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

Vivian Lynn Ly

12 May, 2021

Background and Overview

This is a report on SARS-Cov-2, including some variant analysis (Koyama *et al.*, 2020).

This rmarkdown file will use the 'vcfR'(**???**), 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-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 and recorded deaths due to COVID 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?

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' (**???**) 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' (Wickham, 2016) 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.

Subsections are ok too

Results and Discussion

Figures

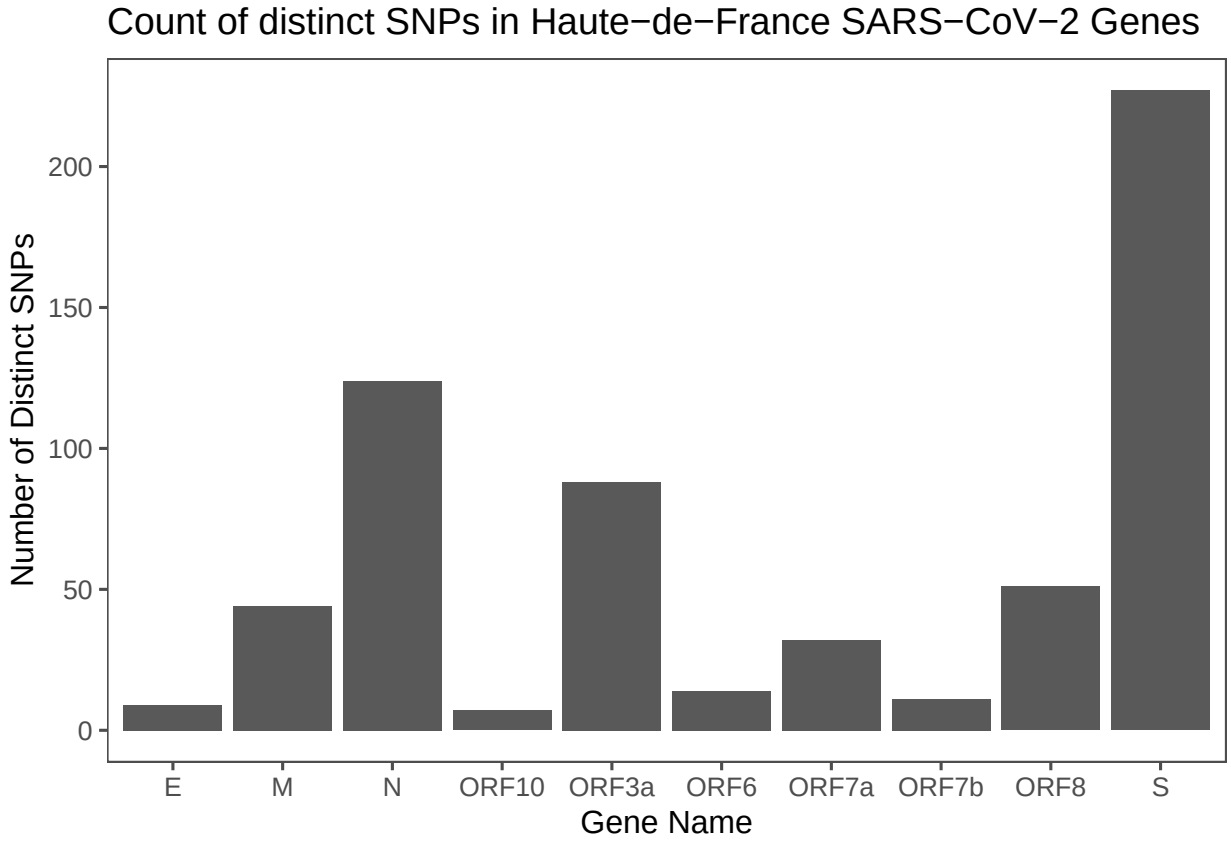

Figure 1: N and S genes have more unique SNPs in the set of samples analyzed.

