Stochastic models in Mathematical Genetics (SC1)

- Simon Myers

- Email: myers@stats.ox.ac.uk

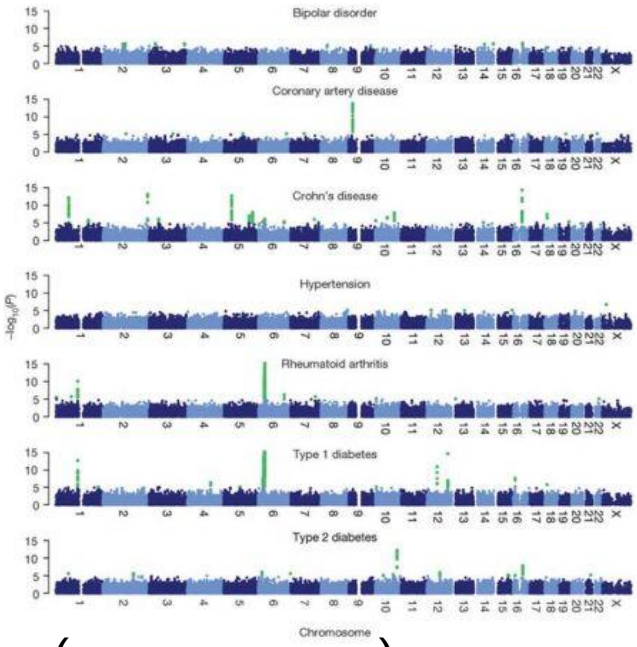
- Course webpage: www.stats.ox.ac.uk/~myers/mathgen.html

- Class problem sheets (7 sheets) are posted online

- Notes are also posted online. (First four weeks.)

Population genetics

Figure 2: Nests with both host and parasitic common cuckoo eggs, illustrating near-perfect mimicry to the human eye. Black arrows identify cuckoo egg.



(Wellcome Trust, 2010)

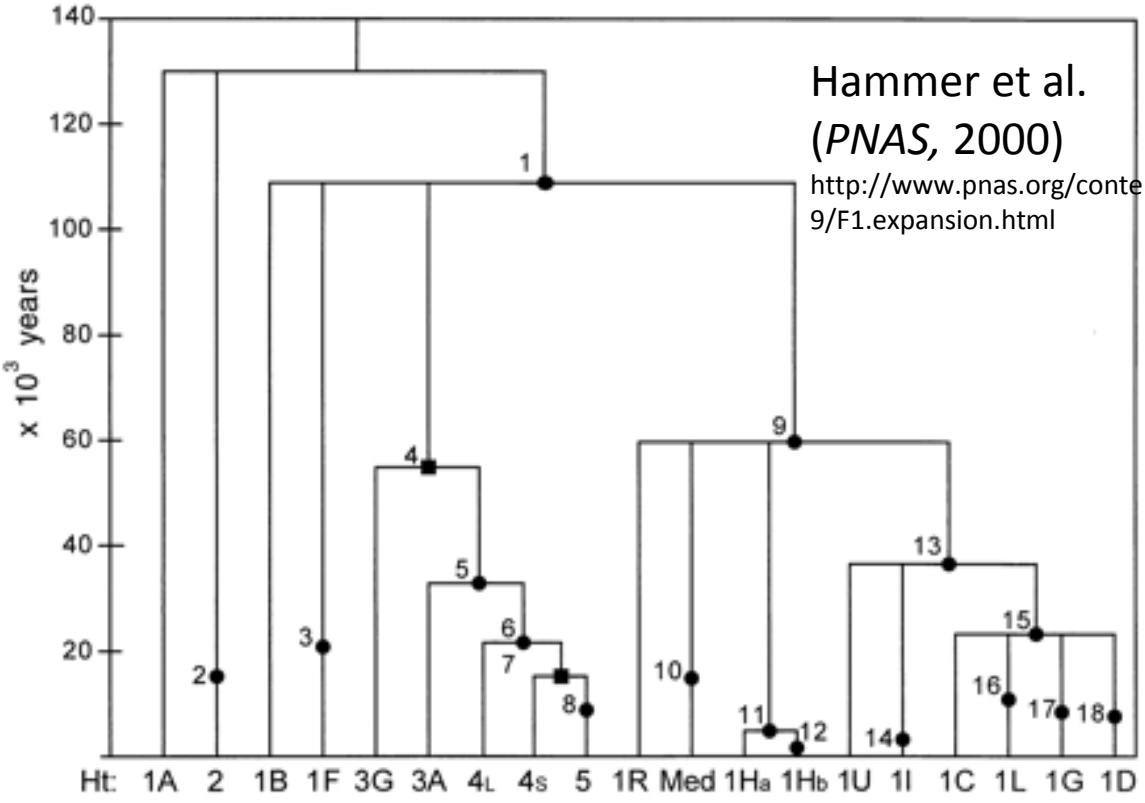
All organisms differ – due to *genetic variation*

- Unrelated people differ more than relatives
- Still, people share characteristics
  - Hair colour, eye colour, disease susceptibility, colour-blindness, blood group.....

- These characteristics are all the result of genetic variation
- Why, and how, do we expect mutations to be shared?

- The answer comes only by carefully considering models for genetic data, and their implications
- We need to look “back in time” to discover how mutations arise and spread in a population

Y-chromosome "genealogical tree"



Hammer et al. (*PNAS,* 2000)
http://www.pnas.org/content/97/12/6769/F1.expansion.html

- The axis is time in thousands of years, for a sample of 1,371 human males. Tree built using DNA sequences.
- Black circles represent mutations seen in those samples
- Shared characteristics come from rare mutations in the distant past
- This tree and the times on it were inferred based on the *coalescent* model
- Computationally intensive inference (Griffiths and Tavaré, 1994)
- This model, its derivation, its properties and inference under the model are what we will look at first, using : stochastic processes ,and graph theory.

Outline of the course

- Two parts, of 8 lectures each
- Part I (weeks 1-4)
  - The “neutral model”
  - Modelling genetic data
  - Genealogical relationships
  - Mutation patterns in populations
- Part II (weeks 5-8)
  - Extending the neutral model
  - Recombination and “shuffling of genetic material”
  - Natural selection
  - Diffusion process models in genetics

The Wright-Fisher model

- Suppose we are interested in a fragment of DNA, which might look like this:

`AC..AAACGTTTAGCCGAT...GG`

- There are *M* (very similar) copies of this fragment in the whole population
- *M* is often very large (>>1000)
- For now, we view each fragment as an "object", called a *haplotype* or *gene* or *sequence*
  - Could be a few positions as shown above, or the whole Y-chromosome of 58,000,000 letters (bases)
- Our task: model the *history* of these fragments in the population

The Wright-Fisher model

- Fisher, Wright (1930-31)
- “The simplest imaginable inheritance model”
- Models the evolution of a population *forward in time* from one generation to the next
- We then (approximately) go back in time
- Constant size population of *M* haplotypes
- Generations are *discrete,* and *independent*: in a generation, a complete new set of *M* haplotypes is created, and all *M* existing haplotypes die
- Each of the *M* new haplotypes inherits their genetic material from the previous generation, choosing their “parent” *independently* and *uniformly at random*

A picture makes this clearer.

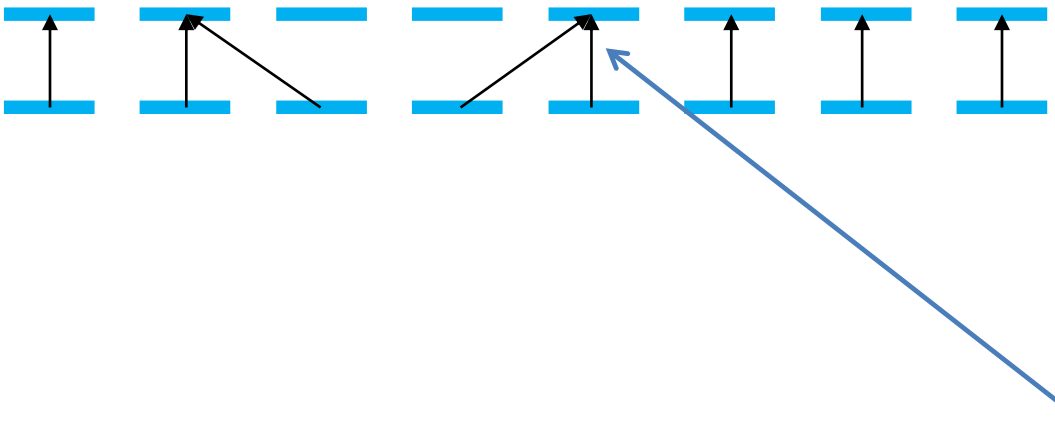
Formally, we form generation k+1 by choosing M "parents" at random in generation k with replacement

If parent of haplotype i in generation k+1 is Z_i

$$P(Z_i = j) = 1/M$$

$$j = 1,....,M$$

Some population members have 0 children, others more than 1 child:



Each haplotype chooses parent in previous generation

If haplotypes **share** a parent back in time, this is called a **coalescence event**

If we continue back in time, eventually a single parent is reached, the **Most Recent Common Ancestor (MRCA) :**

A picture makes this clearer.

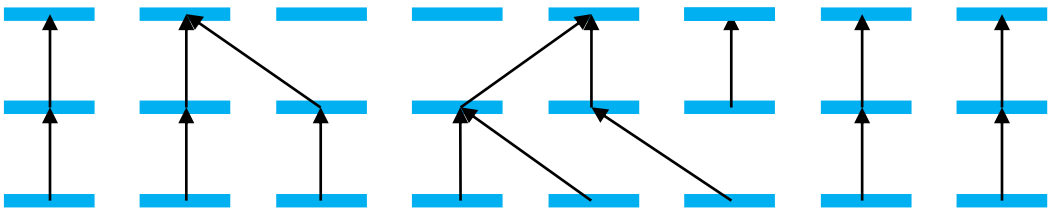
Formally, we form generation k+1 by choosing M "parents" at random in generation k with replacement

If parent of haplotype i in generation k+1 is Z_i

$$P(Z_i = j) = 1/M$$

$$j = 1,\ldots,M$$

Some population members have 0 children, others more than 1 child:



Each haplotype chooses parent in previous generation

If haplotypes **share** a parent back in time, this is called a **coalescence event**

If we continue back in time, eventually a single parent is reached, the **Most Recent Common Ancestor (MRCA) :**

A picture makes this clearer.

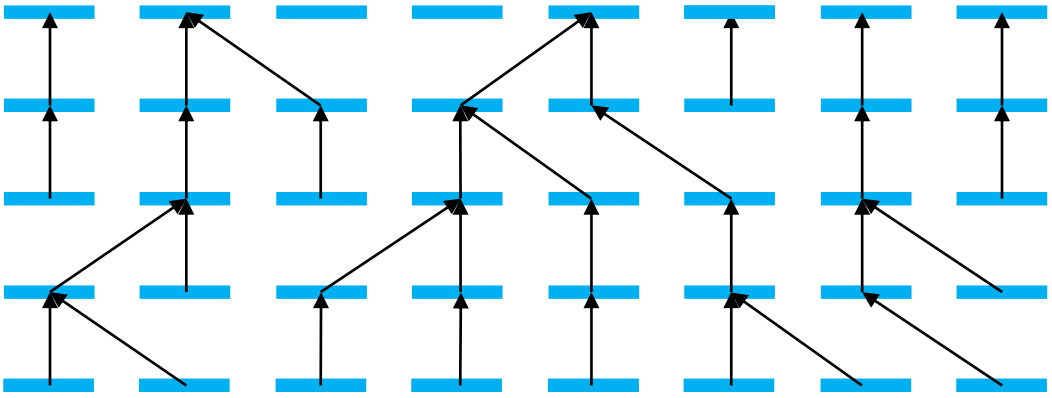
Formally, we form generation k+1 by choosing M "parents" at random in generation k with replacement

If parent of haplotype i in generation k+1 is Z_i

$$P(Z_i = j) = 1/M$$

$$j = 1, \ldots, M$$

Some population members have 0 children, others more than 1 child:



Each haplotype chooses parent in previous generation

If haplotypes **share** a parent back in time, this is called a **coalescence event**

If we continue back in time, eventually a single parent is reached, the **Most Recent Common Ancestor (MRCA) :**

A picture makes this clearer.

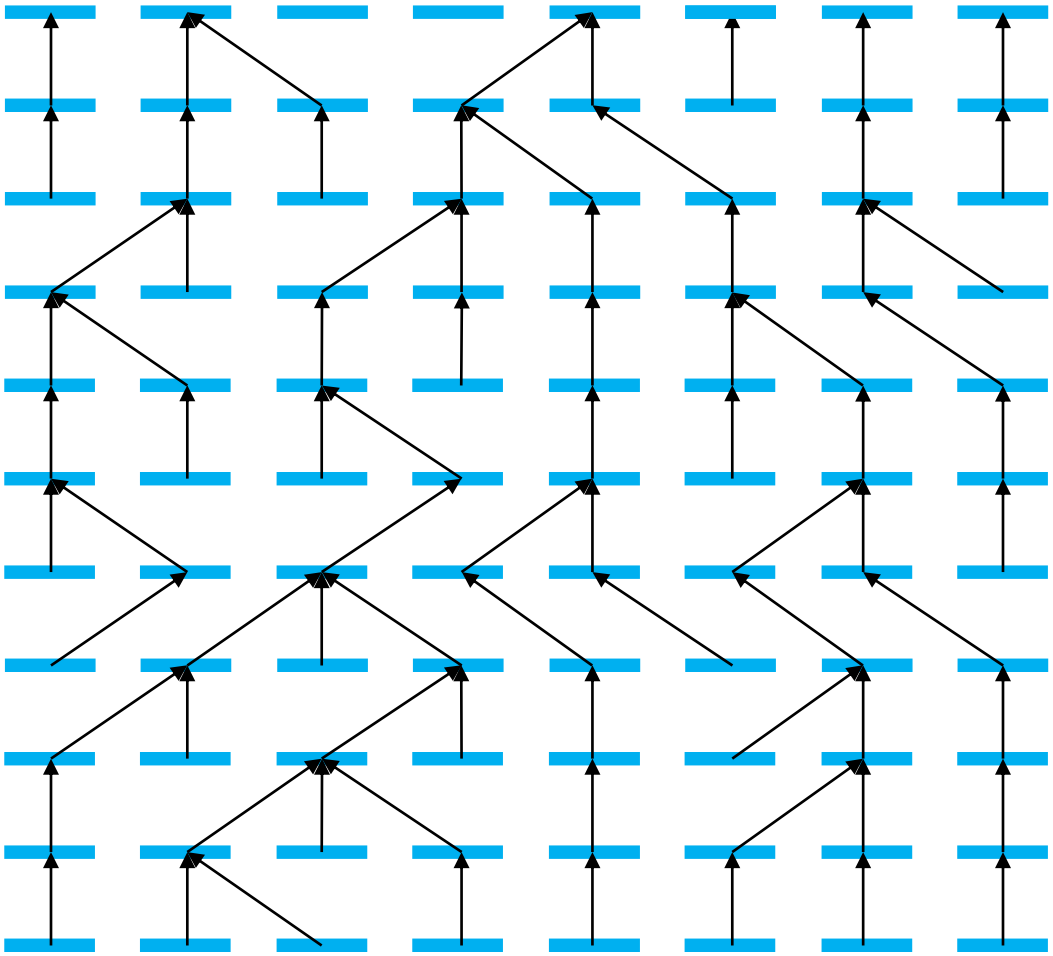
Formally, we form generation k+1 by choosing M "parents" at random in generation k with replacement

If parent of haplotype i in generation k+1 is Z_i

$$P(Z_i = j) = 1/M$$

$$j = 1,\ldots,M$$

Some population members have 0 children, others more than 1 child:



Each haplotype chooses parent in previous generation

If haplotypes **share** a parent back in time, this is called a **coalescence event**

If we continue back in time, eventually a single parent is reached, the **Most Recent Common Ancestor (MRCA) :**

A picture makes this clearer.

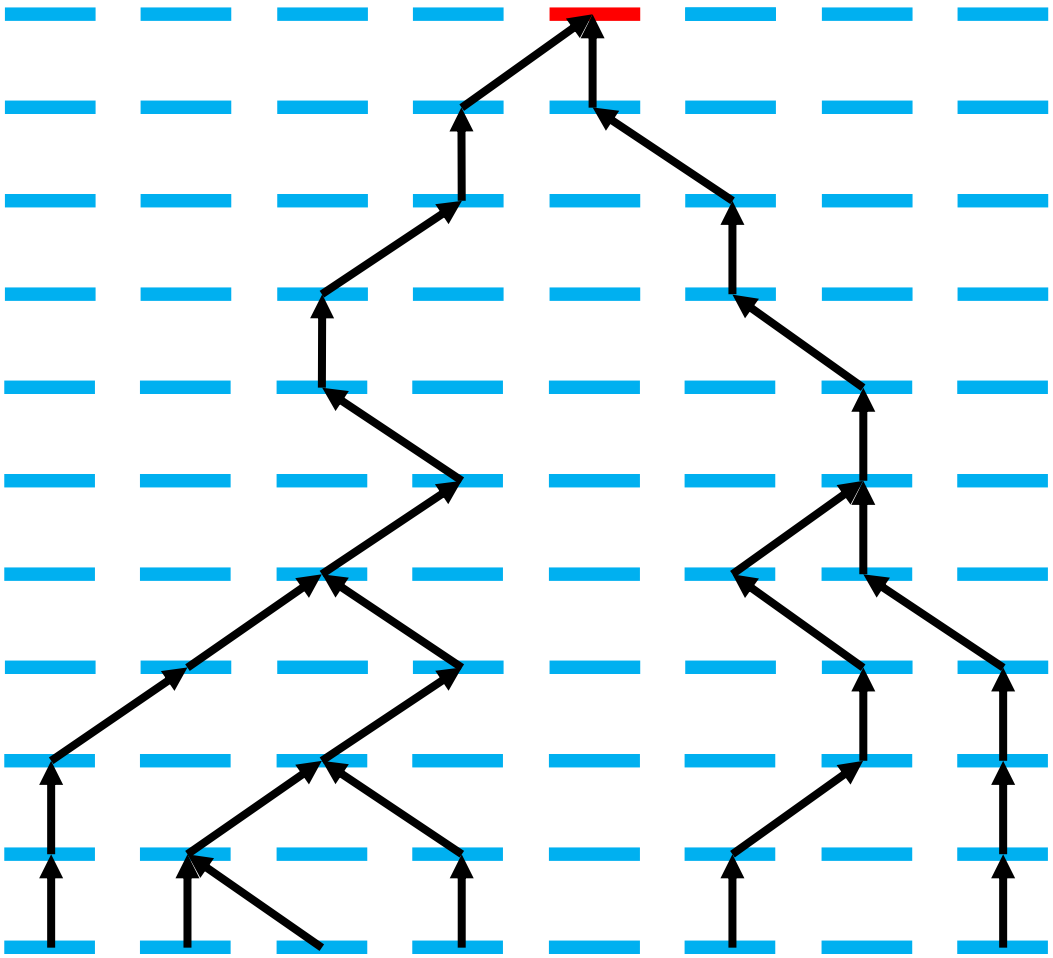
Formally, we form generation k+1 by choosing M "parents" at random in generation k with replacement

If parent of haplotype i in generation k+1 is Z_i

$$P(Z_i = j) = 1/M$$

$$j = 1,\ldots,M$$

Some population members have 0 children, others more than 1 child:



Each haplotype chooses parent in previous generation

If haplotypes **share** a parent back in time, this is called a **coalescence event**

If we continue back in time, eventually a single parent is reached, the **Most Recent Common Ancestor (MRCA) :**

We can examine historical relationships in a **sample**

Looking back in time

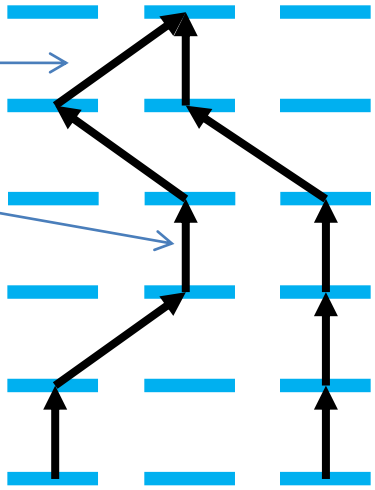
- We seek to understand the distribution of the relationships among individuals (haplotypes) backward in time
  - Why? Our DNA today is inherited from our ancestors
  - Looking at only "real" ancestors means we don't have to keep track of the entire population
- Given there are *i ancestral lineages* at generation k+1, **define** p_{ij} as the probability there are j parents in generation k, j=1,2,...,i
- Our "roadmap" is to proceed as follows:
  - We characterise p_{ij}
  - We assume M is large compared to i
  - This means the population size is big compared to a sample we take from it
  - In this setting, we use p_{ij} to approximate the distribution of the *total* time while exactly i ancestors remain
  - We rescale this time, to measure it in natural units
  - We show that as M becomes large, the whole (rescaled) backward process converges (beautifully) to a limit, called *the coalescent*

Example: two lineages

- Suppose we have two ancestors in generation *k*+1. Then in generation *k* there are 1 or 2 ancestors:

$$p_{21} = 1/M$$

$$p_{22} = (1-1/M)$$



- Generations are independent
- Define τ_2 as the time until a coalescence event occurs, then

$$P(\tau_2 = k) = \left(1-\frac{1}{M}\right)^{k-1}\frac{1}{M} \quad k = 1,2,\ldots$$

- Obviously, this is geometric . What happens if *M* is very large? Note the mean time $E(\tau_2) = M$
- If we measure time in units of *M* generations, the mean time is 1; independent of *M.* Proposition 1.0 shows that as *M* becomes large, this *rescaled* time has an exp(1) distribution.

Example: two lineages

Proposition 1.0
In a Wright-Fisher model with population size *M*, for a sample of size two taken from the population define τ_2 as the time back until a coalescence event occurs. Then setting $T_2=\tau_2/M$ to measure time in units of *M* generations, in the limit as $M\to\infty$, T_2 has an exponential distribution: T_2~exp(1).

Proof From above:

$$P(\tau_2 \le k) = 1 - P(\tau_2 > k)$$

$$= 1 - \left(1 - \frac{1}{M}\right)^k$$

$$P(T_2 \le t) = P(\tau_2 \le Mt) = P\left(\tau_2 \le \lfloor Mt \rfloor\right)$$

$$= 1 - \left(1 - \frac{1}{M}\right)^{\lfloor Mt \rfloor}$$

$$\to 1 - e^{-t} \text{ as } M \to \infty;$$

$$\text{for any } t > 0 : \text{the c.d.f of } \exp(1)$$

Note: by independence of generations, this extends to give the limiting time back until coalescence from any time point where we have two *lineages*.

M and coalescence times in humans and other animals

In humans, it is known that "appropriate" values for *M* are surprisingly small. This is approximation is called the "*effective population size*":

M ≈ 20,000 in Europe

M ≈ 19,000 in East Asia

M < 50,000 for all human populations, highest in Africa

M and coalescence times in humans and other animals

The mean coalescence time for two lineages is just $E(T_2) = 1$ in units of *M* generations, so if we have *G*=28 years per generation, the average ancestry depth for 2 human chromosomes is

$1 \times M \times G$ in years

(20,000-50,000) × 28 =480,000-1,400,000 years

M varies <u>widely</u> across species (Charlesworth, Nature Reviews Genetics 2009):

25,000,000 for *E.coli*

2,000,000 for fruit fly *D. Melanogaster*

<100 for Salamanders (Funk et al. 1999)

Samples of size *n*

- Suppose we are now following the history back in time of a sample of size *n*.
- Measure time backwards and suppose τ generations back, there are $\xi(\tau)$ lineages remaining, $\xi(0)=n$
- Generations are independent, so $\xi(\tau)$ behaves as a Markov process $\{\xi(\tau), \tau = 0, 1, \ldots\}$
  - In other words, given $\xi(0), \xi(1), \ldots, \xi(\tau)$ the distribution of $\xi(\tau+1)$ depends only on $\xi(\tau)$
  - The Markov process is *homogeneous* (does not vary across generations)
- Our p_{ij}'s define the transition matrix P:

$$p_{ij} = P(\xi(\tau+1) = j \,|\, \xi(\tau) = i)$$

- Question – given *i* lineages currently, what is the distribution of the *time* until the next coalescence event?
- At the next coalescence event, how many lineages coalesce?

