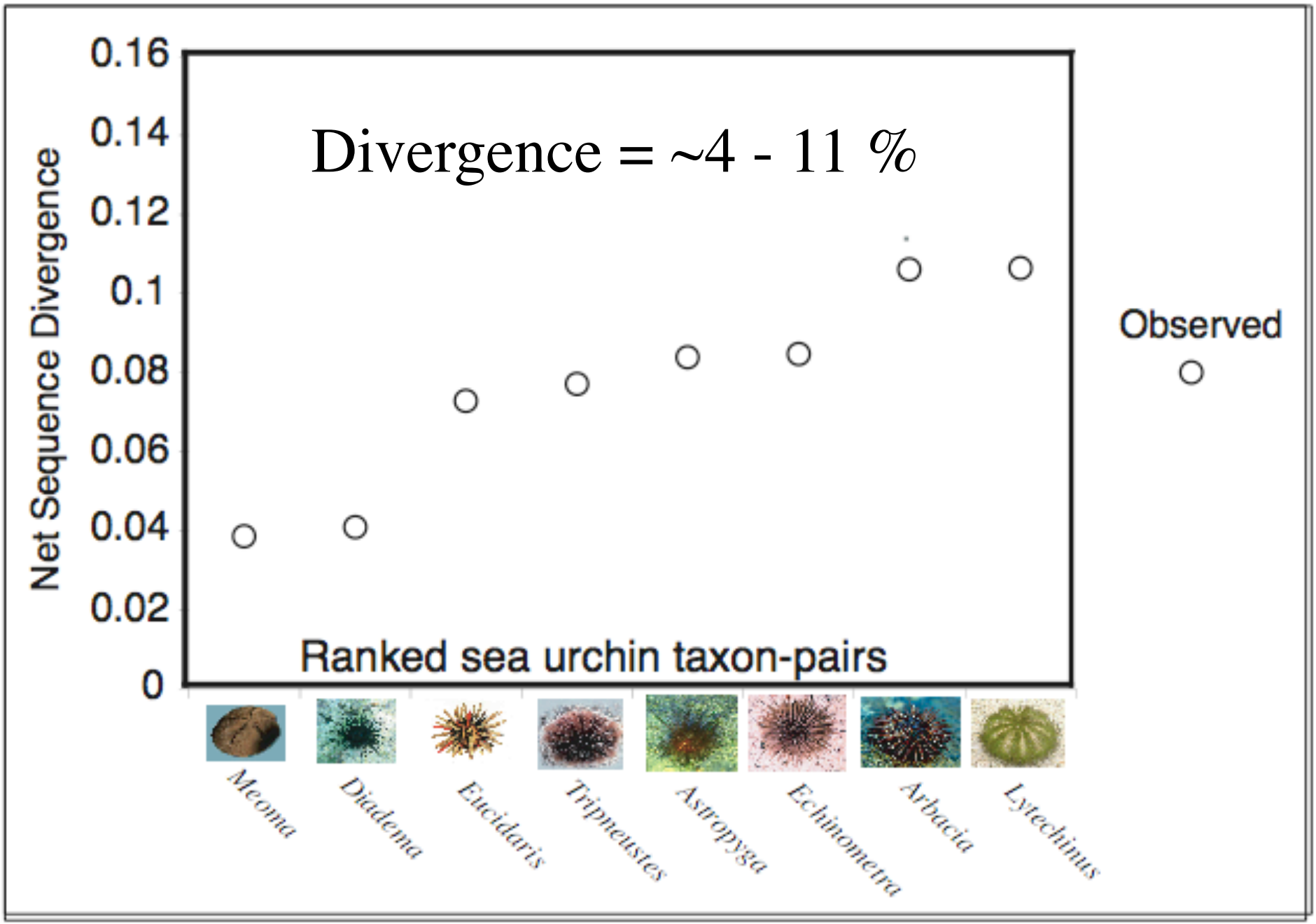
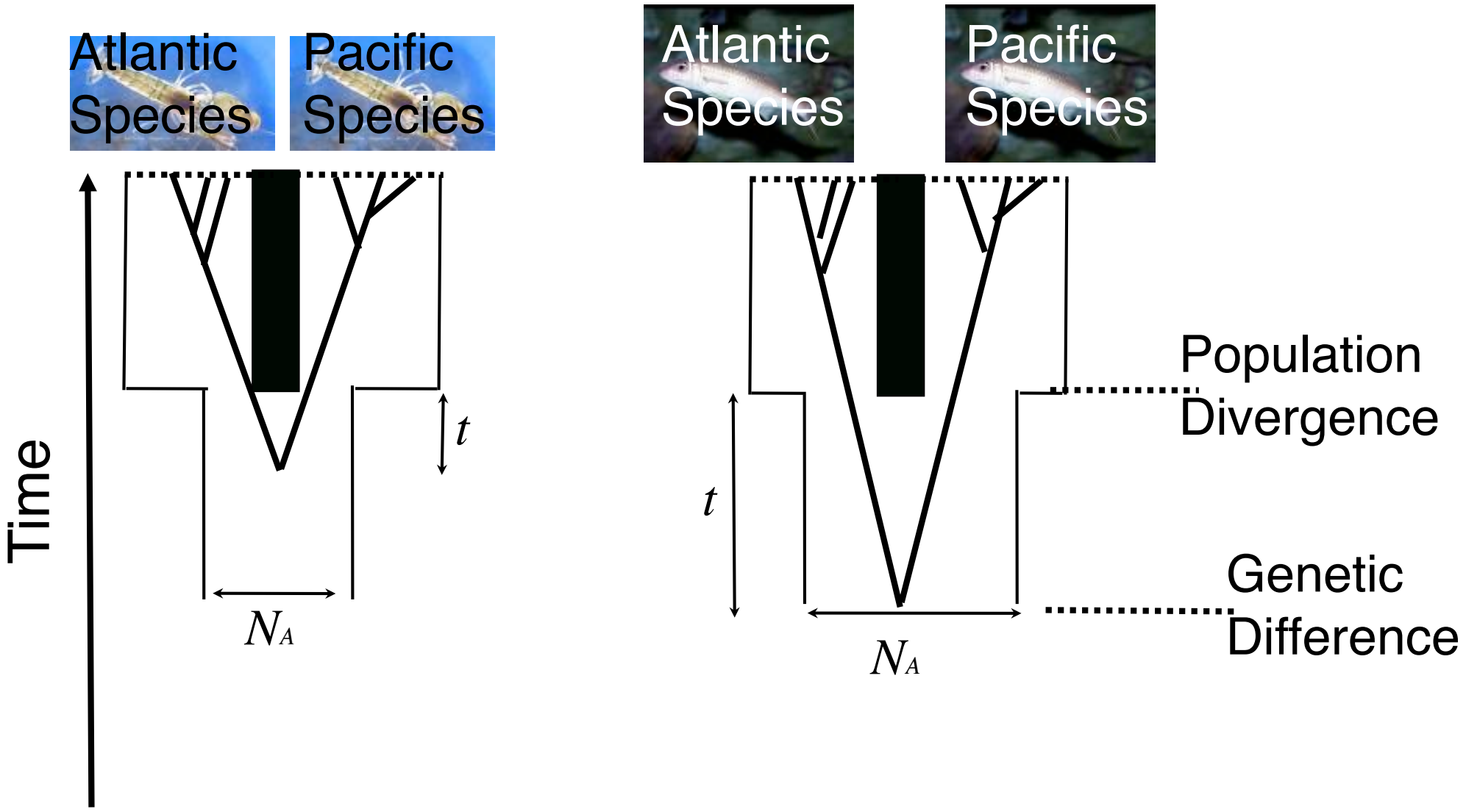
Population Gen Simulation and Inference

Lecture 3:

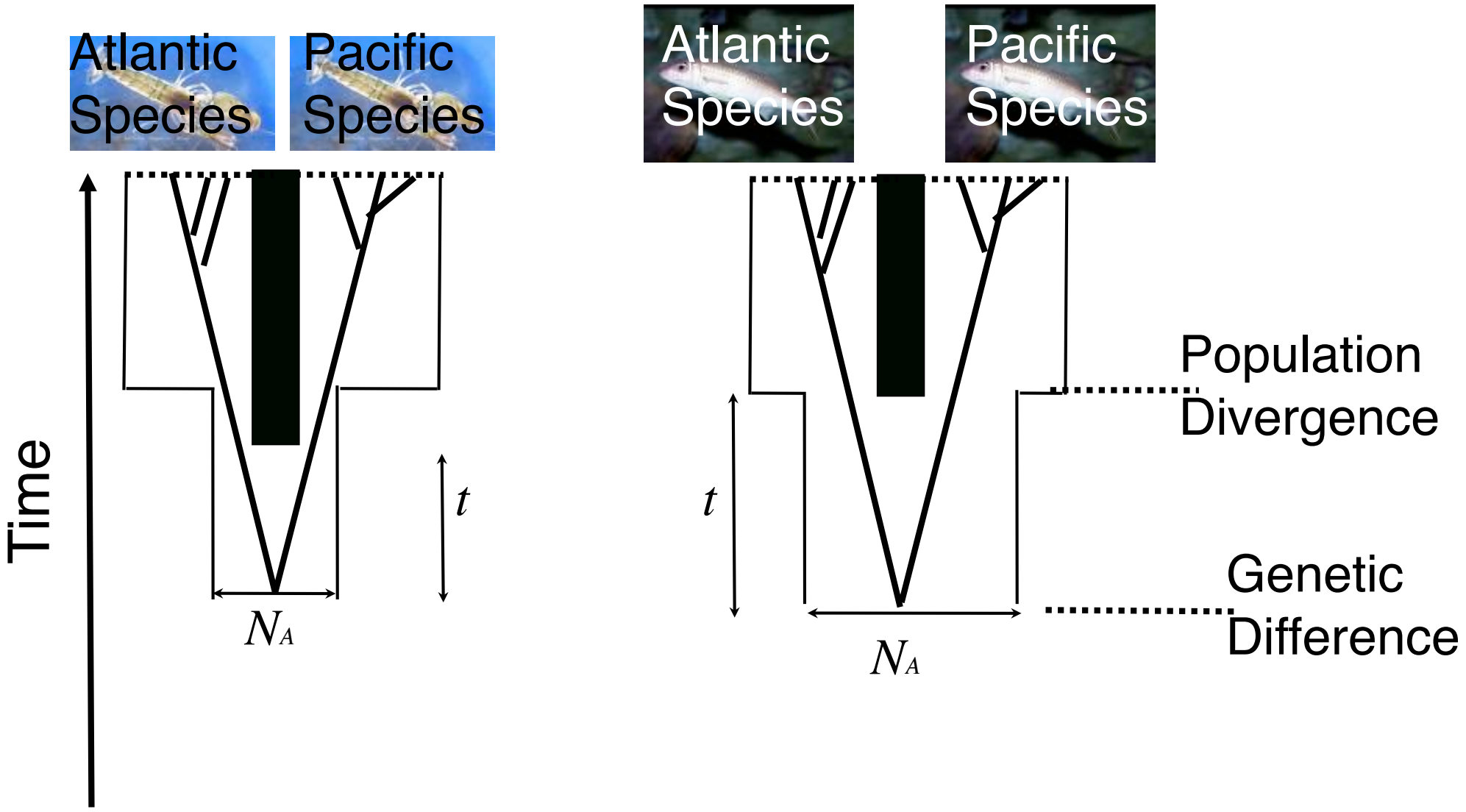
The Coalescent


Divergence = ~4 - 11 %
Net Sequence Divergence
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
Ranked sea urchin taxon-pairs
Observed
Meoma
Diadema
Eucidaris
Tripneustes
Astropyga
Echinometra
Arbacia
Lytechinus

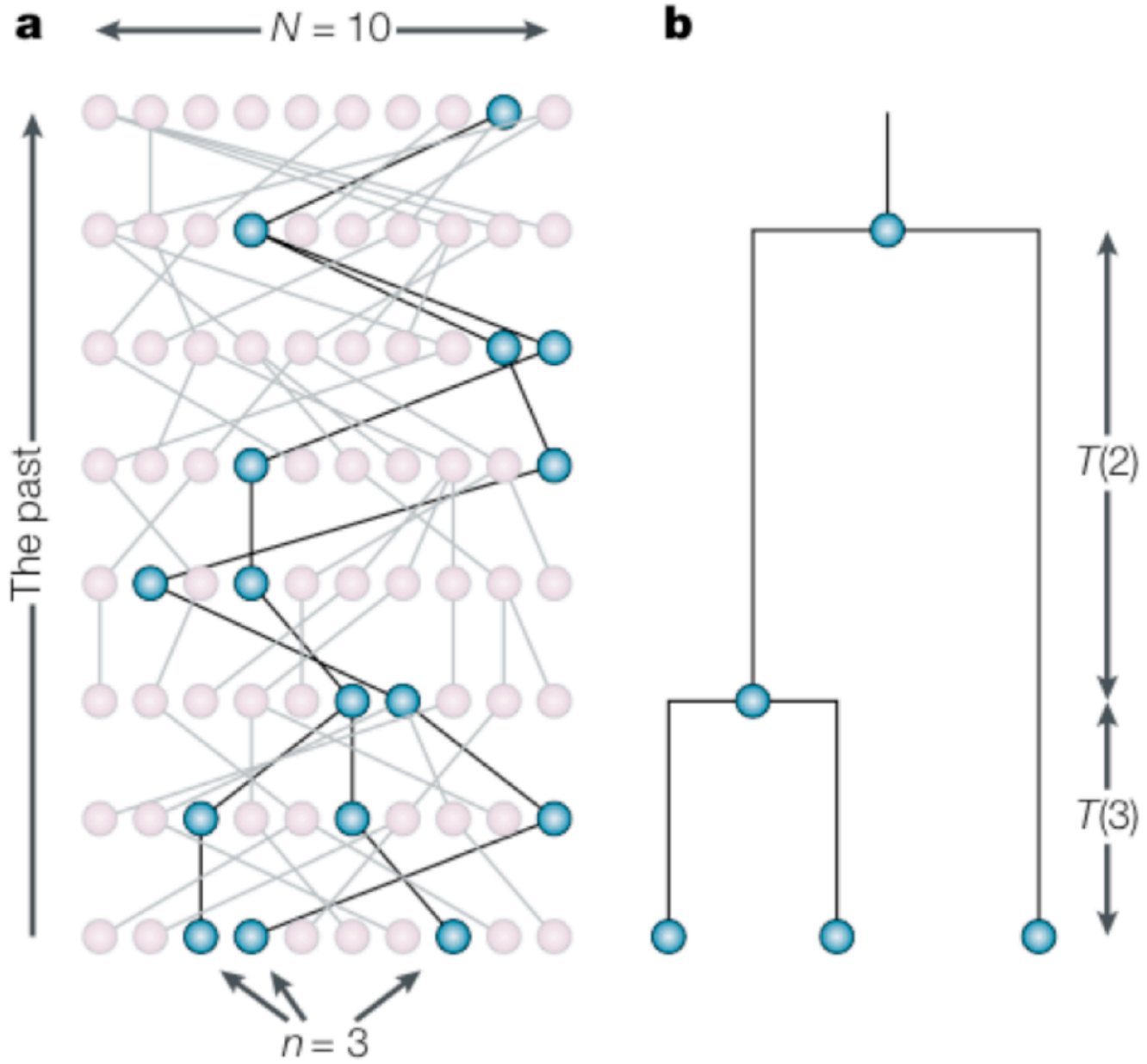
coalescent theory predicts noise



genetic data are noisy



Example of a gene genealogy



Define

1. Gene genealogy
2. Coalescent event
3. Most Recent Common Ancestor (MRCA)
4. Time to MRCA (or TMRCA)

haploid vs Diploid models

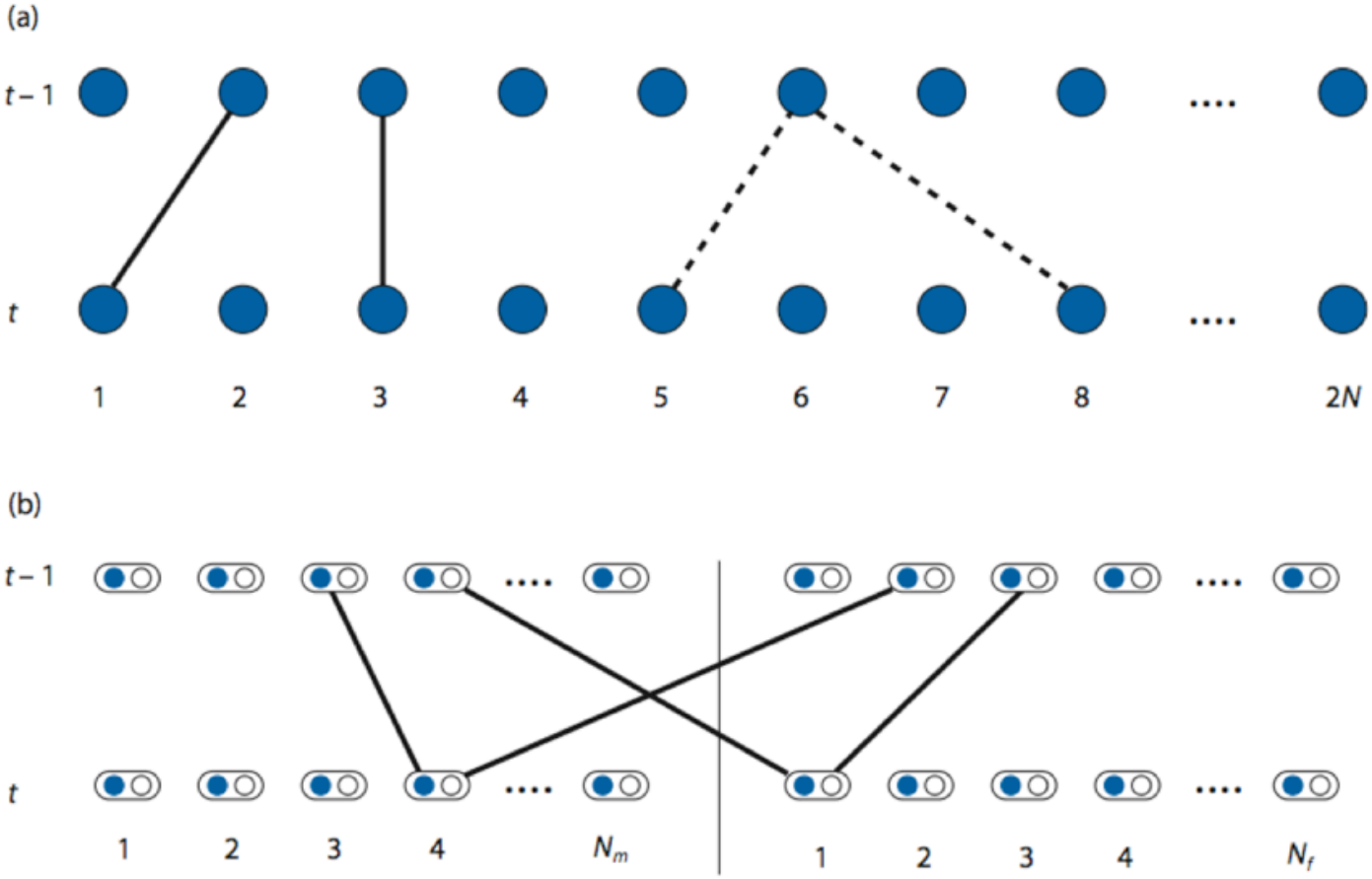


Figure 3.23 Haploid (a) and diploid (b) reproduction in the context of coalescent events. In a haploid population, the probability

NOTE: The haploid model is often used as an approximation for the diploid model. To do so, the number of haplotypes is set equal to the number of chromosomes in the diploid model.

coalescent probabilities per generation

Question

What is the probability that 2 sampled haplotypes have the same parent in the previous generation?

coalescent probabilities per generation

Question

What is the probability that 2 sampled haplotypes have the same parent in the previous generation?

