ANALYSIS OF RESPIRATORY VIRAL MATERIAL FROM NEGATIVE COVID-19 PCR TESTS TO OBSERVE EFFECTS OF 2022-2023 “TWINDEMIC” ON UNIVERSITY CAMPUS

by

Lolo Aboufoul

A thesis submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Master of Science in
Computer Science

Charlotte

2023

Approved by:

_____
Dr. Cynthia Gibas

_____
Dr. Harini Ramaprasad

_____
Dr. Jessica Schlueter

_____
Dr. Nadia Najjar

DEDICATION

I dedicate this work to my parents, who inspire me to be the best version of myself every single day, and to my friends and family, who are the best support system I could ask for.

ABSTRACT

LOLO ABOUFOUL. Analysis of Respiratory Viral Material from Negative COVID-19 PCR Tests to Observe Effects of 2022-23 “Twindemic” on University Campus. (Under the direction of DR. CYNTHIA GIBAS)

Tracking COVID-19 cases has been one of the most important aspects of controlling the pandemic and observing how the virus causing COVID-19 spreads between populations. Tracking cases of the disease helps epidemiologists understand the reason for spreading of the virus and is an important tool in implementing ways of containing the virus and reducing its spread. On a local level, tracking of cases can lead to effective quarantining of people who have been infected by quarantining them as soon as they test positive for the virus. It can also show which areas are most susceptible to being infected and in which areas person to person contact needs to be limited. The first aim of this research is to apply this same method of tracking the presence of the COVID-19 causing virus (SARS-CoV-2) to tracking other potentially infectious viruses and bacteria. By extracting genetic material from samples of on-campus wastewater and negative COVID-19 PCR tests and applying sequencing and analysis techniques on them, we are able to see what other viruses or bacteria are circulating around campus. Furthermore, this method of tracking can be used to make predictions about when a disease will spread around campus in the future. The “twindemic” refers to the spread of both COVID-19 and Influenzas A/B at the same time; observing what is found in negative COVID-19 tests can show the effects of the twindemic and to what extent both viral infections were being spread around the university’s campus.

Various softwares are available that take the results of DNA sequencing (referred to as “reads”) and return the taxonomic classification of each read, therefore revealing what organisms

were originally in the sample. This is the final step of the process known as “sequencing” which involves processing the raw extracted genetic material to eventually obtain the sequence of DNA nucleotides belonging to the sample’s contents. Kraken is a popular metagenomic classification tool that was developed by researchers at Johns Hopkins University’s Center for Computational Biology in 2014. Kraken has, as of September 2022, been eradicated and replaced by Kraken2, which boasts a faster and more compact algorithm. Kraken works by looking up sequences of DNA called “k-mers” of length “k” in a database and selects the least common organism whose DNA contains that k-mer. It then travels down the taxonomic tree in search of the most specific organism that contains the k-mer, eventually matching it to the input DNA sequence. The developers of Kraken inspired researchers at the University of California at Riverside, who developed a metagenomic classification software called CLARK in 2015 to address Kraken’s shortcomings. CLARK works by first building an index using all the possible k-mers of potential target organisms. It then builds a dictionary that associates each k-mer to potential targets. For each input sequence, CLARK’s algorithm searches the index to match the set of k-mers within the input sequence. Whenever an organism contains the k-mer being searched, it receives a “hit”. The organism with the most hits at the end of the search is mapped to the input sequence.

The second aim of this research is to compare the results obtained from using Kraken2 and CLARK to classify the collected samples. Results can be used to see which classification tool is best in accuracy and other metrics. Comparing both CLARK and Kraken2 can also show whether one tool is better at detecting the presence of certain organisms than the other. Comparing these widely popular tools can help suggest if there is a need for the development of other software that overcomes any shortcomings these softwares may possess.

ACKNOWLEDGMENTS

I would like to acknowledge several people who have been instrumental in the completion of this project; without their help, I would not have been able to accomplish this. Thank you to everyone at the UNC-Charlotte COVID-19 Wastewater Processing Lab for their help with extraction of samples and being ready to help at any time. Thank you to my thesis committee members: Dr. Cynthia Gibas, Dr. Harini Ramaprasad, Dr. Jessica Schlueter, and Dr. Nadia Najjar, for their support and guidance, and for being amazing mentors who I will always look up to. Thank you to William Taylor for his mentorship and guidance throughout this entire experiment. Thank you to Samuel Kunkleman for his help with writing our Kraken and CLARK code. A big thank you to Lauren R. Brazell for her guidance, support, and encouragement every step of the way; I have learned so much from Lauren and will always be grateful for her presence in my professional and personal life. Last but not least, thank you to my family and friends, who I could not imagine my life without.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction

The human virome includes viruses, bacteriophages, and virus-derived elements that can infect human cells and often cause illness. Depending on how various cell types react to the virome, there can either be positive or negative effects on human health. Taken together, viruses can shape the overall health of human tissues and therefore the human body itself. Research on the impacts of the human virome is limited, meaning the impacts of all viruses in the virome on human cells are not well understood. There is a lack of robust databases that contain virome genetic information, and the relatively small size of the viral genome can limit our knowledge base [1]. Furthering the amount of research done on viral genetic material can forward research into various disease-causing viruses, and can lead to further development of treatment for these diseases.

To better understand the impacts of viral infections in distinct populations, the seasonal circulation of different viruses should be investigated. This is challenging to document, as collecting clinical data regarding viral infections in a timely and accurate way from all over the world is exceptionally difficult [3]. Consistent and accurate reporting between nations in itself is difficult, and the challenges of testing may lead to results that show a negative result but are actually positive, called “false negatives”, or tests that show a positive result but are actually negative, called “false positives”. Further, this approach relies completely on individuals who go in for testing and those testing results being reported. However, tracking the spread of an infection is necessary to understand the harmful effects the pathogen may have on a given population. Finding patterns in infection cycles can show what seasons of the year a particular virus infects large groups of people. When several communities report large outbreaks of a

certain virus this can lead to an epidemic, whereas an outbreak of the same virus across different countries may lead to a pandemic state.

As an example of a pandemic-causing virus, coronaviruses have been studied extensively. This family of viruses are typically known to infect birds and mammals [2]. The novel human coronavirus, SARS-CoV-2, is responsible for the respiratory infection known as COVID-19 and led to the worldwide global pandemic that has had countless large-scale effects on the political and economic states of entire nations, as well as small-scale impacts on businesses and the mental health of individual people. Tracking the spread of COVID-19 remains important because of the fast-spreading nature of the virus as well as the varying effects of new variants. One infected individual can infect several others in the same community, and those individuals can spread the infection exponentially. The dominant transmission route for COVID-19 is via airborne viral particles, meaning infections can occur much faster between individuals [4] than for viruses that are transmitted through direct fluid contact. Due to the rapid transmission and wide range of harmful effects viral infections like COVID-19 can cause, tracking infection rates is a necessary epidemiological tool for reducing the infection rate.

Viral tracking methods are not limited to COVID-19 and have been used to some degree for hundreds of years, most recently with molecular diagnostics the most effective measure. Surveillance of other respiratory viruses besides SARS-CoV-2 was a topic of concern throughout the pandemic. Observing the spread of other viruses was important to see if there would be a peak in the spread of COVID-19 along with other harmful illnesses. Viral disease tracking can be beneficial on a small scale, such as on a university campus, to inform students and faculty what may be spreading in the area they live or work in or to inform on preventive measures for the spread of infection. On a broader scale, counties and larger communities benefit from tracking

which diseases are in circulation as this impacts K-12 schools and business operations. Predictive models can be created from the data collected, which can lead to informing inhabitants of a certain area when a certain virus is in high circulation in the community or when scientists predict it will increase.

1.2 Background on DNA, RNA, and Sequencing

Deoxyribonucleic acid, orDNA, is a double-stranded molecule which is composed of a sequence of 4 nucleotides: adenine, thymine, cytosine, and guanine (A, T, C, and G). The series of nucleotides in each strand are bound vertically by phosphodiester bonds [7], and both strands are bound together by hydrogen bonds. Each nucleotide has an opposite base pair that it is able to bond with: adenine molecules bind to thymine, while cytosine and guanine molecules bind together. Ribonucleic acid, or RNA, is a single-stranded molecule of genetic material that is used to provide genetic information for processes such as catalyzing reactions and protein synthesis. RNA is single-stranded and contains similar nucleotides to DNA with the exception of the thymine nucleotide being replaced by uracil (U). RNA is an easily degradable molecule, and the presence of uracil rather than thymine contributes to this ease [24]. RNA is degraded as soon as the organism that made it is done with its use, which is why ease of degradability is important to be mindful of [25]. DNA is converted to RNA by DNA transcription, which involves the splitting of the molecule's double helix structure, and each DNA nucleotide gets translated to its corresponding base pair [8].

DNA sequencing is the process of finding the exact order of nucleotides in a given sample of DNA [9]. All DNA molecules consist of the same four nucleotides; the difference in each DNA molecule is the unique order of these nucleotides. The difference in the order of nucleotides determines the information contained within the DNA molecule. The DNA sequence

from one fragment is referred to as a "read". Differences in DNA sequence can be as large-scale as determining whether an organism is a human or a chimpanzee, or can be as small as determining if a human will have green or blue eyes. DNA sequencing is important because it can show exactly what organism is found in a sample down to the nucleotide level. In relation to this work, DNA sequencing can detect what viruses and bacteria are found in samples of wastewater and human COVID-19 negative nasopharyngeal swabs.

1.3 Sequencing protocols

DNA sequencing technologies can be classified depending on how advanced the technology is and how recently it has been developed. First-generation sequencing is considered the "gold standard" and was first developed by Fred Sanger; it is used because of its high accuracy and robustness as well as its ease of use compared to other protocols generated at the time [9]. Sanger's method involved using monomers of DNA called deoxyribonucleotides (dNTPs), which are the individual nucleotides that, when linked together, form one whole strand of DNA. Dideoxyribonucleotides (ddNTPs), a specific form of the dNTP monomers that are called chain terminators, are labeled with a colored dye and are spiked into the sequencing reaction at a small concentration to produce fragments that terminate at that specific base [10]. A DNA primer is a short strand of DNA that can bind to the original sample's DNA and serve as a starting point for the DNA polymerase to attach to and build the opposing strand of the double helix.

A popular method for amplifying genetic material is the polymerase chain reaction, or PCR [6]. PCR involves combining DNA primers, DNA polymerase, and the four base nucleotides in a chamber which is heated to drive forward the reaction. The DNA strand is "photocopied" as the original strand is denatured (split apart) and primers attach to the loose

ends, while DNA polymerase works to add new nucleotides to the unzipped ends of the strand. DNA amplification is necessary in several sequencing protocols to ensure enough genetic material is available for the sequencing platform to detect the presence of nucleotide base pairs.

During the process of Sanger sequencing, DNA polymerase will continue to add to the original DNA strand until a ddNTP is bound to the strand. Once this occurs, there can be no more nucleotides added to keep building the new DNA. This step is repeated several times; the goal is for one ddNTP molecule to be added to every position in the original DNA molecule [10]. The resulting reactions were originally run on a polyacrylamide gel to be visualized [9] although new technologies use capillary polymers and automated detection.

One drawback to Sanger sequencing is that only one fragment of DNA can be sequenced at a time with a maximum of 384 samples on a single system. This leads to costly and time consuming projects for sequencing large genomes. Newer technologies referred to as "next-generation sequencing", or NGS, are able to sequence millions of DNA fragments in one reaction on a single run [11]. These methods allow for multiple different experiments to be performed in parallel through molecular barcoding, or sequencing only certain parts of the organism's genome using targeted sequencing [11].

One widely used next-generation sequencing technology is Oxford Nanopore sequencing. Nanopore sequencing uses a nanoscale protein pore ("nanopore") to sequence single molecules in real-time [5]. A voltage is applied to an electrolyte solution which allows for an ionic current to run through the nanopore so that negatively charged single-strand molecules of RNA or DNA move to the opposite, positively charged side. The changes in the ionic current during translocation match the nucleotide sequence present on the sensing region of the nanopore, leading to the ability to determine the base compositions of the RNA or DNA.

Illumina sequencing is another popular next-generation sequencing technique. Illumina sequencing involves multiple copies of the DNA strand being made, known as amplification, while adhered to a surface [12]. First, the DNA sample is prepared by ligation, or attachment, of adapters to each strand. Adapters are short nucleotide sequences that are attached to the ends of DNA fragments and are combined with primers to amplify, or generate more of, the DNA sample [13]. Ligated single-stranded molecules of DNA are bound to the inside of a flow cell, and several rounds of amplification of the DNA are completed to generate clusters of the input sequence. After the subsequent attachment of DNA polymerase and primers, an image is captured of the nucleotide bases after a laser excites the fragment of DNA. The cycle of image capture is repeated, and the resulting data is able to be analyzed [14].

The slight differences for each protocol are what cause a difference in resulting yields. For example, Oxford Nanopore sequencing can sequence longer reads, while Illumina sequencing requires fragmentation and yields shorter reads [14]. This is because Oxford Nanopore sequencing is a single molecule sequencing protocol and has the sensitivity to detect base nucleotides without first amplifying the genetic molecule and creating clusters of molecules [44]. Illumina sequencing is also responsible for 90% of the world's sequencing data [15]. Currently, the most widely used methods of sequencing are next-generation sequencing methods. In the following sections, the steps of sequencing are discussed.

1.3.1 Steps of sequencing

Although there are several different protocols and methods to accurately complete DNA sequencing, there are several general steps that almost all next-generation sequencing methods utilize. First, a DNA or RNA sample must be collected. This could be, for example, through a nasal swab. The next step is to extract the genetic material from the sample. This can be done

through several different methods. An example technique for RNA extraction is the Qiagen protocol for extraction of RNA from blood or bodily fluids (Qiagen, Germantown, MD). After the cells have been lysed and opened, various reagents are added to the sample which help in isolating the RNA and removing cellular debris and excess proteins. Afterwards, the sample is treated using the DNAse enzyme, which removes the DNA in the sample and retains the RNA necessary for sequencing.

If the starting genetic material is RNA, most methods require that it is first converted to a specific type of DNA called complementary DNA, or cDNA [16]. This is done by the reverse transcription of the RNA molecule which creates a complementary DNA strand from the RNA molecule. Because the amount of viral material that is extracted from the sample is often very low concentration, a single-primer amplification process, known as sequence-independent single primer amplification, or SISPA [41] is used to amplify the starting material for sequencing. SISPA is also the process in which RNA is converted to cDNA by reverse transcription. cDNA is more stable than RNA, which is why the conversion step is necessary.

The next step in sequencing that is common to all next-generation protocols is referred to as library preparation [15]. The DNA or cDNA is fragmented into smaller sections and specific adapters are ligated to the ends of these fragments. DNA fragmentation, which is the process of fragmenting the DNA strand into smaller pieces. This is necessary because some sequencing protocols are not able to sequence extremely large DNA molecules and need them to be split into smaller fragments. Barcodes are added during library preparation to ensure that each sample of DNA can be uniquely identified. Barcodes are short stretches of nucleotides that are attached to each DNA strand and allow it to be differentiated from other samples in the same library when all samples are pooled together. These adapters are sequencing platform dependent. In the case of

Illumina sequencing, adapters are also able to prevent the movement of molecules when adhered to a solid surface, such as a sequencing flow cell.

The final step in the process is to sequence the DNA. The methods of loading the prepared libraries differ based on the sequencing platform that is used. For Illumina sequencing, the prepared library is loaded onto a device called a "flow cell" which in turn gets inserted to a sequencing instrument, and each nucleotide is "read" individually [17]. Adapters on both ends of the DNA molecule attach the fragments to the surface of the flow cell in a bridge format to allow for bridge amplification [12]. The bridge strands are amplified and clusters of DNA molecules are formed. The input DNA sequence is read using image classification. A DNA strand complementary to the input sample DNA is synthesized as part of the sequencing process. As the complementary strand is synthesized, the newly added base pairs emit colored signals, as they are fluorescently labeled. These colored signals are recognized by the sequencing instrument, and the input DNA sequence is extrapolated before being stored in an output file. After the nucleotide sequence is complete, a series of bioinformatics softwares are used to further analyze the DNA and much can be deduced from the resulting sequence. It is at this point that the organism in each sample is able to be identified.

1.4 Bioinformatics Tools and Classification Softwares

The processing and analysis of sequencing reads involves the use of different software designed for specific platforms and applications. For example, some software is better suited for long reads rather than short reads and vice versa. Other factors, such as how many reads match the reference genome, also help determine the quality of a certain bioinformatic software. Metagenomic classification involves a taxonomic identity to as many reads as possible within a

dataset. Ranking these different softwares and seeing which one fits a particular experiment the best will help deliver the most efficient experimental results.

1.4.1 Metagenomic Classification Tools

The classification tool Kraken is popular because of its accuracy and speed [18]. First developed in 2014, Kraken utilizes a database that contains a sequence of DNA with a user-determined number of nucleotides, called a "k-mer", where "k" represents the number of nucleotides in the genome. The Kraken database also contains the least common ancestor of the organisms which contain this k-mer sequence. The software classifies input DNA sequences by searching the entire database for any sequence that contains the k-mer, and then searches its least common ancestors to see which organism the input sequence originates from. If the input k-mer is not efficiently matched with any of the organisms in the database, then Kraken will return the output as unclassified. This feature in particular helps Kraken lower its rate of incorrect classifications when compared to other software. Kraken was later followed by Kraken2, which was developed to reduce Kraken's vast use of device memory [20]. This allows for larger data sources to be used and analyzed.

Kraken2 differs from Kraken in that it uses a compact hash table to store reference sequences, which it uses to compare references to input sequences. A hash table is a data structure that contains keys and value pairs. It is much faster to use to look up data that is stored in a hash table than a list. A hash table will take an input sequence, apply a hash function to the sequence, and look up that result to find the matching value, which in the case of Kraken2 is the taxa that matches the input. A hash table takes the same amount of time to run despite the number of inputs, whereas a list's runtime increases as the amount of inputs increases. For these reasons, Kraken2 is currently the standard classification tool used.

CLARK is a metagenomic classification tool that was developed by researchers at the University of California at Riverside in 2015 [21]. CLARK, an abbreviation for CLAassifiers based on Reduced K-mers, was developed to address shortcomings in previously established classification methods such as Kraken. Operating at a comparable speed but using less disk space, CLARK opts to classify reads at a taxonomic level more precisely. The algorithm begins by building an index containing all possible k-mers of target sequences. After creating this index, the algorithm builds a dictionary that maps input k-mers to possible targets. For each input sequence, CLARK searches the index to find a set of k-mers that matches the inputs. When an input k-mer matches one from the index, that target receives a "hit". The target from the index with the most hits is matched as the organism the input sequence originates from. CLARK works to reduce "noise" in the index by removing common k-mers between targets to reduce the number of targets that can potentially receive the same number of hits.

Other sequence classification tools include Kaiju, which was introduced after Kraken in 2016 and attempted to more efficiently classify sequencing results [19]. Kaiju uses protein-level classification due to its higher accuracy in results it produces. Genomes of bacteria and viruses contain large amounts of varying protein genes. Kaiju works by searching its own genome database for MEMs, or maximum exact matches, to proteins found in the input reads. The label assigned to each read is found by starting the search for the matching organism at the least common ancestor of all organisms in the database. Kaiju has been confirmed to outperform Kraken in both sensitivity and precision of results even when the input reads were obtained using different sequencing methods.

1.4.2 Comparison of Metagenomic Classification Tools

The word “metagenomics” is defined as relating to the study of the structure of entire genetic sequences to allow for better understanding of the relationship between diseases and organisms that naturally inhabit the human body. Selective sequencing of certain microbial genomes can further allow for studying the interactions between humans and microbes. Certain bioinformatic classification tools can be used to observe different taxonomic levels of microbial samples, which means the samples can be studied at even greater levels of specificity.

Several studies have been conducted that compare various metagenomic classification tools, which is an important topic because certain classifiers are meant to work better on genetic material from certain organisms than others. One such study compares five different classification tools (Centrifuge, CLARK, Kaiju, Kraken2, and Genome Detective) in their ability to recognize viruses in respiratory clinical samples [22]. The study reported Kaiju outperformed all other classifiers in terms of sensitivity, but Centrifuge and Genome Detective outperformed in terms of selectivity. Sensitivity is the ability of a software to classify a sample as positively matching a sequence in the reference table, whereas selectivity is the ability of the software to detect if a sequence does not match one in the reference table (detecting negatives). The same study also included various results from other studies, including the amount of misclassification of taxa yielded by each tool. Centrifuge and a slight variation of Kraken2 called KrakenUniq had the smallest rates of misclassification. The study posits that this is likely due to inclusion of the human genome in their reference databases rather than just viral or bacterial genomes.