Samples of size *n*

We consider the transition matrix:

$$p_{ij} = P(\xi(\tau+1) = j \,|\, \xi(\tau) = i) = P(i \to j)$$

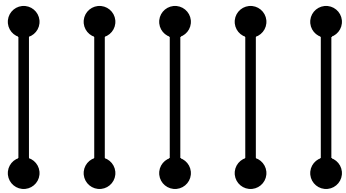
so the probability of no coalescence, $i \to i$:

$$p_{i,i} = \left(\frac{M-1}{M}\right) \times \left(\frac{M-2}{M}\right) \times \ldots \times \left(\frac{M-i+1}{M}\right)$$

$$= \left(1-\frac{1}{M}\right) \times \left(1-\frac{2}{M}\right) \times \ldots \times \left(1-\frac{i-1}{M}\right)$$

$$= 1 - \frac{1}{M}\sum_{k=1}^{i-1} k + O\left(M^{-2}\right)$$

$$= 1 - \frac{\binom{i}{2}}{M} + O\left(M^{-2}\right)$$



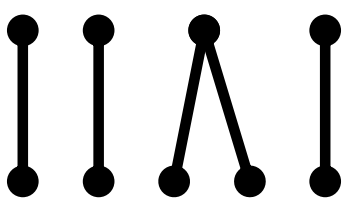
$i = 5, i \to i$

and the probability one pair of lineages coalesce,

so $i \to i$-1, is :

$$p_{i,i-1} = \left(\frac{1}{M}\right) \times \left(\frac{M-1}{M}\right) \times \ldots \times \left(\frac{M-i+2}{M}\right) \times \binom{i}{2}$$

$$= \binom{i}{2} \times \left(\frac{1}{M}\right) \times \left(1 - O\left(M^{-1}\right)\right) = \frac{1}{M}\binom{i}{2} + O\left(M^{-2}\right).$$



$i = 5, i \to i-1$

Samples of size *n*

What is the probability that more than one pair of lineages coalesce *at the same time*? This is

$$P(\xi(\tau+1) < i-1 \,|\, \xi(\tau) = i) = 1 - p_{i,i} - p_{i,i-1}$$

$$= 1 - \left[1 - \frac{\binom{i}{2}}{M} + O\left(M^{-2}\right) \right] - \frac{\binom{i}{2}}{M} + O\left(M^{-2}\right)$$

$$= 0 + O\left(M^{-2}\right)$$

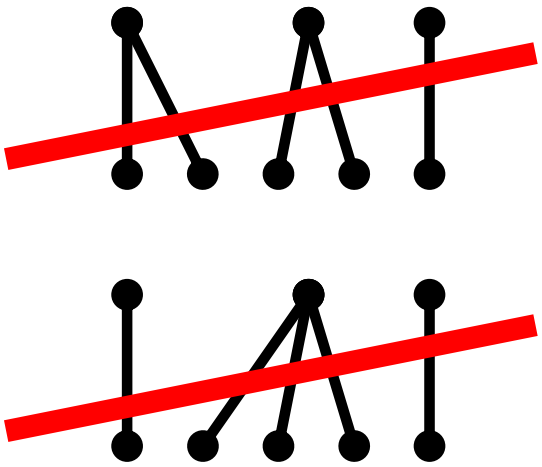
Putting this together, supposing *M* is large and we have *i* lineages, in a single generation, to $O\left(M^{-2}\right)$

$$p_{i,i} \approx 1 - \frac{\binom{i}{2}}{M}$$

$$p_{i,i-1} \approx \frac{\binom{i}{2}}{M}$$

$$p_{i,j} \approx 0$$

for any $j < i-1$

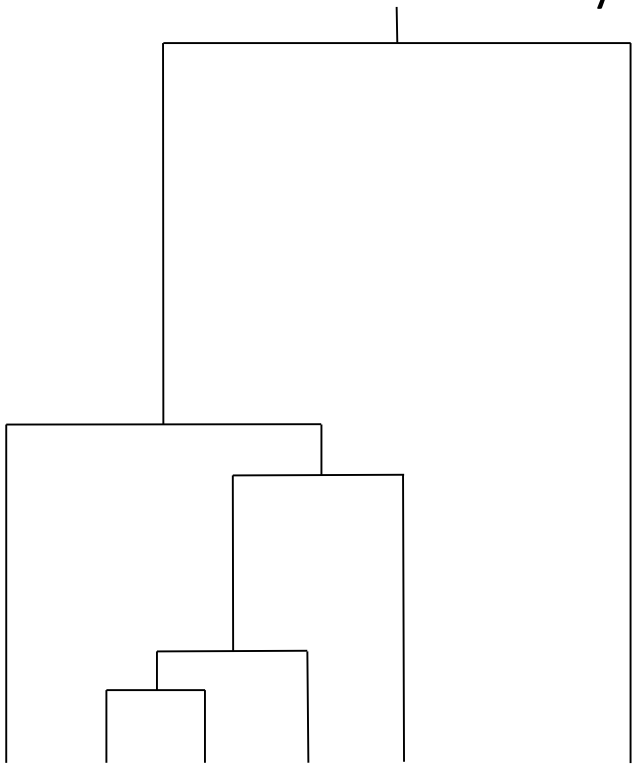


1. In words, asymptotically, only one pair of lineages can coalesce at a time – we have a binary tree.
2. By symmetry, each time a coalescence occurs, all pairs of lineages are equally likely to coalesce
3. To characterise the asymptotic distribution of trees, we then just need to derive the distribution of *times between coalescence events.* First, the answer.

The coalescent

(Kingman, Stochastic processes and their application, 1982)

- As for the two lineage case, for the general case we will measure time in unit of *M* generations
- We will show that the ancestral tree distribution in the Wright-Fisher model converges to the *coalescent*
- This is one of the greatest discoveries in population genetics
- **The coalescent limit actually applies much more widely than just the W-F case (e.g., models with non-discrete gens)**
- The coalescent can be adapted to include many realistic features, some which we will see (mutation, population size changes, recombination) and others we will not (population splits and merges, migration.....)
- The coalescent is a distribution on binary trees:



The coalescent

(Kingman, Stochastic processes and their application, 1982)

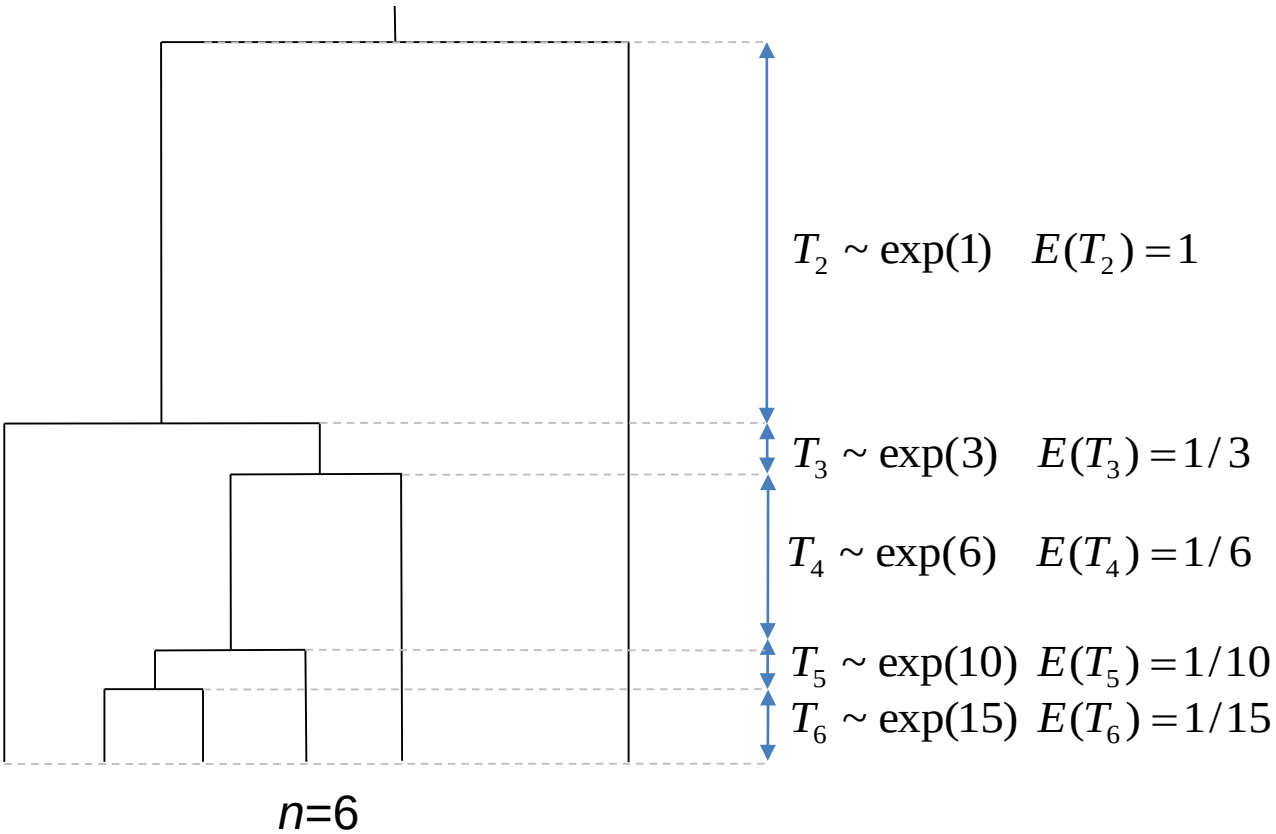
Definition 1.1

The *coalescent* is a distribution on binary trees. Starting with *n* lineages, pairs of lineages coalesce backward in time until a single common ancestor is reached. Defining times T_n, T_{n-1},...,T_2 while *n*,*n-1* ,..., 2 ancestors remain, the times T_j are independent and exponentially distributed:

$$T_j \sim \exp\left[\binom{j}{2}\right] \qquad f_j(t) = \binom{j}{2} e^{-\binom{j}{2}t}, \; t > 0 \qquad E(T_j) = \frac{2}{j(j-1)}$$

At the time of coalescence from *j* to *j*-1 lineages, a pair of lineages is chosen at random from the *j*(*j*-1)/2 possibilities and coalesces.

The coalescent:



The coalescent limit

(Kingman, Stochastic processes and their application, 1982)

Proposition 1.2
In the Wright-Fisher model, as the population size *M* converges to infinity, if time is measured in units of *M* generations then the distribution on ancestral trees for a sample of *n* sequences converges to the coalescent.

Proof:
We previously showed that as $M \to \infty$, only one pair of lineages coalesce at a time, so the limiting tree is binary.

In the Wright-Fisher model, sequences choose parents at random, so obviously all pairs of lineages are equally likely to be the one to coalesce at a coalescence event

It remains to show only that the coalescent gives the correct distribution on the rescaled times T_j while *j* edges remain in the tree. Suppose a sample has *j* ancestors at some time in the past. Recall we showed the probability *j* ancestors remain in the previous generation is given by

$$p_{jj} = 1 - \frac{\binom{j}{2}}{M} + O\left(M^{-2}\right)$$

If τ_j is the total time while *j* ancestors remain then

$$P(\tau_j > k) = \left[p_{jj}\right]^k = \left[1 - \frac{\binom{j}{2}}{M} + O\left(M^{-2}\right)\right]^k \quad \text{(independence)}$$

The coalescent limit

(Kingman, Stochastic processes and their application, 1982)

Proposition 1.2
In the Wright-Fisher model, as the population size *M* converges to infinity, if time is measured in units of *M* generations then the distribution on ancestral trees for a sample of *n* sequences converges to the coalescent.

Proof:

Now we rescale time in units of *M* generations. Set $T_j = \tau_j / M$.

We need to obtain the cdf of T_j in the limit $M \to \infty$.

Let $t \geq 0$ be any non-negative real.

$$P(T_j \leq t) = P(\tau_j \leq Mt) = 1 - P(\tau_j > \lfloor Mt \rfloor)$$

$$= 1 - \left[1 - \frac{\binom{j}{2}}{M} + O\left(M^{-2}\right) \right]^{\lfloor Mt \rfloor} \approx 1 - \left[1 - \frac{\binom{j}{2} t}{Mt} + O\left((Mt)^{-2}\right) \right]^{Mt}$$

$$\to 1 - e^{-\binom{j}{2} t} \text{ as } M \to \infty.$$

This is just the c.d.f of an $\exp\left[\binom{j}{2}\right]$ random variable.

$$E(T_j) = \frac{2}{j(j-1)}$$

Properties of the coalescent

The coalescent is the limit of a range of models. From now on, we will work using only this model (until week 6).

The coalescent describes what evolutionary history looks like in populations

We will see it allows us to study variation, its main use.

It also allows us to understand population history

What properties does it predict?

We will ask two things in particular:

1. How deep are genealogies in time? How variable are these depths?

2. What is the distribution of tree shapes under the coalescent – e.g. are they approximately symmetrical?

Times in the coalescent

In the coalescent the number of lineages decreases from *n* to 1. The time at which the final coalescence takes place is called the **time to the most recent common ancestor** (TMRCA)

$$W_n = T_n + T_{n-1} + \ldots + T_2$$

1.3 Mean and variance of the TMRCA

Immediately:

$$E(W_n) = E(T_n) + E(T_{n-1}) + \ldots + E(T_2)$$

$$= \frac{2}{n(n-1)} + \frac{2}{(n-1)(n-2)} + \ldots + \frac{2}{2\times 1}$$

$$= \sum_{j=2}^{n} \frac{2}{j(j-1)} = \sum_{j=2}^{n} \frac{2}{(j-1)} - \sum_{j=2}^{n} \frac{2}{j}$$

$$= 2\left(1 - \frac{1}{n}\right)$$

Times in the coalescent

1.3 Mean and variance of the TMRCA

Also, the times T_j are independent, so

$$Var(W_n) = Var(T_n) + Var(T_{n-1}) + \ldots + Var(T_2)$$
$$= \ldots$$

(problem sheet 2)

As sample size becomes very large: $n \to \infty$

$$E(W_n) = 2\left(1 - \frac{1}{n}\right) \to 2 \text{ as } n \to \infty$$

$$Var(W_n) < \infty$$

We can interpret this as saying the expected time to coalescence of the *whole population* is finite, with mean 2. We can build a tree for the whole population. In units of generations:

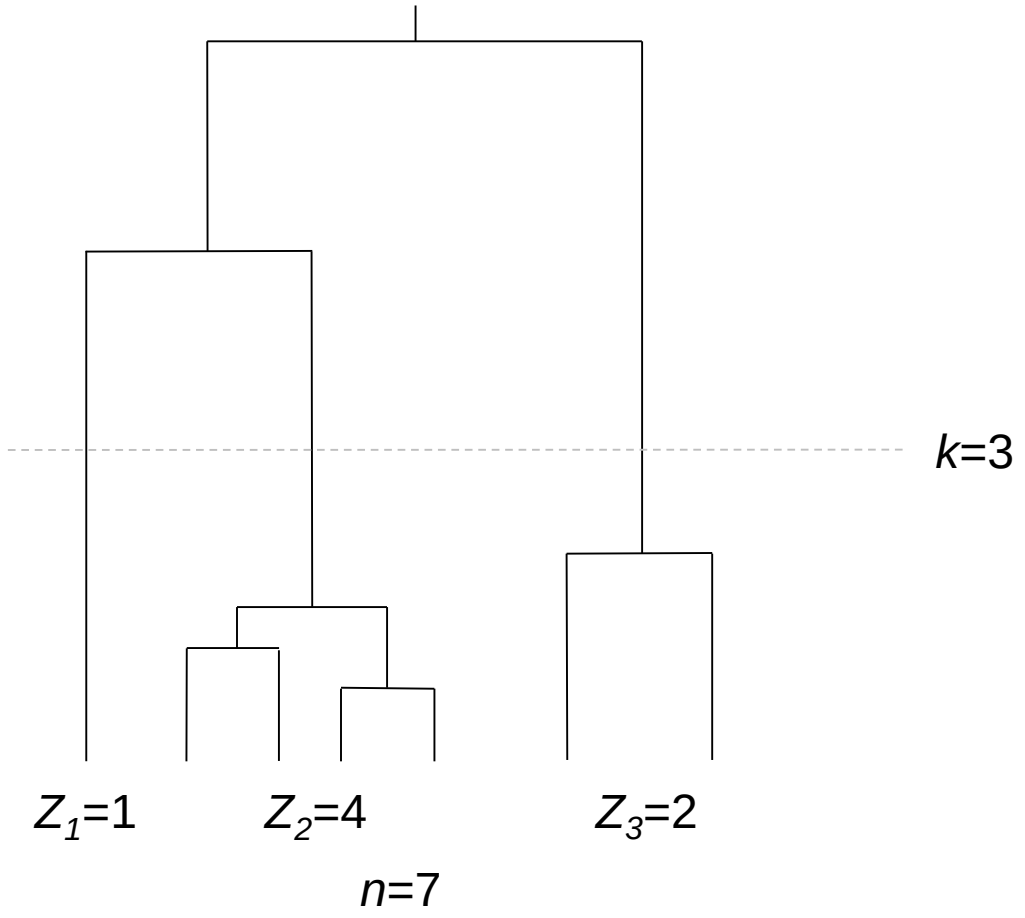
$$E(W_\infty) = 2M \text{ generations}$$

$$\approx 40{,}000 \approx 900{,}000 \text{ years for humans}$$

The “shape” of the coalescent

We could draw some tree shapes and ask “which is more likely”?

We need to be a bit more precise about what we mean. To do this, consider the number of descendants $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_k)$ of each of the lineages when *k* ancestors remain



What is the distribution of $\boldsymbol{Z}$? ANSWER: It is uniform on the possibilities

The “shape” of the coalescent

Proposition 1.4

Suppose we have a sample of size *n* sequences, and that at some time back in the past, there are *k* sample ancestors. Then the number of descendants $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_k)$ of each lineage has a uniform distribution on partitions of *n* :

$$\text{For } z_1 + z_2 + \ldots + z_k = n,\ z_i \geq 1 \text{ for all } i$$

$$P(z_1, z_2, \ldots, z_k) = \binom{n-1}{k-1}^{-1} \qquad (1.4.1)$$

Proof

We will use (backward) induction. The result is trivial for *k*=*n.* For the induction, suppose it is true for *k*≥*m* say. We only need to prove the hypothesis for *k*=*m*-1.

(If so, then the hypothesis is obviously true for *k*=2,....,*n*)

The "shape" of the coalescent

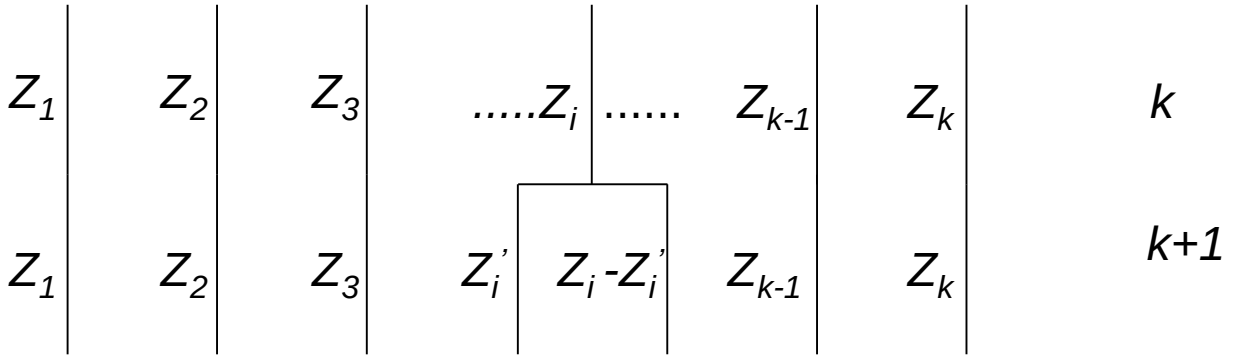
Proposition 1.4 proof ctd:

Clearly, when we have *k* lineages,

(i) some pair of lineages coalesced to go from *k*+1 to *k* lineages

(ii) each of the *k* current lineages, say *i,* has probability 1/*k* of being the one that coalesced last (i.e. branches next)

(iii) Conditional on *i* branching next, we can write down a condition on the number of descendants of the *k*+1 lineages:



(iv) By the induction hypothesis, the probability of any configuration while *k*+1 ancestors remain is known, so using (iii):

$$P_k\left(z_1, z_2, \ldots, z_k \middle| i \text{ branches}\right)$$

$$= \sum_{z_i'=1}^{z_i-1} P_{k+1}\left[z_1, z_2, \ldots, z_i', z_i - z_i', , \ldots, z_k\right]$$

$$= \sum_{z_i'=1}^{z_i-1} \binom{n-1}{k+1-1}^{-1} = (z_i - 1)\binom{n-1}{k}^{-1}$$

(v) Last, we need to sum over *i* according to (ii):

The "shape" of the coalescent

Proposition 1.4 proof ctd:

$$
\begin{aligned}
& P_k(z_1, z_2, \ldots, z_k) \\
&= \sum_{i=1}^{k} P_k(z_1, z_2, \ldots, z_k \,|\, i \text{ branches}) P(i \text{ branches}) \\
&= \sum_{i=1}^{k} P_k(z_1, z_2, \ldots, z_k \,|\, i \text{ branches}) \times \frac{1}{k} \\
&= \sum_{i=1}^{k} (z_i - 1) \binom{n-1}{k}^{-1} \times \frac{1}{k} \\
&= \binom{n-1}{k}^{-1} \frac{(n-k)}{k}, \text{ since } \sum_{i=1}^{k} z_i = n \\
&= \binom{n-1}{k-1}^{-1}, \text{ completing the proof}
\end{aligned}
$$

Example: how many copies of a mutation that occurs when k=2 are present in the sample?

The same as the number of descendants of one of the lineages when k=2 (Sheet 1, Question 5)

Conclusion: coalescent trees are not very symmetric

The "shape" of the coalescent

Note: We didn't use times in the last proof – so Proposition 1.4 holds more generally, **for *any* binary tree with random coalescence**.

Corollary: Suppose there are k lineages and let Z be the number of descendants of one particular lineage. What is the distribution of Z?

Answer: This is just the marginal distribution of Z_1 in the previous proof. If $Z_1=z$, then the number of descendants of lineages 2,....,k must form a partition of n-z, into k-1 boxes. Thus

$$P_k(Z=z) = P_k(Z_1=z)$$

$$= \binom{n-1}{k-1}^{-1} \times (\#\text{ of partitions of } n-z \text{ into } k\text{-}1)$$

$$= \frac{\binom{n-z-1}{k-2}}{\binom{n-1}{k-1}}$$

Simulating the coalescent

- We'd like to apply all this theory to the real world!

- In practice, we can usually only learn about history by looking at patterns of mutation in data

- One thing we'd like to be able to do is simulate the coalescent to see if patterns "match up" with expectations
  - If so, happy. If not, refine model or infer new model parameters
  - Several free programs do this, e.g. makesamples, "ms" (R.R. Hudson)

- How can we simulate the coalescent?
  - We must simulate exponential times T_n, T_{n-1},..., T_2 between coalescence events (and record these)
  - Then, at each coalescence event we must sample a random pair, record the answer, remove the original pair and replace with a new label to mark the coalesced pair
  - Problem 4 on the sheet

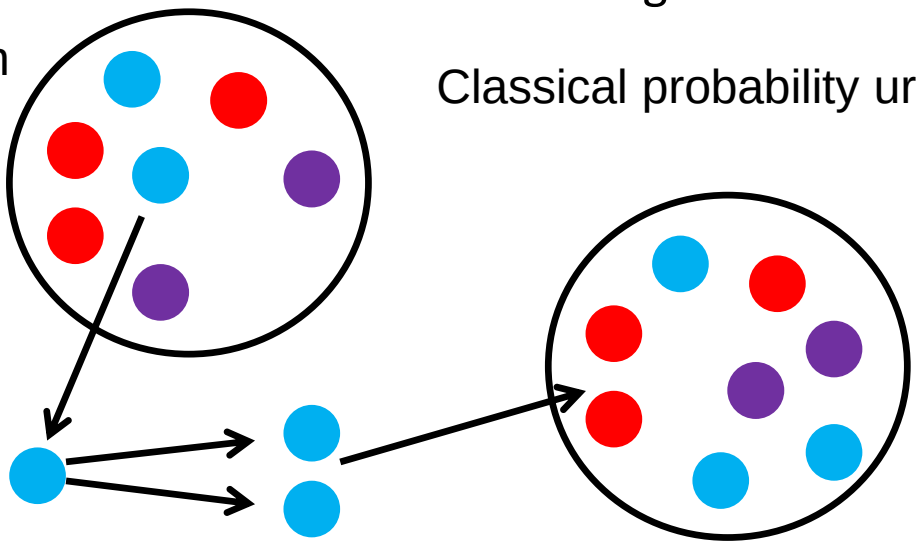
Urn models (supplementary!)

