GGG-201D Problem Set 3

Ethan Holleman

5/10/2021

Problem 1

Part A

Generate a scatter plot that shows the properties of selection in large populations. The x-axis should be frequency of the advantageous allele and range from 0 to 1. The y-axis should be the change in frequency of the advantageous allele after one generation of selection. Perform calculations in steps of 0.01 for each of the following six (1a, 1b, 1c, 2a, 2b, 2c) scenarios: (1) the homozygous deleterious genotype has a selection coefficient of 0.1 and the advantageous allele is (a) recessive, (b) dominant or (c) additive; (2) the homozygous deleterious genotype has a selection coefficient of 0.25 and the advantageous allele is (a) recessive, (b) dominant or (c) additive.

```
# Function to calculate allele freq after selection given freq 2 alleles
# and selection coefficients of all genotypes
allele_freq_after_selection <- function(freq_a, freq_t, Saa, Sat, Stt){

  aa <- freq_a^2 * (1-Saa)
  at <- freq_a * freq_t * (1 - Sat)
  at2 <- 2 * freq_a * freq_t * (1 - Sat)
  tt <- freq_t^2 * (1 - Stt)

  (aa + at) / (aa + at2 + tt)

}
allele_freq_after_selection(0.4, 0.6, 0, 0.2, 0.4)  # check function is working
```

```
## [1] 0.4631579
```

```
# Infer selection coeffiencts if advantageous allele is recessive
sc_ressesive <- function(deleterious_sc){

  c(0, deleterious_sc, deleterious_sc)

}
# Infer selection coeffiencts if advantageous allele is dominant
sc_dominant <- function(deleterious_sc){

  c(0, 0, deleterious_sc)

}
# Infer selection coeffiencts if advantageous allele is additive
sc_additive <- function(deleterious_sc){
```

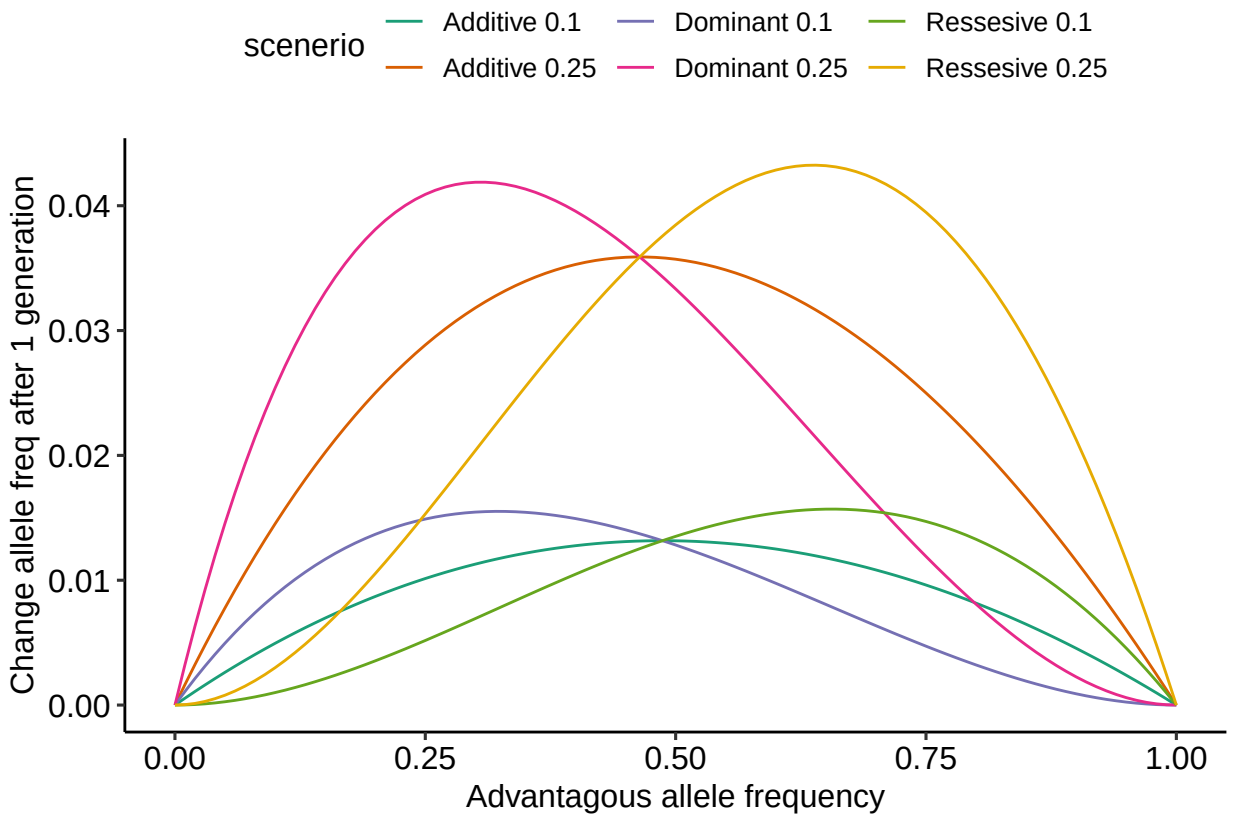
```
  c(0, 0.5 * deleterious_sc, deleterious_sc)

}
```

```
# New function that calculates allele freq in next generation given freq
# of two alleles but takes in the deleterious allele selection coefficient
# and a function to define other selection cofficient values

genotype_aware_afas <- function(freq_a, freq_t, sc_del_allele, sc_func){

  sc_vals <- sc_func(sc_del_allele)
  Saa <- sc_vals[1]
  Sat <- sc_vals[2]
  Stt <- sc_vals[3]

  allele_freq_after_selection(freq_a, freq_t, Saa, Sat, Stt)

}
```

```
scenerio_df <- function(del_sc, genotype_func, genotype_name){
  scenerio_name <- paste(genotype_name, as.character(del_sc))
  df <- data.frame(advan_allele_freq=seq(0, 1, 0.01))
  post_1_gen <- list()
  for (i in 1:length(df$advan_allele_freq)){
    freq_a <- df[i, ]
    freq_t <- 1 - freq_a
    post_1_gen[[i]] <- genotype_aware_afas(freq_a, freq_t, del_sc, genotype_func) - freq_a

  }
  df$scenerio <- scenerio_name
  df$freq_gen_2 <- unlist(post_1_gen)
  df
}
```

```
# ugly
df.1 <- scenerio_df(0.1, sc_ressesive, "Ressesive")
df.2 <- scenerio_df(0.1, sc_dominant, "Dominant")
df.3 <- scenerio_df(0.1, sc_additive, "Additive")
df.4 <- scenerio_df(0.25, sc_ressesive, "Ressesive")
df.5 <- scenerio_df(0.25, sc_dominant, "Dominant")
df.6 <- scenerio_df(0.25, sc_additive, "Additive")

big.df <- rbind(df.1, df.2, df.3, df.4, df.5, df.6)
```

```
# Finally, plot everything
library(ggplot2)
library(ggpubr)
library(RColorBrewer)

ggplot(big.df, aes(x=advan_allele_freq, y=freq_gen_2, color=scenerio)) +
  geom_line() + theme_pubr() + scale_color_brewer(palette = "Dark2") +
  labs(x='Advantagous allele frequency', y='Change allele freq after 1 generation')
```



