Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Charlie Decker April 10, 2024

Dual-Database Improvement of Metagenomic Viral Read Classification: A Respiratory Virus Case Study

by

Charlie Decker

Dr. Anne Piantadosi
Adviser

Biology

Dr. Anne Piantadosi

Adviser

Dr. Katia Koelle

Committee Member

Dr. Timothy D. Read

Committee Member

2024

Dual-Database Improvement of Metagenomic Viral Read Classification: A Respiratory Virus Case Study

By

Charlie Decker

Dr. Anne Piantadosi

Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Biology

2024

Abstract

Dual-Database Improvement of Metagenomic Viral Read Classification:
A Respiratory Virus Case Study
By Charlie Decker

The COVID-19 pandemic has underscored the necessity for precise identification of viral pathogens to inform clinical and public health responses effectively, especially with respiratory viruses with overlapping clinical presentations. Metagenomics, a powerful tool for the genetic profiling of complex microbial communities, has emerged as a promising solution. Utilizing high-throughput sequencing, metagenomics enables the unbiased identification of pathogens in clinical samples, offering a broad-spectrum diagnostic approach that transcends the capabilities of targeted PCR tests. This study introduces a metagenomic pipeline designed to enhance the detection and classification of viral samples, employing a combination of Kraken for initial viral read classification and BLASTN for subsequent validation.

This project's objectives were twofold: first to develop and test the dual database approach, and second to assess the efficacy of this pipeline in identifying known respiratory viruses in samples previously tested negative for COVID-19 using BinaxNOW antigen tests.

The results revealed that the pipeline successfully identified the presence of various respiratory viruses in the samples, including parainfluenza viruses 2 and 3, rhinoviruses A and C, and influenza B, showcasing its superior performance over traditional diagnostic methods. Notably, the pipeline reduced false classifications, a critical advantage in the clinical setting where accurate pathogen identification directly influences treatment decisions and infection control measures.

Dual-Database Improvement of Metagenomic Viral Read Classification: A Respiratory Virus Case Study

By

Charlie Decker

Dr. Anne Piantadosi

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Biology

2024

Acknowledgements

I would like to thank:

Dr. Piantadosi, thank you for taking me on in my final year of my undergraduate education and for giving me an opportunity to work on an interesting project and collaborate with great people.

Dr. Bombin, thank you for providing me with an introduction to metagenomic data processing and support in learning many different tools and languages all at once.

Dr. Bixler, thank you for your knowledge in preprocessing and coding, as well as advice on running BLAST.

Table of Contents

Introduction

Introduction to Metagenomics and mNGS

Metagenomics encompasses the comprehensive analysis of genetic material extracted from environmental samples such as soil, water, and air but also from human specimens, including blood, tissue, and swabs. This field diverges from conventional genetic studies, which concentrate on singular organisms or specific gene sequences. Instead, metagenomics assesses the entirety of genomic content within a sample, capturing the collective genetic signature, facilitating the unguided exploration and characterization of microorganisms, providing a panoramic view of microbial diversity.

The advent of Metagenomic Next Generation Sequencing (mNGS) has significantly advanced this field. While Next Generation Sequencing (NGS) refers to a suite of high-throughput techniques capable of sequencing DNA and RNA much more quickly and cheaply than the traditional Sanger sequencing, mNGS specifically applies these NGS technologies to metagenomic samples. This enables the simultaneous analysis of billions of DNA fragments from multiple organisms within a single sample without the need for prior knowledge about their genetic makeup (Slatko et al.). The transition from targeting specific genes or fragments to an untargeted, comprehensive analysis of all nucleic acids present underscores the leap from NGS to mNGS. This approach not only highlights mNGS's high-throughput capabilities but also its utility in exploring microbial diversity at an unprecedented scale.

Central to the success of NGS and mNGS are the development of sequencing technologies such as Illumina and Nanopore Sequencing. Illumina Sequencing technology, renowned for its sequencing-by-synthesis (SBS) method that offers unmatched accuracy and speed, works by tracking the addition of labeled nucleotides as the DNA chain is copied in a massively parallel fashion (Slatko et al.). Nanopore sequencing, on the other hand, allows for real-time sequencing of long DNA molecules by detecting changes in electrical conductivity as nucleotides pass through a protein nanopore, offering advantages in speed and flexibility. This technology supports a broad spectrum of applications, from genomic to epigenomic analyses, across various organisms. The depth and breadth of analysis possible these sequencing tools underscores their significance in fields ranging from medicine to agriculture (Slatko et al.).

Classification tools like BLAST and Kraken are essential in the metagenomic toolkit (CCB.jhu.edu). BLAST (Basic Local Alignment Search Tool) is a program that compares primary biological sequence information such as DNA and RNA. The BLAST database is populated with data from a variety of sources, mostly research studies and projects that involve sequencing of DNA or proteins that are then uploaded by the study's authors. BLAST uses a heuristic algorithm that enables researchers to identify potential genetic matches and evolutionary relationships via sequence homology (Metagenomics.wiki). Kraken further complements this by leveraging exact alignment of k-mers to assign taxonomic labels to metagenomic DNA sequences with remarkable speed and accuracy. A k-mer refers to all of a sequence's subsequences of length k, serving as a fundamental unit in sequence analysis for efficient and accurate classification. The Kraken database is made up of known k-mers (k = 31 by default) and is used by mapping sequence k-mers to the lowest common ancestor (LCA) of all genomes known to contain a given k-mer. This achieves classification speeds significantly faster than previous methods, and over 200 times faster than BLAST. Together, BLAST and Kraken embody the progress in bioinformatics tools designed to manage and interpret the expansive data produced by mNGS, facilitating a deeper understanding and broader application of microbial genomics in various fields. (Wood and Salzberg; CCB.jhu.edu).

Challenges of mNGS

The implementation of mNGS in clinical and research settings faces significant challenges. One primary issue is in read classification, which is highly sensitive and varies greatly with the pipeline used, emphasizing the need for precision in our tools (Wood et al.). While Kraken's speed and accuracy in assigning taxonomic labels to metagenomic sequences are commendable, its propensity for lower specificity underlines a critical challenge. This limitation is particularly significant in clinical and diagnostic settings, where the precise identification of pathogens is paramount. Another challenge is the large memory consumption of tools like Kraken and BLAST, especially given the massive data inputs required for metagenomic sequencing (Wood et al.). Additionally, the complexity of handling vast datasets produced by technologies like Illumina poses significant bioinformatics challenges (Lema et al.).

The accuracy and sensitivity of pathogen classification are critical, yet achieving high levels in both areas is challenging due to factors like genetic diversity and incomplete databases which

can lead to misclassification (Gupta et al.). While mNGS holds immense potential for advancing our understanding of viral communities, addressing these technical and biological complexities will require ongoing refinement of existing methodologies.

mNGS in COVID-19 Research

The COVID-19 pandemic has posed an unprecedented challenge to global public health systems, highlighting the crucial role of rapid and accurate diagnostic tools in pandemic management. The reliance on PCR and antigen tests for SARS-CoV-2, while critical, have their limitations, particularly in detecting other viral pathogens in individuals who test negative for COVID-19. This singular focus on SARS-CoV-2 may miss co-infections or alternative viral pathogens, contributing to a significant gap in our understanding of respiratory infections during the pandemic. At-home antigen tests, which many people use to quickly check for COVID-19 infection at home, underscore a significant challenge: individuals who test negative for COVID-19 at home may not seek further testing for other viral pathogens, potentially overlooking alternative causes of their symptoms.

Metagenomic Next Generation Sequencing (mNGS) offers a promising solution to this challenge. Unlike traditional diagnostic methods that target specific pathogens, mNGS provides a comprehensive overview of all genetic material in a sample, enabling the detection of a wide array of pathogens (Priya et. al.). This capability makes mNGS particularly useful in cases where patients exhibit symptoms of viral infections but test negative for COVID-19, especially when using at-home antigen tests. By leveraging the comprehensive sequencing capabilities of mNGS, we can gain a better understanding of the illnesses present in these individuals, even if they choose not to seek further testing after a negative at-home test result. Additionally, the study of COVID-negative swabs using mNGS has revealed the presence of SARS-CoV-2 in a significant percentage of cases initially diagnosed as negative by rtPCR, emphasizing the technology's potential to enhance our response to the pandemic by identifying missed cases of COVID-19 as well as alternative viral pathogens (Alteri et. al.).

mNGS may be able to significantly improve our understanding and management of respiratory infections during the COVID-19 pandemic. By identifying alternative viral pathogens in COVID-negative samples, we can improve patient care and inform public health strategies

(Waggoner et. al.). As we continue to navigate this global health crisis, broad-spectrum pathogen detection methods like mNGS will be crucial in expanding our diagnostic capacity and enhancing our response to respiratory diseases.

Using a Dual-Database Approach

We posit that the integration of a dual-database approach, which initially employs Kraken for viral read identification and classification and subsequently confirms with BLASTN, will substantially enhance the accuracy of viral read classification. We also anticipate that this increased precision will be particularly evident when assessing real-world datasets.

Our first aim focuses on the development and optimization of the metagenomic pipeline, directly building upon the computational methodology established by Dr. Andrei Bombin. His approach utilized an advanced pipeline for processing sequencing reads to ensure accurate pathogen identification through mNGS. (Piantadosi et al.) Sequencing reads underwent comprehensive quality control measures, then were run through KrakenUniq. Subsequently, reads flagged as potential human pathogenic viruses were further scrutinized using blastn.

We aim to implement a taxonomical approach for comparing the two outputs, aiming to classify the read as the lowest common ancestor of the two. We also aim to use multiple top results from both Kraken and Blast to identify the most specific common ancestor shared among them. A pivotal component of this aim will be to critically assess the pipeline's proficiency in data handling and processing, especially when juxtaposed against the current lab methodology that singularly relies on Kraken.

Our second aim revolves around assessing and furthering the enhancement of this pipeline. We plan to generate simulated reads, encompassing both known viral and non-viral sequences, and run these simulated datasets through our pipeline to gauge rates of false identification, measure accuracy, and test sensitivity. With this, we can conduct a comparative analysis with standalone Kraken classifications for a robust identification of improvements and addressing any emergent discrepancies. Based on these trials, we can add additional optimization.

Lastly, our third aim seeks to apply our metagenomic pipeline to tangible, real-world biological data. The core of this aim is to analyze COVID-19 metagenomic sequencing datasets with our

refined pipeline, with a focus on discerning non-COVID sequences on COVID-negative Binax swabs.

Methods

Pipeline Construction

Computational Environment

The pipeline is executed on an c5n.9xl AWS EC2 instance, ensuring scalable computing resources. Required libraries and dependencies include BBMap, Trimmomatic, fastp, KrakenUniq, BLAST+, seqtk, and the ete3 Python library. The metagenomic pipeline developed for this study is available on GitHub at https://github.com/valiantseal/metagen/tree/main/metagenClass/v1.5, providing open access to the code and documentation necessary for replication and further development.

Sample Preparation and Quality Control

The pipeline begins with raw sequencing data in compressed FASTQ format, derived from paired-end sequencing. prepInput.sh generates a unique directory for each sample and employs BBMap's clumpify.sh (Version 39.06) to deduplicate the data. Following deduplication, Trimmomatic (Version 0.39) is utilized for quality trimming and adapter removal using parameters such as a 4-base wide sliding window requiring a minimum average quality of 15, and leading and trailing base removal if below a quality of 3. The trimmed reads are then saved as paired-end FASTQ files for further processing.

