Analysis of SARS-CoV-2 Infection in New York City in the Early Stages of the COVID-19 Pandemic

Maggie Walsh

10 May, 2021

Background and Overview

In early 2020, the world was forever changed by the spread of the coronavirus disease (COVID-19), which was declared a global pandemic by the World Health Organization (WHO) in early March. New York City quickly emerged as the epicenter in the United States with 203,000 confirmed cases between the months of March and May (Thompson *et al.*, 2020). Although news of the virus was widely overlooked initially, people quickly became aware of the insidious nature of the SARS-CoV-2 virus. The SARS-CoV-2 virus was hard to track due to it's variable incubation time and it's ability to spread via asymptomatic carriers. The only way to fight this virus early on was through limiting person to person interaction. New York City was placed under a statewide lock-down on March 22, 2020. The number of confirmed cases peaked a week later (Thompson *et al.*, 2020). As the machinery of our society was grinding to a halt, researchers raced to understand the SARS-CoV-2 virus in order to prevent further devastation. My goal for this report was to synthesize some of that research, and generate my own analysis that explains what was going on in New York City during the onset of the pandemic. My analysis started with a look into SARS-CoV-2 case numbers in New York, followed by a look at changes in population mobility during this time. I also used a set of bioinformatics tools to analyze sequence data taken from nasopharyngeal swabs of people in New York City from early March to late April. Twelve samples were sequenced and variants were called against a reference genome to determine SNPs for analysis. The common consensus among researchers is that the rise in case numbers in New York City was the result of uncontrolled community spread after many different travelers (mainly from Europe) introduced the virus to the population (Bushman *et al.*, 2020; Gonzalez-Reiche *et al.*, 2020; Maurano *et al.*, 2020). My findings corroborate this story. My goal for this report was to demonstrate the importance of our public health infrastructure, and show how bioinformatics tools can be used to investigate phenomena and generate a compelling narrative that contributes to our understanding of complex and multifaceted events such as the SARS-CoV-2 pandemic.

Methods

Sample Sequencing and Variant Calling

I downloaded sequences to a remote server from the NCBI database using the accession ID, utilizing fasterq-dump from the NCBI SRA toolkit. Samples were trimmed using the trimmomatic tool (Bolger *et al.*, 2014). My work flow utilized the Burrows Wheeler Aligner (BWA) to map reads against the reference genome (Li and Durbin, 2009) and SAMtools and BCFtools to sort and process the mapped reads (Li *et al.*, 2009). Specific steps of my pipeline can be found in the code folder of this repository. Parts of the pipeline approach are based on the pipeline described in the Data Carpentry Genomics lessons, which are made available under a CC-BY 4.0 license. I executed the bash pipeline on the USF server using a makefile to output vcf files for analysis.

Analysis

Reading in vcf files

I analyzed the output from the bash pipeline in RStudio using a series of functions designed to format the vcf files into tabular data. The vcf files were read into R using the vcfR package (Knaus and Grünwald, 2017).

Analysis of tabular data

I manipulated tabular data using the dplyr package. I made figures using the package ggplot2 with additional formatting using the packages ggthemes and gghighlight.

Additional data sources

I accessed additional data sources using the COVID19 R package (Guidotti and Ardia, 2020). Case numbers originated from the Oxford COVID-19 GOVERNMENT RESPONSE TRACKER data set (Hale *et al.*, 2020). Mobility data was from the Google Mobility Reports. I also used SARS-CoV-2 clade definitions from the NextClade project.

Results

Case Numbers

I started my assessment of case numbers by looking through a wider lens. I wanted to look at New York State as a whole, and compare its confirmed case numbers to the rest of the country. As seen in Figure 1, the state of New York was far beyond other states in terms of positive case numbers early on in the pandemic. However, by the Summer of 2020 it seemed that the exponential growth in cases was stalled, and New York was no longer the state with the most confirmed cases by August. However, case numbers in New York did skyrocket again in the Winter months. To get an exact look at New York City, I narrowed my scope (Figure 2). It appears that the trends in case numbers in New York City followed a very similar pattern as seen in the state as a whole. Next, I narrowed in on my period of interest, the time frame in which my samples were collected (Figure 3). Between March 1st and April 10th, case numbers increased at alarming rates, going from almost no confirmed cases to tens of thousands of cases in a matter of weeks. This sharp incline was seen even after officials ordered a statewide lock-down.