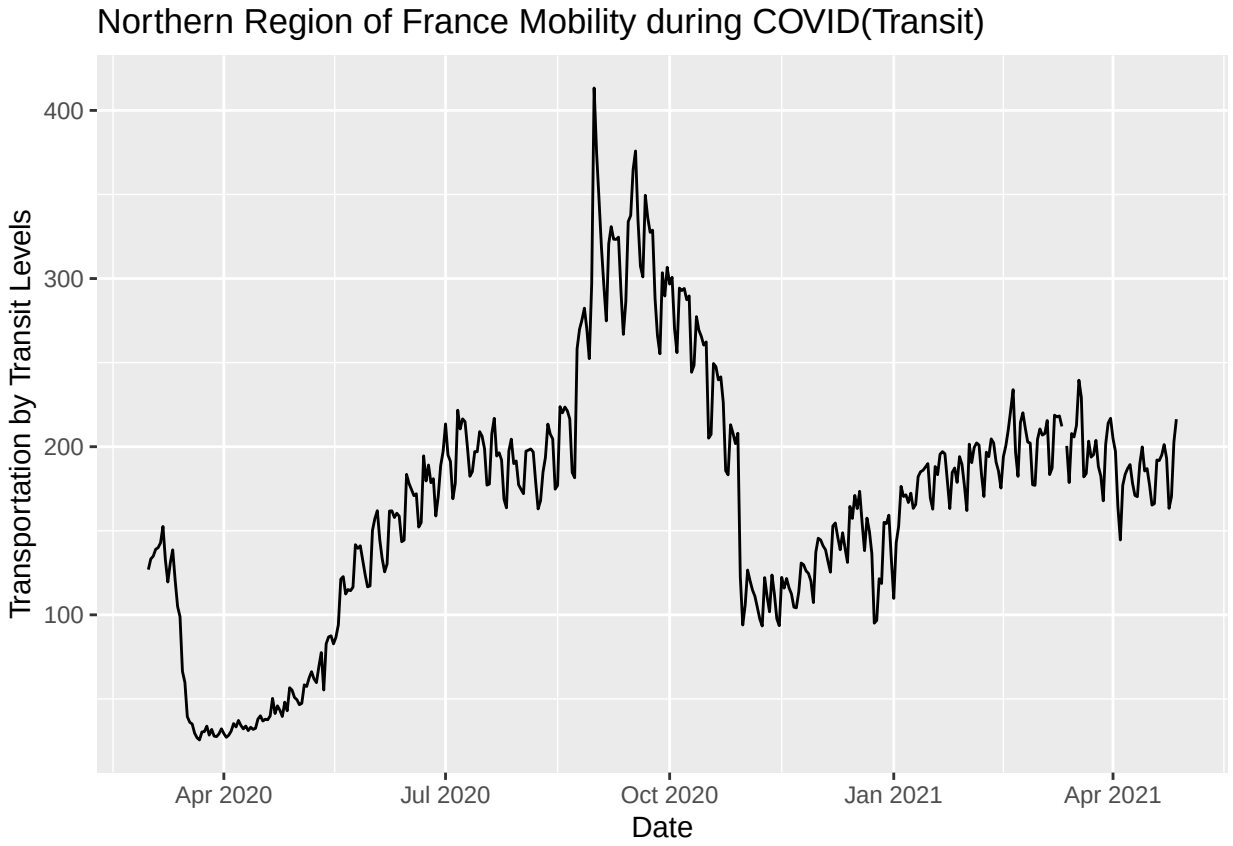

Figure 2: Using Covdata package to analyze transit mobility levels in the Northern Region of France during COVID through apple mobility data.