Instead of the whole coalescent tree, suppose we only wish to simulate a sample from the number of descendants of *k* lineages in a sample of size *n*

- Urn models are a classical tool in probability theory
- Also offer efficient simulation frameworks in genetics

Classical Grecian urn

Classical probability urn

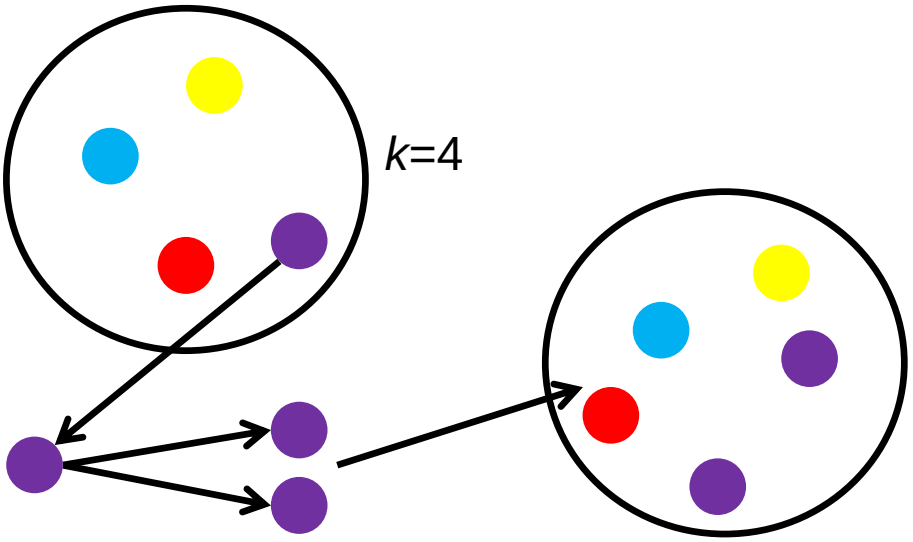


- In general, they are probability distributions on sets of coloured balls, sitting in an urn
- We remove balls, and add balls, according to specified rules
  - This makes simulation trivial
- Balls often represent other things (e.g., lineages)

- There's a nice urn model representation of the distribution of descendants of *k* lineages
- Uses the only thing we needed for our induction proof – when there are *k* lineages, each is equally likely to branch forward in time to give *k*+1 lineages. Gives an algorithm:

Urn model representation

1. Begin with *k* balls in an urn, of different colours
2. Take out a ball at random from those in the urn
3. Replace this ball with two of the same colour
4. Repeat 2 and 3 until there are *n* balls – then stop



- k initial balls represent lineages, and balls sampled represent lineages that branch forward in time

- Viewing it this way, our uniform distribution on partitions:

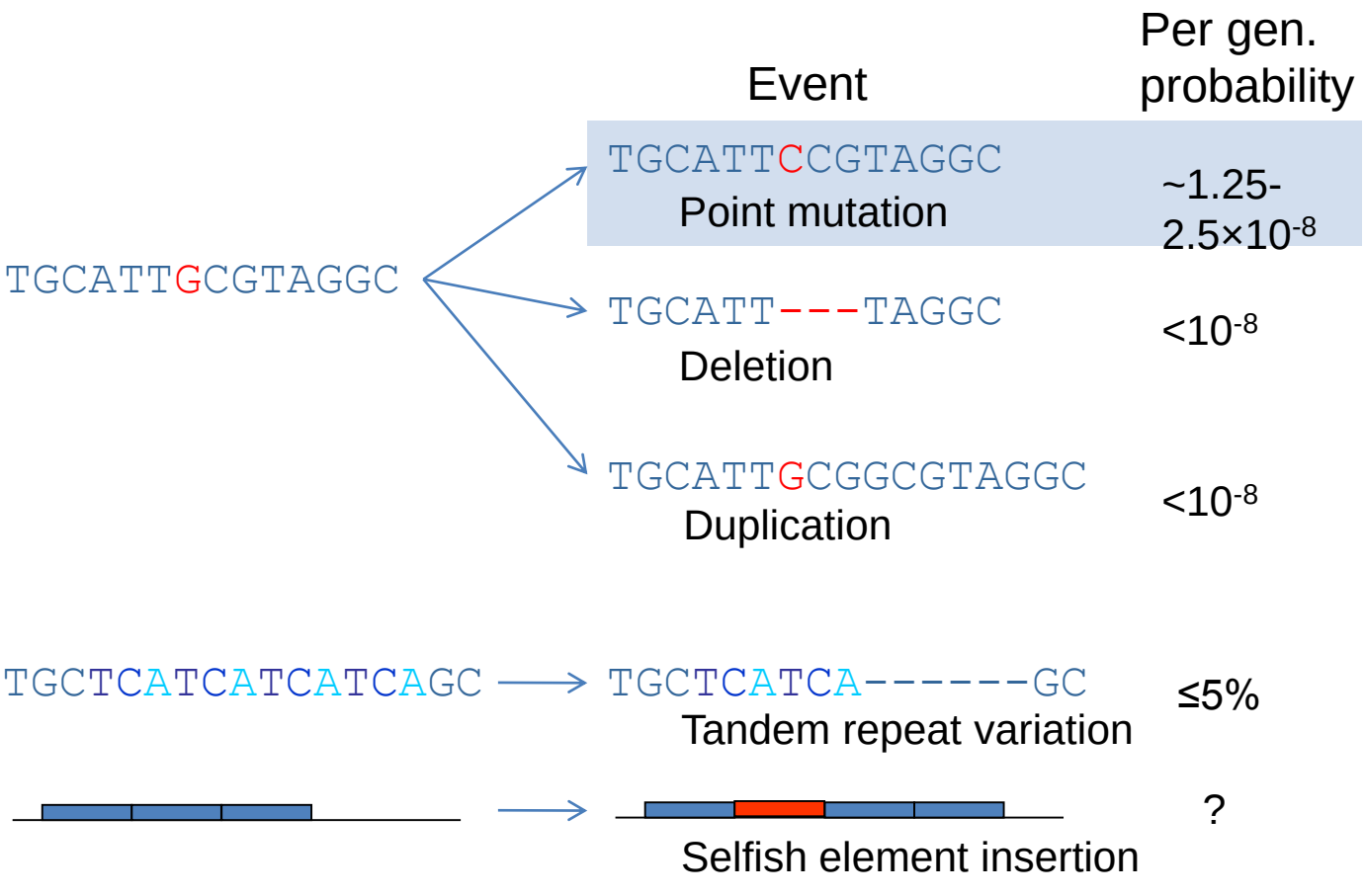
$$\text{for } z_1 + z_2 + \ldots + z_k = n,\ z_i \geq 1 \text{ for all } i$$

$$P(z_1, z_2, \ldots, z_k) = \binom{n-1}{k-1}^{-1} \qquad (1.4.1)$$

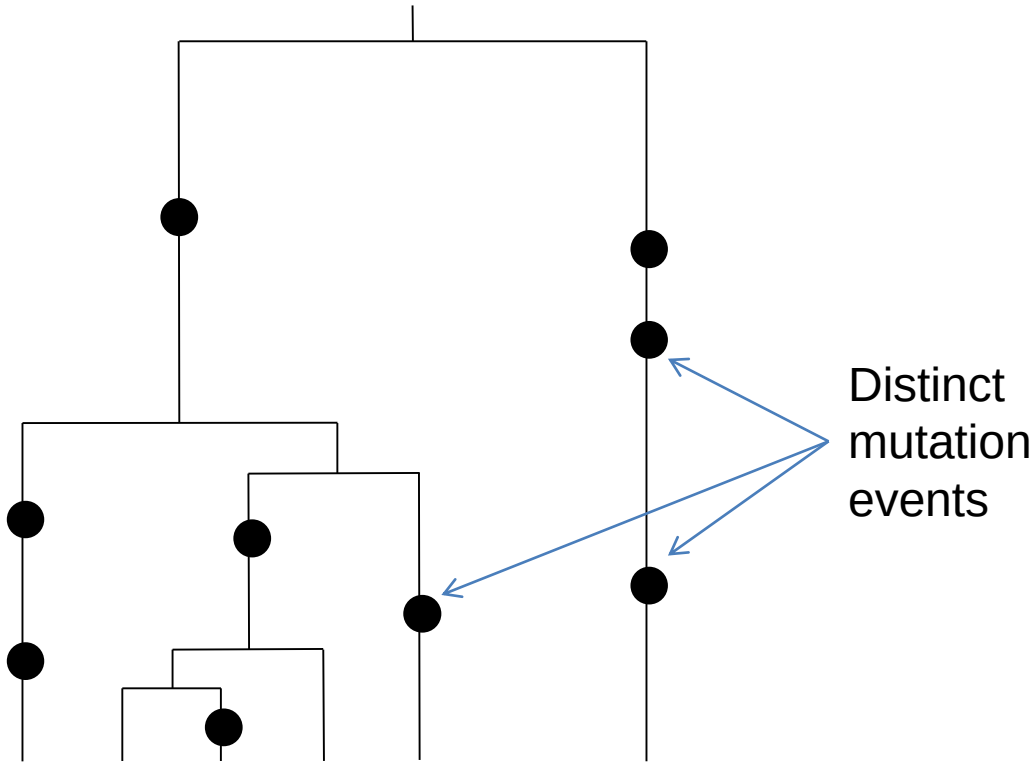
is just a classical result in probability theory.

2.0 Mutation

- In practice, we can only really learn about population history using mutation patterns
  - We can't just look – slow pace of change in populations
  - In any case, histories of interest are usually...historical
  - We have to infer what happened by looking at the patterns of mutations in samples from the population
  - This will be the subject of much of the rest of the course
- We won't need to know much about the details in some cases, but it helps to have an idea
- What is mutation? DNA can change in a variety of ways:



Mutation in the coalescent

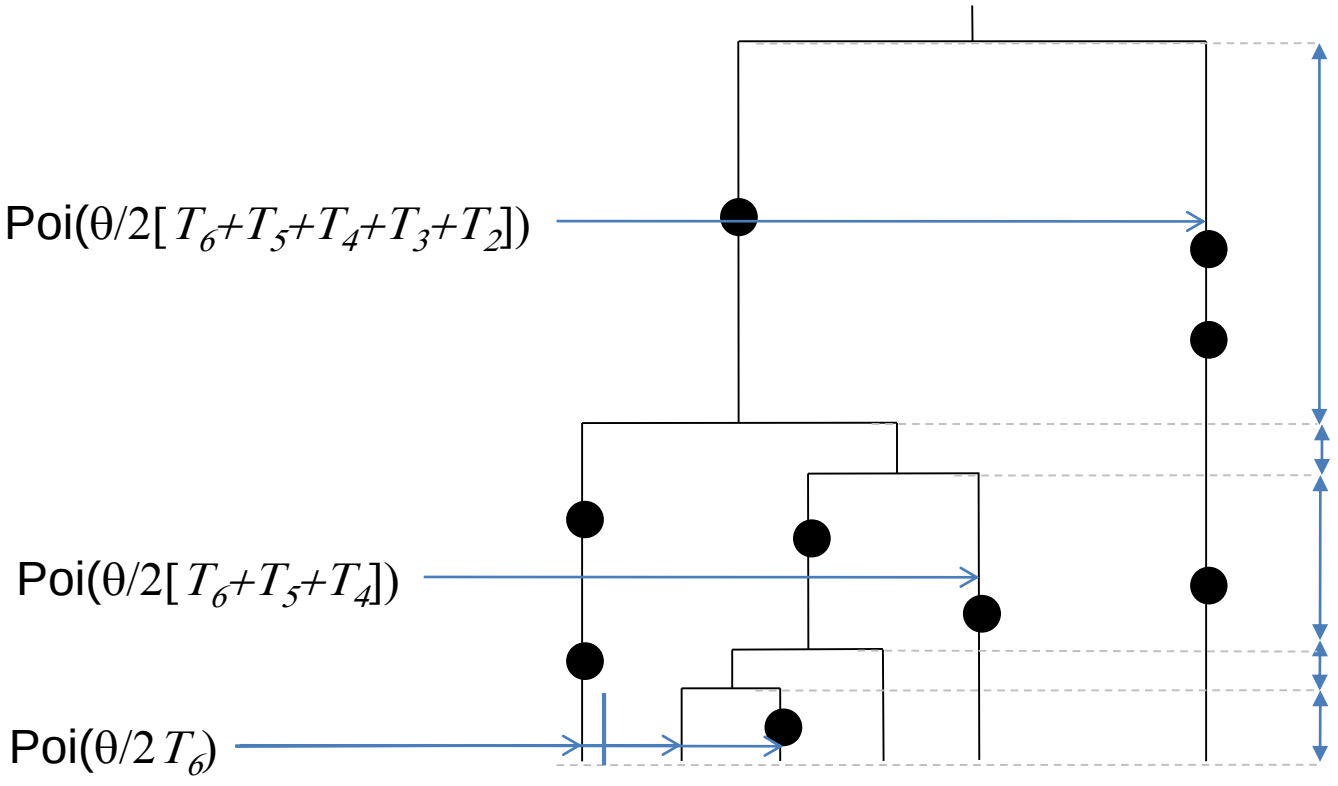


• In order to develop a model for genetic variation, we need to include mutation
• We extend the coalescent to allow mutation
• Recall edges represent ancestral lineages back in time
• So: mutations in ancestors can be represented on the edges (as circles)
• The descendants of a mutant edge inherit that mutation (unless another mutation reverses it)

• Assume that with constant probability μ per generation, there is a mutation (e.g. $\mu = 2.5\times10^{-8}$).
• In coalescent time, there are $M\,\mu$ mutations per unit time
• We model this by taking a parameter $\theta = 2M\,\mu$

• Mutations happen (in the limit as $M\to\infty$) continuously along edges, according to a Poisson process of rate $\theta/2$ on each edge

Mutation in the coalescent

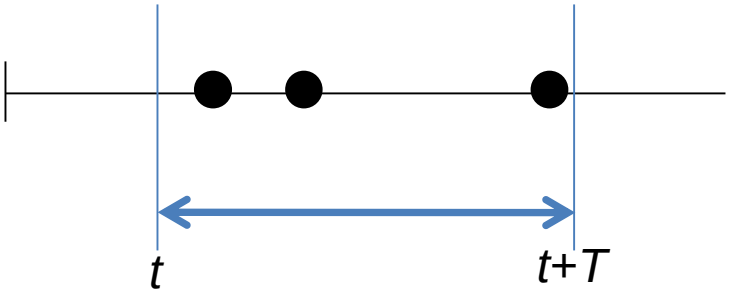


??? Poisson process???

• Let us refresh our memories, with a simple characterisation of a Poisson process

Definition 2.1 Poisson process.

A Poisson process $N(t)$ of rate λ is a continuous time process counting events in time, such that the number of events $E(T)=N(T+t)-N(t)$ in any time interval $[t,t+T)$ of length T has a Poisson distribution with mean λT, independently of all other time intervals.



$E(T)$~Poi(λT)

$N(t)=0$, $N(t+T)=3$
$E(T)=3$

The number of mutations

Sums of independent Poisson random variables are Poisson

Define $S = M_n + M_{n-1} + \ldots + M_2$ to be the total number of mutations for a sample of size n, where M_j is the number of mutations while j ancestors.

Proposition 2.2

S has probability generating function

$$f_n(z) = \prod_{j=1}^{n-1} \left(1 - \frac{(z-1)\theta}{j}\right)^{-1}$$

Further, for each j, M_j is independent, and has a geometric distribution with parameter $p_j = (j-1)/(\theta + j - 1)$.

Proof.

By properties of Poisson processes, since total edge length $T = nT_n + (n-1)T_{n-1} + \ldots + 2T_2$,

given $T_n, T_{n-1}, \ldots, T_2$ we have

$$S \sim Pois\left(\theta / 2\left[nT_n + (n-1)T_{n-1} + \ldots + 2T_2\right]\right)$$

The number of mutations

Proposition 2.2 proof ctd.

Thus

$$f_n(z) = E(z^S) = E_T\left(E\left(z^S \middle| T\right)\right)$$

$$= E\left(\exp\left[(z-1)\theta T/2\right]\right) \text{ from the Poisson p.g.f.}$$

Recall that the independ. $T_j \sim \exp(j(j-1)/2)$

$$f_n(z) = E\left(\exp\left[(z-1)\theta\left(nT_n + (n-1)T_{n-1} + \ldots + 2T_2\right)/2\right]\right)$$

$$= E\left(\prod_{j=2}^{n} \exp\left[(z-1)\theta j T_j/2\right]\right)$$

$$= \prod_{j=2}^{n} E\left(\exp\left[(z-1)\theta j T_j/2\right]\right) \text{(independ.)}$$

$$= \prod_{j=2}^{n} M_{T_j}\left((z-1)\theta j/2\right) \text{ (m.g.f of } T_j\text{)}$$

$$= \prod_{j=2}^{n}\left(1 - \frac{(z-1)\theta j}{2j(j-1)/2}\right)^{-1} \quad (\text{since } T_j \sim \exp(2/j(j-1)))$$

$$= \prod_{j=2}^{n}\left(1 - \frac{(z-1)\theta}{j-1}\right)^{-1} \quad \text{the required result}$$

Further the *jth* term in this product corresponds to mutations in time T_j, so gives the p.g.f of M_j.

The number of mutations

Proposition 2.2 proof ctd.

We are essentially done, as this is the p.g.f of a geometric random variable. Indeed, setting

$$P\left(M_j = k\right) = \left(\frac{\theta}{\theta + j - 1}\right)^k \left(\frac{j-1}{\theta + j - 1}\right), k = 0,1,\ldots$$

$$Q_j(z) = E\left(z^{M_j}\right) = \sum_{k=0}^{\infty} z^k \left(\frac{\theta}{\theta + j - 1}\right)^k \left(\frac{j-1}{\theta + j - 1}\right)$$

$$= \left(\frac{j-1}{\theta + j - 1}\right)\left(1 - \frac{z\theta}{\theta + j - 1}\right)^{-1}$$

$$= \left(1 - \frac{(z-1)\theta}{j-1}\right)^{-1}$$

It is worth pointing out an interpretation here

- Mutations happen independently for each epoch, while j ancestors
- Consider mutations, or coalescences, "events"
- While j ancestors remain, the probability the next event is a mutation is $\theta/(\theta+j-1)$.
- Otherwise, it is a coalescence with probability $p_j = (j-1)/(\theta+j-1)$,and we move to a state with $j-1$ ancestors

The mean/variance of mutation counts

Using the p.g.f of M_j

$$Q_j'(1) = \frac{\theta}{j-1}, \quad Q_j''(1) = 2\left(\frac{\theta}{j-1}\right)^2 \text{ so}$$

$$E(M_j) = Q_j'(1), \mathrm{Var}(M_j) = Q_j''(1) - Q_j'(1)^2 + Q_j'(1)$$

$$E(M_j) = \frac{\theta}{j-1}, \quad \mathrm{Var}(M_j) = \left(\frac{\theta}{j-1}\right)^2 + \frac{\theta}{j-1}$$

This immediately gives expectation and variance for the total number of mutations *S*:

$$E(S) = \sum_{j=2}^{n} \frac{\theta}{j-1} = \theta \sum_{j=1}^{n-1} \frac{1}{j}$$

$$\mathrm{Var}(S) = \sum_{j=2}^{n} \left(\frac{\theta}{j-1}\right)^2 + \frac{\theta}{j-1} = \theta \sum_{j=1}^{n-1} \frac{1}{j} + \theta^2 \sum_{j=1}^{n-1} \frac{1}{j^2}$$

As $n \to \infty$

$$E(S) \sim \theta \log n, \qquad \mathrm{Var}(S) \sim \theta \log n$$

This motivates *Watterson's estimator* (1975) of θ:

$$\hat{\theta} = \frac{S}{\sum_{j=1}^{n-1} \frac{1}{j}}$$

This *moment estimator* is unbiased, and consistent as $n \to \infty$

Example: Estimation of population size

- As we have discussed, the coalescent is a limit under very general population assumptions
- Time in units of *M* generations, where *M* is the "effective population size"
- Estimation of *M* allows us to calibrate into years, to understand time depth.
- I have given *M* estimates for humans
- **The data only give information directly on $\theta = 2M\mu$**
- To infer *M* or μ, we must know (or assume) the other. This idea is how *M* is generally estimated

Example 1 (Zhao et al., PNAS, 2000): In sequence data for 128 human chromosomes sampled worldwide, 75 variant sites were identified. If the mutation rate per DNA base per generation is 2.3×10^{-8}, and 9,901 bases were sequenced, estimate the human effective population size

Example 1 (Zhao et al., PNAS, 2000): In sequence data for 128 human chromosomes sampled worldwide, 75 variant sites were identified. If the mutation rate per DNA base per generation is 2.3×10^{-8}, and 9,901 bases were examined, estimate the human effective population size

Solution:

Watterson's estimator:

$$\hat{\theta} = \frac{S}{\sum_{j=1}^{n-1} \frac{1}{j}} = \frac{75}{\sum_{j=1}^{127} \frac{1}{j}} = 13.82$$

Because we are given μ, we can estimate M:

$$\theta = 2M\mu = 2\times9901\times2.3\times10^{-8}\times M = 4.55\times10^{-4}\times M$$

$$\hat{M} = 2195.7\times\hat{\theta} = 30{,}353$$

This is a fairly typical value for a worldwide human sample. Finally...how does one get μ? Two ways:

- Chimpanzee genome comparisons
- Direct measurement in families

Note: Watterson's estimator does *not* use all the information in data for θ.

Example: Time conditional on number of mutations

- Genealogy depth is stochastically variable
- Longer trees have more mutations (segregating sites) on average
- The distribution of tree depth is altered given the number of segregating sites seen in data

Example 2 (Dorit et al., Science, 1995): In sequence data for 38 human Y chromosomes sampled worldwide, no variant sites were identified at the ZFY locus. If the mutation rate at this gene per generation is 1.96×10^{-5}, and generations last 20 years, derive an equation for the expected TMRCA conditional on this data and a population size N

Solution

If there is no variation in the sample, this means there are no mutation events in the coalescent history of the sample:

$$M_n = M_{n-1} = \ldots = M_2 = 0$$

We can consider the times T_j while j ancestors remain. Recalling that over time T_j, the number of mutations on each of the j edges is independently Poisson with mean $\theta T_j/2$, we have:

$$P(M_j = 0 \mid T_j = t_j) = e^{-j\theta t_j/2} \text{ and so}$$

$$f_j(t_j \mid M_j = 0) = \frac{P(M_j = 0 \mid T_j = t_j) f_j(t_j)}{P(M_j = 0)}$$

$$\propto e^{-j\theta t_j/2} \binom{j}{2} e^{-j(j-1)t_j/2} \text{ using the pdf of } T_j$$

$$\propto \exp\left[-\frac{j(j-1+\theta)}{2} t_j \right] \text{(up to a constant)}$$

Thus **conditional on M_j=0, T_j has the exponential distribution with a (reduced) mean**

$$E(T_j \mid M_j = 0) = \frac{2}{j(j-1+\theta)}$$

Solution continued

Finally, we can give the expected TMRCA (in years) conditional on no mutations:

$$W_n = T_n + T_{n-1} + \ldots + T_2$$

$$E(W_n \mid S = 0) = \sum_{j=2}^{n} E(T_j \mid M_j = 0)$$

$$= \sum_{j=2}^{n} \frac{2}{j\left(j-1+\theta\right)} \text{(in } N \text{ gens)}$$

$$= 20N \sum_{j=2}^{38} \frac{2}{j\left(j-1+2N\times 1.96\times 10^{-5}\right)} \text{(in years)}$$

Tabulating, we see our knowledge of no mutations (in 729 bp) does not have a huge effect:

Population size *N*	Mean TMRCA given no variation	Mean TMRCA unconditionally
2,500	92,000	97,000
5,000	173,000	195,000
10,000	313,000	389,000

Solution continued

Dorit *et al.* made an error, writing:

$$P(S=0 \mid W_n)$$

$$= \prod_{j=2}^{38} \frac{j-1}{j-1+1.96\times 10^{-5}\times W_n} \text{ (in generations)}$$

This led to strange conclusions – 95% CI of (0,800,000 years) and estimate of 270,000 years

This in turn led to a number of rapid critical responses, e.g. “Estimating the age of the common ancestor of men from the ZFY locus” Donnelly, Tavare, Balding and Griffiths, *Science* 1996

These data are actually compatible with a very wide range of times.

Humans are not very variable – on average 1 mutation every 1,000bp between 2 human chromosomes.

Supplement: distribution of number of mutations

- We derived the p.g.f of the total number *S* of mutations. What is the full distribution of *S*?
- We can apply the Gamma function property that for real *z*, $\Gamma(z+1)=z\Gamma(z)$:

$$f_n(z)=\sum_{j=1}^{\infty} z^j P(S=j)$$

$$\prod_{j=1}^{n-1}\left(1-\frac{(z-1)\theta}{j}\right)^{-1}=\prod_{j=1}^{n-1}\left(\frac{j}{j+(1-z)\theta}\right)$$

$$=(n-1)\frac{(n-2)!\Gamma(1+(1-z)\theta)}{\Gamma(n+(1-z)\theta)}$$

$$=(n-1)\beta(n-1,1+(1-z)\theta) \quad \text{defn of "}\beta-\text{fun."}$$

$$=(n-1)\int_0^1 x^{n-2}(1-x)^{(1-z)\theta}\,dx \quad (\text{properties of } \beta)$$

$$=(n-1)\int_0^1 x^{n-2}(1-x)^{\theta}\exp\left[-z\theta\ln(1-x)\right]dx$$

$$=\sum_{j=0}^{\infty}\frac{z^j(n-1)}{j!}\int_0^1 x^{n-2}(1-x)^{\theta}\left[-\theta\ln(1-x)\right]^j dx$$

so if $\lambda(x)=-\theta\ln(1-x)$ and $f(x)=(n-1)x^{n-2}, 0<x<1,$

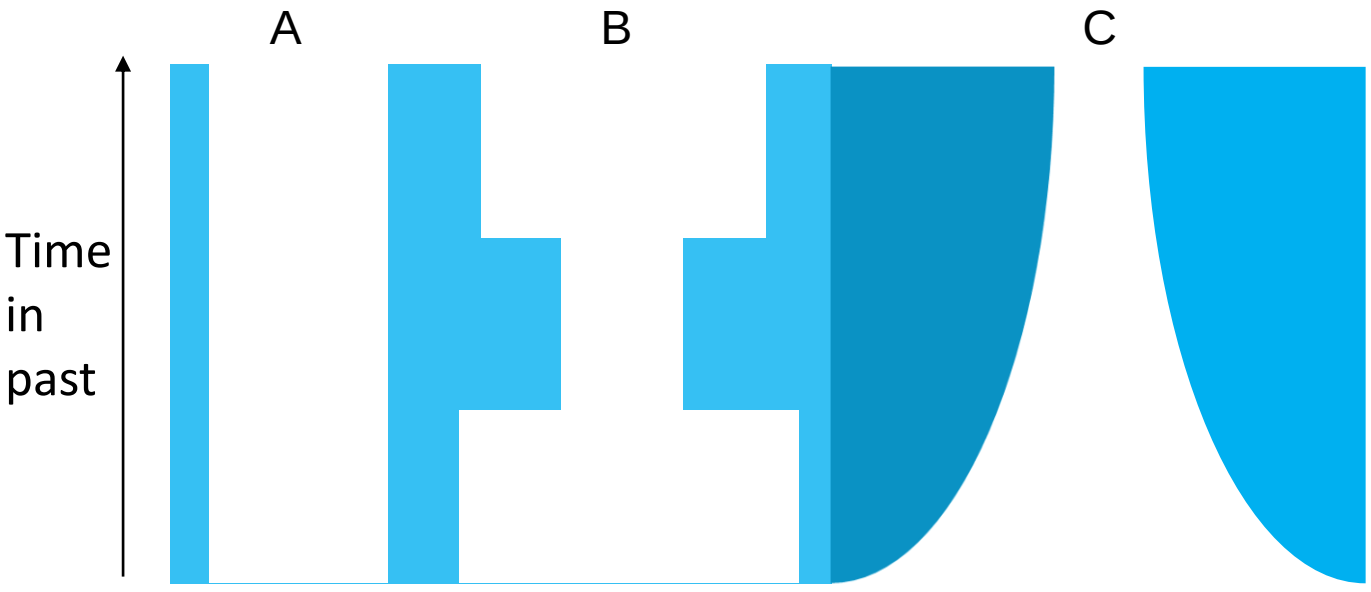
$$P(S=j)=\int_0^1 f(x)\frac{e^{-\lambda(x)}\left[\lambda(x)\right]^j}{j!}\,dx$$

Variable size populations

- Real populations don't have constant size. Suppose the population size a time t in the past is $N(t)=N(0)\nu(t)$
- We need a "clock" – as before, measure time t in units of $N(0)$ generations, $N(0)$ now present day size
- We will extend the coalescent to this setting
- Recall that while j ancestors remain in a Wright-Fisher model, the probability, i.e. "rate" at which coalescence occurs is $j(j\text{-}1)/2M$ *per generation*
- In the new setting, the new per generation coalescence rate is

$$\frac{j(j-1)}{2N(t)} = \frac{j(j-1)}{2\nu(t)} \times \frac{1}{N(0)}$$

- Measuring time in units of $N(0)$ generations, while j ancestors, coalescence occurs at rate $\frac{j(j-1)}{2\nu(t)}$



Variable size populations

Definition 2.3
The *coalescent with variable population size* is a distribution on binary trees. Starting with *n* lineages, randomly chosen pairs of lineages coalesce backward in time until a single common ancestor is reached. Suppose the relative population size at time *t* in the past is *v*(*t*). While *j* edges remain at time *t*, coalescence events occur with instantaneous rate *j*(*j*-1)/2*v*(*t*). Equivalently, defining times T_n, T_{n-1},...,T_2 while *n*,*n-1* ,..., 2 ancestors remain:

$$P\left(T_j > t \middle| T_n + T_{n-1} + \ldots + T_{j+1} = s\right) = \exp\left[-\binom{j}{2} \int_s^{s+t} \frac{1}{v(u)}\, du \right] \quad (2.3.1)$$

Comments

1. The standard coalescent case is *v*(*t*)≡1

2. Equation 2.3.1 can be derived directly as the Wright-Fisher limit

3. Intuitively, in (2.3.1), if there are *j* lineages from time *s* to time *t*+*s*, the coalescence rate changes from *j*(*j*-1)/2*v*(*s*) to *j*(*j*-1)/2*v*(s+*t*)

   This is the reason for the integral term, which "averages out" the coalescence rate

Variable size populations

How do we, e.g., simulate the coalescent with variable size?

