Completed Comparative Analysis of SARS-CoV-2 Haut-de-France Viral Sequenced Samples

Vivian Lynn Ly

20 May, 2021

Background and Overview

This rmarkdown file will use the `vcfR` (Knaus and Grünwald, 2016), `ggplot2` (Wickham, 2016), `dplyr` (Wickham *et al.*, 2021), and `covdata` (Healy, 2020) packages in order to analyze positive sequenced SARS-CoV-2 samples from the Haute-de-France after it has been processed:

Haute-de-France directly translates to the North of France is comprised of 5 cities: Oise, Aisne, Nord, Pas-de-Calais, and Somme. Towards the beginning of the pandemic back in March 2020, France recorded about only about 30 confirmed cases of the Corona-virus. However, the few confirmed cases of the Corona-virus does not explain the huge spike in COVID-10 related deaths. More analysis is required to understand how cases and reported deaths have increased.

More specifically, this pandemic has tremendously affected the wine industry, suburban neighborhoods, holidays, rituals, and most importantly France's economy. The objective of this analysis is to thoroughly comprehend how variants of SARS-CoV-2 has affected hospitalizations due to COVID-19 and recorded deaths due to COVID-19 by analyzing mobility information and variant tracking of genes.

How has mobility and gene variation in Haute-de-France positive sequenced samples affect the cases of hospitalization and deaths due to COVID-19?

Methods

See the set of tutorials on the vcfR package website.

You may also want to use any of a range of different COVID data packages and data sources:

- https://kjhealy.github.io/covdata/
- https://github.com/como-ph/oxcovid19
- https://ropensci.org/blog/2020/10/20/searching-medrxivr-and-biorxiv-preprint-data/
- https://covidtracking.com/data/api
  - `readr::read_csv("https://api.covidtracking.com/v1/states/daily.csv")`

In order to compare how the Corona-virus has progressed within 2019-2021, I extracted information of cases and deaths related to SARS-CoV-2 variants from covdata's database. The `vcfR` (Knaus and Grünwald, 2016) package then will be used visualize, manipulate, and filter quality in vfcR files after it has been processed through the makefile within the git repository. Afterwards, I will use `ggplot2` and `ggthemes` (`???`) to create visual representations of the processed samples from the Haute-de-France. The `dplyr` (Wickham *et al.*, 2021) package will be used to further filter information from the provided datasets.

This data is available from the following URL: https://www.ncbi.nlm.nih.gov/bioproject/724410

Accession Number: PRJEB43269 Sequencing Technology Platform: ILLUMINA

Results

In Figure 1, after sequencing all unique SNP locations within each gene across all samples, the S, N, and ORF3a genes contained the most SNPs location, indicating possible regions where mutations could occur. The S gene consisted of the most unique SNP locations, followed by the N gene and then the ORF3a gene. When the number of COVID-19 cases in France was extracted and constructed onto a line plot, the number of cases at the beginning of the epidemic in France was deficient and exponentially increased through time. On the other hand, when the number of deaths in France was recorded and then graphed onto a line plot, the number of deaths was extraordinarily high and then drastically fell before consistently rising during July 2020. The pattern illustrated on the diagram was coherent to the mobility levels by transit recorded from March 2020 until the present. Simultaneously, there were very low transportation levels during April 2020; however, during July 2020, these mobility levels continued to increase until their peak during October 2020. These observations were similar to the mobility levels of transportation by walking and driving. As transportation mobility levels increase, higher variation was observed as we tracked the variation of the S, N, and ORF3a genes on a point diagram. Over the different collection dates, there is a higher variation of SNPs alterations from the previous collection date.

Discussion

The mobility trends correlate to the cases of COVID-19 and deaths that occur. However, this still does not explain the very low cases of COVID-19 in the beginning despite huge spike in deaths. The changes in variation of the genes demonstrate how SARS-CoV-2 is able to mutate and become more transmissible over the different collection dates. The amount of mutations that occur amongst the genes demonstrate the number of possible mutations that have affected the SARS-CoV-2 variant. After research it is possible that the low cases in the beginning was due to lack of testing kits available and complicated methods to reach medical attention.

SARS-CoV-2, severe acute respiratory syndrome Corona-virus two, created a new infectious disease that originated back in December 2019 was initially detected from an extensive seafood and animal market in Wuhan, China, within the Hubei province. Scientists used a phylogenetic analysis from patients who demonstrated symptoms of the Corona-virus, including troubled breathing, pain in the chest area, and pale to blue skin tone, to investigate the genomic sequences of the novel Corona-virus. After full-length analysis of genomic sequences, scientists discovered the common origin of SARS-CoV-2 to the original SARS virus, which had affected China back in 2003. The International Committee classified this virus as severe acute respiratory syndrome coronavirus (SARS-CoV-2). SARS-CoV-2 genome consists of 14 Open Reading Frames (ORFs) that encode approximately 27 proteins. (Wang *et al.*, 2020) The most important protein responsible for the binding to receptors of the host cell is the spike surface glycoprotein that the S-gene encodes. The essential roles of this glycoprotein facilitate the selection of a host, mediates the receptor binding, and membrane fusion to the host.