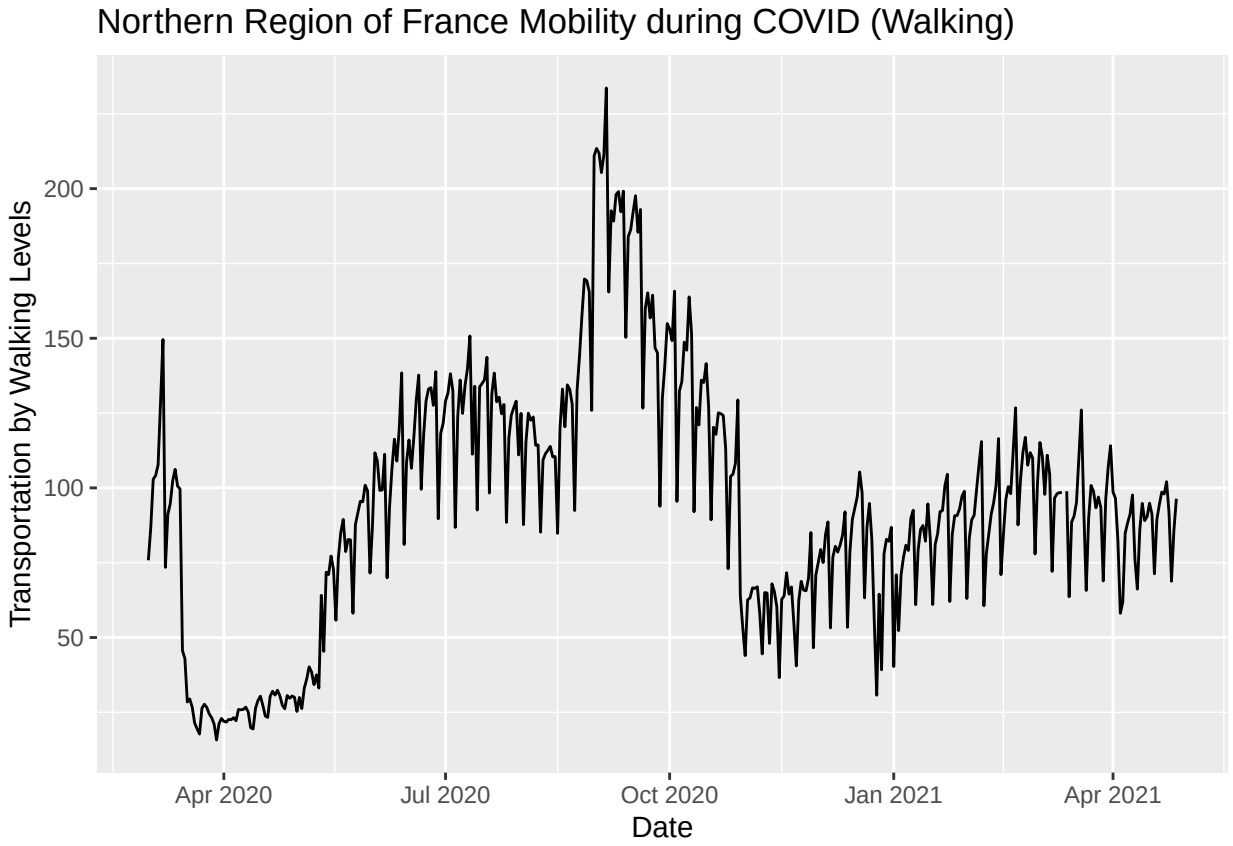

Figure 3: Using Covdata package to analyze walking mobility levels in the Northern Region of France during COVID through apple mobility data.

