8-2018

The Oral Microbiome of Site-Specific Dental Plaque in Health and Disease

Abdullah Abood
Clemson University, abdullah.abood@outlook.com

THE ORAL MICROBIOME OF SITE-SPECIFIC DENTAL PLAQUE IN HEALTH AND DISEASE

A Thesis
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
Microbiology

---

by
Abdullah Abood
August 2018

Accepted by:
Dr. Vincent P. Richards, Committee Chair
Dr. Christina Wells
Dr. J. Antonio Baeza

ABSTRACT

According to the National Institutes of Health, dental caries is the leading chronic disease of children in the United States. Dental caries is biofilm-mediated, multifactorial and dynamic. Research using culturing techniques and high throughput 16S rRNA amplicon sequencing unraveled the taxonomic complexity of mixed microbial communities (microbiome) in dental biofilms (plaque) and their abundance differences. However, 16S rRNA sequencing fails to resolve taxonomic assignment beyond genus level for certain taxa, which is problematic in identifying potential antagonistic species within the same genus. The presented work addressed current shortcomings in dental microbiome research. First, dental plaque samples used in this study were collected from either caries-free (PF) teeth or caries-active teeth with lesions in the enamel layer (PE). This site-specific collection method provides a better understanding of the role of specific organisms and biological processes as teeth transition from health to disease. Second, deep sequencing was used to produce whole genome metagenomic data, i.e. complete or semi complete genomes drafted from mixed bacterial communities, potentially enhancing bacterial species detection, identifying rare species, and providing the gene content of the samples and their metabolic potential. Overall, the objective of this study was to provide species level taxonomic classification and metabolic potential of mixed microbial communities in plaque collected from site-specific dentition. Two different approaches to analyze whole genome metagenomic data were used and

compared. (i) Read based taxonomic classification and supervised assembly where short reads are taxonomically classified prior to genome assembly. (ii) Contig based taxonomic classification and unsupervised assembly where an assembler is used to assemble reads into contigs directly. The contigs produced are then classified taxonomically. The read based taxonomic classification and supervised assembly approach outperformed the latter in an assessment of taxonomic assignment accuracy using a mock metagenomic data set. The taxonomic profiles for PF and PE reported by both approaches were virtually identical however their distributions showed variation. The taxonomic inter-sample similarities were reflected in the gene content information as both approaches reported minor metabolic potential differences between PF and PE. Noticeably, both approaches reported significantly enriched biological processes involved in sugar transport and metabolism in PE.

DEDICATION

I would like to dedicate the work presented here to refugees everywhere in the world.

It will always get better…..

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Vincent P. Richards for his mentorship. I would like to thank Dr. Christina Wells and Dr. Antonio J. Baeza for their continuous support throughout my graduate studies.

TABLE OF CONTENTS

Page

TITLE PAGE i

ABSTRACT ii

DEDICATION iv

ACKNOWLEDGMENTS v

LIST OF TABLES vii

LIST OF FIGURES viii

CHAPTER

I. CHAPTER ONE 1

Introduction 1
Methods 7
Results 14
Discussion 20
Future Direction 25

APPENDICES 38

A: Molysis kit assessment 38
B: Sequencing evaluation 39

REFERENCES 27

LIST OF TABLES

Table Page

1 Basic information of the results of the read based taxonomic classification approach 35

2 Basic information of the results of the contig based taxonomic classification approach 35

3 List of the 31 core shared taxa between PF and PE reported by both approaches and their % abundance and total proportion of the sample. Green signifies abundance increase/decrease shown consistently in both approaches while red signifies inconsistent abundance increase/decrease reported. 36

4 Significantly enriched GO terms in both approaches using FDR corrected p-values. The domain (BP) denotes a Biological Process, (MF) denotes a Molecular function, and (CC) denotes a Cellular Component. Bolded signifies notable terms**.** 37

LIST OF FIGURES

Figure Page

1 Workflow showing the steps and programs used in (i) read based taxonomic classification and supervised assembly and (ii) contig based classification and unsupervised assembly. 31

2 Results reported from read based taxonomic classification approach for the 10 most abundant species. The % relative abundance calculations are based on the most abundant 100 taxa in this approach. 32

3 Results reported from contig based taxonomic classification approach for the 10 most abundant species. The % relative abundance calculations are based on the most abundant 48 taxa in PF and 42 taxa in PE..... 33

4 Results of the pipeline assessment step with column (1) representing the actual proportions of the mock community, column (2) representing the proportions reported using the read based taxonomic classification approach, and column (3) representing the results reported using the contig based taxonomic classification. 34

Introduction

The human oral cavity is home to around 700 species, making it one of the most complex and diverse human microbiomes [1-3]. Anatomically, the oral cavity can be divided into two types of functional surfaces: soft (mucosal) and hard (teeth) [4]. The human oral microbiome plays a significant role in health and disease [5-8] specifically dental caries and periodontitis [4, 9-12]. Dental caries is a biofilm-mediated and multifactorial disease characterized by the dissolution and demineralization of tooth structure due to acid [13]. The disease still stands as the most common chronic disease in children in the United States [14].

Oral health affects people physically and psychologically [9]. Early tooth loss caused by dental decay has been associated with the failure to thrive, impaired speech development, and reduced self-esteem [9]. Over 50% of children in the United States between the ages of 5 and 9 have at least one cavity or filling and that number increases to 78% among 17 year olds [9]. Early Childhood Caries, a rampant form of dental caries, affects approximately 23% of preschoolers less than six years of age in the US and can be observed in toddlers as young as 12 months [15]. Thus, there is a clear need to enhance our understanding of this disease.

Previous studies using culturing techniques, microarray and Multi Locus Sequence Typing (MLST) identified *S. mutans* and *lactobacilli* as major etiological agents in carious lesions [16-19]. *S. mutans* and *lactobacilli* are acidogeneic bacteria capable of fermenting sugars, mainly sucrose [20].

Excessive ingestion of sugars causes an overproduction of acid creating a low pH environment that promotes lesions in the enamel and the underlying dentin layer [9, 21]. Recently, microbiome research using high throughput 16S rRNA amplicon sequencing was used to identify a number of bacteria such as *Bifidobacterium*, *Propionibacterium* and *Scardovia* associated with diseased microflora in addition to *S. mutans* [9, 22, 23]. The reported commensal components of the biofilm were *Firmicutes*, *Actinobacteria*, and *Streptococcus* species [23-25].

Early studies demonstrated that healthy dental plaque was a highly diverse community that included *S. mutans* at a low frequency. On the other hand, caries active teeth were dominated by acidogenic and aciduric bacteria, which create an acidic environment that is uninhabitable by non-aciduric bacteria and therefore less diverse [21, 22]. These results suggest a relationship between plaque microbial composition and the oral health of the host. However, a recent study combining high throughput 16S rRNA amplicon sequencing and site-specific sampling showed that species richness is not necessarily decreased in caries active teeth [26]. The same study also reported *S. mutans* was present at a low frequency, despite the diseased status of teeth in some patients [26]. These findings indicate that there may be more to caries development than the presence of *S. mutans*.

While progress in dental microbiome research has been made, there are shortcomings that the work presented in this thesis address. First**,** previous

dental microbiome research has focused on results from whole dentition [27], pooling plaque from both healthy and diseased tooth surfaces and potentially providing an inaccurate representation of microbial communities associated with specific dental health conditions. In the present study, I obtained samples from caries-free (PF) and early- stage diseased teeth (PE) of the same individual. Comparing healthy and diseased in this manner provide insight into the specific organisms and biological processes that are involved in the transition from health to disease.

Second, high throughput 16S rRNA amplicon sequencing has enabled scientists to examine the bacterial complexity of dental plaque. Nevertheless, for certain taxa, the 16S rRNA gene fails to resolve the taxonomic assignment beyond the genus level [26]. Characterizing the bacterial components of health and disease at the genus level potentially introduces a problem in composition accuracy. For example, the *Streptococcus* genus has been shown to be present in both healthy and diseased dental plaque, thus failure to resolve the genus to a species level is problematic. Deep sequencing of mixed bacterial samples known as Whole Genome Shotgun (WGS) Metagenomics is made possible through Advances in sequencing technology. There are two advantages to using WGS metagenomic sequencing in comparison with high throughput 16S rRNA amplicon sequencing. First, superiority in resolving genera into species as WGS metagenomic sequencing increases sequencing coverage (the number of unique reads that belong to a region of the genome) and the possibility of drafting

complete genomes from a community. Second, WGS metagenomic sequencing provides the ability to attain the gene content information of species within the community. It is imperative to obtain a complete picture of the bacteria present in dental plaque. However, dental plaque bacterial composition information should be coupled with metabolic potential information in order to gain insight of the potential functions of the community as a whole. This information can only be obtained with analysis of the gene content information of the community. For example, alkali-producing bacteria that counteract the effect of acidogeneic bacteria use a variety of pH buffering mechanisms. Examples of these mechanisms are utilization of urease to produce ammonia from urea [24] or producing ammonia from arginine using the arginine deiminase system (ADS) [1].

