2023

Comparison of Taxonomy Assignment and Strain Detection in Urobiome Communities Using 16S rRNA Sequencing and Shotgun Metagenomic Sequencing

Delaney Sauer

LOYOLA UNIVERSITY CHICAGO

COMPARISON OF TAXONOMY ASSIGNMENT AND STRAIN DETECTION IN UROBIOME COMMUNITIES USING 16S rRNA SEQUENCING AND SHOTGUN METAGENOMIC SEQUENCING

A THESIS SUBMITTED TO

THE FACULTY OF THE GRADUATE SCHOOL

IN CANDIDACY FOR THE DEGREE OF

MASTER OF SCIENCE

PROGRAM OF BIOINFORMATICS

BY

DELANEY SAUER

CHICAGO, IL

MAY 2023

For my dad, who may not know much about UTIs,
but was integral to the completion of this project

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Advancements in sequencing technologies have enabled scientists to gain insight into the microbes that inhabit the human body, including the urinary tract. Cataloging the bacteria that inhabit the urinary tract has primarily relied on amplification and sequencing of specific variable regions of the 16S rRNA gene enabling genus-level taxonomic identification. Recently, shotgun metagenomic sequencing has been employed such that bacterial taxonomy as well as the functionality that they encode can be inferred. In this study, we compare taxonomies assigned by 16S sequencing and shotgun metagenomic sequencing of the urinary microbiota (urobiome) of females with and without a clinical diagnosis of a urinary tract infection (UTI). Rather than target specific variable regions of the 16S rRNA gene, we employed long-read sequencing technology which captures all nine variable regions such that species-level taxonomic assignments can be made. First, we characterize the bacterial constituents of the urobiomes of the two cohorts from the full-length 16S sequence. To assess the power of full-length rather than single variable regions, we computationally derived short-reads and compared these predictions to our full-length analyses. Next, we compared the results of the taxonomic predictions from full-length 16S to those of shotgun metagenomic sequencing of a subset of our samples. We found that long-read sequencing created more accurate taxonomic classifications than shotgun sequencing and single variable regions. Both sequencing approaches suggest that multiple strains of species colonize the urobiome.

CHAPTER ONE

INTRODUCTION

Introduction to the Urinary Microbiome

Microbiome research has rapidly expanded over the last ten years, and for good reason: lower cost of sequencing, more tools for analysis, and the keen public's interest provide scientists means for microbial investigation. Microbial research can take place anywhere there is a community of bacteria, whether that be in soil, water, or animals (1). The microbiota (microbial community of a particular niche) can affect the host's state and, in the case of humans, could possibly lead to disease or lack thereof. Because the scientific community can confirm that the microbiota can directly correlate with health, the human microbiota is a rapidly expanding area of microbial research (2–5). We also know that different bacteria inhabit different areas of the human body, which leads to the separate microbial analysis of different sites such as the gut, the mouth, the skin, or even the eye (3,6–8).

The urinary tract microbiota is a low biomass environment. For many years, it was assumed that the urinary tract of healthy humans was sterile because bacteria from urine could not be cultured under standard laboratory conditions, but with Expanded Quantitative Urine Culture (EQUC) procedure, Hilt et al. were able to confirm that bacteria did live in the bladder of asymptomatic ("healthy") female participants (9). This study used transurethral catheterized urine, avoiding any contamination that voided urine might incur; prior analysis of urine collection methods found that transurethral catheterized urine was representative of urine from the bladder (10). The EQUC method isolated several species from asymptomatic females,

including *Lactobacillus gasseri, Corynebacterium coyleae, Streptococcus anginosus, Actinomyces neuii, Staphylococcus epidermidis,* and more (9). High-throughput 16S rRNA gene sequencing of bladder urine samples identified the same species as the EQUC procedure, in addition to fastidious bacterial species that are notoriously difficult to culture (9). The studies of Wolfe et al. and Hilt et al. opened the door for more research into the urinary microbiome or urobiome (the genomic content of the urinary microbiota) (9,10). Subsequent studies confirmed that the bladders of both symptomatic and asymptomatic individuals contain bacteria (11–14).

Urinary Tract Infections and Antibiotic Resistance

Acute urinary tract infections (UTI) are the most treated outpatient infection and healthcare-associated bacterial infection, and they increase in frequency and severity with age (15). The stigma surrounding UTIs can cause emotional distress and ultimately decrease quality of life. When a UTI is contracted, common symptoms include painful and frequent urination, difficulty beginning urine stream, and blood in the urine (16). Relieving these symptoms and/or seeking treatment is time consuming and costly for the patient. In cases where the UTI was developed in the hospital, it often leads to a price increase of a thousand dollars or more from their previous bill ($876 for tests and medications, and up to $10,197 for increased ICU stays) (17). Even after preventative steps and treatments have been taken, it is very common (25%-30%) for UTIs to reoccur, exponentially increasing the cost and burden on the patient (16). While UTIs can sometimes resolve themselves on their own, many patients need treatment involving antibiotics (18). Treatments can depend on whether the infection is considered uncomplicated or complicated, or if the bacteria causing the infection is antibiotic resistant, which is rising in prevalence (18). UTIs are considered complicated if the patient has underlying health conditions that may affect treatment such as age, pregnancy, or diabetes. At first,

treatment resistance was seen mainly in complicated UTIs, but now it can be found in many uncomplicated cases (19). UTIs caused by antibiotic resistant bacteria are now considered a public health threat (20). Despite this rise of antibiotic resistance in UTI treatment, developing new drugs to combat infections is not increasing (19).

The most frequent (75%-95% of uncomplicated urinary tract infections) bacterial agent of acute UTIs is uropathogenic *Escherichia coli* (UPEC), although several other species of bacteria have been shown to cause UTIs as well (4,21,22). However, research has not been able to identify the specific mechanisms that lead to *E. coli* colonization and subsequent infection, as *E. coli* can be present in the bladder of asymptomatic individuals as well (11,21–23). This suggests that UTIs are disorders involving "multiple variables" (4). Until the discovery of a distinct urinary microbiota, the prevailing hypothesis was that *E. coli* strains from the gut are introduced to and colonize the urinary tract leading to acute UTIs (21,23). Antibiotic resistance has been observed in ESBL-producing UPEC strains that can be common in hospital/long term care settings (12). These strains can inactivate antibiotics and convey multidrug resistance via plasmids (24). Another species that has been shown to cause UTIs is *Klebsiella pneumoniae*, which is associated more with healthcare-related infection than uncomplicated UTIs (25). Identification of infections caused by *Klebsiella* is an important step to stop spread of infection, because *Klebsiella* strains that produce Carbapenemase (a β-lactamase) can be deadly (26). Other uropathogenic species include *Proteus mirabilis*, *Pseudomonas aeruginosa*, and *Enterococcus faecalis* (22).

Surveys of the female urobiome have associated several bacterial species with the lack of lower urinary tract symptoms, including species of *Lactobacillus*, *Gardnerella*, and *Staphylococcus* (27). Some members of the urobiome can provide protection from UPEC and

other uropathogens, For instance, *Lactobacillus* strains have been shown to inhibit growth of uropathogenic strains *E. coli, K. pneumoniae*, and *E. faecalis* (28). The Thomas-White dataset, which aimed to characterize the asymptomatic urinary microbiome, found species within the *Staphylococcus, Lactobacillus, Streptococcus, Gardnerella, Bifidobacterium,* and *Bacillus* genera as well as finding *Escherichia* strains (23).

Discovery of Microbial Taxonomy with 16S rRNA Sequencing Data

Complex microbial communities have shaped life on Earth, and recent developments in high-throughput sequencing technology make investigating these communities feasible. Metagenomics is the study of sequences derived from many organisms in a complex community, including the human microbiota. This captures bacteria that cannot be cultured in the lab, e.g., fastidious bacteria of the urinary tract. Currently, there are two main approaches for sequencing microbial communities: 16S rRNA gene sequencing and shotgun metagenomic sequencing.

16S rRNA gene sequencing has been used to investigate microbial taxonomy for almost half a century. At the foundation of this technology is the fact that all prokaryotes (bacteria and archaea) encode for the16S rRNA gene. In the late 1970s and into the 1980s, Carl Woese investigated the 16S rRNA gene sequence and its ability to assign taxonomy (29). The 16S rRNA gene sequence is composed of 9 regions of variation (referred to as variable regions and denoted by a capital V) dispersed between regions of conservation. These variable regions can be translated to the taxonomy of a given prokaryote.

The general steps for 16S rRNA gene sequencing include obtaining and extracting DNA from the sample, performing polymerase chain reaction (PCR) amplification of the 16S rRNA gene sequences in the sample, and sequencing the PCR amplicons (Figure 1 A-C). This is followed by subsequent computational analysis of the sequencing reads produced. The advent of

high-throughput sequencing technologies enabled researchers to sequence all 16S rRNA PCR amplicons from the community in a cost- and time-effective manner. High-throughput sequencing technologies are limited in the length of the reads it can produce, with current short-read sequencing platforms limited to reads of 300 bp (and 600 bp when using the paired-end format). Thus, short-read technologies such as Illumina can only sequence specific regions of the 16S rRNA gene sequence (for example, the V1-V3 regions). Prior research has found that different regions are better at determining taxonomy than others for certain microbiomes, as discussed in subsequent chapters (30,31).

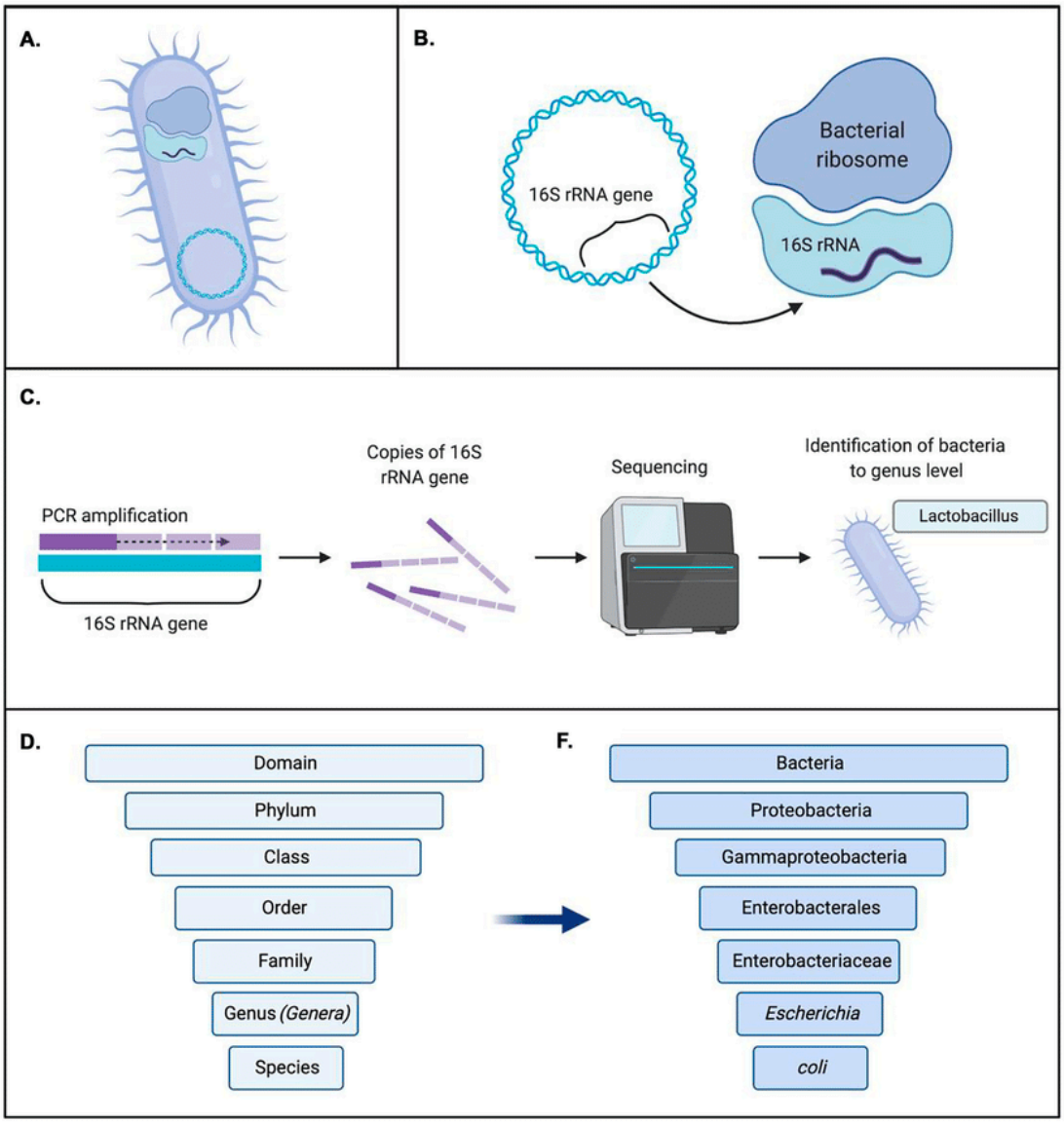

Figure 1. The general schema of 16S rRNA gene sequencing. The 16S gene is amplified, sequenced, and identified **(**Image from (32). A represents a microbe, B shows the 16S rRNA gene codes for the 16S rRNA subunit, C shows the 16S gene being amplified into many copies, and those copies being sequencing and identified. D and F shows the order of phylogenetic classification to the species level.

After sequencing, taxonomy is assigned to the reads (Figure 1 C, D, F). For the last ten to fifteen years, reads from high-throughput 16S rRNA gene surveys were clustered into Operational Taxonomic Units (OTUs). An OTU is not a specific species or strain, rather it is a consensus sequence from a cluster calculated from sequence similarity. Commonly, the OTU

approach groups sequences greater or equal to 97% similarity together, signifying a single taxonomic unit. Sequences from strains of the same genus are expected to have at least 95% similarity and those from the same strain are 99% (33). Although OTUs was an important step in 16S data analysis because it made the process of taxonomic assignment a computationally lighter task by clustering millions of sequences into thousands of groups, it is not sensitive enough to classify the depth of data we can sequence now (34).

The problem with OTUs is the concept of sequence similarity because it does not account for evolutionary distances (34). The 97% species-level classification is an estimation, and it has been proven to be too high *and* too low for some species of bacteria. For instance, some species of bacteria have a 99% similarity and other species (like *E. coli*) can have closer to 5% variance within several copies of the 16S gene (34). In these cases, some bacteria would be combined into one consensus sequence, erasing species-level diversity, and other species would be split into multiple clusters, which would represent false species-level diversity. Because of the clustering and similarity thresholds summarizing sequences instead of deciphering them, it would be unreasonable to conclude species-level diversity from OTUs.

The ability to analyze 16S sequencing data entered a new era with the concept of Amplicon Sequence Variants (ASVs). The concept of ASVs was introduced in 2013, and they attempt to denoise sequences without removing any of the actual biological diversity that can be lost during clustering (35,36). Where OTUs used clustering to attempt to remove sequencing errors, ASV-creating pipelines remove errors through machine learning algorithms. Thus, ASVs should represent actual species, and possibly strains.

The software tool DADA2 implements this ASV strategy (37). In their paper, Callahan et al. claimed that DADA2 can detect variation down to a single nucleotide (37). They provided

benchmarks for DADA2, comparing it to the frequently used 16S rRNA sequence analysis tools UPARSE (OTU clustering), MED (OTU clustering using entropy scores), mothur (OTU clustering using the average linkage algorithm), and QIIME using UCLUST (OTU clustering). DADA2 was made specifically for data generated by the Illumina short-read sequencing platform (the prevalent technology used for 16S sequencing surveys), by basing the denoising algorithm on Illumina self-reported errors. Other steps in the DADA2 core algorithm include pairwise alignments of the sequences via the Needleman-Wunsch algorithm, an error model creation that considers errors within and between reads, an abundance p-value that indicates if the sequence was likely to be created by errors or is present in the sample, and then the divisive partitioning algorithm where all sequences are compared to each other using the error rates and abundance p-values created previously (37). Callahan et al. further benchmarked DADA2 by examining previously annotated human vaginal samples and mouse gut samples and found that DADA2 produced more accurate taxonomic classifications and less incorrect classifications than methods based on OTUs clustering at the 97% similarity threshold (Callahan et al., 2016). Figure 2 included in the DADA2 GitHub repository shows the accuracy obtained by DADA2 relative to OTU clustering at 97% (via mothur) with simulated data (https://benjjneb.github.io/dada2/).

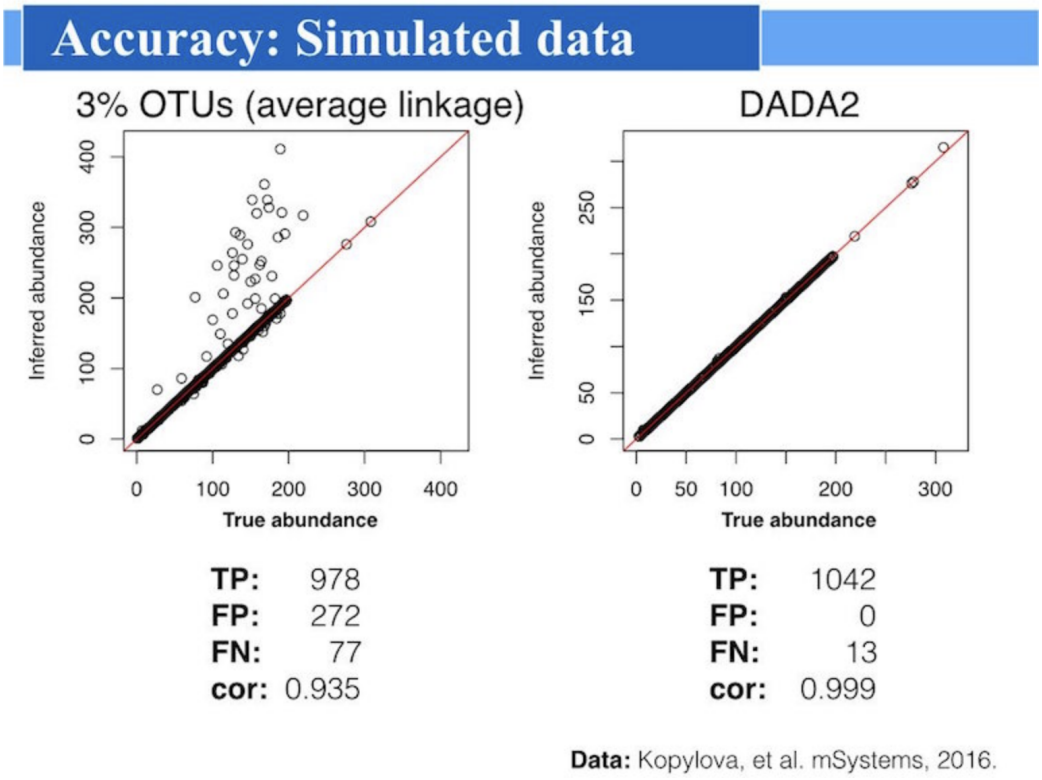

Figure 2**.** Error rates of taxonomy assigned compared between 3% average-linkage (97% similarity) OTUs (the mothur taxonomic identification scheme) and DADA2, showing that DADA2 had a higher correlation to the true classifications. (Image from https://benjjneb.github.io/dada2/.)

More recently, the invention of long-read sequencing provides the ability to sequence all nine of the variable regions of the ~1500 bp 16S rRNA gene sequence. In general, most long-read sequencing can generate reads 1000s of nucleotides long, which is significantly longer than the 300 bp length currently possible with Illumina sequencers (38). Currently, there are two long-read sequencing platforms: PacBio SMRT (Single Molecule Real-Time) platform and ONT (Oxford Nanopore). The SMRT technology was created in 2011 and bought by Pacific Biosciences (PacBio) (39). Previously, the largest drawback to long-read sequencing was high error rates, but the PacBio system has reduced errors to 0.05%, compared to 11-15% from 2015 (38,40). PacBio outperforms ONT regarding time (40) and read length (41). Furthermore, ONT

has higher error rates than PacBio sequencing, but that gap is closing as both technologies advance (41). The process of PacBio sequencing is shown in Figure 3.

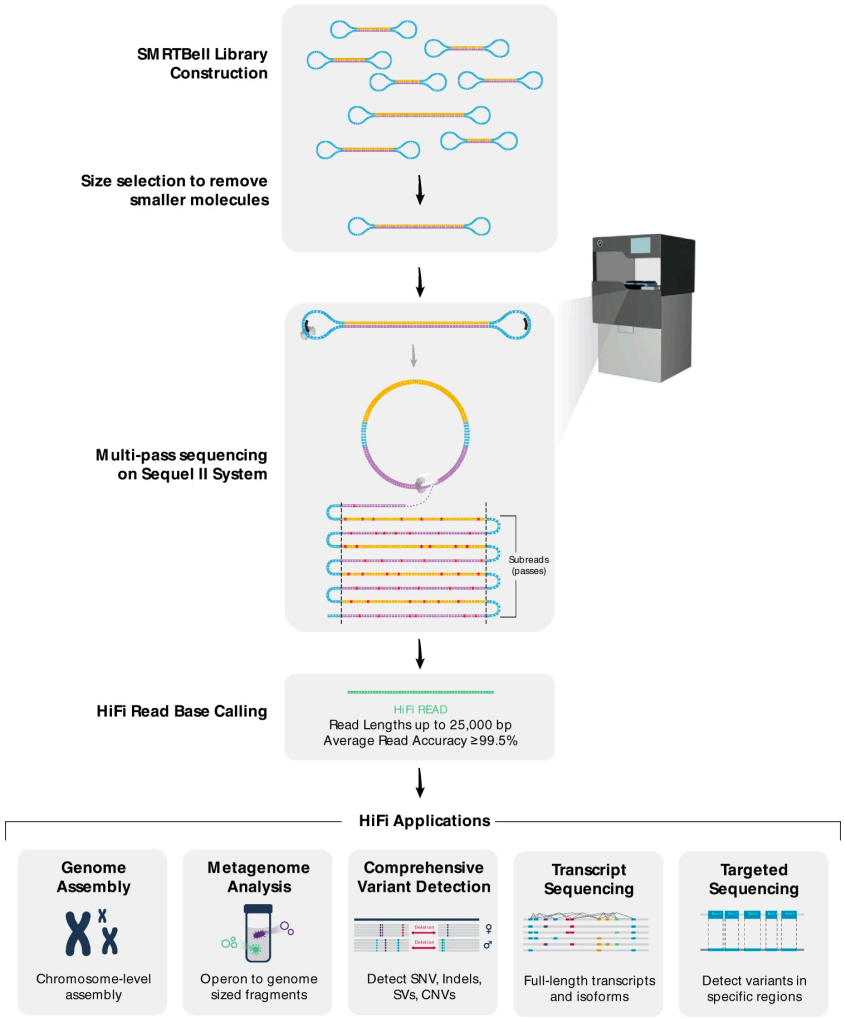

Figure 3. The process of PacBio long-read sequencing using the SMRTBell Library Construction. (Image from 38.)

