SARS-CoV-2 Gene Mutations across Various Geographic Regions

Jonathan Leshchinsky, Faneela Tahir, Jacqueline Vargas, Gyaban Essilfie-Bondzie

Introduction

The COVID-19 pandemic caused by the SARS-CoV-2 virus has had a significant impact on global health. As the virus spreads across different geographic regions, it may undergo mutations in its genetic code resulting in the emergence of new viral variants. In understanding the frequency and distribution of these mutations, we are better able to monitor the spread and evolution of the virus, better informing the legislature of public health policies.

SARS-CoV-2 is a single-stranded, positive-sense[1] RNA virus with a mutation rate of approximately $1*10^{-6}$ - $2*10^{-6}$ mutations per nucleotide per replication cycle[2]. This mutation rate is relatively low when compared with other RNA viruses such as HIV and hepatitis C virus[2]. Nevertheless, SARS-CoV-2 viral mutation rate contributes to its ability to evade human immune response. The viral genome consists of several genes including the spike (S), nucleocapsid (N), and RNA-dependent RNA polymerase (nsp12) genes; mutations in these genes can affect viral transmissibility, infectivity as well as its response to treatments or vaccines.

The present study aims to investigate whether there are significant differences in the frequency and distribution of mutations in the S, N, and nsp12 genes of SARS-CoV-2 isolates collected at specific dates in the USA and UK. Specifically, we display three (3) hypotheses for investigating the data collected: The frequency and distribution of gene mutations in the S, N, and nsp12 genes of SARS-CoV-2 isolates differ significantly between each of the three genes; the frequency and distribution of gene mutations in the S, N, and nsp12 genes of SARS-CoV-2 isolates differ significantly between the US and UK geographic regions; the frequency and distribution of gene mutations in the S, N, and nsp12 genes of SARS-CoV-2 isolates differ significantly for the isolate collection dates. Addressing each of these hypotheses will be useful in understanding the genetic diversity of SARS-CoV-2 for each of the observed data.

Methods & Materials

The data used contains a cumulative count of mutations detected in three genes of the SARS-CoV-2 virus across isolates obtained from the USA and UK geographical origins. The dataset consists of 5 columns. The first column is “Isolate” which is the ID of every isolate; the second column is “Collection_date” which dates the exact day the isolates were collected in the format MM-DD-YYYY; isolates were collected over the course of 1 year (April, July, October 2020 & January, 2021). The third column is “Geo_origin” which states the geographical origin of the isolates, either the United States (USA) or the United Kingdom (UK). Isolates originated from the US and the UK; the fourth column is “Muts_counts” and denotes the number of mutations observed per gene per isolate. Mutation count ranged from 0 to 8. The fifth column is “Gene” and names the specific SARS-CoV-2 gene where the mutation was observed. The genes of interest in this study were the S, N, and the nsp12 genes.

For each hypothesis tested, we present a null and alternative hypothesis for mutation count and gene; mutation count and geographic region; mutation count and collection date, respectively.

1. H_0 : There is no significant difference in the frequency and distribution of SARS-CoV-2 mutations across the S, N, and nsp12 genes.
    a. H_a : There is a significant difference in the frequency and distribution of SARS-CoV-2 mutations across the S, N, and nsp12 genes.
2. H_0 : There is no significant difference in the frequency and distribution of SARS-CoV-2 mutations across the USA and UK geographic regions.
    a. H_a : There is a significant difference in the frequency and distribution of SARS-CoV-2 mutations across the USA and UK geographic regions.

3. H_0 : There is no significant difference in the frequency and distribution of SARS-CoV-2 mutations across April, July, October 2020 and January 2021 collection dates.
    a. H_a : There is a significant difference in the frequency and distribution of SARS-CoV-2 mutations across April, July, October 2020 and January 2021 collection dates.

Violin Plot R-Studio code

```
ggplot(covid, aes(x=Gene, y=Muts_counts,color=Gene))+
geom_violin(aes(fill=Gene))+ facet_grid(rows=vars(Geo_origin))+
  labs(x="Types of Genes", y="Number of Mutations")+ theme_bw()+
  theme(axis.title = element_text(face="bold", size=13, color="purple"), axis.text =
element_text(face="bold", size=11, color="black"), strip.text =
element_text(face="bold", size=13, color="purple"), legend.position = "none")
```

Our first hypothesis noted that there was a significant difference in the frequency and distribution of SARS-CoV-2 mutations across the USA & UK. Hence, we chose a violin plot to demonstrate the number of mutations per gene-of-interest, per geographic region. Wider areas of the plot correlate to the frequency of isolates that have that specific number of mutations in the gene. For example, it could be said that a large frequency of US isolates only had 1 mutation in the S gene.