The answer: we can use a *coupling* of times with the standard coalescent case.

Idea: we transform time into new units. Define

$S_j = T_n + T_{n-1} + \ldots + T_j,\ S_{n+1} = 0$ (coalescence times)

Proposition 2.3
In the variable population size coalescent with relative population size $v(t)$ at time t in the past, if time is rescaled by setting

$$t' = \int_0^t \frac{1}{\nu(u)}\, du$$

then the transformed times S_n', S_{n-1}',...,S_2' at which coalescence events occur are distributed according to the standard coalescent with constant size population

Comments

1. Note that transformed time increases more quickly when the population size is small
2. We invert the transformation to give each S_j
3. To recover coalescence times, we take differences: $T_j = S_j - S_{j+1}$

Variable size populations

Proposition 2.3
In the variable population size coalescent with relative population size $v(t)$ at time t in the past, if time is rescaled by setting

$$t' = \int_0^t \frac{1}{\nu(u)}\,du$$

then the transformed times S_n', S_{n-1}',...,S_2' at which coalescence events occur are distributed according to the standard coalescent with constant size population

Proof
Define the untransformed coalescence times S_n, S_{n-1},...,S_2.
Restating (2.3.1) in terms of these times, we have:

$$P\left(S_j > s+t \middle| S_{j+1} = s\right) = \exp\left[-\frac{j(j-1)}{2}\int_s^{s+t}\frac{1}{\nu(u)}\,du\right] \quad \forall s,t>0 \Leftrightarrow$$

$$P\left(S_j > s_j \middle| S_{j+1} = s_{j+1}\right) = \exp\left[-\frac{j(j-1)}{2}\int_{s_{j+1}}^{s_j}\frac{1}{\nu(u)}\,du\right] \quad \forall s_j > s_{j+1} > 0$$

and it is S.T.P

$$P\left(S_j' > s_j' \middle| S_{j+1}' = s_{j+1}'\right) = \exp\left[-\frac{j(j-1)}{2}\int_{s_{j+1}'}^{s_j'} du\right] \quad \forall s_j' > s_{j+1}' > 0$$

Now it is clear the transformation is well defined, so for every positive t'=s_j' there is a corresponding *untransformed* t=s_j.
Further the transformation is increasing so $s_j' > s_{j+1}' \Rightarrow s_j > s_{j+1}$

Variable size populations

Proposition 2.3

In the variable population size coalescent with relative population size $v(t)$ at time t in the past, if time is rescaled by setting

$$t' = \int_0^t \frac{1}{v(u)}\,du$$

then the transformed times S_n', S_{n-1}',...,S_2' at which coalescence events occur are distributed according to the standard coalescent with constant size population

Proof

Using this reverse transformation, for any $s_j' > s_{j+1}' > 0$:

$$P\left(S_j' > s_j' \middle| S_{j+1}' = s_{j+1}'\right) = P\left(S_j > s_j \middle| S_{j+1} = s_{j+1}\right)$$

$$= \exp\left[-\frac{j(j-1)}{2}\int_{s_{j+1}}^{s_j}\frac{1}{v(u)}\,du\right]$$

$$= \exp\left[-\frac{j(j-1)}{2}\left(\int_0^{s_j}\frac{1}{v(u)}\,du - \int_0^{s_{j+1}}\frac{1}{v(u)}\,du\right)\right]$$

$$= \exp\left[-\frac{j(j-1)}{2}\left(s_j' - s_{j+1}'\right)\right] = \exp\left[-\frac{j(j-1)}{2}\int_{s_{j+1}'}^{s_j'}du\right]$$

A key use of this idea is in simulation of histories under this model (and inference).

Simulation

• Simulation under a variable size model can be accomplished simply, by the following:

1. Simulate coalescence times S_n', S_{n-1}',...,S_2' under the neutral coalescent. Set $S_{n+1}'=0$, and then:

$$T_j' = S_j' - S_{j+1}' = -\binom{j}{2}\log(U_j), \quad j = n, n-1, \ldots, 2$$

where the U_j's are i.i.d $U(0,1)$ random variables.

2. Convert these back to *untransformed* times S_n, S_{n-1},...,S_2 using

$$S_j' = \int_0^{S_j} \frac{1}{\nu(u)}\,du$$

3. Given times, coalescence events are easy to sample, and <u>mutation event counts have the usual Poisson distribution given tree times</u> (note the mutation process in each ancestral lineage is independent of the population size).

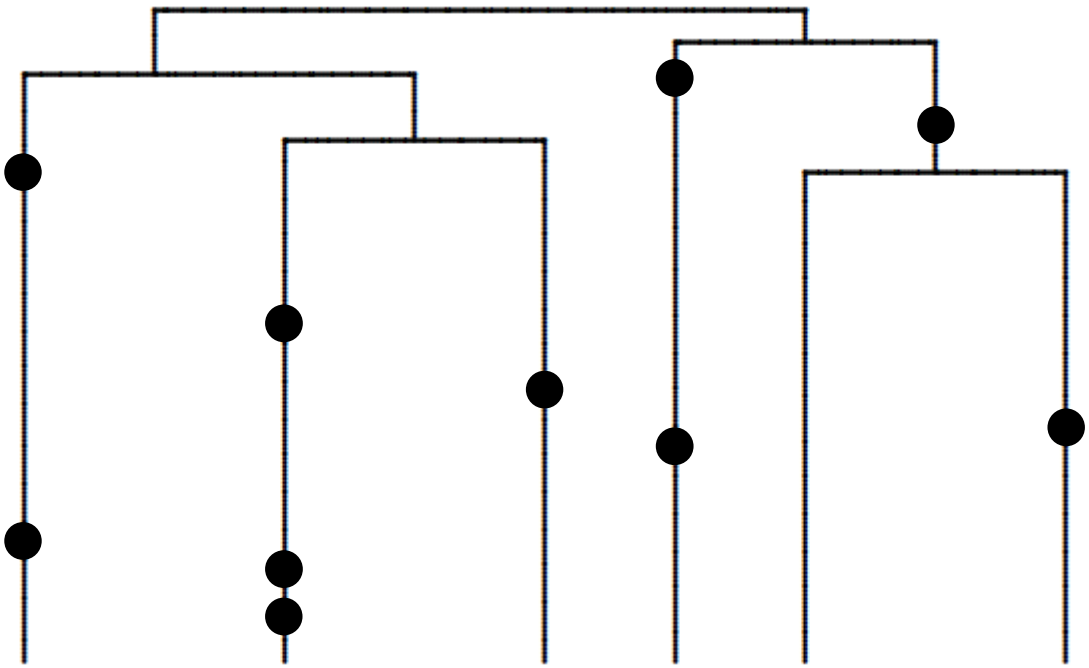
Example

In *exponential expansion* $\nu(t)=\exp(-\beta t)$, so

$$S_j' = \int_0^{S_j} e^{\beta t}\,du = \frac{1}{\beta}\left(e^{\beta S_j} - 1\right) \Rightarrow S_j = \frac{1}{\beta}\log\left(1+\beta S_j'\right)$$

(sheet 2 question 4)

“Star-like” genealogies



Exponential expansion (or expansion generally) makes times relatively shorter in the top parts of the tree

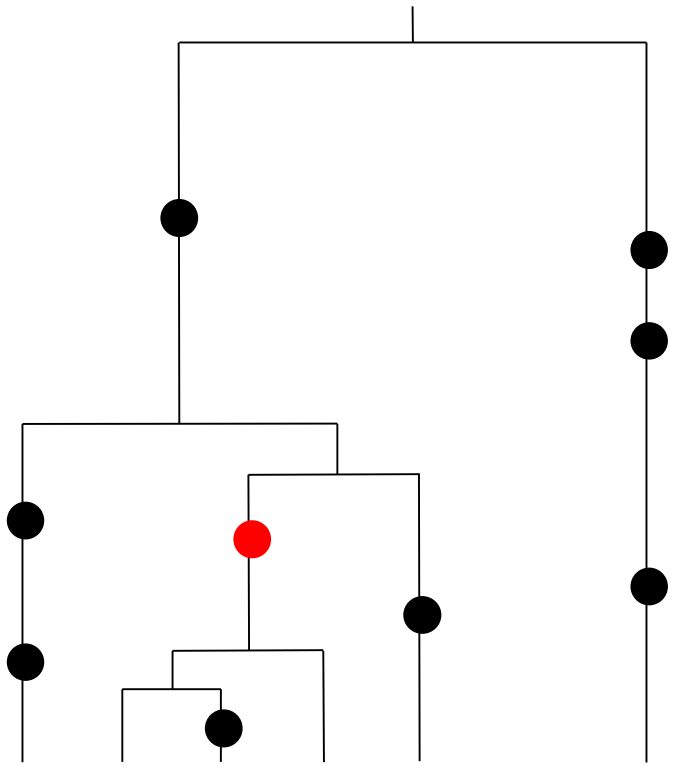
Question: What is the effect of variation in population size on genetic variation data?

This could offer us a way to learn about population sizes in the distant past

To do that, we need to think about the *frequency spectrum* of mutations

We start by thinking about single mutations

3.0 The spread of diversity



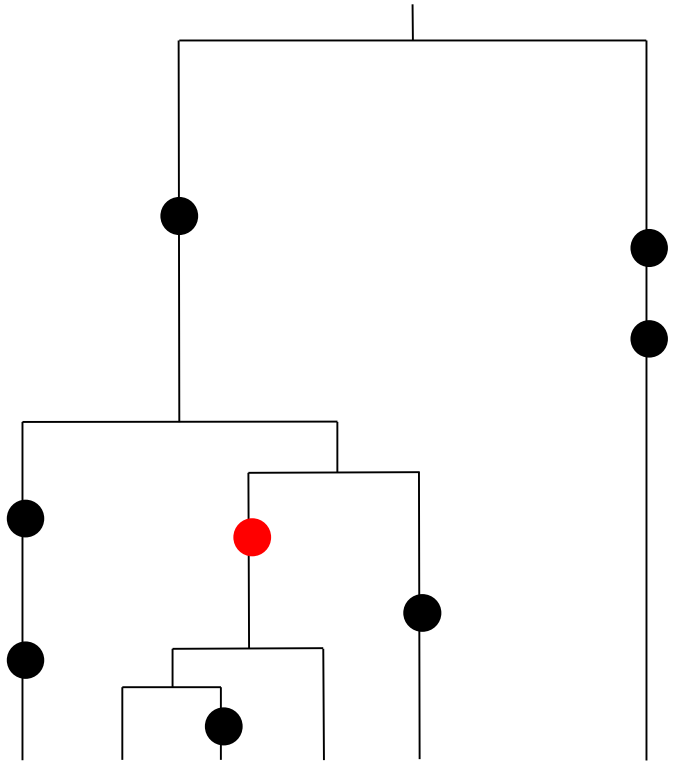
The red mutation happens while there are *k*=4 lineages remaining.

It spreads and is seen in 3 of 6 sample members – the descendants of the lineage on which it occurs

More generally, the shape of the coalescence tree (Proposition 1.4 corollary) tells us that for any mutation that occurs while *k* ancestors remain from an initial sample size of *n*, the probability of *b* descendants is in general:

$$p_{nk}(b) = \frac{\binom{n-b-1}{k-2}}{\binom{n-1}{k-1}} \qquad p_{6,4}(3) = \frac{\binom{6-3-1}{4-2}}{\binom{6-1}{4-1}} = \frac{1}{10}$$

3.0 The spread of diversity



Almost always in practice, we only see diversity patterns.

We know n=6, b=3 but we do not know the number of lineages when the mutation occurred

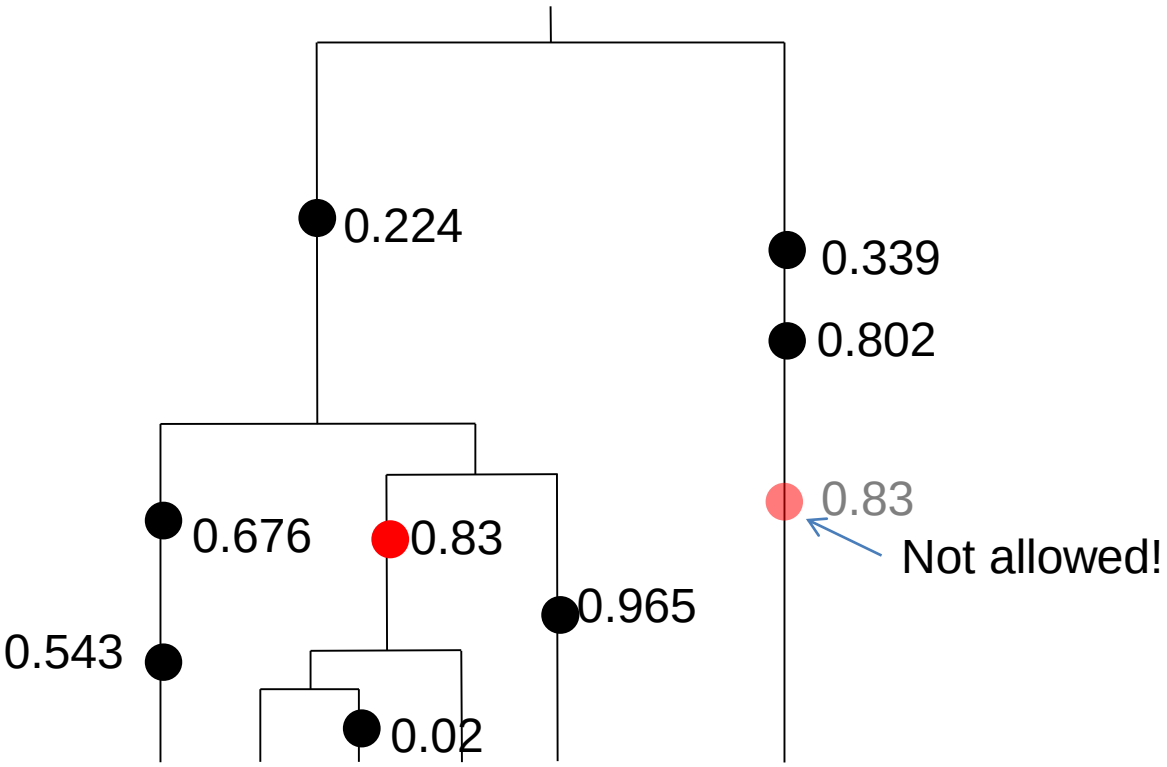
What is the *unconditional* probability, q_{nb}, for a site which varies, that we observe b mutant copies in a sample of size n? This is called the (expected) *frequency spectrum* of mutations

$$q_{nb} = \sum_{k=2}^{n} P(b \text{ mutant copies} \mid \text{while } k \text{ lineages}) \mathrm{P}(k \text{ lineages})$$

$$= \sum_{k=2}^{n} p_{nk}(b) \mathrm{P}(k \text{ lineages})$$

We can also ask, how old is a mutation seen in b of n copies?

The infinite-sites model



Strictly, we need to make an assumption here.

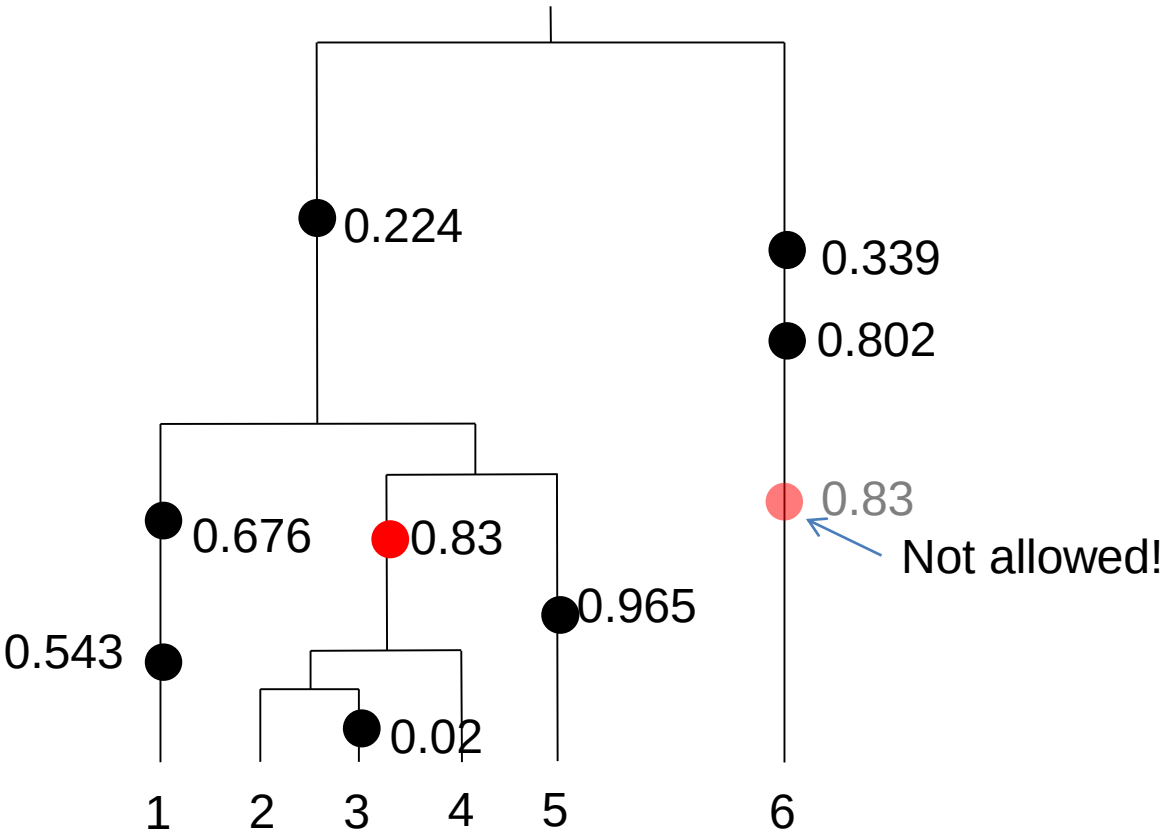
If we see a mutation in some sample members but not others, we assume it is the result of one historical event, not e.g. two identical independent mutation events in different ancestors

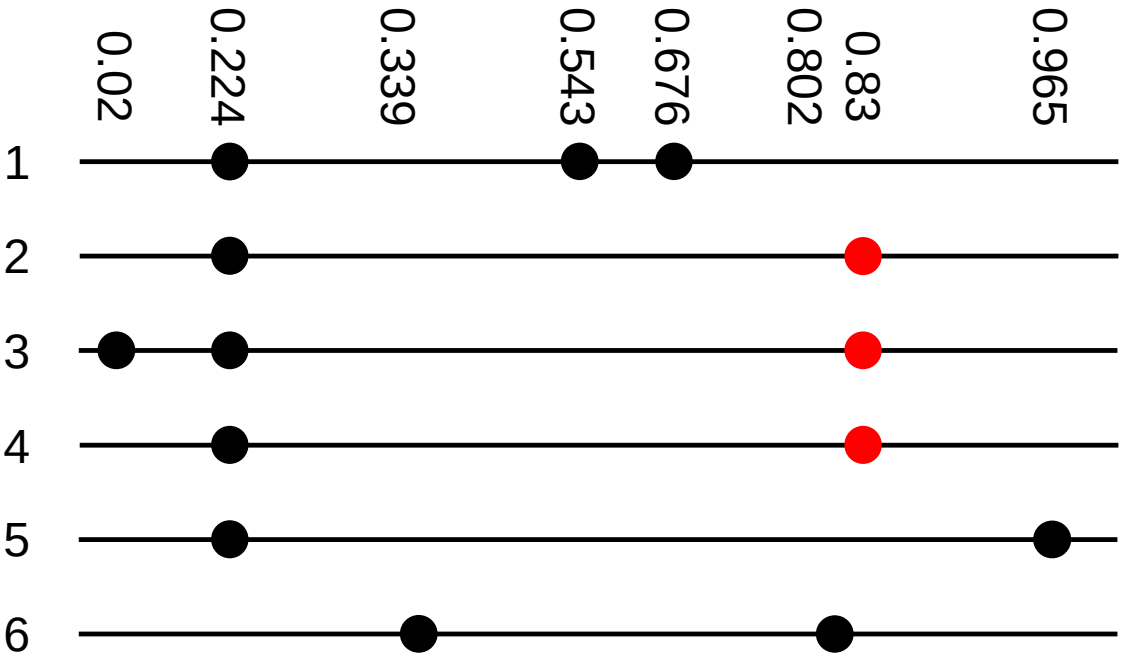
Specifically – *mutations always occur at a position never before mutant.* This is called the infinitely-many-sites model

In this model, each individual site has a vanishingly small probability of mutating (but a region has a non-zero rate)

Without loss of generality, label mutations using independent uniform random variables in [0,1] (i.e. labels always unique)

The infinite-sites model





3.0 The spread of diversity

$$q_{nb} = \sum_{k=2}^{n} P(b \text{ mutant copies} \mid \text{while } k \text{ lineages}) \mathrm{P}(k \text{ lineages})$$

$$= \sum_{k=2}^{n} p_{nk}(b) \mathrm{P}(k \text{ lineages})$$

The only thing we must work out is the probability a mutation observed in a sample occurs while *k* lineages, given only that the mutation segregates in the sample. Suppose the mutation occurs at *x* in [0,1].

We will not (for now) make any assumptions about times in the coalescent tree – so we are in the setting of the coalescent with variable population size

It helps to write the following

$$P(k \text{ lineages} \mid \text{mutation at } x) = \lim_{\delta x \to 0} P(k \text{ lineages} \mid \text{mutation in } [x, x+\delta x))$$

$$= \lim_{\delta x \to 0} \frac{P(\text{mutation in} [x, x+\delta x) \text{ while } k \text{ lineages})}{\sum_{k=2}^{n} P(\text{mutation in} [x, x+\delta x) \text{ while } k \text{ lineages})} = \lim_{\delta x \to 0} \frac{P(I_k = 1)}{\sum_{k=2}^{n} P(I_k = 1)}$$

where I_k=1 if a mutation occurs in [x,x+δx) while k ancestors

That is, we consider the probability of exactly one mutation occurring, in a region containing *x.* The number of mutations in [x,x+δx) while k ancestors is Poisson with mean $kT_k\theta\delta x/2$, so

$$P(I_k = 1) = \exp(-\theta \delta x k T_k / 2) \theta \delta x k T_k / 2$$

$$= \theta \delta x k T_k / 2 + o(\delta x)$$

3.0 The spread of diversity

$$q_{nb} = \sum_{k=2}^{n} P(b \text{ mutant copies} \mid \text{while } k \text{ lineages}) \mathrm{P}(k \text{ lineages})$$

$$= \sum_{k=2}^{n} p_{nk}(b) \mathrm{P}(k \text{ lineages})$$

We can then sum over the distribution of T_k to give unconditionally:

$$P(I_k = 1) = \theta \delta x k E(T_k)/2 + o(\delta x)$$

As $\delta x \to 0$, at most one mutation occurs, so

$$P(k \text{ lineages} \mid \text{mutation at } x)$$

$$= \lim_{\delta x \to 0} \frac{\theta \delta x k E(T_k)/2 + o(\delta x)}{\sum_{k=2}^{n} \theta \delta x k E(T_k)/2 + o(\delta x)}$$

$$= \frac{k E(T_k)}{\sum_{k=2}^{n} k E(T_k)}$$

and finally:

$$q_{nb} = \sum_{k=2}^{n} p_{nk}(b) \mathrm{P}(k \text{ lineages})$$

$$= \frac{\sum_{k=2}^{n} \frac{\binom{n-b-1}{k-2}}{\binom{n-1}{k-1}} k E(T_k)}{\sum_{k=2}^{n} k E(T_k)}, \quad 0 < b < n$$

Example: constant size population

For a constant size population, recall:

$$E(T_k) = \frac{2}{k(k-1)} \Rightarrow q_{nb} = \frac{\sum_{k=2}^{n} \frac{\binom{n-b-1}{k-2}}{\binom{n-1}{k-1}} \frac{2k}{k(k-1)}}{\sum_{k=2}^{n} \frac{2k}{k(k-1)}}, \quad 0 < b < n$$

$$= \frac{\sum_{k=2}^{n} \frac{\binom{n-b-1}{k-2}}{\binom{n-1}{k-1}} \frac{1}{(k-1)}}{\sum_{k=1}^{n-1} \frac{1}{k}} = \frac{\sum_{k=2}^{n} \binom{n-b-1}{k-2} \beta(k-1, n-k+1)}{\sum_{k=1}^{n-1} \frac{1}{k}} \text{ (exercise)}$$