Despite comparing the mobility of three different types of transportation to the number of recorded cases and deaths, it is still unclear how during the beginning of the epidemic in France, there is a large number of deaths recorded even though cases confirmed were deficient. Further investigation leads to the finding of this article called, "Underdetection of cases of COVID-19 in France threatens epidemic control." Towards the beginning, when France entered its first lockdown due to the Corona-virus, approximately 9 out of 10 symptomatic cases were undetected because of a lack of encouragement to seek healthcare for suspected cases. (Pullano *et al.*, 2021) Several weeks later, the healthcare system still seemed insufficient in detecting cases. This issue is possibly due to the lack of testing kits, or individuals with symptoms of the Corona-virus did not feel it necessary to get tested. More examination of testing kits France had during the epidemic is required. For France to improve the detection of SARS-CoV-2, more aggressive and efficient testing is required to lift the restrictive measures within Europe. Approximately 92% of individuals who had suspected cases of COVID-19 required a prescription for a medical test to confirm, but only 31% of those individuals were willing to consult a doctor. (Pullano *et al.*, 2021) The process of acquiring a test and having it prescribed

by a doctor is complicated and inconvenient. The overcomplication of this process possibly explains why the state had such low detection of confirmed cases observed even though high numbers of deaths were recorded. Eventually, throughout the epidemic, large-scale communication campaigns were implemented to increase awareness for individuals to seek healthcare facilities despite mild symptoms. More robust surveillance of cases associated with the Corona-virus detection improved data accuracy and a stronger correlation between cases and deaths demonstrated in Figure 1 and 2.

After testing conditions had improved in France, the detection of SARS-CoV-2 variants were more closely analyzed to reveal that approximately variants from SARS-CoV-2 infections increased 36% transmission advantage responsible for most infections by the end of week 7 in 2021. (Haim-Boukobza *et al.*, 2021) The increase in the transmission rate is consistent with our analysis of the variation of the S gene in Figure 7. Throughout the different collection dates, we can see that the variation increases. Since the S gene encodes the spike glycoprotein that controls the transmissibility of SARS-CoV-2, it is expected that the cases of COVID-19 increased throughout the epidemic until vaccination was established. Understanding the rapid spread of SARS-CoV-2 provided critical insights into how variants of this infectious disease mutated over time.

According to the New York Times, France Corona-virus Map and Case Count, approximately 5,882,800 cases of the novel Corona-virus have been confirmed, and 107,280 related deaths have been reported in the entire country of France. In the Northern Region of France, Haute de France, 8,724 deaths have been recorded. (Allen *et al.*, 2020) Examining the subpopulations that have been hospitalized from these infectious diseases, 15% of these subpopulations passed away quickly after admission, and 85% of the subpopulation died at more extended periods. (Salje *et al.*, 2020) Notably, France emphasizes the importance of seeking medical attention when necessary, making availability to testing and care more convenient. Without herd immunity, it will not be easy to control the coronavirus cases amongst the population efficiently, given how often the genes in association to the SARS-CoV-2 can mutate. The variation of these genes are visually demonstrated in Figures 7, 8, and 9.

Figures

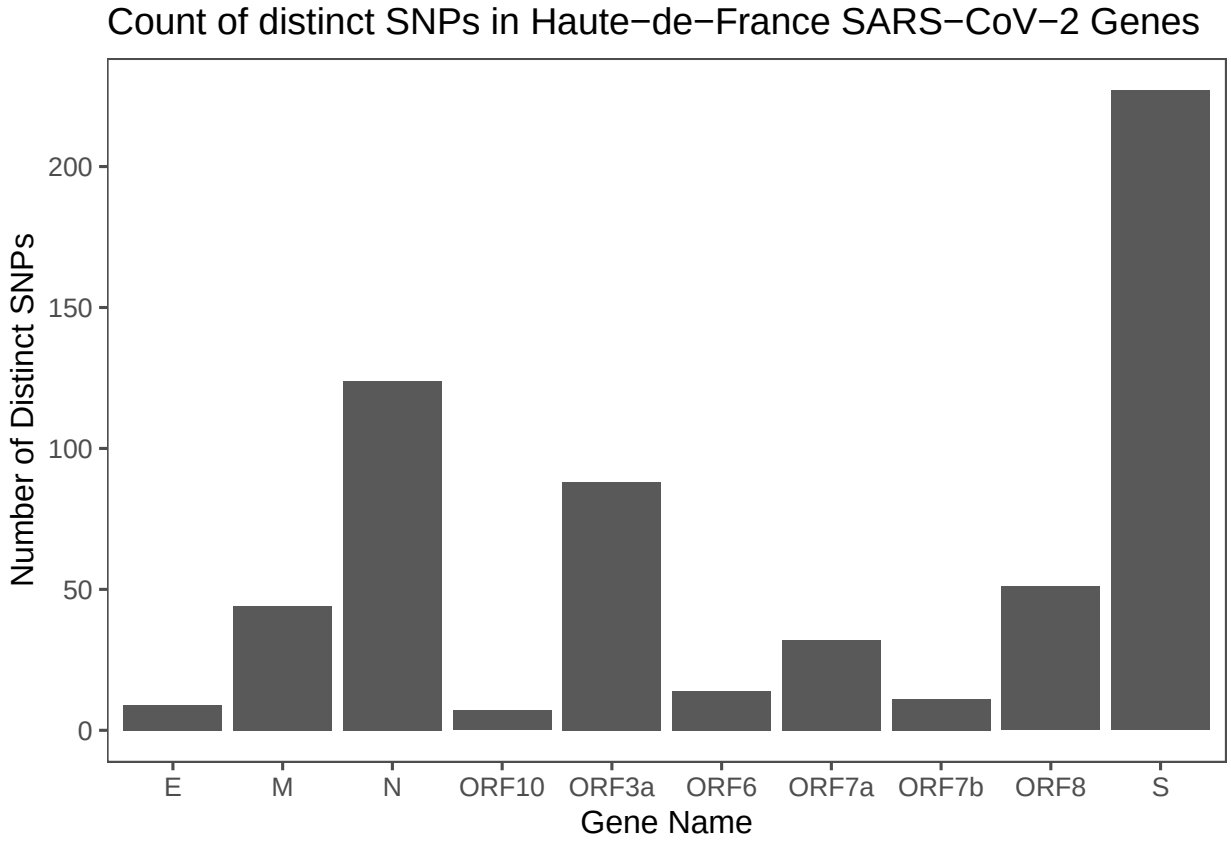

Figure 1: According to this figure which captures the unique SNPs locations within genes across all the sequenced samples, the N gene, S gene, and ORF3a gene contain the most SNPs location susceptible to mutation.