Another such study compares 20 metagenomic classification tools and compares the results they yield [23]. When looking at the speed of classification, Kraken and CLARK are both extremely fast after the reference database has been loaded. Kraken2 had a faster computational time compared to CLARK, however, both took up significant amounts of memory. This further

proves the usefulness of both Kraken and CLARK is constrained by the amount of memory needed to run them.

The goal of the experiment detailed in this thesis is to determine which respiratory viruses have been circulating on UNC Charlotte campus during the 2022-2023 “twindemic”, which refers to the increased spread of both the COVID-19 causing virus SARS-CoV-2 and the Influenza virus. Based on experimental results, we can observe which viruses were the cause of respiratory symptoms reported by students seeking a Covid test at the UNC Charlotte Health Center , and by extension which viruses potentially posed the greatest threat to people on the university's campus. In addition, we compared our results using two different genomic classification software packages and to evaluate which software might be most informative for clinical samples using our protocols. The main goal is to identify respiratory viruses present in clinical negative samples obtained from the on-campus student health center and observe the presence of respiratory viruses besides SARS-Cov-2.

CHAPTER 2: MATERIALS AND METHODS

2.1 Obtaining Samples and Time Series

In order to observe the effects of both Covid-19 and Influenza (also known as the twindemic), the start date of sample collection must coincide with the increase in influenza incidence. It is commonly known that cases of influenza and the common cold rise during colder months. Influenza surprised us during the initial stages of the COVID-19 pandemic because cases of influenza decreased due to COVID imposed restrictions, which limited person to person contact, had people wearing a mask and in turn limited the exposure to respiratory viruses. An article published in January 2022 explained how the Center for Disease Control (CDC) warned of a potential rise in influenza cases once again in the winter of 2022 coinciding with relaxed social distancing, mask wearing and return to normal school and business operations. According to the CDC, this is likely due to decreased immunity in people after being socially distanced for two years. In addition to this, the usual unpredictability of peak influenza season makes it difficult to decide whether the twindemic will actually take place.

The samples used in this experiment come from the University of North Carolina at Charlotte's on-campus Student Health Center. Since the goal of the experiment is to observe the effects of the twindemic on a college campus, negative COVID-19 tests that are directly from students or employees who self-report respiratory symptoms is the best way to measure this. Tests were obtained from September of 2022 through February of 2023. All tests showed PCR "negatives" for COVID-19 by the on-campus Student Health Center laboratory. Additionally, the tests were either negative for Influenza A and/or B, or were not further tested following the negative COVID-19 result. COVID-19 tests at the Student Health Center are conducted using Cepheid's $Xpert^{®}$ Xpress SARS-CoV-2/Flu plus diagnostic tests (Cepheid, Sunnyvale, CA)

[31]. This test is a rapid, real-time RT-PCR test that detects SARS-CoV-2 or Flu nucleic acid from a sample. The specimen sample comes from a human nasal swab and is loaded onto a cartridge. The cartridge contains lysis, binding, elution, and wash reagents as well as three various kinds of beads. These reagents are what is needed to lyse and break apart the nucleic acids in the sample. The cartridge containing the added sample is then loaded onto Cepheid's $GeneXpert^{®}$ Xpress Instrument. Then, results showing if the sample is positive for SARS-CoV-2 or Flu A/B are presented.

As shown in Table 1, we chose 90 samples spread over the course of a six month time period, starting from September 6, 2022 and ending on February 17, 2023. Six controls were included in the extraction process. The three negative controls are labeled "Extraction Negative", "SISPA Negative", and "Library Preparation Negative". The controls are composed of nuclease free water that is added at each respective step and treated as another sample. These controls were included to ensure no contamination occurred between samples during any of those three steps of sequencing, and were expected to have no detection of viral genetic material. Furthermore, three synthetic viral RNA controls were added as positive controls to confirm the accuracy and reliability of the metagenomic classification tools we used later on. These three controls were $Twist^{®}$ Bioscience's Synthetic Respiratory Virus controls, and were expected to be positive for each respective virus. They are labeled as "Influenza A Positive", "Influenza B Positive", and "Covid Positive" on Table 1.

2.2 RNA extraction

RNA was extracted using the Qiagen $QIAmp^{®}$ Viral RNA protocol [26] (Qiagen, Germantown, MD) for RNA extraction was used for a total of 58 samples throughout the course of the six month time observation period. Briefly, the $QIAmp^{®}$ Viral RNA protocol involves

adding 200 µL of the raw sample to a 2 mL tube. 800 µL of AVL buffer is added to the sample. This buffer aids in the lysing or rupturing of the cells found in the sample. The sample is then filtered in increments of 630 µL followed by two different washes, AW1 and AW2. This increases the purity of the final extracted RNA. The sample is resuspended with 60 uL buffer AVE. This protocol is outlined in Figure 1 below.

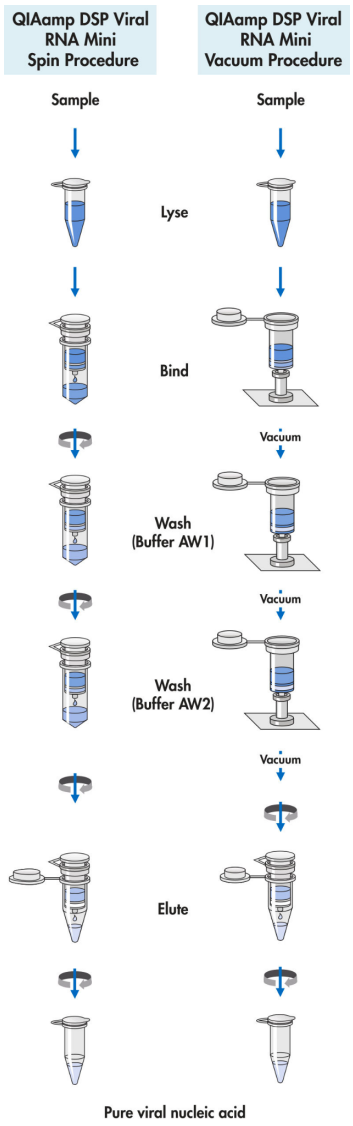


Figure 1: The above diagram was obtained from the Qiagen *QIAmp*® Viral RNA protocol [26] and shows the protocol in detail. For this experiment, samples were lysed and then absorbed into the membranes by centrifuging instead of by vacuum.

The second protocol used for RNA extraction was the Ceres $Nanotrap^{®}$ Microbiome A Automated Protocol (Ceres Nanosciences, Manassas, VA) [27] with $MagMAX^{™}$ kit and the $KingFisher^{™}$ Apex (Thermo Fisher Scientific, Waltham, MA). This protocol was used to extract a total of 32 out of 90 samples from September 2022 through February 2023. Samples extracted using the $KingFisher^{™}$ Apex are labeled in Table 1 with "KF". The Qiagen $QIAmp^{®}$ Viral RNA protocol and the Ceres $Nanotrap^{®}$ Microbiome A Automated Protocol is that the latter is mostly automated, while the former involves manual transfer of reagents and samples. The $KingFisher^{™}$ Apex instrument works by adding the samples and reagents to a plate of 96 wells. Along with the sample plate, a 96-well plate containing 1000 μL per well of 80% ethanol solution and a 96-well plate containing 500 μL per well of a wash buffer are prepared. The sample plate is prepared by adding 400 μL of raw sample along with a series of reagents.

2.3 Sequencing RNA

In this experiment, the extracted samples were sequenced using on the $Illumina^{®}$ NextSEq 2000 (Illumina, Inc., San Diego, CA) [28] using the $Twist^{®}$ Bioscience Respiratory Viral Research Panel protocol [48] (Twist Bioscience, South San Francisco, CA). First, RNA was amplified using a sequence-independent, single-primer amplification process known as SISPA. For this experiment, we used the SISPA protocol with reagents and steps detailed by Moreno and O'connor from the University of Wisconsin-Madison [45].

Library prep was then completed using $Twist^{®}$ Bioscience Library Preparation EF Kit 2.0 for ssRNA Virus Detection protocol [46]. Steps 1 and 2 were omitted due to the creation of cDNA in the SISPA step, but fragmentation, end-repair, dA-tailing, adapter ligation, and amplification were performed according to the protocol. The DNA libraries were then subject to

target enrichment and hybridization, following Twist Bioscience's Target Enrichment Standard Hybridization v2 protocol [47]. Targets used in this experiment were from the *Twist*® Respiratory Viral Research Panel [42], which includes 29 common human respiratory viruses (Fig. 2). In brief, these targets were added to the samples after the library preparation step was completed and were fused or "hybridized" with the input DNA. Targets attach to complementary segments on the input DNA and are then amplified as part of the target enrichment process.

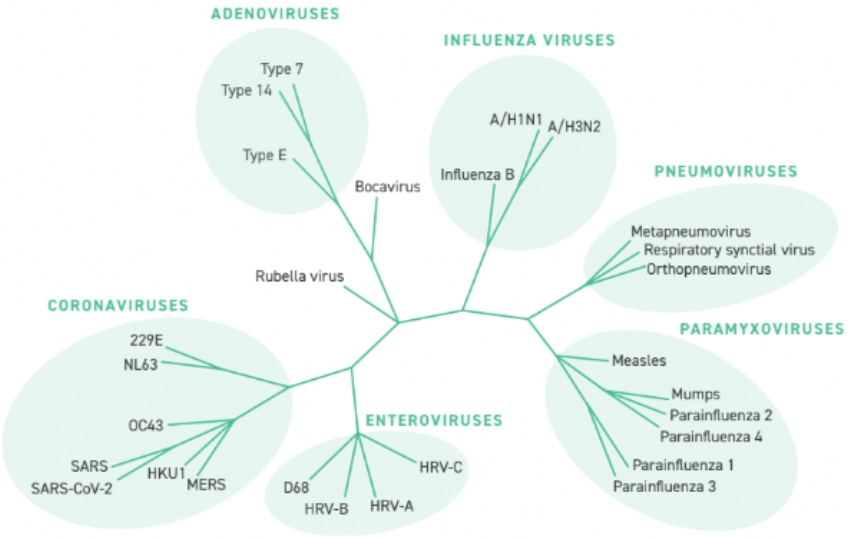


Figure 1. Taxonomic tree of viruses covered on the Twist Respiratory Virus Research Panel spanning the major respiratory viral clades.

Figure 2: The above diagram shows the 29 different targets used in this experiment. All viral targets were from the 'Twist Respiratory Virus Research Panel' from *Twist*® Bioscience.

As a means to check the quality of the samples and ensure there were enough amounts of genetic material to continue to the next steps, the DNA concentration of samples was checked after completion of RNA extraction, SISPA, library preparation, and hybridization of targets using the *ThermoFisher*®Qubit 4 Fluorometer (Thermo Fisher Scientific, Waltham, MA) [49] and the *Agilent*®TapeStation (Agilent Technologies, Santa Clara, CA) [29] quality control

standard protocols described by their respective manufacturers. $ThermoFisher^{®}$Qubit 4 Fluorometer was used after the completion of RNA extraction, SISPA, library preparation, and target hybridization to check the amount of RNA or DNA in the samples depending on the library preparation step. Specifically, the Qubit™ 1X dsDNA High Sensitivity and Broad Range Assay kit was used, as our genetic material had been converted to cDNA at this point after the completion of SISPA. The $Agilent^{®}$ High Sensitivity D1000 ScreenTape® assay was used after target enrichment was completed before samples were loaded onto the flow cell. The instrument used to quantify samples was the $Agilent^{®}$ 4150 TapeStation, which allowed for loading 16 samples (prepared libraries) at once (see Appendix A for library quantification results). All samples were pooled together regardless of DNA concentration values returned after running quality control methods, and subsequently loaded onto the flow cell following Illumina's Custom DNA Prep protocol. The prepared library was then sequenced using the $Illumina^{®}$ NextSeq 2000 instrument. The flow cell used was the NextSeq 1000/2000 P1 flow cell, and the NextSeq instrument ran for a total of 322 cycles on our samples to obtain reads in both directions.

2.4 Processing Sequencing Results

After sequencing was completed using the $Illumina^{®}$ NGS protocol, the resulting reads were processed as follows. Raw reads in the form of FASTQ files were transferred from the NextSeq 2000 instrument to a local directory on our computer. Further manipulation of the FASTQ files was completed off the machine on our personal computers. Two metagenomic classification softwares were used to classify the samples, Kraken and CLARK. Figures 3 and 4 outline the different parameters for running Kraken and CLARK. In Figure 3, a for loop is shown to loop through all output FASTQ files from our sequencing run. In the command calling the

Kraken tool, the database containing viral reference sequences and species names is specified. Samples are specified as paired-end, with 36 threads. Samples unclassified by the software are labeled as such, as are samples that were classified. An output file is generated for all classified samples as well as a report file. The database used to map sample IDs to the corresponding scientific name was built prior to the classification and stored in a file path in a different directory. The taxonomic information in the Kraken database comes from the Reference Sequence Database from the National Center for Biotechnology Information (NCBI). Since the NCBI database updates periodically, for the purpose of our experiment we built and accessed the database on April 10, 2023. After results were classified using Kraken, abundance estimates were calculated using the Bracken software. Bracken (Bayesian Re-estimation of Abundance with KrakEN) is a supplementary tool used after Kraken classification has been completed to compute the abundance of species in a given sample of genetic material [30]. Abundance of a species is calculated as the amount of reads that were able to be classified as a specific species from the total reads obtained as the result of sequencing a sample. Bracken uses the classification assigned by Kraken to compute the relative abundances of each organism (in this experiment, virus) found in the samples.

In Figure 4, a function is shown that details how CLARK was used to classify the samples. First, each sample FASTQ file containing the left and right reads is converted to a text file for each paired-end read. Each sample is then classified using the classify_metagenome.sh CLARK script, with the "-P" flag representing the fact that the input text files are paired-end reads. Finally, abundance estimates per sample are calculated, with the value of "10" representing the percentage abundance estimate threshold we want to observe.

```
module load kraken2
module load blast/2.11.0+

fastqs_dir=/projects/gibas_lab/UNCC-viral-panel/combined_samples/WW_CN_032323/clinical_samples

mkdir -p /scratch/laboufou/kraken_CN_out/

for i in `ls $fastqs_dir/* | grep R1` #for file in fastqs_dir directory that contain pattern of characters R1
do
name=`basename $i | cut -d_ -f1` #name variable is set to basename (file name without path) with delimiter "_" and
extracting first field (f1)
kraken2 \
    --db /scratch/laboufou/utilities/kraken2-custom-DB \
    --threads 36 \
    --paired \
    --classified-out "/projects/gibas_lab/UNCC-viral-
panel/combined_samples/WW_CN_032323/kraken_CN_out/classified_CN_reads_${name}#.fastq" \
    --unclassified-out "/projects/gibas_lab/UNCC-viral-
panel/combined_samples/WW_CN_032323/kraken_CN_out/unclassified_CN_reads_${name}#.fastq" \
    --output "/scratch/laboufou/kraken_CN_out/CN_krakenout_${name}.txt" \
    --report "/scratch/laboufou/kraken_CN_out/CN_krakenreport_${name}" \
    $i ${i/R1/R2}
done
```

Figure 3: The code snippet above shows the commands used to classify the paired end reads yielded from the sequencing run. The outputs yielded by the Kraken classification were then input to the Bracken software to calculate relative abundances.

```
function analyze_sample() # sample
{
    sample="$1"
    echo "anlyzing sample $sample"
    name="$(echo ${sample} | cut -d_ -f1)"
    if [[ -z "${name}" ]]; then echo "sample ${sample} has no name"; exit 1; fi

    echo "combining R1 for sample ${sample}"
    cat /projects/gibas_lab/UNCC-viral-panel/combined_samples/WW_CN_032323/wastewater_samples/${sample}_R1_001.fastq >
${clark_outdir}/${name}_WW_samples.R.txt

    echo "combining R2 for sample ${sample}"
    cat /projects/gibas_lab/UNCC-viral-panel/combined_samples/WW_CN_032323/wastewater_samples/${sample}_R2_001.fastq >
${clark_outdir}/${name}_WW_samples.L.txt

    echo "classifying metagenome for sample ${sample}"
    classify_metagenome_results=${clark_outdir}/clark_results_tables/${name}_WW_CLARK_results
    ./classify_metagenome.sh -P ${clark_outdir}/${name}_WW_samples.{R,L}.txt  -R ${classify_metagenome_results}

    echo "estimating abundance for sample ${sample}"
    ./estimate_abundance.sh -a 10 -F ${classify_metagenome_results}.csv -D /projects/enviro_lab/software/clark_viruses_db/ >
${abundance_results}/${name}_abundances_results.csv
}
```

Figure 4: The code snippet above shows the commands used to classify the sequencing results using CLARK software.

CHAPTER 3: RESULTS AND DISCUSSION

3.1 Analysis of Samples Using Kraken and Bracken

Samples were collected from the UNC Charlotte Health Center starting on September 6, 2022 and ending on February 17, 2023. Out of all 90 clinical negative samples input onto the flow cell, all except four yielded results after the Bracken abundance estimates were completed. Two samples were sequenced but were not included in the analysis or visualization of this experiment due to not having an exact collection date. Therefore, 84 samples were used in the analysis of this experiment.

CLARK and Kraken classify samples in similar ways. Both softwares require building a database, sample classification, and abundance estimation. First, the viral database was built, using information from the NCBI Reference Sequence Database. Then, the classification is run on the paired end samples and abundance estimates are calculated. This results in a file containing the viral species identified in each sample, along with its relative abundance in each sample.

Table 1 shows all samples that were sequenced, including the date of sample collection. In Table 2, the results of sequencing are shown. Each sample is shown with its corresponding collection date. The four samples that did not produce any results after Bracken abundance estimates are shown in bold on Table 2, despite some yielding a high number of reads. The three positive controls (Flu A, Flu B, and Cov Positive) were used to observe if the metagenomic classification tools we used would correctly identify each virus. The three negative nuclease free water (NFW) controls were used to ensure no contamination occurred during each of the three main processes of sequencing: RNA extraction, SISPA, and library preparation.

Sample Layout

	1	2	3	4	5	6
A	9/6/2022_1	9/29/2022_4	10/14/2022_1	11/7/2022_1	12/5/2022_2	1/2/2023_1
B	9/27/2022_1	9/30/2022_1	10/17/2022_1	11/8/2022_1	12/6/2022_1	1/3/2023_1_KF
C	9/27/2022_2	10/3/2022_1	10/20/2022_1	11/9/2022_1	12/7/2022_1	1/6/2023_1
D	9/27/2022_3	10/5/2022_1	10/24/2022_1	11/10/2022_1	12/8/2022_1	1/6/2023_2
E	9/28/2022_1	10/6/2022_1	11/1/2022_1	11/11/2022_1	12/12/2022_1	1/9/2023_1_KF
F	9/29/2022_1	10/7/2022_1	11/2/2022_1	11/28/2022_1	12/13/2022_1	1/10/2023_1
G	9/29/2022_2	10/10/2022_1	11/3/2022_1	12/1/2022_1	12/20/2022_1	1/10/2023_2
H	9/29/2022_3	10/12/2022_1	11/4/2022_1	12/5/2022_1	12/21/2022_1	1/10/2023_3

	7	8	9	10	11	12
A	1/11/2023_1	2/14/2023_1	2/7/2023_1	10/4/2022_2_KF	Date N/A	1/27/2022_1_KF
B	1/13/2023_1	2/16/2023_1	2/6/2023_1	10/5/2022_2_KF	12/8/2022_2_KF	1/25/2022_2_KF
C	Influenza A Positive	2/16/2023_2	9/20/2022_1_KF	10/12/2022_2_KF	12/9/2022_1_KF	1/25/2022_3_KF
D	Influenza B Positive	2/15/2023_1	9/26/2022_1_KF	10/17/2022_2_KF	Date N/A	2/15/2022_2_KF
E	Covid Positive	2/14/2023_2	9/26/2022_2_KF	11/3/2022_2_KF	12/14/2022_1_KF	2/17/2022_3_KF
F	Extraction Negative	2/17/2023_1	9/26/2023_3_KF	11/3/2022_3_KF	12/20/2022_2_KF	2/16/2022_3_KF
G	SISPA Negative	2/17/2023_2	9/28/2022_2_KF	11/4/2022_2_KF	1/25/2022_1_KF	2/14/2022_3_KF
H	LibPrep Negative	2/10/2023_1	10/4/2023_1_KF	11/7/2022_2_KF	1/24/2022_1_KF	2/17/2022_4_KF

Table 1: Plate layout for sequencing. Each cell contains a unique sample identified by date. Multiple samples collected from the same date are uniquely identified by a "1", "2", "3", or "4". Samples extracted with the *KingFisher*™ Apex are labeled with a "KF"; all other samples were extracted using the Qiagen *QIAmp*® Viral RNA protocol.

