Discovering new viral lineages and estimating their abundance in wastewater

by

Isaac Ellmen

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Science
in
Biology

Waterloo, Ontario, Canada, 2022

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Wastewater surveillance of SARS-CoV-2 has emerged as a critical tool for tracking the spread of COVID-19. In addition to estimating the relative case numbers using qPCR, SARS-CoV-2 genomic RNA can be extracted from wastewater and sequenced. The sequenced genomes provide information about which lineages, in particular which variants of concern (VOCs) are present in a community. Wastewater RNA sequencing data has two distinct challenges: First, the genomes are highly fragmented and the alignments often have poor genome coverage. Second, the samples are comprised of a mixture of genomes so mutations cannot be directly attributed to a single lineage. In this thesis, I explore methods to overcome these two challenges to extract useful information from the samples. First, I look at the problem of determining the relative abundance of VOCs. Most existing techniques only consider mutations which are unique to a particular VOC which massively reduces the amount of usable data. I introduce a new technique which extends mean and median frequencies over shared mutations in order to make use of the huge pool of shared mutations. Next, I investigate strategies for designing single-amplicon sequencing methods. I look at selecting single amplicons which are well-conserved and rich in information. I also design a single amplicon which is capable of amplifying multiple coronaviruses. I conclude the SARS-CoV-2 work by providing a technique which can identify novel lineages and sublineages from wastewater sequencing runs. Finally, I show that the techniques for analyzing SARS-CoV-2 in wastewater can also be applied to an important plant pathogen, the Tomato Brown Rugose Fruit Virus.

Acknowledgements

I would like to thank my co-supervisors, Dr. Trevor Charles and Dr. Jozef Nissimov. Researching a virus during an active pandemic has been an exciting but tumultuous journey and I have been very grateful for their support. I would also like to thank the members of my commitee, Dr. Andrew Doxey and Dr. Brendan McConkey, who helped me bridge the gap between computers and biology. This research would not have been possible without the funding support of Mitacs, the University of Waterloo, Health Canada, and the province of Ontario.

I leaned heavily on members of the Charles Lab and Metagenom Bio. Special thanks go out to Delaney Nash, Alyssa Overton, and Jenn Knapp from UW and Dr. Michael Lynch, Dr. Jiujun Cheng, and Rebecca Co from Metagenom. I also appreciated the lively discussions in the Nissimov Lab meetings. I learned a lot from the members of Ontario's Wastewater Surveillance Initiative as well as the broader wastewater sequencing community in Canada.

Finally, I would like to thank my parents and my partner, Anya Forestell, for their support over the last two years.

Dedication

This thesis is dedicated to my parents, Eugene Ellmen and Naomi Overend, and to my partner, Anya Forestell. Their support over the last two years has been tremendous and I am very grateful to them for listening when I needed to talk through my ideas.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

COVID-19 coronavirus disease 2019 1

KNN k-Nearest Neighbours 32

LP linear program 18

NMF non-negative matrix factorization 33

PCA principal component analysis 33

PCR polymerase chain reaction 1

PMMoV pepper mild mottle virus 5

PWM position weight matrix 63

qPCR quantitative polymerase chain reaction 2

RBD receptor binding domain 15

RdRp RNA-dependent RNA polymerase 25

SARS-CoV-2 severe acute respiratory syndrome coronavirus 2 1

SNP single nucloetide polymorphism 3

ToBRFV tomato brown rugose fruit virus 40

VOC variant of concern 1

WHO World Health Organization 4

WSI Wastewater Surveillance Initiative 8

WWTP wastewater treatment plant 1

List of Symbols

$|x|$ Absolute value of x: Equal to x if x is non-negative, or $-x$ if x is negative. 12

$\min_x$ Min over x: Find a value of x which minimizes the equation which follows. 12

$\sum_i$ Sum over i: Add up what follows over each element i. The set of all i is given by the context, although will usually refer to mutations or lineages in this work. 12

Chapter 1

Introduction

The coronavirus disease 2019 (COVID-19) pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has caused millions of deaths worldwide. The global response to the pandemic has led to unprecedented levels of genomic surveillance in order to track the spread and evolution of the virus. These efforts have led to the detection of many SARS-CoV-2 sublineages, some of which have increased transmissibility and virulence. The World Health Organization has identified a number of concerning SARS-CoV-2 sublineages over the course of the pandemic, which are termed variants of concern (VOCs). Owing to their vastly increased transmissibility, VOCs now represent virtually all COVID-19 infections. Tracking the emergence and spread of VOCs has been a critical tool in public health decision making, including the scheduling of lockdowns in Ontario [1]. Typically, VOCs are identified using polymerase chain reaction (PCR) tests for certain diagnostic mutations on clinical samples or, ideally, genomic sequencing of those samples. While clinical sequences remain the most reliable method for identifying VOCs, they can be slow and are expensive to produce, due to the volume of sequencing required for adequate coverage.

When some viruses (such as SARS-CoV-2) infect an individual, they make their way into the digestive tract, and some viral particles are passed in human waste. Viruses which are passed in the stool of humans usually make their way to wastewater treatment plants (WWTPs). Samples from WWTPs therefore contain RNA from infected individuals throughout the community which is served by the plant. Detecting viruses in wastewater had previously been used to monitor Polio [19]. Wastewater surveillance has two clear benefits which have fuelled a massive amount of interest during the COVID-19 pandemic.

- First, it is an efficient method for testing an entire community. Since the wastewater

contains RNA from many infected individuals, testing the wastewater is akin to testing an aggregate sample of a random subsample of the community. For certain tasks such as estimating abundance, wastewater has been shown to agree well with clinical testing, using only one test per time frame rather than hundreds or thousands.

- Second, wastewater tests often provide results which are a few days faster than clinical tests. This is likely because SARS-CoV-2 is shed in the feces before it causes symptoms which prompt individuals to get tested.

In addition to estimating case counts from quantitative polymerase chain reaction (qPCR) testing of wastewater RNA, it is also possible to identify VOCs using genomic sequencing. Typically RNA is extracted, amplified using one of the ARTIC amplicon primer sets [48], and sequenced on an Illumina or Nanopore machine. As with qPCR testing, wastewater sequencing provides a rich source of data which contains genomic fragments from a large number of individuals in a community. Unlike clinical sequencing, a single wastewater sample can be used to estimate the relative abundance of each of the VOCs in a community. However, wastewater sequence analysis has two distinct challenges compared to clinical sequencing.

- First, the samples are highly mixed, and it is usually not possible to determine assemblies of full genomes. This means that variant abundances must be estimated using the frequency of mutations on independently drawn reads rather than just the presence or absence of mutations across the genome.

- Second, the RNA which is extracted from wastewater is usually highly fragmented and often poor quality. It is common for samples to have highly uneven coverage of the genome which can lead to large gaps, sometimes more than 50%. Wastewater samples also typically have higher Ct values (lower RNA concentration) and so are more sensitive to dropout of amplicons, usually due to mutations on the primer binding sites.

The majority of the work presented in this thesis will cover a novel approach for estimating the relative abundances of VOCs from wastewater sequencing data. I will provide an overview of some common techniques for solving the problem and then introduce a new technique which abstracts some of these ideas. I have implemented the approach by creating a tool called Alcov, and I will show examples of real world cases which highlight the strengths of the tool. I also explore a new technique for identifying novel lineages in wastewater which I use to find known lineages/sublineages of SARS-CoV-2. Finally, I will show some work on extending the techniques for SARS-CoV-2 wastewater sequencing to other viruses in other environmental samples.

1.1 History

In December of 2019, a new respiratory virus was discovered in Wuhan, China which quickly spread across the globe. The virus, SARS-CoV-2, was quickly sequenced and the sequenced genome was made public. The sequence shared striking similarities to the original SARS-CoV virus but had some modifications, including on the spike protein which likely resulted in increased transmissibility. Despite efforts to contain the spread, COVID-19 was declared a global pandemic on March 11, 2020 [33].

There was substantial interest in using genomic epidemiology to track mutations in the virus as it spread across the world. There are two main objectives in tracking mutations. First, benign, random mutations act as fingerprints to identify particular viral lineages. These can be used to determine where and how a particular strain is spreading, such as through contact in hospitals. Second, mutations occasionally induce some phenotypic change. The most concerning of these changes are an increase in transmissibility, virulence, or immune evasion.

Many global projects were established to perform collaborative genomic epidemiology for SARS-CoV-2. GISAID is a database for sharing viral sequences, annotated with metadata such as where and when the virus was isolated [38]. Nextstrain is a web application which shows the phylogenetic tree of all known SARS-CoV-2 sequences and tracks which lineages are present in different locations [18]. The website outbreak.info provides figures for SARS-CoV-2 lineages including heatmaps which show the frequency with which each lineage contains each mutation [20]. Pangolin and Nextclade are tools which can be used to identify which lineage a given SARS-CoV-2 genome belongs to [34, 4].

SARS-CoV-2 mutations are typically written as:

`[gene]:[wildtype amino acid][location on protein][mutant amino acid]`

for nonsynonymous mutations and

`[wildtype nucleotide][genomic location][mutant nucleotide]`

for synonymous single nucloetide polymorphisms (SNPs).

While the original goal of tracking SARS-CoV-2 genomes was to monitor its spread, the focus quickly shifted to monitoring concerning mutations or lineages. Identifying mutations which increase transmissibility is typically done by noticing that mutants with the mutation consistently out-compete strains which lack the mutation in the same population. Perhaps the first example of a mutation which incurred increased transmissibility was the mutation S:D614G. Mutations on the spike protein (S) are often significant because the spike protein

binds to human receptors and changes in binding affinity can readily affect transmissibility. By late 2020, lineages with S:D614G were dominant globally.

On May 31 2021 the lineage B.1.1.7 was named Alpha and declared the first "variant of concern" (VOC) by the World Health Organization (WHO). Alpha contained a number of unique mutations, including S:N501Y which increased the spike's binding affinity to human ACE2. Alpha's transmissibility was estimated to be about 1.5x wildtype SARS-CoV-2 so it quickly became dominant in the UK and abroad. More VOCs rapidly followed including the Beta, Gamma, Delta, and Omicron.

Because of previous infections and a significant effort to vaccinate the global population, non-VOC strains of SARS-CoV-2 have essentially been eradicated in humans. The most common task now is to identify emerging VOCs and their subvariants, such as BA.5 in Omicron. As SARS-CoV-2 mutates and spreads, its VOC definitions are constantly changing. For instance, Alpha, the original VOC has not been downgraded to a "Variant Being Monitored". Similarly, the names for different lineages are flexible. In this thesis I will use the terms "variant" and "lineage" interchangeably. A subvariant (or sublineage) is simply a descendant of a larger variant, however subvariants are occasionally referred to as just variants if they begin to represent a substantial number of all cases, such as BA.5.

1.2 SARS-CoV-2 basics

SARS-CoV-2 is an RNA virus with a length of approximately 30,000 bp. It is a coronavirus and is genetically similar to other human coronaviruses such as SARS and MERS [52]. Like other coronaviruses, it is an enveloped, positive-sense, single-stranded RNA virus [27]. Most of the SARS-CoV-2 genome codes for ORF1ab, a non-structural polyprotein which contains the RNA-dependent RNA polymerase [52]. SARS-CoV-2 contains 4 structural proteins: spike (S), nucleocapsid (N), membrane (M) protein, and envelope (E). Importantly, the spike protein binds to human ACE2 which facilitates viral entry into cells. Because it is a receptor binding protein, S has been the primary target for most COVID-19 vaccines. Mutations in S have been meticulously studied. Some S mutations, such as N:501Y are known to alter binding affinity, creating inherently more transmissible/virulent variants [26]. Others have the potential to induce immune escape by avoiding binding to wildtype antibodies while still binding to ACE2 [51].

Chapter 2

Sequencing SARS-CoV-2 in wastewater

I would like to preface this section by reminding the reader that I am a bioinformatician. While I do my best to stay up to date with how the data is generated, I was not actually responsible for performing any of the wet-lab work. Still, it is important to know how the samples are processed, as this can affect the final analysis. With that in mind, here I will describe a common workflow for sequencing SARS-CoV-2 in wastewater, and some common methods for analyzing the resulting data.

2.1 Sample processing and sequencing

Samples of wastewater are typically drawn from WWTPs or occasionally from "passive samplers" at locations such as university campuses. Sometimes viral genomes are extracted from the wastewater using magnetic beads. The RNA is then extracted and reverse-transcribed to yield cDNA.

It is common to run qPCR on the cDNA prior to sequencing to determine a relative abundance of SARS-CoV-2 in the sample. Usually qPCR is also run on the pepper mild mottle virus (PMMoV), a pepper virus which is fairly uniformly found in wastewater due to its presence in human food. PMMoV has previously been identified as the most abundant virus in human feces, and does not fluctuate substantially due to its presence in a wide variety of human foods, such as hot sauces and curries [22]. The ratio of SARS-CoV-2 concentration to PMMoV concentration gives a measure of the SARS-CoV-2 signal in a

community which can be used as a proxy for case counts. Additionally, the raw Ct values can be used as a cutoff since high Ct valued samples often produce very poor sequencing data.

The cDNA is then prepared for sequencing by amplifying the SARS-CoV-2 genome in a series of overlapping amplicons which are typically about 400 bp long. The most common amplicon primer scheme is the ARTIC v3 or v4 primer set which divide the genome into 98 amplicons [32]. The amplicons are then sequenced, often using 2x250 Illumina reads and the generated fastqs are processed.

2.2 SARS-CoV-2 data analysis

The analysis pipeline for SARS-CoV-2 sequencing data depends on the specific objective. A typical clinical sequencing pipeline looks like:

- Read quality filtering
- Adapter trimming
- Read pairing
- Primer trimming
- Alignment to a reference
- Generation of a consensus genome
- Phylogenetic placement of genome
- Flagging unusual mutations

As mentioned earlier, there are two key differences between wastewater and clinical sequencing. First, whereas clinical samples are typically seeded by a single virion and so contain little variation, wastewater samples often contain many distantly related lineages. This means that we cannot take consensus genomes since they conceal any lineages that are present at under 50% frequency. Second, the data is very low quality and often contains large gaps in genome coverage. Because of these considerations, we typically avoid any quality filtering since this typically results in much better coverage.

2.3 Wastewater data preprocessing

In general, sample preparation and data preprocessing should be chosen to maximize the amount of usable data and minimize the bias in that data. Often times the biases can be hard to identify and require careful attention. Here, I will describe each step of our preprocessing pipeline, and include examples of where there can be issues with bias. All steps in this pipeline were optimized for Illumina MiSeq reads. These steps are carried out by a Python script which is available in the appendix.

2.3.1 Adapter trimming and read pairing

I use the tool cutadapt [29] for adapter trimming. Previously, I used SeqPrep [43] for adapter trimming and read pairing and only considered paired reads. By throwing out unpaired reads and implementing a quality score cutoff for read pairing, SeqPrep was implicitly filtering low-quality reads. Since coverage and quality are so variable, and since all analysis methods must be tolerant to local mutations anyway, it was found that retaining low-quality reads ultimately leads to more reliable predictions.

2.3.2 Alignment to the reference genome

RNA viruses like SARS-CoV-2 mutate very rapidly which can be a challenge for alignment software. I use the tool minimap2 [25] for aligning the paired (and unpaired) reads to the SARS-CoV-2 reference genome, NC_045512.2 [52]. Previously, I used bwa [24] for aligning the reads. The choice of alignment methods can be very important since large deletions are important mutations for calling certain VOCs such as Omicron. It was found that bwa could not tolerate 9 bp deletions which caused it to discard all reads containing the N:DEL31/33 mutation, an important diagnostic for Omicron. Note that discarded reads usually fail silently, and manifest as either missing coverage on that location, or (even worse), only non-mutated reads at that location. Because of this, in mixed samples, an alignment program which selectively discards reads for only one VOC will cause all subsequent analysis to underestimate the prevalence of that VOC. Qualitatively, minimap2 seems to be able to capture all reads for all known VOCs in Ontario, possibly due to its concave gap cost for long deletions [25].

2.3.3 Read conversion, sorting, and indexing

Samtools is used for converting the .sam files to the binary .bam filetype to save space. It is then used to sort the reads and index them so that they can be efficiently queried by read processing software such as variant callers.

2.3.4 VOC estimation

The sorted bam files are passed directly to Alcov, a tool that I created for determining the compositions of different VOCs in the sample. Alcov accepts either a single bam file or a text file containing a list of bam files, as well as (optionally) a list of VOCs to look for. Alcov was written in Python and uses different convex optimization packages for estimating variant abundances. The theory and implementation behind Alcov will be discussed at length in the next section.

2.4 Ontario's wastewater sequencing initiative

Much of the work in this thesis was done in support of Ontario's Wastewater Surveillance Initiative (WSI). The WSI was established in 2020 to monitor SARS-CoV-2 in Ontario wastewater in support of public health [1]. Funding for the WSI covers qPCR surveillance for detection and quantification of SARS-CoV-2 as well as sequencing for analyzing lineages. Our lab at the University of Waterloo has participated in the WSI, through routinely sequencing wastewater from about a third of the 170 locations which are included in the initiative. The analysis of the data is completed collaboratively with the University of Western Ontario. Part of the analysis which is reported back to the partners is done using Alcov, the tool that I created which is discussed in the next chapter.

Chapter 3

VOC abundance estimation with Alcov

Estimating the relative abundances of different VOCs is one of the central goals of wastewater sequencing. Knowing the relative abundance of each VOC allows us to determine when new VOCs enter a community, and how quickly a given VOC begins to dominate. These are keys pieces of information for public health decision making such as planning lockdowns. The challenge is similar to estimating the composition of 16S/18S or metagenomic sequencing samples. One of the key differences is that the defining mutations are sparser than reads from across different species so it is uncommon to capture multiple defining mutations on a single k-mer or even a single read. This means that k-mer based analyses essentially reduce to mutation based analyses but are less resilient to sequencing errors or locally prevalent mutations. Another key difference is that the SARS-CoV-2 reads are amplified using PCR amplicons which often produce highly variable read depths across the genome. Variability in amplification can be caused by RNA degradation, inefficient primer binding, or random variation during the early rounds of PCR. Any analysis which assumes independence of reads will over-estimate the certainty of a prediction based on duplicate reads, and will have a heavy skew towards well-amplified regions (even though these regions may not contain more useful data). Finally, a key consideration with regards to SARS-CoV-2 is the wealth of available mutation data associated with defined, named, lineages. Analysis which is based on lineage-defining mutations is much easier to justify (and often equivalent to) compared to k-mer or read-alignment methods.