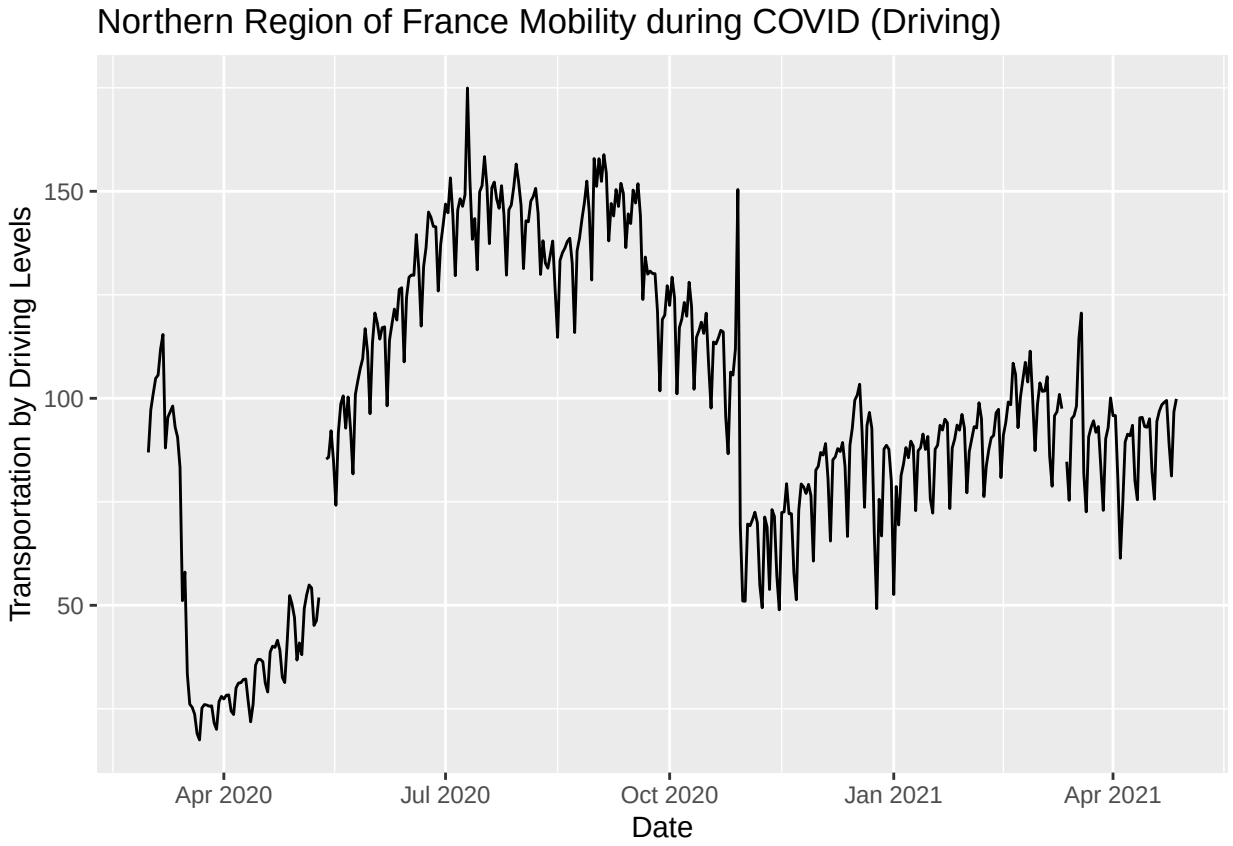

Figure 4: Using Covdata package to analyze driving mobility levels in the Northern Region of France during COVID through apple mobility data.

