The coalescent process

Introduction

■ Random drift can be seen in several ways

- **Forwards** in time: variation in allele frequency
- **Backwards** in time: a process of inbreeding//coalescence

Allele frequencies

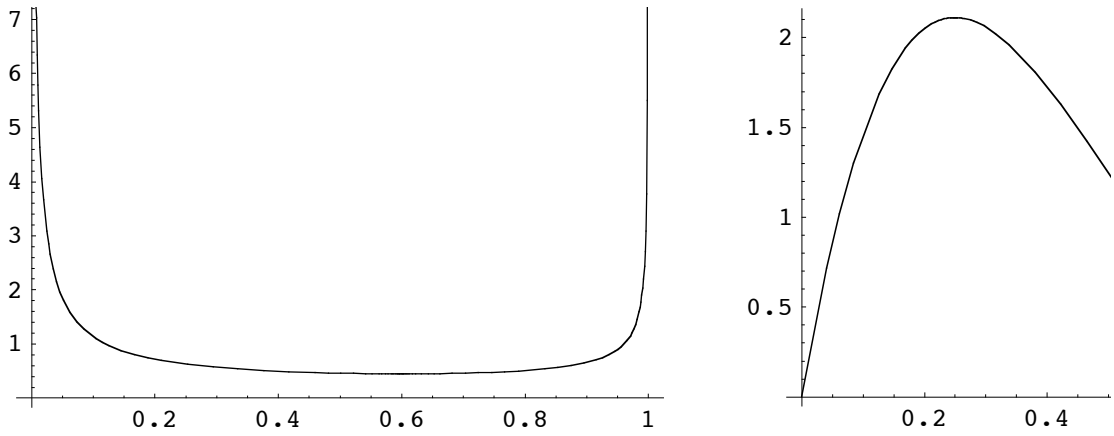
Random variation in reproduction causes random fluctuations in allele frequency:

$$\text{var}(p) = \frac{pq}{2N_e}$$

After many generations, the distribution can be approximated by a *diffusion*.
With random drift and mutation (P→Q at rate μ, Q→P at rate ν) the equilibrium distribution is:

$$\text{prob}(p) \sim p^{4N_e\nu-1}\, q^{4N_e\mu-1}$$

The left-hand plot shows the distribution of p for $N_e = 2,500$, $\nu = 2.5\times10^{-5}$, $\mu = 5\times10^{-5}$; the right-hand plot is for $N_e = 20{,}000$

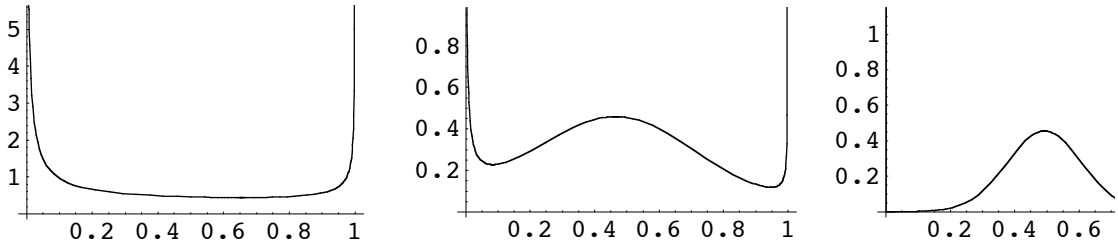


The diffusion approximation can also include other forces, such as selection and migration. For example, the equilibrium distribution under mutation, random drift, and selection is:

$$\text{prob}(p) \sim p^{4N_e\nu-1}\, q^{4N_e\mu-1}\, \bar{w}^{2N_e}$$

With heterozygote advantage (fitnesses 1-s;1:1-s), $\overline{W}^{2N_e} = 1 - s(p^2 + q^2) \sim \text{Exp}[-2N_e\, s(p^2 + q^2)]$

With $N_e = 2,500$, $\nu = 2.5\times10^{-5}$, $\mu = 5\times10^{-5}$, and s=0.0001, 0.001, 0.004 (left to right):



💡 The key parameters are $N_e\,\mu$, $N_e\,\nu$, $N_e\,s$, which give the strength of drift *relative* to mutation and selection.

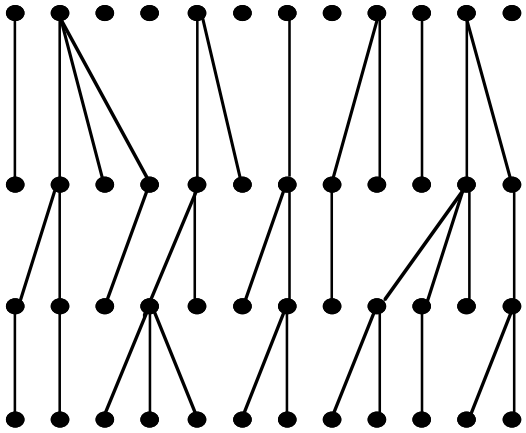
■ **Further reading:** Kimura, *The neutral theory of molecular evolution,* Chap.3

Identity by descent

■ Definition

Wright (1921, 1922), Haldane & Moshinsky (1939), Cotterman (1940) and Malécot (1948) developed the idea of *identity by descent.*

Two genes are *identical by descent* if they descend from the *same* gene in some ancestral population.



■ **Note:**

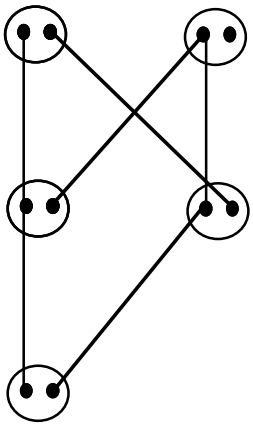
- *Identity by descent* is distinct from *identity in state*
- *i.b.d.* is defined relative to some ancestral *reference* population.
- Identity measures can extend to *many* genes; usually, however, we just deal with identity between *pairs* of genes.

This is related to *variance* of allele frequency, *correlation* between genes, and *homozygosity*

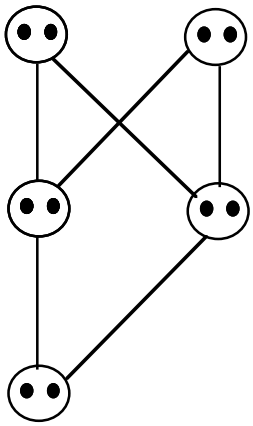
- Relationships among many genes are better thought of in terms of *coalescence* of lineages in a genealogy.

■ The probability of identity by descent is easily calculated for pedigrees

e.g. brother-sister mating



Genes are NOT ibd
in this case



Probability of identity
by descent is 1/4

In general, the probability that two distinct genes in a diploid individual are i.b.d. is $f = \sum_{\text{loops}} (\frac{1}{2})^{n-1} (1 + f_A)$, where the sum is over all loops in the pedigree, n is the number of individuals in the loop, and f_A the identity between genes in the common ancestor.

Note that the random element here is in segregation, not reproduction

■ The increase in i.b.d. with random mating

▪ Wright-Fisher model

Suppose that there are $2N_t$ individuals in a haploid population. In the next generation, there are $2N_{t+1}$, drawn randomly from all $2N_t$ possible parents.

On this scheme, individuals produce a number of offspring which is close to a Poisson distribution.

The Wright-Fisher model also applies to a random-mating diploid population, provided that individuals are as likely to mate with themselves as with anyone else.

Then, the probability that two genes are i.b.d. from the previous generation is $1/2N_t$:

$$f_{t+1} = \frac{1}{2N_t} + \left(1 - \frac{1}{2N_t}\right) f_t \qquad f_0 = 0$$

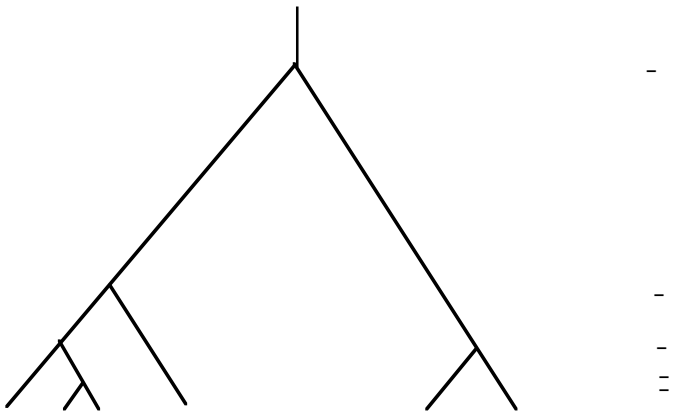
$$h_{t+1} \equiv 1 - f_t = \left(1 - \frac{1}{2N_t}\right) h_t \qquad \text{hence } h_t = \prod_{i=0}^{t-1} \left(1 - \frac{1}{2N_i}\right)$$

With constant population size, h_t declines by (1-1/2N) per generation - approximately, as ~exp(-t/2N).
The typical **timescale** for inbreeding and random drift is 2N generations.

With fluctuating sizes, h_t declines (approximately) as $\exp\left(-\left(\sum_{i=0}^{t-1} \frac{1}{2N_i}\right)\right) = \exp(-t/2N_H)$ where N_H is the *harmonic mean* population size.

Coalescence

The ancestry of a sample of *neutral* genes has a simple statistical distribution:
the chance that any two lineages *coalesce* is $\frac{1}{2N_e}$ per generation



More precisely:
- suppose that each gene leaves v descendants
- As N→∞, the probability that any pair of lineages coalesce, per generation, tends to $\frac{\text{var}(v)}{2N}$
i.e. $N_e = N/\text{var}(v)$

The coalescent process refers to this limit
- equivalent to the diffusion approximation

An influential idea:
- DNA sequences are best described by their genealogy

- a variety of mutation models can be superimposed
- tracing back samples of alleles
    - speeds up simulations
    - gives statistical tests on sampled data

■ **References**

Hudson, R. (1990). Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.* 7, 1-44.
Hudson, R. (1993). The how and why of generating gene genealogies. In *Mechanisms of molecular evolution*, ed. Takahata N & Clark AG, pp 23-36.
Donnelly, P. and S. Tavaré. (1995). Coalescents and genealogical structure under neutrality. *Ann. Rev. Genet.* 29, 401-421.
Rosenberg, N. A., and M. Nordborg, 2002 Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nature Reviews Genetics **3:** 380-390.

■ Properties of the coalescent process

The time during which there are k lineages is exponentially distributed with expectation $\frac{1}{\lambda} = \frac{2 N_e}{k(k-1)/2}$:

$$P(t_k) = \text{Exp}[-\lambda t_k]\, \lambda\, \text{d}t_k \quad \text{where } \lambda = \frac{k(k-1)}{4 N_e}$$

■ **The genealogy is dominated by the deepest split.**

The expected depth of the tree is:

$$2 N_e \left(\frac{2}{k(k-1)} + \frac{2}{(k-1)(k-2)} \cdots \frac{1}{6} + \frac{1}{3} + 1 \right) =$$
$$2 N_e \left(\left(\frac{2}{k-1} - \frac{2}{k} \right) + \left(\frac{2}{k-2} - \frac{2}{k-1} \right) + \cdots \left(\frac{2}{2} - \frac{2}{3} \right) + \left(\frac{2}{1} - \frac{2}{2} \right) \right) =$$
$$2 N_e \left(\left(1 - \frac{2}{k} \right) + 1 \right) \sim 4 N_e \text{ for large } k$$

Thus, the tree collapses to 2 lineages in $\sim 2 N_e$ generations; these take another $2 N_e$ generations to coalesce
Hence, pairwise measures are **uninformative**

■ **The expected length of the genealogy is $\sim 4 N_e$ Log[1.78 k]**

The expected length of the tree is:

$$2 N_e \left(k \frac{2}{k(k-1)} + (k-1) \frac{2}{(k-1)(k-2)} \cdots \frac{4}{6} + \frac{3}{3} + 2 \right)$$
$$= 2 N_e \left(\frac{2}{k-1} + \frac{2}{k-2} + \cdots \frac{2}{3} + \frac{2}{2} + \frac{2}{1} \right)$$
$$= 4 N_e \sum_{j=1}^{k-1} \frac{1}{j}$$
$$\sim 4 N_e \text{ Log}[1.78\, k] \text{ for large } k$$

The distribution of length is highly variable:

The dots show the quantiles at 0.001, 0.01, 0.1, 0.9, 0.99, 0.999.