Mobility Reports

I wanted to use mobility data to gauge how New York City residents were adapting during this time. I chose to visualize transit mobility because I know that a large proportion of the population uses public transit to get around (Figure 4). Transit mobility experienced a sharp decline between March and April of 2020, diving over 80% below baseline levels by mid April. A similar trend can be seen in almost all of the mobility categories (Table 1). Interestingly, there is a clear inverse relationship between case numbers and citizen mobility.

Variant Analysis

For my variant analysis, I compiled a list of distinct SNPs for each named gene in the SARS-CoV-2 genome (Table 2). I found a total of 7 different distinct SNPs in protein-coding regions. Three of those were found

in the nucleocapsid (N) gene, while one distinct SNP was found in each of the remaining genes, S, M, ORF3a, and ORF10. To determine the significance of the variants found in each of my samples, I generated a matrix showing similarity between each sample and each of the known SARS-CoV-2 clades (Table 3). I used this information to generate a heat map which visualizes these similarities (Figure 5). The heat map singles out two samples (SRR14232249 and SRR14232250) as genetically distinct from the rest, with more similarity to the 20B and 20I/501Y.V1 clades. The remaining samples showed more similarity to the 20C and 20H/501Y.V2 clades. It is important to note that clade 20I/501Y.V1 is a descendant of clade 20B, and clade 20H/501Y.V2 is a descendant of clade 20C (Bedford *et al.*, 2021). This means that samples that are similar to 20B or 20C will also show some similarity to their descendants.

Discussion

Contextualizing Case Data

Results from my case data analysis suggests that the number of positive SARS-CoV-2 infections grew exponentially in New York between the months of March and April. However it is important to note that these case numbers represent the number of *confirmed* cases, not the total number of cases. It is possible that the rapid incline in positive case numbers (seen in figures 1-3) was more to do with testing capacity than an increased prevalence of the virus. Multiple research studies have indicated that the SARS-CoV-2 virus was circulating in New York and other US cities long before public health officials originally thought (Bushman *et al.*, 2020; Gonzalez-Reiche *et al.*, 2020). Because of the stealthy nature of SARS-CoV-2, it is likely that asymptomatic or mildly ill individuals were able to infect others without much alarm (Oran and Topol, 2020). Furthermore, the incubation period of SARS-CoV-2 has been reported as variable, with an average of about 5 days until the onset of symptoms, and it may take even longer for someone to begin to experience obvious illness due to Covid-19 (Wu *et al.*, 2020). Within that period, individuals would still go about their normal lives, unaware that they were infected and shedding the virus. This explains why early testing is a crucial part of containment, especially in asymptomatic carriers.

Unfortunately, when it came to early testing in the United States, severe bottlenecks through the CDC and FDA limited local public health authorities ability to quickly and effectively detect and trace the virus in their communities (Cohen, 2020). This means that the SARS-CoV-2 virus was able to operate under the radar and spread widely and rapidly, inevitably reaching travel hubs like New York. The virus, once introduced, was in all likelihood able to easily spread via community infection at least in part due to the heightened population density associated with urban environments (Rocklöv and Sjödin, 2020).

The delay in the availability of tests eliminated the possibility for early detection, which allowed the virus to spread uncontrollably right under our noses. This is why it is extremely important to be skeptical of early case reports, and understand that initial reports grossly underestimated the prevalence of SARS-CoV-2 in New York and the rest of the US in the first few months of 2020. These underlying factors help to explain how SARS-CoV-2 was spreading early on, and why confirmed case numbers soared in New York in March, even after physical distancing initiatives were put in place.

Using Variant Analysis to Explain Outbreak Origins

Although the explanation above describes how SARS-CoV-2 spread within New York City, it does not explain how the virus reached the city. Variant analysis can be an efficient way of tracking the origins of particular viral disease outbreaks. I used variants detected in my samples to identify SARS-CoV-2 genomes genetically related to two distinct clade lineages, the 20B lineage and the 20C lineage (Figure 5, Table 3). Both clades were circulating in Europe during the period of interest (Hadfield, 2020). This suggests a link between the strains of SARS-CoV-2 in Europe and those known to have caused some of the early infections in New York City. I acknowledge that my sequence analysis had some limitations. I used a small sample size (12) that came from the same collection site. Therefore, it is likely that my samples contain much less genetic diversity

than what was present in all of New York City during that time. However, my findings do echo the results of other phylogenetic analyses that also tie the majority of introduction events in New York city to travel from Europe (Bushman *et al.*, 2020; Gonzalez-Reiche *et al.*, 2020; Maurano *et al.*, 2020).