Part B

Four of the six plots from above are highly asymmetric. Explain the biological reason behind these asymmetric patterns.

Biological significane

Problem 2

You sequence a 5.6 kb locus in 5 diploid individuals and observe 11 segregating sites. What is your estimate of theta in this population? What property of the expected coalescent tree is this estimate based on? What is your estimate of coalescent Ne assuming a mutation rate of 10^-8 per bp per generation?

```
locus_len <- 5.6e3
gene_copies <- 5*2  # each diploid contributes 2 gene copies
seg_sites <- 11  # observed 11 segregating sites
```

```
theta <- seg_sites / sum(1 / 1:(gene_copies-1))
```

```
mutation_rate <- 1e-8
message(paste('Theta =', round(theta, 2)))
```

```
## Theta = 3.89
```

Ultimately, this estimation is based on the expectation that the number of segregating sites is determined by the coalescent tree length (number of generations) which in turn is a property of both population size and the number of gene copies that are analyzed.

```
Ne <- theta / (4 * mutation_rate * locus_len)
message(paste('Ne = ', round(Ne, 2)))
```

```
## Ne =  17358.68
```

Problem 3

Explain the difference between coalescent effective population size (Ne) and instantaneous Ne. What is one way to estimate coalescent Ne? What is one way to estimate instantaneous Ne?

The coalescent effective population size is the size of a Wright-Fisher population that would be predicted to have the same amount of genetic variation as the actual population. This is similar but distinct from instantaneous Ne because instantaneous Ne is based on genetic drift. Therefore, an actual population with a coalescent and instantaneous effective size of 50 would harbor the same amount of genetic diversity as a Wright-Fisher population of size 50 and be experiencing the same amount of genetic drift as a Wright-Fisher population of this same size. This also means that from generation to generation a population could be experiencing a large degree of genetic drift and therefore have a small instantaneous effective size, but if the populations overall genetic diversity was maintained, possibly through interactions with meta-populations, coalescent effective size could be much larger.

Coalescent effective population size can be estimated by calculating Tajima's theta which utilizes Pi, the average number of pairwise nucleotide differences at a given loci and a mutation rate. Instantaneous effective population size can be estimated though a maximum likihood approach. This first involves collecting individuals from two generations (G1 and G2) and genotyping individuals in these groups at multiple loci. Then, calculate the probability of observing data given the assumptions that define a Wright-Fisher population.

