EBIO 3080 Lab Week 10: Coronavirus Genetics

Kayden Adams, Sarah Conway, Yousef Al Obaidan, Dylan Oh, Julia Thompson

29 October 2020

Part 1: G Allele Frequency at Codon 614

This section will evaluate the change in allele frequency over time for the derived G allele at codon 614. In the complete, worldwide data set, there are **114404** observations, **15353** of which were of the D allele and **99052** were of the G allele. This resulted in a relative frequency of **0.134** for the D allele and **0.866** for the G allele.

Worldwide Counts

The following code and plots are for the entire data set:

```r
# Importing csv file
D614G <- read.csv("D614G.continent.csv", header = TRUE,
                  sep = ",", fileEncoding = "UTF-8-BOM")

# Calculating relevant values
n <- nrow(D614G)
DGfreq <- table(D614G$Allele) / n
DGworld_counts <- table(D614G[, c(1, 6)])
DGworld_months <- as.integer(labels(DGworld_counts)$Month_Since_Start)
DGworld_n <- DGworld_counts[1,] + DGworld_counts[2,]
Dworld_freq <- DGworld_counts[1,] / DGworld_n
Gworld_freq <- DGworld_counts[2,] / DGworld_n

# Combining data into a data frame for ggplot2
DGworld_plot1 <- data.frame(months = DGworld_months,
                            cts = DGworld_counts[1,],
                            freq = Dworld_freq)
DGworld_plot2 <- data.frame(months = DGworld_months,
                            cts = DGworld_counts[2,],
                            freq = Gworld_freq)

# Plotting count of alleles vs time
fig1 <- ggplot(DGworld_plot1, aes(x = months, y = cts)) +
    geom_line(aes(color = "#25b2ef")) +
    geom_line(data = DGworld_plot2, aes(color = "#e86868")) +
    geom_point(size = 3, color = "#25b2ef") +
    geom_point(data = DGworld_plot2, size = 3, pch = 1, color = "#e86868") +
    scale_color_identity(name = "Amino Acid",
                         breaks = c("#25b2ef", "#e86868"),
                         labels = c("D", "G"),
                         guide = "legend") +
    ggtitle("Worldwide") +
    xlab("Months Since Start of Pandemic") +
    ylab("Count of Alleles") +
    theme(plot.title = element_text(color = "#2c3136", size = 16,
                                    family = "karla"),
          axis.title.x = element_text(color = "#2c3136", size = 10,
                                      family = "karla"),
          axis.title.y = element_text(color = "#2c3136", size = 10,
                                      angle = 90, family = "karla"),
          legend.title = element_text(color = "#2c3136", size = 9,
                                      family = "karla"),
          legend.position = "top",
          legend.justification = "right",
          legend.margin = margin(0,0,0,0),
          legend.box.margin = margin(-10,0,-10,-10),
          legend.key.size = unit(0.8, "line")) +
    scale_x_discrete(breaks = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11),
                     limits = c("1", "2", "3", "4", "5", "6", "7", "8", "9",
                                "10", "11"))
print(fig1)
```