Conclusion

What do these results tell us about the outbreak of SARS-CoV-2 in New York City? I believe that there is ample evidence to suggest that the outbreak that occurred in New York in March of 2020 was a direct result of an insufficient government response to various warning signs. By December of 2019, news of the emerging SARS-CoV-2 virus coming out of China's Wuhan Province was beginning to circulate. By late January, the WHO was already ringing alarm bells about the potential for the SARS-CoV-2 virus to spread globally, declaring the issue a "public health emergency of international concern" (Organization and others, 2020). The WHO urged countries to prepare for the possibility of an outbreak, and emphasized that the emerging virus could be contained through the effective early detection and isolation of cases, contact tracing, as well as through the implementation of social distancing measures. The United States did not seem to heed these warnings. In early February the Trump administration responded by restricting travel to and from China. However, this intervention did not stop the SARS-CoV-2 virus from entering the US via Europe and quickly spreading, while the scarcity of tests made it impossible to track and contain. New York was especially vulnerable to this outbreak because of its status as a travel hub, as well as being densely populated. Additionally, reluctance to act decisively also occurred at the municipal level. Mayor Bill de Blasio showed a disregard for the severity of the virus early on, and did not implement any public safety measures until late March, when community transmission was already widespread. This analysis demonstrates how a poor public health infrastructure and lack of decisive government action contributed to the early spread of SARS-CoV-2 in the United States, especially in New York City. In the future, federal, state, and local governments must invest in public health resources and prepare to act quickly and decisively in the event of another serious public health crisis.

