**Introduction**

The rapid and accurate identification of microbial communities is a crucial tool in many different fields. In medicine, there is a need for metagenomic identification pipelines to detect novel viruses that could be transmitted via blood transfusion (Sauvage & Eloit 2016). New clinical diagnosis tools also utilize metagenomics to help detect diseases (Govender et al. 2021). In public health, COVID-19 tests have been performed using metagenomic identification workflows (Carbo et al. 2020). Food safety has begun to rely on these techniques too, utilizing them to ensure the quality of various products (Billington et al. 2022). Regardless of the context, it is clear that timely and precise metagenomic identification pipelines would be an advantage to these varied applications. While such workflows may consist of many parts, like quality control, assemblers, classifiers, and data visualization, it is the classifiers that perform taxonomic classification and determine which species are present in a particular sample.

           Initially, most metagenomic classification was accomplished with markers (Ye et al. 2019). Marker-based analysis, sometimes referred to as 16S or DNA barcoding, relies on sequencing and comparing highly conserved DNA barcode genes to identify the species present in a sample (Jovel et al. 2016, Badotti et al. 2017, Edgar 2018). For prokaryotes, the most commonly used gene is 16S, but other types of organisms like fungi or eukaryotes use ITS (Ye et al. 2019, Badotti et al. 2017). While marker-based techniques can be useful to quickly determine what species are present in a sample, their abundance estimates can be biased if the organisms present possess differing concentrations of the barcode genes (Edgar 2018, Wang et al. 2015). This makes it more challenging to accurately assess the exact composition of a large, diverse community. However, promising full-genome approaches have been developed that help eliminate this bias, making them the new go-to tool for metagenomic classification.

           Over the past few years, whole-metagenomic shotgun techniques have become more dominant (Ye et al. 2019). Shotgun analysis relies on the full sequencing of everything present in a sample, then comparing it to a vast genome or protein library to identify which species are present (Jovel et al. 2016, Wang et al. 2015). Most shotgun classification is strictly DNA-based, which means that identification is completed by matching sample sequence nucleotides against genome libraries, similar to NCBI’s BLASTn tool (Ye et al. 2019, Jovel et al. 2016). However, there are also protein-based tools, which instead rely on extensive protein libraries and translate sample sequences into amino acids for comparison, similar to NCBI’s BLASTx tool (Ye et al. 2019). Although shown to be more sensitive than the typical DNA-based approaches, protein-based classification has been found to be more computationally intensive, as it must consider all six potential reading frames of a given sequence (Ye et al. 2019). Such techniques are best used in specific circumstances where high precision is a necessity, such as to distinguish the difference between variants of the same sequence.

           DNA-based shotgun classification can be performed on either short-read or long-read sequence data (Pearman et al. 2020). The more common method is short-read data, as long-read data has been historically associated with higher error rates (Portik et al. 2022). However, with recent technological developments like the advent of “third generation” sequencing, the length and quality of long-read data has substantially improved, making it a more promising alternative (Pearman et al. 2020, Portik et al. 2022). Although long-read data is still prone to more errors than short-read data, the advantage of longer sequences for classification is evidenced by their superior accuracy, matching that of protein-based shotgun techniques (Ye et al. 2019, Pearman et al. 2020, Portik et al. 2022). In general, long-read classification software takes more time to run and is more resource intensive than short-read, but this difference seems to be shrinking (Ye et al. 2019). Most classification software is tailored specifically for either short-read or long-read data, so it is important to note what a particular program prefers.

           Raw sequence data is often fed directly into a classification program, but this is not a required procedure. To help aid with identification, short-read sequences can first be assembled into longer contigs by matching up their overlapping regions (Huang 1992). Recent research has shown that feeding these longer, assembled contigs into classifiers significantly improves identification accuracy, making it a viable solution to enhance short-read data when long-read techniques cannot be utilized (Tran & Phan 2020). This process of contig classification enables short-read data to potentially reach the accuracy of long-read data without relying on more expensive sequencing techniques (Tran & Phan 2020). It should therefore be possible to combine assemblers and classifiers into a metagenomic identification pipeline for short-read data that can complete with long-read techniques.