Sequencing Data Processing

Upon preprocessing, filterMerge.sh is called, which uses fastp (Version 0.23.2) for additional quality control and merging of overlapping reads. fastp filters based on quality, length, and adapter content, merging reads with an overlap into a single sequence to simplify the dataset and enhance the reliability of the match to the reference. Parameters include a quality threshold of Q20 for base filtering and a minimum length of 50 bases for read acceptance. Post-merging, FqToFa.sh converts the quality-controlled FASTQ files to FASTA format, crucial for the compatibility with downstream bioinformatics tools, using Seqtk.

KrakenUniq Classification

krakenUniq.sh employs KrakenUniq (Version 0.5.8), which compares sequences against a viral database. KrakenUniq outputs both the classified reads and a comprehensive report, providing an overview of the taxonomic distribution within the sample. The subsequent sortKraken.R script is executed in parallel, extracting the taxonomic IDs for each read. If a taxonomic ID is a BioProject ID (i.e. not in the NCBI database), it is replaced with its NCBI parent ID. The getSampKrakReads.sh further refines the selection, isolating only the viral reads as identified and classified by Kraken.

BLASTN Homology

The runBlastNt.sh script conducts a BLAST search of the viral reads using the blastn program from the NCBI BLAST+ suite (Version 2.12.0) against a nucleotide collection database, with parameters set to optimize for viral sequences, such as an E-value cutoff of 1e-5 and a word size of 11. This homology search is parallelized with blastNtV4Par.sh for enhanced performance, ensuring each subset of the dataset is processed simultaneously to expedite the analysis. The results include many metrics, including taxonomy IDs (staxids), subject titles (stitle), E-value (evalue), and bit score (bitscore) for the top 10 matches of each read. Following BLAST, blastFiltTopSamp.R filters the results to retain only the top 3 unique, non-synthetic matches for each read, thus ensuring the highest confidence in taxonomic assignment.

Post-Processing

The final step involves tax.py, a Python script that integrates the data from both KrakenUniq and BLAST, determining the lowest common ancestor (LCA) to create a consensus classification. For each read, the six Kraken IDs from the forward and reverse sequences are compared with the six BLAST IDs from the forward and reverse sequences. The script utilizes the NCBI taxonomy database to align IDs to a standardized taxonomy, ensuring consistent classification levels. For each set of taxonomic IDs, the function calculates the LCA by finding the most recent common ancestor in the taxonomic hierarchy that encompasses all IDs. The final LCA is then outputted, providing a consensus taxonomic classification for each read.

Overall, the pipeline was enhanced via the incorporation of additional quality control steps, establishing a consensus classification between Kraken and BLAST for viral read identification,

and the introduction of an assessment module. Prior to these contributions, the pipeline primarily utilized Kraken for sorting non-viral reads and focused on viruses of interest, with subsequent BLAST analysis for viral reads.

Pipeline Assessment

Pre-Pipeline

Initially, the preprep.sh script prepares a dataset for pipeline testing by simulating reads from diverse viral genomes. Using InSilicoSeq (version 2.0.1), we generated 3,000 reads for each of 30 different viruses, spanning 10 viral groups, employing the MiSeq error model to mimic real sequencing errors accurately. The selected viruses represent a broad spectrum of viral pathogens to ensure comprehensive testing of the pipeline's identification capabilities. These include:

- Orthomyxoviridae: Influenza A (Alphainfluenzavirus influenzae), Influenza B (Betainfluenzavirus influenzae), Influenza D (Deltainfluenzavirus influenzae);
- Paramyxoviridae: Measles (Morbillivirus hominis), Mumps (Orthorubulavirus hominis), Human Parainfluenza Virus (Respirovirus laryngotracheitidis);
- Caliciviridae & Enterovirus: Norovirus (Norwalk virus), Sapporo-like virus (Sapporo virus), Poliovirus (Enterovirus C);
- Herpesviridae: Herpes Simplex Virus 1 (Simplexvirus humanalpha1 HSV-1), Herpes Simplex Virus 2 (Simplexvirus humanalpha2 HSV-2), Chickenpox/Shingles (Varicellovirus humanalpha3 VZV);
- Retroviridae: HIV-1 (Human Immunodeficiency Virus Type 1), HIV-2 (Human Immunodeficiency Virus Type 2), HTLV-1 (Primate T-lymphotropic virus 1);
- Papillomaviridae: Human Papillomavirus Type 18 (Alphapapillomavirus 7), Human Papillomavirus Type 16 (Alphapapillomavirus 9), Human Papillomavirus Type 6 (Alphapapillomavirus 10);
- Orthoflavivirus: Dengue Virus (Orthoflavivirus denguei), Zika Virus (Orthoflavivirus zikaense), West Nile Virus (Orthoflavivirus nilense);
- Alphavirus: Chikungunya Fever Virus (Chikungunya virus), Eastern Equine Encephalitis Virus

(Eastern equine encephalitis virus), Ross River Fever Virus (Ross river virus);

Coronaviridae: COVID-19 Virus (Severe acute respiratory syndrome-related coronavirus SARS-CoV-2), MERS Virus (Middle East Respiratory Syndrome Coronavirus MERS-CoV), Human Coronavirus HKU1 (Human coronavirus HKU1);

Reoviridae: Rotavirus (Rotavirus A), Colorado Tick Fever Virus (Colorado tick fever coltivirus), Rotavirus C (Rotavirus C).

The FASTAs for these viruses were sourced using the efetch utility from NCBI's database, then processed into simulated reads with InSilicoSeq to create a realistic and diverse dataset for pipeline testing.

Post-Pipeline

Taxanalysis.py compares the pipeline-generated identifications against the reference library. The analysis involves comparing the taxonomic IDs at the species and family levels between the pipeline's classifications and Kraken with the reference labels. The script defines a Positive Identification (PID) when the pipeline's classification matches the reference label at the corresponding taxonomic level, indicating accurate classification. Conversely, a Negative Identification (NID) denotes a mismatch at the same taxonomic level, reflecting a misclassification.

Binax Swab Analysis

Sample Collection

BinaxNOW COVID-19 Ag Card rapid antigen tests were collected from individuals with upper respiratory symptoms as part of an ongoing study by the Emory University Rapid Acceleration of Diagnostics (RADx) team. Test cartridges that were negative for SARS-CoV-2 were included in this study.

Sequencing

The collected swabs underwent RNA extraction (Qiagen) followed by library preparation with random hexamer cDNA synthesis (NEB) and Nextera library construction (Illumina). The

samples were then sequenced using the Illumina platform with 150bp paired-end reads, a method chosen for its high-throughput capability and accuracy.

During the sequencing run on February 9th, 2024, it was observed that the yield of nucleic acid sequences was insufficient for comprehensive analysis, falling significantly below the expected threshold for reliable metagenomic classification. To address this issue and ensure the integrity and accuracy of our pathogen identification process, a decision was made to conduct a duplicate sequencing run on March 1st, 2024.

Additionally, during the sequencing runs on June 6th, 2023, and August 4th, 2023, an unforeseen procedural oversight occurred with the use of carrier RNA (cRNA) in the RNA extraction process. Typically, extraction kits recommend the addition of carrier RNA, comprising molecules of poly-A, to improve the binding efficiency of RNA to the extraction column, thereby enhancing yield. Although this is beneficial in many contexts, for metagenomic sequencing projects like ours, it introduces a significant volume of non-target RNA, leading to the generation of unwanted sequence data. cRNA was then depleted for subsequent runs before sequencing.

Pipeline Processing

Raw fastq sequences were processed through the previously described mNGS pipeline. This involved quality control measures, deduplication, adapter trimming, merging of overlapping reads, and conversion to FASTA format for compatibility with downstream bioinformatics tools. The sequences were then subject to viral read identification and classification using KrakenUniq and homology searches via BLASTN against a nucleotide collection database. The final consensus classification, based on the lowest common ancestor (LCA) method, integrated data from both KrakenUniq and BLAST to ensure accurate taxonomic assignment.

Classification Validation

Initially, for each sample where a virus was detected, a reference-based assembly was attempted using viral-ngs (Version 2.1.19.0-rc119). This involved using a corresponding reference genome of the identified virus to construct a consensus genome assembly. Assemblies deemed to be of suboptimal quality—exhibiting low unambiguous length and insufficient depth—prompted a

subsequent homology search using MegaBLAST, which aimed to identify a more homologous reference genome that could potentially yield a higher quality assembly. The homologous reference genomes identified through MegaBLAST were then employed to reassemble the viral genomes.

Results

Pipeline Assessment Using Synthetic Data

Our assessment of the newly developed metagenomic pipeline reveals its conservative nature in pathogen identification, as demonstrated by its performance on a set of synthetic viral reads. The pipeline was evaluated for its ability to correctly identify viral pathogens at both the species and family levels, comparing its performance against the standalone use of Kraken.

We tested this pipeline on a diverse set of 30 viruses across 10 different groups, generating 3,000 reads for each virus. We assessed the Positive Identification (PID) rates, which correspond to classifications accurately matched at the species and family levels, as well as the Negative Identification (NID) rates, denoting misclassifications at these same levels. These identification rates reflect classifications that are directly comparable to the reference library at or below the species and family taxonomic levels. Classifications that occur at higher taxonomic ranks than these are not included in the PID or NID counts but still contribute to the overall assessment of the pipeline's conservative nature in pathogen identification. Unclassified reads are included in neither PID or NID, but are represented by their omission from the 3000 total reads generated for each virus.

Notably, the pipeline demonstrated a lower rate of negative identifications across the board, with zero negative IDs at the species level, indicating a reduced likelihood of false classifications. This is a critical feature for clinical and research settings, where falsely identifying a pathogen that isn’t there could have significant implications.

Furthermore, the pipeline showed higher specificity at the species and family levels compared to Kraken alone (with the exceptions of Dengue and Chikungunya viruses). For example, for the 3000 simulated reads of influenza A (Alphainfluenzavirus influenzae): 2919 of the reads were identified as influenza A or a subspecies of influenza A by Kraken, while the pipeline found 2950. The same goes for the family level, as Kraken was able to determine that 2919 of the reads belonged to the family Orthomyxoviridae, while the pipeline found 2950 reads to be classified as such. This enhancement underscores the pipeline's ability to classify viral sequences more specifically and conservatively than using only one tool.

In addition, our pipeline was tested with 30,000 reads from non-viral samples. These were generated from 10 different genomes of various eukaryotic, prokaryotic, and fungal reference genomes. It accurately identified all these samples as non-viral, demonstrating its reliability and discernment in distinguishing between viral and non-viral genetic material.

The computational efficiency of the pipeline is underscored by the real-time processing metrics, with the majority of scripts, such as 'prepInput', 'filterMerge', and 'FqtoFa', demonstrating rapid processing. Notably, 'runBlastNt' requires the most significant processing time, which is reflective of its exhaustive comparison against nucleotide databases. However, this step is integral for the accuracy and specificity of the pipeline, justifying the increased time investment. 'sortKraken' and 'tax' scripts also exhibit longer processing times, which may be attributed to the complex computational tasks they perform, such as sorting large data outputs and executing taxonomic classification algorithms.