Figures

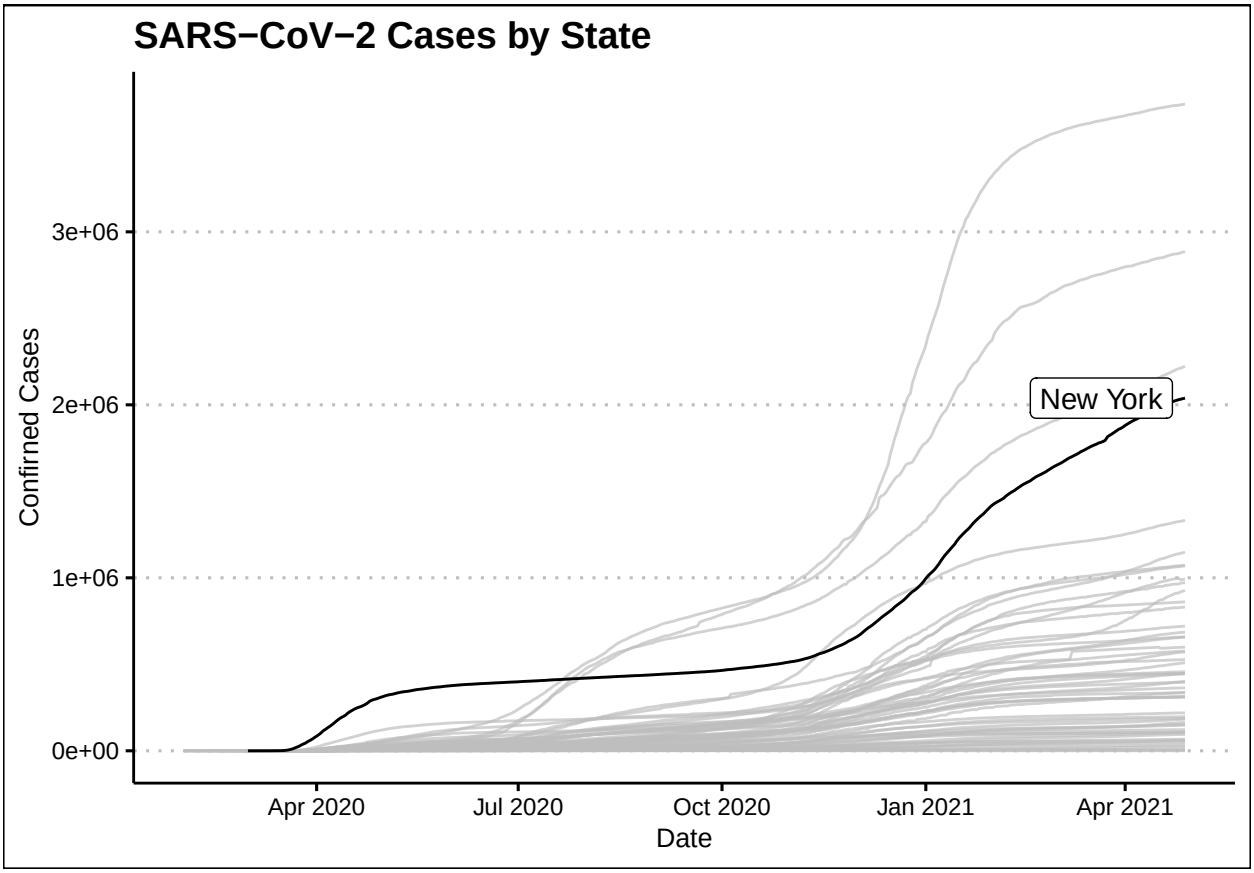


Figure 1: Plot of confirmed SARS-CoV-2 cases in the past year. New York state was the hardest hit early on in the pandemic having a majority of the nations confirmed cases. The steep rise in cases was contained in subsequent months. In recent winter months, New York, alongside other states, saw another steep rise in cases.