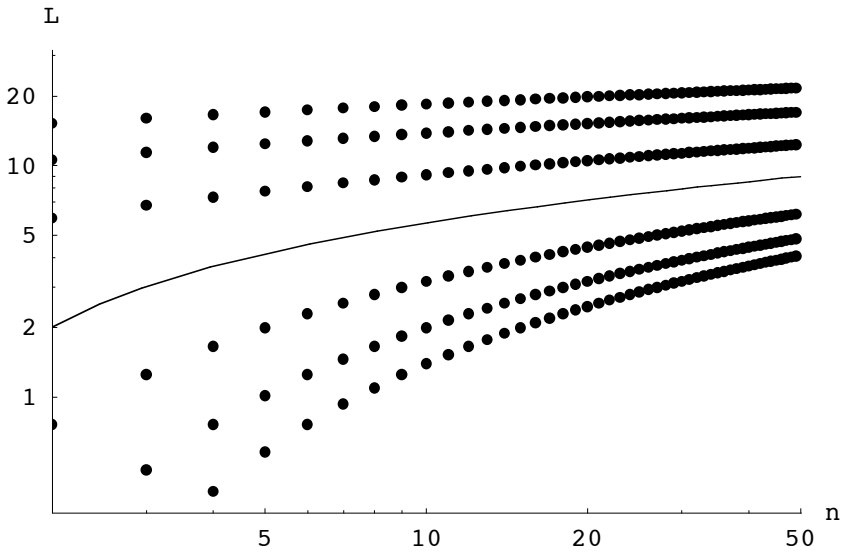
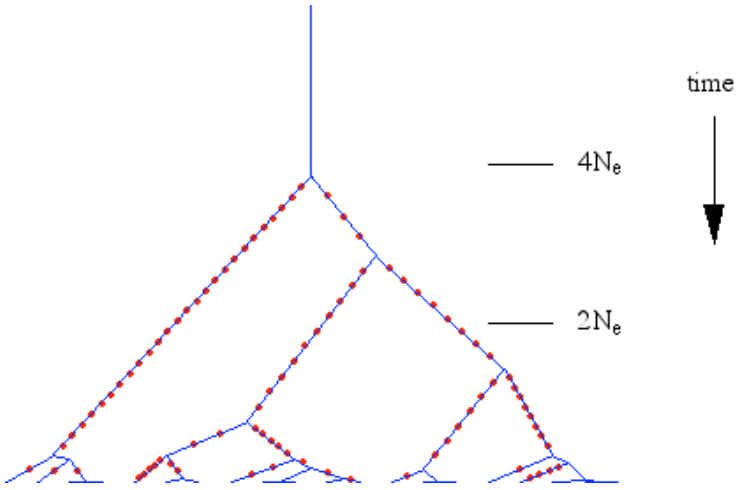


Figure 1

■ Fluctuating population size

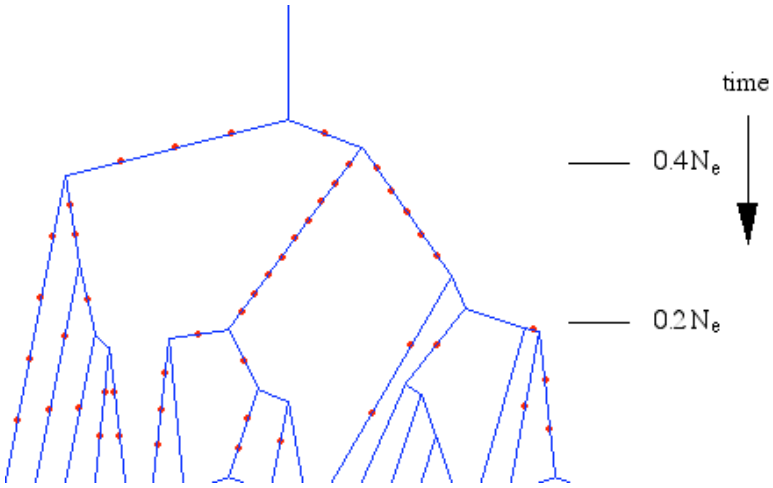
Changes in N_e cause changes in timescale

The standard coalescent



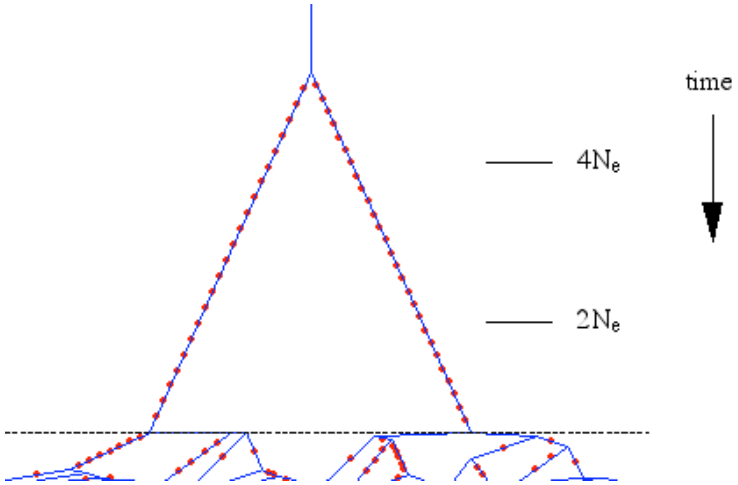
Expanding populations → "star phylogeny"

exponential growth: popl'n was 10% of the current size at T_{MRCA}



Population bottlenecks → burst of coalescence

a bottleneck equivalent to 2 N_e 'ordinary' generations of drift



■ **Changing timescales**

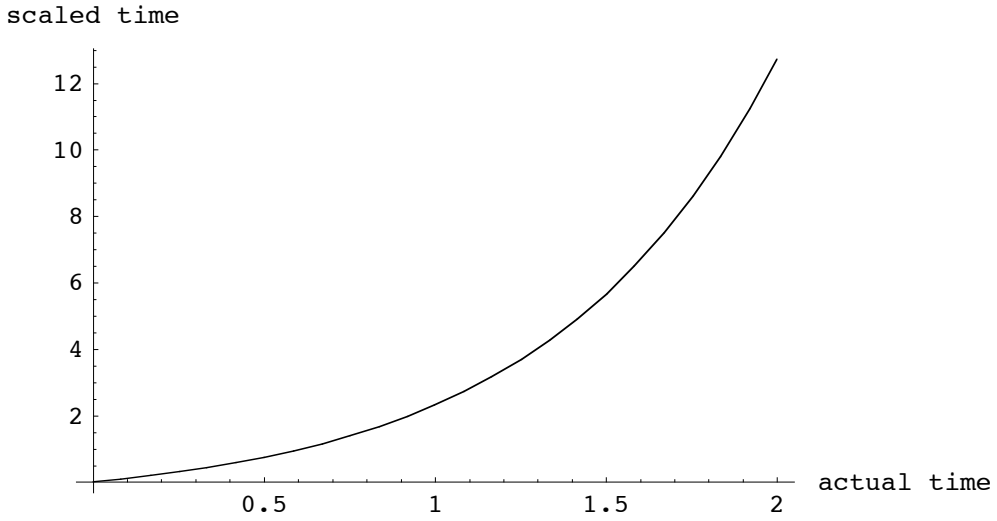
The "scaled time" is a measure of the total amount of genetic drift that has occurred:

$$T = \int_0^t \frac{dt}{2\,N(t)}$$

For a constant population size, $T = t/(2N)$. If the population is growing at a rate λ, and the present size is N_0, then $N = N_0\, e^{-\lambda t}$, and so:

$$T = \int_0^t \frac{e^{\lambda t}}{2\,N_0}\,dt = \frac{1}{2\,N_0\,\lambda}\,(e^{\lambda t} - 1)$$

The parameter λ is a measure of the amount of population growth over the current timescale set by population size, $2N_0$. Here is the transformation for $\lambda = 1.5$

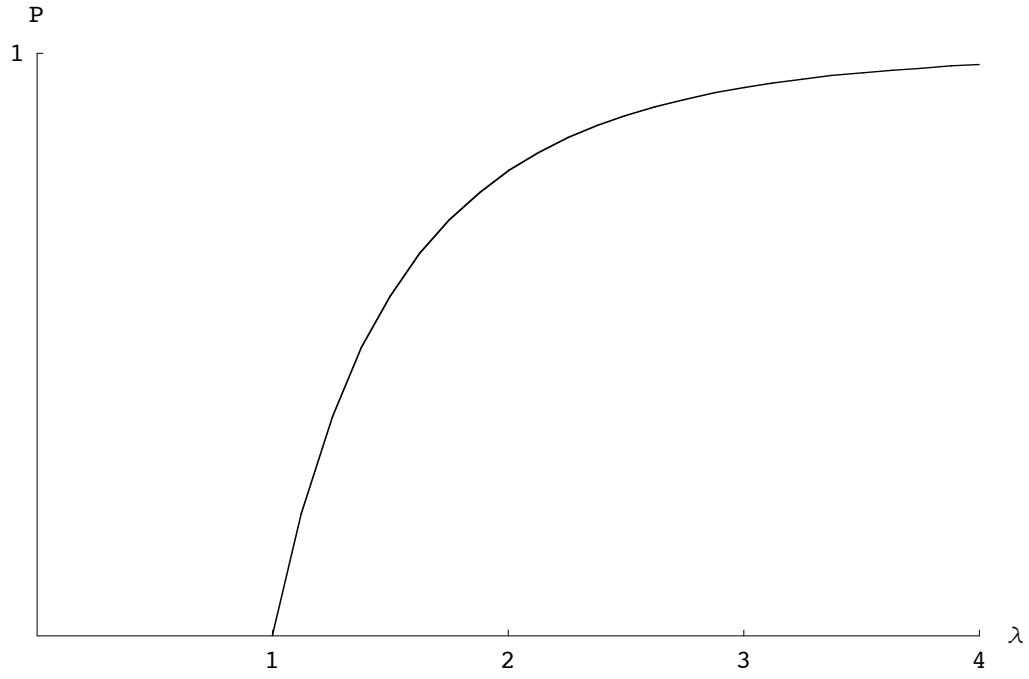


■ Branching processes

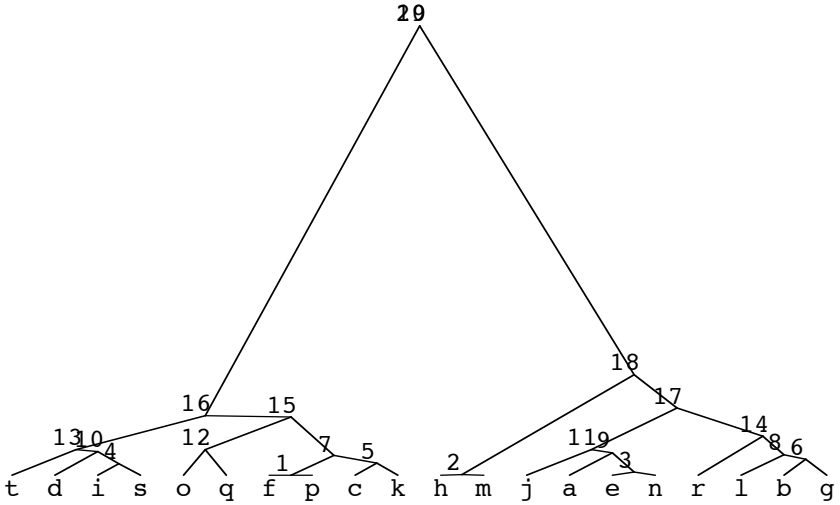
The coalescent process only applies to samples from a large population