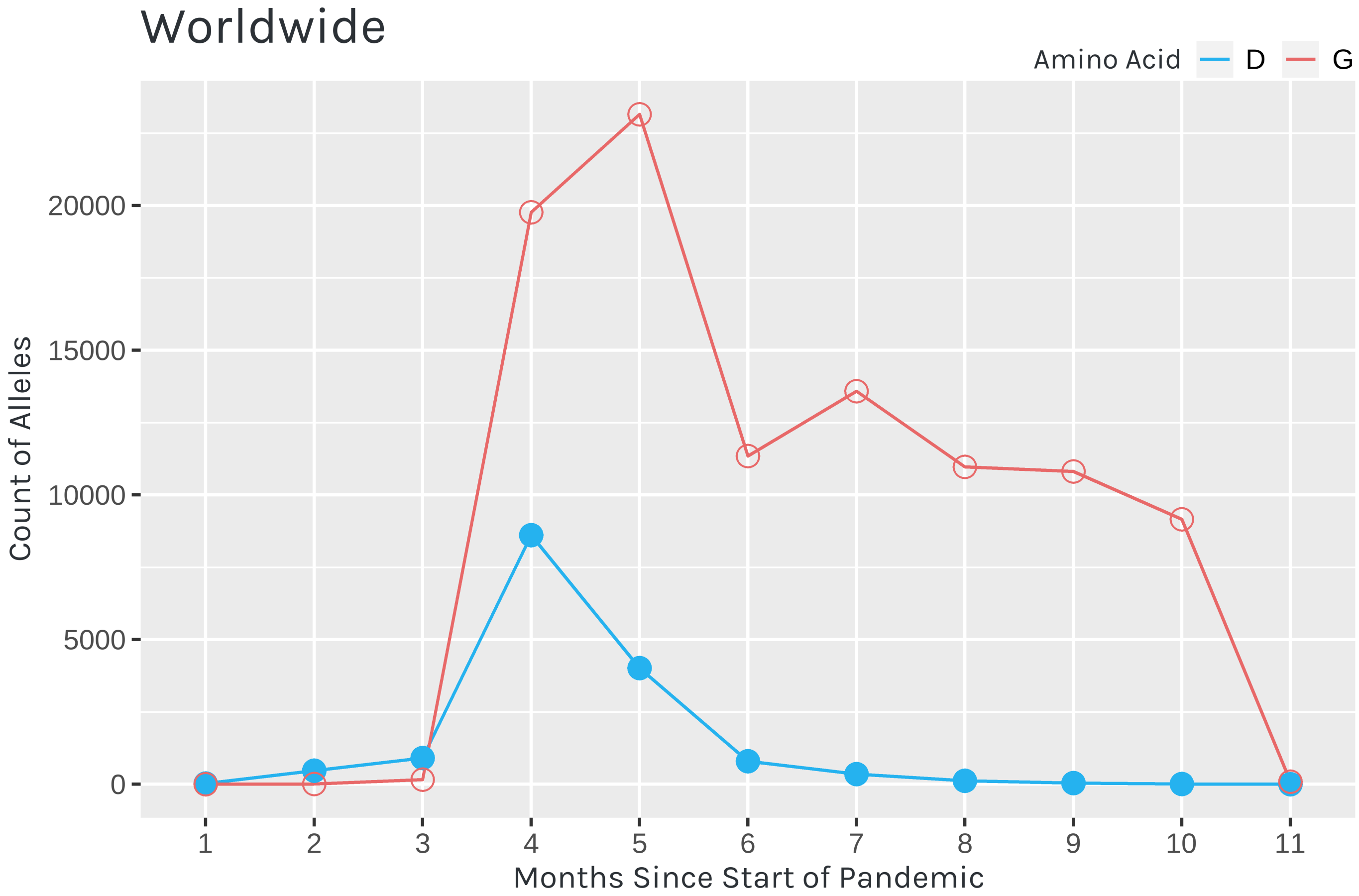

Figure 1: Number of observations of both the ancestral D and derived G alleles for every month since the start of the COVID-19 pandemic. The count of each allele spiked significantly four months after the first observation, and both have generally decreased since. The G allele surpassed the D allele in propagation after the third month, and reached a much higher total count.

```
# Plotting relative frequency of alleles vs time
fig2 <- ggplot(DGworld_plot1, aes(x = months, y = freq)) +
    geom_line(aes(color = "#25b2ef")) +
    geom_line(data = DGworld_plot2, aes(color = "#e86868")) +
    geom_point(size = 3, color = "#25b2ef") +
    geom_point(data = DGworld_plot2, size = 3, pch = 1, color = "#e86868") +
    scale_color_identity(name = "Amino Acid",
                         breaks = c("#25b2ef", "#e86868"),
                         labels = c("D", "G"),
                         guide = "legend") +
    ggtitle("Worldwide") +
    xlab("Months Since Start of Pandemic") +
    ylab("Relative Allele Frequency") +
    theme(plot.title = element_text(color = "#2c3136", size = 16,
                                    family = "karla"),
          axis.title.x = element_text(color = "#2c3136", size = 10,
                                      family = "karla"),
          axis.title.y = element_text(color = "#2c3136", size = 10,
                                      angle = 90, family = "karla"),
          legend.title = element_text(color = "#2c3136", size = 9,
                                      family = "karla"),
          legend.position = "top",
          legend.justification = "right",
          legend.margin = margin(0,0,0,0),
          legend.box.margin = margin(-10,0,-10,-10),
          legend.key.size = unit(0.8, "line")) +
    scale_x_discrete(breaks = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11),
                     limits = c("1", "2", "3", "4", "5", "6", "7",
                                "8", "9", "10", "11"))
print(fig2)
```

