CS 68: BIOINFORMATICS

Prof. Sara Mathieson
Swarthmore College
Spring 2018

Outline: Mar 21

- Neutral theory of evolution
- Measures of sequence diversity
- Probability distributions and expected value
- The Coalescent

Notes:

- Office hours today are canceled
- Population genetics reading is posted
- Interested in being a ninja? Let me know
- Midterms will be handed back on Friday

Recap last time

Recap last time

- Wright-Fisher model of evolution; discrete time (measured in generations)

Recap last time

- Wright-Fisher model of evolution; discrete time (measured in generations)

- Assumptions (for now):
  - *constant population size*
  - *random mating*
  - *the two chromosomes for each individual choose their parents independently*
  - *mutations are* ***neutral*** *(i.e. not selectively advantageous or deleterious)*

Recap last time

- **Wright-Fisher** model of evolution; discrete time (measured in generations)

- **Assumptions** (for now):
  - *constant population size*
  - *random mating*
  - *the two chromosomes for each individual choose their parents independently*
  - *mutations are **neutral** (i.e. not selectively advantageous or deleterious)*

- **Genetic drift**: changes in allele frequencies are due to random chance, not selection

Recap last time

- All neutral genetic variation will eventually die out or become fixed in the population (question: so why do we observe variation?)

- The probability of fixation for a new mutation is $1/(2N)$ where N is the population size

- In general the fixation probability is f_0, the initial frequency of the mutation in generation 0

- Question: how is genetic drift affected by the population size N? What consequences might this affect have?

Recap last time

- All neutral genetic variation will eventually die out or become fixed in the population (question: so why do we observe variation?)

Intermediate frequencies can persist for many generations, selection, admixture, any deviations from neutrality

- The probability of fixation for a new mutation is $1/(2N)$ where N is the population size

- In general the fixation probability is f_0, the initial frequency of the mutation in generation 0

- Question: how is genetic drift affected by the population size N? What consequences might this affect have?

Recap last time

- All neutral genetic variation will eventually die out or become fixed in the population (question: so why do we observe variation?)

Intermediate frequencies can persist for many generations, selection, admixture, any deviations from neutrality

- The probability of fixation for a new mutation is $1/(2N)$ where N is the population size

- In general the fixation probability is f_0, the initial frequency of the mutation in generation 0

- **Question: how is genetic drift affected by the population size N? What consequences might this affect have?**

The lower the population size, the greater the chance new mutations will fix, even weakly deleterious ones. This can lead to what would typically be rare traits reaching high frequency.

Brief detour to Hardy-Weinberg

Hardy-Weinberg expectations

- If we have two alleles, A and a, then each individual can have *genotype* AA, Aa, or aa

Hardy-Weinberg expectations

- If we have two alleles, A and a, then each individual can have *genotype* AA, Aa, or aa

- We say that AA and aa are *homozygous* and Aa is *heterozygous*

Hardy-Weinberg expectations

- If we have two alleles, A and a, then each individual can have *genotype* AA, Aa, or aa

- We say that AA and aa are *homozygous* and Aa is *heterozygous*

- If the genotype at this *locus* (site) is responsibly for a *Mendelian* (think: binary) *phenotype* and A is *dominant*, then AA and Aa will have the same phenotype

Hardy-Weinberg expectations

- If we have two alleles, A and a, then each individual can have *genotype* AA, Aa, or aa

- We say that AA and aa are *homozygous* and Aa is *heterozygous*

- If the genotype at this *locus* (site) is responsibly for a *Mendelian* (think: binary) *phenotype* and A is *dominant*, then AA and Aa will have the same phenotype

- In that case we would call aa recessive

Hardy-Weinberg expectations

- If we have two alleles, A and a, then each individual can have *genotype* AA, Aa, or aa

- We say that AA and aa are *homozygous* and Aa is *heterozygous*

- If the genotype at this *locus* (site) is responsibly for a *Mendelian* (think: binary) *phenotype* and A is *dominant*, then AA and Aa will have the same phenotype

- In that case we would call aa recessive

- If aa is disease causing or *deleterious*, this can reduce the frequency of a through selection

Hardy-Weinberg expectations

- If we have two alleles, A and a, then each individual can have ***genotype*** AA, Aa, or aa