If all genes are observed, we have a **branching process**

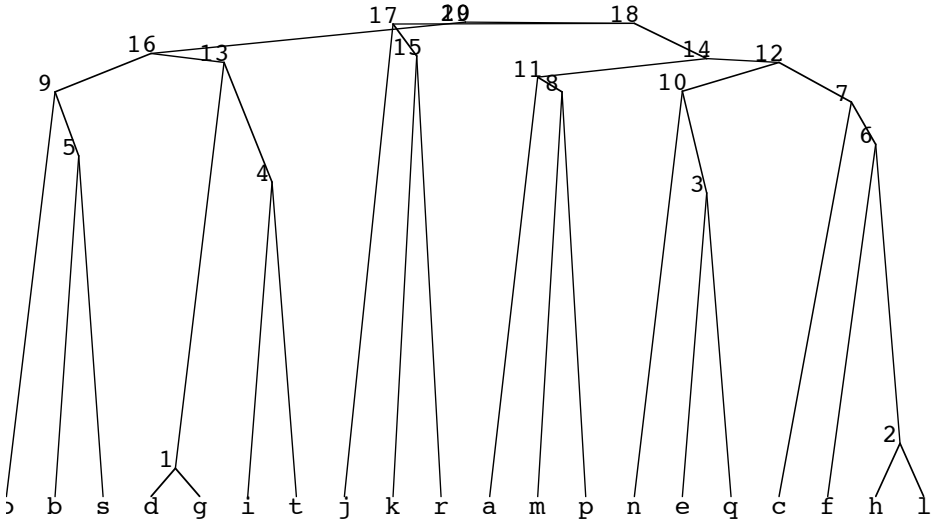
e.g. discrete time: # of offspring i follows a Poisson distribution with $E[i] = \lambda$



More generally, for $\lambda \sim 1$, $P \sim 2\,(\lambda - 1) / \mathrm{var}(i)$



coalescent



sample from a branching process
λ = 1.1

Mutation

■ Infinite alleles

Assuming that every mutation generates a new allele, the probability of identity in allelic state ("homozygosity") is F = $\sum_t f_t\,(1-\mu)^{2t}$, where f_t is the distribution of coalescence times.

$$F \sim E[e^{-2\mu t}] = \int_0^\infty e^{-2\mu t}\, f_t\, dt = \int_0^\infty e^{-2\mu t}\, e^{-t/2N_e}\, \frac{dt}{2N_e} = \frac{1}{1+4N_e\,\mu}$$

Identity coefficients, F, can easily be calculated by going back in time one generation:

$$F = (1-\mu)^2 \left(\left(1-\frac{1}{2N_e}\right)F + \frac{1}{2N_e}\right) \Rightarrow F = \frac{(1-\mu)^2}{2\,\text{Ne}\,(1-(1-\frac{1}{2\,\text{Ne}})\,(1-\mu)^2)} \sim \frac{1}{1+4N_e\mu}$$

Identity coefficients are *generating functions* for the distribution of coalescence times:

$$F \sim E[e^{-2\mu t}] \quad \therefore F = 1 \text{ when } \mu = 0$$

$$\frac{dF}{d\mu} \sim E[-2t\,e^{-2\mu t}] \quad \therefore \frac{dF}{d\mu} = -2E[t] \text{ when } \mu = 0$$

$$\frac{d^2F}{d\mu^2} \sim E[4t^2\,e^{-2\mu t}] \quad \therefore \frac{d^2F}{d\mu^2} = 4E[t^2] \text{ when } \mu = 0$$

■ More general models of mutation

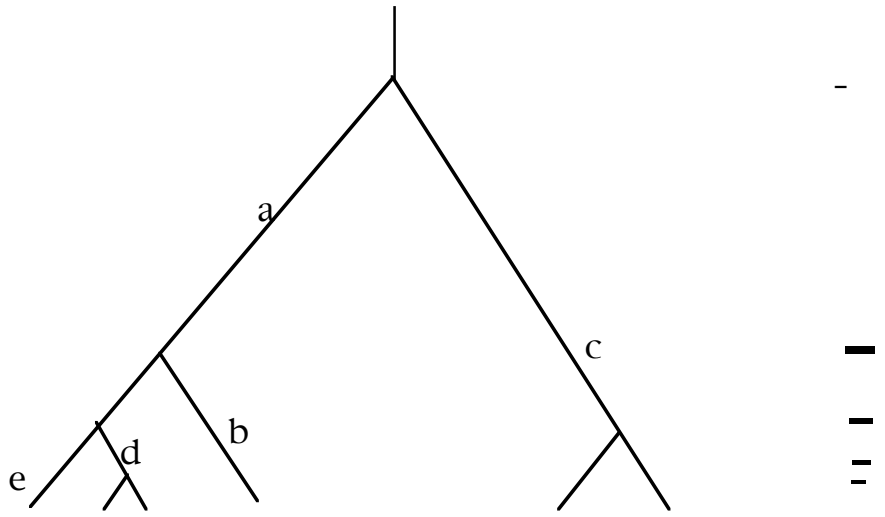
Bases mutate at rate μ, and change to A, T, G, C with equal probability
Probability of identity in state of two genes is:

$$F = E\left[\frac{1}{4}\,(1-e^{-2\mu t}) + e^{-2\mu t}\right]$$

■ Infinite sites

For DNA sequences, the 'infinite sites' model is more appropriate: each mutation is at a new site in the sequence.
Two alleles may differ by mutations at 1, 2… sites - giving a measure of the time for which they have been diverging.

If there are mutations on every internal branch, the genealogy can be reconstructed:



Gene	1	2	3	4	5	6
Mut ' n						
a	1	1	1	1	0	0
b	0	0	0	1	0	0
c	0	0	0	0	1	1
d	0	1	1	0	0	0
e	1	0	0	0	0	0

To root the tree, we must know which mutations are derived - which requires an outgroup

Any pair of sites which carried *all four* combinations is incompatible with a tree
- recombination
- multiple mutations

The mean pairwise diversity, π, is just E[2μ t] = $4 N_e \mu$

The number of segregating sites, n_s, in a sample is proportional to the total *length* of the tree: $E[n_s] = \mu L$, where $L = \sum_{j=1}^{k} j t_j$

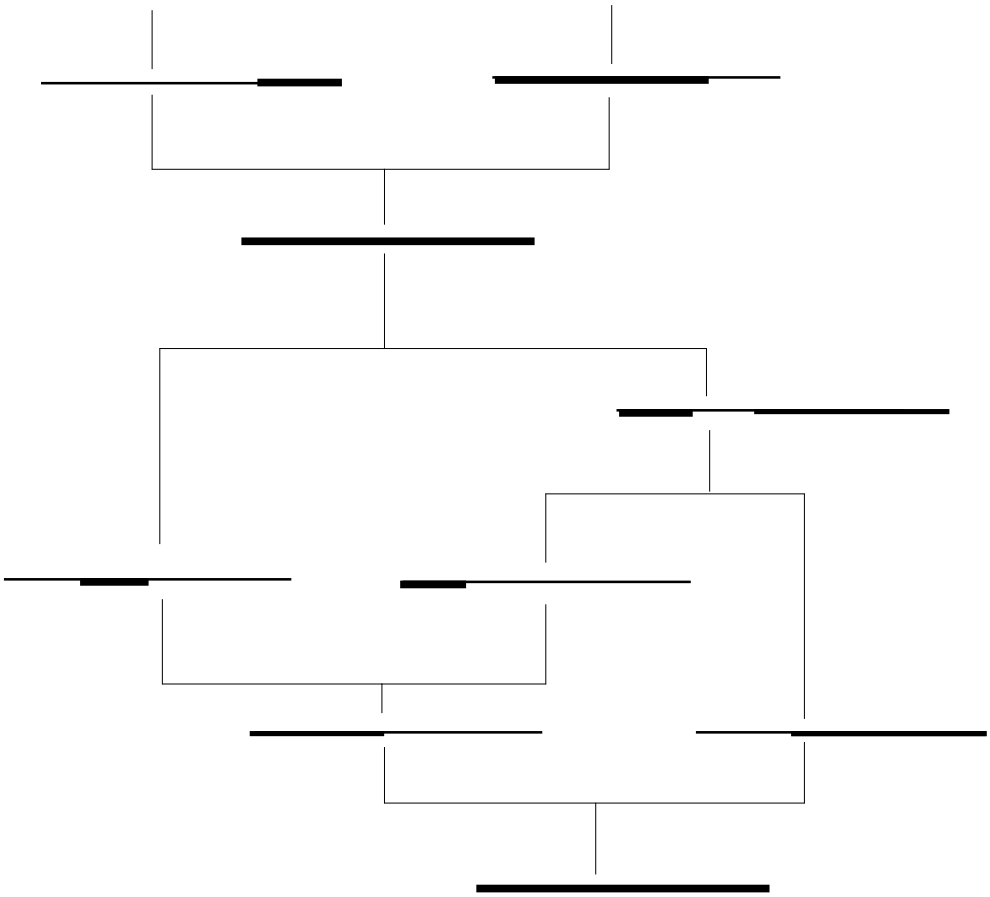
$$E[n_s] = E[\mu L] = 4 N_e \mu \left(\frac{1}{(k-1)} + \frac{1}{(k-2)} \ldots \frac{1}{3} + \frac{1}{2} + 1 \right) \sim 4 N_e \mu \text{Log}[1.78\ k]$$

Under neutrality, we expect a definite relation between the # of segregating sites and the pairwise diversity

Recombination

■ Ancestral graphs

With sexual reproduction, genomes have multiple ancestors.
Ancestry is described by an *ancestral graph:*



Coalescence amongst k lineages at a rate $\frac{k(k-1)}{2}\ \frac{1}{2N_e}$
Recombination at a rate kr

Pattern depends on $R = 2N_e r$

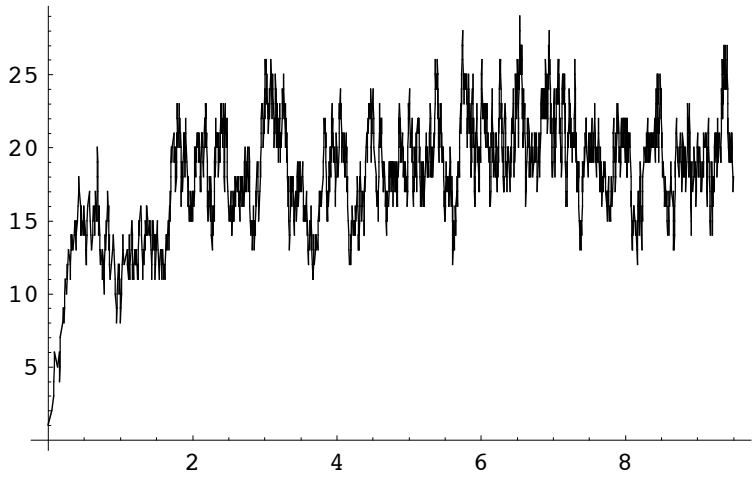
Each recombination generates a pair of unique *junctions*
Junctions can disappear if they meet eachother in a coalescence

At any time, any one genome is distributed across several ancestral lineages

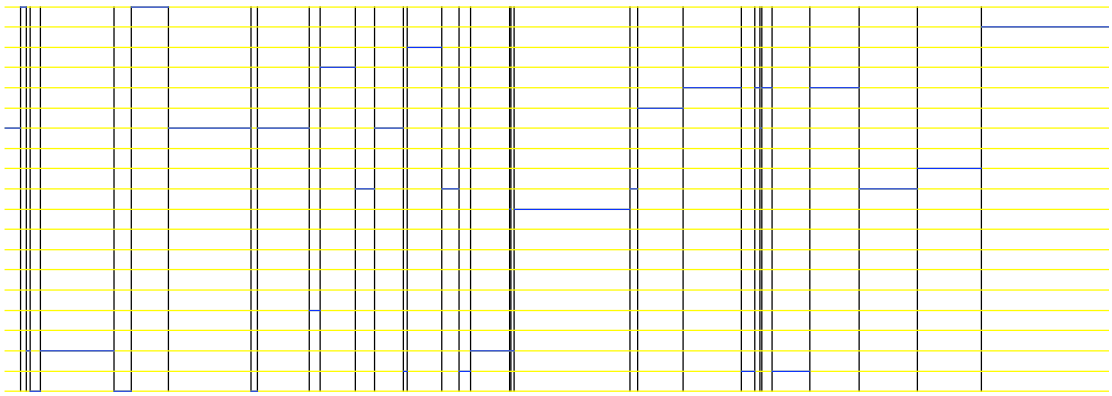
$$1 + R - \frac{R^2}{3} + \frac{13}{54} R^3 + O(R^4) \quad \text{(Derrida \& Jung – Muller 1999)}$$