Lastly, thus far, no gold standard has been adopted to provide taxonomic classification and gene content of whole genome metagenomic data. Researchers potentially utilize two approaches to obtain taxonomic assignments, either taxonomically classifying short reads directly or assembling reads into contigs first then taxonomically classifying contigs. Each approach has potential downfalls. (i) Classifying short reads directly prior to assembly could lead to erroneous hits in the database, this is because as the sequence length shortens, the sequence potentially matches multiple hits in the database resulting in random taxonomic assignments. (ii) Classifying contigs with a taxonomic classifier is preceded with metagenomic assembly risking the possibility of hybrid

species reads being incorporated into the same contig, which leads to inaccurate taxonomic assignment. The same applies for obtaining gene content information. Researchers can classify short reads taxonomically prior to assembly (supervised assembly). On the other hand, short reads could be assembled into contigs directly (unsupervised assembly). The former is more advantageous in well-characterized communities like the oral cavity since the majority of taxa are known and are present in databases. The latter, however, is potentially more beneficial in drafting novel genomes and thus more applicable in communities with little or no information in databases.

Here, I utilized and compared two approaches in analyzing mixed bacterial communities. (i) Read based taxonomic classification and supervised assembly where short reads are taxonomically classified prior to genome assembly. I utilized taxonomic classification as the step preceding assembly. Previous research used coverage depth, sequence composition, mate-pair information, and other criteria instead [28-30]. (ii) Contig based taxonomic classification and unsupervised assembly where an assembler is used to assemble reads into contigs directly. The contigs produced are then classified taxonomically. Both approaches provide a species level resolution and the gene content information of the community.

Overall, the objective of this study was to provide species level taxonomic classification and metabolic potential of mixed microbial communities in plaque collected from site-specific dentition. Two different approaches to analyze whole

genome metagenomic data were used and compared: The read based taxonomic classification and supervised assembly approach outperformed the latter in an assessment of taxonomic assignment accuracy using a mock metagenomic data set. The taxonomic profiles for PF and PE reported by both approaches were virtually identical however their distributions showed variation. The taxonomic inter-sample similarities were reflected in the gene content information as both approaches reported minor metabolic potential differences between PF and PE. Noticeably, both approaches reported significantly enriched biological processes involved in sugar transport and metabolism in PE.

Methods

Study subject

Following Richards' *et al* [26], two site-specific supragingival dental plaque samples were collected from the same child. Informed consent was obtained from parents or legal guardians of the child under a protocol approved by the Institutional Review Board of the University of Florida Health Science Center (#272-2010). The selection process excluded children who received antibiotics within 3 months of their study visit, who were taking any medication, or had orthodontic appliances. The samples were collected from the same child who was categorized as caries active [mean number decayed teeth (DT=0); mean number of missing and filled teeth (MFT>0)]. One plaque sample was collected from caries free teeth (PF) and the other from caries active teeth with lesions in the enamel layer (PE).

Sample collection

Following Richards' *et al* [26], the patient halted any oral hygiene practices 8 hours prior to sample collection. Samples were collected separately from: (i) tooth surfaces that were caries-free (PF) and (ii) active, enamel carious lesions (PE). Each plaque sample was obtained by pooling material from at least two different tooth sites of similar health condition (site-specific) using sterile periodontal curettes.

Sample QC

Absolute quantitative PCR (qPCR) was used to assess the amount of human DNA in samples relative to bacterial and fungi prior to sequencing. Genomic DNA was extracted from plaque using ZR Fecal DNA Kit™ (Zymo Research, CA, USA) in accordance with supplier's specifications. The targeted genes of interest were 16S rRNA V4 region gene for bacteria (primer sequence Forward: GTGCCAGCAGCCGCGGTAA Reverse: GGACTACCAGGGTATCTAAT) and β-actin gene for humans (Forward: GCGTTACACCCTTTCTTGACA Reverse: CGCATCTCATATTTGGAATGACT). Two replicates were used per sample. The samples were prepared for qPCR using SYBR Green Master Mix (BioRad, CA, USA) in accordance with supplier's specifications.

Sample enrichment

The remaining volume of plaque was treated with MolYsis™Basic5 kit (Molzym Life Science, Bremen, Germany) to remove human DNA and enrich microbial DNA for sequencing. The enrichment kit efficiency was assessed on two plaque samples prior to application on the study samples. The assessment compared the amount of human DNA after and before treatment with Molysis using absolute quantitative PCR (qPCR) using the same steps mentioned in the section above.

Genomic DNA extraction and sequencing

Genomic DNA was extracted from treated plaque using ZR Fecal DNA Kit™ (Zymo Research, CA, USA) in accordance with supplier's specifications. DNA volume was measured on Qubit 3.0 Flourometer (Thermo Fisher Scientific, MA, USA) using dsDNA High Sensitivity assay kit (Invitrogen, CA, USA). Library preparation was completed using Nextera XT 150bp Paired-End reads kit (Illumina, CA, USA) at the Weill Cornell Medical College Genomics and Epigenomics Core Facility. The library was sequenced at the same facility using one lane on the Illumina HiSeq 4000 NGS platform. The raw reads in Fastq files were evaluated for quality using the program FastQC version 0.11.5 [available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/] (See appendix B). Reads with ambiguous bases and sequencing adapters were trimmed using the program ea-utils version 1.1.2 [available at https://github.com/ExpressionAnalysis/ea-utils].

Read based taxonomic classification and supervised assembly pipeline

A workflow of this pipeline can be found in figure 1. Reads were assigned to a taxon using the program Kaiju version 1.6.1 [31] in Maximum Exact Match mode. In this program, reads are first translated into amino acids sequences that are then searched against a protein database to find exact matches. I made use of a custom database within our laboratory built using 95,021 bacteria and 992 fungi genome sequences and the human genome. From NCBI Refseq, 94,664

bacteria and 249 fungal assembled and annotated genome sequences were obtained. An additional 743 fungal and assembled and annotated genome sequences were obtained from GenBank at NCBI. To these, the following were added. (i) 261 *Streptococcus mutans* genomes whose reads were obtained from SRA at NCBI (these genomes represented additional strains not contained in refseq genome database). Reads were assembled using A5-miseq [32] and annotated using Prokka [33]. (ii) 96 Streptococcus genomes (oral isolates) representing 12 species that were isolated and sequenced by our laboratory (unpublished data). These isolates were sequenced on the Illumina Miseq platform using v2 chemistry (250bp reads, Paired End) and assembled and annotated as above. The speed and precision of classification, coupled with a low memory requirement are considered the main advantages for using Kaiju over other classification software such as Kraken [31]. The reads were taxonomically classified and placed in their respective species directories (bins). Unclassified reads and reads classified at the genus level were removed from abundance calculations. In order to obtain a normalized measure of species abundance, read coverage was used as proxy for species counts. The number of assigned reads to a specific species was divided by the average genome length of that species. The Whole Genome Shotgun (WGS) database at NCBI was utilized to acquire an average genome length for each species. The abundance calculation was based on the equation below:

Equation 1

$$\frac{(\overline{R} * \#R)}{\overline{G}}$$

Where

$\overline{R}$ = Average read length in the entire sample. A constant number obtained from the assembler

#R = Number of reads assigned to a specific species (bin)

$\overline{G}$ = Average genome length of a specific species in Mb

Next, The individual species (bins) and the unclassified reads were assembled using the program Spades version 3.11.0 [34]. The quality of assembly was assessed using Quast version 4.5 [35]. The open reading frames (ORFs) of the assembled genomes and the unclassified contigs were called using the annotation pipeline Prokka version 1.12 [33]. The amino acid sequences were extracted as input for the program Interproscan version 5 [36] that assign Gene Ontology (GO) terms, which describe gene functions along three domains (molecular functions, cellular component, and biological process). Enrichment of GO terms between samples was determined with a Fisher exact test implemented using GOaTools [available at https://github.com/tanghaibao/goatools]. P values were corrected for multiple testing using the False Discovery Rate (FDR) [37] (α=0.05).