Binax Swab Analysis

Pipeline Viral Identification

All Binax swab sequencing data was run through the pipeline, in addition to being tested for COVID (all negative), Flu A (all negative), and Flu B via PCR, as demonstrated in Table 3. Samples sequenced on February 9th were found to have insufficient read quality and were thus re-sequenced on March 1st. Detected viruses include parainfluenza viruses 2 and 3, rhinoviruses A and C, influenza B, and mastadenovirus. In addition to these human pathogens, multiple bacteriophage (pahexavirus, *Staphylococcus* phage, and *Escherichia* phage) were detected in many of the samples. Pahexavirus phage was present in every sample, suggesting reagent contamination, while other phage as well as mosaic viruses were found in various samples, suggesting the presence of other microbes. Other human viral pathogens were also detected, but in substantially lower quantities (less than 50 positive reads total), including dengue virus and Marburg virus, which we posit have no clinical significance.

Classification Validation

With many of the samples classified as having viral material, we moved to validate these pathogens using reference-based genome assembly. Reference genomes were procured from RefSeq, though the low quality of initial assemblies from these genomes necessitated the use of MegaBLAST to find higher homology references. These data are presented in Table 4. Coverage plots, separated by virus, are also provided (Figures 4-9).

PCR, Positive RPM, and RNA

The presence of COVID-19 and Influenza A are consistently the same for both PCR and the pipeline – neither tool detected either virus, which lends more confidence to the pipeline's abilities. With Influenza B, three samples tested positive via PCR: 7224J, 7125O, and 7137A. Samples 7224J and 7125O both tested positive for Influenza B by the pipeline, but for 7137A, the first round of sequencing, which was repeated due to low quality reads, tested negative; the subsequent sequencing round tested positive. This highlights the need for good quality sequencing, but also reinforces the efficacy of the pipeline, as it was able to pick up on the presence of influenza B in all the quality samples.

The PCRs raise an additional question, which is whether the number of positives per million reads (Positive RPM) correlates negatively to the PCR Ct. With such a small number of samples, this is difficult to determine. We see that the sample with the lowest Ct (10.7), 7224J, has the highest positive RPM (1742); the second lowest Ct (22.8), sample 7125O, has the second highest positive RPM (mean 879); and the highest Ct (33), sample 7137A, has the lowest positive RPM (0 for one sequencing run and 5 for the other). This would seem to suggest that the pipeline does in fact have a dose-dependent response based on the amount of viral material present in the sample, as we would expect.

The final question we wish to address with these data is the effect of carrier RNA depletion on data quality. We find that there was a higher RPM in the depleted samples compared to the undepleted samples, though this difference was not statistically significant (paired t-test $p = 0.06$) (Figure 10). However, the first round of sequencing of sample 7137A yielded no significant viral classification, while the second round did, as previously mentioned, and is corroborated by the

positive Influenza B PCR. This lends some credibility to the idea that a lack of carrier RNA depletion could affect the results of this pipeline.

Reference-Based Assemblies

The reference-based genome assemblies served to corroborate the identifications of respiratory viruses by the pipeline. For most samples, these assemblies seem to properly line up to their references, with depths greater than 15 (for at least one of the sequencing runs) and unambiguous lengths that are comparable to the reference genome. For samples 7224J and 7137A, however, we have low confidence in the classification by the pipeline. This lines up with the low positive RPM, however, as both samples were found by the pipeline to have their respective viruses in very low quantities. Future directions for this pipeline will involve investigating the exact cutoff values for positive RPMs for determining whether a virus is of clinical significance.

Discussion

Pipeline Assessment

Our newly developed metagenomic pipeline emphasizes a precise and conservative approach in pathogen identification, as evidenced by its performance on a diverse selection of viral sequences (Figure 2). The pipeline demonstrated a high level of specificity in correctly identifying viral pathogens at the species and family taxonomic levels. Notably, it consistently delivered zero negative identifications (NIDs) across all viral groups, underscoring its high selectivity and diminished likelihood of false classifications (Table 1). Such a feature is particularly crucial in clinical and epidemiological contexts, where the accurate detection of pathogens is paramount.

There was an absence of significant differences in the PID rates across various viral groups when utilizing the pipeline. This consistency in PID rates suggests a uniform performance of the pipeline, regardless of viral diversity, which is indicative of its robust and adaptable nature. However, we did observe variances in the standard deviation within these groups. The fluctuations in standard deviation may point to intrinsic properties of the viral genomes or indicate varying degrees of genetic similarity within the taxonomic categories assessed.

Binax Swab Analysis

Pipeline Viral Identification

The development and application of a metagenomic pipeline for pathogen identification in COVID-negative Binax swabs has unveiled the presence of several respiratory viruses, including parainfluenza viruses 2 and 3, rhinoviruses A and C, influenza B, and mastadenovirus (Figure 3). This discovery underscores the complexity of respiratory viral infections and their potential overlap in clinical presentations with COVID-19, raising important considerations for clinical diagnosis, epidemiology, and public health.

The symptoms caused by parainfluenza viruses, rhinoviruses, influenza B, and mastadenovirus share commonalities with those of COVID-19, including fever, cough, and difficulty breathing. Parainfluenza viruses are known to cause upper and lower respiratory illnesses, including croup,

bronchitis, and pneumonia, particularly in children (Centers for Disease Control and Prevention, 2022; Kristina Herndon, 2023). Rhinoviruses are the most frequent cause of the common cold, presenting symptoms such as cough, sneezing, runny nose, and sore throat (Centers for Disease Control and Prevention, 2023). Influenza B can lead to cough, fatigue, fever, and muscle aches, among other symptoms, typically milder than those caused by influenza A but still significant for public health (Groth, 2023).

The correct identification of respiratory viruses, including those similar in symptomatology to COVID-19, is critical for effective treatment, infection control, and public health planning. The seasonal prevalence of viruses like rhinoviruses, which peak during fall and spring, often overlaps with flu season, complicating the differentiation from COVID-19, a virus with year-round transmission (Centers for Disease Control and Prevention, 2023). The shift in transmission dynamics of these pathogens, potentially influenced by COVID-19 mitigation strategies such as mask-wearing and social distancing, underscores the complexity of managing respiratory diseases in the current era (Kim et al., 2020). Furthermore, the occurrence of co-infections with SARS-CoV-2 indicates the necessity of a comprehensive diagnostic approach, emphasizing the importance of broad surveillance and accurate pathogen identification to prevent misallocation of healthcare resources and to ensure the appropriateness of public health responses (Kim et al., 2020).

Given the overlapping symptomatology and potential for co-infection, it is imperative to advocate for the inclusion of a broad spectrum of respiratory pathogens in routine diagnostic screenings. The use of metagenomic next-generation sequencing (mNGS) pipelines, like the one developed in this project, can facilitate the simultaneous detection of multiple pathogens, offering a powerful tool for improving diagnostic accuracy and informing public health interventions.

References

Alteri C, Cento V, Antonello M, Colagrossi L, Merli M, Ughi N, Renica S, Matarazzo E, Di Ruscio F, Tartaglione L, Colombo J, Grimaldi C, Carta S, Nava A, Costabile V, Baiguera C, Campisi D, Fanti D, Vismara C, Fumagalli R, Scaglione F, Epis OM, Puoti M, Perno CF. Detection and quantification of SARS-CoV-2 by droplet digital PCR in real-time PCR negative nasopharyngeal swabs from suspected COVID-19 patients. PLoS One. 2020 Sep 8;15(9):e0236311. doi: 10.1371/journal.pone.0236311. PMID: 32898153; PMCID: PMC7478621.

Centers for Disease Control and Prevention. (2022, December 22). Clinical overview of human parainfluenza viruses (hpivs). Centers for Disease Control and Prevention. https://www.cdc.gov/parainfluenza/hcp/clinical.html#print

Centers for Disease Control and Prevention. (2023, March 8). Rhinoviruses: Common colds. Centers for Disease Control and Prevention. https://www.cdc.gov/ncird/rhinoviruses-common-cold.html#print

Cleveland Clinic. (2024, March 19). What's the difference between RSV, the flu and covid-19? https://health.clevelandclinic.org/rsv-vs-covid-vs-flu

"COVID-19 Testing." National Institutes of Health, 2020. https://covid19.nih.gov/covid-19-testing

Czubak J, Stolarczyk K, Orzeł A, Frączek M, Zatoński T. Comparison of the clinical differences between COVID-19, SARS, influenza, and the common cold: A systematic literature review. Adv Clin Exp Med. 2021 Jan;30(1):109-114. doi: 10.17219/acem/129573. PMID: 33529514.

Groth, L. (2023, November 10). Influenza B: How different is it?. Health. https://www.health.com/condition/flu/what-is-influenza-b

Gupta, Ankit, Aditya S. Malwe, Gopal N. Srivastava, Parikshit Thoudam, Keshav Hibare, Vineet K. Sharma. "MP4: a machine learning based classification tool for prediction and functional

annotation of pathogenic proteins from metagenomic and genomic datasets." BMC Bioinformatics, vol. 23, 2022, 4.

Hedberg P, Karlsson Valik J, van der Werff S, et al. Clinical phenotypes and outcomes of SARS-CoV-2, influenza, RSV and seven other respiratory viruses: a retrospective study using complete hospital data. Thorax 2022; 77:1-10.

"How to Choose Your Metagenomics Classification Tool." CCB.jhu.edu. https://ccb.jhu.edu/software/choosing-a-metagenomics-classifier/

Kim D, Quinn J, Pinsky B, Shah NH, Brown I. Rates of Co-infection Between SARS-CoV-2 and Other Respiratory Pathogens. JAMA. 2020;323(20):2085–2086. doi:10.1001/jama.2020.6266

Koskinen A, Tolvi M, Jauhiainen M, Kekäläinen E, Laulajainen-Hongisto A, Lamminmäki S. Complications of COVID-19 Nasopharyngeal Swab Test. JAMA Otolaryngol Head Neck Surg. 2021 Jul 1;147(7):672-674. doi: 10.1001/jamaoto.2021.0715. PMID: 33914064; PMCID: PMC8085764.

Kristina Herndon, R. (2023, July 20). When the flu may actually be a case of Parainfluenza. Verywell Health. https://www.verywellhealth.com/what-is-parainfluenza-770639

Lema, Niguse K., Mesfin T. Gemeda, and Adugna A. Woldesemayat. "Recent Advances in Metagenomic Approaches, Applications, and Challenges." Current Microbiology, vol. 80, 2023.

Liang, Qiaoxing, Paul W Bible, Yu Liu, Lai Wei. "DeepMicrobes: taxonomic classification for metagenomics with deep learning." NAR Genomics and Bioinformatics, vol. 2, 2020.

"Metagenomic Next Generation Sequencing: How Does It Work and Is It Coming to Your Clinical Microbiology Lab?" ASM.org. https://asm.org/Articles/2019/November/Metagenomic-Next-Generation-Sequencing-How-Does-It

"Metagenomics - BLAST." Metagenomics.wiki. https://www.metagenomics.wiki/tools/blast

"Next-Generation Sequencing (NGS)" Illumina.com. https://www.illumina.com/science/technology/next-generation-sequencing.html

Parainfluenza: Causes, symptoms, diagnosis & treatment. Cleveland Clinic. (n.d.). https://my.clevelandclinic.org/health/diseases/24522-parainfluenza

Piantadosi A, Shariatzadeh N, Bombin A, Arkun K, Alexandrescu S, Kleinschmidt-DeMasters BK, Solomon IH. Double-stranded RNA immunohistochemistry as a screening tool for viral encephalitis. Am J Clin Pathol. 2023 Aug 1;160(2):210-219. doi: 10.1093/ajcp/aqad039. PMID: 37141170; PMCID: PMC10392367.

