CS 68: BIOINFORMATICS

Prof. Sara Mathieson
Swarthmore College
Spring 2018

Outline: Mar 26

- Recap the Coalescent
- Using the coalescent to detect deviations from neutrality
- Tajima's D test statistic

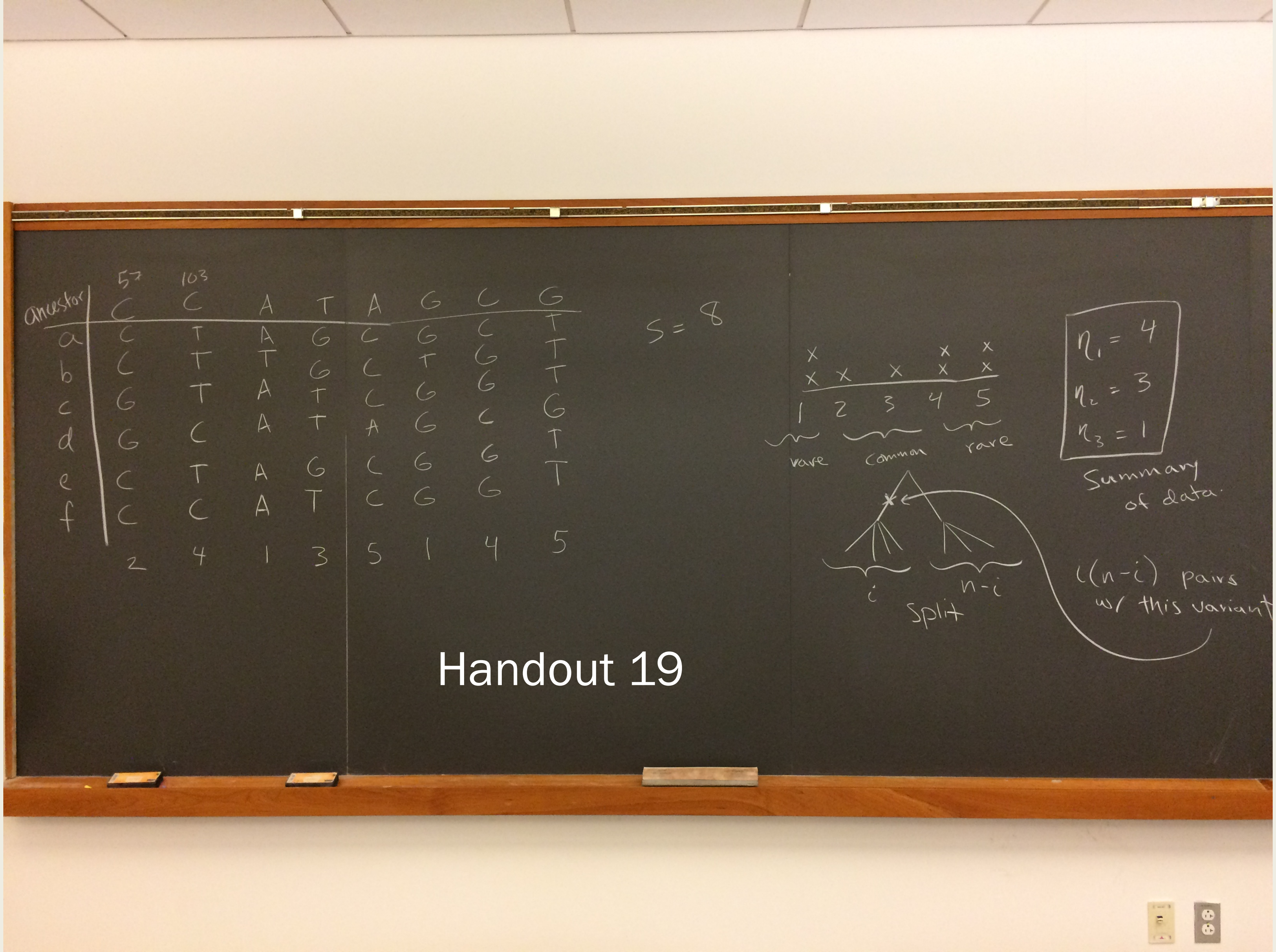
Notes:

- Office hours TODAY 3-5pm
- Handout 20, 1(a) error: y cannot be 0

Logistic Notes

- cc your partner when communicating with me about the lab
- I have been getting a lot of great & duplicate questions over email – unless your question applies only to a specific issue with your code, use Piazza!
- Please be on time to class and lab, not only affects you but your partner as well
- Let me know if you have any partner issues

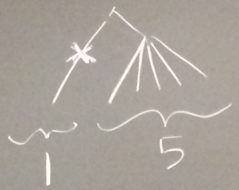
Finish Handout 19


57
103
ancestor C C A T A G C G
a C T A G C G C T
b C T T G C T G T
c G T A T C G G T
d G C A T A G C G
e C T A G C G G T
f C C A T C G G T
2 4 1 3 5 1 4 5
S = 8
1 2 3 4 5
rare
common
rare
$n_1 = 4$
$n_2 = 3$
$n_3 = 1$
Summary of data.
i
n-i
Split
i(n-i) pairs w/ this variant

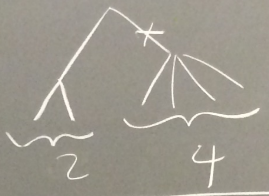
Handout 19

Handout 19

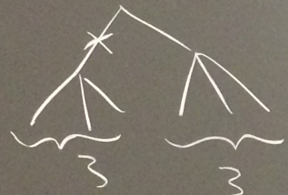
$n_1 = 4$ "1/5 split"



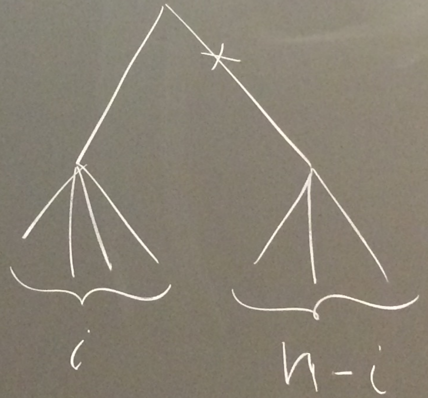
$n_2 = 3$ "2/4 split"



$n_3 = 1$ "3/3 split"