Contig based taxonomic classification and unsupervised assembly pipeline

A workflow diagram for this pipeline can be found in Figure 1. The reads were assembled directly (unsupervised assembly) using metaSpades 3.11.0 [38], a metagenomic assembler found within the spades assembly toolkit. The produced contigs were classified taxonomically using Kaiju following the same steps in the read based approach. The unclassified contigs and the contigs classified at genus level were removed from abundance calculations. In order to obtain the number of alignments assigned to the contigs in each bin, the program Burrows-Wheeler Aligner (BWA) version 0.7.17 [39] was used to align the trimmed unassembled reads in the fastq files to the contigs within assembled taxonomically classified bins. The BWA-MEM algorithm was used in this step following the developers' recommendation [available at http://bio-bwa.sourceforge.net/] as it is faster and more accurate than its counter part BWA-SW and supports longer read alignments contrary to BWA-backtrack. The number of taxa with 1X coverage was low. Therefore, bins with at least 1,000,000bp were included in determining relative abundance. The number of alignments in each of these bins was calculated from the produced SAM files using BBmap version 37.75 [available at http://sourceforge.net/projects/bbmap]. Each alignment was treated as a read assigned to that bin for purposes of determining relative abundance. Once the number of alignments was obtained, the abundance calculations were conducted using the same equation in the read based method. Open reading frames calling along with GO terms assignments,

and GO term enrichment tests were conducted using the same methods in the read based approach.

Pipeline assessment

A 10-taxa mock metagenome with known read proportions, originally used to evaluate metagenomic assembly in a study by *Mende et al.* [40], was utilized to asses the accuracy of taxonomic classification following the same methods for both read and contig based approaches. Figure 6 provides a list of the taxa along with their proportions in the mock sample.

Results

Sample quality and assessment

Previous research reported human DNA contamination values of ~90% in oral plaque samples [41]. These alarming contamination levels could lead to skewing the coverage of the sequencing process towards human DNA rather than microbial. Therefore, genomic DNA was extracted from both samples and Quantitative Polymerase Chain Reaction (qPCR) was used to assess the level of human contamination in the samples. The qPCR results showed approximately 13% human contamination in plaque collected from teeth with lesions in the enamel layer (PE) in comparison with 2% of human DNA in plaque collected from caries free teeth (PF). These results might not significantly affect the overall coverage of microbial DNA, albeit, presence of unwanted human DNA was reported which warranted the use of microbial enrichment to avoid loss of coverage to human contamination. Molysis enrichment kit claims to decrease the level of human DNA contamination in mixed samples to <1%. The qPCR results of the Molysis efficiency test showed that plaque samples collected from caries free teeth started with 3-4% human contamination that was lowered in both cases to <1% human contamination (Appendix A). The qPCR results were confirmed after sequencing as both samples showed approximately 0.03% of the reads (31,191 PF and 47,892 in PE) assigned to the human genome.

Pipeline assessment

The results of the pipeline assessment are shown in figure 4. In the read based taxonomic classification approach, approximately 6% of the reads were either unclassified or assigned to taxa different from the 10 taxa comprising the mock sample relative to approximately 17% of the alignments In the contig based taxonomic classification approach. Overall, the resulting proportions of the 10 taxa reported by the latter were mostly lower and less precise than those reported by the read based approach.

Read based taxonomic classification and supervised assembly pipeline

Refer to table 1. The number of reads remaining after the trimming process for the PF sample was 125,326,716. Of those reads, 32% were not resolved to species level and 5.7% were unclassified using the read based classification method. The number of reads remaining after the trimming process for the PE sample was 166,763,119. Of those reads, 24.1% were not resolved to species level and 5.9% were unclassified using the read based classification method. The number of taxa classified in PF was 13,919 and in PE 13,733. In PF, approximately 5% of the taxa (685 bins of 13,919) were singletons, which denote that only one read was assigned to the taxon. In fact, ~95% of the of the taxa (13,154 bins) were assigned <1000 reads each. In PE, 5% of the taxa (700 bins of 13,733) were singletons as well with approximately 92% of the taxa (12,642 bins) assigned <1000 reads each. Abundance calculation analysis was restricted to taxa that had at least 1X coverage based on Equation 1 in the

methods. The number of taxa that showed at least 1X in PF was 100 species, which is exactly the same in PE. Approximately 74,000,000 reads (59% of total reads) were assigned to the first 100 taxa in PF while only 4,000,000 reads (3% of total reads) were assigned to the rest of the taxa (13,819). A Similar pattern was observed in PE where ~112,000,000 reads (67% of total reads) were assigned to the first 100 taxa while the rest of the taxa (13,633) received a total of 4,500,000 reads (2.7% of total reads). Figure 2 shows the most abundant taxa in both plaque samples along with percentage relative abundance. The two plaque samples shared 70 different taxa albeit with differences in abundance. In PF, the proportion of the 70 shared taxa comprised 98.6% of the total sample abundances while the proportion of the 30 unique taxa comprised only 1.4%. In PE, the proportion of the 70 shared taxa comprised 98.8% of the total sample abundances while the proportion of the 30 unique taxa was only 1.2%.

Overall, both PF and PE shared similar profiles with minor taxonomic composition differences, albeit at mostly differing abundances. PF contained more *Prevotella* and *Neisseria* species, while PE contained more *Leptotrichia* and *Actinomyces* species. Both health conditions were dominated by *Actinobaculum* sp. oral taxon 183 with relative abundance of ~41% in PF and ~52% in PE. *Campylobacter jejuni* was uniquely associated with PE (~200,000 reads in PE compared to 530 reads in PF). *Prevotella nigrescens* was uniquely associated with PF (106,000 reads compared to 673 reads in PE).

Contig based taxonomic classification and unsupervised assembly pipeline

Refer to table 2. The total number of contigs produced by the assembly program metaSpades for PF was 69,175 with N50 value of 6,800 bp and for PE 57,242 with N50 value of 8,802 bp. Approximately 28% of the contigs were not resolved to species level in PF and ~6% were unclassified while ~30% of the contigs were not resolved to species level and ~8% were unclassified in PE. The number of taxa classified in PF was 3,397 compared to 3,335 in PE. The number of taxa with 1X coverage was low. Therefore, bins with at least 1,000,000bp were included in abundance analysis. The resultant number of taxa was 48 in PF and 42 in PE with 31 shared taxa between them.. In PF, the proportion of the 31 shared taxa was 85.3% of the total sample abundances while the proportion of the 19 unique taxa was 14.7%. In PE, the 31 shared taxa comprised 86.8% of the sample while the proportion of the 11 unique taxa was 13.2%. Similar to the results in the read based taxonomic classification approach, *Actinobaculum* sp. oral taxon 183 was the most abundant with relative abundance of 13% in PF and 17% in PE. Overall, PF contained more *Prevotella* species (*P. nigrescens, P. oris, P oral taxon 317, and P. HMSC073D09*) while PE contained more *Streptococcus* species (*S. mitis and S. sp. SK643*).

Shared taxa reported by both approaches

Refer to table 3. The number of identical taxa reported by both approaches in both PF and PE (core taxa) was 31. The proportion of those

shared 31 taxa comprised 93.1% of total sample abundances in read based PF, 94.8% in read based PE, 83% in contig based PF, and 86.8% in contig based PE. Noticeably, the proportion of total sample abundances of unique taxa in PF or PE reported by both approaches was low.

For the most part, despite the differences in taxa abundances reported by both approaches, a conserved trend of abundance increase/decrease as transition from health to disease occurred was observed. In other words, if the taxon was reported to be increased in health and decreased in disease or vice versa in the read based approach, the relationship was consistent in the contig based approach for the majority of the taxa (Table 3).

Gene calling

The two approaches reported comparable numbers of open reading frames in both PF and PE. In the supervised assembly ~129,000 genes were called in PF and decreased to ~99,000 in the unsupervised assembly approach. The same pattern was observed in PE, where ~235,000 genes were called by the supervised assembly approach compared to ~217,000 in the unsupervised assembly approach.

Gene Ontology (GO) terms enrichment

Significantly enriched GO terms along with their description can be found in table 4. The test was conducted to compare the enriched terms (significantly

increased gene copy numbers) in PE relative to PF. No significantly enriched terms in PF relative to PE were reported. Overall, there were 15 enriched GO terms in the read based (supervised assembly) approach in comparison to only 3 in the contig based (unsupervised assembly) approach. The majority of the terms identified by both approaches were broad. However, two GO terms in the read based (supervised assembly) approach associating with carbohydrate and sugar transport and metabolism were enriched in PE relative to PF (GO:0008643 and GO:0009401). Comparatively, a single GO term in the contig based (unsupervised assembly) approach associating with sugar transport and metabolism was enriched in PE relative to PF (GO:0009401). In addition, the other two significantly enriched terms in PE show oxidoreductase activity, which might reflect, increased sugar catabolism.

Discussion

This study evaluated two microbiome analysis approaches of whole genome metagenomic data produced from deep sequencing of supragingival dental plaque collected from tooth sites with clinically different caries status in a single child. These approaches could enhance understanding of bacterial profile and metabolic potential changes as the tooth transitions from health to disease. The study utilized two paired samples from the same child and provided important findings. In both read based and contig based approaches, the transition from health to disease did not strongly affect the taxonomical composition of the samples, however the abundance of most taxa changed. The changes were not exclusively associated with the change from PF to PE or vice versa, but also with using one analysis approach or the other.

