Published data for 21 autosomal, dinucleotide loci, selected as independent loci.

Location, haploid sample size (n) and Mean Expected Heterozygosity (H)

Population	Latitude/Longitude	n	H
Mbuti Pygmies (sub-Saharan Africa)	1°N, 29°E	30	0.816
Biaka Pygmies (sub-Saharan Africa)	4°N, 17°E	68	0.807
Yorubans (sub-Saharan Africa)	6-10N°N, 2-8°E	50	0.796
Palestinians (central Israel)	32°N, 35°E	102	0.785
Makrani (Pakistan)	26°N, 62-66°E	50	0.791
Han (China)	26-39°N, 108-120°E	90	0.704
French (Europe)	46°N, 2°E	58	0.762

Central African Pygmies, Biaka and Mbuti

21 independent microsatellite loci

Length	Biaka	Mbuti
16	2	
15	1	
14	2	
13		1
12	3	
11	13	3
10		
9		
8	3	7
7	4	7
6	33	2
5		
4	1	
3	8	1
2	2	5
1		4
0		
Total	72	30

Length	Biaka	Mbuti
13	2	5
12		1
11	1	4
10	27	1
9	6	4
8	5	10
7	8	
6	23	5
5		
4		
3		
2		
0		
Total	72	30

Mutation rates and effective population sizes

Estimates of relative sizes, mutation rates (θ_1, θ_2), and effective population sizes (N_1, N_2) for pairs of populations. Dinucleotide microsatellite mutation rate taken as 1.52×10^{-3} from Zhivotovsky et. al. (2003).

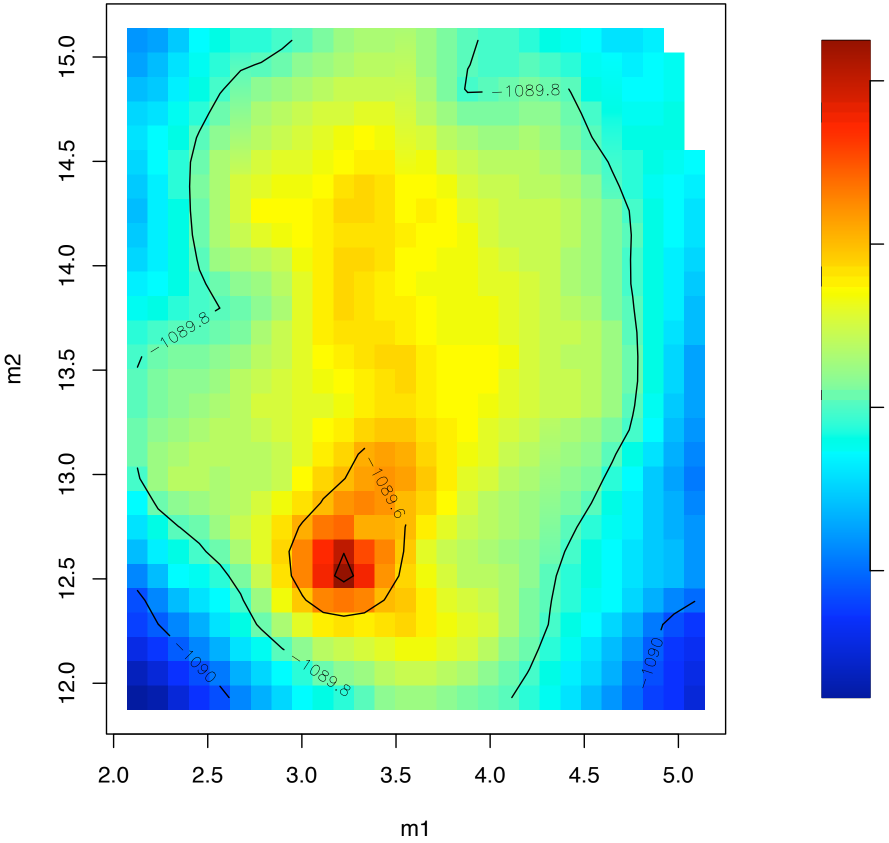
Populations	q_1, q_2	θ_1, θ_2	N_1, N_2
Mbuti	0.725	13.56	4459
Biaka	0.275	5.14	1692
Mbuti	0.55	9.90	3257
Yorubans	0.45	8.10	2664
Mbuti	0.65	11.77	3870
Palestinians	0.35	6.33	2084
Mbuti	0.65	10.85	3570
Makrani	0.35	5.85	1924
Mbuti	0.725	11.89	3911
Han	0.275	4.51	1484
Mbuti	0.725	12.32	4054
French	0.275	4.68	1538

Populations	q_1, q_2	θ_1, θ_2	N_1, N_2
Biaka	0.3	4.35	1431
Yorubans	0.7	10.15	3339
Biaka	0.5	7.15	2352
Palestinians	0.5	7.15	2352
Biaka	0.5	7.20	2368
Makrani	0.5	7.20	2368
Biaka	0.65	7.73	2544
Han	0.35	4.17	1370

Estimates of migration parameters for population pairs.
m_1, m_2 are scaled backward rates,
v_1, v_2 are rates per lineage per generation in units of 10^{-3}

Populations	q_1, q_2	m_1, m_2	v_1, v_2
Mbuti	0.725	2.2	0.18
Biaka	0.275	19.3	1.57
Mbuti	0.55	10.0	0.84
Yorubans	0.45	9.0	0.76
Mbuti	0.65	2.5	0.21
Palestinians	0.35	7.9	0.66
Mbuti	0.65	4.1	0.37
Makrani	0.35	7.8	0.71
Mbuti	0.725	2.1	0.19
Han	0.275	7.0	0.65
Mbuti	0.725	1.6	0.14
French	0.275	8.7	0.78

Populations	q_1, q_2	m_1, m_2	v_1, v_2
Biaka	0.3	43.9	0.460
Yorubans	0.7	13.3	1.39
Biaka	0.5	6.6	0.70
Palestinians	0.5	5.7	0.61
Biaka	0.5	6.6	0.70
Makrani	0.5	4.4	0.46
Biaka	0.65	4.5	0.57
Han	0.35	6.4	0.82

