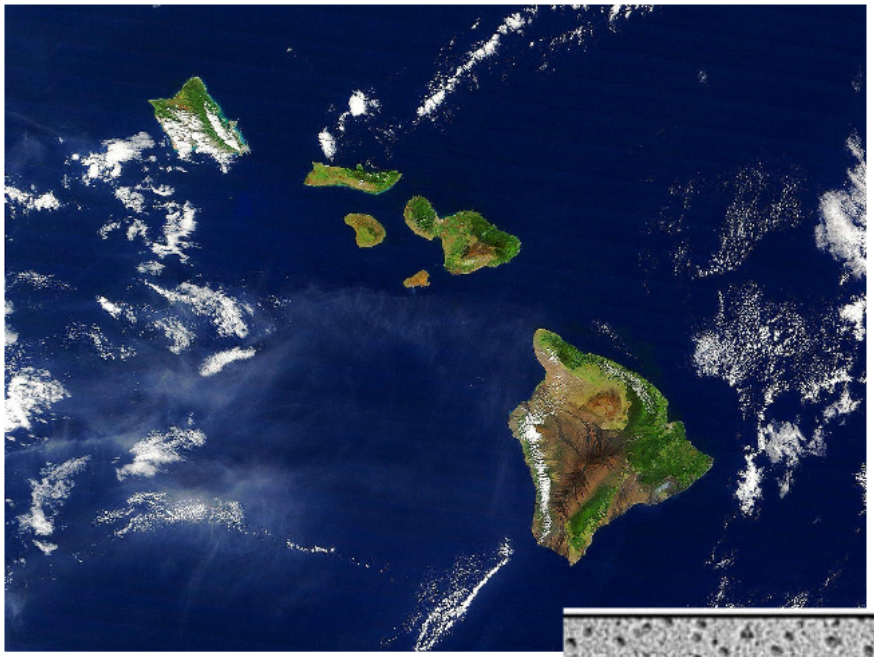
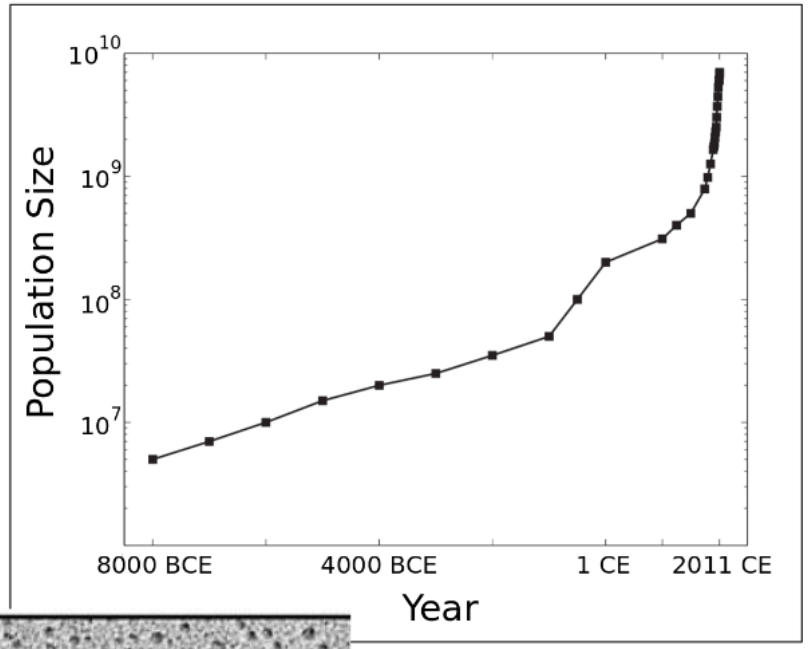
# Extending the coalescent

Mark Reppell

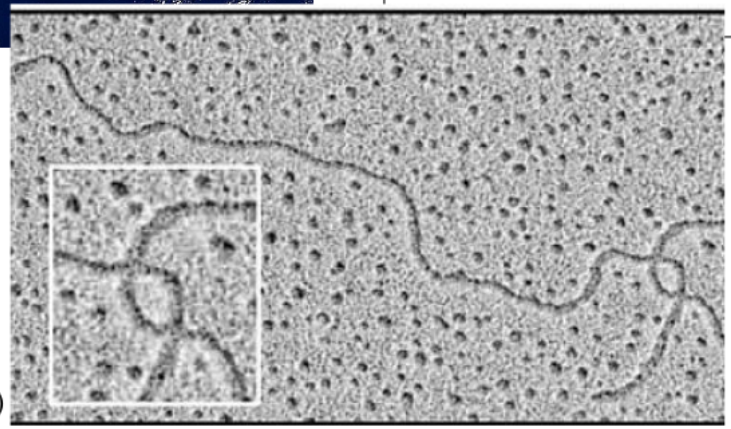
July 21, 2015

# Real populations often violate coalescent assumptions






Keinan and Clark (2012)

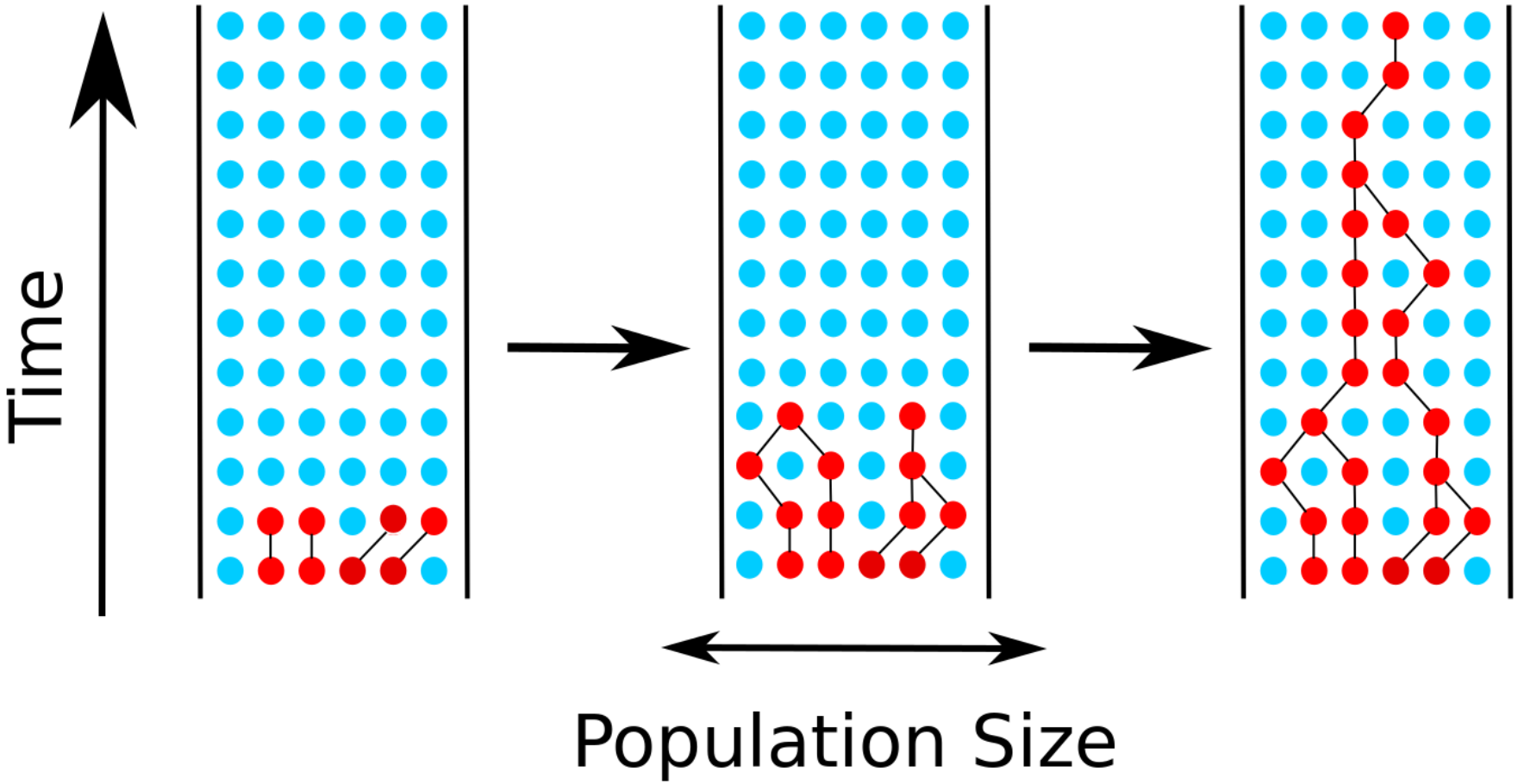


Neil Hunter (2012)