Central to the success of taxonomic classification of 16S survey data is the choice of database used to assign taxonomy to the sequences. Popular databases used to assign taxonomy, specifically to 16S rRNA gene sequences, include NCBI, SILVA, RDP (Ribosomal Database Project), and GreenGenes. The NCBI 16S database is manually curated and composed of all

organisms submitted to the NCBI database (42). It has been shown to be successful at assigning 16S taxonomy (30,42). There are externally created databases (labeled as contributed databases) that use the RefSeq database from the NCBI database that are formatted for use with DADA2, but those databases are not maintained by the DADA2 team (https://zenodo.org/record/4735821#.Y35IbezMLvU). SILVA, RDP, and GreenGenes are officially maintained with formatting specific for use with DADA2. While GreenGenes is an important database in the context of the history of 16S survey studies, it has not updated its taxonomy since 2013 (42). At the time of writing, that was almost ten years ago. Furthermore, in more recent benchmarking studies, GreenGenes classifications are not as accurate as other database options (43,44). Although the RDP database was updated more recently than GreenGenes (14 August 2020), it does not outperform GreenGenes (45). In a 2018 study, the RDP database miscalled 5% of genus classifications when used with mothur, and Greengenes miscalled 3.4% of genera when used with QIIME2 (note, these databases were not compared directly against each other because they are formatted for different tools) (43). SILVA is the largest database of the three maintained for use with DADA2, and it shares more taxonomic assignments with NCBI than the other two databases (42). The current SILVA release v138.1 has over 9.4 million sequences to compare for taxonomic assignment for ASV sequences (46). The SILVA v138 release also includes species level classifications, whereas RDP does not, and the GreenGenes species level classifications would probably be incorrectly assigned due to the lack of updates to the database.

Taxonomic Classification using Shotgun Metagenomic Sequencing

Shotgun metagenomic sequencing is another sequencing strategy used to explore microbial communities. Shotgun sequencing gets its name from the general methodology of

capturing the entire genome: the genome is broken into small fragments, sequenced, and then reconstructed computationally. In contrast to 16S sequencing, shotgun sequencing can be used to sequence any organism, including eukaryotes and viruses which do not contain the 16S rRNA gene. Shotgun sequencing provides chromosomal and extrachromosomal information from a sample, so more than just taxonomic information can be inferred (47). It has been proven previously that shotgun sequencing can detect species that 16S rRNA sequencing cannot identify (48). Because more data can be extracted from the sample, species level classifications can be made and observing strain level variation is possible (49).

The concept for shotgun sequencing was proposed at a similar time as 16S rRNA sequencing; it was first proposed in 1979 by Rodger Staden (39). The concept of the approach was to use bacterial vectors to clone sections of a bacteria's DNA, and computationally combine those cloned vectors to create a scaffold of contigs, and further assemble those scaffolds into a completed genome (50). An important distinction to make here is that Staden proposed this method of breaking down and rebuilding within the context of the singular genome of which the scientist already knew the taxonomic classification. The introduction of this technology led to a boom of full-genome phage and bacteria sequences being created in the 1980s (39). Shotgun metagenomic sequencing was another advancement that happened in the 2010s: instead of breaking up and reconstructing a singular genome, the approach was applied to entire bacterial communities. Each bacterial cell's DNA is fragmented and sequenced, then computationally reconstructed into genomes, and it falls on the computational algorithm to determine the taxonomies represented by the DNA sequences.

Taxonomy is assigned to metagenomic assembled genome sequences (MAGs) or bins of sequences (assembled or raw reads) likely to belong to the same genome or strain via database

comparisons. Before the creation of shotgun metagenomics-specific taxonomy software, the best way to assign taxonomy to metagenomic assemblies was by comparison to the NCBI database via the BLAST algorithm (51). As the throughput of sequencing technologies increased, use of BLAST became infeasible. As a result, several tools were developed. For example, the software tool Kraken was introduced in 2014 as a k-mer based approach to taxonomy assignment (51). Kraken relied on exact-match k-mers instead of attempting to find exact matches in sequences to assign taxonomy from a database, which shortened the length of time for alignment and lightened the computational load (51). In 2019, Kraken2 was released, and the biggest change to the tool was a reduction of memory usage of 85% (52). When compared to 16S taxonomic assignment algorithms, Kraken2 outperformed QIIME2's classification scheme in "computational requirements, runtime, and accuracy" (53).

Preliminary Work in Strain-Level Classifications

In general, strain-level diversity of bacteria in human microbiomes, whether that be in the urinary tract, gut, skin, or anywhere, could be the difference between a symptomatic or an asymptomatic disease state. For instance, specific strains in the urinary tract can be connected to pathogenicity and antibiotic resistance (18). There have been attempts to extract species and strain diversity from long-read 16S data, and the preliminary results are promising. A 2019 paper by Johnson et al found that PacBio sequencing of the V1-V9 regions could resolve strain differences in sequences taken from the GreenGenes database, and provided more specificity than 16S sequences that covered the V1-V3 regions (33). The strains were determined to be strains of the same species if they had 99% similarity with another sequence (33). It should be noted, though, that they found no differences in the taxonomic classification at the genus level for 16S sequencing V1-V3 regions, V1-V9 regions, and shotgun metagenomic sequencing,

suggesting that previous variable region 16S sequencing was adequate for taxonomic assignment if species-level classification was not necessary (33). The Johnson study used OTUs as the taxonomic identification strategy, and they claim that a 99% similarity OTU could resolve strain level variation and 97% could classify genera level (33). They conclude their paper by saying that long-read sequencing (specifically PacBio) could be a valuable tool in investigating strain level variation, but it is hindered by its high error rates and the best way to continue research would be to develop computational tools that can decipher the actual biological data through the errors. It should also be noted that this study focused more on the issue that polymorphisms of the 16S gene present in short-read sequencing and did not focus on a particular organism or microbiome.

There are several tools that can be used to extract strain level variation from shotgun metagenomic sequencing, although none have truly outperformed others thus far. There are tools (MEGAN and MetaMaps) that utilize reference genomes to compare experimental genomes to find differing strains, similarly to taxonomy assignment in 16S rRNA gene sequencing (54,55). This approach is not as successful as others because many species and strains found in shotgun data have no reference to compare to (a recent meta-analysis study found about 77% of strains found across 150,000 genomes did not have a reference) (56). The tools that attempt to resolve down to the strain level are, as Anyansi et al write in a paper comparing current computational tools, “still in their infancy” (57). In this way, it seems as if the computational tools are playing catch-up to the extreme wealth of information we can gain from utilizing shotgun metagenomic sequencing. Alternatively, a de novo assembly tool, which does not depend on reference genome sequences, must be used. STRONG is a pipeline that constructs bins that represent species found in a sample, and then differentiates strains within the bins or MAGs (58). STRONG also utilizes

DESMAN (De Novo Extraction of Strains from Metagenomes) to further refine the bins/MAGs. When comparing the two tools, DESMAN identifies less strains than STRONG, which Quince et al use to indicate that their graph-based assembly approach is more effective than the DESMAN read-based approach alone (58).

Scope of Thesis

In this thesis, I examine the urinary microbiome with PacBio long-read data, simulated short-read variable region data, and shotgun metagenomic data. I also examine if strain-level diversity can be found within the samples. We sequenced the urobiomes of 14 females without a clinical UTI diagnosis (UTI negative), the urobiomes of 33 females with a clinical UTI diagnosis (UTI positive) as well as 2 mock community samples. The mock communities are our positive controls and helped gauge the accuracy of our methods. In Chapter 2, I investigate the urinary microbiome with PacBio long-read 16S rRNA gene sequencing. I examine the differences between the taxonomies found in the UTI positive and UTI negative samples. I also determine if strains can be detected using long-read sequencing. In Chapter 3, I computationally parse variable regions of the long-read 16S data to compare taxonomy found from variable regions and full 16S sequences. In Chapter 4, I use shotgun sequencing to assign taxonomy and resolve strains to UTI positive and UTI negative samples. These investigations of the urobiome will characterize the community and provide insight into taxonomy assignment and resolving strains with several types of microbial sequencing.

CHAPTER TWO

LONG-READ SEQUENCING OF THE URINARY MICROBIOME

Introduction

Taxonomic characterization of urobiomes is largely based on short-read amplification of variable regions of the 16S rRNA gene. Short-read surveys have profiled the urobiome of males and females, and more specifically continent females, females with urinary incontinence, and individuals with UTIs (10,14,23,27,30,59). These studies have created a general description of the commonly found phyla and genera in the urobiome, but given the limitations of short-read sequencing, they are unable to resolve constituents to the species level. Until the invention of long-read sequencing, species-level resolution has only been possible with culture-based studies (14,60).

Examining species-level diversity is important because many of the genera that are commonly found in the urobiome include species classified as uropathogens as well as nonpathogenic "commensal" members. For example, *Staphylococcus*, which is commonly identified in the urobiomes of females with and without lower urinary tract symptoms, includes the "commensal" *S. epidermidis* as well as the UTI-associated species *S. aureus*, which is considered an emerging cause of UTIs in some patient populations (14,60–63). *Lactobacillus* is another example of this phenomenon: within the genus, it has been shown that different species can have different associations with urinary symptoms. Culture-based characterization of urobiomes of women with/without UUI (urge urinary incontinence) found that *L. gasseri* is

detected more frequently in the UUI cohort, while *L. crispatus* is more frequently detected in the control samples (14,60)

L. crispatus can inhibit or kill uropathogens and has been associated with a healthy female urobiome, and clinical *L. gasseri* strains also have been found to kill urogenital pathogens, despite being found more frequently in the urobiome of females with UUI symptoms (64,65). A clinical *L. jensenii* strain was also found to be bactericidal for *E. coli* (66). Furthermore, the presence or absence of these species is not the only factor in possible commensal effects as variation in *Lactobacillus* inhibition strength has been detected both between species and even strains of the same species (28).

While prior sequence-based investigations of the urobiome have been limited to short-reads, primarily a residual of technological limitations, high-throughput full-length 16S rRNA gene sequencing surveys have now become feasible. Full-length sequences provide greater resolution, providing species-level and potentially strain-level resolution of bacteria within complex communities (33,67). At the onset of this project, this technology had yet to be applied to profiling the urobiome. In November 2022, the first report of long-read 16S sequencing of urinary samples was published (68). This study compared culture-based analysis and long-read 16S sequencing of midstream voided urine samples collected from 20 European females without lower urinary tract symptoms.

Here, we utilize long-read 16S sequencing of 33 UTI positive and 14 UTI negative urobiome samples to achieve species level classification. Additionally, we produced long-read 16S sequencing of two mock communities of urinary isolates as a critical proof-of-concept for the technology for species-level identification of urinary isolates as well as detection of species variation.

Methods

Fecal PacBio Data

Because analysis of urobiome samples with PacBio data has not been done before, preexisting PacBio long-read data was collected and pushed through the DADA2 pipeline as a proof of concept. Five human fecal PacBio SMRT samples were used (accession numbers: SRR8557463.1, SRR8557464.1, SRR8557465.1, SRR8557466.1, SRR8557467.1; these five samples were the first five entries from BioProject accession number PRJNA521754) (69). The human fecal PacBio data was downloaded to a remote server using SRA toolkit function fastqdump (https://github.com/ncbi/sra-tools). The RScript created for PacBio analysis was derived from the script Callahan et al. developed for the analysis of samples in this BioProject (https://benjjneb.github.io/LRASManuscript/LRASms_fecal.html). The first step was removal of primers and an initial length filter. Then, the main steps of DADA2 were run, which include dereplicating sequences, learning an error model, denoising, and creation of a frequency table of ASVs. Taxonomy was assigned to the frequency table using the SILVA v138 training set and chimeras were removed.

Mock Communities

Mock communities were included to add a measure of accuracy. To create the mock communities, five urinary strains of *E. coli*, one urinary strain of *P. mirabilis*, one urinary strain of *S. epidermidis*, and one urinary strain of *E. faecalis* were grown. The *E. coli* strains were grown using BHI media, and the *Proteus*, *Staphylococcus*, and *Enterococcus* strains were grown using LB. Freezer (-80°C) stocks of all seven strains were streaked on 1.7% agar plates of the corresponding media. Plates were incubated overnight at 35°C with 5% CO_2. A single colony from each plate was added to 1 mL of the corresponding liquid media and incubated overnight at

35°C with 5% CO_2. DNA was extracted from these cultures using the DNeasy Blood and Tissue Kit (Qiagen), following the manufacturer's protocol for Gram-positive bacteria with the following exceptions: we used 230 µl of lysis buffer (180 µl of 20 mM Tris-Cl, 2 mM sodium EDTA, and 1.2% Triton X-100 and 50 µl of lysozyme) in step 2 and altered the incubation time in step 5 to 10 min. DNA concentrations were then quantified using a Qubit Flourometer. Two mock communities were created by combining different quantities of DNA (Table 1). Sequencing was conducted as described below for the samples.

Species/strain	**Mock Community #1 (relative abundance)**	**Mock Community #2 (relative abundance)**
E. coli UMB1180 (B1)	100 ng (14.96%)	400 ng (34.07%)
E. coli UMB1162 (B2)	100 ng (14.96%)	50 ng (4.26%)
E. coli UMB1225 (D)	100 ng (14.96%)	300 ng (25.55%)
E. coli UMB0103 (F)	100 ng (14.96%)	100 ng (8.52%)
E. coli UMB1220 (B2)	100 ng (14.96%)	50 ng (4.26%)
E. faecalis	54.3 ng (8.13%)	100 ng (8.52%)
P. mirabilis	60 ng (8.98%)	120 ng (10.22%)
S. epidermidis	54 ng (8.08%)	54 ng (4.60%)

Table 1. List of species, quantities of DNA, and relative abundances in each mock community. For *E. coli* strains, the phylogroup is listed in parentheses.

Sample Selection and Sequencing

The urobiome samples used in this study were provided by the Loyola Urinary Education and Research Collaborative (LUEREC), a part of Loyola Stritch School of Medicine. There were 33 urine samples from females with a clinical diagnosis of UTI, and 14 urine samples from females with no lower urinary tract symptoms (hereto referred to as UTI negative). The samples

were collected using transurethral catheterization as part of previously approved IRB protocols (IRB #: 207102, 204195, 206449, 207152, 209545, 204133). At the time of collection, AssayAssure (10% by volume) was added to the urine sample and stored at -80°C.

DNA was extracted from each urine sample using the Norgen Urine DNA Isolation Kit following the manufacturer's protocol with one exception: because we started with 500uL of urine (rather than 1.75mL), we adjusted the volume of the Binding Solution used accordingly. Serving as a negative control, nuclease free water was extracted using this same protocol.

PacBio SMRT sequencing was conducted at the University of Maryland facility. 10 uL of extracted DNA of each sample was sent to the University of Maryland where library prep was completed following the PacBio procedure. Briefly, a Qubit florometer was used to measure the DNA concentration, then the DNA was normalized to 500 pg/microliter in elution buffer. The DNA was amplified using the following primers: "AGRGTTYGATYMTGGCTCAG" (27F) and "RGYTACCTTGTTACGACTT" (1492R). A SMRTbell library was constructed from the PCR products and sequenced using the PacBio HiFi platform.

Analysis

We ran all samples through DADA2 with the same parameters. For the filterAndTrim step, the default DADA2 parameters were used (maxN=0, rm.phix=FALSE, maxEE=2) and the minimum length was set to 1000 bp and the maximum length was 1600 bp. After primers were removed and reads were trimmed, the sequences were dereplicated. Next was the error estimation step, which is critical to the DADA2 algorithm. The specific parameter for PacBio data "PacBioErrfun" was used for the error estimation model. Next, the DADA algorithm was run on the sequences. The assignTaxonomy() function from DADA2 was used to assign taxonomy with the SILVA v138 database (70).

After DADA2 was finished assigning taxonomy and creating an ASV table, the ASV table was used along with a metadata table to run the tool decontam (v 1.18), which uses statistical inferences to sort real ASVs from contaminants (71). In decontam, we used a 0.5 threshold filter, which was more aggressive than the default. The aggressive filter identifies all sequences in the negative control as a contaminant, removing any sequence present in the negative control in the experimental data. A presence-absence table from the metadata table was created, and from there a chi-square test was performed and P scores for each ASV were generated. Based on the P score value, an ASV was classified as a contaminant or a real species found. A data.frame is outputted so that the sequences found to be contaminants from the UTI positive and UTI negative samples could be removed from the original ASV table.

To represent the taxonomy found, ASV tables and taxonomy assignments were converted to .csv files and read back into RStudio. The taxonomic identifications were assigned to the ASV tables, replacing the actual sequences with species level classifications. With the modified ASV table, relative abundances of the species within samples could be created using the make_relative() function from the funrar package in R (72). To plot the abundances and relative abundances, the ASV tables had to be reshaped by the melt() function from the reshape2 package (73). Taxonomy plots were created using the package ggplot2 and arranged into figures using the ggpubr package (74,75). For the Shannon diversity metric, the diversity() function was used, which is provided in the base R language.

Results

Fecal Sample Analysis

First, we conducted a proof of concept for our analysis pipeline using five samples from the human fecal microbiome. This specific data was chosen because it was the proof of concept

data for PacBio data analysis in the DADA2 paper (76). In their proof of concept work, they highlighted the ability to utilize PacBio data with DADA2 and to detect strain level variation in the sample. Table 2 shows how many reads of the fecal samples passed through each step of the analysis pipeline using DADA2 (see table description). The average number of reads to make it through the DADA2 pipeline without being filtered out is about 60-64%.

	Number of Original Reads	**Number of Reads after Primers Removed**	**Number of Reads after Filtering Step**	**Number of Reads after the Denoising Steps**	**Percent Kept Throughout Pipeline**
SRR8557463	14675	11838	9447	9401	64.06%
SRR8557464	25306	19923	16052	15937	62.98%
SRR8557465	24657	18663	14939	14761	59.87%
SRR8557466	22799	18663	14939	14761	64.74%
SRR8557467	16315	12641	10115	9733	59.66%

Table 2. Numbers of reads after each step of DADA2: the Primers columns refers to how many reads were kept after the removePrimers() function where the user defines the primers used to sequence the reads to be removed. “Filtered step” refers to the reads remaining after the filterAndTrim() function, which filters reads of a certain quality score and trims the remaining sequences to a user-set length. “Denoised” refers to the number kept after the learnErrors() and dada() functions, where the PacBio error rates were applied to the experimental data and ASVs were inferred.

After it was confirmed that this proof of concept PacBio data could successfully run through our DADA2 analysis pipeline, we wanted to investigate if strain-level variation could be identified in the PacBio data. Figure 4 shows the abundances of the individual *E. coli* ASVs found in each fecal sample. There were 14 different ASVs found, which is the same result as the

Callahan paper using the data from this BioProject. In the paper, they concluded that this represented 6-7 strains (69).

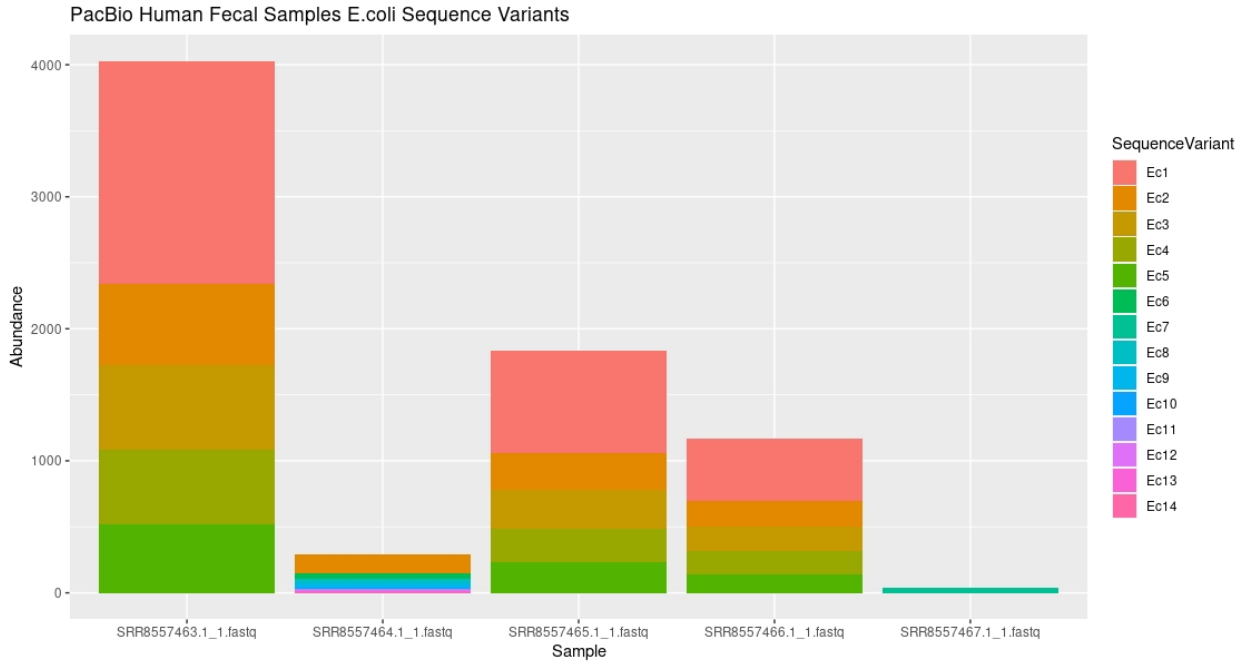

Figure 4**.** Stacked bar chart showing the raw abundances of *E. coli* sequence variants in the fecal PacBio samples.

