**Introduction**

Characterizing viral genomes, or viromes, is essential for public health officials and research regarding pandemic preparedness. The COVID-19 pandemic has highlighted significant shortcomings in our ability to respond to the emergence of highly contagious and deadly microbial threats. However, COVID-19 is not the first pandemic of this century caused by an emerging pathogen, and it is unlikely to be the last. In the past 20 years, several high-impact pathogens have either emerged or re-emerged, including three novel coronaviruses: Severe Acute Respiratory Syndrome (SARS) in 2003, Middle East Respiratory Syndrome (MERS) in 2012, and the current COVID-19 pandemic (SARS-CoV-2). Bioinformatics and NGS techniques can rapidly analyze high-throughput sequencing data from environmental, human, and animal samples to identify novel and known viruses for the early identification of pathogens. High-throughput sequencing has been pivotal in detecting novel plant pathogens and understanding their genetic makeup, which is instrumental in disease management and control measures​ [8, 15]. Furthermore, the integration of machine learning with bioinformatics has enhanced the ability to detect and classify viruses, even those with highly divergent sequences. Tools like VirHunter [19] and VirFinder [15] utilize deep learning to streamline the discovery of novel viruses, significantly reducing the need for manual curation and enabling more efficient analysis of large sequencing datasets. The bioinformatics pipeline described in this paper presents a straightforward yet powerful method for extracting meaningful ecological and taxonomic insights. It is tailored for applications in bioinformatics where understanding viral diversity and abundance is crucial. This pipeline facilitates the analysis of large sequencing datasets, contributing significantly to pandemic preparedness efforts.

**Strategy and Algorithmic Design**

The overall strategy is to deliver a straightforward, efficient, and meaningful analysis of high-throughput sequencing data from metagenomic samples, with a specific focus on viral communities. The script initially reads sequencing data from FASTQ files and reference sequences from FASTA files, which are standard formats in bioinformatics for storing sequencing reads and reference sequences, respectively. By leveraging these formats, the script ensures compatibility with most sequencing datasets. To maintain uniformity and avoid case-sensitivity issues during subsequent analysis, all sequences are converted to uppercase.

Sequences from the FASTQ file are filtered based on quality scores to ensure that only high-quality data is used for further analysis. Filtering sequences based on quality scores is crucial to eliminate low-quality reads that could introduce errors in downstream analyses. By setting a quality threshold, the script ensures that only high-confidence data is used, improving the reliability of the results [1, 2]. For instance, thresholds such as a minimum average quality score of 20 are commonly used in various studies and tools [12]. Quality scores are converted from ASCII to numeric values. Once converted, the average quality score for each sequence is calculated and then compared against a predefined threshold. If the quality scores are below this predefined threshold as defined in the command-line prompt, they are discarded and only those above the threshold are kept for further analysis. These filtered sequences are then used as input for k-mer creation.

K-mers are substrings of length k derived from the input sequences and is a fundamental step in various bioinformatics algorithms, including sequence alignment, error correction, and genome assembly [4]. The choice of k-mer size is crucial: larger k-mers provide higher specificity but can be computationally intensive, while smaller k-mers are computationally simpler but may lead to ambiguous alignments [4, 10]. After the sequences have been filtered for quality, each sequence is broken down into overlapping k-mers of a specified length. For example, for a sequence "GATTACA" and k=3, the resulting k-mers would be "GAT", "ATT", "TTA", "TAC", and "ACA". This method ensures that all potential subsequences are considered for alignment and further analysis. The creation and analysis of k-mers allow for the efficient identification of sequence similarities and differences, which is essential for the alignment and taxonomic classification steps. This approach leverages the efficiency of k-mer indexing and the specificity provided by carefully chosen k-mer lengths. In viral metagenomics, where the goal is often to distinguish closely related viral sequences, a k-mer size of 5 to 7 is commonly used. This size provides a good balance, offering enough specificity to distinguish between different sequences while maintaining manageable computational requirements [10]. These k-mers are then aligned to reference genomes from the FASTA file using the Burrows-Wheeler Transform (BWT) algorithm, which is known for its efficiency in sequence alignment.