Priya Edward, Andrew S Handel, Metagenomic Next-Generation Sequencing for Infectious Disease Diagnosis: A Review of the Literature With a Focus on Pediatrics, Journal of the Pediatric Infectious Diseases Society, Volume 10, Issue Supplement_4, December 2021, Pages S71–S77, https://doi.org/10.1093/jpids/piab104

Reiner Benaim, A., Sobel, J. A., Almog, R., Lugassy, S., Ben Shabbat, T., Johnson, A., Eytan, D., & Behar, J. A. (2021, April 13). Comparing covid-19 and influenza presentation and trajectory. Frontiers. https://www.frontiersin.org/articles/10.3389/fmed.2021.656405/full

Shi Y, Peng JM, Qin HY, Du B. Metagenomic next-generation sequencing: A promising tool for diagnosis and treatment of suspected pneumonia in rheumatic patients with acute respiratory failure: Retrospective cohort study. Front Cell Infect Microbiol. 2022 Aug 3;12:941930. doi: 10.3389/fcimb.2022.941930. PMID: 35992169; PMCID: PMC9381725.

Slatko BE, Gardner AF, Ausubel FM. Overview of Next-Generation Sequencing Technologies. Curr Protoc Mol Biol. 2018 Apr;122(1):e59. doi: 10.1002/cpmb.59. PMID: 29851291; PMCID: PMC6020069.

Waggoner JJ, Vos MB, Tyburski EA, et al. Concordance of SARS-CoV-2 Results in Self-collected Nasal Swabs vs Swabs Collected by Health Care Workers in Children and Adolescents. JAMA. 2022;328(10):935–940. doi:10.1001/jama.2022.14877

Wood, Derrick E., and Steven L. Salzberg. “Kraken: ultrafast metagenomic sequence classification using exact alignments.” Genome Biology 15.3 (2014): R46. https://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-3-r46

Wood, Derrick E., Jennifer Lu, and Ben Langmead. “Improved metagenomic analysis with Kraken 2.” Genome Biology, vol. 20, 2019.

Figures

Virus	Kraken Species PID	Kraken Species NID	Kraken Family PID	Pipeline Species PID	Pipeline Species NID	Pipeline Family PID
Alphainfluenzavirus influenzae (Orthomyxoviridae) TaxID: 2955291 (Accession AF144300.1)	2919	0	2919	2950	0	2950
Betainfluenzavirus influenzae (Orthomyxoviridae) TaxID: 2955465 (Accession AF101982.1)	2923	0	2923	2952	0	2952
Deltainfluenzavirus influenzae (Orthomyxoviridae) TaxID: 2955744 (Accession JQ922305.1)	2914	0	2914	2940	0	2940
Morbillivirus hominis (Paramyxoviridae) TaxID: 3052345 (Accession NC_001498.1)	2939	0	2942	2960	0	2960
Orthorubulavirus hominis (Paramyxoviridae) TaxID: 3052556 (Accession AB543336.1)	2928	0	2928	2950	0	2950
Respirovirus laryngotracheitidis (Paramyxoviridae) TaxID: 3049952 (Accession NC_003461)	2893	0	2902	2950	0	2950
Norwalk virus (Caliciviridae & Enterovirus) TaxID: 11983 (Accession FJ692500.1)	2916	0	2916	2948	0	2948
Sapporo virus (Caliciviridae & Enterovirus) TaxID: 95342 (Accession AF182760.1)	2933	0	2933	2968	0	2968
Enterovirus C (Caliciviridae & Enterovirus) TaxID: 138950 (Accession V01149.1)	750	0	2929	2956	0	2956
Simplexvirus humanalpha1 (HSV-1) (Herpesviridae) TaxID: 3050292 (Accession NC_001806.2)	2832	1	2935	2926	0	2926
Simplexvirus humanalpha2 (HSV-2) (Herpesviridae) TaxID: 3050293 (Accession NC_001798.2)	2237	1	2909	2936	0	2936
Varicellovirus humanalpha3 (Herpesviridae) TaxID: 3050294 (Accession X04370.1)	2923	0	2923	2940	0	2940
HIV-1 (Retroviridae) TaxID: 11676 (Accession NC_001802.1)	2776	0	2907	2918	0	2918
HIV-2 (Retroviridae) TaxID: 11709 (Accession NC_001722.1)	2822	0	2901	2956	0	2956
Primate T-lymphotropic virus 1 (Retroviridae) TaxID: 194440 (Accession AF033817.1)	2910	0	2915	2946	0	2946
Alphapapillomavirus 10 (Papillomaviridae) TaxID: 333754 (Accession NC_001355)	2936	0	2941	2954	0	2954
Alphapapillomavirus 9 (Papillomaviridae) TaxID: 337041 (Accession NC_001526.4)	2903	0	2906	2938	0	2938
Alphapapillomavirus 7 (Papillomaviridae) TaxID: 337042 (Accession NC_001357)	2912	0	2932	2956	0	2956
Orthoflavivirus nilense (Orthoflavivirus) TaxID: 3048448 (Accession NC_009942.1)	2881	0	2925	2952	0	2952
Orthoflavivirus zikaense (Orthoflavivirus) TaxID: 3048459 (Accession NC_035889.1)	2926	0	2931	2956	0	2956
Orthoflavivirus dengue (Orthoflavivirus) TaxID: 3052464 (Accession NC_002640)	2936	0	2937	2608	0	2608
Chikungunya virus (Alphavirus) TaxID: 37124 (Accession NC_004162.2)	2914	0	2950	2412	0	2950
Eastern equine encephalitis virus (Alphavirus) TaxID: 11021 (Accession NC_003899.1)	2904	1	2927	2948	0	2948
Ross river virus (Alphavirus) TaxID: 11029 (Accession NC_075016.1)	2918	0	2928	2958	0	2958
Human coronavirus HKU1 (Coronaviridae) TaxID: 290028 (Accession NC_006577.2)	2881	0	2915	2948	0	2948
SARS-CoV-2 (Coronaviridae) TaxID: 694009 (Accession NC_045512)	2563	0	2602	2817	0	2838
MERS-CoV (Coronaviridae) TaxID: 1335626 (Accession NC_019843.3)	2802	1	2919	2952	0	2952
Rotavirus A (Reoviridae) TaxID: 28875 (Accession NC_011500)	2916	0	2916	2954	0	2954
Rotavirus C (Reoviridae)	2841	0	2841	2950	0	2950

TaxID: 36427 (Accession NC_007543)						
Colorado tick fever coltivirus (Reoviridae) TaxID: 2748762 (Accession NC_004181)	2890	0	2912	2938	0	2938

Table 1: Comparative Analysis of Pathogen Identification by Kraken and the Enhanced Metagenomic Pipeline. This table presents a summary of species and family-level pathogen identification rates (PIDs) and negative identification rates (NIDs) for selected viruses, demonstrating the performance of the original Kraken tool versus our enhanced metagenomic pipeline.

Script	Real Time (s) per Million Reads
prepInput	42.1
filterMerge	3.1
FqtoFa	0.6
krakenUniq	2.5
sortKraken	57.9
getSampKrakReads	0.6
splitReads	0.6
runBlastNt	2380.1
blastFiltTopSamp	26.4
tax	59.8

Table 2: Computational Performance of Metagenomic Pipeline Components. Runtime Efficiency of Pipeline Scripts. The table details the real-time processing duration required for each script within the metagenomic pipeline per million reads.

Collaborator ID	Piantadosi Lab ID	Sequencing Date	Flu A PCR Ct	Flu B PCR Ct	cRNA Depleted?	Reads	Viruses Detected	Viral RPM
ADH6697	EHC_C19_7126P	6/6/2023	U	U	N	3,965,391	NA	
	EHC_C19_7126P_L3	3/1/2024	U	U	Y	2,344,551	NA	
ADH6701	EHC_C19_7127Q	6/6/2023	U	U	N	2,891,815	NA	
	EHC_C19_7127Q_L3	3/1/2024	U	U	Y	1,097,083	NA	
AEP6533	EHC_C19_7117G	6/6/2023	U	U	N	2,666,880	NA	
	EHC_C19_7117G_L3	2/9/2024	U	U	Y	175,727	NA	
	EHC_C19_7117G_L3	3/1/2024	U	U	Y	856,240	NA	
AEP6674	EHC_C19_7119I	6/6/2023	U	U	N	4,582,782	Rhinovirus A	420
	EHC_C19_7119I_L3	2/9/2024	U	U	Y	239,277	Rhinovirus A	7180
AEP6681	EHC_C19_7123M	6/6/2023	U	U	N	2,908,135	NA	
	EHC_C19_7123M_L3	2/9/2024	U	U	Y	56,420	NA	
	EHC_C19_7123M_L3	3/1/2024	U	U	Y	240,003	NA	
AEP6755	EHC_C19_7144H	6/6/2023	U	U	N	3,692,204	NA	
	EHC_C19_7144H_L3	3/1/2024	U	U	Y	1,663,254	NA	
AEP6758	EHC_C19_7219E	8/4/2023	U	U	N	1,534,221	NA	
AEP6807	EHC_C19_7234T	8/4/2023	U	U	N	1,041,928	NA	
AEU6718	EHC_C19_7132V	6/6/2023	U	U	N	1,933,722	NA	
	EHC_C19_7132V_L3	3/1/2024	U	U	Y	3,408,615	NA	
AEU6744	EHC_C19_7140D	6/6/2023	U	U	N	3,495,102	NA	
	EHC_C19_7140D_L3	3/1/2024	U	U	Y	1,494,282	NA	
AEU6746	EHC_C19_7141E	6/6/2023	U	U	N	2,007,292	NA	
	EHC_C19_7141E_L3	3/1/2024	U	U	Y	5,451,728	NA	
AGN6535	EHC_C19_7115E	6/6/2023	U	U	N	4,851,208	Parainfluenza 2	815
	EHC_C19_7115E_L3	2/9/2024	U	U	Y	316,658	Parainfluenza 2	41347
	EHC_C19_7115E_L3	3/1/2024	U	U	Y	595,889	Parainfluenza 2	3052
AGN6545	EHC_C19_7116F	6/6/2023	U	U	N	8,331,344	NA	
	EHC_C19_7116F_L3	2/9/2024	U	U	Y	287,376	NA	
AGN6545	EHC_C19_7116F_L3	3/1/2024	U	U	Y	1,442,794	NA	
AGN6685	EHC_C19_7118H	6/6/2023	U	U	N	2,120,434	Parainfluenza 3	9048
	EHC_C19_7118H_L3	2/9/2024	U	U	Y	58,149	Parainfluenza 3	35426
	EHC_C19_7118H_L3	3/1/2024	U	U	Y	254,364	Parainfluenza 3	1902
AGN6688	EHC_C19_7120J	6/6/2023	U	U	N	3,894,710	NA	
	EHC_C19_7120J_L3	2/9/2024	U	U	Y	1,618,020	NA	
	EHC_C19_7120J_L3	3/1/2024	U	U	Y	4,329,258	NA	
AGN6692	EHC_C19_7121K	6/6/2023	U	U	N	3,901,096	NA	
	EHC_C19_7121K_L3	2/9/2024	U	U	Y	73,787	NA	
	EHC_C19_7121K_L3	3/1/2024	U	U	Y	524,436	NA	
AGN6693	EHC_C19_7131U	6/6/2023	U	U	N	3,086,900	NA	
	EHC_C19_7131U_L3	3/1/2024	U	U	Y	3,553,983	NA	
AGN6695	EHC_C19_7124N	6/6/2023	U	U	N	3,036,288	Rhinovirus A	1216

