PHYLOGENETIC CLASSIFICATION OF LONG READ SEQUENCES

by

KEVIN CHIN-WEI CHAN

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

BACHELOR OF SCIENCE

in

THE FACULTY OF SCIENCE

(Honours Computer Science, Microbiology and Immunology)

We accept this thesis as conforming to the required standard

……	……
Dr. Steven Hallam	Dr. Martin Hirst
Thesis Supervisor	Thesis Nominee

……
Mr. Connor Morgan-Lang
Project Supervisor

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

APRIL 2019

Abstract TreeSAPP (Tree-based Sensitive and Accurate Protein Profiler) is an analysis pipeline designed to functionally and taxonomically classify protein and nucleotide sequences using marker genes and phylogenetic methods. Currently, TreeSAPP supports short read sequencing data (e.g. Illumina), but does not support long reads from newer sequencing platforms (e.g. Nanopore). Therefore, ten isolate datasets sequenced using Oxford Nanopore Technologies were aligned to reference sequences of five single-copy phylogenetic marker genes. Of the four aligners tested (minimap2, GraphMap, LAST and SNAP), minimap2 performed the best when judged by raw and weighted averages of taxonomic distance of alignments to their optimal placements, which is crucial for phylogenetic inference. Minimap2 was subsequently integrated into the long read workflow of TreeSAPP, and was tested on the same datasets and a mock community. While the workflow performed well with isolate datasets, poor recall was demonstrated with the mock community, suggesting required improvements in TreeSAPP's linear model for taxonomic inference, or for higher resolution nucleotide reference packages.

Importance Short read sequencing information pose several challenges for downstream bioinformatic analyses, such as sequencing error, non-uniform coverage of samples, computational time complexity and resolving repetitive regions. With the advent of cost-effective long read sequencing technologies, many of these problems are alleviated through contiguous sequences encoding full length open reading frames. Despite this benefit, relative to short reads, long reads have high error and insertion/deletion rates, with the potential to limit their utility in marker gene classification. To resolve this dilemma, TreeSAPP requires a separate workflow for long read sequences.

Table of Contents

1. Introduction

1.1. Sequencing Nucleic Acids

Microbes are the invisible majority of living things on Earth and interconnected microbial communities drive nutrient and energy flow through networks of metabolite exchange that are critical for ecosystem functions and services in natural and engineered environments, including our own bodies. Despite their numerical abundance and critical role in the world, our understanding of these networks has been limited by the inability to cultivate most microorganisms in laboratory settings. Over the past decade, high-throughput sequencing has begun to overcome this cultivation gap by providing researchers with vast amounts of nucleic acid (DNA and RNA) sequence information sourced directly from environmental samples that can be used to understand microbial interaction networks based on gene prediction and pathway reconstruction. When initially describing microbial communities, researchers are typically interested in which microorganisms are found within a given environment from a taxonomic perspective and what are they are capable of doing from a metabolic or functional perspective (1).

A common approach to identifying the taxonomic composition of microbial communities is through amplicon sequencing of the small subunit ribosomal RNA (SSU or 16S rRNA) gene (2), as this conserved essential marker gene can be used as a fingerprint to infer phylogenetic relationships between organisms (3–5). However, this approach lacks the ability to discover and annotate functions from the community without resorting to transitive or inferential methods that try and predict traits based on cultured representatives sharing the same 16S rRNA gene.

Therefore, an increasingly popular approach to capture metabolic potential is to conduct whole genome shotgun sequencing at the individual (single-cell amplified genomes), population or community (metagenome assembled genomes) levels of organization (6). Using this type of sequence information we are better able to link microbial agents to specific metabolic processes and reconstruct taxonomic profiles using single copy phylogenetic marker genes to place assembled groups of contigs on to reference trees (7).

Whole genome sequencing technologies have improved greatly since the advent of the Sanger sequencing method (8), from second generation sequencing (SGS) (9) to the more recent third generation sequencing (TGS) (10). A comparison of read lengths, sequencing templates, pricepoint and throughput for each of the technologies is described in **Table 1**. In SGS, several technologies emerged as popular platforms, but comparing the cost and throughput of each platform, Illumina sequencing has become the most successful (11), dominating the present market. In Illumina sequencing, both ends of a DNA fragment attach to a flow cell, and fluorescent, reversible dye-terminators are introduced and incorporated into a complementary strand. This process, known as bridge amplification, generates clonal copies of a fragment in a cluster, and imaging of the fluorescence is used to call the current base. Reads from Illumina sequencing are typically 150 bp in length and have low error rates (<2%) with extremely high throughput (12). In TGS, two platforms are currently used: Pacific Biosciences (PacBio), which uses Single Molecule Real Time sequencing (13), and Oxford Nanopore Technologies (ONT), which passes DNA molecules through graphene pores and measures changes in electric current to infer the nucleic acid sequence (14).

	Sanger reads	Illumina reads	Oxford Nanopore reads
1	>800 bp	2x150 bp	>6-8 kbp
2	Template using pool of molecules	Template clonally derived from single molecule	Template is single nucleic acid sequence
3	$0.00125 / bp	$6.67 x 10^-7 / bp	$4.75 x 10^-9 / bp
4	Highest quality, relatively low throughput	High quality, extremely high throughput	Poor quality, comparitively low throughput

Table 1. Comparison of reads generated from Sanger sequencing, Illumina sequencing (SGS), and Oxford Nanopore sequencing (TGS). The typical length of reads, templates for sequencing, cost per base pair, base pair quality and throughput are depicted.

A major challenge for both SGS and TGS is analysis of the generated data. Numerous bioinformatics pipelines exist to meet the needs of researchers, depending on the type of data and research goal (15). Thus, large quantities of short read data pose challenges in the typical steps of bioinformatics pipelines, such as read mapping, assembly, and annotation (16). Sequencing error, non-uniform coverage of samples, computational time complexity and resolving repetitive regions are some of the many challenges associated with short read data (17). With a shorter read length, any base call errors potentially have larger impact, and software may not scale appropriately with the high volume of data. Further, repetitive regions longer than the sequenced repeating unit typically cannot be resolved. TGS has partially alleviated some of these challenges. Whereas SGS requires extensive library preparation and amplification, TGS can utilize and sequence single molecules, which removes the need for amplification (10). TGS also generates longer reads, exceeding 6-8 kbp on average (18), which helps to resolve repeat regions more accurately, though these reads also come with a significantly higher error rate. In PacBio reads, the error rate ranges from 13% to 15% (19), whereas in ONT reads, this range is from 5% to 15% (20). With the rapid development of TGS, longer read lengths and reduced error rates

will be achieved, making TGS a better candidate over SGS for future studies. Thus, we should focus on adapting existing analysis software to TGS platforms.

