Polyploidy Workshop
Day 4

Patrick Monnahan

University of Minnesota, USA

Overview

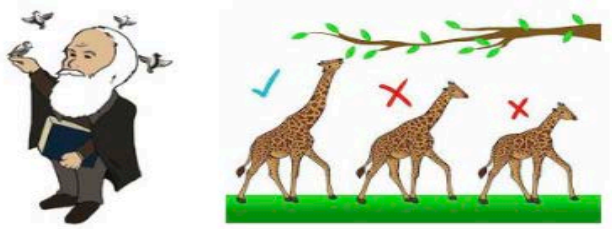
- Part 1: The response to selection
- Part 2: Fixation probability of beneficial alleles
- Part 3: Probability of adaptation VS probability of fixation
- Part 4: Linked selection
- Bonus: Basic NGS analysis of polyploid data

R script

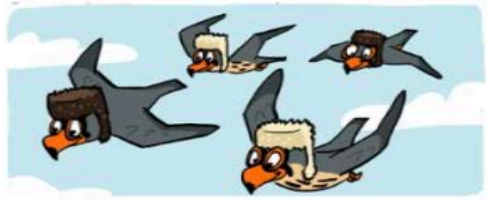
- Open Day7_PolyploidyWorkshop2018.R in a **text editor.**
  - DO NOT TRY TO RUN THE ENTIRE SCRIPT!!
- Copy and Paste entire section named “Define Useful Functions”
  - If you haven’t installed the necessary packages yet, try install.packages(“PACKAGE_NAME”)
  - The questions in the handout will often contain hints of which function to use
- The remaining section, named “Class examples,” utilizes these functions to answer the questions
  - R gurus: See if you can figure out how to use the functions without consulting “Class examples”

4 forces of evolution (i.e. allele frequency change)

Natural selection



Migration



Random drift

Mutation

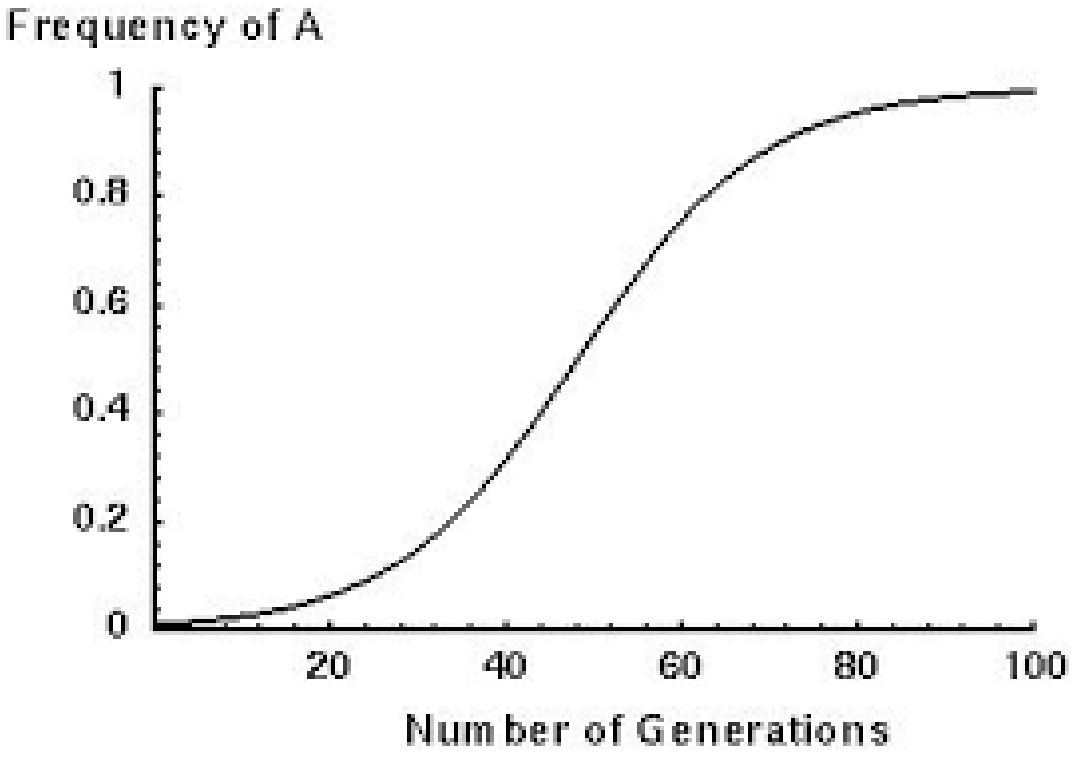
https://www.eli5.ml/topic/evolution/

4 tenets of natural selection

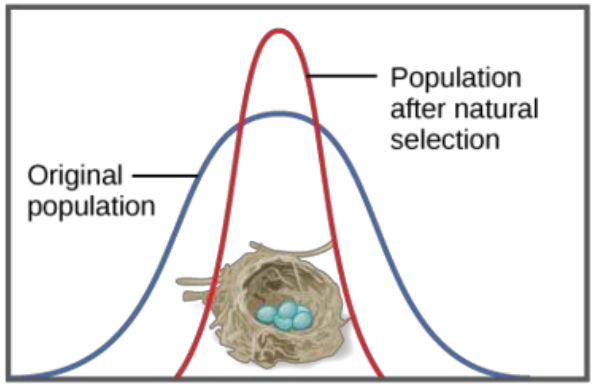
- Individuals vary
- Variation is (partially) heritable
- More individuals are produced each generation than can survive
- Differential survival and reproduction based on individual differences