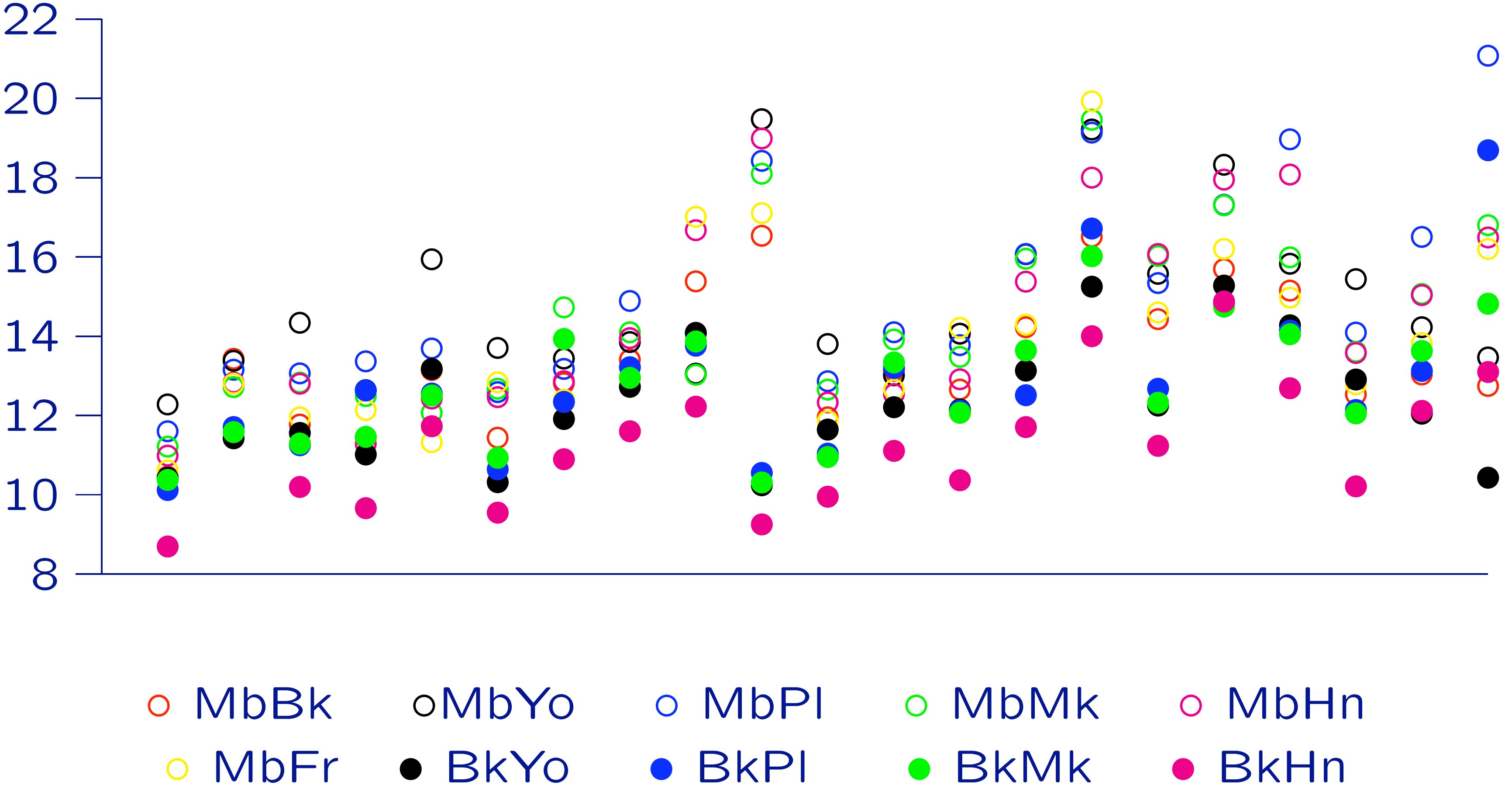


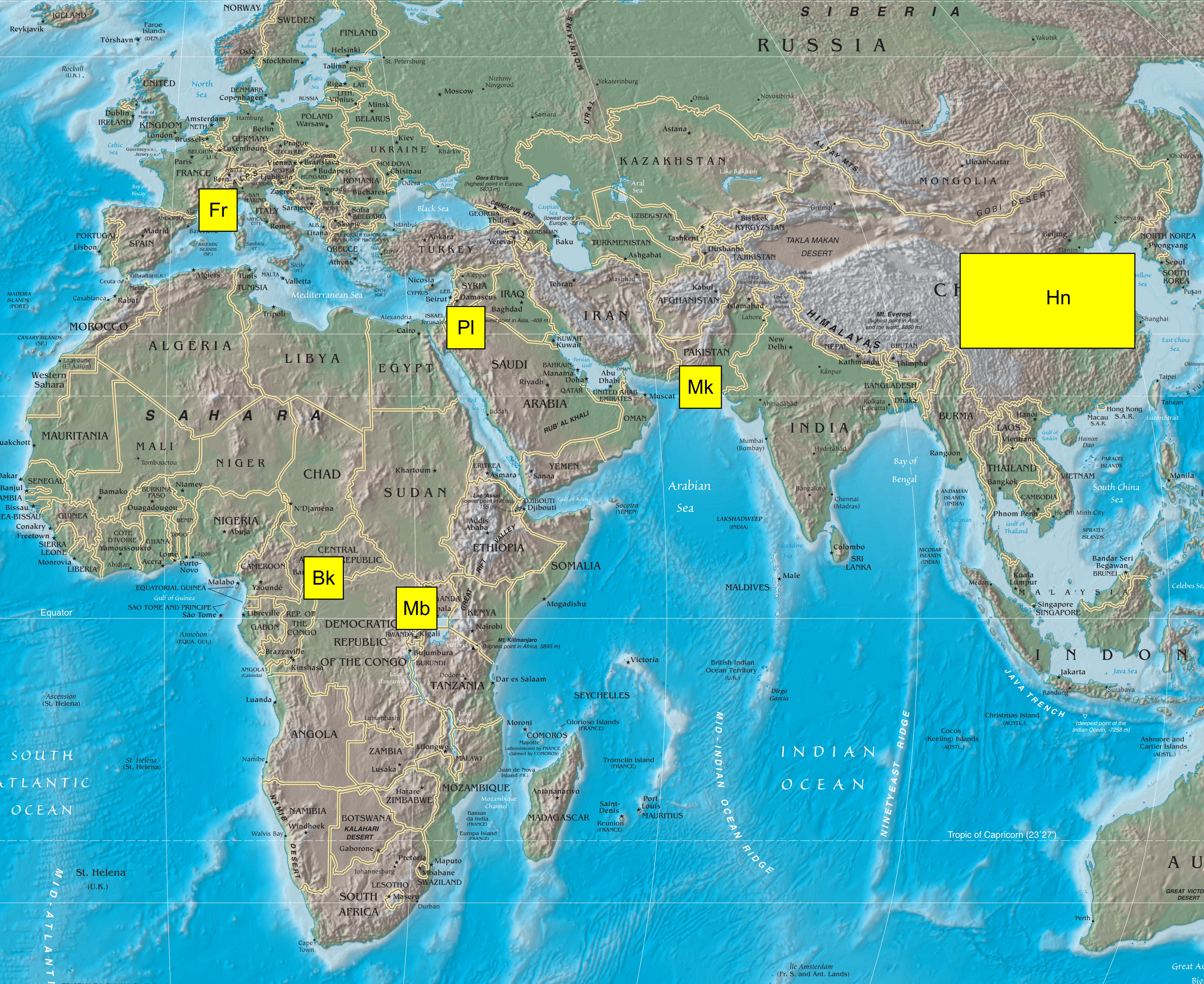
Likelihood contours for migration rates.

Mbuti, Biaka $m_1 = 3.3$, $m_2 = 12.6$.

Time to the most recent common ancestor (TMRCA)

Time units in 10^4 years. Loci in increasing order of range of allele size.