For these reasons, the community has largely focused the analysis of wastewater data on the analysis of mutation frequencies, that is the number of times a mutation is observed divided by the total read depth at that position on the genome. In this section I

will describe two common approaches to using mutation frequencies for estimating VOC abundances and introduce a model which abstracts these to use more available data.

3.1 Estimating VOC abundances by averaging *lineage defining* mutations

The problem of translating mutation frequencies to VOC frequencies essentially boils down to coming up with an estimate of the VOC frequencies which best explains the mutation frequencies. A naive approach to this is, for each VOC, to simply take the average frequency of each of the mutations which it contains. The issue is that some VOCs contain shared mutations, for instance every VOC now contains the mutation S:D614G. Taking a simple average will, in general, cause the program to over-estimate the abundances. The most common solution to this problem is to generate a set of *lineage defining* mutations which are relatively unique to a given VOC. I say relatively because, with over 1 million sequenced genomes, and several known VOCs, mutations are almost always found across multiple lineages. Still, it is usually possible to come up with a set of mutations which are present in at least, say 95% of a given lineage and at most 5% of all other lineages. The Pango network maintains a list of lineage defining mutations which are used for calling clinical sequences, available at (https://github.com/cov-lineages/constellations). These mutations are neither completely unique, nor exhaustive for each variant, but are typically associated with each VOC and this list is commonly used to determine the lineage-defining mutations.

The word average can refer to different measures. The most commonly used measure is the *mean*, which divides the sum of the entries by the number of entries. An alternative is the *median*, which sorts the entries and then considers only the middle entry (or mean of the two middle entries if there are an even number). The choice between mean and median is important in cases where there is poor, variable coverage. When a variant is present at a low frequency, it is common that only one or two mutations will be amplified by PCR and detected in the sequencing run. Conversely, it is relatively common for a locally prevalent lineage to contain certain mutations which are not normally associated with that VOC. If the bioinformatician would like to make their analysis sensitive to mutations which only appear at a low frequency so that they can detect low-frequency variants then they may wish to use the mean. If, instead, they want to make the analysis resilient to sequencing errors or local mutations, and ignore outlier mutations, then they may wish to use the median. In practice, both of these methods are used.

3.2 Considering shared mutations

The requirement of uniqueness for each mutation can cause issues in the analysis. If the mutation is not actually unique, then including it can erroneously support the presence of the wrong VOC. On the other hand, there are a large number of important mutations which are not unique but can still provide evidence for a given VOC if used in the right context. As seen in Figure 3.1, simply monitoring the (shared) mutation S:N501Y tells the story about the rise of Alpha, Delta, and Omicron.

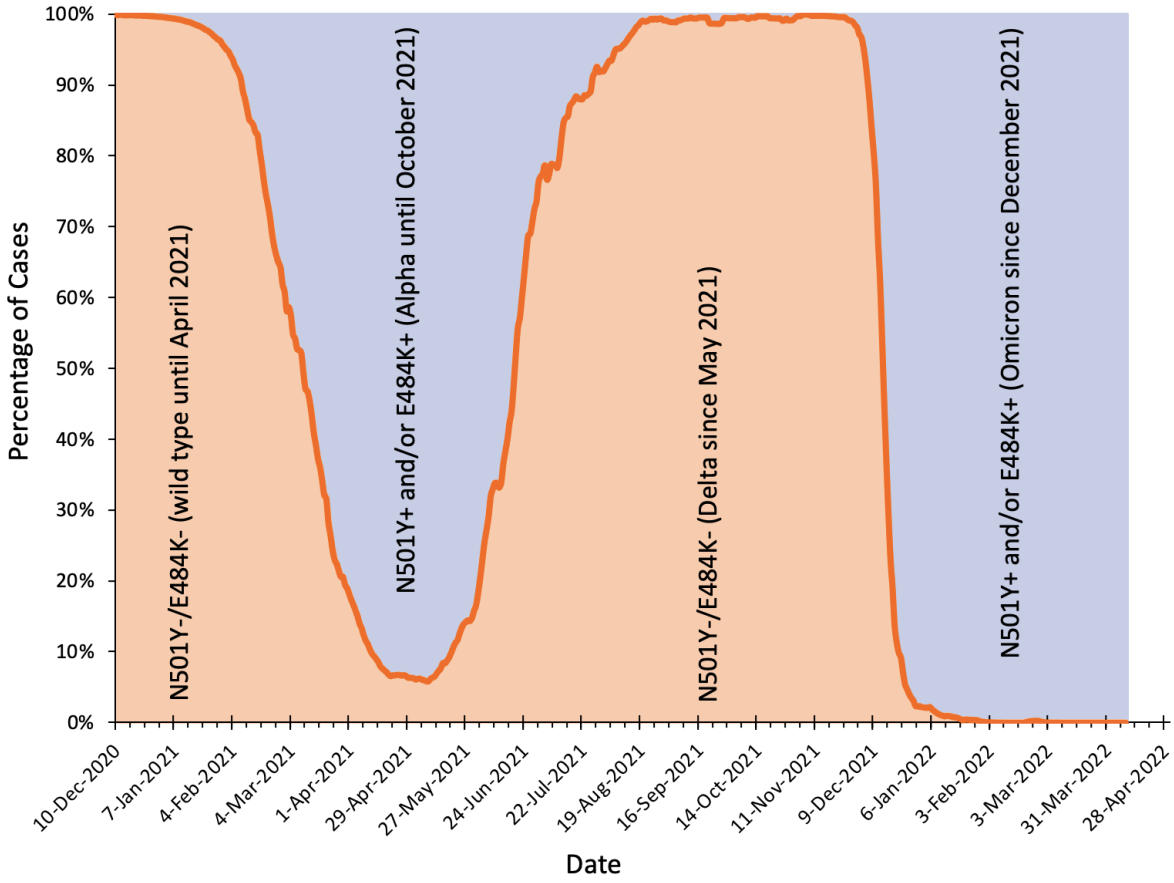


Figure 3.1: Prevalence of S:N501Y in Ontario over time. Taken from the Ontario Science Table Dashboard [45]

In order to proceed, we will consider the mathematical formalization of the problem.

I have explained what each equation means, and have included a short description of the meaning of $\sum_i$, $\min_x$, and $|x|$ in the glossary under *List of Symbols*. We can consider the problem formulated as the following:

Determine an estimate for the relative abundance of each VOC such that the predicted mutation frequencies best match the observed mutation frequencies.

Let y_i denote the observed frequency of mutation i and $x_{i,j}$ denote whether lineage j contains mutation i. Finally, let β_j denote the predicted relative abundance of lineage j in the sample. Then the predicted frequency of mutation i is

$$\sum_j \beta_j x_{i,j} \tag{3.1}$$

We can cast the problem of determining the lineages as determining a set of β_j to minimize the total difference between the observed mutation rates and the predicted mutation rates:

$$min_\beta \sum_i |y_i - \sum_j \beta_j x_{i,j}| \tag{3.2}$$

or similarly

$$min_\beta \sum_i (y_i - \sum_j \beta_j x_{i,j})^2 \tag{3.3}$$

Equation 3.3 is a well studied problem in optimization and statistics, known as ordinary least squares (OLS). OLS can be solved in closed form [39] or using quadratic program solvers. Because of its relation to linear regression, efficient solvers are also packaged into common machine learning libraries such as scikit-learn [35]. Viewed as a regression, we are optimizing the parameters of a linear model to best account for each mutation which acts as a data point. One advantage of this formulation is that we can only include the mutations, or data points, for which we have coverage in a given sample.

This was the first model which powered Alcov.

To better understand how the model is minimizing the difference between observed and predicted mutation frequencies, we can consider only the unique mutations for each lineage. In this case, we simply optimize the relative abundance of each lineage independently. For

a given lineage j, we minimize the squared difference between the predicted frequency of the mutations i that j contains and the observed mutation frequencies y_i.

$$min_{\beta_j} \sum_i (y_i - \beta_j)^2 \tag{3.4}$$

We can take the derivative of 3.4 with respect to β_j which gives:

$$2 \sum_i (y_i - \beta_j) = 0 \tag{3.5}$$

and so:

$$\beta_j = 1/n \sum_i y_i \tag{3.6}$$

where n is the number of unique mutations which lineage j contains.

Put simply, considering only the unique mutations, the least square model will be minimized by taking the mean frequency of all unique mutations for each lineage. This shows that the model is a sensible abstraction which extends the idea of a mean frequency over all shared mutations. The following section, adapted from the Alcov manuscript [13], shows how Alcov is able to predict sensible lineage abundances on real-world data.

3.3 Applying Alcov to real-world wastewater sequencing data (adapted from Alcov medRxiv preprint)

3.3.1 Data collection

In order to test the predictions made by Alcov, we searched for "SARS-CoV-2 wastewater" on the Sequencing Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra). Seven SRA studies were found, of which three had sufficient data to make predictions. One recent sample from each study was downloaded and preprocessed using the steps described above before being passed into Alcov for VOC abundance prediction. The specific samples used were SAMN18310570, SAMN18378816, and SAMN18915228. These samples were each analyzed as a case study. The VOC mutation information Alcov uses was downloaded from https://outbreak.info/compare-lineages on May 12, 2021.

3.3.2 Frankfurt, Germany - December 2020 (SAMN18310570)

The first sample analyzed was collected from Frankfurt in December 2020 [3]. This sample acted as a negative control since there were not widespread VOCs in Germany at that time. As expected, Alcov predicts negligible abundances for all of the variants of concern 3.2. The slightly non-zero values are likely due to sequencing/reverse transcription errors, or low levels of sequence variation within that community.

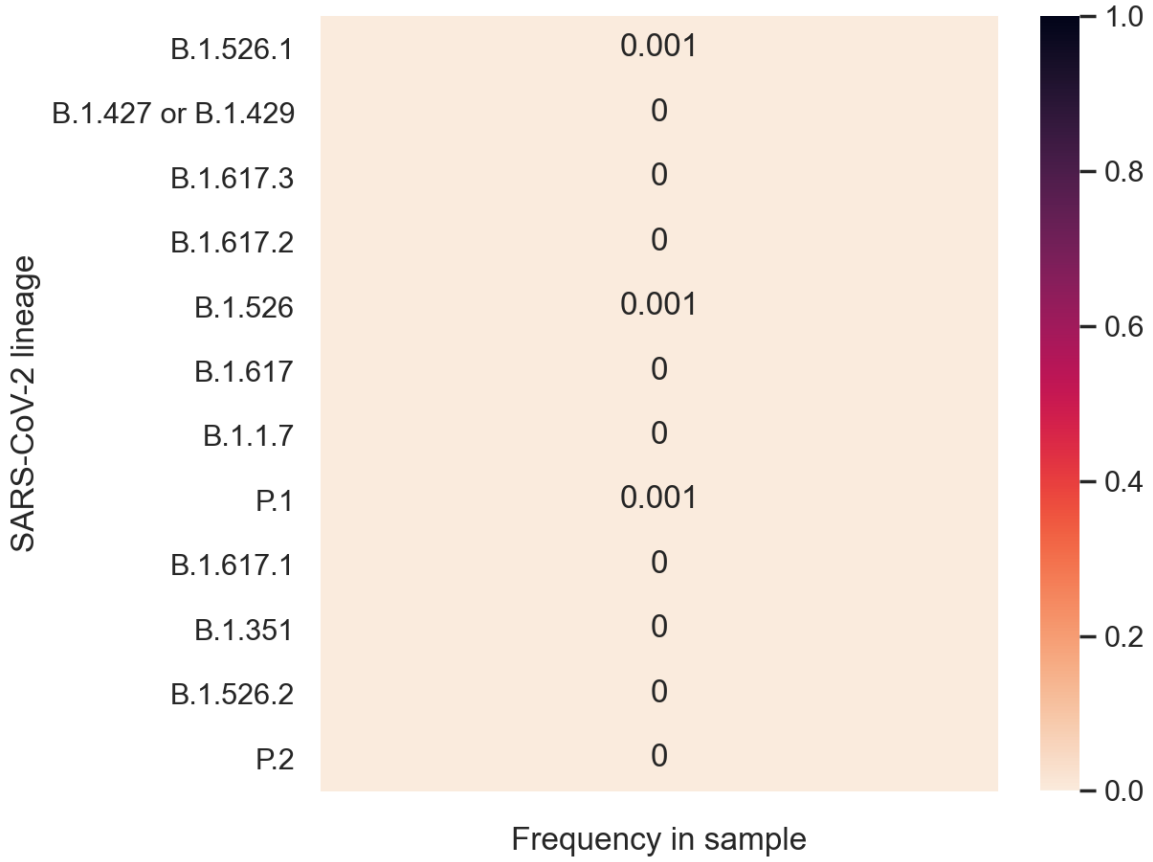


Figure 3.2: Predicted lineages in Frankfurt (December 2020). Lineages are coloured according to their predicted frequency which is also labelled. As expected, the frequency of each VOC lineage is negligible.

3.3.3 New York City, New York, USA - March 2021 (SAMN18378816)

The second sample analyzed was collected in New York City (NYC) in March 2021 [42]. Only a 332-nucleotide region of the receptor binding domain (RBD) was sequenced which covered six known mutations present in VOCs (S:L452R, S:S477N, S:T478K, S:E484K/Q, and S:N501Y). Because multiple VOCs were identical with respect to these mutations, some were merged before making the prediction. Two VOCs were predicted by Alcov in this sample, both of which are consistent with clinical data [20]. B.1.1.7 was predicted to account for 43% of the sample whereas some mixture of B.1.526.1, B.1.427, and B.1.429 was predicted to account for 22% of the sample 3.3. It is very likely that the latter portion was due to B.1.526.1, which is a strain that originated in NYC and has been observed in high prevalence in NYC clinical samples [20]. Similarly, the B.1.1.7 proportion is in line with clinical sampling data.

3.3.4 Newport, Oregon, USA - March 2021 (SAMN18915228)

The third sample analyzed was collected in Newport in March 2021. The sample had many regions with low coverage which made it challenging to analyze by looking at only the most common characteristic mutations. Alcov predicted a split between 62% B.1.417/B.1.429 and 38% B.1.351 3.4. The B.1.417/B.1.429 prediction makes sense since these VOCs originate from California which is adjacent to Oregon and are present in clinical samples from Lincoln county which includes Newport [20]. The B.1.351 prediction is surprising since it is rare in Oregon and, to our knowledge, has not been discovered in Lincoln. Mutations which were not included in the prediction due to insufficient read depth were analyzed manually, and supported the B.1.351 prediction. This example may demonstrate some of the utility of Alcov for wastewater identifying VOCs that have yet to be captured in clinical sequencing. That said, it is unlikely that the proportions in this sample are representative of the larger community, and may have been exaggerated due to stochasticity in PCR amplification.

3.4 Issues with the least square model

While the least square model has a satisfying theoretical basis, and often makes reasonable lineage calls, it has a tendency to call a high number of low but non-zero lineage abundances.

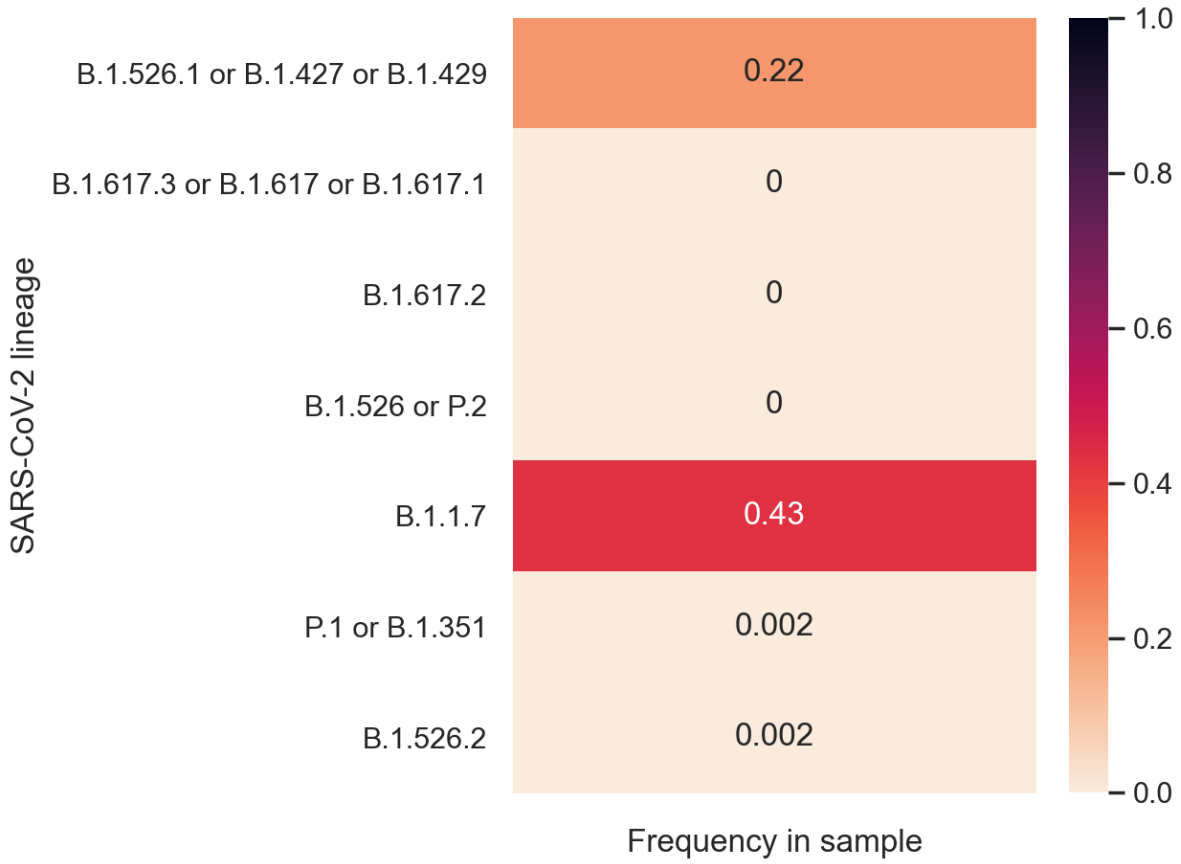


Figure 3.3: Predicted lineages in NYC (March 2021). Because of limited sequencing coverage, some VOCs were indistinguishable and so automatically merged. The B.1.1.7 and B.1.526.1 predictions agree with clinical data.