As the results showed, the number of identical taxa reported by both approaches in both PF and PE (core taxa) was 31. The proportion of those shared 31 taxa comprised the majority of both PF and PE in both approaches. For the most part, despite the differences in taxa abundances reported by both approaches, a conserved trend of abundance increase/decrease as transition from health to disease occurred was reported. In other words, if the taxon was reported to be increased in health and decreased in disease, or vice versa, in the read based approach, the relationship was consistent in the contig based approach for the majority of the taxa (Table 3). While a majority proportion of each sample was occupied by the same taxa, GO term analysis using both

approaches reported enriched biological processes involved in sugar transport and metabolism in the disease stage (PE). In order to account for which taxa were potentially responsible for the metabolic differences, the fold changes in abundance were calculated (Table 3). The taxa that showed noticeably high increase in abundance within PE relative to PF were considered potential contributors to metabolic differences between the two samples, especially sugar transport and metabolism. For both approaches, the most abundant taxon obtained in both PF and PE is *Actinobaculum* sp. oral taxon 183. This taxon was shown in multiple dental plaque microbiome studies as associated with both health and disease albeit at higher proportions in disease [42-44]. The same pattern was also shown here as the abundance of this taxon was increased in PE. *Actinomyces gerencseriae* showed a noticeably increased abundance in PE (Table 3). This increase was consistent with previous findings in literature as this taxon was shown to be present in both health and disease albeit at higher abundance in the latter [45]. Both of these taxa were shown to associate with increased sugar metabolism. A study by *Keller et al.* reported that as sugar intake is decreased, the proportion of *Actinobaculum sp*. Oral taxon 183 is decreased [42]. A study by *Dame-Teixeira et al*. investigating gene expression in root caries showed enzymes involved in glycolysis were highly expressed in *Actinomyces gerencseriae* [46]. Thus, significantly enriched GO terms involved in sugar transport and metabolism in PE could be attributed to the abundance increase of *Actinobaculum* sp. Oral taxon 183 and *Actinomyces gerencseriae.*

Noticeably, in the contig based taxonomic classification approach, the proportions of the core taxa in both PE and PF were lower than that reported by the read based taxonomic classification approach in the same samples. These proportion differences were consistent with the results reported by the pipeline assessment with many taxa reporting lower abundance when contigs are classified compared to the reads and an overall lower proportion. Reasons for this discrepancy could be attributed to the metagenomic assembly step, which potentially introduced hybrid-taxa contigs that reduce classification accuracy [47, 48]. The erroneous contigs caused a propagation of inaccurate taxonomic classification. Based on both the experimental and mock data results, the read based taxonomic classification approach for this dataset outperformed the contig based taxonomic classification approach in abundance reporting accuracy. Varying results due to method differences in taxonomic classification of whole genome metagenomic data have been previously reported [49] suggesting the necessity of using standardized datasets for method comparison prior to propagating misleading results.

From a biological standpoint, the results of this study were for the most part consistent with the literature. For example, *Lautrpoia mirabilis, Streptococcus sanguinis*, *abiotrophia defective,* and *Neisseria* oral taxon 14 were reported previously as commensal oral microflora [50-53], which was consistent with the results reported in this study. *Leptotrichia* species, opportunistic

pathogens able to ferment carbohydrates and produce lactic acid [54], were shown to be associated uniquely with disease.

Oddly, *Prevotella nigrescens*, implicated in periodontal disease [55], was reported as uniquely associated with PF. *Actinomyces naeslundii was* decreased in abundance in PE compared to PF contrary to what was previously reported[26]. In the read based taxonomic approach*, Streptococcus mutans* was shown to occupy <1% of the data in both PF and PE samples. Given that previous research has shown that *S. mutans* is a major culprit in dental caries [17], these results were striking and potentially support the claim made by *Richards et al.* and *Gross et al* [22, 26] that *S. mutans* role might not be of importance until the caries have reached a progressive state. *Campylobacter jejuni,* a major human pathogen associated with foodborne gastrointestinal illness [56], was reported to be associated with PE. While several *Campylobacter* species are implicated in human periodontal disease -such as *C. rectus*, *C. gracilis*, and *C. showae* [57]- *C. jejuni* was not shown to associate with dental disease. In the contig based taxonomic classification approach, *Streptococcus mitis*, a major commensal bacterial in the oral cavity [58], was reported as uniquely associated with disease

Overall, the objective of this study was to provide species level taxonomic classification and metabolic potential of mixed microbial communities in plaque collected from site-specific dentition. Two different approaches to analyze whole genome metagenomic data were used and compared. (i) Read based taxonomic

classification and supervised assembly where short reads are taxonomically classified prior to genome assembly. (ii) Contig based taxonomic classification and unsupervised assembly where an assembler is used to assemble reads into contigs directly. The contigs produced are then classified taxonomically. The read based taxonomic classification and supervised assembly approach outperformed the latter in an assessment of taxonomic assignment accuracy using a mock metagenomic data set. The taxonomic profiles for PF and PE reported by both approaches were virtually identical however their distributions showed variation. The taxonomic inter-sample similarities were reflected in the gene content information as both approaches reported minor metabolic potential differences between PF and PE. Noticeably, both approaches reported significantly enriched biological processes involved in sugar transport and metabolism in PE.

Future direction

Future studies should focus on investigating role of fungal colonization in dental caries. The prevalence of antibiotic and antifungal resistance raises the need to explore the inter-microbial trans-kingdom interactions to treat those poly-microbial infections, or to prevent them from occurring in the first place. Microbiome studies focusing on the bacterial components are considered incomplete therefore should be referred to as bacteriome studies [4]. The fungal components of the oral microbiome have not been studied as extensively as its bacterial counterparts because of their rarity, difficultly to extract DNA from their cells, and most are uncultivable using current culturing techniques [59]. The advent of Next Generation Sequencing (NGS) technology has provided the ability to detect fungal species. Candida species are the most common in humans with a 150 species, most prevalently *C. albicans,* occupying different niches [60]. *C. albicans* is found in both hyphae and yeast form and can grow both aerobically and anaerobically allowing for diverse and complex inter-kingdom associations. *C. albicans* and bacteria associate with each other in antagonistic or synergistic manner [60, 61]. Examples of these polymicrobial interactions include adhesion and invasion, antimicrobial resistance, and quorum sensing [4, 59]. Synergistically, *C. albicans* has been shown to stimulate the growth of *S. mutans*, render *S. aureus* less susceptible to antibiotics, and release oxygen tension from *S. gordonii* allowing it to survive [59]. Antagonistically, *C. albicans* hyphae are killed by *P. aeruginosa* quorum sensing molecules[59].

Inter-kingdom interactions are important in the onset and progression of disease. Most importantly, they play a role in severe Early Childhood Caries (ECC). ECC a rampant form of dental caries is and is one of the most common infectious diseases affecting children worldwide [61]. The disease is initiated as a result of interaction between *C. albicans* and *S. mutans* on tooth surfaces in the presence of sucrose leading to biofilm formation [61]. *S. mutans* spp. lacks the ability to adhere to mucosal surfaces thus *C. albicans* acts as a bridging organism for attachment aiding *S. mutans* in evading salivary flow. Although acid production is considered the direct cause of enamel erosion, without biofilm formation acid production will not cause caries. *S. mutans-C. albicans* association is facilitated by the production of exopolysachrides (EPS), which are the building blocks of biofilms [62]. EPS are produced by bacterial exoenzymes [glucotransferases (Gtfs)] through fermentation of sugars [62]. The production of EPS boosts the ability of both organisms to colonize teeth. This renders the biofilm as a diffusion-limiting matrix, thus allowing the accumulation of additional acidogenic and aciduric bacteria, which in turn increases the production of acid[61] thus leading to dental caries.