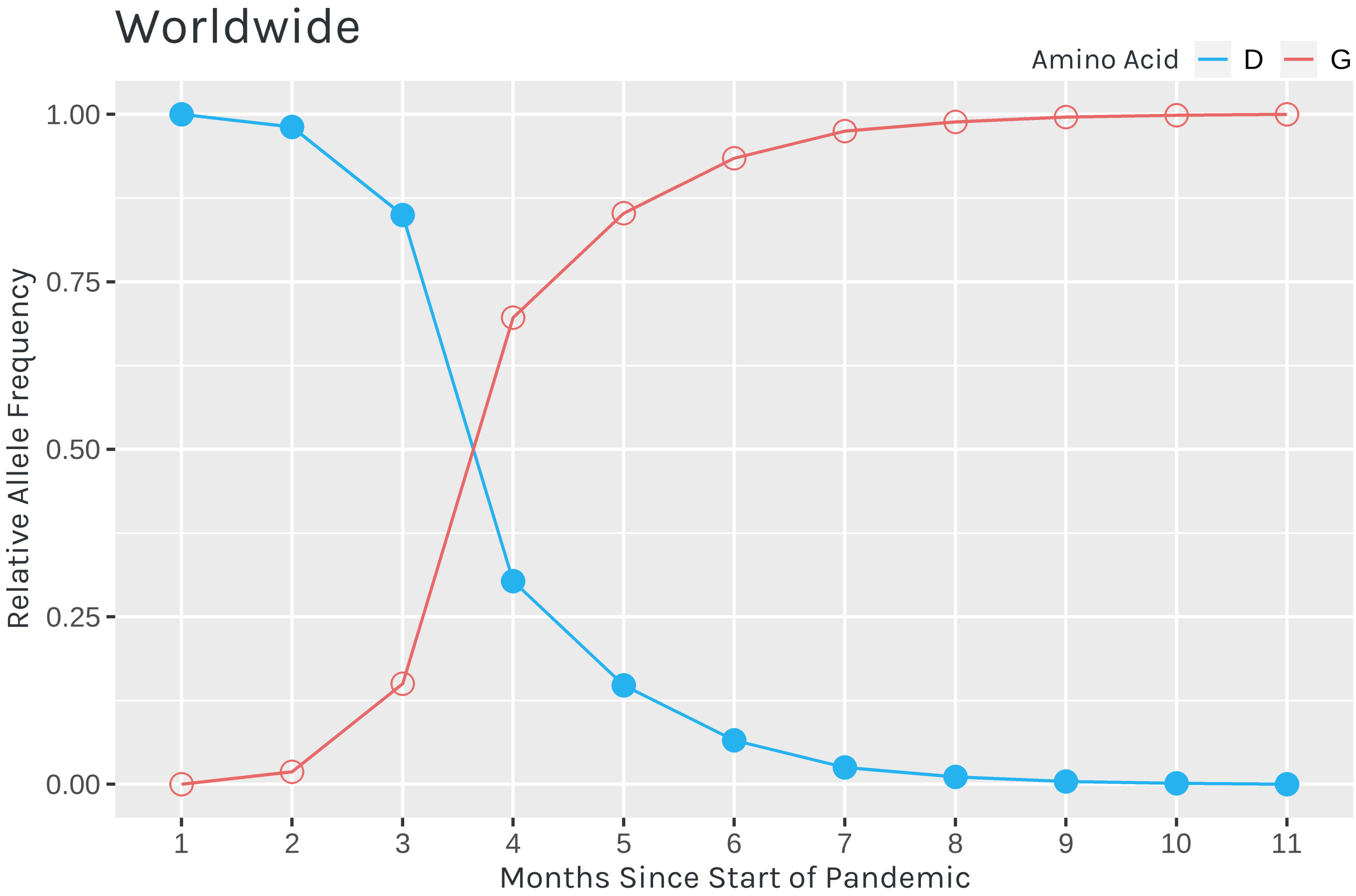

Figure 2: Relative frequencies of both the ancestral D and derived G alleles for every month since the start of the COVID-19 pandemic. At the outset of the pandemic, the viral genome at codon position 614 was observed to only be the D allele, but between months 3 and 4 after the initial outbreak, G became more common, and by month 11 it was the only observed allele at that position.

North American Counts

The following code and plots are specific to North America:

```r
# Calculating relevant values for North America
NAm <- subset(D614G, Continent == "North-America")
n_NAm <- nrow(NAm)
DGNAm_counts <- table(NAm[, c(1, 6)])
DNAm_counts <- DGNAm_counts[1,]
GNAm_counts <- DGNAm_counts[2,]
NAm_months <- as.integer(labels(DGNAm_counts)$Month_Since_Start)
DGNAm_n <- DNAm_counts + GNAm_counts
DNAm_freq <- DNAm_counts / DGNAm_n
GNAm_freq <- GNAm_counts / DGNAm_n

DGNAm1 <- data.frame(months = NAm_months,
                     cts = DNAm_counts,
                     freq = DNAm_freq)
DGNAm2 <- data.frame(months = NAm_months,
                     cts = GNAm_counts,
                     freq = GNAm_freq)
rownames(DGNAm1) <- 1:nrow(DGNAm1)
rownames(DGNAm2) <- 1:nrow(DGNAm2)

# Plotting count of alleles vs time for North America
fig3 <- ggplot(DGNAm1, aes(x = months, y = cts)) +
    geom_line(aes(color = "#25b2ef")) +
    geom_line(data = DGNAm2, aes(color = "#e86868")) +
    geom_point(size = 3, color = "#25b2ef") +
    geom_point(data = DGNAm2, size = 3, pch = 1, color = "#e86868") +
    scale_color_identity(name = "Amino Acid",
                         breaks = c("#25b2ef", "#e86868"),
                         labels = c("D", "G"),
                         guide = "legend") +
    ggtitle("North America") +
    xlab("Months Since Start of Pandemic") +
    ylab("Count of Alleles") +
    theme(plot.title = element_text(color = "#2c3136", size = 16,
                                    family = "karla"),
          axis.title.x = element_text(color = "#2c3136", size = 10,
                                      family = "karla"),
          axis.title.y = element_text(color = "#2c3136", size = 10,
                                      angle = 90, family = "karla"),
          legend.title = element_text(color = "#2c3136", size = 9,
                                      family = "karla"),
          legend.position = "top",
          legend.justification = "right",
          legend.margin = margin(0,0,0,0),
          legend.box.margin = margin(-10,0,-10,-10),
          legend.key.size = unit(0.8, "line")) +
    scale_x_discrete(breaks = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11),
                     limits = c("1", "2", "3", "4", "5", "6", "7",
                                "8", "9", "10", "11"))
print(fig3)
```