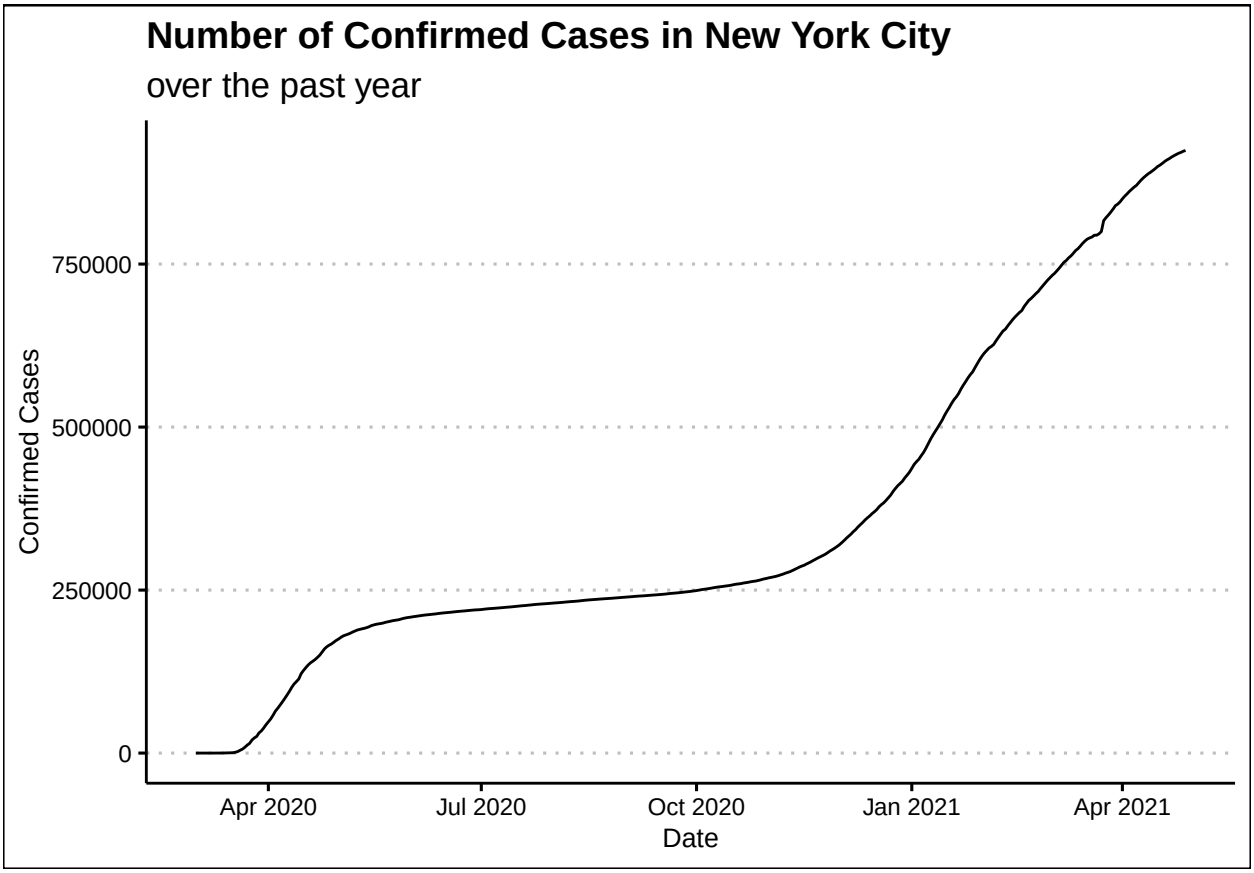


Figure 2 Looking specifically at New York City, we can see a similar trend as the state case numbers shown above. New York City was responsible for a large portion of the nations case numbers during the first wave of the pandemic.