	EHC_C19_7124N_L3	2/9/2024	U	U	Y	318,457	Rhinovirus A	102717
	EHC_C19_7124N_L3	3/1/2024	U	U	Y	3,331,452	Rhinovirus A	28294
AGN6696	EHC_C19_7125O	6/6/2023	U	22.8	N	3,271,646	Influenza B	129
	EHC_C19_7125O_L3	2/9/2024	U	22.8	Y	81,147	Influenza B	2378
	EHC_C19_7125O_L3	3/1/2024	U	22.8	Y	363,421	Influenza B	131
AGN6698	EHC_C19_7128R	6/6/2023	U	U	N	3,198,244	NA	
	EHC_C19_7128R_L3	3/1/2024	U	U	Y	3,126,522	NA	
AGN6699	EHC_C19_7129S	6/6/2023	U	U	N	3,813,462	Rhinovirus A	481
	EHC_C19_7129S_L3	3/1/2024	U	U	Y	4,349,347	Rhinovirus A	393
AGN6702	EHC_C19_7130T	6/6/2023	U	U	N	2,077,262	Rhinovirus A	819
	EHC_C19_7130T_L3	3/1/2024	U	U	Y	13,844,494	Rhinovirus A	4952
AGN6717	EHC_C19_7133W	6/6/2023	U	U	N	2,484,640	NA	
	EHC_C19_7133W_L3	3/1/2024	U	U	Y	6,387,440	NA	
AGN6723	EHC_C19_7134X	6/6/2023	U	U	N	3,049,650	Rhinovirus C	9648
	EHC_C19_7134X_L3	3/1/2024	U	U	Y	12,790,887	Rhinovirus C	29423
AGN6725	EHC_C19_7135Y	6/6/2023	U	U	N	1,438,524	NA	
	EHC_C19_7135Y_L3	3/1/2024	U	U	Y	1,404,583	NA	
AGN6729	EHC_C19_7136Z	6/6/2023	U	U	N	3,244,013	Parainfluenza 3	13949
	EHC_C19_7136Z_L3	3/1/2024	U	U	Y	2,618,215	Parainfluenza 3	70576
AGN6734	EHC_C19_7137A	6/6/2023	U	33.0	N	3,496,442	NA	
	EHC_C19_7137A_L3	3/1/2024	U	33.0	Y	1,293,554	Influenza B	5
AGN6742	EHC_C19_7138B	6/6/2023	U	U	N	3,290,867	NA	
	EHC_C19_7138B_L3	3/1/2024	U	U	Y	4,754,229	NA	
AGN6743	EHC_C19_7139C	6/6/2023	U	U	N	2,664,966	NA	
	EHC_C19_7139C_L3	3/1/2024	U	U	Y	3,912,338	NA	
AGN6748	EHC_C19_7142F	6/6/2023	U	U	N	3,753,324	NA	
	EHC_C19_7142F_L3	3/1/2024	U	U	Y	4,722,446	NA	
AGN6750	EHC_C19_7143G	6/6/2023	U	U	N	3,250,073	NA	
	EHC_C19_7143G_L3	3/1/2024	U	U	Y	5,185,105	NA	
AGN6761	EHC_C19_7145I	6/6/2023	U	U	N	3,119,186	Parainfluenza 3	3177
	EHC_C19_7145I_L3	3/1/2024	U	U	Y	9,961,100	Parainfluenza 3	140113
AGN6768	EHC_C19_7220F	8/4/2023	U	U	N	1,086,900	NA	
AGN6784	EHC_C19_7223I	8/4/2023	U	U	N	1,252,493	Influenza B	1416
AGN6785	EHC_C19_7224J	8/4/2023	U	10.7	N	931,717	Influenza B, Mastadenovirus E	1742 (Flu), 54 (Mast)
AGN6788	EHC_C19_7226L	8/4/2023	U	U	N	738,478	NA	
AGN6794	EHC_C19_7227M	8/4/2023	U	U	N	1,111,963	Parainfluenza 3	1248
AGN6798	EHC_C19_7230P	8/4/2023	U	U	N	1,434,279	NA	
AGN6805	EHC_C19_7231Q	8/4/2023	U	U	N	1,371,573	NA	
AMT6523	EHC_C19_7114D	6/6/2023	U	U	N	3,844,230	NA	
	EHC_C19_7114D_L3	2/9/2024	U	U	Y	250,355	NA	
	EHC_C19_7114D_L3	3/1/2024	U	U	Y	1,965,753	NA	
AMT6694	EHC_C19_7122L_L2	6/6/2023	U	U	N	3,953,898	NA	

	EHC_C19_7122L_L3	2/9/2024	U	U	Y	122,273	NA	
	EHC_C19_7122L_L3	3/1/2024	U	U	Y	539,288	NA	
AMT6781	EHC_C19_7222H_L2	8/4/2023	U	U	N	79,731	NA	
APD6801	EHC_C19_7229O_L2	8/4/2023	U	U	N	800,105	NA	
APD6809	EHC_C19_7232R_L2	8/4/2023	U	U	N	982,100	NA	
APD6810	EHC_C19_7233S_L2	8/4/2023	U	U	N	1,288,364	NA	
ASJ6519	EHC_C19_7113C	6/6/2023	U	U	N	3,374,029	NA	
	EHC_C19_7113C_L3	3/1/2024	U	U	Y	4,032,146	NA	
ASJ6770	EHC_C19_7221G_L2	8/4/2023	U	U	N	1,850,568	NA	

Table 3: Detection of Viral Sequences in Negative Binax Now COVID-19 Antigen Test Swabs. Summary of viral sequences detected using the bioinformatics pipeline on swabs that tested negative with the Binax Now COVID-19 Antigen Test. Flu and COVID PCRs were run on some samples and compared with the pipeline's viral identification.

ID	Date	Viruses Detected	Positive RPM	Reference Genome	UA Assembly Length	Assembly Length	Reference Genome Length	Assembly Depth
EHC_C19_7224J	8/4/2023	Mastadenovirus E	54	Mastadenovirus E strain HAdV-E/USA/3477/2015/P4H4F4 RefSeq (Accession KY996446.1)	1550	35131	35949	0.6
EHC_C19_7115E_L3	3/1/2024	Parainfluenza 2 virus	3052	Parainfluenza 2 RefSeq (Accession NC_003443)	15567	15567	15646	249.3
EHC_C19_7118H_L3	3/1/2024	Parainfluenza 3 virus	1902	Parainfluenza 3 RefSeq (Accession NC_075446.1)	15327	15327	15462	69.4
EHC_C19_7136Z_L3	3/1/2024	Parainfluenza 3 virus	70576	Parainfluenza 3 RefSeq (Accession NC_075446.1)	15367	15367	15462	2719.1
EHC_C19_7145I_L3	3/1/2024	Parainfluenza 3 virus	140113	Parainfluenza 3 RefSeq (Accession NC_075446.1)	15360	15361	15462	5377.5
EHC_C19_7227M	8/4/2023	Parainfluenza 3 virus	1248	Parainfluenza 3 RefSeq (Accession NC_075446.1)	15143	15216	15462	26.1
EHC_C19_7137A_L3	3/1/2024	Influenza B virus	5	Influenza B virus genome ASM3108572v1 (Accession GCA_031085725.1)	172	242	14623	0.2
EHC_C19_7223I	8/4/2023	Influenza B virus	1416	Influenza B virus genome ASM3108572v1 (Accession GCA_031085725.1)	14553	14555	14623	55.7
EHC_C19_7224J	8/4/2023	Influenza B virus	1742	Influenza_B_CA_15_2018_with_spacers	13569	15098	15098	36.7
EHC_C19_7125O_L3	3/1/2024	Influenza B virus	131	Influenza B virus genome ASM3108572v1 (Accession GCA_031085725.1)	12240	13344	14623	16.9
EHC_C19_7119I	6/6/2023	Rhinovirus A	420	Rhinovirus A51 strain ATCC VR-1161 (Accession FJ445136.1)	6026	6995	7152	34.9
EHC_C19_7124N_L3	3/1/2024	Rhinovirus A	28294	Rhinovirus A RefSeq (Accession NC_001617.1)	6617	7122	7152	1540.2
EHC_C19_7129S_L3	3/1/2024	Rhinovirus A	393	Rhinovirus A30 strain ATCC VR-1140 (Accession FJ445179.1)	4255	6972	7093	22.8
EHC_C19_7130T_L3	3/1/2024	Rhinovirus A	4952	Rhinovirus A54 strain SC176 (Accession KY369875.1)	7103	7103	7104	354.5
EHC_C19_7134X_L3	3/1/2024	Rhinovirus C	10543	Rhinovirus C11 strain CL-170085 (Accession EU840952.2)	7053	6941	7108	279.3

Table 4: Summary Statistics of Reference-Based Assembly. Summary of assembly statistics for individual viruses detected, demonstrated unambiguous (UA) assembly length, assembly length, reference genome length, and depth, as well as positive reads per million (positive RPM).