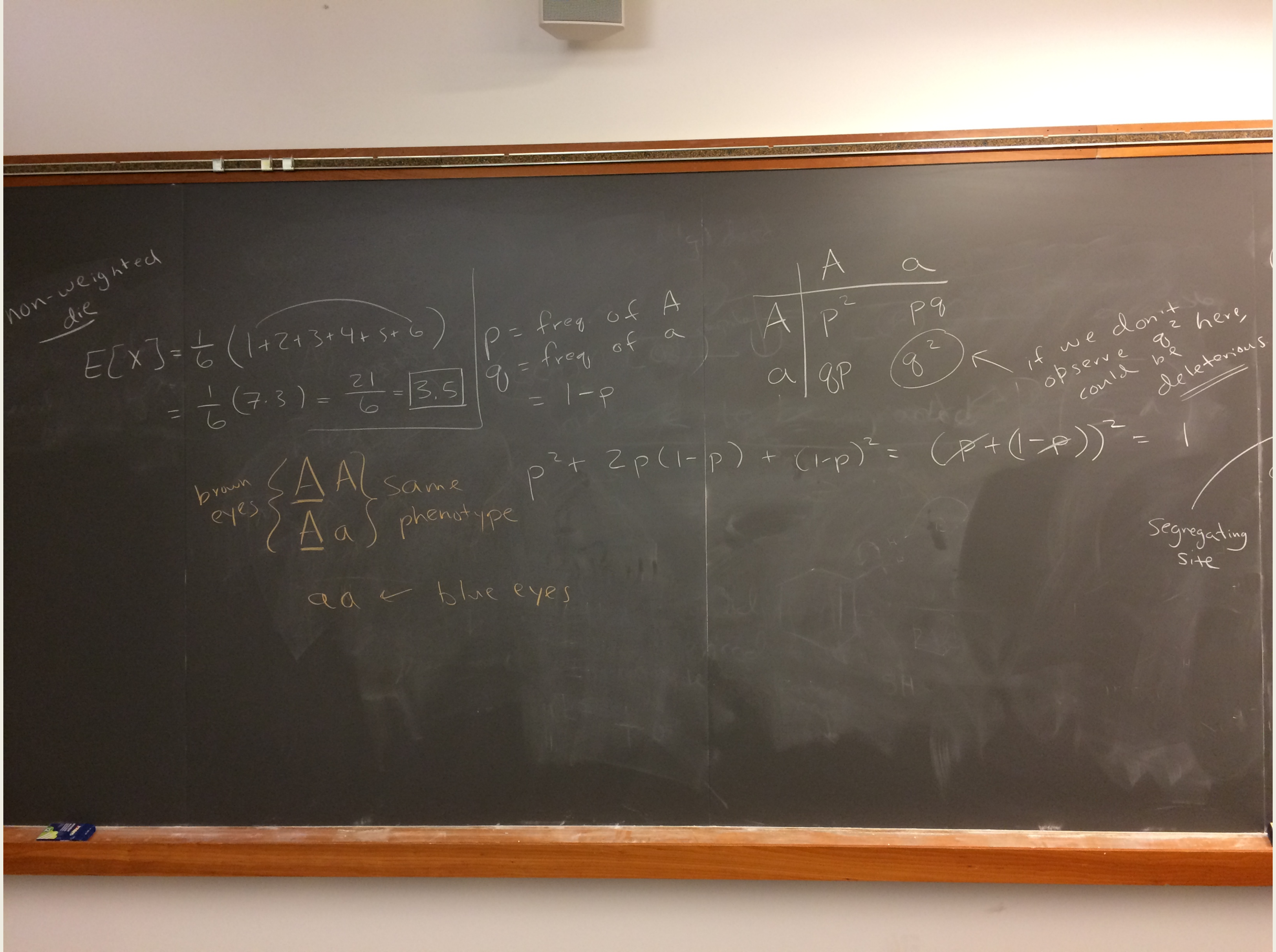
- We say that AA and aa are ***homozygous*** and Aa is ***heterozygous***

- If the genotype at this *locus* (site) is responsibly for a ***Mendelian*** (think: binary) ***phenotype*** and A is ***dominant***, then AA and Aa will have the same phenotype

- In that case we would call aa recessive

- If aa is disease causing or *deleterious*, this can reduce the frequency of a through selection