Negative Control Analysis

Before analyzing the mock communities, the negative control sample was run through our DADA2 analysis pipeline. This sample showed a high relative abundance of *E. coli* ASVs (Figure 5). Other genera of note identified in the negative control were *Sphingomonas* and *Aerococcus*, which have both been observed in the urobiome previously (77). The ASVs that were not *E. coli* were found at a low abundance, mostly at a count of one.

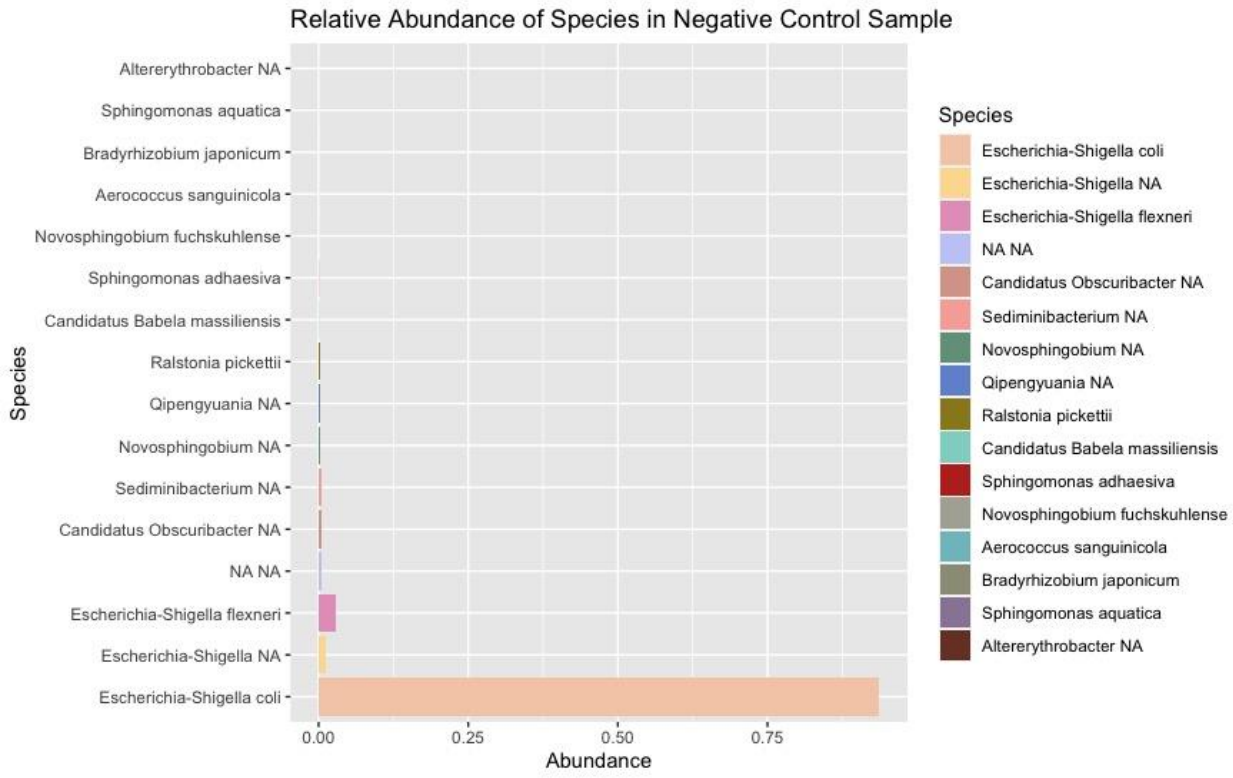

Figure 5**.** Relative abundance of all species classified from the negative control sample.

Mock Communities

Next, I analyzed the mock communities with DADA2. We could assess the quality of the data through tracking the reads through DADA2 and comparing those percentages kept to the percentages kept of the fecal data. Overall, the percentages kept through the pipeline were similar to the fecal proof of concept data. Also, we found that our mock community samples contained on average more reads than the fecal samples: the average number of reads in the fecal data was 20,750 and the average number of reads in the mock communities was 47,588.5 (Table 3).

	Number of Original Reads	Number of Reads After Primers Removed	Numbers of Reads after Filter Step	Number of Reads after Denoising Steps	Percent Kept Throughout Pipeline
Mock 1	50499	45772	28845	22443	44.44%
Mock 2	44678	40907	27611	24836	58.89%

Table 3**.** Shows the number of reads in the mock community samples after each step of DADA2.

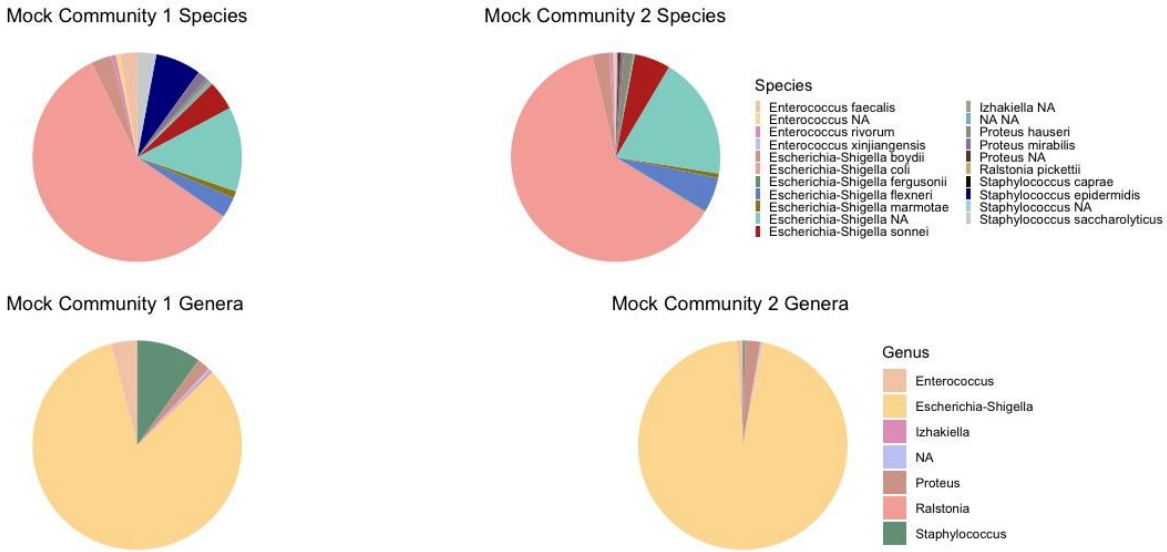

Figure 6**.** Pie charts showing the relative abundances of classifications for both mock communities at the species and genus levels.

In Figure 6 (top), we see that DADA2 with PacBio reads found all of the species within the mock communities; *Escherichia coli* (Mock 1: n=12,941, Mock 2: n=15,435)*, S. epidermidis* (Mock 1: n=1,558, Mock 2: n=20)*, E. faecalis* (Mock 1: n=586, Mock 2: n=87)*,* and *P. mirabilis* (Mock 1: n=328, Mock 2: n=119) are all present. Most of the other species' classifications found are within the genus of a species present, like *S. saccharolyticus*, *E. rivorum*, *P. hauseri*, *E.*

flexneri, *E. boydii*, *E. marmota*e, and *E. fergusonii*. Furthermore, many of the *Escherichia* species fall under the umbrella of *E. coli* in addition to the classification of *Escherichia Shigella* NA (Mock 1: n=2,903, Mock 2: n=4679); these ASVs could be classified to the genus, but not the species. (Note, the SILVA database combines the genera *Escherichia* and *Shigella*.)

When we consider the genus level rather than the species level, we find that most of the predictions are for the genera of the species included in the mock communities (Figure 6 bottom). There are only two incorrectly classified genera (*Ralstonia* and *Izhakiella*), and their abundance is very low compared to the correct classifications (relative abundances of 0.041% and 0.18% for Mock 1 and Mock 2, respectively). Species of these two genera, *R. pickettii* and *Izhakiella* NA, were detected in the species level analysis at low relative abundances (Figure 6 top). When looking at the genus level, the small differences between species are resolved and the classifications are truer to the biological communities.

Next, I looked specifically into the *E. coli* variants found within the mock communities. Originally, there were five strains added to the mock communities from four different phylotypes. We also know that *E. coli* can have seven different copies of the 16S rRNA gene sequence, which could alter the amount of 16S variants found (78). Among the 5 *E. coli* strains in the mock communities, 45 different sequence variants were identified. This may represent roughly the number of *E. coli* variants found (5 strains per 7 copies could equal 35 sequence variants). However, the mock communities also include *Proteus*, which also belongs to the family Enterobacteriaceae, which could influence classifications.

Analysis of UTI Positive Samples

Next, the 33 urobiomes from females with a diagnosed UTI were analyzed. 2,109 ASVs were classified after DADA2, and 32 were filtered from the decontam package as contaminants

(based upon the negative control sample), leaving 2,077 "true" ASVs. One sample, 7803 did not have any ASVs identified. Overall, the most abundant ASV in the UTI samples was *E. coli* (n=273,417, 48.4%). In addition to *E. coli*, other species of the genus "*Escherichia Shigella*" that were identified in the samples included *Escherichia Shigella flexneri* (2.4%)*, boydii* (1.7%)*, sonnei* (1.7%)*,* and NA (3%). The second most abundant species level classification was *Aerococcus sanguinicola* (n=26,104, 4.6%)*. Klebsiella aerogenes* (n=24,302, 4.3%) and *K. pneumoniae* (n=20,284, 3.6%) were also found in the UTI positive data. Other abundant species include *P. aeruginosa* (n=22,829, 4%) and *E. faecalis* (n=22,380, 3.8%). Figure 7 shows that when *Klebsiella, Aerococcus, Pseudomonas,* or *Enterococcus* genera were found in a sample, they were the majority of the classifications for that sample.

Some genera that were less abundant than the previously mentioned species but still prevalent in the samples were *Delftia tsuruhatensis* (4.1%)*, Rhizobium* NA (3.4%)*, Ralstonia pickettii* (1.5%)*, Lactobacillus iners* (0.5%)*,* and *Gardnerella vaginialis* (0.4%)*,* which were abundant in varying levels throughout the samples but never was the majority of the classifications for a sample. Figure 7 (top) shows a heatmap for the 50 most abundant species for UTI positive data and a stacked bar chart for the top 20 most abundant genera assigned. Figure 7 shows that the species and genera assignments are highly variable within the samples.

Other than examining relative abundances of species found, it is also important to note patterns found within the samples, classified as urotypes. A urotype is the dominant taxon in a urobiome (60). In our UTI positive samples, we found 5 distinct urotypes, samples that were dominated by a uropathogen: *E. coli, Klebsiella, Pseudomonas, Enterococcus,* and *Aerococcus* (Figure 2.4 bottom). A sixth urotype is suggested in which no species was a majority of the ASVs; we call this the "mixed" urotype. The most abundant urotype was mixed, with 13

samples. There are several common genera across these samples: *Ralstonia, Delftia*, and *Rhizobium* as well as less common genera like *Bifidobacterium* (sample 7714) and *Lactobacillus* (7772). The *Escherichia* urotype was the next most common, with 12 out of 33 samples being dominated by *E. coli* (7728, 7531, 7660, 7672, 7676, 7707, 7715, 7720, 7771, 7775, 7785, and 7791). The *Aerococcus* urotype was seen in samples 7758 and 7805, with relative abundances of specifically *A. sanguinicola* at 97.80% and 60.49% respectively. The other uropathogen-dominated urotypes were found in one sample each. The *Klebsiella* urotype was seen in samples 7651 and 7674. The *Pseudomonas* urotype was only seen in one sample, 7714, with a relative abundance of 78.94% *P. aeruginosa*. The *Enterococcus* urotype was seen in sample 7654 with 95.37% *E. faecalis* relative abundance.

Figure 7**.** Above: a heatmap of relative abundances of the top 30 most abundant species in the UTI positive samples. Below: a stacked bar chart illustrating the differences of abundances of genera within the samples. Note, sample 7803 in both graphs is empty because after DADA2 denoising, only 6 sequences remained, so no taxonomy was assigned to the sample.

Figure 8 examines the samples that did not have a high relative abundance of *E. coli* (<n=20 *E. coli* ASVs). Here, we see the four uropathogen urotypes (*Aerococcus, Klebsiella, Pseudomonas,* and *Enterococcus*) as well as the mixed urotype. Within the mixed urotype, samples contain a combination of the species *Delftia tsuruhatensis*, *Rhizobium* NA, and *Ralstonia* NA genera (it is worth noting that *R. pickettii* itself was not seen in every sample). Many of those classifications remain at the genus level, and do not extend to the species level (indicated by the classification of NA following a genus name) (Figure 8).

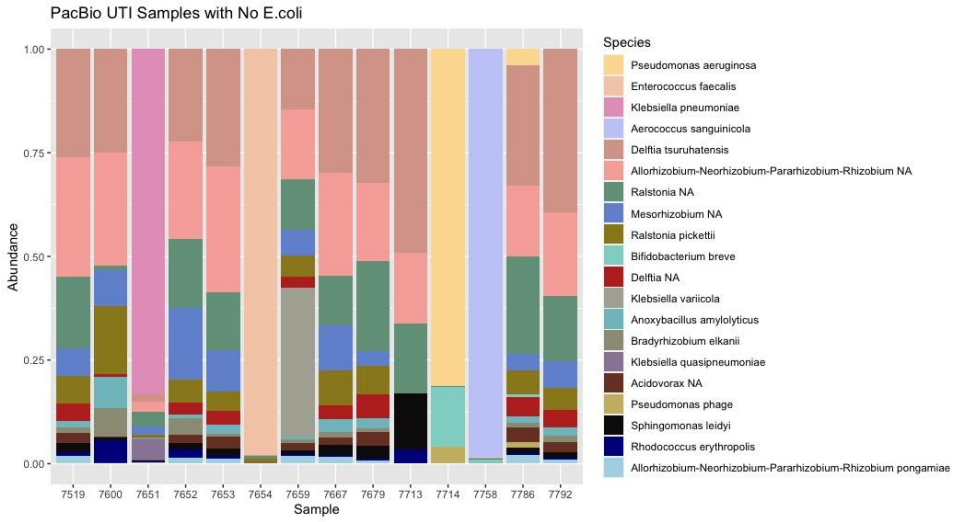

Figure 8. A stacked bar chart of the species composition of samples that had less than 20 *E. coli* ASVs assigned. Sample 7803 was removed because no taxonomic classifications were assigned.

We also examined the specific *E. coli* sequence variants of the UTI positive samples. 393 unique *E. coli* sequence variants were found within the UTI positive samples. The number of sequence variants per sample ranged from 0 (sample 7600) to 71 (sample 7720) (Figure 9). From comparing the raw counts of ASVs to the number of sequence variants found, we can see that the

number of sequence variants does not directly correlate to the number of ASVs classified (ie, more counts means more ASVs). For example, sample 7772 had 23 *E. coli* sequence variants with *E. coli* having a relative abundance >30%, but sample 7771, which had only 3 sequence variants, had *E. coli* at a relative abundance of over 90% (Figure 9).

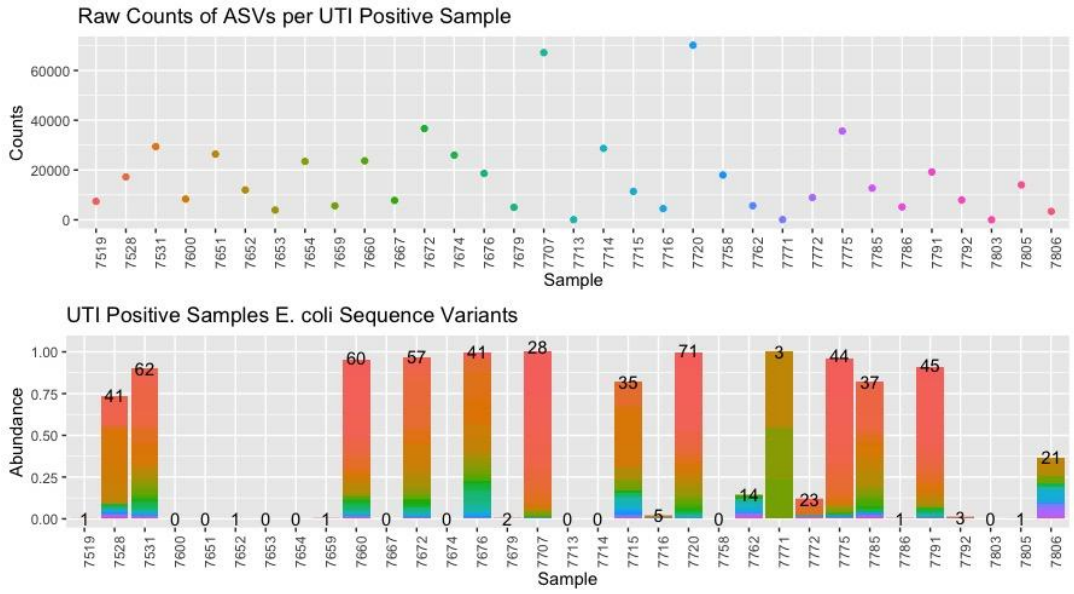

Figure 9. Above: Scatterplot of raw counts of ASVs assigned to the UTI positive data. Below: stacked bar chart of the *E. coli* sequence variants with the counts of sequence variants on the bars per sample.

Analysis of UTI Negative Samples

Next, we classified the taxonomies of the 14 urobiomes from asymptomatic (UTI negative) females. There were 912 ASVs found. Decontam classified 7 of those as contaminants, so after processing there were 905 ASVs. While *E. coli* was found in the UTI negative samples, it was not the most abundant taxa (Figure 10 top). The most abundant taxa were *Ralstonia pickettii* (n=115,018, 48.3%) followed by, *Escherichia Shigella dysenteriae* (n=38,170, 16%) and *Escherichia Shigella coli* (n=19,582, 8.2%). After the two *Escherichia* species was *Delftia tsuruhatensis* (n=16,278, 6.8%). The next most abundant taxa were only classified at the genus

level and not to the species (*Rhizobium* NA 5%*, Ralstonia* NA 4.3%*, Delftia* NA 1%). With the exception of sample 5461, many of the UTI negative samples resembled each other. Sample 5461 was dominated by *Escherichia* with a relative abundance of 96.88%, while the average relative abundance of *Escherichia* in the other UTI negative samples was 0.69% (Figure 10).

Figure 10**.** Above: a heatmap of relative abundances of the top 30 most abundant species in the UTI negative samples. Below: a stacked bar chart illustrating the differences of abundances of genera within the samples.

There were only two urotypes observed in the UTI negative data, mixed and *Escherichia*. All of the samples except for 5461 were of the mixed urotype, being mainly composed of *Ralstonia, Delftia, Rhizobium,* and *Mesorhizobium*. The previously four listed genera were present in every UTI negative sample, except for sample 5461. *Lactobacillus,* a previously

established healthy urobiome community member, was found in 4339, 4368, 4646, 4668, and 4814 (60). *Bradyrhizobium elkanii*, *Mesorhizobium jarvisii*, *Sphingomonas leidyi* were also all found in every sample except for sample 5461. Sample 5461 has a relative abundance of 96.88% *E. coli*, and stands out from the other urotype observed in the dataset.

The *E. coli* variants of the UTI negative samples were also plotted (Figure 11). Only 22 sequence variants were found in the UTI negative samples, which is less than the mock communities and UTI positive samples. Furthermore, almost all of the *E. coli* variants in the UTI negative samples were found only in sample 5461 (n=19,407). In fact, it contained all the sequence variants found within the UTI negative group. In fact, the sample has a higher abundance of *E. coli* ASVs than some UTI positive samples (sample 5461 *E. Coli* abundance: 96.88%, sample 7519: 0.12%, sample 7772: 12.22%, sample 7758: 14.40%) and more *E. coli* sequence variants than 20 of the UTI positive samples (Figure 10). Although less abundant than sample 5461, samples 4646 (n=49) and 4979 (n=26) had *E. coli* sequence variants as well.

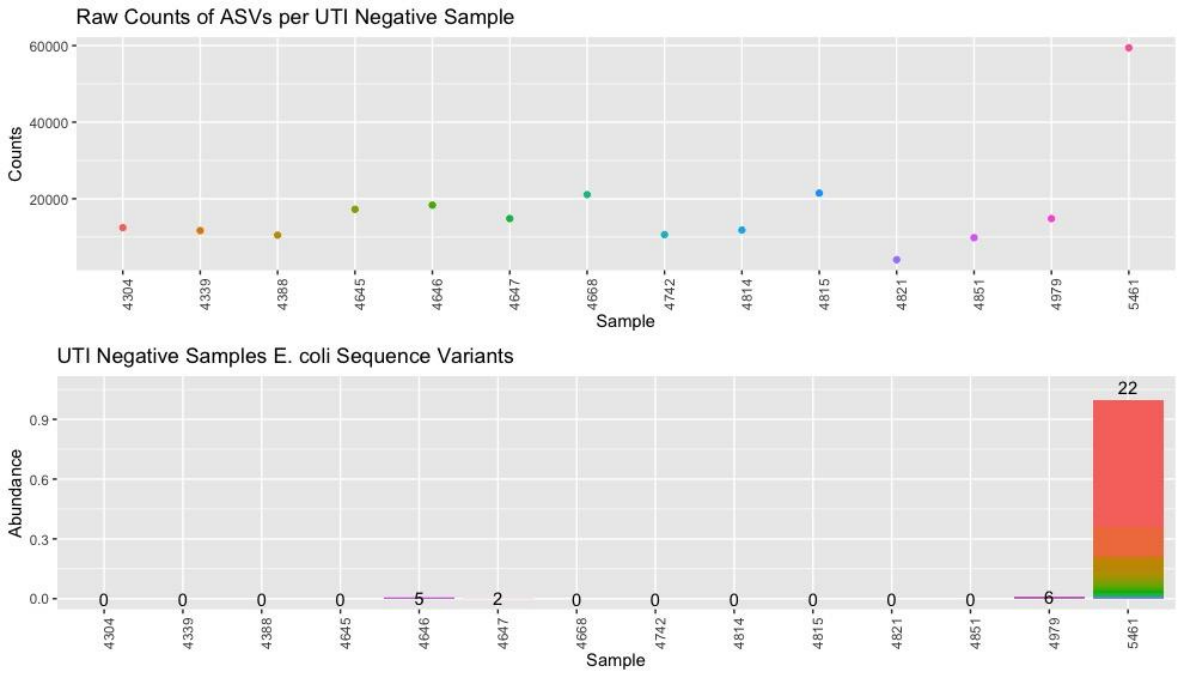

Figure 11. Above: Scatterplot of raw counts of ASVs assigned to the UTI negative data. Below: stacked bar chart of the *E. coli* sequence variants with the counts of sequence variants on the bars per sample.