Fr
Pl
Mk
Hn
Bk
Mb
SIBERIA
RUSSIA
KAZAKHSTAN
MONGOLIA
GOBI DESERT
TAKLA MAKAN DESERT
HIMALAYAS
IRAN
AFGHANISTAN
PAKISTAN
INDIA
SAUDI ARABIA
TURKEY
EGYPT
LIBYA
ALGERIA
MOROCCO
SAHARA
MALI
NIGER
CHAD
SUDAN
ETHIOPIA
SOMALIA
NIGERIA
DEMOCRATIC REPUBLIC OF THE CONGO
TANZANIA
KENYA
ANGOLA
ZAMBIA
NAMIBIA
BOTSWANA
SOUTH AFRICA
MOZAMBIQUE
MADAGASCAR
UNITED KINGDOM
FRANCE
SPAIN
GERMANY
POLAND
UKRAINE
NORWAY
SWEDEN
FINLAND
Mediterranean Sea
Black Sea
Caspian Sea
Arabian Sea
Bay of Bengal
South China Sea
INDIAN OCEAN
SOUTH ATLANTIC OCEAN
Equator
Tropic of Capricorn (23°27')
MID-INDIAN OCEAN RIDGE
NINETYEAST RIDGE
JAVA TRENCH

DNA sequences and Gene Trees

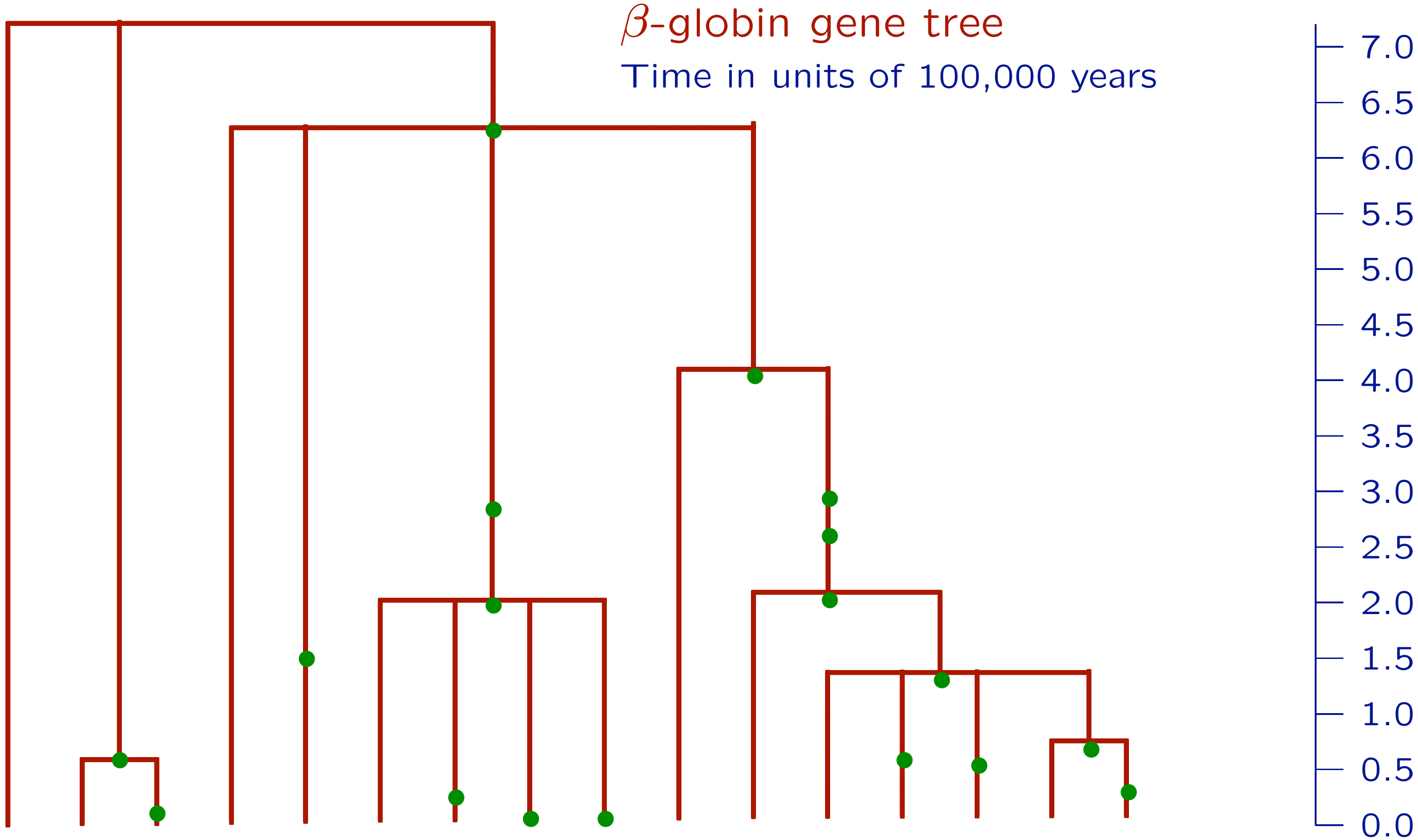
β-globin data sequences

Root	C	T	T	T	A	C	C	T	T	C	T	G	G	G	C	A	G	T	T	Freq
A1	A	T	T	T	A	C	C	T	G	C	T	G	G	G	C	A	G	T	G	104
A2	A	T	T	C	A	C	C	T	G	C	T	G	G	G	C	A	G	T	G	1
A3	A	T	T	T	A	C	C	T	G	C	T	A	G	G	C	A	G	T	G	8
A4	A	T	T	T	G	C	C	T	G	C	T	G	G	G	C	A	G	T	G	1
B1	A	T	T	T	A	C	C	T	T	C	T	G	G	G	C	T	G	T	T	79
B2	C	T	T	T	A	C	C	T	T	C	T	G	G	G	C	A	G	T	T	18
B3	A	T	T	T	A	C	C	T	T	C	T	G	G	G	C	A	G	T	T	9
B4	C	T	T	T	A	C	C	T	T	C	T	G	G	G	C	A	G	C	T	3
B9	A	T	T	T	A	C	C	T	T	C	T	G	G	G	A	A	G	T	T	2
B11	C	T	T	T	A	C	C	T	T	C	T	G	G	G	C	A	A	C	T	1
C1	A	C	C	T	A	T	G	T	T	C	C	G	G	G	A	A	G	T	T	48
C2	A	T	C	T	A	T	G	T	T	C	C	G	G	G	A	A	G	T	T	9
C3	A	T	C	T	A	T	C	T	T	C	C	G	G	G	A	A	G	T	T	10
C7	A	T	C	T	A	T	G	T	T	C	C	G	G	A	A	A	G	T	T	19
D1	A	T	C	T	A	T	G	T	T	T	C	G	C	G	A	A	G	T	T	13
D2	A	T	C	T	A	T	G	T	T	T	C	G	G	G	A	A	G	T	T	1

Harding R. M., Fullerton S. M., Griffiths R. C., Bond J., Cox M. J., Schneider J. A., Moulin D., and Clegg J. B. (1997).

Archaic African *and* Asian lineages in the genetic ancestry of modern humans

American Journal of Human Genetics, **60**, 772–789.



World	18	3	1	9	79	104	8	1	1	2	10	9	48	19	1	13
Pygmies	4	.	.	1	6	9	1	.	.	1	.	.	.	.	.	.
Gambia	6	3	.	2	5	8	2	.	.	.	.	.	1	.	1	.
Kenya	8	.	1	6	9	12	5	.	.	.	.	.	.	.	.	1
Mongolia	.	.	.	.	3	3	.	.	1	.	.	2	4	6	.	3
Amerind	.	.	.	.	2	15	.	.	.	.	.	6	22	.	.	1
PNG	.	.	.	.	12	1	.	.	.	.	7	.	.	4	.	.
Sumatra	.	.	.	.	10	8	.	.	.	.	.	1	14	6	.	.
UK	.	.	.	.	16	23	.	.	.	1	.	.	.	2	.	4
Vanuatu	.	.	.	.	16	25	.	1	.	.	3	.	7	1	.	4

A gene tree constructed from fungal data of Carbone and Kohn (2001)

Three species:
Sclerotinia sclerotiorum, *S. trifoliorum* and *S. minor*.