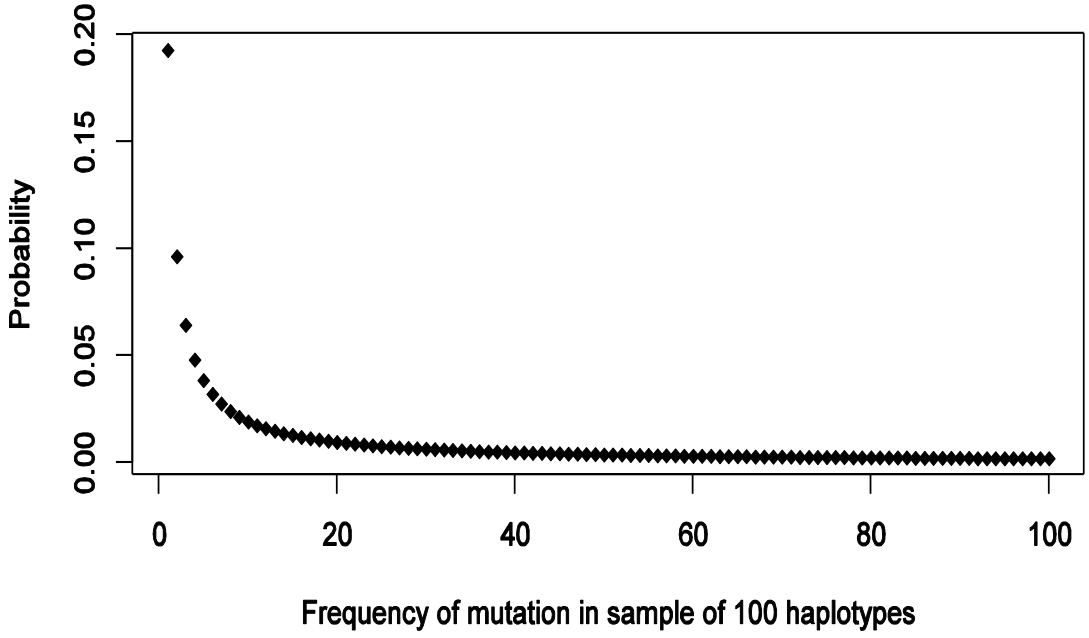
$$= \frac{\sum_{k=2}^{n} \binom{n-b-1}{k-2} \int_0^1 x^{k-2}(1-x)^{n-k}\,dx}{\sum_{k=1}^{n-1} \frac{1}{k}}$$

$$= \frac{\int_0^1 (1-x)^{b-1} \sum_{k=2}^{n} \binom{n-b-1}{k-2} x^{k-2}(1-x)^{(n-b-1)-(k-2)}\,dx}{\sum_{k=1}^{n-1} \frac{1}{k}}$$

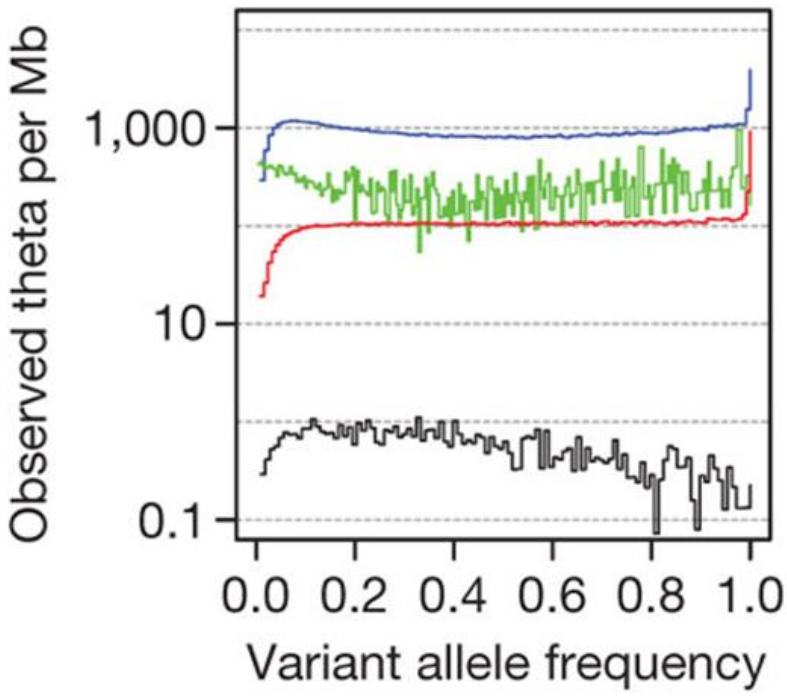
$$= \frac{\int_0^1 (1-x)^{b-1}dx}{\sum_{k=1}^{n-1} \frac{1}{k}} = \frac{b^{-1}}{\sum_{k=1}^{n-1} k^{-1}} \text{ (details of algebra are exercise)}$$

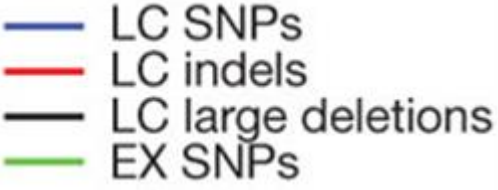
Real data vs. predictions

Our constant size model predicts there are more rare than common mutations:



What do we see for real populations? A remarkable match for a new worldwide dataset with 6.5 million mutations:





"A map of human genetic variation from population level sequencing"
Nature Oct. 2010

Errors cluster at extreme frequencies

Supplement: how old is my mutation?

• What is the expected age of a mutation if it is seen in *b* copies out of *n*?

• We use the same basic idea as before, and condition on when it occurs. Let the age be ξ_{nb}. If the mutation occurs while *k* ancestors, its age is obviously uniform across the period while *k* ancestors:

$$\xi_{nb} = UT_k + T_{k+1} + \ldots + T_n$$

where *U* is uniform on (0,1) and independent of the T_i's.

• Applying the same argument as before, we condition on when the mutation occurs and define an indicator I_k=1 if a mutation occurs in a small interval [*x*,*x*+*d*x) containing *x* while *k* ancestors

$$E(\xi_{nb}) = \lim_{\delta x \to 0} \frac{\sum_{k=2}^{n} E(\xi_{nb} \mid I_k = 1, b \text{ copies})\, p_{nk}(b) P(I_k = 1)}{\sum_{k=2}^{n} p_{nk}(b) P(I_k = 1)}$$

$$= \frac{\sum_{k=2}^{n} p_{nk}(b) k E\left(\left[T_k / 2 + T_{k+1} + \ldots + T_n\right] T_k\right)}{\sum_{k=2}^{n} p_{nk}(b) k E(T_k)}$$

We can consider the constant size case again

Example: constant size popn

$$E(\xi_{nb}) = \frac{\sum_{k=2}^{n} p_{nk}(b) k E\left(\left[T_k/2 + T_{k+1} + \ldots + T_n\right] T_k\right)}{\sum_{k=2}^{n} p_{nk}(b) k E(T_k)}$$

$$= \frac{\sum_{k=2}^{n} \frac{\binom{n-b-1}{k-2}}{\binom{n-1}{k-1}} \left(k \binom{k}{2}^{-1} \left[\frac{2}{k} - \frac{2}{n} \right] + \frac{k}{2} \left[2 \binom{k}{2}^{-2} \right] \right)}{2b^{-1}}$$

$$\ldots$$

$$= \frac{2b}{n-b} \sum_{j=b+1}^{n} j^{-1}$$

The algebra missed out is tedious – it relies on certain combinatoric identities. For more (but not quite full) details, see RCG's notes, linked to on the webpage

The age of a mutation at frequency *x* in the entire population

We just set $b = \lfloor nx \rfloor$ and let $n \to \infty$, so $b/n \to x$ and

$$E(\xi_x) = \lim_{n\to\infty} \frac{2\lfloor nx \rfloor}{n - \lfloor nx \rfloor} \sum_{j=\lfloor nx \rfloor + 1}^{n} j^{-1} = \lim_{n\to\infty} \frac{2x}{1-x} \left[\log n - \log nx \right]$$

$$= \lim_{n\to\infty} \frac{2x}{1-x} \left[\log n - \log n - \log x \right]$$

$$= \frac{-2x}{1-x} \log x$$

REFS:

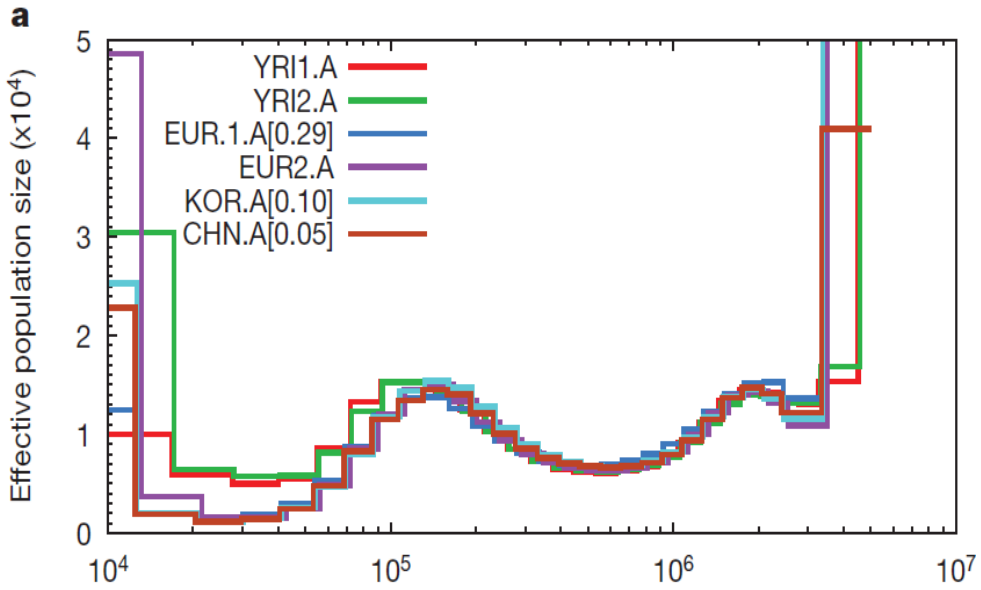
Kimura and Ohta (1973)
Griffiths and Tavare (1998)
Wiuf and Donnelly (1999)

Practical implications

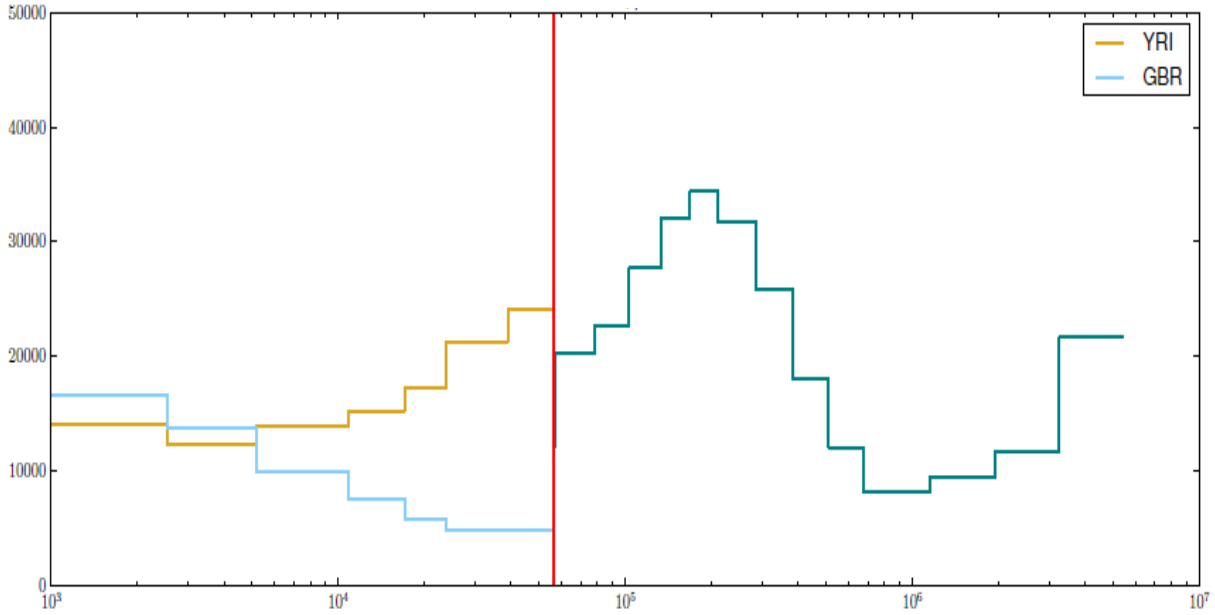
1. This theory is very important in practice!
   - We have seen coalescence times, and hence the frequency spectrum, are affected by historical population size
   - So – we can use the former to infer the latter (e.g. Adams and Hudson, ***Genetics*** 2004, Williamson et al. ***PNAS*** 2005)
   - But there are always multiple possible histories exactly matching an observed spectrum (Myers, Fefferman and Patterson ***Theor. Pop. Biol.*** 2008)
2. The age of a mutation "should" fit with its frequency
   - Selectively advantageous mutations can spread more quickly to high frequency
   - Essentially all the approaches to find real selection, in humans and other species, use this idea
   - Look for mutations which appear young, but are at high frequency

Estimates of ancient human population size

Li and Durbin (Nature, 2011)

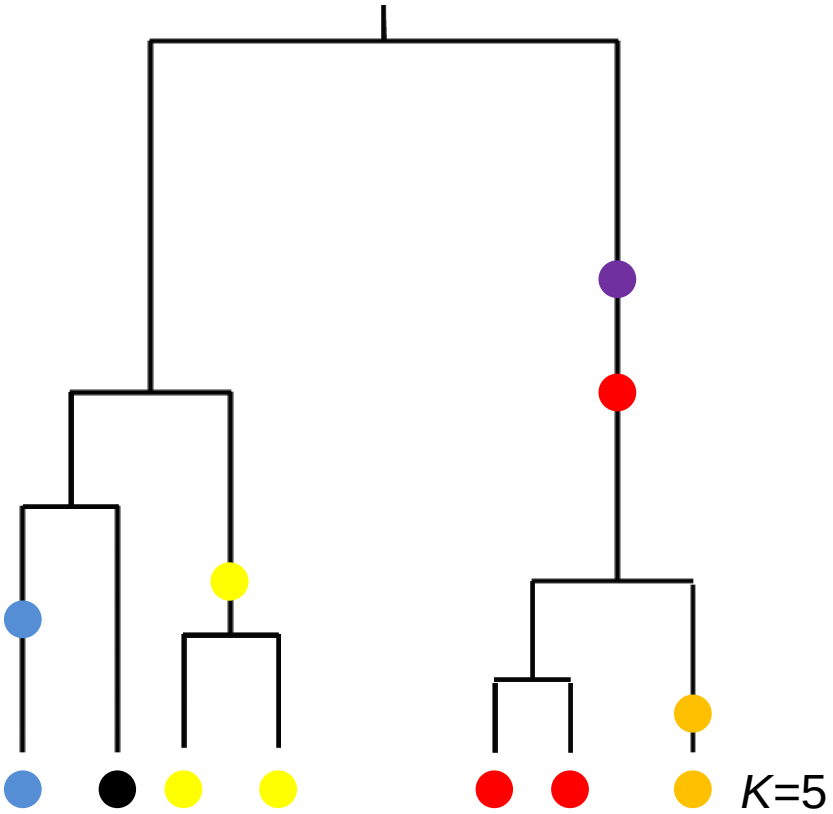


(Marie Forest, Jonathan Marchini, me, unpublished, building trees)



Split: About 80-120,000YBP

4.0 The number of different types



We have talked about the number of segregating sites as a measure of diversity

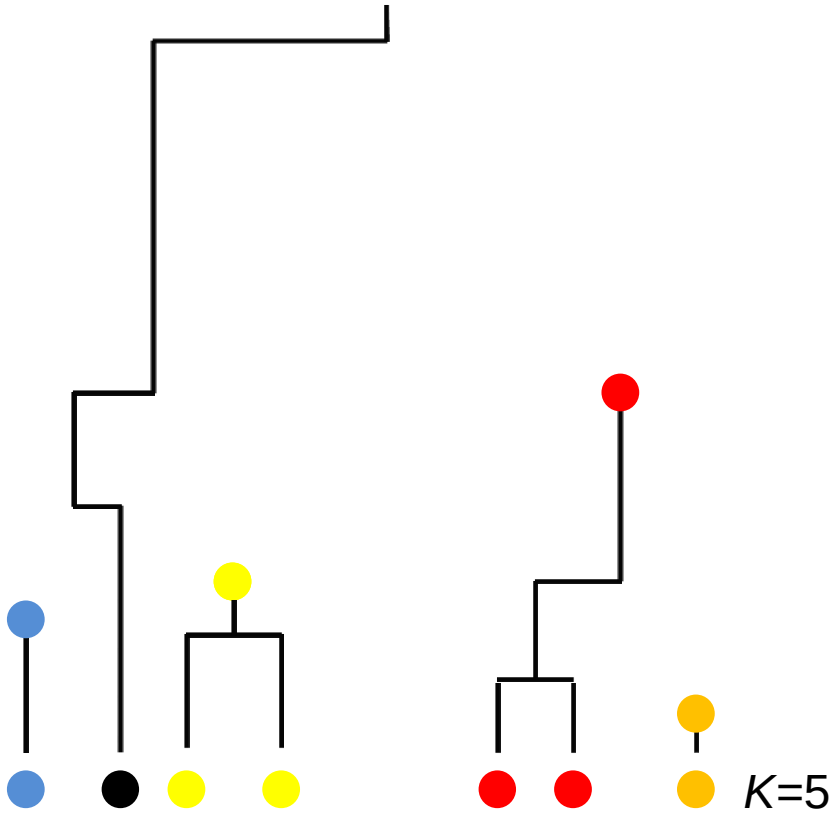
Another natural measure of diversity is the number of *distinct haplotypes K* in a sample. How does this behave?

It is helpful to us to understand the distribution of this number

First, a **definition**. We say the *infinitely-many-alleles* model holds if every mutation makes a new type, never seen before in the population

Note, the infinitely-many-sites model is different from, but implies, infinitely-many-alleles

Following “non-mutant” lines



We will derive the mean, variance and p.g.f of *K,* the number of distinct alleles.

• Looking back in time, view *alleles* (distinct types) as created at mutation events

• To count alleles, we follow the tree, allowing coalescence events, until we see any mutation event – then we know that mutant ancestor passes on a unique type

• We view this as a *death process*: lines “die”, through either mutation or coalescence

• The last line to be lost always represents some final type

Following "non-mutant" lines

Proposition 4.1

Under the infinite-alleles model of mutation for the standard coalescent with mutation rate θ, the number of alleles *K* in a sample of size *n* can be written

$$K = I_n + I_{n-1} + ... + I_2 + 1$$

where the indicator variables I_j are independent and

$$P(I_j = 1) = \frac{\theta}{\theta + j - 1}$$

Proof

Consider following the coalescent history of the sample back in time, allowing lineages to coalesce, and "killing" lineages that mutate, until one lineage remains, at which point the process terminates.

The number of lineages clearly decreases monotonically from *n* to 1. While *j* lineages remain, we are tracing the history of a random sample of *j* lineages in the population, so coalescence occurs at rate *j*(*j*-1)/2 and mutation as a Poisson process of total rate $j\,\theta/2$.

Define I_j=1 if the *j*th lineage is lost by mutation and I_j=0 otherwise. The I_j*'s* are clearly independent. Denoting M_j to be the number of mutations while *j* ancestors in the coalescent:

$$P(I_j = 1) = 1 - P(I_j = 0) = 1 - P(M_j = 0) = 1 - \frac{j-1}{\theta + j - 1} = \frac{\theta}{\theta + j - 1}$$

(*Propn 2.2*)

From the previous discussion, each lineage lost by mutation adds one extra allele, and the last line remaining is an allele, so

$$K = I_n + I_{n-1} + ... + I_2 + 1$$

Following “non-mutant” lines

$$K = I_n + I_{n-1} + ... + I_2 + 1$$

$$P(I_j = 1) = \frac{\theta}{\theta + j - 1}$$

As a corollary, it is straightforward to calculate the mean, variance and p.g.f of *K*:

$$E(K) = E(I_n) + E(I_{n-1}) + ... + E(I_2) + 1$$

$$= 1 + \sum_{j=2}^{n} \frac{\theta}{\theta + j - 1} = 1 + \sum_{j=1}^{n-1} \frac{\theta}{\theta + j}$$

$$Var(K) = \sum_{j=2}^{n} Var(I_j) = \sum_{j=2}^{n} \frac{\theta(j-1)}{(\theta + j - 1)^2}$$

$$= \sum_{j=1}^{n-1} \frac{\theta j}{(\theta + j)^2}$$

By definition of the p.g.f:

$$f_K(z) = E(z^K) = E(z^{I_n + I_{n-1} + ... + I_2 + 1})$$

$$= z \prod_{j=2}^{n} E\left(z^{I_j}\right) = \prod_{j=1}^{n} \left(\frac{j-1}{\theta + j - 1} + \frac{\theta z}{\theta + j - 1} \right)$$

$$= \frac{(\theta z)^{(n)}}{(\theta)^{(n)}}$$ using rising factorials

$$x^{(n)} = x(x+1)\ldots(x+n-1); 1^{(n)} = n!$$

Rates in the coalescent

• A nice, powerful way to think of the coalescent is in terms of event rates. As usual we think backwards in time

• While *j* lineages remain, the total coalescence rate is *j*(*j*-1)/2

• We can think of this as each pair of lineages coalescing, independently, at rate 1

• Similarly, while *j* lineages, the total mutation rate is $j\,\theta/2$, so on each lineage, mutation occurs independently at rate $\theta/2$.

• The rate at which some event occurs is the sum of all the rates, and the probability of each type of event can be obtained by the relative rate.

Example 1: In our death process representation of generating alleles, while *j* lineages (*j*>1):

$$\text{Death rate} = \theta j/2 + j(j-1)/2$$

$$P(I_j = 1) = P(\text{next event mutation}) = \frac{\theta j/2}{\theta j/2 + j(j-1)/2}$$

Example 2: In the general coalescent, the probability the next event is a mutation on lineage *i* say is:

$$\text{Event rate} = \theta j/2 + j(j-1)/2$$

$$P(\text{lineage } i \text{ mutates}) = \frac{\theta/2}{\theta j/2 + j(j-1)/2} = \frac{\theta}{j(j-1+\theta)}$$

The distribution of *K*

We can get the distribution of the number of alleles by expanding the p.g.f:

$$f_K(z) = E(z^K) = \frac{(\theta z)^{(n)}}{(\theta)^{(n)}} = \sum_{k=1}^{n} z^k P(K=k)$$

We use an identity involving *Stirling numbers of the first kind*:

$$x^{(n)} = x(x+1)\ldots(x+n-1) = \sum_{k=1}^{n} \left| S_k^n \right| x^k$$

$$\frac{(\theta z)^{(n)}}{(\theta)^{(n)}} = \sum_{k=1}^{n} \frac{\left|s(n,k)\right|(\theta z)^k}{(\theta)^{(n)}}$$

$$P(K=k) = \frac{\left|s(n,k)\right|\theta^k}{(\theta)^{(n)}}, k = 1,2,\ldots,n$$

If we observe *k* alleles, we can obtain the m.l.e of the mutation rate.

$$l(k) = \log\left[P(K=k)\right] = k\log\theta - \log\prod_{j=1}^{n}(\theta + j - 1) + \text{const}$$

$$\frac{\partial l}{\partial \theta} = \frac{k}{\theta} - \sum_{j=1}^{n} \frac{1}{j-1+\theta} \Rightarrow$$

$$k = \sum_{j=1}^{n} \frac{\hat{\theta}}{j-1+\hat{\theta}} = E(K \mid \theta = \hat{\theta})$$

Thus, the m.l.e. is the *first moment estimator*.

Large samples

We can deduce asymptotic behaviour for the number of alleles:

$$E(K) = 1 + \sum_{j=1}^{n-1} \frac{\theta}{\theta + j} = 1 + \theta \sum_{j=1}^{n-1} \frac{1}{j} - \sum_{j=1}^{n-1} \frac{1}{j(j+\theta)}$$

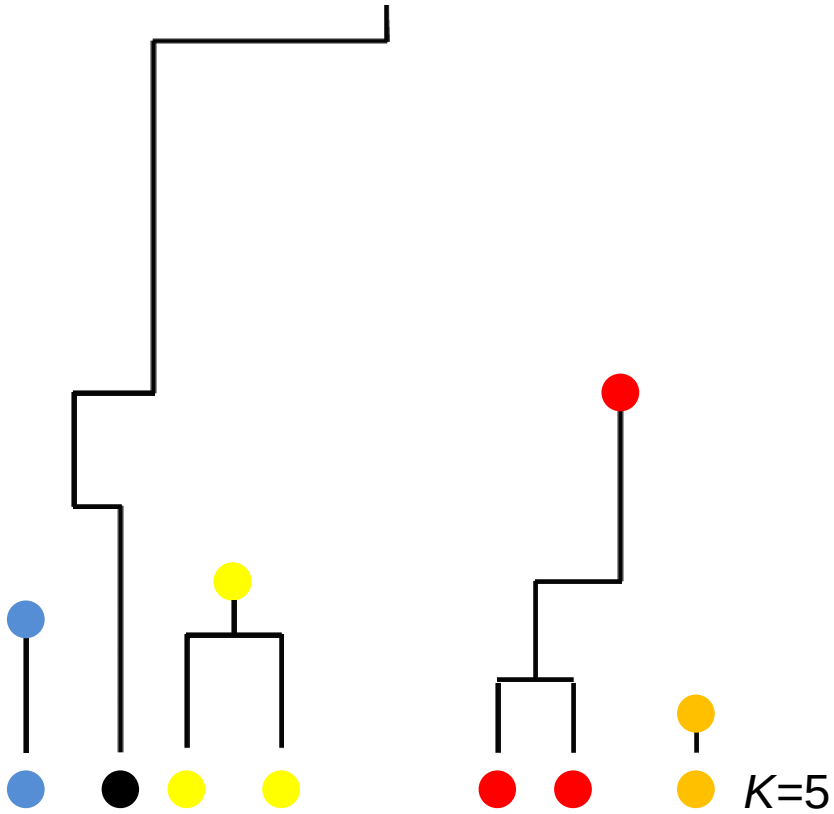
$$E(K) \sim \theta \log n \text{ as } n \to \infty.$$

$$Var(K) = \sum_{j=1}^{n-1} \frac{\theta j}{(\theta + j)^2}$$

$$Var(K) \sim \theta \log n \text{ as } n \to \infty$$

Asymptotically, almost all segregating sites uniquely define a new type in the sample and the number which do not is finite.