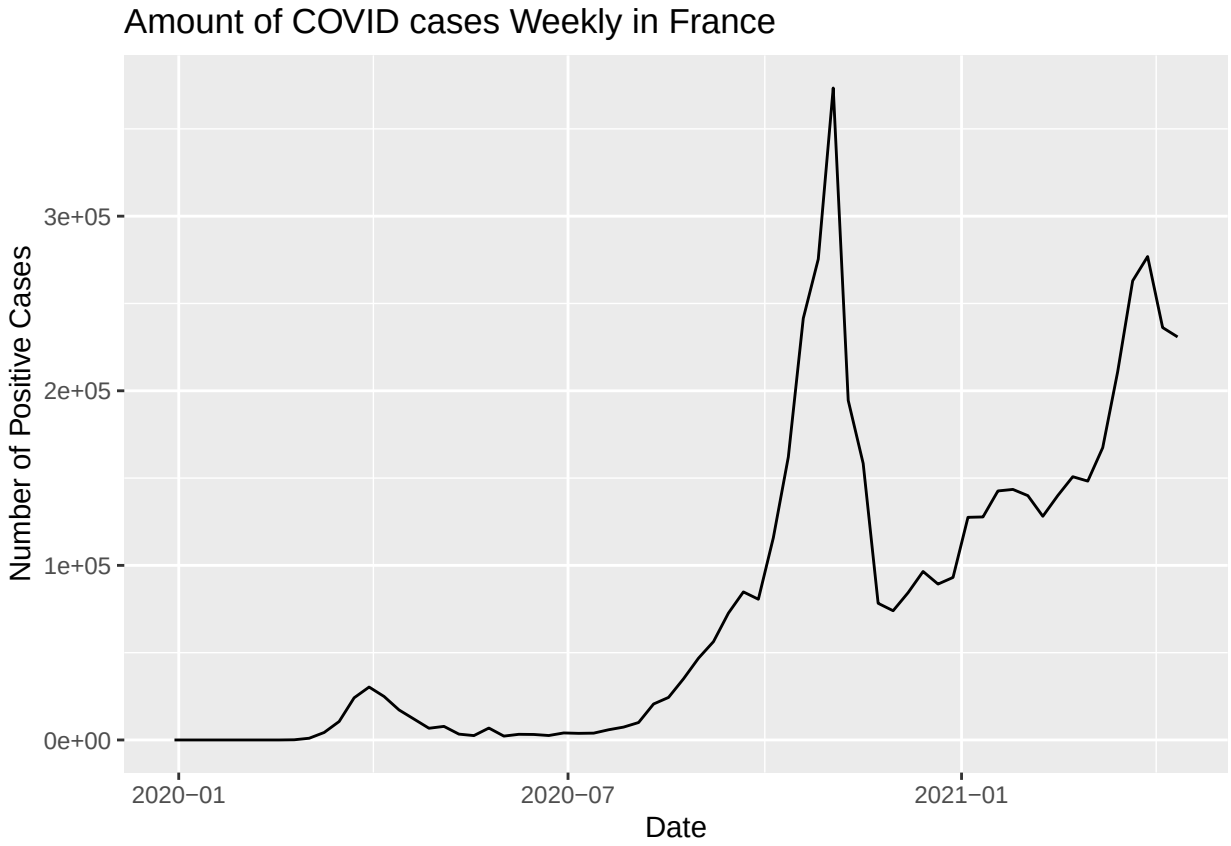

Figure 2: This figure records the amount of positive sampled COVID-19 cases in the whole country of France. The figure also includes statistics from the Northern region of France. Two main characteristics to notice here is that there is no huge spike in cases towards April 2020. On the other hand, there is a large number of cases between 2020-07 and 2021-01.