■ Example: R = 50

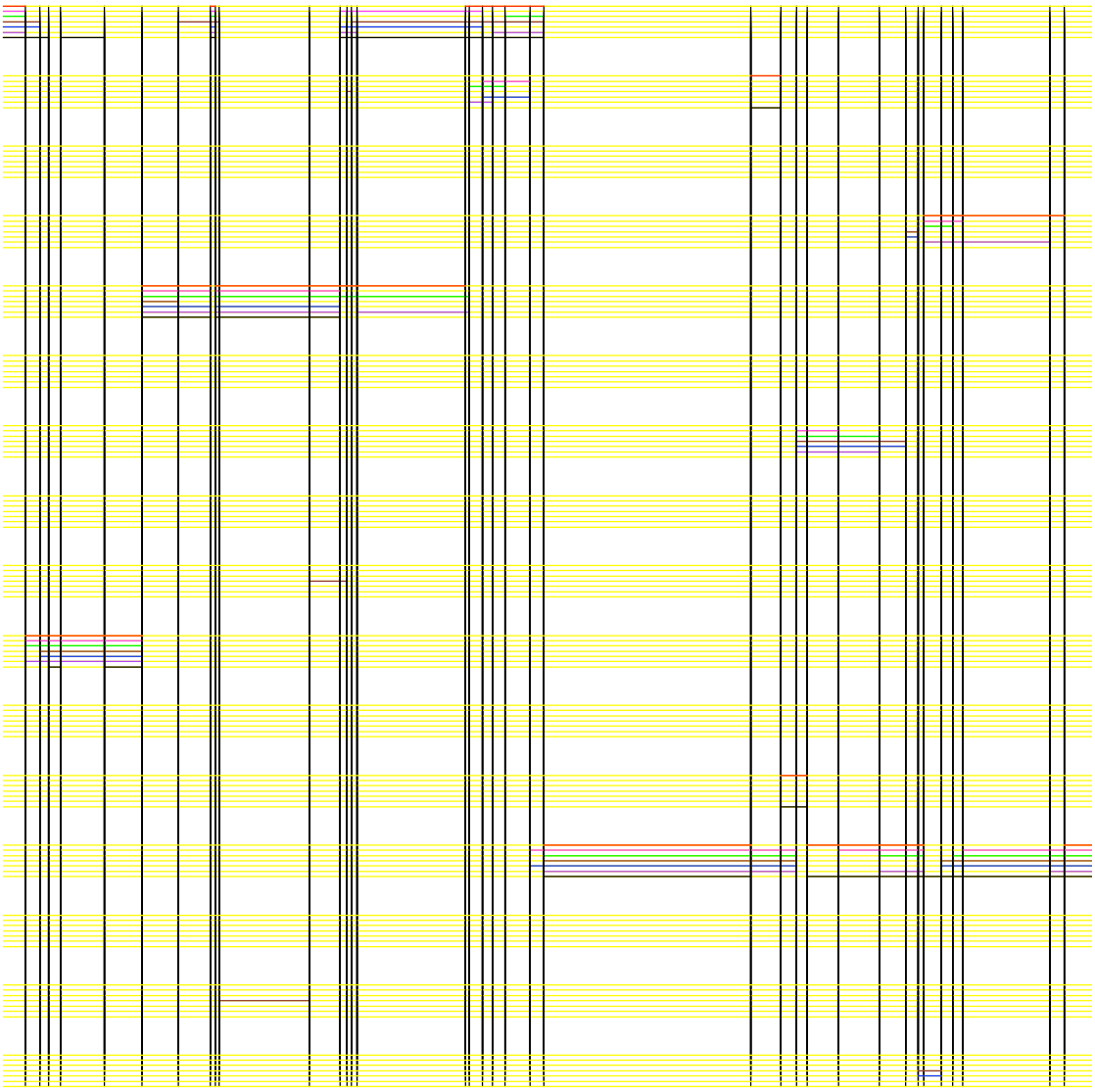
Number of ancestral lineages:



A typical sample, with 18 ancestors:

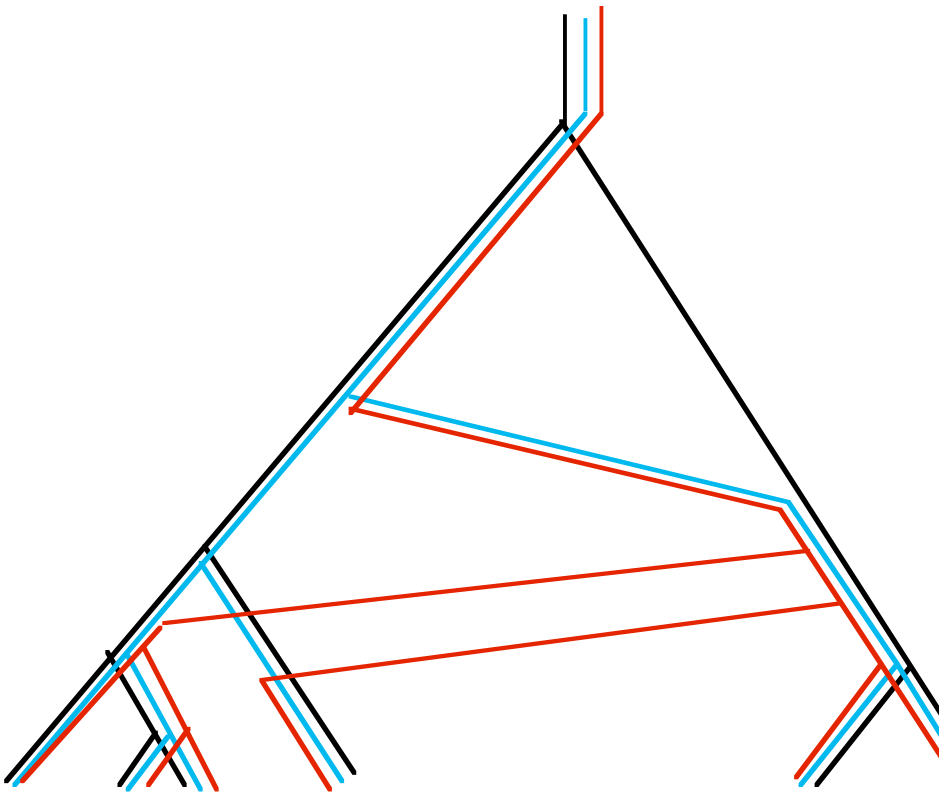


Six sampled genomes represented by colours $\left(\frac{t}{2N_e} = 0.6\right)$:



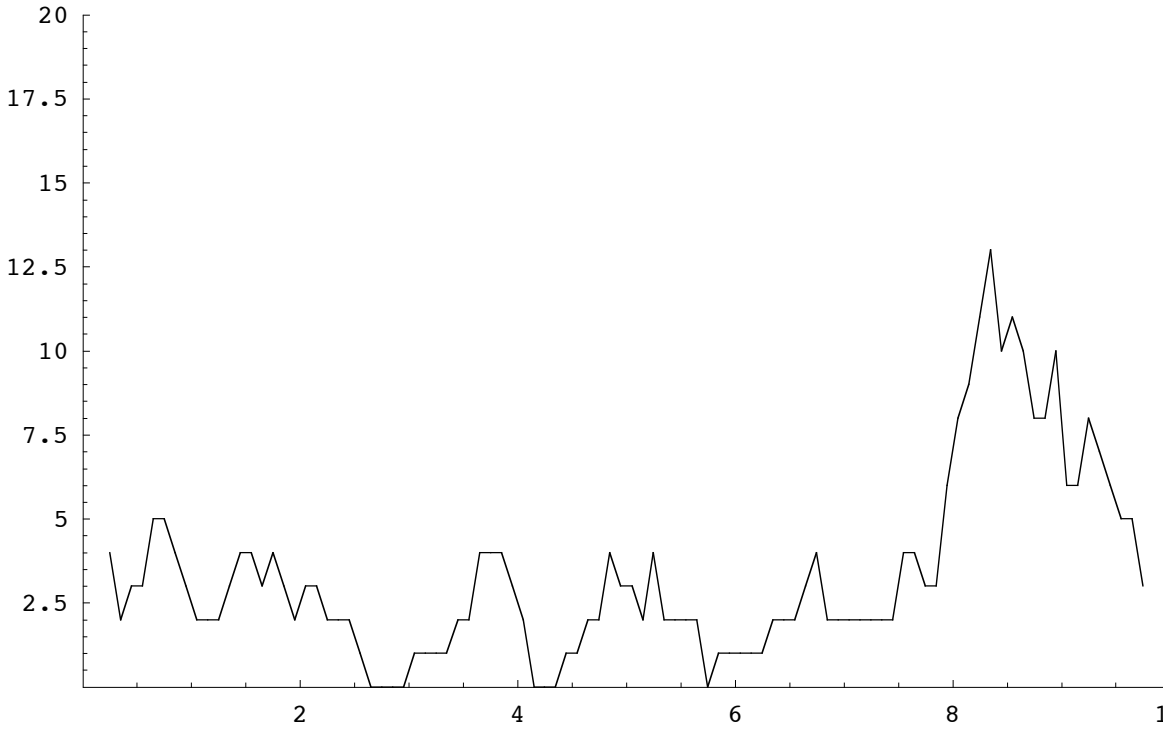
■ Looking along the genome....

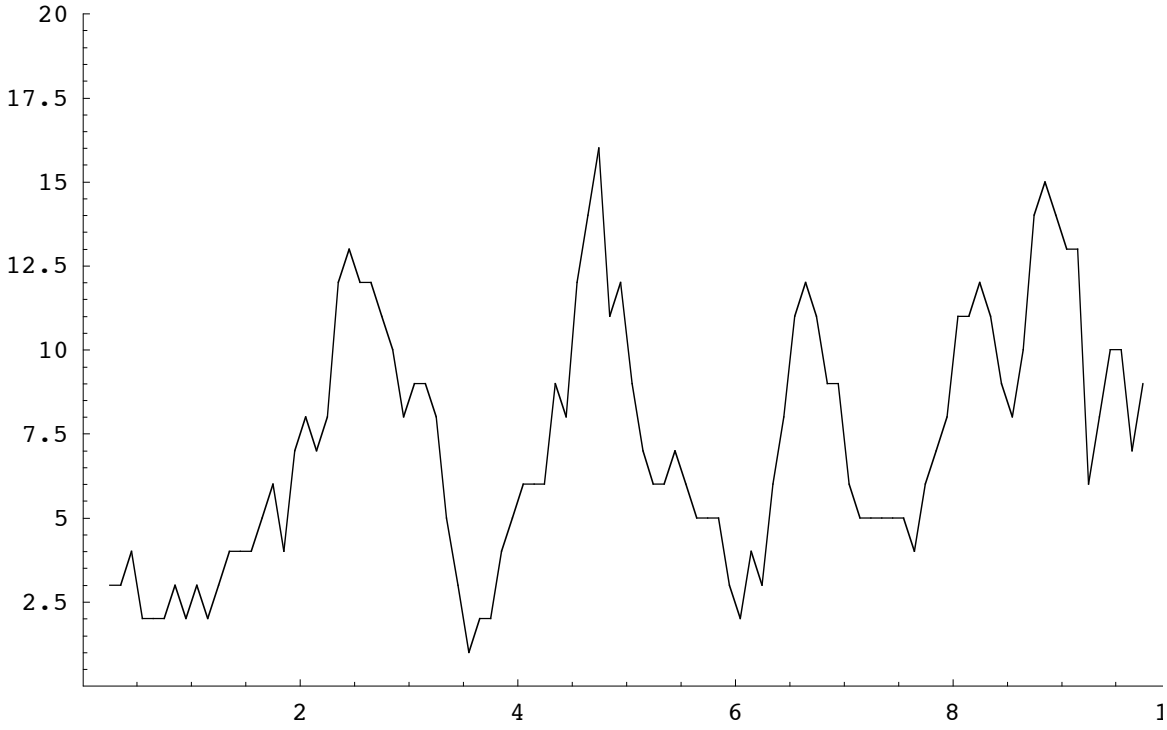
Different regions have different genealogies:

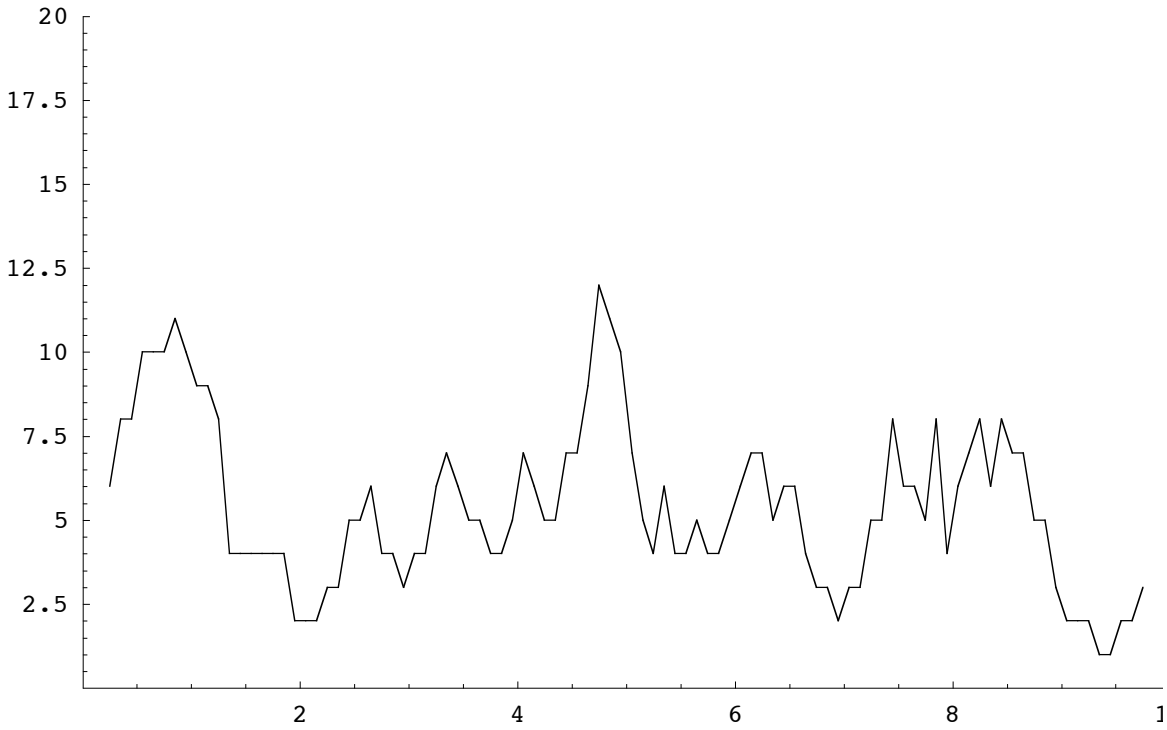


■ **Patterns of diversity vary along the genome:**

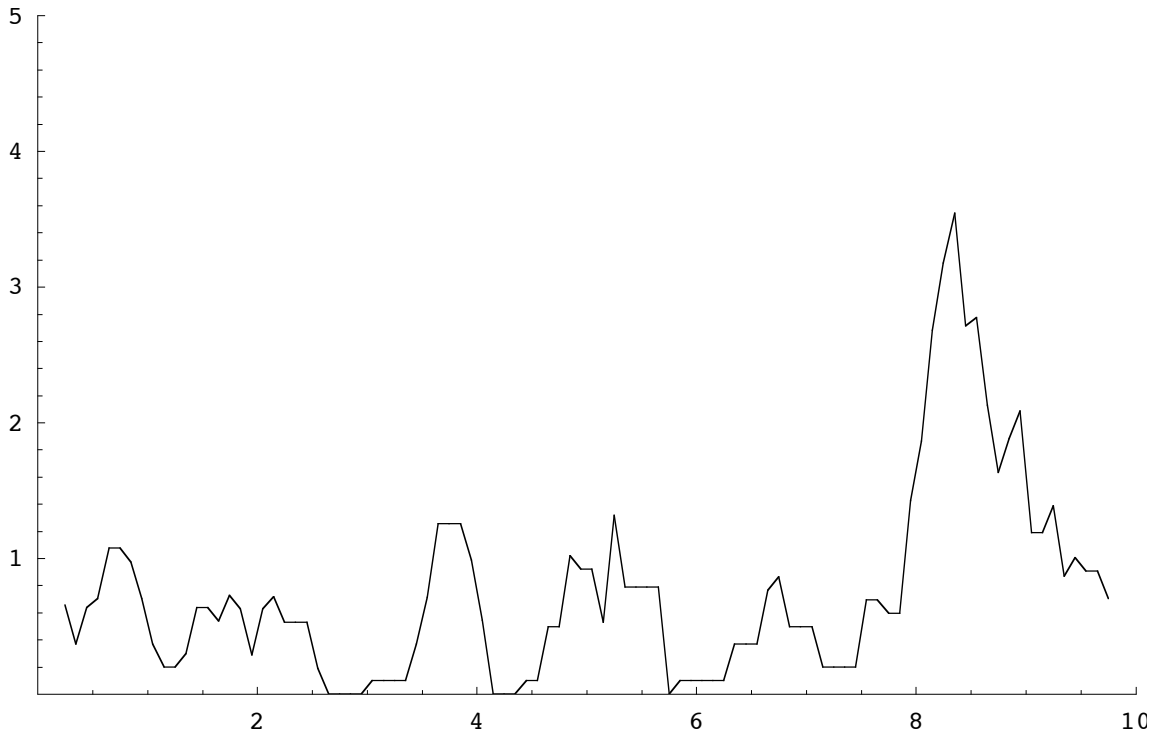
Numbers of segregating sites (20 sampled genomes; $\theta = 4\, N_e\, \mu = 30$; sliding window width 0.5)

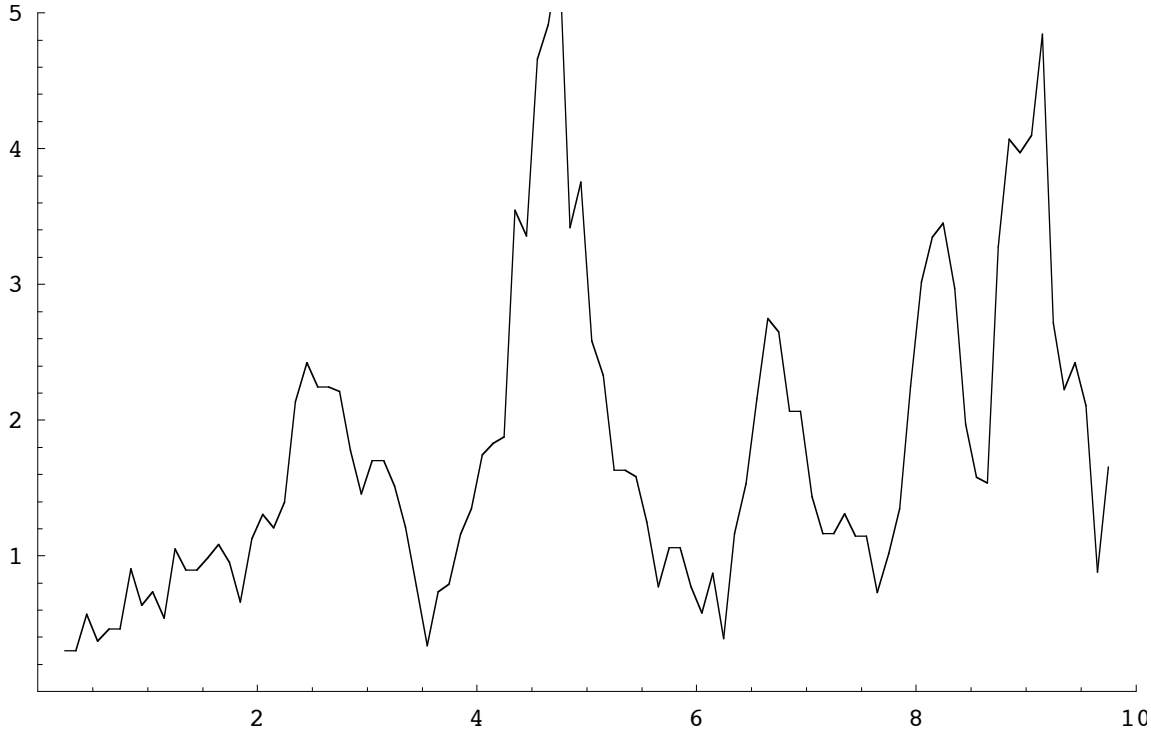


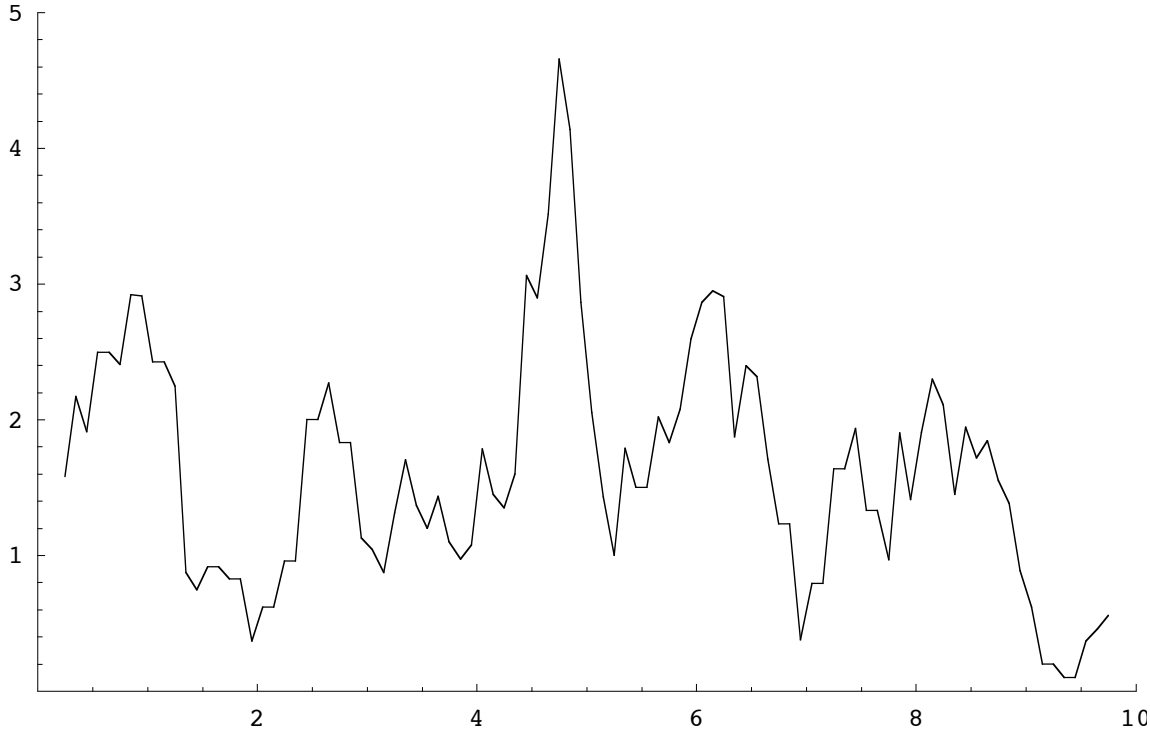




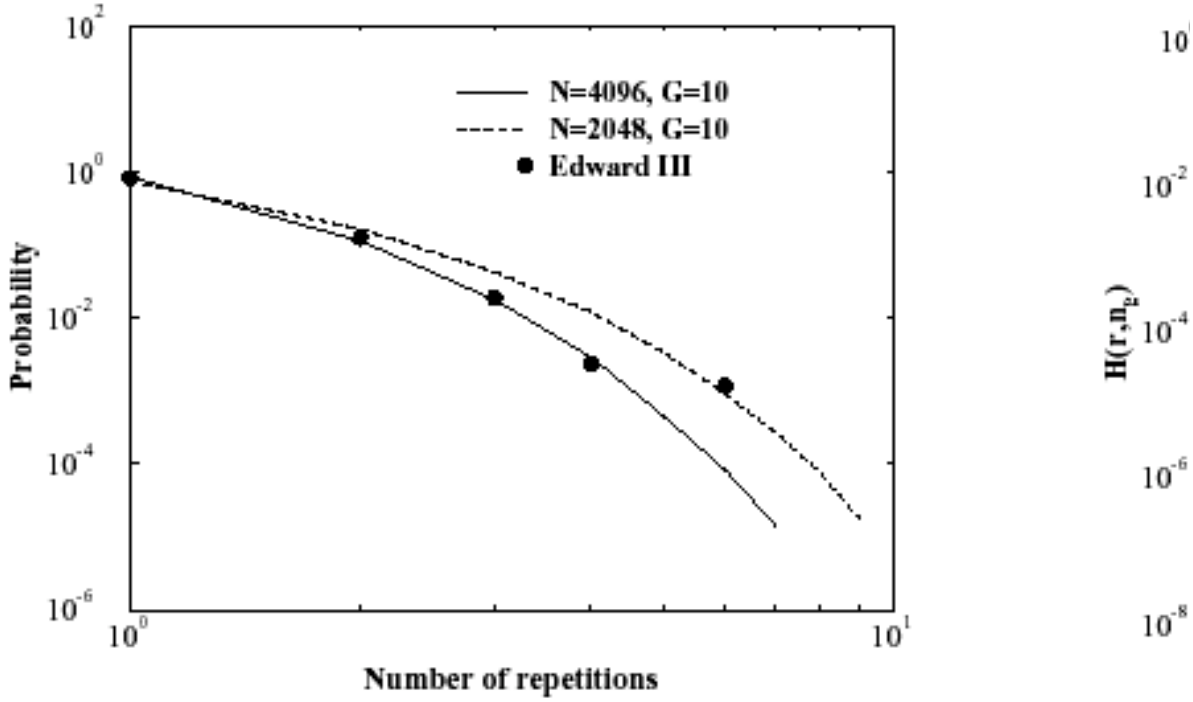
Mean number of pairwise differences:


