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){
```

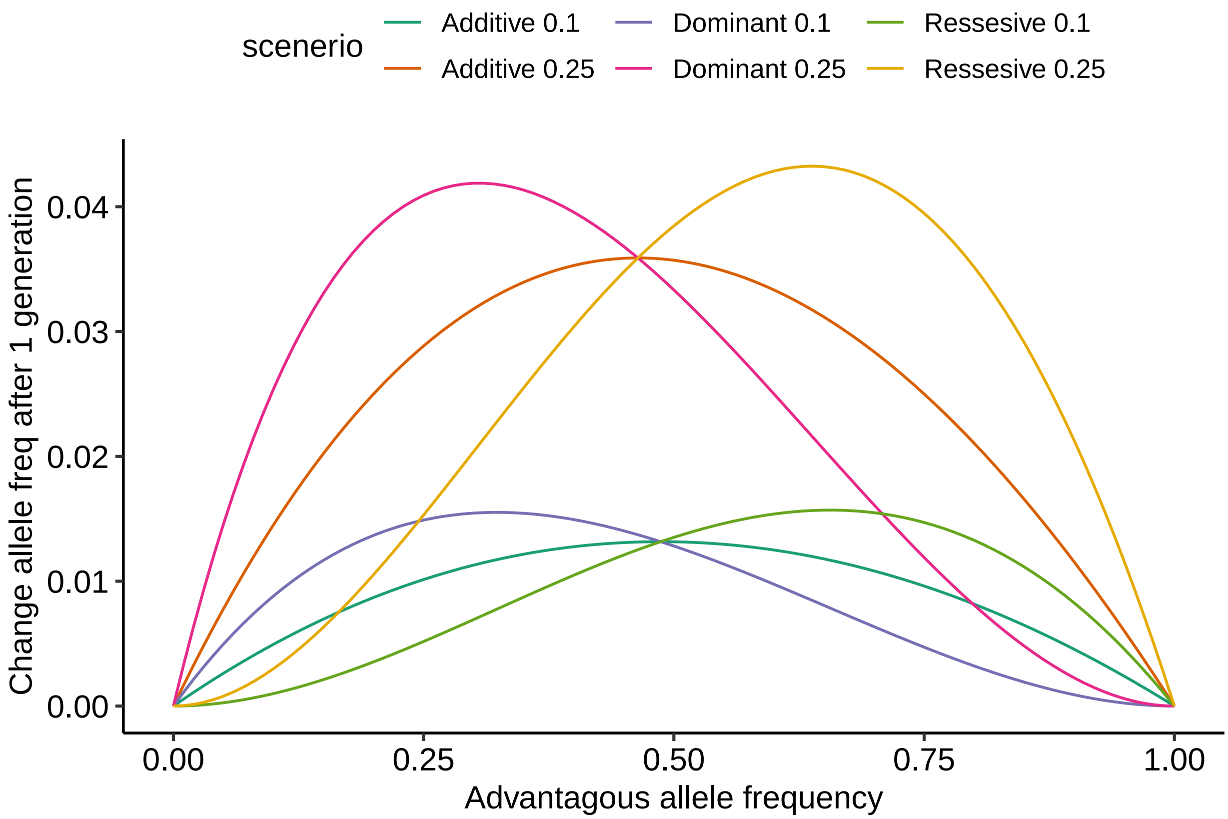
```r
  c(0, 0.5 * deleterious_sc, deleterious_sc)

}
```

```r
# 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)

}
```

```r
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
}
```

```r
# 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)
```

```r
# 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.

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 <- 10e-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 coalensent 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)
message(paste('Ne = ', round(Ne, 2)))
```

```
## Ne =  9720858.47
```

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?

Effective number of individuals in idealized population that has same magnitude of genetic drift as actual population (how much allele freq change across time in actual populations)

Find wright fisher population that we would expect to have that same level of genetic drift (instanous or short term effective population size Ne)

Another way to define is from coalensent theory persepctive

Effective size of a population is number of individuals in idealizes pop that has the same amount of genetic variation as the actual population (vs genetic drift degree). This is colacent or long term effective size.

From generation to generation a population could be experiencing a lot of genetic drift and therefore have low instanous effective population size but for whatever reason (meta populations salmon) could maintain genetic diversity and have larger long term effective population size.

Population 1: Hard to count sequence a bunch from a locus and calulate Pi = 2.4 diff / 1200 bp pi is also Tajimia theta = 4N u where u is a mutation rate

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.

Yes because that is how the normal distrabution works.