The BWT algorithm transforms a sequence into a permutation that is easier to compress and search by reordering characters based on lexicographic rotations of the sequence [5]. This transformed sequence is then used to build an index coupled with a suffix array to allow for efficient searching of substrings within the sequence. By using the index, the script can efficiently find occurrences of k-mers from the query sequences within the reference genome sequences. By reducing the search space and leveraging the sorted order of the transformed sequence, BWT allow for rapid alignment. This transformation significantly enhances the efficiency of aligning short reads to a reference genome, a critical step in high-throughput sequencing analysis [5]. By enabling rapid and accurate alignment, the BWT algorithm helps to identify and characterize viral sequences in metagenomic samples, facilitating a deeper understanding of the underlying viral diversity and abundance [9]. Due to its efficiency, scalability, and straightforward implementation, BWT was chosen as the best algorithm for alignment due to time constraints. The aligned sequences were then used to extract taxonomic information from the sequence headers in the FASTA file.

Taxonomic classification is a crucial step in metagenomic analysis, enabling the identification and categorization of organisms present in a sample [6, 20]. A regular expression (regex)-based method was used to parse headers and extract relevant taxonomic information. For instance, headers within the FASTA file used in this analysis included the virus name "Severe acute respiratory syndrome coronavirus 2," followed by the strain/isolate "SARS-CoV-2/human/USA/CA-LACPHL-AY06582/2020" and the accession number "2697049.9243343." The extracted information is mapped to the aligned sequences to allow for classification of each sequence according to its parsed header information. The mapping was implemented using dictionaries that allow efficient look-up and association. For instance, a dictionary was used where the keys are sequence identifiers, and the values contain the taxonomic information.

To understand the ecological dynamics of viral communities within the metagenomic samples, various diversity metrics were calculated. These key metrics include abundance, Shannon diversity index, Simpson diversity index, richness, and evenness. Abundance indicates how common or rare each taxon is within the sample. The Shannon diversity index measures the overall diversity of the viral community, considering both the number of different taxa and the evenness of their distribution [17]. The Simpson diversity index reflects the probability that two randomly selected sequences will belong to different taxa [18]. Richness represents the total number of distinct taxa present in the sample, while evenness assesses how evenly the sequences are distributed among the different taxa [14]. These metrics provide a comprehensive view of the viral community structure, helping researchers understand the ecological relationships and dynamics within the sample [11].

Lastly, the script outputs a .csv and .txt file that captures the key metrics calculated as well as information about the viral sequences identified in the metagenomic sample included the virus name, strain/isolate and the ascension number that was extracted during the taxonomic classification step.

**Analysis of Results**

The proportion of each virus strain ranges from 1.608321-1.613056, indicating its relative abundances within the sample. The overall Shannon diversity Index metric was calculated to be 4.127134 with a Simpson Diversity Index of 0.983871, indicating a high diversity within the viral community and is very diverse, with a low probability that two randomly selected sequences will belong to the same taxon. Overall, there is a richness of 62, or 62 distinct viral taxa identified in the sample and the Evenness is a perfect score of 1.0, indicating the sequences re evenly distributed among the different taxa, with no one taxon dominating the sample. Understanding the diversity and distribution of viral strains can help in monitoring and controlling outbreaks. High diversity and even distribution might indicate widespread transmission and the potential for emerging new variants.

**Limitations and Future Directions**

There are still limitations regarding this script as the results are based on a specific metagenomic sample, and the findings might not be generalizable to all environments or populations. While the script filters sequences based on quality scores, low-quality data might still impact the results. Additionally, the choice of quality threshold can influence which sequences are included or excluded [2]. Sequencing errors can also introduce noise into the data, potentially inflating diversity metrics or creating false positives in taxonomic classification and can arise from various sources, including PCR amplification biases, sequencing platform errors, and sample contamination [7, 12]. Another limitation to consider is the accuracy of taxonomic classification since it relies on the reference database that was used. Incomplete or biased reference databases can lead to misclassification or underestimation of viral diversity [6]. To address these limitations, in the future it would be important to conduct similar analyses on a broader range of samples from different environments or populations to validate the findings and provide a more comprehensive picture of viral diversity. Improving the algorithm used for sequencing alignment may also significantly enhance the accuracy of identifying viral sequences. Techniques such as multiple-sequence alignment, which considers evolutionary relationships between sequences, can provide more accurate alignments [13]. There are also hybrid approaches such as Kraken2 or MetaPhlAn, that combine different alignment algorithms and classification methods, such as using both k-mer based and alignment-based approaches, can provide complementary strengths and improve overall accuracy [21]. It would also be interesting to explore integrating machine learning techniques into this metagenomic analysis to improve the accuracy and efficiency of taxonomic classification. Machine learning models, trained on large datasets, can learn complex patterns and make more accurate predictions about the taxonomic identity of sequences [3].