$\boxed{n = 6}$



$$S = \sum_{i=1}^{\lfloor n/2 \rfloor} n_i$$

$$\boxed{\pi = \frac{1}{\binom{n}{2}} \sum_{i=1}^{\lfloor n/2 \rfloor} \underbrace{i(n-i)}\, n_i}$$

pairs that have one differen

mutation w/ "$i/(n-i)$" split

erence

π = avg. # of pairwise differences

$$\pi = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} k_{ij}$$

# of differences between i & j

$$\boxed{k_{ef} = 2}$$

Recap the Coalescent

Coalescent Theory

- The Coalescent (usually attributed to Kingman, 1982) is a mathematical model for the evolution and genealogical history of a population

Coalescent Theory

- The Coalescent (usually attributed to Kingman, 1982) is a mathematical model for the evolution and genealogical history of a population

- The Coalescent can be derived from the Wright-Fisher model, but also several other discrete-time models (i.e. the Moran model)

Coalescent Theory

- The Coalescent (usually attributed to Kingman, 1982) is a mathematical model for the evolution and genealogical history of a population

- The Coalescent can be derived from the Wright-Fisher model, but also several other discrete-time models (i.e. the Moran model)

- We assume the population size N is large

Coalescent Theory

- The Coalescent (usually attributed to Kingman, 1982) is a mathematical model for the evolution and genealogical history of a population

- The Coalescent can be derived from the Wright-Fisher model, but also several other discrete-time models (i.e. the Moran model)

- We assume the population size N is large

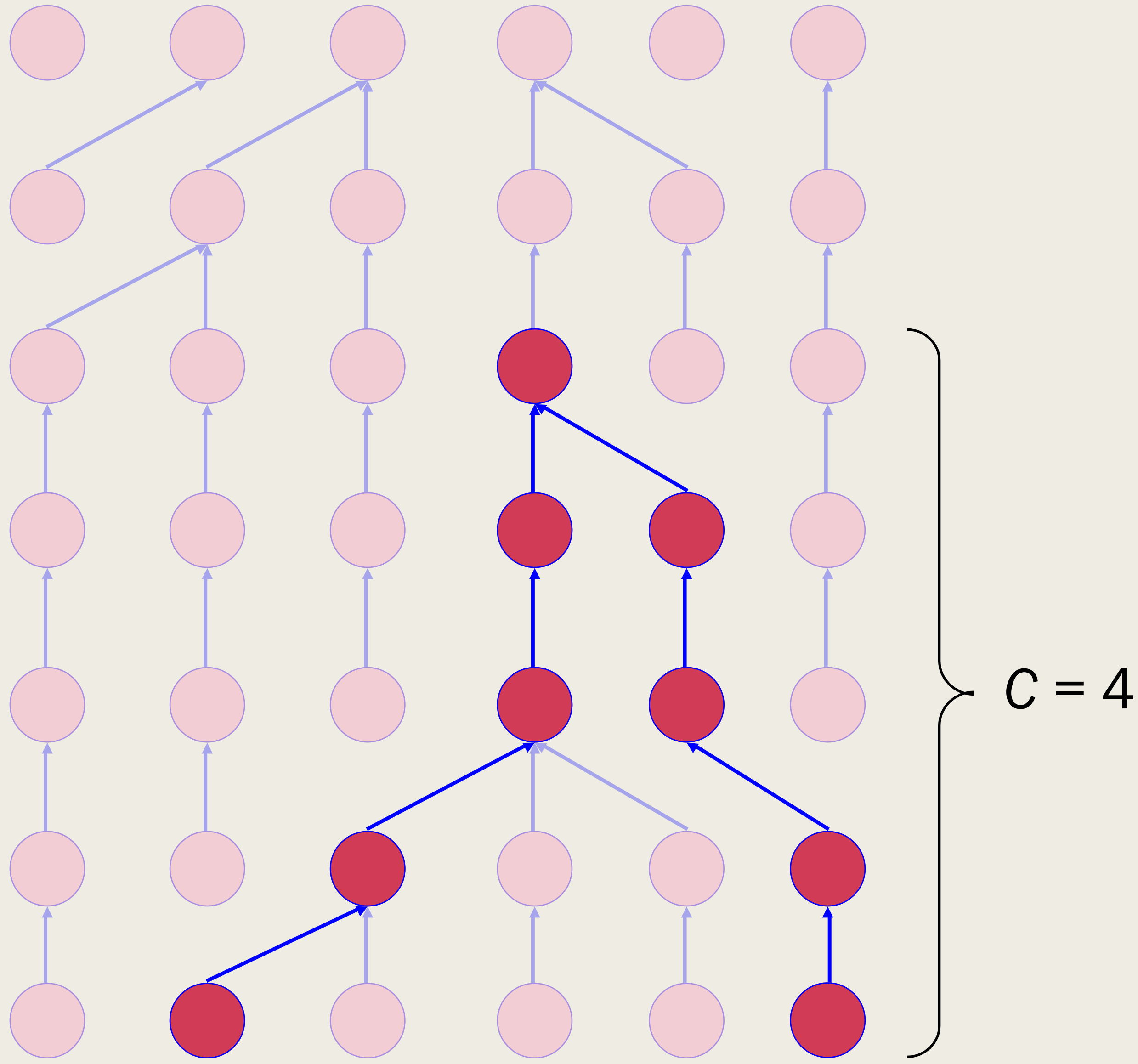
- We rescale time where 1 unit in coalescent time = $2N$ generations

Coalescent Theory

- The Coalescent (usually attributed to Kingman, 1982) is a mathematical model for the evolution and genealogical history of a population
- The Coalescent can be derived from the Wright-Fisher model, but also several other discrete-time models (i.e. the Moran model)
- We assume the population size N is large
- We rescale time where 1 unit in coalescent time = $2N$ generations
- Rescaling time allows us to work with numbers that are on order 1 (avoiding numerical issues that arise with very small numbers) and we also avoid a factor of $2N$ in every formula

Coalescent derivation from the Wright-Fisher model

Probability two samples *coalesce* after g generations:



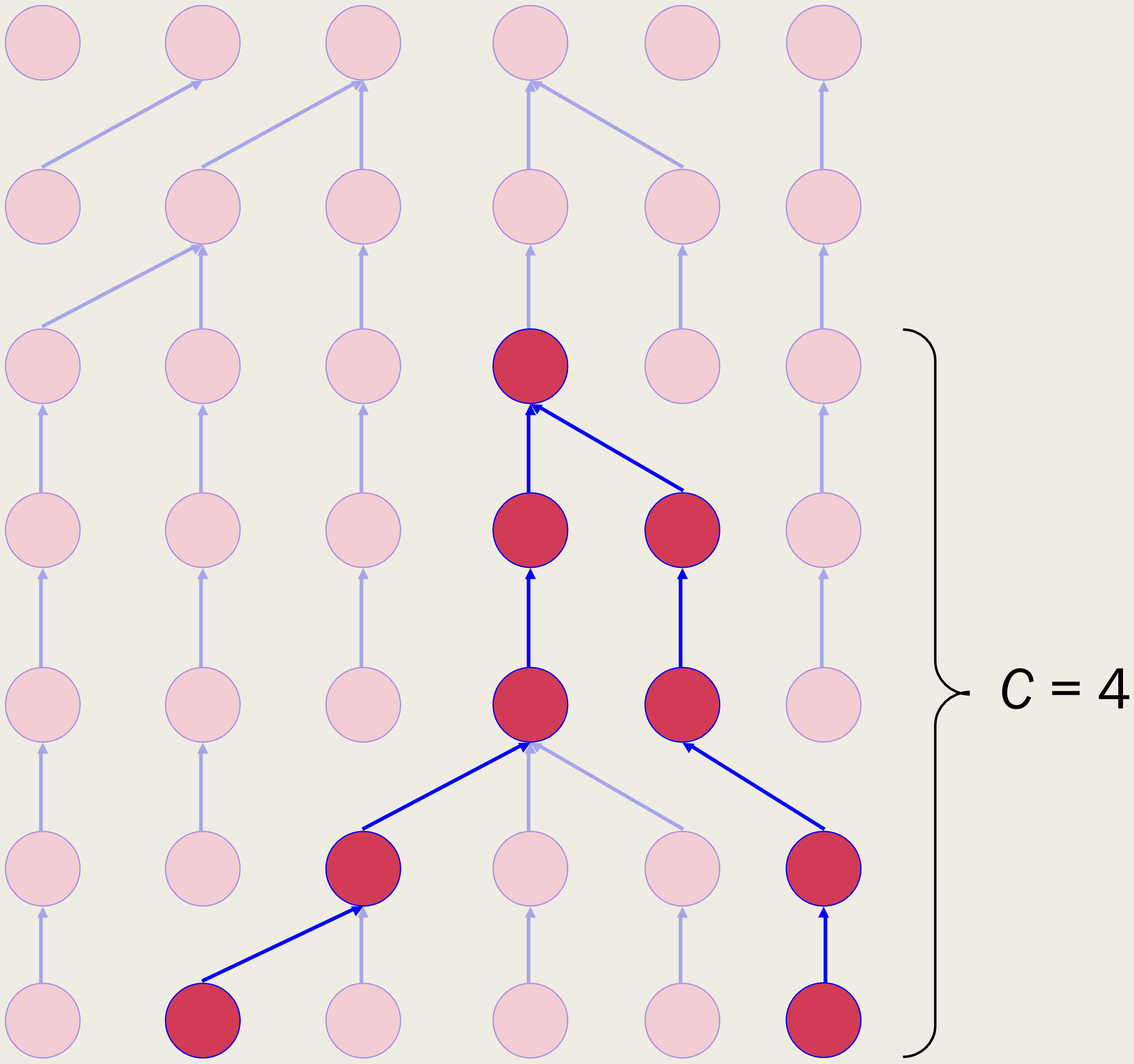
Population size $2N=6$, sample size $n = 2$

Coalescent derivation from the Wright-Fisher model

Probability two samples *coalesce* after g generations:

$$P_C(g) = \left(1 - \frac{1}{2N}\right)^{g-1} \frac{1}{2N}$$

[Geometric distribution]



Population size $2N$=6, sample size n = 2

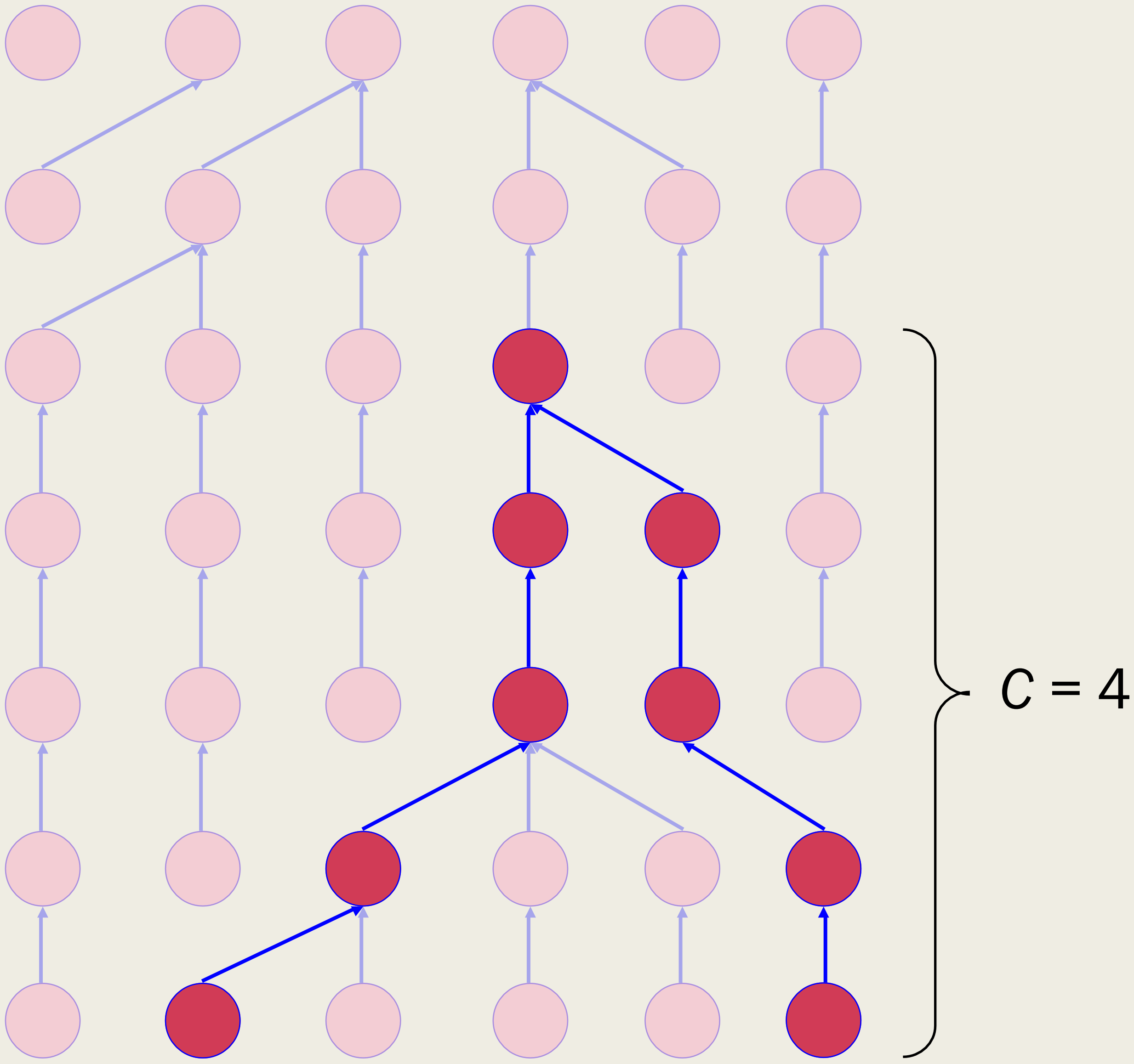
Coalescent derivation from the Wright-Fisher model