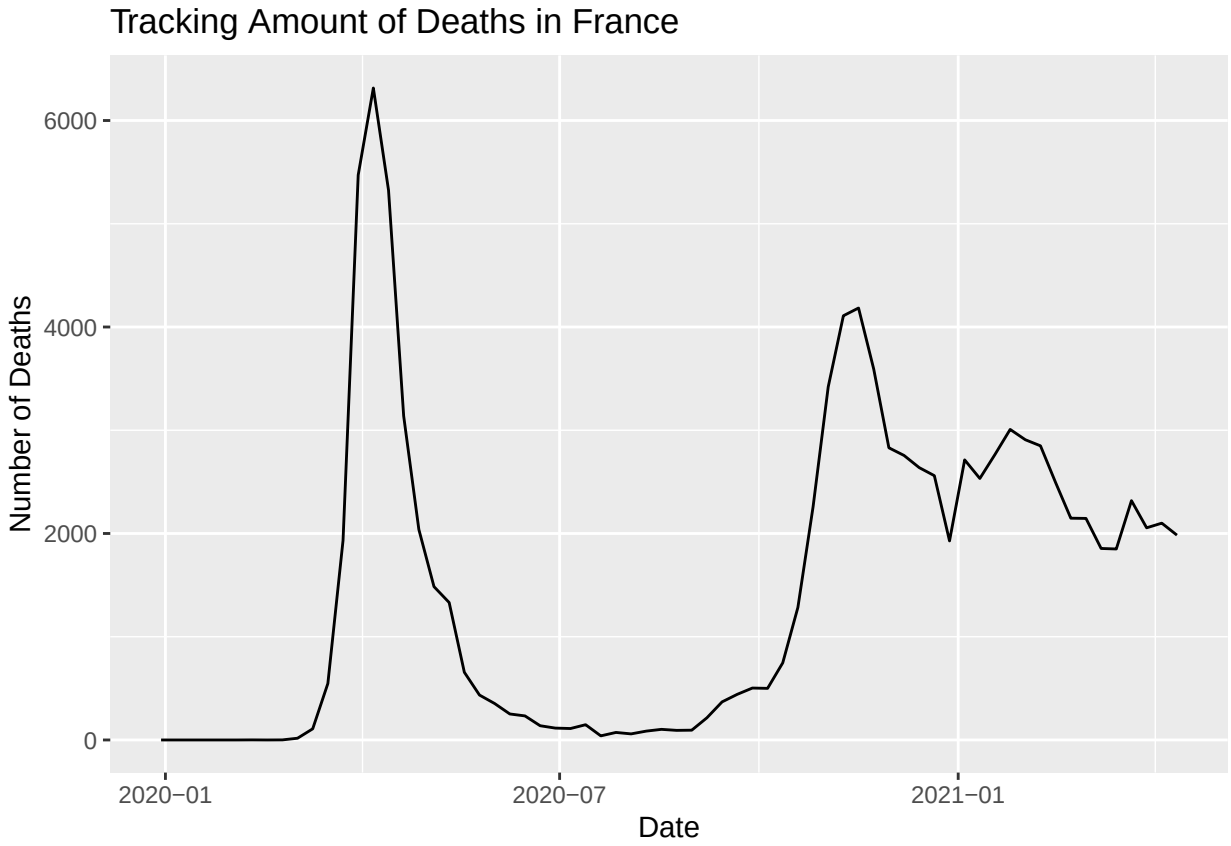

Figure 3: The figure tracks the amount of deaths that have occurred between 2020 and 2021 during the COVID-19 epidemic in France. As we can observe from this figure, there is a huge spike in deaths starting from March-April 2020 and then there is a sudden drop before exponentially increasing between July 2020 - January 2021.

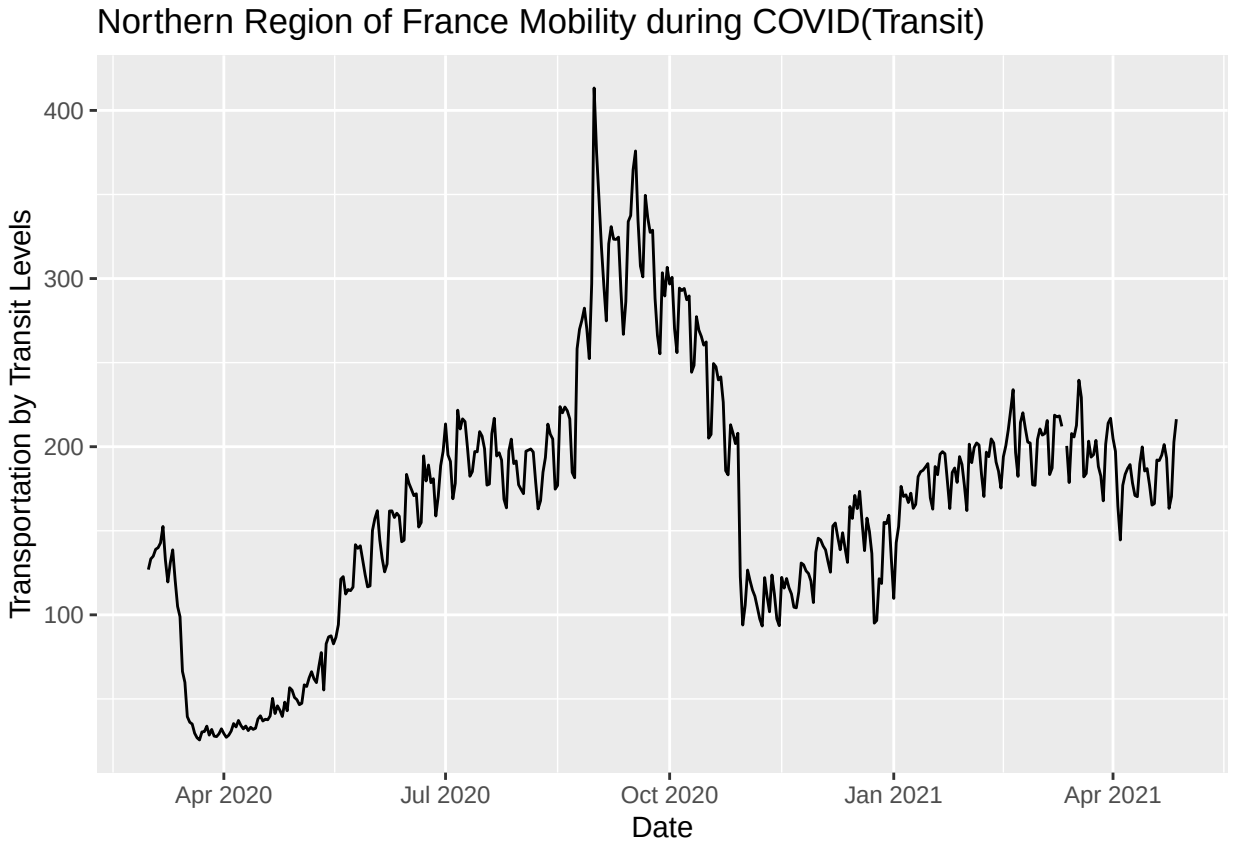

Figure 4: In this figure that diagrams transportation by transit mobility from the apple mobility dataset, there are two significant peaks that occur from July 2020 - October 2020. Compared to the visualization that illustrates the cases and numbers of deaths, the mobility by transit transportation correlates with the the trend shown in Figure 8 and Figure 9. However, this does not explain how the the amount of COVID-19 cases towards the beginning of this epidemic does not correlate to the amount of deaths that occurred. Further analysis is performed to examine this characteristic.