Sequencing Results

Sample Date	Raw Read Counts	Number of Reads Classified by Kraken	Number of Reads Classified by CLARK
9/26/2022_3	248732	1164	0
9/26/2022_2	211422	843	0
9/26/2022_1	175150	6530	0
9/27/2022_1	119566	573	0
9/27/2022_2	175214	1326	0
9/6/2022_1	667392	23527	211322

9/27/2022_3	484254	82958	119830
9/28/2022_2	451038	48983	34661
9/28/2022_1	110332	3108	0
9/20/2022_1	187294	1089	0
9/29/2022_2	149940	1207	0
9/29/2022_4	89598	547	0
9/29/2022_1	106142	736	0
9/29/2022_3	90330	779	0
9/30/2022_1	1432426	510454	482573
10/3/2022_1	91490	848	0
10/4/2022_2	2862	74	0
10/4/2022_1	114388	390	0
10/5/2022_1	110232	843	0
10/5/2022_2	13254	27	729
10/6/2022_1	169470	19989	23889
10/7/2022_1	95936	3453	0
10/10/2022_1	112108	1572	0
10/12/2022_1	314424	85173	79745
10/12/2022_2	7796	30	0
10/14/2022_1	1025802	65788	52851
10/17/2022_1	16858562	7377164	6533194
10/17/2022_2	12984	19	0
10/20/2022_1	373832	2898	0
10/24/2022_1	2623220	438398	227740
11/1/2022_1	1351826	280176	266217
11/2/2022_1	566122	4626	0
11/4/2022_2	14628	44	0
11/3/2022_3	17064	20	0

11/3/2022_2	4728	60	0
11/3/2022_1	646770	4994	0
11/4/2022_1	292916	2524	0
11/7/2022_2	13002	51	722
11/7/2022_1	334688	2680	0
11/8/2022_1	300064	2663	0
11/9/2022_1	626006	6517	0
Date N/A	47346	362	0
11/10/2022_1	0	0	0
11/11/2022_1	1403130	440548	393677
1/9/2023_1	1036902	512260	512157
11/28/2022_1	1401990	189094	211594
12/1/2022_1	378494	923	0
12/5/2022_1	131424	118	7544
12/5/2022_2	155938	3085	0
12/6/2022_1	5965084	151824	0
12/7/2022_1	318730	1137	0
12/8/2022_1	381418	1323	0
12/8/2022_2	53156	462	0
12/9/2022_1	38762	222	0
12/12/2022_1	1895718	793946	768055
12/13/2022_1	937820	4210	0
Date N/A	13752	62	0
12/14/2022_1	21684	122	0
12/20/2022_2	47196	285	0
12/20/2022_1	621202	1385	0
12/21/2022_1	1285326	11905	0
1/3/2023_1	2400	0	0

1/2/2023_1	16	0	0
1/6/2023_2	3667612	1819694	1819089
1/6/2023_1	8398	10	0
1/10/2023_1	15580	4693	4962
1/10/2023_3	44626	153	0
1/10/2023_2	28426	18	0
1/11/2023_1	42374	1274	0
1/13/2023_1	22534	126	0
1/24/2023_1	59916	340	0
1/25/2023_3	58602	274	0
1/25/2023_1	6402	14	0
1/25/2023_1	6216	18	0
1/27/2023_1	5202	0	0
2/6/2023_1	850254	292409	280607
2/7/2023_1	175520	3136	0
2/10/2023_1	86700	15268	13977
2/14/2023_1	11800	47	0
2/14/2023_3	14650	2900	2034
2/14/2023_2	17664	132	0
2/15/2023_1	11786	55	0
2/15/2023_2	8180	726	851
2/16/2023_3	6104	17	0
2/16/2023_1	9678	48	0
2/16/2023_2	12300	87	0
2/17/2023_1	29122	108	0
2/17/2023_4	3182	13	0
2/17/2023_2	211630	14180	12738
2/17/2023_3	3806	13	0

Covid Positive	190	92	49
Influenza A Positive	29222	35	0
Influenza B Positive	35752	50	0
Library Preparation Negative	166	12	79
SISPA Negative	39084	1235	0
Extraction Negative	26516	110	0

Table 2: Sequencing details for all samples. Each sample is listed with its corresponding collection date, along with how many reads were yielded from the sequencing run using the NextSeq 2000. These values were already trimmed after being checked for quality with the built-in Illumina DRAGEN (Dynamic Read Analysis forGENomics) analysis software. The number of reads classified by Kraken or CLARK or both is also shown.

All controls yielded classification results from Kraken, while only the “Covid_Positive” and “Library_Preparation_Negative" controls yielded classification results using CLARK. Observing Figures 5 and 6, it is evident that some contamination occurred across positive and negative controls. Kraken and CLARK showed different classifications for the “Covid_Positive” control, as CLARK classified it as SARS-related coronavirus, which was expected, while Kraken classified it as a different virus altogether. Contamination could have occurred at any step during the sequencing process. Using one 96-well plate to sequence all our samples was likely a factor in making contamination so widespread amongst all controls.

Kraken Detection of Respiratory Viruses in Positive and Negative Controls

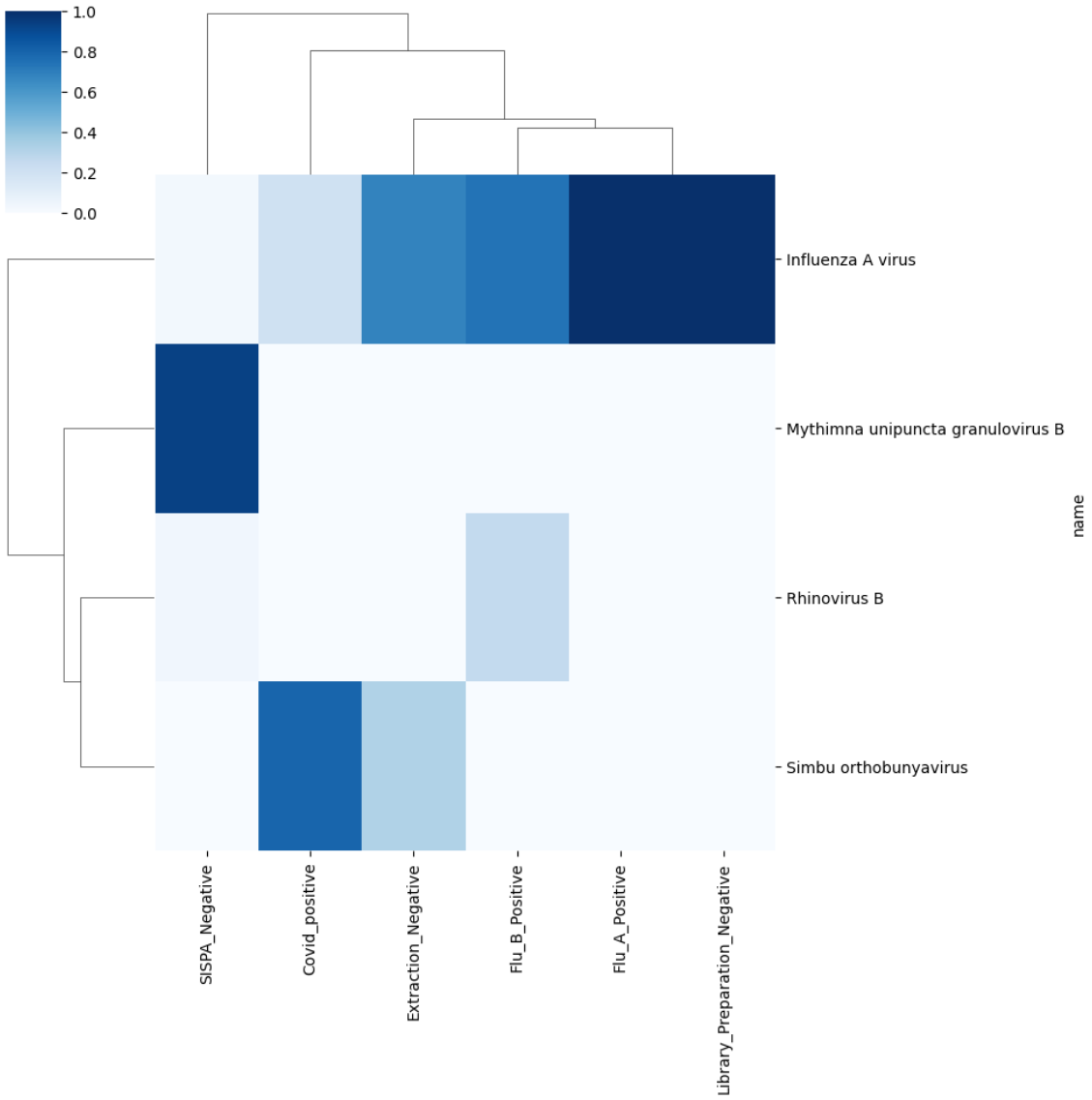


Figure 5: The above heatmap shows detection of viruses using the Kraken metagenomic across all six controls. The key in the upper right corner shows how the varying levels of color represent the percentage of abundance each virus was detected in each sample.

CLARK Detection of Respiratory Viruses in Positive and Negative Controls

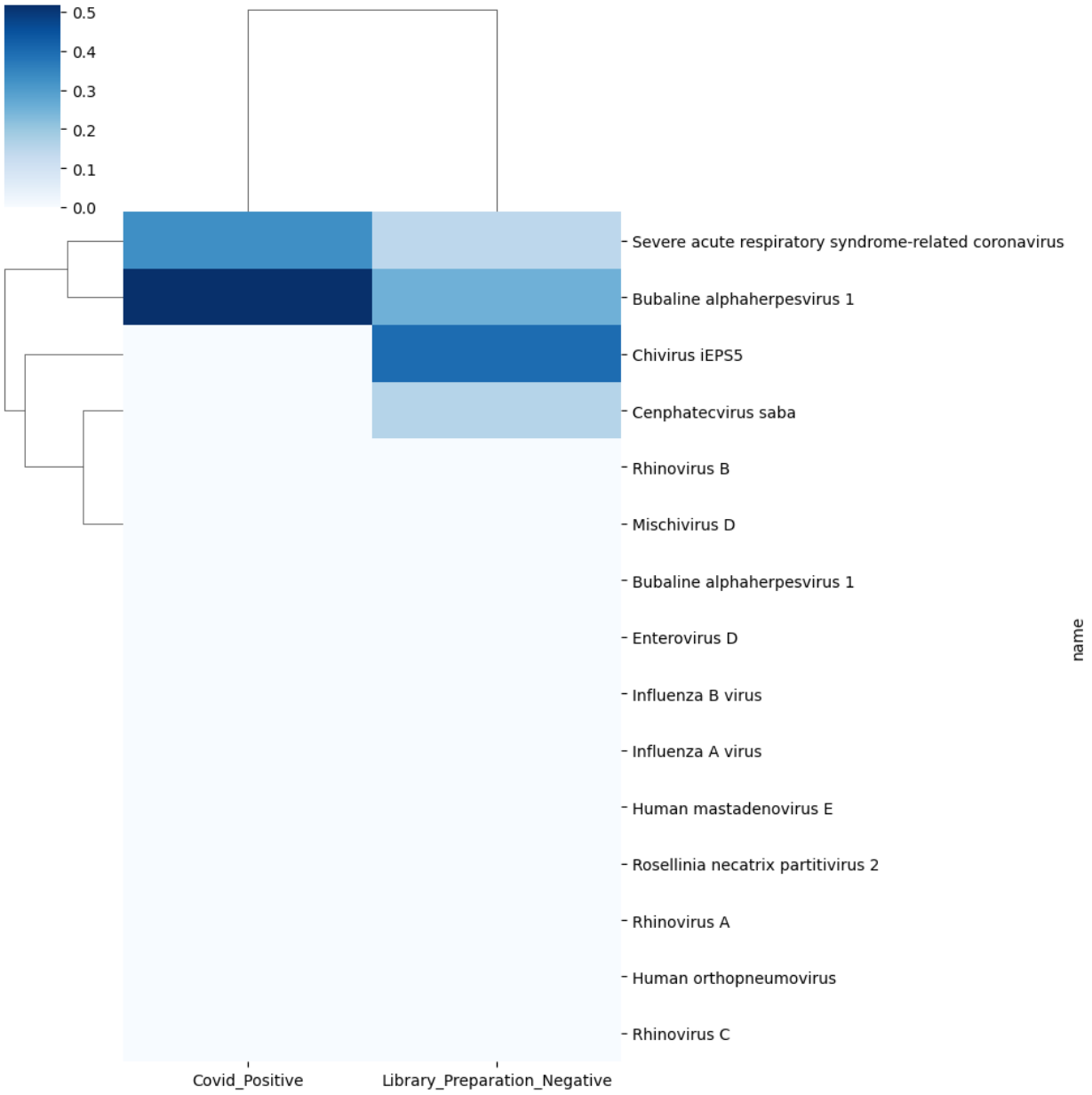


Figure 6: Showing detection of viruses using the CLARK metagenomic classifier across two controls. The rest of the controls were not classified by CLARK. The key in the upper right corner shows how the varying levels of color represent the percentage of abundance each virus was detected in each sample.

All viruses with a relative abundance estimate of less than 10% were removed from the following visualizations. This pruning step removes visual skewing for low abundance and provides results that show only the viruses that were highly abundant across all samples. It is

important to note that Kraken has an issue with classifying taxa that may not actually be in the samples, a phenomenon known as “phantom taxa”. A study completed at UNC-Charlotte showed that the Kraken algorithm would classify samples as taxa that did not actually exist in those samples [40]. In order to combat this phenomenon, all viruses with less than 10% abundance were removed from results yielded by Kraken before analysis.

Figures 7-9, show that *BeAn* 58058 virus was commonly detected in samples from September through November. SARS-related coronavirus was also commonly detected in samples from those months. Out of fifteen samples observed from September, all showed some detection of *BeAn* 58058, with six of those samples showing a relative abundance of over 50%, making it the most abundant virus in those samples. In October, eleven out of the fifteen samples observed showed some detection of *BeAn* 58058, with three of those samples showing a relative abundance of over 70% for that virus. In November, the relative abundance of *BeAn* 58058 rose, as it was detected in nine out of thirteen samples from that month, with six of those nine samples having a relative abundance of higher than 70%. Three samples in December showed average abundance for this virus, all being below 40%. *BeAn* 58058 is a virus originally found in an *Oryzomis* rodent in Brazil [32] and is able to infect both humans and animals. It is also found in post-mortem COVID-19 patients [33]. Furthermore, it is a variant of *Vaccinia virus*, which is highly genetically similar to the virus that causes smallpox [34]. Figure 7 shows the number of samples that have high detection of *BeAn* 58058 per month. It shows how the number of cases peaks in September and begins to steadily decrease throughout December through February.

Number of Samples with Abundance of *BeAn* 58058 Greater than 10%

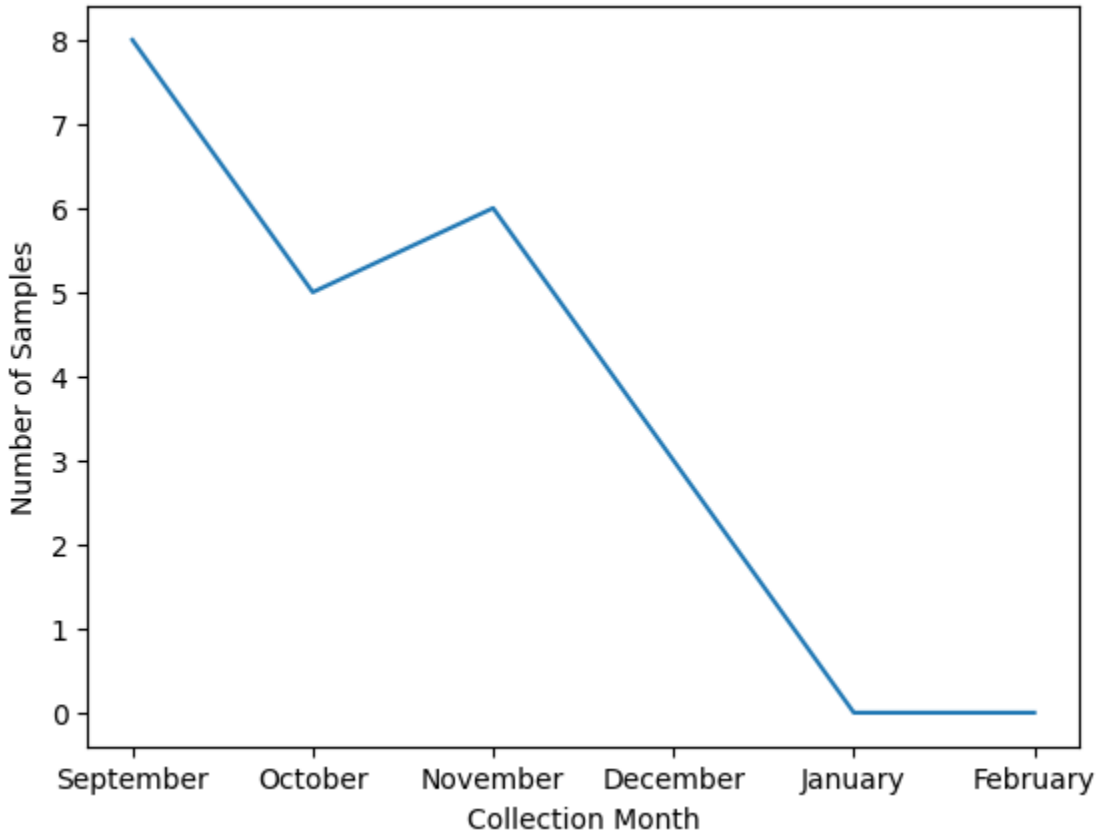


Figure 7: The number of samples with an abundance estimate of over 10% for *BeAn* 58058 across the six month observation period is shown in the Figure above.

In Figures 9-12, a high detection of severe acute respiratory syndrome-related coronavirus is shown. In September, three samples showed an abundance estimate of higher than 20% for SARS-CoV. In October, six samples showed an abundance estimate of over 20%, with four samples showing an abundance estimate of 100%. In November, four samples showed an abundance estimate percentage of 50% or higher, including two which showed a 100% abundance estimate for this virus. Severe acute respiratory syndrome-related coronavirus is a species of virus that consists of the strain of coronavirus that is responsible for causing the COVID-19 illness. While this might be considered surprising as the samples collected tested negative for SARs-CoV-2, we have to consider that the probe sets were designed to detect not

only Covid-19 but also other seasonable coronaviruses. Another explanation is that the sample testing presented a false negative and the individual was actually positive for SARS-CoV-2. Figure 8 shows the number of samples that showed abundance estimates for SARS-related coronavirus across all six months. It is clear that most months had a high number of samples with SARS-Cov detection. This confirms the fact that despite other viruses being in circulation on campus, coronaviruses are still a large factor in causing students and faculty to be ill.