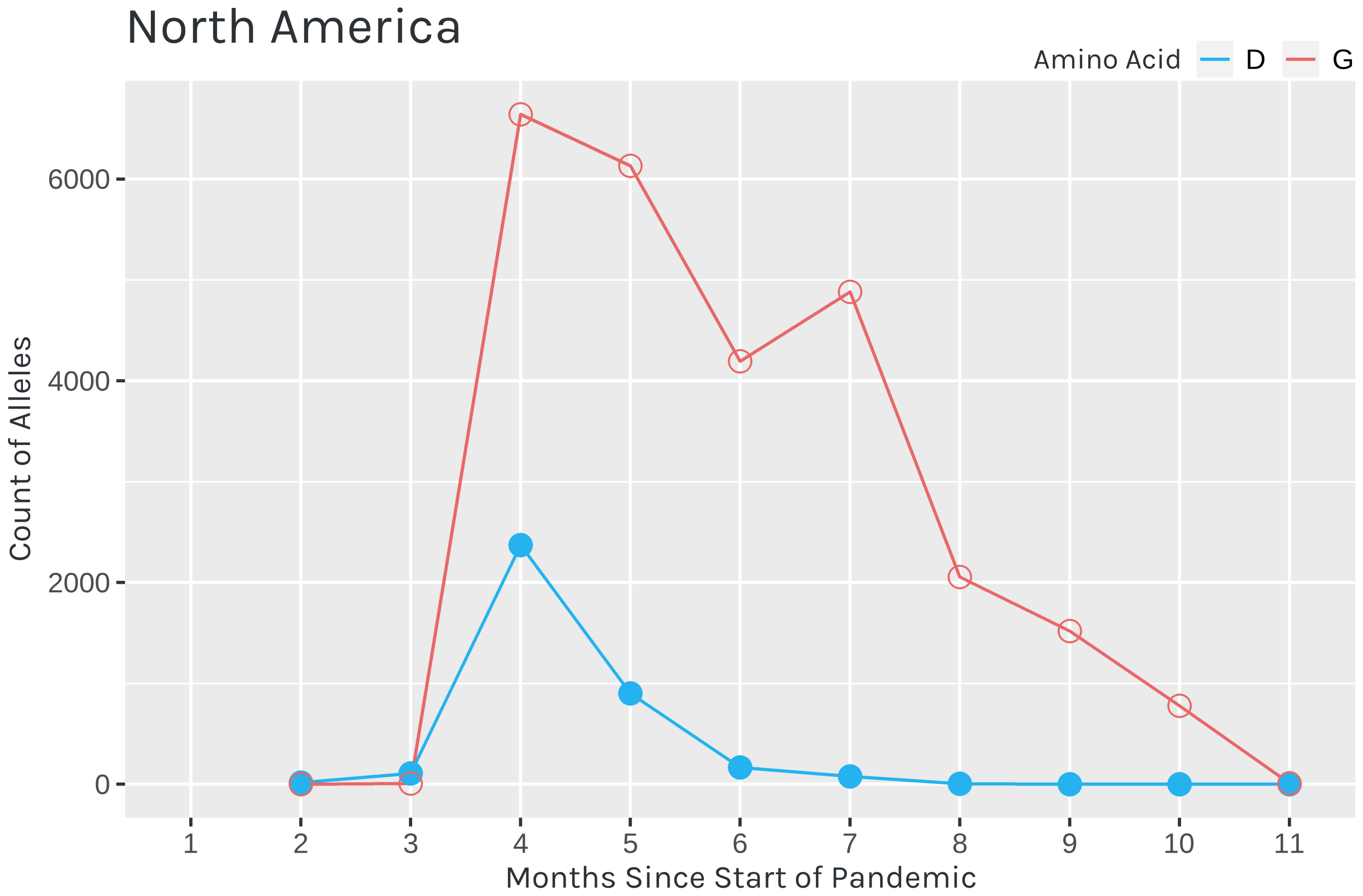

Figure 3: Number of North American observations of both the ancestral D and derived G alleles for every month since the start of the COVID-19 pandemic. The pattern seen worldwide generally also held in North America, with 1 month lag in observations/spread of the virus. Moreover, the observations for the G allele declined much faster than it did worldwide, suggesting that another viral locus arose and became dominant in North America earlier than average.

```
# Plotting relative frequency of alleles vs time for North America
fig4 <- ggplot(DGNAm1, aes(x = months, y = freq)) +
    geom_line(aes(color = "#25b2ef")) +
    geom_line(data = DGNAm2, aes(color = "#e86868")) +
    geom_point(size = 3, color = "#25b2ef") +
    geom_point(data = DGNAm2, size = 3, pch = 1, color = "#e86868") +
    scale_color_identity(name = "Amino Acid",
                         breaks = c("#25b2ef", "#e86868"),
                         labels = c("D", "G"),
                         guide = "legend") +
    ggtitle("North America") +
    xlab("Months Since Start of Pandemic") +
    ylab("Relative Allele Frequency") +
    theme(plot.title = element_text(color = "#2c3136", size = 16,
                                    family = "karla"),
          axis.title.x = element_text(color = "#2c3136", size = 10,
                                      family = "karla"),
          axis.title.y = element_text(color = "#2c3136", size = 10,
                                      angle = 90, family = "karla"),
          legend.title = element_text(color = "#2c3136", size = 9,
                                      family = "karla"),
          legend.position = "top",
          legend.justification = "right",
          legend.margin = margin(0,0,0,0),
          legend.box.margin = margin(-10,0,-10,-10),
          legend.key.size = unit(0.8, "line")) +
    scale_x_discrete(breaks = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11),
                     limits = c("1", "2", "3", "4", "5", "6", "7",
                                "8", "9", "10", "11"))
print(fig4)
```