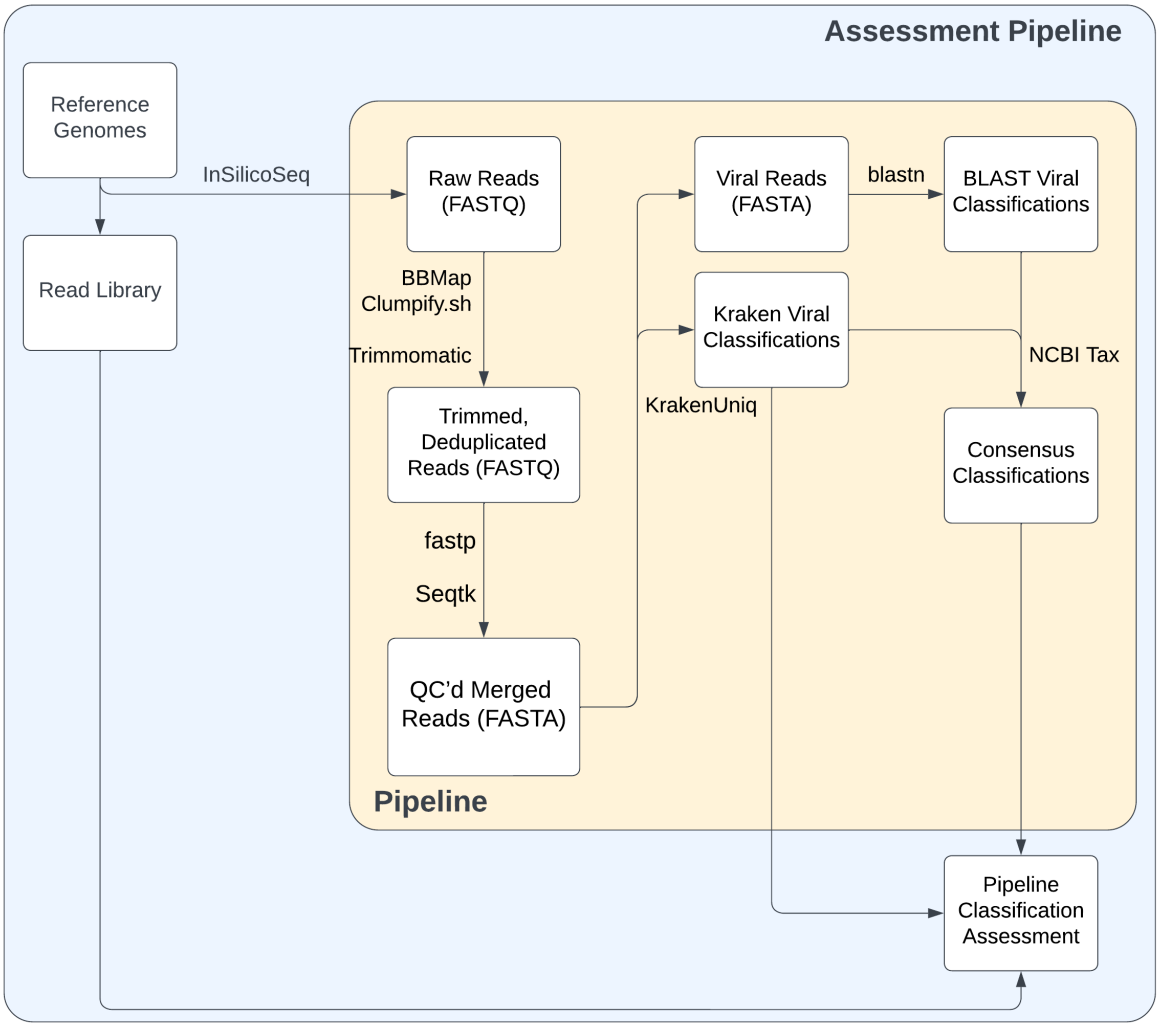

Figure 1: Enhanced Metagenomic Pipeline Workflow for Viral Pathogen Identification. In the main pipeline (orange), raw paired-end sequencing reads are prepared and quality-controlled, including deduplication with BBMap's clumpify.sh and trimming via Trimmomatic. Quality checks and merging of reads are conducted with fastp and Seqtk, followed by FASTA conversion. KrakenUniq classifies these processed reads against a viral database. Homology-based validation is performed using BLAST, with top matches filtered for further analysis. The pipeline concludes with the integration of Kraken and BLAST classifications, applying a Python script to establish consensus classifications by determining the lowest common ancestor (LCA) for each read using NCBI Taxonomy data. Outside of the main pipeline (blue), InSilicoSeq simulates reads from a range of reference genomes and the pipeline readies the dataset for viral pathogen identification. After the core pipeline runs, the Pipeline Classification Assessment uses

the generated read library to compare the pipeline's output against reference labels, defining Positive Identifications (PID) when matches at the species or family level are confirmed, or Negative Identifications (NID) when misclassifications occur, thus evaluating the pipeline's accuracy and efficacy in pathogen identification.

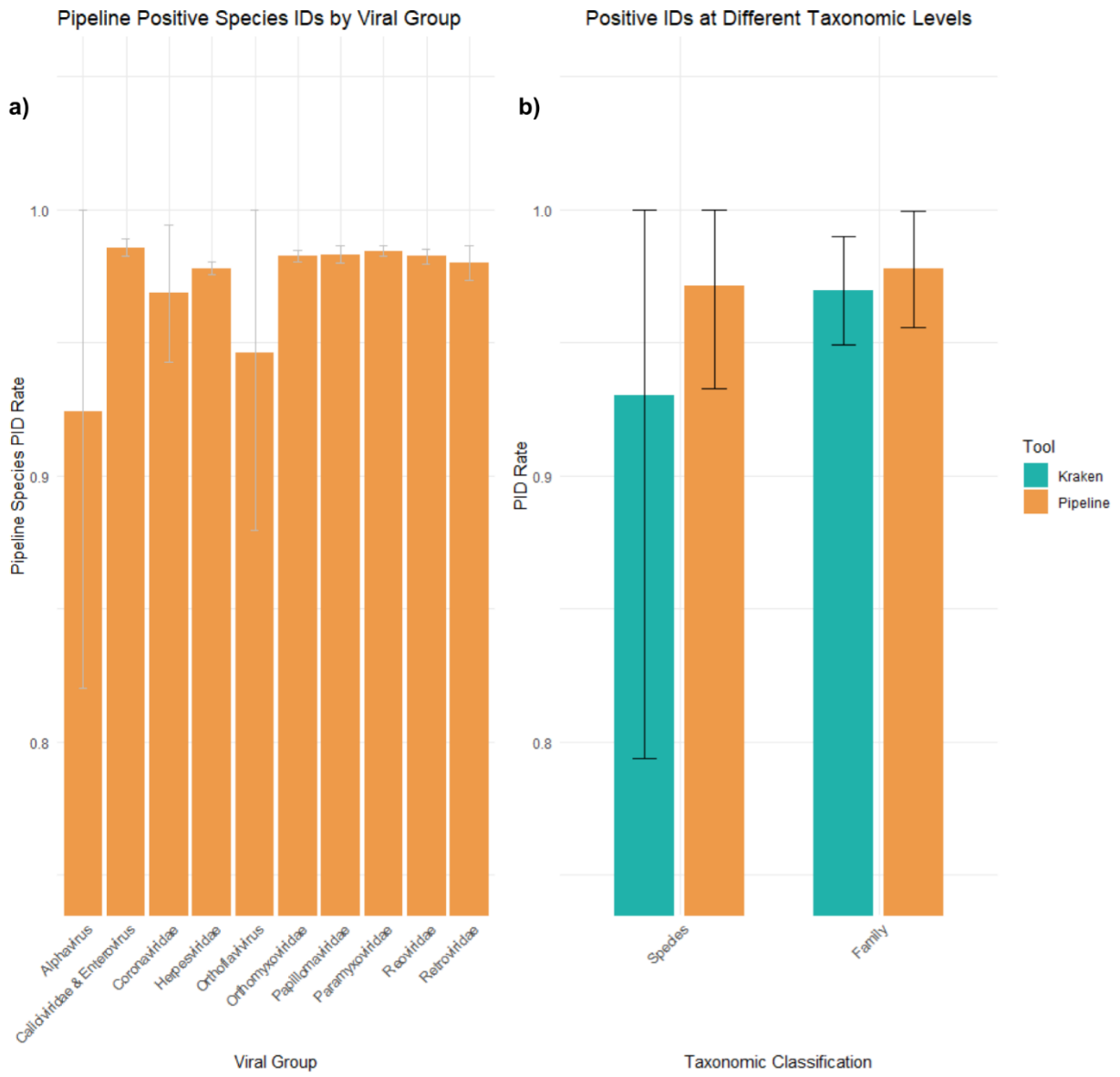

Figure 2: Comparative Analysis of Pathogen Identification Across Viral Groups and Tools. The graphs illustrate the Positive Identification (PID) rates obtained from two analytical tools across various viral groups (a) and taxonomic classifications (b). In graph (a), each bar represents the average PID rate for a viral group as identified by the Pipeline tool. In graph (b), the PID rates are shown at two different taxonomic levels, Species and Family, with the blue bars indicating Kraken tool results and the orange bars indicating Pipeline tool results. Error bars represent the standard deviation, indicating the variation in PID rates within each group.

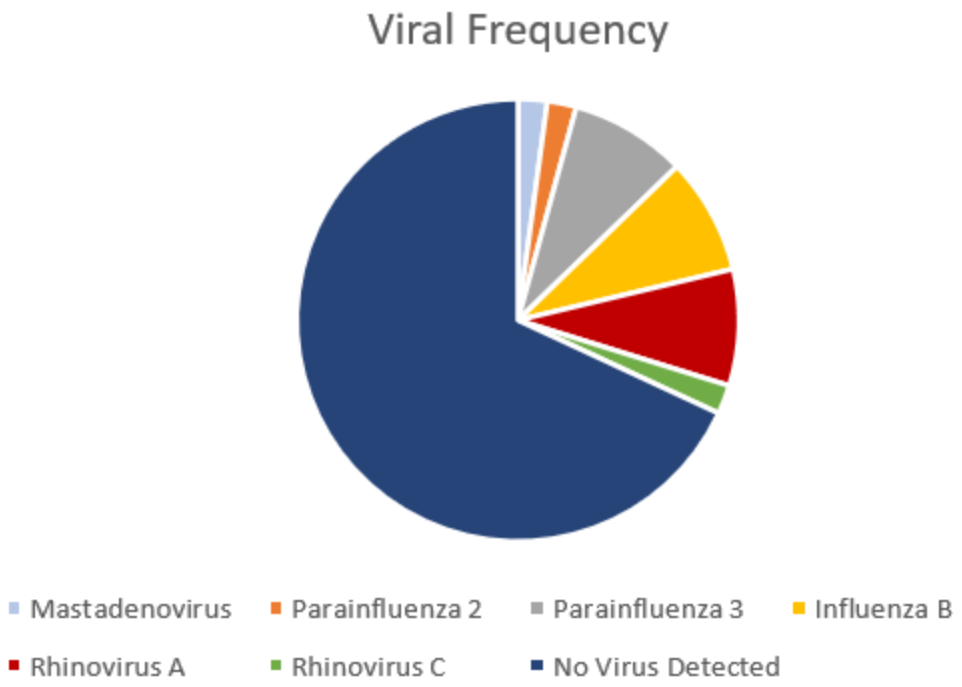

Figure 3: Viral Frequency. Proportion of 48 total unique samples containing each of the classified viruses. 1 sample had human mastadenovirus E, 1 sample had parainfluenza 2, 4 samples had parainfluenza 3, 4 samples had influenza B, 4 samples had rhinovirus A, 1 sample had rhinovirus C, and the other 34 samples had no detected viruses.