The central objective of this project is to explore DNA-based shotgun classification techniques and determine which tools are the most accurate, to facilitate the creation of a commercially-scalable end-to-end metagenomic identification pipeline for CIAN Diagnostics. I hypothesize that long-read classification will prove the most accurate, as they are at the forefront of recent developments. Contig classification will probably be a close second, as this method has been shown to enhance the performance of standard short-read classification. I anticipate that short-read classification will have the worst performance, on account of the other techniques having been developed to be superior. It serves almost as a baseline. Different software will be utilized for contig and short-read classification to enable a broader comparison. Performance metrics such as precision, accuracy, F1 score, F0.5 score, and abundance estimates will be analyzed to enable comparison. Higher performance metric scores shall indicate whether a certain method or tool is superior to another. My hypothesis will be evaluated based on a comparison of these performance metrics.

**Specific Aims**

The first specific aim of this proposal is to determine whether a raw long-read, contig, or raw short-read classification approach is most accurate. This will be accomplished with assembled contigs from a colleague and raw read data provided by CIAN Diagnostics. Contigs are expected to be in a FASTA format and raw reads in FASTQ. Data will be run through four selected classifiers, all of which accept either format and produce a CSV of identification results (Wood 2020, Kim 2021, Marcelino 2022, BugSeq 2023). Because the true composition of each dataset is known, accuracy can be gauged by comparing performance metrics like precision, accuracy, and abundance. F1 and F0.5 scores, comparative metrics calculated from the aforementioned precision and accuracy, may also be used. Although most classifier review papers rely on raw reads, past research has shown that assemblers are an effective way to improve classifier performance (Tran & Phan 2020). I expect raw long-reads to be superior, if only marginally, as this is the direction in which most recent research has been trending. Raw short-reads, although anticipated to perform the worst, are included as a baseline against which to judge the other approaches. To keep things simple, only DNA-based classifiers have been considered for this study. It should be noted, however, that protein-based classifiers are a promising alternative (Ye et al. 2019). MMseqs2, a protein-based classifier, has proved to be quite accurate, though it does have comparatively demanding requirements (Mirdita et al. 2021). In the event that something goes wrong with the DNA-based approach or one wishes to explore further, this would be a valuable avenue to investigate.

The second specific aim is to determine which selected classifier tools perform the best: Kraken 2, CCMetagen, Centrifuge, or BugSeq. This is also to be accomplished via a comparison of the performance metrics, such as F1 score and abundance. Raw short-reads will be run through Kraken 2, CC Metagen, and Centrifuge, while raw long-reads go through BugSeq. Contigs can be tested with all four classifiers. Since Kraken 2 and Centrifuge are widely utilized as benchmarks, Kraken 2 for being ultra-fast and Centrifuge for being able to run on a desktop computer, they seemed like obvious choices to include (Wood & Salzberg 2014, Kim et al. 2016). CCMetagen is a more recent tool that aims to enhance the identification of eukaryotic and fungal species without significantly compromising the identification of prokaryotic ones (Marcelino et al. 2020). Although it claims to be superior to Kraken 2 and Centrifuge, I will determine whether that holds true within the context of this study (Marcelino et al. 2020). BugSeq is a new cloud-based tool for long-read classification that is remarkably simple to use and excels at species-level classification with highly-accurate data (Fan et al. 2021, Portik et al. 2022). I anticipate that BugSeq will perform the best, followed by CCMetagen, Kraken 2, and Centrifuge. Other classifiers are available for both short-read and long-read classification, such as CLARK, ProPhyle, and MEGAN-LR (Ounit et al. 2015, Brinda et al. 2017, Huson et al. 2018). In the event that the selected tools do not work as hoped, these may be viable substitutes.