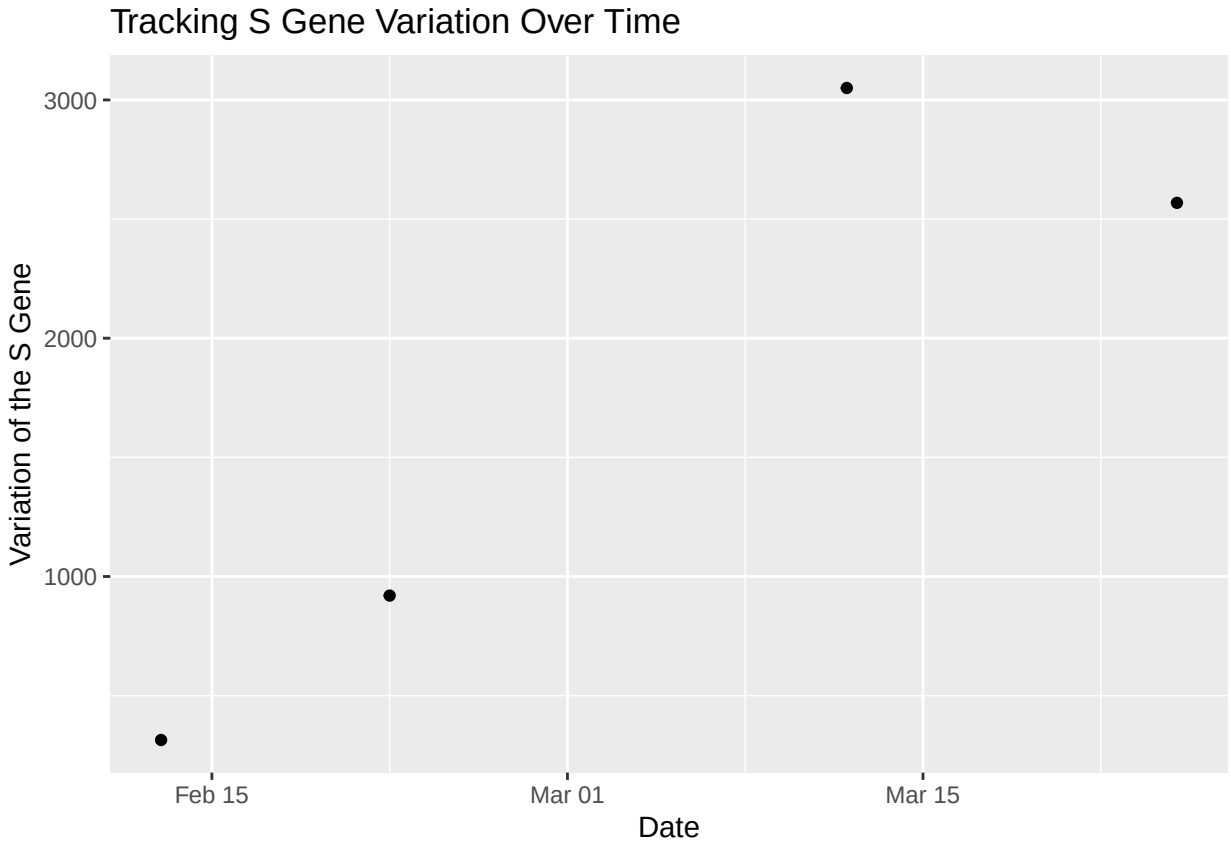

Figure 5: Looking at the variation of the S-gene of Sars-Cov-2 Haute-de-France positive sequenced samples from the beginning of the pandemic until now.