Scatterplot Linear Regression R-Studio code

```
ggplot(covid, aes(x=Collection_date, y=Muts_counts)) + geom_point()+
geom_smooth(method=lm) + theme_bw()+
  labs(x="Collection date", y="Number of Mutations") +
  theme(axis.title = element_text(face="bold", size=13, color="purple"), axis.text =
element_text(face="bold", size=11, color="black"))
```

Our second hypothesis analyzed the significant difference in frequency and distribution of SARS-CoV-2 mutations across the different ranges of collection dates. Thus, the data visualization method we employed was a scatter plot with linear regression. This selection was based on its ability to determine whether there was any relationship between mutation counts and the collection date. The inclusion of the linear regression model allowed us to determine the goodness of fit and evaluate how well the model fits the observed data.

Histogram #1 R-Studio code

```
ggplot(covid, aes(x=Muts_counts,
fill=as.factor(Muts_counts)))+geom_histogram(binwidth=1)+ theme_bw()+
facet_grid(rows=vars(Gene)) +
labs(x="Number of Mutations", y="Number of Isolates") +
  theme(axis.title = element_text(face="bold", size=13, color="purple"), axis.text =
element_text(face="bold", size=11, color="black"), strip.text =
element_text(face="bold", size=13, color="purple"),legend.title=element_blank(),
legend.text=element_text(face="bold", size=10), legend.position="top" )
```

In line with the third hypothesis, a histogram was utilized as an approach in understanding the spread of mutations across the three different genes (N,S, nsp12) with reference to the total number of isolates analyzed. A histogram will allow for easy identification of the genes that exhibit high and low mutation counts and will provide insight into their relative frequency.

Histogram #2 R-Studio code

```
ggplot(covid, aes(x=Muts_counts,
fill=as.factor(Muts_counts)))+geom_histogram(binwidth=1)+ theme_bw()+
facet_grid(rows=vars(Geo_origin)) +
 labs(x="Number of Mutations", y="Number of Isolates") +
 theme(axis.title = element_text(face="bold", size=13, color="purple"), axis.text =
element_text(face="bold", size=11, color="black"), strip.text =
element_text(face="bold", size=13, color="purple"), legend.title=element_blank(),
legend.text=element_text(face="bold", size=10), legend.position="top" )
```

Similar to the violin plot, the histogram analyzes mutation count per geographic region over the total number of isolates collected. This will allow us to easily identify the genes exhibiting higher and lower mutation count.