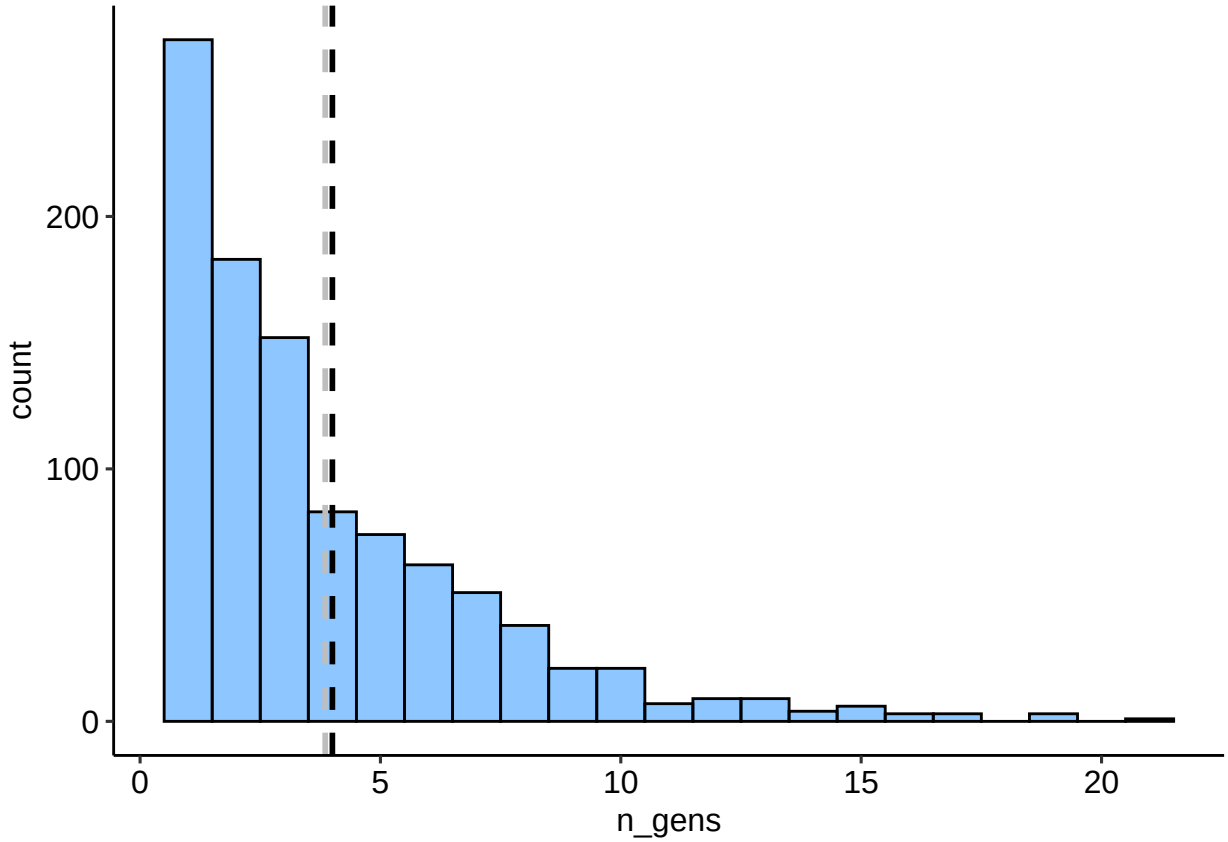
Problem 4

The expected time to the first coalescent event of four gene copies is 2N/6 generations before the present but the actual time could be much more or much less. If you sample 100 sets of four gene copies, each set has an actual time to the first coalescent event. Do you expect the number of sets that have an actual first coalescent before 2N/6 to be approximately equal to the number of sets to have an actual first coalescent after 2N/6? Explain your answer.

```
n_gene_copies <- 4

sample_gen <- function(n_gene_copies){

  coalesent_key <- 0  # is random number is this then coalensent event
  n_gens <- 0 # number generations
  while (coalesent_key != 1){
    coalesent_key <- sample(1:n_gene_copies, 1)
    n_gens <- n_gens + 1
  }
  n_gens
}
```

```
simulation <- function(n_runs, n_gene_copies){

  n_gens_list <- list()
  for (i in 1:n_runs){
    n_gens_list[[i]] <- sample_gen(n_gene_copies)
  }
  data.frame(n_gens=unlist(n_gens_list))
}
# Run the simulation 100 times and plot
df <- simulation(1000, n_gene_copies)
```

```
expectation <- (1 / n_gene_copies) ^-1
ggplot(df, aes(x=n_gens)) + geom_histogram(binwidth = 1, color='black', fill='dodgerblue', alpha=0.5) +
                color = "black", size=1) +
geom_vline(xintercept=mean(df$n_gens), linetype="dashed",
                color = "grey", size=1)
```



Grey line is expectation and black line is observed mean time to most recent common ancestor.

```
freqs <- table(df)
less <- sum(freqs[(1:length(freqs) < expectation)])
greater <- sum(freqs[(1:length(freqs) >= expectation)])

message(paste('Area under curve < expectation', less))
```

```
## Area under curve < expectation 605
```

```
message(paste('Area under curve > expectation', greater))
```

```
## Area under curve > expectation 395
```

The probability of observing a coalescent event between any two gene copies will decrease with the number of generations. This makes intuitive sense sense because each generation represents another chance of coalescence. Therefore the distribution with observe is not symmetric and most of the area under the curve occurs before our expectation of 2N/6.