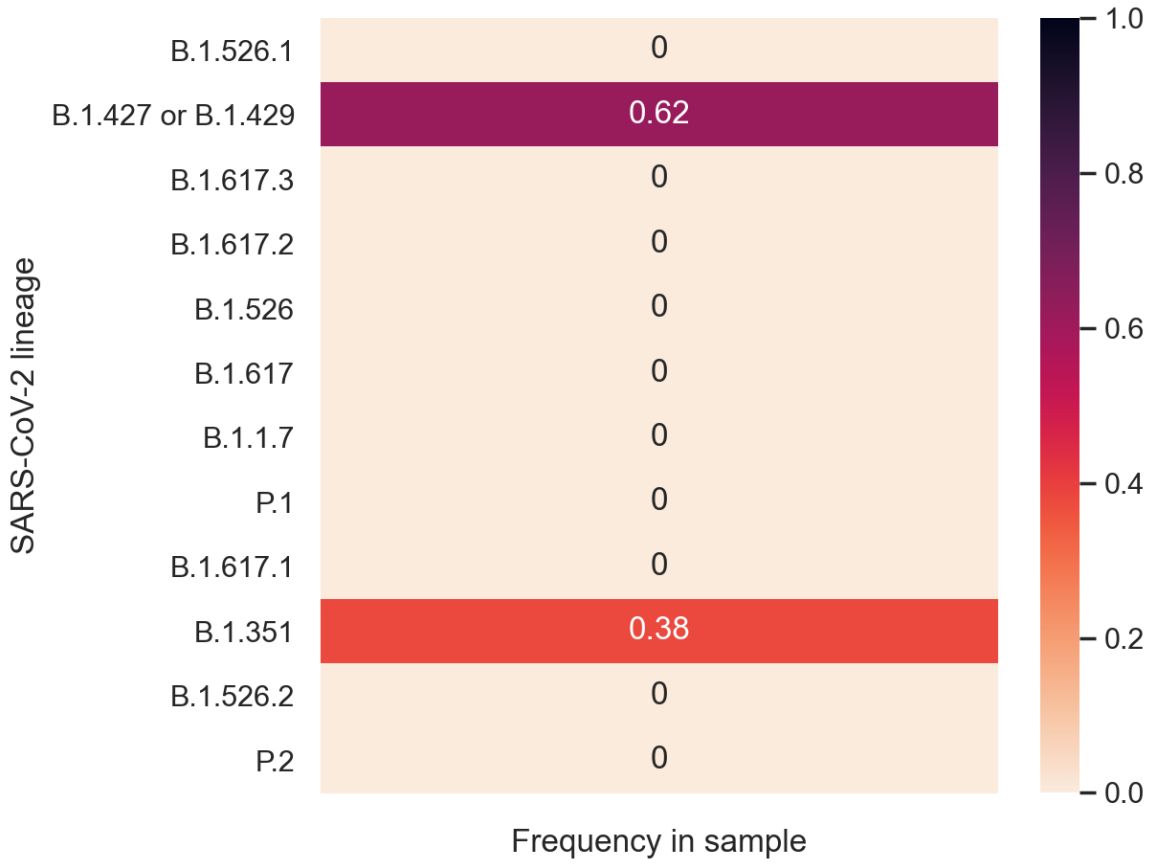


Figure 3.4: Predicted lineages in Newport (March 2021). The B.1.427/9 lineages originated in California which is adjacent to Oregon. The B.1.351 prediction is high compared to clinical data but is well supported across multiple mutations in the sample.

To understand this, consider two lineages, one which contains mutations a, b, and c and another which contains mutations c, d, and e. Suppose we observe mutation a at 80%, mutation b at 90%, mutation c at 100%, and mutations d and e and 0%. We might guess that we should predict lineage 1 at 90% and lineage 2 at 0% (since it is missing 2/3 of its defining mutations). However, in this case the squared error is 200 $((90-80)^2+(100-80)^2)$. The model will notice that predicting some small amount of lineage 2 (say 2%) actually decreases the squared error since it reduced the error on mutation c from 100 $((100-90)^2)$ to 64 $((100-(90+2))^2)$ while only increasing the squared error on mutation d and e by a total of 8 $(2\times(2-0)^2)$.

This property of tending to predict a large number of non-zero values is known as a tendency to predict *dense* solutions and is well known for least squares models. Normally, this is not an issue since the small non-zero values have only a small impact on the predictions made by the model. However, since the parameters learned by our model represent lineage abundances, non-zero values should be predicted only when they are well-supported to avoid erroneously suggesting the presence of concerning lineages. We can achieve this by returning to our error metric and encouraging our model to predict *sparse* solutions.

3.5 Minimizing the ℓ^1 norm

As discussed in *Convex Optimization* [7] minimizing equation 3.3 tends to give sparse solutions. Instead of minimizing the squares or ℓ^2 norm, we can minimize the absolute value of the error or ℓ^1 norm. Minimizing the ℓ^1 norm can be achieved by casting equation 3.2 as an optimization problem called a linear program (LP) [7]:

$$\begin{aligned} \underset{\beta_i,\, t_i}{\text{minimize}} \quad & \sum_i t_i \\ \text{subject to} \quad & -t_i \le y_i - \sum_j \beta_j x_{i,j} \le t_i \end{aligned} \tag{3.7}$$

To understand why 3.7 minimizes the absolute error, think of each t_i as the error over mutation i. Then the program is trying to find a set of relative abundances (β_i) and errors (t_i) such that the sum of the errors are minimized.

The LP can be solved efficiently using standard LP optimizers such as the simplex algorithm. I implemented 3.7 in Python using the Google OR tools library [36].

As with the least squares model, we can gain insight into how the model is predicting lineage abundances by considering how it makes predictions using only unique mutations.

For a given lineage j, we minimize the difference between the predicted frequency of the mutations i that j contains and the observed mutation frequencies y_i.

$$min_{\beta_j} \sum_i |y_i - \beta_j| \tag{3.8}$$

We can take the derivative of 3.8 with respect to β_j which gives:

$$\sum_i sign(y_i - \beta_j) = 0 \tag{3.9}$$

where $sign$ is the sign function (1 if $y_i - \beta_j$ is greater than 0, 0 if $y_i - \beta_j$ is 0 and -1 if $y_i - \beta_j$ is less than 0)

In order for 3.9 to be true, we require that there are an equal number of y_i which are greater than β_j as y_i which are less than β_j. In other words, we require that β_j be the median frequency of its unique mutations.

This is a very satisfying result. We saw that we could abstract the mean frequency of unique mutations to shared mutations by minimizing the squared difference over all mutations. We then observed that this tended to produce dense solutions, a well-known but undesirable property for predicting lineage abundance. We observed in the literature that minimizing the ℓ^1 norm instead of the ℓ^2 norm tends to produce dense solutions. By restricting the new formulation to only unique mutations, we arrive at the median frequency. In other words ℓ^1 minimization leads to a generalization of the median frequency in exactly the same way that ℓ^2 minimization leads to a generalization of mean frequency. The relationship between the four methods is shown in 3.5

It is now clear why ℓ^1 minimization tends to produce sparser solutions than ℓ^2 minimization. Consider a lineage with 10 unique mutations, 9 of which are observed at 0%, and 1 of which is observed at 100%. We may deduce that the tenth mutation is not actually unique and that the lineage is likely not present in our sample. The median (ℓ^1) method takes the middle value, which is 0% and predicts none of the lineage. The mean (ℓ^2), however, takes the average of the ten frequencies and predicts that the lineage is present at 10%.

In general, the ℓ^1 minimization is much more resilient to errors which can arise from locally unique sublineages, PCR bias, contamination, or other sources. Conversely, the ℓ^2 minimization tends to be more sensitive to low abundances of a new lineage, where most of the characteristic mutations frequencies are 0% (unamplified).

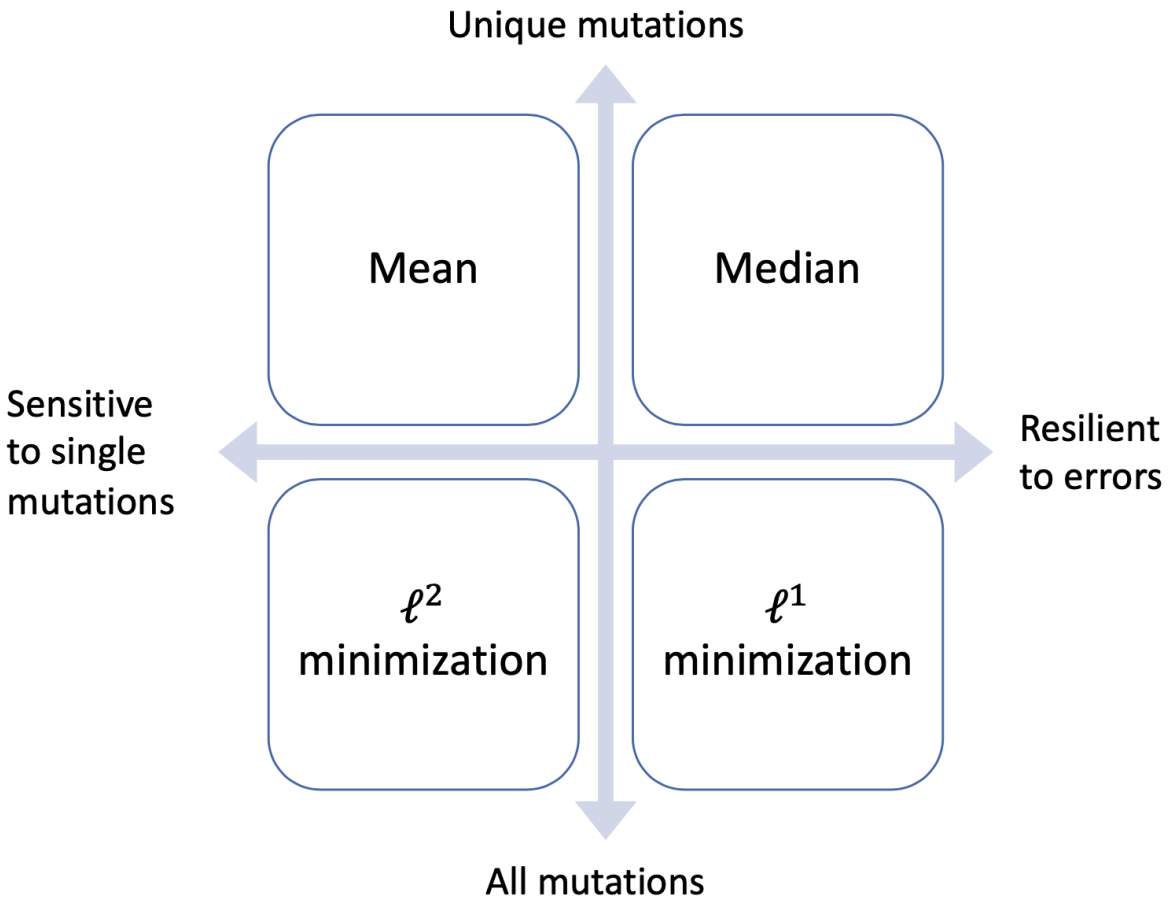


Figure 3.5: Comparing the four methods. Mean and median can each be abstracted to include shared mutations by minimizing the error of predicted and observed mutation frequencies. Methods on the right are less sensitive to individual mutations, and so are more resilient to errors. Alcov supports all four methods, using the *unique* and *l2* flags.

Because of its resilience to errors, the ℓ^1 minimization model forms the basis for Alcov2.

As discussed earlier, the way this model is actually minimized is by casting it as an LP. Casting the problem as an LP has a few practical advantages.

First, we can ensure that each predicted lineage abundance is greater than 0:

$$\beta_j \geq 0 \tag{3.10}$$

This ensures that the model will not try to "invent" a new lineage by subtracting one lineage from another.

Second, we can set additional constraints such as:

$$\sum_j \beta_j \leq 1 \tag{3.11}$$

or

$$\sum_j \beta_j = 1 \tag{3.12}$$

to explicitly ensure that the relative abundances add up to at most or exactly 100%.

Finally, we can set weights to prioritize certain mutations over others. An example of where we may wish to set a higher weight is a deletion of 1 or more amino acids since a deletion of one amino acid typically requires a deletion of three nucleotides, and because nucleotide deletions are less common to appear by chance than nucleotide substitutions. This can be easily achieved using by minimizing the weighted error:

$$min_\beta \sum_i w_i |y_i - \sum_j \beta_j x_{i,j}| \tag{3.13}$$

where w_i is the relative weight of mutation i. Alcov2 does not currently use weighted mutations, but it may be a promising feature to add, especially for identifying low abundance lineages.

3.6 Generating mutation lists

A core requirement of mutation frequency methods is the availability of curated lists of mutations for each VOC. As with other SARS-CoV-2 resources, there are a variety of sources for mutation lists and no one standard has been accepted by the community. There are complexities in generating these lists. The variant definition files:

- must decide which mutations are sufficiently abundant across the lineage
- must decide which mutations are sufficiently unique across other lineages

- can change over time as variants evolve

A key challenge is the requirement of uniqueness, which necessarily changes as variants evolve. A classic example of this is the introduction of subvariants. When omicron first emerged, it contained many unique mutations, however as subvariants evolved, there became fewer and fewer unique mutations per variant.

As the pandemic evolved, I switched my source of mutation information many times. First, I manually curated a list of a few variants from official sources such as the WHO. Next, I used the download function from the outbreak.info heatmaps. Unfortunately, the download feature was removed, so I switched to a json file supplied by covariants.org. The covariants json file did not update when Omicron subvariants emerged so I switched to using the cov-lineages/constellations GitHub repository. At the time of writing (June 2022), there are three current challenges with this source.

- The repository is missing some important Omicron subvariants.
- The definitions use an inconsistent syntax for representing mutations which makes it difficult to parse.
- Most (but not all) shared mutations are omitted from the definitions.

Going forward, it would make sense to generate our own mutation lists by parsing genomes from GISAID. The most difficult part of maintaining a repository of mutations is ensuring relative uniqueness, which is not an issue for Alcov, since it uses shared mutations. In fact, including more shared mutations would likely result in better predictions, as long as all major variants which are present in a sample have definitions in the database.

Chapter 4

Single amplicon sequencing methods

As previously discussed, wastewater sequencing of SARS-CoV-2 often produces alignments with sparse coverage. This is particularly problematic when trying to detect emerging variants for two reasons:

- First, with very low coverage, it is important to recover highly diagnostic mutations such as large deletions which may only be spanned by one or two amplicons.
- Second, amplicons are typically optimized for known variants. New variants sometimes have mutations on the primer binding sites which reduce PCR amplification and make the variants harder to detect.

A partial solution to these problems is to sequence only a single amplicon. This ensures that there will be deep coverage of the corresponding region on the genome which can be selected to be highly diagnostic. Single amplicon sequencing is not a new idea. Indeed, it is widely used in 16S/18S ribosomal sequencing for community analysis [5]. Additionally, single amplicon sequencing is very similar to the widely used qPCR quantification of variants. Instead of designing a novel primer for each variant and comparing it to a generic primer in the same region, we can simply take the frequency of reads which contain the diagnostic mutation(s).

A final application of single amplicons is that, carefully chosen, they can enable sequencing of multiple coronaviruses. Such an approach could be used to compare the ratio of SARS-CoV-2 to milder coronaviruses in the population. It could also be used to identify novel coronaviruses by sequencing environmental samples around known spillover hazards like bat caves.

4.1 Some amplicons are better conserved in wastewater

A consideration in qPCR primer design is ensuring that the target site is well conserved in wastewater. As one might expect, RNA is generally unstable in wastewater. This is due to a number of factors, including the presence of RNA-degrading enzymes [21]. Figure 4.1 shows three samples from the same run, amplified using two different versions of ARTIC primers (v3 and v4). There is a wide range of coverage, from only a few amplicons to almost 100%.

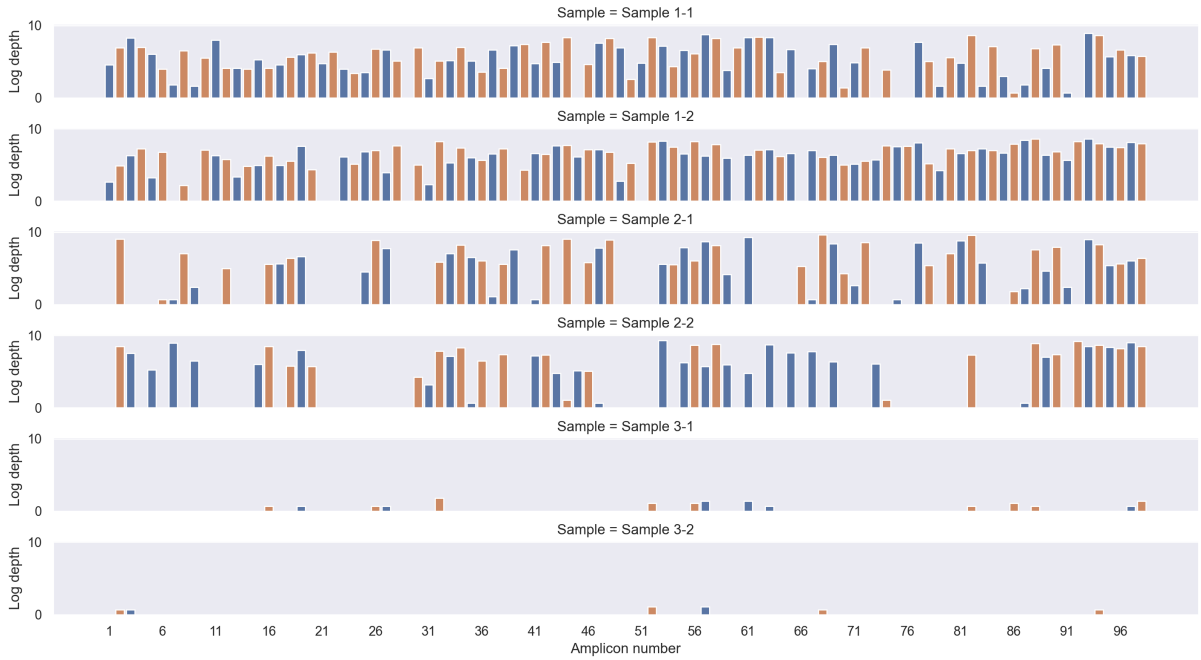


Figure 4.1: Coverage across three different samples from the same sequencing run, repeated with two different ARTIC primer versions each. The level of RNA degradation (and so coverage) varies markedly. Overall coverage is highly dependant on the initial RNA concentration which can be relatively estimated using qPCR.