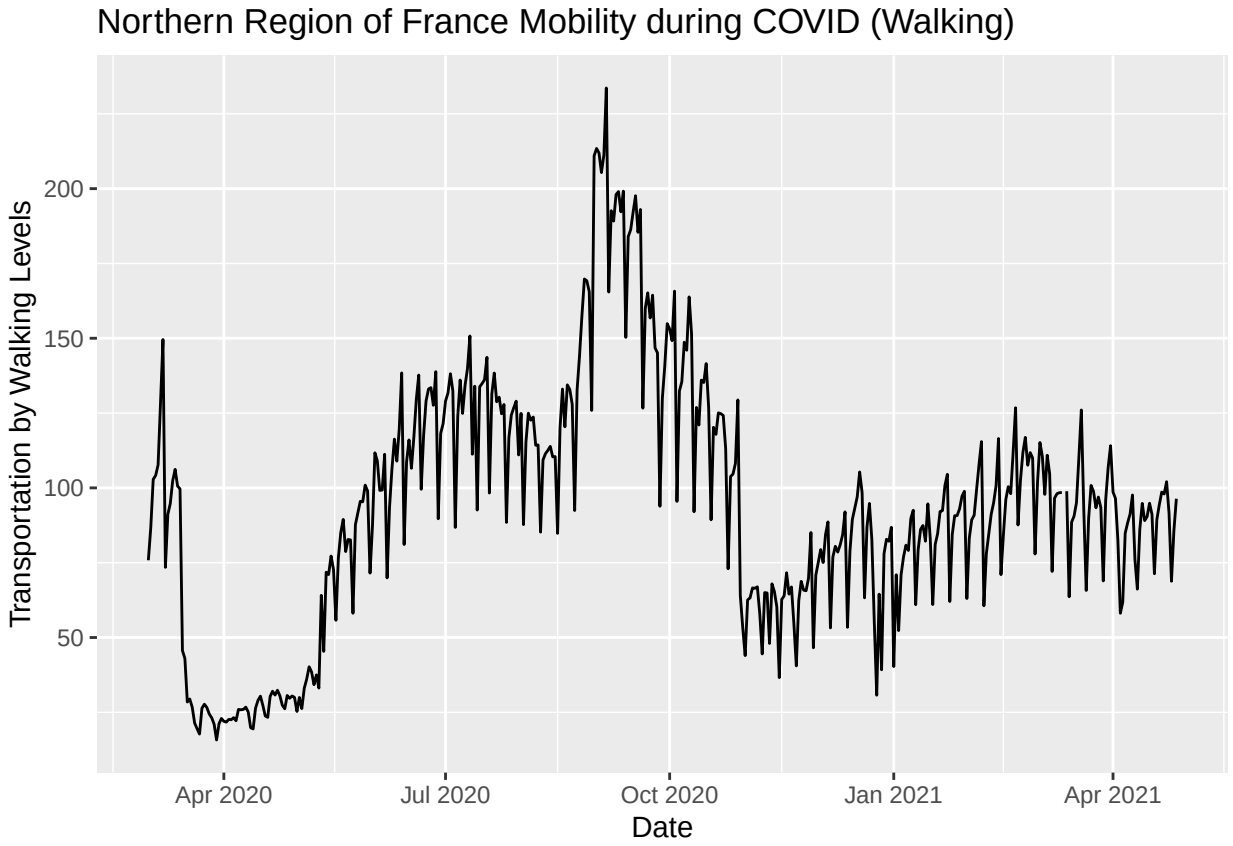

Figure 5: The figure illustrates transportation by walking mobility levels from the apple mobility data examining the Nord-Pas-de-Calais region of France. Similar to the previous figure, this diagram demonstrates a correlation between transportation by walking and the number of cases and deaths related to COVID-19. Unlike the previous figure, Figure 2, this visual shows us further indication of how there could have been a spike earlier on in the number of recorded deaths related to COVID-19. As we can observe from this figure, there are three prominent peaks that occur in March 2020, July 2020, and October 2020. However, this information still does not correlate with how low the numbers of confirmed Corona-virus cases there were in the beginning when the epidemic first began.