Probability two samples *coalesce* after g generations:

$$P_C(g) = \left(1 - \frac{1}{2N}\right)^{g-1} \frac{1}{2N}$$

Don't choose the same parent for g-1 generations

[Geometric distribution]



Population size $2N$=6, sample size n = 2

Coalescent derivation from the Wright-Fisher model

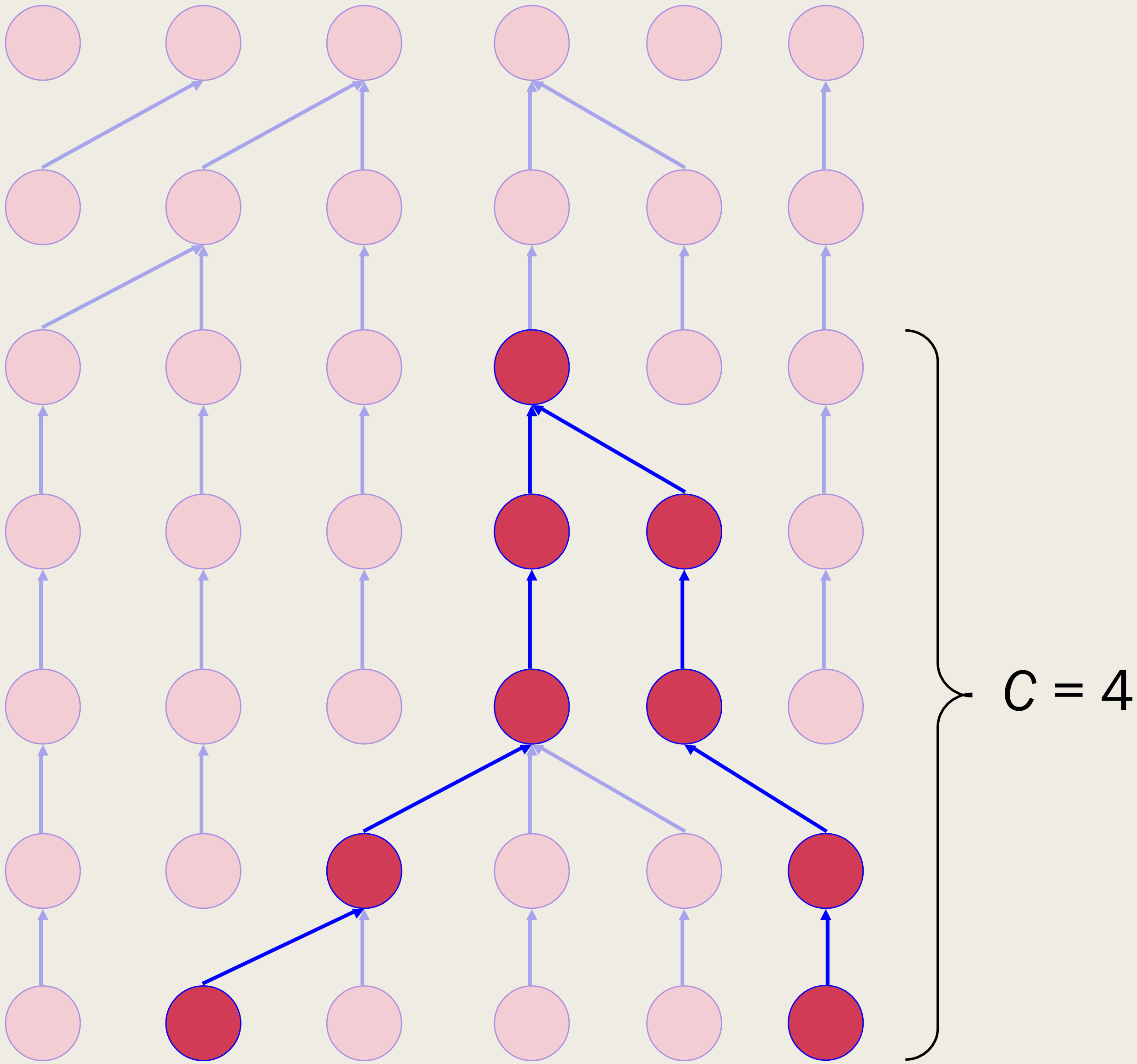
Probability two samples *coalesce* after *g* generations:

$$P_C(g) = \left(1 - \frac{1}{2N}\right)^{g-1} \frac{1}{2N}$$

Don't choose the same parent for *g*-1 generations

Choose same parent in the g^{th} generation

[Geometric distribution]