- In four words:
  - Heritable variation in fitness

Selection

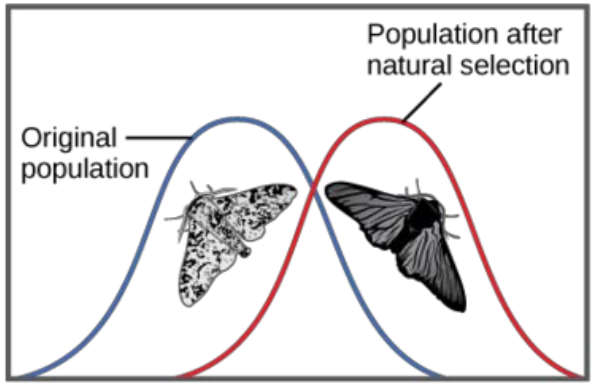


(a) **Stabilizing selection**



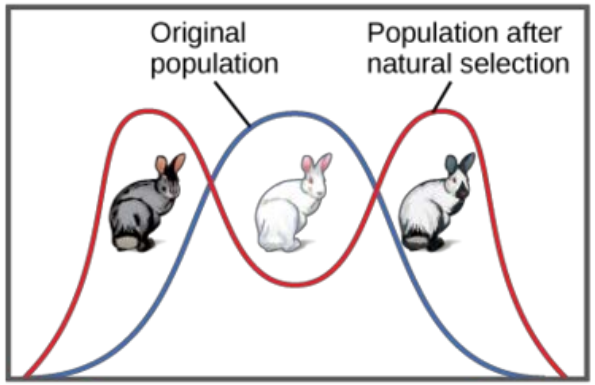
Robins typically lay four eggs, an example of stabilizing selection. Larger clutches may result in malnourished chicks, while smaller clutches may result in no viable offspring.

(b) **Directional selection**



Light-colored peppered moths are better camouflaged against a pristine environment; likewise, dark-colored peppered moths are better camouflaged against a sooty environment. Thus, as the Industrial Revolution progressed in nineteenth-century England, the color of the moth population shifted from light to dark, an example of directional selection.

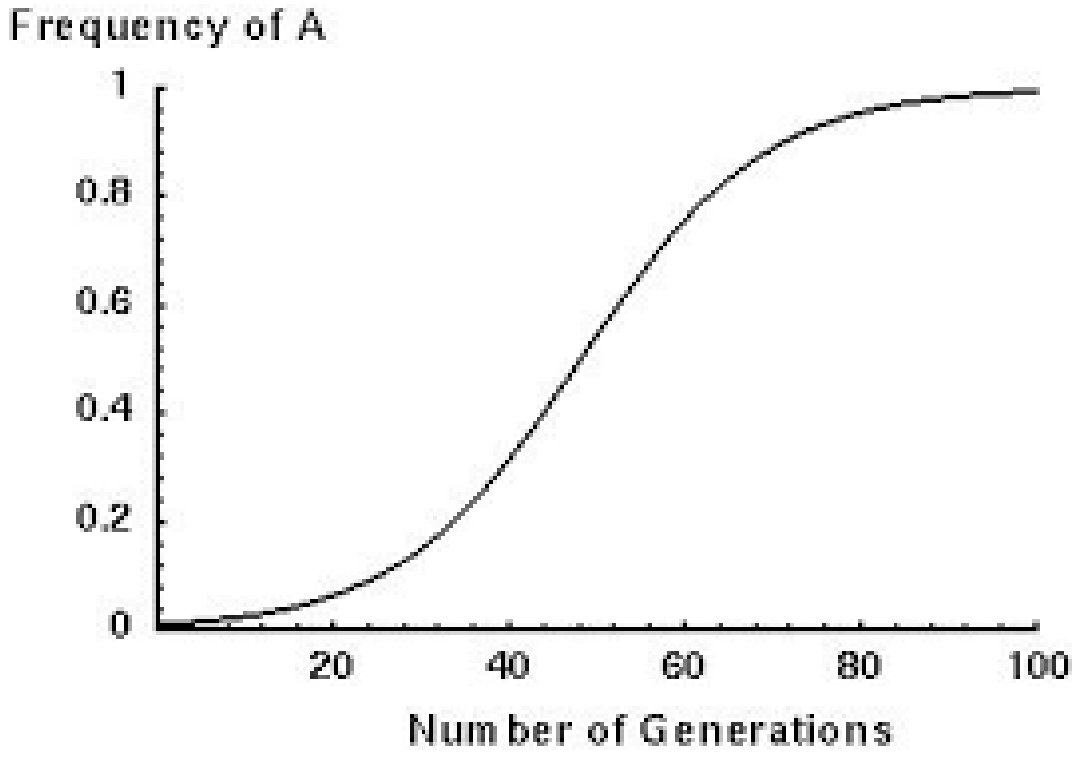
(c) **Diversifying selection**



In a hyphothetical population, gray and Himalayan (gray and white) rabbits are better able to blend with a rocky environment than white rabbits, resulting in diversifying selection.

https://courses.lumenlearning.com/boundless-biology/chapter/adaptive-evolution/

What determines the rate at which *A* goes to fixation?



Let’s find out!