Related question: What is the probability you and I have the same birthday?

coalescent probabilities per generation

Question

What is the probability that 2 sampled haplotypes have the same parent in the previous generation?

$$Pr(2 \text{ coalesce to one}) = \frac{1}{\#\text{ of haplotypes}}$$

so for diploids:

$$Pr(2 \text{ coalesce to one}) = \frac{1}{2N}$$

coalescent probabilities per generation

Question

What is the probability that 2 sampled haplotypes have the same parent in the previous generation?

$$Pr(2 \text{ coalesce to one}) = \frac{1}{\#\text{ of haplotypes}}$$

so for diploids:

$$Pr(2 \text{ coalesce to one}) = \frac{1}{2N}$$

Question

What is the probability 3 sampled genes coalesce in the same generation?

coalescent probabilities per generation

Question

What is the probability that 2 sampled haplotypes have the same parent in the previous generation?

$$Pr(2 \text{ coalesce to one}) = \frac{1}{\#\text{ of haplotypes}}$$

so for diploids:

$$Pr(2 \text{ coalesce to one}) = \frac{1}{2N}$$

Question

What is the probability 3 sampled genes coalesce in the same generation?

$$Pr(3 \text{ coalesce to one}) = \frac{1}{2N}\frac{1}{2N} = \frac{1}{4N^2}$$

coalescent probabilities

Implications

- If N is large and the number of lineages under consideration is small, the probability of 2 coalescent events per generation is very small relative to single coalescent events per generation.
- Suggests the approximation that we only consider *pairwise* (i.e. 2-way) coalescent as being possible.
- Thus, if there are k lineages to consider, there are $\binom{k}{2}$ possible lineages we can join, the total probability of a coalescent event in a generation is:

$$\binom{k}{2} \times \frac{1}{2N}$$

and the mean waiting time is then the reciprocal $\frac{2N}{1}$

$$\binom{k}{2} = \frac{k!}{2!(k-2)!} * \frac{1}{2N}$$

number of lineages

$$\binom{k}{2} = \frac{k!}{2!(k-2)!} * \frac{1}{2N}$$

$$\binom{2}{2} = \frac{2!}{2!(2-2)!} * \frac{1}{2N} = \frac{1}{2N}$$

number of lineages

mean waiting time = $\frac{2N}{1}$

$$\binom{k}{2} = \frac{k!}{2!(k-2)!} * \frac{1}{2N}$$

$$\binom{2}{2} = \frac{2!}{2!(2-2)!} * \frac{1}{2N} = \frac{1}{2N}$$

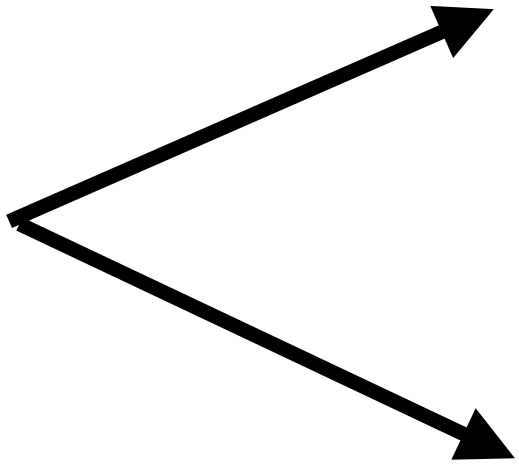
mean waiting time = $2N$

$$\binom{6}{2} = \frac{6!}{2!(6-2)!} * \frac{1}{2N} = \frac{15}{2N}$$

mean waiting time = $\frac{2N}{15}$

$$\binom{k}{2} = \frac{k!}{2!(k-2)!} * \frac{1}{2N}$$

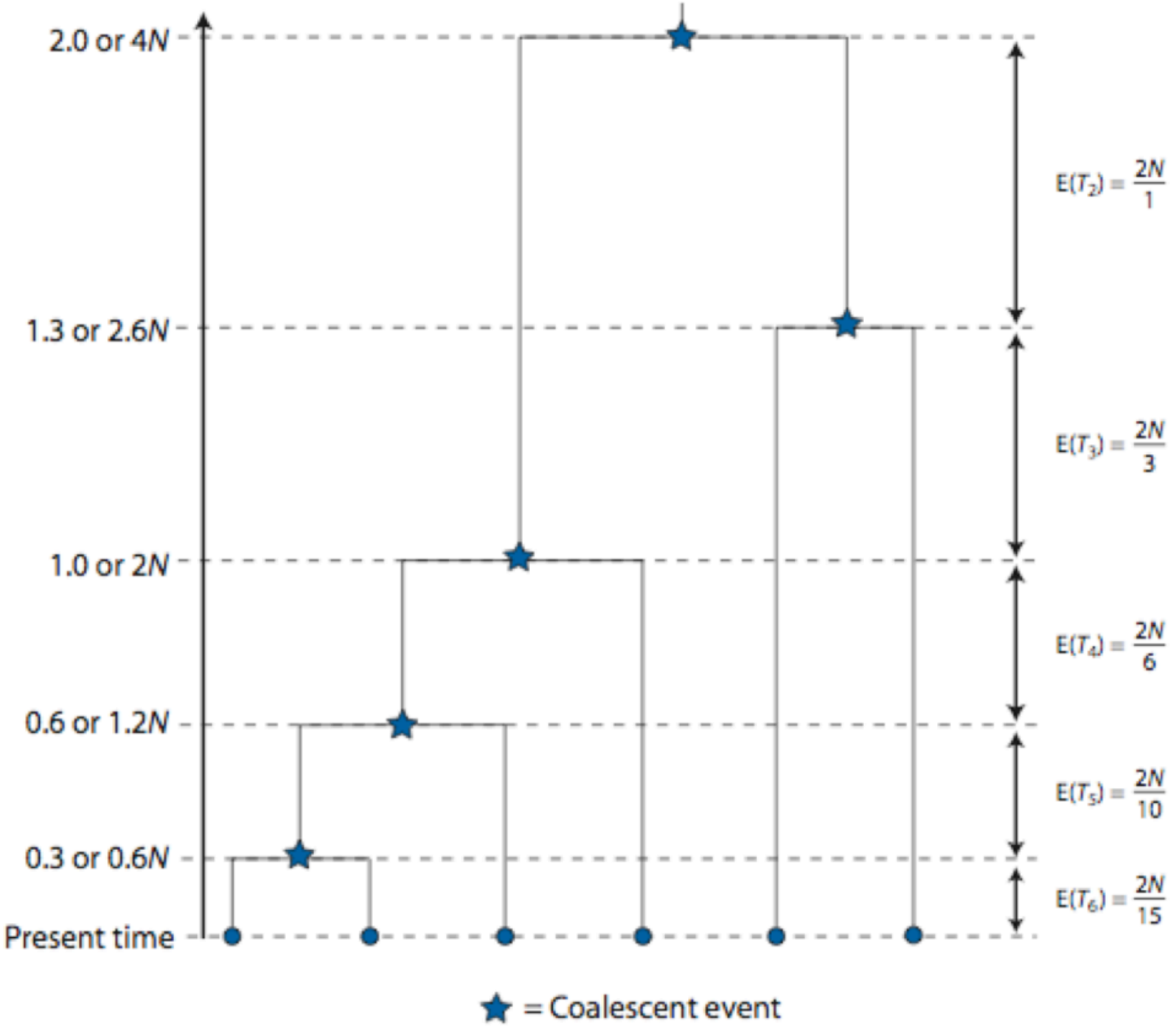
can be approximated by the exponential distribution



mean waiting time = $2N$

mean waiting time = $\frac{2N}{15}$

The *expected* rate of coalescent events



times between coalescent events are smaller when th
of lineages (for instance, toward the tips of the tree)
the last two lineages remain. Using the exponentia
now write down five easy steps for generating gene
size n:

the basic coalescent algorithm

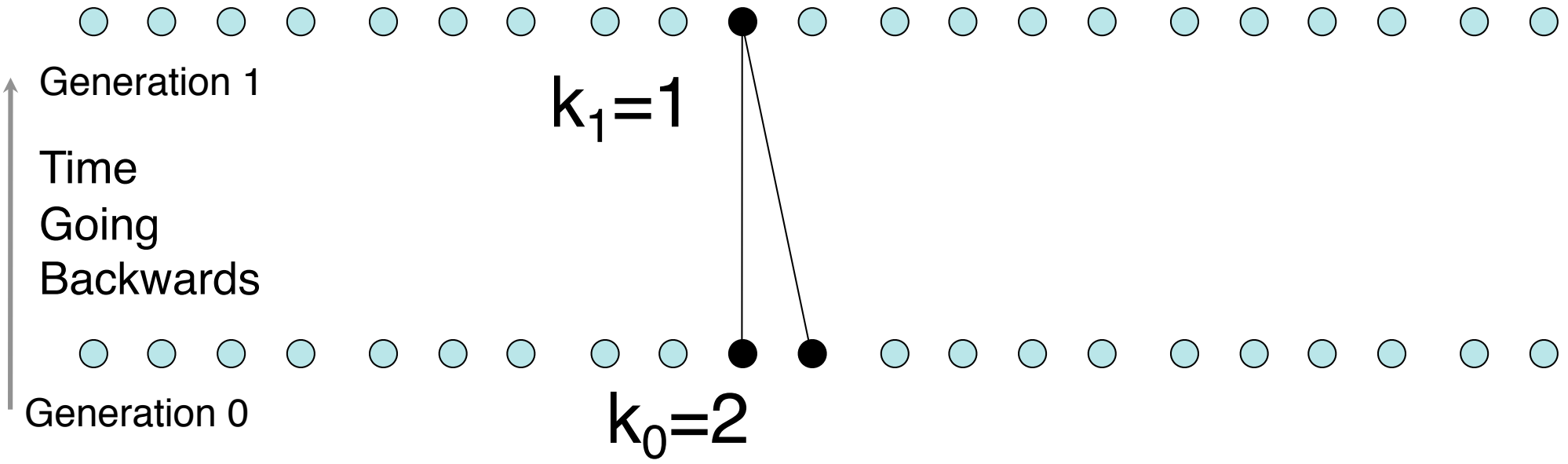
1. Start with $i = n$ chromosomes.
2. Choose a time until the next coalescence from an exponential distribution with parameter $x = i(i-1)/2$.
3. Choose two chromosomes at random to coalesce.
4. Merge the two lineages that were chosen and set $i \rightarrow i - 1$.
5. If $i > 1$, go to step 2; if not, stop.

As an example, **FIGURE 6.2** shows one possible genealogy generated by following the above steps for $n = 5$. Starting with $i = n = 5$, the process

from Hahn page 115

probability of 2 gene copies descending from a single copy …

$$\Pr(k_1=1 \mid k_0=2) = 1/2\mathbf{N}$$

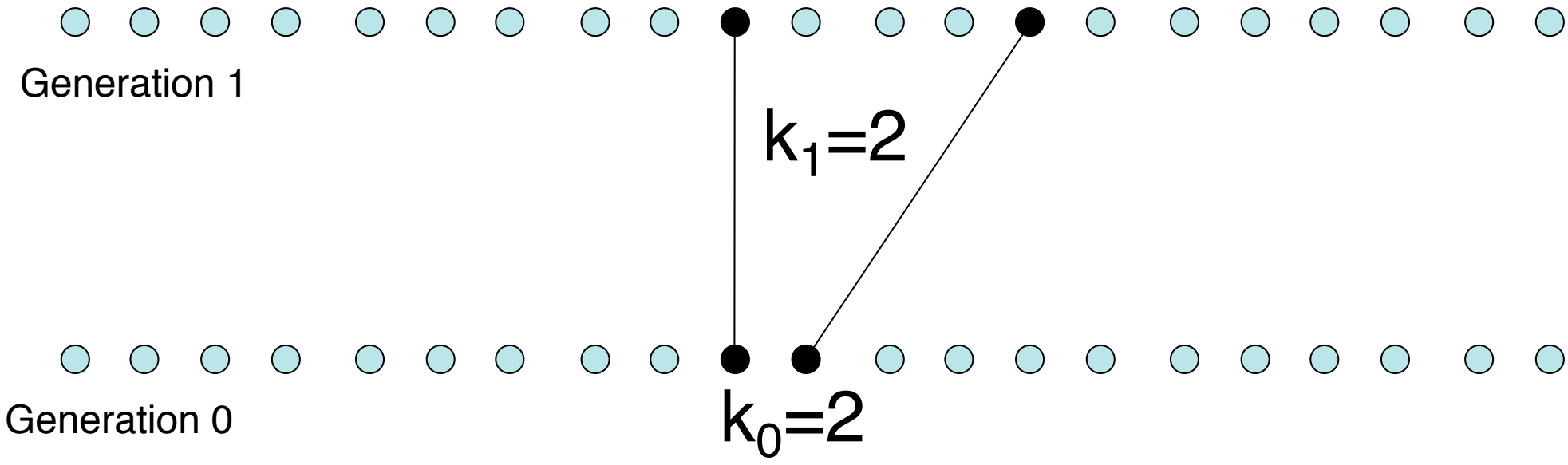


Pop size = **N** (2 gene copies per individual)

Therefore, probability of 2 gene copies descending from different copies

$$\Pr(k_1=2 \mid k_0=2) = 1-1/2\mathbf{N}$$

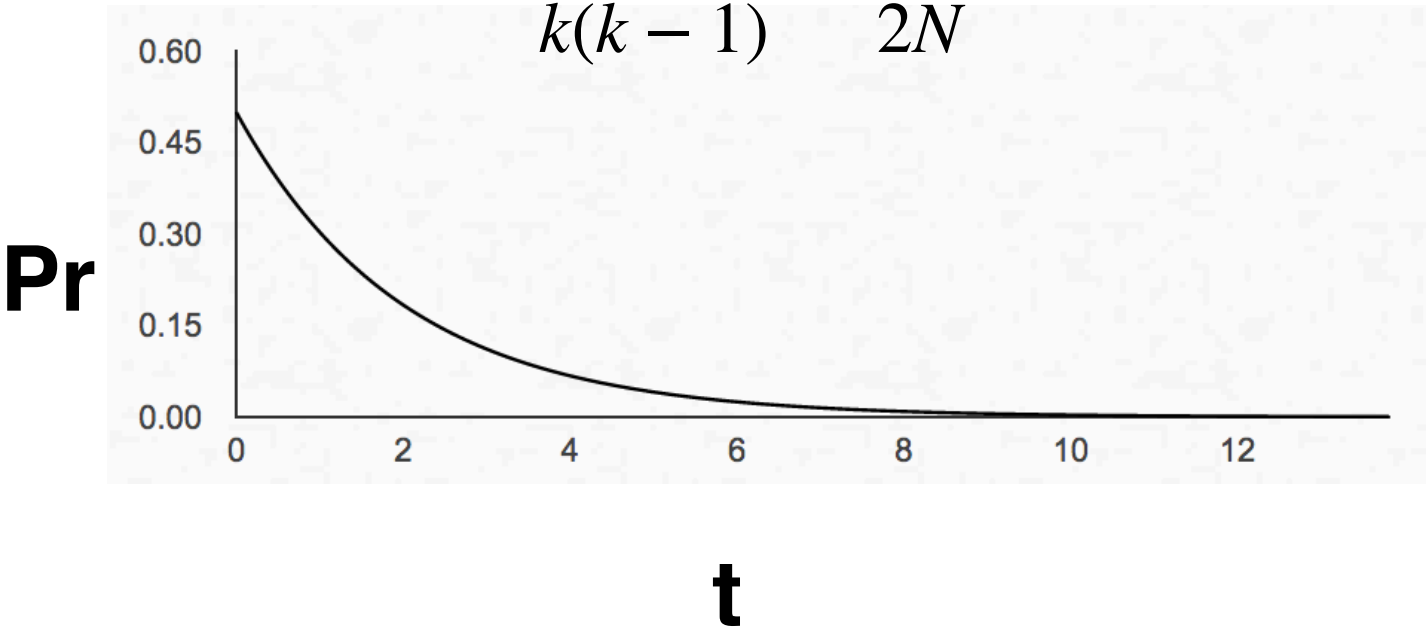
(law of probability)

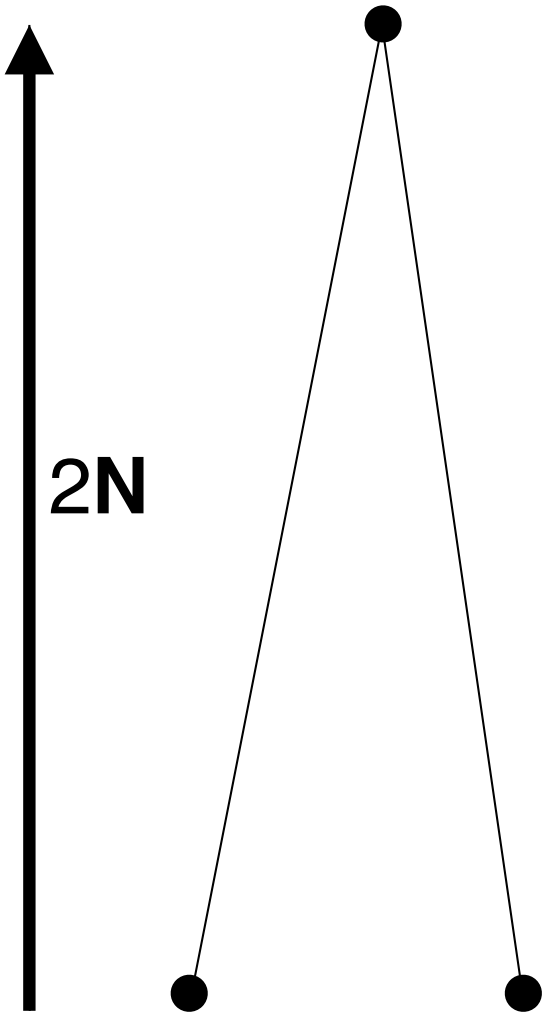


Pop size = **N** (2 gene copies per individual)

Number of generations since 2 randomly picked gene copies have common ancestor approximates an exponential distribution with a mean of 2**N (**run of the mill Poisson process)

$$\frac{4N}{k(k-1)} = \frac{4N}{2N} = 2N$$



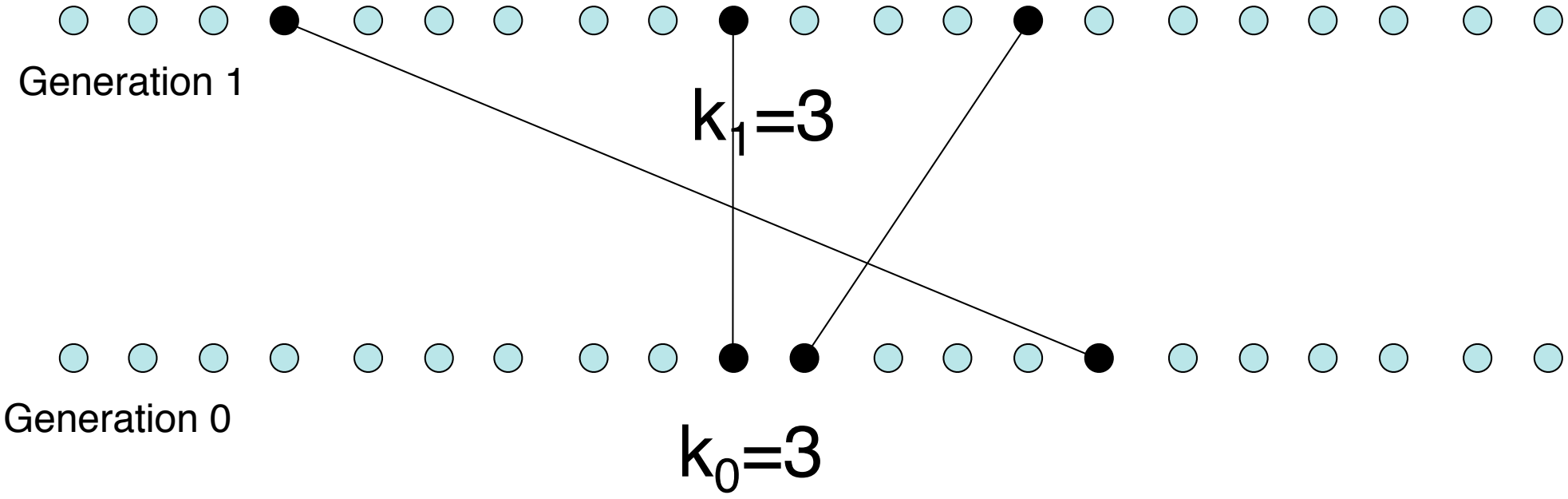


$Pr(k_1=3 \mid k_0=3)$ is

probability of 2 copies having different parents (1-1/2**N**), multiplied by probability of the 3rd having different parents than the other two (1-2/2**N**)