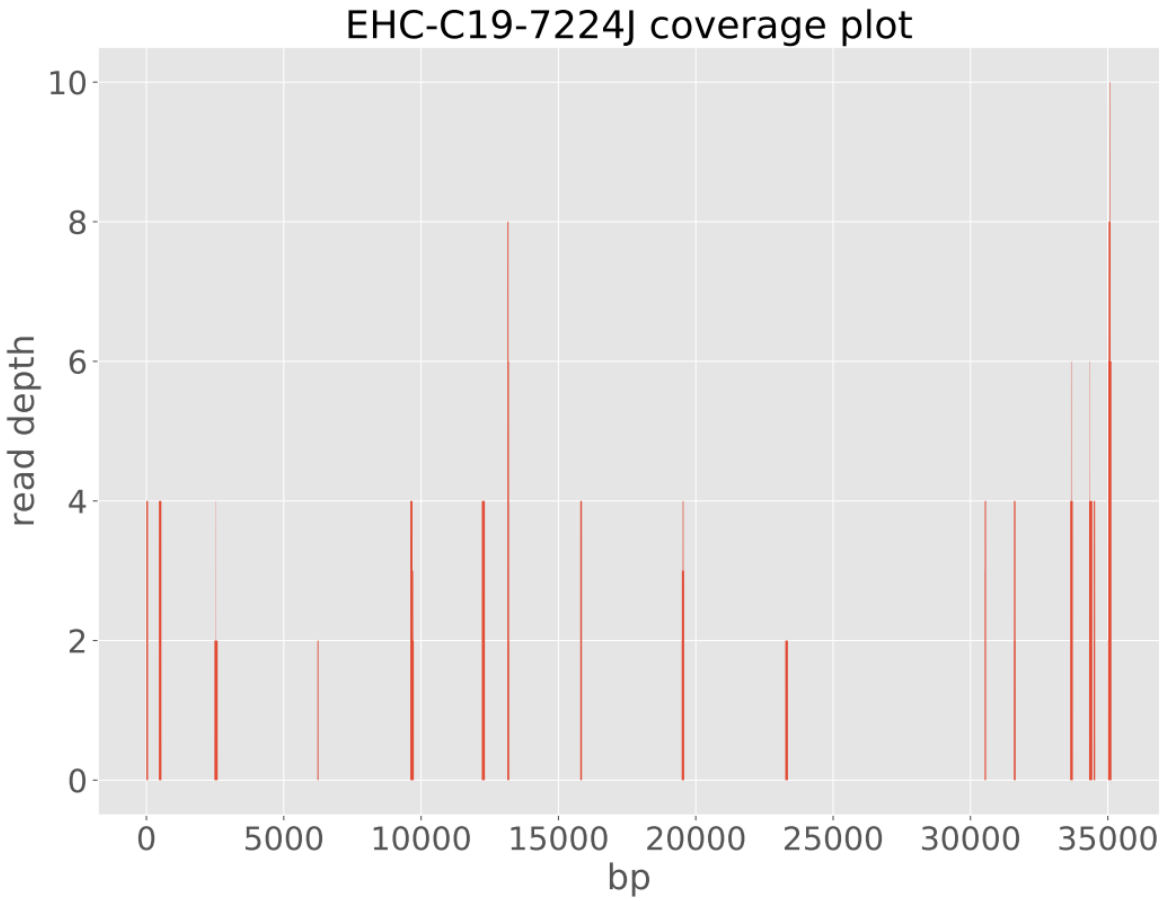

Figure 4: Coverage Plot for Mastadenovirus E. The self-coverage plot of human mastadenovirus E for sample 7224J.

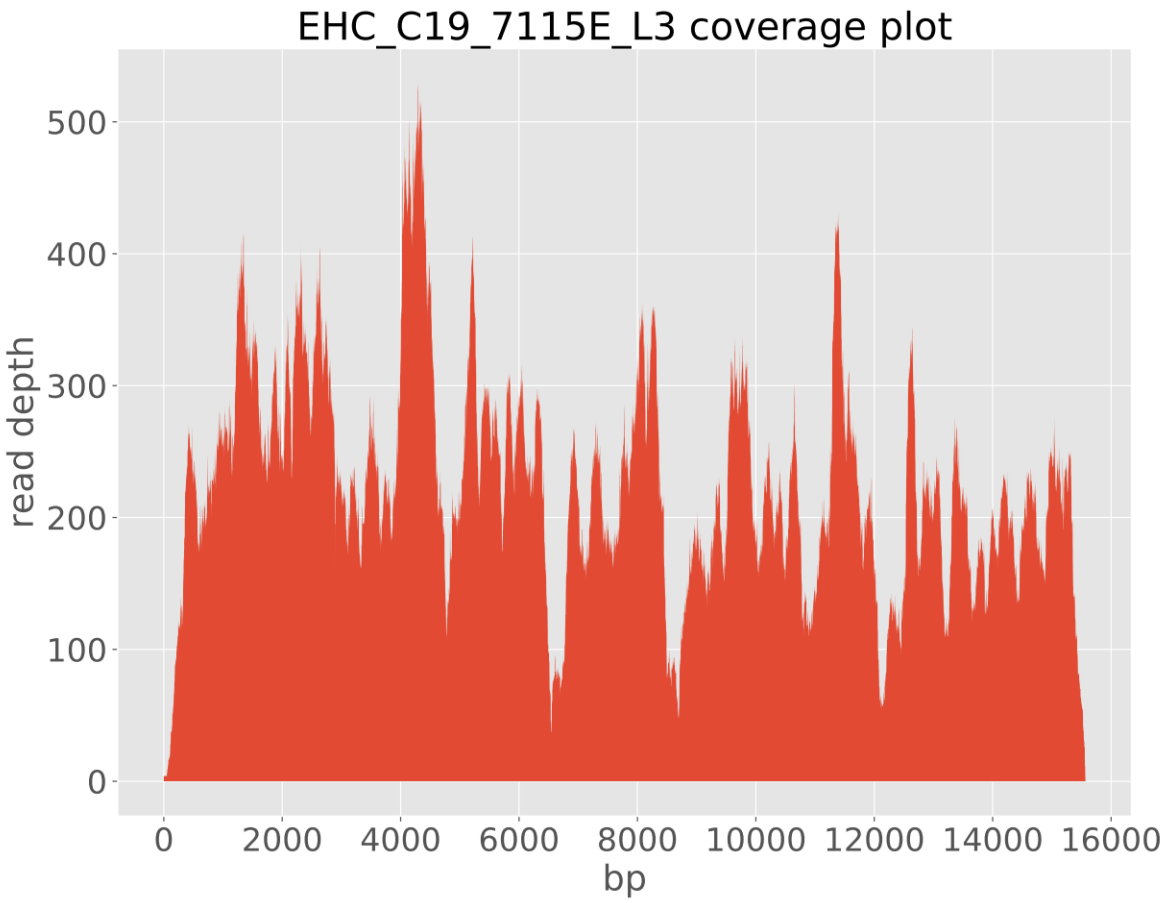

Figure 5: Coverage Plots for Parainfluenza Virus 2. The self-coverage plot of human parainfluenza virus 2 for sample 711E.

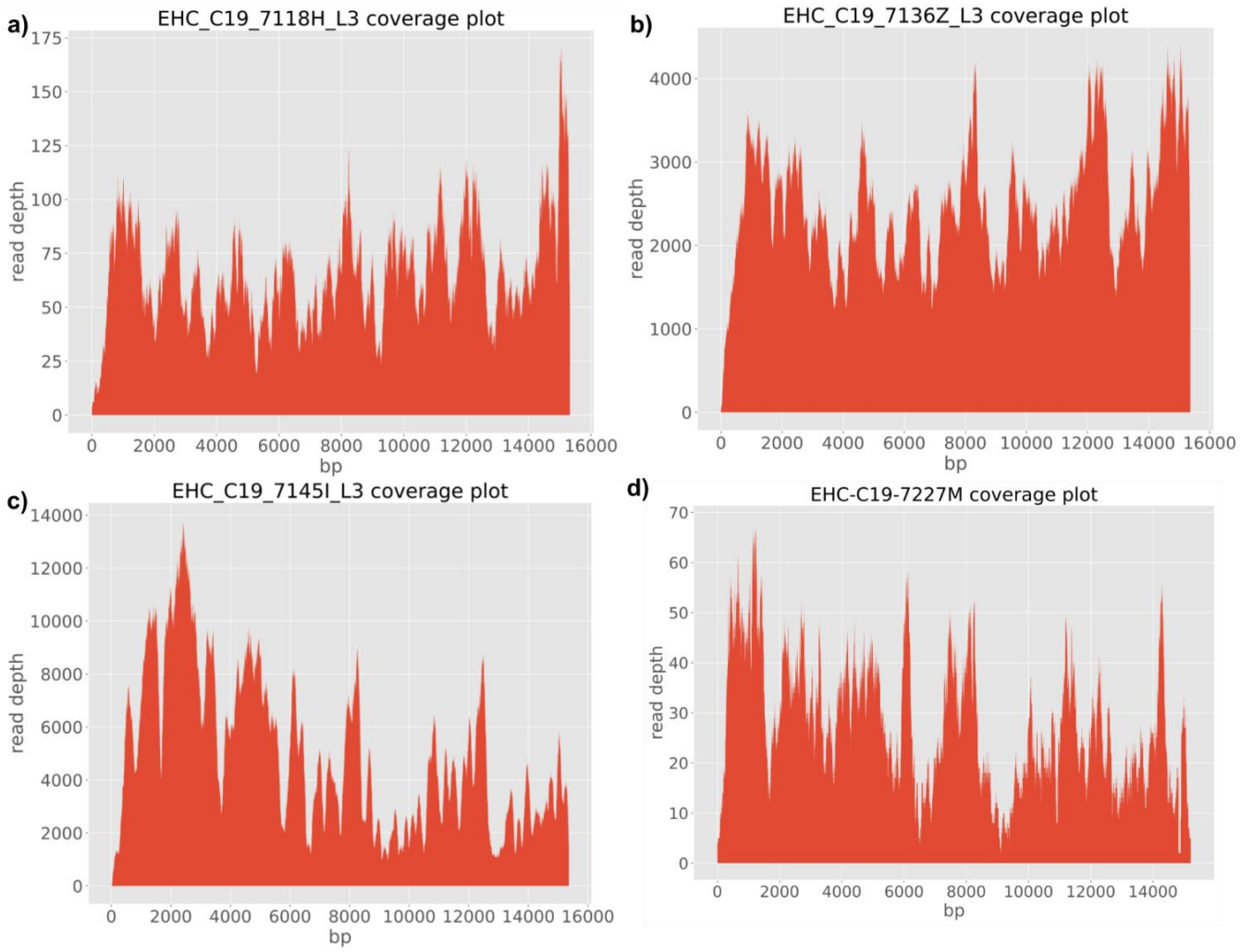

Figure 6: Coverage Plots for Parainfluenza Virus 3. The self-coverage plots of human parainfluenza virus 3 for samples 7118H (a), 7136Z (b), 7145I (c), and 7227M (d).