It is generally accepted that the N gene on SARS-CoV-2 tends to be amplified well, although the reason for this is unclear. I wondered if the reliability of some amplicons over others may be due to greater stability in the secondary structure of that region. While it is difficult to predict secondary structure over long RNA sequences [41], a broad predictor of secondary structure is GC content since G-C bonds are stronger than A-U bonds [9].

I took data from all five samples in one of our runs and plotted the log(depth+1) (similar to Ct) vs. GC content for each amplicon. The value log(depth+1) was chosen instead of log(depth) so that amplicons which were not amplified were given a log(0+1)=0 whereas log(0) is undefined. I created a scatter plot with a best-fit line using seaborn in Python. I first analyzed all amplicons which was heavily influenced by amplicons with total dropout. I then looked at only the amplicons which had some coverage, in order to compare only the level of amplification.

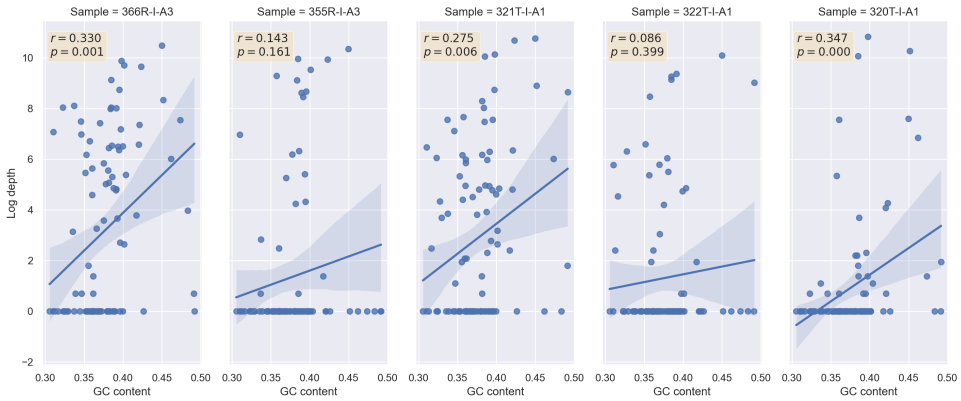
In both cases all five samples showed a positive correlation between GC content and log(depth). I calculated the Pearson correlation coefficient and p-value (probability that the values are independent of one another) for each sample using SciPy. The correlation is relatively weak within each sample, however I also combined the samples, essentially taking replicates for each amplicon. The combined samples had a coefficient of 0.22 and a p-value of 5.03×10^{-7}, well below p=0.05.

Note that this suggests an explanation for the stability of the N gene in wastewater, since it has a very high GC content relative to the rest of the genome, as seen in Figure 4.3. The GC content of the N gene is 47.2% compared to 38.0% for the whole genome. This result is supported in work by Cao et al., who mapped the architecture of SARS-CoV-2 RNA in the viral capsid [8]. They were able to plot the predicted number and strength of interactions across the genome, which showed the highest peak in the N gene.

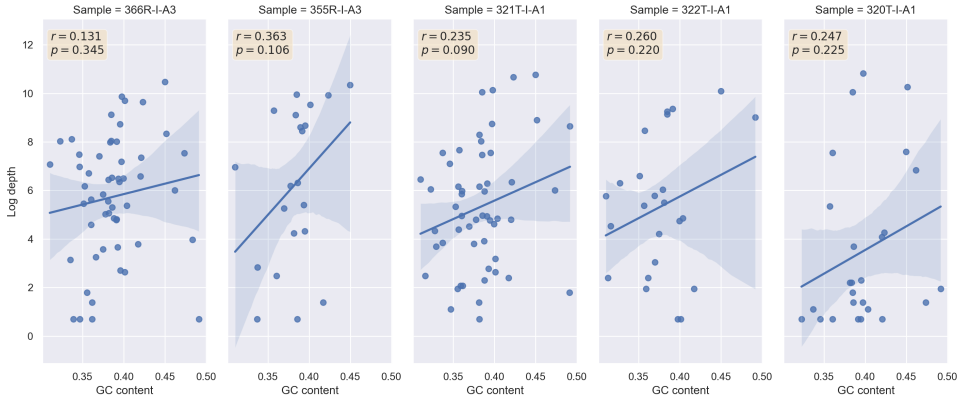
4.2 Selecting an amplicon which can distinguish many VOCs

Mutations in different VOCs are not evenly distributed across the genome. The RNA-dependent RNA polymerase (RdRp) is relatively stable across different variants and even across different viruses. Conversely, the spike protein is a hotbed for novel mutations. RNA viruses have famously high mutation rates, partially due to the low fidelity of the RdRp. The high mutation rate confers a few evolutionary benefits, including a rapid capacity to adapt to new hosts and evade immunity [12]. This is especially important for the spike protein, since it interacts with the host-specific receptor proteins and is the primary target for vaccines (as well as a key target for natural immunity).

Because of the regional clustering of mutations on the genome, it is important to identify a region which is highly discriminatory of multiple VOCs. In order to quantify this, I wrote a script to compare the amplified region of each VOC/VOI, for each ARTIC amplicon. I plotted how many of the variants had a unique profile for each amplicon. This number



(a) Including amplicons which had no coverage.



(b) Excluding amplicons which had no coverage.

Figure 4.2: Log depth vs. GC content of each amplicon. Pearson correlation coefficient and p-value shown for each plot. There is a positive correlation, possibly due to increased secondary structure. 4.2a includes amplicons which had no coverage and 4.2b omits them.

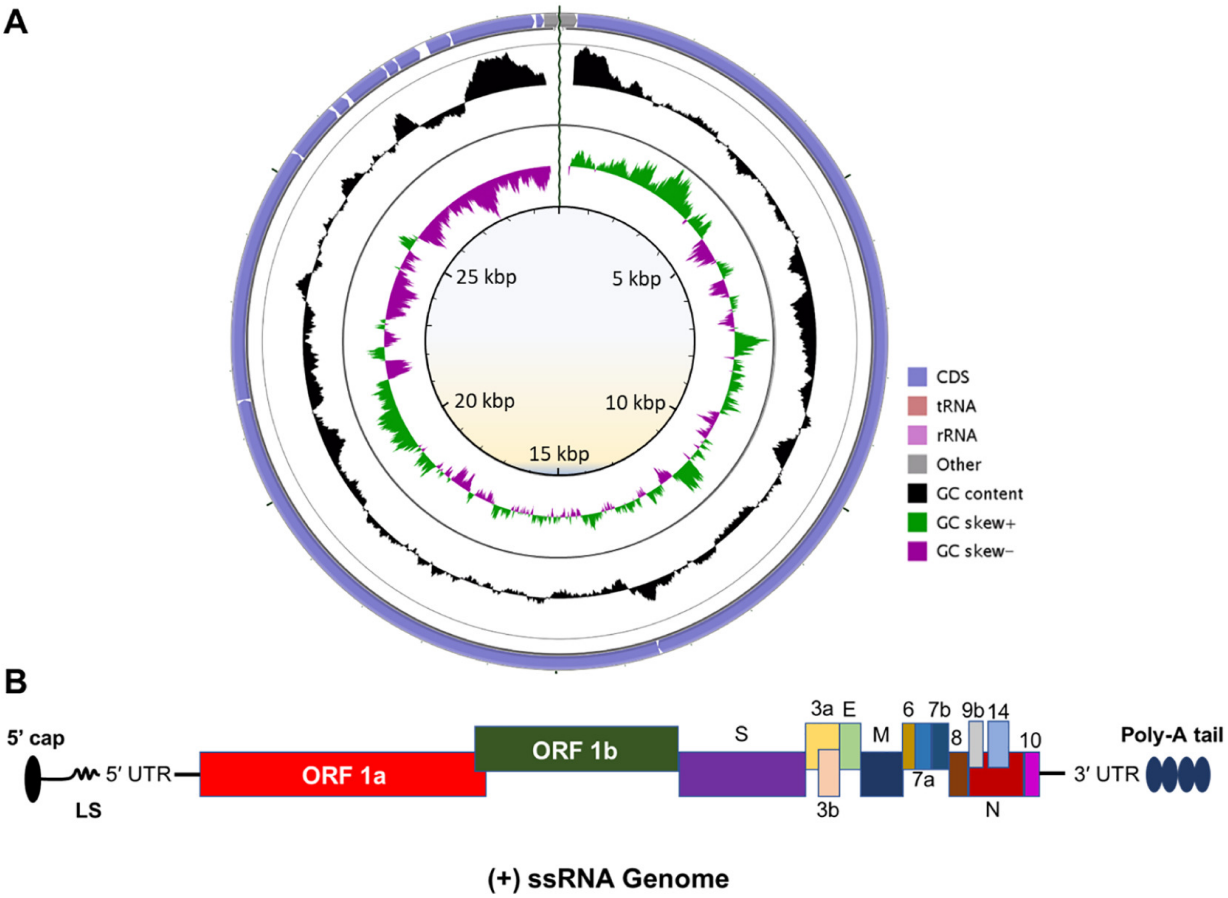


Figure 4.3: GC content across the SARS-CoV-2 genome. Note that the N gene (top, just left of centre) has a very high GC content relative to the rest of the genome which could increase its stability in wastewater. Taken from [31].

represents how many variants can be distinguished by sequencing that region. At the time there were 4 VOIs and 4 VOCs for a total of 8 variants.

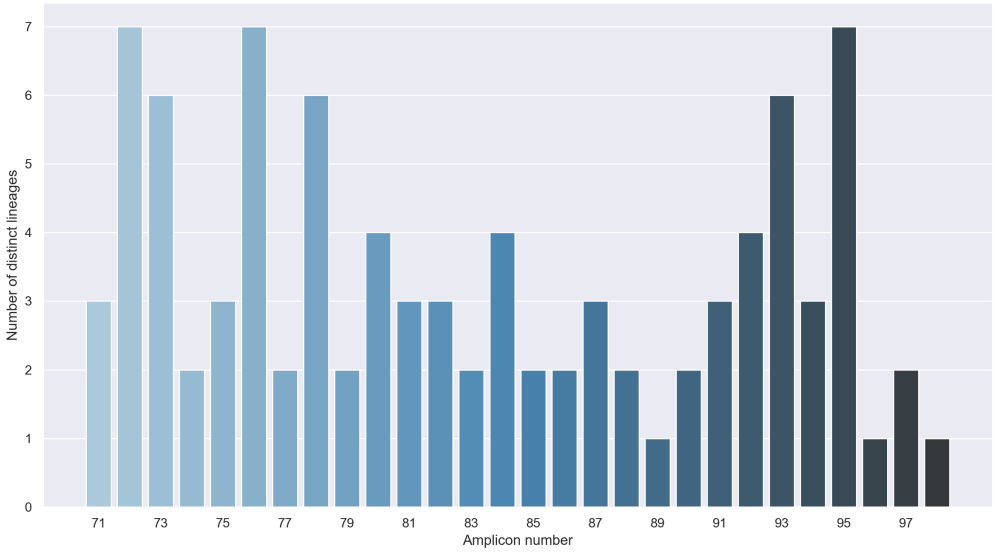


Figure 4.4: Number of variants which can be distinguished by each amplicon. Amplicons which cover the RdRp were omitted since mutations over the RdRp are sparse. Three amplicons (72, 76, 95) are able to distinguish 7/8 of the variants which were tested.

As seen in Figure 4.4, there were three amplicons which could distinguish 7/8 of the variants. Amplicon 76 was an attractive candidate because it covers the RBD of the spike protein which is known to be physiologically important. Additionally, we identified that moving the forward primer of amplicon 76 slightly upstream allowed it to cover position 417 of the spike which allowed it to distinguish all 8 variants. Amplicon 95 was also an attractive candidate because it covers the N gene which (as discussed) is typically well conserved in wastewater.

4.3 Designing an amplicon to target multiple viruses

There are many human diseases caused by coronaviruses including SARS, MERS, SARS-CoV-2, and several "common colds". During COVID-19, wastewater surveillance of SARS-

CoV-2 proved to be a very useful tool for monitoring the spread of SARS-CoV-2. Going forward it may be useful to be able to use wastewater for monitoring the spread of coronaviruses *in general*.

While it would be useful to be able to sequence the spike of a wide variety of coronaviruses, the genetic diversity of the spike makes it impossible for a set of primers to bind to a diverse set of coronaviruses. Designing amplicons for multiple viruses is an inverse problem to selecting an amplicons for distinguishing viral lineages. Previously, we needed a region that was sufficiently different across different lineages so as to be able to distinguish them. Now, there are so many mutations between different species that instead we are looking for a region which is sufficiently similar that the same PCR primers can bind to all the viruses. As noted in the previous section, the RdRp tends to be quite stable, and so could be a good target. In fact, there has previously been an amplicon designed for 11 coronaviruses which targets the RdRp [44]. The primer pair was designed before SARS, let alone SARS-CoV-2, as well as many bioinformatic methods, so we decided to design a new pair. In particular, we wanted a primer pair which is capable of amplifying a wide range of betacoronaviruses including SARS, SARS-CoV-2, and common cold viruses.

Because there is so much diversity within even closely related viruses, it is usually not possible to find a 20 bp site which is perfectly conserved across all viruses in a given set. A way to overcome this is to design primers with *degeneracy*, that is to allow some of the positions on the primer to be a mixture of nucleotides. When degenerate primers are synthesized, all of the different possible combinations are included. The degeneracy of a primer is the number of possible combinations for that degenerate primer. For example, the sequence `WAB` has degeneracy 6 because W (A or T) has degeneracy 2, A has degeneracy 1, and B (C, G, or T) has degeneracy 3 and $2 \times 1 \times 3 = 6$. Because the number of possible combinations is multiplicative, including a large number of degenerate positions in a primer can vastly reduce the yield of the particular primer which binds to each virus as well as the specificity of the set. Therefore it is important to design primers with low degeneracy.

I used the tool PrimerProspector [50] to identify a series of candidate degenerate primers which bind across most betacoronaviruses. I created a tool to call the core functions of PrimerProspector, and then select the best candidate primers. One of the difficulties was updating PrimerProspector to use Python3. Some of the updates I had to make to PrimerProspector to make it Python3 compatible are described in appendix A.

The tool automatically identifies the least degenerate primers (which can still bind to all genomes of interest) and provides a shorthand for calling PrimerProspector to simulate sequencing reads using a selected amplicon. Currently, we performed manual filtering of candidates, but the tool could be extended to filter based on other features such as

GC content. It may also be possible to construct phylogenetic trees from the simulated amplicons using each candidate amplicon and tree similarity metric to find the amplicon which most accurately reproduces the phylogeny from whole genomes.

I ran the tool to find a primer pair capable of amplifying a broad range of betacoronaviruses. I downloaded all betacoronavirus genomes on RefSeq, and used the tool to predict primers and select the least degenerate pair spaced 200 bp to 450 bp apart. The pair is shown in Table 4.1 along with the pair from [44]. Interestingly, the primers amplify an almost identical region of the RdRp. The forward primer binds to the exact same location and the reverse primers bind only 6 bp upstream. Notably however, the new primers have significantly lower degeneracy which may lead to more efficient binding to betacoronaviruses and less off-target binding.

Table 4.1: Comparison of primers from [44] with the least degenerate primers found using the new tool.

	Stephensen et al.	Newly generated
Forward primer	ACTCARWTRAATYTNAAATAYGC	CARATGAATYTKAARTATGC
Reverse primer	TCACAYTTWGGATARTCCCA	TTAGGRTARTCCCAACCCAT
Degeneracy	136	20

Chapter 5

Deriving lineages from wastewater sequencing data

Up to this point, the central problem I have been trying to address is how to determine the relative abundances of known lineages in a sample. Because SARS-CoV-2 is so well-characterized, this is usually the most important question to answer. We usually want to track the emergence of variants which have been found in other parts of the world but are just making their way into Canada. However, it would be very useful if did not need to rely on existing lineage definitions. The known lineage definitions are usually not specific to regional variants, so the predictions which rely on them are usually slightly inaccurate. Additionally, it is difficult to track the emergence of new variants in wastewater if we are only looking for variants which have already appeared in clinical samples. What if we could have identified Omicron before it finished evolving? Finally, it would be very useful to be able to use wastewater as a source of discovering all the variants of a poorly characterized virus, such as a pathogen which has just recently been discovered.

As with abundance estimation, the challenges for lineage discovery are two-fold: the data is usually composed of a mixture of different lineages and there are often substantial gaps. This means that taking a consensus genome from a wastewater sample will usually have large gaps (which are called to the reference) and may contain mutations from multiple lineages. However, the frequencies of mutations which belong to a particular variant will be correlated both within a sample and across multiple samples. Therefore, if we find mutations which are correlated over many samples we may deduce that they belong to the same lineage. I discovered a nice way to solve this problem mathematically which I will describe after a brief introduction to representing data as vectors.

5.1 Casting data as vectors

A vector is an ordered collection of data. Vectors look like this:

$$v = \begin{bmatrix} 5 \\ -2 \\ 3.7 \end{bmatrix} \tag{5.1}$$

The number of elements in a vector is called the dimension. In 5.1, the dimension of v is 3.

Vectors can be used to represent data. For instance, if an image has n pixels, we can represent it as the RGB values of each pixel in an $n\times3$-dimensional vector.

We can also represent wastewater sequencing samples as vectors, where each value corresponds to a possible mutation. The values range between 0 and 1 where 0 means we did not observe the mutation and 1 means we observed it at 100%.

5.2 Imputing data from amplicon dropouts

All of our data vectors must have the same dimension, so we must have a value for the frequency of each mutation. This is sometimes problematic because we sometimes get dropout of amplicons and have no information about the mutations that those amplicons span. To solve this, I used a technique called data imputation where a model can guess what the missing data should be. The simplest form of data imputation is to just guess the same value for each missing point, for instance 0. This can be an issue when you have an amplicon which frequently drops out since it will fill in that value over many samples.

A more sophisticated technique for data imputation uses k-Nearest Neighbours (KNN). KNN imputation has been used to fill in gaps in DNA microarray data [46]. KNN finds the k most similar data points (samples) and fills in the missing mutations with the average of their frequencies. This means that the missing mutation frequencies will tend to get filled in with values from samples with similar lineage abundances. I used the scikit-learn implementation of KNN imputation and used the default value of $k = 5$.