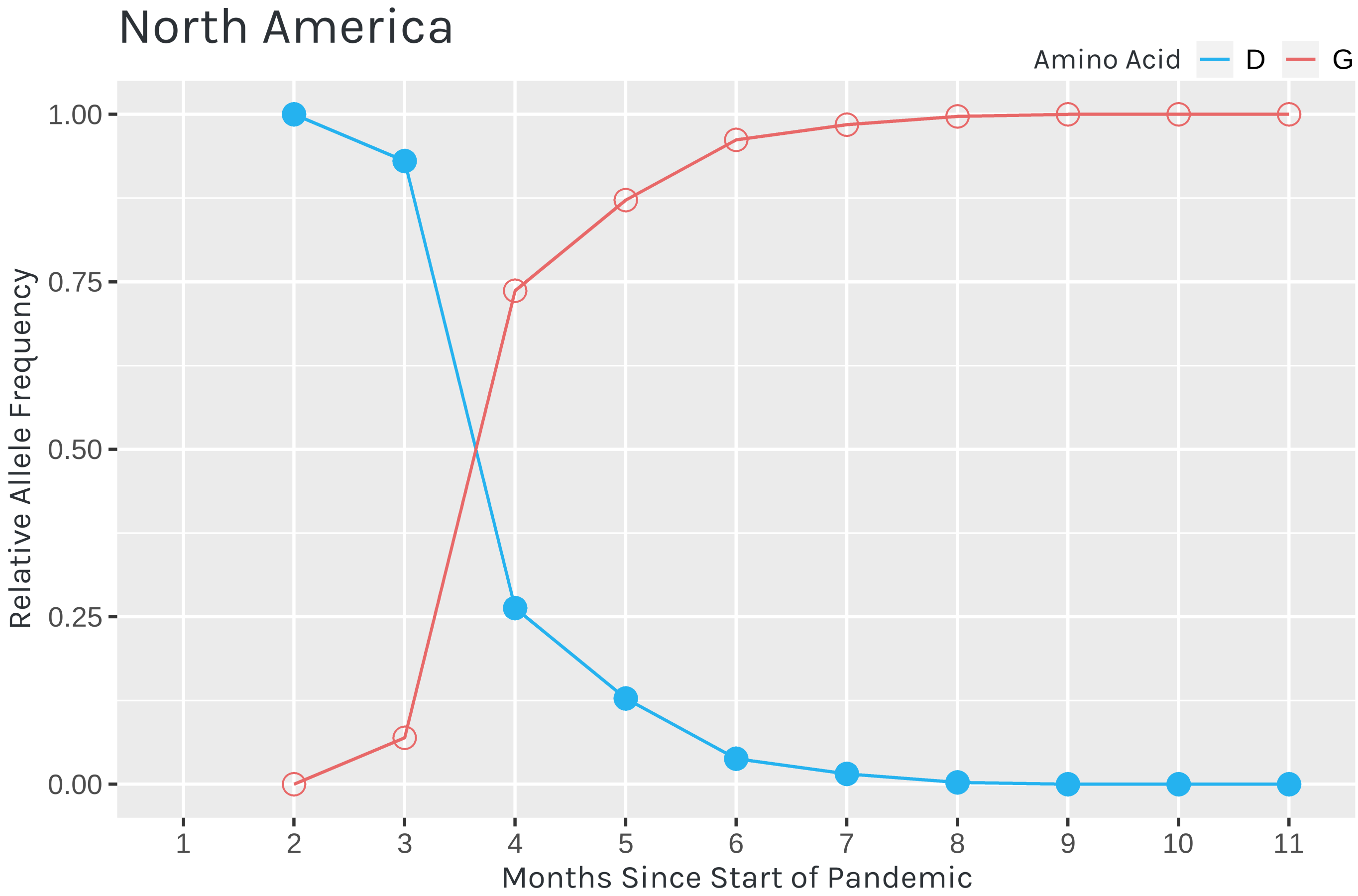

Figure 4: Relative frequencies of both the ancestral D and derived G alleles in North America for every month since the start of the COVID-19 pandemic.

African Counts

The following code and plots are specific to Africa:

```r
# Calculating relevant values for Africa
Af <- subset(D614G, Continent == "Africa")
n_Af <- nrow(Af)
DGAf_counts <- table(Af[, c(1, 6)])
DAf_counts <- DGAf_counts[1,]
GAf_counts <- DGAf_counts[2,]
Af_months <- as.integer(labels(DGAf_counts)$Month_Since_Start)
DGAf_n <- DAf_counts + GAf_counts
DAf_freq <- DAf_counts / DGAf_n
GAf_freq <- GAf_counts / DGAf_n

DGAf1 <- data.frame(months = Af_months,
                    cts = DAf_counts,
                    freq = DAf_freq)
DGAf2 <- data.frame(months = Af_months,
                    cts = GAf_counts,
                    freq = GAf_freq)
rownames(DGAf1) <- 1:nrow(DGAf1)
rownames(DGAf2) <- 1:nrow(DGAf2)

# Plotting count of alleles vs time for Africa
fig5 <- ggplot(DGAf1, aes(x = months, y = cts)) +
    geom_line(aes(color = "#25b2ef")) +
    geom_line(data = DGAf2, aes(color = "#e86868")) +
    geom_point(size = 3, color = "#25b2ef") +
    geom_point(data = DGAf2, size = 3, pch = 1, color = "#e86868") +
    scale_color_identity(name = "Amino Acid",
                         breaks = c("#25b2ef", "#e86868"),
                         labels = c("D", "G"),
                         guide = "legend") +
    ggtitle("Africa") +
    xlab("Months Since Start of Pandemic") +
    ylab("Count of Alleles") +
    theme(plot.title = element_text(color = "#2c3136", size = 16,
                                    family = "karla"),
          axis.title.x = element_text(color = "#2c3136", size = 10,
                                      family = "karla"),
          axis.title.y = element_text(color = "#2c3136", size = 10,
                                      angle = 90, family = "karla"),
          legend.title = element_text(color = "#2c3136", size = 9,
                                      family = "karla"),
          legend.position = "top",
          legend.justification = "right",
          legend.margin = margin(0,0,0,0),
          legend.box.margin = margin(-10,0,-10,-10),
          legend.key.size = unit(0.8, "line")) +
    scale_x_discrete(breaks = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11),
                     limits = c("1", "2", "3", "4", "5", "6", "7",
                                "8", "9", "10", "11"))
print(fig5)
```