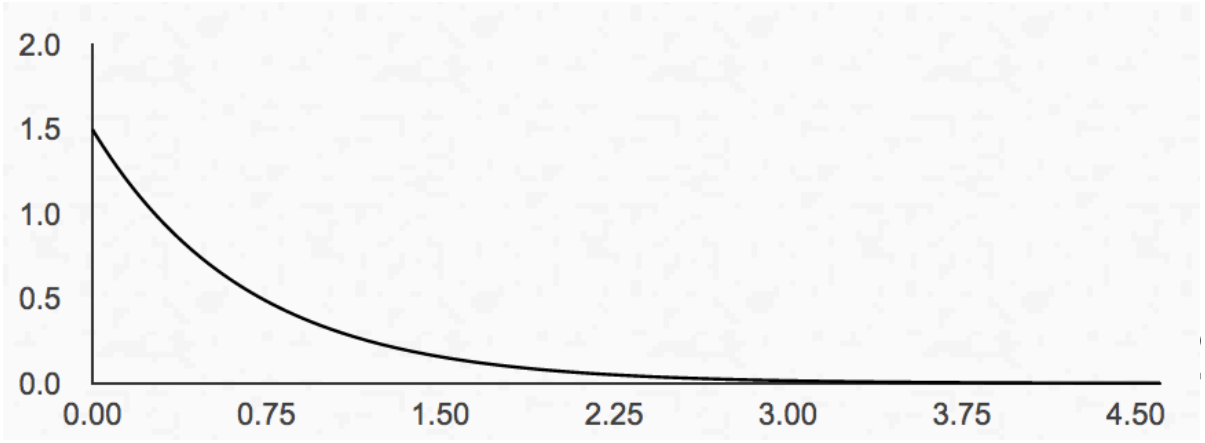
$$Pr(k_1=3 \mid k_0=3) = (1-1/2\mathbf{N})x(1-2/2\mathbf{N})$$

$$= 1-3/2\mathbf{N} + 2/2\mathbf{N}^2 \text{ (can ignore 2nd b/c N is large)}$$

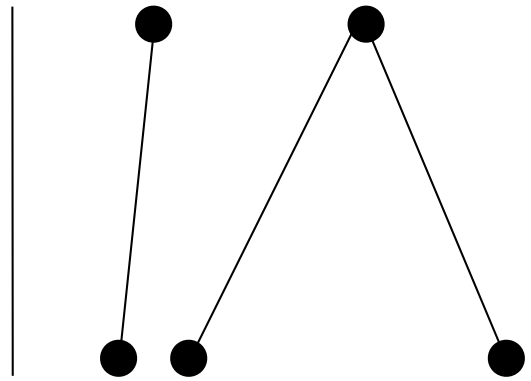


Waiting time (t) for k_0= 3 copies having k_t=2 ancestors has an exponential distribution with a mean of 2**N/3** generations

$$\frac{4N}{k(k-1)} = \frac{4N}{3(3-1)} = \frac{2N}{3}$$



N/3



Simulating The Coalescent

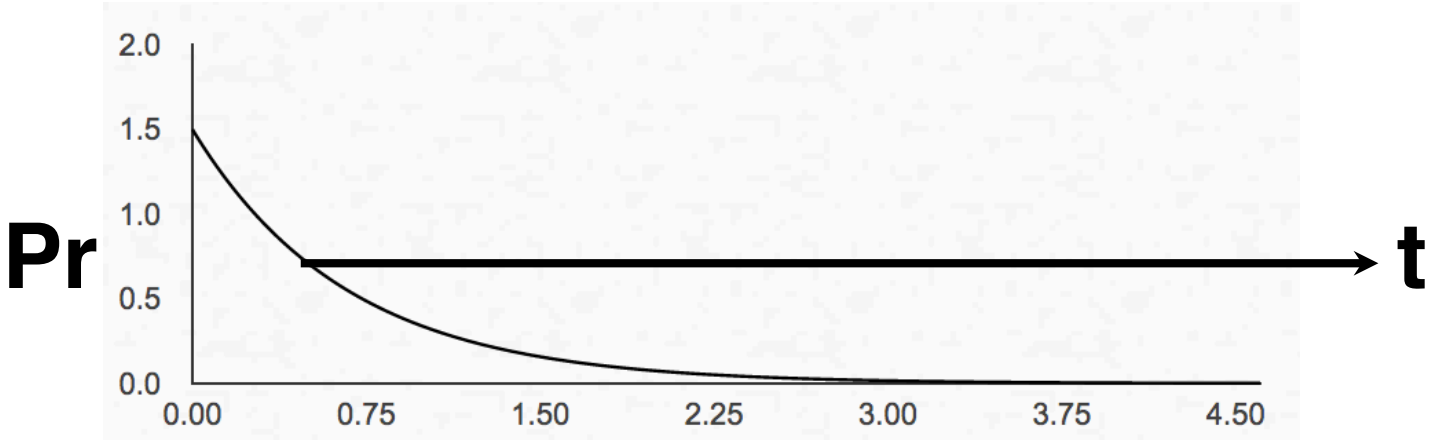
k = sample size;

N = population size

t (time) = 0; go backwards;

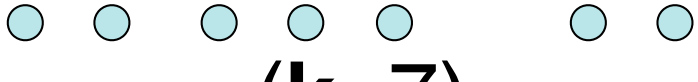
1. randomly draw **t** from exponential

with mean = $\frac{4N}{k(k-1)}$



Simulating The Coalescent

k = sample size

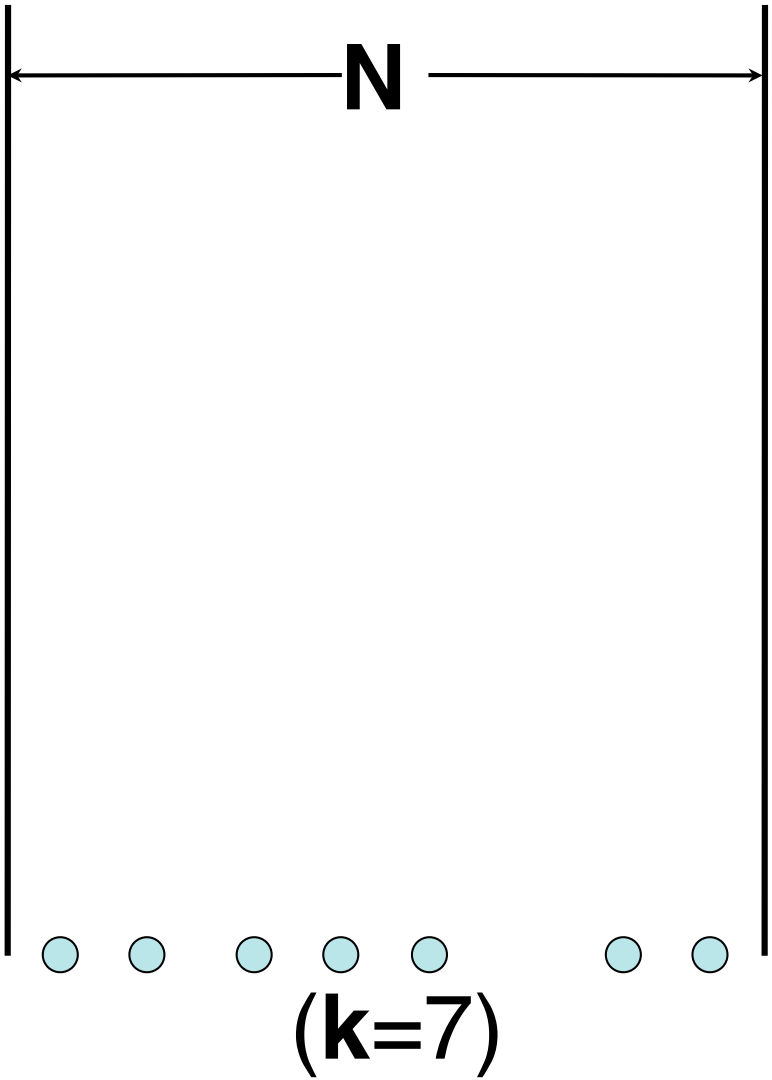


(**k**=7)

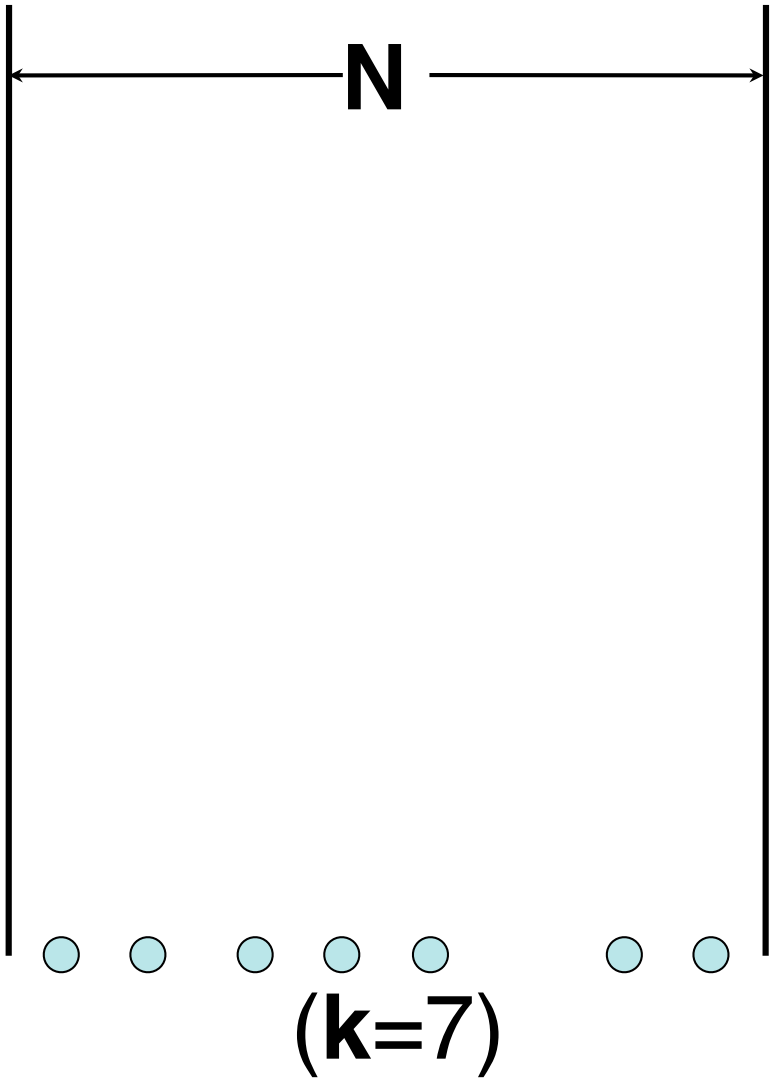
Simulating The Coalescent

N = population size

k = sample size



Simulating The Coalescent



T = time (go backwards)

N = population size

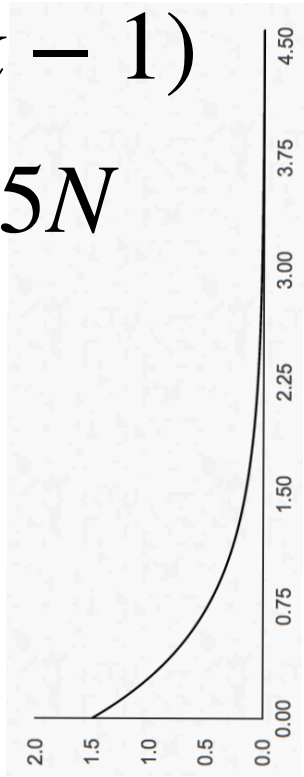
k = sample size

Simulating The Coalescent

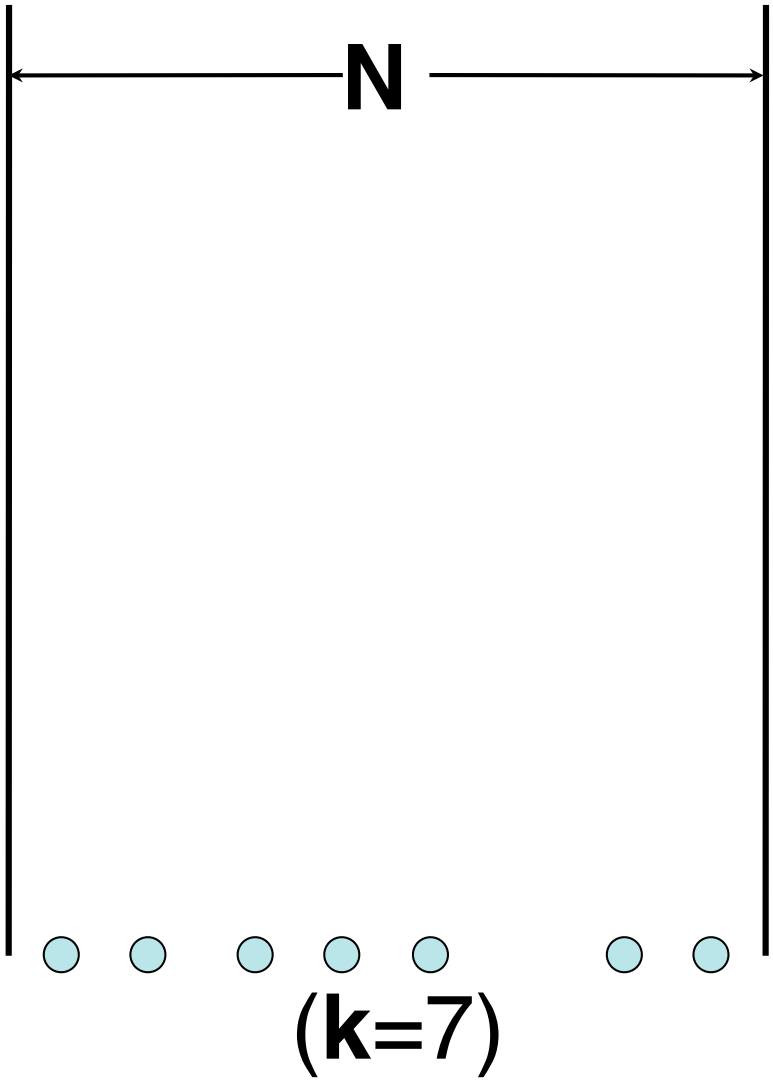
randomly draw **t** from exponential with

$$\text{mean} = \frac{4N}{k(k-1)}$$

$$\frac{4N}{42} = 0.095N$$



Pr

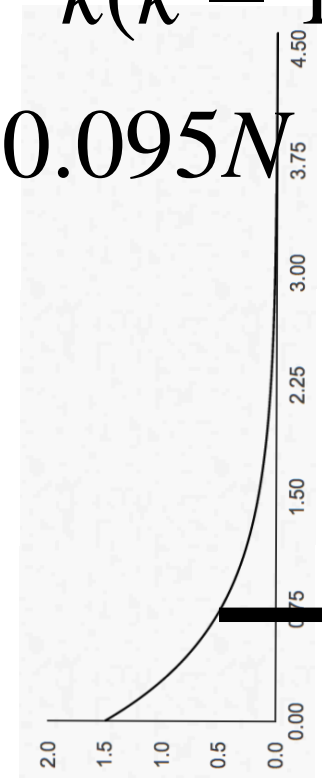


Simulating The Coalescent

randomly draw **t** from exponential with

$$\text{mean} = \frac{4N}{k(k-1)}$$

$$\frac{4N}{42} = 0.095N$$



Pr

N

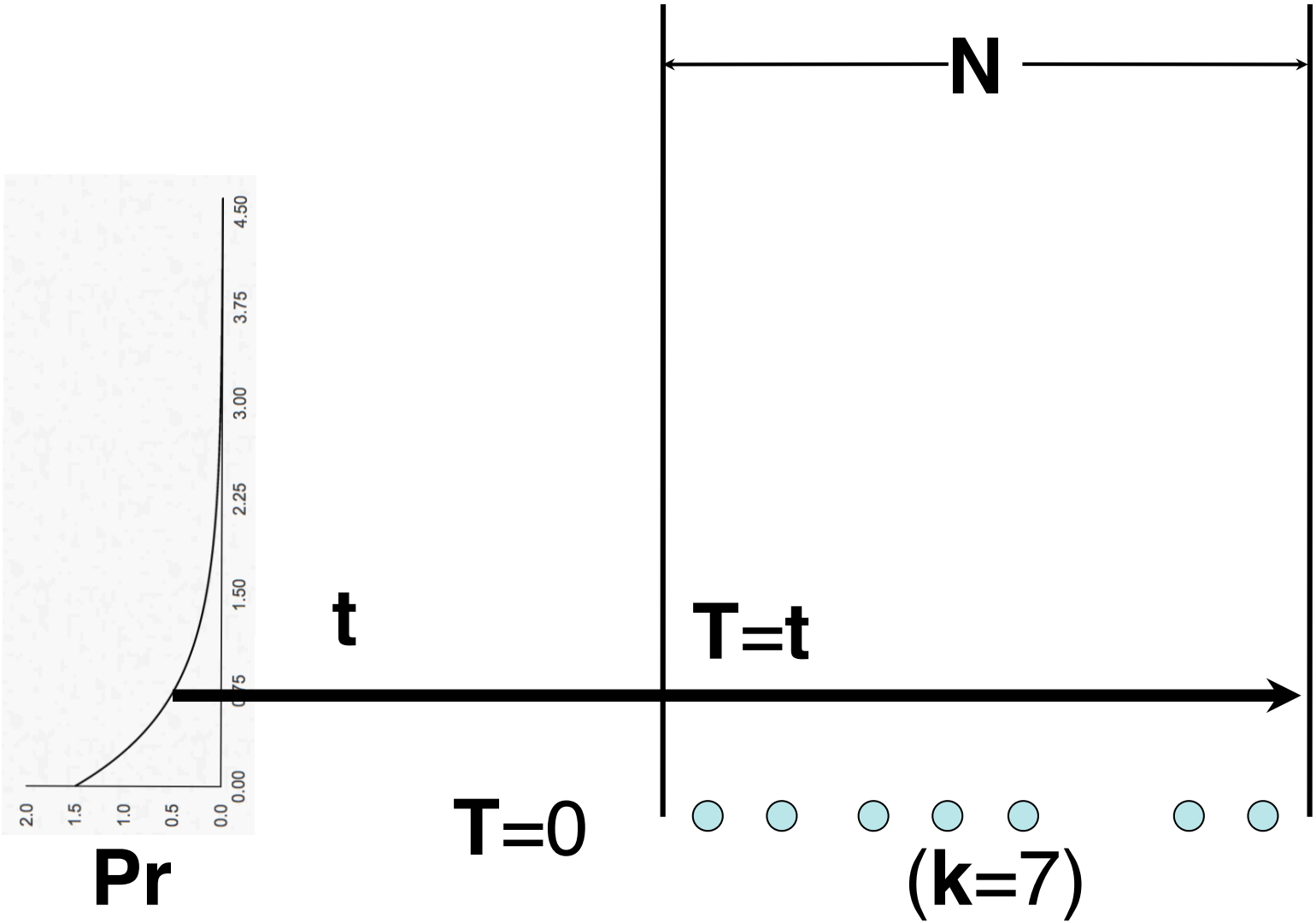
t

T=0

(k=7)

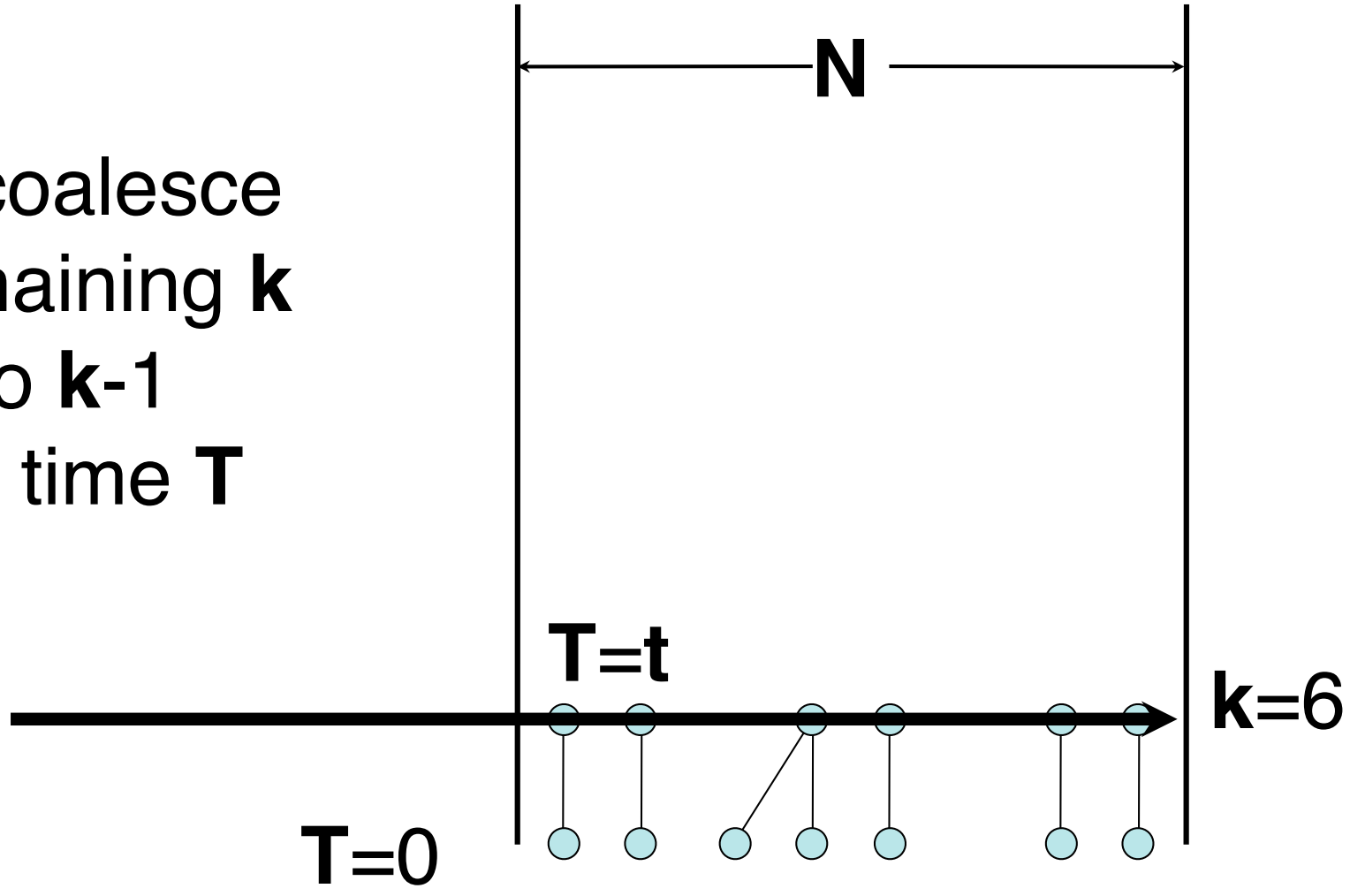
Simulating The Coalescent

set **T = T+ t**



Simulating The Coalescent

Randomly coalesce
2 of the remaining **k**
lineages into **k**-1
Lineages at time **T**