5.3 PCA and NMF

As discussed, we cannot assume that each variant will be present at the consensus level. Additionally, the frequencies of mutations in a sample are highly variable, so each sample may not show an accurate snapshot of the lineages it contains. However, we know that on average, the frequency of a given mutation in our sample will just be the sum of the abundances of the lineages which contain that mutation. This means that frequencies of mutations which are contained in the same lineage will be correlated across different samples. Our task then is to determine which mutation frequencies tend to be correlated (increase and decrease together) across all samples, and guess that those mutations form a lineage.

There is a technique called principal component analysis (PCA) which solves a version of this problem. PCA is widely used in machine learning, including microbiome analysis [6]. It finds which values are the best correlated and groups them together. It then finds the next best group which is orthogonal to the first group. Then the next best, and so on. Loosely, each of these groups of values (mutations) can be interpreted as a lineage. There are a few problems with the assumptions made by PCA which are addressed by a similar technique called non-negative matrix factorization (NMF). NMF has been used to find similar types of patterns in image data [23]. The three main advantages of NMF over PCA are:

- NMF requires the learned values to be positive. This is helpful since the samples are positive combinations of mutations in each lineage and so the learned vectors will tend to correspond to lineage definitions.
- NMF does not require the learned vectors to be orthogonal. Essentially, this means that the learned vectors can correspond to similar lineages such as different omicron lineages.
- NMF minimizes a mixture of the ℓ_1 and ℓ_2 differences of the reconstructed samples. As discussed in the Alcov chapter, incorporating ℓ_1 loss leads to sparse solutions. Since there are tens of thousands of possible mutations and each lineage only contains tens to hundreds, the lineage definitions should be sparse.

5.4 A framework for finding conserved lineages

The strategy I devised for extracting lineage definitions from our wastewater samples is as follows:

1. The samples are run through Gromstole, a tool for extracting mutation frequencies from SARS-CoV-2 reads.
2. The frequencies of each mutation are recorded for each sample, where mutations with a coverage of less than 20 reads are omitted.
3. The samples are cast as vectors where each entry corresponds to the frequency of each of the observed mutations.
4. Missing data in each of the sample vectors is imputed using KNN imputation with $k = 5$.
5. NMF is run on the samples to find n components where n is the number of lineages which are thought to be present.
6. The learned NMF vectors are normalized so that the highest value of each is 1.
7. For each lineage vector, all mutations which have a corresponding value of at least 0.25 are included in that lineage's definition.
8. The mutations are applied to the reference genome of SARS-CoV-2 to create fastas which can be fed into downstream phylogenetic analysis.

5.5 Discovering SARS-CoV-2 VOCs

5.5.1 Finding lineages in simulated reads

A group at McGill University created a simulated set of wastewater reads, available at (https://github.com/suskraem/ww_benchmark). The methodology for creating the reads and the proportions of different variants in each samples were not released. We were informed that the dataset contains simulated reads from four major SARS-CoV-2 lineages (BA.1, BA.2, Delta, a "deltacron" recombinant) as well as a synthetic lineage which contains random mutations. The dataset contains 100 samples, including different combinations of the five lineages, some with simulated amplicon dropouts.

I ran the method on the 100 simulated samples and looked for 5 NMF components (corresponding to five lineage definitions). The five predicted lineages were run through Nextclade [4] since Pangolin struggles with recombinant lineages. As expected, the five predicted lineages were predicted to be BA.1.18 (BA.1), AY.4 (Delta), BA.2.3 (BA.2), B (undetermined synthetic), and XS (Deltacron). Accession numbers for all genomes which were used to simulate the reads were provided on GitHub. I downloaded the genomes from GISAID and ran them with the predicted genomes through Nextclade. Nextclade predicted a range of BA.1, BA.2, and Delta sublineages. All deltacrons were predicted to be XS which corresponded to the predicted lineage. The synthetic genome was not included since it was not uploaded to GISAID. I downloaded the alignment from Nextclade and built a tree using Seaview [15]. The predicted lineages are highlighted in yellow and clearly cluster with the 4 major lineages, as shown in Figure 5.1. The predicted lineage, "lineage4" clusters distinctly alone, as predicted for a synthetic genome.

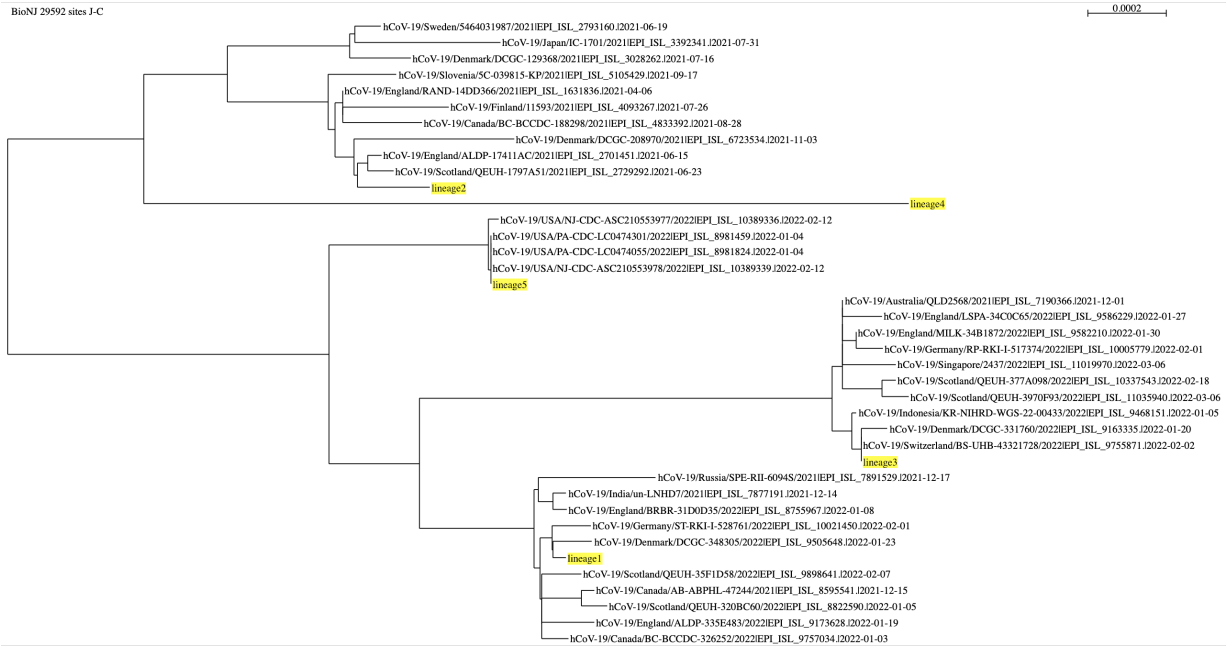


Figure 5.1: Phylogenetic tree showing how the predicted lineages cluster with the isolates that were used to simulate the reads. Predicted lineages are highlighted in yellow.

5.5.2 Finding major VOCs across all samples

I ran the method on a set of 1026 samples from our routine sequencing of Ontario wastewater. I used NMF with 3 components which correspond to 3 lineages. I looked for three because there were three major lineages which were dominant in Ontario over that time period. In testing the model I found that looking for too few lineages tends to predict hybrid, "merged" lineages, and looking for too many tends to predict incomplete pieces of sublineages.

I applied the mutations to the SARS-CoV-2 reference genome to create a fasta with each of the predicted lineages. I ran Pangolin [34] on the predicted lineages to assign a lineage to each of them. Pangolin also runs a tool called Scorpio which assigns lineages and provides a confidence score for the particular lineage call. The predicted lineages for Pangolin and Scorpio along with the Scorpio support values are shown in Table 5.1.

Table 5.1: Lineage assignments for each of the predicted lineages from all samples. The predicted lineages were all highly abundant in Ontario. Both Pangolin and Scorpio agree on all three and Scorpio indicates strong support for the predictions.

Isolate	lineage	scorpio call	scorpio support
lineage1	BA.2	Omicron (BA.2-like)	0.97
lineage2	BA.1.1	Omicron (BA.1-like)	0.88
lineage3	B.1.617.2	Delta (B.1.617.2-like)	0.92

The three predictions were BA.2, BA.1.1, and B.1.617.2. All three of these make sense. B.1.617.2 is the parent lineage for all delta sublineages which were dominant in Ontario before being replaced by Omicron (BA.1.1). Eventually BA.2, another Omicron sublineage, replaced BA.1.1. Together, these give an accurate snapshot of the most significant lineages in Ontario between Fall 2021 and Spring 2022.

I downloaded the frequency of each mutation for the three predicted lineages from outbreak.info and compared them to the learned mutation values. Figure 5.2 shows the values for the spike mutations next to the frequency with which those values are observed in clinical sequences. In general, they agree well. The lineages on outbreak.info do not include synonymous mutations or insertions. Some mutations may represent legitimate local variation (like S:A222V), albeit with poor coverage and so a small sample size. The confidence and accuracy of the predictions decreases with each consecutive lineage which makes sense because the components in NMF are ranked according to their relative importance.

Figure 5.3 plots the values for the N gene. All mutations which are predicted to be significant agree with the outbreak.info data, including the variable presence of N:G215C

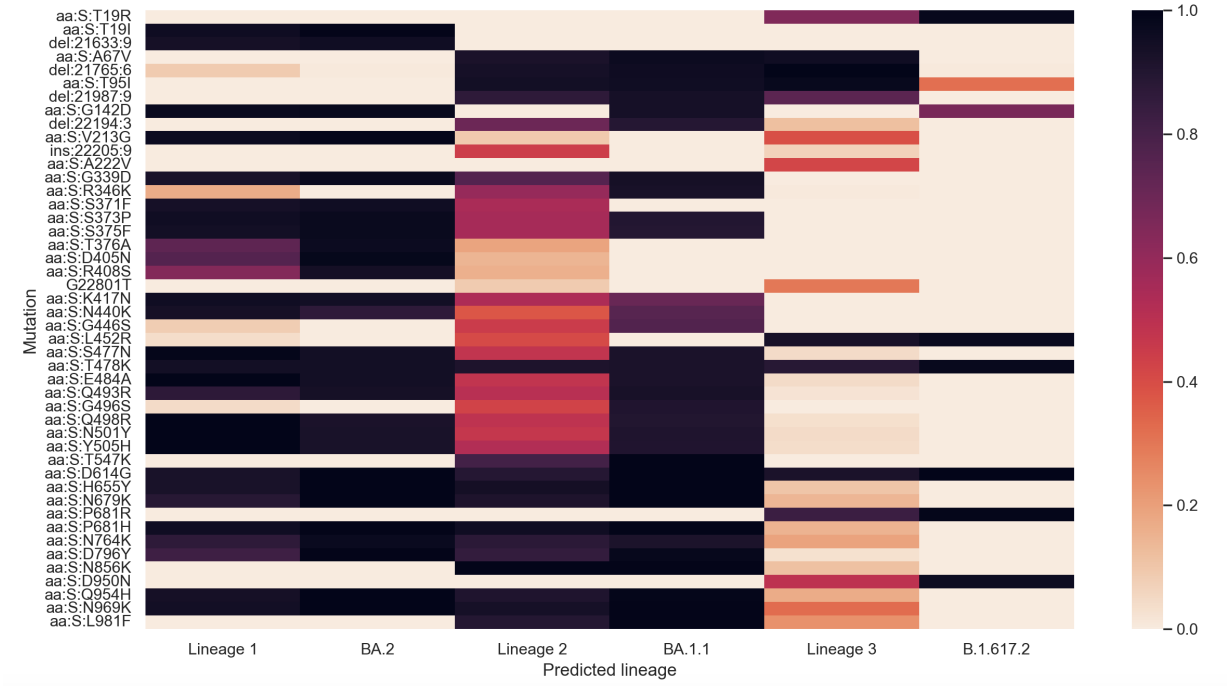


Figure 5.2: Heatmap showing the learned spike mutation values of the predicted lineages next to the frequency with which those mutations are observed in the corresponding lineage according to outbreak.info.

in Delta. While all values greater than 0.25 are simply included in the lineage definitions, stricter cutoffs could be used to determine which mutations are the most diagnostic *in wastewater* which would exclude N:G215C since it is not always contained in Delta sequences. The N gene carries fewer mutations than S and has much better coverage which probably leads to increased accuracy in the predicted lineages. Lineages 1 and 2 are identical since BA.2 and BA.1.1 have the same version of the N gene.

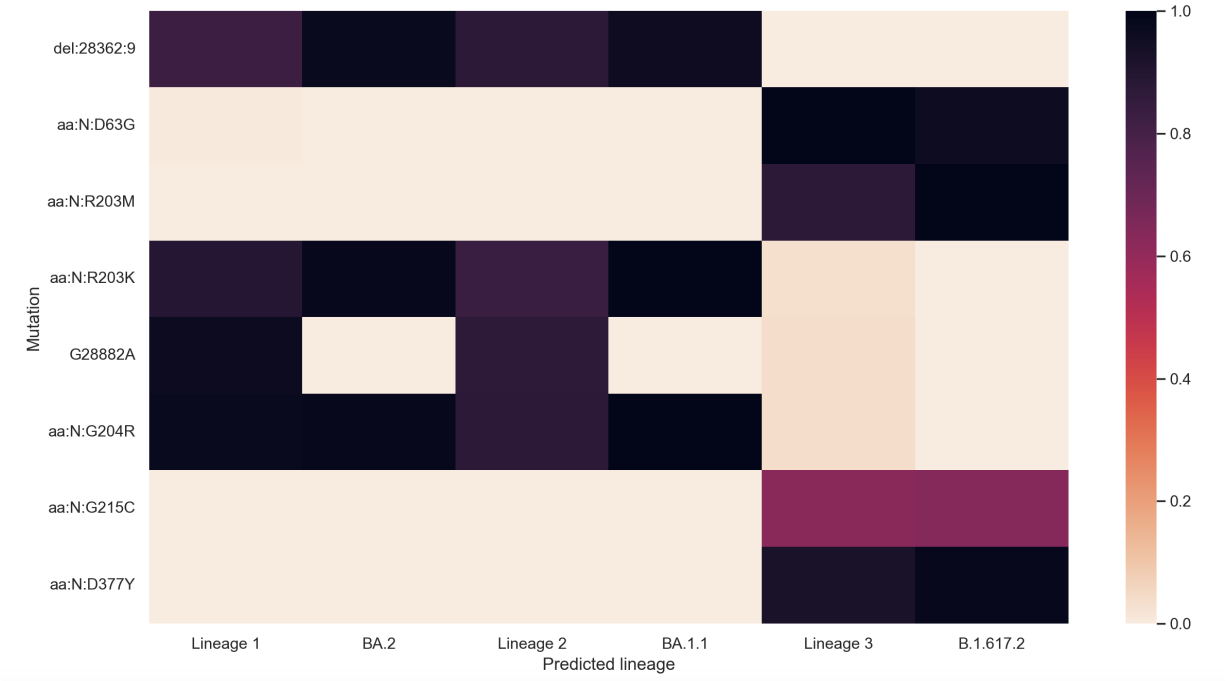


Figure 5.3: Heatmap showing the learned N gene mutation values of the predicted lineages next to the frequency with which those mutations are observed in the corresponding lineage according to outbreak.info.

5.5.3 Finding VOC subvariants in a single run

I also ran the method on a single run with samples from across Ontario in late June. The lineage predictions are shown in Table 5.2.

Using samples from a single run, the method was able to accurately predict the two major Omicron lineages in Ontario at the time, BA.2 and BA.5. Surprisingly, the method

Table 5.2: Lineage assignments for each of the predicted lineages from a single run in late June. The method accurately predicts BA.5 and BA.2 including sublineages which have been found in clinical samples with very high scorpio support.

Isolate	lineage	scorpio call	scorpio support
lineage1	BA.5.2.1	Omicron (BA.5-like)	0.97
lineage2	BA.2.12.1	Omicron (BA.2-like)	0.97

was able to pick up all mutations with enough accuracy to predict specific sublineages of the two. Both of these sublineages have been identified in Ontario at the time according to outbreak.info, although the prevalence of BA.5.2.1 in clinical cases is lower than we are likely seeing in wastewater.

I plotted the predicted values of the spike mutations for the two lineages, which are shown in 5.4. BA.2 and BA.5 are very similar which can pose a challenge for the method but it was able to identify distinguishing mutations such as S:F486V.

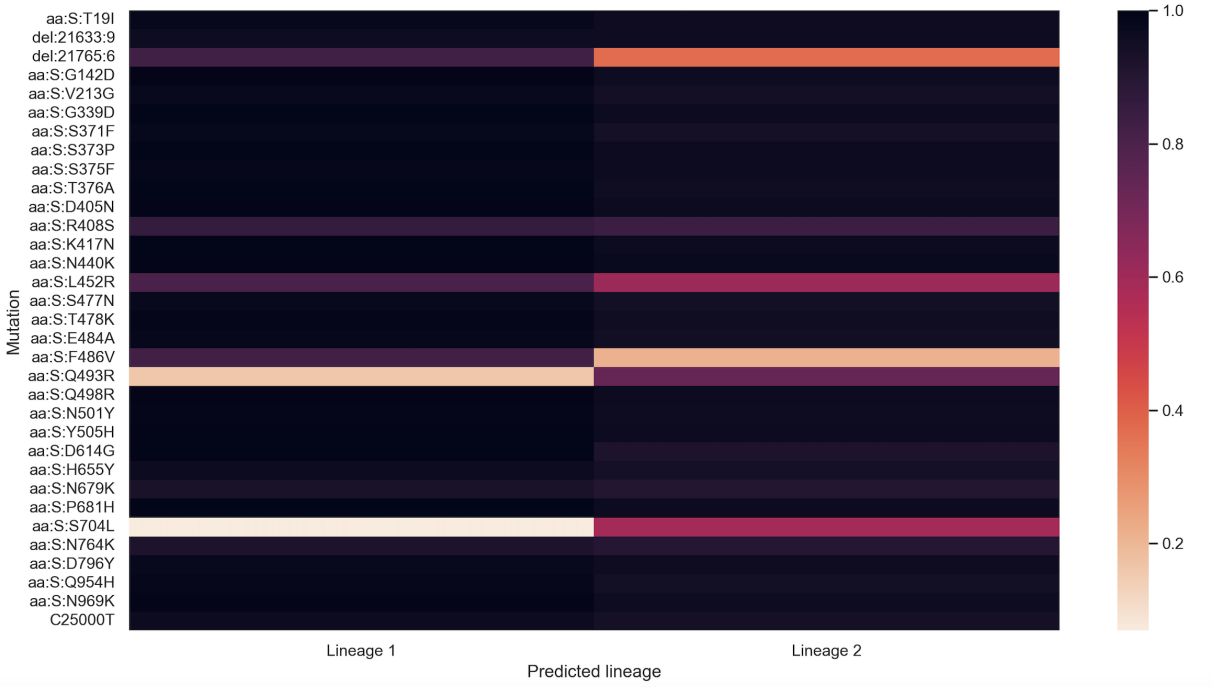


Figure 5.4: Heatmap showing the learned spike mutation values of the predicted lineages in the single run

It is worth remembering that sublineages are notoriously difficult to call in wastewater, even with known mutations to look for. This is because they are usually only differentiated by a few mutations and the frequency of those mutations can vary wildly from sample to sample. Here, we are looking at all mutations and discovering lineages with no prior knowledge of what to look for. Surprisingly, we are able to identify the mutations so accurately that not only can we discover the major lineages which are present in a run, but we can actually identify the specific sublineages which are most abundant. The accuracy likely comes from the ability of the method to pool information from multiple samples which works to smooth some of the noise within individual samples.