Population size $2N$=6, sample size n = 2

Coalescent derivation from the Wright-Fisher model

- We will make use of the Taylor series for e^{-x} around $x = 0$:

$$e^{-x} = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$$

- We will only use the first 2 terms:

$$e^{-x} \approx 1 - x$$

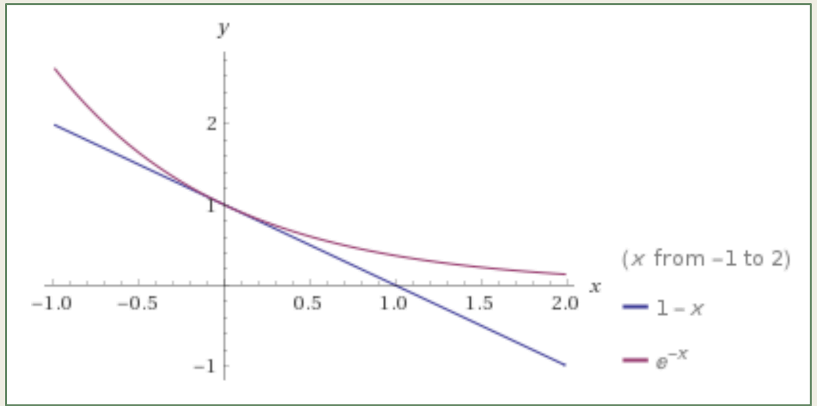
Coalescent derivation from the Wright-Fisher model

- We will make use of the Taylor series for e^{-x} around $x = 0$:

$$e^{-x} = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$$

- We will only use the first 2 terms:

$$e^{-x} \approx 1 - x$$



Created using WolframAlpha

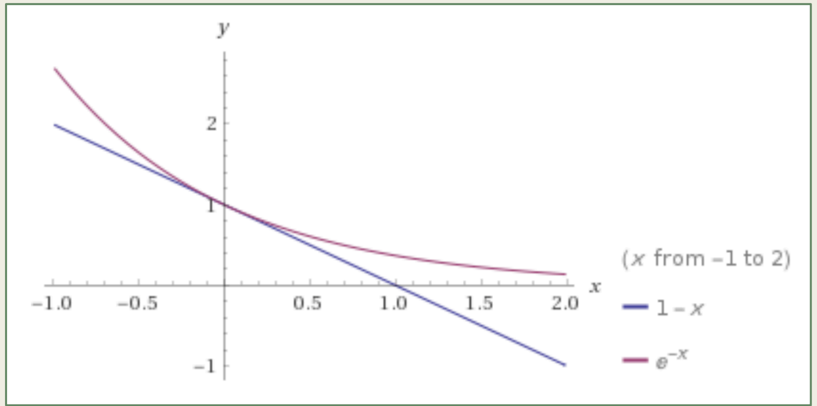
Coalescent derivation from the Wright-Fisher model

- We will make use of the Taylor series for e^{-x} around $x = 0$:

$$e^{-x} = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$$

- We will only use the first 2 terms:

$$e^{-x} \approx 1 - x$$



Created using WolframAlpha

- This allows us to rewrite our geometric coalescent probability

$$P_C(g) = \left(1 - \frac{1}{2N}\right)^{g-1} \frac{1}{2N}$$

- as (drop the -1 since *g* is large):

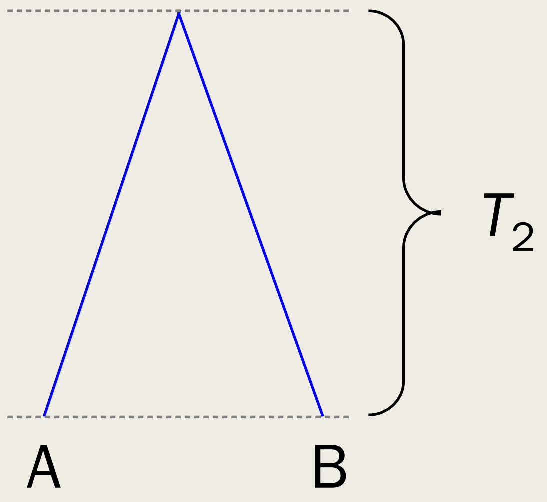
$$P_C(g) \approx \frac{1}{2N} e^{-\frac{g}{2N}}$$

Correction!

Coalescent for $n = 2$

- We let 1 coalescent unit = $2N$ generations, and let our new variable be t

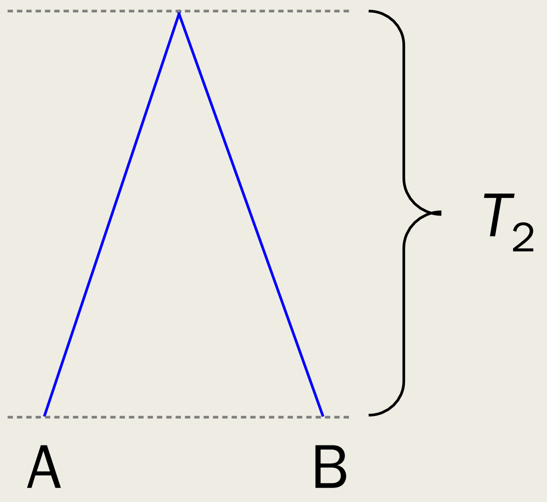
- We let T_i be a random variable representing the time when there are i lineages



Coalescent for $n = 2$

$$P_{T_2}(t) = e^{-t}$$

- We let 1 coalescent unit = $2N$ generations, and let our new variable be t
- We let T_i be a random variable representing the time when there are i lineages
- For n=2, this gives us an exponential distribution with parameter 1