Comparison of UTI and UTI Negative Samples

The taxonomic assignments of UTI positive and UTI negative samples vary, but they share *E. coli*, *R.pickettii*, and *D. tsushatensis* in their most abundant classifications. Examining the genera of the communities shows that in UTI positive samples, not only is there a higher abundance of sequences and ASVs in general, but there are previously identified pathogenic genera including *Klebsiella, Aerococcus, Pseudomonas, and Enterococcus.* In the UTI negative samples, *R. pickettii* was present in every sample except for 5461, which was majority *E. coli*. *E. coli* was the third most abundant ASV in the UTI negative samples, after *R. pickettii* and *Escherichia Shigella dysenteriae,* although almost all of the abundance is found in sample 5461 (n=19,407). When taxa are collapsed to the genus level, the differences within *Escherichia Shigella* are resolved and the genus becomes the second most abundant in the UTI negative samples. After *Escherichia Shigella*, the third most abundant genus is *Delftia*, which is much more widespread across the UTI negative samples (Figure 8).

While individual taxonomic composition is important, the main similarities and differences between the UTI positive and UTI negative samples are best shown by their urotypes. The UTI positive samples produced six urotypes (*Escherichia, Aerococcus, Klebsiella, Pseudomonas, Enterococcus*, and mixed), and the UTI negative produced two (mixed and *Escherichia*). The difference in number of urotype indicates that the UTI negative samples are more similar to other samples in the group, whereas in the UTI positive samples the composition varies greatly. In the mixed urotype in both UTI positive and UTI negative samples, *Ralstonia*, *Delftia*, *Rhizobium*, and *Mesorhizobium* are shared across all samples. The members of the

mixed urotype do not vary greatly between UTI positive and UTI negative samples, so it is interesting that some of the patients are symptomatic for UTIs, and others are not.

To examine the differences in ecological diversity within the UTI positive and UTI negative groups more in depth, Shannon diversity scores were assigned to the samples and those scores were plotted on histograms (Figure 12). Generally, the UTI positive samples showed more diversity than the UTI negative samples shown by the Shannon Diversity scores. The UTI negative scores range from 1.4-2.4, whereas the UTI positive samples have a wider range (0.8-4.2).

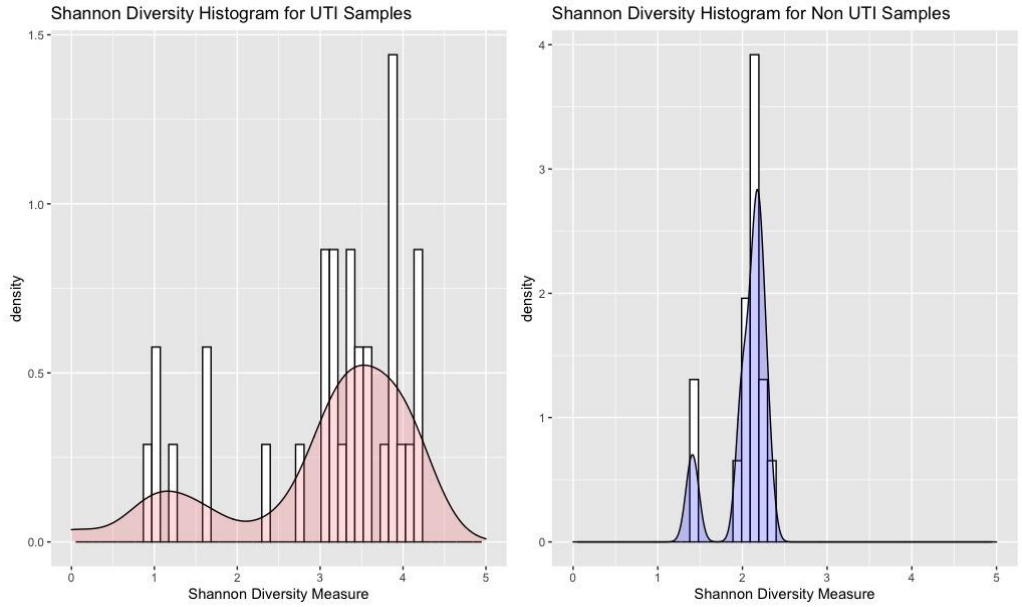

Figure 12**.** Left: Shannon diversity score of UTI positive data. Right: Shannon diversity score of UTI negative data. Density is the number of samples assigned a certain Shannon diversity score.

Given the relevance of *E. coli* to UTI symptoms, we next compared the *E. coli* sequence variants among the UTI positive and UTI negative samples. In total, 410 *E. coli* sequence variants were identified among the 27 *E. coli*-containing urobiome samples (23 UTI positive and 4 UTI negative). The majority (70.73%) of these sequence variants are found only in one sample.

The most abundant sequence variant was identified in 12 of these samples, all UTI positive samples. While it was found in all these samples, it was typically found in low relative abundance (<5%). There were no sequence variants only found and found in all the UTI negative *E. coli*-containing urobiomes. This suggests that there is not a pathogenic and non-pathogenic *E. coli* "type." Sequence variants observed in more than one of the UTI negative samples were also observed in more than one of the UTI positive samples.

Discussion

In this chapter, we investigated the urobiomes of females that had UTIs and those that did not. We found species and genera associated with the urobiome from previous research, and other species that have not been documented as often (22,28,60,66,79,80). Utilizing PacBio sequencing also allowed us to examine possible strain-level diversity with sequence variants.

UTI Positive Sample Taxonomy and Uropathogens

The most abundant ASV in the UTI positive samples was *E. coli*, which is not surprising given it is the primary cause of acute UTIs (21,22). However, not all the UTI positive samples included *E. coli* (Figure 2.6). *A. sanguinicola* was the second most abundant ASV in the UTI positive samples, which has been investigated as a uropathogenic species (79). It is worth noting that *A. sanguinicola* was a contaminant present in the negative control sample, but because the samples were computationally decontaminated and the contaminants removed, that specific ASV was removed from the experimental data. *A. sanguinicola* and *A. urinae* have both been found in urobiomes associated with UTIs, but of the two, only *A. sanguinicola* was found in our UTI positive samples (79,81). *A. christensenii* was the only other *Aerococcus* species found in the UTI positive data (sample 7772), which has been linked with polymicrobial infection in the

vagina (82). Interestingly, *A. christensenii* (n=462) was found in sample 7772 at a higher abundance than *A. sanguinicola* (n=4).

The next most abundant ASV was *Klebsiella aerogenes*. *K. aerogenes* was previously named *Enterobacter aerogenes*, but was renamed as it more closely resembles *Klebsiella pneumonaie* than the *Enterobacter* genus (83). While it has not yet been connected to UTIs, it has been connected to bloodstream infection (84). The next three most abundant ASVs are known uropathogens: *P. aeruginosa*, *E. faecalis,* and *K. pneumoniae* (25,80,85).

Lactobacillus species, many of which are known to have strong commensal effects on uropathogens (28), were found in some of the UTI positive samples. *L. iners* was the most abundant *Lactobacillus* ASV found in the UTI positive data, but it was only found in sample 7772 (n=2,911). It was the most abundant taxonomic classification in the sample with a relative abundance of 32.56%. The next most abundant *Lactobacillus* ASV was *L. jensenii*, found in samples 7653 (0.23%), 7715 (0.4%), 7772 (0.04%), 7805 (0.07%), and 7806 (1.37%). *L. cristpatus* (n=30)*, L. fornacalis* (n=30)*,* and *L. gasseri* (n=35) were all found in trace amounts.

UTI Negative Samples and the Lack of *E. coli*

In a 2019 review article, *Lactobacillus, Gardnerella,* and *Streptococcus* were the most common genera found in asymptomatic female urobiomes (60,77). All three genera were present in the ASV classifications of our UTI negative data, with *Lactobacillus* and *Gardnerella* classifications more common than *Streptococcus*. While *Lactobacillus* strains have previously shown commensal effects on uropathogens, our data presented showed that their presence was not the deciding factor on if *E. coli* was present in the community. Similar to the UTI positive samples, *L. iners* was the most abundant *Lactobacillus* species in the UTI negative data. It was found in five samples at low abundances (sample 4339 - 0.06%, 4338 - 6.94%, 4646 - 0.45%,

4668 - 2.67%, and 4814 - 0.02%). In the samples where *L. iners* was found in the UTI negative data, there was no *E. coli* found except for one, sample 4646. In contrast to other Lactobacilli, *L. iners* is the only species that has not been associated with antimicrobial properties as it does not produce either H_2O_2 or lactic acid, rather producing a cytotoxin (86). Furthermore, in the vaginal microbiota, *L. iners* has been associated with bacterial vaginosis (87). *L. iners* was recently identified as a dominant member of some asymptomatic female urobiomes (68).

The next most common *Lactobacillus* was *L. gasseri*, which was found in sample 4646 (0.03%), 4668 (0.03%), 4851 (0.06%), and 4749 (0.2%). *L. crispatus* was the next most abundant with 108 ASVs classified, and the other *Lactobacillus* species were found in trace amounts (*L. jensenii* n=76, *L. paragasseri* n=34, *L. taiwanensis* n=7. and *L. kitasatonis* n=2). *E. coli* was found in samples that contained *L. crispatus* and *L. gasseri*, showing that the presence of a *Lactobacillus* species alone will not completely destroy *E. coli*, as has been previously observed (14,28,60). Other previously found community members for the healthy urobiome were found in low abundance. Another known healthy urobiome community member is *Gardnerella vaginalis*, which we found in samples 4339 (0.1%), 4388 (0.18%), 4668 (0.2%), 4851 (0.36%), and 4979 (0.69%) (60). Although *G. vaginalis* has been seen in the healthy urobiome, recent study shows that it can be present in recurrent UTIs (77,88). Streptococcus has also been associated with the healthy urobiome (60). Two *Streptococcus* species were found, *S. thermophilus* and *S. alactolysticus,* at low abundance (n=9 and n=5 ASVs across all UTI negative samples respectively).

Most of the UTI negative samples had a low abundance of *E. coli*, except for one, sample 5461 (Figure 11). The individual sample 5461's sequence variant abundances are plotted in Figure 11, showing the sample had all 22 sequence variants found in the UTI negative group.

After consulting the metadata for the sample, it was confirmed that this patient had gone to the hospital for a different reason, and after a urine test, it was found that they had a large amount of *E. coli* in their urinary tract. This patient had substantial amounts of *E. coli* in their urinary tract but was asymptomatic, so a clinical UTI diagnosis was not made. With the *E. coli* variants present in sample 5461, as well as samples 4646 and 4979, we confirm earlier studies finding *E. coli* in the "healthy" urobiome (60,89).

Similarities Between UTI Positive and UTI Negative Taxonomies: *Ralstonia* and *Delftia*

There were some major similarities found in the UTI positive and UTI negative taxonomies, notably the presence of *R. pickettii* and *D. tsuruhatensis*. *Ralstonia*, in particular the species *R. pickettii* and *R. solanacearum*, were once classified as *Pseudomonas* species but they were removed and reclassified as a new genus *Ralstonia* due to phenotypic properties (90). It should be noted that two ASVs classified as *R. pickettii* were present in the mock communities after decontamination, but in the experimental samples they are present at significantly higher abundances. We hypothesize that either *Ralstonia* is present in the urobiome (despite not previously being isolated via culture methods) or it was introduced via contamination in the lab or collection method. *R. insidiosa* has previously been mentioned in regard to contamination of metagenomic samples (91). Contamination could be introduced via the collection method (catheterization here), the processing of the sample (collection tube, addition of AssayAssure), the DNA extraction kit, the library preparation kit, or the sequencing itself. Negative controls for the library prep and sequencing, however, did not produce reads thus removing the last two possibilities from our list. *R. pickettii* was found at varying relative abundances throughout the UTI positive and negative samples. The other possibility is that *R. pickettii* is a urobiome community member that has not been extensively documented, possibly because of its lack of

ability to be cultured with the EQUC procedure (9). In a recently published paper examining the female asymptomatic urobiome using 16S sequencing (V1-V8), *R. mannitolilytica* was found and notably, it could not be found by culturomics (68). *R. mannitolilytica* was also identified in long-read 16S sequencing of urinary samples and assigned to one of the observed urotypes (68). Shotgun metagenomic sequencing of urinary samples has also found *Ralstonia* (no species assignment) in UTI positive samples (92).

A similarity between *R.pickettii* and *D. tsuruhatensis* is their growing prevalence in serious hospital-related infections; strains of *Delftia* have proven to be fatal in some cases like *Ralstonia* (93). That furthermore begs the question of why those species are found in the urobiomes of patients without UTIs at such high abundances. Their presence in our UTI data is also interesting because our data was not collected from hospital related UTIs. A characteristic that *Delftia* has that *Ralstonia* does not have is its proximity to UTIs; while no UTIs directly caused from *D. tsuruhatensis* have been recorded as of 2022, UTIs caused by its close genetic cousin *D. acidovorans* have (94). Like *D. tsuruhatensis*, *D. acidovorans* is a rare opportunistic pathogen that is most commonly found in immunocompromised patients and is resistant to several antibiotics (95). These species are extremely difficult to tell apart; in one study of an infant with a *D. tsuruhatensis* infection in the lungs, the MALDI-TOF result was the pathogen was *D. acidovorans* and the 16S rRNA sequencing result was it was *D. tsuruhatensis* (96). Note, this species has also been called *Comamonas acidovorans* and *Pseudomonas acidovorans* previously, which only adds to the possibilities of these species being confused for another (97). Both *Delftia* species have been found in water and soil, as well as on medical equipment (97). In a case of *D. acidovorans* causing a complicated UTI, it was hypothesized that the pathogen was introduced via long-term urinary catheter (97). A different *D. tsuruhatensis* infection was

similarly introduced via intravenous catheter (94). Both *Delftia* species discussed and *Ralstonia* species have much in common: they are both originally environmental pathogens, they have a history of causing infection in a healthcare setting, both show patterns of antibiotic resistance, and they can cause serious infection. They are also both prevalent species in both the UTI and UTI negative data.

Differences Between the UTI Positive and UTI Negative Samples

While the UTI positive and UTI negative samples shared abundances of *R. pickettii* and *D. tsuruhatensis* particularly within the mixed urotype, the datasets had differences in overall taxonomic composition. The main difference can be seen in the number of urotypes: the UTI positive had six urotypes and the UTI negative data had two, and if sample 5461 was removed, the UTI negative samples would have one urotype: mixed. This shows that potentially uropathogenic species often take over the urobiomes they inhabit and are not present in asymptomatic urobiomes. *Klebsiella* was found in two samples, 5461 (n=2) and 4821 (n=1), although its presence is probably a false classification because the abundance is so low. *Enterococcus* (n=12) and *Aerococcus* (n=4) were also found in trace amounts across all UTI negative samples. *Pseudomonas* (n=293) was found in a higher abundance than the other uropathogens, making up 1.7% of sample 4979. *Escherichia* was the most commonly found potentially uropathogenic genus in the UTI negative samples, with an abundance of 675 across all samples except for 5461, which had an abundance of 80,999. Note, this value is for *Escherichia* genus classifications, as *E. coli* itself was not found in every sample.

When *E. coli* sequence variants were examined, we found that the UTI positive had 393 distinct *E. coli* ASVs and the UTI negative had 22. 17 of the sequence variants were only found in UTI negative samples (5461, 4646, and 4647), although one variant was not seen in all the

UTI negative samples and not in any UTI positive samples. This shows that there is a difference in the *E. coli* strain composition between uropathogenic *E. coli* and other *E. coli* that can exist in the urobiome without incidence of infection, but there is not a distinct type or sequence that we can identify from our data. A further direction of this study could be investigating these sequence variants more, and determining indicators that may make one sequence variant uropathogenic and another a non-infectious community member.

Conclusion

In this chapter, we investigated the urinary microbiome of patients that were UTI positive and UTI negative using PacBio long-read data. Utilizing PacBio sequencing allowed for our classifications to reach the species level with relative accuracy, as shown by our mock community results. Many of the UTI positive samples had a large abundance of uropathogenic *E. coli*, but we also found that some UTI positive samples did not contain large amounts of *E. coli* or any *E. coli* at all. In those samples, *R.pickettii* and *D.tsuruhatensis* were abundant among many other classifications, resulting in a mixed urotype. Other pathogens like *Aerococcus, Enterococcus, Klebsiella,* and *Pseudomonas* were found to take over the majority of classifications, resulting in a more species-dominant urotype than the mixed samples. In the samples of UTI negative patients, many of the ASVs were classified as *R. pickettii* and *D. tsuruhatensis* in a mixed urotype composition. We also found that some of the UTI negative patients had *E. coli* in their urinary microbiomes, conferring with prior studies finding *E. coli* in the “healthy” urobiome.

CHAPTER THREE

LONG-READ SEQUENCING VERSUS SINGLE VARIABLE REGION 16S rRNA SEQUENCING

Introduction

With the decreased cost of long-read sequencing, its application has expanded from whole genome sequencing projects to also include amplicon sequencing studies, such as 16S surveys. Long-read sequencing can capture all nine variable regions in contrast to short-read sequencing, which only captures one or a few variable regions. The majority of 16S surveys conducted to date, however, relied on short-read sequencing. Given the limited amplicon length of short-read sequencing, prior studies identified the best variable region to sequence for a microbial community by comparing the taxonomic resolution of single variable regions whole 16S rRNA gene sequences (98). Previously utilized variable regions for the urobiome include V1-V3, V4-V6, and V4 alone, with the later variable regions (V6-V9) not being recommended or utilized as often (11,14,99). General guidelines for 16S surveys of urobiome samples have been recommended (100).

Sequence amplicons for commonly used primers for short-read 16S rRNA sequencing were compared for their ability to identify taxa of the urobiome (99). In this study, amplicon sequences were computationally generated from full length 16S rRNA gene sequences from urobiome whole genome sequences (101). DADA2 was used to identify ASVs, and species were predicted using both a BLCA classifier and naive Bayes algorithm. They found that the database used had the biggest impact on taxonomy found. They also found that multiple variable regions

(i.e., V1-V3 or V2-V3) are better for taxonomic accuracy than single variable regions (i.e., V4), and the taxonomy found from these regions was relatively accurate.

Recently, computationally derived variable region amplicons derived from full-length long-read 16S rRNA gene sequences (rather than whole genome sequences) have also been compared, but for the gut microbiota (31). Furthermore, this study additionally conducted short-read sequencing. They used the QIIME pipeline to create OTUs clustered at 97% and used the SILVA (v132) database to assign taxonomy (31). They found that the taxonomies assigned to the long-read 16S rRNA amplicon sequences and the artificially created variable regions were more different than the artificially created variable regions and the short-read sequencing variable regions, which were more similar. They concluded that it was not necessarily the type of technology that would provide more information about the community, but the length of the amplicon that was more effective in resolving taxonomy.

The precedent set by both papers provides a solid foundation to attempt to mimic variable region data from our PacBio urobiome data and compare the taxonomy assigned to the taxonomy found from the full length 16S reads. Comparing taxonomy found from full-length sequences and short-read sequences can reveal bacterial genera that have been previously over or under-represented in previous urinary microbiome research that resulted from the type of sequencing used. It is important to note that the term “single variable 16S sequencing” may refer to multiple variable regions (i.e., V1-V3) in this chapter, but because it does not include the entire 16S gene, it will be referred to as single variable.

Methods

Multiple Sequence Alignment

Two variable regions, V1-V3 and V4, were computationally parsed from our PacBio reads using the methodology developed in the Hoffman et al. paper (30). The PacBio reads were first converted from fastq format to fasta format using SeqKit (https://github.com/shenwei356/seqkit). The fasta format files were then aligned to the *E. coli* 16S rRNA gene sequence (Accession No. EU014689.1) using MUSCLE (v5) with the "-clwstrict" parameter specified (102). Alignment to the *E. coli* reference sequence is necessary to identify the location of the PCR primers and thus amplified sequence for the variable regions. The hybridization sites were identified within these alignments for the following primers sequences: V1: 27F (AGAGTTTGATCCTGGCTCAG), V3: 534R (ATTACCGCGGCTGCTGG), V4: 515F (GTGCCAGCMGCCGCGGTAA) or 806R (GGACTACHVGGGTWTCTAAT). Based upon these hybridization sites, the expected amplicon sequence was parsed from the full-length sequence for each read. The fasta format files had to be converted back to fastq file format for DADA2 analyses.

DADA2 Analysis

The samples were run through DADA2 in the same order as detailed in the previous chapter's methods. Because the sequences were changed to fasta format for the multiple sequence alignment, the quality scores associated with the sequences were lost. Therefore, the DADA2 error model algorithm could not be used. Sequences that did not align to the complete reference gene region were removed. We also removed the ASVs that had a count of 1 because the DADA algorithm removes singletons. Next, taxonomy was assigned within the pipeline using the assignTaxonomy() function with the SILVA database, as detailed in Chapter 2.

Data Visualization

To create the heatmaps (Figures 3.4 and 3.5), ASV tables of all the combined ASVs between UTI positive and UTI negative data were created. The values in the tables were made into relative abundances per patient using the make_relative() function from the R package "funrar" (72). The melt() function from the "reshape2" package was used to reformat the tables for visualization using "ggplot2" (73,75).

Statistics

To investigate the differences between the V1-V3 and PacBio taxonomic classifications and the differences between the V4 and PacBio taxonomic classifications, relative abundances were calculated. Paired t-tests were used for the V1-V3 data, and the Wilcoxon Rank sum test was used for the V4 data (correct for lower number of samples). Because the ASV abundances were highly variable from patient to patient, an FDR (or Benjamini-Hochberg) correction was used to normalize the p-values that were outputted from the tests. The "reshape2" package was used to manipulate the data and "ggplot2" was used for visualization (73,75). The aforementioned statistical analyses were conducted in R.

Results

Because short-read 16S sequencing cannot reliably classify to the species level, our comparison of long-read and computationally generated variable region amplicons will be at the genus level. The results are presented in the following order: Mock communities, V1-V3 individual results for UTI positive and negative data, V4 individual results for UTI positive and negative data, a comparison of the relative abundances of V1-V3 and V4 taxonomic classifications, and the comparisons of the V1-V3 region to PacBio data and V4 region to PacBio data.