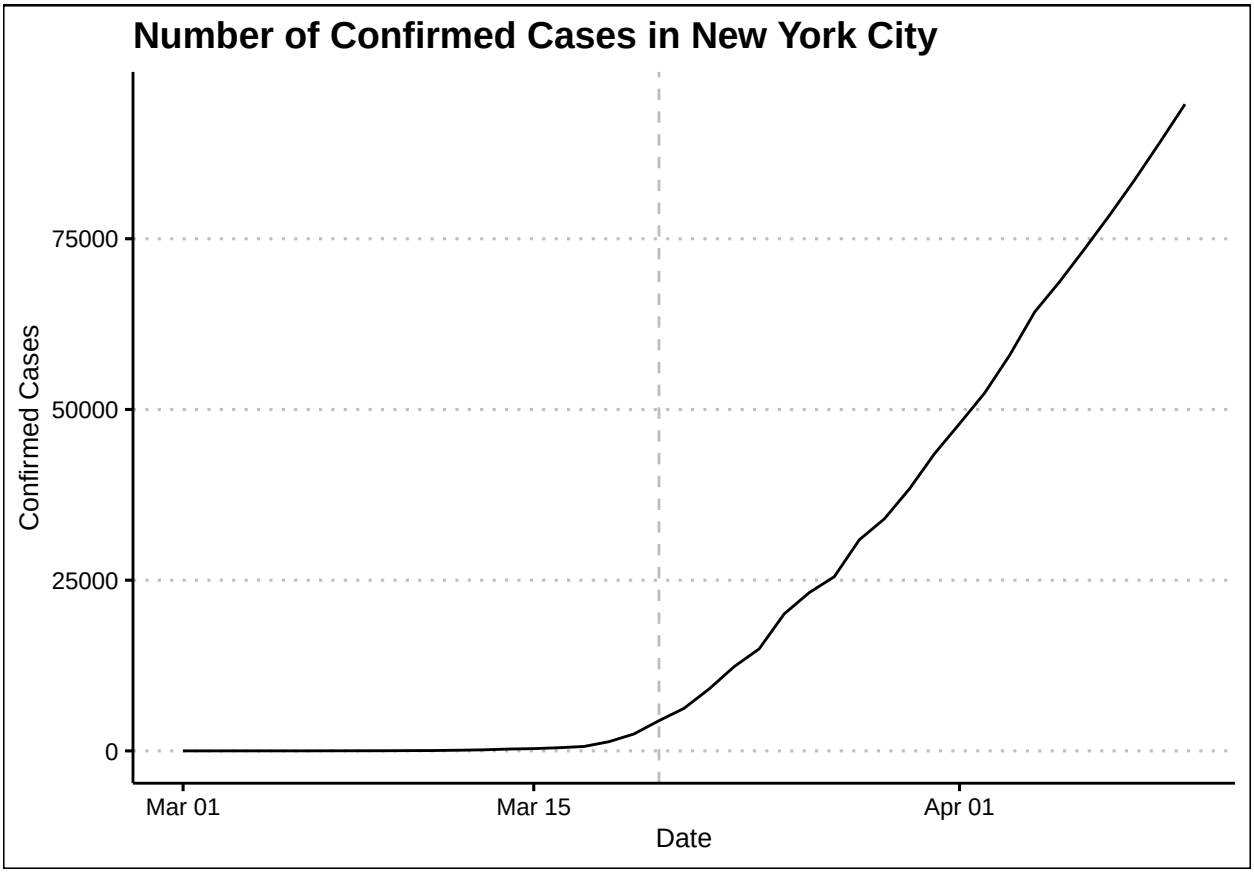


Figure 3 Going for a zoomed-in look, we can see how rapidly new cases of SARS-CoV-2 infection were being recorded during a short few-week period in New York City. This steep incline