The third specific aim is to determine how results are affected with or without a quality control step. FastQC is the quality control tool of choice. It takes a FASTQ input and produces a ZIP with the relevant analysis data (Andrews 2019). This is a relatively popular tool that I have some limited experience with, so it seemed like a good choice for this project (Sprang et al. 2021). While it stands to reason that quality-controlled data will perform better across the board, there is still value in assessing the discrepancy. Some tools may do particularly poorly with bad data, while others less so. Noting such details may assist researchers in choosing the programs that will work best with the data they have. I anticipate that the quality-controlled data will perform better by a large margin, on account of how crucial this step is always hailed. Another popular quality control tool is FastP (Sprang et al. 2021). In the event that something goes wrong with FastQC, it would be a good alternative.

As a whole, I anticipate that the long-read classification with BugSeq will have the best accuracy, followed by contig classification with CCMetagen. The F1 scores and abundance ratios should reflect this or the hypothesis will be rejected. There are numerous alternative programs that can be used in place of the ones I have selected. CLARK and ProPhyle are potential alternatives for Kraken 2 and Centrifuge respectively (Ye et al. 2019). CLARK was designed to compete with Kraken, and ProPhyle can run on a desktop similar to Centrifuge (Ounit et al. 2015, Brinda et al. 2017). MEGAN-LR could work in place of BugSeq, as it is also designed specifically for long-read classification (Huson et al. 2018). A protein-based approach could also be adopted too, ideally with MMseqs2 (Mirdita et al. 2021). For quality control alternatives, FastP is another popular tool that could be utilized (Sprang et al. 2021). I anticipate that the project will work as intended, but it is reassuring to have viable alternative tools at our disposal.

**Citations**

Andrews, S. (2019). *FastQC Manual*.<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

Badotti, F., de Oliveira, F. S., Garcia, C. F., Vaz, A. B. M., Fonseca, P. L. C., Nahum, L. A., Oliveira, G., & Góes-Neto, A. (2017). Effectiveness of ITS and sub-regions as DNA barcode markers for the identification of Basidiomycota (Fungi). *BMC Microbiology*, *17*(1), 42.<https://doi.org/10.1186/s12866-017-0958-x>

Billington, C., Kingsbury, J. M., & Rivas, L. (2022). Metagenomics Approaches for Improving Food Safety: A Review. *Journal of Food Protection*, *85*(3), 448–464.<https://doi.org/10.4315/JFP-21-301>

Brinda, K., Salikhov, K., Pignotti, S., & Kucherov, G. (2017). *ProPhyle: A phylogeny-based metagenomic classifier using Burrows-Wheeler Transform*.

<https://zenodo.org/record/1045427#.ZASm2HZKhEY>

BugSeq, B. Inc. (2023, January 11). *BugSeq Manual*.<https://docs.bugseq.com/input/>

Carbo, E. C., Sidorov, I. A., Zevenhoven-Dobbe, J. C., Snijder, E. J., Claas, E. C., Laros, J. F. J., Kroes, A. C. M., & de Vries, J. J. C. (2020). Coronavirus discovery by metagenomic sequencing: A tool for pandemic preparedness. *Journal of Clinical Virology*, *131*, 104594.<https://doi.org/10.1016/j.jcv.2020.104594>

Edgar, R. C. (2018). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics*, *34*(14), 2371–2375.<https://doi.org/10.1093/bioinformatics/bty113>

Fan, J., Huang, S., & Chorlton, S. D. (2021). BugSeq: A highly accurate cloud platform for long-read metagenomic analyses. *BMC Bioinformatics*, *22*(1), 160.<https://doi.org/10.1186/s12859-021-04089-5>

Govender, K. N., Street, T. L., Sanderson, N. D., & Eyre, D. W. (2021). Metagenomic Sequencing as a Pathogen-Agnostic Clinical Diagnostic Tool for Infectious Diseases: A Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies. *Journal of Clinical Microbiology*, *59*(9), e02916-20.<https://doi.org/10.1128/JCM.02916-20>