Chapter 6

Analysis of Tomato Brown Rugose Fruit Virus

The COVID-19 pandemic has fuelled many innovations in genomic epidemiology, especially in wastewater surveillance. There is now considerable interest in extending some of the techniques used to analyze SARS-CoV-2 in wastewater to other viruses in wastewater as well as other environmental samples. There are a number of human viruses which are being investigated. Polio has been detected in wastewater even before COVID-19 [19]. There are also significant seasonal viruses like influenza or other coronaviruses which cause the "common cold" (along with rhinoviruses, adenoviruses, and enteroviruses) .

In addition to monitoring human viruses, we can also detect plant viruses in wastewater. PMMoV is commonly used as a baseline to compare changing Ct values of SARS-CoV-2 against. A related virus, the tomato brown rugose fruit virus (ToBRFV) is currently wreaking havoc on tomato plants across the world. The virus is particularly harmful to tomatoes grown in hydroponic greenhouses where the virus quickly spreads through frequent touches and water systems [17].

Given that the ToBRFV is so pathogenic, it is perhaps surprising that it is highly abundant in wastewater. ToBRFV was found to be highly abundant in California wastewater [10]. We discovered ToBRFV in a metagenomic shotgun sequencing run of Ontario wastewater while looking for SARS-CoV-2, as shown in Figure 6.1.

The global interest in ToBRFV, the existence of distinct clades [49], and its presence in municipal wastewater motivated us to apply the SARS-CoV-2 monitoring tools to study it.

Figure 6.1: Plot produced by Kaiju [30], showing all plant-infecting viruses discovered in the shotgun sequencing sample. ToBRFV (left) is highly abundant. The sample was taken from a WWTP in Waterloo in late 2021.

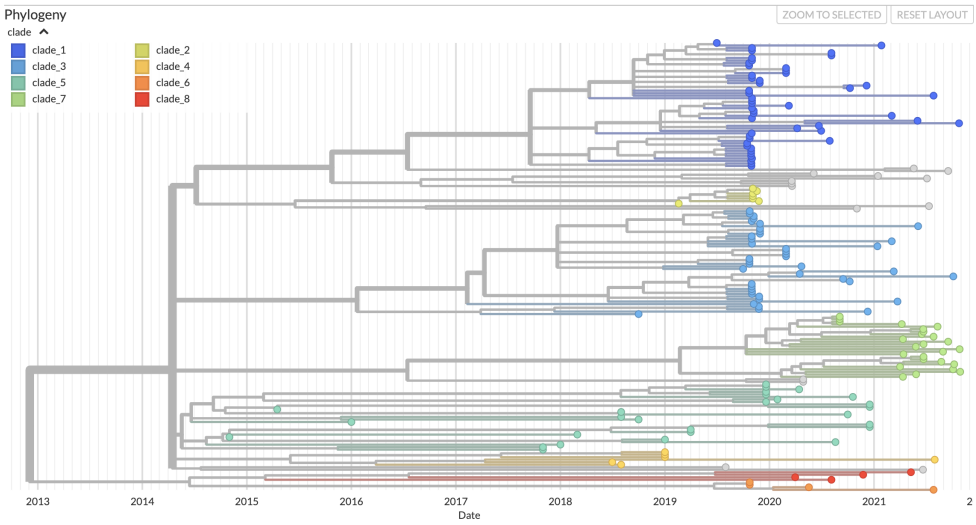
6.1 ToBRFV is a global pandemic of a different kind

ToBRFV is a Tobamovirus in the same genus as Pepper Mild Mottle Virus and Tobacco Mosaic Virus. It has a very compact genome which is only about 6 kb and encodes three genes: an RNA-dependent RNA polymerase, a movement protein, and a coat protein. ToBRFV was first discovered in Israel in 2014 and has since spread across the world [16]. There are a large number of sequenced genomes from Europe, but it has also been sequenced in North America, South America, Africa, and Asia. Importantly, there are at least 8 distinct clades (major lineages) which cluster geographically [49], suggesting widespread local transmission (Figure 6.3). ToBRFV often causes brown rugose symptoms but can also cause yellowing of fruit and necrosis, as shown in Figure 6.2.

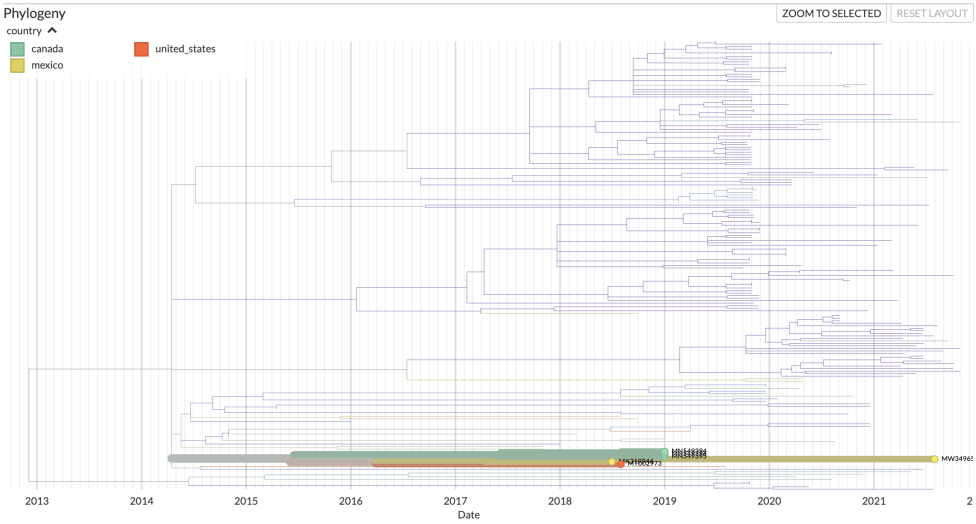
Figure 6.2: The effects of ToBRFV on tomato plants, taken from [11]

6.2 Generating a set of amplicons

Following the example set by SARS-CoV-2, we first set out to create an amplicon tile which could amplify the genome in roughly 400 bp amplicons. I used Primal Scheme [37], the same tool used by the ARTIC network to generate the SARS-CoV-2 amplicon panel.



(a) The eight distinct clades.



(b) North American genomes cluster together in clade 4.

Figure 6.3: The global phylogeny of ToBRFV from Nextstrain [49].

I followed the instructions for using the Primal Scheme CLI. I downloaded all ToBRFV genomes from the Nextstrain build [49] and checked them for sequence similarity using the Clustal Omega web interface [14, 28]. I then used the `primalscheme multiplex` command with the default parameters which produced a set of 20 primer pairs.

The primer scheme was ordered and tested on wastewater from different sites in Ontario. Figure 6.4 shows the coverage from one of our runs. The primer set works extremely well. Whereas the ARTIC amplicon set for SARS-CoV-2 routinely has wide gaps in coverage, all of our ToBRFV runs had 100% coverage of our amplicons. Furthermore, the runs had fairly consistent depth across amplicons, and consistent depth of each amplicon between samples. This suggests that the differences in amplification are due to RNA structure or PCR bias as opposed to random degradation. The disparity between SARS-CoV-2 coverage and our ToBRFV coverage is probably two-fold. First, the Ct values are much higher in the wastewater. Second, the RNA is probably less degraded, possibly because it is protected by a more robust capsid than SARS-CoV-2 [22]. It is again striking just how abundant ToBRFV is in wastewater, given that it is typically regarded as a very harmful pathogen.

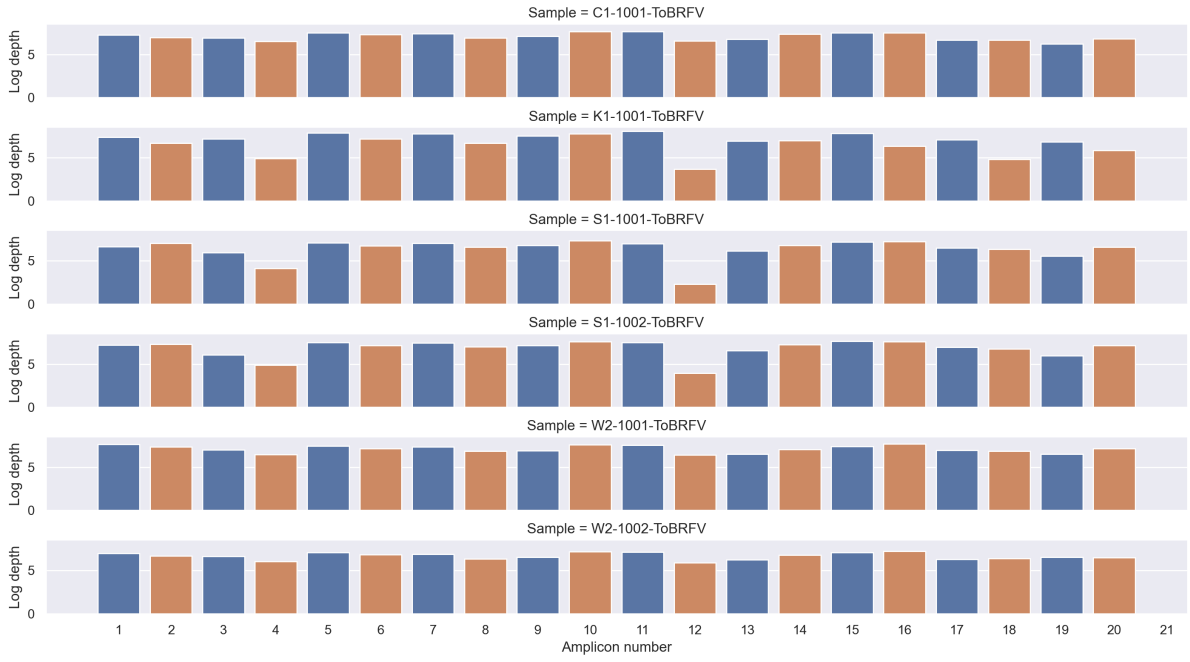


Figure 6.4: Log read depth for each ToBRFV amplicon. All amplicons had non-zero coverage in all samples.

6.3 Phylogenetic placement of consensus genomes

As with SARS-CoV-2, a first step for analysis of sequenced viral RNA is to treat it as a single isolate and generate a consensus genome. After aligning the reads to the reference, NC_028478 [40], I used samtools and bcftools to create the variant call files (VCFs) and accompanying fastas. A popular method for assigning a lineage to a SARS-CoV-2 isolate is Pangolin. In its standard mode, Pangolin uses PangoLEARN [34], a decision tree which uses the presence/absence of mutations to call lineages. Pangolin can also run UShER [47] which quickly places the sequence in the broader SARS-CoV-2 phylogenetic tree and then calls a lineage based on its placement.

I used UShER directly to place our samples from one run in the tree from the Nextstrain build. Interestingly, the samples formed a unique clade which was fairly close to the root (Figure 6.5). This may be partially due to the diversity in each sample leading to mixed mutations which largely get called to the reference base, creating consensus genomes that contain few mutations.

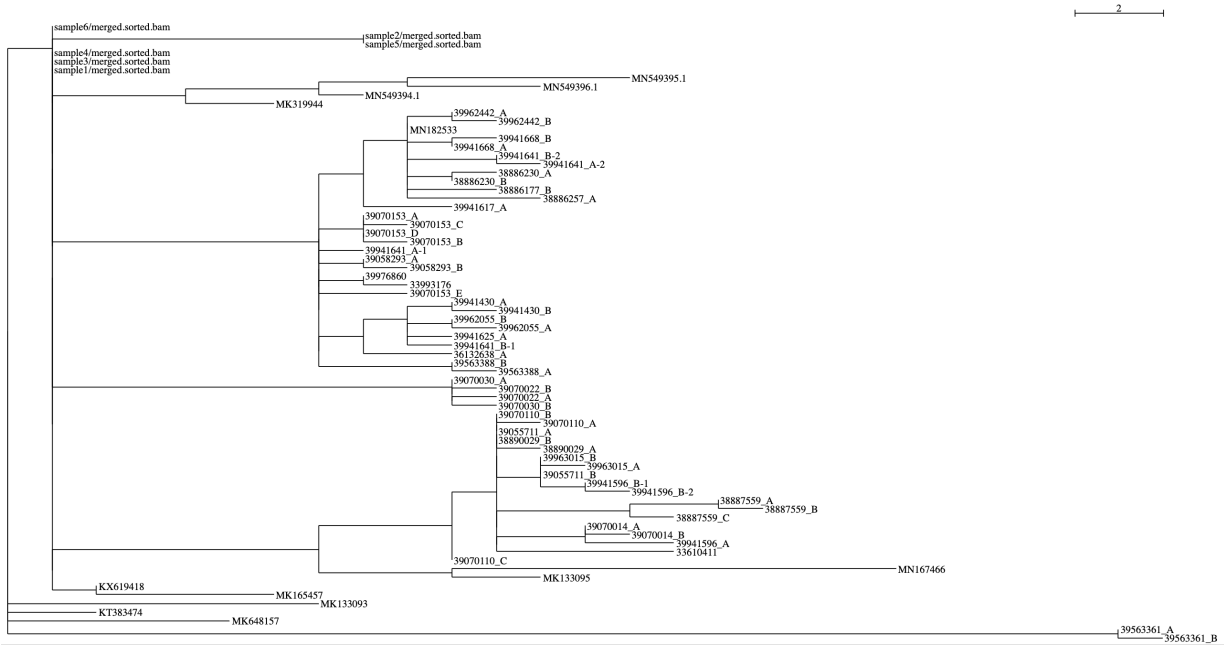


Figure 6.5: ToBRFV tree with consensus genomes from a wastewater placed as `sample[n]/merged.sorted.bam`. The wastewater samples (top of figure) cluster together, mostly near the root.

We had co-op students in web development spend a month with our lab to develop a web tool. I was largely responsible for managing their project and we decided to have them create a tool for placing ToBRFV genomes into the tree. The tool runs UShER and then renders the tree along with information about each node. Figure 6.6 shows an example output. The tool is currently hosted at (https://tobrfv-lineages.netlify.app/) although the site may not be stable going forward.

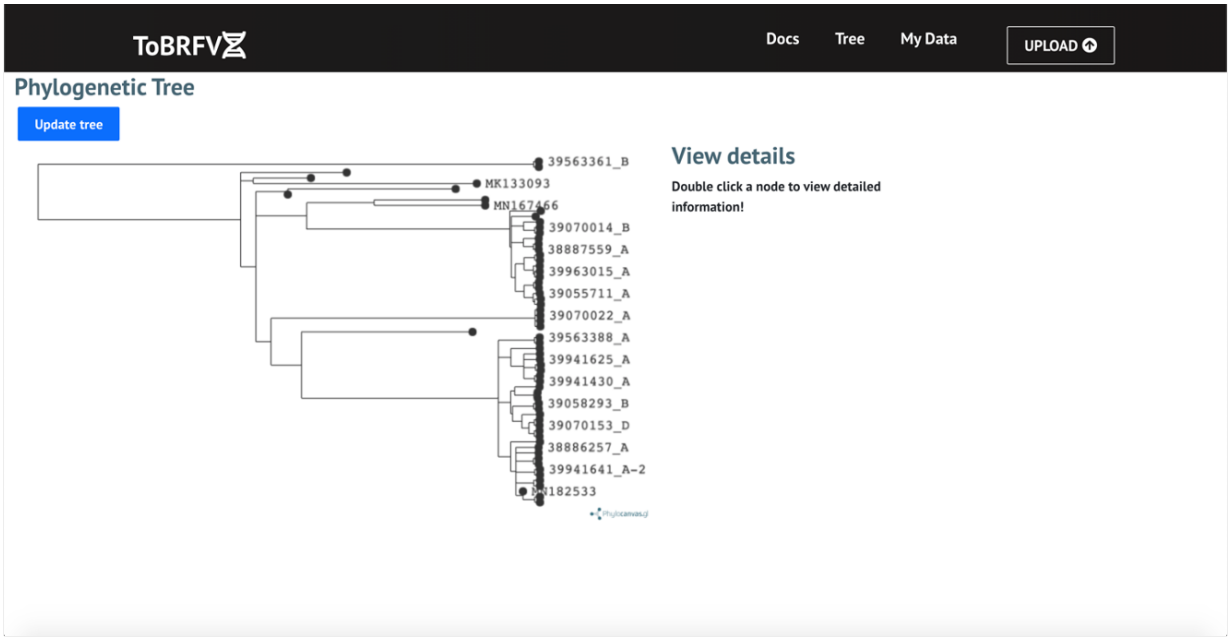


Figure 6.6: A screenshot of the output of our ToBRFV placement web tool.