Part 1: goals

- (Re)introduce the basic theory behind selection on a single locus
- Understand the main parameters that govern the response to selection
- Extend the theory to tetraploids

Part 1: Important definitions and assumptions

- Fitness
  - Simplest def'n = probability of surviving and reproducing
  - Not allowing for selection on fecundity (# of offspring)
  - Fitness is constant over time
- Infinite population size
  - Or at least, very large…
  - Also, assuming that N = Ne
- Random mating
  - No inbreeding, no assortative mating, etc…
- Only considering bi-allelic loci

Notes on notation

- We will now designate the different alleles at a locus via upper and lower case letters
  - E.g. Diploids: AA, Aa, aa
    - Tetraploids: AAAA, AAAa, AAaa, Aaaa, aaaa
  - Different letters represent different loci
- Some lettered variables are used multiple times in population genetics

Variable	Def'n 1	Def'n 2
s	Selfing rate	Selection coefficient
Rho	Measure of differentiation	Population recomb. rate
c	Ploidy level	Recombination fraction

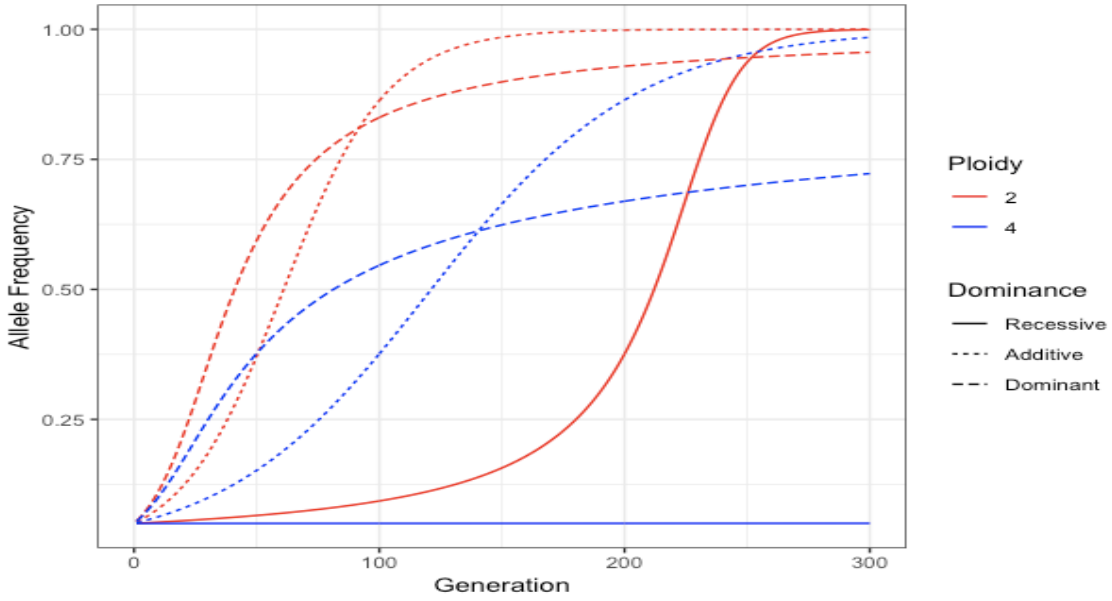
Part 1: Typos/Corrections

- Question 7: should read Q6 and **not** Q81
- Question 16: should read Q16 and **not** Q89