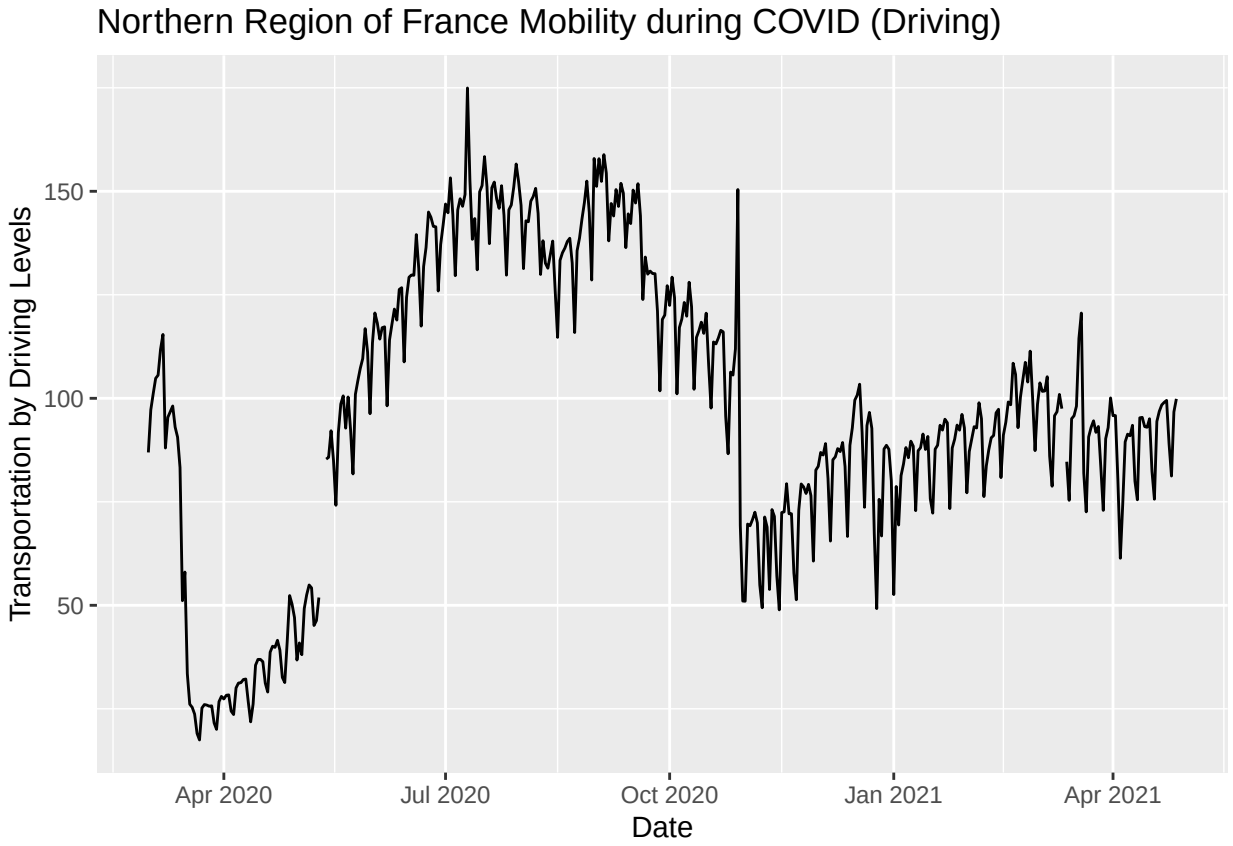

Figure 6: This figure demonstrates tranportation by dirivng mobility levels over the course of 2020 until 2021 during the Corona-virus epidemic in the Northern Region of France. From the visual illustrations of this figure, we can come to the conclusion that driving is much more preferable than taking the transit or walking. Additionally, given the patterns of the information presented in this diagram, we can associate this information with the numbers of confirmed cases and deaths related to the novel Corona-virus.

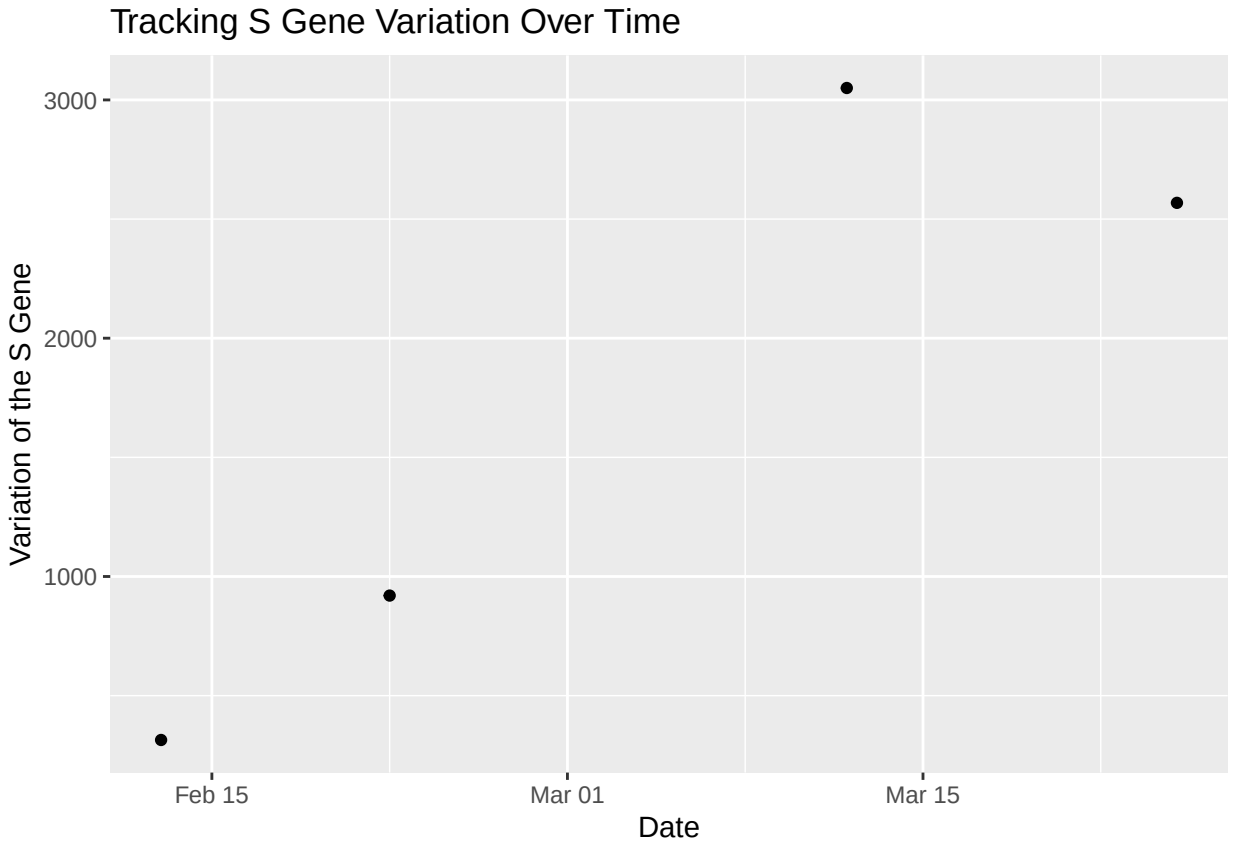

Figure 7: This figure demonstrates the amount of variation amongst the s-gene of the SARS-CoV-2 Haute-de-France positive sequenced samples based on the collection dates. As we can see over the period of the different collection dates the variation of S-gene increases. This indicates a susceptibility of possibly more mutations that can allow SARS-CoV-2 to become even more transmissible.