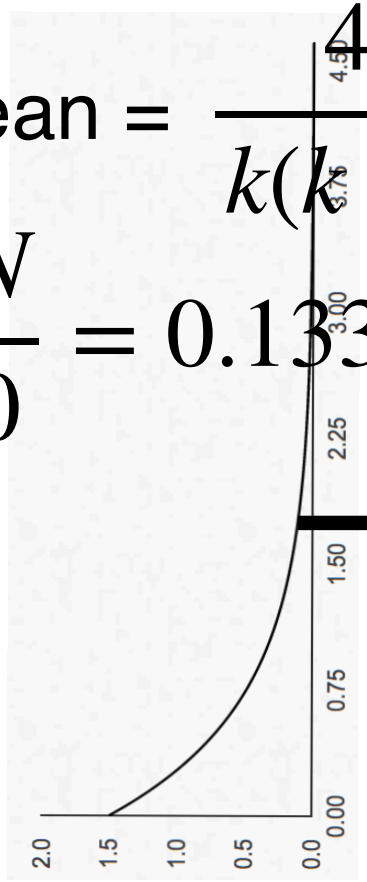


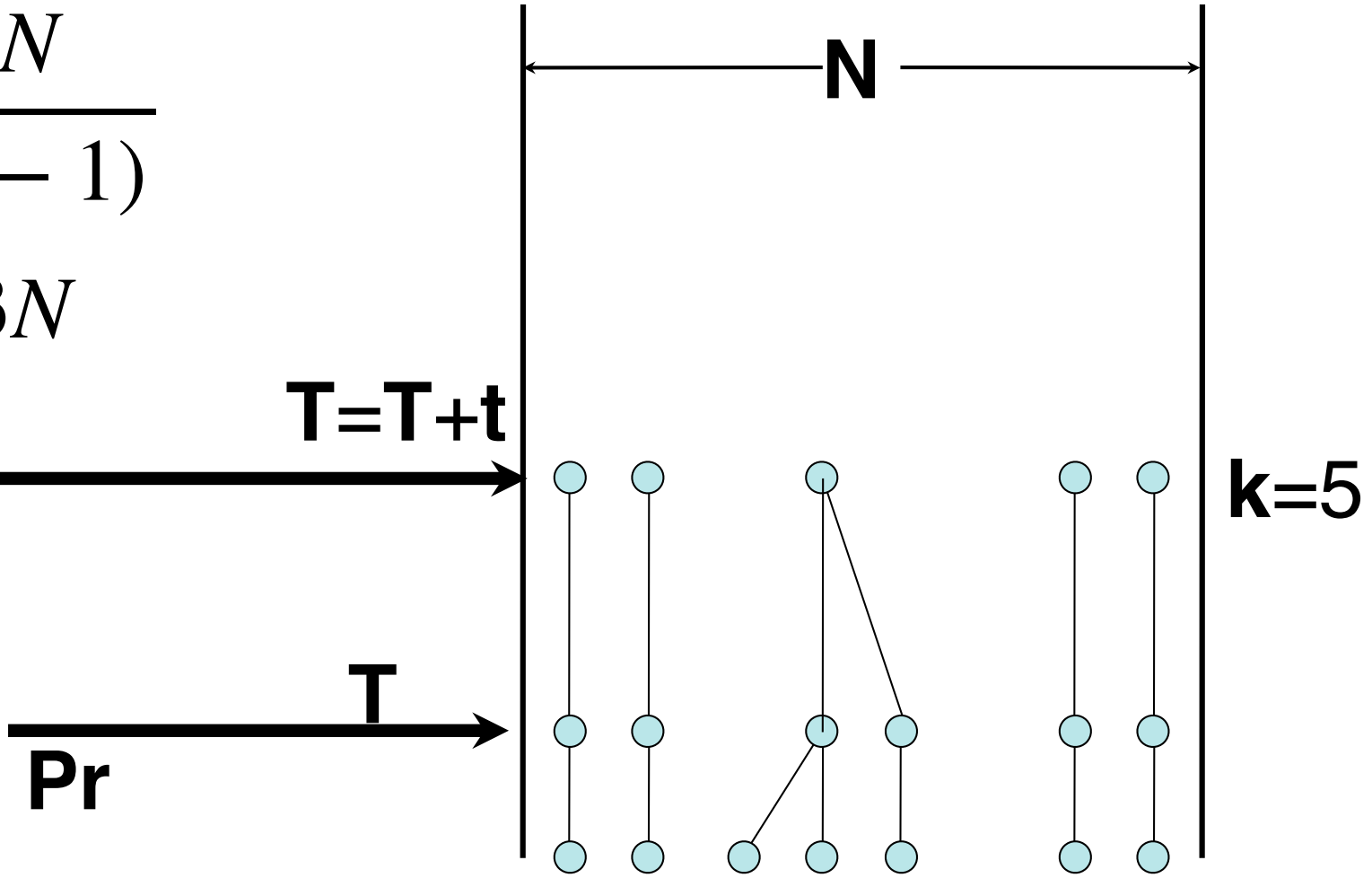
Simulating The Coalescent

randomly draw **t** from exponential with

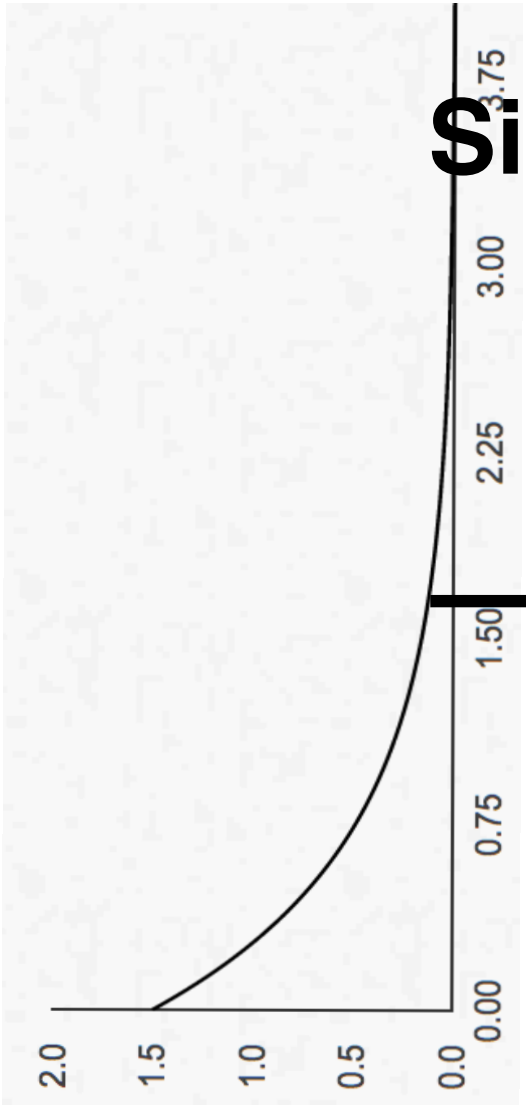
$$\text{mean} = \frac{4N}{k(k-1)}$$

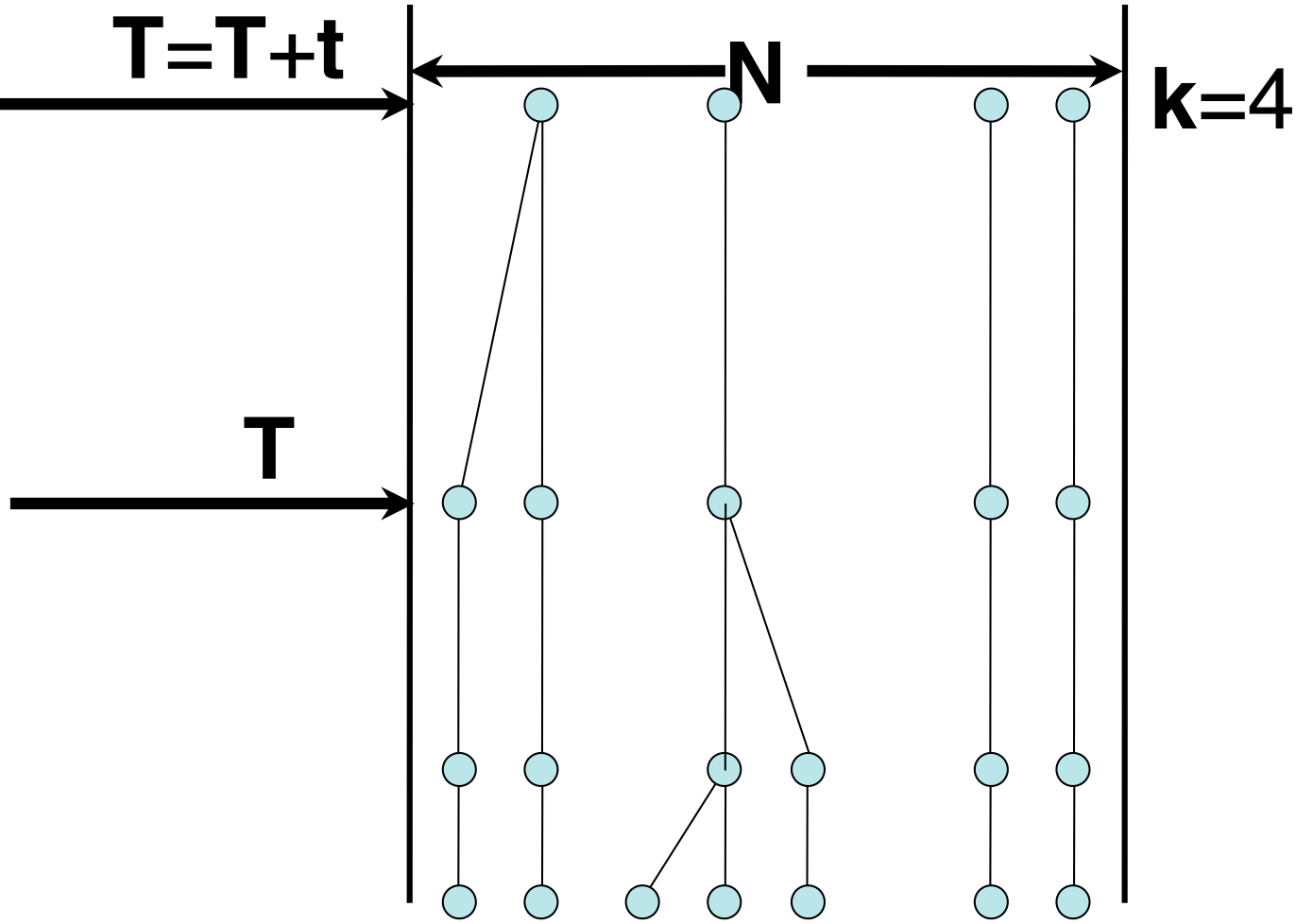
$$\frac{4N}{30} = 0.133N$$





Simulating The Coalescent



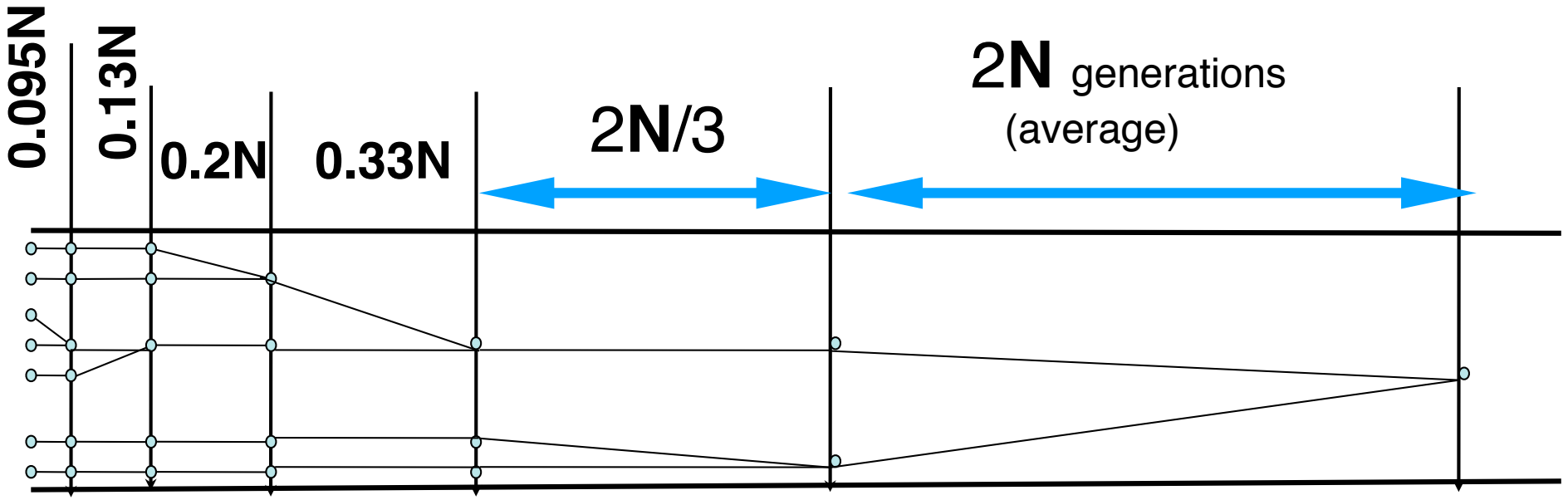


randomly draw **t** from exponential with

mean = $\dfrac{4N}{k(k-1)}$

$\dfrac{4N}{20} = 0.2N$

Simulating The Coalescent



properties of the coalescent

Large Variance

9 randomly generated realizations of gene trees from the coalescent process, all with 20 tips & drawn to same scale (Figure 26.5 from Felsenstein book)

coalescent process leads to highly variable gene genealogies

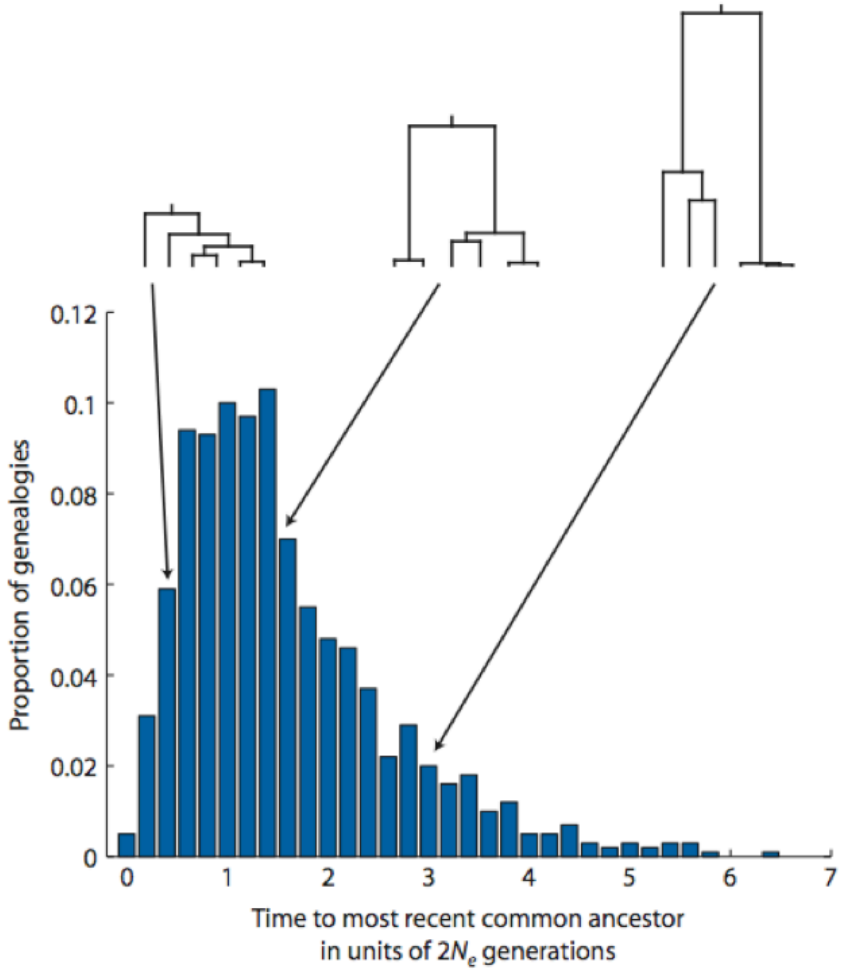
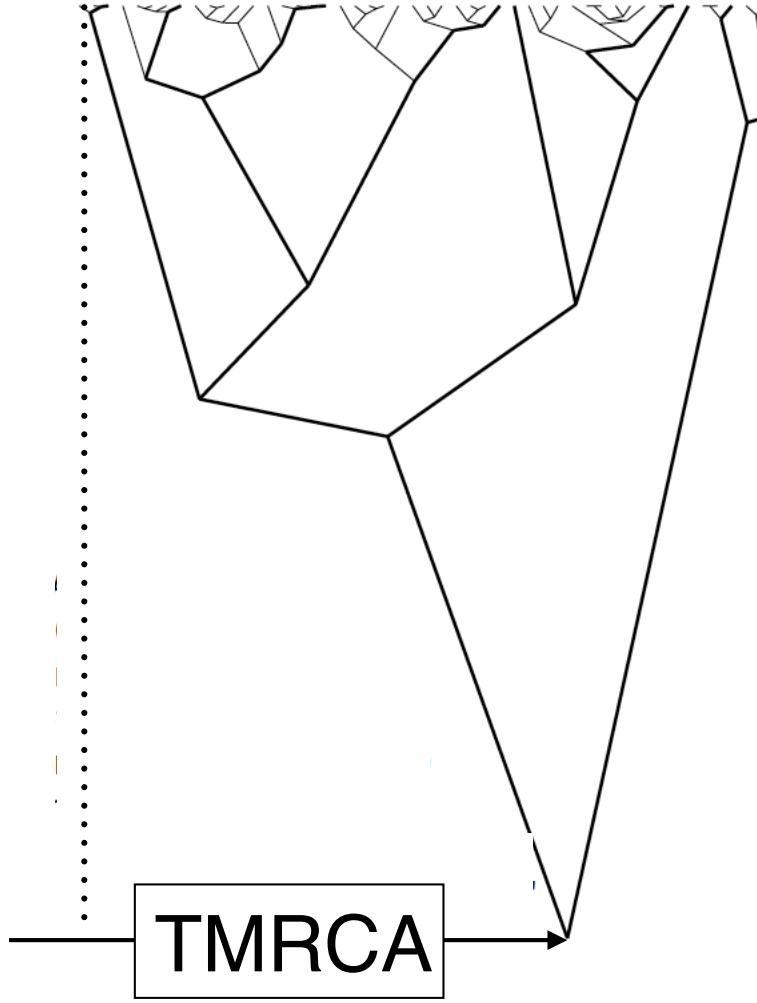