BEGIN PART 1

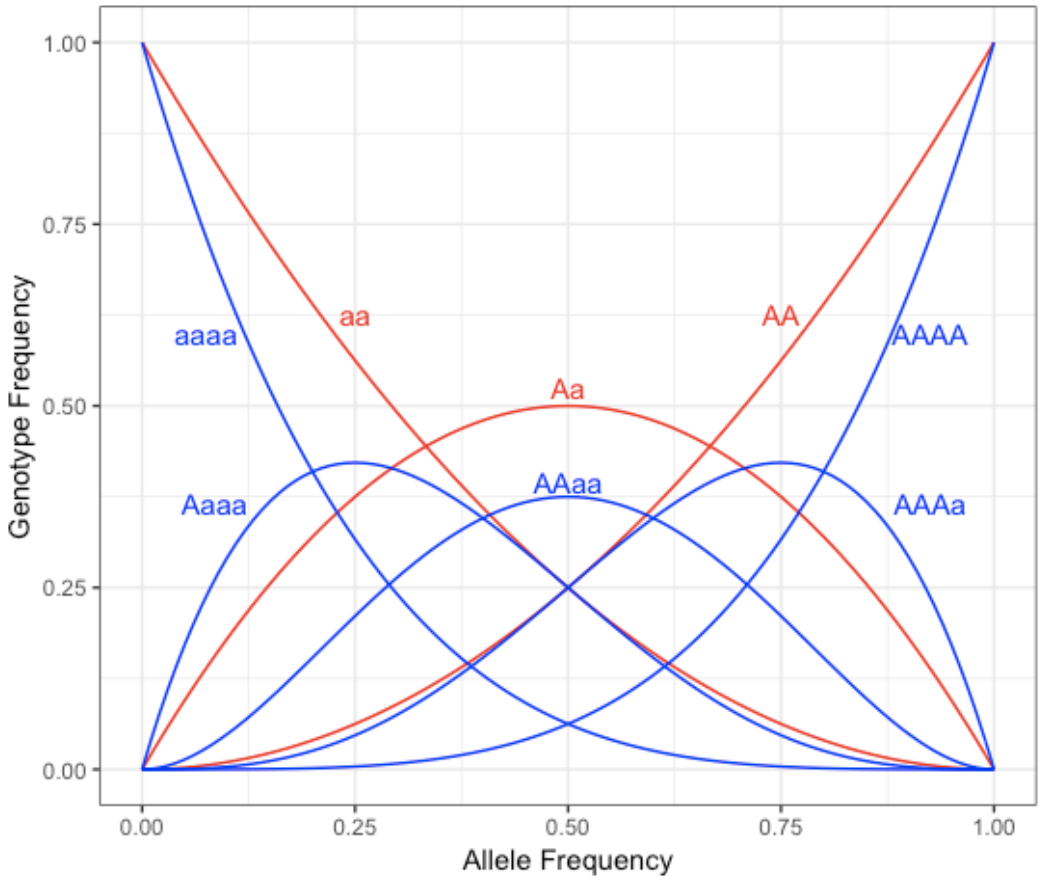
Part 1 summary

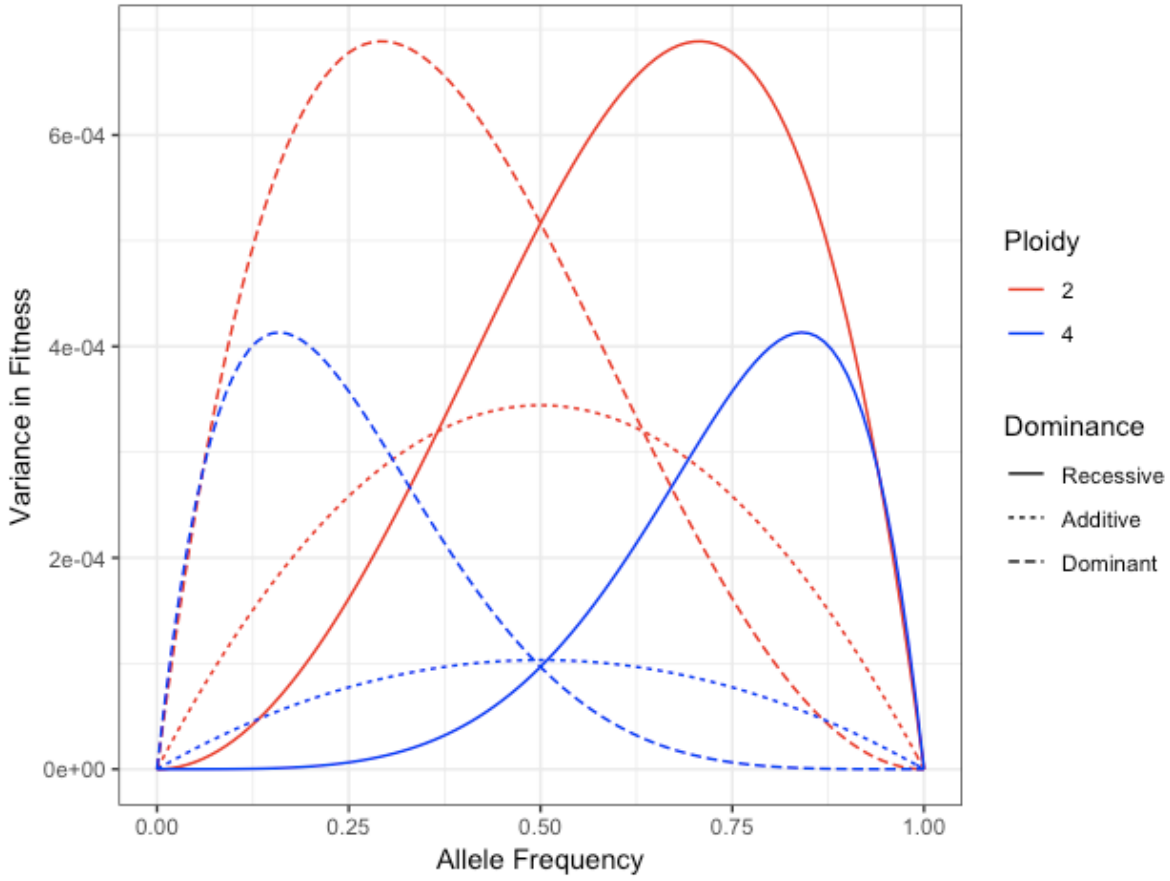
- What determines the rate at which a beneficial allele goes to fixation?
  - The strength of selection (i.e. the selection coefficient)
    - This relates directly to ***relative*** fitness of genotypes (not ***absolute*** fitness)
  - Dominance coefficient(s)
  - PLOIDY!!



Part 1 summary

- Ploidy effect is due to different numbers/frequencies of genotypes
  - Which directly translates to the variance in fitness (which underlies selection response