Mock Communities

The genera in the mock communities were *Escherichia* (with five separate strains, four phylotypes)*, Staphylococcus, Proteus, and Enterococcus* at varying concentrations. As shown in Figure 13, the long-read sequencing captured the four genera tested. Ten genera were assigned within the V1-V3 region, with the four known community members being the most abundant. The additional six genera (*Ralstonia, Enterobacteriaceae* NA, *Aerococcus, Mesorhizobium, Delftia,* and *Rhizobium*) made up 0.15% of Mock Community 1 and 0.18% of Mock Community 2 (Figure 13). Only *Ralstonia* was identified by the PacBio sequencing. Taxa that could not be classified at the genus level were rare: 5 out of 13,429 ASVs for Mock Community 1 (0.04%) and 2 out of 10,852 ASVs of Mock Community 2 (0.02%) were classified as "NA".

Fourteen genera were identified by the V4 data not included in the mock communities. Five of these genera, *Ralstonia,* Enterobacteriaceae NA, *Delftia*, *Rhizobium*, and *Mesorhizobium*, were also identified during our V1-V3 analysis. The V4 classifications included additional misclassifications, many of which were not able to be classified to the genus level. In Mock Community 1, 481 ASVs were classified as "NA" for their genus (Enterobacterales NA, Morganellaceae NA, Enterobacteriaceae NA, Staphylococcaceae NA, Bacilli NA, Lactobacillales NA, and Bacillaceae NA) making up 1.7% of classifications. In Mock Community 2, 61 ASVs were NA at the genus level, 0.2% of classifications. Both values are greater than the mock community NA abundances from the V1-V3 regions (0.04% and 0.02% respectively.)

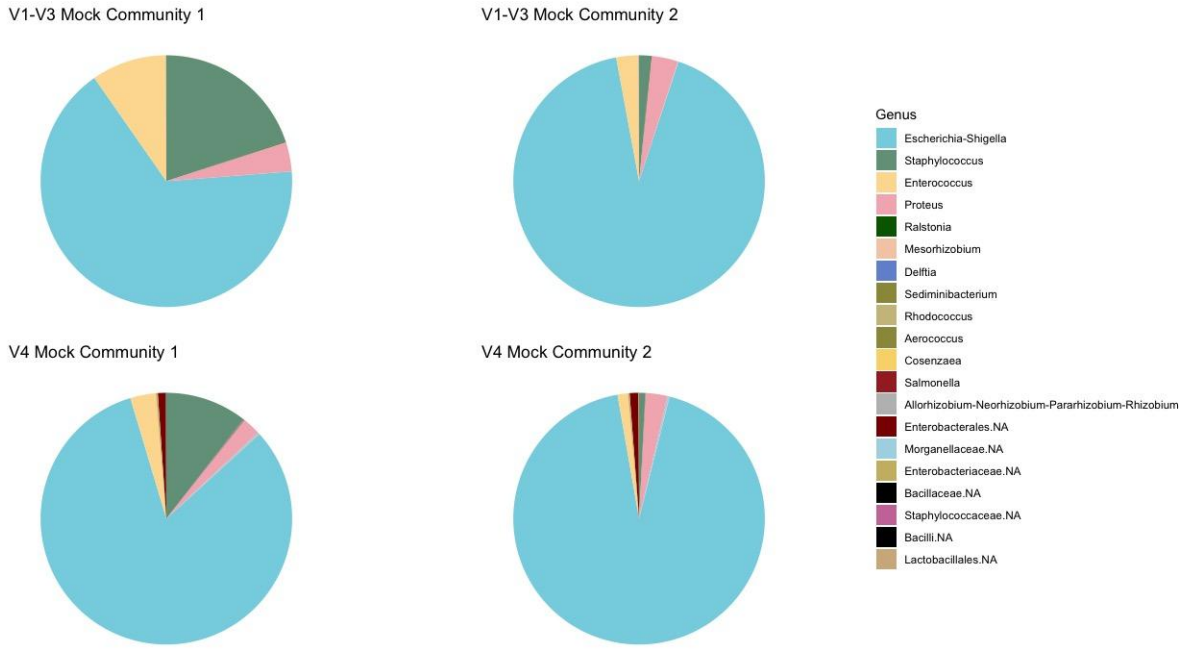

Figure 13. Pie charts showing abundances of ASVs in mock community samples according to the V1-V3 regions (top) and the V4 region (bottom).

V1-V3 Classifications of Urobiome Samples

Analysis of the 33 UTI positive samples using the V1-V3 data lead to the identification of 106 taxonomic classifications with the most abundant taxon being *Escherichia Shigella* (n=286,365, 37.34% of ASVs), and the next most abundant taxon being *Ralstonia* (n=204,770 26.62% of ASVs). The next most abundant genus was *Aerococcus*, which has been associated with UTIs before (79). *Aerococcus* was mainly found in sample 7758 (n=51,181, 98.99%) and sample 7805 (n=12,296, 57.34%). The next most abundant ASV classification was *Delftia,* which was found in almost every UTI positive sample (n=29,483, 3.83%). The next most abundant genera have all been associated with causing UTIs before: *Pseudomonas* (3.18%)*, Klebsiella* (2.98%)*,* and *Enterococcus* (n=22,741, 2.95%) (25,80,85). We found *Lactobacillus* (0.45%), *Gardnerella* (0.42%), and *Sphingomonas* (0.31%), all of which have been observed in the healthy urobiome (60,101).

The V1-V3 data also showed a pattern of one microbe often dominating a community. As Figure 14 (right) shows, UTI samples dominated by *Klebsiella, Pseudomonas, Aerococcus,* or *Enterococcus* (samples 7651, 7654, 7674, 7714, 7758, and 7803) had little or no ASVs assigned for the genera *Escherichia* or *Ralstonia*. Note, sample 7674 is assigned the *Klebsiella* urotype because although *Klebsiella* is not a majority of ASVs, it also includes a significant number of ASVs assigned to Enterobacteriaceae NA and *Klebsiella* is a member of Enterobacteriaceae.

In the UTI negative samples (Figure 14 left), we found an abundance of *Ralstonia* (n=126,801, 49.64%) ASVs. The second most abundant ASV was *Escherichia* (n=72,586, 27.65%). The majority of *Escherichia* ASVs came from sample 5461, which had 72,092 *Escherichia* ASVs and a relative abundance of 99.88%. The third most abundant genus was *Delftia* (n=17,450, 7.76%). Other frequently identified taxa include *Rhizobium* (n=13,069, 4.98%), *Mesorhizobium* (n=7,060, 2.69.%), and *Candidatus Obscuribacter* (n=5,290, 2.01%). *Lactobacillus* is another important community member in UTI negative samples, and it was the 7th most abundant ASV (n=2,238, 0.85%). *Sphingomonas* was the next most abundant classification (n=1,706, 0.65%). In the UTI negative data, we observed two urotypes: mixed and *Escherichia* (sample 5461).

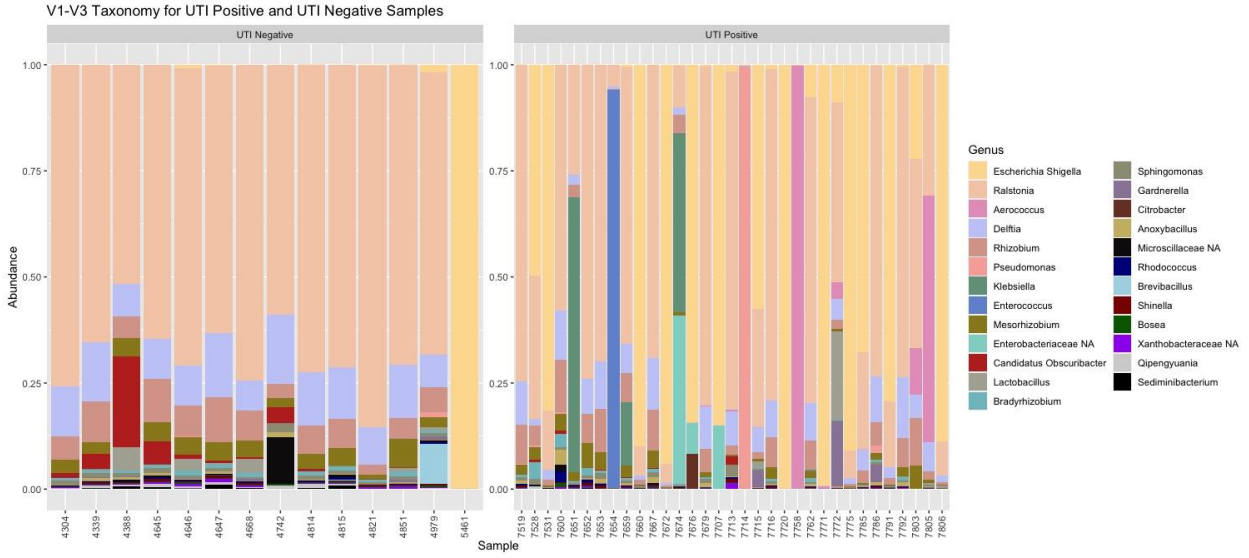

Figure 14**.** Relative abundances of UTI positive and UTI Negative samples for the V1-V3 region. The top 25 genera from both sets of classifications are plotted.

V4 Classifications of UTI Positive and Negative Samples

Next, we examined the UTI positive samples' predictions based upon the V4 region. The most abundant ASV in the UTI positive samples was *Escherichia Shigella* (n=416,844, 44.24%). The next most abundant ASV was *Ralstonia* (n=226,832, 24.08%). The third most abundant ASV was *Aerococcus* (n=84,187, 8.93%). The next top ASVs include *Klebsiella* (5.54%)*, Delftia* (3.72%)*, Rhizobium (*2.84%*), Pseudomonas* (2.78%), and *Enterococcus* (2.58%). *Mesorhizobium* was another frequently identified genus (1.15%). *Lactobacillus* (0.39%)*, Sphingomonas* (0.31%), and *Gardnerella* (0.38%) were also found at similar concentrations to the V1-V3 classifications.

Next, we examined the V4 classifications for the UTI negative samples. In the V4 UTI negative samples, there were 124 distinct taxonomic classifications. The most abundant classification was *Ralstonia* (n=144,189, 47.49%), followed by *Escherichia Shigella* (n=81.684 26.9%), and *Delftia* (n=24,261, 8.0%). Again, the majority of *Escherichia Shigella* ASVs were

in sample 5461 and without the sample, the relative abundance drops dramatically from 26.9% to 0.29%. *Rhizobium* (5.45%) and *Mesorhizobium* (2.92%) were next, followed by *Candidatus Obscuribacter* (2.1%). *Lactobacillus* (0.83%) and *Sphingomonas* (0.74%). *Gardnerella* was less abundant (0.11%).

In the UTI positive samples, we found *Escherichia, Klebsiella, Enterococcus, Pseudomonas,* and *Aerococcus* urotypes, as well as the general mixed urotype. In the UTI negative dataset, we found the mixed and *Escherichia* urotypes, with the mixed urotype often dominated by *Ralstonia*.

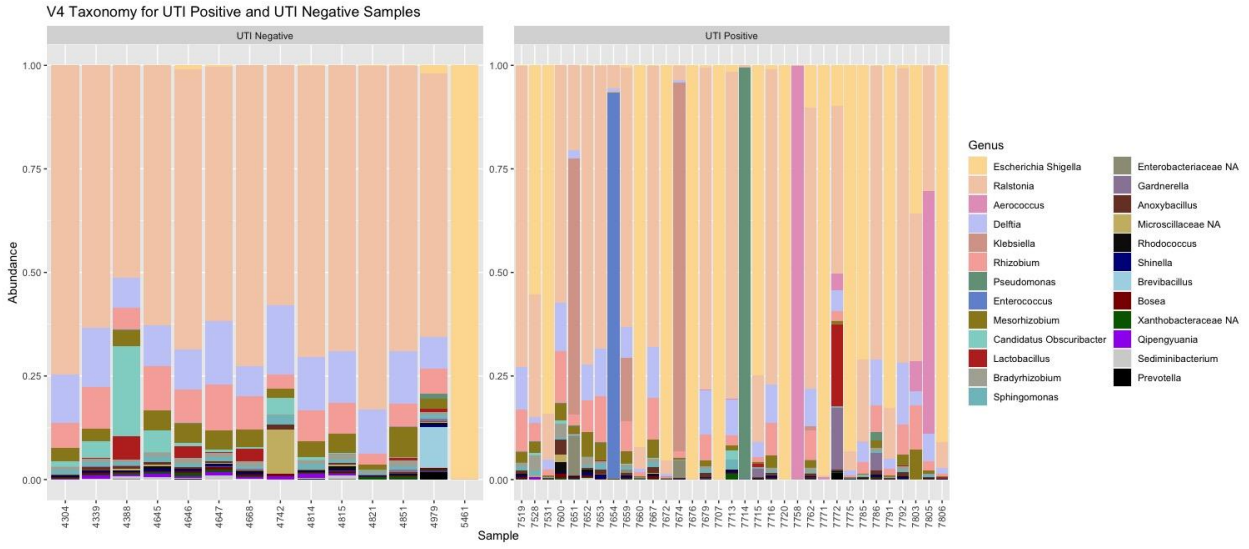

Figure 15. Relative abundances of UTI positive and UTI Negative samples for the V4 region. The top 25 genera from both sets of classifications are plotted.

Comparison of V1-V3, V4 and Full-length 16S Predictions for the UTI Positive Data

Generally, in the UTI Positive data, there were not many differences between the V1-V3, V4, and PacBio classifications. Across all three sets of data analyzed and for all samples examined, 190 different taxa were identified. Because our statistical analysis required non-zeroes as abundances for comparison, 118 taxa (72 were dropped) were compared between V1-V3 and

PacBio, and 123 taxa (63 dropped) were compared between V4 and PacBio. We found that in the V1-V3 results, only one genus, *Ralstonia*, was overrepresented, meaning that it was found more frequently in the V1-V3 results than in the PacBio results. *Rhizobium, Acidovorax, Delftia,* and Burkholderiaceae NA were found to be underrepresented in the V1-V3 results. In the V4 results, *Ralstonia* and Oxalobacteraceae NA were overrepresented, and the same four genera identified underrepresented in the V1-V3 were found to be underrepresented in the V4 data. While the predicted taxa between the two tested variable regions do differ from PacBio, most of the taxa in the UTI positive samples remain significantly similar to PacBio.

Overrepresented V1-V3	**Underrepresented V1-V3**	**Overrepresented V4**	**Underrepresented V4**
Ralstonia	*Rhizobium*	*Ralstonia*	*Rhizobium*
	Acidovorax	*Oxalobacteraceae* NA	*Acidovorax*
	Delftia		*Delftia*
	Burkholderiaceae NA		*Burkholderiaceae* NA

Table 4. Over- and underrepresented genera of variable region classifications compared to PacBio classifications for the UTI positive data.

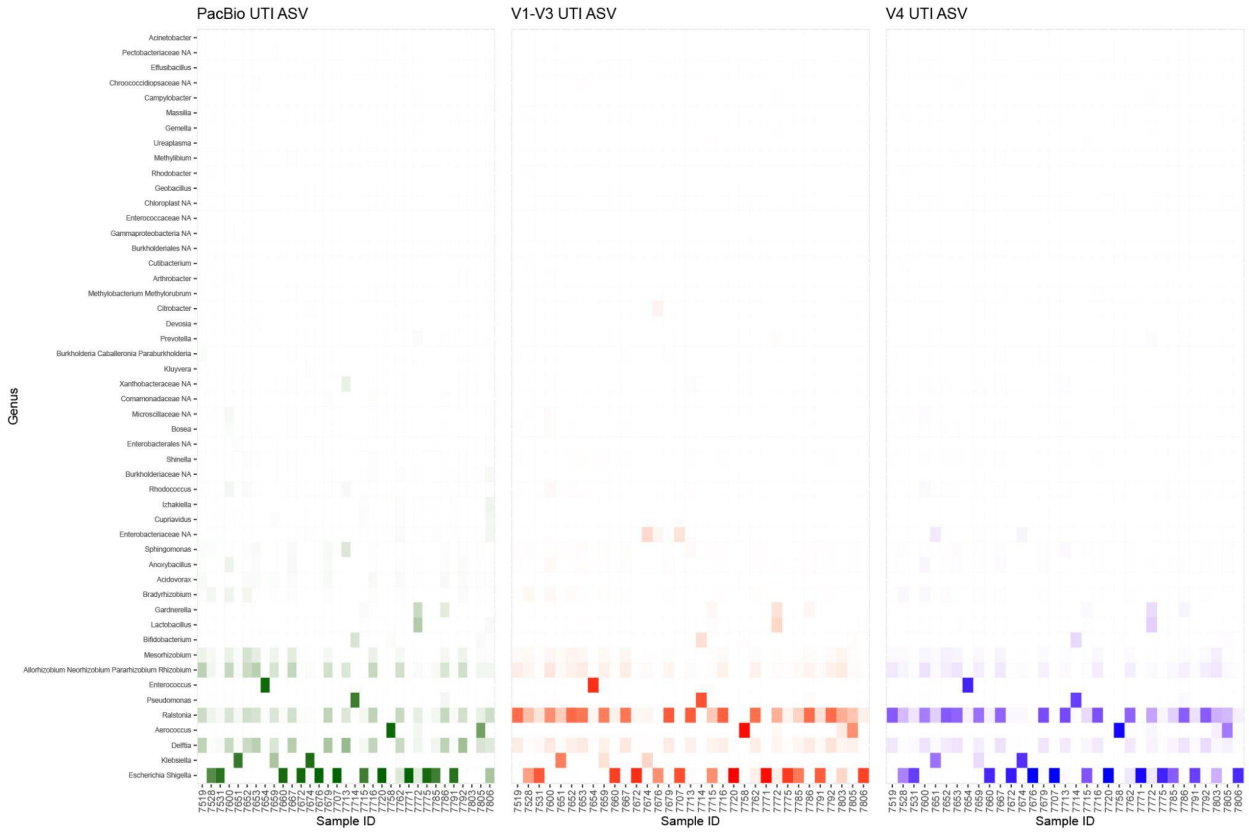

Figure 16. Heatmap of relative abundances of taxa in the UTI Positive samples according to the PacBio data (green), V1-V3 data (red), and V4 data (blue). The top fifty taxa shown are derived from the top 50 most abundant taxa from the PacBio classifications. Darker squares indicate higher relative abundance than lighter squares.

Comparison of V1-V3, V4, and PacBio Predictions for the UTI Negative Data

In the UTI negative data, many more genera were found to be differentially predicted between the PacBio and computationally generated variable regions. In the statistical comparison, 123 taxa (67 dropped) were compared for our V1-V3 versus PacBio analysis, and 131 (59 dropped) taxa were compared between V4 and PacBio. For the V1-V3 classifications, there were 11 genera identified in significantly different abundances than they were in the PacBio analysis (Table 5). Only two of these genera are underrepresented in the V1-V3 data: *Ralstonia* and *Acidovorax*. For the V4 classifications, there were 20 significantly different

classifications from the PacBio data (Table 5). Similar to the V1-V3 data, most (n=18) of these genera are overrepresented in the V4 classifications. The two underrepresented genera were *Ralstonia* and *Acidovorax*, the same two genera that were underrepresented for the V1-V3 region. Table 5 shows the over- and underrepresented genera from the V4 data regarding the PacBio predicted abundances. Of note is the overrepresentation of *Escherichia Shigella*; both variable regions examined are reporting more ASVs for *Escherichia Shigella* than were identified in the PacBio sequencing.

Overrepresented V1-V3	**Underrepresented V1-V3**	**Overrepresented V4**	**Underrepresented V4**
Bosea	*Ralstonia*	*Rhodococcus*	*Ralstonia*
Bradyrhizobium	*Acidovorax*	*Geobacillus*	*Acidovorax*
Xanthobacteraceae NA		*Burkholderiales* NA	
Qipengyuania		*Shinella*	
Devosia		*Anoxybacillus*	
Sphingomonas		*Rhizobiaceae* NA	
Candidatus Obscuribacter		*Bradyrhizobium*	
Anoxybacillus		*Candidatus Obscuribacter*	
Escherichia Shigella		*Devosia*	
		Bosea	
		Xanthobacteraceae NA	
		Sphingomonas	
		Escherichia Shigella	

		Methylobacterium Methylorubrum	
		Qipengyuania	
		Renibacterium	
		Rhizobiacaea NA	
		Methylibium	
		Oxalobacteraceae NA	
		Sediminibacterium	

Table 5. Over- and underrepresented genera of variable region classifications compared to PacBio classifications for the UTI negative data.