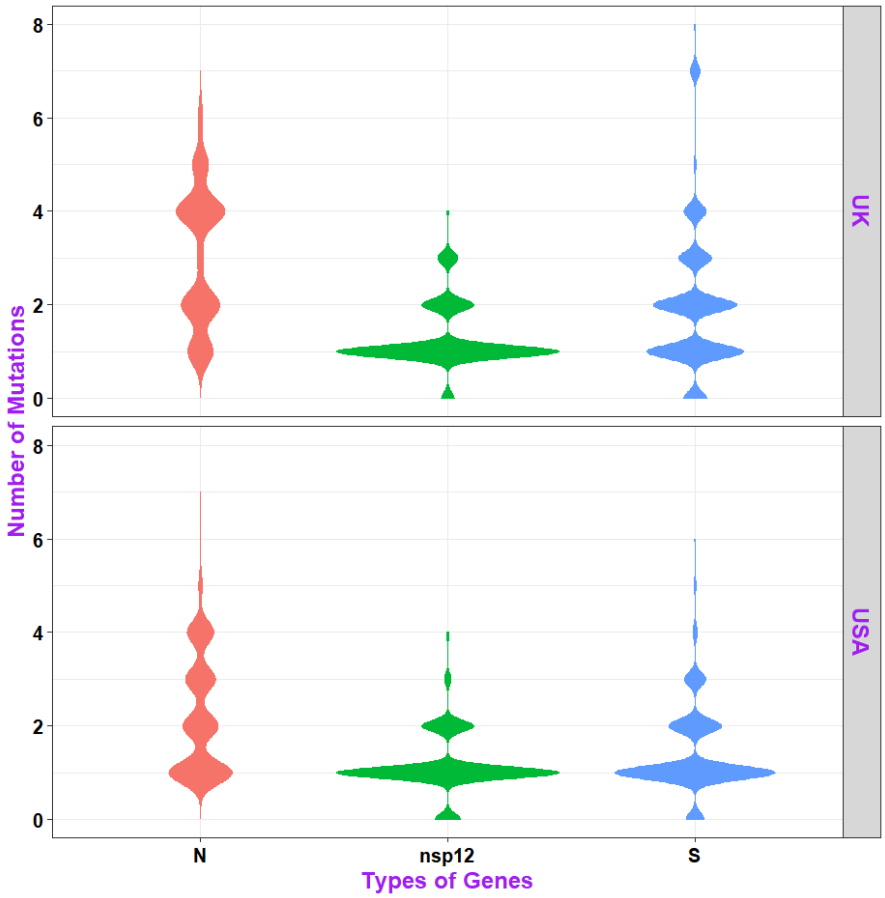
Statistical Analysis R-Studio code

```
group3_dataset<-read.csv("group3_dataset.csv")
aov(Muts_counts~Geo_origin,data=group3_dataset)
myaov<-aov(Muts_counts~Geo_origin,data=group3_dataset)
summary(myaov)
aggregate(group3_dataset$Muts_counts, by=list(group3_dataset$Geo_origin), FUN=mean)
aggregate(group3_dataset$Muts_counts, by=list(group3_dataset$Geo_origin), FUN=sd)
group3_dataset<-read.csv("group3_dataset.csv")
aov(Muts_counts~Gene,data=group3_dataset)
summary(myaov)
group3_dataset<-read.csv("group3_dataset.csv")
aov(Muts_counts~Collection_date,data=group3_dataset)
summary(myaov)
```

An ANOVA (Analysis of Variance) t-test for statistical analysis was conducted to provide insight into whether there are significant differences in means among the groups being tested in each of the three (3) hypotheses. Further, ANOVA compares the variability between groups and

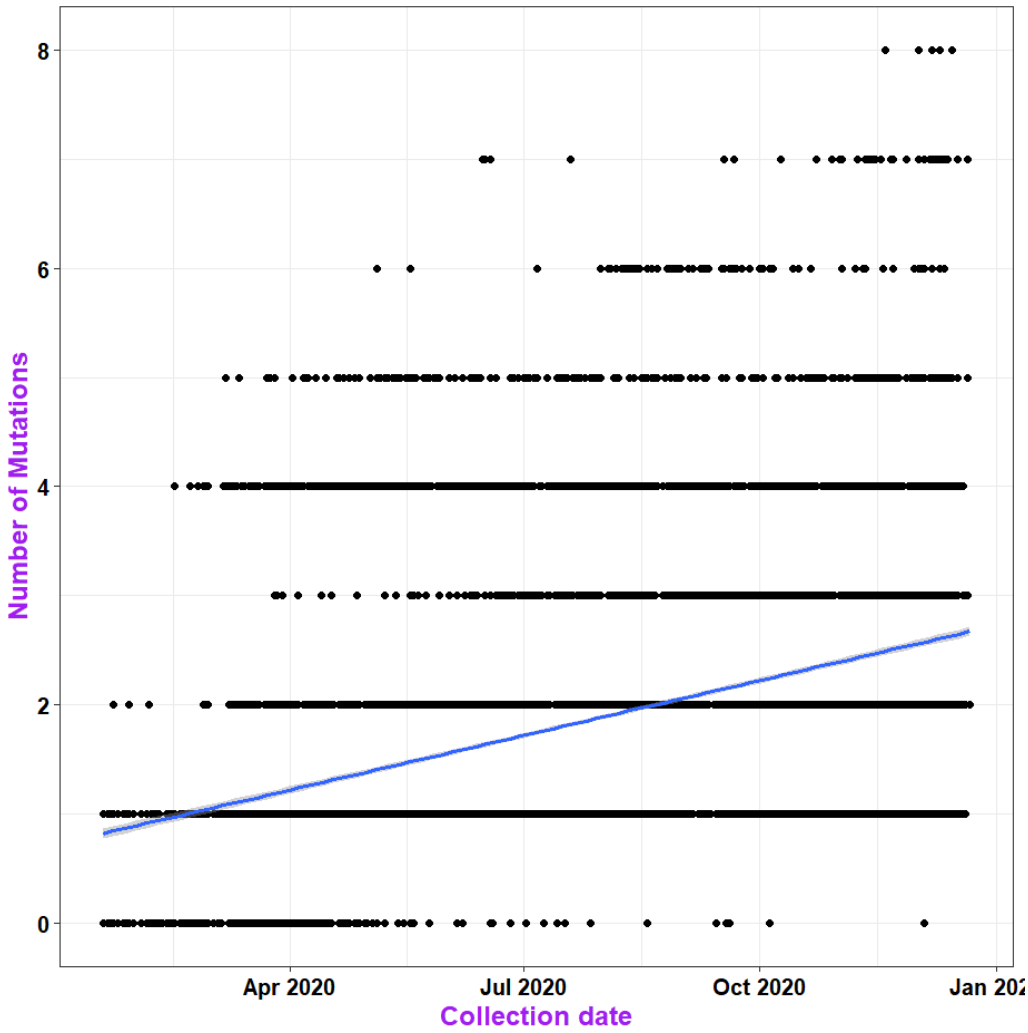
allows for the assessment of whether or not the observed differences are due to random variation or significant group difference.