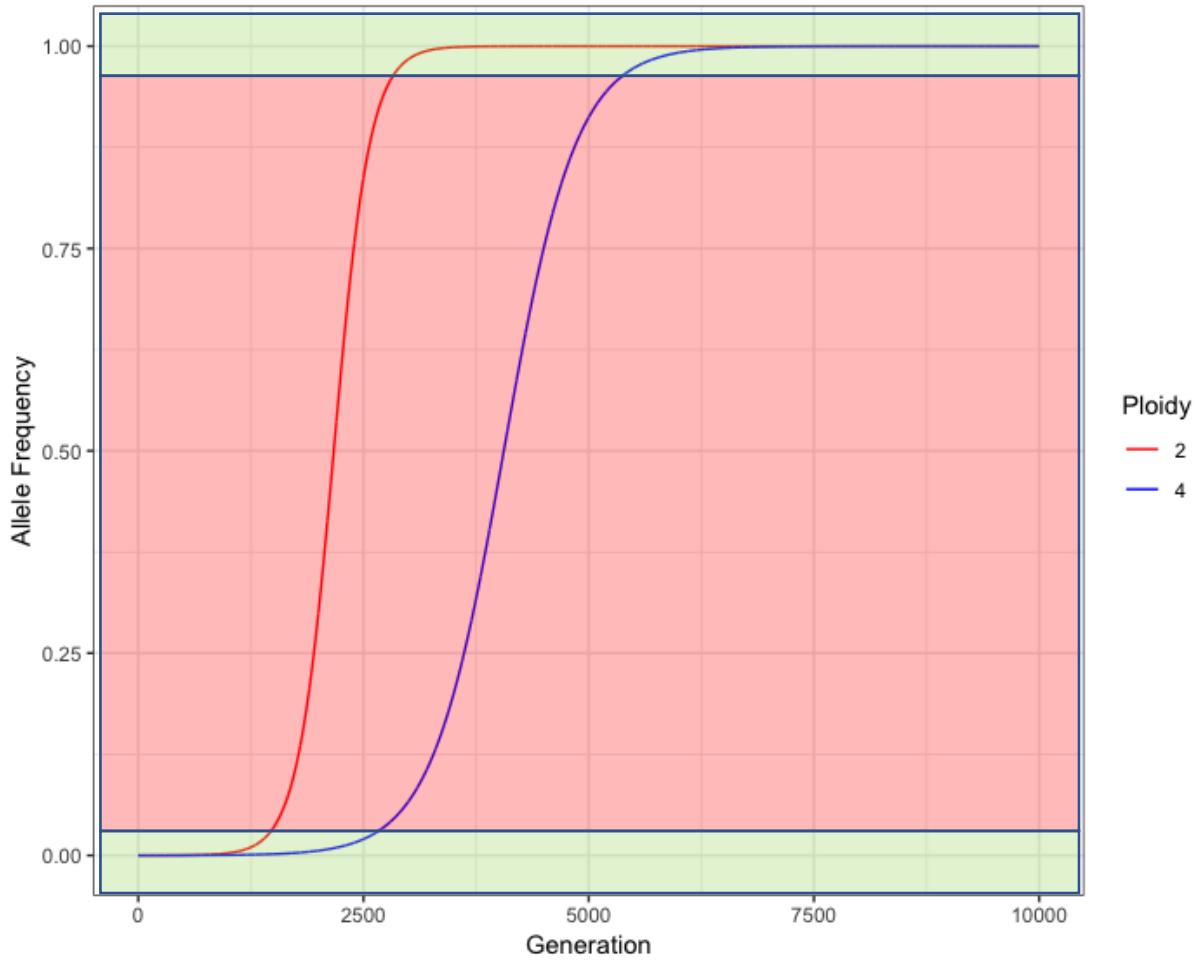


These plots can be generated with commands contained in the “EXTRA” section of the Rscript

Part 2: Fixation Probability

- Selection is most effective in red region
  - Less effective in green
- Why??

- Lower green portion is crucial for the ultimate fixation of the beneficial allele
  - Must survive stochastic/random loss



Part 2: Fixation Probability

- If this seems confusing, consider question 6.
    - Even the most fit genotypes (corresponding to the genotypes with the beneficial allele) have a low individual probability of survival
    - If the beneficial mutation arises as a single copy in the population, there is a substantial chance that individual bearing it will not reproduce

Part 2: Goals

- Unify the ***deterministic*** dynamics of allele frequency change introduced in Part 1 with the ***stochastic*** dynamics that occur near the boundaries of loss and/or fixation

- Understand what determines the ultimate fixation probability of a beneficial allele

- Determine if the fixation probability differs between ploidy levels

- Stress the importance of recognizing the particular parameterization of ***s*** and ***h***

Part 2: Typos/Corrections

- Question 21: Should read Q20 and **not** Q93
- Question 22: Should read Q 21 and **not** Q94.
  - Also, should read Q16 and **not** Q89
-

BEGIN PART 2

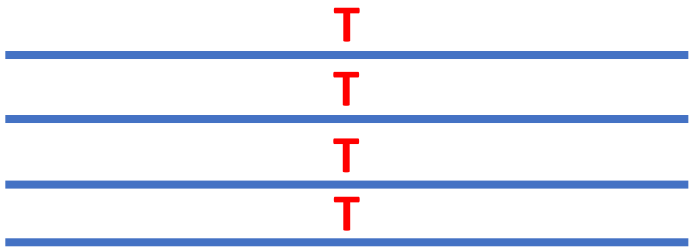
Part 2: Summary

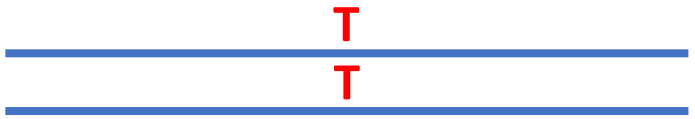
- A **selection coefficient, s**, of 0.1 – 0.2 would be considered by most as "strong" selection.
  - However, even if these mutations are completely dominant, the probability that they escape stochastic loss from the population is not that high
    - Fixation probability = $2h_cs = 0.2 – 0.4$

- If the **selection coefficient, s**, is the same across ploidies then the most important is the **dominance** in ***SINGLE COPY.***
  - *In other words, the fitness of the **Aa** and **Aaaa** genotypes for diploids and tetraploids, respectively.*

Part 3: Rate of Adaptation VS Rate of Fixation

- Results so far are for a single beneficial allele ALREADY PRESENT in the population
- If not already present, we must wait for mutation to introduce a beneficial allele
- If mutation from T to A is beneficial, will the waiting time for the A mutation be greater in tetraploids or diploids??