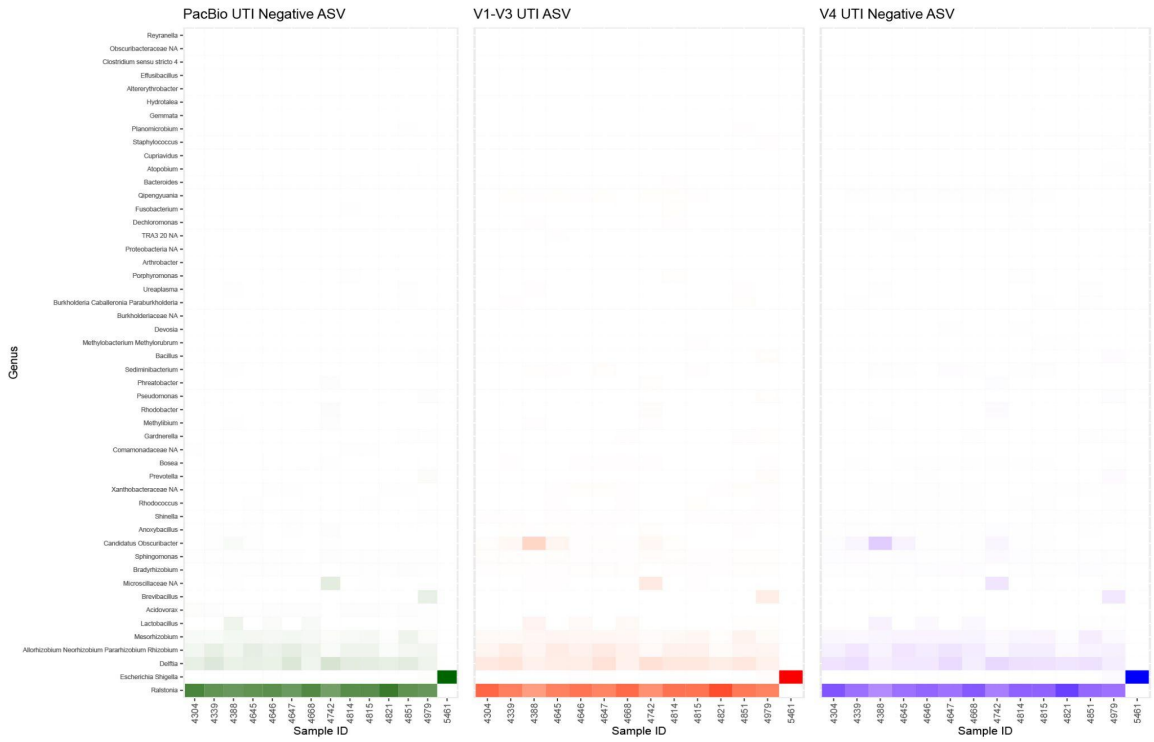

Figure 17. Heatmap of relative abundances taxa in the UTI negative samples according to the PacBio data (green), V1-V3 data (red), and V4 data (blue). The top fifty taxa shown are derived from the top 50 most abundant taxa from the PacBio classifications. Darker squares indicate higher relative abundance than lighter squares

Discussion

Examining the differences between the taxonomic profiles predicted by different variable regions provides insight into the best variable regions for accurate characterization of the urobiome. Furthermore, comparison of the short-read and full-length sequence taxonomic assignments enables us to assess the cost-benefit of long-read sequencing. Prior assessments of the strengths/limitations of the different variable regions for urobiome data has only examined genomes from isolated urinary taxa, with few representatives of different species or strains (30). Based upon this prior work, we expected our computationally simulated 16S variable regions to be similar to the PacBio results and fairly similar to the taxonomic predictions using other variable regions (30). Our results supported that finding. The longer variable region (V1-V3) more closely resembled the PacBio data, but the taxonomies found between variable regions (V1-V3 and V4) were still similar.

The UTI positive samples for PacBio, V1-V3, and V4 all produced the same 6 urotypes. Five of those were dominated by known uropathogens, *Escherichia, Aerococcus, Enterococcus, Pseudomonas,* and *Klebsiella*, that were easily detected in all three analyses. While the most uropathogenic classifications remain the same, the less abundant taxa within them vary greatly. This presents a problem especially for the mixed urotype samples in which several low abundance taxa were not predicted by the variable region or not identified in the PacBio analysis. Because the causal agent of these UTI symptoms is unclear, taxonomic misclassifications increase the noise within these samples.

In the UTI negative data, we found that *Escherichia Shigella* was overrepresented in both V1-V3 and V4 classifications. This was not an issue in the UTI positive data, probably because of the stronger *Escherichia* signal due to many samples containing a large abundance of *E. coli*.

This presents an issue for previous asymptomatic urobiome studies that utilized the V4 region, as overrepresentation of *Escherichia* and the additional noise generated from sequencing the V4 region itself creates skewed results. Interestingly, *Escherichia* was not overrepresented in the V1-V3 region, showing a more accurate representation of the taxonomy of the community.

The Hoffman paper concluded that while the V4 region did not classify correct taxonomy as much as V1-V3 and V2-V3 regions, it was still a valid classification region depending on the depth of taxonomic resolution a study required (30). They found that V4 (with the NCBI 16S database) correctly classified 52 out of 78 at the species level, which they decided is "reasonable" (30). Their study does not capture the variation that is inherent in real urine samples, though, which would lower accuracy due to additional noise in the sequences (30). We found that while our data generally supported their results of V4 being a decent classifier, it was also contingent on the type of community (UTI positive or negative) sequenced. The V4 region results were more similar to the PacBio results for the UTI positive data, than they were for the UTI negative data. This indicates that variable region selection should be based on the type of community sequenced. The UTI positive samples had a much higher abundance of sequences able to be classified, and those samples had less noise because they were often dominated by a single uropathogen. We thus do not recommend V4 sequencing of UTI negative urobiome samples.

Conclusion

In this chapter, we reclassified the PacBio sequences using single variable region data that was computationally created. The most accurate classifications to the PacBio data were the V1-V3 regions of UTI positive data, with 5 significantly different taxa across all classifications. We found that the V1-V3 region more accurately predicted the taxonomies found by the PacBio

sequencing, especially regarding the UTI negative samples. The V4 taxonomic classifications were not as accurate as the PacBio taxonomy, and if the PacBio is treated as the gold standard in this study, therefore cannot be recommended to survey the urobiome. We also found that in UTI negative samples, taxonomy was much more variable than the UTI positive samples, indicating that a lower biomass will create classifications with more noise. Because of this, and the overrepresentation of *Escherichia* in UTI negative samples, we do not recommend V4 sequencing of asymptomatic urobiomes. We recommend V1-V3 sequencing over V4 sequencing for the urinary microbiome for more accurate taxonomic classifications, but ultimately PacBio provides the most accurate taxonomic classifications for 16S gene surveys.

CHAPTER FOUR

CHARACTERIZING STRAIN-LEVEL VARIATION OF URINARY TRACT INFECTIONS USING SHOTGUN METAGENOMIC SEQUENCING

Introduction

In contrast to 16S rRNA gene sequencing, shotgun metagenomics sequences all of the genomic material in a sample. The goal with utilizing shotgun sequencing in the context of the urobiome is to derive more information about the community itself; shotgun sequencing can provide insight into the presence genes other than the 16S rRNA as well as extrachromosomal components and non-prokaryotic constituents like fungi and bacteriophages. Some work has been done characterizing the urobiome with shotgun metagenomic sequencing data. Initial urobiome analyses with shotgun data support generally the major UTI pathogens that have been identified by 16S rRNA gene surveys and isolates including: *Escherichia, Klebsiella, Pseudomonas, Enterobacter and Citrobacter* (92). This study also found that *Gardnerella vaginalis* was represented more in shotgun reads, and in general *Lactobacillus* and *Provotella* species were found more commonly in women (92). Many urobiome studies using shotgun metagenomics are directly interested in applying the technology to healthcare due to the nature of the data; shotgun metagenomic sequencing provides more information than just the taxonomy of the community, including the functionality encoded by the microbiota (103,104).

A benefit to using shotgun metagenomic is that fine-scale variation can be resolved from the sequences. As mentioned in the introduction, none of the contemporary tools for assigning strain level diversity to shotgun metagenomic data are perfectly accurate. Previous work in our

lab has identified STRONG as the best option available (105). In examining urobiome data sets from asymptomatic individuals, strains of the same species could be identified. This finding was pivotal in determining that even though the urinary tract is a low biomass community, especially in asymptomatic individuals, there can be multiple different strains of the same species inhabiting this niche and they can be identified via shotgun metagenomic sequencing.

In this chapter, we use shotgun metagenomic sequencing to assign taxonomy and find strain-level variation in a subset of the samples examined in Chapters 2 and 3 by 16S rRNA sequencing, including 15 UTI positive and 2 UTI negative samples as well as the two mock communities. We compare the taxonomies predicted here to those from full-length 16S rRNA gene sequencing from Chapter 2. Additionally, we conducted strain-level detection via STRONG and examined the strain-detection capabilities of both shotgun metagenomics and full-length 16S rRNA gene sequencing approaches.

Methods

Samples and Sequencing

For our shotgun metagenomic sequencing, we sequenced 15 of the 33 UTI positive samples and 2 of the 14 UTI negative samples giving a total of 17 samples to be analyzed. These samples include: 4821, 5461, 7531, 7651, 7654, 7672, 7676, 7707, 7714, 7720, 7771, 7772, 7775, 7785, 7786, 7791, and 7803. These samples were chosen because the extractions produced sufficient DNA for shotgun sequencing without amplification. The same DNA extractions used for the 16S PacBio sequencing (Chapter 2) were used for shotgun metagenomic sequencing. The 17 samples were sequenced at MIGS (Pittsburgh, PA). Libraries were prepared by MIGS using the Illumina DNA Prep kit and IDT 10bp UDI indices, and sequenced on an Illumina NextSeq 2000, producing 2x151bp reads.

From previous work done in the lab with this data, we knew that these samples contained significant amounts of host DNA from analysis done with MGRast (106). Bowtie2 (v2.4.5) was used to align the reads to a human genome using the Bowtie2 precomputed human host genome GRCh38; aligned reads were removed from subsequent analyses (107).

Taxonomic classification

The tool Kraken2 (v2.08-beta) was used to classify taxonomy with the filtered shotgun metagenomic data (52). It was used with the pre-assembled database Standard-16, found at the tool's GitHub repository (https://benlangmead.github.io/aws-indexes/k2), which was last updated on 9/26/22. The only change to default parameters in the Kraken2 command was the addition of "--use-names" which switched the output from a Kraken2 database taxonomic ID to the scientific name of the species.

Examining Strain-Level Diversity

As mentioned above STRONG (STrain Resolution ON Graphs) was chosen to specifically examine strain-level diversity in the shotgun metagenomic data. Its intended purpose is to compare strains that appear in all samples in an input group, but for our purposes, we compared samples against each other to extract strains present in each sample. Within STRONG, the user defines which algorithms to use to assemble and bin reads. We used SPAdes (v3.14), the default for STRONG, with read length 150 and k=77 to assemble the reads. We chose concoct (v1.0) to be the binning algorithm to create the bins, with a contig size of 1000 (108). After STRONG successfully created bins, the sequences from the bins were compared against the GenBank nr/nt database via BLAST to assign taxonomy (109).

Statistical Analysis

Statistics determining significantly over or under-represented taxa were repeated from Chapter 3 using the Wilcoxon sum rank test.

Results

Mock Community

The two mock community data sets were merged for our analysis here. In total, 53 taxonomic species-level classifications were made that ranged from a general classification of Bacteria NA (n=42), to as specific as species. Because shotgun sequencing has been shown to not capture relative abundance, we focus here instead on what classifications were true to the known community members. When Kraken2 cannot resolve a species-level classification, it classifies the sequence as the next known taxonomic order (53).

Classification	**Count**	***E. coli***	***E. faecalis***	***P. mirabilis***	***S. epidermidis***
Bacillales NA	4	-	-	-	+
Bacteria NA	42	+	+	+	+
Citrobacter amalonaticus	3	-	-	-	-
Citrobacter freundii	2	-	-	-	-
Citrobacter koseri	1	-	-	-	-
Citrobacter NA	3	-	-	-	-
Citrobacter sedlakii	1	-	-	-	-
Enterobacter bugandensis	1	-	-	-	-

Enterobacter cloacae	1	-	-	-	-
Enterobacter hormaechei	2	-	-	-	-
Enterobacter NA	1	-	-	-	-
Enterobacterales NA	116	+	-	+	-
Enterobacteriaceae NA	3278	+	-	-	-
Enterococcus faecalis	21	-	+	-	-
Enterococcus NA	16	-	+	-	-
Escherichia albertii	9	-	-	-	-
Escherichia coli	3754	+	-	-	-
Escherichia fergusonii	6	-	-	-	-
Escherichia marmotae	4	-	-	-	-
Escherichia NA	239	+	-	-	-
Gammaproteobacteria NA	61	+	-	+	-
Klebsiella grimontii	1	-	-	-	-
Klebsiella NA	6	-	-	-	-
Klebsiella pneumoniae	10	-	-	-	-
Klebsiella quasipneumoniae	1	-	-	-	-
Kosakonia NA	2	-	-	-	-

Lactiplantibacillus NA	2	-	-	-	-
Lactobacillus jensenii	1	-	-	-	-
Lelliottia amnigena	1	-	-	-	-
Morganellaceae NA	5	-	-	+	-
Pluralibacter gergoviae	1	-	-	-	-
Proteobacteria NA	16	+	-	+	-
Proteus columbae	1	-	-	-	-
Proteus mirabilis	94	-	-	+	-
Pseudescherichia vulneris	1	-	-	-	-
Raoultella terrigena	1	-	-	-	-
Salmonella enterica	7	-	-	-	-
Salmonella NA	5	-	-	-	-
Shigella boydii	4	-	-	-	-
Shigella dysenteriae	3	-	-	-	-
Shigella flexneri	5	-	-	-	-
Shigella NA	3	-	-	-	-
Shigella sonnei	1	-	-	-	-
Staphylococcaceae NA	5	-	-	-	+
Staphylococcus aureus	4	-	-	-	-

Staphylococcus capitis	3014	-	-	-	-
Staphylococcus caprae	1	-	-	-	-
Staphylococcus epidermidis	134	-	-	-	+
Staphylococcus haemolyticus	14	-	-	-	-
Staphylococcus hominis	3	-	-	-	-
Staphylococcus NA	111	-	-	-	+
Staphylococcus saccharolyticus	1	-	-	-	-
Staphylococcus xylosus	1	-	-	-	-

Table 6. Table of species-level classifications of the mock communities. Pluses and minuses indicate a correct classification or if the classification is a higher order of the species.

When discussing the taxonomy found in the mock communities, it is not as simple as saying that a species was there or it was not, as when Kraken2 cannot classify a species, it moves the classification up a taxonomic order. Because of this, many of the sequences from the mock communities were classified at the Family or Order taxonomic levels (Table 6). For example, every taxonomic level of *E. coli* was classified starting with *Escherichia*, then *Enterobacteriaceae*, Enterobacterales, Gammaproteobacteria, Proteobacteria, and Bacteria (Table 6). From Enterobacterales up until Bacteria, these classifications are shared with *Proteus mirabilis*. We can conclude that the sequences classified to *Enterobacteriaceae* probably belong to *E. coli*, but the other classifications within Enterbacterales, Gammaproteobacteria, and Proteobacteria could belong to either of the two species, although it is important to note that a

significantly more volume of *E. coli* was added to the mock communities during preparation (Chapter 2 Methods). Other than the higher-level classifications of *E. coli*, many species level classifications were made within the family *Enterobacteriaceae* itself: 4 other species were classified within *Escherichia* including *Escherichia* NA, 5 species of *Shigella*, 5 species of *Citrobacter*, 4 species of *Enterobacter*, 2 species of *Salmonella,* 4 species of *Klebsiella*, and 1 species within the genera *Koskania, Lelliottia, Pluralibacter,* and *Pseudoescherichia* each. We can assume that all of those classifications were misclassifications of the *E. coli* sequences as they are within the same family.

Staphylococcus was also misclassified at the species level. Instead of *Staphylococcus epidermidis* (n=134, 1.22%) being the most abundant taxon in the *Staphylococcus* genera, it was *Staphylococcus capitis* (n=3,014, 27.34%). Other *Staphylococcus* species misclassifications include *S. aureus, S. caprae,* and *S. hominis.* Unlike *E. coli* and *P. mirabilis*, all taxonomic levels of *S. epidermidis* were not represented: Staphylococcaceae and Bacillales are the family and order of the genus and are included in our classifications, but higher levels like the class (Bacilli) and phylum (Firmicutes) are not. *E. faecalis* is also a member of the class Bacilli and phylum Firmicutes. *Enterococcus* interestingly did not have species-level misclassifications, just the more general classification of *Enterococcus* NA. *E. faecalis* was practically absent from the mock community classifications (n=21).

For the mock communities, we conclude that the taxonomic classifications from the 16S rRNA gene sequencing (PacBio) were more accurate than those from the shotgun metagenomic data. PacBio predicted the taxonomy of the mock communities with only 0.2% of ASV abundance being misclassifications, and the shotgun data's misclassifications made up 64.8% of the predictive taxonomy assigned.

UTI Positive Sample Taxonomy

The genus and family taxonomic classifications were derived for each UTI positive sample (Figure 18). Five different urotypes, characterized by the identity of the predominant microbe(s) can be identified (110). The first urotype includes samples that are dominated by *E. coli* (Figure 18 left, pink). Samples 7707, 7720, and 7771 have >50% of its taxonomic calls attributed to *E. coli*. We also believe that samples 7775, 7785, and 7791 are likely dominated by *E. coli*. For these three samples, we see *E. coli* as well as Enterobacteriaceae (Figure 18 left, lavender), the taxonomic family for *E. coli*. The second, third, and fourth urotypes have a single representative in our samples, dominated by *Klebsiella* (Figure 18 left, gray) - sample 7651, dominated by *Pseudomonas* (Figure 18 left, olive) - sample 7714, and dominated by *Proteus* (Figure 18 left, red) - sample 7803. The final urotype is the mixed urotype, meaning that no one genus dominates the sample. The remaining samples are assigned to this urotype. It is worth noting that two of these mixed urotypes are in fact dominated by *Ralstonia* (samples 7672 and 7772).

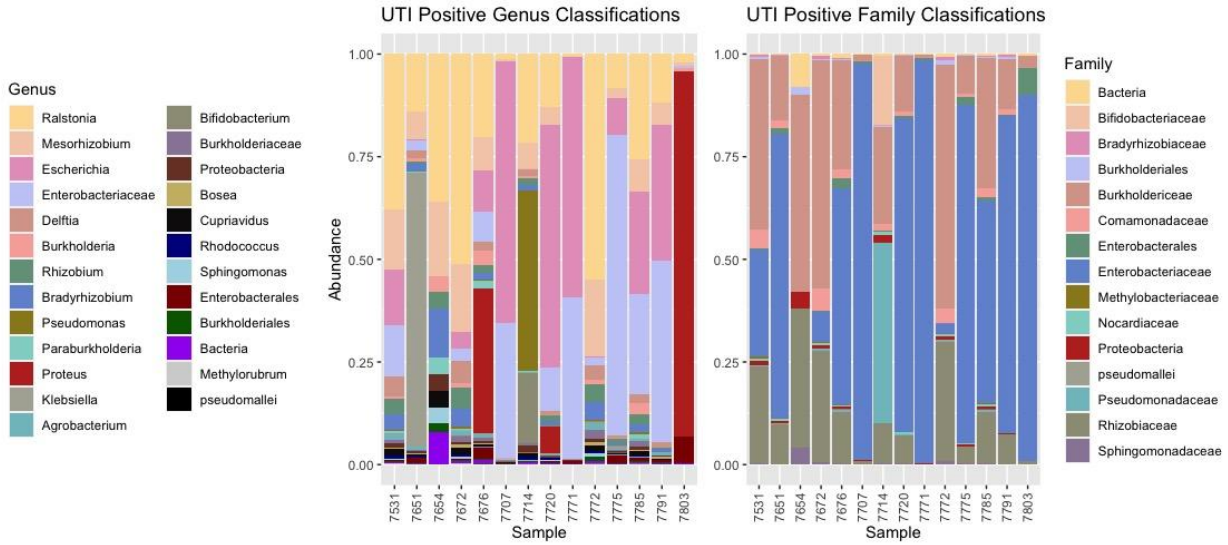

Figure 18. Stacked bar chart of the relative abundances of classifications in the UTI positive samples, genus-level classifications on left and family-level classifications on the right.

In addition to considering the genus calls by Kraken2, we also examined the taxonomic family-level identifications (Figure 18 right). Nine of these samples are dominated by Enterobacteriaceae (Figure 18 right, dark blue) - Samples 7651, 7676, 7707, 7720, 7771, 7775, 7785, 7791, and 7803. These samples include the six we assigned to the *E. coli* urotype (samples 7707, 7720, 7771, 7775, 7785, and 7791) as well as the *Klebsiella* urotype (sample 7651) and *Proteus* urotype (sample 7803). Both *Klebsiella* and *Proteus* are members of the family Enterobacteriaceae. Additionally, we find sample 7676 to be dominated by Enterobacteriaceae. When referring back to the genus classifications for this sample (Figure 18 left), it was a mixed urotype including taxonomic classifications of *Proteus*, *E. coli*, and Enterobacteriaceae. Sample 7714 is dominated by Pseudomonadaceae (Figure 4.1 right, gray), the family for *Pseudomonas*. Samples 7654, 7672, and 7772 are dominated by the family Burkholdericeae (Figure 18 right, peach), the family for *Ralstonia*. Sample 7531 is mixed at the family level, containing Burkholderiaceae, Enterobacteriaceae, and Rhizobiaceae (Figure 18 right, teal).

UTI Negative Taxonomy

There were only two samples sequenced from the UTI negative dataset, samples 4979 and 5461. While sample 4821 is clearly a mixed urotype, not dominated by any one genus (Figure 19 left) or family (Figure 19 right), sample 5461 is dominated by *Ralstonia*. This sample also contains Enterobacteriaceae (Figure 19 right, dark blue), albeit at a low relative abundance. The two UTI negative samples' urotypes resemble the UTI positive samples mixed urotypes dominated by Ralstonia (Figure 19, samples 7672 and 7772), both when considering the genus and family classifications.

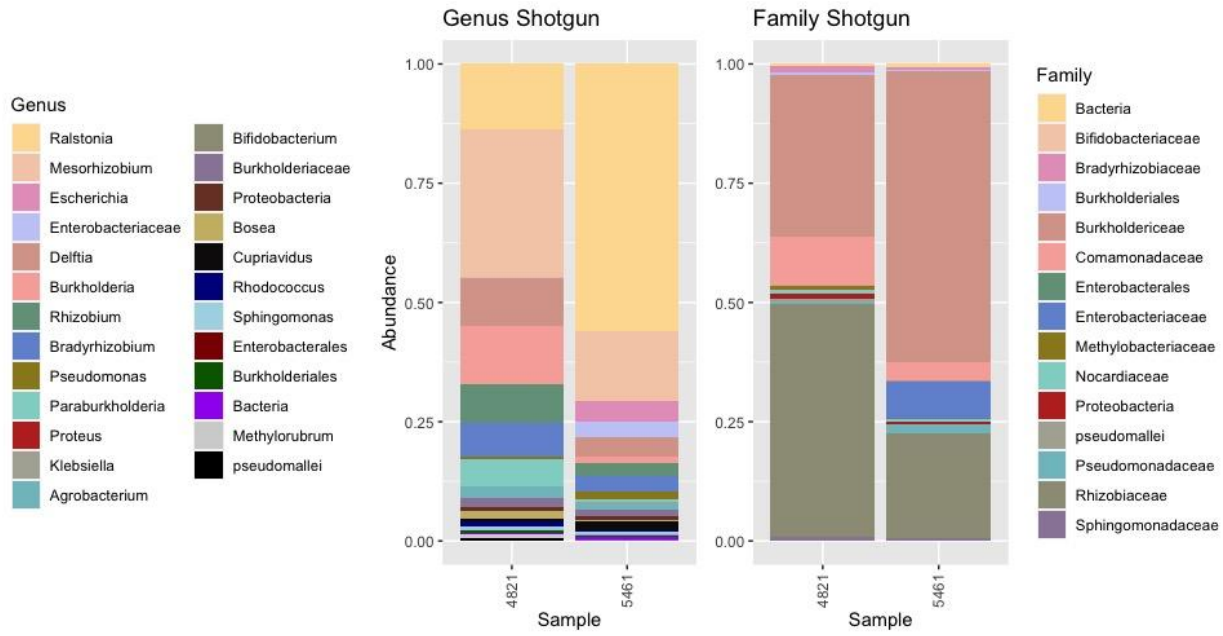

Figure 19. Stacked bar chart of the relative abundances of classifications in the UTI negative samples, genus-level classifications on left and family-level classifications on the right. Note, only two samples from the UTI negative cohort were sequenced using shotgun metagenomic sequencing.