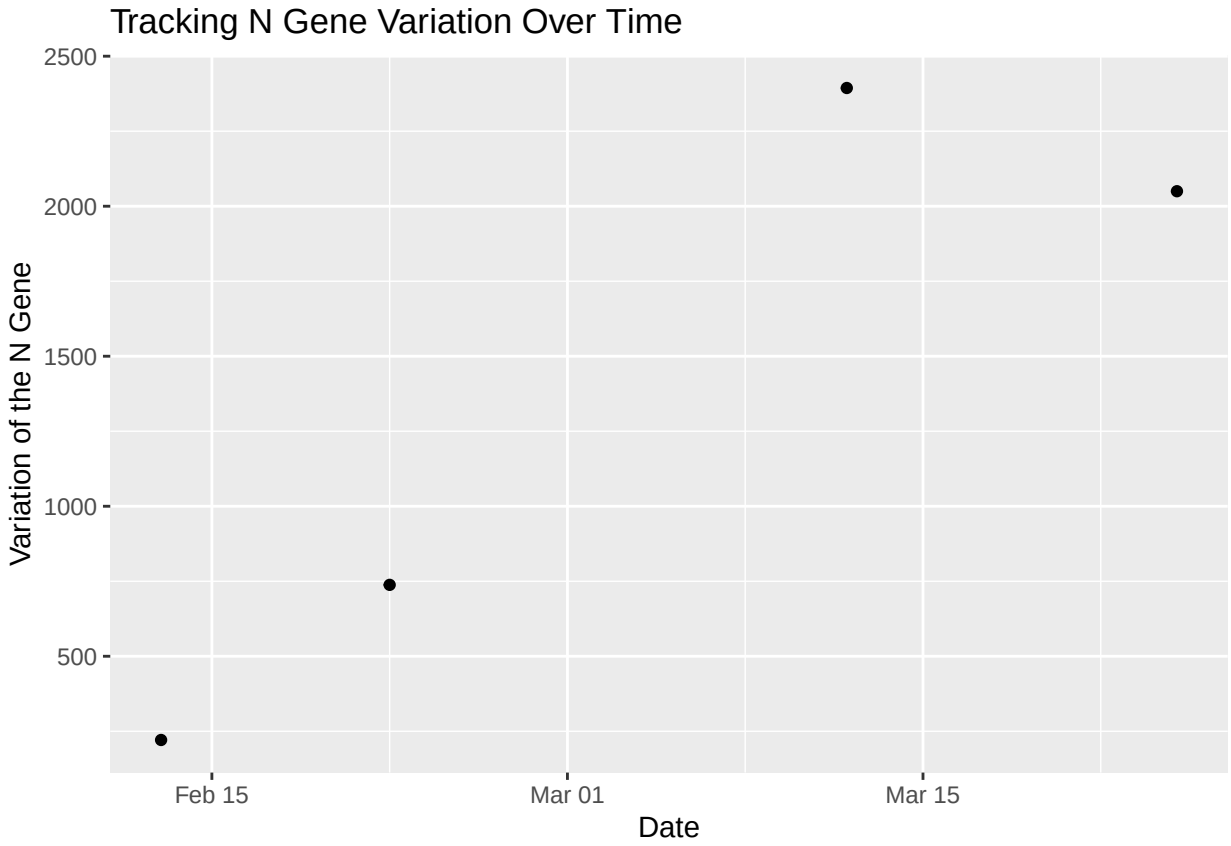

Figure 6: Looking at the variation of the N-gene of Sars-Cov-2 Haute-de-France positive sequenced samples from the beginning of the pandemic until now.