Part 3: Goals

- Understand how differences in the mutational and selection processes differ between ploidies
  - And, the implications this has for the evolutionary process
- Consider how the ***equilibrium*** allele frequencies differ for ***deleterious*** alleles across ploidy
  - And, how an environmental shift could utilize this variation
- Think deeply on our use of ***s*** and ***h*** up to this point.
  - And, what types of mutations these would (or would not) represent

Part 3: Typos/Corrections

- Question 23: Should read Q16 and Q23 instead of Q89 and Q96
- Question 24: Should read Q24 and **not** Q97
- Question 26: Should read 25 and **not** 98
- Question 27: Should read 25 and 26 and **not** 98 and 99, respectively

BEGIN PART 3

Part 3: Summary

- Whether or not adaptation occurs more quickly in one ploidy versus another will depend on whether adaptation is ***mutation-***limited or ***selection***-limited

- Although the fixation of individual alleles may occur more quickly in diploids on average, beneficial alleles are introduced at twice the rate in tetraploids.

- Again, the dominance in ***single-copy*** is a critical parameter.

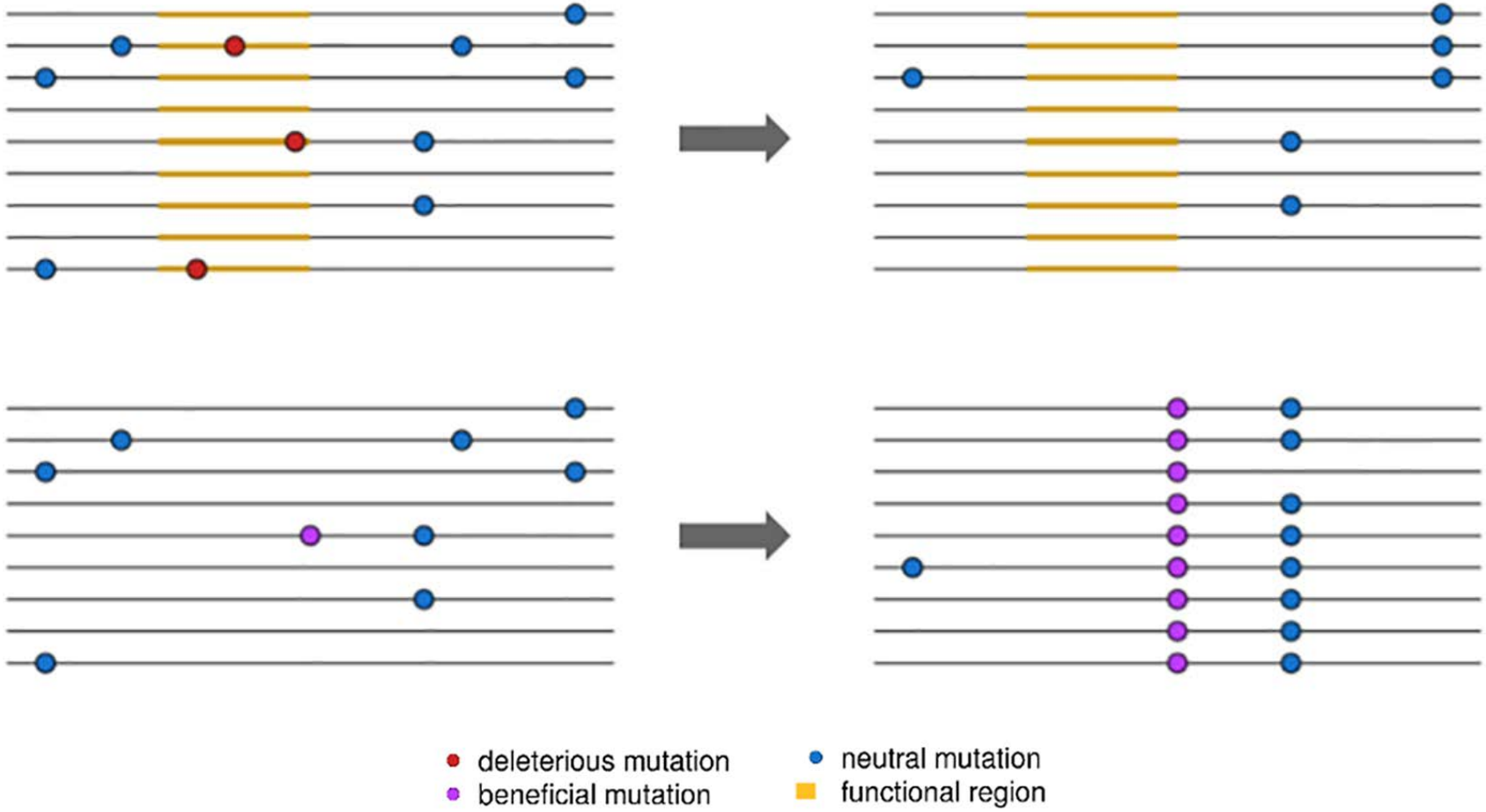
Part 4: Linked selection

Before selection: Diversity is high

After selection: Diversity is low

Background Selection

Genetic Hitchhiking/
Selective Sweeps



Josephs and Wright 2016

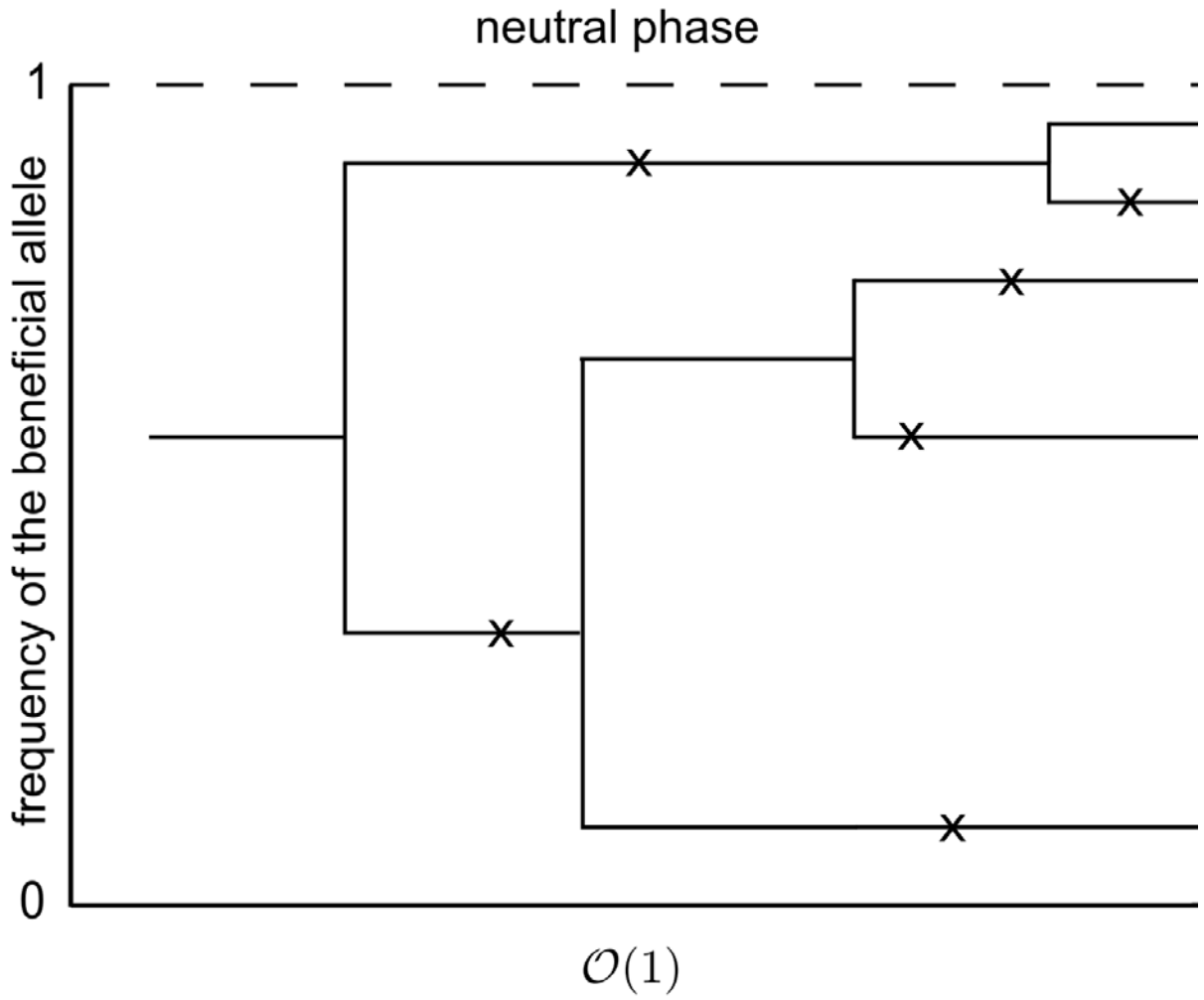
Part 4: Some controversy

- Some believe that linked selection is so pervasive that the entire genome is affected

- If so, this has implications for using neutral theory as our null expectation for essentially all population genomic analysis
  - i.e. Genetic **Drift** vs Genetic **Draft** as null hypothesis

- Personal opinion: Important lesson to keep in mind, but reality is likely not as extreme as the loudest voices might lead you to believe