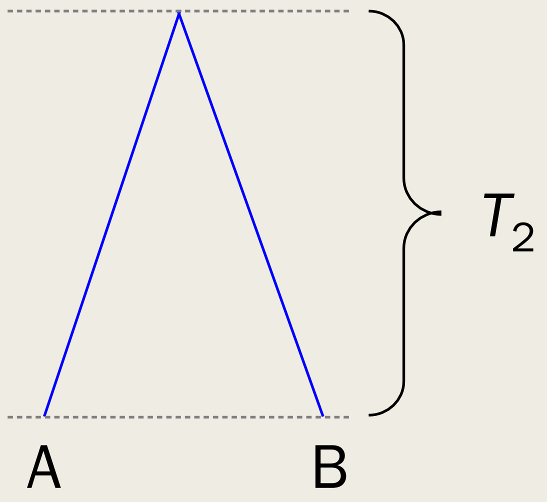


Coalescent for n = 2

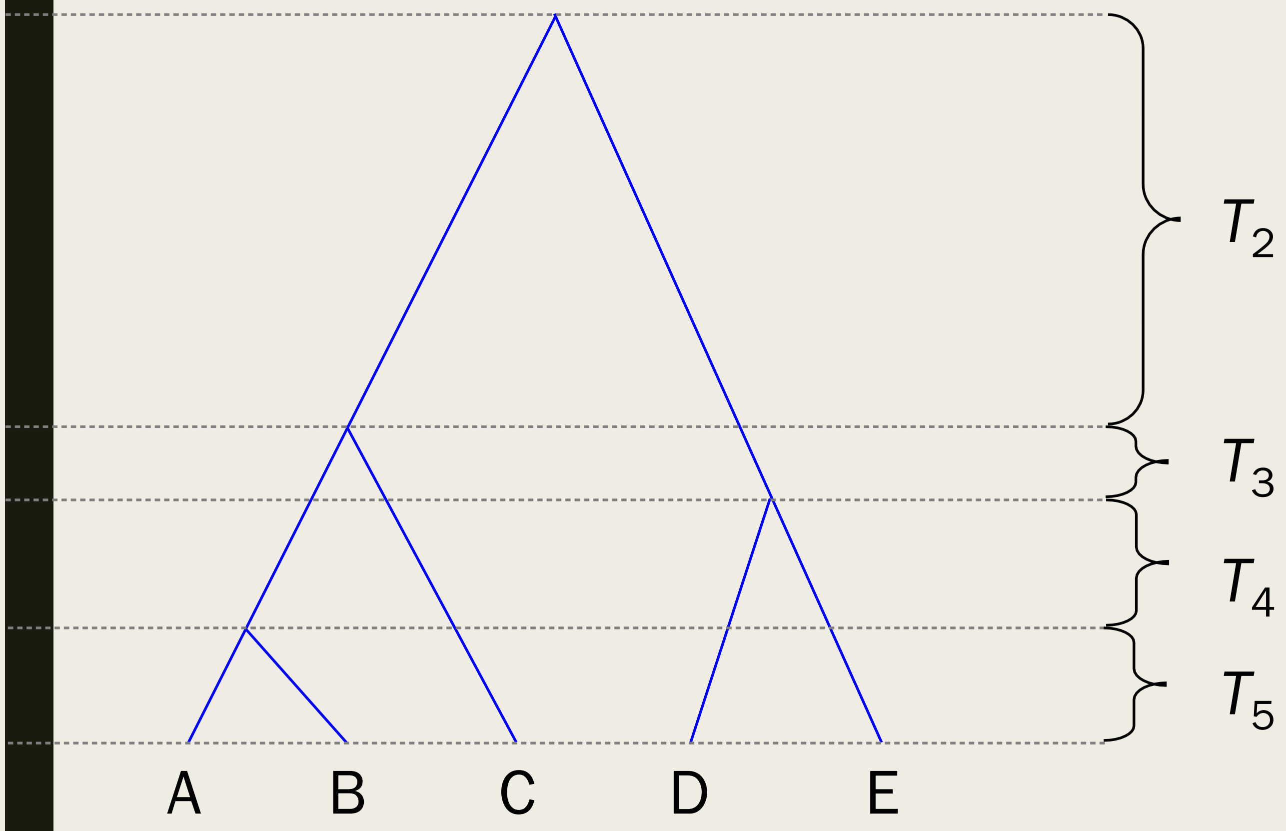
- We let 1 coalescent unit = 2N generations, and let our new variable be t
- We let T_i be a random variable representing the time when there are i lineages
- For n=2, this gives us an exponential distribution with parameter 1
- The expected time for 2 lineages to coalesce is 1 coalescent unit of time => 2N generations