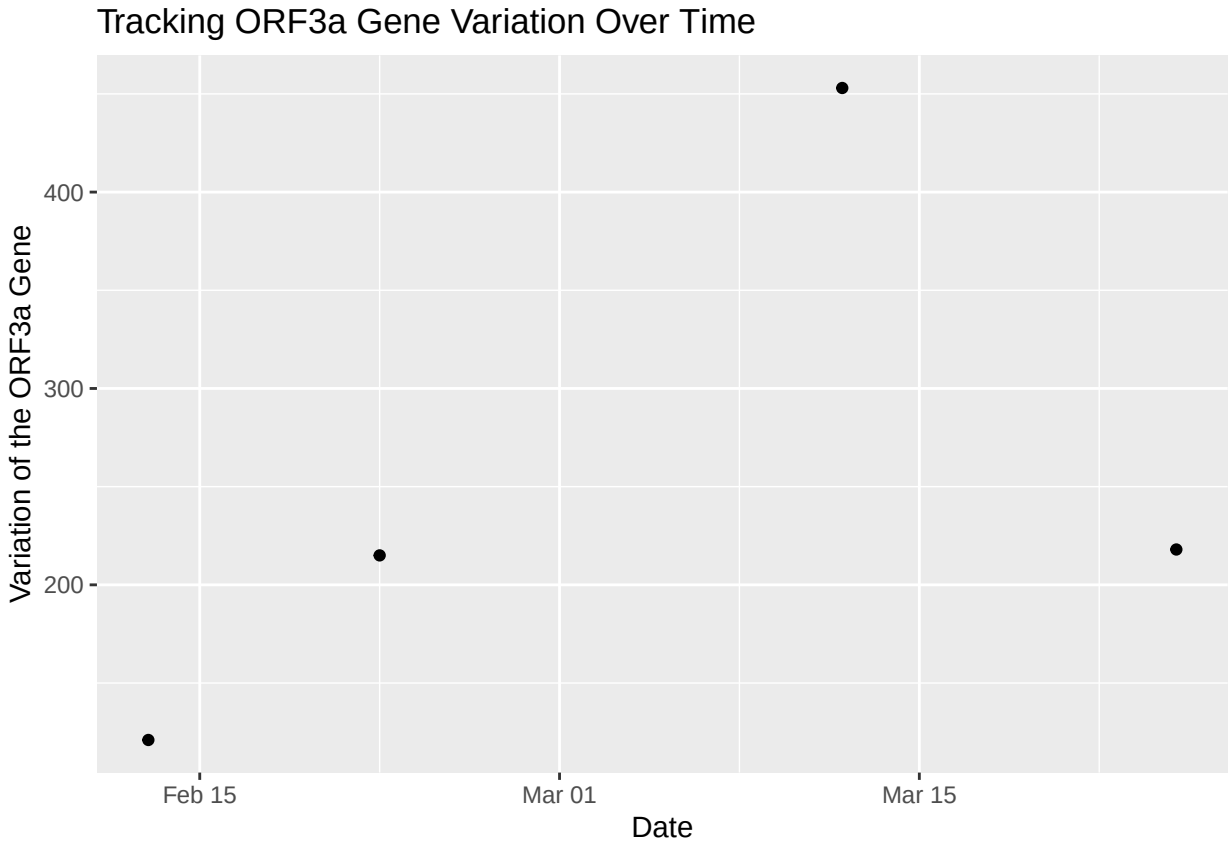

Figure 7: Looking at the variation of the ORF3a gene of Sars-Cov-2 Haute-de-France positive sequenced samples from the beginning of the pandemic until now.

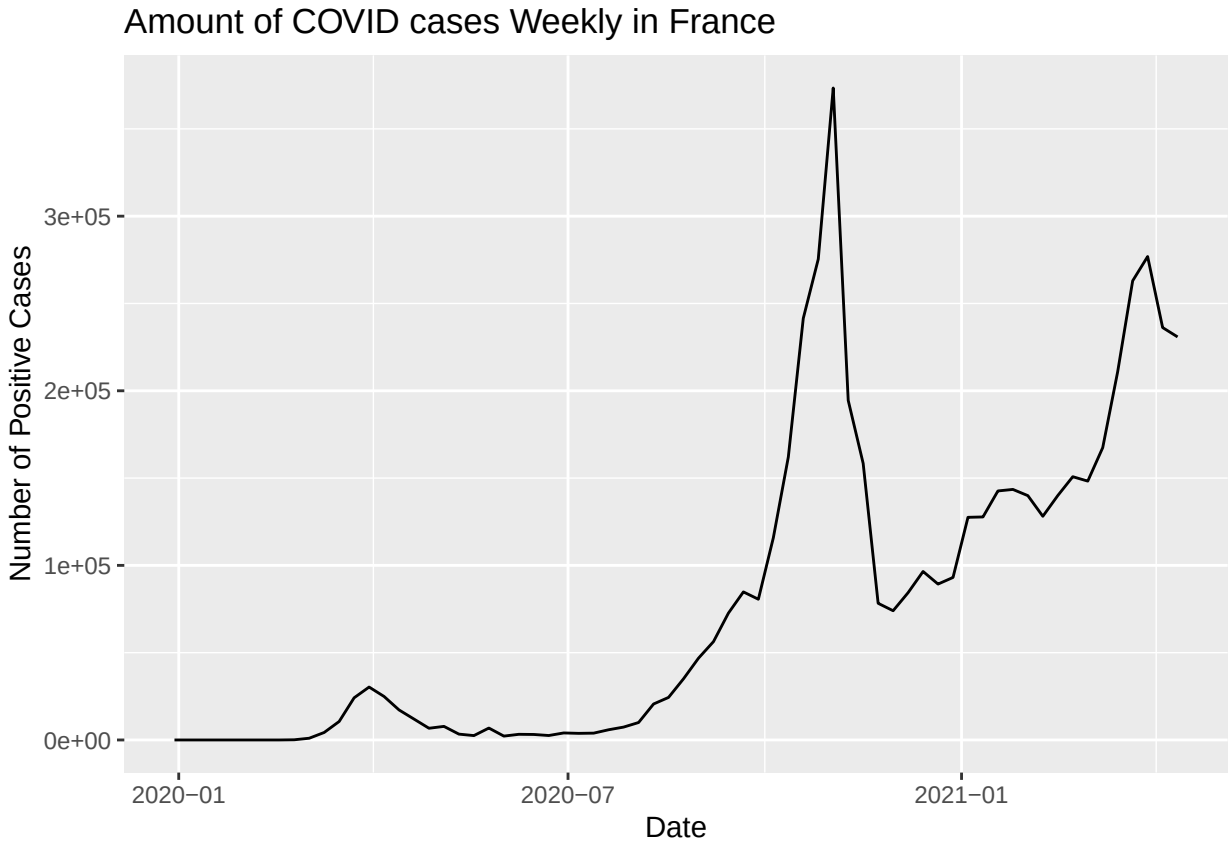

Figure 8: Tracking the number of COVID cases over time.