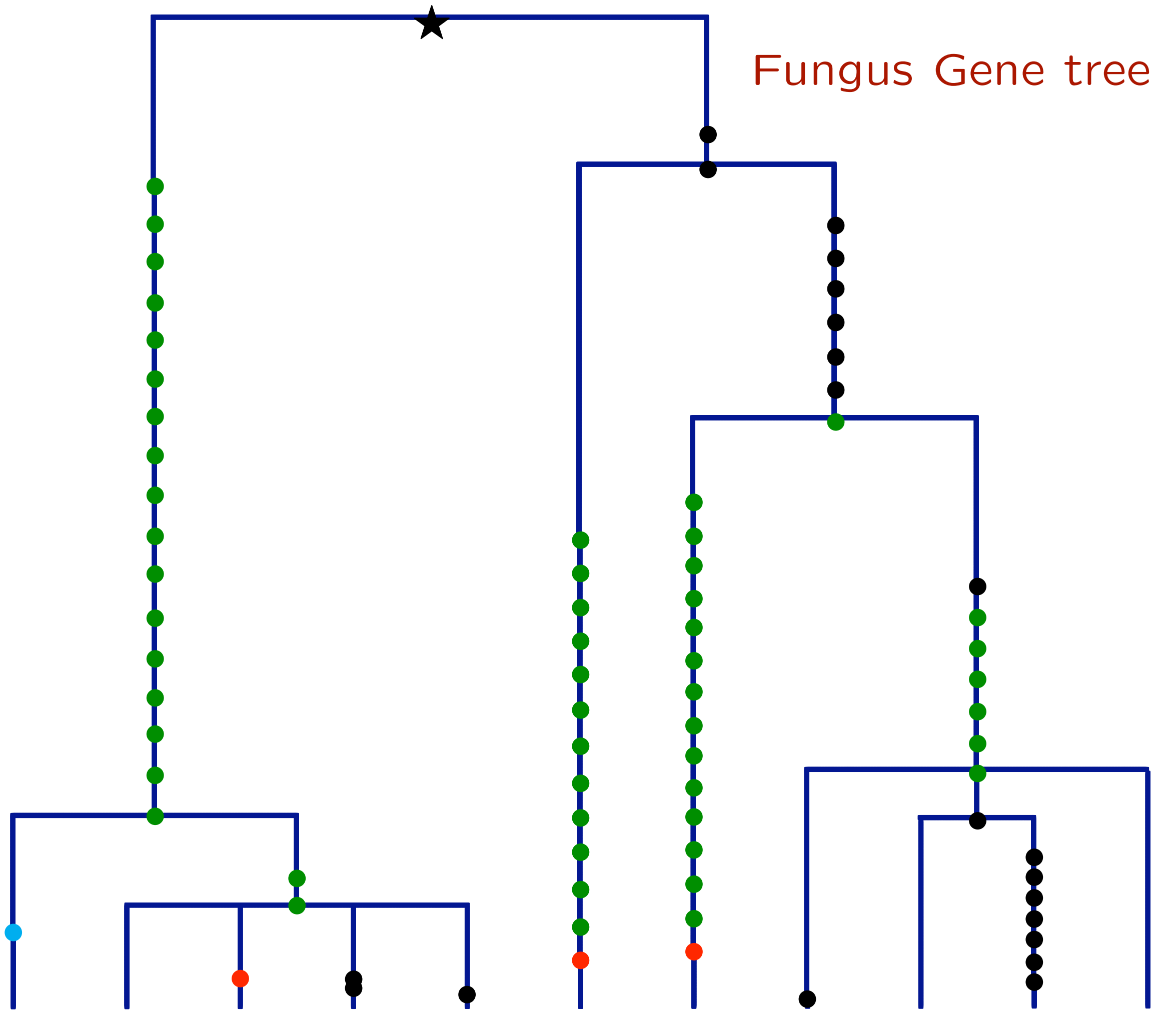
There are three sampling areas for these species, Temperate, Sub-tropical, and Wild.

From the original paper $\theta = 11.8$.

Backward migration matrix

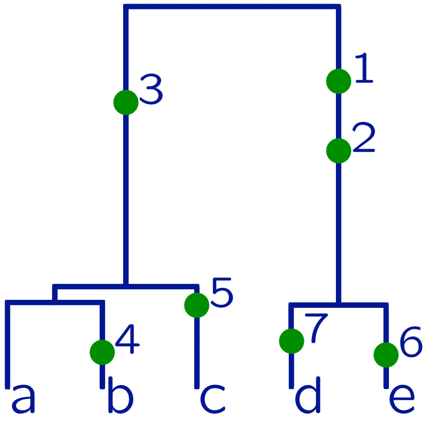
	T	S	W
T	-	2.0	1.1
S	0.5	-	0.4
W	1.0	1.3	-

Fungus Gene tree

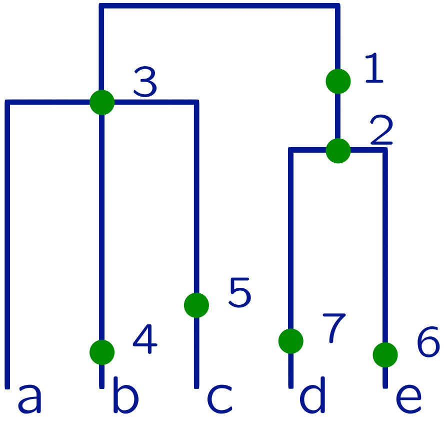


T	·	80	51	1	·	7	29	·	4	·	·
S	·	121	1	12	11	·	·	1	1	1	1
W	64	·	·	·	·	·	·	·	·	·	·

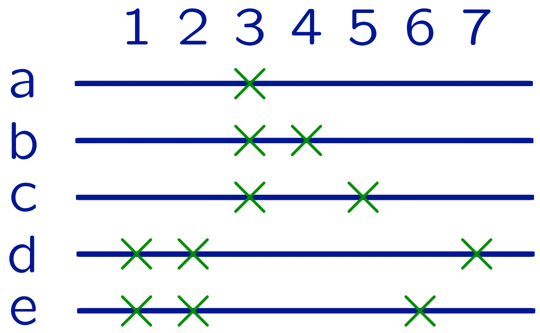
Coalescent tree.



Gene tree.



Mutation pattern on sequences



Gene tree ≡ Mutation pattern

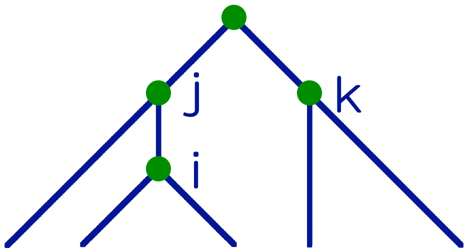
DNA sequences and Gene trees.

In a sample of n sequences suppose there are s segregating sites, corresponding to s mutations. Label the mutations $1, 2, \ldots, s$ and let $O_1, \ldots, O_s$ be the sets of sequences containing mutations $1, 2, \ldots, s$.

The sets $O_1, \ldots, O_s$ are partially ordered by inclusion, that is, for $i \neq j$ either

$$O_i \subset O_j, O_j \subset O_i, \text{ or } O_i \cap O_j = \phi.$$

The partial ordering is easy to see from a tree.



$O_i \subset O_j$, $O_j \cap O_k$ is empty.