Comparison of PacBio to Shotgun Metagenomic Sequencing Taxonomy

Next, we compared the taxonomic classifications by PacBio to those from the shotgun metagenomic sequencing via Kraken2 analysis. First, we considered the species-level classifications. As shown in Table 7, the number of shotgun taxonomic classifications exceeded those made by PacBio. Moreover, many of these shotgun taxonomic calls were not identified by PacBio. Those taxonomic classifications made only by PacBio were often represented in the shotgun taxonomic calls by close relatives, e.g., *Bradyrhizobium elkanii* was identified via PacBio for sample 4821 and, while it was not identified by Kraken in the shotgun analysis, 17 other *Bradyrhizobium* species were. As is shown in Table 7, in most cases the majority of PacBio genus classifications were found both in the PacBio and Shotgun predictions. The two largest exceptions are Sample 5461 in which >60% of the PacBio ASVs were for taxonomies that were not identified by the shotgun data analysis. Also in sample 7771, >40% of the PacBio ASVs were for taxonomies that were not identified by the shotgun data analysis. Strikingly, there were several samples in which >80% of the shotgun sequence data was assigned to a taxonomic classification that was not predicted by the PacBio analysis: Samples 4821, 5461, 7676, and 7775. These samples exemplify a striking discord in the power of genus-level taxonomic classification by these two approaches.

Sample	# Taxa Identified by Both PacBio & Shotgun	# Taxa Identified by Only PacBio	# Taxa Identified by Only Shotgun
4821	8	5	767
5461	5	1	195
7531	14	17	131
7651	13	22	59

7654	3	12	22
7672	11	11	280
7676	3	4	128
7707	3	0	64
7714	6	2	181
7720	3	5	52
7771	2	1	39
7772	17	31	80
7775	12	10	70
7785	12	19	136
7791	6	21	19
7803	0	0	16

Table 7. Species level classification counts for shotgun and PacBio

Because the shotgun analysis by Kraken2 is limited in its ability to distinguish between species with significant genomic similarity, we chose to reduce the granularity of our comparison and instead look at the genus-level classifications made by PacBio and the shotgun data (Table 8) (53). Again, we see that most of the genera identified by PacBio are also identified by Kraken2. Similar to that observed when considering species-level classifications, we see that Kraken2 has identified many taxa that were not classified from the PacBio 16S data. In total, 28 unique genera were identified only by the PacBio analysis in 11 of the 16 samples. For all but one of these samples, the relative abundance of these "PacBio only" genera consisted of <1% of the relative abundance. For sample 7772, however, these "PacBio only" genera consisted of 4.05% of the relative abundance. 1.63% of these calls were for the genus *Prevotella*. The

remaining 9 genera that were only identified by the PacBio analysis were all found with a relative abundance <1%.

Sample	# Taxa Identified by Both PacBio & Shotgun	# Taxa Identified by Only PacBio	# Taxa Identified by Only Shotgun
4821	10	0	369
5461	2	0	104
7531	12	8	76
7651	12	10	29
7654	5	1	12
7672	10	2	144
7676	3	1	68
7707	1	0	42
7714	4	4	100
7720	2	1	36
7771	1	0	25
7772	16	10	45
7775	12	2	35
7785	13	8	36
7791	8	5	10
7803	0	0	13

Table 8. Genus-level classifications for shotgun and PacBio

The Wilcoxon sum rank test was used to identify significantly different taxonomic classifications between the PacBio and shotgun analyses both at the species level as well as at the genus level. All of the significantly different taxonomic classifications were overrepresented in

the shotgun classifications relative to the PacBio classifications (Table 9). For the species comparison, Bacteria NA, *M. terrae,* and *R. insidiosa* were the significantly overrepresented species. *R. insidiosa* was only identified in one of the PacBio analyses, sample 7772. In contrast, it was identified in all of the shotgun data sets, albeit at a low relative abundance in all samples (<1.5%). The species classification *M. terrae* was not found in any of the PacBio classifications but was found in all 16 of the shotgun data sets (n=10,036). The majority of *M. terrae* classifications were found in sample 4821 (n=7,068, 23.71% of classifications in that sample). The relative abundance of *M. terrae* for sample 7672 was 7.06%; all others were below 5% (0.02%-4.87%).

For the genus-level comparison, *Mesorhizobium* was significantly overrepresented in the shotgun data (Table 9). While *M. terrae* was not identified at the species-level for the PacBio data, other members of this genus were detected in 9 of the 16 samples. The abundance of *Mesorhizobium* in this full-length 16S data, however, was often less than 1%, with the exception of sample 7651 which had an abundance of 2.13%. *Mesorhizobium* was identified in all shotgun data sets. In the shotgun data analyses, the relative abundance of *Mesorhizobium* ranged from 0.10% (sample 7803) to 28.80% (sample 4821), with an average of 8.48% across all 16 samples (Figures 20 and 21).

Species Comparison	**Genus Comparison**
Bacteria NA	Bacteria NA
Mesorhizobium terrae	*Mesorhizobium*
Ralstonia insidiosa	

Table 9. Classifications that were significantly overrepresented in the shotgun classifications compared to the PacBio classifications.

In the shotgun classifications for UTI positive samples, we classified five urotypes: samples dominated by *E. coli, Klebsiella, Pseudomonas,* and *Proteus*, as well as a mixed urotype. In the PacBio classifications, four of the five urotypes were represented, and at different abundances. First, the *Proteus* urotype seen in the shotgun classifications (sample 7803) was not seen at all in the subset of PacBio classifications. Sample 7803 did not have any ASVs classified at all in PacBio, as after it was denoised with DADA2 there were only 6 sequences remaining. The other four urotypes were represented in both datasets: instead of 3 samples (7707, 7720, 7771) being dominated by *E. coli*, 9 samples (7531, 7672, 7676, 7707, 7720, 7771, 7775, 7785, and 7791) were dominated by *E. coli* in the PacBio classifications. Sample 7676 had an abundance of *Proteus* classifications in the shotgun data, although it was classified as mixed because there were not enough to be a majority of the sample, and those *Proteus* classifications appear to be classified as *Escherichia* in the PacBio classifications (Figure 4.5). For the *Klebsiella* urotype, the shotgun and PacBio both classify sample 7651 to have a majority of *Klebsiella*. The *Pseudomonas* urotype was also maintained in both datasets in sample 7714. As for the mixed urotype, most of the mixed urotype shotgun classifications were classified as *E. coli* in PacBio. Samples 7531, 7654, 7672, 7676, and 7772 were all considered the mixed urotype in the shotgun classifications, and they all follow the *E. coli* urotype pattern in the PacBio data.

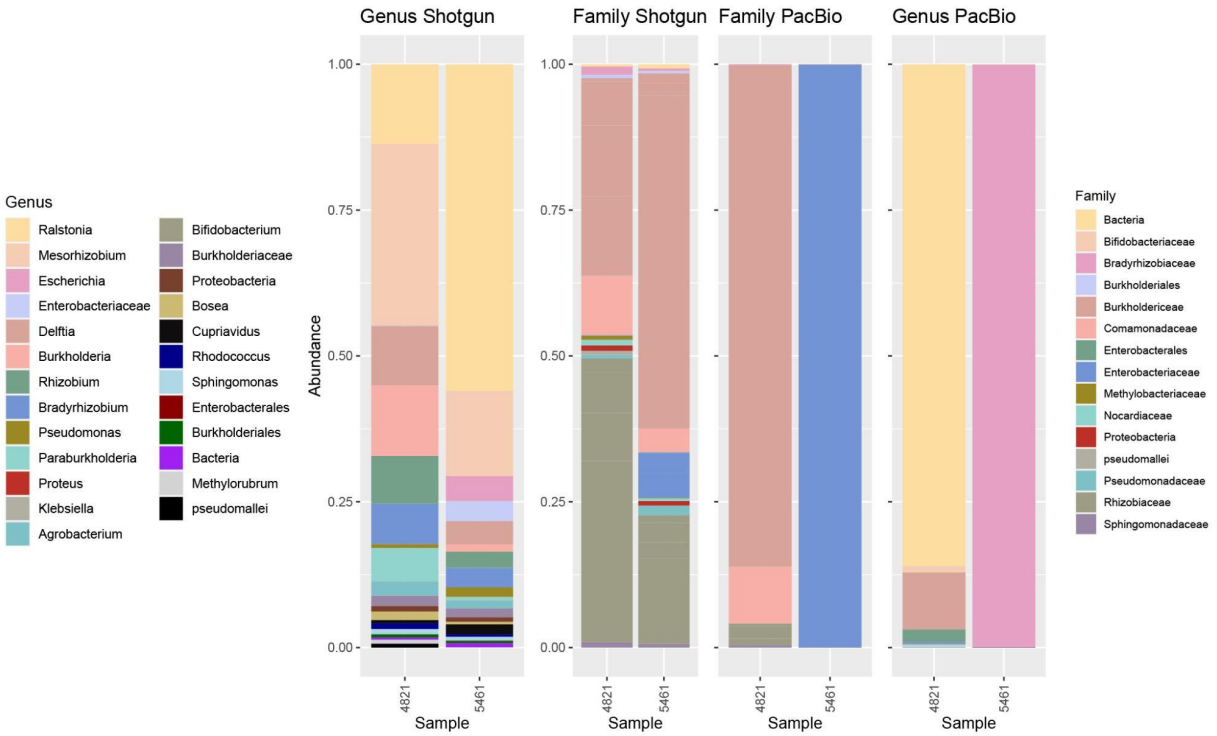

Figure 20. Stacked bar chart of genus and family level classifications for shotgun and PacBio data for UTI negative samples.

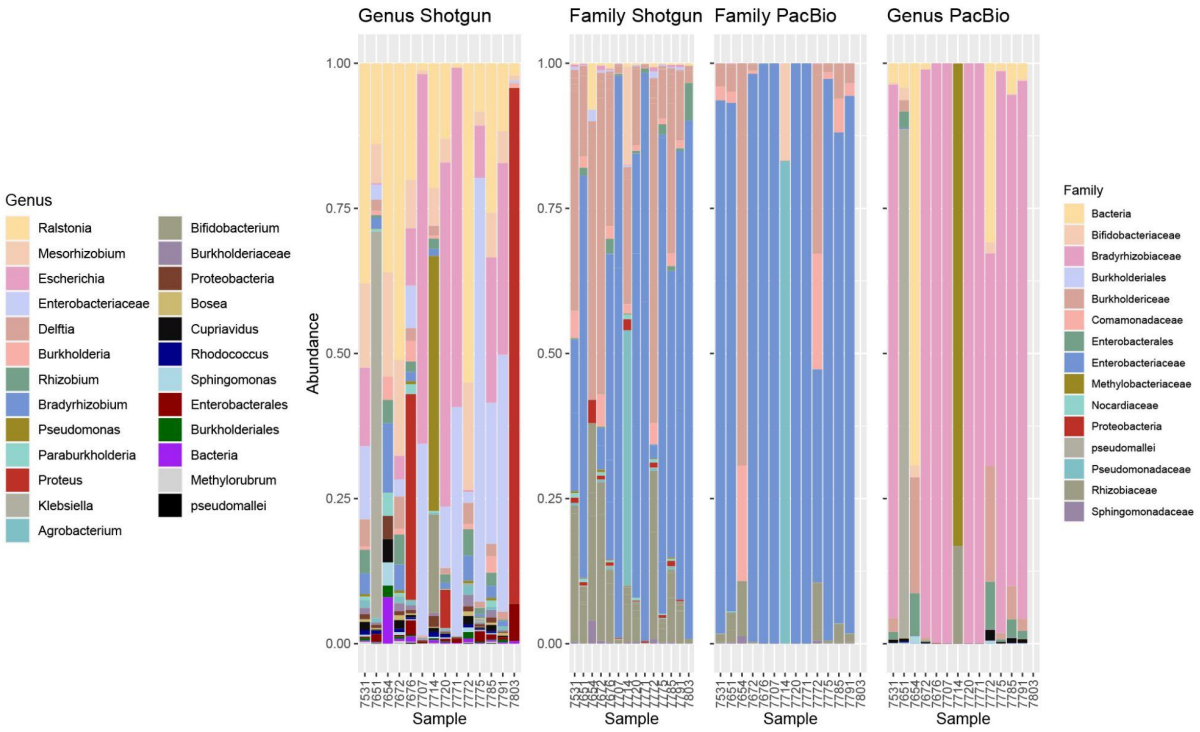

Figure 21. Stacked bar chart comparing genus and family level classifications between shotgun and PacBio data for UTI positive samples.

Strain Level Classification of Mock Communities

Next STRONG was run on the shotgun metagenomic data sets to resolve strain level diversity. First the mock communities were compared against each other in STRONG as a proof of concept. STRONG successfully found all species added to the mock communities: *E. faecalis*, *P. mirabilis, S. epidermidis,* and *E. coli*. Specifically, with the *E. coli* bins, four strains were found (Figure 22). It is important to note that five different strains of *E. coli* were included in the mock communities, representative of four different *E. coli* phylotypes. STRONG only reports four different strains (Figure 22). We thus posit that STRONG was unable to resolve the two

closely related phylogroup B2 strains *E. coli* UMB1162 and *E. coli* UMB1220. This proved that STRONG could detect strain diversity within a species.

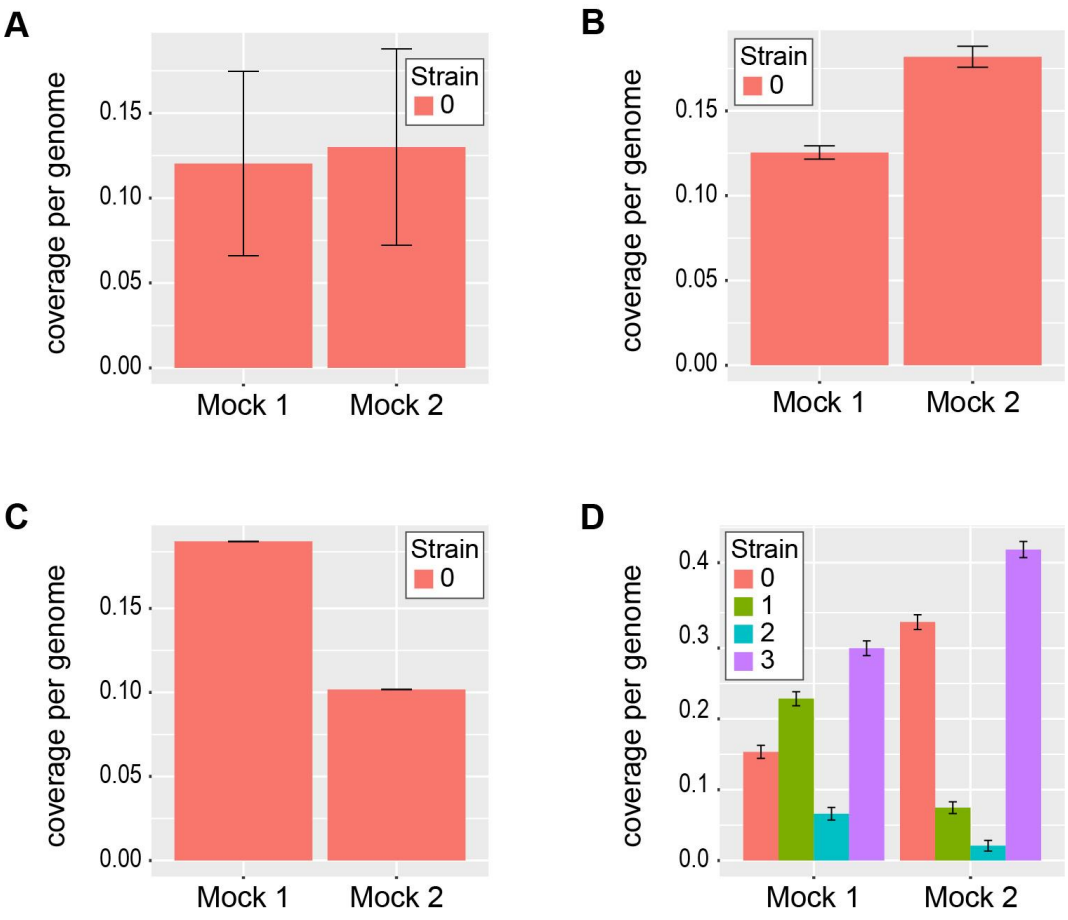

Figure 22. Strain-level detection of constituents of the two mock communities. (A) *E. faecalis*, (B) *P. mirabilis*, (C) *S. epidermidis*, and (D) *E. coli*.

Strain Level Classifications of Experimental Samples

Out of the 16 urobiome samples sequenced via shotgun metagenomics, five binned successfully. Four of those were UTI positive samples and one was UTI negative. As shown in the figure below, of the five samples that binned successfully, all of them were determined to only have one bin and one strain within them. Four of the five samples are classified as *E. coli* (including the UTI negative sample 5461), and the fifth was classified as *Bifidobacterium breve*

(Figure 23). Below are the STRONG results and the PacBio sequence variants associated with the stain classifications (Figure 23 and 24). As shown in Figure 24, there does not appear to be a correlation between the number of sequence variants found and the strain classification from STRONG. The number of *E. coli* or *B. breve* sequence variant classifications range from 22 to 71 (Figure 24). Figure 24 also illustrates that even though only one strain was assigned from STRONG, only three of the samples had a sequence variant that was the majority of the classifications within the specific species (samples 5461, 7714, and 7720).

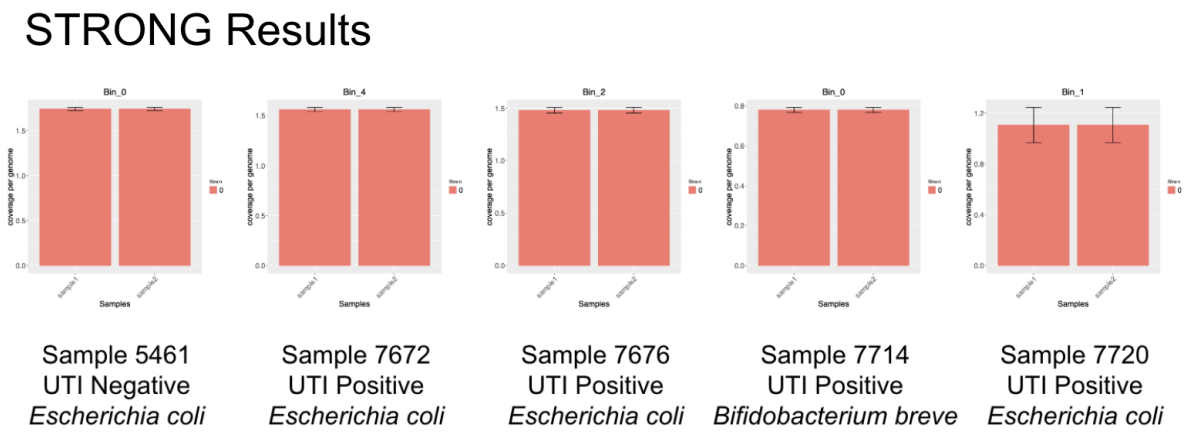

Figure 23. Bins created by STRONG for all samples.

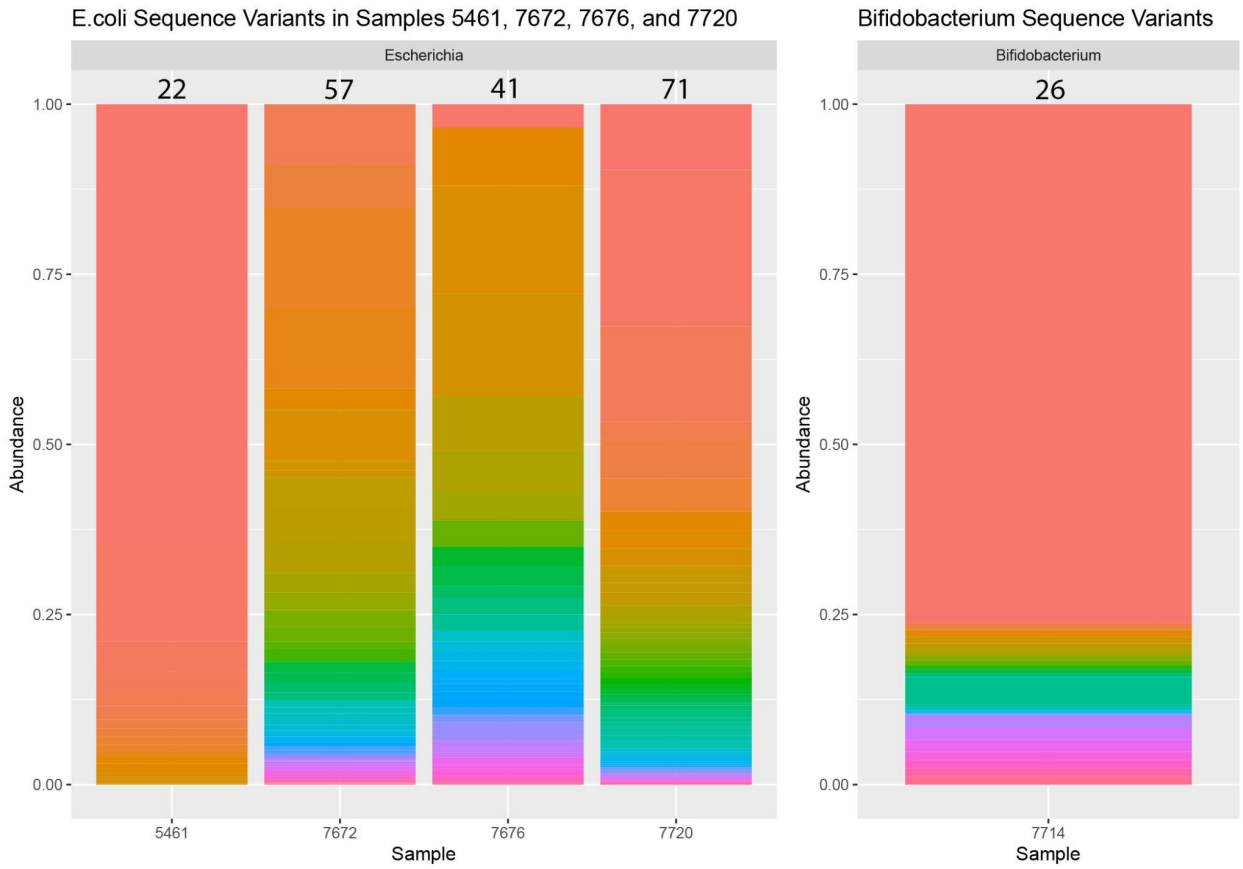

Figure 24. Stacked bar charts showing the relative abundances of *E. coli* and *B. breve* sequence variants from the samples that binned in STRONG. Number of sequence variants is included on the top of the bars per sample.