- If most alleles either become fixed or die out, that means eventually everyone will either be aa or AA. This is called the *loss of heterozygosity*



non-weighted die

$$E[X] = \frac{1}{6}(1+2+3+4+5+6)$$

$$= \frac{1}{6}(7 \cdot 3) = \frac{21}{6} = \boxed{3.5}$$

p = freq of A

q = freq of a

= 1-p

	A	a
A	p^2	pq
a	qp	q^2

if we don't observe q^2 here, could be deleterious

$$p^2 + 2p(1-p) + (1-p)^2 = (\cancel{p} + (1-\cancel{p}))^2 = 1$$

brown eyes $\left\{ \begin{array}{l} \underline{A}A \\ \underline{A}a \end{array} \right\}$ same phenotype

aa ← blue eyes

Segregating site

Neutral theory of evolution

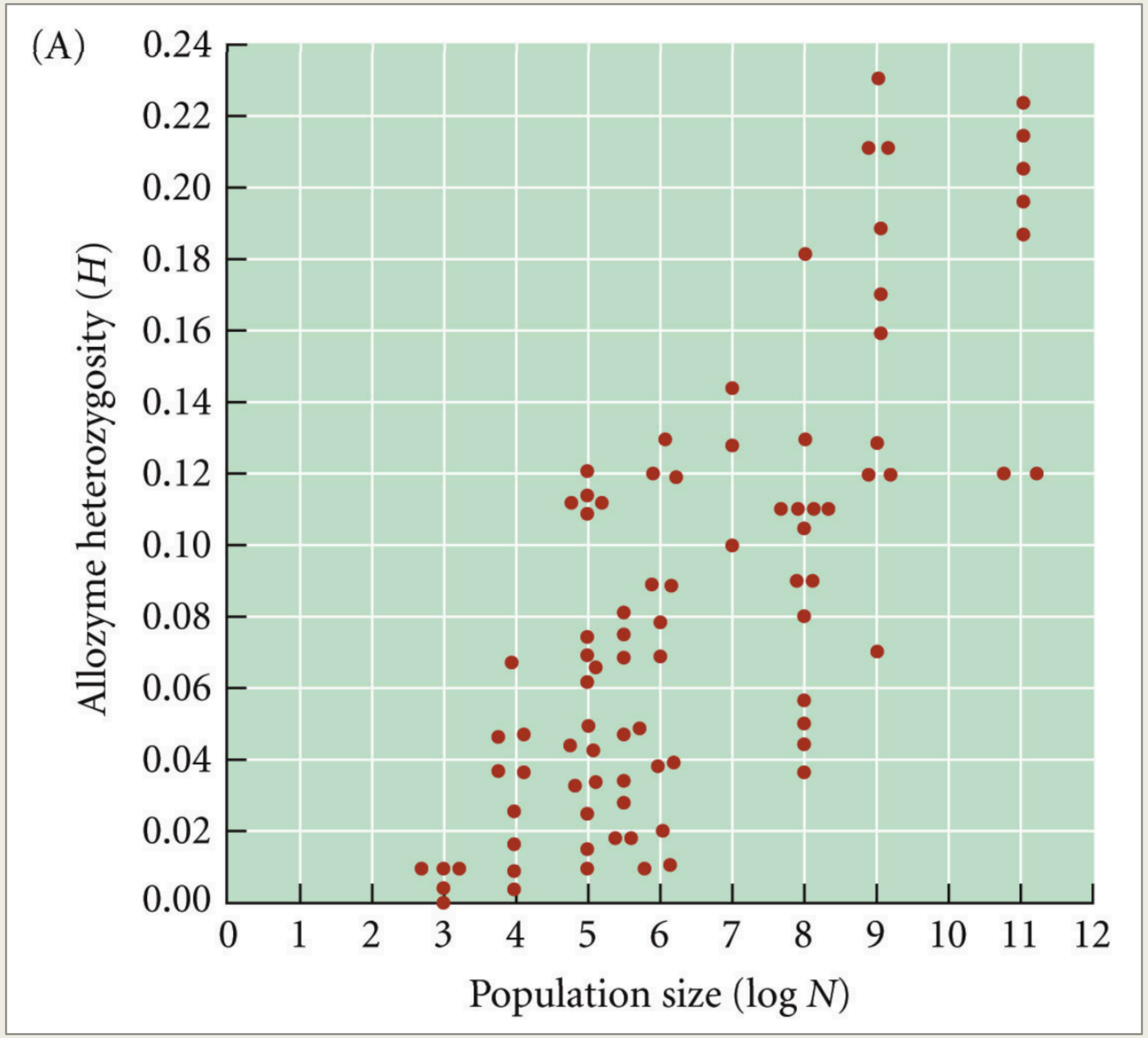
Neutral theory of evolution

- Kimura 1968

- Claim: most genetic variation is neutral

- Consistent with the idea that population size is responsible for the level of variation in a population