1.2 TreeSAPP

Phylogenetic marker genes are conserved nucleic acid sequences encoding core cellular functions that are widely conserved and seldomly transferred between lineages that be used to classify microorganisms at different taxonomic ranks (Domain, Phylum, Class, Order, Family, Genus, Species, Strain) (22). The Tree-based Sensitive and Accurate Protein Profiler (TreeSAPP) (21) is a pipeline which utilizes functional and phylogenetic marker genes to classify and place sequenced (meta)genomic data onto cognate phylogenetic reference trees. TreeSAPP is able to work with both protein and nucleic acid sequence information. Currently, TreeSAPP is optimized for processing contigs generated from assemblies of SGS data. First, open reading frames (ORFs) are extracted using Prodigal (23), followed by identification of homologs and functional marker genes via hidden markov models (HMMs) with hmmsearch of the HMMER suite (24). These homologs are then aligned against reference sequences using hmmalign (24) to form a multiple sequence alignment, which allows for classification of sequences in a reference tree using RaxML (25). A visualization of the TreeSAPP workflow is depicted in **Figure 1**.

However, analyzing TGS reads with the current workflow in TreeSAPP comes with challenges. Firstly, ORF predictions using Prodigal can be inaccurate due to high and non-uniform error distributions error rates in TGS data (26). Indeed, the primary error modes in TGS data are insertions and deletions, which can lead to frameshift or premature stop codons. Secondly, non-uniform error distributions in combination and longer read lengths characteristic

of TGS sequences make it harder to align to reference sequences (26). But considering the advantages and growing popularity of TGS, TreeSAPP must support long reads to extend its versatility and longevity, with additional support for assemblies using TGS data. Further, ultra-long reads may also eliminate the need for assembly (10), thus TreeSAPP requires an option to analyze TGS reads directly. Several aligners have been designed to account for this by implementing novel algorithms tuned to the characteristics of TGS reads, and may thus be more suitable for these stages of the pipeline.

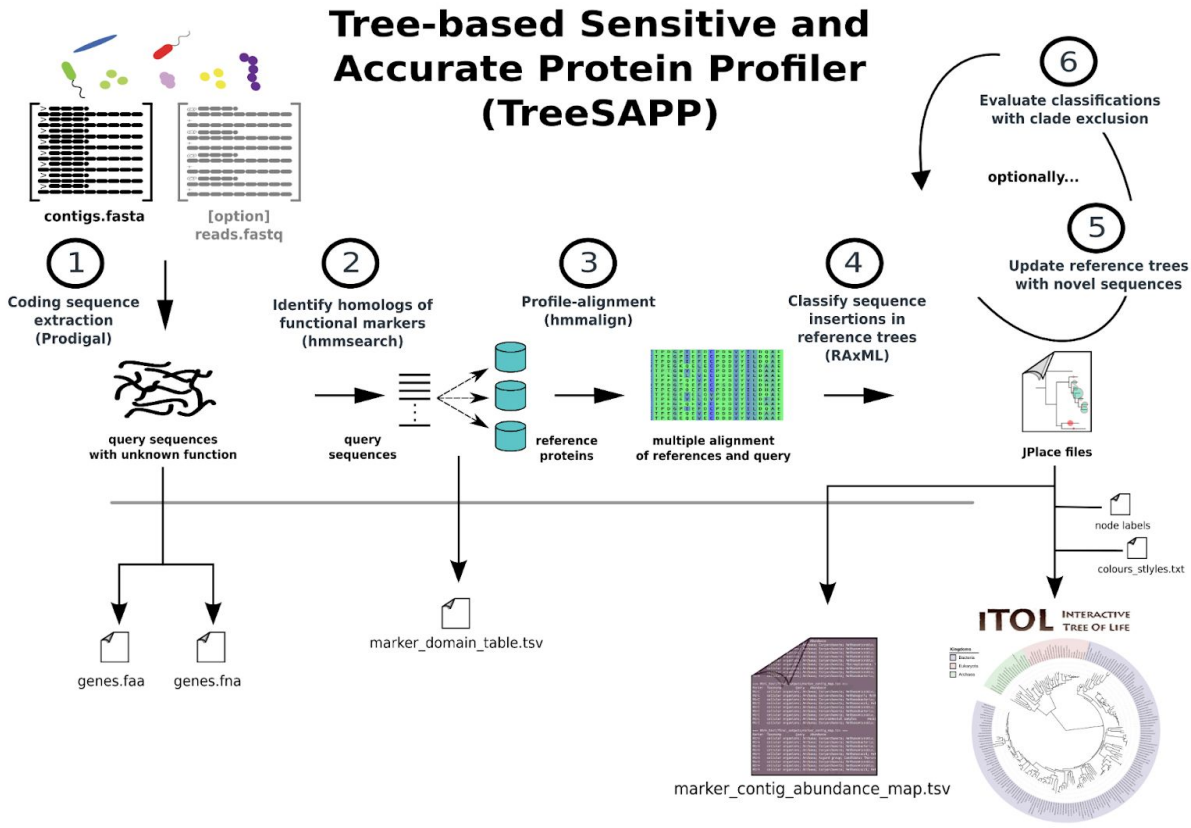

Figure 1. Illustration of the general TreeSAPP workflow, obtained from (21).

1.3 Alignment Tools

Generally, homologs are identified by aligning query sequences to a reference set of sequences and attempting to find the optimal locations/sequences in the reference set for each query, based on the aligner's heuristics or scoring schemes (27). The reference set is typically larger than the query sequences, with typical use cases including genomes, and contigs or scaffolds from assemblies (27).

Minimap2. The minimap2 algorithm (28) is the latest aligner developed by the authors of popular tools for SGS data, including BWA (29) and Samtools (30). Minimap2 can align DNA or mRNA reads, and has support for both TGS and SGS data. While minimap2 has been shown to outperform other popular aligners using the human genome as a reference (28), it has not been tested against marker gene references. Given this, and together with the growing use of minimap2 by the research community, minimap2 proved to be a strong candidate to include in this comparison. In minimap2, minimizers, which are representative k-mers from a group of adjacent k-mers (31), are used and indexed into a hash table, which is a data structure allowing for fast lookup and retrieval. Seeds, the starting points for alignments, are extracted from minimizers by taking those appearing in the reference sequence(s) below a frequency threshold. Seeds are then chained together by dynamic programming and serve as mapping positions for the query sequences, with the alignment extending from the leftmost seed.