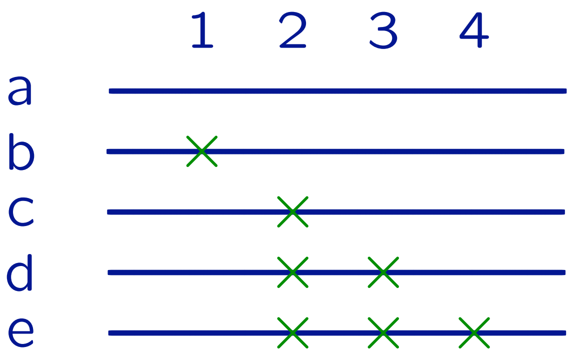
Gusfield's algorithm. Gusfield (1991). Efficient algorithms for inferring evolutionary trees. *Networks.* **21**, 19–28.

- Represent duplicate columns in the incidence matrix as a single column with a label corresponding to the identical columns, for example (1,6,8).

- Considering each column as a binary number, sort the numbers into decreasing order, with the largest number in column 1.

- Construct paths from the leaves to the root in the gene tree by labelling nodes by mutation column labels, and reading vertices in paths from the right to the left where 1's occur in rows.

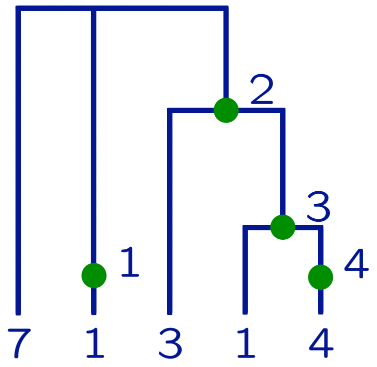
DNA sequences



Incidence matrix

	1	2	3	4
a	0	0	0	0
b	1	0	0	0
c	0	1	0	0
d	0	1	1	0
e	0	1	1	1

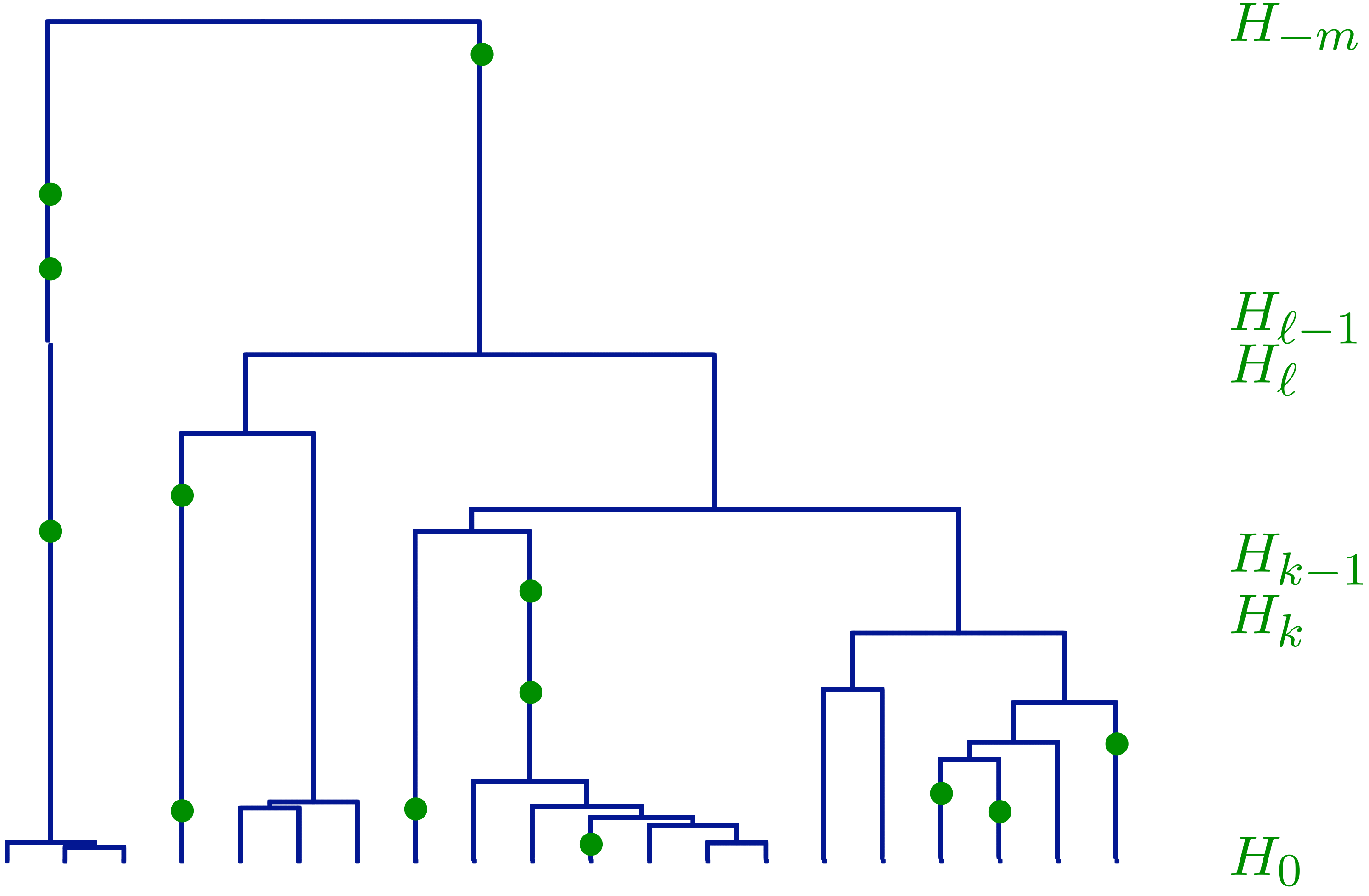
Hammer's Y tree



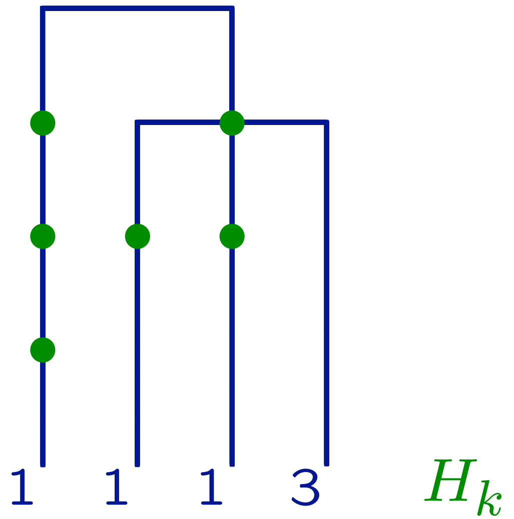
Theorem

The configuration of mutations on sequences is equivalent to a gene tree.

Coalescent History Process



History states H_k are Gene trees $(\mathcal{T}, \mathbf{n})$



Tree $\mathcal{T}$. Multiplicity of lineages $\mathbf{n} = (1, 1, 1, 3)$.