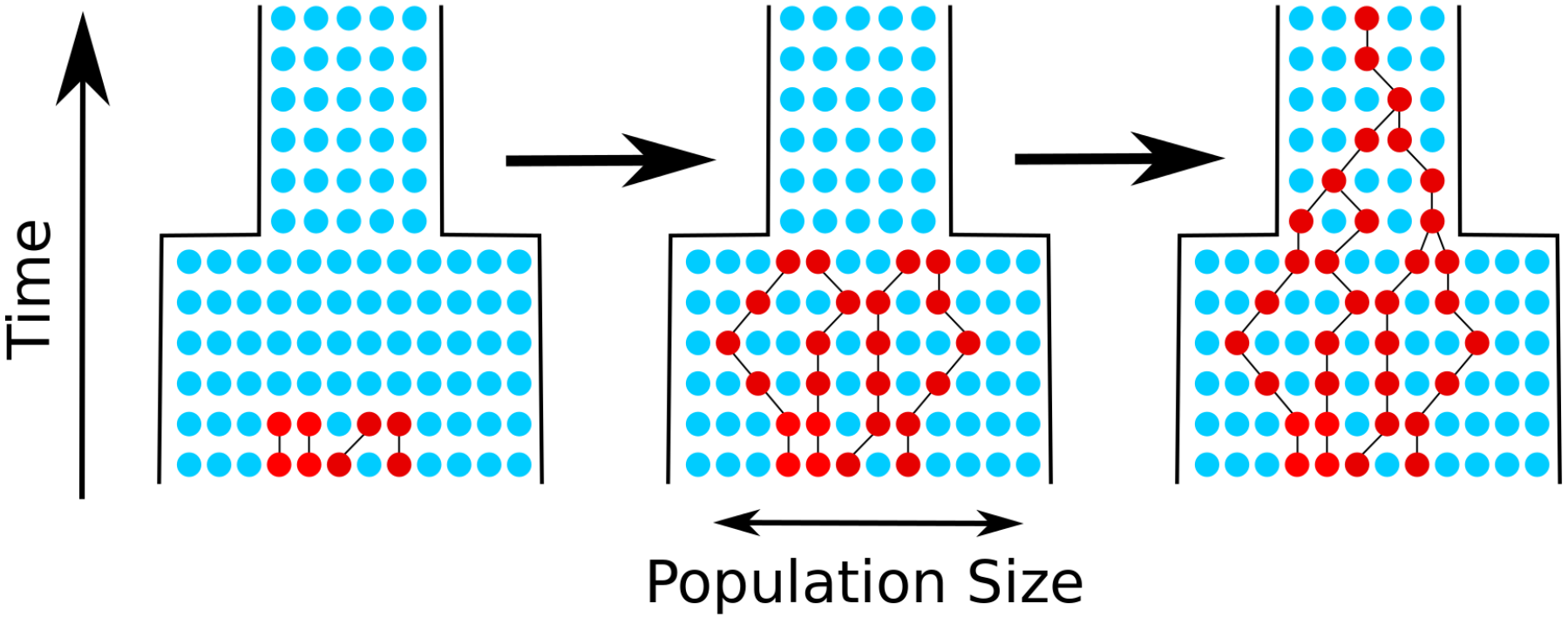
## Extensions to the coalescent

- Variable population size
- Population structure
- Recombination

# Finding common ancestors in a constant size population

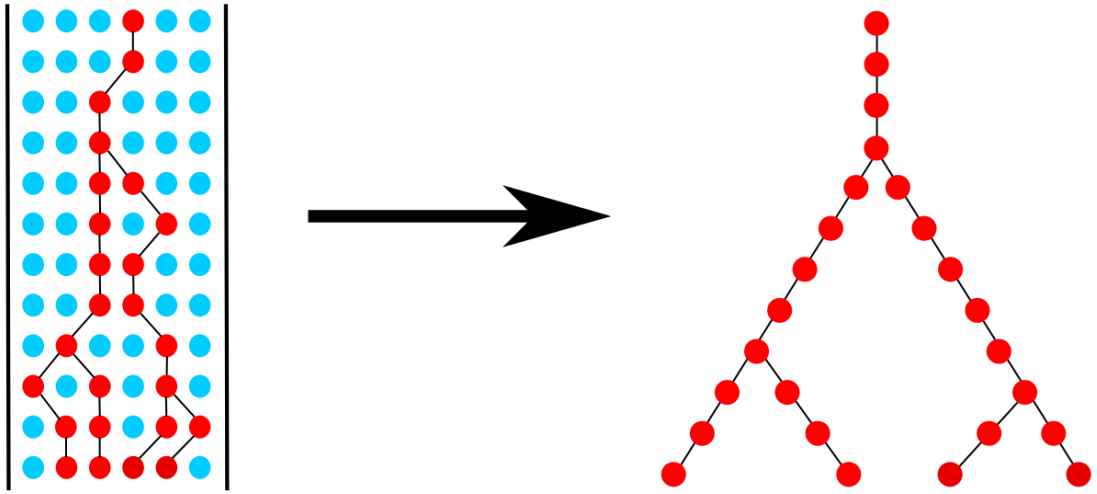


# Finding common ancestors in a variable size population

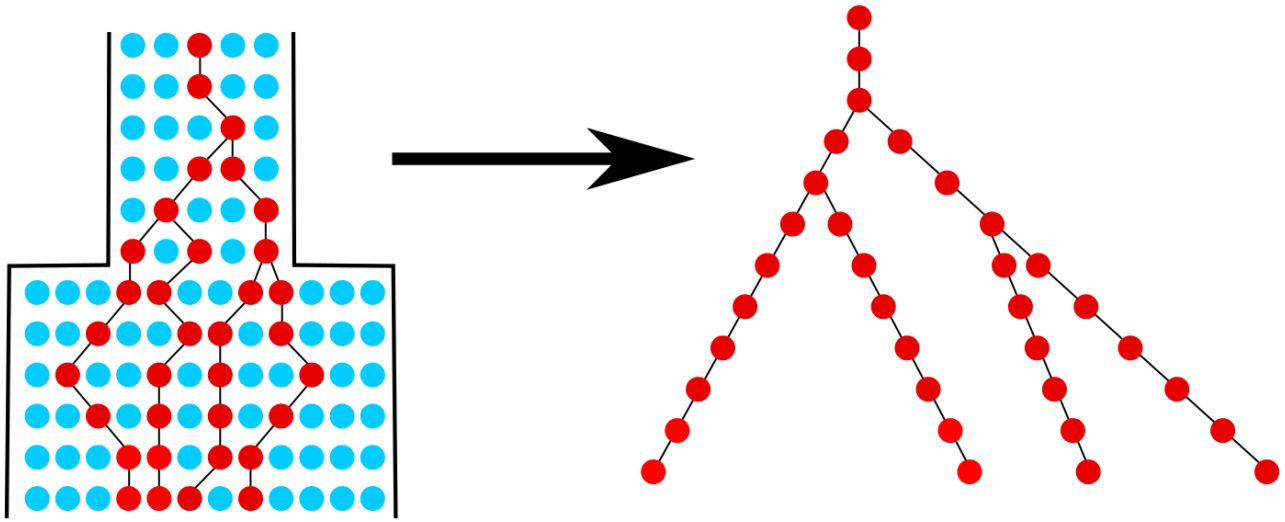



The probability of finding a common ancestor in the preceding generation changes when the population size changes