$$P_{T_2}(t) = e^{-t}$$

$$E[T_2] = \int_0^\infty te^{-t}dt = 1$$



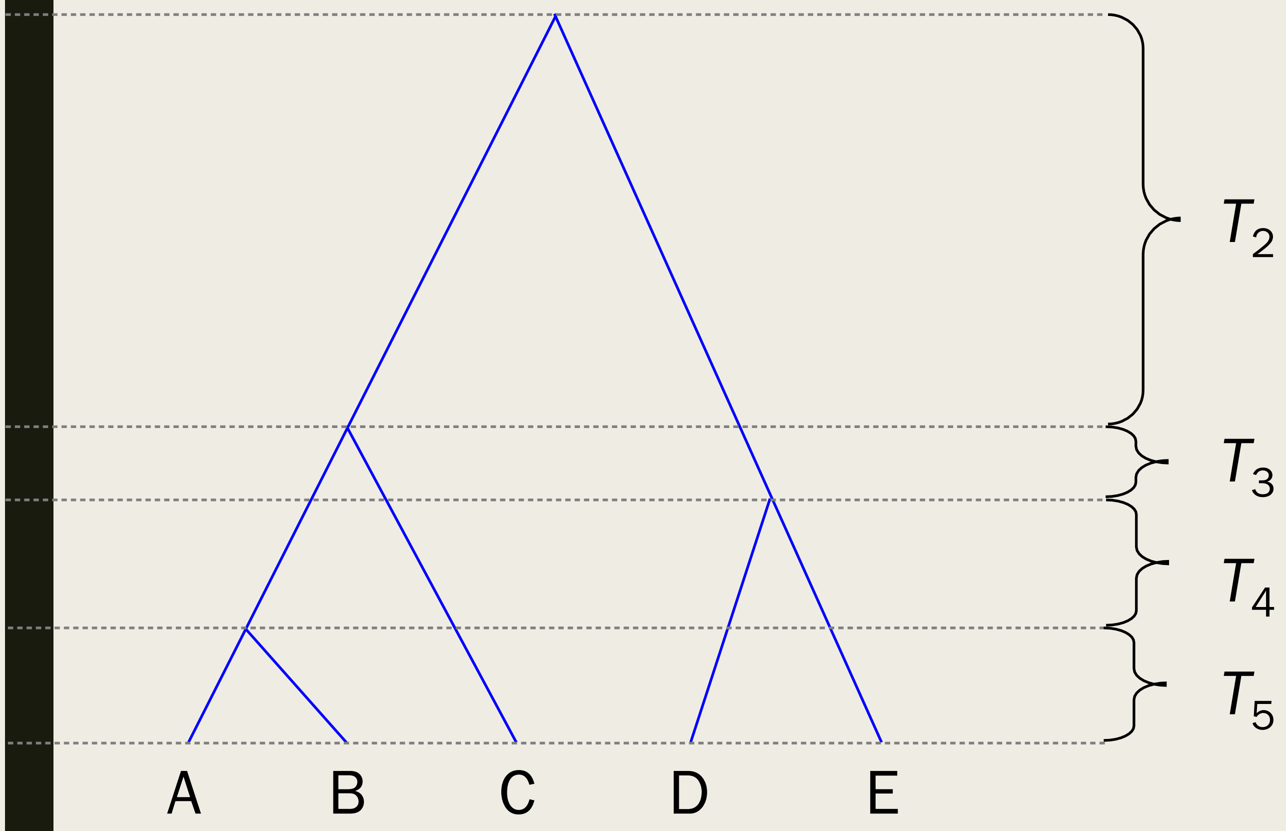
The Coalescent



- The larger our sample size n, the more pairs we have that can coalesce right away
- In general, the time when there are *i* lineages is also exponentially distributed with parameter *i*(*i*-1)/2 (*i* “choose” 2)

$$P_{T_i}(t) = \binom{i}{2} e^{-\binom{i}{2} t}$$

The Coalescent

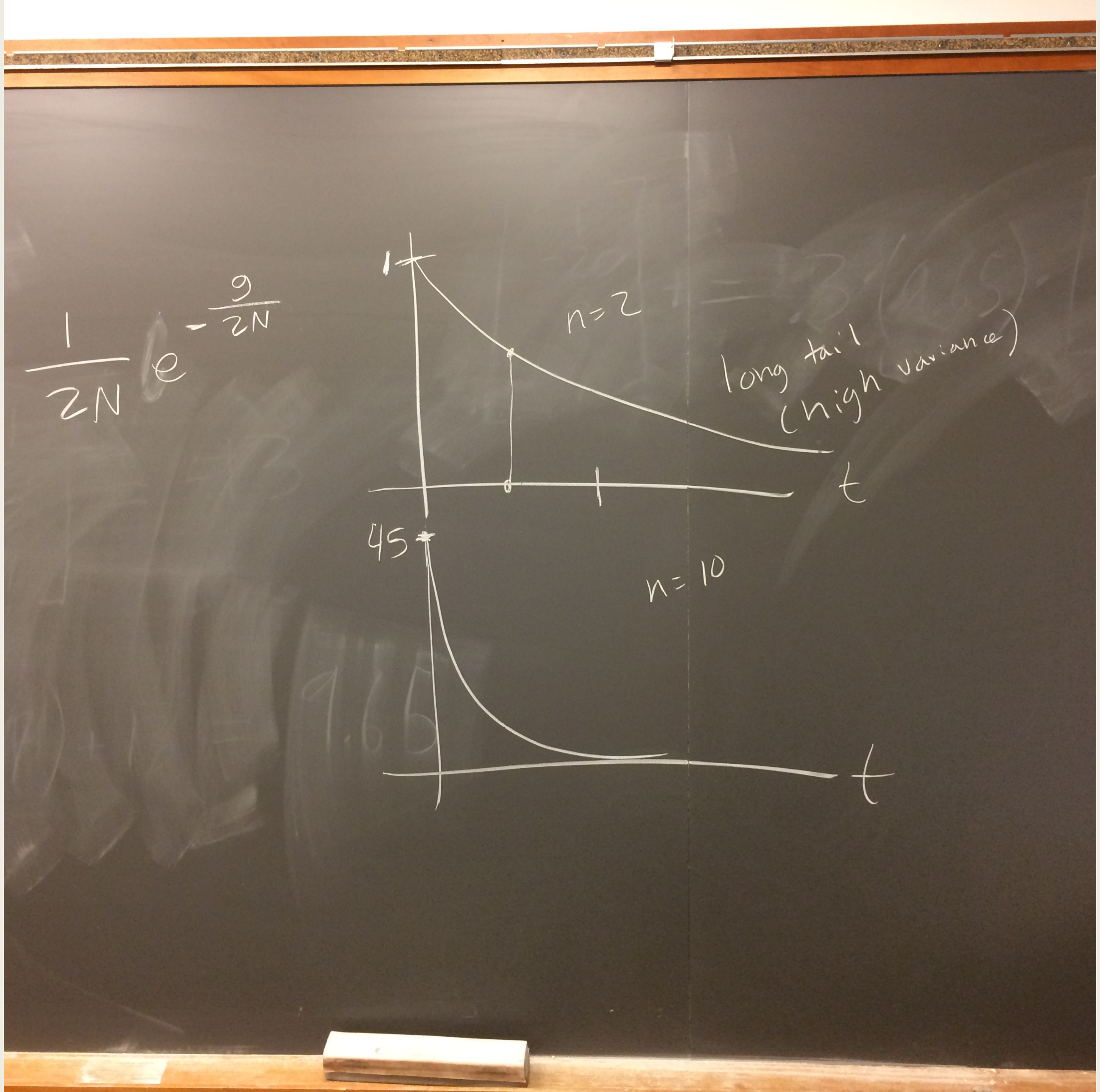


- The larger our sample size n, the more pairs we have that can coalesce right away
- In general, the time when there are *i* lineages is also exponentially distributed with parameter *i*(*i*-1)/2 (*i* "choose" 2)

$$P_{T_i}(t) = \binom{i}{2} e^{-\binom{i}{2} t}$$

- Expected value (think: weighted average, mean)

$$E[T_i] = \int_0^\infty t \binom{i}{2} e^{-\binom{i}{2} t} dt = \frac{1}{\binom{i}{2}}$$


$\frac{1}{2N} e^{-\frac{g}{2N}}$
1
n=2
long tail
(high variance)
t
45
n=10
t

Deviations from neutrality: Tajima’s D

Tajima’s D

- We often say a site/locus is “**neutral**” if it has no positive or negative effect on **fitness**

- More generally, “neutral” means agreeing with our Wright-Fisher model assumptions (constant population size, mutations have no consequences, random mating, etc)

Tajima’s D

- We often say a site/locus is “neutral” if it has no positive or negative effect on fitness

- More generally, “neutral” means agreeing with our Wright-Fisher model assumptions (constant population size, mutations have no consequences, random mating, etc)