likely reflects a greater rate of testing. The dashed line indicates when New York State issued a statewide mandatory stay at home order.

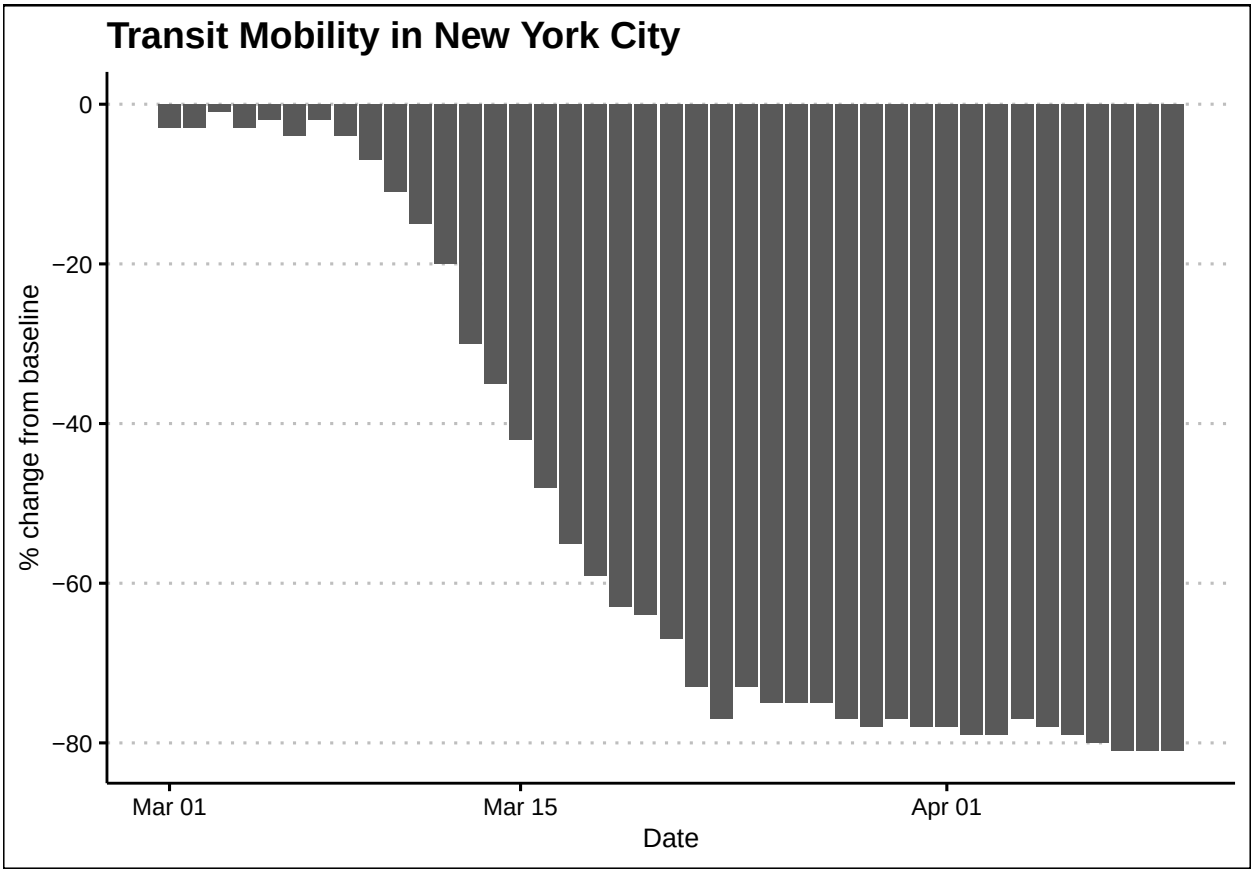
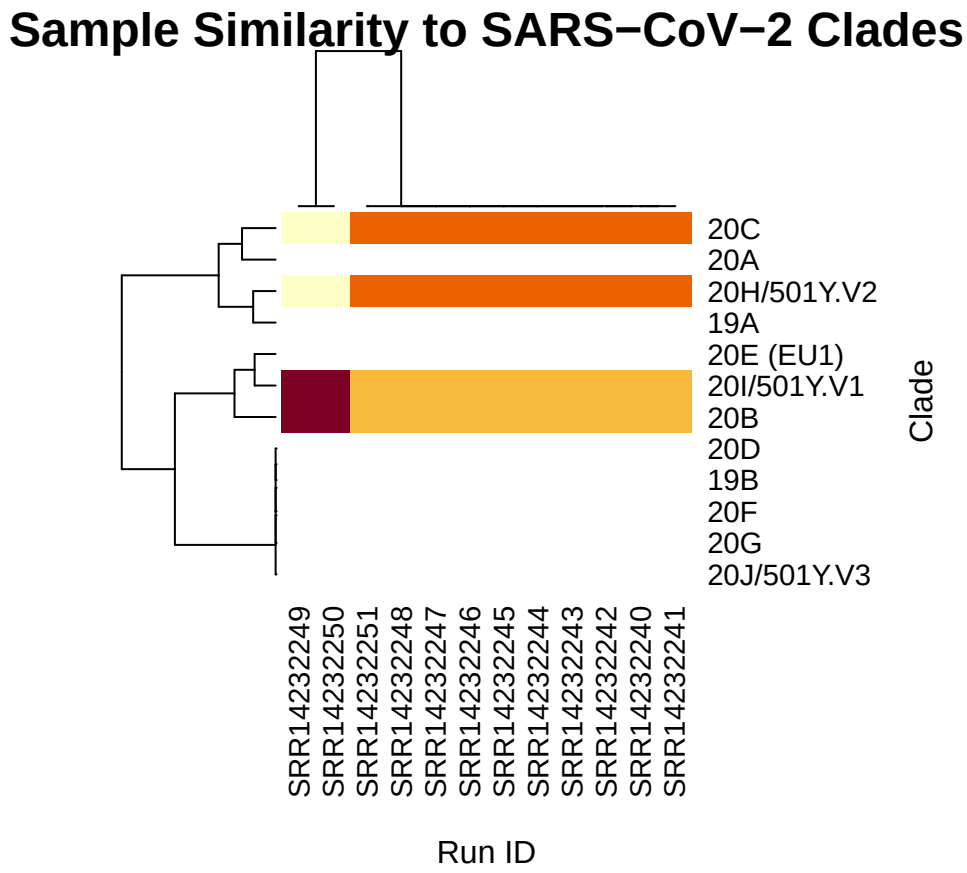