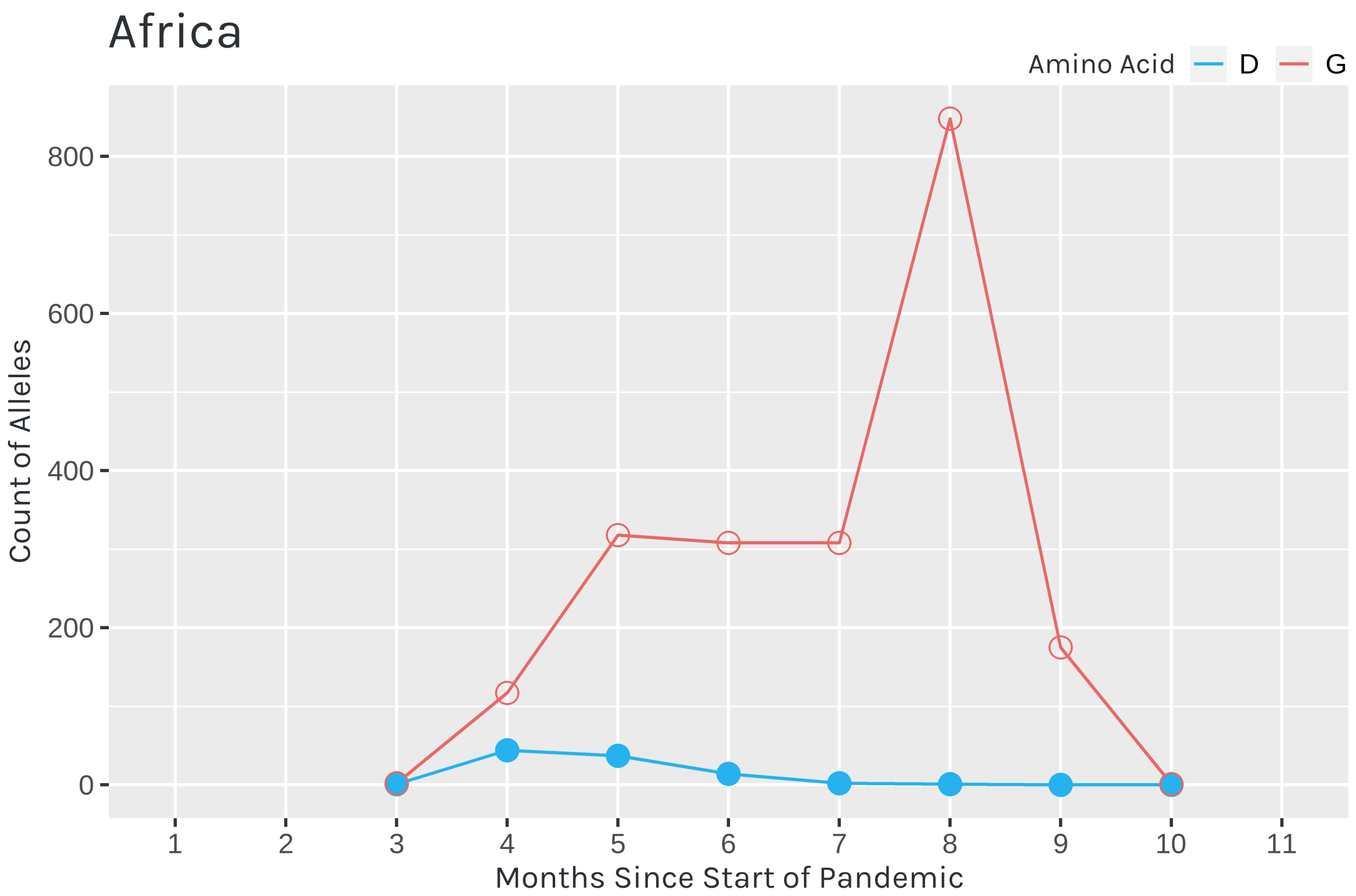

Figure 5: Number of African observations of both the ancestral D and derived G alleles for every month since the start of the COVID-19 pandemic. There is a two-month lag in African observations of the viral genome, and by the time observations began the prevalence of the D allele at codon 614 had already been greatly reduced compared to the G allele. Moreover, the peak of the G allele count occurred 8 months after the start of the pandemic, a stark contrast to both the worldwide and North American observations. Both alleles hit their minimum one month before the latter two as well.

```
# Plotting relative frequency of alleles vs time for Africa
fig6 <- ggplot(DGAf1, aes(x = months, y = freq)) +
    geom_line(aes(color = "#25b2ef")) +
    geom_line(data = DGAf2, aes(color = "#e86868")) +
    geom_point(size = 3, color = "#25b2ef") +
    geom_point(data = DGAf2, size = 3, pch = 1, color = "#e86868") +
    scale_color_identity(name = "Amino Acid",
                         breaks = c("#25b2ef", "#e86868"),
                         labels = c("D", "G"),
                         guide = "legend") +
    ggtitle("Africa") +
    xlab("Months Since Start of Pandemic") +
    ylab("Relative Allele Frequency") +
    theme(plot.title = element_text(color = "#2c3136", size = 16,
                                    family = "karla"),
          axis.title.x = element_text(color = "#2c3136", size = 10,
                                      family = "karla"),
          axis.title.y = element_text(color = "#2c3136", size = 10,
                                      angle = 90, family = "karla"),
          legend.title = element_text(color = "#2c3136", size = 9,
                                      family = "karla"),
          legend.position = "top",
          legend.justification = "right",
          legend.margin = margin(0,0,0,0),
          legend.box.margin = margin(-10,0,-10,-10),
          legend.key.size = unit(0.8, "line")) +
    scale_x_discrete(breaks = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11),
                     limits = c("1", "2", "3", "4", "5", "6", "7",
                                "8", "9", "10", "11"))
print(fig6)
```