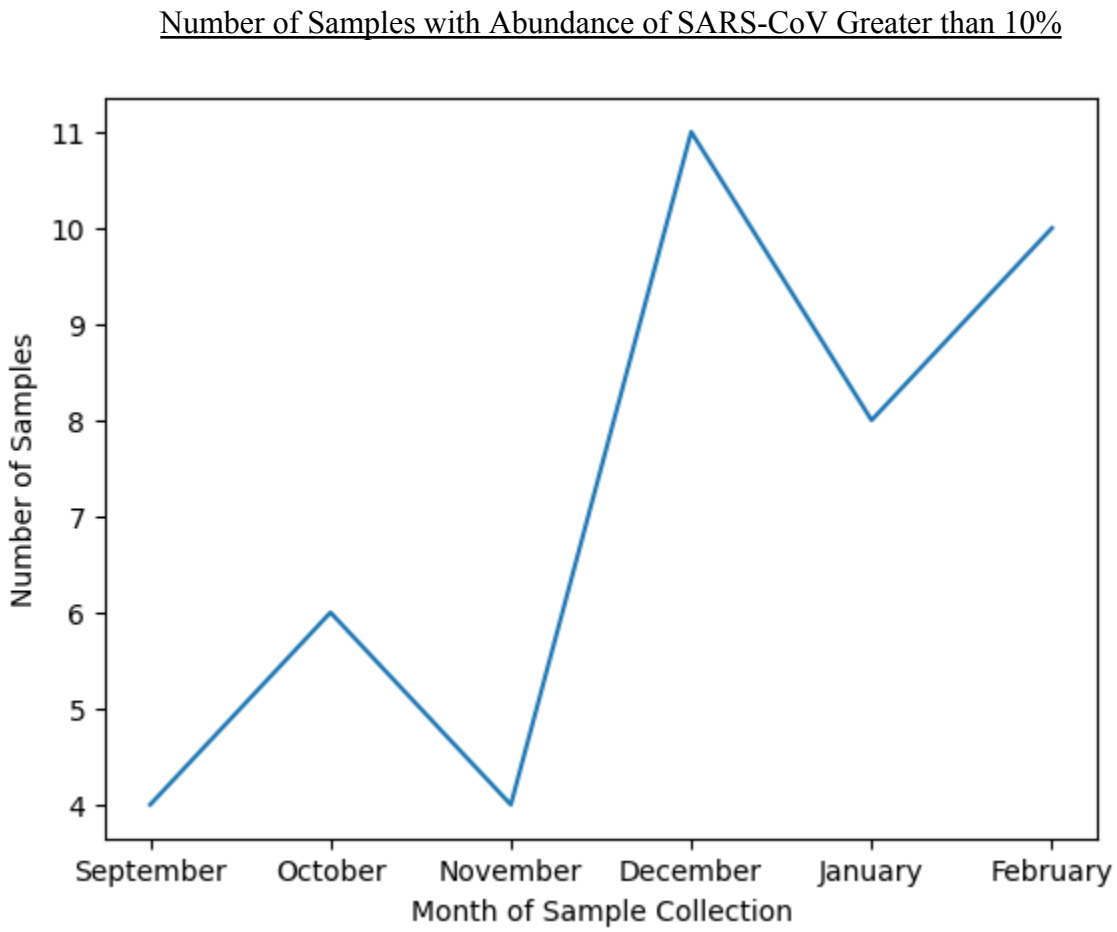


Figure 8: The number of samples with an abundance estimate of over 10% for SARS-related coronavirus across the six month observation period is shown in the Figure above.

Kraken Detection of Respiratory Viruses in Clinical Negative COVID-19 PCR Tests - September 2022

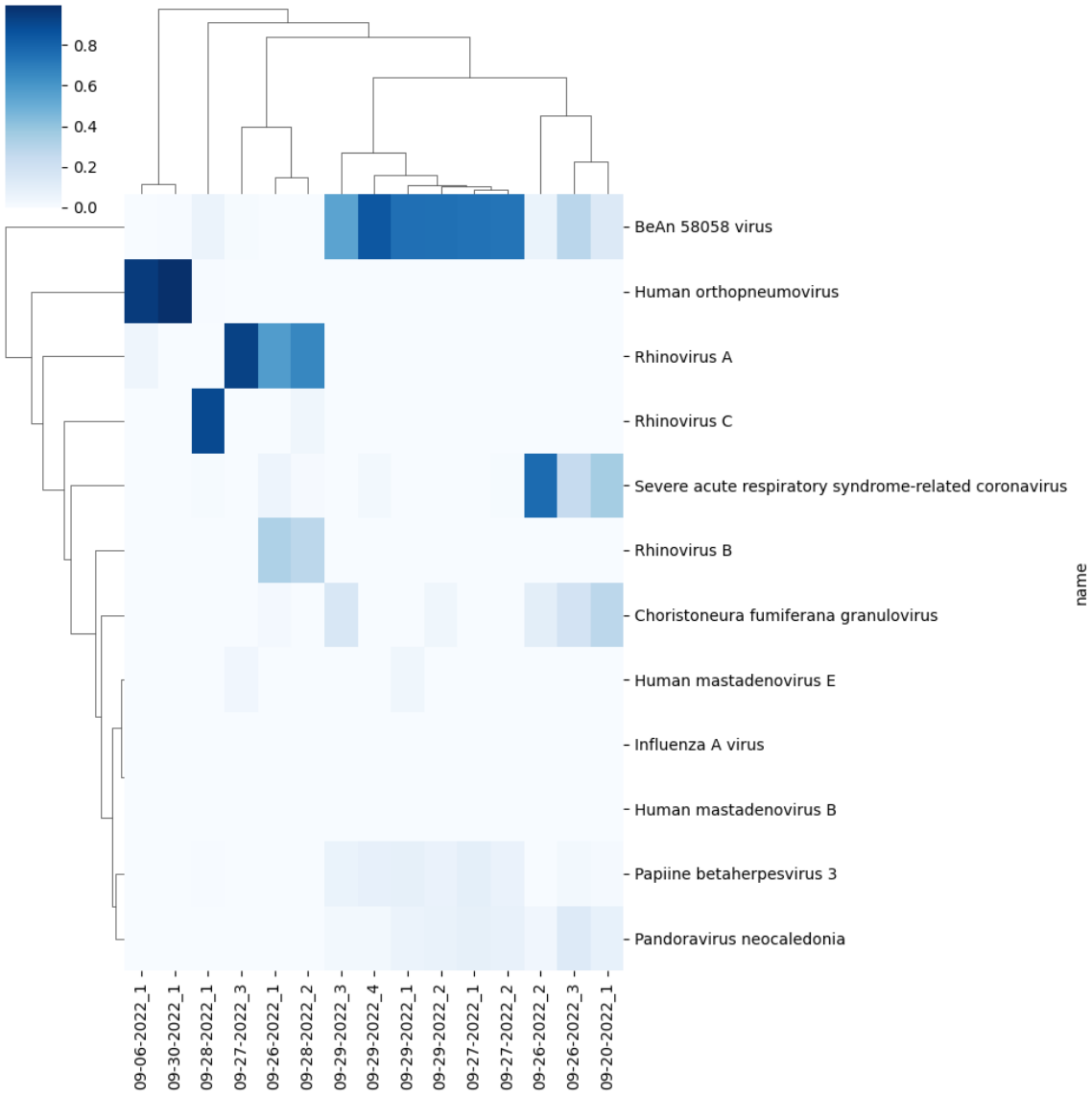


Figure 9: The above heatmap shows detection of viruses using the Kraken metagenomic classifier with more than 10% abundance per sample across samples collected from September 2022. Each tick on the x-axis represents a sample, labeled with the specific date it was collected from. The key in the upper right corner shows how the varying levels of color represent the percentage of abundance each virus was detected in each sample.

Kraken Detection of Respiratory Viruses in Clinical Negative COVID-19 PCR Tests - October 2022

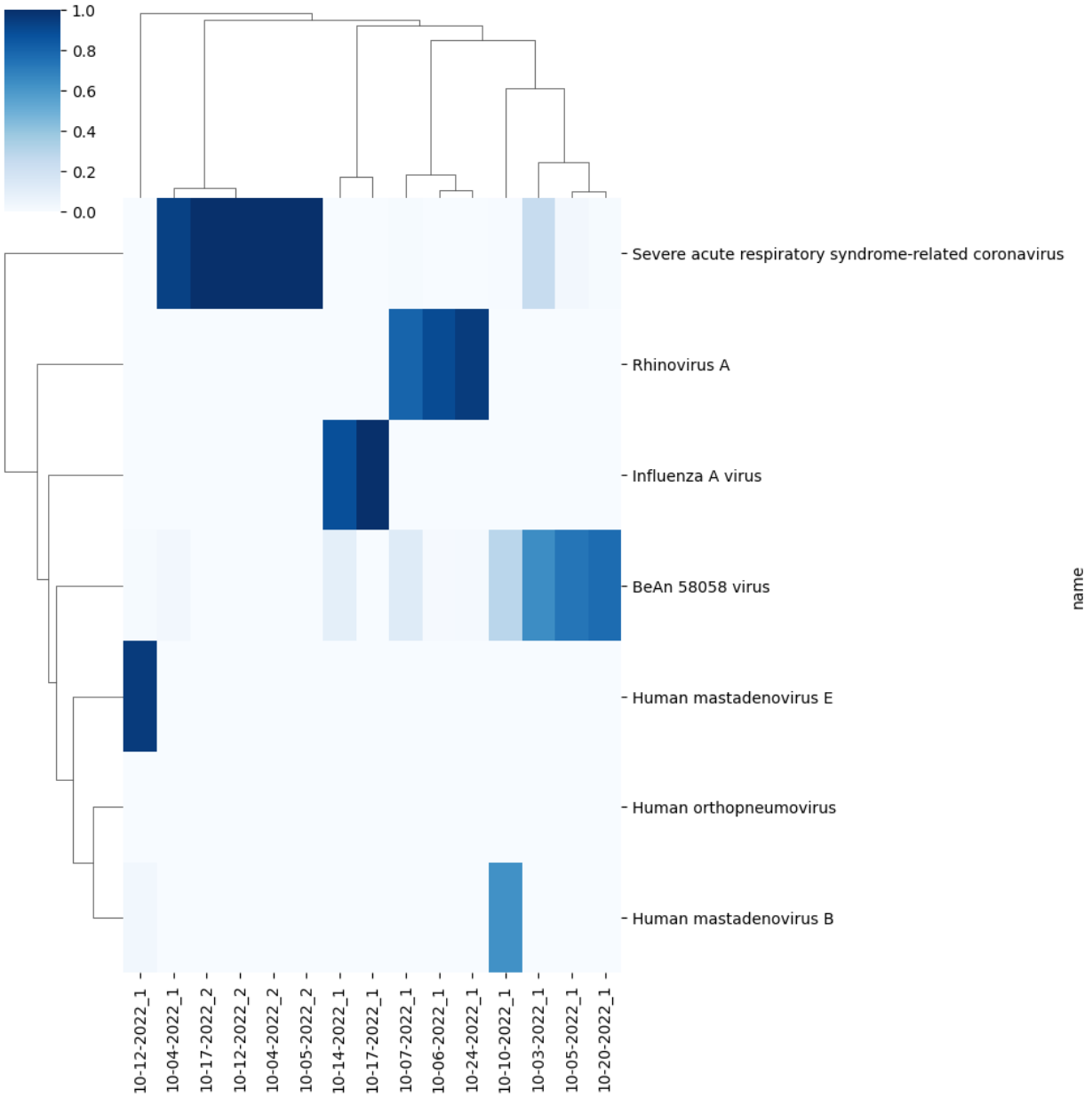


Figure 10: The above heatmap shows detection of viruses using the Kraken metagenomic classifier with more than 10% abundance per sample across samples collected from October 2022. Each tick on the x-axis represents a sample, labeled with the specific date it was collected from. The key in the upper right corner shows how the varying levels of color represent the percentage of abundance each virus was detected in each sample.

Kraken Detection of Respiratory Viruses in Clinical Negative COVID-19 PCR Tests - November 2022

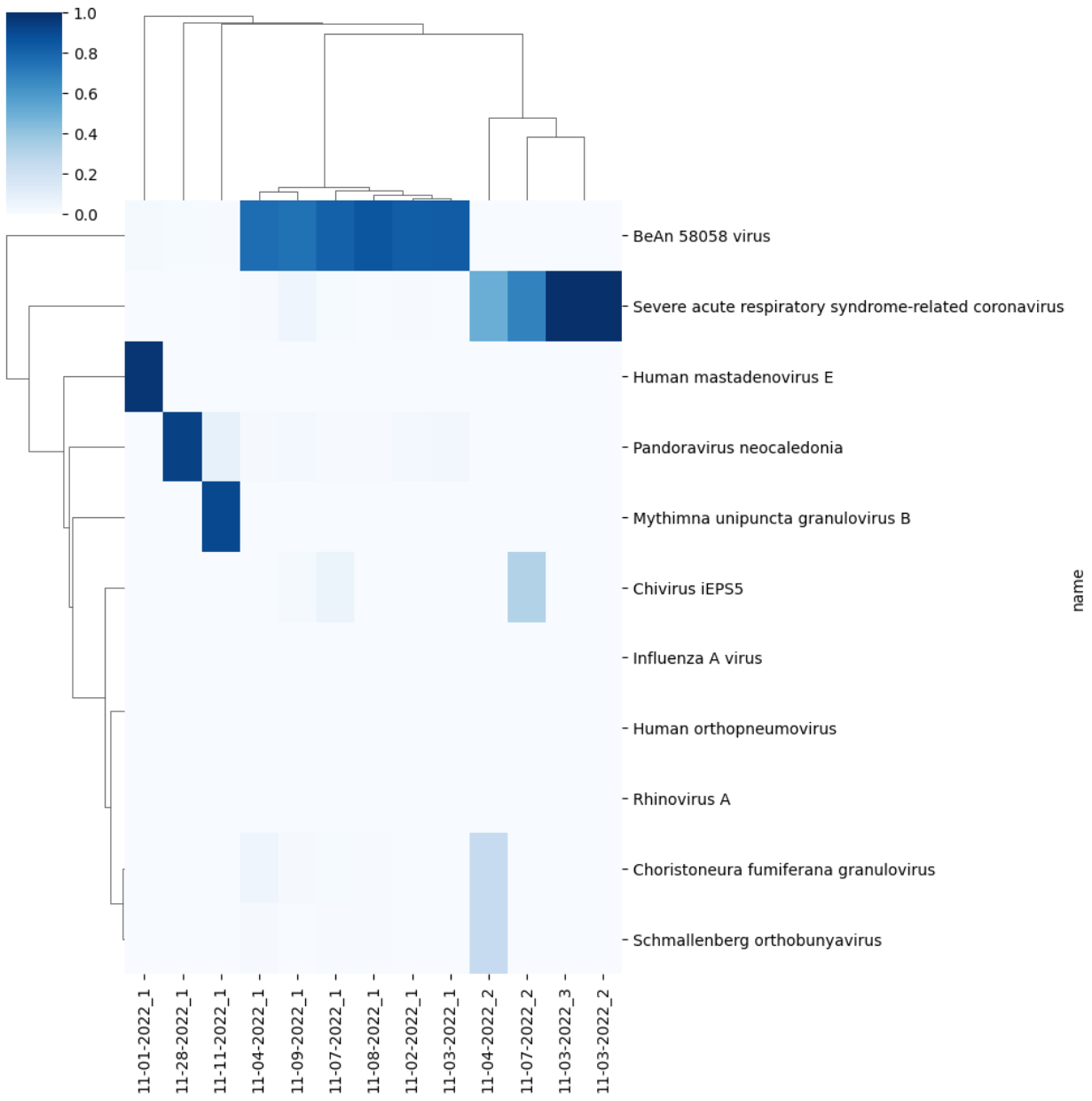


Figure 11: The above heatmap shows detection of viruses using the Kraken metagenomic classifier with more than 10% abundance per sample across samples collected from November 2022. Each tick on the x-axis represents a sample, labeled with the specific date it was collected from. The key in the upper right corner shows how the varying levels of color represent the percentage of abundance each virus was detected in each sample.

Kraken Detection of Respiratory Viruses in Clinical Negative COVID-19 PCR Tests - December 2022

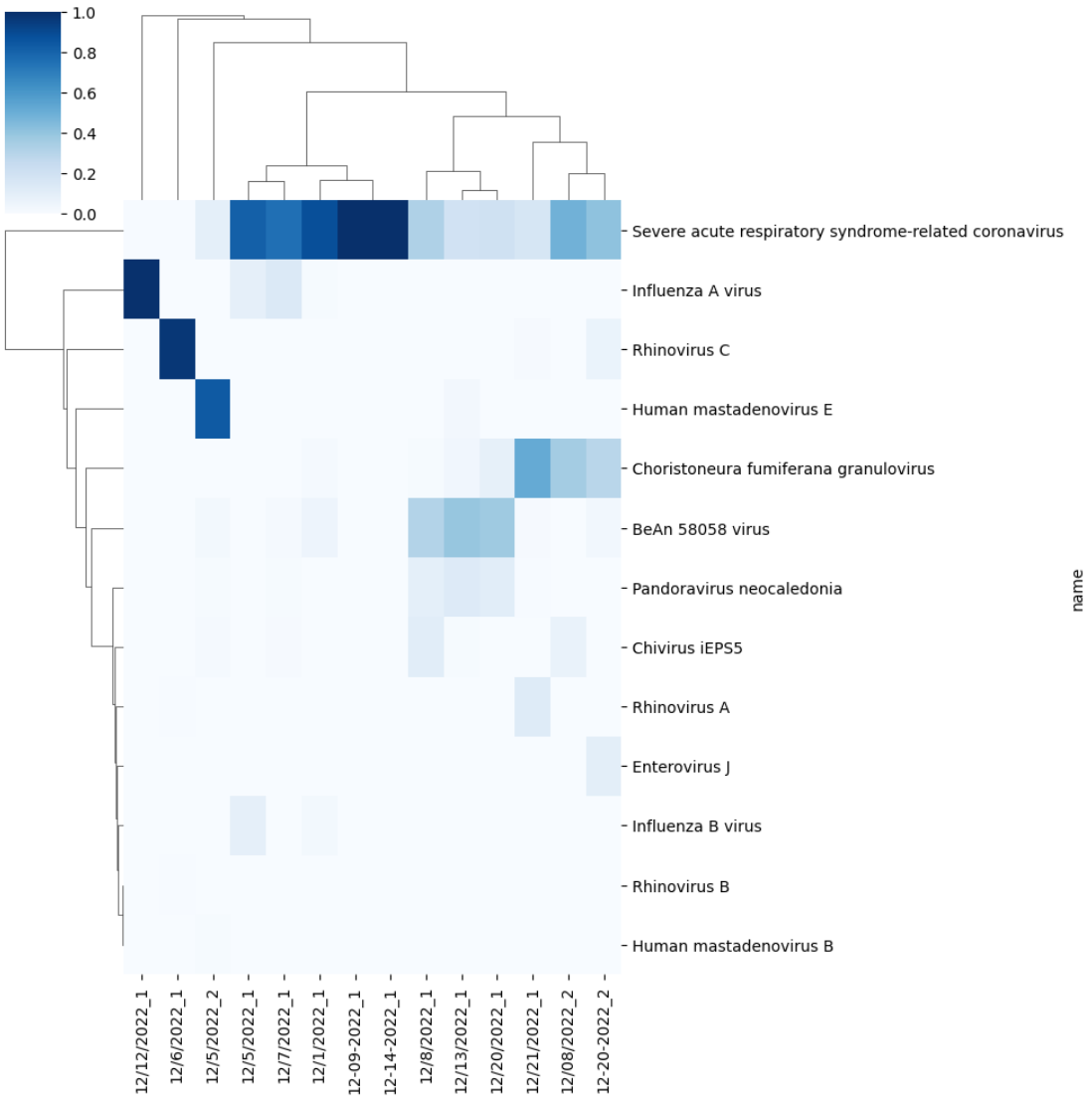


Figure 12: The above heatmap shows detection of viruses using the Kraken metagenomic classifier with more than 10% abundance per sample across samples collected from December 2022. Each tick on the x-axis represents a sample, labeled with the specific date it was collected from. The key in the upper right corner shows how the varying levels of color represent the percentage of abundance each virus was detected in each sample.

Another virus common to samples collected from September and October is rhinovirus A. Three samples from both October and September showed an abundance of higher than 20% for rhinovirus A. This virus is a member of the *Enterovirus* genus and is responsible for

respiratory tract infections that can potentially hospitalize children [35]. Rhinovirus A was detected in November, but with a low abundance, while another species. Rhinovirus C detected in one sample each from September and December is another species of the same virus. However, it has been found that children who are infected with rhinovirus C may experience more severe respiratory infection than those infected with other strains. In general, infections caused by HRVs (human rhinoviruses) are referred to as the “common cold” due to the mild respiratory infections they usually cause. Figures 13 and 14 show the number of samples that showed an abundance of rhinovirus and other viruses from the *Enterovirus* genus.