GraphMap. GraphMap (32) is a mapper designed for error prone TGS data, and supports SGS data. Since GraphMap uses a novel vertex-centric algorithm, its differences between previous algorithmic approaches (27) make it a suitable candidate for comparison. GraphMap follows a 5-stage procedure. First, regions in the reference where a seed can align to

are selected as suitable positions. Next, a minimizer index is built, using gapped qgrams for Levenshtein distance (32) for seed construction, as this is more suitable for the frequent indels characteristic of TGS (32). GraphMap then constructs a directed acyclic graph (DAG) with k-mers from the query sequence as vertices, and edges form links between k-mers. The DAG is then traversed and the longest existing path is extended in a greedy fashion to form an "alignment anchor". With error prone reads, a single query is typically covered by several anchors. Thus, a "longest common subsequence in k-length substrings" is applied on all anchors to extend each one, and further filtering is applied to refine alignments.

Local Alignment Search Tool (LAST). LAST (33) builds upon the hash-based seed and extend algorithm introduced in the popular aligner BLAST (34). However, the BLAST algorithm uses fixed length seeds, whereas LAST uses adaptive seeds, which can shorten or lengthen until the number of matches in the target sequence is less than or equal to a specified frequency threshold (33). Adaptive seeds have been shown to perform better against databases (or references) where the nucleic acid composition is non-uniform, as is the case in the majority of references (33). More recently, the authors of LAST have developed a training module, LAST-TRAIN (35), designed to infer the optimal parameters using LAST for a given query and reference by iteratively adjusting score parameters and aligning the query to the reference until the parameters converge. Because of the recent improvements in LAST, the proven effectiveness of the seed and extend algorithms with scoring schemes, and its popularity, LAST was included in this comparison.

In this thesis, I evaluate 4 aligners: minimap2, GraphMap, SNAP (36) and LAST, to determine the optimal aligner for TreeSAPP to use aligning TGS reads to marker gene reference

sequences. I begin by aligning reads from each dataset against each constructed marker gene reference, and score alignments from each aligner by alignment contiguity and phylogenetic distance. Following this, I extend the analysis by integrating the optimal aligner into TreeSAPP, and evaluate the performance on a mock community. In determining the optimal aligner, I hope to incorporate it into TreeSAPP's existing workflow, and allow TreeSAPP to work with the latest generation of sequencing data, thus extending its utility and longevity.

2. Methods

2.1 Dataset retrieval

Ten datasets were retrieved from the European Nucleotide Archive to serve as inputs for each aligner. Criteria for each dataset was set by using the Query Builder tool, and the following parameters were set: "Domain" was set to "Read", and "Instrument model" was set to MinION, GridION, or PromethION to account for various ONT sequencers. However, only MinION datasets matching the previous criteria were found. Datasets were further filtered by selecting from six different taxonomic ranks at the class level for a breadth of taxonomic inputs. These classes included: Gammaproteobacteria, Alphaproteobacteria, Bacilli, Flavobacteria, Clostridia, and Actinobacteria. Lastly, the "Library Source" for each dataset was chosen to be genomic. A summary of metadata for each dataset is illustrated in **Table 2**.

2.2 Marker genes evaluated

Five nucleotide marker genes were used as references in the comparison, including: recombinase A (recA), the β subunit of bacterial RNA polymerase (rpoB), ribosomal protein S8 (rps8), ribosomal protein L10 (RPL10), and CTP synthase (pyrG). All marker gene sequences were gathered from the Functional Gene Pipeline and Repository (37). Sequences were filtered

to have a minimum HMM coverage of 60-90% of the total HMM length for each respective marker to be reliability placed in the reference tree. Each set of sequences were then clustered at 99% using USEARCH (38) with the following options: "-cluster_fast example_reference.fasta -id 0.99 -centroids". Therefore, each sequence corresponds to a prokaryotic species, thus each marker gene reference encapsulates a large portion of all sequenced prokaryotes.

	Dataset ID	Read Count	Taxonomy	Class	recA	rpoB	rps8	pyrG	RPL10
1	DRR129261	13966	Pseudomonas fulva	Gammaproteobacteria	Y	Y	N	Y	N
2	SRR7647306	588555	Pseudomonas aeruginosa	Gammaproteobacteria	Y	Y	N	Y	Y
3	SRR7690687	208566	Pseudomonas sp. Leaf58	Gammaproteobacteria	Y	Y	N	Y	N
4	SRR3191595	30412	Agrobacterium tumefaciens	Alphaproteobacteria	Y	Y	Y	Y	N
5	ERR2109177	6460	Enterococcus faecium	Bacilli	Y	N	Y	Y	Y
6	SRR7820386	440867	Bergeyella cardium	Flavobacteria	N	N	N	N	N
7	SRR5344355	22867	Anaerotignum lactatifermentans	Clostridia	N	N	N	N	N
8	ERR2724039	772381	Mycobacterium tuberculosis	Actinobacteria	Y	N	Y	Y	Y
9	SRR2671868	55663	Bacillus cereus ATCC 10987	Bacilli	Y	N	Y	Y	Y
10	SRR7743080	208000	Streptococcus pyogenes	Bacilli	Y	N	Y	Y	Y

Table 2. Summary of metadata for each of the 10 datasets analyzed. Each dataset was chosen based on the class-level classification of the reported taxonomy to get a breadth of input sequences for testing. All datasets were of isolates sequenced on MinION machines. A binary classification of whether the species were present in each set of marker gene reference sequences is denoted with "Y" (present) or "N" (absent). To evaluate TreeSAPP's accuracy of taxonomic estimates, the species in datasets SRR7820386 and SRR5344355 were explicitly chosen to be unrepresented in the marker gene references.