***EVOLUTION 2e*, Figure 10.9 (Part 1)**



From: Graham Coop, https://gcbias.org/2016/09/21/population-genetics-undergrad-class/

Probability distributions and expected value

Discrete probability distribution

- Let *X* be a random variable that can take on values x_1, x_2, …, x_k
- Example: a die that can take on values 1,2,3,4,5,6

Discrete probability distribution

- Let X be a random variable that can take on values $x_1, x_2, \ldots, x_k$
- Example: a die that can take on values 1,2,3,4,5,6
- If we rolled the die many times and took the average, we would have an estimate of the expected value

Discrete probability distribution

- Let X be a random variable that can take on values $x_1, x_2, \ldots, x_k$
- Example: a die that can take on values 1,2,3,4,5,6
- If we rolled the die many times and took the average, we would have an estimate of the expected value
- Let p_i = the probability of observing value x_i
- Example: p_1=0, p_2=1/6, p_3=1/6, p_4=1/6, p_5=1/6, p_6 = 1/3

Discrete probability distribution

- Let *X* be a random variable that can take on values $x_1, x_2, \ldots, x_k$
- Example: a die that can take on values 1,2,3,4,5,6
- If we rolled the die many times and took the average, we would have an estimate of the expected value
- Let p_i = the probability of observing value x_i
- Example: p_1=0, p_2=1/6, p_3=1/6, p_4=1/6, p_5=1/6, p_6 = 1/3
- We should check that the sum of the probabilities of all possible values is 1

$$\sum_{i=1}^{k} p_i = 1$$

Discrete probability distribution

- Let X be a random variable that can take on values x_1, x_2, …, x_k
- Example: a die that can take on values 1,2,3,4,5,6
- If we rolled the die many times and took the average, we would have an estimate of the expected value
- Let p_i = the probability of observing value x_i
- Example: p_1=0, p_2=1/6, p_3=1/6, p_4=1/6, p_5=1/6, p_6 = 1/3
- We should check that the sum of the probabilities of all possible values is 1