References

1. Huang, X., et al., *A Highly Arginolytic Streptococcus Species That Potently Antagonizes Streptococcus mutans.* Appl Environ Microbiol, 2016. **82**(7): p. 2187-201.
2. Paster, B.J., et al., *The breadth of bacterial diversity in the human periodontal pocket and other oral sites.* Periodontol 2000, 2006. **42**: p. 80-7.
3. Hutter, G., et al., *Molecular analysis of bacteria in periodontitis: evaluation of clone libraries, novel phylotypes and putative pathogens.* Microbiology, 2003. **149**(Pt 1): p. 67-75.
4. Raja, M., A. Hannan, and K. Ali, *Association of oral candidal carriage with dental caries in children.* Caries Res, 2010. **44**(3): p. 272-6.
5. Farrell, J.J., et al., *Variations of oral microbiota are associated with pancreatic diseases including pancreatic cancer.* Gut, 2012. **61**(4): p. 582-8.
6. Filkins, L.M., et al., *Prevalence of streptococci and increased polymicrobial diversity associated with cystic fibrosis patient stability.* J Bacteriol, 2012. **194**(17): p. 4709-17.
7. Seymour, G.J., et al., *Relationship between periodontal infections and systemic disease.* Clin Microbiol Infect, 2007. **13 Suppl 4**: p. 3-10.
8. Simon-Soro, A., et al., *Microbial geography of the oral cavity.* J Dent Res, 2013. **92**(7): p. 616-21.
9. Duran-Pinedo, A.E. and J. Frias-Lopez, *Beyond microbial community composition: functional activities of the oral microbiome in health and disease.* Microbes Infect, 2015. **17**(7): p. 505-16.
10. Ghannoum, M.A., et al., *Characterization of the oral fungal microbiome (mycobiome) in healthy individuals.* PLoS Pathog, 2010. **6**(1): p. e1000713.
11. He, J., et al., *RNA-Seq Reveals Enhanced Sugar Metabolism in Streptococcus mutans Co-cultured with Candida albicans within Mixed-Species Biofilms.* Front Microbiol, 2017. **8**: p. 1036.
12. Whitmore, S.E. and R.J. Lamont, *Oral bacteria and cancer.* PLoS Pathog, 2014. **10**(3): p. e1003933.
13. Pitts, N.B., et al., *Dental caries.* Nat Rev Dis Primers, 2017. **3**: p. 17030.
14. *Oral health in America: a report of the Surgeon General.* J Calif Dent Assoc, 2000. **28**(9): p. 685-95.
15. Allison, D.L., et al., *Candida-Bacteria Interactions: Their Impact on Human Disease.* Microbiol Spectr, 2016. **4**(3).
16. Peterson, S.N., et al., *Functional expression of dental plaque microbiota.* Front Cell Infect Microbiol, 2014. **4**: p. 108.
17. Loesche, W.J., *Role of Streptococcus mutans in human dental decay.* Microbiol Rev, 1986. **50**(4): p. 353-80.
18. Forssten, S.D., M. Bjorklund, and A.C. Ouwehand, *Streptococcus mutans, caries and simulation models.* Nutrients, 2010. **2**(3): p. 290-8.

19. Stecksen-Blicks, C., *Lactobacilli and Streptococcus mutans in saliva, diet and caries increment in 8- and 13-year-old children.* Scand J Dent Res, 1987. **95**(1): p. 18-26.
20. Takahashi, N. and B. Nyvad, *The role of bacteria in the caries process: ecological perspectives.* J Dent Res, 2011. **90**(3): p. 294-303.
21. Marsh, P.D., *Dental plaque as a biofilm and a microbial community - implications for health and disease.* BMC Oral Health, 2006. **6 Suppl 1**: p. S14.
22. Gross, E.L., et al., *Beyond Streptococcus mutans: dental caries onset linked to multiple species by 16S rRNA community analysis.* PLoS One, 2012. **7**(10): p. e47722.
23. Wade, W.G., *The oral microbiome in health and disease.* Pharmacol Res, 2013. **69**(1): p. 137-43.
24. Huang, S.C., R.A. Burne, and Y.Y. Chen, *The pH-dependent expression of the urease operon in Streptococcus salivarius is mediated by CodY.* Appl Environ Microbiol, 2014. **80**(17): p. 5386-93.
25. Hwang, G., et al., *Candida albicans mannans mediate Streptococcus mutans exoenzyme GtfB binding to modulate cross-kingdom biofilm development in vivo.* PLoS Pathog, 2017. **13**(6): p. e1006407.
26. Richards, V.P., et al., *The microbiome of site-specific dental plaque of children with different caries status.* Infect Immun, 2017.
27. Johansson, I., et al., *The Microbiome in Populations with a Low and High Prevalence of Caries.* J Dent Res, 2016. **95**(1): p. 80-6.
28. Wu, Y.W. and Y. Ye, *A novel abundance-based algorithm for binning metagenomic sequences using l-tuples.* J Comput Biol, 2011. **18**(3): p. 523-34.
29. Wu, Y.W., et al., *MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm.* Microbiome, 2014. **2**: p. 26.
30. Dick, G.J., et al., *Community-wide analysis of microbial genome sequence signatures.* Genome Biol, 2009. **10**(8): p. R85.
31. Menzel, P., K.L. Ng, and A. Krogh, *Fast and sensitive taxonomic classification for metagenomics with Kaiju.* Nat Commun, 2016. **7**: p. 11257.
32. Coil, D., G. Jospin, and A.E. Darling, *A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data.* Bioinformatics, 2015. **31**(4): p. 587-9.
33. Seemann, T., *Prokka: rapid prokaryotic genome annotation.* Bioinformatics, 2014. **30**(14): p. 2068-9.
34. Bankevich, A., et al., *SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.* J Comput Biol, 2012. **19**(5): p. 455-77.
35. Gurevich, A., et al., *QUAST: quality assessment tool for genome assemblies.* Bioinformatics, 2013. **29**(8): p. 1072-5.
36. Jones, P., et al., *InterProScan 5: genome-scale protein function classification.* Bioinformatics, 2014. **30**(9): p. 1236-40.

37. Benjamini, Y.H., Yosef, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the Royal Statistical Society, Series B, 1995. **57**(1): p. 289-300.
38. Nurk, S., et al., *metaSPAdes: a new versatile metagenomic assembler.* Genome Res, 2017. **27**(5): p. 824-834.
39. Li, H. and R. Durbin, *Fast and accurate long-read alignment with Burrows-Wheeler transform.* Bioinformatics, 2010. **26**(5): p. 589-95.
40. Mende, D.R., et al., *Assessment of metagenomic assembly using simulated next generation sequencing data.* PLoS One, 2012. **7**(2): p. e31386.
41. Liu, B., et al., *Deep sequencing of the oral microbiome reveals signatures of periodontal disease.* PLoS One, 2012. **7**(6): p. e37919.
42. Keller, M.K., et al., *Oral microbial profiles of individuals with different levels of sugar intake.* J Oral Microbiol, 2017. **9**(1): p. 1355207.
43. Lif Holgerson, P., et al., *Maturation of Oral Microbiota in Children with or without Dental Caries.* PLoS One, 2015. **10**(5): p. e0128534.
44. Mark Welch, J.L., et al., *Biogeography of a human oral microbiome at the micron scale.* Proc Natl Acad Sci U S A, 2016. **113**(6): p. E791-800.
45. Brailsford, S.R., et al., *The predominant Actinomyces spp. isolated from infected dentin of active root caries lesions.* J Dent Res, 1999. **78**(9): p. 1525-34.
46. Dame-Teixeira, N., et al., *Actinomyces spp. gene expression in root caries lesions.* J Oral Microbiol, 2016. **8**: p. 32383.
47. Pignatelli, M. and A. Moya, *Evaluating the fidelity of de novo short read metagenomic assembly using simulated data.* PLoS One, 2011. **6**(5): p. e19984.
48. Charuvaka, A. and H. Rangwala, *Evaluation of short read metagenomic assembly.* BMC Genomics, 2011. **12 Suppl 2**: p. S8.
49. Peabody, M.A., et al., *Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities.* BMC Bioinformatics, 2015. **16**: p. 363.
50. Peterson, S.N., et al., *The dental plaque microbiome in health and disease.* PLoS One, 2013. **8**(3): p. e58487.
51. Kreth, J., Y. Zhang, and M.C. Herzberg, *Streptococcal antagonism in oral biofilms: Streptococcus sanguinis and Streptococcus gordonii interference with Streptococcus mutans.* J Bacteriol, 2008. **190**(13): p. 4632-40.
52. Wolfgang, W.J., et al., *Neisseria oralis sp. nov., isolated from healthy gingival plaque and clinical samples.* Int J Syst Evol Microbiol, 2013. **63**(Pt 4): p. 1323-8.
53. Aas, J.A., et al., *Defining the normal bacterial flora of the oral cavity.* J Clin Microbiol, 2005. **43**(11): p. 5721-32.
54. Eribe, E.R.K. and I. Olsen, *Leptotrichia species in human infections II.* J Oral Microbiol, 2017. **9**(1): p. 1368848.
55. Zhang, Y., et al., *Population-Genomic Insights into Variation in Prevotella intermedia and Prevotella nigrescens Isolates and Its Association with Periodontal Disease.* Front Cell Infect Microbiol, 2017. **7**: p. 409.