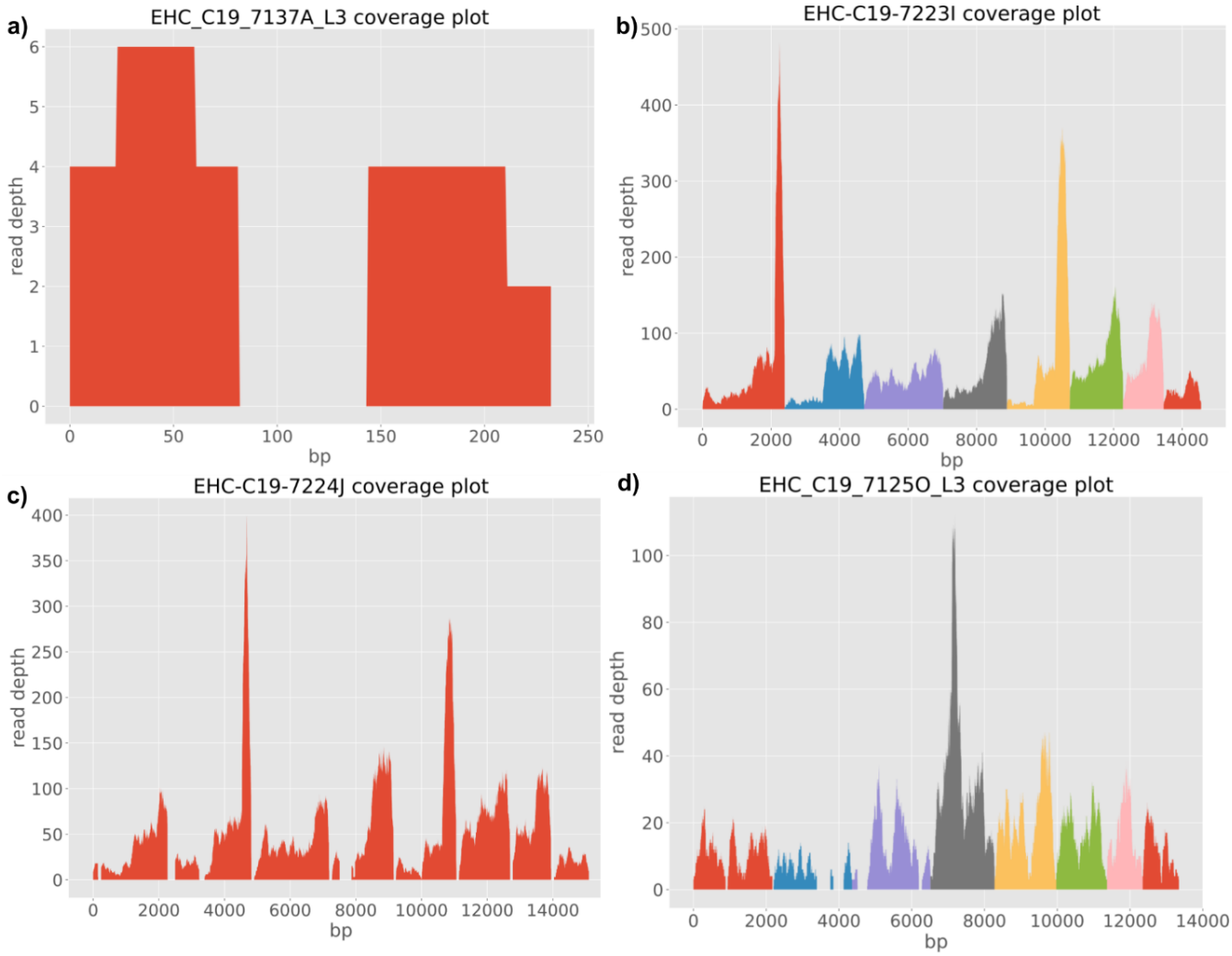

Figure 7: Coverage Plots for Influenza B Virus. The self-coverage plots of influenza B virus for samples 7137A (a), 7223I (b) 7224J (c), and 7125O (d).

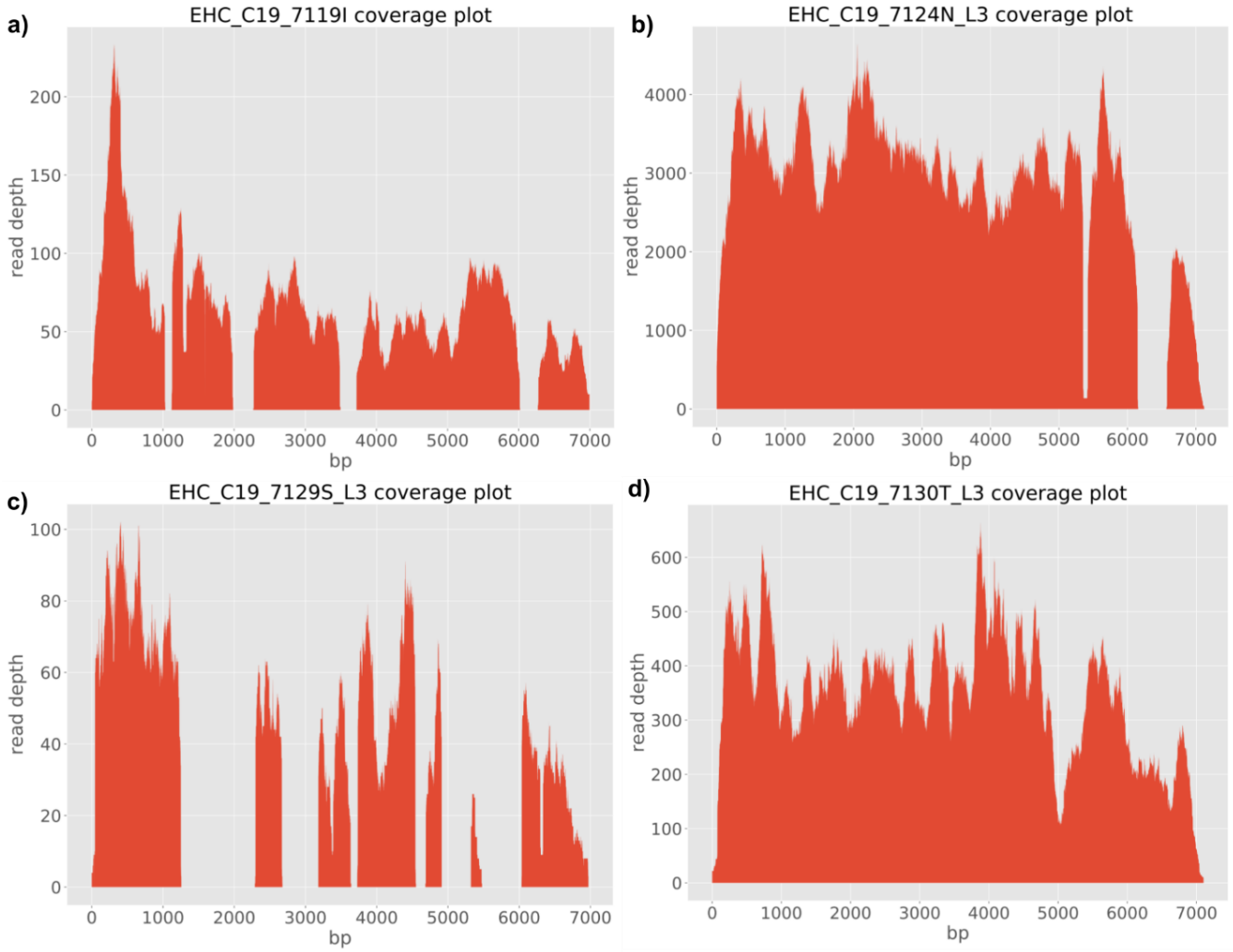

Figure 8: Coverage Plots for Rhinovirus A. The self-coverage plots of human rhinovirus A for samples 7119I (a), 7124N (b), 7129S (c), sample 7130T (d).

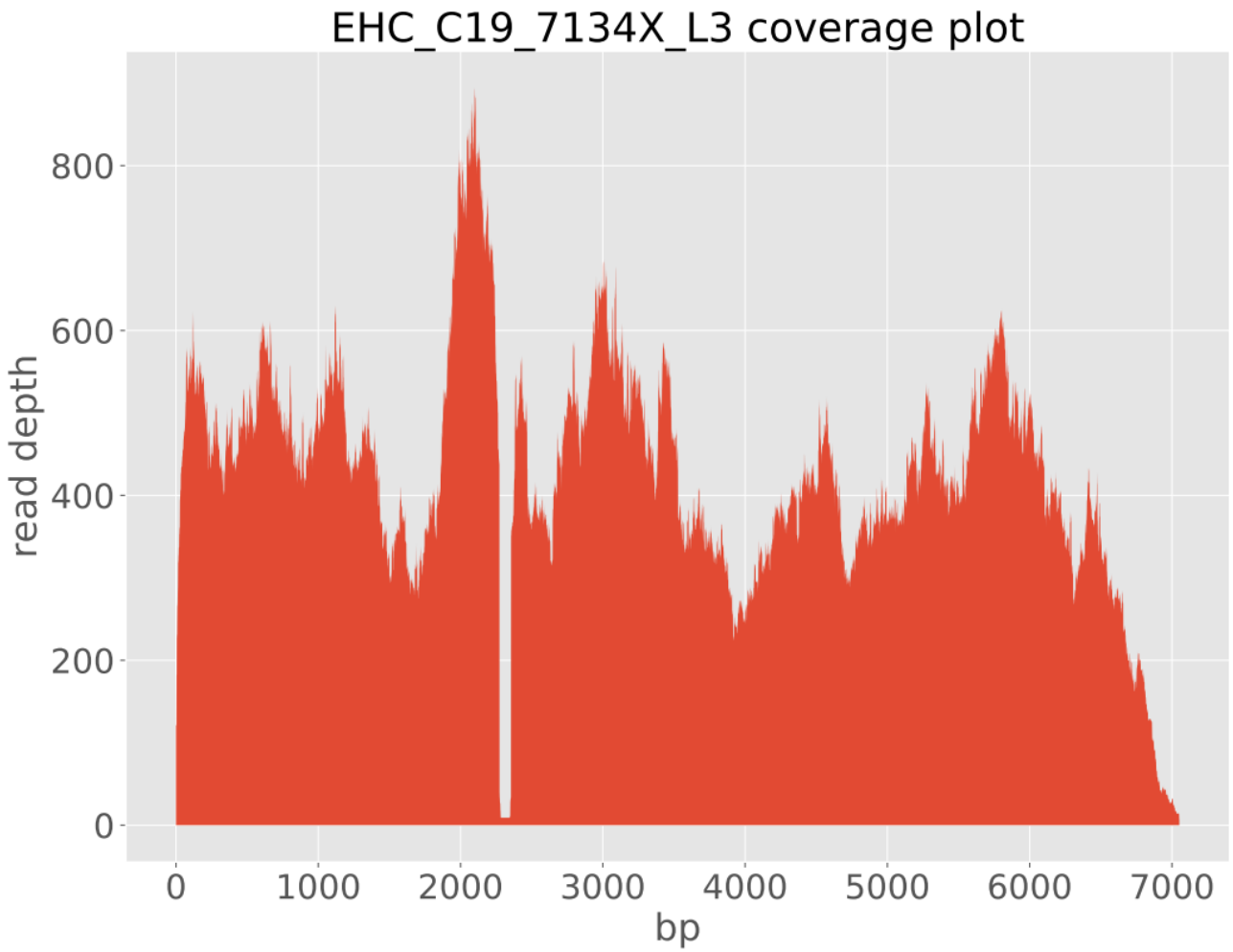

Figure 9: Coverage Plots for Rhinovirus C. The self-coverage plot of human rhinovirus C for sample 7134X.

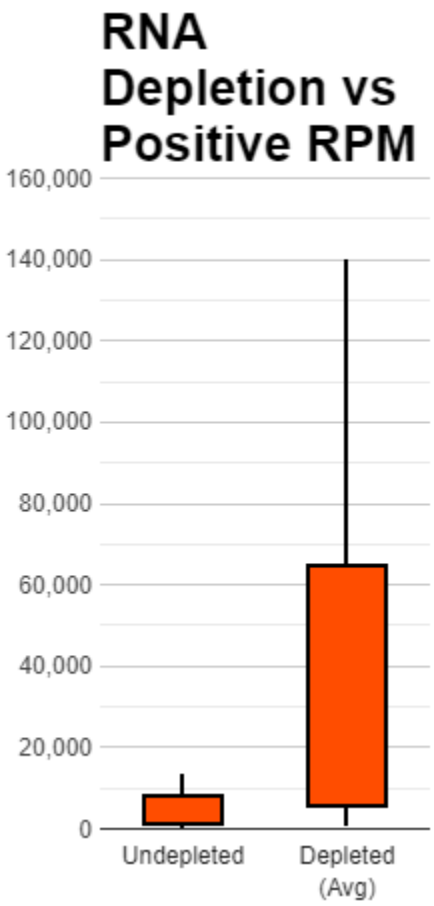

Figure 10: RNA Depletion vs Positive RPM. Candlestick plot demonstrating the difference between paired undepleted and depleted samples.