Number of Samples with Abundance of Rhinovirus Greater than 10%

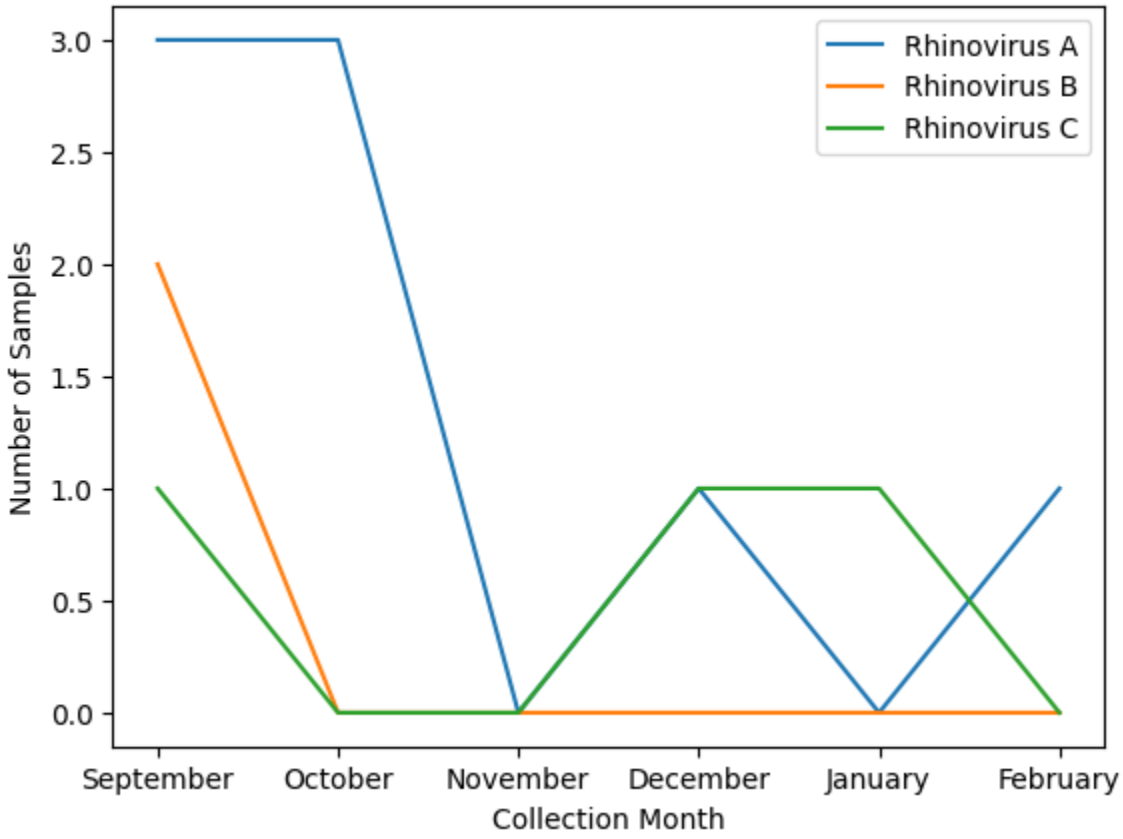


Figure 13: The number of samples with an abundance estimate of over 10% for three species of rhinovirus across the observed six months is shown in the Figure above.

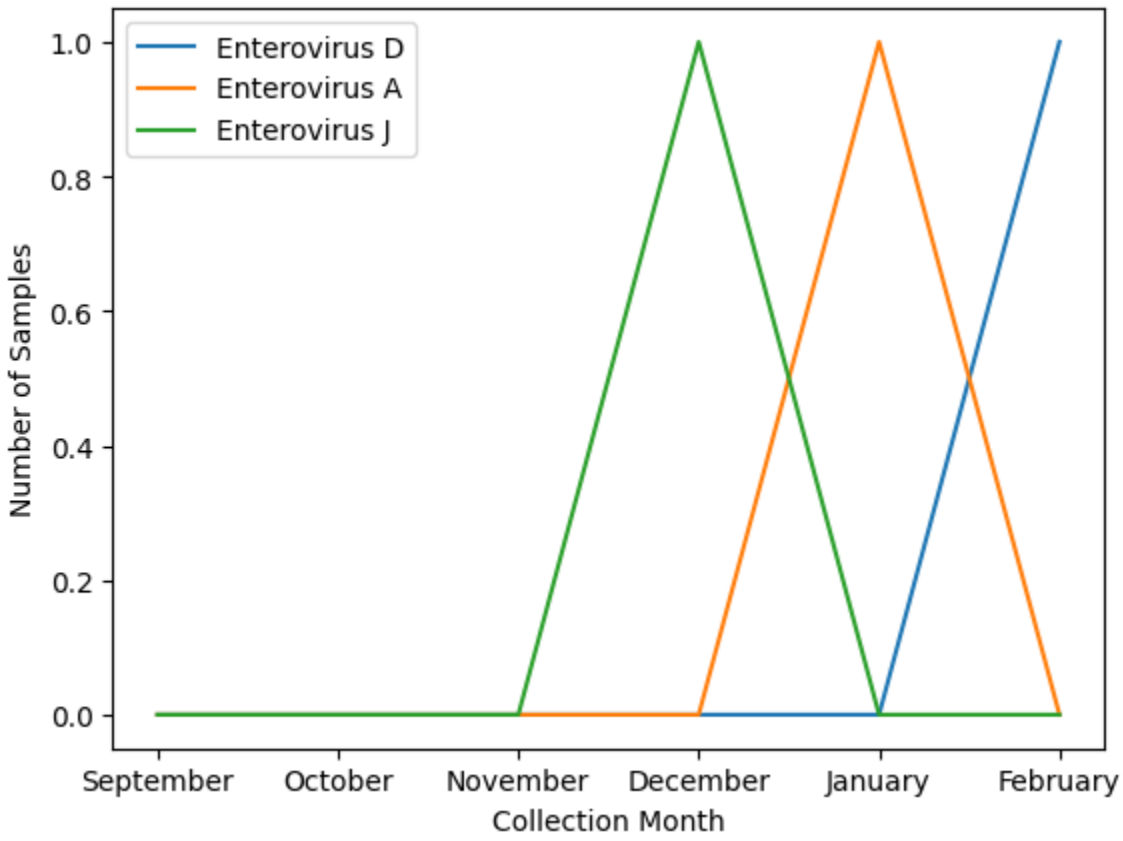


Figure 14: The number of samples with an abundance estimate of over 10% for three species of *Enterovirus* across the observed six months is shown in the Figure above.

Two samples from September show high levels of abundance for human *Orthopneumovirus*, with both samples showing an abundance estimate of over 95%. This virus was also detected in samples from November, but none had significant abundance levels. Also known as respiratory syncytial virus, human *Orthopneumovirus* is spread by nasal or oral secretions through direct contact or by the spread of droplets through air [36]. Observing data collected weekly by the state department of health and human services on the spread of various respiratory viruses from June 11, 2022- June 3, 2023 [43], we can see that RSV infections rise significantly during September and October of 2022. In fact, the peak of RSV positive cases is

November 5, 2023, as shown in Figure 15. After the steady increase in positive RSV cases from September 3, 2023 through November 5, 2023, cases begin to drop significantly throughout the month of November, from 1072 positive cases in the week of 11/5/2023 to 877 positive cases in the week of 11/12/2023.

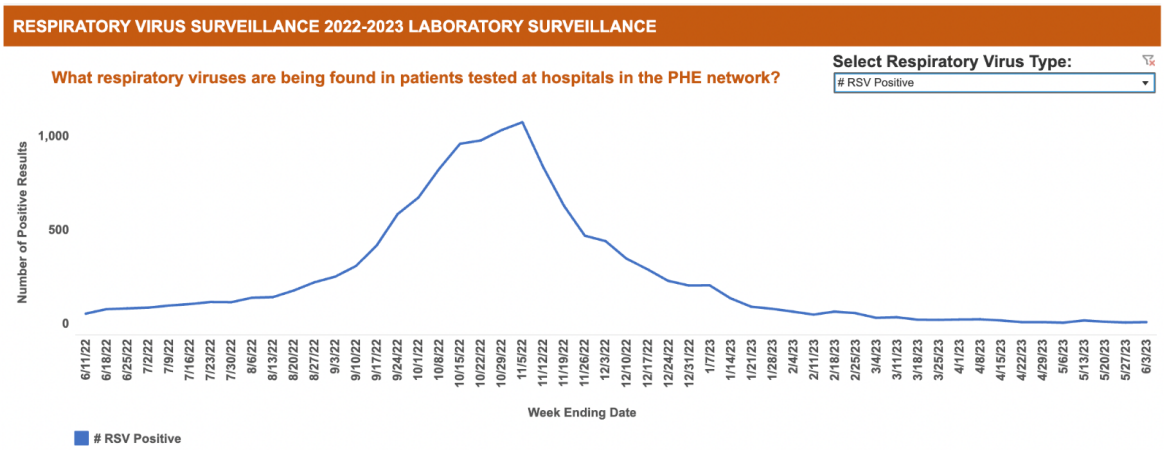


Figure 15: The above Figure was obtained from the North Carolina Department of Health and Human Services Detailed Respiratory Virus Surveillance Dashboard [43]. The line graph shows the number of positive RSV cases per tracking week, starting on June 11, 2022 through June 3, 2023.

Various strains of human mastadenovirus were detected in samples collected from September through December. Strain E showed high abundance in one sample from October, November, and December. In September, there was some abundance of both strains E and B across all samples, but was too low to be noted. Human *Mastadenoviruses*, or HAdVs, consist of viruses from one of seven species, A-G and can cause a wide range of infections, from gastroenteritis to respiratory illnesses such as the common cold [37]. This virus also spreads by nasal or oral secretions, mainly through exchange of air droplets. Figure 16 shows the number of samples per month that showed an abundance of three different species of *Mastadenovirus* greater than 10%.

Number of Samples with Abundance of *Mastadenovirus* Greater than 10%

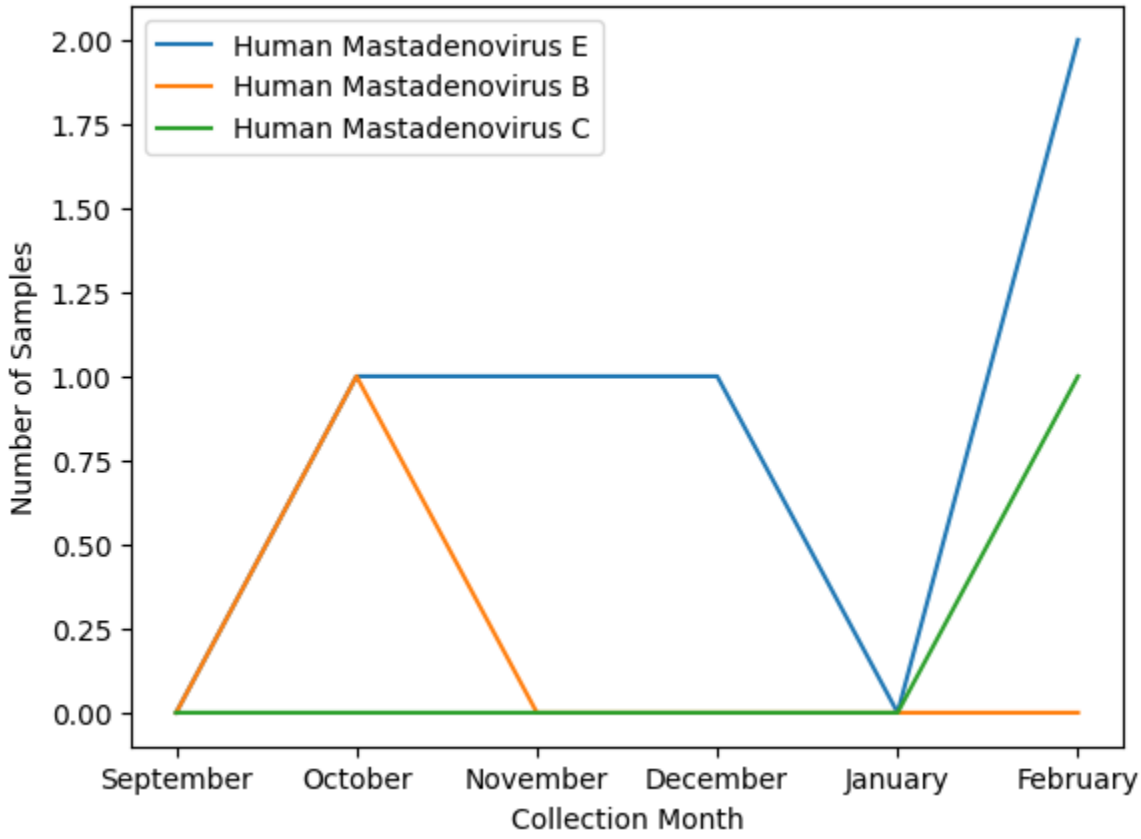


Figure 16: The number of samples with an abundance estimate of over 10% for three species of *Mastadenovirus* across the observed six months is shown in the Figure above.

One sample from November showed heavy detection of *Pandoravirus neocaledonia*. *Pandoravirus* is a genus of giant virus, and is the second largest known virus in size [38]. This sample from November 28, 2023 showed a 93% abundance of *Pandoravirus*. This virus, similar to most giant viruses, does not impact humans in any known harmful way. Another sample from November 11, 2023 showed high abundance of a type of granulovirus. *Mythimna unipuncta granulovirus B* is part of the *Betabaculovirus* genus of bacteria, which are commonly found in arthropods, which serve as their natural host, from which the virus lives [39].

Observing Figure 12, the most common virus in abundance in December's samples was SARS-CoV. Eleven out of fourteen samples observed for that month showed an abundance estimate of over 15% per sample. This is similar to the trend observed in October and November, although December has the highest amount of samples with an abundance of SARS-CoV out of all months observed during 2022. Human *Mastadenovirus E,* influenza A, and rhinovirus C were detected across three different samples collected in December.

The detection of rhinoviruses and mastadenoviruses all proves the circulation of other viruses besides SARS-CoV-2 on campus. The detection of other viruses besides SARS-related coronavirus is a sign of the ongoing "twindemic", in which SARS-CoV-2 was not the only virus rapidly spreading. High detection of the virus BeAn 58058 was not expected, yet symptoms of infection from this virus usually present as those from a common cold. As abundance estimates per virus are observed from January and February, similar trends in terms of common viruses detected should continue.

Kraken Detection of Respiratory Viruses in Clinical Negative COVID-19 PCR Tests - January 2023

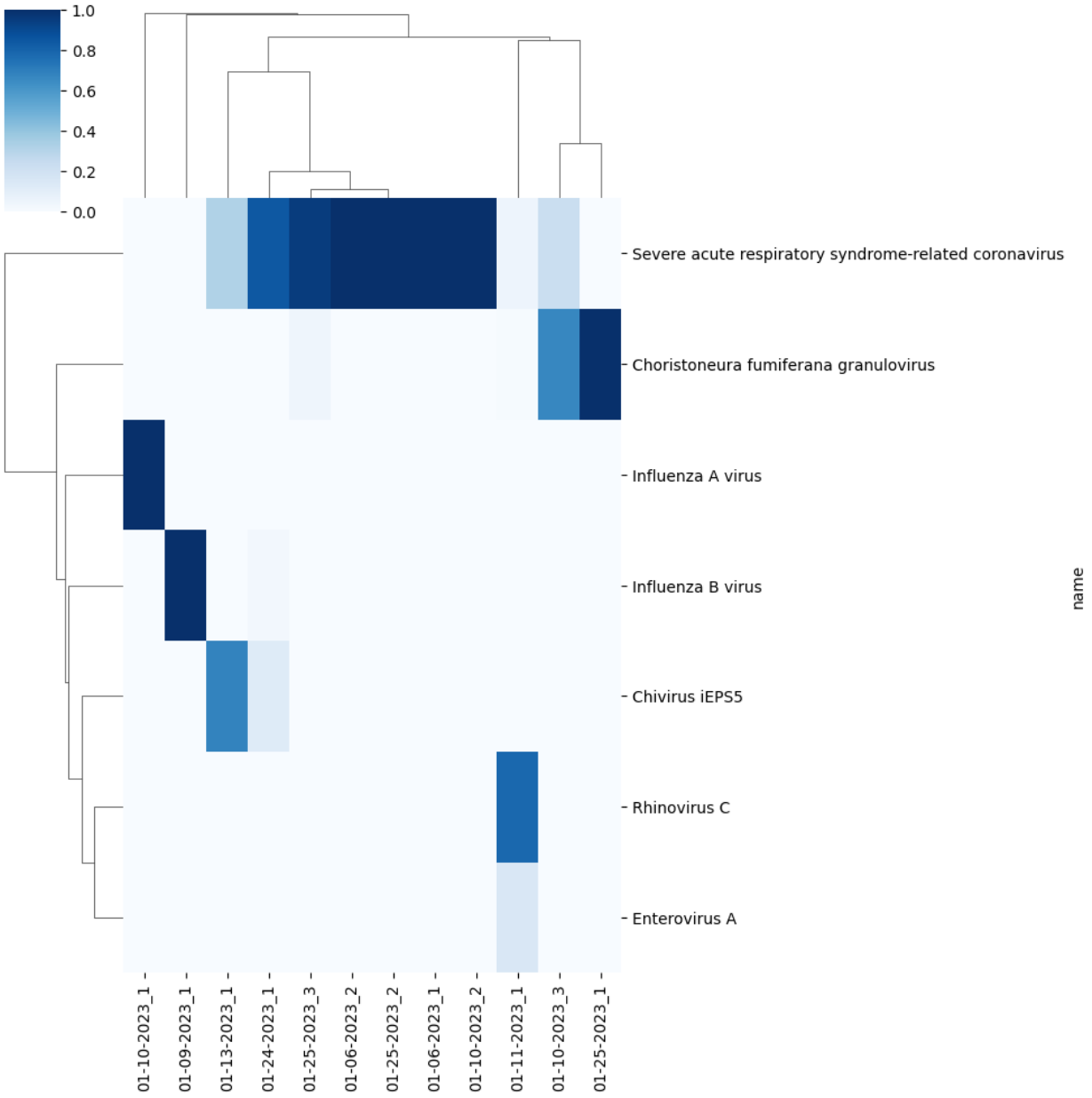


Figure 17: The above heatmap shows detection of viruses using the Kraken metagenomic classifier with more than 10% abundance per sample across samples collected from January 2023. Each tick on the x-axis represents a sample, labeled with the specific date it was collected from. The key in the upper right corner shows how the varying levels of color represent the percentage of abundance each virus was detected in each sample.

Kraken Detection of Respiratory Viruses in Clinical Negative COVID-19 PCR Tests - February 2023

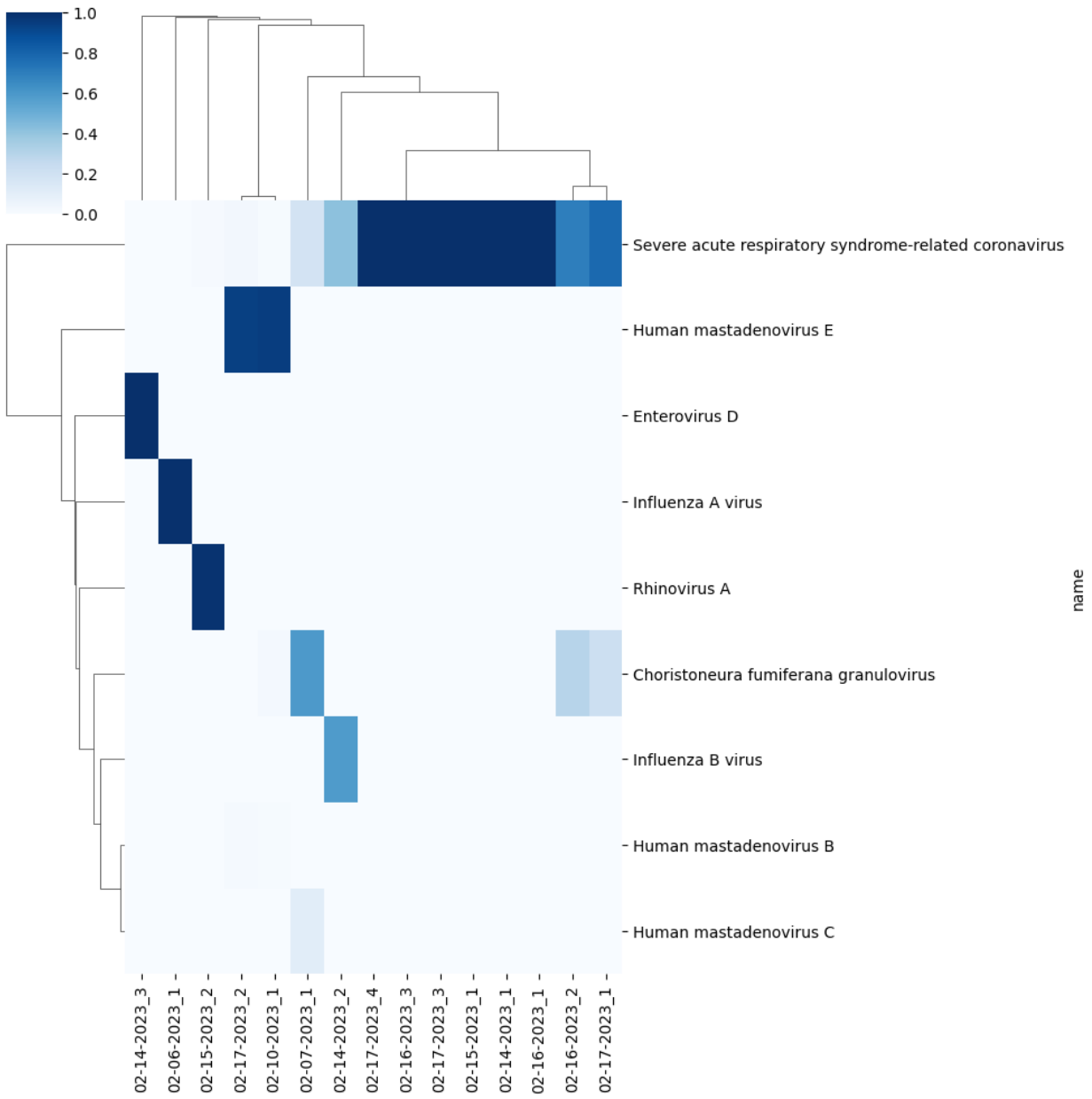


Figure 18: The above heatmap shows detection of viruses using the Kraken metagenomic classifier with more than 10% abundance per sample across samples collected from February 2023. Each tick on the x-axis represents a sample, labeled with the specific date it was collected from. The key in the upper right corner shows how the varying levels of color represent the percentage of abundance each virus was detected in each sample.