2.3 Aligner evaluation

During initial stages of testing, a variety of parameter combinations were tested for each aligner. However, SNAP and LAST were not considered in the comparison. Repeated attempts to run SNAP failed to produce any outputs, despite varying several parameters. Further parameter configuration would have required editing of SNAP's source code, which was beyond the scope of this thesis. Similarly, attempts to train LAST parameters on reference sequences never ran to completion due to errors. Nonetheless, with minimap2 and GraphMap, the different

combinations tested did not result in noticeable changes in outputs. Thus, minimap2 was ran with the preset option "-x map-ont", and GraphMap was ran with default parameters as per both respective author's recommendations when aligning ONT data. Version 2.12-r849-dirty of minimap2 and v0.5.2 GraphMap were tested.

By default, minimap2 and GraphMap output mappings are judged by internal scoring criteria. However, minimap2 outputs a primary mapping with five secondary mappings per sequence while GraphMap outputs a single mapping. Therefore, the alignment with the highest mapping quality was retained for each sequence, with unaligned sequences (corresponding to a mapping quality of 0) filtered out.

The resulting alignments for each aligner were then assessed based on contiguity and phylogenetic distance. Contiguity was measured by taking the length of the alignment divided by the length of the reference sequence aligned to. The average bacterial gene is approximately 330 amino acids (39), which translates to roughly 1 kbp in length. Given that the average length of a TGS read exceeds 6-8 kbp on average (18), it is possible in theory to cover the full length of a marker gene. Therefore, longer and more contiguous were scored more favorably.

Phylogenetic distance for each alignment was scored using the taxonomic distance between the reference sequence aligned to and the optimal placement for the dataset used. Although the organism sequenced for each dataset is known, it is not guaranteed to appear in any of the five tested marker gene references. Therefore, the full taxonomic lineage for the reported organism for each dataset, and for each sequence of each marker gene reference was determined. The optimal placement for each dataset and marker gene reference combination was then calculated by finding the lowest common ancestor (LCA) between the reported organism and

each reference sequence of a marker gene, and taking the LCA with the deepest lineage. The phylogenetic distance for each alignment was calculated by counting the number of taxonomic ranks which differ between the reference sequence aligned to, and the optimal placement up to the minimum length of the former and latter. Taxonomic lineages and LCA calculations were determined using the Joint Genome Institute's taxonomy server (40).

To obtain a single score for ranking each alignment, a cumulative taxonomic distance (CTD) was used with the following formula:

$$\sum_{i=0}^{k} P_i \times i$$

where i is the distance from the optimal taxonomic rank (determined by the lowest common ancestor of the optimal rank and the assigned taxon) and P is the proportion of query sequences aligning to the reference sequence assigned at that distance i. A CTD value of 0 indicates all sequences were assigned to their optimal taxonomy and larger values indicate sequences were incorrectly classified.

2.4 Mock community analysis

A commercially-available mock community sequenced with ONT GridION machines from the ZymoBIOMICS Microbial Community Standard was obtained (41), and the R10 dataset was analyzed. To evaluate the sensitivity and specificity of sequence classification by the long read workflow, Matthew's Correlation Coefficient (MCC) was calculated. However, annotated genomes for the organisms in the community were unavailable, but required for construction of the confusion table. To address this need, PacBio draft assemblies of isolates were obtained (42) and annotated with PROKKA (43).

The optimal placement for each organism was calculated for each marker gene reference based on methods described earlier and mock community reads were then mapped to the assemblies with minimap2. Reads which had overlapping coordinates to marker genes of interest annotated by PROKKA were kept as a set for comparison (denoted as *S*), and assigned a taxonomy at their respective optimal placements. If we let *T* denote the set of reads identified by TreeSAPP, then consider read *i*. A true positive classification was defined if read *i* was contained in *S* and *T*, aligned to the same marker gene, and was within a taxonomic distance of two from the optimal placement of *i*; subsequently, false negatives composed all reads in *S* but not in *T*, or if the taxonomic distance from the optimal placement was greater than two. False positives were defined as reads in *T*, but not in *S*, and true negatives were defined as reads not in both *S* and *T*.

3. Results

3.1 Aligner comparison

Figure 2 illustrates the difference between proportions of query sequences mapped at a distance from their optimal ranks when comparing the two mappers. If query sequences were mapped to a reference organism which had a mean alignment length of less than 20%, then these sequences were removed, as they were likely to be spurious alignments. When sequences were aligned using GraphMap, there were higher proportions of reads at a further distance from the optimal taxonomic rank for each marker gene, suggesting that GraphMap either was less accurate overall, or retained a larger number of poor quality alignments. Indeed, the majority of reads appear to concentrate between distances of 3 to 7. On the other hand, the large majority of query sequences had a distance of 0 from the optimal rank when mapping with minimap2,

implying that most query sequences aligned to the reference sequences most closely related to the report organism for each dataset. However, for both mappers some marker genes performed more poorly overall than others e.g. rps8. The majority of query sequences in GraphMap and roughly 12% of reads in minimap2 had taxonomic distances between 6 and 7, suggesting that this reference package did not have representative taxonomic lineages for some organisms, which is indeed consistent with **Table 2**.