Results:



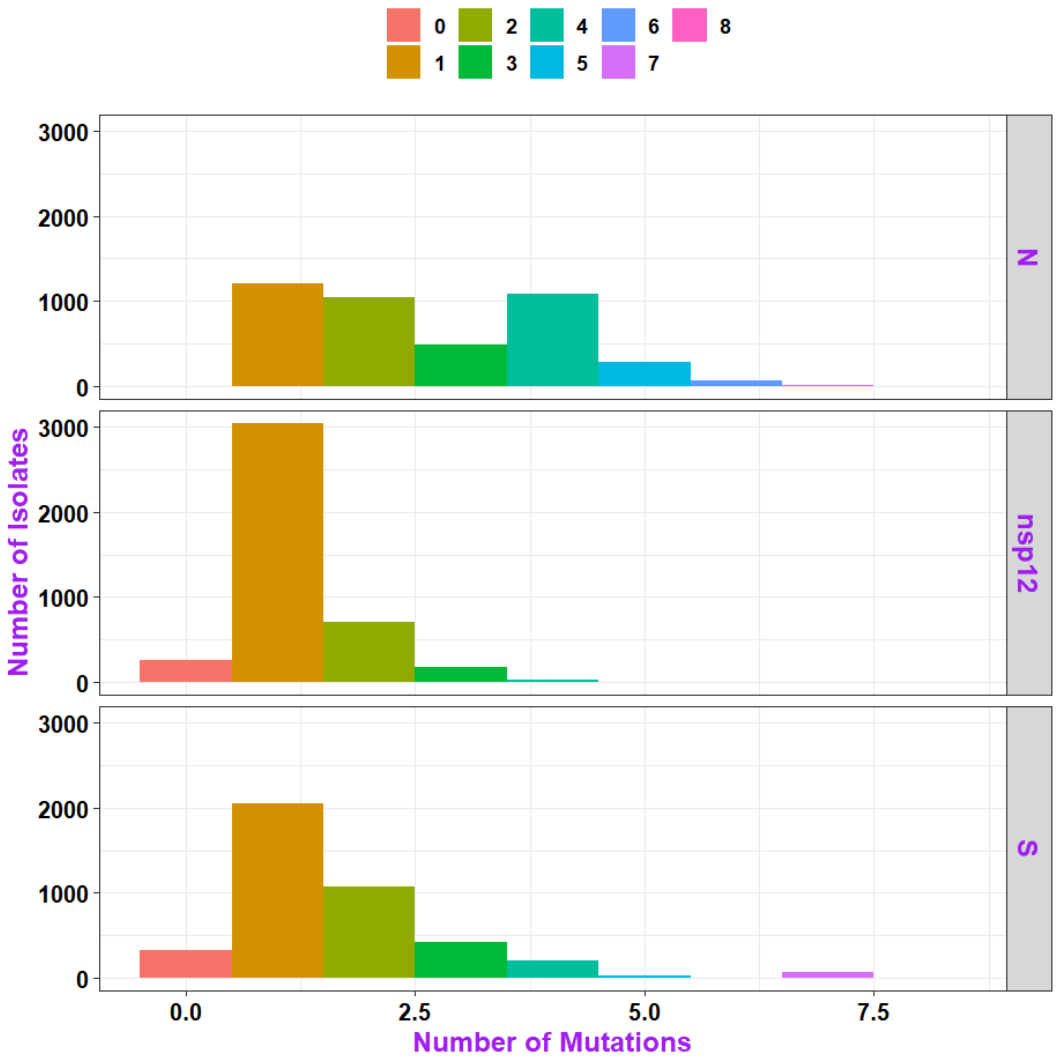
Through analysis of the violin plot, we observed the density of specific genes (x-axis) against their relative number of mutations (y-axis) per geographic region, USA and UK. In genes found in the USA, we observe that the S, N, nsp12 genes have a higher frequency of mutations around 1. Conversely, in the UK region, the N gene observed an average mutation count of around 4, while the density of the mutation count of the nsp12 gene is higher at approximately 1. Lasty, the S gene had the highest density mutation count at approximately 1 and 2. These findings suggest