# Variable population sizes alter genealogies

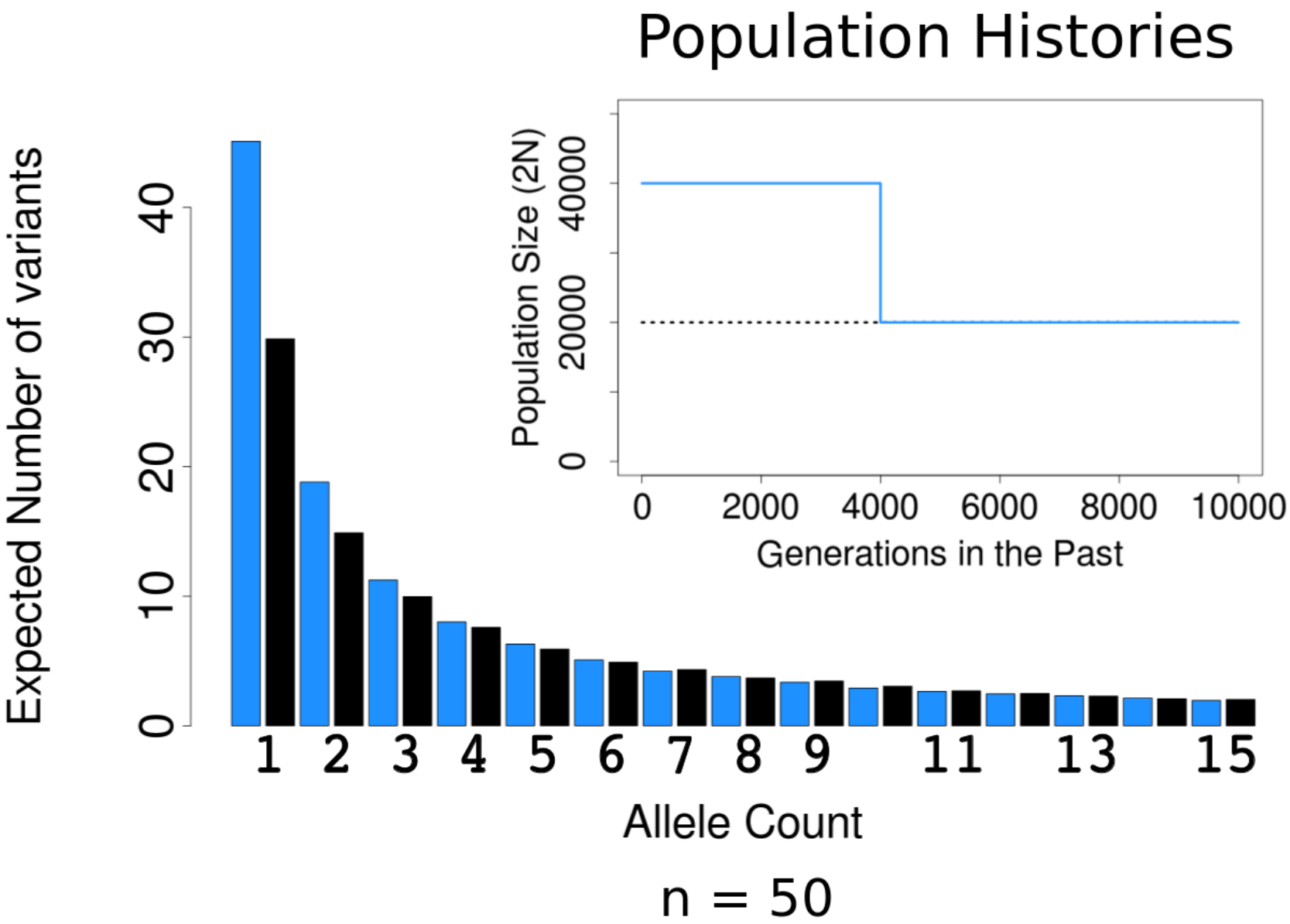


It takes longer for chromosomes to find common ancestors in larger populations



Recent expansion results in proportionally longer external branches

# Sample SFS reflect genealogy structure



# The coalescent with continuous growth

Let $N(t)$ be a population's size at time $t$, define $\lambda(t) = \frac{N(t)}{N(0)}$ and $\Lambda(t) = \int_0^t \frac{1}{\lambda(s)} ds$

We can use $\Lambda(t)$ to "stretch" time from a constant population with size $N(0)$ to one following a function defined by $N(t)$

Griffiths and Tavaré (1994) introduced how to use $\Lambda(t)$: if $T_2, T_3, \ldots, T_n$ are the waiting times while $2, 3, \ldots, n$ lineages remain in a genealogy, and $v_k = T_n + \ldots + T_k$ is the cumulative time until $k-1$ lineages remain:

$$P(T_{k-1} > t | V_k = v_k) = exp\left\{ -\binom{k-1}{2} (\Lambda(t+v_k) - \Lambda(v_k)) \right\}$$

# Coalescent algorithm with continuous growth

## Previous algorithm

**1)** Start with $k = n$

**2)** Simulate waiting time $T_k$ to next event, $T_k \sim \text{Exp}\left(\binom{k}{2}\right)$

**3)** Choose pair of lineages $(i, j)$ uniformly among $\binom{k}{2}$ possible pairs

**4)** Merge $i$ and $j$ into single lineage, and decrease sample size by one $k \rightarrow k - 1$

**5)** If $k \geq 2$ go to **2)**, otherwise stop

**6)** For each branch along genealogy, with length $l$

**a)** draw X mutations, $X \sim Pois(\frac{\theta l}{2})$

**b)** select location of each mutation along sequence uniformly

Update steps **1)** and **2)**:

**1)** Start with $k = n$ and $v_{n+1} = 0$

**2a)** Simulate $t_k^*$ from standard coalescent with $t_k^* \sim Exp\left(\binom{k}{2}\right)$

**b)** Solve for $t_k$ in

$$t_k^* = \Lambda\left(t_k + v_{k+1}\right) - \Lambda\left(v_{k+1}\right)$$

**c)** Use $t_k$ as the time until the next coalescent event

**d)** Update $v_k = v_{k+1} + t_k$

# Exponential growth example

Exponential growth (backwards in time) can be defined as

$$N(t) = N(0)e^{-\beta t}$$

Then $\lambda(s) = e^{-\beta s}$ and $\Lambda(t) = \frac{1}{\beta}\left(e^{\beta t} - 1\right)$ and

$$P(T_{k-1} > t | V_k = v_k) = exp\left\{ -\binom{k-1}{2}\frac{1}{\beta}e^{\beta v_k}\left(e^{\beta t} - 1\right)\right\}$$

So, following the previous slide's algorithm we "stretch" times from the standard coalescent, $t^*$, using

$$t_{k-1} = \frac{1}{\beta} log\left(1 + \beta t^*_{k-1} e^{-\beta v_k}\right)$$

# Important Notes

Under the model of continuous growth, coalescent times are no longer independent of each other!