56. Schnee, A.E. and W.A. Petri, Jr., *Campylobacter jejuni and associated immune mechanisms: short-term effects and long-term implications for infants in low-income countries.* Curr Opin Infect Dis, 2017. **30**(3): p. 322-328.
57. Han, X.Y., J.J. Tarrand, and D.C. Rice, *Oral Campylobacter species involved in extraoral abscess: a report of three cases.* J Clin Microbiol, 2005. **43**(5): p. 2513-5.
58. Engen, S.A., et al., *The oral commensal Streptococcus mitis activates the aryl hydrocarbon receptor in human oral epithelial cells.* Int J Oral Sci, 2017. **9**(3): p. 145-150.
59. Baker, J.L., et al., *Ecology of the Oral Microbiome: Beyond Bacteria.* Trends Microbiol, 2017. **25**(5): p. 362-374.
60. Loesche, W.J., *Microbiology of Dental Decay and Periodontal Disease*, in *Medical Microbiology*, S. Baron, Editor. 1996, University of Texas Medical Branch at Galveston
The University of Texas Medical Branch at Galveston.: Galveston (TX).
61. de Carvalho, F.G., et al., *Presence of mutans streptococci and Candida spp. in dental plaque/dentine of carious teeth and early childhood caries.* Arch Oral Biol, 2006. **51**(11): p. 1024-8.
62. Kim, D., et al., *Candida albicans stimulates Streptococcus mutans microcolony development via cross-kingdom biofilm-derived metabolites.* Sci Rep, 2017. **7**: p. 41332.

Figures

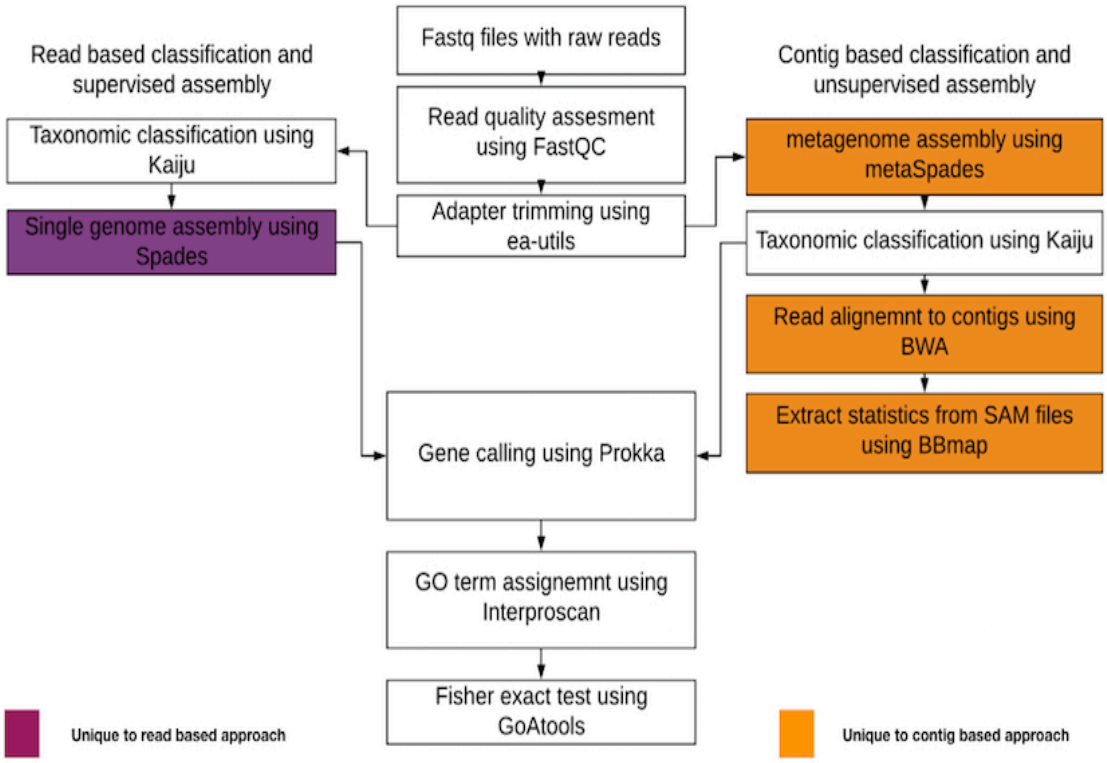


Figure 1: Workflow showing the steps and programs used in (i) read based taxonomic classification and supervised assembly and (ii) contig based classification and unsupervised assembly.

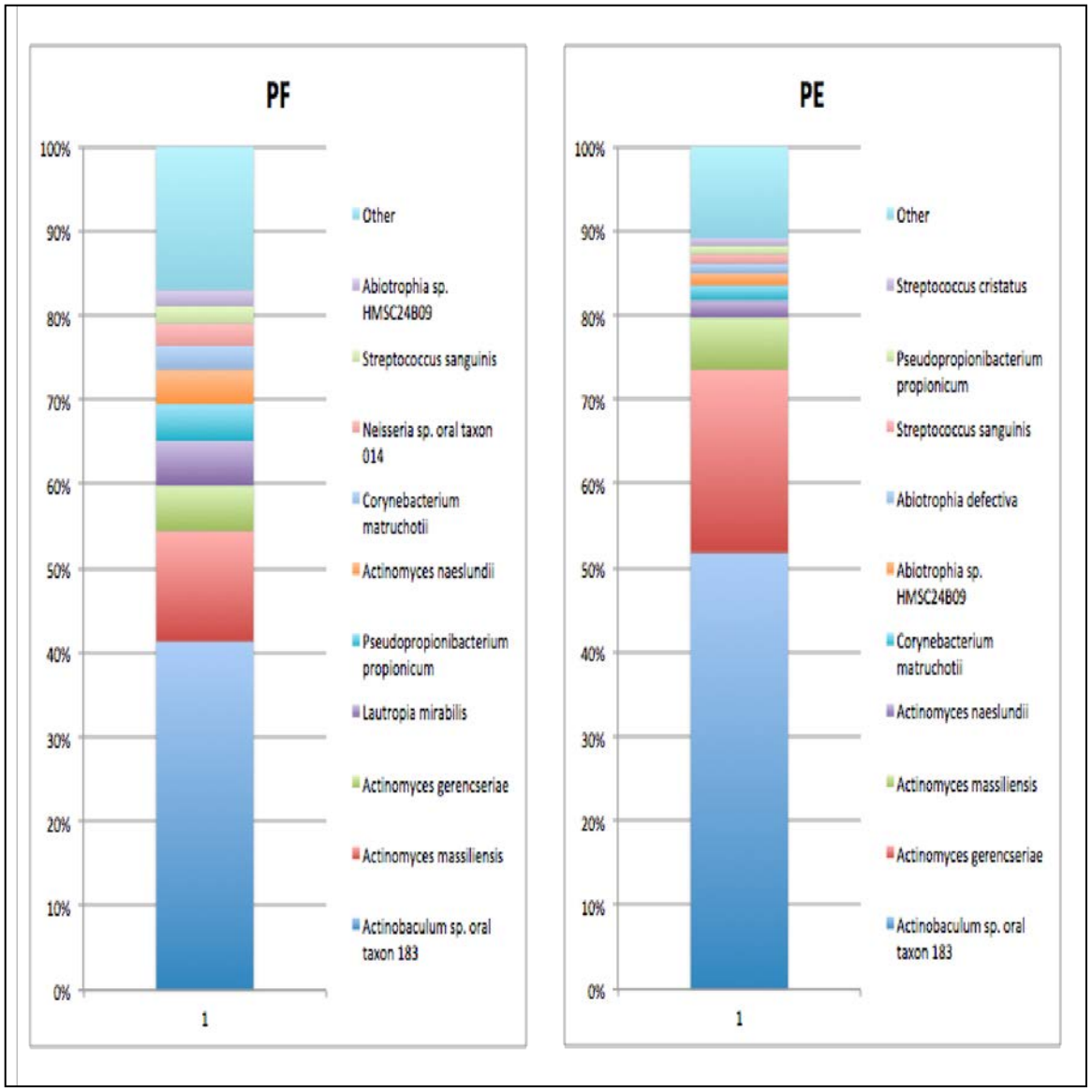


Figure 2: Results reported from read based taxonomic classification approach for the 10 most abundant species. The % relative abundance calculations are based on the most abundant 100 taxa in this approach.