- Deviations from neutrality could mean that any of these assumptions are wrong

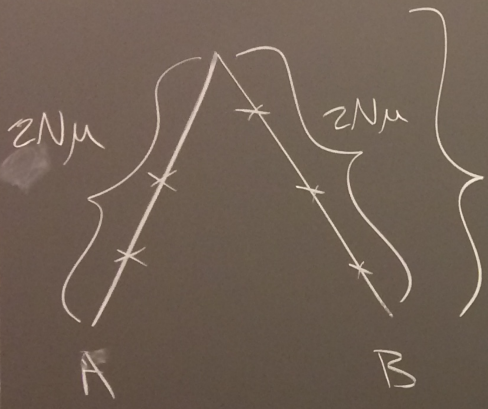
- We will focus on two of them: allowing variable population size and allowing mutations with different selective advantages/disadvantages

Tajima’s D

- We often say a site/locus is “**neutral**” if it has no positive or negative effect on **fitness**
- More generally, “neutral” means agreeing with our Wright-Fisher model assumptions (constant population size, mutations have no consequences, random mating, etc)
- Deviations from neutrality could mean that any of these assumptions are wrong
- We will focus on two of them: allowing variable population size and allowing mutations with different selective advantages/disadvantages
- Tajima’s D (1989) is a test statistic that compares different measures of sequence diversity that should be the same under neutrality
- If they are not the same, we can further investigate the causes

μ = mutation rate per base per generation

<u>humans</u>: $\mu = 1.25 \times 10^{-8}$ * assumption



$E[T_2] = 1$ or $2N$ generations

$$E[k_{AB}] = 4N\mu = \theta$$

$$\boxed{E[\pi] = 4N\mu = \theta}$$

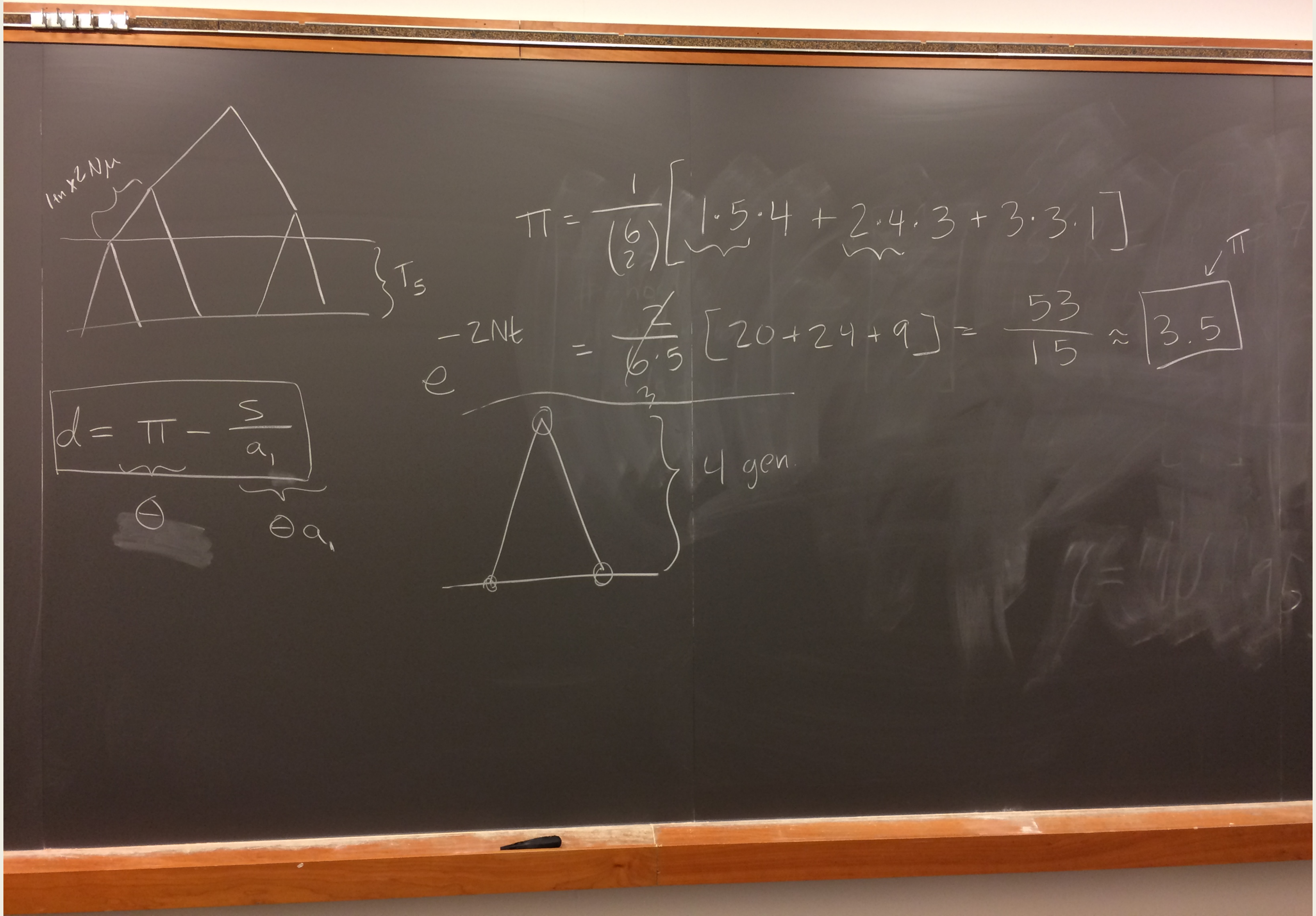
$$E[S] = \{\text{total branch length}\} \cdot 2N\mu$$

$$E[S] = \left(\sum_{i=n}^{2} E[T_i] \cdot i\right) 2N\mu$$

$$= \left(\sum \frac{2}{\cancel{i}(i-1)} \cancel{i}\right) 2N\mu$$

$$E[S] = \underbrace{\left(\sum_{i=1}^{n-1} \frac{1}{i}\right)}_{a_1} \underbrace{4N\mu}_{\theta}$$

This should be a 4, not a 2!


1m x 2Nμ
T5
π = 1/(6 2) [1·5·4 + 2·4·3 + 3·3·1]
= 2/(6·5) [20+24+9] = 53/15 ≈ 3.5
π
e^-2Nt
4 gen.
d = π − S/a1
θ
θa1