Figure 3.27 The distribution of times to a MRCA (or genealogy heights) for 1000 replicate genealogies starting with six lineages ($k = 6$). The distribution of total coalescence

properties of the coalescent

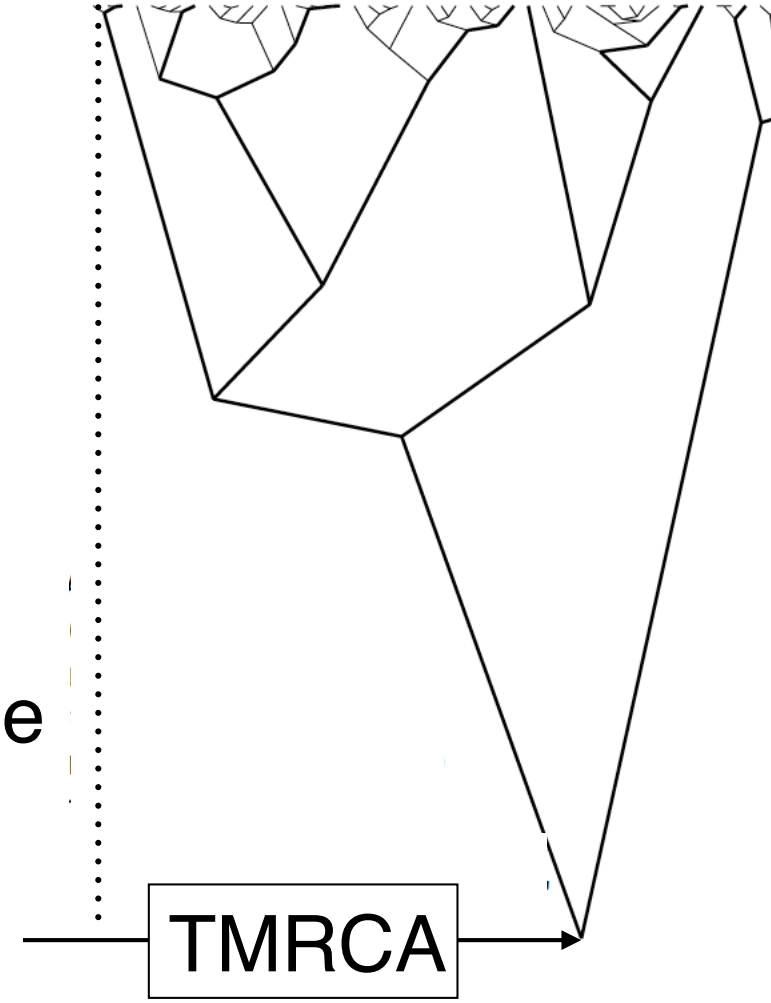
For any k samples,
TMRCA sample = $4N(1 - 1/k)$



properties of the coalescent

For any k samples,
TMRCA sample = $4N(1 - 1/k)$

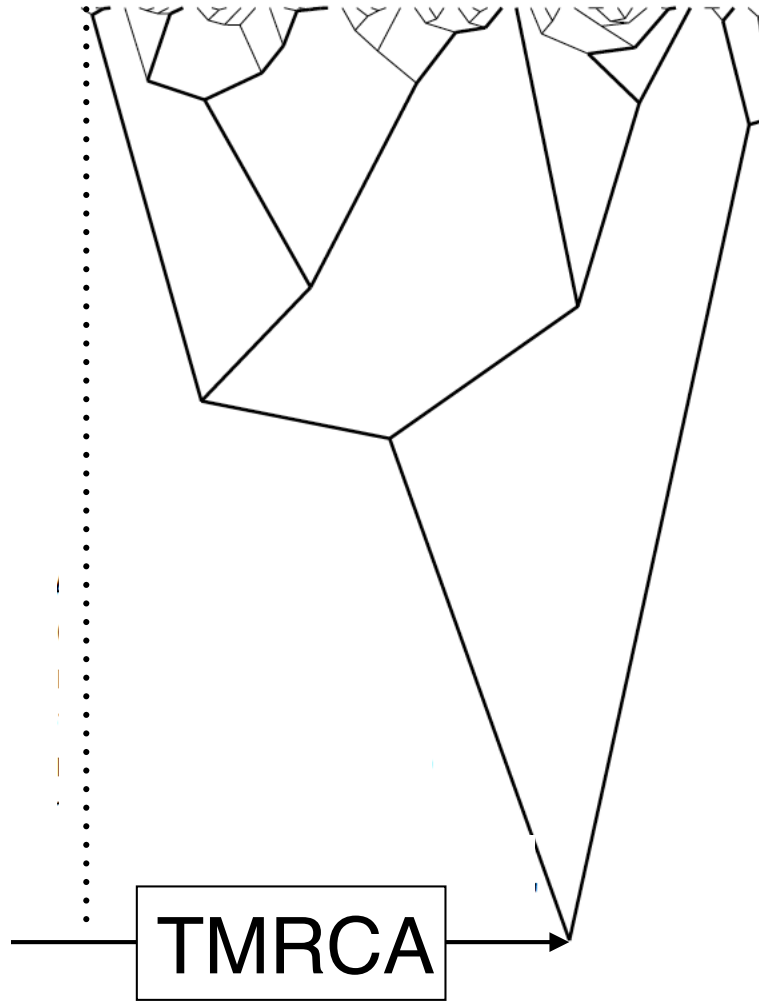
Half this time (~2**N**) is waiting for the last 2 coalescent events



properties of the coalescent

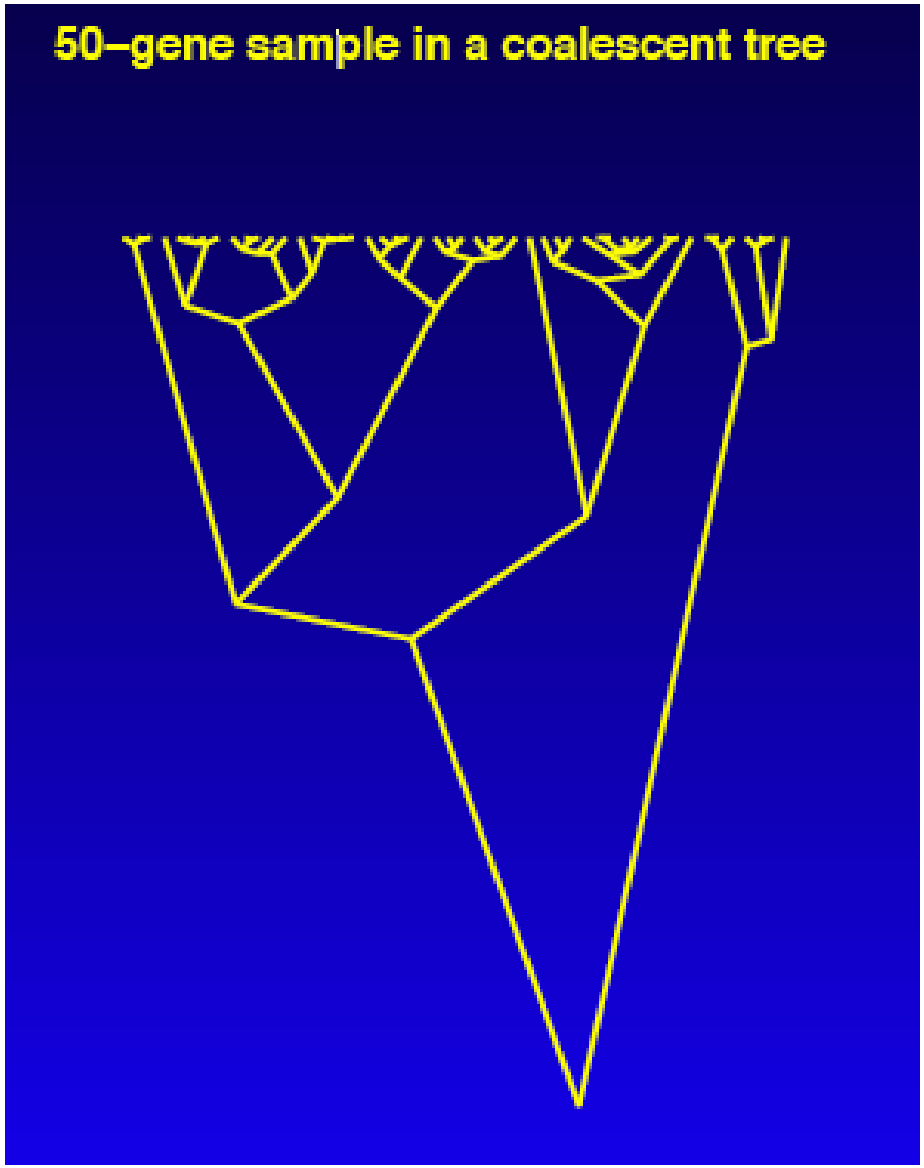
For any k samples,
TMRCA sample = $4N(1 - 1/k)$

$$Var = 16N^2(1 - 1/k)^2$$



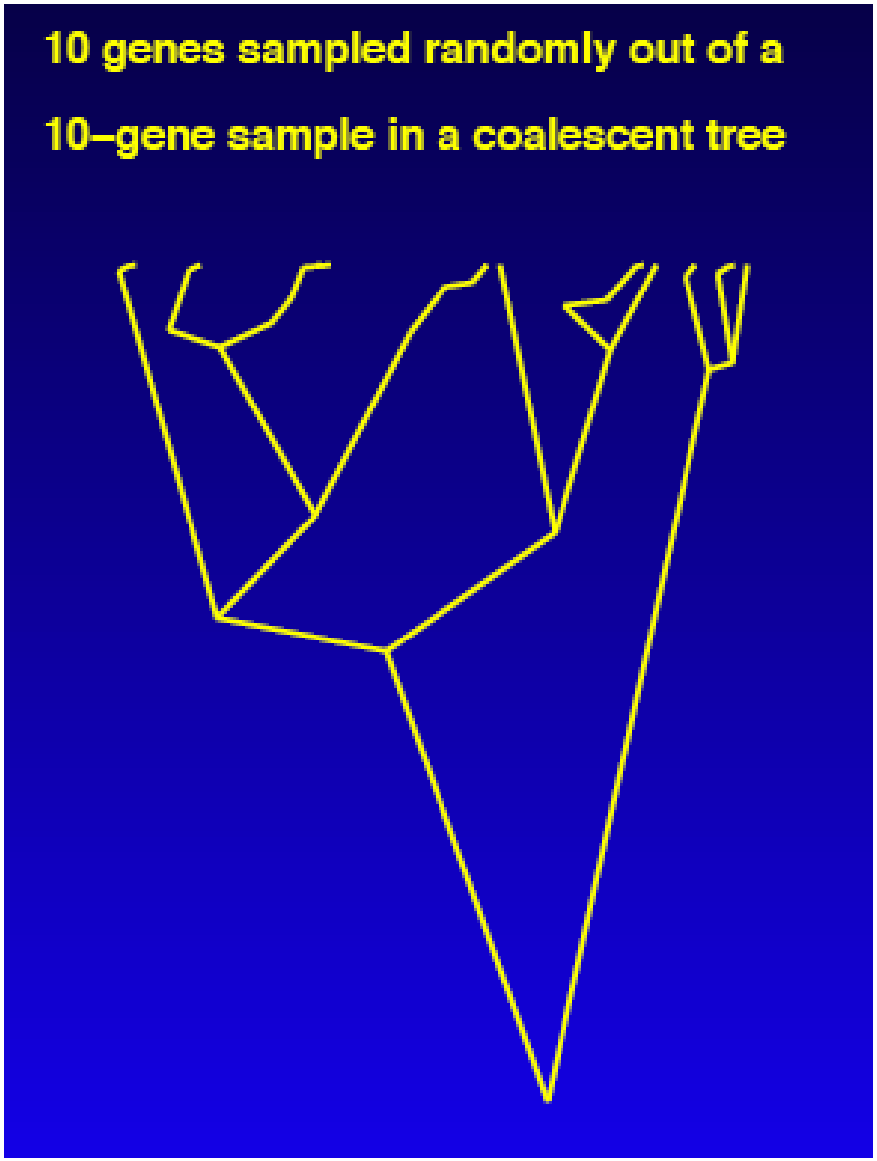
MOST of the variance/noise comes from the last coalescent event

properties of the coalescent



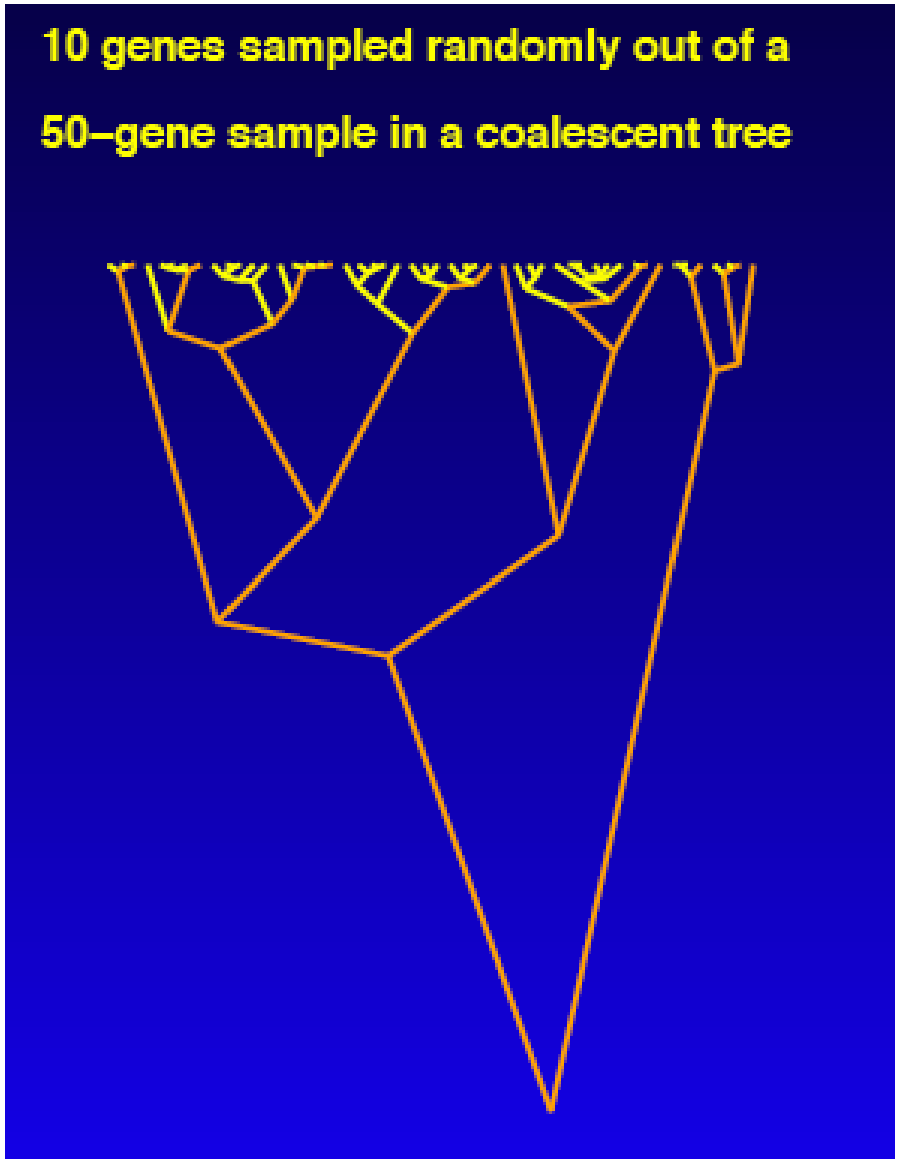
Samples of
k = 10 has most of the information

properties of the coalescent



Samples of
k = 10 has most of the information

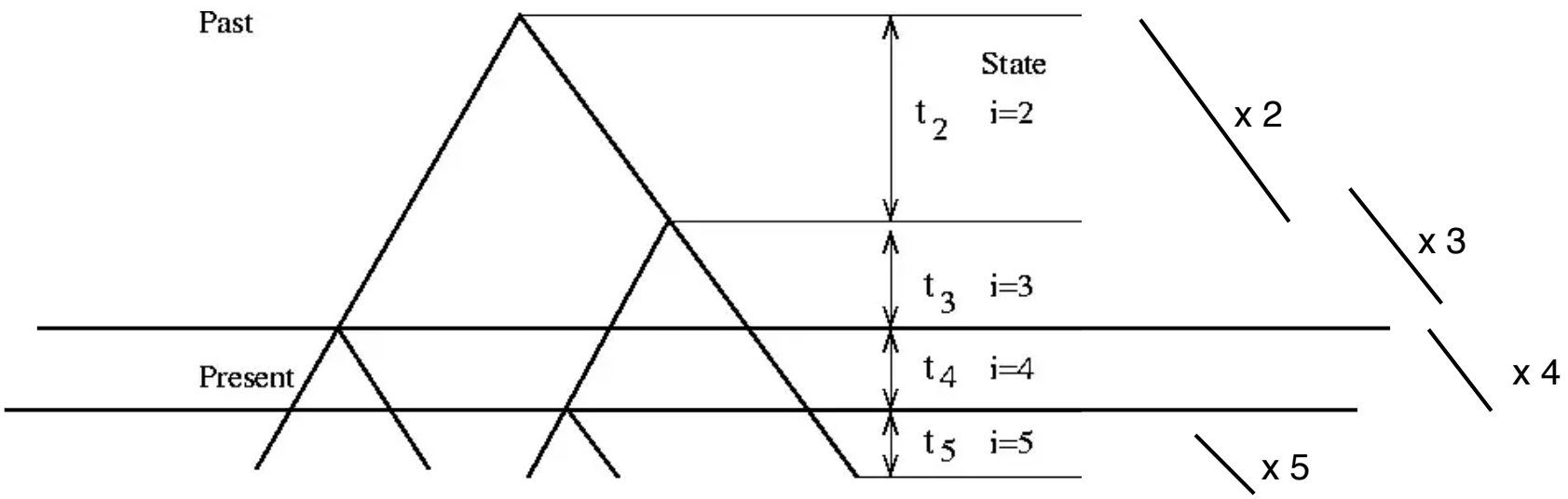
properties of the coalescent



Samples of
k = 10 has most of the information

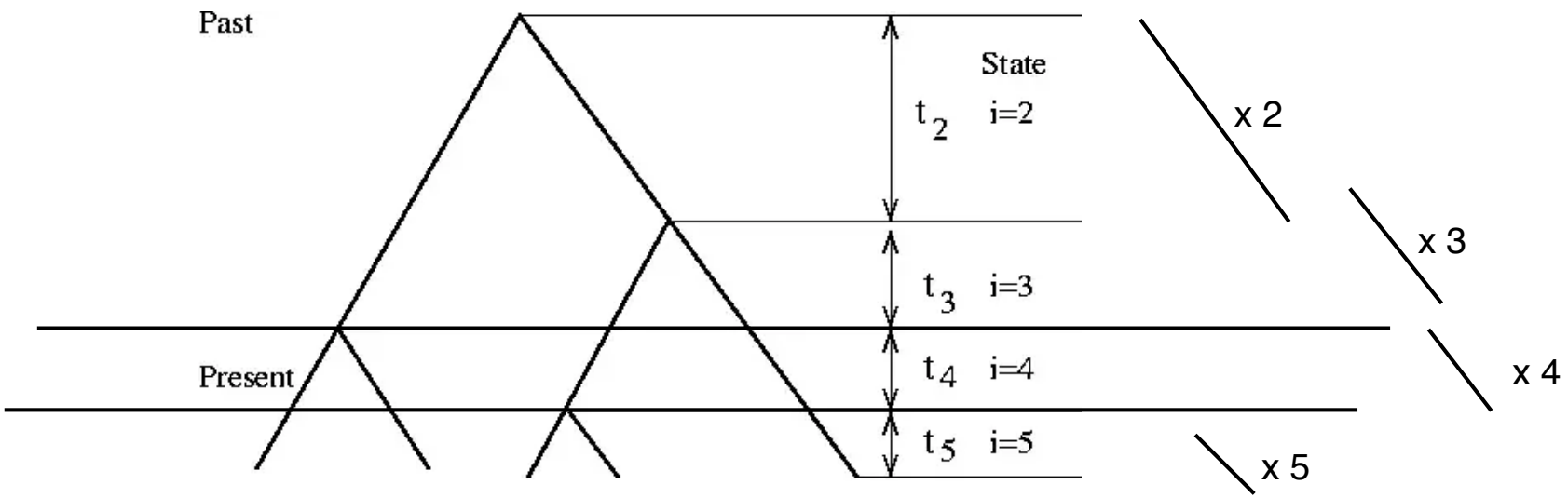
For any k samples,
the total tree length is

$$E(T_{total}) = \sum_{k=2}^{n} k \frac{4N}{k(k-1)}$$



For any k samples,
the total tree length is

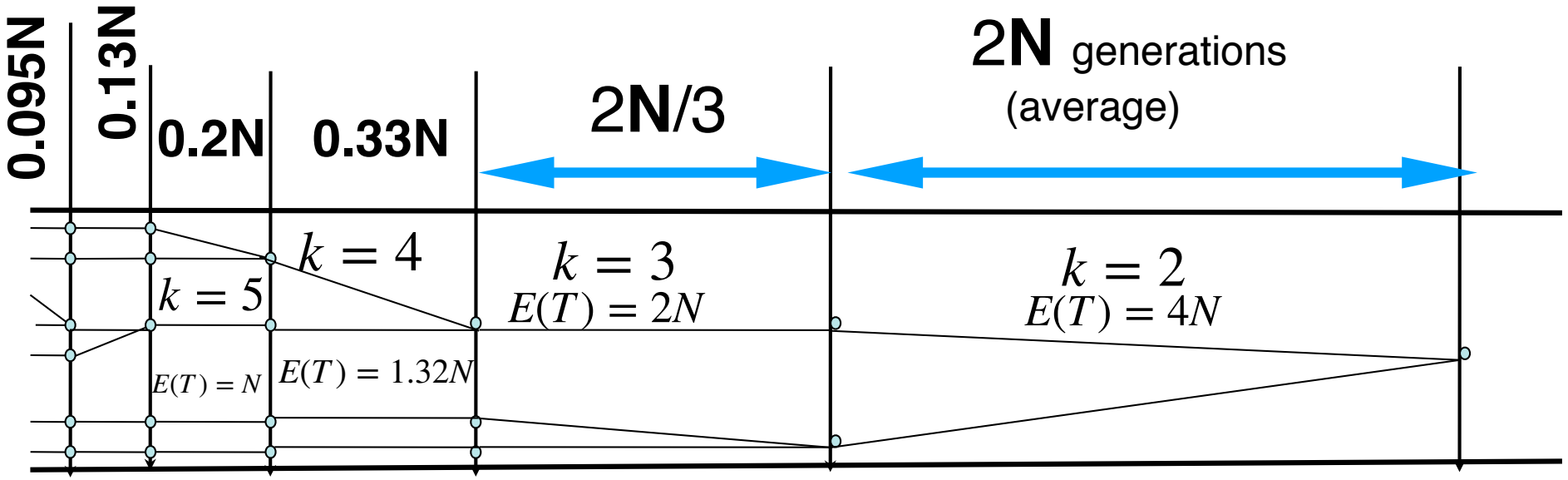
$$E(T_{total}) = \sum_{k=2}^{n} k \frac{4N}{k(k-1)}$$



given that the time that there is *k* lineages before k-1 lineages = $\frac{4N}{k(k-1)}$, we can sequentially multiple these times by *k*.