# A tangent on exponential random variables

If $X_i \sim Exp(\beta_i)$ for $i \in 1, .., k$, are independent and $Y = min(X_i)$

$$P(Y > t) = P(X_1, ..., X_k > t) = P(X_1 > t)P(X_2 > t)...P(X_k > t)$$

$$= e^{-\beta_1 t} \times e^{-\beta_2 t} \times \cdots \times e^{-\beta_k t} = e^{-\sum_{i=1}^{k} \beta_i t}$$

so $Y \sim Exp(\sum_{i=1}^{k} \beta_i)$, the sum of the rates of $X_i$

---

The probability that $Y = X_j, i.e. X_j$ is the smallest $X_i$, is

$$P(X_j < min(X_i), i \in 1, .., k, j \neq i) = P\left(X_j < Z \sim Exp(\sum_{i, i \neq j} \beta_i)\right)$$

$$= \int_0^\infty \int_0^z f(x_j) f(z) dx_j dz = \int_0^\infty \int_0^z \beta_j e^{-\beta x_j} \sum_{i, i \neq j}^{k} \beta_i e^{-\sum_{i, i \neq j}^{k} \beta_i z} dx_j dz = \frac{\beta_j}{\sum_{i}^{k} \beta_i}$$

# The coalescent with population substructure

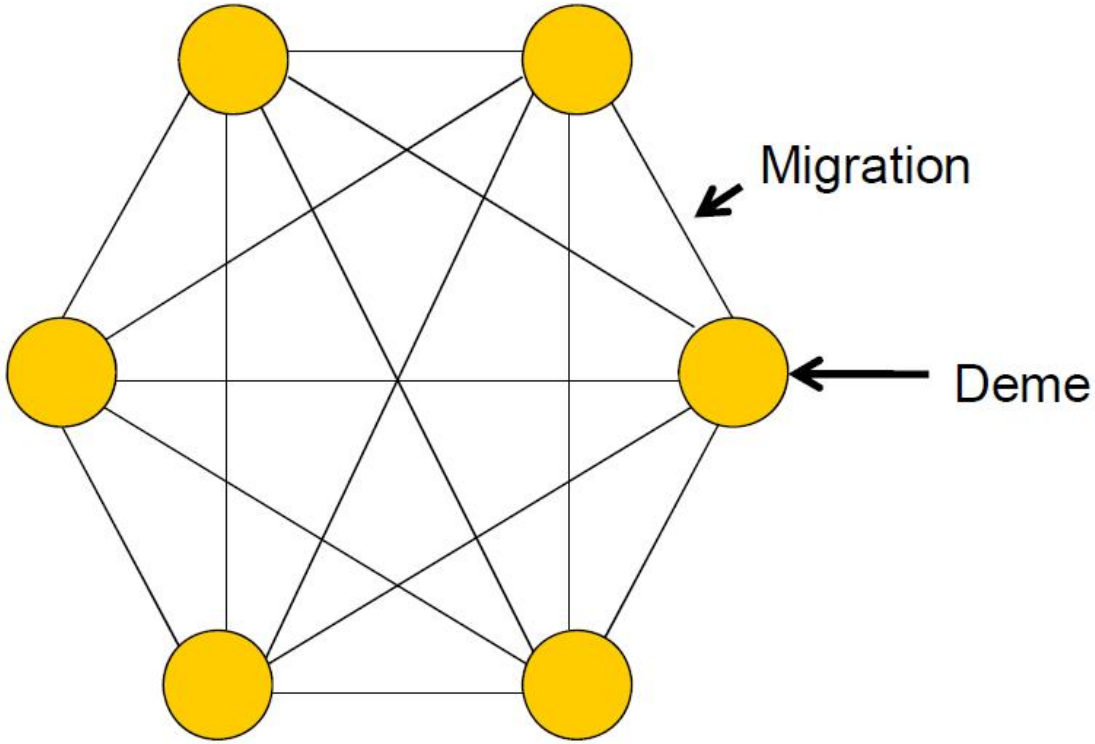
# Population structure

Population structure refers to violations of the random mating assumption

In the coalescent context, this is equivalent to sample lineages having different probabilities of coalescing with each other

A subpopulation from which all samples have the same behavior is referred to as a **deme**

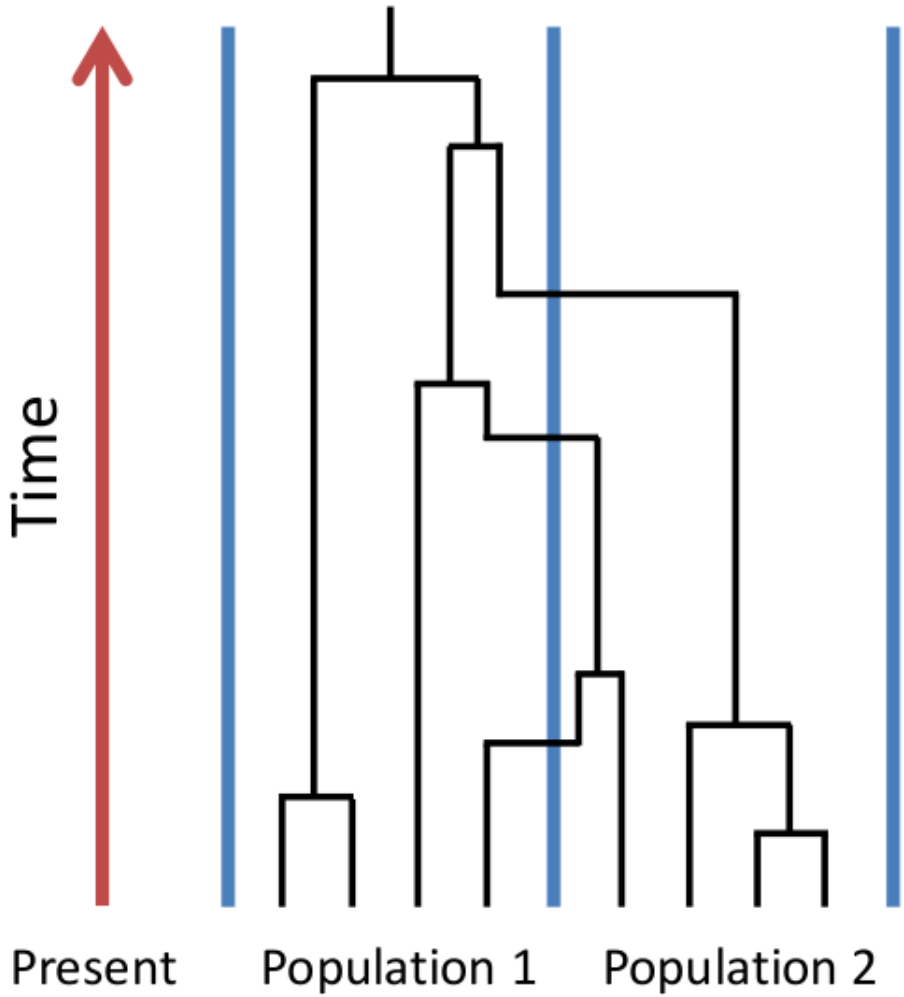
# The island model




Samples migrate between all pairs of demes at identical rate

Within each deme we assume random mating

# Coalescent and the island model




Two possible events: migration or coalescence

A sample can only coalesce with a lineage in the same deme

MRCA only occurs once final samples are in the same deme

For $d$ demes, minimum of $d - 1$ migrations before MRCA

# Migration as an exponential process

Define $m$ as the proportion of migrants leaving/joining a deme each generation. When each deme is size $2N$ and there's a total of $k = \sum_{i=1}^{d} k_i$ samples remaining, with $k_i$ samples in deme $i$

$$P(\text{migration in next generation}) = (1 - (1 - m)^k)$$

Then the probability that the next migration occurs in generation $j$ has probability

$$P(T_M = j) = \left[(1 - m)^k\right]^{j-1} \left(1 - (1 - m)^k\right)$$

The distribution function of a geometric random variable with rate $(1 - (1 - m)^k)$