5
4
3
2
1
2
4
6
8
10


5
4
3
2
1
2
4
6
8
10



■ **Pedigrees - or an infinitely long genome**



Probability of ancestor repetitions in the genealogical tree of the king Edward III. The continuous and dashed lines show simulations of F[r] in a closed population with 2^{11} and 2^{12} individuals for our model.

Distributio
13, 15, 17, 1

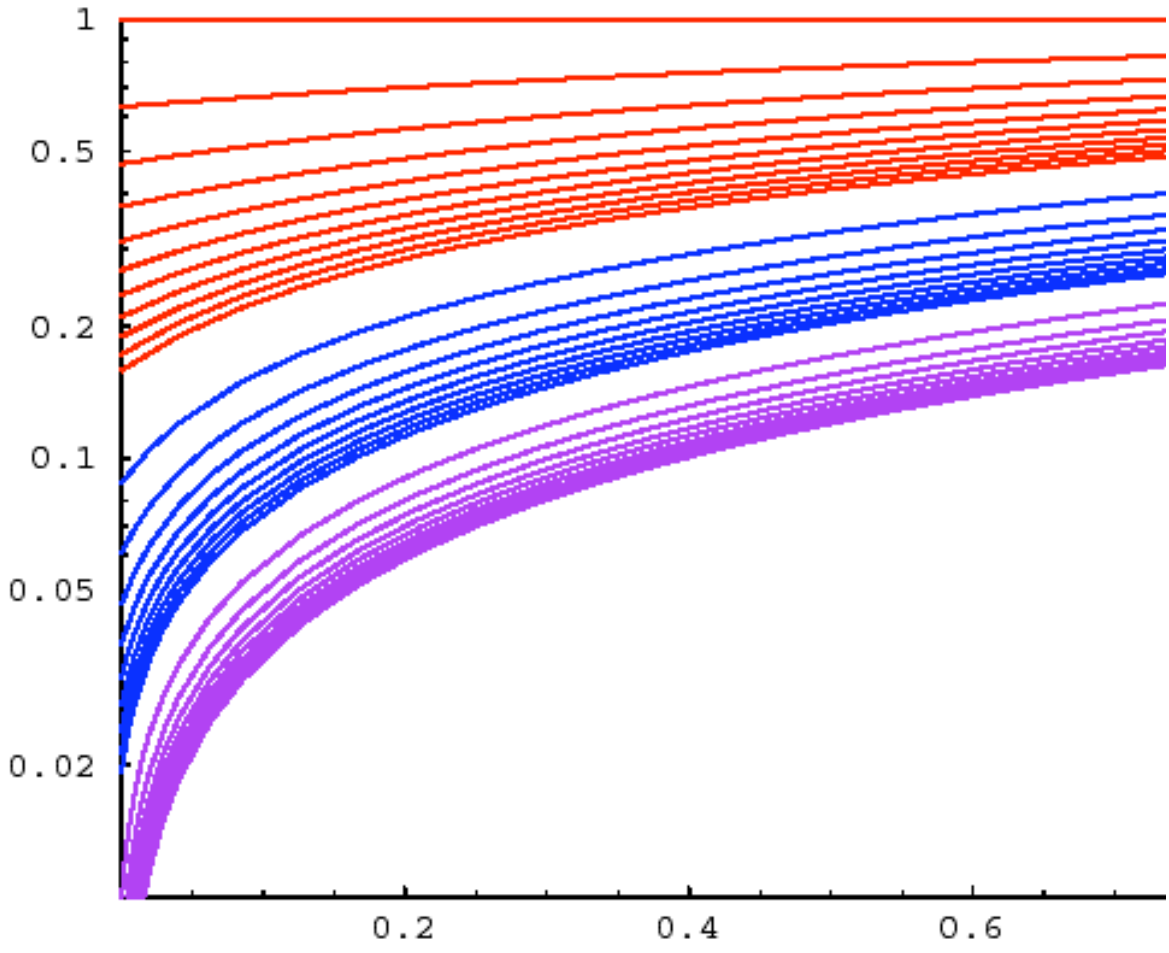
Derrida, B., S. C. Manrubia, and D. H. Zanette. 1999. Statistical properties of genealogical trees. Physical Review Letters 82:1987-1990.

■ Forwards in time

What is the fate of a single ancestral genome?
In an infinitely large population, this is a branching process.

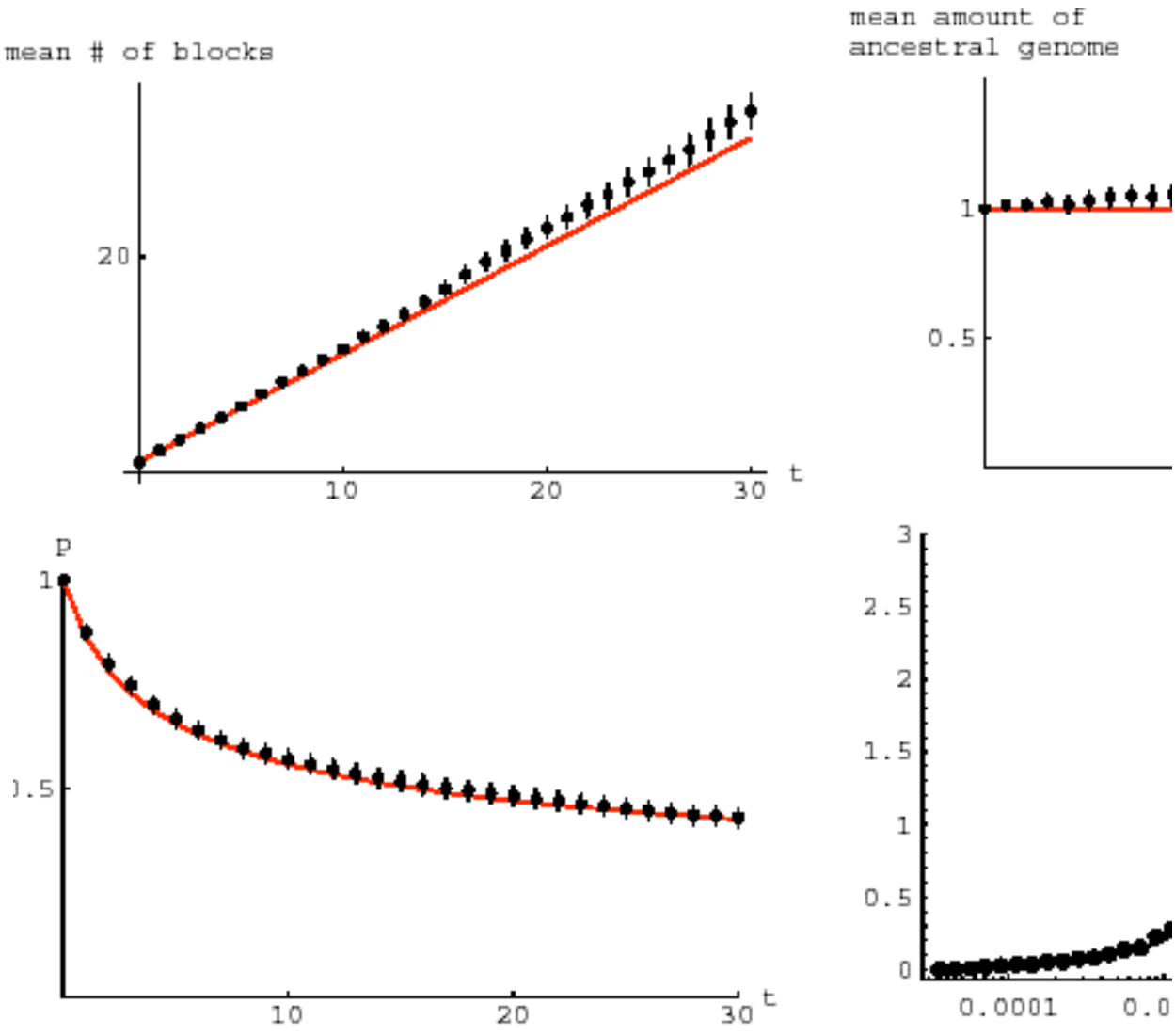
The chance that the *pedigree* will survive is ~ 80%

Any finite piece of genome is certain to be lost - but very slowly



The probability of survival of a neutral genome (S = 0) as a function of map length, R. From top to bottom, the
curves show P_t [R] for t = 0, 1, 2... 10; 20, 30...100; and 200, 300...1000 generations.

The distribution of blocks of genome that remain after 50 generations; map length
R = 1. The two panels show two random realisations of this process. Each line represents one genome.



The increase in mean block number over time (±1 standard error), compared with the expectation

$1+Rt$. (b) The mean amount of ancestral material over time , compared with the constant expectation

R. (c) The probability of survival, P, compared with the value calculated from Eq. 2. (d) The distribution of

block sizes at time $t = 30$ compared with the expectation. (R=1).