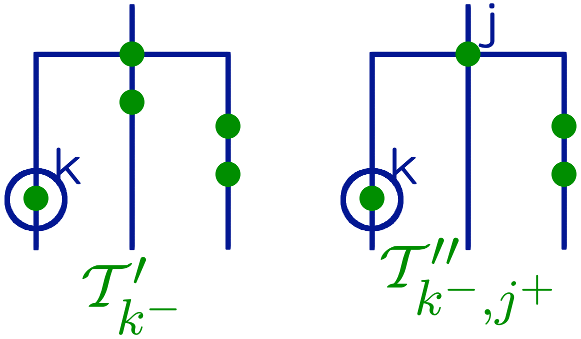
Probability of a gene tree

$$p(\mathcal{T},\mathbf{n}) \;=\; = \frac{(n-1)}{(n-1+\theta)}\sum_{k:n_k\geq 2}\frac{(n_k-1)}{n-1}p(\mathcal{T},\mathbf{n}-\mathbf{e}_k)$$

$$+\; \frac{\theta}{(n-1+\theta)}\sum_{k}\frac{1}{n}p(\mathcal{T}'_{k-},\mathbf{n})$$

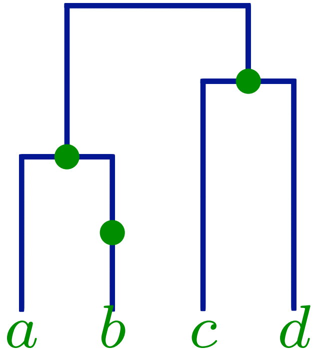
$$+\; \frac{\theta}{(n-1+\theta)}\sum_{k\to j}\frac{(n_j+1)}{n}p(\mathcal{T}''_{k-,j+},\mathbf{n}'')$$

Removing a mutation



Proposal distribution for gene trees $\hat{p}(H_{j-1} \mid H_j)$.

Choose a gene in H_{j-1} uniformly from the possible genes which may change by coalescence or mutation.



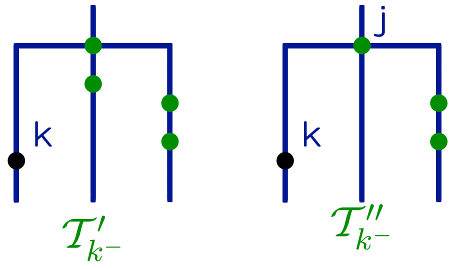
Choose from b, c, d each with probability $1/3$.

(i) b: Remove the mutation on lineage b.

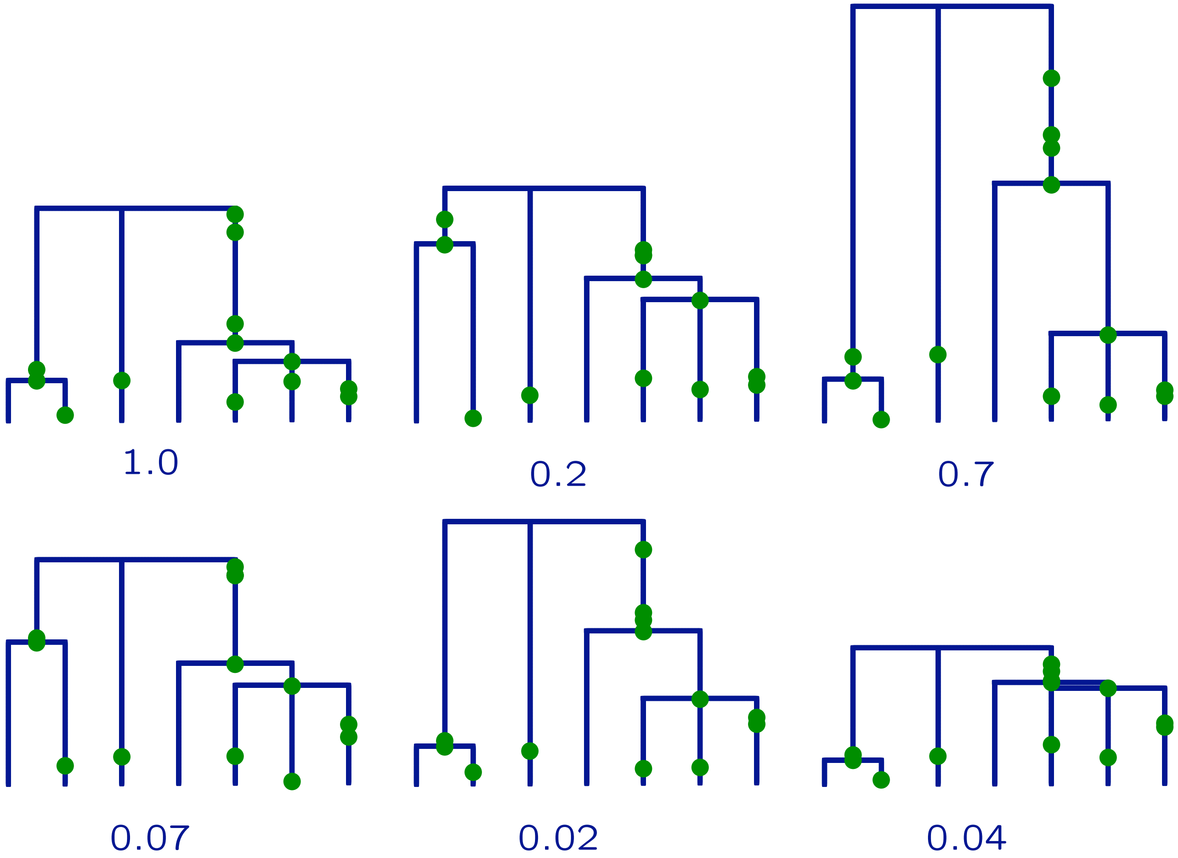
(ii) c or d: Coalesce the two lines c and d.

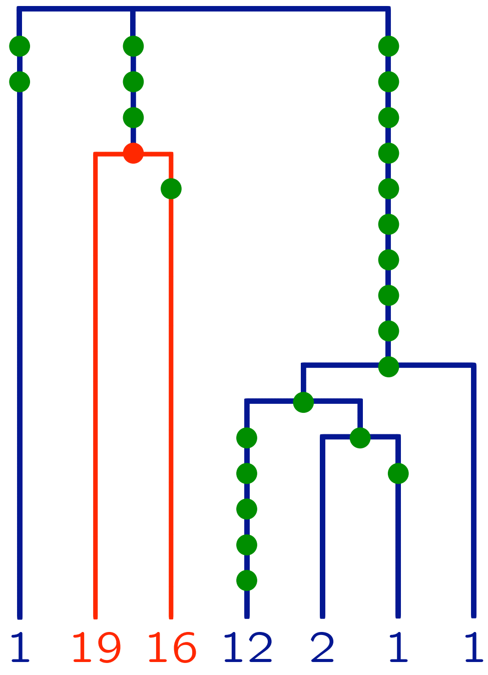
Proposal distribution and importance weights for gene trees

H_{i-1}	Proposal distribution	Importance weights
$(\mathcal{T}, \mathbf{n} - \mathbf{e}_k)$	$\dfrac{n_k}{n_\circ}$	$\dfrac{n_\circ}{n_k} \cdot \dfrac{n_k - 1}{n - 1 + \theta}$
$(\mathcal{T}_{k^-}, \mathbf{n})$	$\dfrac{1}{n_\circ}$	$\dfrac{n_\circ}{n} \cdot \dfrac{\theta}{n - 1 + \theta}$
$(\mathcal{T}''_{k^-}, \mathbf{n} + \mathbf{e}_j)$	$\dfrac{1}{n_\circ}$	$\dfrac{n_\circ}{n} \cdot \dfrac{(n_j + 1)\theta}{n - 1 + \theta}$