that the distribution of mutation count for the observed genes have more variation in the UK region.



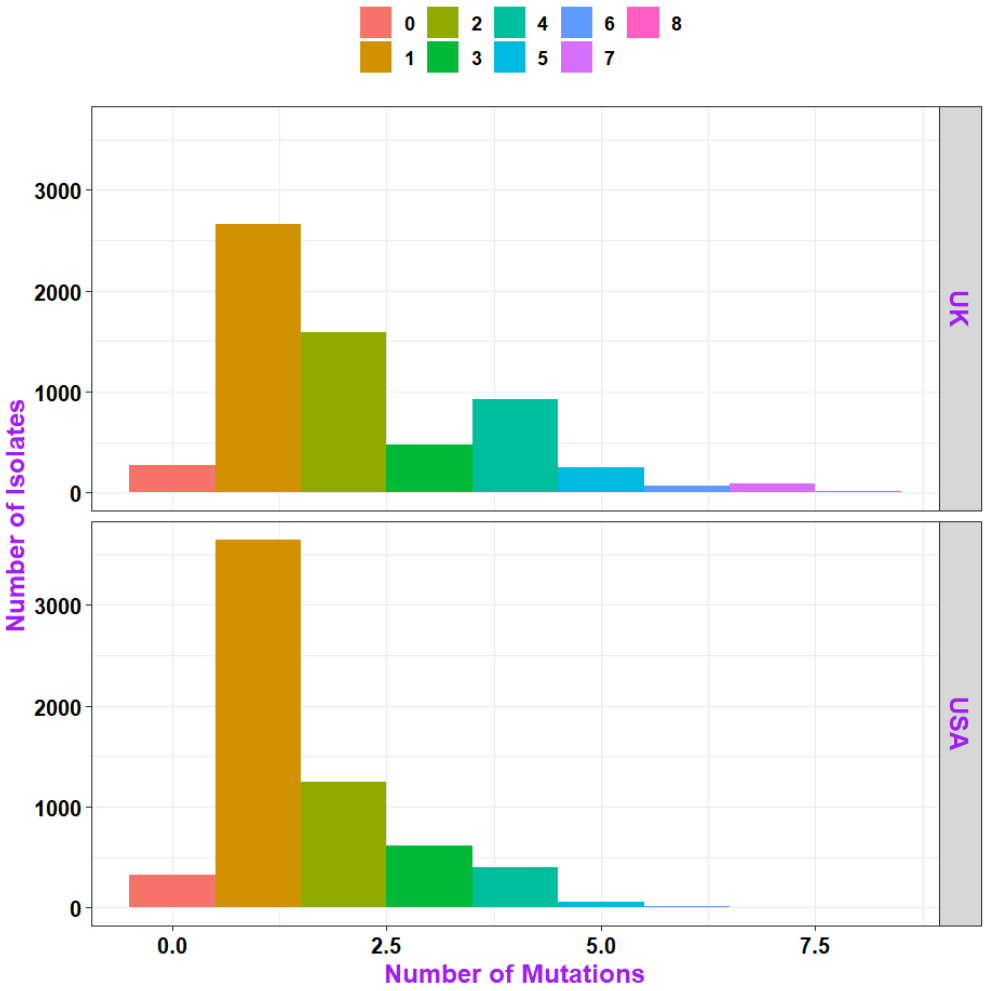
Based on the scatterplot and linear regression model, it seems that the number of mutations in isolates gradually increased as time went on. For example, nearing the end of 2020 a large portion of the isolates collected were observed to have 6-8 mutations in either of three SARS-CoV-2 genes. At the beginning of the year, however, isolates only had 0-2 mutations in their genes. These findings do make sense because the virus was able to continuously replicate as time went on. Given their fast replication cycle, mutations are more likely to accumulate; the