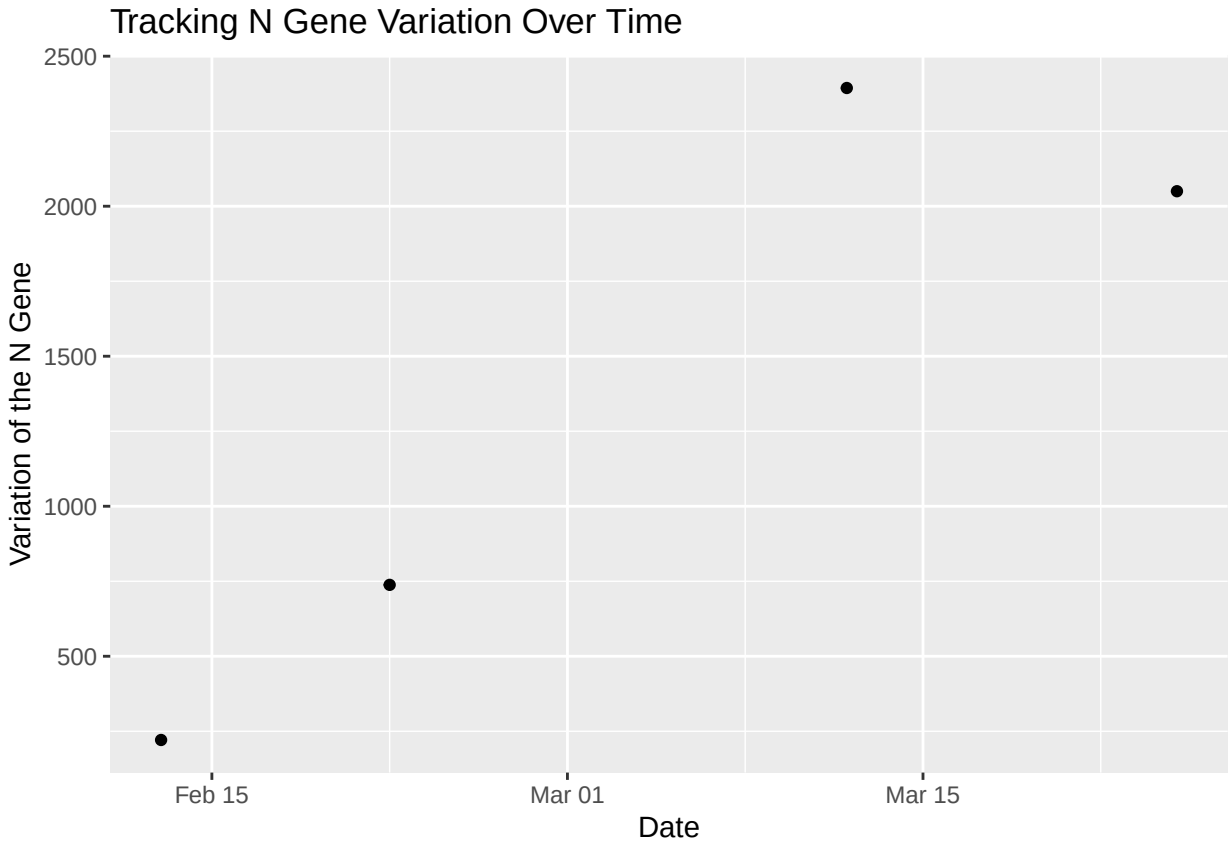

Figure 8: This diagram outlines the variation of the N-gene because it has the second to most unique locations of SNP which indicates that there are more locations where mutations could occur. After observing this figure, we can see an upward trend in variation of the N gene which is in agreement with how the cases of Corona-virus and deaths related to SARS-CoV-2 increases as time continues.