**References**

1. Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
2. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114-2120. doi:10.1093/bioinformatics/btu170
3. Boulesteix, A. L. (2010). Overview of methods for pattern recognition and machine learning in bioinformatics. In Bioinformatics Methods and Protocols (pp. 3-42). Humana Press. doi:10.1007/978-1-60327-241-4\_1
4. Compeau, P. E. C., Pevzner, P. A., & Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. Nature Biotechnology, 29(11), 987-991. doi:10.1038/nbt.2023
5. Ferragina, P., & Manzini, G. (2000). Opportunistic data structures with applications. In Proceedings of the 41st Annual Symposium on Foundations of Computer Science (pp. 390-398). IEEE. doi:10.1109/SFCS.2000.892127
6. Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. Genome Research, 17(3), 377-386. doi:10.1101/gr.5969107
7. Kunin, V., Engelbrektson, A., Ochman, H., & Hugenholtz, P. (2010). Wrinkles in the rare biosphere: Pyrosequencing errors can lead to artificial inflation of diversity estimates. Environmental Microbiology, 12(1), 118-123. doi:10.1111/j.1462-2920.2009.02051.x
8. Kutnjak D, Tamisier L, Adams I, Boonham N, Candresse T, Chiumenti M, De Jonghe K, Kreuze JF, Lefebvre M, Silva G, et al. A Primer on the Analysis of High-Throughput Sequencing Data for Detection of Plant Viruses. Microorganisms. 2021; 9(4):841. <https://doi.org/10.3390/microorganisms9040841>
9. Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25(14), 1754-1760. doi:10.1093/bioinformatics/btp324
10. Li, H., Yu, C., & Ye, C. (2016). Methods and algorithms for de novo assembly of short sequence reads. In Computational and Statistical Methods for Gene Interaction and Protein Interaction Networks (pp. 1-32). Springer, New York, NY. doi:10.1007/978-1-4939-3575-3\_1
11. Magurran, A. E. (2004). Measuring Biological Diversity. Blackwell Publishing.
12. Minoche, A. E., Dohm, J. C., & Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. Genome Biology, 12(11), R112. doi:10.1186/gb-2011-12-11-r112
13. Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of Molecular Biology, 302(1), 205-217. doi:10.1006/jmbi.2000.4042
14. Pielou, E.C. (1966) The Measurement of Diversity in Different Types of Biological Collections. Journal of Theoretical Biology, 13, 131-144.
15. Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A., & Sun, F. (2020). VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome, 8(1), 1-15. doi:10.1186/s40168-020-00845-3
16. Satam H, Joshi K, Mangrolia U, Waghoo S, Zaidi G, Rawool S, Thakare RP, Banday S, Mishra AK, Das G, et al. Next-Generation Sequencing Technology: Current Trends and Advancements. Biology. 2023; 12(7):997. <https://doi.org/10.3390/biology12070997>
17. Shannon, C. E. (1948) [A mathematical theory of communication](https://en.wikipedia.org/wiki/A_mathematical_theory_of_communication). The Bell System Technical Journal, 27, 379–423 and 623–656.
18. Simpson, E.H. (1949) Measurement of Diversity. Nature, 163, 688.  
    http://dx.doi.org/10.1038/163688a0
19. Sukhorukov G, Khalili M, Gascuel O, Candresse T, Marais-Colombel A, Nikolski M. VirHunter: A Deep Learning-Based Method for Detection of Novel RNA Viruses in Plant Sequencing Data. Front Bioinform. 2022 May 13;2:867111. doi: 10.3389/fbinf.2022.867111. PMID: 36304258; PMCID: PMC9580956.
20. Tang, P., & Chiu, C. (2010). Metagenomics for the discovery of novel human viruses. Future Microbiology, 5(2), 177-189. doi:10.2217/fmb.09.120
21. Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology, 15(3), R46. doi:10.1186/gb-2014-15-3-r46