# Migration as an exponential process

Using the properties of the geometric distribution

$$P(T_M \leq j) = 1 - \left[1 - \left(1 - (1-m)^k\right)\right]^j \approx 1 - \left[1 - \left(1 - (1 - mk)\right)\right]^j$$
$$= 1 - (1 - mk)^j$$

Define compound migration parameter $M = 4Nm$, and scale time by deme size $t = j/2N$

$$\lim_{2N \to \infty} 1 - \left(1 - \frac{kM}{4N}\right)^{2Nt} = 1 - e^{\frac{-Mkt}{2}}$$

Note that the rate of migrations is linear in $k$, whereas the coalescent rate is quadratic

# Adding additional exponential processes to the coalescent algorithm

**1)** Specify all events that can occur

**2)** Determine exponential rate of each type of event

**3)** Use the properties of the exponential distribution to determine

- **a)** Time until the next event of *any* type occurs
- **b)** The type of event that occurs at the first event

**4)** If MRCA reached stop, else back to **1)**

# Simulating with the island model

In model with $d$ total demes, if $k_i$ is the number of lineages in deme $i$, the coalescent rate, $c_i$, and migration rate $b_i$ are

$$c_i = \binom{k_i}{2} \qquad b_i = \frac{Mk_i}{2}$$

Then total rate until the next event in *any* deme is $\sum_{i=1}^{d}[c_i + b_i]$

To simulate:

**1)** Draw time until next event using total event rate

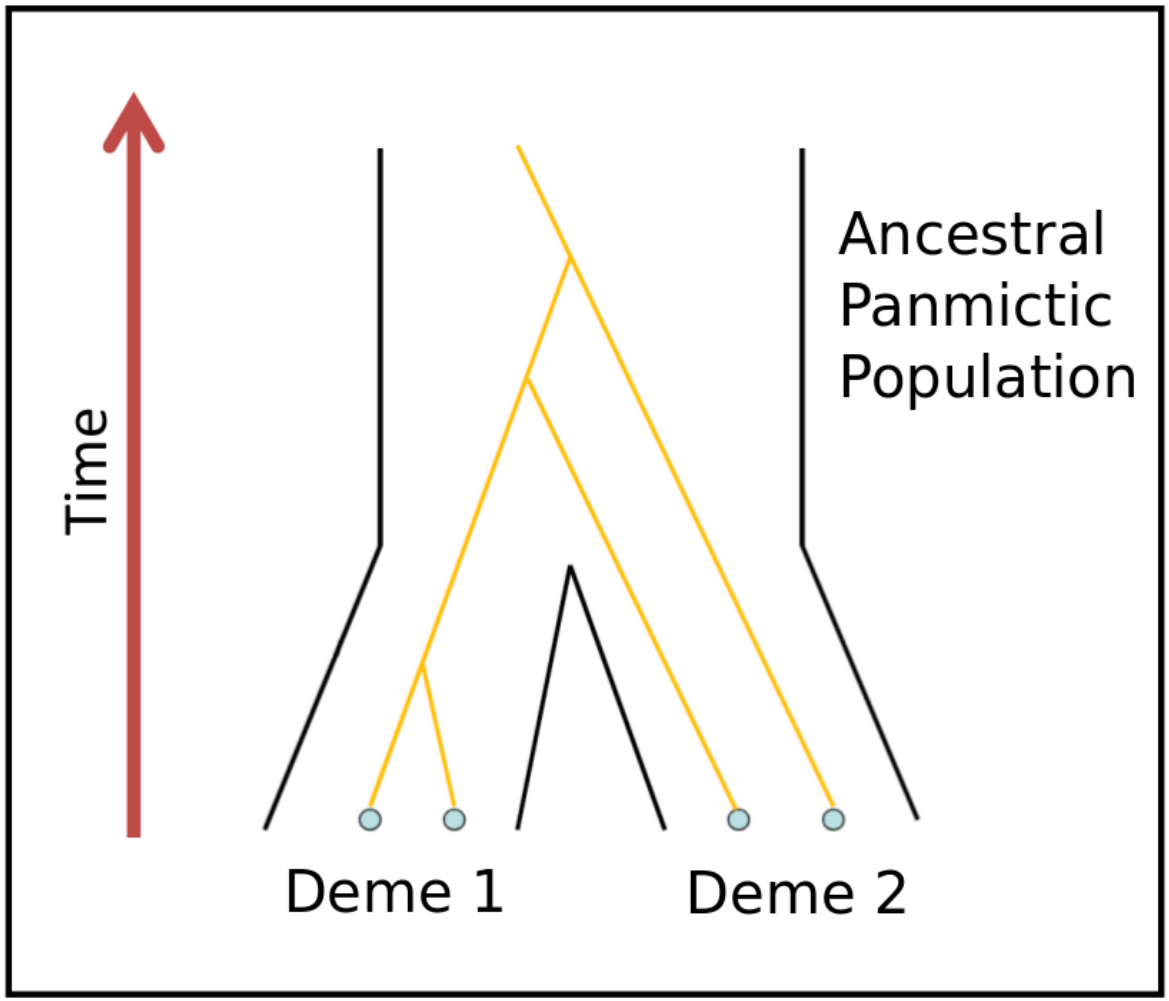
**2)** $P(\text{coalescence, deme i}) = \frac{c_i}{\sum c+b}$, $P(\text{migration, deme i}) = \frac{b_i}{\sum c+b}$

**3)** If event is coalescence, choose 2 lines in deme $i$ and coalesce

**4)** If event is migration, choose line in deme $i$ and destination deme $(\neq i)$ with equal probability among all options