Simulated gene trees with relative likelihoods conditional on tree topology

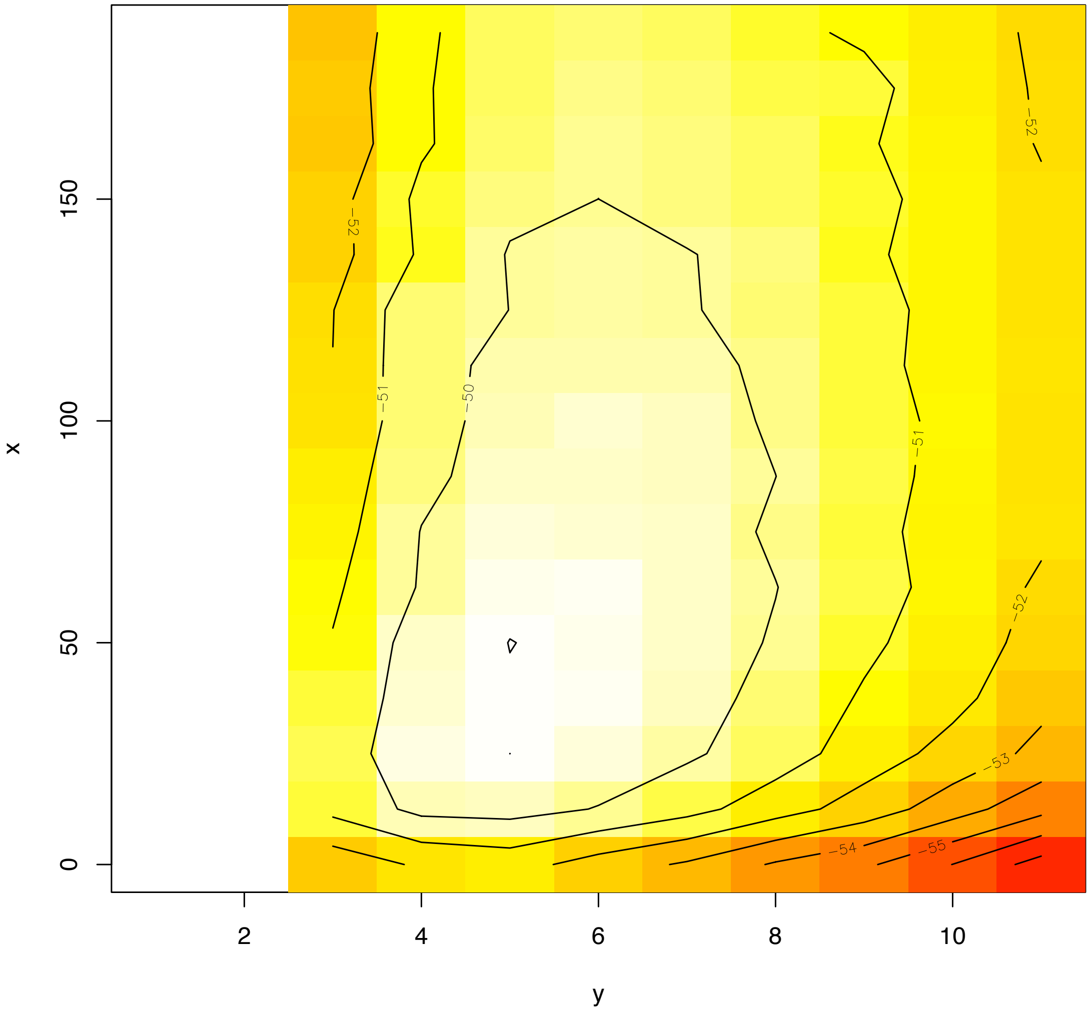




Gene tree with a selected mutation ●
Graham Coop, 2004

The population is subdivided by the changing selected site frequency back in time.

Likelihood contours, y-axis θ, x-axis σ



Coalescent tree simulation

1. A sample of n genes has b mutant genes with the selected site and $n - b$ non-mutant genes.

2. Simulate X, the population frequency at current time, conditional on the configuration $(b, n - b)$.

3. Simulate the trajectory of the mutation frequency $\{X(t)\}$ using the reversed diffusion conditional on absorption at zero, using a birth and death approximation of the diffusion process.

4. Simulate coalescent trees using variable population size models with relative frequencies $\{X(t)\}$, $\{1 - X(t)\}$.

5. Use importance sampling to find the likelihood of a gene tree with a selected mutation.

Two locus diffusion process model, type space $E_A \times E_B$

Generator

$$\mathscr{L} = \sum_{(i,j)\in E_A\times E_B} L_{ij}\frac{\partial}{\partial x_{ij}},$$

where

$$L_{ij} = \frac{1}{2}\sum_{(k,l)\in E_A\times E_B} x_{ij}(\delta_{ik}\delta_{jl}-x_{kl})\frac{\partial}{\partial x_{kl}}+b_{ij}(\boldsymbol{x})+\frac{1}{2}\rho\,(x_{i\cdot}x_{\cdot j}-x_{ij})$$

ρ is the population-scaled recombination rate and

$$b_{ij}(\boldsymbol{x}) = \frac{\theta_A}{2}\sum_{k\in E_A} x_{kj}(P^A_{ki} - \delta_{ki}) + \frac{\theta_B}{2}\sum_{l\in E_B} x_{il}(P^B_{lj} - \delta_{lj})$$

PIM model proposal distribution, $P_{kl}^A = P_l^A, P_{kl}^B = P_l^B$
Griffiths, Jenkins, Song (2008).

$$\begin{aligned}\widehat{\pi}[(i,j) \mid \boldsymbol{n}] \;=\; & \frac{1}{\mathcal{N}'}\Big\{\, n_{ij} + \theta_A P_i^A \widehat{\pi}_B[j \mid \boldsymbol{n}_B] + \theta_B P_j^B \widehat{\pi}_A[i \mid \boldsymbol{n}_A] \\ & + \frac{1}{2}\rho \left(\frac{n+\theta_A}{n+1+\theta_A} + \frac{n+\theta_B}{n+1+\theta_B} \right) \widehat{\pi}_A[i \mid \boldsymbol{n}_A] \widehat{\pi}_B[j \mid \boldsymbol{n}_B] \Big\},\end{aligned}$$

where

$$\mathcal{N}' = n + \theta_A + \theta_B + \frac{1}{2}\rho \left(\frac{n+\theta_A}{n+1+\theta_A} + \frac{n+\theta_B}{n+1+\theta_B} \right).$$

Proposal distributions have been extended to models with migration as well in GJS (2008).