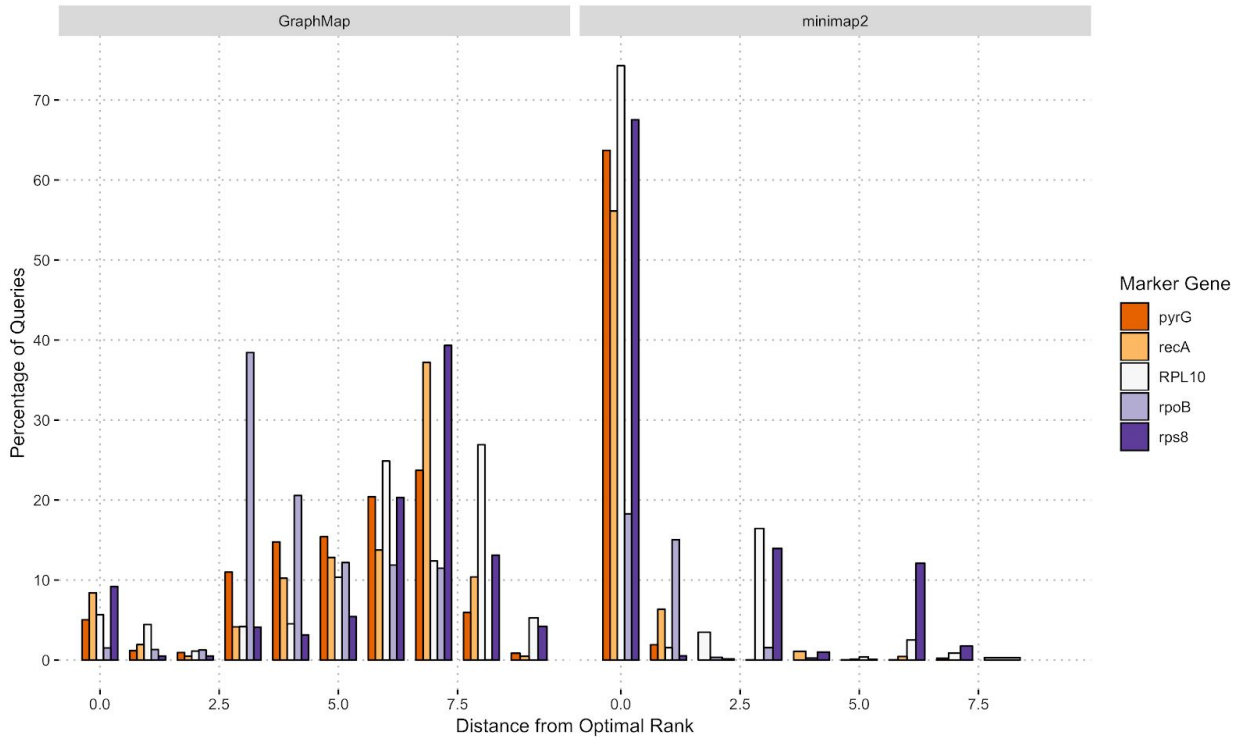

Figure 2. Percentage of query sequences at each distance from their optimal taxonomic rank, between GraphMap and minimap2. Each bar represents the proportion of the total number of mapped reads across all datasets, for a single marker gene at a given taxonomic distance. Optimal taxonomic ranks were determined by taking the deepest LCA between the reported organism in each dataset and each taxonomic lineage in the marker gene reference package. Unclassified query sequences, or query sequences which aligned to a reference organism with a mean alignment of less than 20% were filtered out.

In contrast to ranking alignments solely by distance from the optimal placement, the CTD metric factors in the proportion of query sequences mapping at a given distance as well. By

taking the sum over all alignments for each dataset and marker gene combination, this serves as a weighted average, since there may be outlier points which can shift the scores upwards. In **Figure 3**, CTD scores are shown for each aligner. Here, GraphMap generally has a smaller range of values, in contrast to minimap2. However, the distribution of data points in GraphMap appear to be clustered around values ranging from 4 to 6. In comparison, the majority of data points for minimap2 are near 0, with the exception of the marker gene rpoB.

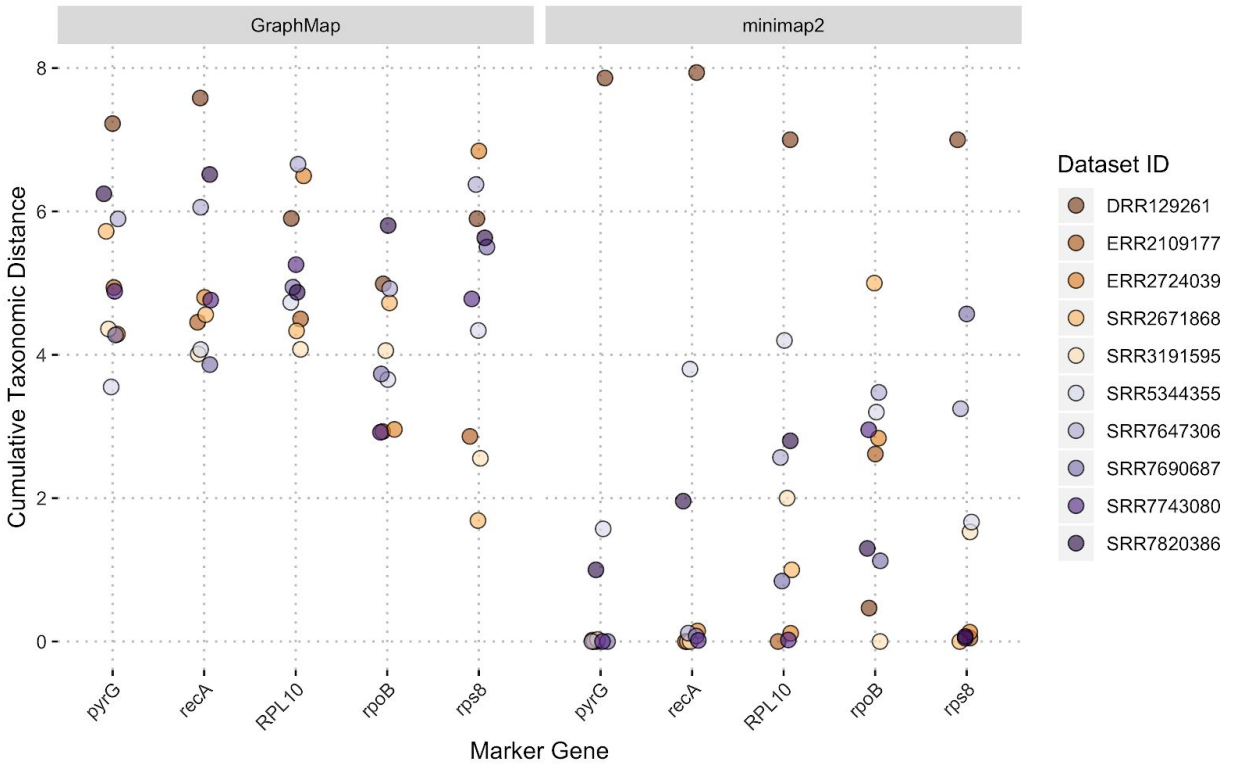

Figure 3. Comparison of cumulative taxonomic distance scores for each marker gene between GraphMap and minimap2. Each dataset was individually mapped to a reference set of marker gene sequences, and cumulative taxonomic distances were calculated by taking the sum of the distance from the optimal placement multiplied by the proportion of mapped query sequences at that distance. Unclassified query sequences were removed prior to plotting.

Of interesting note in minimap2 are the scores for the *Pseudomonas fulva* dataset (DRR129261), as these appear to encompass the majority of extreme values in comparison to all

other datasets. Further exploration revealed that the majority of query sequences in this dataset mapped to numerous unrelated reference sequences compared to dataset organism, despite the majority of reference packages containing this organism. However, relative to other datasets, dataset DRR129261 had lower sequence quality, which likely resulted in the large number of short, low quality alignments observed.