Figures 17 and 18 show a pattern of high abundance of SARS-CoV in samples that was seen beginning from 2022 and continues through January and February of 2023. Influenza A and influenza B are also found in high abundance in two samples from January, while one sample from February shows a high abundance of influenza A. Influenza A virus is responsible for seasonal epidemics of influenza. It is easily spread between human hosts and can easily mutate as it spreads. Low immunity amongst the population is a significant reason as to why it easily spreads and can make most people who are infected ill. Influenza A was also highly abundant in one sample from December of 2022, while influenza B showed low abundance across all samples from December.

In January of 2023, one sample showed an abundance estimate of 78% for rhinovirus C. One sample from February showed a high abundance for rhinovirus A. Various strains of rhinovirus were detected in September, October, and November, with rhinovirus A being commonly abundant in samples from September and October. In February, two samples showed a high abundance estimate for mastadenovirus E. Various types of human mastadenovirus were highly detected in a few samples from October and November, and mastadenoviruses B and C were low in abundance in samples from February.

As mentioned earlier, the trend of SARS-related coronavirus being the most commonly abundant virus in samples from December of 2022 continues in samples from January and February of 2023. The end of many restrictions on large gatherings and face covering requirements could have caused a faster and wider spread of this virus. Other respiratory viruses in circulation included human mastadenoviruses and rhinoviruses. Some samples showed detection of influenza A, which is fast spreading and could be dangerous as most are not immune to a new strain. The end of masking and other restrictions also aided in the spread of these

viruses and caused them to be highly abundant in the observed samples. This assumption can be used to further the hypothesis that a twindemic or tripledemic of various respiratory diseases was taking place on UNC-Charlotte campus during the end of 2022 and the beginning of 2023.

3.2 Analysis of Samples Using CLARK

CLARK Detection of Respiratory Viruses in Clinical Negative COVID-19 PCR Tests - September 2022 through January 2023

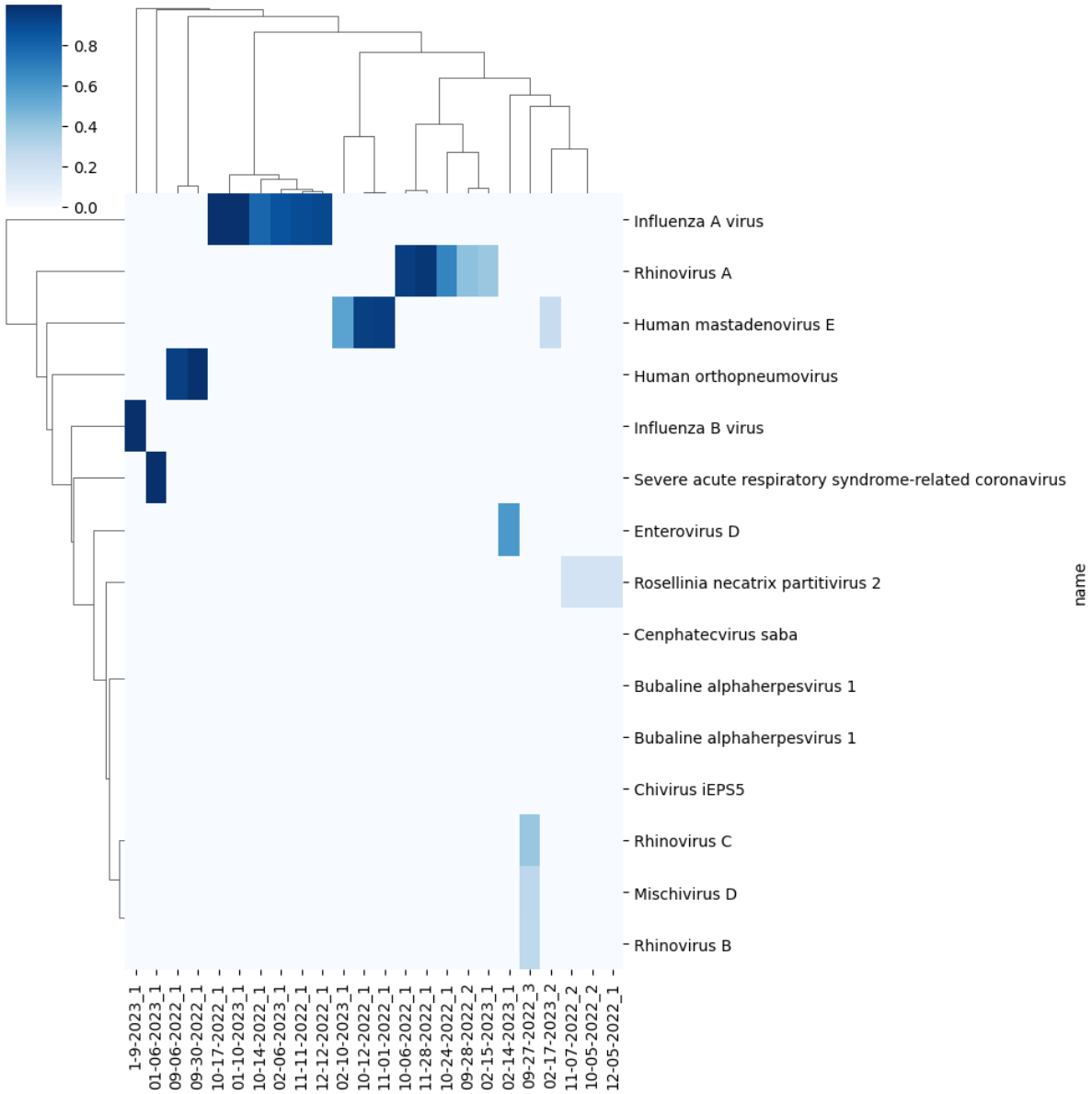


Figure 19: The above heatmap shows detection of viruses using the CLARK metagenomic classifier. Each tick on the x-axis represents a sample, labeled with the specific date it was

collected from. The key in the upper right corner shows how the varying levels of color represent the percentage of abundance each virus was detected in each sample.

Kraken Detection of Respiratory Viruses in Clinical Negative COVID-19 PCR Tests - September 2022 through January 2023

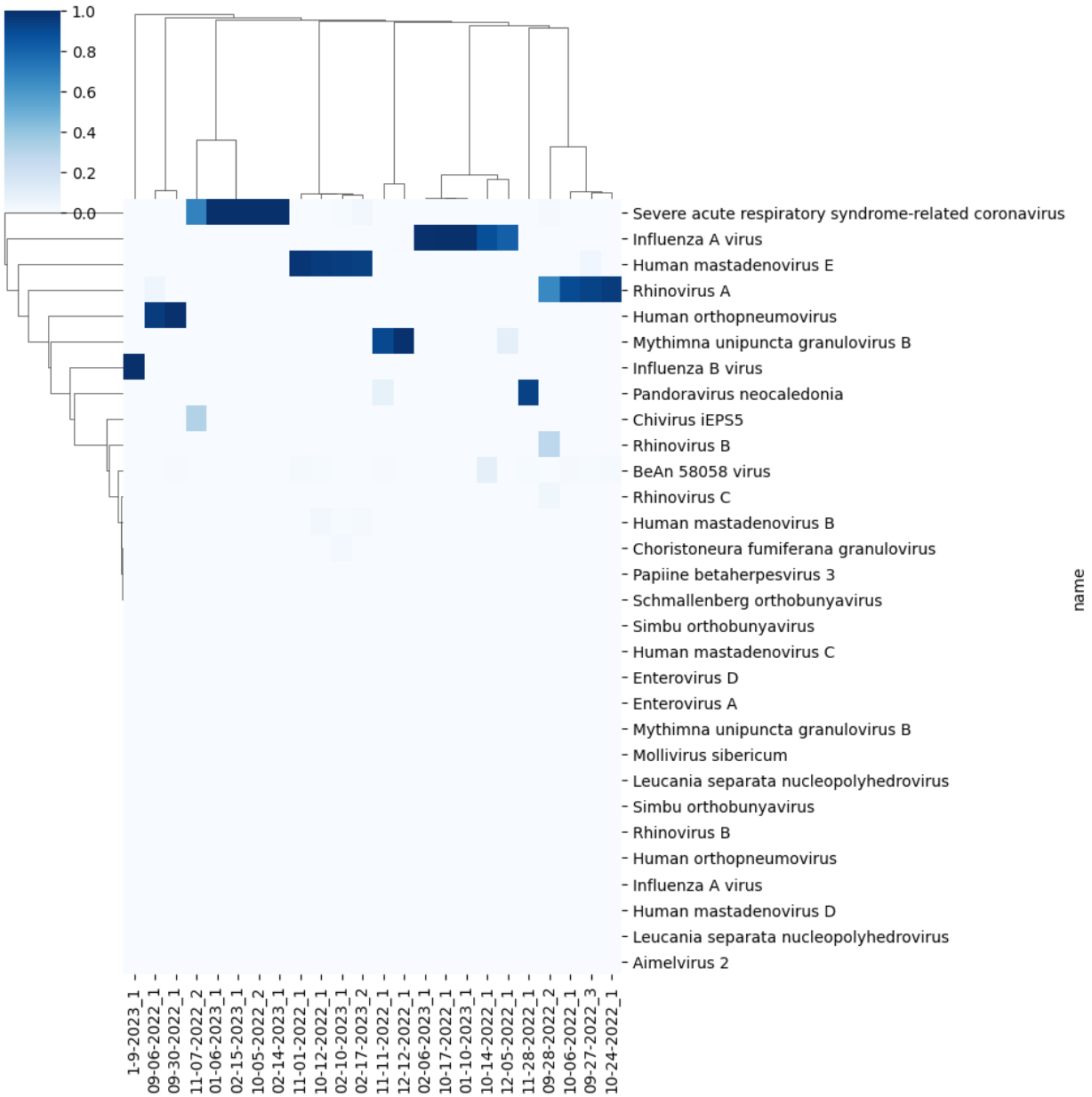


Figure 20: The above heatmap shows the classification of viruses using Kraken that were also classified by CLARK. Each tick on the x-axis represents a sample, labeled with the specific date it was collected from. The key in the upper right corner shows how the varying levels of color represent the percentage of abundance each virus was detected in each sample.

Figure 19 shows abundance estimates of viruses from samples that were classified using the CLARK metagenomic classification tool. Several differences can be seen in results from CLARK classification and the previously discussed classification using Kraken. CLARK classifies significantly less samples than Kraken, and therefore produces less results to draw conclusions from. This could be due to several reasons. Firstly, the CLARK algorithm may be performing a separate "quality check" on input reads before classification takes place. This would explain the low amounts of viruses being detected; the algorithm classifies a large number of reads as low quality and does not classify them as any taxa. In an attempt to control Kraken's issue of classifying "phantom taxa", as previously mentioned, all viruses with less than 10% abundance were removed from results yielded by Kraken before analysis. This could potentially not have been a strict enough filter on the samples and more thorough vetting of viruses and abundances may have been necessary.

CLARK classification results showed high detection of *Influenza A virus*, rhinovirus A, and human mastadenovirus E. *Influenza A virus* was most commonly detected by CLARK in samples from October, November, December, January, and February. Only six samples showed an abundance of influenza A. This contrasts results from classification using Kraken, where December, January, and February showed a combined 3 samples that had an abundance estimate for influenza A. However, CLARK and Kraken are consistent in not showing any abundance for *Influenza A virus* during September. Kraken classification results also showed samples across all six observed months having high abundance of SARS-CoV, while CLARK classification only showed one sample from January having a high abundance of SARS-related coronavirus.

Rhinovirus A was heavily detected by CLARK in samples from September, October, and November. This is consistent with Kraken classification results as far as samples from September

and October. Kraken results from November did not show high abundance of rhinovirus A at all, while CLARK results classified a sample from 11/28/2022 with a 97.2% abundance estimate for rhinovirus A. That same sample showed a 93% abundance estimate for *Pandoravirus neocaledonia*. It is possible this virus and rhinovirus A are similar in genetic sequence that Kraken misclassified it as such due to both virus's similarities. A potential reason for this may be that Kraken misclassified Human mastadenovirus E was detected in one sample from November and one sample from October in classification results obtained from Kraken. CLARK classification results showed three samples with an abundance estimate of over 50% for human mastadenovirus E from October, November, and December. Two samples from September showed an abundance estimate of over 90% for human *Orthopneumovirus* according to CLARK classification results. These samples were collected on 9/6/2022 and 9/30/2022. Kraken classification results showed extremely similar abundance estimates (over 90%) for human *Orthopneumovirus* from samples collected on those same exact dates. This is one instance of an exact match of results from both classification tools.

Kraken classification is reliable due to its wide-spread use and regard amongst classification tools. It is able to run on a large amount of samples, making it widely popular. The goal of CLARK is to run faster than Kraken and be able to classify the same amount of samples in less time and using less disk space. The results shown above yielded by Kraken cannot be deemed inaccurate, as some similarities in viruses detected were also shown by CLARK. Overall, based on results from both classification tools, it can be concluded that the presence of certain viruses, such as SARS-CoV, rhinovirus, and mastadenovirus were the most commonly spread viruses on UNCC campus between September of 2022 and February of 2023.

CHAPTER 4: IMPACTS AND FUTURE WORK

4.1 Impacts of Project

The local impacts of this project are possibly the most important when considering the size of the study. Since a total of 90 samples were observed, and 85 samples were included in the classification analysis, the sample size for this study is relatively small, but does provide enough power to draw some conclusions. The results obtained can mostly be used to make assumptions and draw conclusions about circulating viruses only on UNC-Charlotte campus. Furthermore, people who are tested at the on-campus student health center must be UNC-Charlotte students or faculty, which means they will certainly be moving around campus and contributing to the circulation of any virus they are carrying.

Further local impacts of this study include the results being used as a proxy for viral circulation in areas around campus or even regionally. Results from this study can be used to infer what may be circulating in nearby neighborhoods and locations. This is because students and faculty may leave campus during their time of infection, before or during the time when they are symptomatic. Finding what viruses are circulating on campus can be telling about what viruses are circulating in areas close to campus. This can give more information about what viruses are spreading in schools around campus, or other areas where there are large gatherings of people, giving more room for circulation of viruses and infection.

The impacts of this study can be mirrored with the impacts of an ongoing study being conducted by the North Carolina Department of Health and Human Services (NCDHHS). Their ongoing respiratory virus surveillance project tests patients who have entered the emergency department at hospitals in the hospital-based public health epidemiologist program, or the PHE. Using a multiplex PCR assay, the laboratories at PHS hospitals are able to test for seven different

respiratory viruses, including those that were commonly found in the samples used for this study. Looking at Figure 21, the number of weekly positive cases of influenza, adenovirus (of which mastadenovirus is a subgroup), and rhinovirus is shown beginning from June of 2022 until May 27, 2023. There are several similarities between the study being conducted by the NCDHHS and this project, including testing symptomatic patients for respiratory viruses other than SARS-CoV-2. The correlation between the presence of RSV in campus samples and the rise in positive RSV cases across the state in September and October of 2022 was previously mentioned. Observing Figure 21, the number of positive cases of influenza A begins to rise from October 15, 2022 and peaks in the week of December 3, 2022. Our results do not show a high abundance of influenza A throughout the six month time series, however we do see few cases of influenza A and B from December through February.

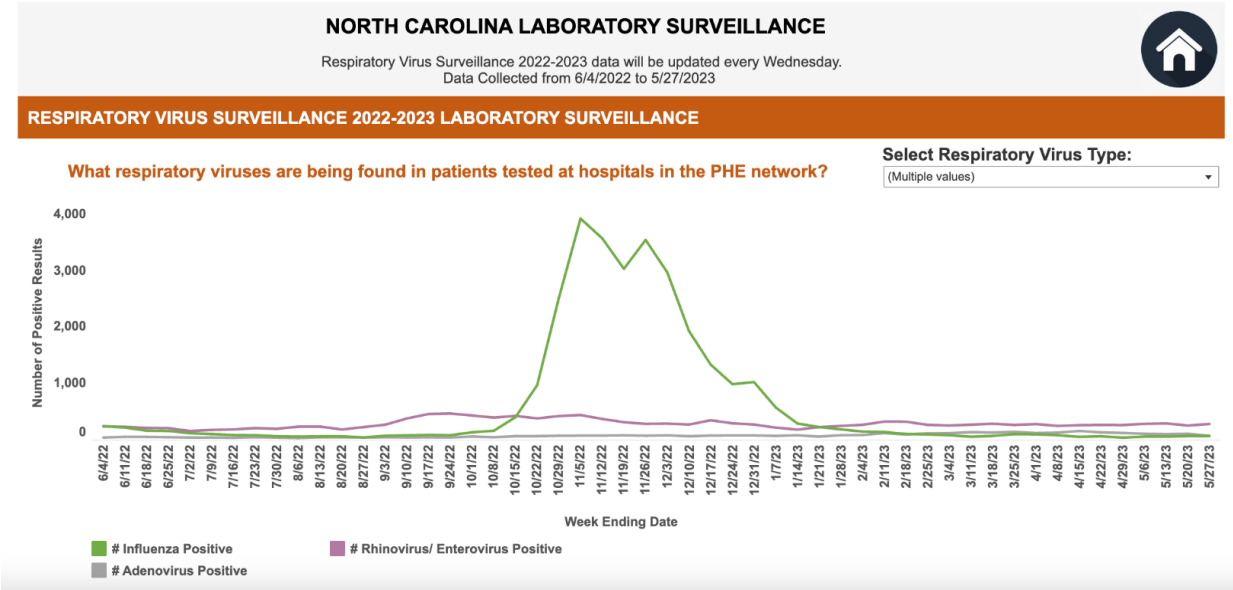


Figure 21: The above Figure was obtained from the North Carolina Department of Health and Human Services Detailed Respiratory Virus Surveillance Dashboard [43]. The line graph shows the number of positive cases per tracking week of various respiratory viruses, such as influenza, adenovirus, and rhinovirus, which were also found in a large number of on campus COVID-19 negative samples.

4.2 Future Work

Due to limited time given to complete the sequencing and analysis of samples, a follow-up study could be conducted using samples collected over a time period longer than six months. This would allow for a wider range of information to be collected for analysis. Sequencing samples to include, for example, a full year of collection would be useful to see if the currently observed trends in circulating viruses continues. Furthermore, the effects of the "twindemic" can then be observed across a larger time period.

We could also increase the number of samples being observed per month. For example, instead of observing 10-15 samples from each month within a six month period, observing double that amount of samples per month would allow for a more in-depth analysis of what viruses are in circulation. Observing more samples could also show if viruses that were only detected in a few samples using a small sample size are detected more often. This could potentially expand the list of top viruses detected across all samples or even confirm if some of the low-abundance viruses are more common..