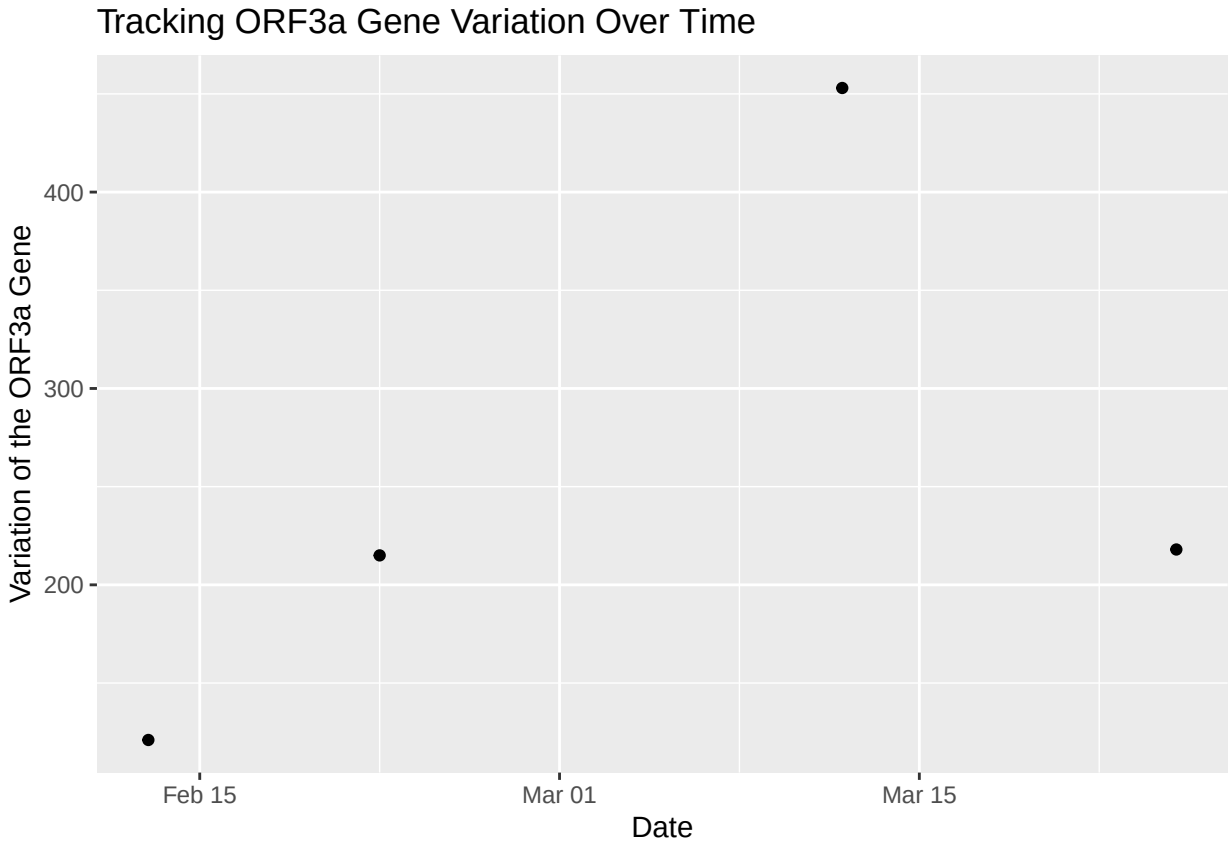

Figure 9: Looking at the variation of the ORF3a gene of SARS-CoV-2 Haute-de-France positive sequenced samples from the beginning of the pandemic until now. The upward trend over the course of the collection dates are consistent with the cases and deaths related to SARS-CoV-2.

Tables

Gene Name	Start	End	Length
S	21563	25384	3821
ORF3a	25393	26220	827
E	26245	26472	227
M	26523	27191	668
ORF6	27202	27387	185
ORF7a	27394	27759	365
ORF7b	27756	27887	131
ORF8	27894	28259	365
N	28274	29533	1259
ORF10	29558	29674	116

Table 1: Gene names, locations, and lengths in the SARS-CoV-2 genome. Higher SNP counts in the S and N genes may be related to the larger size of these genes.

Gene	Reference	Alternate	Count
S	C	A	1181
S	A	G	1086

Gene	Reference	Alternate	Count
S	C	T	1058
S	T	G	634
S	A	T	631
S	G	C	567
S	TTTATTA	TTTA	510
S	ATACATGT	AT	460
S	G	A	263
S	G	T	234

Table 2: This table documents the occurrence of certain nucleotide modifications within the S gene when compared to the reference sequence. The highest modifications of nucleotide alternations is C -> A and A -> G. These large amount of nucleotide modifications indicates there are more instances of mutations that could possibly occur.

Gene	Reference	Alternate	Count
N	G	A	1627
N	G	C	1464
N	C	T	1016
N	A	T	577
N	T	A	558
N	G	T	99
N	T	C	33
N	A	G	12
N	C	A	6
N	T	G	5

Table 3: Examines the nucleotide alternations that occur among the N gene when compared to the reference sequence. It it important to consider how many alternations occur to determine how many possible mutations could occur.

Gene	Reference	Alternate	Count
ORF3a	G	T	463
ORF3a	C	T	439
ORF3a	G	A	27
ORF3a	G	C	20
ORF3a	A	T	16
ORF3a	A	G	13
ORF3a	T	C	13
ORF3a	A	C	7
ORF3a	TGTTA	T	3
ORF3a	C	G	2

Table 4: The ORF3a gene is a viral nucleic acid sequence within the SARS-CoV-2 genetic information. This table measures the top occurrence of alternate nucleotides compared to the reference at the ORF3a gene. Similar to the previous the previous table, the larger the number of alternations, the higher chance there is for mutation.

Sources Cited

Allen,J. *et al.* (2020) France coronavirus map and case count. *The New York Times*.

Haim-Boukobza,S. *et al.* (2021) Detecting rapid spread of sars-cov-2 variants, france, january 26-february 16, 2021. *Emerging Infectious Diseases*, **27**, 1496.

Healy,K. (2020) Covdata: COVID-19 case and mortality time series.

Knaus,B.J. and Grünwald,N.J. (2016) VcfR: An r package to manipulate and visualize VCF format data. *BioRxiv*.

Pullano,G. *et al.* (2021) Underdetection of cases of covid-19 in france threatens epidemic control. *Nature*, 134.

Salje,H. *et al.* (2020) Estimating the burden of sars-cov-2 in france. *Science*, **369**, 208.

Wang,H. *et al.* (2020) The genetic sequence, origin, and diagnosis of sars-cov-2. *European Journal of Clinical Microbiology & Infectious Diseases*, **39**, 1629.

Wickham,H. (2016) Ggplot2: Elegant graphics for data analysis Springer-Verlag New York.

Wickham,H. *et al.* (2021) Dplyr: A grammar of data manipulation.