**5)** End if MRCA achieved, else update event rates and return to **1)**

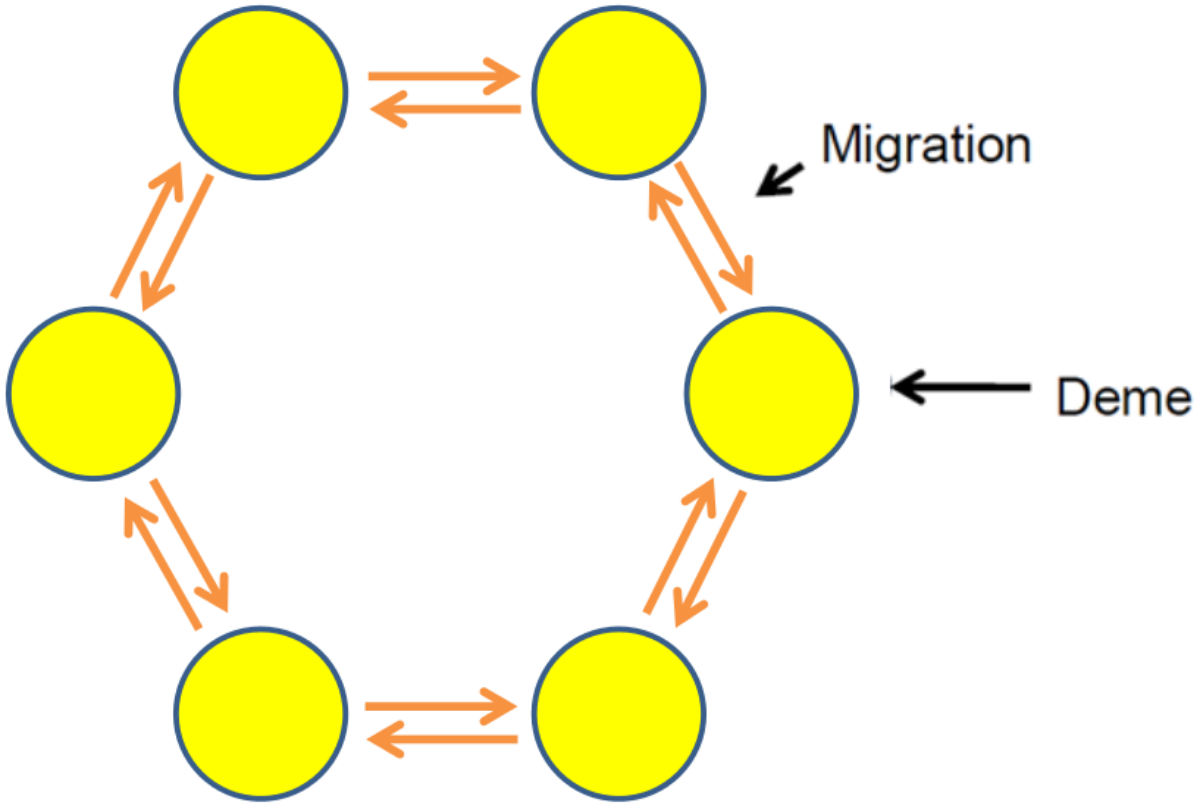
# The divergence (trouser) model




Alternate model where demes merge into panmictic populations in the past

Coalescent probabilities change over time

# The circular stepping stone model




Adds a spatial component to basic island model, migration only occurs between adjacent demes

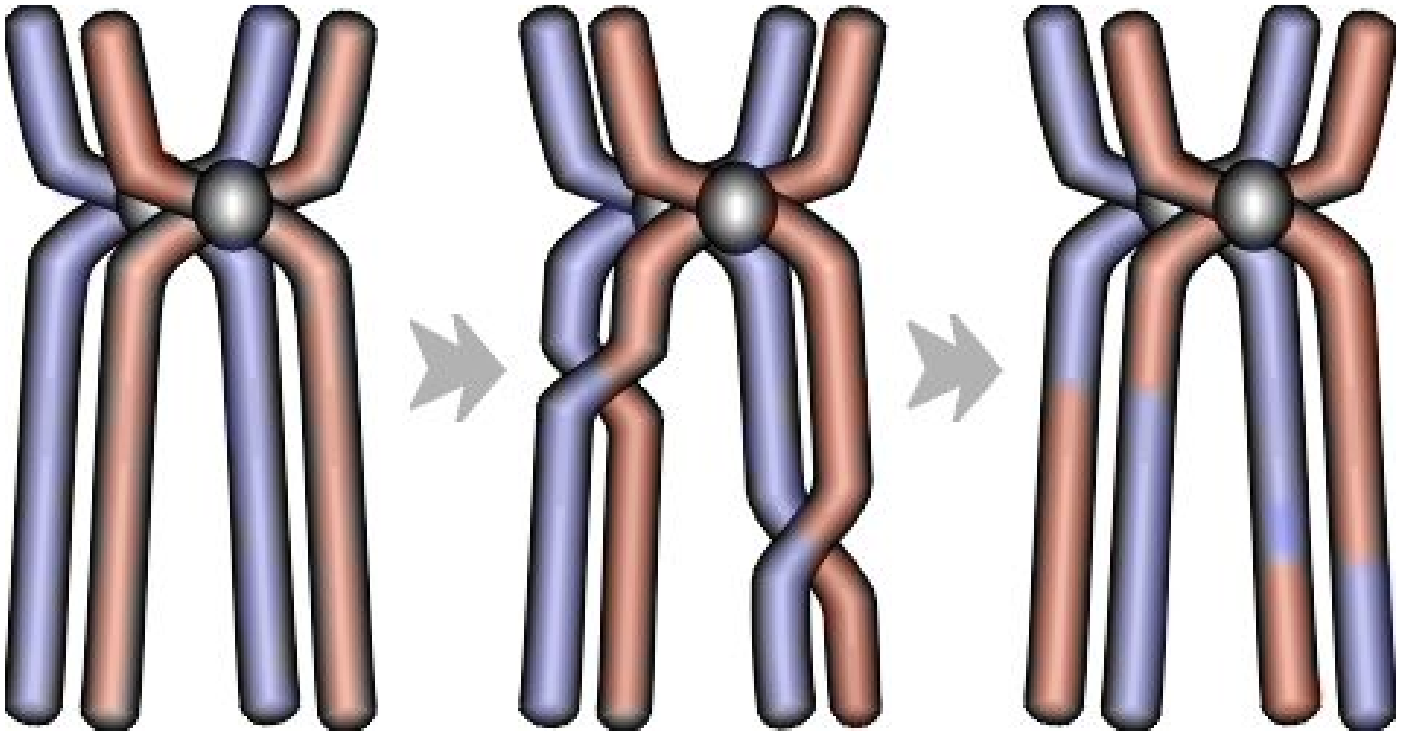
# Final thoughts on population structure

In a structured population coalescence is more likely with members of the same deme, creating clusters of similar samples within a larger overall sample

In simple models all demes behave the same and behavior doesn't change with time, but these assumptions can be relaxed

In a model with a spatial component, closer populations are more genetically similar, a feature often observed in real data