Supplement: Multiplicity of alleles



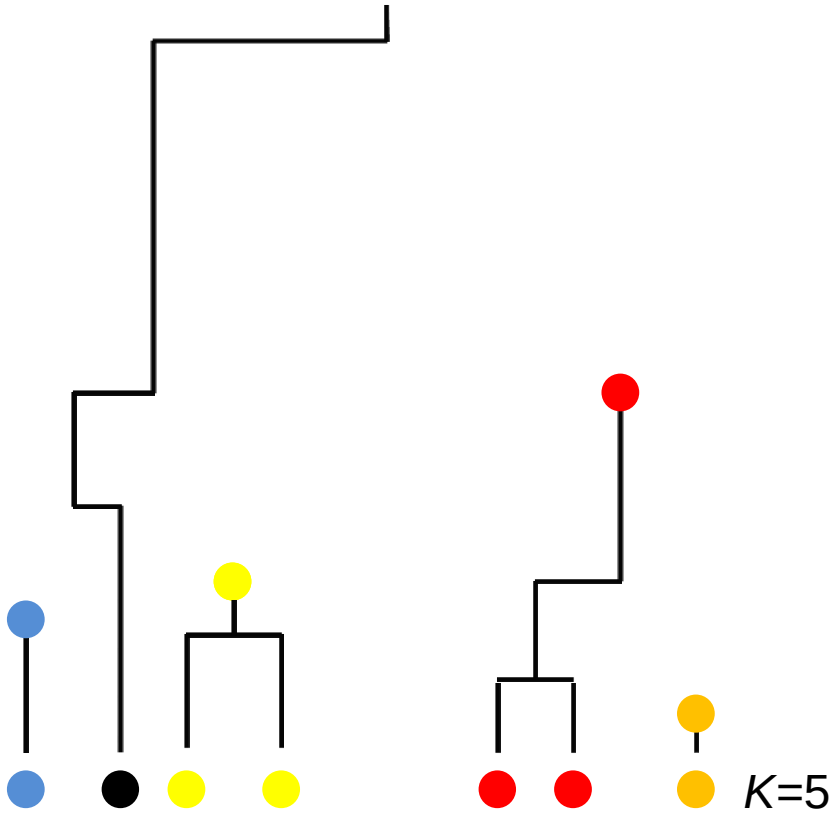
• Suppose we are interested in the full distribution of the number of alleles *and* their frequencies in the sample.

• We will construct an urn model, *Hoppe's urn*, to sample from this.

• Note: the death process shown above defines both the alleles (colours) and how many copies of each is in the sample
• At coalescence events, pairs of lineages coalesce at random
• All lineages are associated with colours

• IDEA: We reverse time in the death process, so new types are "born"

Supplement: Multiplicity of alleles



- Backward in time: While *j* of *n* lineages remain,

$$P(\text{death via mutation}) = \frac{\theta}{\theta + j - 1}$$

- Forward in time, we start with 1 lineage, and while *j*:

$$P(\text{allele born via mutation}) = \frac{\theta}{\theta + j}$$

$$P(\text{particular lineage splits}) = \frac{j}{\theta + j} \times \frac{1}{j} = \frac{1}{\theta + j}$$

- At mutation events, we add a new "colour" to the tree
- At lineage branches, the number of copies of chosen colour increases by 1

Supplement: Hoppe's urn

• We have effectively derived an urn representation

• Represent alleles by balls of different colours in an urn, similarly to the "descendants" urn model we earlier introduced

• We add an extra detail. There's an extra "mutation" ball, of mass θ relative to the other balls with mass 1, and chosen with probability proportional to its mass

Definition (Hoppe's urn model):
Hoppe's urn model constructs a sample of allelic types and multiplicities for *n* haplotypes under the infinite-alleles model

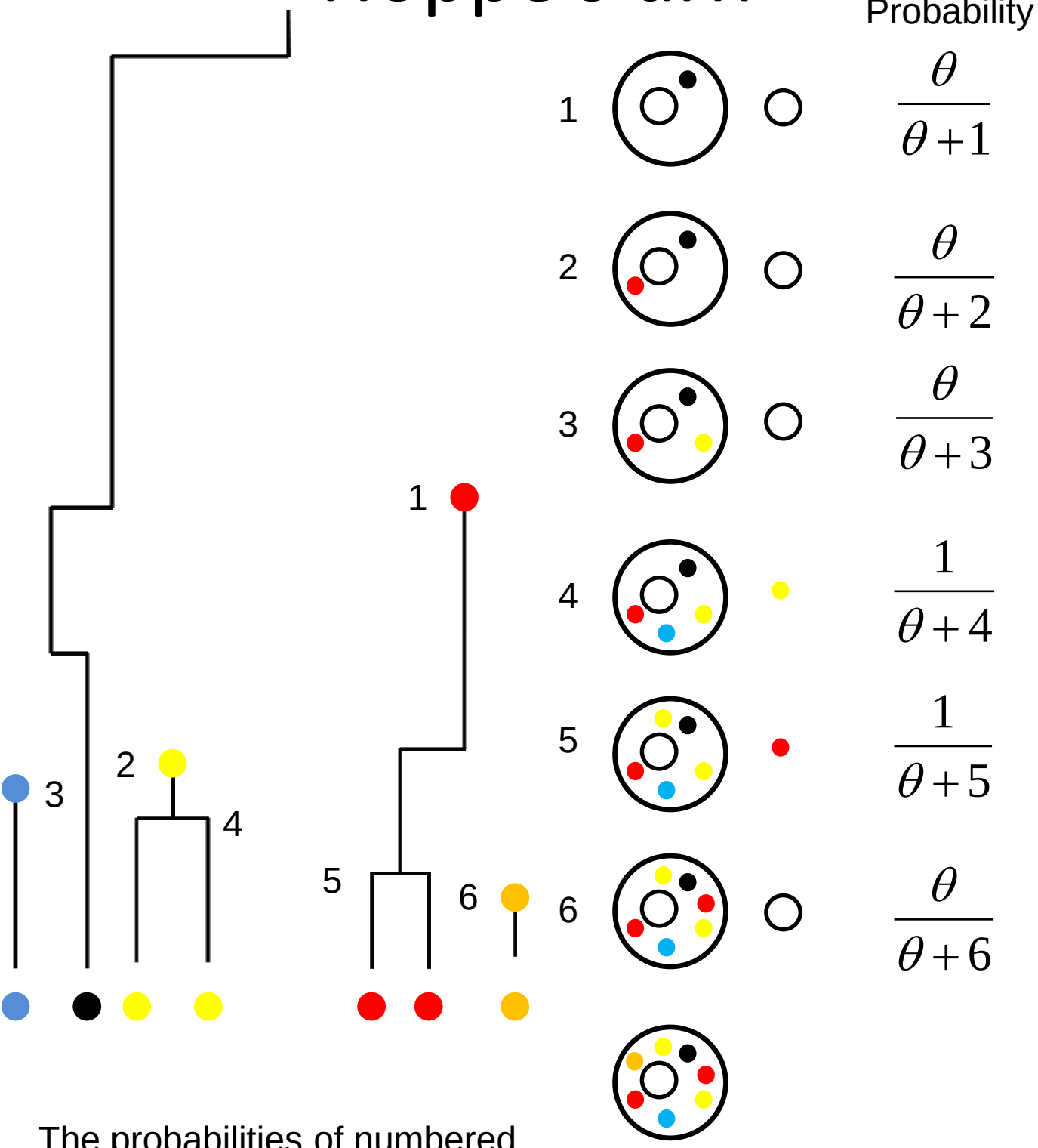
1. Begin with a white and a coloured ball, of mass θ and 1.

2. While *j* non-white balls of mass 1, pull out one of the *j*+1 balls with probability proportional to its mass. If the white ball, replace in the urn and add in a single ball of a new colour. If a coloured ball, replace in the urn and add in an additional ball of the *same* colour

3. When there are *n* non-white balls, stop.

The number of different colours is the number of haplotypes in the sample, and the multiplicity of each colour the multiplicity of these types, summing to *n*.

Supplement: birth/death and Hoppe's urn



The probabilities of numbered events are identical in the urn and the genealogy

Ewen's sampling formula

- Define $\alpha(j)$ to be the number of types occurring at frequency j in the sample for j=1,2,..,n. Then if $K=k$: $$\sum_{j=1}^{n} \alpha(j) = k, \quad \sum_{j=1}^{n} j\alpha(j) = n$$
- *Definition: Ewens' sampling formula* gives the probability of the sample configuration:

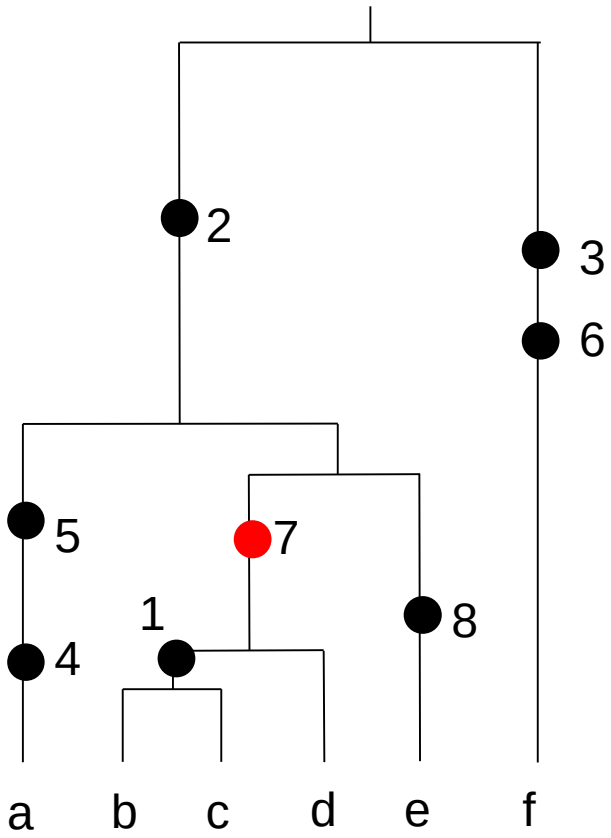
$$P(\alpha(1), \alpha(2), \ldots, \alpha(n)) = \frac{\theta^k}{(\theta)^{(n)}} \frac{n!}{\prod_{j=1}^{n} j^{\alpha(j)} \alpha(j)!}$$

- This can be proved inductively from the urn model
- Note (n,K) is sufficient for θ

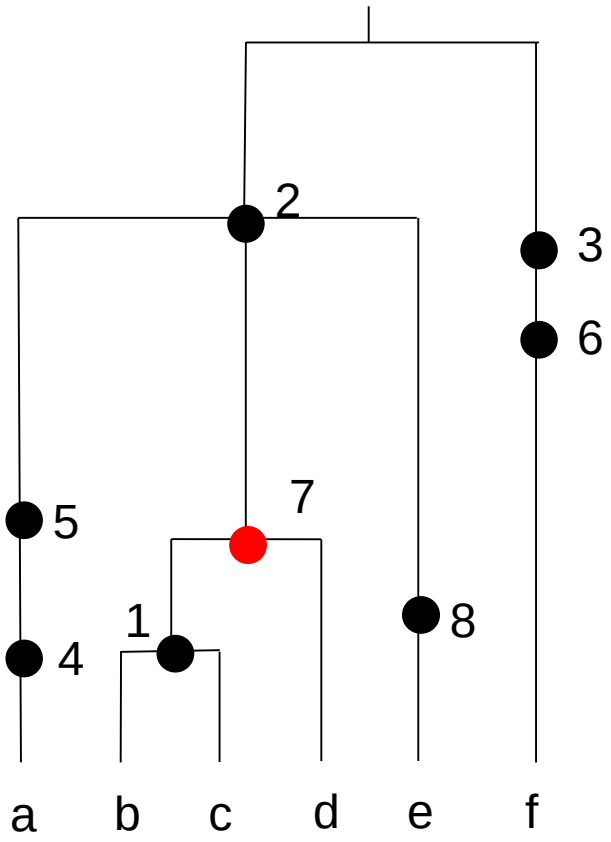
5.0 Gene trees!

- Coalescent trees are not, in general, unique given variation data
- We'd like a historical representation of a sample that is "well defined", but reflects historical relationships among samples
- The solution is to construct a *gene tree*
- We again assume infinite-sites: each mutation occurs at a position never before mutant

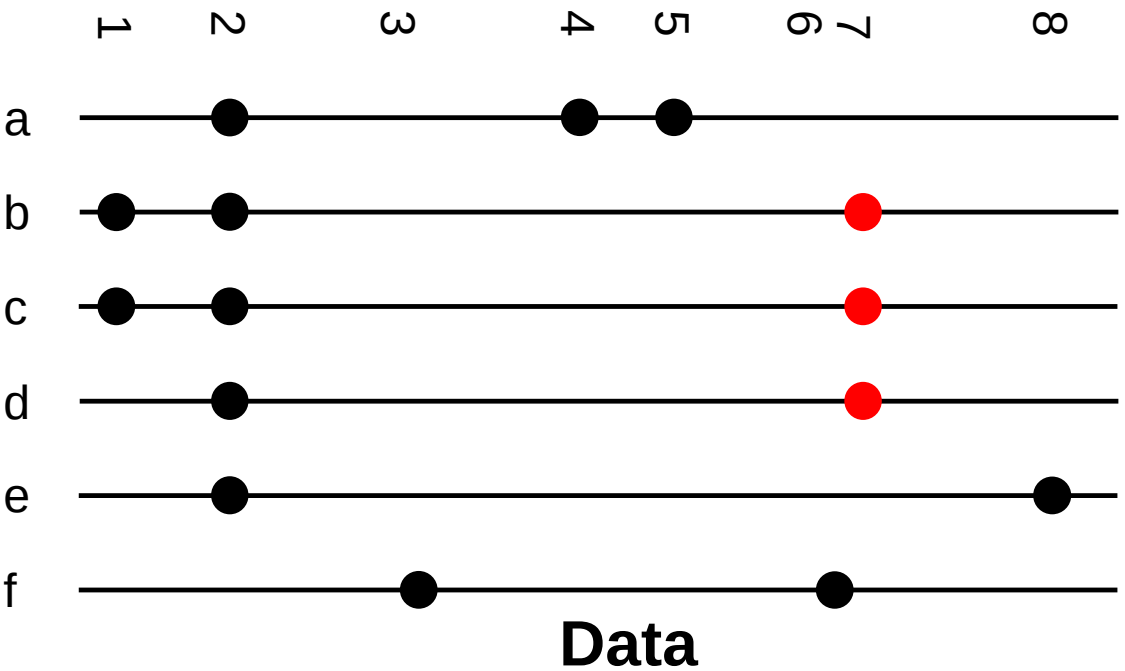
Example gene tree



Coalescent tree



Gene tree



Data

5.0 Gene trees!

- In a gene tree, vertices represent mutations
- These are our information, from variation data
- In general, the tree is not binary and a vertex can have any number of descendants
- We often cluster identical sequences and allow multiplicities on the tips of the tree
- Lineages below a mutation inherit the mutation
- We will show
  1. The data and the gene tree are exactly equivalent
  2. One can check infinite-sites "compatibility" by deriving a necessary and sufficient condition for a gene tree to exist
- To begin constructing a tree, think of our data as binary, with the mutant type denoted by 1, so the "ancestral" type is 0. We define an $n \times s$ ***incidence matrix S***
  - Each column represents a ***segregating site***, with the total number of sites the number of mutations s in the sample history
  - *Each row represents a* ***haplotype***

Sequence\Site	1	2	3	4	5	6	7	8
a	0	1	0	1	1	0	0	0
b	1	1	0	0	0	0	1	0
c	1	1	0	0	0	0	1	0
d	0	1	0	0	0	0	1	0
e	0	1	0	0	0	0	0	1
f	0	0	1	0	0	1	0	0

The incidence matrix

Sequence\Site	1	2	3	4	5	6	7	8
a (1)	0	1	0	1	1	0	0	0
b (2)	1	1	0	0	0	0	1	0
c (3)	1	1	0	0	0	0	1	0
d (4)	0	1	0	0	0	0	1	0
e (5)	0	1	0	0	0	0	0	1
f (6)	0	0	1	0	0	1	0	0

$s_{ij} = 1$ if ind. i mutant at site j, $s_{ij} = 0$ otherwise

$1 \le i \le n, 1 \le j \le s$

- We say a sequence is ancestral if it perfectly matches the type of the ancestor
- This corresponds to a row of zeros in the incidence matrix (mutation occur since the ancestor)
- For site *i*, define the set of carriers of the mutation:

$$O_i = \{m : s_{mi} = 1\}; i = 1,2\ldots, s$$

Example above:

$$O_1 = \{2,3\}, O_2 = \{1,2,3,4,5\}, O_3 = O_6 = \{6\},$$
$$O_4 = O_5 = \{1\}, O_7 = \{2,3,4\}, O_8 = \{5\}$$

Notice that in these data, we have the following:

$$O_1 \subseteq O_7 \subseteq O_2, O_3 \subseteq O_6,$$
$$O_4 \subseteq O_5 \subseteq O_2, O_8 \subseteq O_2$$
$$O_i \cap O_j = \Phi \text{ otherwise}$$

Ordering by inclusion

This pattern turns out to be general, and a powerful way to test the infinitely-many-sites assumption with the incidence matrix:

Proposition 5.1

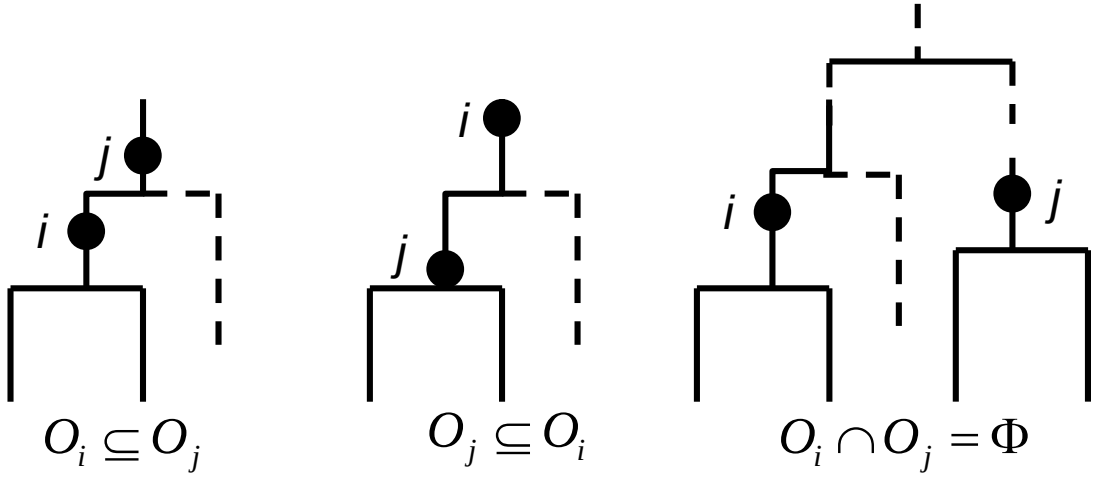
If the infinitely-many-sites model holds, then defining O_i to be the set of individuals in a sample of size *n* carrying the *i*th mutation *i*=1,2,...,*s*, the O_i's are *ordered by inclusion*:

$$\text{for all} 1 \le i, j \le s, \text{either}$$

$$O_i \subseteq O_j, O_j \subseteq O_i, \text{or} O_i \cap O_j = \Phi$$

Proof

Consider the coalescent tree for the sample. For any *i* and *j,* under infinite-sites the *i*th and *j*th mutations occur on tree edges. One of the following must occur: the mutation *i* edge is ancestral to the mutation *j* edge, the opposite occurs, or neither, respectively leading to the three conditions above.



Example

- Is the following dataset, with sequence *c* ancestral, compatible with infinite-sites?

Sequence\Site	1	2	3	4	5	6	7
a (1)	A	G	C	A	C	G	G
b (2)	C	T	T	A	T	A	C
c (3)	C	T	C	A	C	G	C
d (4)	A	T	C	A	T	G	G
e (5)	A	G	C	G	C	G	G

Incidence matrix, noting *c* is ancestral:

Sequence\Site	1	2	3	4	5	6	7
a (1)	1	1	0	0	0	0	1
b (2)	0	0	1	0	1	1	0
c (3)	0	0	0	0	0	0	0
d (4)	1	0	0	0	1	0	1
e (5)	1	1	0	1	0	0	1

	1	5
a (1)	1	0
b (2)	0	1
c (3)	0	0
d (4)	1	1
e (5)	1	0

Check ordering by inclusion. Note that

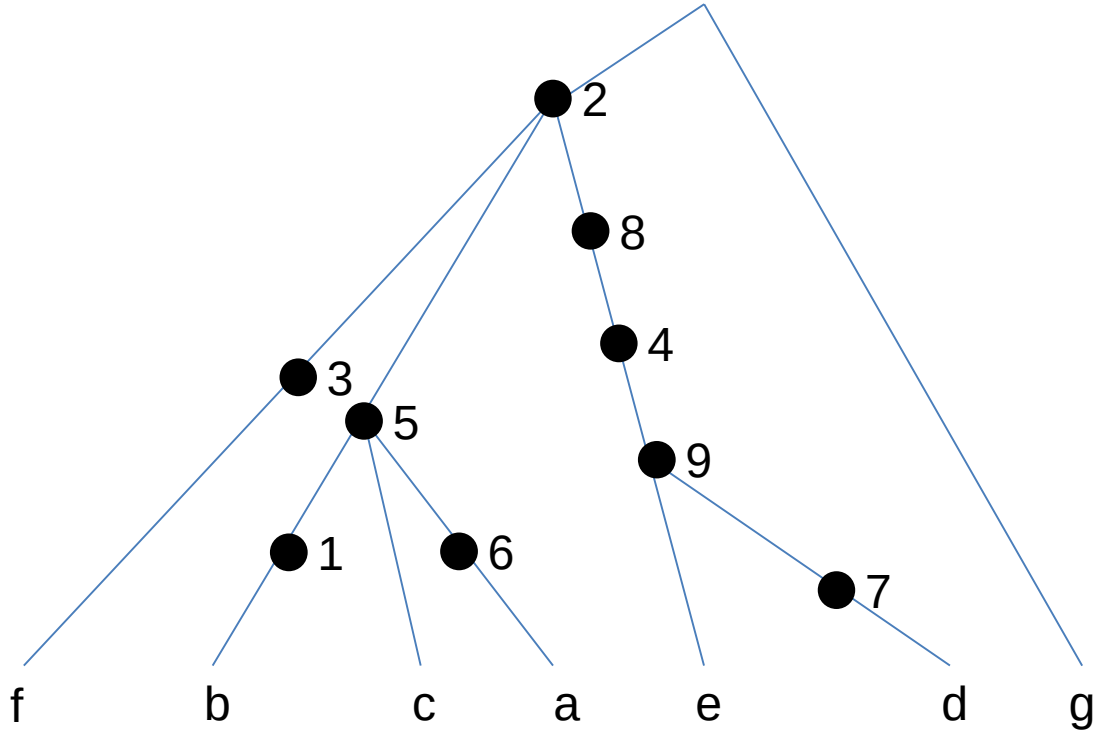
$$O_1 = \{1,4,5\}, O_5 = \{2,4\}, O_1 \cap O_5 \neq \Phi, O_1 \not\subseteq O_5, O_5 \not\subseteq O_1$$

Thus the data are not ordered by inclusion, so **not compatible with infinite-sites**
We will explore this idea more later on.

Note: removing sequence *b* would fix things.

Building gene trees

- Suppose we take a gene tree and trace a “path to the root” for each sequence:



- Denote the root as 0 and go backwards in time:

```
a: 6 5 2 0
b: 1 5 2 0
c: 5 2 0
d: 7 9 4 8 2 0
e: 9 4 8 2 0
f: 3 2 0
g: 0
```

- These “paths to root” are enough to build the gene tree, so equivalent to a gene tree
- We need an algorithm to order mutations from variation data – Gusfield’s algorithm

Gusfield's algorithm

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

Algorithm 5.2

For data compatible with the infinite-sites model, the following algorithm allows the generation of a gene tree based on an incidence matrix consisting of 0's and 1's, with the ancestral type always denoted by 0.

1. Reorder the columns, and column labels, by considering each column as a binary number, and ordering so the columns are decreasing. If duplicate columns occur, choose an arbitrary non-increasing column order.
2. For each sequence, construct a path to the root by reading from right to left in the corresponding row of the incidence matrix, recording mutation labels where 1's occur in rows, and append 0 to this list.
3. Given paths back to the root, use these to draw the gene tree.

Example

A recent common ancestry for human Y chromosomes

Michael F. Hammer, ***Nature*** 1995. 16 sequences, 4 segregating sites seen.

Sequence\Site	1	2	3	4
a (7)	0	0	0	0
b (1)	0	1	0	0
c (3)	1	0	0	0
d (4)	1	0	1	1
e (1)	1	0	0	1

Incidence matrix

1.

Sequence\Site	2	1	4	3
a (7)	0	0	0	0
b (1)	1	0	0	0
c (3)	0	1	0	0
d (4)	0	1	1	1
e (1)	0	1	1	0

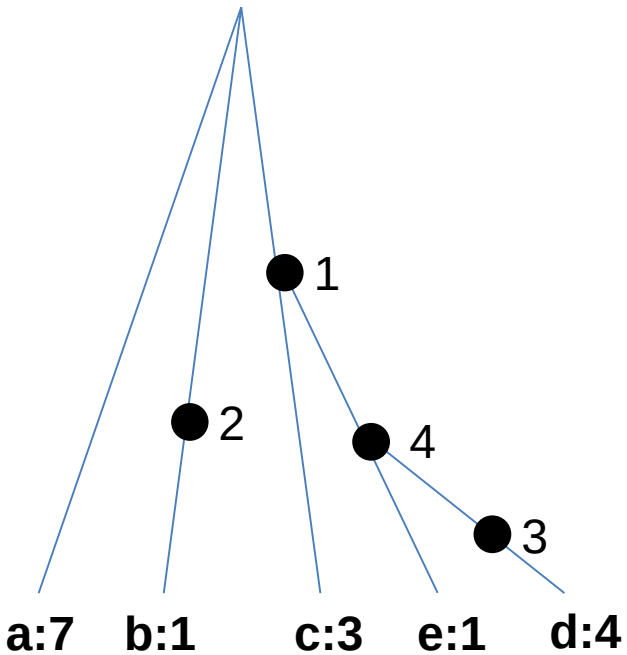
Reordered incidence matrix

2.

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

Paths to root

3.



Gene tree

Variation data ↔ Gene tree

Proposition 5.3

Any variation dataset expressed in the form of an incidence matrix, where the sets of carriers of each mutation are ordered by inclusion, is equivalent to a gene tree.

Notes

1. This implies that ordering by inclusion is both necessary (proposition 5.1) and sufficient for a gene tree to exist, and hence for the data to be consistent with infinite-sites, so this is a complete check
2. Clearly a gene tree can be used to give an incidence matrix, which is automatically compatible with infinite-sites, so we must only prove a gene tree exists given an incidence matrix.
3. We will prove that Gusfield's algorithm correctly produces such a gene tree.

Variation data ↔ Gene tree

Proof of proposition:

We prove Gusfield's algorithm yields a set of paths to root giving a valid gene tree by induction on the number of sequences so far included in the gene tree. We consider constructing the tree, successively adding in sequences.