As mentioned earlier, several discrepancies showed in the comparison of CLARK and Kraken results. According to CLARK results, fewer reads were classified into viral taxa and, in turn, fewer samples were classified. Comparing the algorithms of CLARK and Kraken to see where CLARK deems certain reads as too poor in quality to actually classify would be an interesting question to resolve. This relates to the issue of Kraken and its classification of "phantom taxa". Furthermore, other metagenomic classification tools could be used and compared to both CLARK and Kraken. CLARK was chosen due to its quality of being based off of Kraken, and because it is meant to be a faster and more precise classification tool. However,

other classification tools can also be used to classify samples, and the results yielded can be compared with the current results discussed above.

Different statistical analysis tools could also be used to discover correlations between sample detections. Correlations could be shown across different months or different seasons throughout the year. Computing statistical correlations could give information about similarities in viruses detected that is not easily identifiable by observing a heat map containing the data. A statistical model could be used to see what clusters form in the data, showing which data points are related to each other. This could then show what samples have the same abundance of virus, depending on the cluster they form together.

4.2.1 Applying Project Concept to Wide Scale Project

Using the $Twist^{®}$ Respiratory Viral Research Panel has limited the detection of viruses in this study. A wide scale project can be conducted by potentially testing samples for more viruses and even bacterial presence by not using a targeted sequencing approach. This means more can be detected from a sample and detection is not limited to a set amount of respiratory viruses. This method can be used should bacterial or viral epidemics arise, in order to see what else is spreading in a community besides that particular virus or bacteria. Similar to this study, the information could be used to see what else could be causing illness in a community.

4.2.2 Compare Results with Viral Research Panel Sequences

Some of the viruses that were detected using Kraken and CLARK's classification algorithms were not evident on the $Twist^{®}$ Respiratory Viral Research Panel we targeted. A potential reason for this could be that the viruses classified by both classification tools are very similar in genetic makeup to those we were originally targeting. A way to check this would be to compare the nucleotide sequences of the viruses found in the classified samples with the

nucleotide sequences of the viruses on the viral panel. The similarities between the sequences could explain why the classification tools classified those samples as viruses that were not on the viral panel.

4.2.3 Better Preparation of Controls

One issue with our experiment was the contamination of the controls. Looking at classification results from Kraken, it is evident that all three negative controls were contaminated. The library preparation negative control was also contaminated when looking at the classification results from CLARK. Furthermore, contamination also occurred in the positive controls. Avoiding contamination would be an important point of focus in further experiments. Splitting the samples into two different experimental runs so that there is not as high of a chance of contamination could potentially be helpful.

REFERENCES

1. Zou, S., Caler, L., Colombini-Hatch, S. *et al.* (2016). Research on the human virome: where are we and what is next. *Microbiome* 4, 32. https://doi.org/10.1186/s40168-016-0177-y

2. Kim, D., et. al. (2020). The Architecture of SARS-CoV-2 Transcriptome. *Cell, 181*(4), 914-921. https://doi.org/10.1016/j.cell.2020.04.011

3. Heraud, J., Razanajatovo, N. H., Viboud, C. (2019). Global circulation of respiratory viruses: from local observations to global predictions. *The Lancet Global Health, 7*(8), 982-983. https://doi.org/10.1016/S2214-109X(19)30277-3

4. Zhang, R., et. al. (2020). Identifying airborne transmission as the dominant route for the spread of COVID-19. *Proceedings of the National Academy of Sciences, 117*(26), 14857-14863. https://doi.org/10.1073/pnas.2009637117

5. Wang, Y., et. al. (2021). Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology, 39,* 1348-1365. https://doi.org/10.1038/s41587-021-01108-x

6. Garibyan, L., & Avashia, N. (2013). Polymerase chain reaction. *The Journal of investigative dermatology*, *133*(3), 1–4. https://doi.org/10.1038/jid.2013.1

7. *Phosphoester Formation.* (n.d). LibreTexts Chemistry. Retrieved June 4, 2023 from https://chem.libretexts.org/Courses/Sacramento_City_College/SCC%3A_Chem_309_-_General_Organic_and_Biochemistry_(Bennett)/Text/13%3A_Functional_Group_Reactions/13.10%3A_Phosphoester_Formation

8. Clancy, S., Brown, W. (2008). Translation: DNA to mRNA to Protein. *Nature Education, 1*(1).

9. Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1-8. https://doi.org/10.1016/j.ygeno.2015.11.003

10. *DNA sequencing.* (n.d). Khan Academy. Retrieved June 5, 2023 from https://www.khanacademy.org/science/ap-biology/gene-expression-and-regulation/biotechnology/a/dna-sequencing

11. Bunnik, E. M., & Le Roch, K. G. (2013). An introduction to functional genomics and systems biology. *Advances in Wound Care*, *2*(9), 490-498. https://doi.org/10.1089/wound.2012.0379

12. *Illumina Sequencing Technology.* (n.d). Illumina. Retrieved May 15, 2023 from https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

13. Jiang, H., Lei, R., Ding, S. W., & Zhu, S. (2014). Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC bioinformatics*, *15*, 1-12. https://doi.org/10.1186/1471-2105-15-182

14. McNaughton, A. L., Roberts, H. E., Bonsall, D., de Cesare, M., Mokaya, J., Lumley, S. F., ... & Matthews, P. C. (2019). Illumina and Nanopore methods for whole genome sequencing of hepatitis B virus (HBV). *Scientific reports*, *9*(1), 1-14. https://doi.org/10.1038/s41598-019-43524-9

15. Van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in genetics*, *30*(9), 418-426. https://doi.org/10.1016/j.tig.2014.07.001

16. Ozsolak, F., & Milos, P. M. (2011). Single‐molecule direct RNA sequencing without cDNA synthesis. *Wiley Interdisciplinary Reviews: RNA*, *2*(4), 565-570. https://doi.org/10.1002/wrna.84

17. *NGS overview: from sample to sequencer to results*. (n.d). iRepertoire. Retrieved June 5, 2023 from https://irepertoire.com/ngs-overview-from-sample-to-sequencer-to-results/

18. Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, *15*(3), 1-12. https://doi.org/10.1186/gb-2014-15-3-r46

19. Menzel, P., Ng, K. L., & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature communications*, *7*(1), 11257. https://doi.org/10.1038/ncomms11257

20. Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome biology*, *20*, 1-13. https://doi.org/10.1186/s13059-019-1891-0

21. Ounit, R., Wanamaker, S., Close, T. J., & Lonardi, S. (2015). CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC genomics*, *16*(1), 1-13. https://doi.org/10.1186/s12864-015-1419-2

22. Carbo, E. C., et. al. (2022). Performance of five metagenomic classifiers for virus pathogen detection using respiratory samples from a clinical cohort. *Pathogens*, *11*(3), 340. https://doi.org/10.3390/pathogens11030340

23. Simon, H. Y., Siddle, K. J., Park, D. J., & Sabeti, P. C. (2019). Benchmarking metagenomics tools for taxonomic classification. *Cell*, *178*(4), 779-794. https://doi.org/10.1016/j.cell.2019.07.010

24. *Uracil.* (2023, June 2). National Human Genome Research Institute. Retrieved March 8, 2023 from https://www.genome.gov/genetics-glossary/Uracil

25. Houseley, J., & Tollervey, D. (2009). The many pathways of RNA degradation. *Cell*, *136*(4), 763-776. https://doi.org/10.1016/j.cell.2009.01.019

26. QIAamp Viral RNA Mini Handbook. (2020). *Qiagen,* Hilden, Germany.

27. Nanotrap® Microbiome A; Automated Protocol with MagMAXTM Kit and the KingFisherTM Apex. (2022, October). *Ceres Nanosciences,* Manassas, Virginia, USA.

28. *NextSeq 1000 & NextSeq 2000 Systems*. (n.d). Illumina. Retrieved May 15, 2023 from https://www.illumina.com/systems/sequencing-platforms/nextseq-1000-2000.html

29. *An Automated Electrophoresis Solution that Delivers DNA and RNA Sample QC*. (n.d.). Agilent. Retrieved May 15, 2023 from https://www.agilent.com/en/product/automated-electrophoresis/tapestation-systems/tapestation-dna-screentape-reagents

30. Lu, J., Breitwieser, F. P., Thielen, P., & Salzberg, S. L. (2017). Bracken: estimating species abundance in metagenomics data. *PeerJ Computer Science*, *3*, e104. https://doi.org/10.7717/peerj-cs.104

31. Hardin, J. (2021) UNC Charlotte Student Health Center 6.15 Cepheid SARS-CoV-2/Flu plus or CoV-2 plus Procedure.

32. Fonseca, F.G, et. al. (1998). Morphological and molecular characterization of the poxvirus BeAn 58058. *Archives of Virology, 143,* 1171-1186. https://doi.org/10.1007/s007050050365

33. Ferravante, C., et. al. (2022). Nasopharyngeal virome analysis of COVID-19 patients during three different waves in Campania region of Italy. *Journal of Medical Virology, 94*(5), 2275-2283. https://doi.org/10.1002/jmv.27571

34. Medeiros-Silva, D.C., et. al. (2009). Clinical signs, diagnosis, and case reports of *Vaccinia virus* infections. *The Brazilian Journal of Infectious Diseases, 14*(2), 129-134. https://doi.org/10.1016/S1413-8670(10)70025-8

35. Sayama, A., et. al. (2022). Comparison of Rhinovirus A–, B–, and C–Associated Respiratory Tract Illness Severity Based on the 5′-Untranslated Region Among Children Younger Than 5 Years. *Open Forum Infectious Diseases, 9*(10). https://doi.org/10.1093/ofid/ofac387

36. Shoham, S. (2022, November 11). *Respiratory syncytial virus.* Johns Hopkins Medicine. https://www.hopkinsguides.com/hopkins/view/Johns_Hopkins_ABX_Guide/540472/all/Respiratory_syncytial_virus

37. Chen, S., & Tian, X. (2018). Vaccine development for human mastadenovirus. *Journal of thoracic disease*, *10*(Suppl 19), 2280–2294. https://doi.org/10.21037/jtd.2018.03.168

38. Pandoravirus. (2023, May 10). In *Wikipedia.* https://en.wikipedia.org/wiki/Pandoravirus

39. Betabaculovirus. (2023, January 30). In *Wikipedia.* https://en.wikipedia.org/wiki/Betabaculovirus

40. Johnson, J., Sun, S., Fodor, A. (2022). Systematic classification error profoundly impacts inference in high-depth Whole Genome Shotgun Sequencing datasets. https://doi.org/10.1101/2022.04.04.487034

41. Reyes, G. R., & Kim, J. P. (1991). Sequence-independent, single-primer amplification (SISPA) of complex DNA populations. *Molecular and cellular probes*, *5*(6), 473–481. https://doi.org/10.1016/s0890-8508(05)80020-9

42. *Respiratory Virus Research Panel*. (n.d.). Twist Bioscience. Retrieved June 6, 2023 from https://www.twistbioscience.com/products/ngs/fixed-panels/respiratory-virus-research-panel

43. Detailed Respiratory Virus Surveillance Dashboard. (2023, May 31). NCDHHS COVID-19 Response. Retrieved June 1, 2023 from https://covid19.ncdhhs.gov/dashboard/respiratory-virus-surveillance

44. Ameur, A., Kloosterman, W. P., Hestand, M. (2019). Single-Molecule Sequencing: Towards Clinical Applications. *Trends in Biotechnology, 37*(1), 72-85. https://doi.org/10.1016/j.tibtech.2018.07.013

45. Gage Moreno, David O'connor 2020. Sequence-Independent, Single-Primer Amplification of RNA viruses. Protocols.io. https://dx.doi.org/10.17504/protocols.io.bhk4j4yw

46. Twist Bioscience Library Preparation EF Kit 2.0 for ssRNA Virus Detection. (2022, November 1). *Twist Bioscience*, South San Francisco, California, USA.

47. Twist Target Enrichment Standard Hybridization v2 Protocol. (2022, April 11). *Twist Bioscience*, South San Francisco, California, USA.

48. Detect Multiple Viral Pathogens From a Single Sample. (n.d.). *Twist Bioscience.* Retrieved June 12, 2023 from https://www.twistbioscience.com/products/ngs/fixed-panels/respiratory-virus-research-panel

49. Fast, accurate, sensitive, and specific quantification of DNA, RNA, and protein. (n.d.). *ThermoFisher Scientific.* Retrieved June 12, 2023 from https://www.thermofisher.com/us/en/home/industrial/spectroscopy-elemental-isotope-analysis/molecular-spectroscopy/fluorometers/qubit.html?gclid=CjwKCAjwhJukBhBPEiwA

niIcNXqaKSYN7uxInx0nCJ_M_zMhu6JmYFrxY37NMIZr77Q0BhNfZWR39hoCBJAQ
AvD_BwE&ef_id=CjwKCAjwhJukBhBPEiwAniIcNXqaKSYN7uxInx0nCJ_M_zMhu6J
mYFrxY37NMIZr77Q0BhNfZWR39hoCBJAQAvD_BwE:G:s&s_kwcid=AL!3652!3!55
5768009100!p!!g!!qubit%20thermofisher!7851124051!80648208414&cid=bid_pca_aqb_
r01_co_cp1359_pjt0000_bid00000_0se_gaw_bt_pur_con

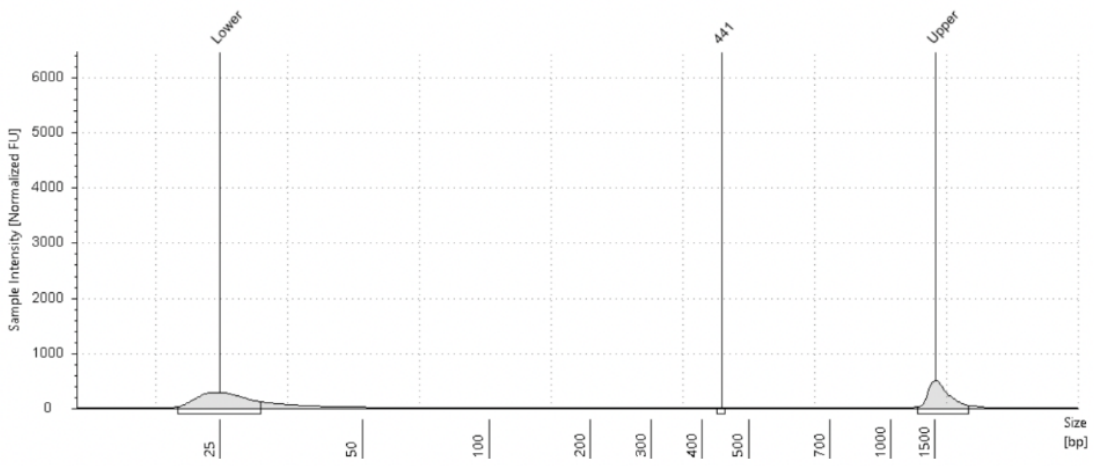
APPENDIX A: *Agilent*® TapeStation Quantification Results

Well	Concentration [pg/µl]	Samples in Well
E2	2.00	9/6/2022_1 9/27/2022_1 9/27/2022_2 9/27/2022_3 9/28/2022_1 9/29/2022_1 9/29/2022_2 9/29/2022_3
F2	48.5	9/29/2022_4 9/30/2022_1 10/3/2022_1 10/5/2022_1 10/6/2022_1 10/7/2022_1 10/10/2022_1 10/12/2022_1
G2	11.3	10/14/2022_1 10/17/2022_1 10/20/2022_1 10/24/2022_1 11/1/2022_1 11/2/2022_1 11/3/2022_1 11/4/2022_1
L2	2160	11/7/2022_1 11/8/2022_1 11/9/2022_1 11/10/2022_1 11/11/2022_1 11/28/2022_1 12/1/2022_1 12/5/2022_1
H2	65.6	12/5/2022_2 12/6/2022_1 12/7/2022_1 12/8/2022_1

		12/12/2022_1 12/13/2022_1 12/20/2022_1 12/21/2022_1
A3	20.1	1/2/2023_1 1/3/2023_1_KF 1/6/2023_1 1/6/2023_2 1/9/2023_1_KF 1/10/2023_1 1/10/2023_2 1/10/2023_3
B3	12.8	1/11/2023_1 1/13/2023_1 Influenza A Positive Influenza B Positive Covid Positive Extraction Negative SISPA Negative LibPrep Negative
C3	1670	2/14/2023_1 2/16/2023_1 2/16/2023_2 2/15/2023_1 2/14/2023_2 2/17/2023_1 2/17/2023_2 2/10/2023_1
D3	1630	2/7/2023_1 2/6/2023_1 9/20/2022_1_KF 9/26/2022_1_KF 9/26/2022_2_KF 9/26/2023_3_KF 9/28/2022_2_KF 10/4/2023_1_KF
E3	1340	10/4/2022_2_KF 10/5/2022_2_KF 10/12/2022_2_KF 10/17/2022_2_KF

		11/3/2022_2_KF 11/3/2022_3_KF 11/4/2022_2_KF 11/7/2022_2_KF
F3	7910	Date N/A 12/8/2022_2_KF 12/9/2022_1_KF Date N/A 12/14/2022_1_KF 12/20/2022_2_KF 1/25/2022_1_KF 1/24/2022_1_KF
G3	–	Ladder
H3	325	1/27/2022_1_KF 1/25/2022_2_KF 1/25/2022_3_KF 2/15/2022_2_KF 2/17/2022_3_KF 2/16/2022_3_KF 2/14/2022_3_KF 2/17/2022_4_KF

E2: red 1



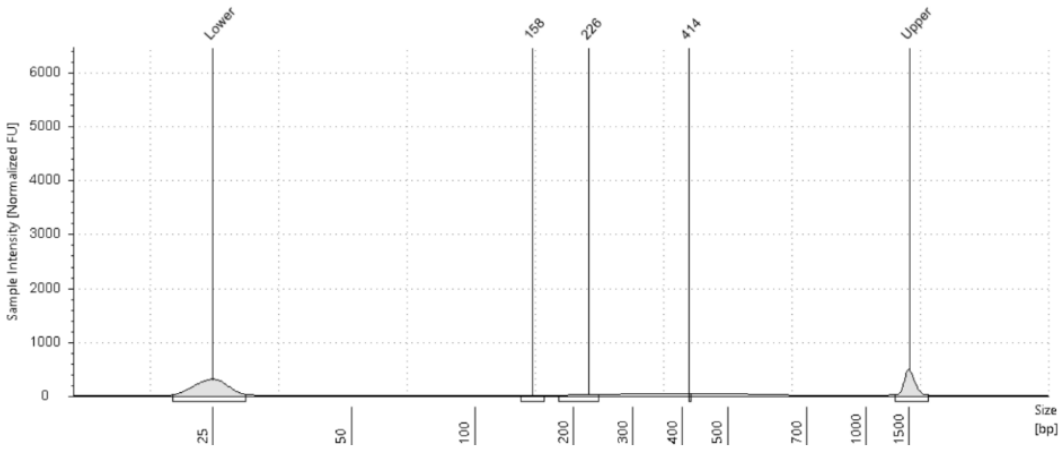
Sample Table

Well	Conc. [pg/µl]	Sample Description	Alert	Observations
E2	2.00	red 1		

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	345	-	21200	-		Lower Marker
441	2.00	-	6.99	100.00		
1500	250	250	256	-		Upper Marker