For any k samples,
the total tree length

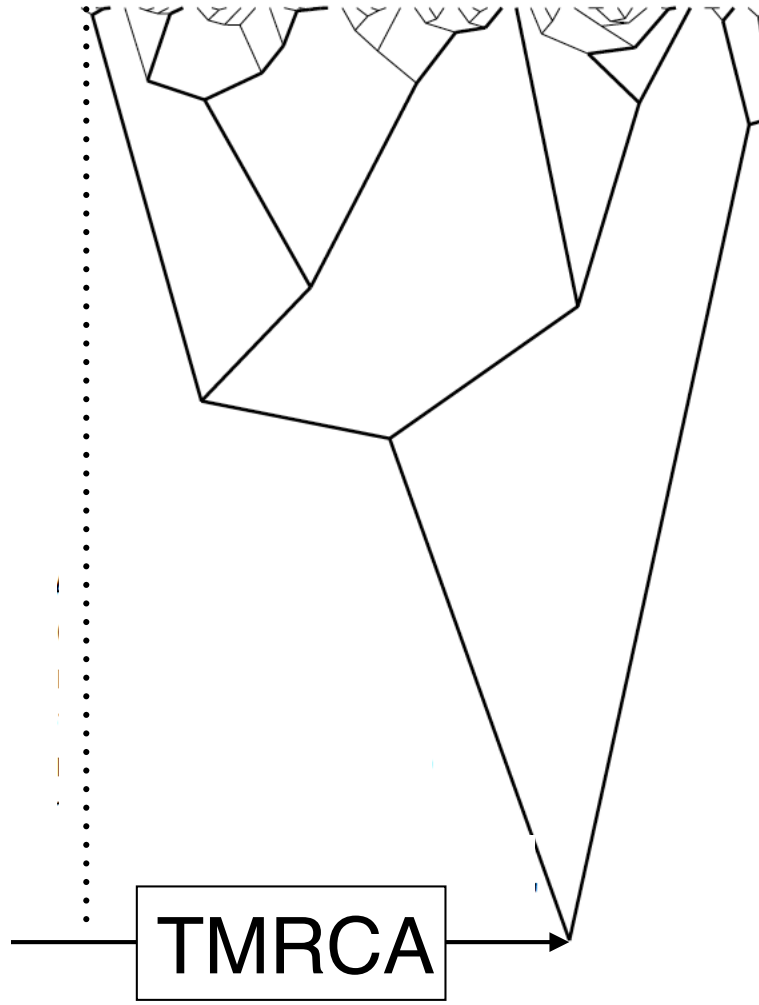
$$E(T_{total}) = \sum_{k=2}^{n} k \frac{4N}{k(k-1)}$$



if we “sprinkle” mutations, then $E(T_{total})$ gives you E(# polymorphic sites)

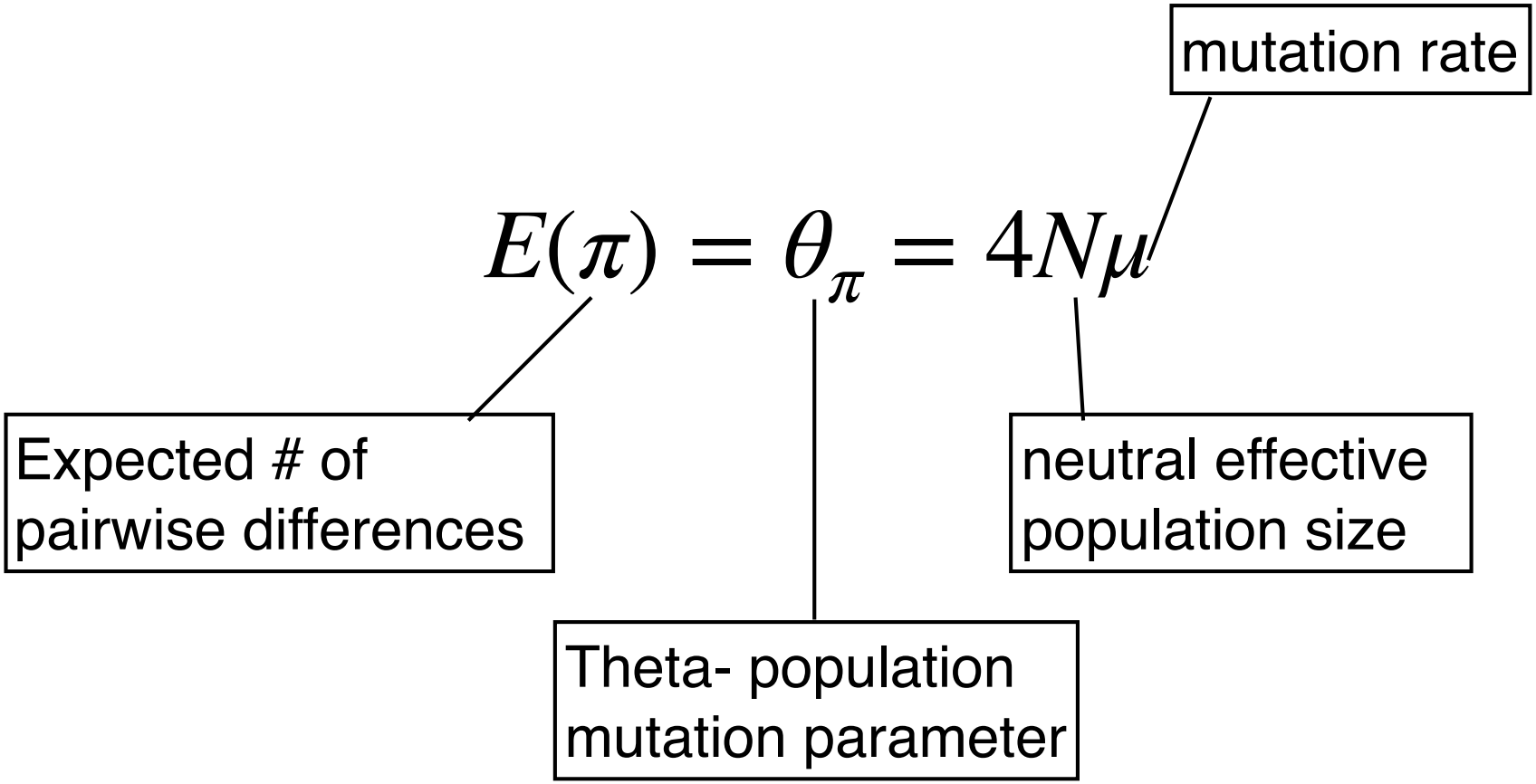
recall

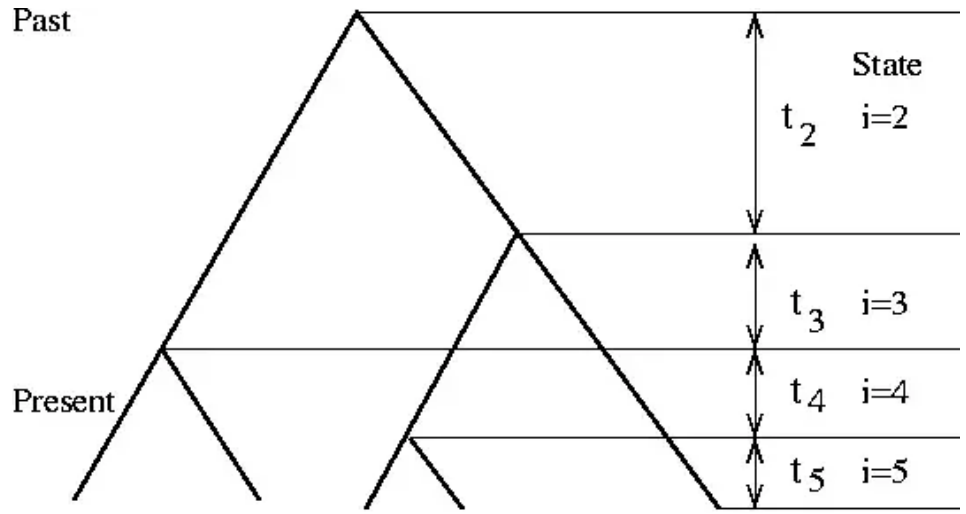
For any k samples,
TMRCA sample = $4N(1 - 1/k)$



if we “sprinkle” mutations, then TMRCA for any two samples becomes $4N\mu$

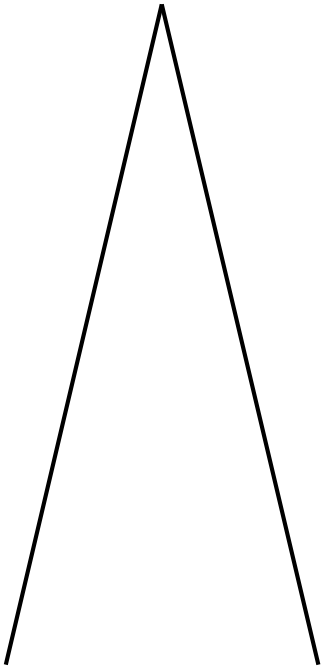
now back to the classic sumstat, average pairwise differences





think about the case of 2 lineages ($k = 2$)

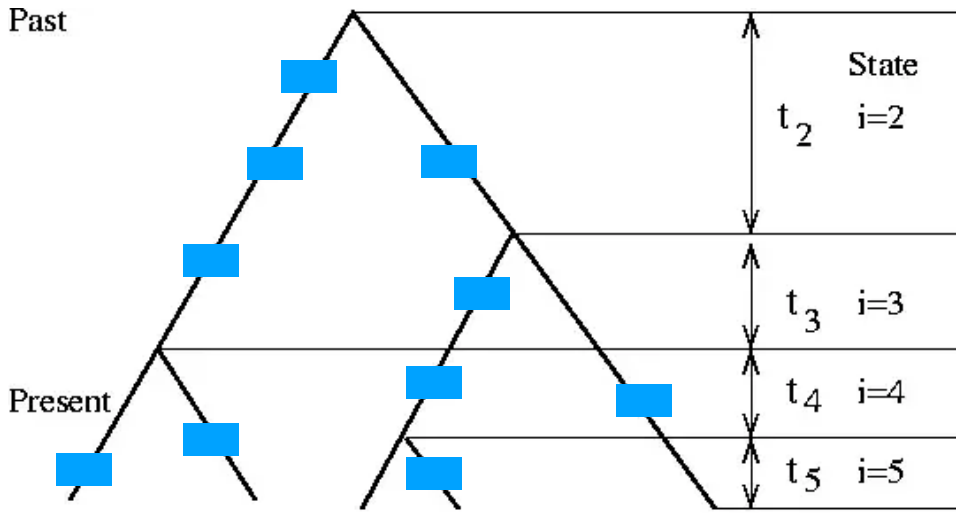
TMRCA sample = $4N(1 - 1/k) = 4N(1 - 1/2) = 2N$



2*TMRCA between any 2 samples

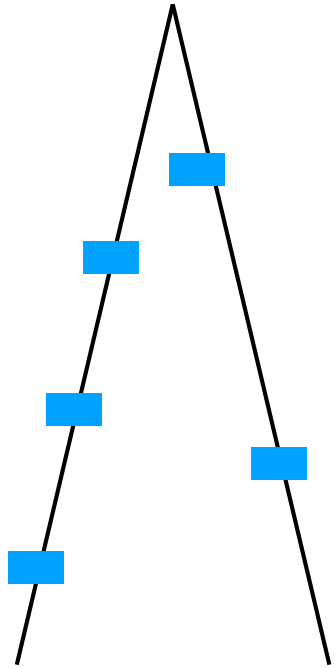
$= 4N$

sprinkling mutations $\longrightarrow E(\pi) = \theta_\pi = 4N\mu$



think about the case of 2 lineages ($k = 2$)

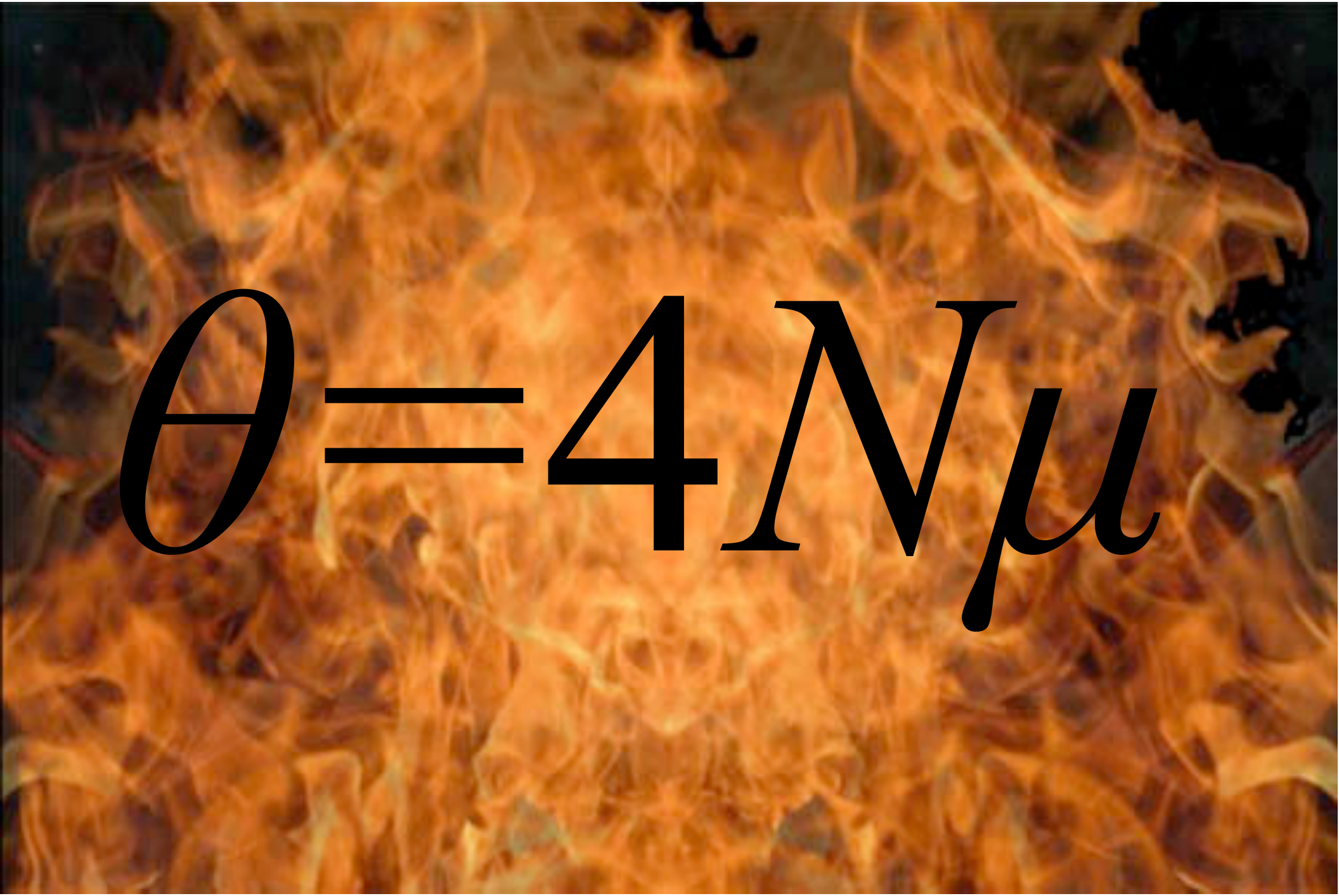
TMRCA sample = $4N(1 - 1/k) = 4N(1 - 1/2) = 2N$



2*TMRCA between any 2 samples

$= 4N$

sprinkling mutations $\longrightarrow E(\pi) = \theta_\pi = 4N\mu$


$\theta = 4N\mu$

S (# of polymorphic sites) estimator of θ

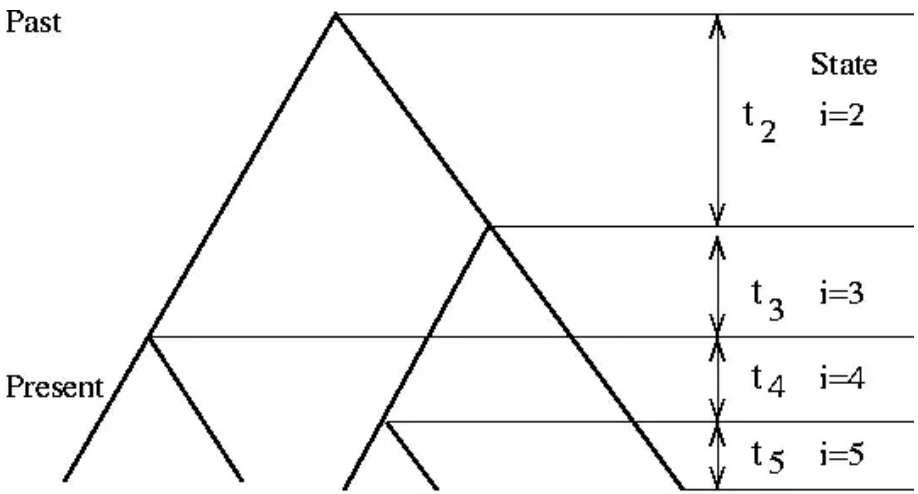
$$\theta_W = S / \sum_{k=1}^{n-1} \frac{1}{k}$$

looks like total tree length!