6.4 Altob: prediction of clade abundances

The final step for extending our SARS-CoV-2 methods to ToBRFV was to build an Alcov equivalent for ToBRFV. I called the ToBRFV version of Alcov, Altob. Altob is publicly available on GitHub at (https://github.com/Ellmen/altob). In order to adapt Alcov, I had to change the hard-coded constants such as the genome sequence for the reference. I also had to update the genes and gene locations. There is an ORF which is completely contained in the RdRp gene on RefSeq which I removed because it can cause ambiguity in the mutation names. I also had to update the lineage (or clade) definitions. This was

challenging because there are no public sources for clade-specific mutations such as outbreak.info or cov-lineages for SARS-CoV-2. Delaney Nash found that the clade specific mutations were listed in Nextstrain and compiled them manually. Unfortunately, most of these mutations are inaccurate due to an issue with processing mutations in their pipeline. I confirmed the error in a private correspondence with the maintainers. Instead, Delaney performed a multiple sequence alignment and manually determined a curated list of conserved mutations within each clade. In the future, it would be more efficient (and likely accurate) to generate these lists automatically.

I ran Altob on the tiled amplicon wastewater runs. The results were fairly consistent across multiple runs. Figure 6.7 shows the predicted clade abundances from one run. Essentially Altob predicts about 50% clade 4 and the rest to be an unknown clade. The clade 4 prediction makes sense since the only known Canadian isolates on Nextstrain all cluster together within clade 4.

6.5 Predicting a lineage definition

I ran the method described in chapter 5. I used all samples from all ToBRFV runs and looked for two lineages. I anticipated that the two lineages would correspond to the clade 4 and "undetermined"" lineages, however both were very similar members of clade 4. I ran the predicted lineages through BLAST looking for all ToBRFV genomes in the nt database. The results for the two predicted lineages were similar. For the first lineage, I sorted the results by percent identity. The highest hit, with a percent identity of 99.91%, was a sequence with accession number MN549394.1 [2]. According to GenBank, this sequence comes from an unpublished report titled "First report of Tomato brown rugose fruit virus on tomato in Canada".

This is another exciting result. Whereas trying to align consensus genomes resulted in an unspecific and novel clade near the root, the NMF-based method is actually able to identify a sequence which clusters well with known Canadian isolates. This shows that there has likely been local transmission within Canada since at least 2019 when MN549394.1 was sampled. It also shows that wastewater can be a valuable tool for deciphering specific viral lineages for a large number of viruses - whether they infect humans, plants, or other hosts.

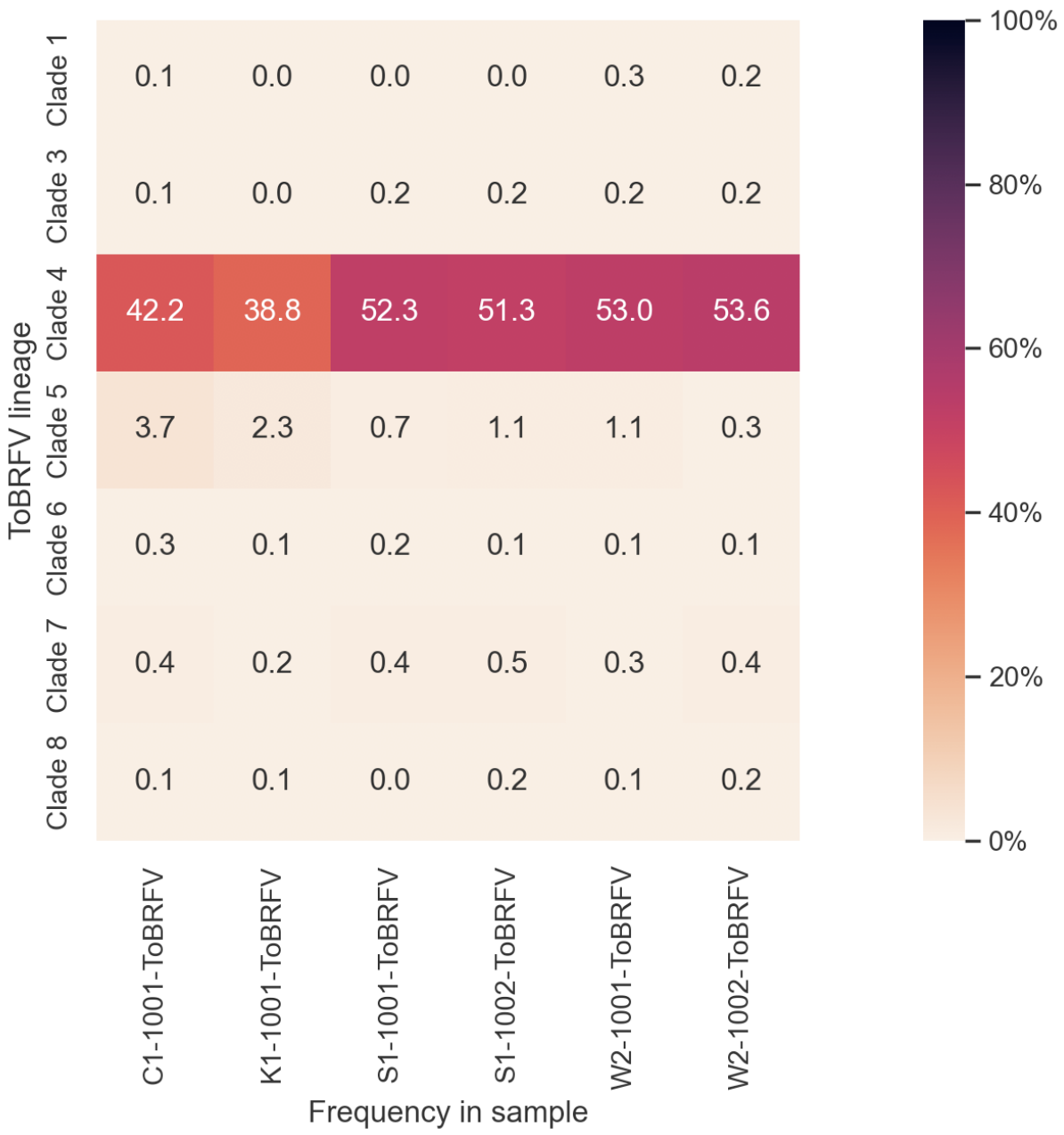


Figure 6.7: Predicted clade abundances from Altob in a wastewater run. Clade 2 was automatically omitted since it was predicted to represent less than 0.1% in all samples.

Chapter 7

Conclusion

Wastewater sequencing is still an emerging field and has progressed rapidly over the past two years. At the beginning of the COVID-19 pandemic, the focus of wastewater sequencing was identifying lineages that could be traced to geographic areas for the purpose of tracking the spread. This was reflected in websites such as Nextstrain where they plotted the breakdown of clades in different countries and showed possible links. With the emergence of VOCs, the task has shifted slightly to tracking the introduction of new variants because of their physiological differences. Being ably to quickly and/or retroactively track emerging VOCs has had a direct impact on public health.

There are a number of questions about viral spread which can be answered by wastewater. The first, and perhaps most obvious, is "How many people are infected?" While it is difficult to answer directly, qPCR data normalized to PMMoV has been used extensively to track increases and decreases in case counts. Because of changes in public health guidance on who should be tested, wastewater is perhaps even more accurate at determining case counts than clinical tests. The next important question, "Which variants of the virus are present?" is a natural fit for sequencing. The question can be broken up into:

1. Which known variants are present?
2. What is the relative abundance of each variant?
3. Are there any new variants?

In this work I mainly focused on (2) and (3). In most cases, (1) is answered by (2), however most abundance estimation methods struggle to confidently call low-abundance

variants. In our first observations of Omicron, we only found a few reads in support of Omicron, but upon further inspection these reads contained two SNPs and a 9 bp deletion. Because Alcov (and other methods) assumed independence of the mutations, the evidence of Omicron appeared to be weaker than when we examined the read pileups.

Chapter 3 introduced Alcov, the tool I created for performing abundance estimation of known variants. Chapter 4 discussed some strategies to get more reliable or generic data using single amplicons instead of amplicon tiles. Chapter 5 explored a new technique for finding novel lineages in the mixed samples. Finally, Chapter 6 applied the above techniques to analyze ToBRFV.

7.1 Summary of findings

7.1.1 Alcov

I created Alcov, a new tool for estimating the relative abundance of known lineages in wastewater sequencing samples. Alcov was one of the first tools for estimating lineages abundance of SARS-CoV-2 in wastewater. The key innovation that distinguishes Alcov from similar abundance estimation methods is its ability to use shared mutations. Most methods avoid including mutations which are found in more than one lineage because they risk "double counting" the mutation. Alcov uses a minimization framework to find an estimate of the lineage abundances which minimize the difference between predicted and observed mutation frequencies. The predicted frequency of a given mutation is the sum of the relative abundance of each lineage which contains that mutation. This means that Alcov is aware that the frequency of a shared mutation may be the result of multiple lineages added up. Shared mutations are particularly relevant when calling sublineages. For instance, there are only a few mutations which distinguish BA.2, BA.4, and BA.5. Alcov is able to use all the shared Omicron mutations to establish the amount of BA.*, then use the shared BA.4/5 mutations to distinguish that proportion from BA.2, and finally use the mutations which distinguish BA.4 and BA.5 to come up with the final estimate. Using many more mutations makes the Alcov predictions more stable.

The original version of Alcov minimized the squared error between predicted and observed mutations. I showed that this essentially reduces to an abstraction of the mean frequency over shared mutations. The newer version of Alcov uses linear programming to minimize the absolute error which I showed reduces to an abstraction of the median over shared mutations. The newer version tends to be more resilient to outliers, which allows it to look for more lineages without making erroneous calls.

7.1.2 Single amplicons

I investigated two cases where single amplicon sequencing may be appropriate.

The first case is when a particular region of the viral genome contains useful information and it is important to ensure coverage of that region. I looked at the relationship between GC content and amplification of different amplicons and suggested that increased RNA secondary structure may lead to some amplicons being better conserved in wastewater. I also performed an analysis of how many variants of interest/concern could be distinguished by different amplicons and identified three high-information amplicons.

The second case is when there are many related viruses that would be worth testing for. I adapted an existing tool to find low-degeneracy primers for amplifying a set of related viruses. I used the tool to find an amplicon for the betacoronaviruses which found an amplicon on the RdRp. I showed that this amplicon is nearly identical to one proposed in the literature, although is much less degenerate.

7.1.3 Finding novel lineages using NMF

I devised a new technique for finding novel lineages in wastewater sequencing data. The technique is based on NMF, which attempts to find lower dimensional representations of data which can still encode enough information to reconstruct the original data. I encoded the samples as vectors where each entry contained the frequency of each possible mutation. I ran the technique on SARS-CoV-2 sequencing data and discovered three potential lineages. I applied the predicted mutations to the SARS-CoV-2 reference genome to create three genomes and analyzed the genomes with Pangolin. The three genomes were predicted to be B.1.617.2 (Delta), BA.1.1, (Omicron), and BA.2 (Omicron). All three had been dominant in some of the samples. I also ran the method on a single sequencing run where it was able to discover sublineages of BA.2 and BA.5.

7.1.4 ToBRFV

I created a new amplicon tile for amplifying ToBRFV RNA. Our lab (not me) used the primers to amplify and sequence ToBRFV RNA in wastewater samples. I oversaw the development of a web tool for placing ToBRFV genomes which was made by co-op students during a brief rotation in the Charles lab. I created Altob, an analogue of Alcov, for determining the relative abundance of different ToBRFV clades in wastewater. Altob predicted the presence of clade 4 which contains known North American sequences. I also

searched for novel lineages in the data which yielded two genomes, one of which matched very closely to a known Canadian sequence.

7.2 Future directions

Wastewater is an incredibly rich source of RNA. Normally, to study a virus, a researcher has to find that virus. In the case of SARS-CoV-2 and ToBRFV this means going out and finding sick people and plants. If a researcher wants to track community trends this means finding a lot of sick people and plants which is difficult and expensive. Fortunately, lots of viral RNA ends up in wastewater and wastewater treatment plants contain a mixture of the wastewater in an entire community.

In addition to improving the methods for abundance estimation, I see two major unexplored research themes: better methods for calling low-abundance lineages and better methods for finding novel lineages.

7.2.1 Calling low-abundance lineages

As I discussed briefly at the start of this chapter, lineage abundance estimation methods are not always designed for calling low-abundance lineages. This can be problematic since one of the biggest advantages of wastewater is its lead time over clinical sequencing. It would be very helpful if wastewater sequencing could function as an early warning system for emerging variants. Unfortunately, we usually have low confidence in the low abundance calls, and by the time a variant is found in higher abundance, it has already been detected in clinical sequences. There need for a tool which can quantify the evidence for a particular lineage in a sample. Such a tool must take into account the increased evidence for multiple mutations on the same read. It must also consider that reads are often duplicated by PCR amplification and so reads which cover the same amplicon are not necessarily independent.

7.2.2 Finding novel lineages

Wastewater sequencing covers RNA from entire communities, and so contains a fairly complete snapshot of the variants which are present. If we could accurately identify novel variants in wastewater, it could help find the next big VOC before it becomes globally dominant. In Chapter 5 I introduced a new method for identifying novel lineages. I

hope that this can form a foundation for future work on the problem. In order to be used practically, the method will have to be adapted to include some kind of confidence measure for the variants it predicts. The expanded method should also be able to determine an appropriate number of components (lineages) to find automatically. It would be especially helpful if there was a method for finding low-abundance lineages (since emerging lineages always start out in low abundance). This is a challenge for NMF and may require re-thinking the loss metric. I am optimistic that methods for determining novel lineages can at least find a home in identifying (1) emerging subvariants of SARS-CoV-2 and (2) local lineages of less-sequenced viruses such as ToBRFV.

References

[1] COVID-19 wastewater monitoring. URL: http://www.ontario.ca/page/covid-19-wastewater-monitoring.

[2] Tomato brown rugose fruit virus isolate Ca1A, complete genome - Nucleotide - NCBI. URL: https://www.ncbi.nlm.nih.gov/nucleotide/MN549394.1?report=genbank&log$=nuclalign&blast_rank=1&RID=E1KJG5U801N.

[3] Shelesh Agrawal, Laura Orschler, and Susanne Lackner. Metatranscriptomic Analysis Reveals SARS-CoV-2 Mutations in Wastewater of the Frankfurt Metropolitan Area in Southern Germany. *Microbiology Resource Announcements*, 10(15), April 2021. doi:10.1128/MRA.00280-21.

[4] Ivan Aksamentov, Cornelius Roemer, Emma B. Hodcroft, and Richard A. Neher. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *Journal of Open Source Software*, 6(67):3773, 2021. Publisher: The Open Journal. doi:10.21105/joss.03773.

[5] Gabriele Berg, Daria Rybakova, Doreen Fischer, Tomislav Cernava, Marie-Christine Champomier Vergès, Trevor Charles, Xiaoyulong Chen, Luca Cocolin, Kellye Eversole, Gema Herrero Corral, Maria Kazou, Linda Kinkel, Lene Lange, Nelson Lima, Alexander Loy, James A. Macklin, Emmanuelle Maguin, Tim Mauchline, Ryan McClure, Birgit Mitter, Matthew Ryan, Inga Sarand, Hauke Smidt, Bettina Schelkle, Hugo Roume, G. Seghal Kiran, Joseph Selvin, Rafael Soares Correa de Souza, Leo van Overbeek, Brajesh K. Singh, Michael Wagner, Aaron Walsh, Angela Sessitsch, and Michael Schloter. Microbiome definition re-visited: old concepts and new challenges. *Microbiome*, 8(1):103, June 2020. doi:10.1186/s40168-020-00875-0.

[6] Evan Bolyen, Jai Ram Rideout, Matthew R. Dillon, Nicholas A. Bokulich, Christian C. Abnet, Gabriel A. Al-Ghalith, Harriet Alexander, Eric J. Alm, Manimozhiyan

Arumugam, Francesco Asnicar, Yang Bai, Jordan E. Bisanz, Kyle Bittinger, Asker Brejnrod, Colin J. Brislawn, C. Titus Brown, Benjamin J. Callahan, Andrés Mauricio Caraballo-Rodríguez, John Chase, Emily K. Cope, Ricardo Da Silva, Christian Diener, Pieter C. Dorrestein, Gavin M. Douglas, Daniel M. Durall, Claire Duvallet, Christian F. Edwardson, Madeleine Ernst, Mehrbod Estaki, Jennifer Fouquier, Julia M. Gauglitz, Sean M. Gibbons, Deanna L. Gibson, Antonio Gonzalez, Kestrel Gorlick, Jiarong Guo, Benjamin Hillmann, Susan Holmes, Hannes Holste, Curtis Huttenhower, Gavin A. Huttley, Stefan Janssen, Alan K. Jarmusch, Lingjing Jiang, Benjamin D. Kaehler, Kyo Bin Kang, Christopher R. Keefe, Paul Keim, Scott T. Kelley, Dan Knights, Irina Koester, Tomasz Kosciolek, Jorden Kreps, Morgan G. I. Langille, Joslynn Lee, Ruth Ley, Yong-Xin Liu, Erikka Loftfield, Catherine Lozupone, Massoud Maher, Clarisse Marotz, Bryan D. Martin, Daniel McDonald, Lauren J. McIver, Alexey V. Melnik, Jessica L. Metcalf, Sydney C. Morgan, Jamie T. Morton, Ahmad Turan Naimey, Jose A. Navas-Molina, Louis Felix Nothias, Stephanie B. Orchanian, Talima Pearson, Samuel L. Peoples, Daniel Petras, Mary Lai Preuss, Elmar Pruesse, Lasse Buur Rasmussen, Adam Rivers, Michael S. Robeson, Patrick Rosenthal, Nicola Segata, Michael Shaffer, Arron Shiffer, Rashmi Sinha, Se Jin Song, John R. Spear, Austin D. Swafford, Luke R. Thompson, Pedro J. Torres, Pauline Trinh, Anupriya Tripathi, Peter J. Turnbaugh, Sabah Ul-Hasan, Justin J. J. van der Hooft, Fernando Vargas, Yoshiki Vázquez-Baeza, Emily Vogtmann, Max von Hippel, William Walters, Yunhu Wan, Mingxun Wang, Jonathan Warren, Kyle C. Weber, Charles H. D. Williamson, Amy D. Willis, Zhenjiang Zech Xu, Jesse R. Zaneveld, Yilong Zhang, Qiyun Zhu, Rob Knight, and J. Gregory Caporaso. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, 37(8):852–857, August 2019. Number: 8 Publisher: Nature Publishing Group. URL: https://www.nature.com/articles/s41587-019-0209-9, doi:10.1038/s41587-019-0209-9.

[7] Stephen P. Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, UK ; New York, 2004.