Fearnhead and Donnelly's (2001) proposal distribution

$$\begin{aligned}\widehat{\pi}_{\mathsf{FD}}[(i,j)\,|\,\boldsymbol{n}] \;=\; & \frac{1}{\mathcal{N}'_{\mathsf{FD}}}\Big\{\, n_{ij} + \theta_A P^A_i \widehat{\pi}_B[j\,|\,\boldsymbol{n}_B] + \theta_B P^B_j \widehat{\pi}_A[i\,|\,\boldsymbol{n}_A] \\ & + \rho\left(\frac{n+\theta_A+\theta_B}{n}\right)\widehat{\pi}_A[i\,|\,\boldsymbol{n}_A]\,\widehat{\pi}_B[j\,|\,\boldsymbol{n}_B]\Big\}\,,\end{aligned}$$

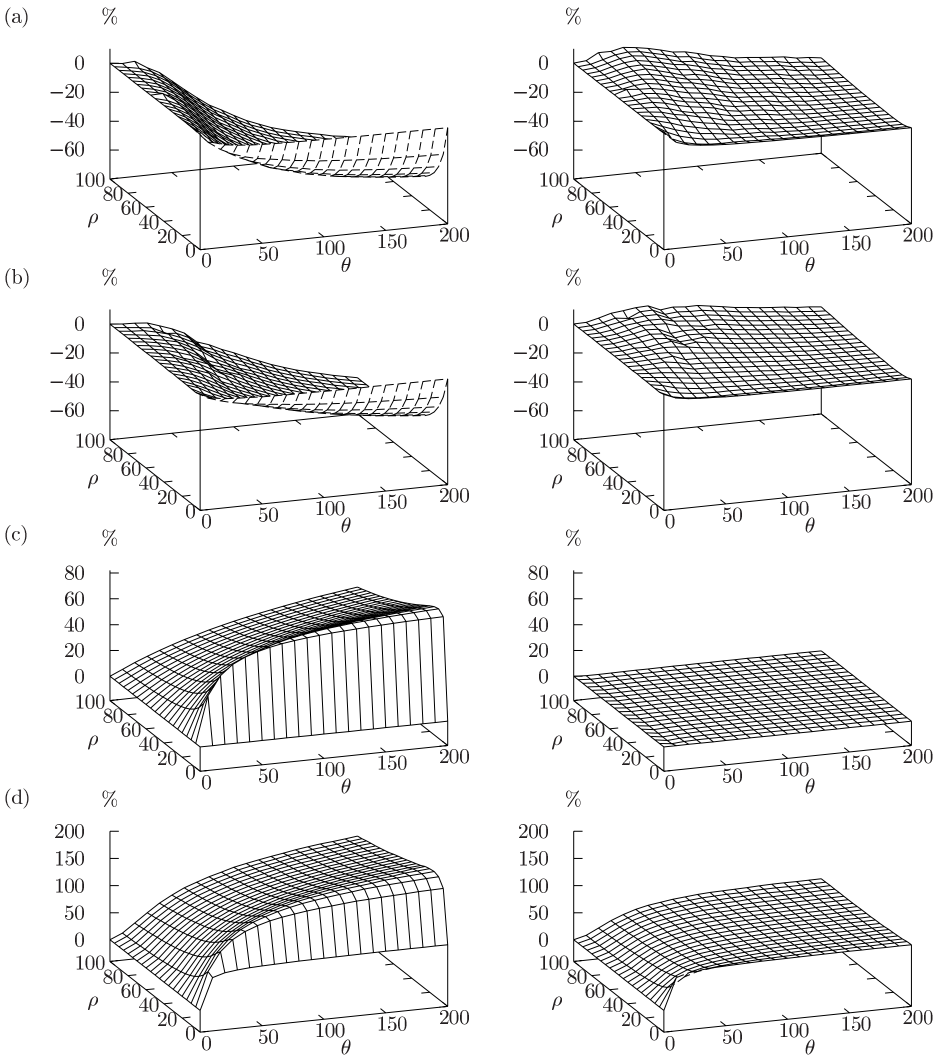
where

$$\mathcal{N}'_{\mathsf{FD}} = n + \theta_A + \theta_B + \rho\left(\frac{n+\theta_A+\theta_B}{n}\right).$$

Infinitely-many-alleles-data

Left Column: Deviation of Fearnhead and Donnelly's distribution $\widehat{\pi}^*_{\mathsf{FD}}[(1,1)|\boldsymbol{n}]$ from the true distribution.

Right Column: Deviation of GJS distribution $\widehat{\pi}^*[(1,1)\,|\,\boldsymbol{n}]$ from the true distribution. For all figures, $\theta_A = \theta_B = \theta/2$. (a) $\boldsymbol{n} = (4,3,2,3)$. (b) $\boldsymbol{n} = (5,4,4,5)$. (c) $\boldsymbol{n} = (0,0,1,0)$. (d) $\boldsymbol{n} = (0,0,2,1)$.


(a)
(b)
(c)
(d)
%
0
−20
−40
−60
80
60
40
20
200
150
100
50
100
80
60
40
20
0
ρ
0
50
100
150
200
θ

References

Barton, N. H., Etheridge, A. M., Strum, A. K. (2003) Coalescence in a random background. *Ann. Appl. Prob.* 14, 754–785.

Birkner, M., Blath, J., Capaldo, M., Etheridge, A., Möhle, M., Schweinsberg, J., Wakolbinger, A. (2005) Alpha-stable branching and beta-coalescents. *Electronic Journal of Probability* 10, 303–325.

Bustamante, C. D., Wakeley, J., Hartl, D. L. (2001) Directional selection and the site-frequency spectrum. *Genetics* 159, 1779–1788.

Carbone, I. Kohn, M. (2001) A microbial population-species interface: nested cladistic and coalescent inference with multilocus data. *Molecular Ecology* 10, 947–964.

Coop, G., Griffiths, R. C. (2004) Ancestral inference on gene trees under selection. *Theoret. Popul. Biol.* 66, 219-232.

De Iorio, M., Griffiths, R. C. (2004) Importance sampling on coalescent histories. I. *Adv. Appl. Prob.* 36, 417–433.

De Iorio, M., Griffiths, R. C. (2004) Importance sampling on coalescent histories. II Subdivided population models. *Adv. Appl. Prob.* 36,434–454.

De Iorio, M., Griffiths, R.C., Lebois, R., Rousset, F. (2005) Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. *Theoret. Popul. Biol.* 68, 41–53.

Fearnhead, P., Donnelly, P. (2001) Estimating recombination rates from population genetic data. *Genetics* 159, 1299–1318.

Griffiths, R.C. (2002) Ancestral inference from gene trees. In: Veuille, M. Slatkin, M. (Eds.), Modern Developments in Theoretical Population Genetics: the Legacy of Gustave Malécot, Oxford University Press, New York, pp. 94–117.

Griffiths, R.C., Griffiths, Y.J. (2006) Ancestral Inference from microsatellite data by sequential importance sampling in subdivided population models. Proceedings of Simulations, Genetics and Human Prehistory - A Focus on Islands, Cambridge University, 29 July - 1 August 2005.

Griffiths, R. C., Jenkins, P. A., Song, Y. S. (2008) Importance sampling and the two-locus model with subdivided population structure. *Adv. Appl. Prob.* 40, 473–500.

Griffiths, R.C., Tavaré, S. (1994) Ancestral inference in population genetics. *Statistical Science* 9, 307–319.

Griffiths, R.C., Tavaré, S. (1999) The ages of mutations in gene trees. *Ann. Appl. Prob.* 9, 567–590.

Gusfield, D. (1991) Efficient algorithms for inferring evolutionary trees. *Networks.* 21, 19–28.

Harding, R.M., Fullerton, S.M., Griffiths, R.C., Bond, J., Cox M.J., Schneider, J.A., Moulin, D., Clegg, J.B. (1997) Archaic African and Asian lineages in the genetic ancestry of modern humans. *Amer. J. Hum. Genet.* 60, 772–789.

Stephens, M., Donnelly, P. (2000) Inference in molecular population genetics. *J. Roy. Statist. Soc.* B 62, 605–655.

Zhivotovsky, L.A., Rosenberg, N.A., Feldman, M.W. (2003) Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. *Amer. J. Hum. Genet.* 72, 1171–1186.