Part 4: Coalescent simulations



For diploids

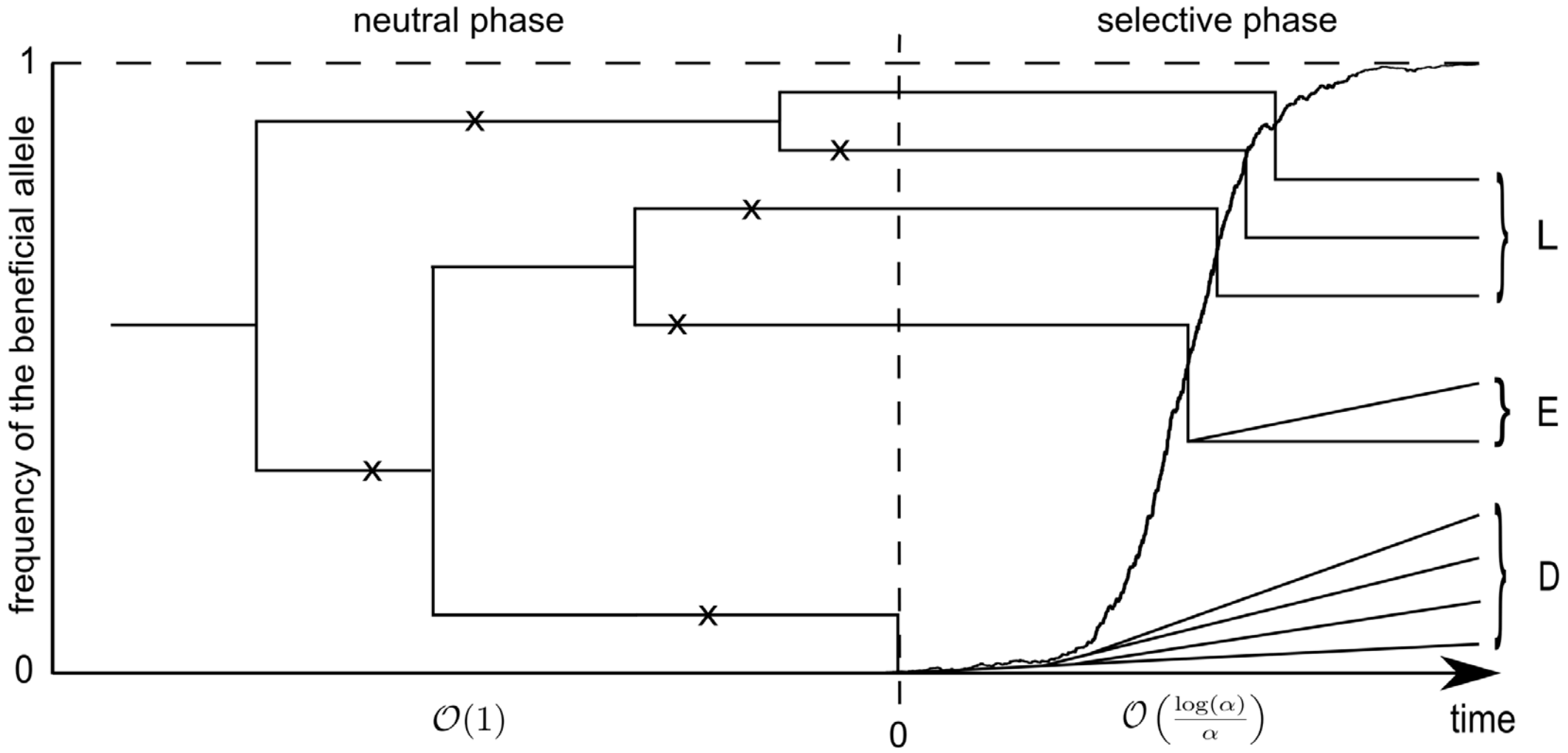
$$\Pr\{\text{Co}\} = \frac{n(n-1)}{2}\frac{1}{2N_e}$$

$$E[t] = \frac{4N_e}{n(n-1)}$$

$$E[T_{MRCA}] = 4N_e \sum_{i=1}^{n-1} \frac{1}{i(i+1)}$$

Bossert S, Pfaffelhuber P (2014)

Part 4: Coalescent simulations



Bossert S, Pfaffelhuber P (2014)

Part 4: Goals

- Gain a basic intuition of how selection on beneficial alleles can impact variation at ***linked*** neutral sites

- Determine the differences between diploids and tetraploids that are relevant to linked selection

- Consider the implications for detecting and comparing selection across ploidies

Part 4: Simulations in R

```
####### Part 4: Linked Selection ########
# Set your parameters here
ploidy = 2
pop_size = 10000 # Keep this value above 100 and below 1000000 (computation time will increase with increasing pop_size)
selection_coeff = 0.1 # Keep this between 0 and 1
dominance = 0.5 # Must be vector of 3 numbers if ploidy = 4. For example, c(0.25, 0.5, 0.75) for an additive allele.
seq_len = 1000000 # Length of sequence that we will simulate with mssel.  Increasing this value will increase computation time.
mutation_rate = 1e-8 # per-base mutation rate; mu
recomb_rate = 1e-8 # per-base recombination rate; r
samp_num = 10 # number of individuals to sample

### HERE YOU MUST ENTER THE PATH TO MSSEL
path_to_mssel = "XXXXXXX"
path_to_mssel = "/Users/pmonnahan/Documents/Research/code/dmc/mssel_modified/mssel"

# Perform stochastic simulations for selection on beneficial allele
new_traj = PloidyForSim2(ploidy, pop_size, selection_coeff, dominance)

# Run mssel
infile = msselRun(N = pop_size, n = samp_num, new_traj, L = seq_len, mu = mutation_rate, r = recomb_rate, ploidy = ploidy, ms =
path_to_mssel)

# calculate population genetic metrics in sliding windows across simulated region
dat = msselCalc(infile, numWindows = 200, rep(ploidy * samp_num, 2), Nsites = seq_len)

# Plotting
ggplot() + geom_line(data = dat, aes(x = bp.end, y = Pi.1), color = "red") + geom_line(data = dat, aes(x = bp.end, y = Pi.2),
color = "blue") + xlab("Position (bp)") + ylab("Diversity")

ggplot() + geom_line(data = dat, aes(x = bp.end, y = Kelly.Z_nS_1), color = "red") + geom_line(data = dat, aes(x = bp.end, y =
Kelly.Z_nS_2), color = "blue") + xlab("Position (bp)")

ggplot() + geom_line(data = dat, aes(x = bp.end, y = Pi.1), color = "red") + geom_line(data = dat, aes(x = bp.end, y = Pi.2),
color = "blue") + xlab("Position (bp)")

ggplot() + geom_line(data = dat, aes(x = bp.end, y = fst)) + xlab("Position (bp)")
```

BEGIN PART 4

Part 4: Summary

- Sweeps may appear “softer” in tetraploids even though the strength of selection is the same
  - Longer fixation times in tetraploids allow for more opportunity to recombine (and thus retain haplotype diversity)

- Greater retention of diversity following a sweep may also result from increased **population-scaled recombination** in tetraploids

- Diversity might recover more quickly in tetraploids following a selective sweep due to the higher mutational input