Compared to minimap2, GraphMap reported far more candidate alignments. However, manual inspection of subsets of these alignments for each marker gene revealed that the majority map to unrelated organisms, which suggests a high false negative rate. Indeed, this is consistent with previous alignment tests (26), which can subsequently inflate distance calculations.

The mean and median taxonomic distance for GraphMap and minimap2 are 5.012 and 5.000, and 1.341 and 1.000 respectively, after removing unclassified sequences and query sequences with a mean alignment length of less than 20%. Furthermore, the mean and median CTD scores for GraphMap and minimap2 are 4.820 and 4.771, and 1.727 and 0.922 respectively. The contrast between taxonomic distances between the two aligners suggests that, on average, minimap2 is able to align query sequences to within 1 to 2 ranks of the optimal placement, which is far more accurate than what was achieved with GraphMap. Therefore, minimap2 was chosen to be integrated into TreeSAPP for the long read workflow.

Each dataset was subsequently run after integrating minimap2 into the TreeSAPP pipeline, and by specifying all 5 marker gene references for use. CTD scores using this pipeline are depicted in **Figure 4**. Overall, the data show pyrG, rpoB and rps8 perform relatively well, with the majority of data points distributed around CTD values of 0 to 2, indicating that most query sequences are placed within two taxonomic ranks of the optimal placement. However,

recA and RPL10 show worse scores, as these values are closer to a CTD of 4. This difference likely arises from the construction of the nucleotide reference packages for both marker genes, which can depend on the quality of the initial sequences gathered as the reference, the parameters used for clustering, and the parameters used when building the reference package. While the mean (2.470) and median (2.085) CTD scores are higher than minimap2 alignment based classification alone across all marker genes and datasets, the TreeSAPP pipeline appears to decrease the range of CTD values for each marker gene, suggesting that the phylogenetic approach TreeSAPP utilizes results in more consistent performances.

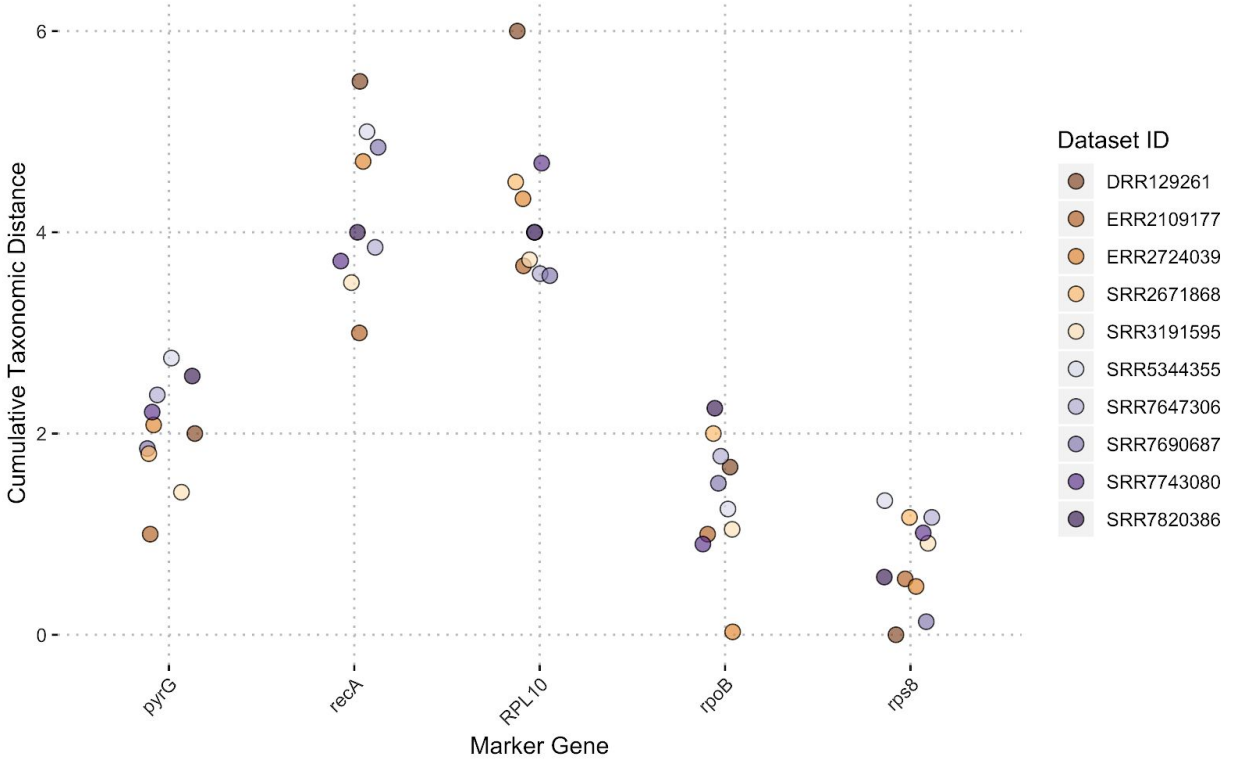

Figure 4. Cumulative taxonomic distance scores for each dataset when classified using TreeSAPP. Each dataset was ran using all 5 marker gene references. Scores were calculated in the same manner as in **Figure 2**.

3.2 Mock community

Table 3 demonstrates the overall performance of the long read workflow of TreeSAPP with the ZymoBIOMICS mock community. The total number of sequences classified by TreeSAPP is 6548 (taken by adding true and false positives), which is considerably less than the total number of reads (33351, taken by adding true positives and false negatives) which were identified by mapping query sequences to the annotated draft assemblies. This suggests that TreeSAPP was unable to identify enough query sequences from the mock community using the marker gene reference sequences before phylogenetic classification. The lack of candidate sequences inherently decreases TreeSAPP's ability to classify sequences, and of greater importance, true positive classifications. This is reflected in a value of 0.234 for Matthew's Correlation Coefficient (MCC). An MCC of 0 indicates no better than random classification, whereas -1 and 1 indicate total disagreement and perfect classification respectively. Therefore, the calculated MCC resolves a relatively weak signal of accurate classification by TreeSAPP.