$$E(T_{total}) = \sum_{k=2}^{n} k \frac{4N}{k(k-1)} = 4N \sum_{k=1}^{n-1} \frac{1}{k}$$

if we “sprinkle” mutations, then $E(T_{total})$ becomes E(*S*) (# polymorphic sites)

$$E(T_{total}) = \mu * 4N \sum_{k=1}^{n-1} \frac{1}{k}$$



adding more lineages increases *S*, but less and less to the total length

demographic events can change the shape of the tree

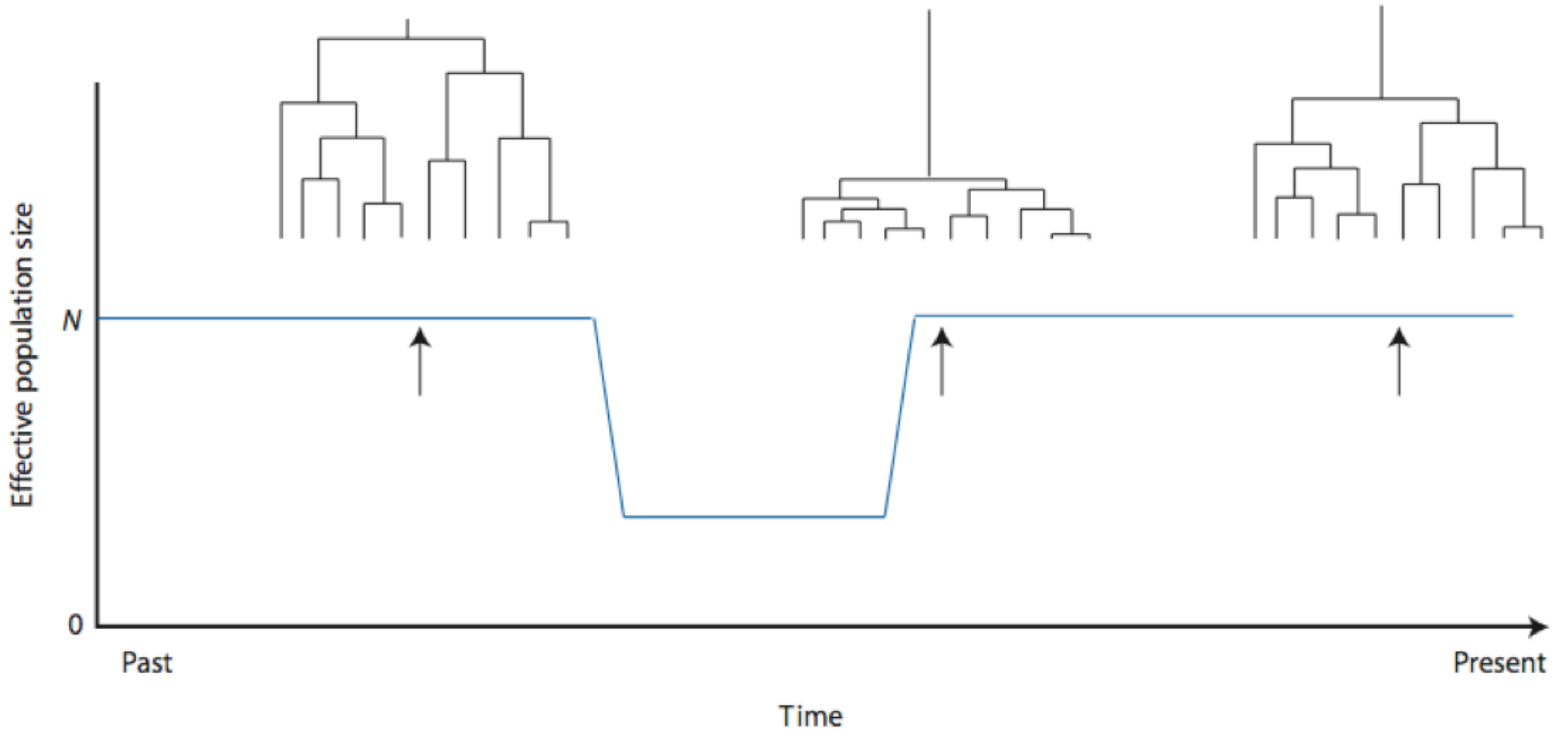
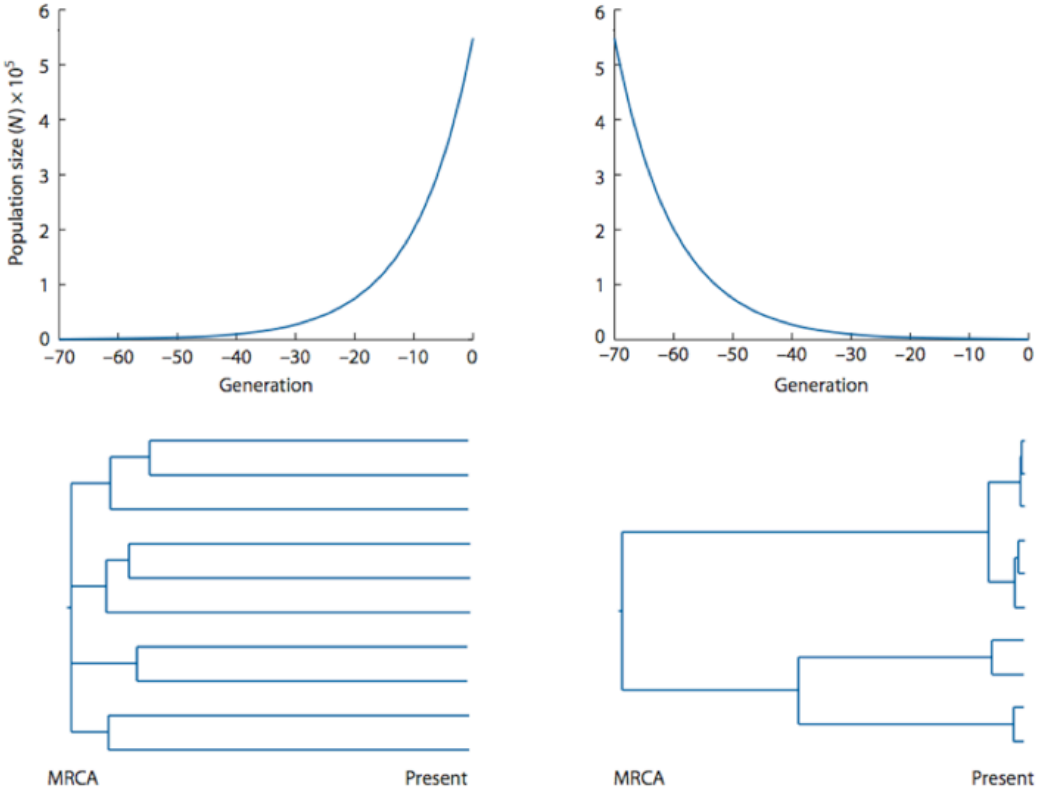


Figure 3.28 The effects of a population bottleneck on gene genealogies. During the bottleneck the chance that two randomly sampled gene copies are derived from one copy in the previous generation $\left(\frac{1}{2N_e}\right)$ increases. This can also be thought of as a

demographic events can change the shape of the tree



NOTE: Tree-shape determines pattern of variation in the population. Population growth (left) will create many alleles that are carried by one individual. Population declines (right) will create many loci carried by 1 or more individuals.

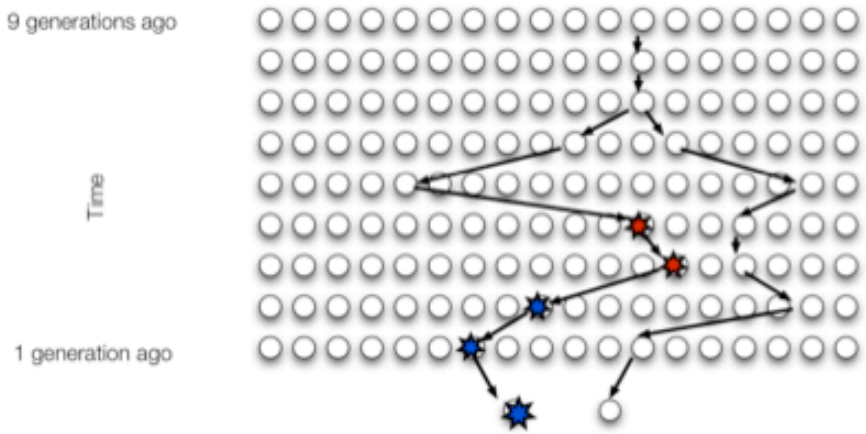
review

Why do we find:

- The coalescent to be faster in small populations than large ones?
- The coalescent to be much faster when k is large then when small?

review: expected number of differences between a random pair of sequences

Differences must arise due to mutation before the coalescent event.



NOTE: We'll assume every mutation gives rise to an observable difference (i.e. infinite sites model)

Example

Sequence 1: TG
Sequence 2: AA

review: expected number of differences between a random pair of sequences

- Differences must arise due to mutation.
- What is the average amount of time before the coalescent event?

review: expected number of differences between a random pair of sequences

- Differences must arise due to mutation.
- What is the average amount of time before the coalescent event?
  - Answer: 2N generations

review: expected number of differences between a random pair of sequences

- Differences must arise due to mutation.
- What is the average amount of time before the coalescent event?
  - Answer: 2N generations
- What is the rate of mutation while there are two lineages?

review: expected number of differences between a random pair of sequences

- Differences must arise due to mutation.
- What is the average amount of time before the coalescent event?
  - Answer: 2N generations
- What is the rate of mutation while there are two lineages?
  - Answer: 2μ mutations per generation.

review: expected number of differences between a random pair of sequences

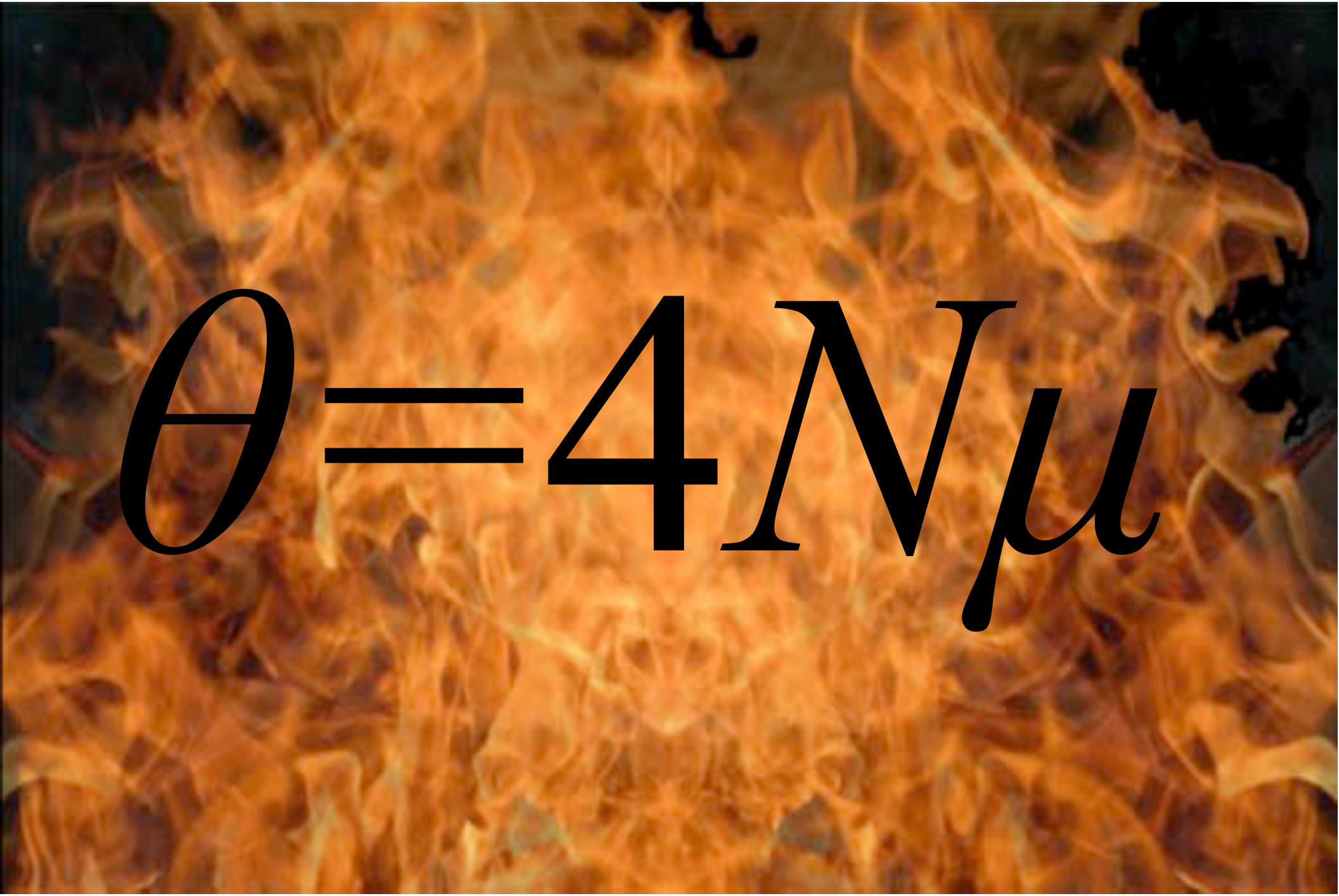
- Differences must arise due to mutation.
- What is the average amount of time before the coalescent event?
  - Answer: 2N generations
- What is the rate of mutation while there are two lineages?
  - Answer: 2μ mutations per generation.
- Rate $\times$ Time $=$ Total expected number.
  - Answer: $2N$ generations $\times 2\mu$ mutations per generation $= 4N\mu$ total differences.

review: expected number of differences between a random pair of sequences

- Differences must arise due to mutation.
- What is the average amount of time before the coalescent event?
  - Answer: 2N generations
- What is the rate of mutation while there are two lineages?
  - Answer: 2μ mutations per generation.
- Rate $\times$ Time $=$ Total expected number.
  - Answer: $2N$ generations $\times 2\mu$ mutations per generation $= 4N\mu$ total differences.

Definition

$\theta = 4N\mu$: sometimes called the "fundamental population genetic parameter"


$\theta = 4N\mu$

review: expected # of SNPs

```
Seq1 ACTAAAGGCG
Seq2 ACTAAAGGCG
Seq3 ACGAAATGCG
Seq4 ACTAAAGGCG
     --*---*---
```

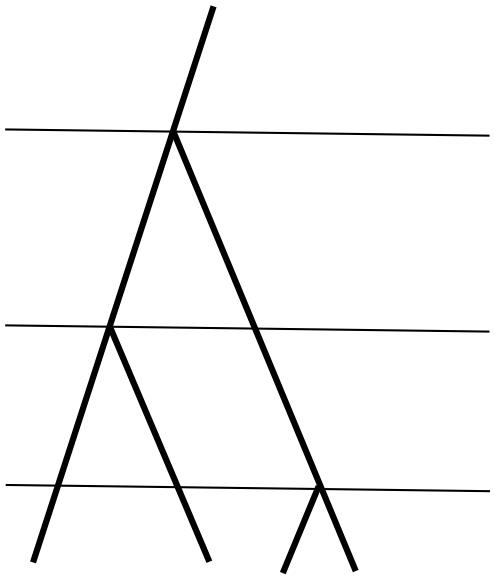
2 segregating sites out of 10.

Question

What is the expected number of segregating sites in the sample?

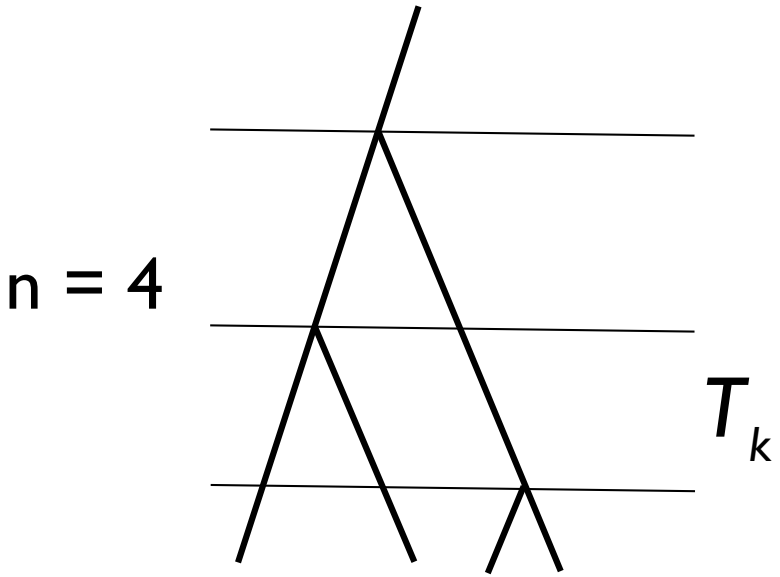
- We can divide the tree up into $n - 1$ segments.

n = 4

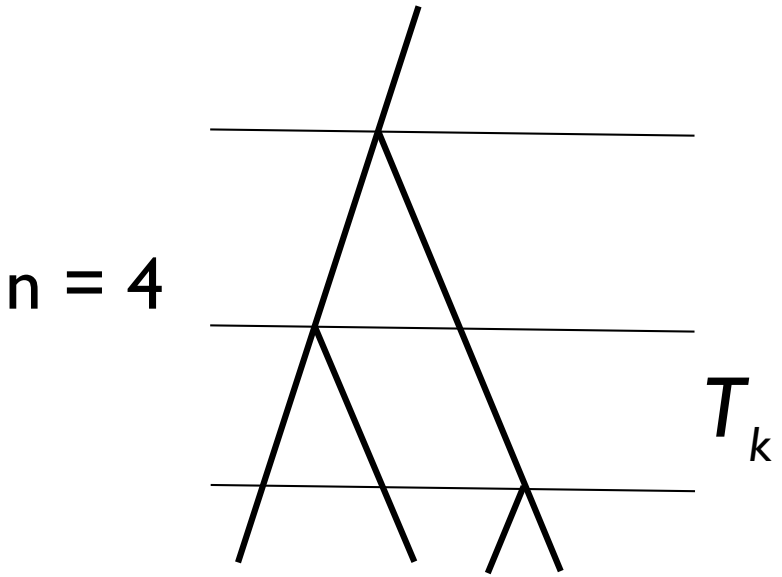


review:expected # of SNPs

- We can divide the tree up into $n - 1$ segments.
- The kth segment has a time length of T_k, we have a rate of mutations arriving at μk during that time interval.



- We can divide the tree up into $n-1$ segments.
- The kth segment has a time length of T_k, we have a rate of mutations arriving at μk during that time interval.
- The expected number of mutations per interval is then $E[T_k]\mu k$.

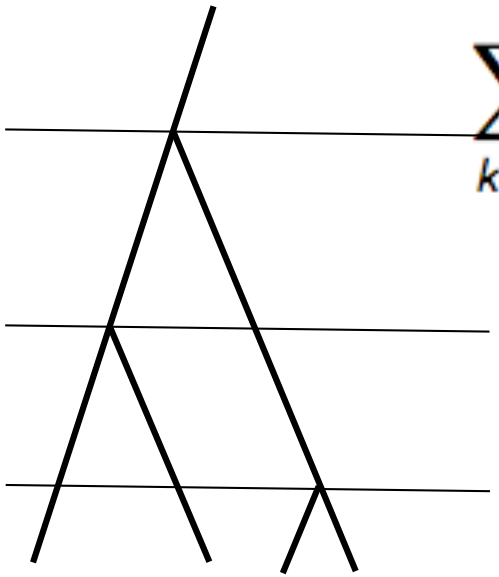


$E[T_k]=\mu k$ (ie μ*3)

review: expected # of SNPs

- We can divide the tree up into $n-1$ segments.
- The kth segment has a time length of T_k, we have a rate of mutations arriving at μk during that time interval.
- The expected number of mutations per interval is then $E[T_k]\mu k$.
- Summing across the intervals we have:

$$\sum_{k=2}^{n} E[T_k]k\mu = \mu \sum_{k=2}^{n} kE[T_k]$$



review: expected # of SNPs

- We can divide the tree up into $n-1$ segments.
- The kth segment has a time length of T_k, we have a rate of mutations arriving at μk during that time interval.
- The expected number of mutations per interval is then $E[T_k]\mu k$.
- Summing across the intervals we have:

$$\sum_{k=2}^{n} E[T_k]k\mu = \mu \sum_{k=2}^{n} kE[T_k]$$

which because $E[T_k] = 2N/\binom{k}{2} = \frac{4N}{k(k-1)}$

review: expected # of SNPs

- We can divide the tree up into $n-1$ segments.
- The kth segment has a time length of T_k, we have a rate of mutations arriving at μk during that time interval.
- The expected number of mutations per interval is then $E[T_k]\mu k$.
- Summing across the intervals we have:

$$\sum_{k=2}^{n} E[T_k]k\mu = \mu \sum_{k=2}^{n} kE[T_k]$$

which because $E[T_k] = 2N/\binom{k}{2} = \frac{4N}{k(k-1)}$

$$= \mu \textstyle\sum_{k=2}^{n} k \frac{4N}{k(k-1)} \quad (1)$$

$$= 4N\mu \textstyle\sum_{k=2}^{n} \frac{1}{k-1} \quad (2)$$

$$= 4N\mu \textstyle\sum_{k=1}^{n-1} \frac{1}{k} \quad (3)$$

Summary of our coalescent-derived results

- Mean number of differences between a pair of sequences $= \theta$
- Expected heterozygosity $= \frac{\theta}{\theta+1}$
- Expected number of segregating sites $= \theta \sum_{i=1}^{n-1} \frac{1}{i}$

Summary of our coalescent-derived results

- Mean number of differences between a pair of sequences $= \theta$
- Expected heterozygosity $= \frac{\theta}{\theta+1}$
- Expected number of segregating sites $= \theta \sum_{i=1}^{n-1} \frac{1}{i}$

Example

If $N = 10000$ and $\mu = 2.5 \times 10^{-8}$ per site and we consider a sequence of length $1000bp$, then $\theta = 4 \times 10^4 \times 2.5 \times 10^{-8} \times 10^3 = 1$.