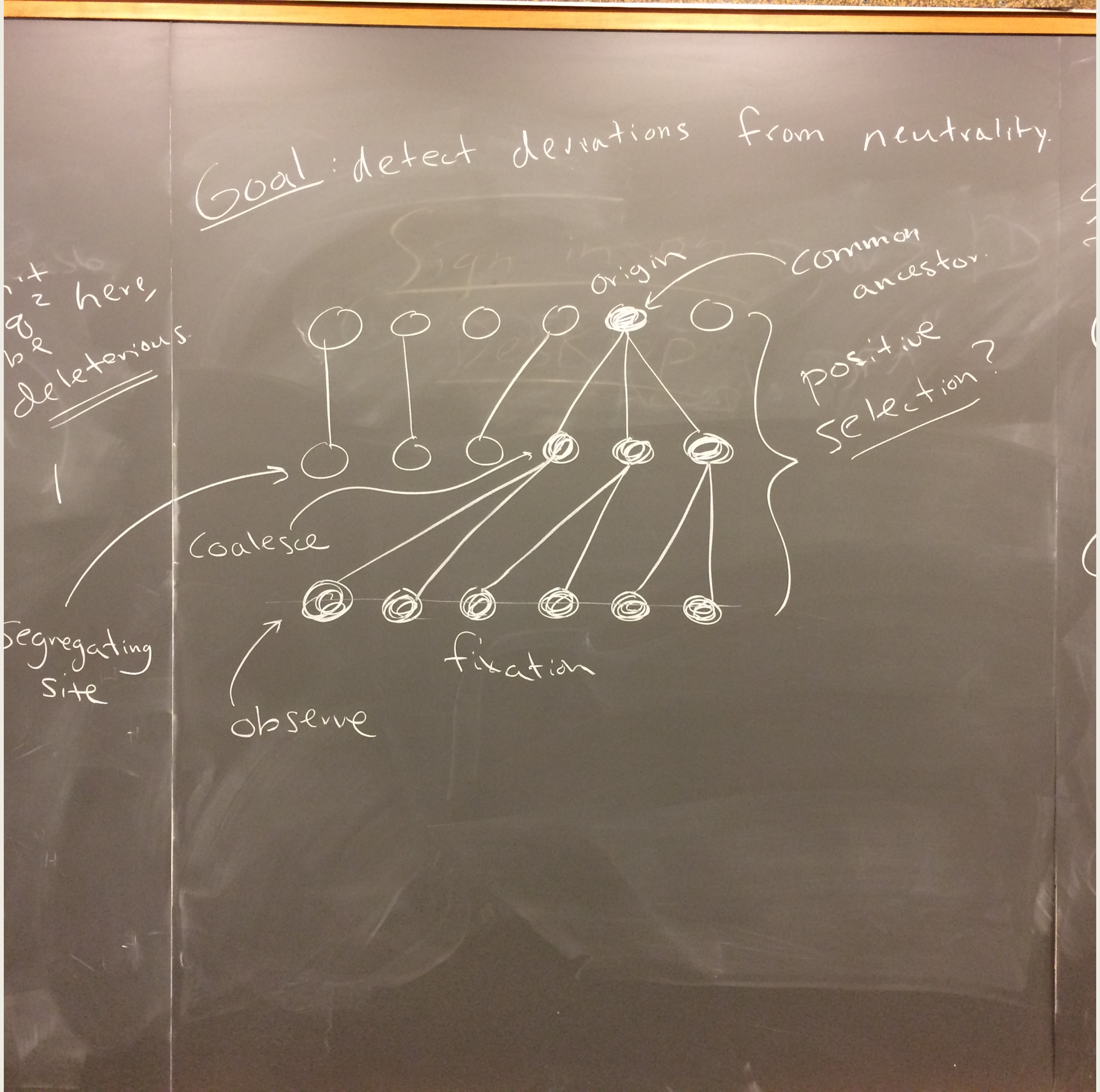
$$\sum_{i=1}^{k} p_i = 1$$

- Compute expectation:

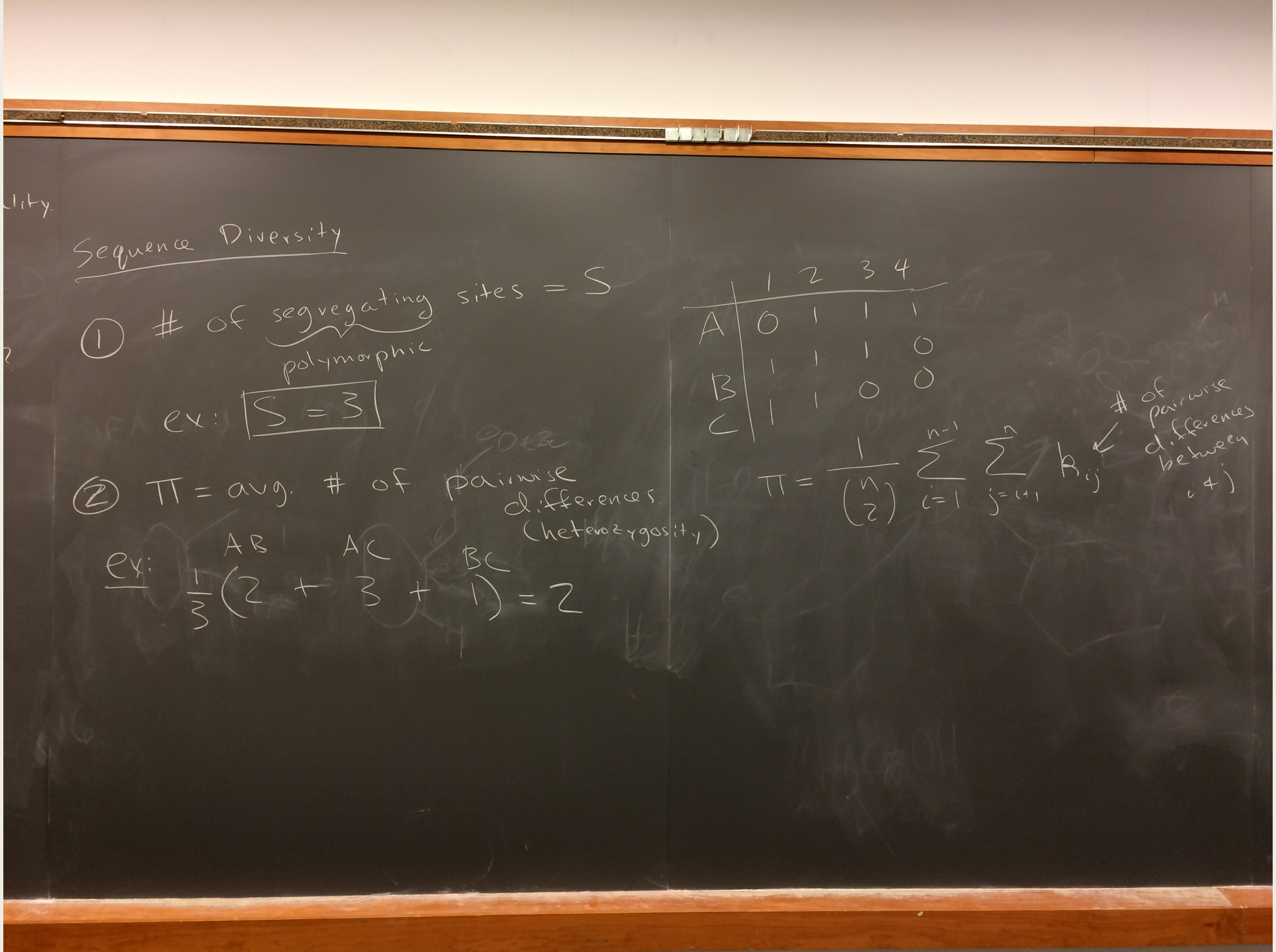
$$E[X] = p_1 x_1 + p_2 x_2 + \cdots + p_k x_k = \sum_{i=1}^{k} p_i x_i$$