Classifications	Scores
True Positives	3929
False Positives	2619
True Negatives	3221774
False Negatives	29422
MCC	**0.234**

Table 3. Confusion table and Matthew's Correlation Coefficient (MCC) for classification of query sequences from the ZymoBIOMICS mock community using the long read workflow of TreeSAPP. True positives are composed of query sequences correctly identified by TreeSAPP, accurate within a taxonomic distance of 2. Conversely, false negatives are the set of query sequences which failed to be identified or had a taxonomic distance greater than 2 by TreeSAPP. False positives encompass query sequences which were identified by TreeSAPP but were not annotated in the reference genome, and true negatives represent query sequences which TreeSAPP correctly did not annotate.

4. Discussion

When studying microbial communities, two common questions researchers explore are "who is there?" and "what are they doing?". TreeSAPP aims to answer both questions through classification of protein and nucleic acid sequence information. By extracting single copy phylogenetic and functional marker genes from query sequences, TreeSAPP is able to place such sequences into marker gene reference trees to infer the taxonomic and functional composition from a given community. While TreeSAPP currently supports SGS (e.g. Illumina) data, several challenges arise from these data, such as resolving repetitive regions, and non-uniform coverage. Yet with the rapid development of TGS, longer read lengths and reduced error rates will be achieved, making TGS a better candidate over SGS for future studies. Therefore, this thesis aimed to provide a long read workflow in TreeSAPP by analyzing four candidate aligners (minimap2, GraphMap, SNAP, and LAST) for integration into the existing pipeline.

As mentioned previously, minimap2 was observed to outperform GraphMap by aligning to reference sequences closely related to, if not exactly at the optimal placement for each dataset. This may be due to the unconventional usage of the aligners: typically, each aligner is used to map query sequences to a reference genome, which is longer in length compared to the individual reads. In each aligner tested except for LAST, seeds from the reference sequence(s) are indexed in a hash table (28, 32, 36), which serves as a rapid look-up table to begin the alignment between two homologous regions. However, there are roughly one thousand reference sequences in each nucleotide reference package. Given that the average bacterial gene is approximately 1 kbp, this provides little room for distinct seeds to be constructed. Indeed,

GraphMap constructs an index by calculating Levenshtein distance for seeds, and scores regions on the reference by finding and clustering seeds (32). But each region may potentially come from a different organism, therefore this aligner may be more suitable when aligning to a single reference genome. A similar case can be made for SNAP: this aligner uses longer seed lengths in its index (36) to reduce false positives, but this likely does not give a strong performance enhancement since the marker gene reference sequences are small to begin with. Furthermore, TGS reads are roughly 6-8 kbp on average, which may complicate edit distance calculations when aligning to a much smaller reference sequence. In the case for LAST, the total length of the reference for each marker gene is likely not long enough for the training module to infer the substitution and gap probabilities, as LAST typically requires a full reference genome to train on (35).

Despite this, the long read workflow of TreeSAPP performed relatively well when analyzing isolate datasets, as indicated by CTD scores in **Figure 3**. However, the performance markedly deteriorates when analyzing a mock community, as indicated by the MCC value in **Table 3**. Upon further inspection, 5319 sequences were filtered during the TreeSAPP pipeline due to low likelihood weight ratios (LWR) in phylogenetic placement, and 2 sequences were filtered due to a combination of LWRs and exceeding tree-specific pendant length distances. This implies a total of 11869 candidate sequences were found when aligning reads to the marker gene reference sequences. Compared to the total number of sequences classified as marker genes of interest (33351) when aligning reads to the reference assemblies (the sum of true positives and false negatives), this leaves 21482 unaligned sequences. In the majority of marker gene

references, the 10 species composing the mock community were not found, but broader taxonomic classifications (i.e. genera) were.

The large difference in sequences aligned between reference assemblies and nucleotide reference packages suggests that minimap2 was unable to map homologous regions in the reduced reference set. One possible reason for this is through the approach used to determine the set of reads which map to a marker gene in the reference assemblies: this was based off reads which overlapped marker gene coordinates determined by PROKKA. However, given the long length of TGS reads, it is possible that the same read may align well to a region of the reference, but poorly in the overlapping regions of marker genes. Subsequently, alignment using marker genes as references may filter these out due to low mapping quality.

Regardless, the number of false positives classified by TreeSAPP is relatively high when compared to the true positives, which can trivially be decreased by increasing the maximum taxonomic distance threshold. However, further exploration of the test data would be required to determine the major sources of error, which can include the insertion/deletion rates or single nucleotide polymorphisms. But, TreeSAPP currently incorporates a linear correlation of taxonomic rank with placement distance for taxonomic estimations (21). The placement distances, combined with clade exclusion analysis, serve as parameters to the linear model. Indeed, a model incorporating the error rates of ONT data, which are markedly higher than SGS data (20) likely would produce a better fit, and should be considered in future versions of TreeSAPP.

The process of building nucleotide reference packages also plays a factor in the classification of sequences. Indeed, this concept has been established before for rapid and

reproducible classification in phylogenetic trees (44). In TreeSAPP, a reference package for a marker gene is composed of a multiple sequence alignment, an HMM, and a file mapping organisms to their full taxonomic lineages. While multiple parameters are available for tuning when building a reference package, this flexibility is designed to encapsulate a compact, representative set of organisms to provide accurate taxonomic annotations. Currently, reference packages are built through trial and error parameter sweeps, and manual curation for roughly one thousand reference sequences. This provides a tractable number of sequences to minimize computational time during phylogenetic tree construction, which is the bottleneck in the pipeline.

However when extrapolating to metagenomic data, this method will likely perform poorly. The sheer number of unknown sequences may not be captured in the reference package, which forces TreeSAPP's linear model for taxonomic estimation to place sequences at the root level. Therefore, one possible alternative is to produce metagenome specific reference packages by starting with alignment based methods. Query sequences could first be aligned to a set of unfiltered reference sequences prior to building a reference package to search for candidate organisms. Given the speed of modern aligners such as BWA (29) and minimap2 (28), this step would require little computational time. Then, once candidate organisms are identified, more sequences from closely related organisms could be added to the nucleotide reference package, thus encapsulating a more comprehensive taxonomic profile.

Further refinements to a reference package can also include assigning taxonomies through iterative runs of TreeSAPP. Initial runs can be used to identify a coarse-grained taxonomic profile, and visualized on the marker gene reference trees. Subsequently, additional reference sequences can be added based on where sequences were placed on the tree(s), and the

reference package can be rebuilt, allowing for higher resolution reference packages. Given that there is potential for an immense number of unknown organisms in a metagenomic sample (45), it is no surprise that TreeSAPP would benefit from more prudent and metagenome specific packages, and should be incorporated in the next release.