F2: red 2



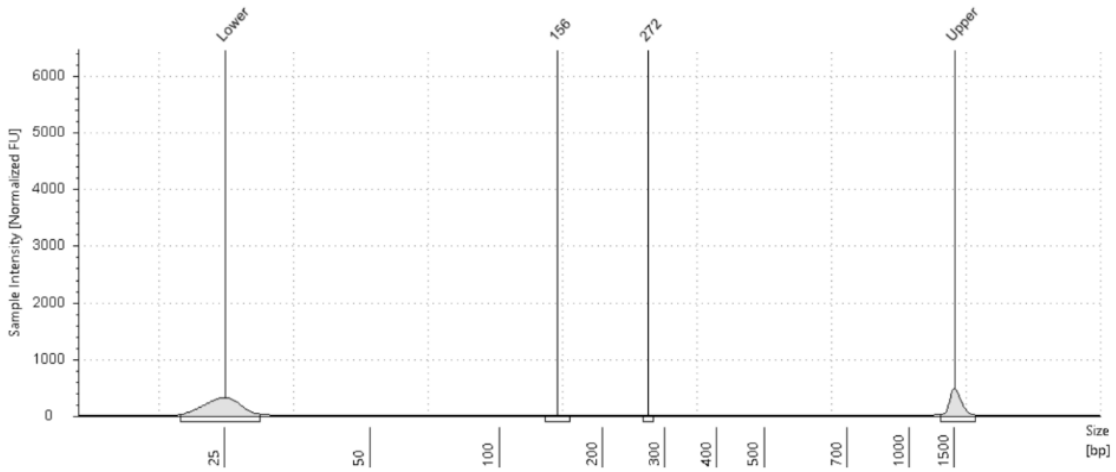
Sample Table

Well	Conc. [pg/µl]	Sample Description	Alert	Observations
F2	48.5	red 2		

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	503	-	31000	-		Lower Marker
158	3.83	-	37.3	7.89		
226	39.3	-	268	81.02		
414	5.38	-	20.0	11.09		
1500	250	250	256	-		Upper Marker

G2: red 3



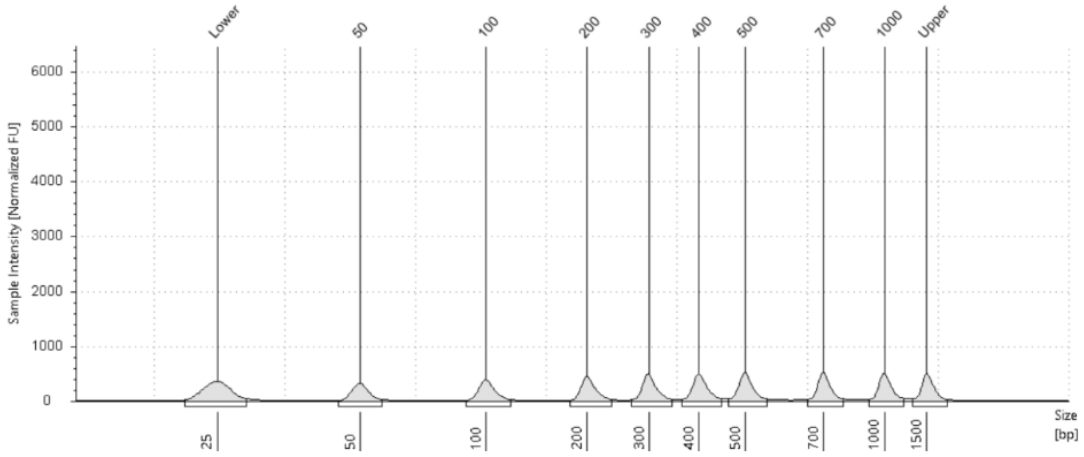
Sample Table

Well	Conc. [pg/µl]	Sample Description	Alert	Observations
G2	11.3	red 3		

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	527	-	32400	-		Lower Marker
156	4.22	-	41.6	37.30		
272	7.09	-	40.1	62.70		
1500	250	250	256	-		Upper Marker

L2: Ladder



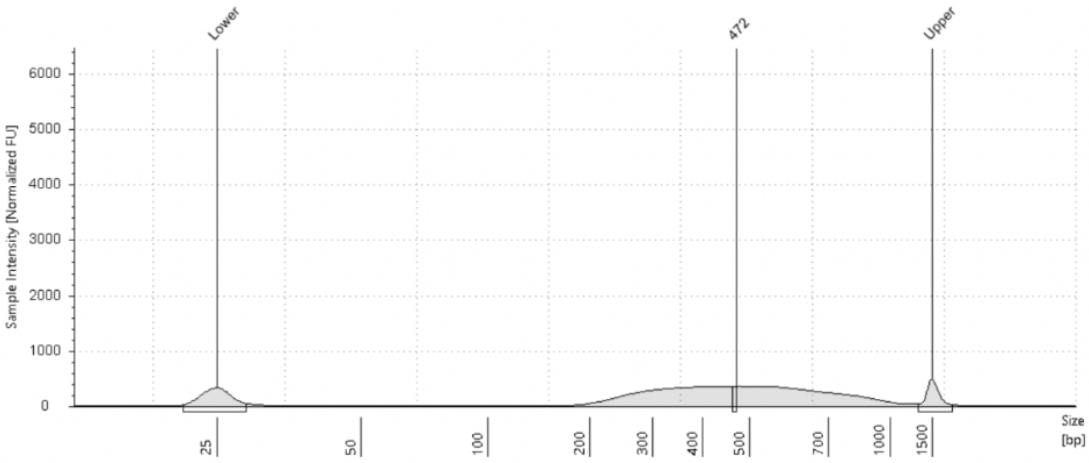
Sample Table

Well	Conc. [pg/µl]	Sample Description	Alert	Observations
L2	2160	Ladder		Ladder

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	410	-	25300	-		Lower Marker
50	225	-	6910	10.40		
100	257	-	3950	11.88		
200	269	-	2070	12.45		
300	284	-	1460	13.15		
400	293	-	1130	13.55		
500	299	-	919	13.83		
700	270	-	593	12.49		
1000	265	-	407	12.25		
1500	250	250	256	-		Upper Marker

H2: red 4



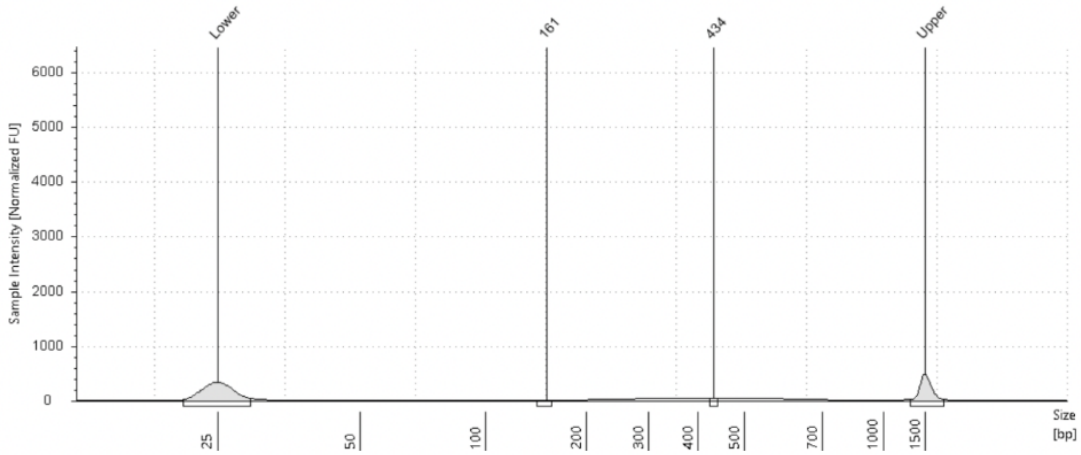
Sample Table

Well	Conc. [pg/µl]	Sample Description	Alert	Observations
H2	65.6	red 4		

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	432	-	26600	-		Lower Marker
472	65.6	-	214	100.00		
1500	250	250	256	-		Upper Marker

A3: red 5



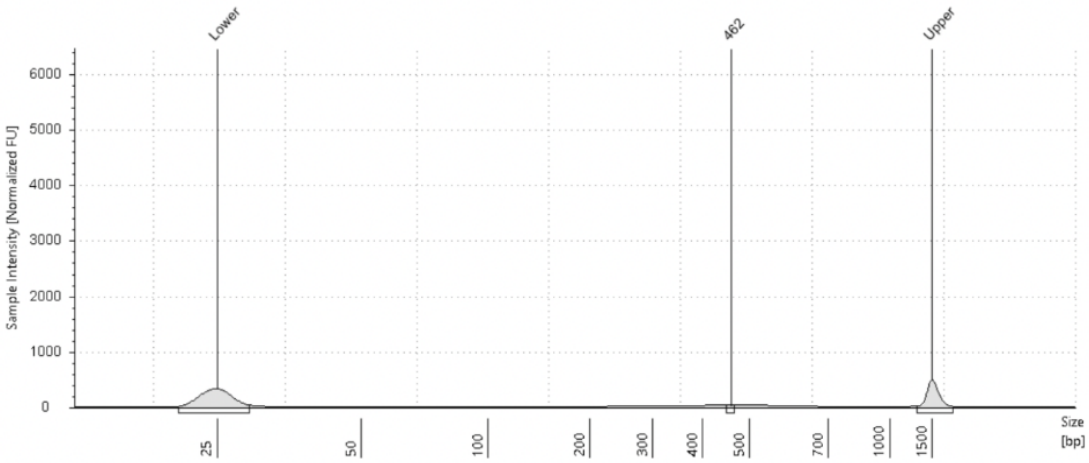
Sample Table

Well	Conc. [pg/µl]	Sample Description	Alert	Observations
A3	20.1	red 5		

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	474	-	29100	-		Lower Marker
161	3.05	-	29.2	15.18		
434	17.1	-	60.4	84.82		
1500	250	250	256	-		Upper Marker

B3: red 6



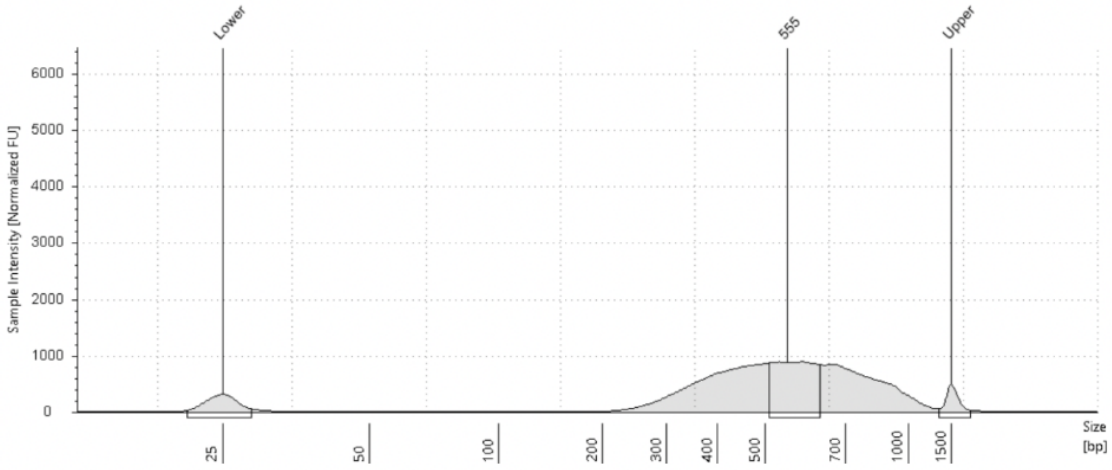
Sample Table

Well	Conc. [pg/µl]	Sample Description	Alert	Observations
B3	12.8	red 6		

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	475	-	29200	-		Lower Marker
462	12.8	-	42.6	100.00		
1500	250	250	256	-		Upper Marker

C3: red 7



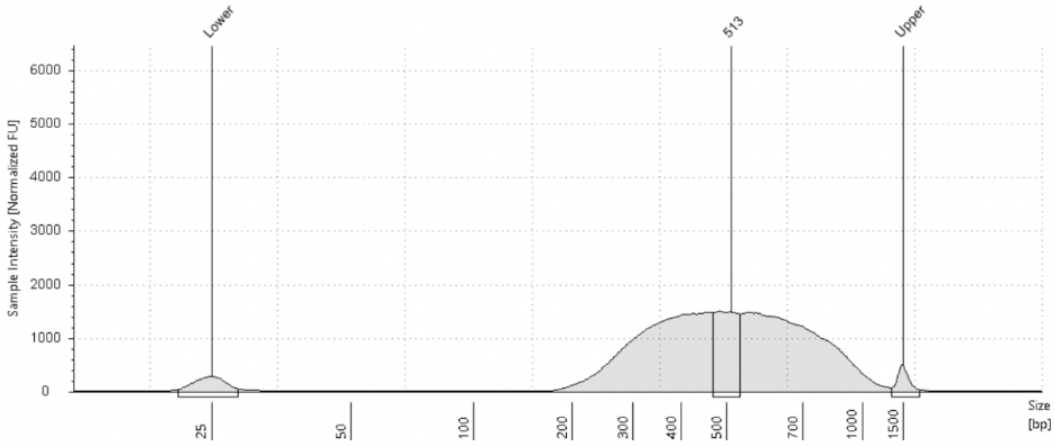
Sample Table

Well	Conc. [pg/µl]	Sample Description	Alert	Observations
C3	1670	red 7		

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	423	-	26000	-		Lower Marker
555	1670	-	4620	100.00		
1500	250	250	256	-		Upper Marker

D3: red 8



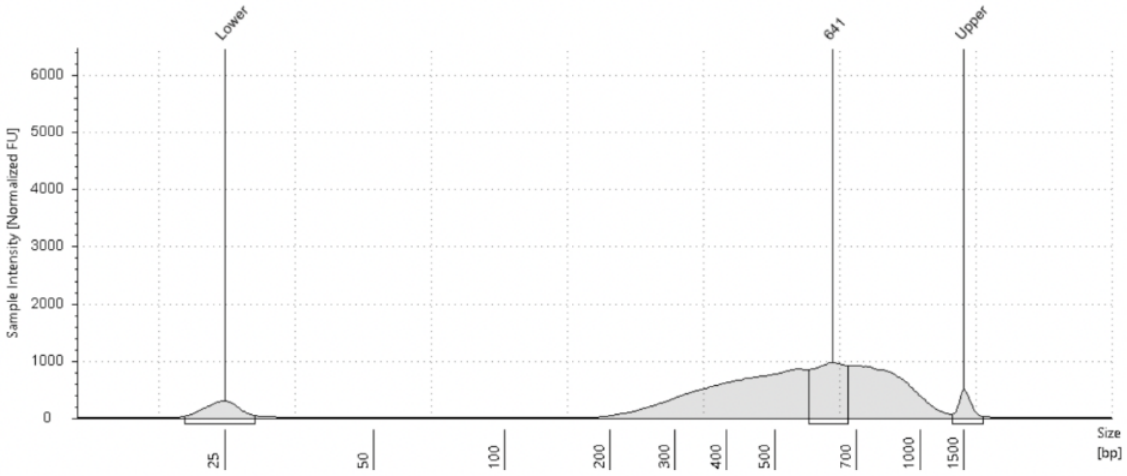
Sample Table

Well	Conc. [pg/µl]	Sample Description	Alert	Observations
D3	1630	red 8		

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	411	-	25300	-		Lower Marker
513	1630	-	4890	100.00		
1500	250	250	256	-		Upper Marker

E3: red 9



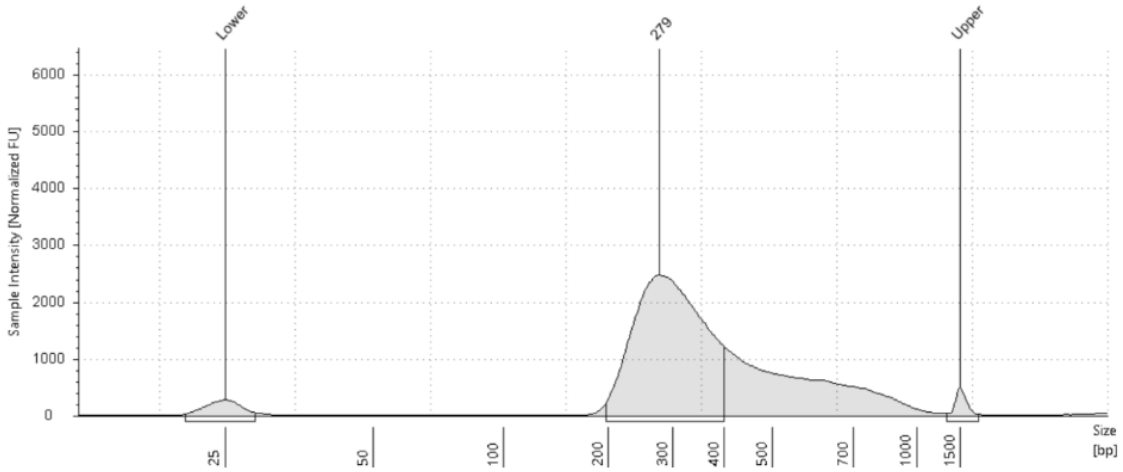
Sample Table

Well	Conc. [pg/µl]	Sample Description	Alert	Observations
E3	1340	red 9		

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	432	-	26600	-		Lower Marker
641	1340	-	3220	100.00		
1500	250	250	256	-		Upper Marker

F3: red 10



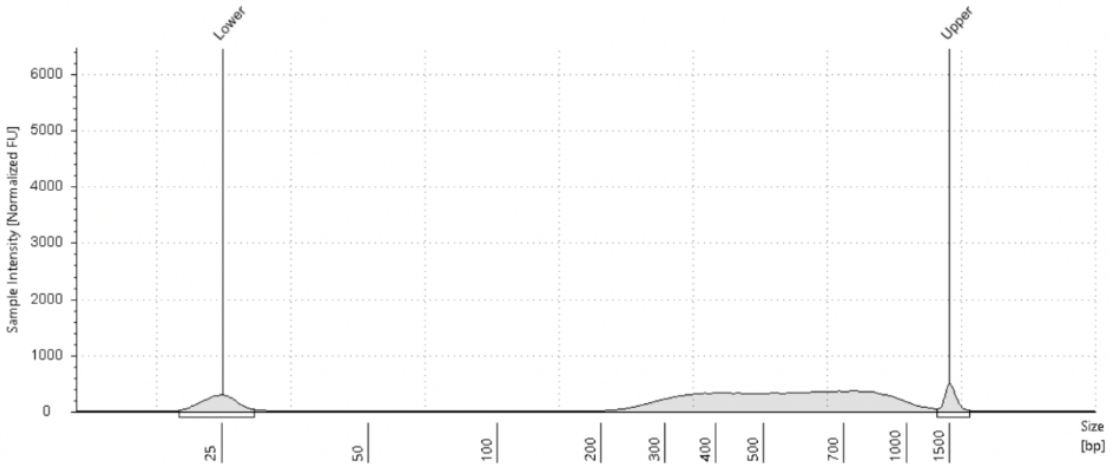
Sample Table

Well	Conc. [pg/µl]	Sample Description	Alert	Observations
F3	7910	red 10		

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	450	-	27700	-		Lower Marker
279	7910	-	43600	100.00		
1500	250	250	256	-		Upper Marker

G3: red 11



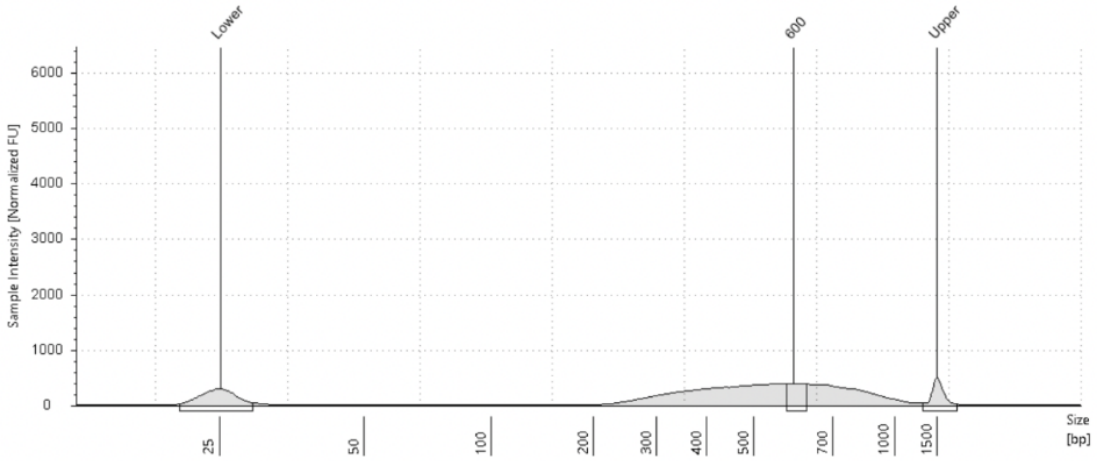
Sample Table

Well	Conc. [pg/µl]	Sample Description	Alert	Observations
G3		red 11		

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	462	-	28400	-		Lower Marker
1500	250	250	256	-		Upper Marker

H3: red 12



Sample Table

Well	Conc. [pg/µl]	Sample Description	Alert	Observations
H3	325	red 12		

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	489	-	30100	-		Lower Marker
600	325	-	835	100.00		
1500	250	250	256	-		Upper Marker