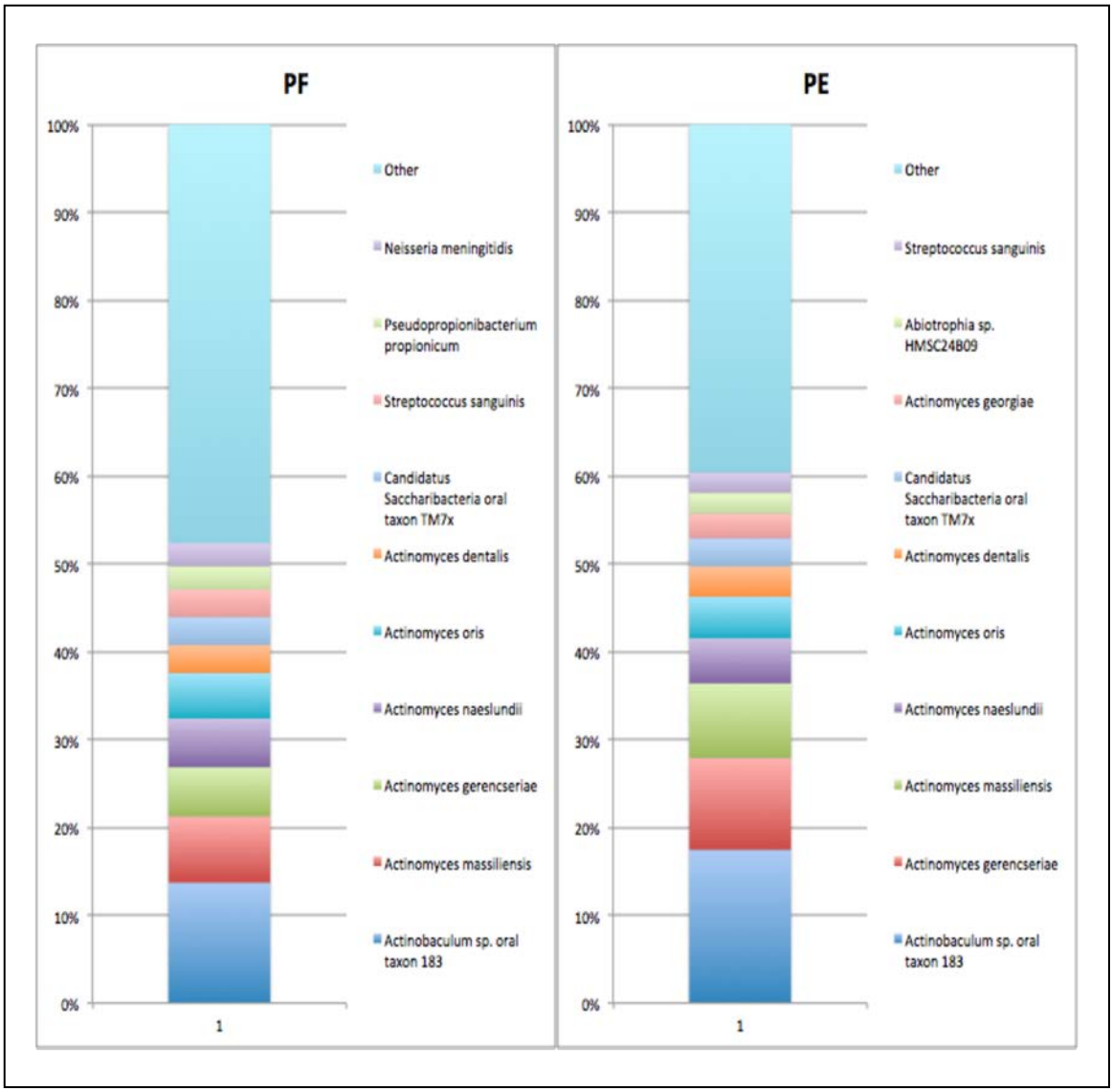


Figure 3: Results reported from contig based taxonomic classification approach for the 10 most abundant species. The % relative abundance calculations are based on the most abundant 48 taxa in PF and 42 taxa in PE

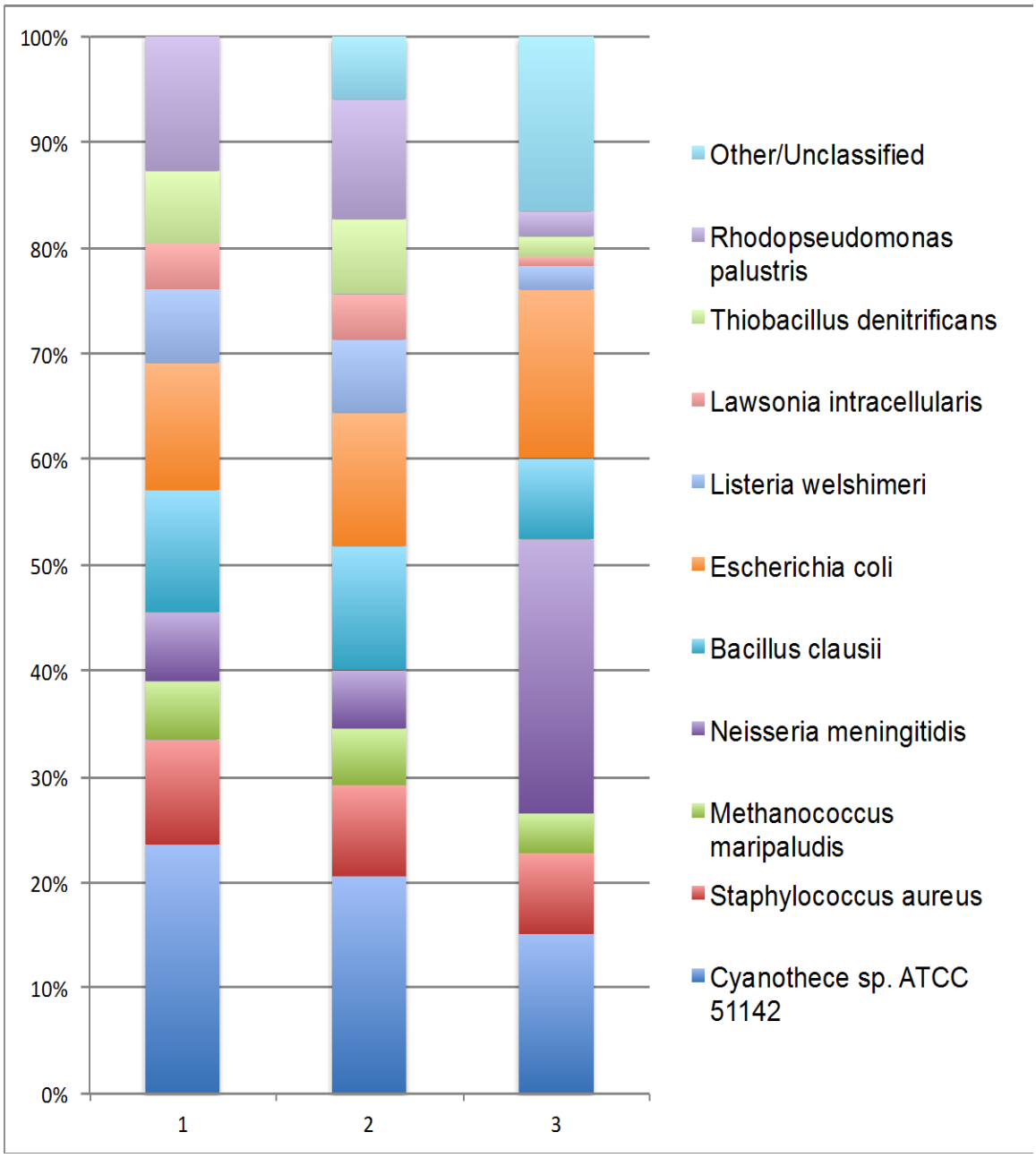


Figure 4: Results of the pipeline assessment step with column (1) representing the actual proportions of the mock community, column (2) representing the proportions reported using the read based taxonomic classification approach, and column (3) representing the results reported using the contig based taxonomic classification.

Tables

Table 1 Basic information of the results of the read based taxonomic classification approach

	PF	PE
Number of reads after trimming	125,326,716	166,763,119
% reads classified to species level	62.3	70
% reads failed to classify to species level	32	24.1
% reads unclassified	5.7	5.9
Number of taxa	13,919	13,733
Number of taxa assigned singletons	685	700
Number of taxa assigned <1000 reads	13,154	12,642
Number of taxa included in abundance analysis	100	100
Number of reads assigned to top 100 taxa	73,982,395	112,312,687
Number of reads assigned to the rest of the taxa	3,981,089	4,475,079

Table 2. Basic information of the results of the contig based taxonomic classification approach

	PF	PE
Number of contigs	69,175	57,242
N50	6800bp	8802bp
% contigsclassified to species level	65.9	61.2
% contigs failed to classify to species level	28.2	30.9
% contigs unclassified	5.9	7.9
Number of taxa	3,397	3,335
Number of taxa included in abundance analysis	48	42

Table 3: List of the 31 core shared taxa between PF and PE reported by both approaches and their % abundance and total proportion of the sample. Green signifies abundance increase/decrease shown consistently in both approaches while red signifies inconsistent abundance increase/decrease reported.