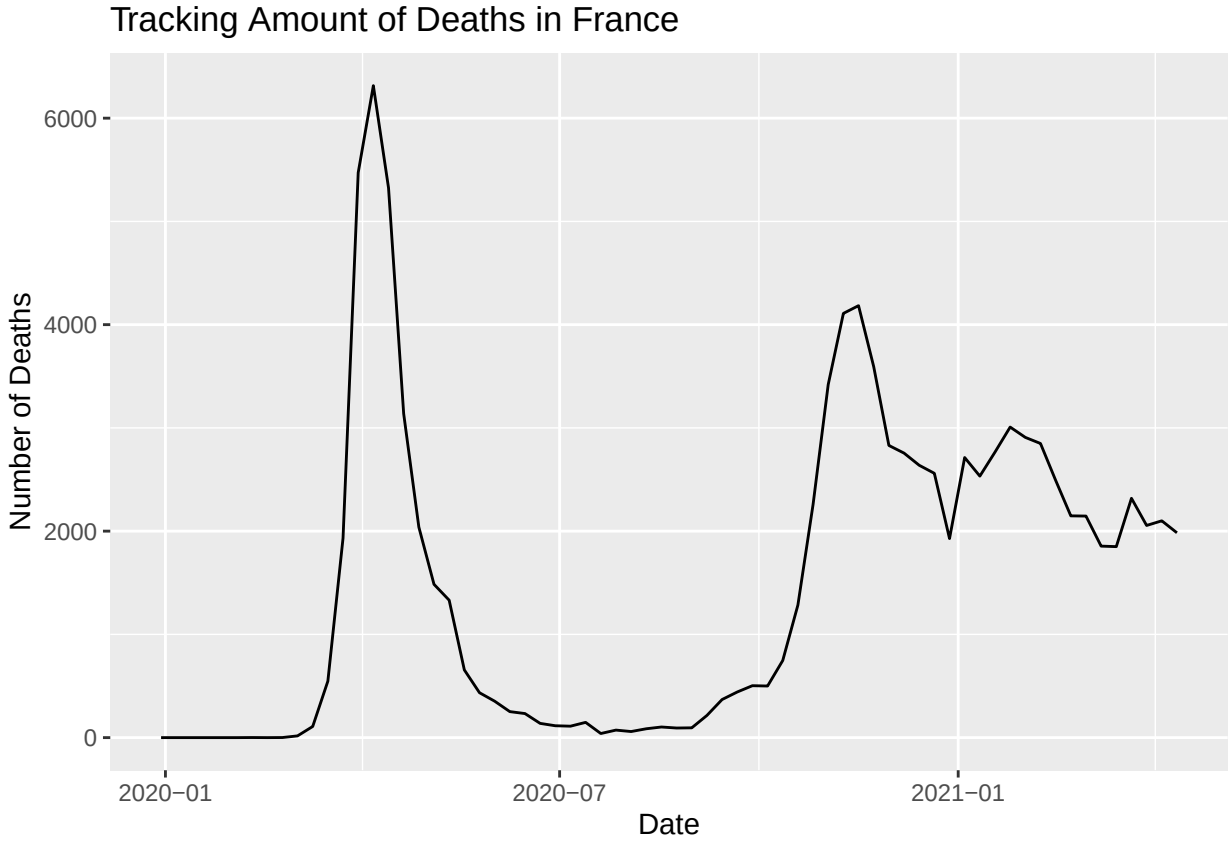

Figure 9: Tracking the number of deaths related to COVID in France.

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
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: Top Occurrence of alternate nucleotides compared to the reference at the S gene.

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: Top Occurrence of alternate nucleotides compared to the reference at the N gene.

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 5: Top Occurrence of alternate nucleotides compared to the reference at the ORF3a gene.

Sources Cited

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

Koyama,T. *et al.* (2020) Variant analysis of sars-cov-2 genomes. *Bulletin of the World Health Organization*, **98**, 495.

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

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