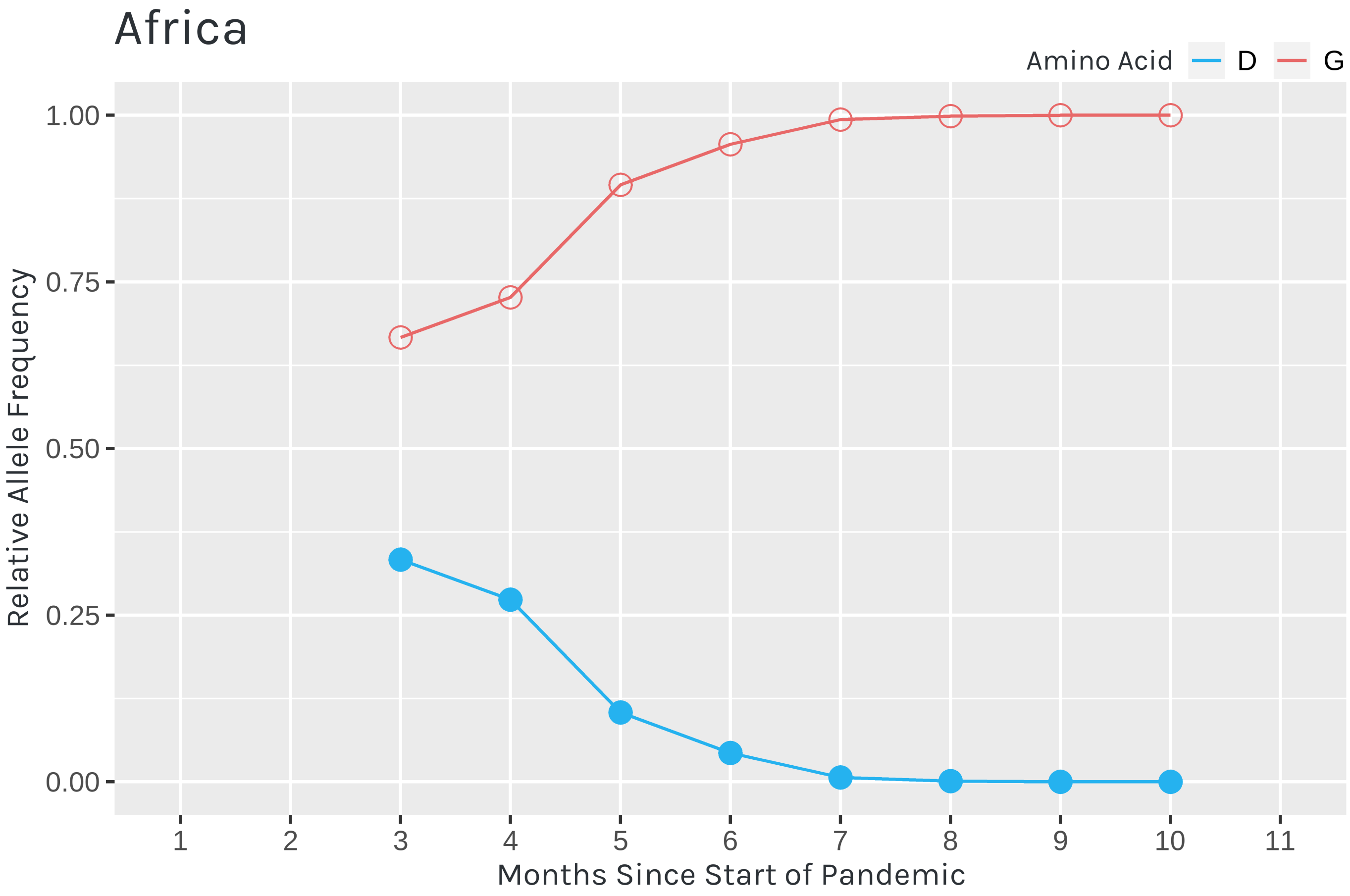

Figure 6: Relative frequencies of both the ancestral D and derived G alleles in North America for every month since the start of the COVID-19 pandemic. Compared to the other observations, the frequency of the G allele in Africa was never lower than the D allele. It also took longer for the difference in their frequencies to become large.

Based on the data above, it can be claimed that the G allele experienced positive selection early in the pandemic. In all three observation regions–North America, Africa, and the world as a whole–the frequency of the G allele compared to the D allele rose dramatically in the first few months after the initial outbreak, and always ended with a frequency of 1 at the end of the observation period. Moreover, the numerical count graphs show the number of observations of the G allele far surpass those of the D allele, with a count of around 850 for G in Africa during month 8 compared to nearly 0 for D, although in other plots both D and G allele counts had peaks at around four months.

From this it can be concluded that both the D and G alleles initially experienced positive selection (with the D allele having a lower selection strength) compared to other loci and not another evolutionary force such as genetic drift because there is such a strong and significant shift in prevalence from D-allele viruses to G-allele viruses in such a short period of time. It was around the peak of both alleles in month 4 that viruses with allele G at codon position 614 began to vastly out-compete those with allele D at the same position, and it may be that both viruses were able to coexist when the scope of the pandemic was still small and more geographically limited. At some tipping point, however, the G-allele virus became dominant as the virus spread and proliferated (it could be that the mutation allowed for faster spread or lower host mortality or some combination of the two and therefore “took over” when the pandemic started to become more serious). Interestingly, both viral types declined in prevalence after month 4, which could be the result of negative selection (another mutation may have lent even more fitness, thus out-competing both D614 and G614 viruses). Despite the frequency graphs indicating that the G allele continued to become more common, the numerical count graphs indicate that both populations were in decline, and thus negative selection acted on both viral types. It is still highly unlikely that genetic drift played a significant role in the patterns observed due to the large population size and because the pattern was seen globally (thereby affecting all populations of the coronavirus), and this conclusion is therefore strongly supported.

Part 2: Comparing D614G to other Loci in the Viral Genome

This section will compare the derived allele frequency of D614G to the derived frequencies of other loci throughout the viral genome in order to determine whether the precipitous rise of the G allele seen in the previous section is typical or not. In this new data set, there are 59 total observations. After calculations, the frequency of the G allele was found to be **0.796**. Four loci within the data set had an equal or higher derived allele frequency than the G allele, and the proportion of loci that had an equal or higher derived allele frequency was **0.068**.

```r
# Importing csv file
var_data <- read.csv("variation_summary_frequency.csv",
                     header = TRUE, sep = ",",
                     fileEncoding = "UTF-8-BOM")
G614freq <- var_data[which(var_data$Aamut == "D614G"), 2]
DAF_higher <- which(var_data$freq >= G614freq)
n_higher <- length(DAF_higher)
emp.p <- n_higher / nrow(var_data)

fig3 <- ggplot() +
    geom_histogram(data = var_data,
                   color = "white", fill = "#e86868",
                   aes(x = freq)) +
    geom_vline(xintercept = as.numeric(G614freq),
               color = "#25b2ef",
               size = 1) +
    ggtitle("Frequency of Derived Allele Frequencies",
            "Compared to the derived allele frequency of G614") +
    xlab("Derived Allele Frequencies") +
    ylab("Number of Sites with \n Derived Allele Frequency") +
    theme(plot.title = element_text(color = "#2c3136", size = 16,
                                    family = "karla"),
          plot.subtitle = element_text(color = "#2c3136", size = 10,
                                       family = "karla"),
          axis.title.x = element_text(color = "#2c3136", size = 10,
                                      family = "karla"),
          axis.title.y = element_text(color = "#2c3136", size = 10,
                                      family = "karla"))
print(fig3)
```