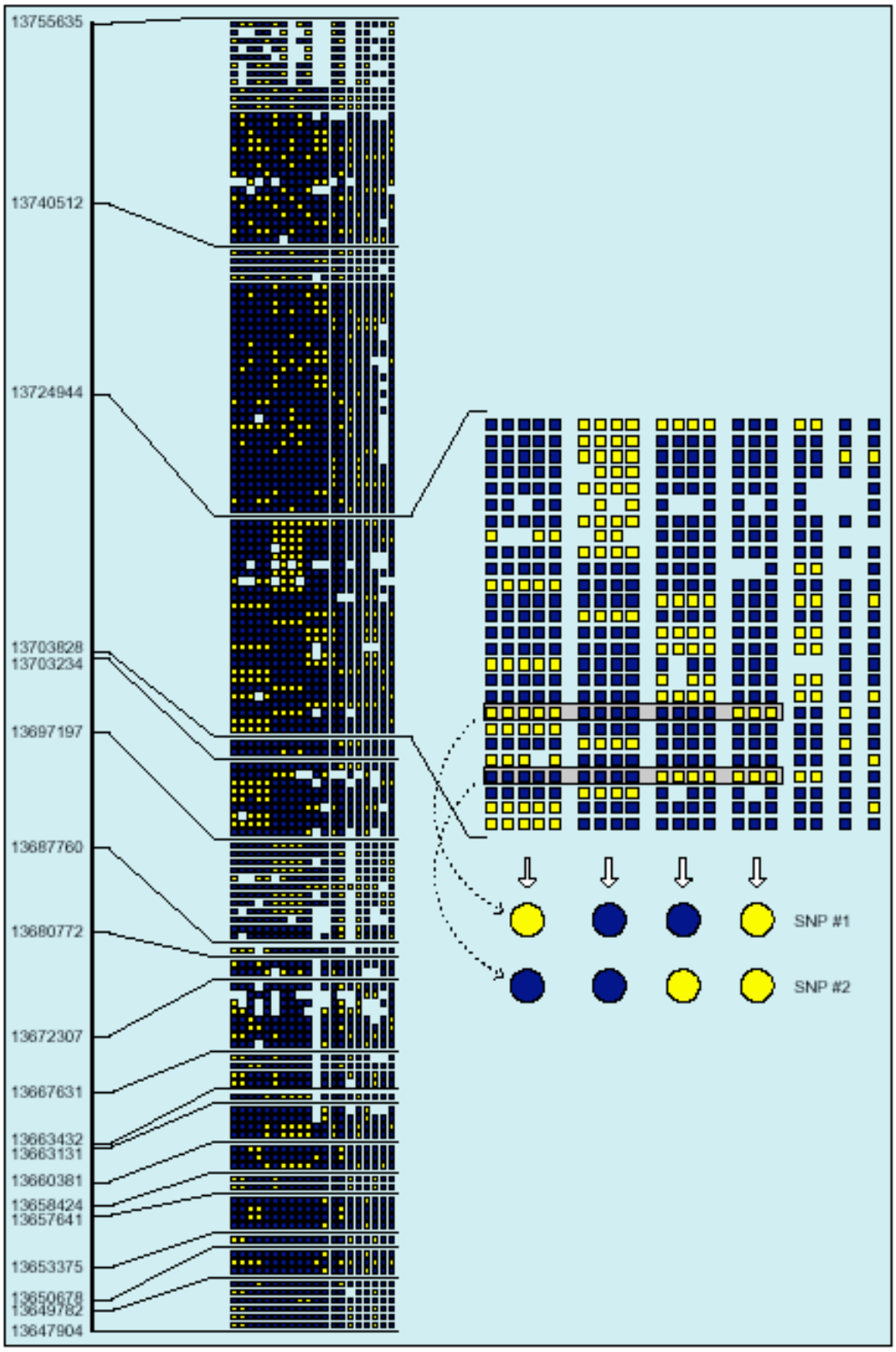
■ What do we see?

What is the relation between the ancestry of segments of genome, and the patterns we see?

Patil et al. 2001 Science 294:1719

21,676,868 bases, 36000 SNPs;
~4000 "blocks" identified; ~2700 SNPs capture ~80% of haplotype variation

What is the actual structure of these 20 chromosomes?


13755635
13740512
13724944
13703828
13703234
13697197
13687760
13680772
13672307
13667631
13663432
13663131
13660381
13658424
13657641
13653375
13650678
13649782
13647904
SNP #1
SNP #2

Selection on linked sites

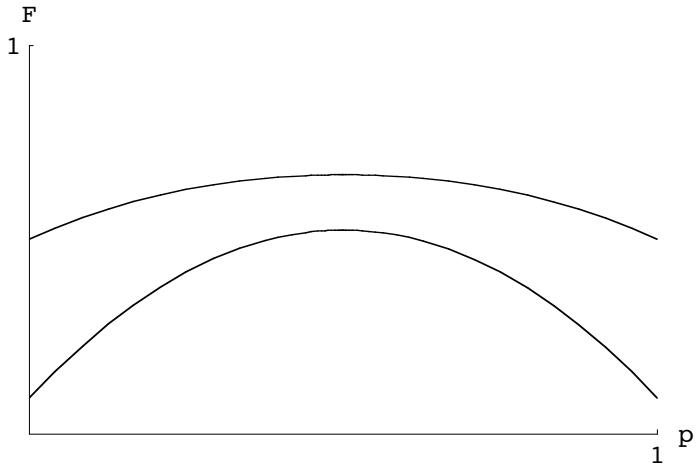
■ Balancing selection

■ Complete linkage

- Kreitman & Aguade (Genetics, 1986) observed excess polymorphism in the Adh region of *D. melanogaster*.
- Hudson, Kreitman & Aguade (Genetics, 1987) introduced the "HKA test" to detect balancing selection.
- A polymorphism with two alleles P, Q divides linked markers into two separate gene pools.
- Eventually, there will be a set of alleles with homozygosity $\frac{1}{(1+4\,\mathrm{N}p\mu)}$ associated with P, and a distinct set associated with Q, with homozygosity $\frac{1}{(1+4\,\mathrm{N}q\mu)}$. The overall homozygosity is:

$$F = \frac{p^2}{1 + 4\,N\,\mu\,p} + \frac{q^2}{1 + 4\,N\,\mu\,q}$$

e.g. 1-F vs p for 4Nμ = 0.1 (bottom), θ=1 (top):



■ Recombination

We must follow identities between genes both associated with P, F_{PP}, both with Q, F_{QQ}, or one with each, F_{PQ}

$$F'_{PP} = (1 - r\,q)^2\,F_{PP} + 2\,r\,q\,(1 - r\,q)\,F_{PQ} + r^2\,q^2\,F_{QQ}$$

Assuming r small:

$$\delta F_{PP} = 2\,r\,q\,(F_{PQ} - F_{PP})$$
$$\delta F_{PQ} = r\,(q\,F_{QQ} + p\,F_{PP} - F_{PQ})$$
$$\delta F_{QQ} = 2\,r\,p\,(F_{PQ} - F_{QQ})$$

The effects of mutation and drift can be found in a similar way.Overall:

$$\delta F_{PP} = -2\,\mu\,F_{PP} + 2\,r\,q\,(F_{PQ} - F_{PP}) + \frac{(1 - F_{PP})}{2\,N\,p}$$
$$\delta F_{PQ} = -2\,\mu\,F_{PQ} + r\,(q\,F_{QQ} + p\,F_{PP} - F_{PQ})$$
$$\delta F_{QQ} = -2\,\mu\,F_{QQ} + 2\,r\,p\,(F_{PQ} - F_{QQ}) + \frac{(1 - F_{PP})}{2\,N\,q}$$

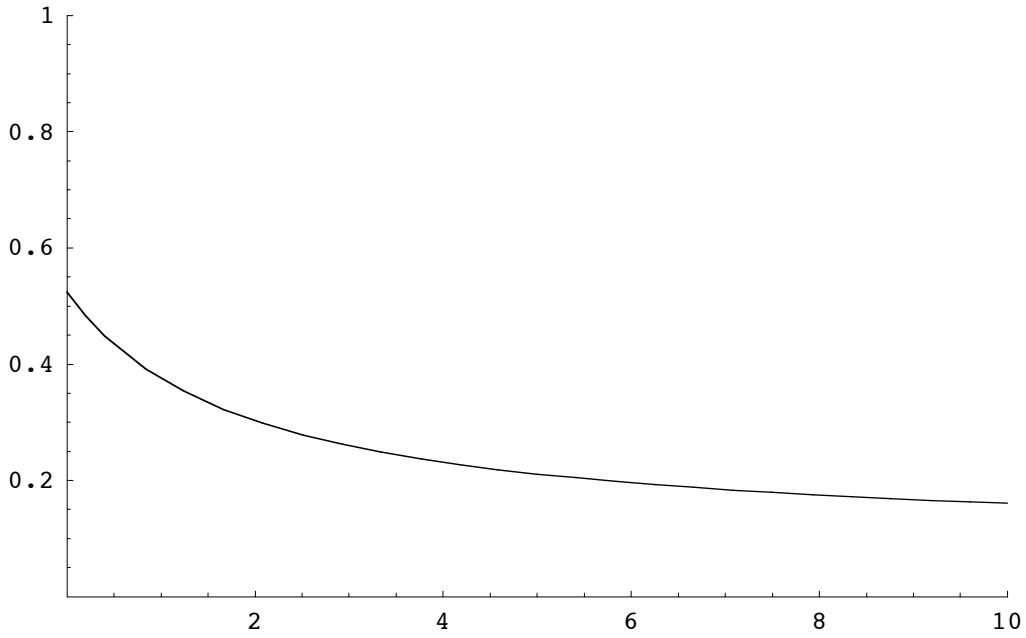
At equilibrium, δF=0. The average F is:

$$\bar{F} = (2 + \rho - 4\,p\,q\,(1 - N\,\mu\,(2 + 3\,\rho + \rho^2)))\,/$$
$$(2 + \rho + 4\,N\,\mu\,(2 + (1 + 4\,pq)\,\rho + p\,q\,\rho^2) + 16\,N^2\,\mu^2\,p\,q\,(2 + 3\,\rho + \rho^2))$$

where ρ=r/μ.

Note that the effect is only over recombination rates of order μ

- **Plot of heterozygosity $(1 - \overline{F})$ against r/μ for 4Nμ = 0.1**




```
ss = Solve[{0 == -2 μ F_PP + 2 r q (F_PQ - F_PP) + (1 - F_PP)/(2 n p),
0 == -2 μ F_PQ + r (q F_QQ + p F_PP - F_PQ),
0 == -2 μ F_QQ + 2 r p (F_PQ - F_QQ) + (1 - F_QQ)/(2 n q)}, {F_PP, F_PQ, F_QQ}];
```

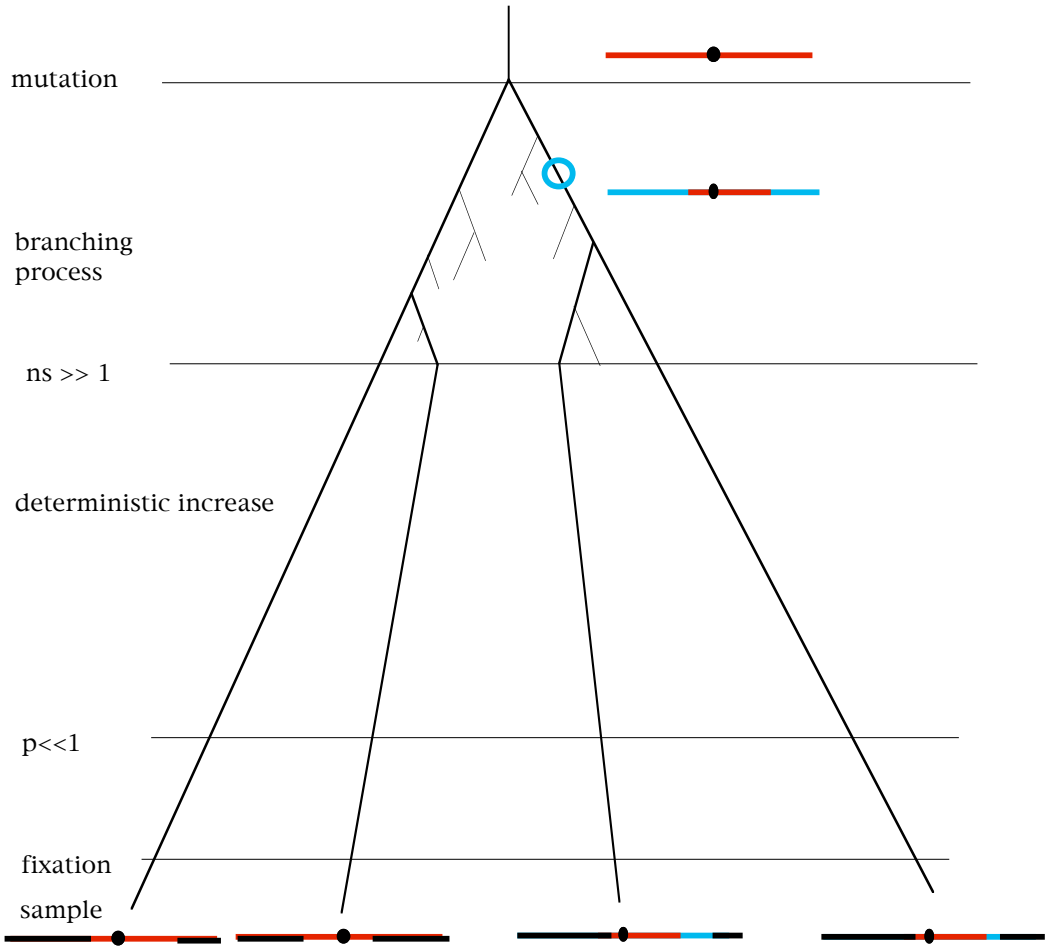
```
({F_PP, F_PQ, F_QQ, p^2 F_PP + 2 p q F_PQ + q^2 F_QQ} /. ss[[1]] /.
     {μ -> γ μ, r -> γ ρ μ, n -> 1 / (γ nn), q -> 1 - p} //
    Cancel) /. nn -> 1 / n // Simplify
```

```
{((2 + ρ) (-1 + 4 n (-1 + p) μ (1 + p ρ))) /
  (-2 - ρ + 16 n^2 (-1 + p) p μ^2 (2 + 3 ρ + ρ^2) +
    4 n μ (-2 + (-1 - 4 p + 4 p^2) ρ + (-1 + p) p ρ^2)),
 (ρ (-1 + 4 n (-1 + p) p μ (2 + ρ))) /
  (-2 - ρ + 16 n^2 (-1 + p) p μ^2 (2 + 3 ρ + ρ^2) +
    4 n μ (-2 + (-1 - 4 p + 4 p^2) ρ + (-1 + p) p ρ^2)),
 ((2 + ρ) (-1 + 4 n p μ (-1 + (-1 + p) ρ))) /
  (-2 - ρ + 16 n^2 (-1 + p) p μ^2 (2 + 3 ρ + ρ^2) +
    4 n μ (-2 + (-1 - 4 p + 4 p^2) ρ + (-1 + p) p ρ^2)),
 -(2 + ρ + 4 p (-1 + n μ (2 + 3 ρ + ρ^2)) - 4 p^2 (-1 + n μ (2 + 3 ρ + ρ^2))) /
   (-2 - ρ + 16 n^2 (-1 + p) p μ^2 (2 + 3 ρ + ρ^2) +
     4 n μ (-2 + (-1 - 4 p + 4 p^2) ρ + (-1 + p) p ρ^2))}
```