# Recombination in the coalescent

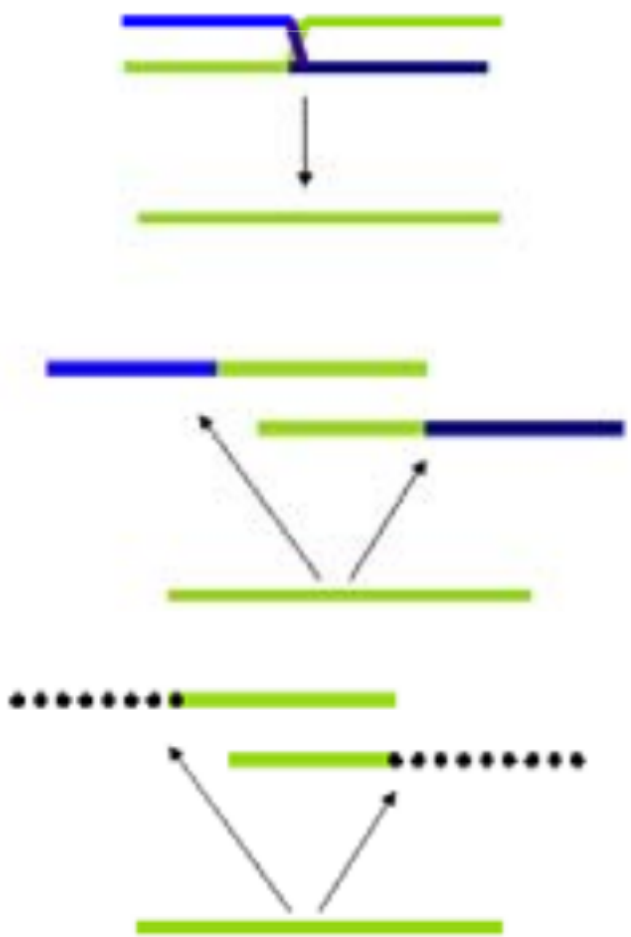


# The impact of a recombination event

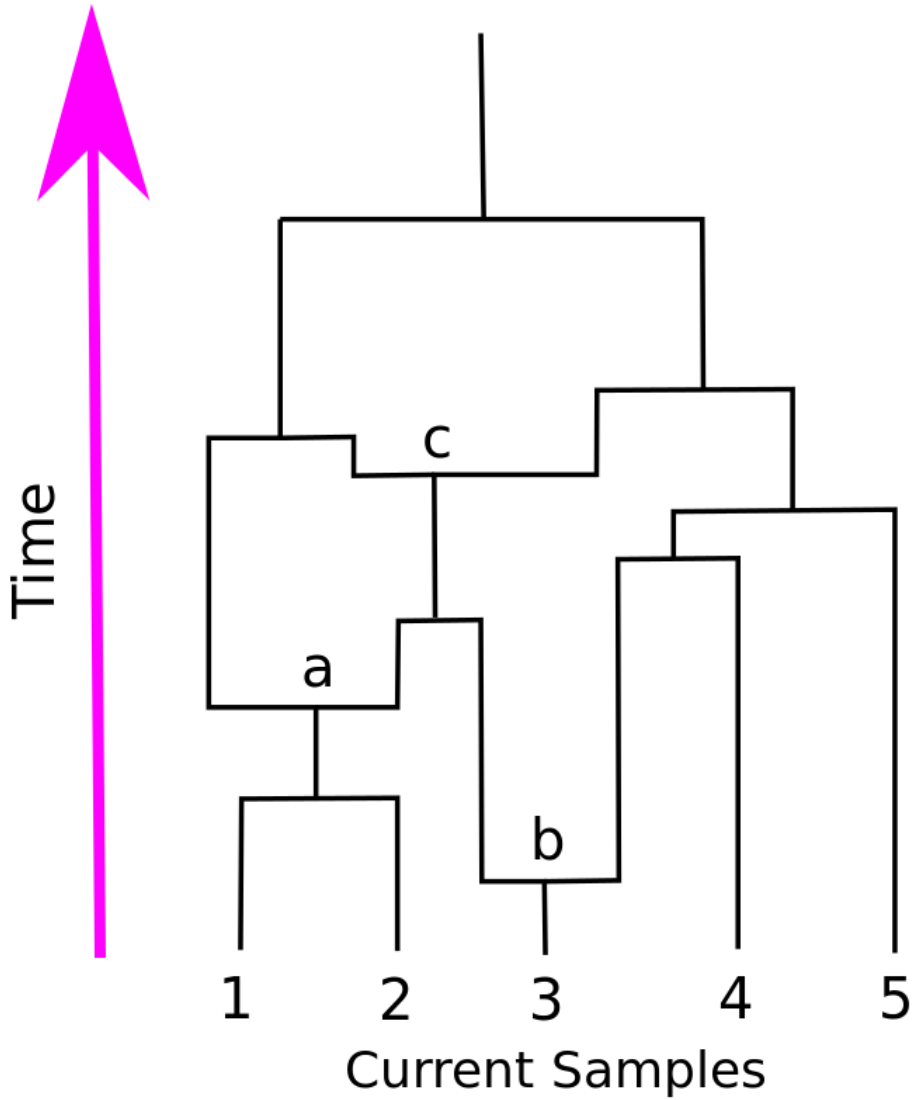
During meiosis, when a cell carries 4 copies of each autosome, matching homologous chromosomes can pass genetic material between them in a recombination event

A resulting chromosome carries both maternal and paternal segments of DNA

Looking backwards in time, the maternal and paternal segments may have different common ancestors and genealogies



# The ARG

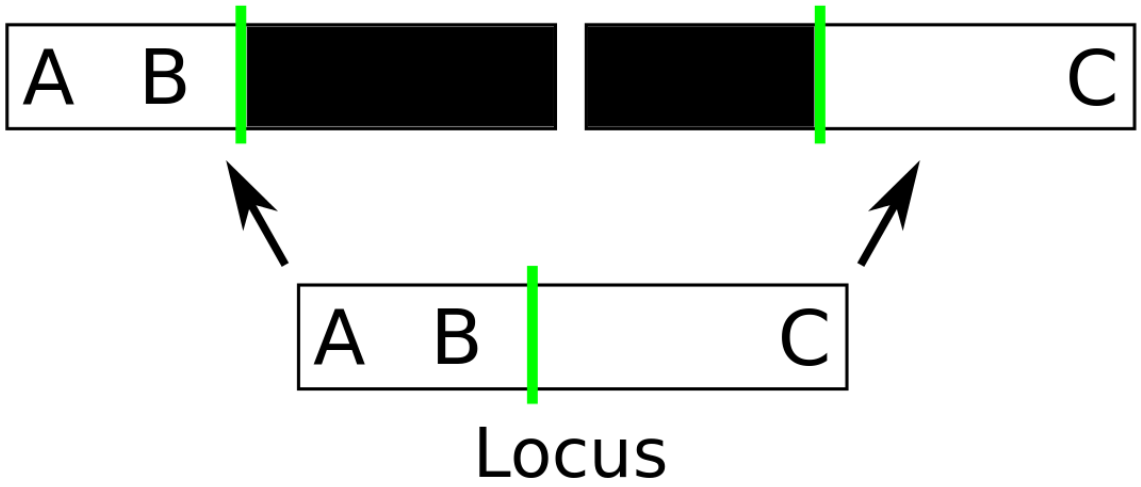



This algorithm constructs an **ancestral recombination graph (ARG)**

a,b, and c represent recombination events

# Recombination events

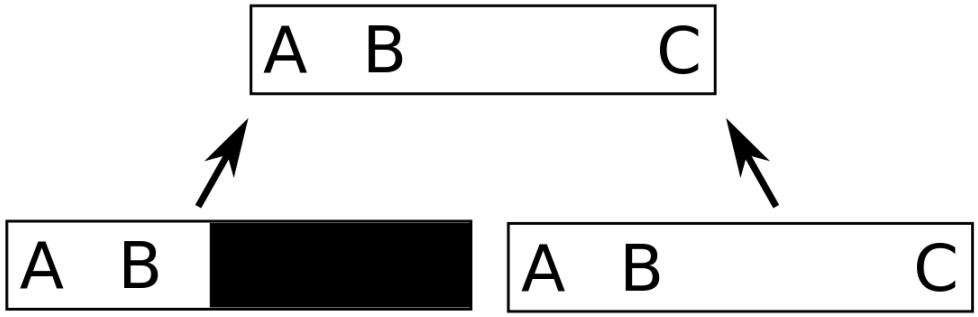
After a recombination event, generated lineages are ancestral to only part of the sequence



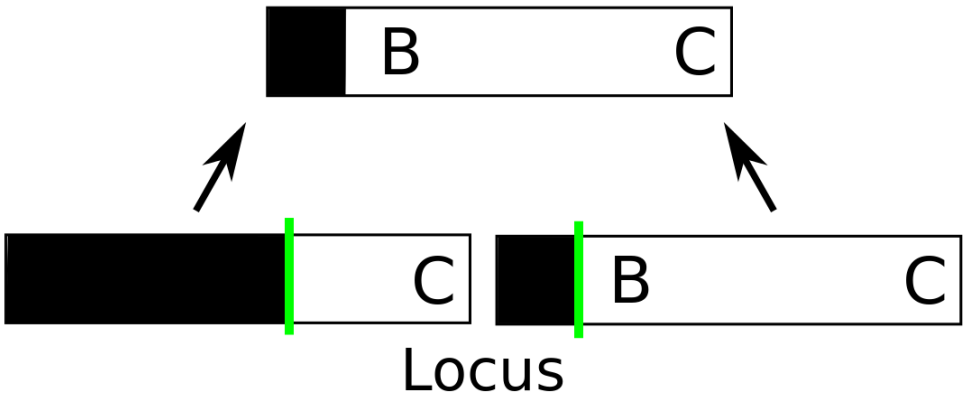

Whether sites along the sequence share ancestry is dependent on their genetic distance