Huang, X. (1992). A contig assembly program based on sensitive detection of fragment overlaps. *Genomics, 14*(1), 18–25. <https://doi.org/10.1016/S0888-7543(05)80277-0>

Huson, D. H., Albrecht, B., Bağcı, C., Bessarab, I., Górska, A., Jolic, D., & Williams, R. B. H. (2018). MEGAN-LR: New algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. *Biology Direct*, *13*(1), 6.<https://doi.org/10.1186/s13062-018-0208-7>

Jovel, J., Patterson, J., Wang, W., Hotte, N., O’Keefe, S., Mitchel, T., Perry, T., Kao, D., Mason, A. L., Madsen, K. L., & Wong, G. K.-S. (2016). Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. *Frontiers in Microbiology, 7*. <https://doi.org/10.3389/fmicb.2016.00459>

Kim, D. (2021). *Centrifuge Manual*.<https://ccb.jhu.edu/software/centrifuge/manual.shtml>

Kim, D., Song, L., Breitwieser, F. P., & Salzberg, S. L. (2016). Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Research*, *26*(12), 1721–1729.<https://doi.org/10.1101/gr.210641.116>

Marcelino, V. R. (2023). *CCMetagen Manual* [HTML].<https://github.com/vrmarcelino/CCMetagen> (Original work published 2018)

Marcelino, V. R., Clausen, P. T. L. C., Buchmann, J. P., Wille, M., Iredell, J. R., Meyer, W., Lund, O., Sorrell, T. C., & Holmes, E. C. (2020). CCMetagen: Comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data. *Genome Biology*, *21*(1), 103.<https://doi.org/10.1186/s13059-020-02014-2>

Mirdita, M., Steinegger, M., Breitwieser, F., Söding, J., & Levy Karin, E. (2021). MMseqs2: Fast and sensitive taxonomic assignment to metagenomic contigs. *Bioinformatics*, *37*(18), 3029–3031.<https://doi.org/10.1093/bioinformatics/btab184>

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), 236.<https://doi.org/10.1186/s12864-015-1419-2>

Pearman, W. S., Freed, N. E., & Silander, O. K. (2020). Testing the advantages and disadvantages of short- and long- read eukaryotic metagenomics using simulated reads. *BMC Bioinformatics, 21*(1), 220. <https://doi.org/10.1186/s12859-020-3528-4>

Portik, D. M., Brown, C. T., & Pierce-Ward, N. T. (2022). Evaluation of taxonomic classification and profiling methods for long-read shotgun metagenomic sequencing datasets. *BMC Bioinformatics*, *23*(1), 541.<https://doi.org/10.1186/s12859-022-05103-0>

Sauvage, V., & Eloit, M. (2016). Viral metagenomics and blood safety. *Transfusion Clinique et Biologique*, *23*(1), 28–38.<https://doi.org/10.1016/j.tracli.2015.12.002>

Sprang, M., Krüger, M., Andrade-Navarro, M. A., & Fontaine, J.-F. (2021). Statistical guidelines for quality control of next-generation sequencing techniques. *Life Science Alliance*, *4*(11), e202101113.<https://doi.org/10.26508/lsa.202101113>

Tran, Q., & Phan, V. (2020). Assembling Reads Improves Taxonomic Classification of Species. *Genes*, *11*(8), 946.<https://doi.org/10.3390/genes11080946>

Wang, W.-L., Xu, S.-Y., Ren, Z.-G., Tao, L., Jiang, J.-W., & Zheng, S.-S. (2015). Application of metagenomics in the human gut microbiome. *World Journal of Gastroenterology, 21*(3), 803–814. <https://doi.org/10.3748/wjg.v21.i3.803>

Wood, D. E. (2020). *Kraken 2 Manual*. GitHub.<https://github.com/DerrickWood/kraken2/wiki/Manual>

Wood, D. E., & Salzberg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3), R46.<https://doi.org/10.1186/gb-2014-15-3-r46>

Ye, S. H., 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>