[8] Changchang Cao, Zhaokui Cai, Xia Xiao, Jian Rao, Juan Chen, Naijing Hu, Minnan Yang, Xiaorui Xing, Yongle Wang, Manman Li, Bing Zhou, Xiangxi Wang, Jianwei Wang, and Yuanchao Xue. The architecture of the SARS-CoV-2 RNA genome inside virion. *Nature Communications*, 12(1):3917, June 2021. Number: 1 Publisher: Nature Publishing Group. URL: https://www.nature.com/articles/s41467-021-22785-x, doi:10.1038/s41467-021-22785-x.

[9] Chi Yu Chan, C Steven Carmack, Dang D Long, Anil Maliyekkel, Yu Shao, Igor B Roninson, and Ye Ding. A structural interpretation of the effect of GC-content on efficiency of RNA interference. *BMC Bioinformatics*, 10(Suppl 1):S33, January 2009. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2648742/, doi:10.1186/1471-2105-10-S1-S33.

[10] Alexander Crits-Christoph, Rose S Kantor, Matthew R Olm, Oscar N Whitney, Basem Al-Shayeb, Yue Clare Lou, Avi Flamholz, Lauren C Kennedy, Hannah Greenwald, Adrian Hinkle, Jonathan Hetzel, Sara Spitzer, Jeffery Koble, Asako Tan, Fred Hyde, Gary Schroth, Scott Kuersten, and Jillian F Ban. Genome Sequencing of Sewage Detects Regionally Prevalent SARS-CoV-2 Variants. 12(1):9, 2021.

[11] Aviv Dombrovsky and Elisheva Smith. *Seed Transmission of Tobamoviruses: Aspects of Global Disease Distribution*. IntechOpen, December 2017. Publication Title: Advances in Seed Biology. URL: https://www.intechopen.com/chapters/undefined/state.item.id, doi:10.5772/intechopen.70244.

[12] Siobain Duffy. Why are RNA virus mutation rates so damn high? *PLOS Biology*, 16(8):e3000003, August 2018. URL: https://dx.plos.org/10.1371/journal.pbio.3000003, doi:10.1371/journal.pbio.3000003.

[13] Isaac Ellmen, Michael D. J. Lynch, Delaney Nash, Jiujun Cheng, Jozef I. Nissimov, and Trevor C. Charles. Alcov: Estimating Variant of Concern Abundance from SARS-CoV-2 Wastewater Sequencing Data. Technical report, medRxiv, June 2021. Type: article. URL: https://www.medrxiv.org/content/10.1101/2021.06.03.21258306v1, doi:10.1101/2021.06.03.21258306.

[14] Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, Julie D Thompson, and Desmond G Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7(1):539, January 2011. Publisher: John Wiley & Sons, Ltd. URL: https://www.embopress.org/doi/full/10.1038/msb.2011.75, doi:10.1038/msb.2011.75.

[15] M. Gouy, S. Guindon, and O. Gascuel. SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. *Molecular Biology and Evolution*, 27(2):221–224, February 2010. URL: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msp259, doi:10.1093/molbev/msp259.

[16] Canadian Food Inspection Agency Government of Canada. Tomato brown rugose fruit virus, June 2019. Last Modified: 2019-11-29. URL: https://inspection.canada.ca/plant-health/invasive-species/plant-diseases/tobrfv/eng/1560266450577/1560266450826.

[17] Innovation Government of Canada. Edition 26, May 2019 - Science.gc.ca. Last Modified: 2020-05-08 Publisher: Innovation, Science and Economic Development Canada. URL: https://science.gc.ca/eic/site/063.nsf/eng/h_97842.html#6.

[18] James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, December 2018. doi:10.1093/bioinformatics/bty407.

[19] T. Hovi, L. M. Shulman, H. Van Der Avoort, J. Deshpande, M. Roivainen, and E. M. De Gourville. Role of environmental poliovirus surveillance in global polio eradication and beyond. *Epidemiology & Infection*, 140(1):1–13, January 2012. Publisher: Cambridge University Press. doi:10.1017/S095026881000316X.

[20] Julia L. Mullen, Ginger Tsueng, Alaa Abdel Latif, Manar Alkuzweny, Marco Cano, Emily Haag, Jerry Zhou, Mark Zeller, Nate Matteson, Kristian G. Andersen, Chunlei Wu, Andrew I. Su, Karthik Gangavarapu, Laura D. Hughes, and Center for Viral Systems Biology. outbreak.info. Publication Title: outbreak.info. URL: https://outbreak.info/.

[21] Rose S. Kantor, Kara L. Nelson, Hannah D. Greenwald, and Lauren C. Kennedy. Challenges in Measuring the Recovery of SARS-CoV-2 from Wastewater. *Environmental Science & Technology*, 55(6):3514–3519, March 2021. Publisher: American Chemical Society. doi:10.1021/acs.est.0c08210.

[22] Masaaki Kitajima, Hannah P. Sassi, and Jason R. Torrey. Pepper mild mottle virus as a water quality indicator. *npj Clean Water*, 1(1):1–9, October 2018. Number: 1 Publisher: Nature Publishing Group. URL: https://www.nature.com/articles/s41545-018-0019-5, doi:10.1038/s41545-018-0019-5.

[23] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, October 1999. URL: http://www.nature.com/articles/44565, doi:10.1038/44565.

[24] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, May 2013. Number: arXiv:1303.3997 arXiv:1303.3997 [q-bio]. URL: http://arxiv.org/abs/1303.3997, doi:10.48550/arXiv.1303.3997.

[25] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, September 2018. doi:10.1093/bioinformatics/bty191.

[26] Yang Liu, Jianying Liu, Kenneth S. Plante, Jessica A. Plante, Xuping Xie, Xianwen Zhang, Zhiqiang Ku, Zhiqiang An, Dionna Scharton, Craig Schindewolf, Vineet D. Menachery, Pei-Yong Shi, and Scott C. Weaver. The N501Y spike substitution enhances SARS-CoV-2 transmission. *bioRxiv*, page 2021.03.08.434499, March 2021. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7986995/, doi:10.1101/2021.03.08.434499.

[27] Jatin Machhi, Jonathan Herskovitz, Ahmed M. Senan, Debashis Dutta, Barnali Nath, Maxim D. Oleynikov, Wilson R. Blomberg, Douglas D. Meigs, Mahmudul Hasan, Milankumar Patel, Peter Kline, Raymond Chuen-Chung Chang, Linda Chang, Howard E. Gendelman, and Bhavesh D. Kevadiya. The Natural History, Pathobiology, and Clinical Manifestations of SARS-CoV-2 Infections. *Journal of Neuroimmune Pharmacology*, 15(3):359–386, 2020. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7373339/, doi:10.1007/s11481-020-09944-5.

[28] Fábio Madeira, Matt Pearce, Adrian R N Tivey, Prasad Basutkar, Joon Lee, Ossama Edbali, Nandana Madhusoodanan, Anton Kolesnikov, and Rodrigo Lopez. Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Research*, page gkac240, April 2022. doi:10.1093/nar/gkac240.

[29] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12, May 2011. Number: 1. URL: https://journal.embnet.org/index.php/embnetjournal/article/view/200, doi:10.14806/ej.17.1.200.

[30] Peter Menzel, Kim Lee Ng, and Anders Krogh. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, 7(1):11257, September 2016. URL: http://www.nature.com/articles/ncomms11257, doi:10.1038/ncomms11257.

[31] Ahmad Abu Turab Naqvi, Kisa Fatima, Taj Mohammad, Urooj Fatima, Indrakant K. Singh, Archana Singh, Shaikh Muhammad Atif, Gururao Hariprasad, Gulam Mustafa Hasan, and Md. Imtaiyaz Hassan. Insights into SARS-CoV-2 genome, structure,

evolution, pathogenesis and therapies: Structural genomics approach. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1866(10):165878, October 2020. URL: https://www.sciencedirect.com/science/article/pii/S092544392030226X, doi:10.1016/j.bbadis.2020.165878.

[32] ARTIC Network. artic-network/primer-schemes: v1.1.1, September 2020. URL: https://zenodo.org/record/4020380, doi:10.5281/zenodo.4020380.

[33] World Health Organization. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020.

[34] Áine O'Toole, Emily Scher, Anthony Underwood, Ben Jackson, Verity Hill, John T McCrone, Rachel Colquhoun, Chris Ruis, Khalil Abu-Dahab, Ben Taylor, Corin Yeats, Louis du Plessis, Daniel Maloney, Nathan Medd, Stephen W Attwood, David M Aanensen, Edward C Holmes, Oliver G Pybus, and Andrew Rambaut. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evolution*, 7(2):veab064, December 2021. doi:10.1093/ve/veab064.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[36] Laurent Perron and Vincent Furnon. OR-Tools, July 2019. URL: https://developers.google.com/optimization/.

[37] Joshua Quick, Nathan D. Grubaugh, Steven T. Pullan, Ingra M. Claro, Andrew D. Smith, Karthik Gangavarapu, Glenn Oliveira, Refugio Robles-Sikisaka, Thomas F. Rogers, Nathan A. Beutler, Dennis R. Burton, Lia Laura Lewis-Ximenez, Jaqueline Goes de Jesus, Marta Giovanetti, Sarah C. Hill, Allison Black, Trevor Bedford, Miles W. Carroll, Marcio Nunes, Luiz Carlos Alcantara, Ester C. Sabino, Sally A. Baylis, Nuno R. Faria, Matthew Loose, Jared T. Simpson, Oliver G. Pybus, Kristian G. Andersen, and Nicholas J. Loman. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nature Protocols*, 12(6):1261–1276, June 2017. Number: 6 Publisher: Nature Publishing Group. URL: https://www.nature.com/articles/nprot.2017.066, doi:10.1038/nprot.2017.066.

[38] Re3data.Org. GISAID. 2012. Publisher: re3data.org - Registry of Research Data Repositories. URL: https://www.re3data.org/repository/r3d100010126, doi:10.17616/R3Q59F.

[39] Alvin C. Rencher and William F. Christensen. *Methods of multivariate analysis*. Wiley series in probability and statistics. Wiley, Hoboken, New Jersey, third edition edition, 2012.

[40] N. Salem, A. Mansour, M. Ciuffo, B. W. Falk, and M. Turina. A new tobamovirus infecting tomato crops in Jordan. *Archives of Virology*, 161(2):503–506, February 2016. doi:10.1007/s00705-015-2677-7.

[41] Jaswinder Singh, Jack Hanson, Kuldip Paliwal, and Yaoqi Zhou. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nature Communications*, 10(1):5407, November 2019. Number: 1 Publisher: Nature Publishing Group. URL: https://www.nature.com/articles/s41467-019-13395-9, doi:10.1038/s41467-019-13395-9.

[42] Davida S. Smyth, Monica Trujillo, Kristen Cheung, Anna Gao, Irene Hoxie, Sherin Kannoly, Nanami Kubota, Michelle Markman, Kaung Myat San, Geena Sompanya, and John J. Dennehy. Detection of Mutations Associated with Variants of Concern Via High Throughput Sequencing of SARS-CoV-2 Isolated from NYC Wastewater. preprint, Infectious Diseases (except HIV/AIDS), March 2021. URL: http://medrxiv.org/lookup/doi/10.1101/2021.03.21.21253978, doi:10.1101/2021.03.21.21253978.

[43] John St. John. jstjohn/SeqPrep, May 2021. URL: https://github.com/jstjohn/SeqPrep.

[44] Charles B. Stephensen, Donald B. Casebolt, and Nupur N. Gangopadhyay. Phylogenetic analysis of a highly conserved region of the polymerase gene from 11 coronaviruses and development of a consensus polymerase chain reaction assay. *Virus Research*, 60(2):181–189, April 1999. URL: https://linkinghub.elsevier.com/retrieve/pii/S0168170299000179, doi:10.1016/S0168-1702(99)00017-9.

[45] Ontario COVID-19 Science Advisory Table. 2022-04-13-Distribution-of-Variant.png (1588×1232). URL: https://covid19-sciencetable.ca/wp-content/uploads/2022/04/2022-04-13-Distribution-of-Variant.png.

[46] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, June 2001. doi:10.1093/bioinformatics/17.6.520.

[47] Yatish Turakhia, Bryan Thornlow, Angie S. Hinrichs, Nicola De Maio, Landen Gozashti, Robert Lanfear, David Haussler, and Russell Corbett-Detig. Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nature Genetics*, 53(6):809–816, June 2021. Number: 6 Publisher: Nature Publishing Group. URL: https://www.nature.com/articles/s41588-021-00862-7, doi:10.1038/s41588-021-00862-7.

[48] John R. Tyson, Phillip James, David Stoddart, Natalie Sparks, Arthur Wickenhagen, Grant Hall, Ji Hyun Choi, Hope Lapointe, Kimia Kamelian, Andrew D. Smith, Natalie Prystajecky, Ian Goodfellow, Sam J. Wilson, Richard Harrigan, Terrance P. Snutch, Nicholas J. Loman, and Joshua Quick. Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore, September 2020. Pages: 2020.09.04.283077 Section: New Results. URL: https://www.biorxiv.org/content/10.1101/2020.09.04.283077v1, doi:10.1101/2020.09.04.283077.

[49] Bart T. L. H. van de Vossenberg, Thikra Dawood, Marek Woźny, and Marleen Botermans. First Expansion of the Public Tomato Brown Rugose Fruit Virus (ToBRFV) Nextstrain Build; Inclusion of New Genomic and Epidemiological Data. *PhytoFrontiers™*, 1(4):359–363, December 2021. Publisher: Scientific Societies. URL: https://apsjournals.apsnet.org/doi/10.1094/PHYTOFR-01-21-0005-A, doi:10.1094/PHYTOFR-01-21-0005-A.

[50] William A. Walters, J. Gregory Caporaso, Christian L. Lauber, Donna Berg-Lyons, Noah Fierer, and Rob Knight. PrimerProspector: de novo design and taxonomic analysis of barcoded polymerase chain reaction primers. *Bioinformatics*, 27(8):1159–1161, April 2011. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr087, doi:10.1093/bioinformatics/btr087.

[51] Brian J. Willett, Joe Grove, Oscar A. MacLean, Craig Wilkie, Giuditta De Lorenzo, Wilhelm Furnon, Diego Cantoni, Sam Scott, Nicola Logan, Shirin Ashraf, Maria Manali, Agnieszka Szemiel, Vanessa Cowton, Elen Vink, William T. Harvey, Chris Davis, Patawee Asamaphan, Katherine Smollett, Lily Tong, Richard Orton, Joseph Hughes, Poppy Holland, Vanessa Silva, David J. Pascall, Kathryn Puxty, Ana da Silva Filipe, Gonzalo Yebra, Sharif Shaaban, Matthew T. G. Holden, Rute Maria Pinto, Rory Gunson, Kate Templeton, Pablo R. Murcia, Arvind H. Patel, Paul Klenerman, Susanna Dunachie, John Haughney, David L. Robertson, Massimo Palmarini, Surajit Ray, and Emma C. Thomson. SARS-CoV-2 Omicron is an immune escape variant with an altered cell entry pathway. *Nature Microbiology*, 7(8):1161–1179, August 2022. Number:

8 Publisher: Nature Publishing Group. URL: https://www.nature.com/articles/s41564-022-01143-7, doi:10.1038/s41564-022-01143-7.

[52] Fan Wu, Su Zhao, Bin Yu, Yan-Mei Chen, Wen Wang, Zhi-Gang Song, Yi Hu, Zhao-Wu Tao, Jun-Hua Tian, Yuan-Yuan Pei, Ming-Li Yuan, Yu-Ling Zhang, Fa-Hui Dai, Yi Liu, Qi-Min Wang, Jiao-Jiao Zheng, Lin Xu, Edward C. Holmes, and Yong-Zhen Zhang. A new coronavirus associated with human respiratory disease in China. *Nature*, 579(7798):265–269, March 2020. doi:10.1038/s41586-020-2008-3.

APPENDICES

Appendix A

Designing Amplicons with viral-amplicons

To design low-degeneracy amplicons for a range of viruses, I built a custom tool which wraps the functionality of PrimerProspector. The tool is publicly available on GitHub at (https://github.com/Ellmen/viral-amplicons).

A.1 Updating PrimerProspector for Python3

PrimerProspector was released in 2011 and so was written using Python2 which is now deprecated. To build a tool using Python3 which calls functions from PrimerProspector, I decided to update it for Python3. This turned out to be more challenging than one might expect. Here I describe some of the changes which were required to update PrimerProspector. The details are also available in the commit history on GitHub (https://github.com/Ellmen/primerprospector3/commits/master).

I started by using the `futurize` Python package which will automatically update some of the common syntax patterns. Examples of this are tokens like `print` and `raise` which used to be statements but are now functions. I also had to replace `from string import upper` imports with `str.upper`.

Next, I had to replace a large dependency of PrimerProspector, cogent. There exists a Python3 compatible version of cogent called cogent3 which I used, however it has a slightly different API. For example, the function `LoadSeqs` was renamed `load_unaligned_seqs` in cogent3. Determining which new functions correspond to the old versions was challenging

and required reading through a substantial amount of source code. To make matters worse, some of the old functionality was not replicated in the new version. In these cases I included a copy of some of the old cogent source code in the new PrimerProspector source, which I manually futurized.

One of the functions, `make_unaligned_seqs` from cogent, used to return a list of dictionaries but now returns a position weight matrix (PWM). I coded a custom function to convert a PWM to a list of dictionaries:

```
def pwm_to_dict(pwm):
    return [{
        'T': pos[0],
        'C': pos[1],
        'A': pos[2],
        'G': pos[3],
    } for pos in pwm]
```

I attempted to update the tests, but they were heavily reliant on the old version of cogent. Instead, I manually ensured that the commands from the PrimerProspector tutorial produced the same output as the original version.

A.2 Using viral-amplicons

Viral-amplicons is run as a CLI tool. The following information is taken from the README.

A.2.1 Installing

Clone the repository and run

```
pip install .
```

This will install the Python library and the CLI.

A.2.2 Usage Example

Generate primer hits and amplicons:

```
cd test_data
amplicons find sars-cov-2.fasta sequences.fasta aligned.fasta
```

Find lowest degeneracy primer pair:

```
amplicons ld
```

Generate simulated reads for the generated amplicon

```
amplicons get_amplicons sequences.fasta 18065f 18307r
```

Glossary

vector A quantity containing values along different dimensions, commonly used in data science to encode data points 32