5. Conclusion

Through evaluation of four aligners, minimap2 outperformed GraphMap when judged by the distance of phylogenetic placement of query sequences from their optimal ranks, and by CTD scores, which serve as averages of distance from the optimal rank, weighted by proportion. The remaining two aligners likely failed due to the smaller reference set of sequences, which goes against the implicit assumption of the reference set being larger than the query set.

Incorporation of minimap2 into TreeSAPP and testing the long read workflow using isolate datasets generally showed consistent classifications within one to two taxonomic ranks from the optimal placements. However, when applied to a mock community, this results in a marked decrease in the number of (correct) sequences classified, since minimap2 fails to map the diverse set of sequences to the reduced reference set, resulting in a high false negative rate and a low recovery rate. Therefore, future work is required in adjusting the linear model to take ONT error profiles into account, and building higher resolution, and potentially metagenome specific nucleotide reference packages.

References

1. Vollmers J, Wiegand S, Kaster A-K. 2017. Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters! PLoS One 12:e0169662.

2. Hugerth LW, Andersson AF. 2017. Analysing Microbial Community Composition through Amplicon Sequencing: From Sampling to Hypothesis Testing. Front Microbiol 8:1561.

3. Woese CR. 1987. Bacterial evolution. Microbiol Rev 51:221–271.

4. Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A 74:5088–5090.

5. Woese CR, Stackebrandt E, Macke TJ, Fox GE. 1985. A phylogenetic definition of the major eubacterial taxa. Syst Appl Microbiol 6:143–151.

6. Rashid M, Stingl U. 2015. Contemporary molecular tools in microbial ecology and their application to advancing biotechnology. Biotechnol Adv 33:1755–1773.

7. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055.

8. Sanger F, Nicklen S, Coulson AR. 1992. DNA sequencing with chain-terminating inhibitors. 1977. Biotechnology 24:104–108.

9. Heather JM, Chain B. 2016. The sequence of sequencers: The history of sequencing DNA. Genomics 107:1–8.

10. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. 2018. The Third Revolution in Sequencing Technology. Trends Genet 34:666–681.

11. Greenleaf WJ, Sidow A. 2014. The future of sequencing: convergence of intelligent design and market Darwinism. Genome Biol 15:303.

12. Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M. 2012. Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012:251364.

13. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S. 2009. Real-time DNA sequencing from single polymerase molecules. Science 323:133–138.

14. Wasfi A, Awwad F, Ayesh AI. 2018. Graphene-based nanopore approaches for DNA sequencing: A literature review. Biosens Bioelectron 119:191–203.

15. Carriço JA, Rossi M, Moran-Gilad J, Van Domselaar G, Ramirez M. 2018. A primer on

microbial bioinformatics for nonbioinformaticians. Clin Microbiol Infect 24:342–349.

16. Metzker ML. 2010. Sequencing technologies - the next generation. Nat Rev Genet 11:31–46.

17. Paszkiewicz K, Studholme DJ. 2010. De novo assembly of short sequence reads. Brief Bioinform 11:457–472.

18. Bleidorn C. 2015. Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. System Biodivers 14:1–8.

19. Weirather JL, de Cesare M, Wang Y, Piazza P, Sebastiano V, Wang X-J, Buck D, Au KF. 2017. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. F1000Res 6:100.

20. Rang FJ, Kloosterman WP, de Ridder J. 2018. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol 19:90.

21. Morgan-Lang C, McLaughlin R, Zhang G, Konwar K, Armstrong Z, Song Y, Crowe S, Hallam SJ. TreeSAPP: phylogenetic Tree-based Sensitive and Accurate Protein Profiling. In Preparation.

22. Liu W, Li L, Khan MA, Zhu F. 2012. Popular molecular markers in bacteria. Molecular Genetics, Microbiology and Virology.

23. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC

Bioinformatics 11:119.

24. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res 41:e121.

25. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

26. Pavlovikj N, Moriyama EN, Deogun JS. 2017. Comparative analysis of alignment tools for nanopore reads2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).

27. Li H, Homer N. 2010. A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform 11:473–483.

28. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100.

29. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

30. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

31. Roberts M, Hayes W, Hunt BR, Mount SM, Yorke JA. 2004. Reducing storage requirements for biological sequence comparison. Bioinformatics 20:3363–3369.

32. Sović I, Šikić M, Wilm A, Fenlon SN, Chen S, Nagarajan N. 2016. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. Nat Commun 7:11307.

33. Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. 2011. Adaptive seeds tame genomic sequence comparison. Genome Res 21:487–493.

34. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

35. Hamada M, Ono Y, Asai K, Frith MC. 2017. Training alignment parameters for arbitrary sequencers with LAST-TRAIN. Bioinformatics 33:926–928.

36. Matei Zaharia, William J. Bolosky, Kristal Curtis, Armando Fox, David Patterson, Scott Shenker, Ion Stoica, Richard M. Karp, and Taylor Sittler. 2011. Faster and More Accurate Sequence Alignment with SNAP. arXiv:11115572v1.

37. Fish JA, Chai B, Wang Q, Sun Y, Brown CT, Tiedje JM, Cole JR. 2013. FunGene: the functional gene pipeline and repository. Front Microbiol 4:291.

38. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461.

39. Zhang J. 2000. Protein-length distributions for the three domains of life. Trends Genet 16:107–109.

40. Bushnell B. 2014. BBMap: A Fast, Accurate, Splice-Aware Aligner. Conference: 9th

Annual Genomics of Energy & Environment Meeting.

41. Nicholls SM, Quick JC, Tang S, Loman NJ. Ultra-deep, long-read nanopore sequencing of mock microbial community standards.

42. McIntyre ABR, Alexander N, Grigorev K, Bezdan D, Sichtig H, Chiu CY, Mason CE. 2019. Single-molecule sequencing detection of N6-methyladenine in microbial reference materials. Nat Commun 10:579.

43. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069.

44. Boyd JA, Woodcroft BJ, Tyson GW. 2018. GraftM: a tool for scalable, phylogenetically informed classification of genes within metagenomes. Nucleic Acids Res 46:e59.

45. Wooley JC, Godzik A, Friedberg I. 2010. A Primer on Metagenomics. PLoS Computational Biology.