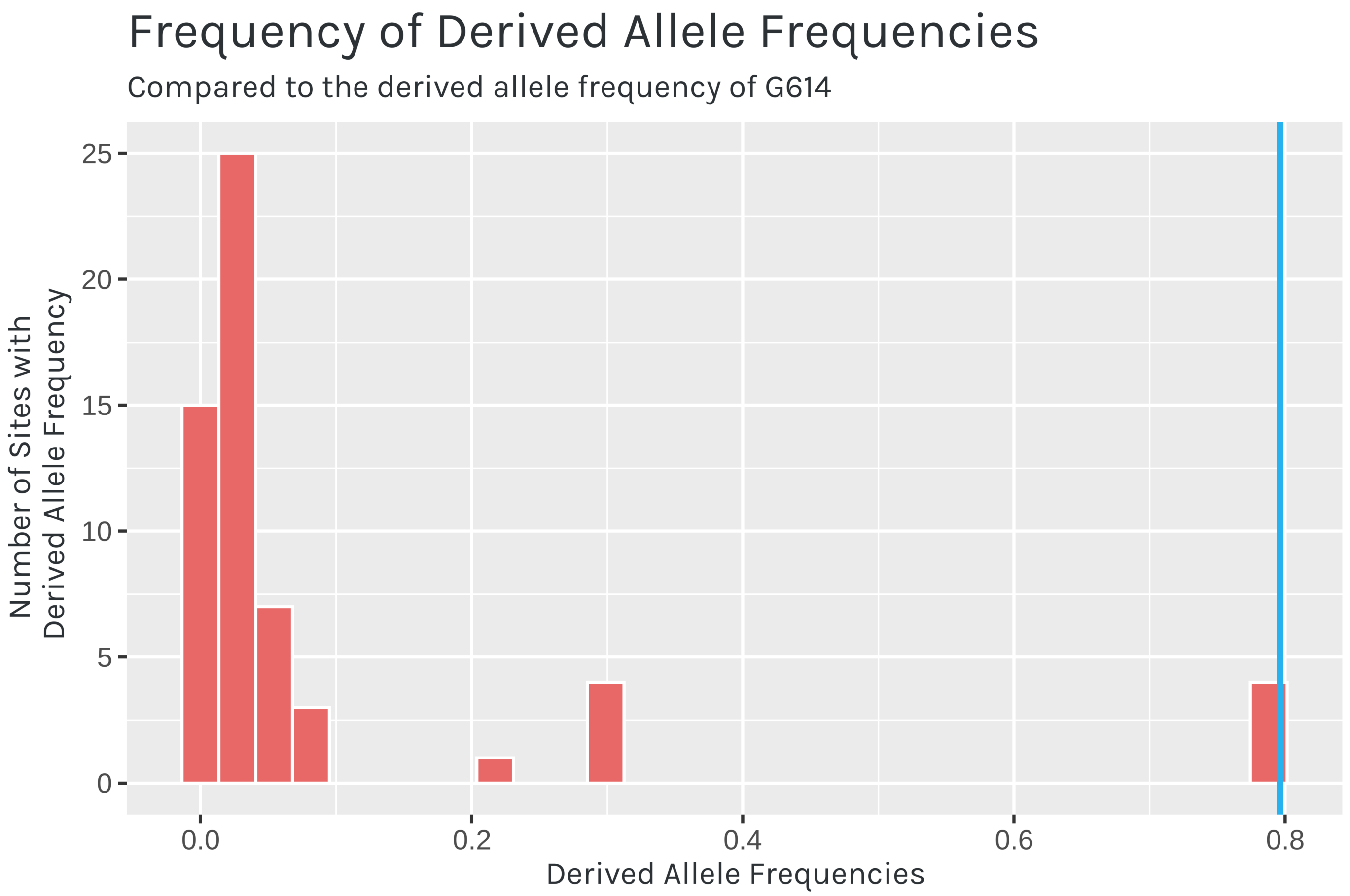

Figure 7: Histogram of the derived allele frequencies for various loci throughout the coronavirus genome, vertical line represents frequency of G614 allele only. Most loci in the viral genome had low allele frequencies, with the notable exception of the G614 allele.

From the graphs, it is clearly unusual for a particular locus to have a derived allele frequency of 0.8, a value which represents an outlier in the above graph. Most of the loci had a DAF of around 0.05. Sites like codon 614 and the 4 loci near it do seem to be under strong selection, which is evidenced by their comparatively high allele frequencies. This claim is supported by the data and can be made with reasonably high confidence. However, all the other loci could be affected either by negative selection, genetic drift, or both. It is difficult to make that claim from the data provided. There could be thousands of mutations that pop up only in certain populations around the world and are quickly removed from the genome through drift without ever spreading further afield just as easily as there could be thousands of mutations that pop up and spread out but are outperformed by endemic viruses with mutations that lend much more fitness. More data are needed to be sure.

Group Member Contributions

Kayden Adams: Helped with part 1 R code and conclusions

Sarah Conway: Helped with claims, conclusions, and figure captions for parts 1 and 2

Yousef Al Obaidan: Part 1 R code

Dylan Oh: Plots; R Markdown report; and R code, writing, and figure captions for part 2

Julia Thompson: Helped with claims and conclusions for part 2

Appendix

Base R Plot Code

```r
# Create the plot
windows(width = 5, height = 8)
par(mfcol = c(2, 1), mar = c(3, 3, 1, 0.2),
    mgp = c(1.75, 0.75, 0), oma = c(1, 1, 1, 1))

# Plot the allele count
plot(DGworld_months, DGworld_counts[1,],
     ylim = c(0, max(DGworld_counts)),
     col = 1, type = "b",
     xlab = "Month Since Start of Pandemic",
     ylab = "Count of Allele",
     main = "Worldwide")

points(DGworld_months, DGworld_counts[2,], col = 2, type = "b")

legend("topright", legend = c("D", "G"), title = "Amino Acid",
       bty = 'n', col = 1:2, pch = 1)

### Plot the frequency of each allele over time:
plot(DGworld_months, Dworld_freq, type = "b",
     xlab = "Months Since Start of Pandemic",
     ylab = "Allele Frequency",
     main = "Worldwide")
points(DGworld_months, Gworld_freq, type = "b", col = 2)

legend("right", legend = c("D", "G"), title = "Amino Acid",
       bty = "n", col = 1:2, pch = 1)

# The same methods were used to plot the graphs for North America and Africa
# and that code will not be included.
```