Core taxa	Read based PF % abundnace	Read based PE % abundance	Fold change	Contig based PF % abundance	Contig based PE % abundance	Fold change
Actinobaculum sp. oral taxon 183	41.36	51.81	1.25	13.72	17.46	1.27
Actinomyces massiliensis	13.10	6.16	0.47	7.53	8.49	1.13
Actinomyces gerencseriae	5.35	21.74	4.06	5.59	10.45	1.87
Lautropia mirabilis	5.29	0.57	0.11	2.36	1.07	0.45
Pseudopropionibacterium propionicum	4.40	0.99	0.23	2.67	1.74	0.65
Actinomyces naeslundii	4.11	2.06	0.50	5.58	5.14	0.92
Corynebacterium matruchotii	2.73	1.79	0.66	1.67	1.50	0.90
Neisseria sp. oral taxon 014	2.70	0.11	0.04	2.19	1.10	0.50
Streptococcus sanguinis	2.07	1.11	0.54	3.11	2.31	0.75
Abiotrophia sp. HMSC24B09	1.93	1.45	0.75	1.07	2.36	2.20
Actinomyces oris	1.70	0.83	0.49	5.17	4.72	0.91
Abiotrophia defectiva	1.31	1.12	0.86	2.06	1.76	0.85
Cardiobacterium hominis	1.29	0.64	0.50	1.74	1.46	0.84
Corynebacterium durum	1.01	0.81	0.80	1.05	1.29	1.22
Actinomyces dentalis	0.65	0.16	0.25	3.20	3.50	1.09
Ottowia sp. oral taxon 894	0.64	0.10	0.16	0.66	0.92	1.41
Aggregatibacter aphrophilus	0.53	0.22	0.41	1.28	1.05	0.82
Kingella denitrificans	0.43	0.05	0.12	1.95	1.21	0.62
Cardiobacterium valvarum	0.43	0.10	0.23	1.31	1.03	0.78
Neisseria meningitidis	0.34	0.06	0.18	2.64	1.24	0.47
Streptococcus cristatus	0.34	0.92	2.73	2.11	1.33	0.63
Selenomonas noxia	0.22	0.05	0.24	1.20	1.19	0.99
Candidatus Saccharibacteria oral taxon TM7x	0.21	0.11	0.54	3.17	3.17	1.00
Rothia dentocariosa	0.20	0.23	1.19	0.69	0.63	0.91
Streptococcus oralis	0.18	0.20	1.08	1.56	1.66	1.07
Actinomyces georgiae	0.18	0.19	1.06	2.23	2.80	1.26
Capnocytophaga sputigena	0.12	0.04	0.30	0.77	0.77	1.01
Aggregatibacter sp. oral taxon 458	0.11	0.12	1.06	1.42	1.22	0.86
Veillonella parvula	0.10	0.72	7.30	1.03	1.74	1.69
Capnocytophaga gingivalis	0.07	0.03	0.45	0.82	0.85	1.03
Streptococcus intermedius	0.07	0.35	5.18	1.46	1.65	1.13
Total sample abundance proportion	**93.14**	**94.85**		**83.01**	**86.80**	

Table 4. Significantly enriched GO terms in both approaches using FDR corrected p-values. The domain (BP) denotes a Biological Process, (MF) denotes a Molecular function, and (CC) denotes a Cellular Component. Bolded signifies notable terms.

Read based approach (supervised assembly)				Contig based approach (unsupervised assembly)			
GO ID	**FDR corrected p-value**	**Domain**	**Description**	**GO ID**	**FDR corrected p-value**	**Domain**	**Description**
GO:0044765	0.002	BP	Single organism transport	GO:0050485	0.004	MF	Oxidoreductase activity, acting on X-H and Y-H to form an X-Y bondwith a disulfide as acceptor
GO:1902578	0.002	BP	Single organism localization	GO:0046992	0.004	MF	Oxidoreductase activity, acting on X-H and Y-H to form an X-Y bond
GO:0005575	0.002	CC	Cellular component	**GO:0009401**	**0.034**	**BP**	**Phosphoenolpyruvate-dependent sugar phosphotransferase system**
GO:0051234	0.002	BP	Establishment of localization				
GO:0006810	**0.002**	**BP**	**Transport**				
GO:0051179	0.002	BP	Localization				
GO:0008150	0.002	BP	Biological process				
GO:0071702	0.008	BP	Organic substance transport				
GO:0005215	**0.01**	**MF**	**Transporter activity**				
GO:0022857	**0.012**	**MF**	**Transmembrane transporter activity**				
GO:0003674	0.018	MF	Molecular function				
GO:0065007	0.022	BP	Biological regulation				
GO:0044699	0.038	BP	Single organism process				
GO:0008643	**0.04**	**BP**	**Carbohydrate transport**				
GO:0009401	**0.048**	**BP**	**Phosphoenolpyruvate-dependent sugar phosphotransferase system**				

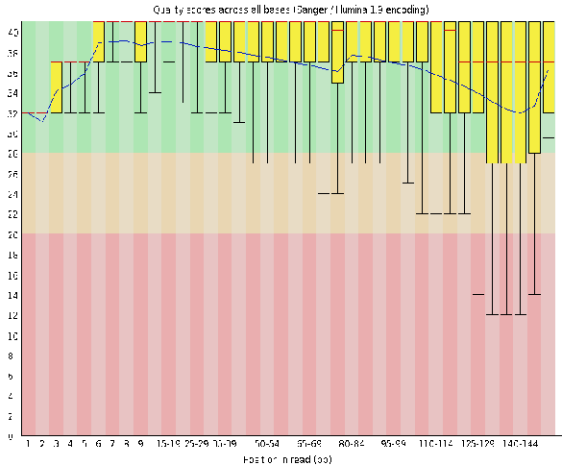
Appendix A

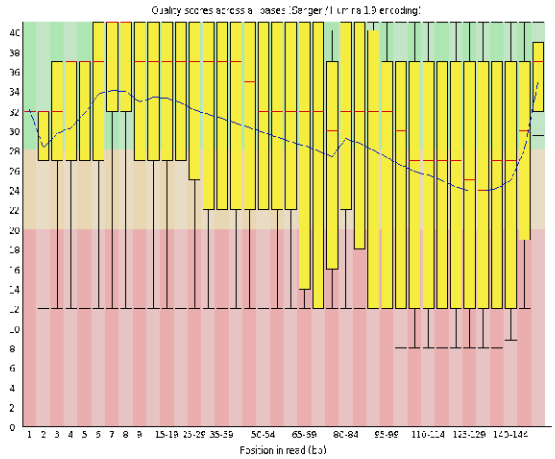
Molysis kit assessment

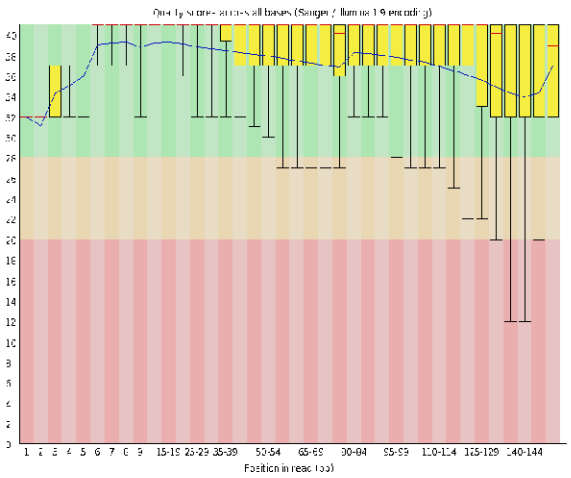
qPCR results of Molysis kit assessment. Green represents human DNA and blue represents bacterial DNA. A and C represent DNA extracted from plaque samples with out treatment with Molysis. B and D represent the same samples DNA after treatment with Molysis treatment. The results show a decrease in human DNA (Green) contamination to less than 1% of the total sample.

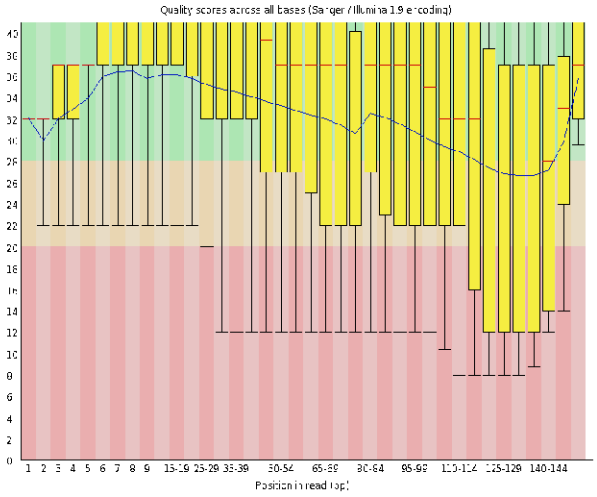
Appendix B

Sequencing evaluation









FastQC results. Top row represents PF (forward reads on the left and reverse reads on the right) and bottom row represents PE (forward reads on the left and reverse reads on the right).