mutations may be beneficial to the virus which causes it to maintain a select number of mutations.



Based on the histogram, the number of mutations expressed per isolate depended on the specific gene being observed. The nsp12 gene saw the largest number of isolates expressing 1 mutation. Similarly, the S gene was the only gene to show isolates that exhibited 7 mutations. The N gene had a significantly high number of isolates that contained 4 mutations. These differences could be attributed to selective pressures that influenced mutations to occur,

potentially depending on what that specific gene was responsible for. Genes involved in transmission should have seen more mutations to amplify the viruses ability in transmission and infection.



Based on the histogram, the number of mutations that isolates expressed depended on the geographical origin. The UK was shown to have isolates that exhibited a wide range of mutation counts; this was not the case for the US since a large majority of isolates only had 1 mutation and the highest mutation was observed at 6. These findings suggest that different environmental pressures on the isolates, such as the climate, could have been crucial in altering virus

transmissibility, causing isolates to accumulate more mutations in order to increase their chances of survival.

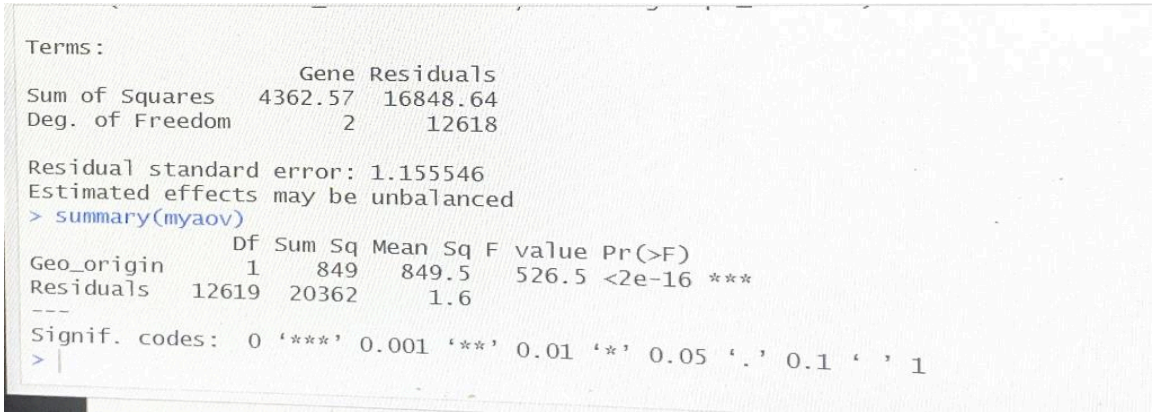
Statistical Analysis:

In order to evaluate our hypotheses regarding potential statistical differences between multiple groups of the dataset, we performed a t-test using ANOVA in R studio. This approach allowed us to obtain important statistical measures including the F-statistic, mean square, degrees of freedom and associated p value per hypotheses tested. Given our focus on the p-value, we interpreted its results in the context of the study performed. When performing the ANOVA in R-Studio, p-values considerably smaller than the predetermined alpha of 0.05 were obtained. The observations from the ANOVA provided compelling support of our alternate hypotheses—the indication of a statistically significant difference in mutation count across each of the three genes analyzed (S, N, nsp12), the USA and UK, and each collection date.