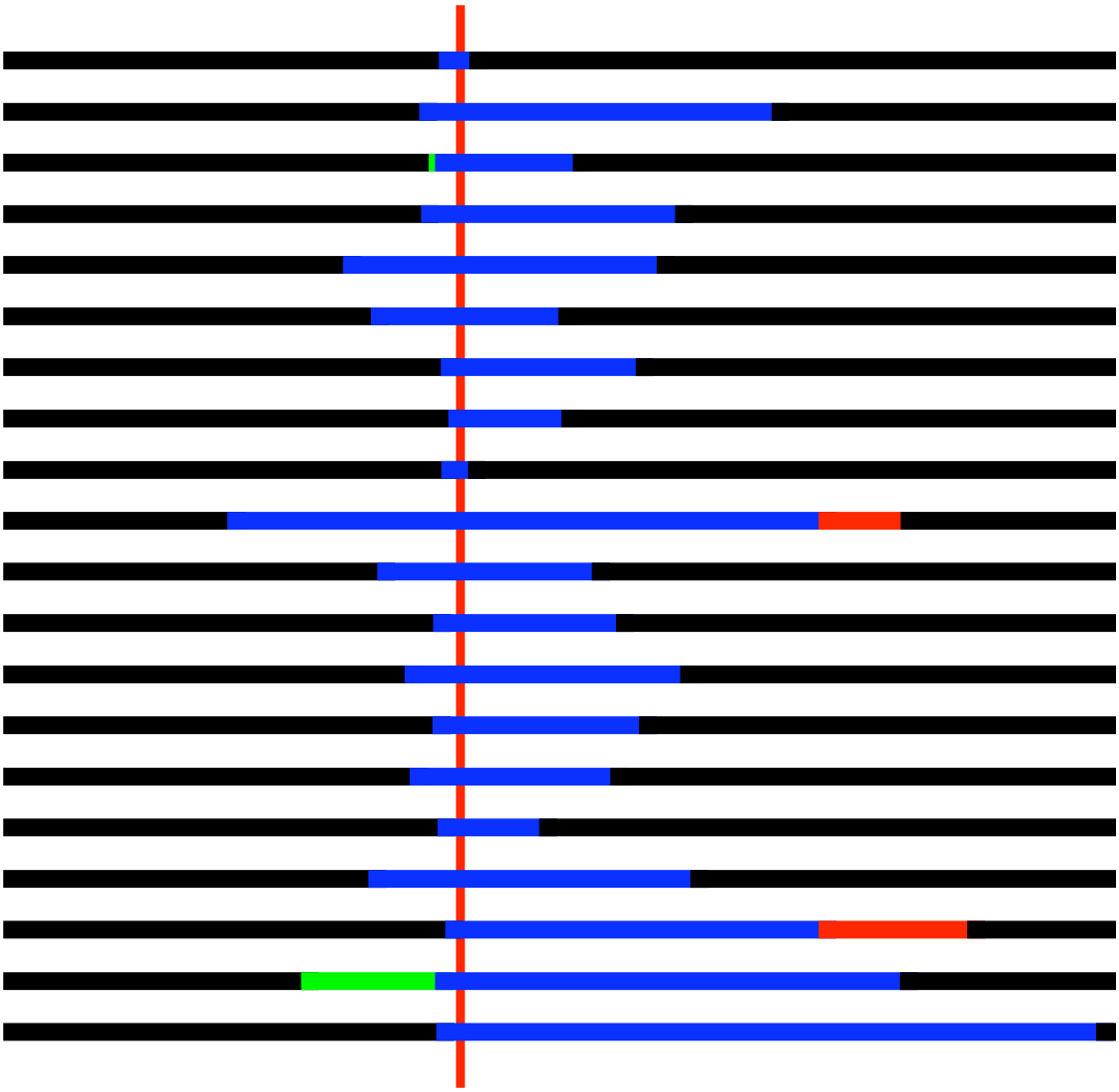
```
Plot[
  1 + (2 + ρ + 4 p (-1 + n μ (2 + 3 ρ + ρ^2)) - 4 p^2 (-1 + n μ (2 + 3 ρ + ρ^2))) /
     (-2 - ρ + 16 n^2 (-1 + p) p μ^2 (2 + 3 ρ + ρ^2) +
       4 n μ (-2 + (-1 - 4 p + 4 p^2) ρ + (-1 + p) p ρ^2)) /.
   {n -> 0.025 / μ, p -> 1 / 2, ρ -> Abs[ρ]}, {ρ, 0, 10},
  PlotRange -> {{0, 10}, {0, 1}}];
```

■ Selective sweeps

Fixation of a single favourable mutation carries with it a segment of linked genome



An example: s = 0.1, $N = 10^5$, sampled when p = 0.1. r = {-0.05, 0.15}
s/Log[2N] = 0.008



Fixation takes $\sim \frac{\text{Log}[2\,N]}{s}$ generations, so a region of $r \sim \frac{s}{\text{Log}[2\,N]}$ has reduced diversity

- **References**

Maynard Smith, J., and J. Haigh. 1974. The hitch-hiking effect of a favourable gene. Genet.Res. 23:23-35.
Hudson, R. B., and N. L. Kaplan. 1988. The coalescent process in models with selection and recombination. Genetics 120:831-840.
Kaplan, N. L., R. R. Hudson, and C. H. Langley. 1989. The hitch-hiking effect revisited. Genetics 123:887-899.
Barton, N. H. 2000. Genetic hitch-hiking. Philosophical Transactions of the Royal Society (London) B 355:553-1562.
Kim, Y., and W. Stephan. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160:765-777.
Gillespie, J. H. 2001. Is the population size of a species relevant to its evolution? Evolution 55:2161-2169.

Monte Carlo methods

- **Generalities**

How can we make inferences from genetic data?
- statistics such as # of segregating sites, pairwise diversity…
- likelihood: the probability of observing the data, given some hypothesis

Statistical inference:
- significance tests
- likelihood
- Bayesian inference

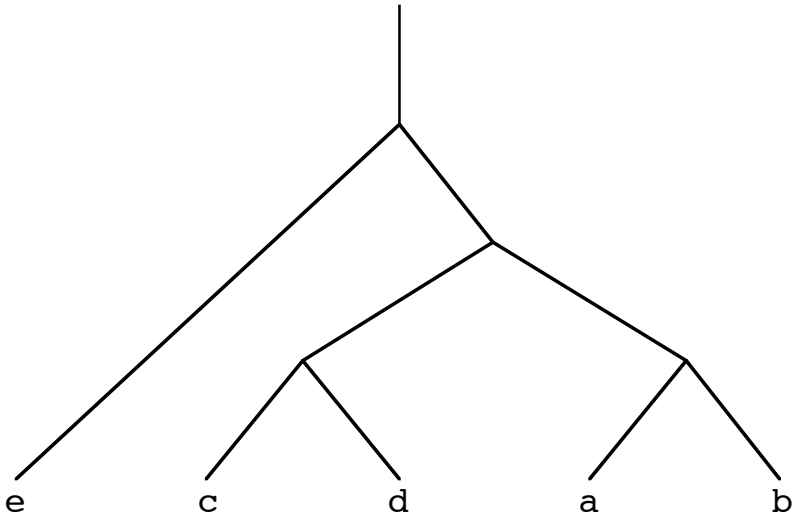
- **Griffiths-Tavare**

Griffiths, R. C., and S. Tavare. 1994. Simulating probability distributions in the coalescent. Theoretical Population Biology 46:131-159.

We observe some configuration of mutations:

	1	2	3	4	5	6	7	8	9
a	0	0	0	1	1	0	1	0	0
b	0	0	0	1	1	0	0	1	1
c	1	1	1	0	0	1	0	0	0
d	1	1	1	0	0	0	0	0	0
e	0	0	0	0	0	0	0	0	0

This configuration was produced by this genealogy:



This rooted genealogy cannot be fully reconstructed, because there were no mutations along the branchs leading down to *e* and to {a,b,c,d}

- **The algorithm (exact version):**

  - Work back along the genealogy, until the most recent mutation or coalescence

  - Sites can only lose a mutation if that mutation is represented only in one leaf; let there be J such sites. (In the example above, sites 6,7,8,9 are singletons; J=4).

  - A pair of lineages can only coalesce if they carry the same set of mutations; let there be K such pairs. In the example, there are no such possibilities: K=0.

  - With n lineages, the rate of events is $\lambda_n = n\frac{\theta}{2} + \frac{n(n-1)}{2}$; a sum is taken over these events, with the appropriate probability, and expressed in terms of the probabilities of the simpler configurations generated by loss of a mutation or coalescence.

  - This sum over J+K possible previous configurations is wighted by the overall weight $\frac{1}{\lambda}$:

$$P[S] = \frac{1}{\lambda_n}\left(\sum_{j=1}^{J}\frac{\theta}{2}P[S_j^*] + \sum_{k=1}^{K}P[S_k^*]\right) \text{ where } \lambda_n = n\frac{\theta}{2} + \frac{n(n-1)}{2}$$

  S_j^* represents deletion of the j' th singleton site from S, and S_k^* the coalescence of the k' th pair.

  This algorithm becomes extremely slow for large numbers of mutations and lineages.

- **Monte Carlo version:**

  A Monte Carlo estimate can be made by sampling possible paths back through the genealogy, with relative probability $\frac{\phi}{2}$ for possible losses of mutations, and 1 for possible coalescences:

$$P[S] = \left(\frac{\theta}{\phi}\right)^m E\left[\prod \frac{1}{\lambda_i}\left(\frac{\phi}{2}J_i^* + K_i^*\right)\right]$$

  where J_i^* is the number of possible losses of mutations, K_i^* the number of possible coalescences,
  m the number of segregating sites, and i the current # of lineages

The parameter ϕ can be chosen arbitrarily: it should take a value which minimises the variance of the estimator. Note that while $\phi=\theta$ seems natural, it does not give an optimal estimator.

- **Other applications:**

Joint estimation of recombination and mutation ($4 N_e r,\ 4 N_e \mu$) :

Kuhner, M. K., J. Yamato, and J. Felsenstein. 2000. Maximum likelihood estimation of recombination rates from population data. Genetics 156:1393-401.
Fearnhead, P., and P. Donnelly. 2001. Estimating recombination rates from population genetic data. Genetics 159:1299-1318.

Estimation of population structure:

Beerli, P., and J. Felsenstein. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proceedings of the National Academy of Sciences (U.S.A.) 98:4563-4568