- Mean number of differences between a pair of sequences of length 1000 bp $= 1$ difference.
- Expected heterozygosity $= \frac{\theta}{\theta+1} = 1/2$
- If $n = 10$, expected number of segregating sites $= \theta \sum_{i=1}^{n-1} \frac{1}{i} = 2.928$

Summary of our coalescent-derived results

- Mean number of differences between a pair of sequences $= \theta$
- Expected heterozygosity $= \frac{\theta}{\theta+1}$
- Expected number of segregating sites $= \theta \sum_{i=1}^{n-1} \frac{1}{i}$

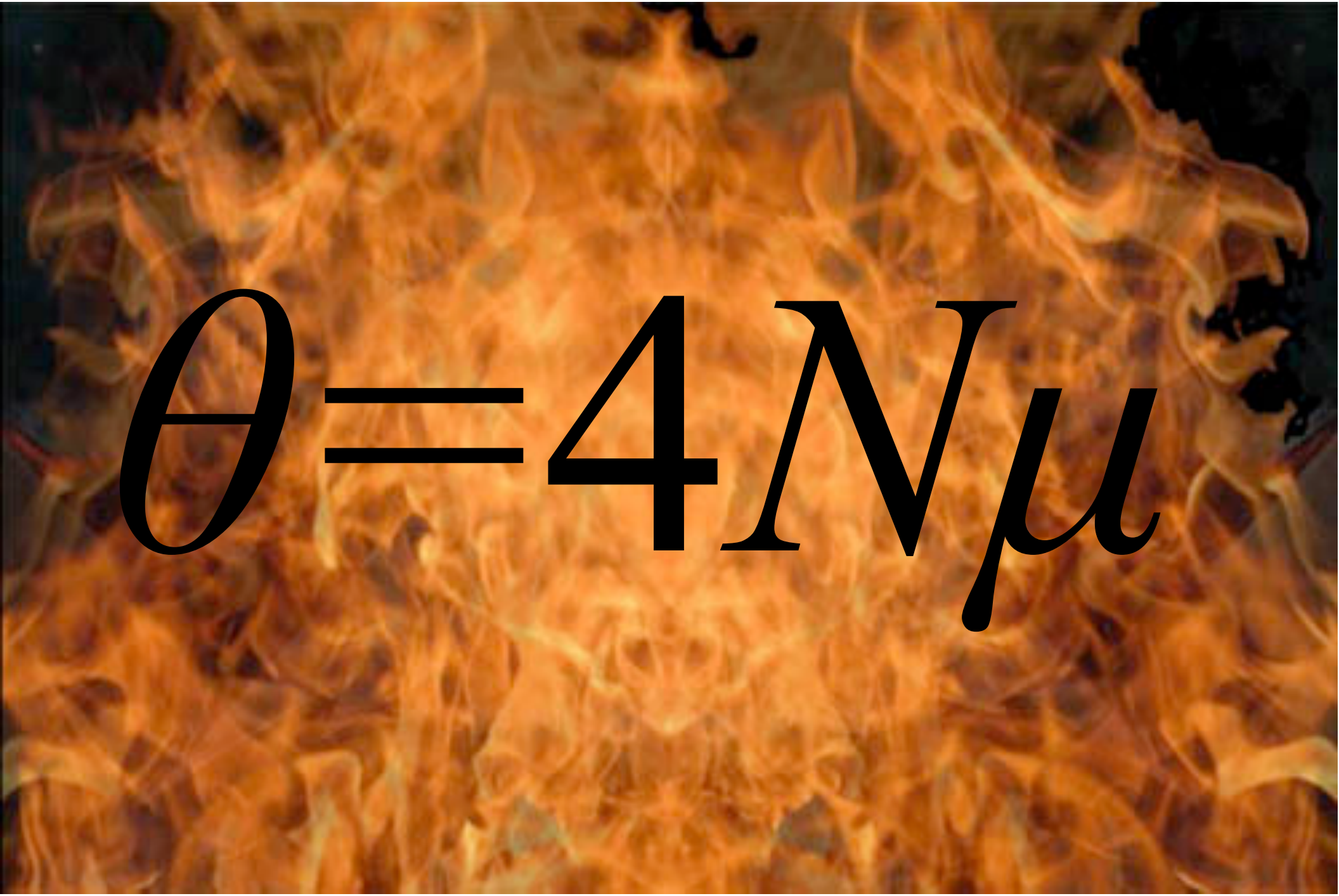
Example

If $N = 10000$ and $\mu = 2.5 \times 10^{-8}$ per site and we consider a sequence of length $1000bp$, then $\theta = 4 \times 10^4 \times 2.5 \times 10^{-8} \times 10^3 = 1$.

- Mean number of differences between a pair of sequences of length 1000 bp $= 1$ difference.
- Expected heterozygosity $= \frac{\theta}{\theta+1} = 1/2$
- If $n = 10$, expected number of segregating sites $= \theta \sum_{i=1}^{n-1} \frac{1}{i} = 2.928$

NOTE

The parameter $\theta = 4N\mu$ came up in each of our three results.


$\theta = 4N\mu$

Coalescent theory:

- **Data-driven**: Only concerned with the variation in a *small sample of data* from large population.

Coalescent theory:

- **Data-driven**: Only concerned with the variation in a *small sample of data* from large population.
- **Theoretical insights**: Allows us to see how N and μ perfectly balance each other out in controlling observed diversity.

Coalescent theory:

- **Data-driven**: Only concerned with the variation in a *small sample of data* from large population.
- **Theoretical insights**: Allows us to see how N and μ perfectly balance each other out in controlling observed diversity.
- **Easy math**: We can derive important results with just a few lines of math.

Coalescent theory:

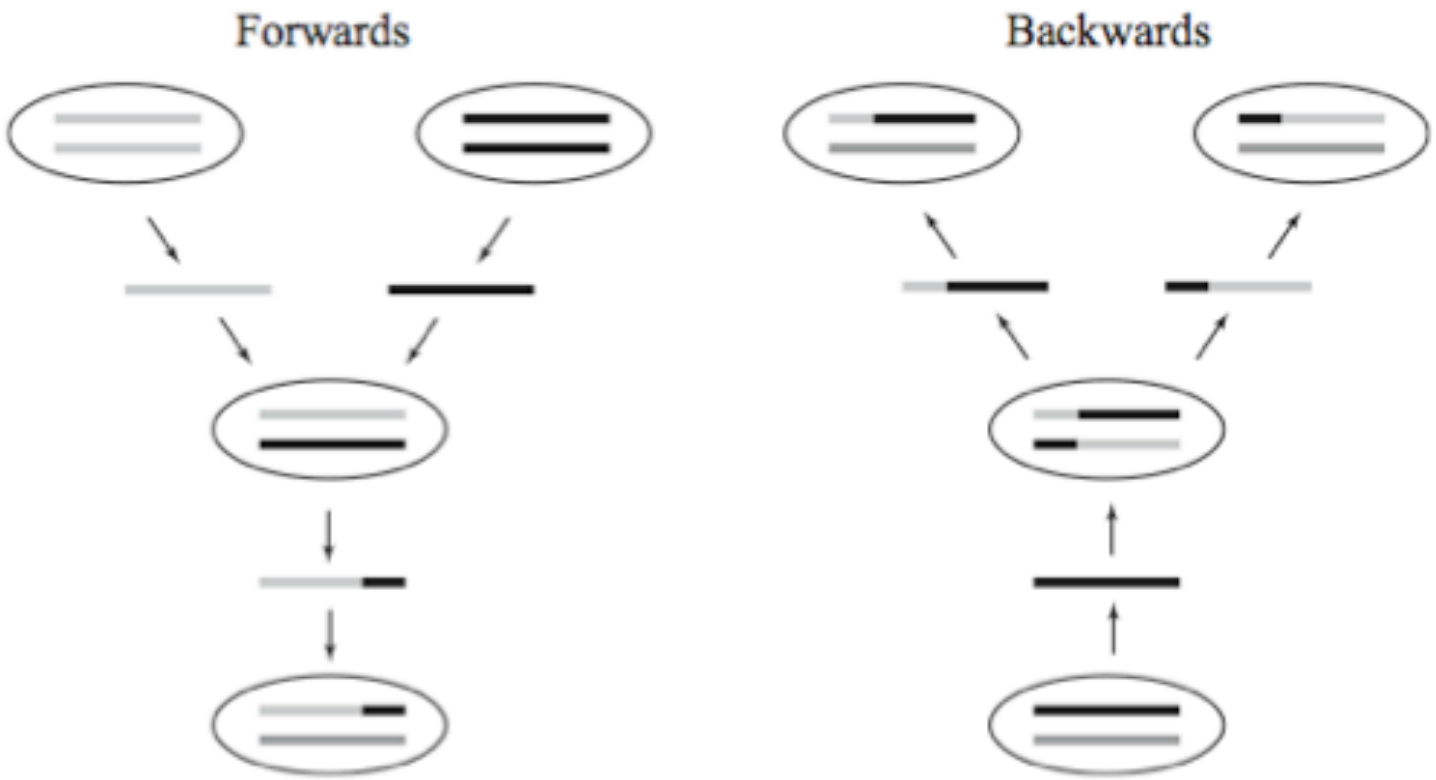
- **Data-driven**: Only concerned with the variation in a *small sample of data* from large population.
- **Theoretical insights**: Allows us to see how N and μ perfectly balance each other out in controlling observed diversity.
- **Easy math**: We can derive important results with just a few lines of math.
- **Fast simulation**: No need to simulate the whole population - just your sample.

Coalescent theory:

- **Data-driven**: Only concerned with the variation in a *small sample of data* from large population.
- **Theoretical insights**: Allows us to see how N and μ perfectly balance each other out in controlling observed diversity.
- **Easy math**: We can derive important results with just a few lines of math.
- **Fast simulation**: No need to simulate the whole population - just your sample.
- **Mutation-drift equilibrium**: The results do not change depending on when we sample the population... Implies the population is at an equilibrium. Here we only model mutation and drift - hence this is a *mutation-drift* equilibrium.

Coalescent theory:

- **Data-driven**: Only concerned with the variation in a *small sample of data* from large population.
- **Theoretical insights**: Allows us to see how N and μ perfectly balance each other out in controlling observed diversity.
- **Easy math**: We can derive important results with just a few lines of math.
- **Fast simulation**: No need to simulate the whole population - just your sample.
- **Mutation-drift equilibrium**: The results do not change depending on when we sample the population... Implies the population is at an equilibrium. Here we only model mutation and drift - hence this is a *mutation-drift* equilibrium.
- **Recombination**: Adds a complication - genes break apart at a rate of kr where r is the recombination rate.

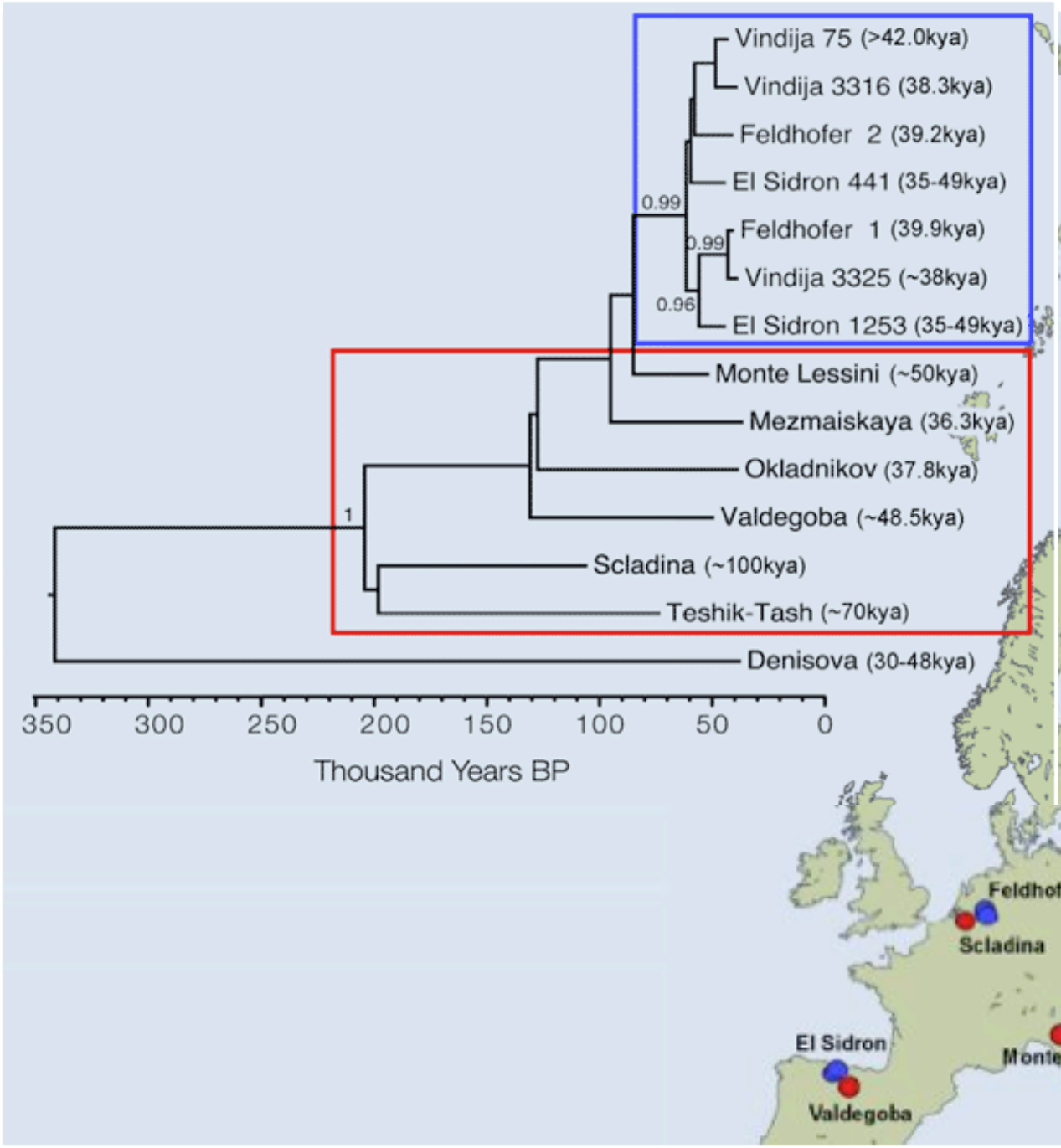


Recombination induces splits in an ancestral line as we look backwards in time.

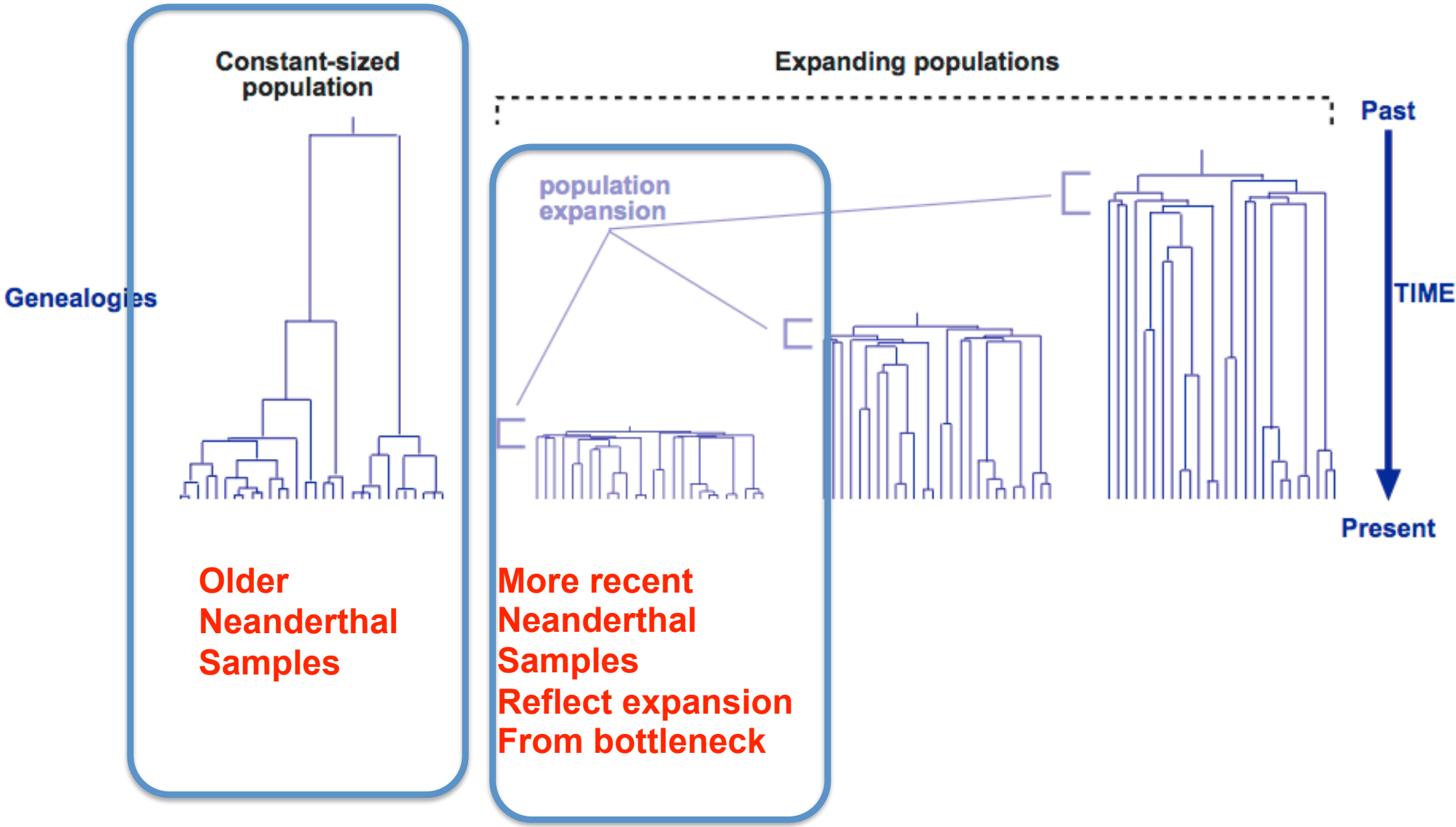
Coal.	Recomb.	Anc. mat.
1	ρ	$\frac{1}{2}\rho+\lambda$
3	$\frac{3}{2}\rho$	$\frac{1}{2}\rho+\lambda$
6	2ρ	$\rho+\lambda$
10	$\frac{5}{2}\rho$	ρ
6	2ρ	ρ
3	$\frac{3}{2}\rho$	ρ
1	ρ	ρ

An example of a coalescent tree with recombination.
Note: We can also few this as a series of trees arrayed along the chromosome.

Neandertal population structure



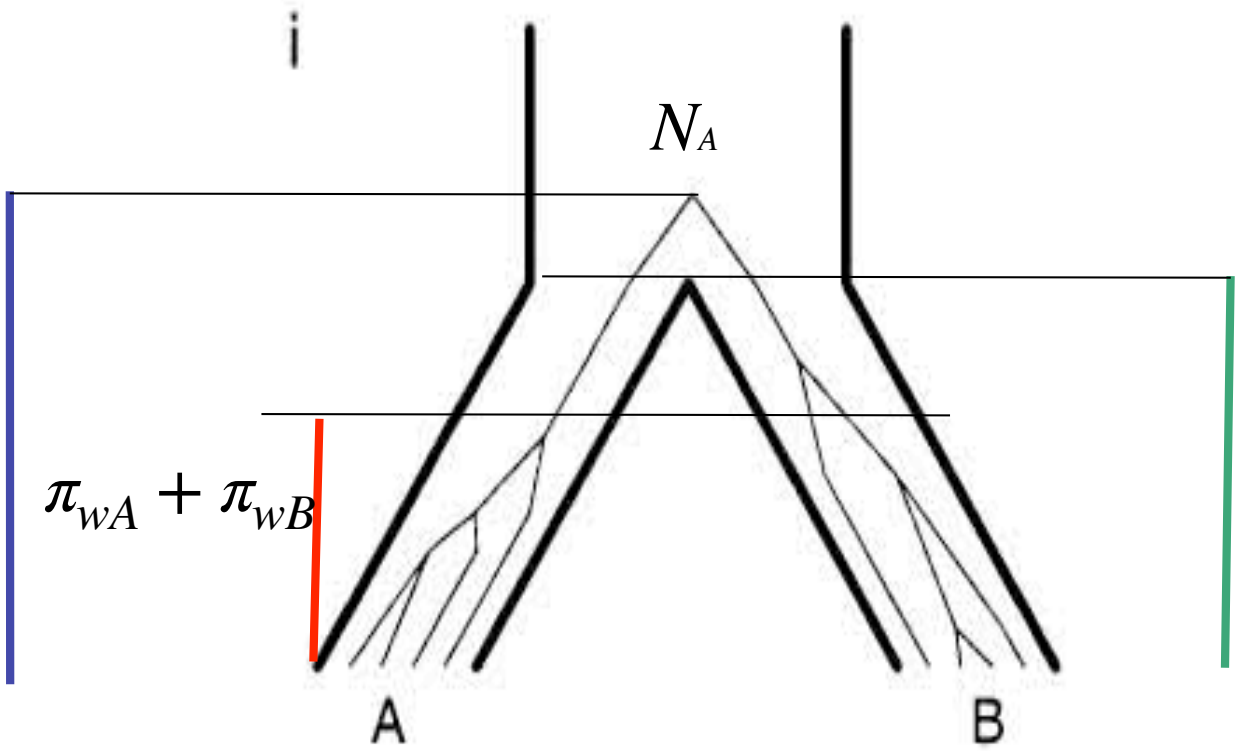
Shape of gene tree reflects different histories



Estimating Divergence time with Nei and Li's π-net

π_b

(aka D_{xy})
ave. pairwise diffs betw populations



π_{net}

$$\pi_{net} = \pi_b - (\pi_{wA} + \pi_{wB})/2$$

Simulation Assignment

Moving away from the forward time SliM (for now), here we'll focus again on msPrime, a state of the art backwards coalescent simulator

go to https://github.com/DRL/SMBE-SGE-2019
go back to "Session1"
2.Introduction_to_msprime.ipynb

Rest of instructions in email to google group

Next Week Marcelo and PipeMaster

preview PipeMaster tutorial

https://github.com/compphylo/compphylo.github.io/blob/master/Oslo2019/PM_files/Dermatonotus_example.md