# Coalescent events

Coalescence events combine the sequence covered by either coalescing lineage

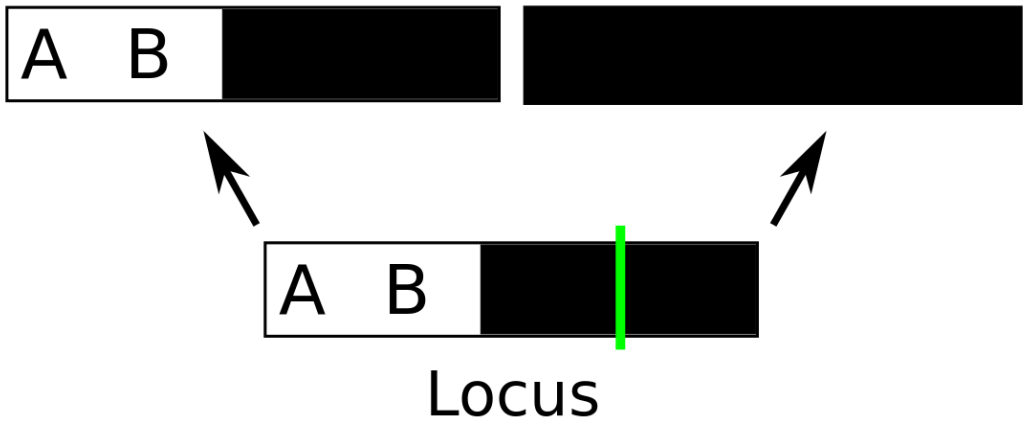



The ancestral sequence contains the union of it's descendant sequences



# Irrelevant recombinations

If a recombination occurs in a segment not ancestral to the current sample, it results in lines that are not ancestral




The rate of relevant coalescent events is actually $c\rho/2$ where $c$ is the proportion of sequence ancestral to the present sample

# Recombination as an exponential process

This should look familiar by now...

Assume the probability that a recombination event occurs in the previous generation at a locus is $r$, then

$$P(\text{No recombination for j generations}) = (1-r)^j$$

A geometric distribution!

Now scale, set $\rho = 4Nr$ and $t = j/2N$ boom! convergence to

$$P(\text{no recombination before time t}) = e^{-\frac{t\rho}{2}}$$

So, recombination events are exponentially distributed with rate $\rho/2$

# Incorporating recombination into our algorithm

**1)** Set $k = n$

**2)** Draw the time to the next event, $T \sim Exp(\frac{k(\rho+k-1)}{2})$

**3)** $P(\text{recombination}) = \frac{\rho}{(\rho+k-1)}$, $P(\text{coalescence}) = \frac{k-1}{(\rho+k-1})$

**4)** If event is a recombination

- **a)** choose upon which of the $k$ line event occurs (uniformly)
- **b)** choose location along haplotype where recombination occurs (uniformly)
- **c)** split lineage at recombination point, $k = k + 1$

**5)** If event is a coalescence

- **a)** choose pair of $k$ lines to coalesce (uniformly)
- **b)** merge lineages, $k = k - 1$

**6)** If $k = 1$ end, otherwise return to **2)**

This process will reach state $k = 1$ with probability 1

# Simulating recombination efficiently

By ignoring irrelevant recombinations we avoid tracking non-ancestral lineages in our ARG, and greatly improve efficiency

Let $c_i$ be the proportion of ancestral sequence present on line $i$. Then when there are $k$ lines remaining the total rate of recombination is $Exp\left(\frac{\rho \sum_{i=1}^{k} c_i}{2}\right)$

Conditional on a recombination event, the probability it occurs along line $i$ is $\frac{c_i}{\sum_{j=1}^{k} c_j}$, and its location is uniformly distributed on the ancestral portion of line $i$

Algorithm now ends when all segment pieces reach their MRCA

# Properties of the ARG

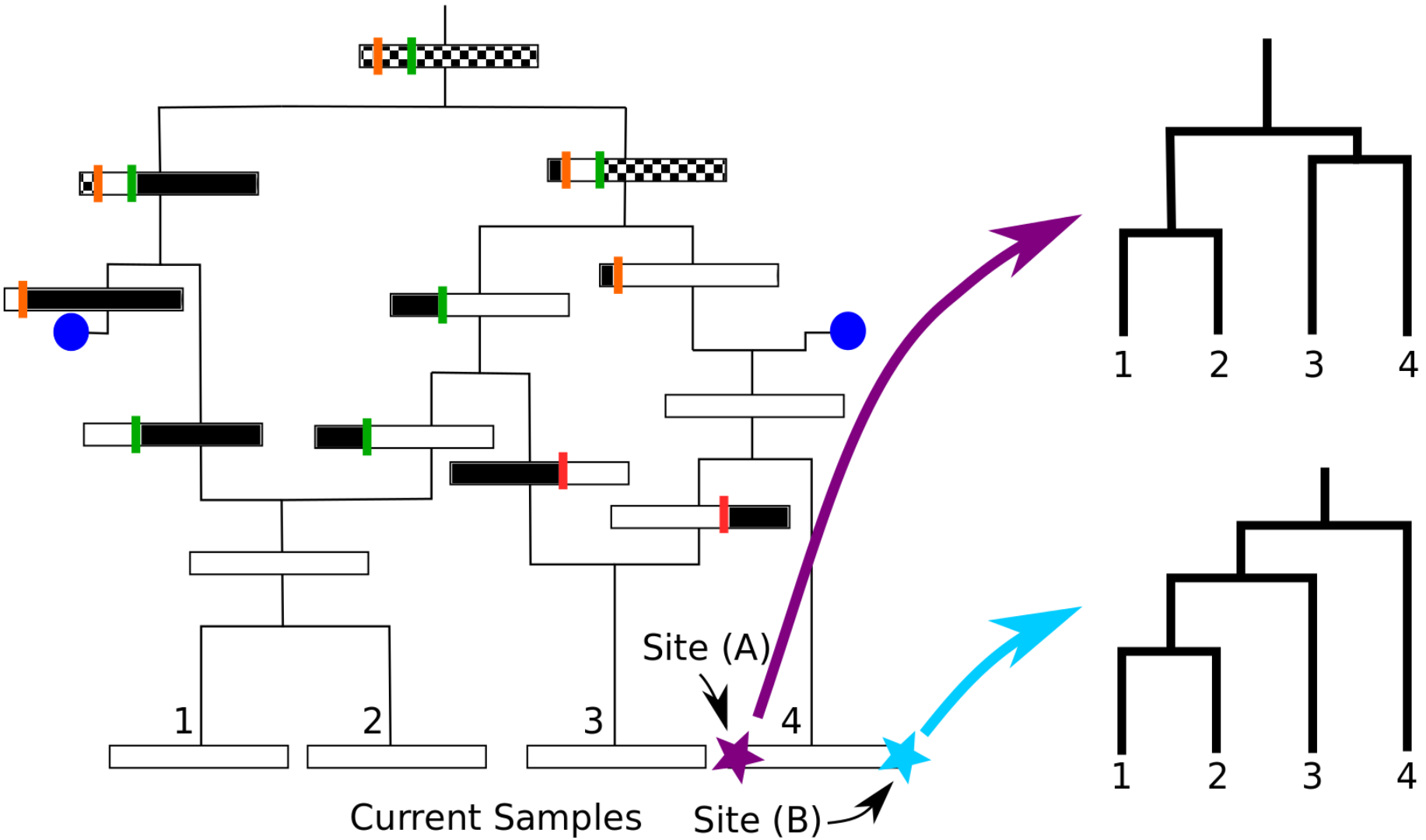
Sequence segments can have different MRCA

Coalescent events reduce the total amount of sequence in ARG, recombination events keep it constant

Every non-recombining sequence segment has a binary tree genealogy that is a sub-graph of the ARG

The space of ARGs is much larger than the (already huge) space of binary tree genealogies

# A more complete visualization of the ARG



# Summary

The basic coalescent makes restrictive assumptions, often violated in real populations

Methods for incorporating variable population sizes, population structure, and recombination make the model more complex, but also more robust

By modeling each process with exponential R.V.s they all "play well" with each other, and although addressed here in isolation, they can all be modeled simultaneously