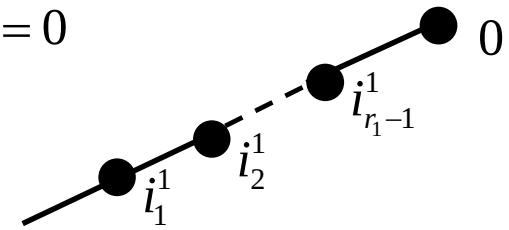
First, assume *wlog* all columns in the incidence matrix are unique (identical columns can be collapsed into one, if present)

After reordering the matrix, viewing each column as a binary number, note that

column $i <$ column $j \Rightarrow$ $O_i \subset O_j$, or $O_i \cap O_j = \Phi$

$O_i \subset O_j \Rightarrow$ column $i <$ column j

The first two algorithm steps obviously lead to a set of sequences of paths to root for each row in the incidence matrix. For the first sequence, we simply add the ordered sequence of mutations that sequence carries, $i_1^1\ i_2^1 \ldots i_{r_1}^1 = 0$



Suppose we have successfully added *k*-1 sequences to the gene tree.

Variation data ↔ Gene tree

Proof ctd:

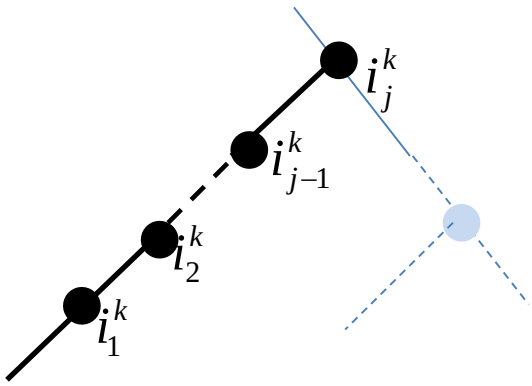
We consider adding the *k*th sequence and Gusfield's algorithm provides an ordered sequence of mutations this sequence carries:

$$i_1^k \; i_2^k \; \ldots i_{r_k}^k = 0$$

Each of these mutations is carried by sequence *k*, so they are not disjoint, and as noted above:

$$\begin{aligned} &\text{column } i_1^k < \text{column } i_2^k < \ldots < \text{column } i_{r_k}^k \Rightarrow \\ &O_{i_1^k} \subset O_{i_2^k} \ldots \subset O_{i_{r_k}^k} = \{1,2,\ldots,n\} \end{aligned} \qquad (5.1)$$

Mutations on this list are either included on the current gene tree or not. Let i_j^k be the first mutation already included in the current gene tree. Then form a new edge containing mutations $i_1^k, i_2^k, \ldots i_{j-1}^k$ and attach it to node i_j^k, to include individual *k* in the gene tree:



We must now only show that the sequence of mutations on the pre-existing path from i_j^k to the root is exactly

$$i_j^k \; i_{j+1}^k \ldots i_{r_k}^k = 0$$

Variation data ↔ Gene tree

Proof ctd:

Note that by equation (5.1):

$$O_{i_j^k} \subset O_{i_{j+1}^k} \ldots \subset O_{i_{r_k}^k} = \{1,2,\ldots,n\} \quad (5.2)$$

We know that some previous sequence *m* carries mutation i_j^k and hence by (5.2), *m* must carry all the mutations

$$i_j^k \; i_{j+1}^k \ldots i_{r_k}^k = 0$$

Because we successfully added sequence *m* in, according to the inductive hypothesis, these mutations all lie on the pre-constructed gene tree, and by (5.2) , since we add mutations in the order specified by Gusfield's algorithm, they lie on the path *upward* from node to the root.

Conversely, any mutation *q* on the path from node i_j^k up to the root is carried by sequence *m,* and since *m* carries i_j^k, ordering by inclusion implies:

$$O_{i_j^k} \subset O_q$$

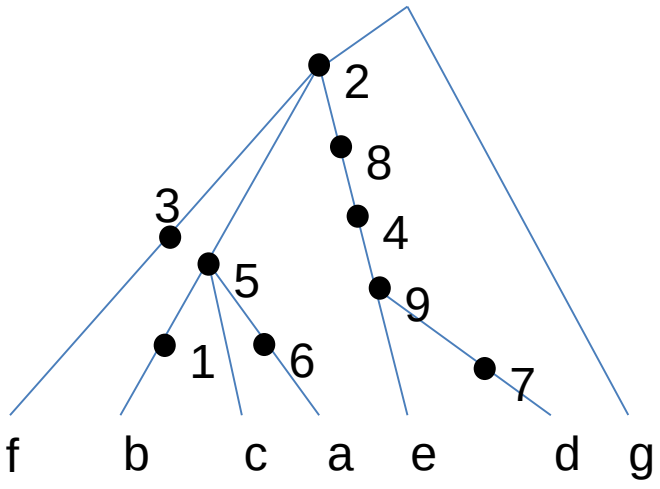
Thus sequence *k* also carries mutation *q*, so for some *r>j*

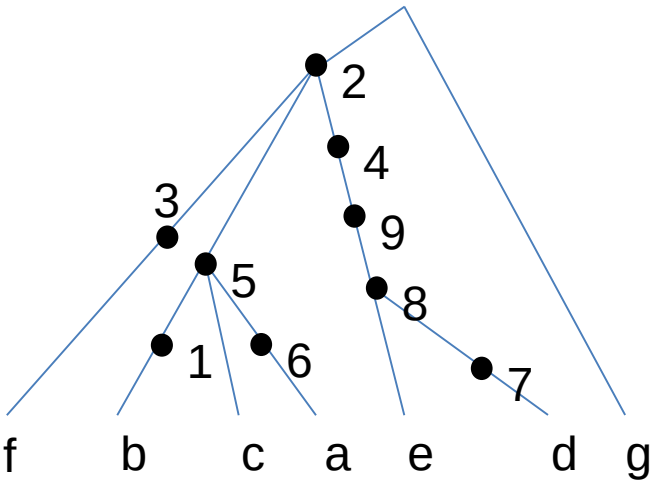
$$q = i_r^k$$

Thus, the mutations on the path from sequence *k* to the root are exactly those carried by sequence *k*, and we successfully add this additional sequence in

Bells and whistles

- Mutations with identical patterns in the sample can be randomly permuted on the edge on which they occur





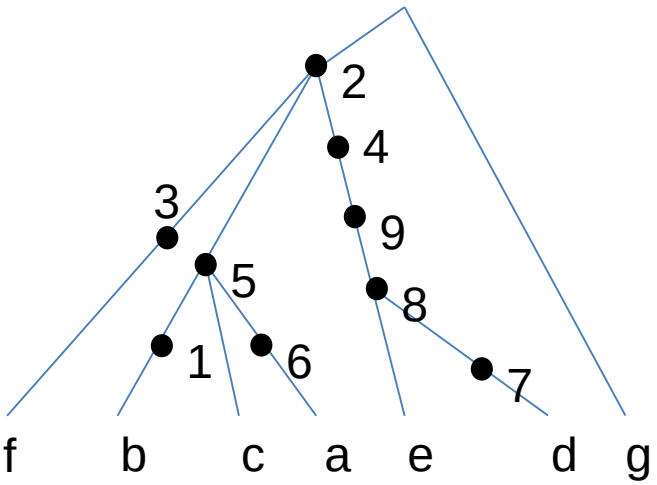
8, 4, 9 all equivalent

- Identical sequences by convention share a single edge, labelled with multiplicity of the sequence
- *Unrooted* trees do not assume we know the ancestral type at each mutation
  - Given say A/G types at a site, we may not be able to infer which is ancestral
  - An unrooted tree incorporates the set of all possible rooted trees.

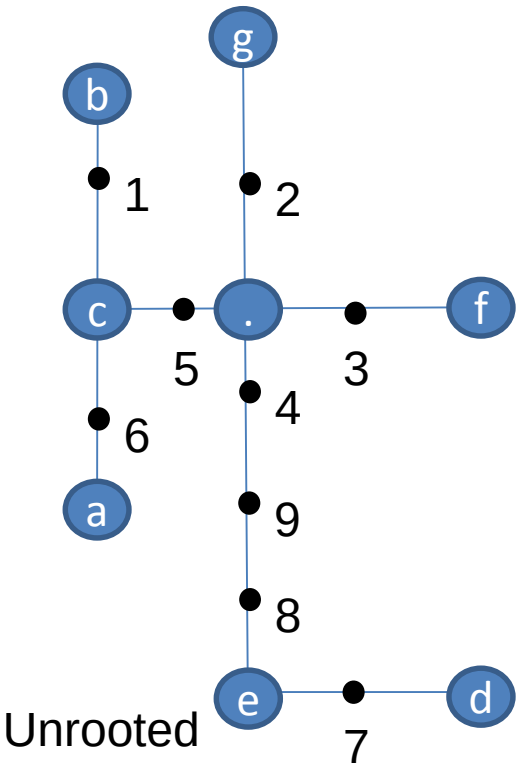
Unrooted trees

- An unrooted tree has sequences (instead of sites) as vertices. Some sequences are inferred in general
- Edges between sequences contain the mutations separating them
- A simple way to construct an unrooted tree from data is to construct a *rooted* tree, then "remove" root
- In general, multiple (rooted) gene trees can give the same unrooted tree.
- Straighten line to root, slide each mutation up from its vertex, and collapse edges with no mutations:

Example 1



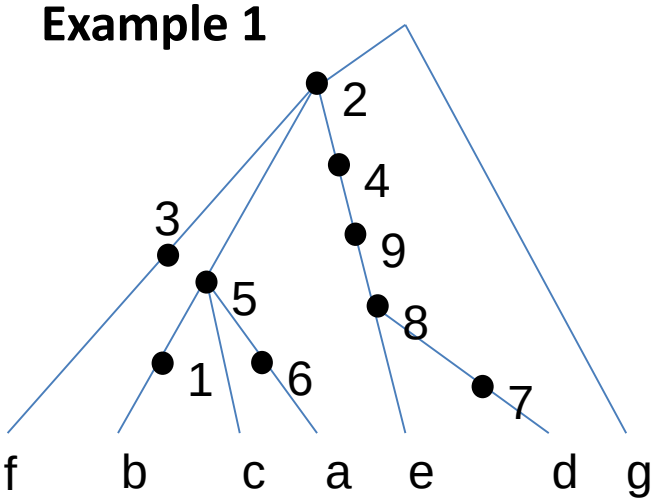
Rooted



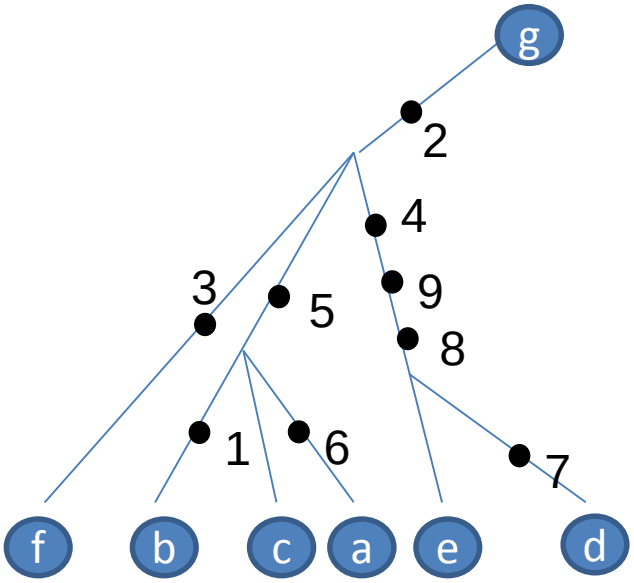
Unrooted

Unrooted trees

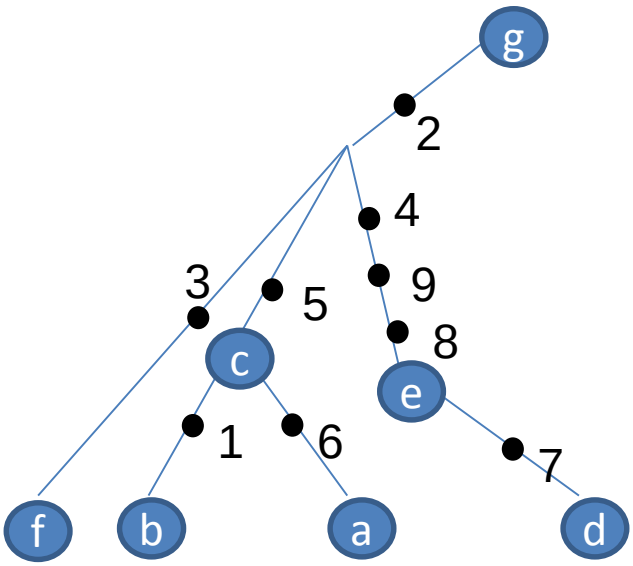
Example 1



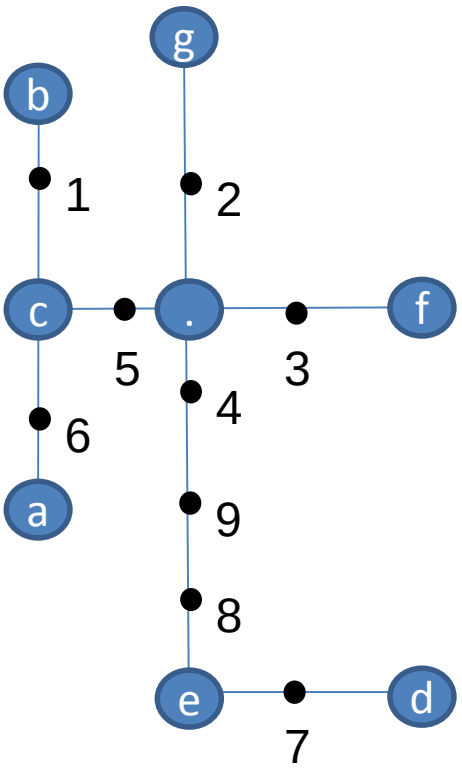
Rooted tree



1. Slide mutations up



2. Remove terminal edges



3. Unrooted tree

Example 2

Mitochondrial DNA sample

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
																			Freq
a	A	G	G	A	A	T	C	C	T	C	T	T	C	T	C	T	T	C	2
b	A	G	G	A	A	T	C	C	T	T	T	T	C	T	C	T	T	C	2
c	G	A	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	T	1
d	G	G	A	G	A	C	C	C	C	C	T	T	C	C	C	T	T	C	3
e	G	G	G	A	A	T	C	C	T	C	T	T	C	T	C	T	T	C	19
f	G	G	G	A	G	T	C	C	T	C	T	T	C	T	C	T	T	C	1
g	G	G	G	G	A	C	C	C	T	C	C	C	C	C	C	T	T	T	1
h	G	G	G	G	A	C	C	C	T	C	C	C	T	C	C	T	T	T	1
i	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	C	C	T	4
j	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	C	T	T	8
k	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	C	5
l	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	T	4
m	G	G	G	G	A	C	C	T	T	C	T	T	C	C	C	T	T	C	3
n	G	G	G	G	A	C	T	C	T	C	T	T	C	C	T	T	T	C	1

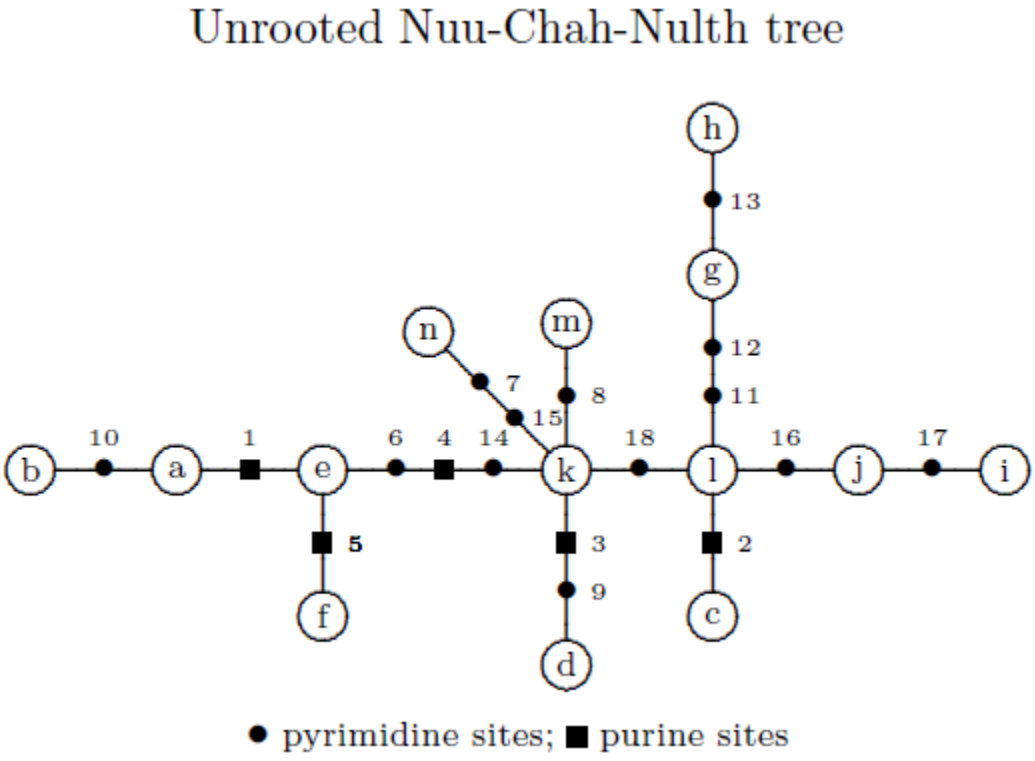
Ward, R. H. Frazier, B. L., Dew, K. and Paabo, S. (1991)
Extensive mitochondrial diversity within a single Amerindian tribe.
Proc. Nat. Acad. Sci. USA **88** 8720-8724.
North American Indian tribe,
the Nuu-Chah-Nulth from Vancouver Island.
$N = 600$ (women).

Example 2

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
a	1	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0
b	1	0	0	1	0	1	0	0	0	1	0	0	0	1	0	0	0	0
c	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
d	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
e	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0
f	0	0	0	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0
g	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1
h	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1
i	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
j	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
k	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
l	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
m	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
n	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0

A rooted tree can be constructed using Gusfield's algorithm from above incidence matrix (root is sequence *k*)

Example 2 continued



For these data I gave one possible choice of ancestral sequences leading to a rooted tree. We could have, e.g., used sequence *l* as ancestral

In general, for an unrooted tree containing *s* mutations, the total number of different sequences on edges (including tips) is *s*+1. Any of these could be the ancestor type.

Hence, there are *s*+1 possible rooted gene trees for a given unrooted tree, in this example 19 rooted gene trees.

Different root choices “toggle” 0 and 1 within columns

Conditions for trees

- The infinite-sites model might be a strong assumption for some species
- Given data, it is of interest to test this model
- Suppose we know ancestral types
- A natural approach is to ask if we can build a rooted gene tree, and hence a coalescent tree.
- If so, we say our data is *compatible* with the infinite-sites model.
- It is easy to prove the following:

Proposition 5.4

A variation dataset expressed in the form of an incidence matrix is compatible with the infinite-sites model if and only if the sets of carriers of each mutation are ordered by inclusion.

Proof

Proposition 5.1 shows necessity of ordering-by-inclusion. The proof of Gusfield's algorithm (Proposition 5.3) shows we can build a gene tree whenever ordering-by-inclusion holds, which immediately implies sufficiency

Conditions for trees

- There is a simple way to test this condition

Corollary 5.5

A variation dataset expressed in the form of an incidence matrix, where ancestral types are coded 0 and mutant types coded 1, is compatible with the infinite-sites model if and only if no pair of sites shows the pattern

$$\begin{matrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{matrix}$$

in any 3 rows of the incidence matrix

Example revisited

- Is the following dataset, with sequence *c* ancestral, compatible with infinite-sites?

Sequence\Site	1	2	3	4	5	6	7
a (1)	A	G	C	A	C	G	G
b (2)	C	T	T	A	T	A	C
c (3)	C	T	C	A	C	G	C
d (4)	A	T	C	A	T	G	G
e (5)	A	G	C	G	C	G	G

Incidence matrix, noting *c* is ancestral:

Sequence\Site	1	2	3	4	5	6	7
a (1)	1	1	0	0	0	0	1
b (2)	0	0	1	0	1	1	0
c (3)	0	0	0	0	0	0	0
d (4)	1	0	0	0	1	0	1
e (5)	1	1	0	1	0	0	1

	1	5
a (1)	1	0
b (2)	0	1
c (3)	0	0
d (4)	1	1
e (5)	1	0

Check new condition.

Note that for sites 1 and 5, rows 1, 2 and 4 respectively give the pattern

$$\begin{matrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{matrix}$$

so these data are incompatible with infinite-sites.

Conditions for trees

Proof of Corollary.

Suppose we see this pattern at two sites, *i* and *j* say and some 3 rows.

	i	j
l	1	0
m	0	1
n	1	1

Clearly

$$O_i \supseteq \{l, n\}, O_j \supseteq \{m, n\}, O_i^c \supseteq \{m\}, O_j^c \supseteq \{l\}$$

$$O_i \not\subset O_j, O_j \not\subset O_i, O_i \cap O_j \neq \Phi$$

so ordering by inclusion does not hold, and the data are incompatible with infinite-sites, proving necessity.

Conversely if the data are incompatible with infinite sites, by the previous proposition for some pair of columns *i*, *j* ordering by inclusion does not hold:

$$O_i \not\subset O_j \Rightarrow \exists l \in O_i, l \notin O_j$$

$$O_j \not\subset O_i \Rightarrow \exists m \notin O_i, m \in O_j$$

$$O_i \cap O_j \neq \Phi \Rightarrow \exists n \in O_i, n \in O_j$$

For columns *i* and *j* and rows *l*, *m*, *n* in the incidence matrix:

	i	j
l	1	0
m	0	1
n	1	1

so the pattern is seen, proving sufficiency

Unknown ancestral types

- If we don't know ancestral types, at each site we can't tell who has the mutation, only who differs
- The incidence matrix is defined up to "toggling" 0-1 status at each site
- The compatibility question becomes whether it is possible to find a toggling to allow a rooted tree

Corollary 5.6

A variation dataset expressed in the form of an incidence matrix, where ancestral types are unknown, is compatible with the infinite-sites model if and only if no pair of sites shows the pattern

$$\begin{matrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{matrix}$$

in any 4 rows of the incidence matrix

Conditions for trees

Proof of Corollary.

Suppose we see this pattern at two sites, *i* and *j* say and some 4 rows.

$$\begin{matrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{matrix}$$

Clearly, toggling 0-1 status at either site this pattern remains. Therefore for any toggling the pattern

$$\begin{matrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{matrix}$$

is seen, and the data are incompatible with a rooted tree and hence infinite-sites. This proves necessity.

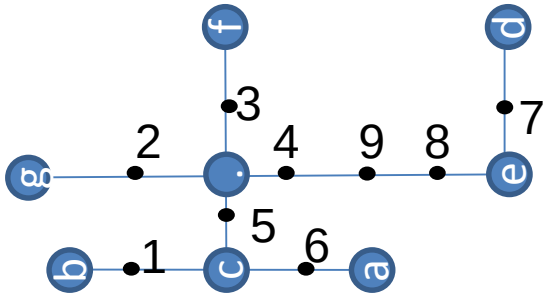
For the converse, suppose there is no such pattern in any pair of columns. Toggle the matrix columns, so the first sequence is a row of zeros (i.e. pick this to be ancestral). Consider columns *i* and *j.* They do not show the pattern $\begin{matrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{matrix}$.

The first row of these two columns is $0 \quad 0$ by construction, so no other 3 rows have the pattern $\begin{matrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{matrix}$.

Hence with this choice of ancestral sequence, there is a rooted gene tree by Corollary 5.3, giving sufficiency.