Figure 4 Bar plot showing percent change in mobility at transit stations in New York, from a baseline set before the pandemic. This metric contradicts the steep rise in cases during this time, unless you account for the variable incubation time of the SARS-CoV-2 virus, and the probability of higher case rates in early 2020 than shown above due to insufficient testing. For more mobility information, see Table 1.




```
## $rowInd
##  [1] 12  9  8  2  6  4 11  7  1 10  3  5
##
## $colInd
##  [1] 10 11 12  9  8  7  6  5  4  3  1  2
##
## $Rowv
## NULL
##
## $Colv
## NULL
```

Figure 5 Heat map showing relationships between clades and sample genotypes. Darker colors represent a greater similarity. Hierarchical clustering points to two genetically distinct groups, with differing clade distinctions. For exact values, see Table 3.

Tables

Week	Reail and Recreation	Grocery and Pharmacy	Parks	Transit Stations	Workplaces	Residential
9	0.000000	4.666667	5.00000	-2.333333	2.666667	0.000000
10	-1.857143	5.857143	10.57143	-4.714286	0.000000	1.285714
11	-35.857143	4.428571	-19.28571	-35.000000	-26.428571	10.571429

Week	Reail and Recreation	Grocery and Pharmacy	Parks	Transit Stations	Workplaces	Residential
12	-75.000000	-31.142857	-55.71429	-68.000000	-60.142857	23.571429
13	-83.000000	-47.857143	-68.00000	-76.428571	-70.285714	28.000000
14	-84.142857	-46.857143	-65.42857	-78.571429	-71.571429	28.285714
15	-87.333333	-50.000000	-71.33333	-81.000000	-79.333333	33.666667

Table 1 Summary of Google mobility data from weeks 9-15 of the year 2020. Values represent the mean percent change from baseline for that week. The table shows a gradual decrease in most mobility factors over the first several weeks of the pandemic. Notably, the Google data shows an increase in mobility to grocery stores during weeks 9-11 which likely reflects the panic fueled stockpiling of necessities, a trend seen across the US which led to a shortage of items like toilet paper. Additionally, there was increasing mobility in residential areas, likely a reflection of people seeking outdoor exercise around their homes during quarantine.

Gene Name	Count
M	1
N	3
ORF10	1
ORF3a	1
S	1

Table 2 Count of distinct SNPs found for each named gene in the SARS-CoV-2 genome. In the samples provided, there were 7 distinct SNPs detected in named genes. Three of those variants were found in the N gene. It is worth noting that the SNP found in the spike protein was identified as the D614G variant which was the first known variant to surpass the wild type allele to become globally dominant.

	40	41	42	43	44	45	46	47	48	49	50	51
19A	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
19B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
20A	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67
20B	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.80	0.80	0.40
20C	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.40	0.40	0.80
20D	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
20E (EU1)	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33
20F	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
20G	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
20H/501Y.V2	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.29	0.29	0.57
20I/501Y.V1	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.50	0.50	0.25
20J/501Y.V3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 3 Matrix showing relationship between samples and clades. Value is the proportion of of SNPs in common over the total number in each clade. Higher values correspond to a greater similarity.

Sources Cited

Bedford,T. *et al.* (2021) Updated nextstrain SARS-CoV-2 clade naming strategy. *Nextstrain*.

Bolger,A.M. *et al.* (2014) Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Bushman,D. *et al.* (2020) Detection and genetic characterization of community-based SARS-CoV-2 infections—new york city, march 2020. *Morbidity and Mortality Weekly Report*, **69**, 918.

Cohen,J. (2020) The united states badly bungled coronavirus testing—but things may soon improve. *Science*, **10**.

Gonzalez-Reiche,A.S. *et al.* (2020) Introductions and early spread of SARS-CoV-2 in the new york city area. *Science*, **369**, 297–301.

Guidotti,E. and Ardia,D. (2020) COVID-19 data hub. *Journal of Open Source Software*, **5**, 2376.

Hadfield,J. (2020) August 2020 update of COVID-19 genomic epidemiology.

Hale,T. *et al.* (2020) Variation in government responses to COVID-19. *Blavatnik school of government working paper*, **31**, 2020–11.

Knaus,B.J. and Grünwald,N.J. (2017) Vcfr: A package to manipulate and visualize variant call format data in r. *Molecular ecology resources*, **17**, 44–53.

Li,H. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, **25**, 1754–1760.

Maurano,M.T. *et al.* (2020) Sequencing identifies multiple early introductions of SARS-CoV-2 to the new york city region. *Genome research*, **30**, 1781–1788.

Oran,D.P. and Topol,E.J. (2020) Prevalence of asymptomatic SARS-CoV-2 infection: A narrative review. *Annals of internal medicine*, **173**, 362–367.

Organization,W.H. and others (2020) Novel coronavirus (2019-nCoV): Situation report, 3.

Rocklöv,J. and Sjödin,H. (2020) High population densities catalyse the spread of COVID-19. *Journal of travel medicine*, **27**, taaa038.

Thompson,C.N. *et al.* (2020) COVID-19 outbreak—new york city, february 29–june 1, 2020. *Morbidity and Mortality Weekly Report*, **69**, 1725.

Wu,D. *et al.* (2020) The SARS-CoV-2 outbreak: What we know. *International Journal of Infectious Diseases*, **94**, 44–48.