$$0 \cdot 1 + \frac{1}{6}(2+3+4+5) + \frac{1}{3} \cdot 6 = 4\frac{1}{3}$$

Measures of sequence diversity



Goal: detect deviations from neutrality.



lity.

Sequence Diversity

① # of $\underbrace{\text{segregating}}_{\text{polymorphic}}$ sites = S

ex: $\boxed{S = 3}$

	1	2	3	4
A	0	1	1	1
B	1	1	1	0
C	1	1	0	0

② π = avg. # of pairwise differences (heterozygosity)

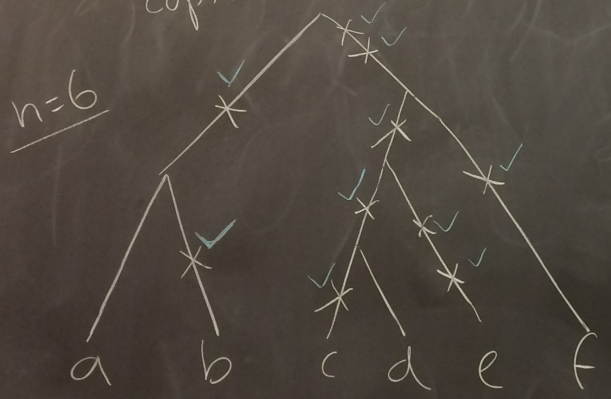
ex: $\frac{1}{3}(\overset{AB}{2} + \overset{AC}{3} + \overset{BC}{1}) = 2$

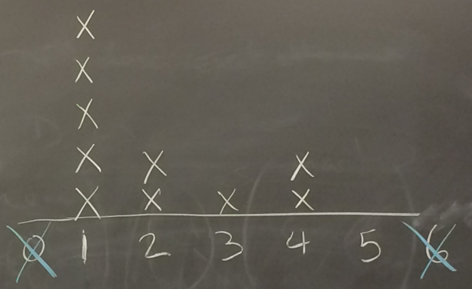
$$\pi = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} k_{ij}$$

k_{ij} ← # of pairwise differences between i + j

③ SFS = site frequency spectrum.

ξ_i = # of sites with $\boxed{i}$ copies of the mutant allele + $\boxed{n-i}$ copies of the ancestral allele




$\cancel{0}$
$\cancel{6}$