Unknown ancestral types and unrooted trees

- Any rooted tree can “build” an unrooted tree
- An unrooted tree can “build” multiple rooted trees.
- Notice an unrooted tree is <u>invariant</u> to 0-1 toggling of sites (because mutations on edges just show differences between sequences).
- Thus:
  1. We can view an unrooted tree as the “ancestral type unknown” equivalent of a (rooted) gene tree
  2. The unrooted tree is unique even if ancestral types are unknown (up to permutation of equivalent mutations)
  3. Corollary 5.6 can be viewed as a condition on the <u>existence of an unrooted tree:</u>



Ancestral types known
Infinite sites
⟺
Rooted tree
⟺
No pattern

$$\begin{matrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{matrix}$$

Ancestral types unknown
Infinite sites
⟺
Unrooted tree
⟺
No pattern

$$\begin{matrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{matrix}$$

6.0 The probability of a dataset

- Suppose we observe some variation data
- What is its likelihood?
- We can equivalently think of this as the probability of a gene tree
- We only consider the infinite-sites case

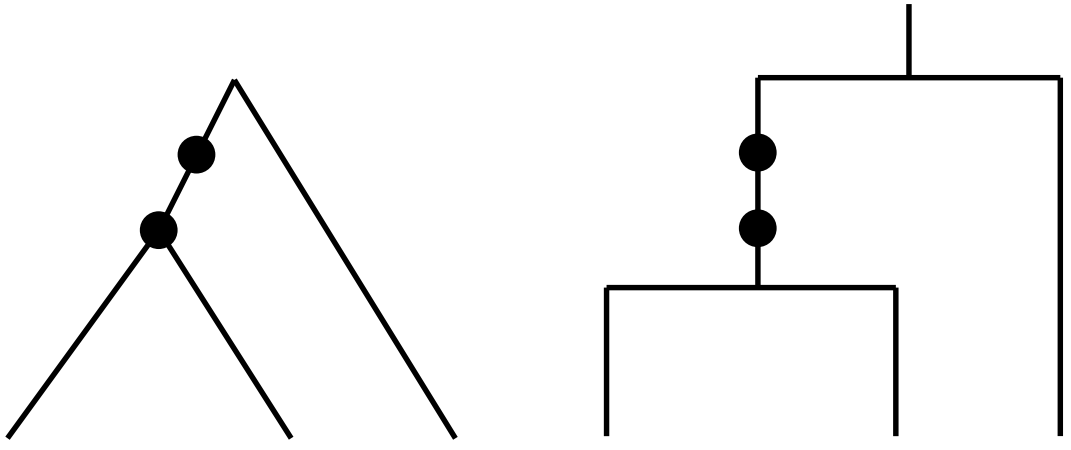
- We begin with a simple example

6.0 Example

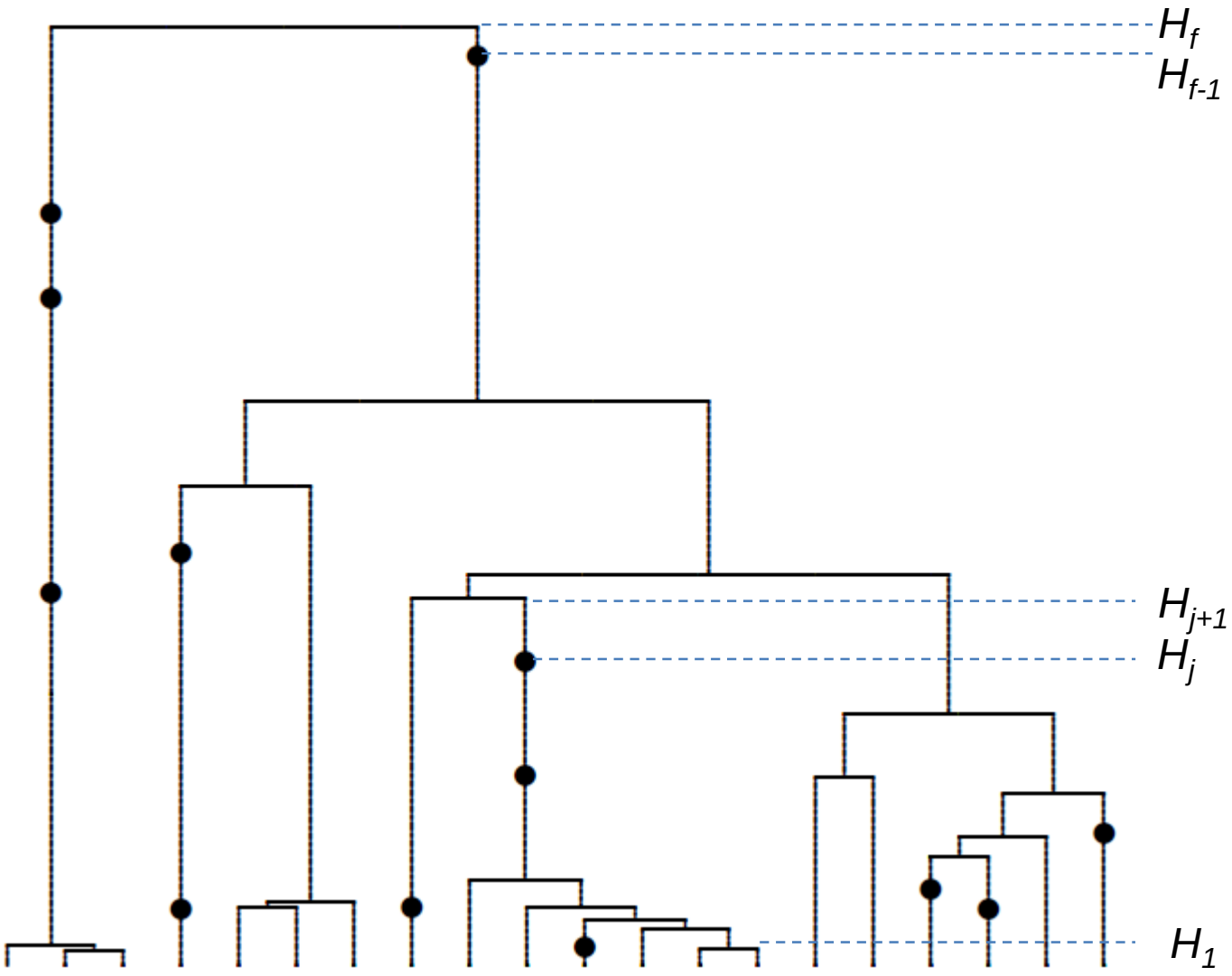
- Consider the following dataset, with 0 ancestral:

Sequence\Site	1	2
a	1	1
b	1	1
c	0	0

- What is the likelihood of the data as a function of θ?
- What is the distribution of the TMRCA conditional on θ and the data?
- First, note there is only one possible coalescent tree:



Coalescent histories



Define the ***history*** of a set of sequences:

$$H = (H_1, H_2, \ldots, H_f)$$

where H_j defines what occurs at the *j*th mutation or coalescence event back in time, i.e. whether this event is a mutation or coalescence event, and which lineage(s) are involved. E.g. H_j shown above is a mutation on lineage 5.

Coalescent histories

Suppose there are k lineages remaining before the jth event. H_j is either a coalescence between two lineages m and n, $C_k(m,n)$ or a mutation on some lineage m, $M_k(m)$:

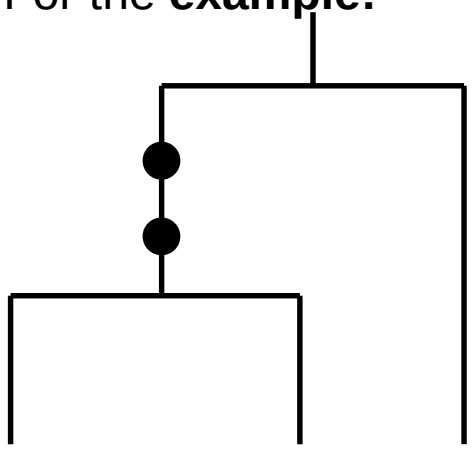
$$\text{Total event rate}: \quad \binom{k}{2} + k\theta/2 = k(k-1+\theta)/2$$

$$P\left[H_j = C_k(m,n)\right] = \frac{1}{k(k-1+\theta)/2} = \frac{2}{k(k-1+\theta)}$$

$$P\left[H_j = M_k(m)\right] = \frac{\theta/2}{k(k-1+\theta)/2} = \frac{\theta}{k(k-1+\theta)} \qquad (6.1)$$

Different events are independent, conditional on the number of edges k remaining, due to the Markov property of Poisson processes.

For the **example:**



H_1 is a coalescence, H_2 and H_3 are mutations, H_4 coales.

$$L = P\left[C_3(1,2), M_2(1), M_2(1), C_2(1,2)\right]$$

$$= \frac{2}{3(3-1+\theta)} \times \frac{\theta}{2(2-1+\theta)} \times \frac{\theta}{2(2-1+\theta)} \times \frac{2}{2(2-1+\theta)}$$

$$= \frac{\theta^2}{6(2+\theta)(1+\theta)^3}$$

Times conditional on history

Having sampled the sample history, suppose there are *k* lineages remaining immediately before the *j*th event, H_j. Events happen as a Poisson process, so times between events are independent and exponential:

$$\text{Total event rate}: \quad \binom{k}{2} + k\theta/2 = k(k-1+\theta)/2$$

$$\text{Let time between events } j-1 \text{ and } j \text{ be } E_j$$

$$E_j \sim \exp\left(k(k-1+\theta)/2\right)$$

$$E\left[E_j\right] = \frac{2}{k(k-1+\theta)}$$

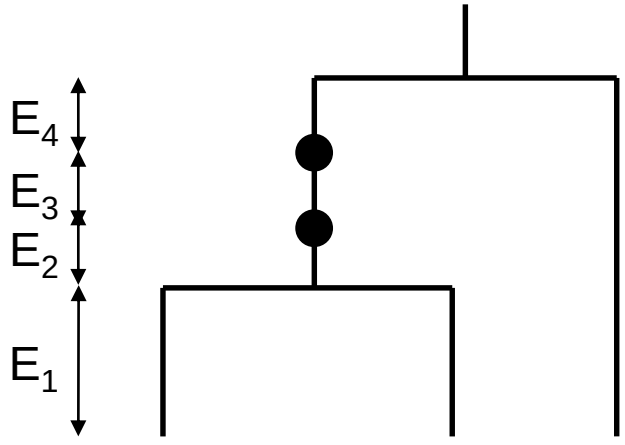
For the **example** we can write the TMRCA as a sum of 4 independent exponentials**.** Then for example:

$$E(T_2+T_3) = \frac{2}{3(3-1+\theta)} + 3\times\frac{2}{2(2-1+\theta)}$$

$$= \frac{2}{3(2+\theta)} + \frac{3}{1+\theta}$$

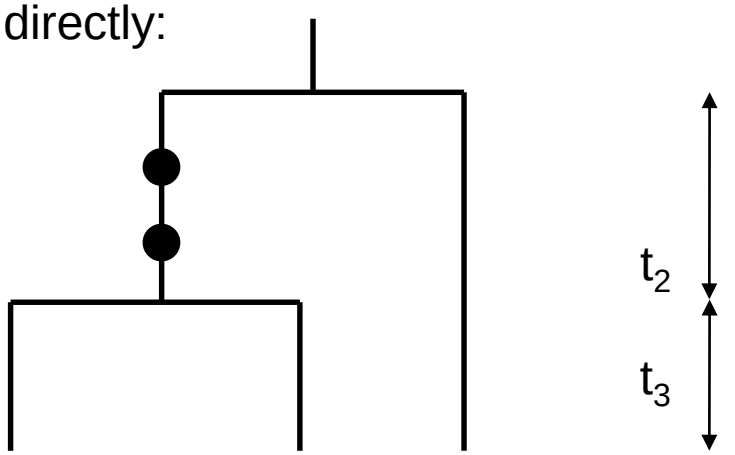
The oldest mutation has expected age

$$\frac{2}{3(2+\theta)} + \frac{2}{1+\theta}$$



Times given data

Alternatively - and equivalently - obtain the joint distribution directly:



$$f(t_3, t_2 \mid D) \propto P(D \mid t_3, t_2)\, f(t_3, t_2)$$

$$= \frac{1}{3}\left[\exp\left(-t_3\theta/2\right)\right]^2 \frac{1}{2}\left(\frac{\theta t_2}{2}\right)^2 \exp\left(-t_2\theta/2\right)\exp\left(-\left[t_3 + t_2\right]\theta/2\right)$$

$$\times 3\exp(-3t_3)\exp(-t_2)$$

This expression integrates to give the likelihood

Normalise to obtain the joint conditional density

The conditional density can be used to give the expected TMRCA:

$$E(T_3 + T_2 \mid D) = \int_0^\infty \int_0^\infty (t_3 + t_2)\, f(t_3, t_2 \mid D)\, dt_3\, dt_2$$

$$= \frac{2}{3(2+\theta)} + \frac{3}{1+\theta}$$

Problem sheet 4 has another example

Complex datasets

- For a general dataset, we have seen there is no unique coalescent tree.
- We can still *sum* over histories
- Given data *D*, define H(*D*) to be the (finite) set of possible histories producing the data
- The likelihood is just:

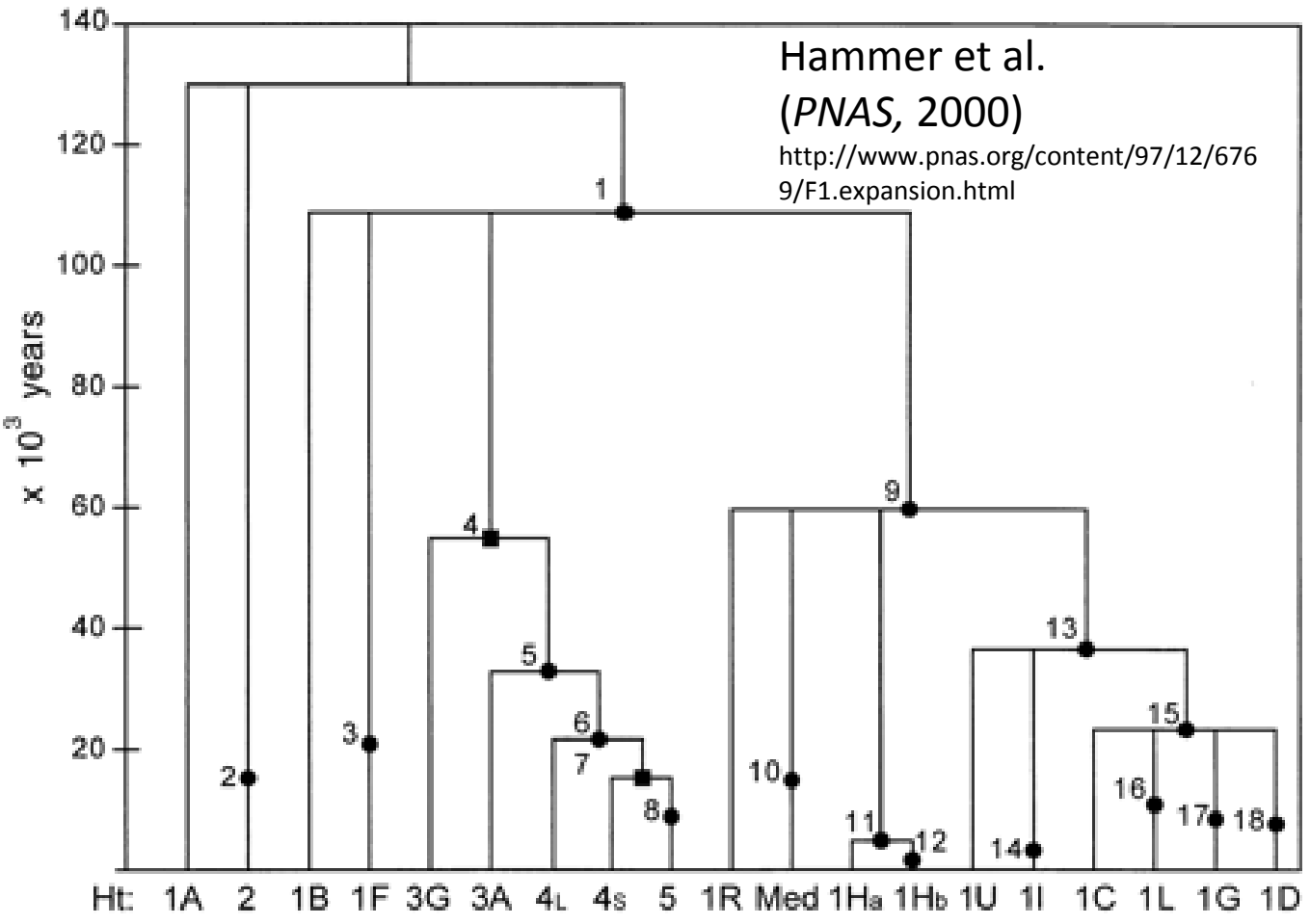
$$L(D) = \sum_{H \in \mathrm{H}(D)} P(H \cap D) = \sum_{H \in \mathrm{H}(D)} P(H)$$

$$= \sum_{H \in \mathrm{H}(D)} P(H_1, H_2, \ldots, H_f)$$

$$= \sum_{H \in \mathrm{H}(D)} \prod_{i=1}^{f} P(H_i \mid k_i)$$

- To obtain expected ages of mutations, average over histories given data

Review



In the Y-chromosome data.

- The model is the constant-size coalescent we derived
- The calibration into years uses M=5,000, estimated as we have seen (and 20 years per generation)
- The structure of the gene tree is drawn using Gusfield's algorithm
- The ages of the mutations are obtained conditional on the data, as we have seen, by "summing" over possible histories

- The supplement describes how this summing was done efficiently, using importance sampling (IS).

Supplement: Importance sampling (IS)

- There is typically an extremely large space of histories to sum...e.g. $n!(n\text{-}1)!/2^n$ trees of n seqs.
- Direct summation often computationally infeasible
- Most histories have a negligible contribution to the likelihood
- Can we add up the "important" terms?
- We use a simple rearrangement to do this

$$L(D) = \sum_{H \in \mathrm{H}(D)} P(H)$$

$$= \sum_{H \in \mathrm{H}(D)} \frac{P(H)}{Q(H)} Q(H)$$

For ANY distribution on histories with p.m.f Q, giving non-zero probabilities over H(D)

$$= E_Q\left[\frac{P(H)}{Q(H)}\right]$$

Q is called a *proposal distribution*

$$\approx \frac{1}{M} \sum_{i=1}^{M} \frac{P(H^i)}{Q(H^i)} \text{ where we sample } H^i \text{ using } Q$$

Importance sampling

$$L(D) \approx \frac{1}{M} \sum_{i=1}^{M} \frac{P(H^i)}{Q(H^i)}$$

*i*th importance weight

- We simulate *M* different histories by sampling each independently using the proposal *Q*
- Calculate the *M* corresponding importance weights and average
- Each importance weight is an i.i.d random variable.
- The previous page shows the mean of the importance weight distribution, sampling under *Q*, is the likelihood we seek.
- The WLLN then implies the likelihood approximation above is exact as M→∞ for *any* valid proposal
- How to pick a "good" proposal distribution, i.e. *Q*?
  - We must be able to write down *Q*
  - Picking a "good" proposal just means trying to make importance weights have low variance
  - In the coalescent setting, that means picking "likely histories" given the data

Supplement: IS

$$L(D) \approx \frac{1}{M} \sum_{i=1}^{M} \frac{P(H^i)}{Q(H^i)}$$

*i*th importance weight

The best (known) scheme for infinite sites is due to Stephens and Donnelly (***JRSS B***, 2000).

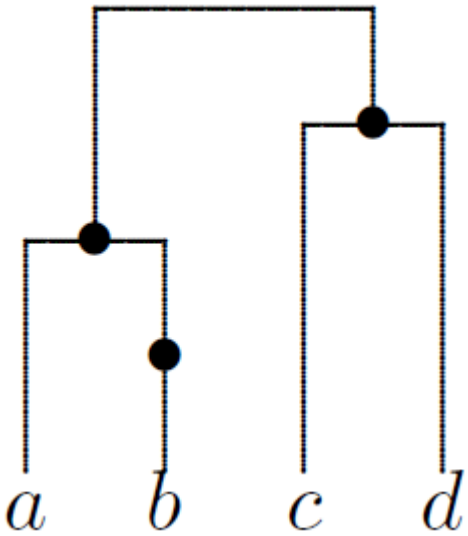
They construct a proposal distribution *Q* on histories as follows.

1. Sample historical events successively back in time.
2. Before event *j*, identify the subset of n_0 lineages to whom the next event could occur
3. Choose one of these lineages uniformly at random: P(lineage *i*)=1/ n_0 and perform the (unique) corresponding mutation or coalescence event
4. Return to step 2 until common ancestor reached

[Note: no θ! The mutation rate comes in to the importance weights only through *P*.]

Example

Consider a dataset corresponding to the below gene tree, and sampling using Stephens' and Donnelly's *Q*.



For the first event in history lineage *b* could mutate, or lineages *c* and *d* might coalesce. Lineage *a* cannot be involved

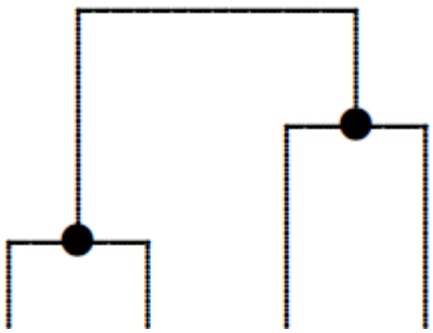
Thus n_0=3 and

$$P[H_1 = M_4(2)] = 1/3$$

$$P[H_1 = C_4(3,4)] = 1/3 + 1/3 = 2/3$$

We choose a first event. Next event chosen same way, until common ancestor reached

If $H_1 = M_4(2)$:



All lineages can have an event, n_0=4:

$$P[H_2 = C_4(1,2) \mid H_1 = M_4(2)] = 1/4 + 1/4 = 1/2$$

$$P[H_2 = C_4(3,4) \mid H_1 = M_4(2)] = 1/4 + 1/4 = 1/2$$

Continuing, for example:

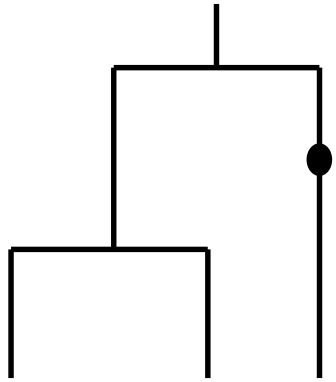
$$Q[H = M_4(2), C_4(3,4), M_3(3), C_3(1,2), M_2(1), C_2(1,2)] = \frac{1}{3} \times \frac{1}{2} \times \frac{1}{3} \times 1 \times 1 \times 1$$

$$= \frac{1}{18}$$

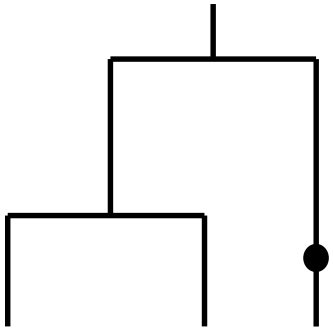
Importance sampling example

Sequence\Site	1
a	0
b	0
c	1

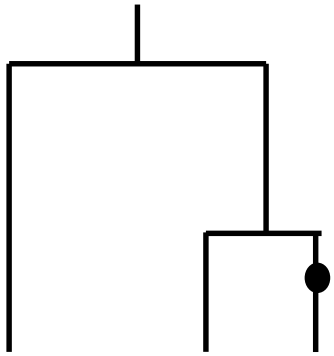
4 possible histories:



History 1: $C_3(1,2),M_2(3),C_2(1,2)$



History 2: $M_3(3),C_3(1,2),C_2(1,2)$



History 3: $M_3(3),C_3(2,3),C_2(1,2)$

History 4: $M_3(3),C_3(1,3),C_2(1,2)$

Initially: Sequence 1 or 2 can coalesce, sequence 3 can mutate so n_0=3. Any one of the 3 sequences can be chosen for the first event, with probability 1/3.

$$Q(\text{Hist 1}) = Q(C_3(1,2),M_2(3),C_2(1,2)) = \frac{2}{3}\times 1\times 1$$

$$Q(\text{Hist 2}) = Q(M_3(3),C_3(1,2),C_2(1,2)) = \frac{1}{3}\times\frac{1}{3}\times 1 = \frac{1}{9}$$

$$Q(\text{Hist 3}) = Q(\text{Hist 4}) = \frac{1}{9}$$

Importance sampling example

Sequence\Site	1
a	0
b	0
c	1

$$Q(\text{Hist 1}) = Q(C_3(1,2), M_2(3), C_2(1,2)) = \frac{2}{3} \times 1 \times 1$$

$$Q(\text{Hist 2}) = Q(M_3(3), C_3(1,2), C_2(1,2)) = \frac{1}{3} \times \frac{1}{3} \times 1 = \frac{1}{9}$$

$$Q(\text{Hist 3}) = Q(\text{Hist 4}) = \frac{1}{9}$$

$$P(\text{Hist 1}) = P(C_3(1,2), M_2(3), C_2(1,2)) = \frac{2}{3(2+\theta)} \times \frac{\theta}{2(1+\theta)} \times \frac{2}{2(1+\theta)}$$

$$P(\text{Hist 2}) = P(M_3(3), C_3(1,2), C_2(1,2)) = \frac{\theta}{3(2+\theta)} \times \frac{2}{3(2+\theta)} \times \frac{2}{2(1+\theta)}$$

$$P(\text{Hist 3}) = P(\text{Hist 4}) = \frac{\theta}{3(2+\theta)} \times \frac{2}{3(2+\theta)} \times \frac{2}{2(1+\theta)}$$

by equation (6.1)

History	Likelihood terms	Prob.	Importance weight
History 1	$\frac{\theta}{3(2+\theta)(1+\theta)^2}$	2/3	$\frac{\theta}{2(2+\theta)(1+\theta)^2}$
History 2	$\frac{2\theta}{9(2+\theta)^2(1+\theta)}$	1/9	$\frac{2\theta}{(2+\theta)^2(1+\theta)}$
History 3	$\frac{2\theta}{9(2+\theta)^2(1+\theta)}$	1/9	$\frac{2\theta}{(2+\theta)^2(1+\theta)}$
History 4	$\frac{2\theta}{9(2+\theta)^2(1+\theta)}$	1/9	$\frac{2\theta}{(2+\theta)^2(1+\theta)}$
Likelihood	$L(D) = \frac{\theta}{3(2+\theta)(1+\theta)}\left[\frac{1}{1+\theta} + \frac{2}{2+\theta}\right]$	-	-

N.B. For any θ value, mean importance weight is true likelihood.
If θ=2, importance weights all identical – scheme is optimal

Glossary of terms used

- Allele: a mutation or combination of mutations forming a distinct type in the population, within the region spanned by a haplotype
- Coalescence event: An event back in time where two or more sequences share a single ancestor
- Effective population size: the size of the Wright-Fisher population (which may change through time) that most accurately models evolutionary history in a real-world population
- Frequency spectrum of mutations: the distribution of the number of mutant copies in a sample of size n over mutations segregating in a sample. We can define the observed frequency spectrum seen in an actual sample, the hypothetical expected frequency spectrum, and the population frequency spectrum (as $n \to \infty$).
- Gene tree. A graphical object representing the history of a sample of sequences, with nodes representing mutations back in time. The type of the ancestor to the sequences corresponds to the top of the tree.
- Haplotype (also loosely referred to as *sequence*, or sometimes *gene*): the DNA sequence of a region of DNA, sometimes interpreted to include only variable positions, and sometimes viewed as a binary sequence of 0's and 1's
- Incidence matrix. A matrix of 0's and 1's representing variation in a sample when there are only two types present at each segregating site. Rows represent sequences, and columns correspond to sites. If known, the ancestral type is often represented by 0 at each site.
- Infinitely-many-sites model. The idea that mutation is rare (true in many species) so that mutations always hit different positions in the genome. This means if a segregating site is observed, it is always the result of a single historical mutation, never two independent, identical mutations at the same position.
- Infinitely-many-alleles model. The related idea that mutations always create new alleles in the population. Thus if two haplotypes are identical, there are no mutations on the history between them before their MRCA. NB – the infinitely-many-sites model implies the infinitely-many-alleles model, so is a special case (under infinite-sites, each mutation is new in the population so trivially defines a new allele).
- Most recent common ancestor (MRCA): the first ancestor in the history of a sample of n sequences who all n sequences are descended from.

Glossary of terms used

- Root sequence: A sequence whose type is identical to that of the MRCA. In the binary infinite-sites model representation of variation, this corresponds to a sequence whose type is all zeros and can be used to define which type is represented as 0, which as 1, at each mutation.
- Segregating site: a mutation seen in some, but not all, members of a sample of size *n*
- Time to the most recent common ancestor (TMRCA): the time back at which the MRCA lived
- Unrooted tree. A graphical object representing the relationships among a sample of sequences, with nodes representing sequences and mutations along edges. The type of the ancestor to the sequences does not need to be known.
- Watterson's estimator: A moment-based estimator of the population scaled mutation rate, based on the number of observed segregating sites in a sample of size *n*.