The following is a t-test result between mutation count and geographic origin (USA and UK)

```
> myaov<-aov(Muts_counts~Geo_origin,data=group3_dataset)
> summary(myaov)
              Df Sum Sq Mean Sq F value Pr(>F)
Geo_origin     1    849   849.5   526.5 <2e-16 ***
Residuals   12619  20362     1.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

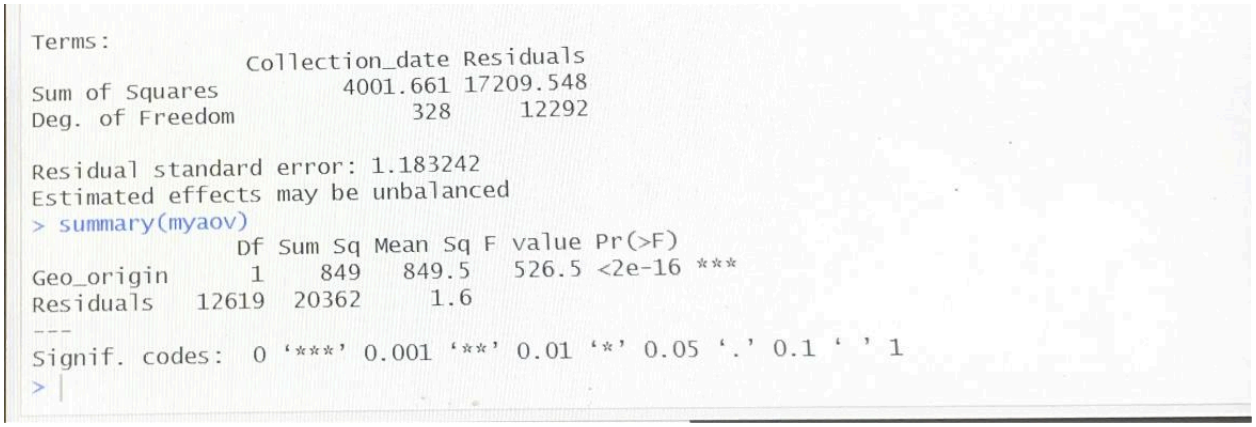
The following is a t-test result between mutation count and observed genes (S, N and NSP12)



```
Terms:
                    Gene Residuals
Sum of Squares   4362.57  16848.64
Deg. of Freedom        2     12618

Residual standard error: 1.155546
Estimated effects may be unbalanced
> summary(myaov)
              Df Sum Sq Mean Sq F value Pr(>F)
Geo_origin     1    849   849.5   526.5 <2e-16 ***
Residuals  12619  20362     1.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Lastly, we present the t- test results for mutation count versus collection date:



```
Terms:
                Collection_date Residuals
Sum of Squares         4001.661 17209.548
Deg. of Freedom             328     12292

Residual standard error: 1.183242
Estimated effects may be unbalanced
> summary(myaov)
              Df Sum Sq Mean Sq F value Pr(>F)
Geo_origin     1    849   849.5   526.5 <2e-16 ***
Residuals  12619  20362     1.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Conclusion:

Based on our statistical analysis using ANOVA, we find compelling evidence to reject the null hypotheses. The p-values we obtained after performing t-tests for all three (3) hypotheses was approximately <2e-16, which is less than the alpha of 0.05. In light of this, we reject our null hypotheses in favor of the alternative hypotheses. Our analysis revealed significant differences in frequency and distribution of gene mutations in the S (spike), N (nucleocapsid) and nsp12 (RNA-dependent RNA polymerase) genes. These differences were also observed between the USA and UK geographic regions and between specific data collection dates (April, July, October 2020, January 2021). The following t-test results aided us to draw the following conclusions:

- There was a significant difference in the frequency and distribution of SARS-CoV-2 mutations across the USA & UK geographic regions.
- There was a significant difference in the frequency and distribution of SARS-CoV-2 mutations across the different genes observed (S, N, nsp12).
- There was a significant difference in the frequency and distribution of SARS-CoV-2 mutations at the given collection dates ranging from April, July, October 2020 and January 2021.

References

1. Naqvi, Ahmad Abu Turab, et al. “Insights into SARS-COV-2 Genome, Structure, Evolution, Pathogenesis and Therapies: Structural Genomics Approach.” *Biochimica et Biophysica Acta. Molecular Basis of Disease*, 1 Oct. 2020, www.ncbi.nlm.nih.gov/pmc/articles/PMC7293463/.
2. Markov, P.V., Ghafari, M., Beer, M. et al. The evolution of SARS-CoV-2. *Nat Rev Microbiol* **21**, 361–379 (2023). https://doi.org/10.1038/s41579-023-00878-2