The classification of a *B. breve* strain in sample 7714 invites investigation into sample 7714 from all results of previous taxonomic sequencing (Figure 23, 24). The most abundant ASV in that sample from the PacBio data is *P. aeruginosa* (n=22,660), followed by *B. breve* (n=4,060). In the shotgun classifications, the most abundant classification was *Pseudomonas* NA (n=1,083) followed by *P. aeruginosa* (n=881) and then *B. breve* (n=515).

One trait that samples 7672, 7676, and 7720 according to PacBio data is that they all have large relative abundances of *E. coli* (ranging from 67.76% to over 99%). According to the shotgun classifications though, *E. coli* was not the majority (>50%) of classifications made,

except for in sample 7720. In Chapter 2, it was determined that sample 5641 had more *E. coli* abundance than some of the UTI positive samples (n=19,407 and *E. dysenteriae* n=38,170), though being collected as a UTI negative sample. In the shotgun classifications though, only 119 classifications of *E. coli* were made in the sample, making up only 3.9% compared to the over 99% relative abundance *Escherichia* classification from PacBio. Similarly to sample 7714, although being less relatively abundant in the shotgun classifications than *Ralstonia* and *Mesorhizobium* species, an *E. coli* strain was classified. In sample 7672, PacBio classified 74.53% of ASVs as *E. coli* (n=27,162), and shotgun sequencing found 3.54% of classifications as *E. coli* (n=228). Sample 7676 classified 67.76% of ASVs as *E. coli* with PacBio (n=11,485) and shotgun classified 8.45% of sequences as *E. coli* (n=143). Lastly, sample 7720 was 89.79% *E. coli* according to PacBio (n=60,152) and 56.24% of classifications in shotgun inferred taxonomy (n=685).

Discussion

Currently in the field of metagenomics, shotgun metagenomic sequencing is considered the gold standard for surveying microbial communities because of its depth of sequencing and ability to examine extrachromosomal components. In our survey of our mock communities, UTI positive, and UTI negative data, we found that the taxonomy predicted is likely overfitting the biological data. We also found that the software tool STRONG successfully found specific strains in urobiome data, although it did not find varying abundances of different strains within the experimental samples.

Mock Community Taxonomic Assignment

The mock community taxonomy found from shotgun sequencing was the first indication of species and genera misclassifications. These misclassifications can be confirmed because we

know exactly what was added to the mock communities, and their relative abundances. While only 4 species were included in these mock communities, Kraken2 generated 53 different taxonomic classifications. These classifications ranged from as general as "Bacteria" to specific species-level. The most abundant classification was *E. coli*, which was correct, but the next most abundant classification was Enterobacteriaceae NA, meaning that only family-level taxonomy could be assigned to those sequences. Because *E. coli* and *P. mirabilis* belong to the family Enterobacteriaceae, these sequences could be from any of the five *E. coli* strains or the *P. mirabilis* strain in either community. The genera and species that were found in the classifications but not in the mock communities were mainly from the family Enterobacteriaceae (*Citrobacter, Enterobacter, Klebsiella, Koskania, Lelliottia, Pluralibacter, Pseudoescherichia, Salmonella,* and *Shigella*), suggesting that Kraken2 incorrectly identified these taxa rather than the *E. coli* or *P. mirabilis* in the mock community.

The majority of *S. epidermidis* was misclassified as *S. capitis*. *S. capitis* is closely related to *S. epidermidis* and a member of the *S. epidermidis* group (111). An interesting result from the mock communities was the apparent lack of *E. faecalis* entirely. *E. faecalis* was classified 21 times, and *Enterococcus* NA was classified 16 with no other *Enterococcus* species being found. The family and order classifications of *Enterococcus*, Enterococcaceae and Lactobacillales, were not classified at all. *Enterococcus* shares its phylum classification with *Staphylococcus*, Firmicutes, which was also not found. The lack of *E. faecalis* invites investigation into the tool used to assign taxonomy: Kraken2 does document that classifications of highly similar species are a shortcoming in the tool, as similarity in species sequences are sometimes too minute to separate (53). In those cases, taxonomy is assigned at the genus level. If this happened to *E. faecalis* sequences, we would expect more classifications as *Enterococcus* NA, but *Enterococcus*

NA (n=16) was classified less than *Enterococcus faecalis* (n=21). A possible explanation for this would be *E. faecalis* being classified simply as Bacteria NA. This lack of classification for *E. faecalis* does not seem to extend to the phylum level, though, because *Staphylococcus* was found in high abundance. This is a finding that impacts the inferred shotgun sequencing taxonomy going forward in our analysis: *Enterococcus* was simply not found by the classifier even though we know it was in the sample.

Generally from the mock communities, we see that the bulk of classifications were either not able to be classified to the genus or species level or were misclassified at the genus or species level completely. Comparing the taxonomy assignment from shotgun sequencing and PacBio sequencing, we see that the PacBio mock community classifications were much closer to the truth.

Experimental Samples Taxonomy Assignment

In the UTI positive samples, we found high abundances of *Ralstonia, Mesorhizobium,* and *Escherichia*, but we also found a high abundance of sequences that were not classified past the family classification of Enterobacteriaceae. There were also samples where previously identified uropathogens predominated, like *Klebsiella* (also an Enterobacteriaceae), *Proteus* (also an Enterobacteriaceae) and *Pseudomonas* (25,112). Other uropathogens like *Enterococcus* and *Aerococcus* were not found in high abundance. In the UTI negative samples, Kraken2 found an abundance of *M. terrae*, followed by *Ralstonia* NA, *Delftia* NA, and *Rhizobium* NA.

Shotgun sequencing reported many more taxonomic classifications than the PacBio data. On average, shotgun sequencing assigned 139.94 more species-level classifications than PacBio. The most different sample between PacBio and shotgun classifications was sample 4821 with 767 more classifications from shotgun than PacBio, which is interesting because that sample

represented the mixed urotype from an asymptomatic urobiome. Although we cannot directly compare abundances between PacBio and shotgun as shotgun classifications do not accurately capture relative abundance, we can infer that the differences between number of taxa assigned show shotgun sequencing overfitting the sequences. The vast difference between the shotgun and PacBio taxonomy is hard to capture, as our statistics can only compare the taxonomic classifications found in abundance in both samples. Another way to examine differences between the sequencing technologies is our classification of urotypes. PacBio determined 2 more urotypes than shotgun (*Enterococcus* and *Aerococcus*), and one less (*Proteus*). Note, the samples that were the *Aerococcus* urotype in PacBio were not sequenced with shotgun sequencing, so a comparison cannot be made. This means that between the sequencing types, two samples changed urotypes: sample 7654 changed from *Enterococcus* to a mixed urotype, and sample 7803 did not have any classified ASVs in PacBio and was classified with shotgun as *Proteus*. The sample 7654 urotype change makes sense because the shotgun classifications barely found any *Enterococcus* in the experimental samples (n=117 across all samples), as well as not being found in high abundance in the mock communities, where we know it was added.

As for similarities between the datasets, we can also examine urotypes. We determined five urotypes in the UTI positive data from shotgun sequencing: *Escherichia, Klebsiella, Pseudomonas, Proteus*, and mixed (which always contained *Ralstonia, Mesorhizobium,* and *Delftia*). Shared between PacBio and shotgun sequencing are the *Escherichia, Klebsiella, Pseudomonas,* and mixed urotype. The similarities in urotype show that although the number of classification varies between the types of sequencing technology, the majority of classifications in 14 of the samples were the same.

The genera identified in our UTI positive samples are similar to those found in a study by Moustafa et. al (92) in which shotgun metagenomic sequencing was conducted for 49 UTI positive samples. In that study, taxa were assigned using the Human Longevity Inc. database using an Expectation Maximization algorithm, and they reported an average of 41 bacterial strains per sample, and samples were split into three clinical groups (92). They found *Escherichia*, *Klebsiella*, *Pseudomonas*, *Enterobacter*, *Citrobacter*, *Acidovorax*, *Ralstonia*, *Aerococcus, Proteus, Gardnerella,* and *Bifidobacterium*, among many others. The classifications of *Ralstonia* and *Bifidobacterium* were supportive of our findings as they have not been extensively examined in regard to the urobiome, so it is encouraging to see them documented in another study examining UTI positive urobiomes. They did not classify it as abundantly as we did, although this comparison is difficult to make because *Ralstonia* is included in all three of their clinical clusters and genera were not reported as a raw count or relative abundance.

In yet another shotgun metagenomic sequencing study of the urobiome, 43 culture positive and 43 culture negative samples were sequenced and classified using Kraken2 resulting in 200 different taxonomic classifications. In the culture positive samples, the species *E. coli, P. aeruginosa, E. faecalis, P. mirabilis, K. pneumoniae*, and *A. urinae* dominated over 90% of all samples (105). All of these classifications were found in high abundance in our shotgun data, except for *Aerococcus urinae* (n=1 in sample 7772). In the culture negative samples, *G. vaginalis, E. coli, K. pneumoniae, Acinetobacter lwoffii, S. agalactiae, S. epidermidis, Bifidobacterium dentium, P. aeruginosa, L. gasseri, L. crispatus, L. jensenii,* and *E. faecalis* were found. In our UTI negative samples, we found 8 of the 12; we did not find *A. lwoffii, S. agalactiae, L. gasseri,* or *B. dentium,* but we did find species within those genera. This paper

shows that even though the same tool was used to assign taxonomy, taxonomic classifications vary at the species level, particularly in UTI negative samples.

Strain Detection

A key objective of this study was to ascertain if UTIs were the result of monoclonal colonization or multiclonal colonization. The analysis of the mock community data using STRONG confirmed that *E. coli* phylogroups could be distinguished from shotgun metagenomic sequencing data. Examination of the UTI positive and UTI negative samples, however, revealed that there are relatively few species sequenced at a depth such that single copy number genes could be identified with sufficient representation such that STRONG could confidently predict a species' presence. Strain diversity within the samples that had enough definition was not found. Four of the five samples binned into *E. coli* strains, which is to be expected given previous knowledge about UTIs and the previous high abundances in the PacBio data.

Conclusion

In this chapter, we examined the classifications created from shotgun metagenomic sequencing of UTI positive and UTI negative samples using Kraken2 and compared taxonomy to the previous PacBio classifications of previous chapters. We also examined strain-level diversity of the samples using STRONG. According to our mock communities, the inferred shotgun taxonomy was not as accurate to the biological truth as the PacBio taxonomy. A large number of sequences were misclassified within the correct family, and many sequences could not be resolved to the genus or species level. We also found that in the experimental data, the shotgun classifications greatly outnumbered the PacBio classifications, with the shotgun classifications having extremely low abundances (n=1). This, in addition to the knowledge from the mock communities, shows that shotgun sequencing is not as accurate in finding taxonomy as full-

length 16S sequencing in the urinary microbiome. And although STRONG did find 4 of 5 strains in the mock community (which shows that it cannot distinguish between phylotypes), it did not successfully find strain-level diversity in the experimental samples. We can assume that different strains of species are present in the experimental samples because of the PacBio sequence variant results, and the lack of strain detection within samples shows that the tool is insufficient for strain-level classification in regard to urobiome samples. Furthermore, strains should have theoretically been detected in all samples as we know that each sample had a biological community, and that was not the case. Shotgun metagenomic sequencing is a powerful tool for surveying the functional capacity and extrachromosomal features of a microbiome, but we found that the taxonomy assigned from current sequencing and tools was not as effective as full-length 16S sequencing.

CHAPTER FIVE

CONCLUSIONS AND FUTURE DIRECTIONS

Comparison of Taxonomy Found between Methods

The first aim of this thesis was to compare the taxonomy found between long-read 16S sequencing to 16S single-variable region sequencing, and long-read 16S sequencing and shotgun metagenomic sequencing. Comparison of 16S single-region sequencing to shotgun sequencing directly was not done because 16S single-region sequencing does not reach the specificity that PacBio or shotgun sequencing has been shown to reach. We found that the best technology for taxonomy assignment for our urinary microbiome data was PacBio sequencing.

In the mock community classifications, PacBio was the most accurate of all three methods. The analysis of the mock community PacBio sequences identified 17 species that were not in the community, most of which fell under the umbrella of *E. coli*. At the genus level, 2 genera were misclassifications, which only made up 0.2% of abundance. In the mock community analysis with single variable region data, the V1-V3 classifications were more accurate to the biological truth and the PacBio classifications. The V4 classifications found all of the correct taxonomy as well, but there were more additional classifications not added to the communities and more ASVs were classified as NA at the genus level. Analysis of the shotgun metagenomic sequencing of the two mock communities identified all four of the mock community species in addition to 49 other classifications. These additional taxonomic predictions include higher-order taxonomic predictions as when species-level taxonomy cannot be assigned by Kraken, the next order of certain taxonomic classification is. Interestingly, Kraken did not identify the *E. faecalis*

strain at all. This is concerning as *E. faecalis* is a known uropathogen, frequently associated with recurrent UTIs, and also is frequently missed during culture-based diagnoses (113).

In the UTI positive data, PacBio sequencing found abundances of previously documented uropathogens, including *E. coli* (the most abundant ASV), *A. sanguinicola*, *K. pneumoniae*, *P. aeruginosa,* and *E. faecalis*. These uropathogens defined the 6 urotypes assigned within the UTI positive data: *Escherichia, Aerococcus, Klebsiella, Pseudomonas,* and *Enterococcus,* as well as a more general "mixed" urotype where no genus was the majority of classifications. Both variable regions found the same six urotypes as PacBio sequencing in the UTI positive samples. Within the mixed urotype, though, PacBio found more *Delftia* and *Rhizobium*, and both variable regions both found more *Ralstonia*. Shotgun metagenomic sequencing of a subset of the UTI positive urobiomes classified many more taxa than PacBio. The largest difference was seen in sample 4821, with 767 classifications made in shotgun but not in PacBio. Shotgun sequencing found 5 urotypes: *E. coli*, *Klebsiella*, *Pseudomonas*, *Proteus*, and mixed, 4 of which were shared with PacBio. The PacBio-classified *Aerococcus* urotype was not sequenced via shotgun sequencing. The PacBio-classified *Enterococcus* urotype was classified as mixed urotype in the shotgun data. This concurs with the observation made in the analysis of the mock communities: Kraken2 fails to identify urinary strains of *Enterococcus*.

PacBio 16S sequencing of the UTI negative data identified two urotypes: *Escherichia* and mixed. The *Escherichia* urotype was only observed in sample 5461, which had enough *E. coli* abundance to be considered a UTI, but the female participant was asymptomatic. The urotype for the UTI negative samples, similar to those from the UTI positive samples, showed a similar composition of *Ralstonia, Delftia,* and *Rhizobium* across UTI positive and UTI negative samples.

The variable region data and shotgun data all found the two urotypes, and sample 5461 was always the only sample that could be classified as *Escherichia*.

Comparison of Strain Level Diversity

The second aim of this thesis was to examine the strains found in the urinary microbiome with long-read sequencing and shotgun metagenomic sequencing. There was more preliminary work on strain diversity using shotgun sequencing, but with the depth and specificity that PacBio sequencing now possesses, determining strains via PacBio is worth investigating. Before addressing the similarities and differences in strain level diversity found, it is important to note that the single variable region data will not be included in this discussion. Short-read 16S rRNA gene sequence classifications cannot resolve species, let alone strains.

The discussion of strain-level variation between PacBio and shotgun sequencing is not as binary as taxonomic assignment. The mock community classifications from both types of sequencing show quite different results. In the PacBio classifications, 45 *E. coli* sequence variants were found, representing 5 distinct strains of *E. coli*. Considering that *E. coli* has seven different copies of the 16S gene this suggests that perhaps the "purified" *E. coli* isolates were not as pure as thought, including one or more very closely related strains and thus additional variants. STRONG identified four strains of *E. coli*. As the mock communities contained five different strains, representative of four different phylogroups, this result suggests that STRONG cannot distinguish between phylogroups. Thus, strain-level variation could be detected, but not phylotype-variation within the strains.

In the PacBio UTI positive results, there were 393 distinct *E. coli* sequence variants. While some samples only had a few *E. coli* sequence variants, some had many more than the mock communities. For instance, sample 7720 included 71 different *E. coli* sequence variants.

While STRONG did identify a strain of *E. coli* in this sample, it did not identify more than one strain. Based on observations from the analysis of the mock communities, we hypothesize that this sample contains more than one strain, although these strains are likely from the same phylotype. Per PacBio sequencing, sample 7720 is an *E. coli* urotype, dominated by *E. coli*. Thus, the inability to distinguish between strains is not likely due to underrepresentation of the different strains in the sequencing data. Our investigation illuminates the limitations of STRONG to distinguish similar genomes.

Future Directions and Recommendations

When I started this project, based on previous literature I assumed that shotgun metagenomic sequencing would be more effective at taxonomic assignment and strain level detection. This was not the case. In the mock communities, PacBio-based taxonomic classifications were more accurate than the shotgun. In the urobiome data, *Enterococcus*, a proven uropathogen, was not found in the shotgun classifications, when I knew it was present in the samples and over 50% of ASVs in sample 7654. Based on these results, I can conclude that full-length 16S sequencing is better at classifying taxonomy than shotgun sequencing. Furthermore, more work needs to be done regarding the lack of *Enterococcus* being found in the shotgun data if shotgun metagenomics is being used to find uropathogens.

In our third chapter, I compared PacBio sequencing to single variable region 16S sequencing. This is not because I thought that single variable regions could provide as much insight as PacBio (i.e., look for strains), but I wanted to see the similarities and differences in the taxonomies found. Single variable region analysis identified many taxa that were not present in the mock communities, for which we know the constituents, as well as in the urobiome samples, including taxa that have not been isolated from the urobiome. These false positive results could

be misleading for researchers and generate noise when trying to determine the microbial "signal" of lower urinary tract symptoms or urinary tract health. Now that PacBio runs are economical and error rates have been drastically reduced and there are software tools for analyzing PacBio long-read 16S rRNA sequence data, future taxonomic surveys of the urobiome as well as other ecological niches should rely on full-length sequencing rather than short-read.

I next attempted to find strain-level diversity in urobiome communities. The results from the mock communities indicated that shotgun sequencing could differentiate strains, because four of the five *E. coli* phylotypes were found as well as all the other species in the community. However, it did not continue its success with the experimental samples. Five of the 16 shotgun samples were binned by STRONG, meaning that it identified a species present. It did not identify more than one strain. Given the number of sequence variants found from the PacBio sequencing per species, we hypothesize that there are more than one strain present in many of our samples. I found that sequence variants do not directly correlate to strains in the PacBio classifications, as we included 5 *E. coli* phylotypes in the mock communities and we found 45 *E. coli* sequence variants. Future research into how to take these sequence variants and ascertain how many strains are present is needed. Given that this is only possible with full-length 16S rRNA sequencing, this is a new challenge for bioinformatics.

Lastly, it is important to address the genera and species found in the urobiome samples. Perhaps the most interesting taxonomic classification across all three chapters was the consistent presence of *Ralstonia* in the mixed urotype samples. *Ralstonia* has been observed in urobiomes with 16S and shotgun metagenomic sequencing, but it has yet to be isolated via culture methods from urine samples. Future studies should investigate the role of *Ralstonia* in the urobiome, as maybe it has not been previously widely discovered because it cannot be cultured using EQUC

procedure. We also found an abundance of *Delftia* in UTI negative samples in the PacBio classifications specifically. Common uropathogens were identified: *E. coli, P. aeruginosa, P. mirabilis, S. epidermidis, A. sanguinicola,* and *E. faecalis* (found primarily with PacBio).

In conclusion, future studies into the taxonomy of complex communities should avoid short-read sequencing all together. PacBio far outperforms single variable region sequencing, and the mock communities were instrumental in highlighting the limitations of individual variable regions. While I originally thought that shotgun metagenomic sequencing would generate the best representation of the diversity within the sample, there were a lot of false positive calls. Furthermore, I thought that shotgun metagenomics would allow me to detect strains, but this was not possible with STRONG on the urobiomes despite our success with the mock communities. Improved bioinformatics tools are needed to distinguish strains, particularly when the strains have significant genomic similarity. An outstanding question surrounds the question of the presence of *R. pickettii*. This species was seen in all the samples but in low abundance in the negative control. This leads us to ask if it is a contaminant or a true member of the urobiome.

BIBLIOGRAPHY

1. Cullen CM, Aneja KK, Beyhan S, Cho CE, Woloszynek S, Convertino M, et al. Emerging Priorities for Microbiome Research. Front Microbiol. 2020;11:136.

2. Askarova S, Umbayev B, Masoud AR, Kaiyrlykyzy A, Safarova Y, Tsoy A, et al. The Links Between the Gut Microbiome, Aging, Modern Lifestyle and Alzheimer's Disease. Front Cell Infect Microbiol. 2020;10:104.

3. Durack J, Lynch SV. The gut microbiome: Relationships with disease and opportunities for therapy. J Exp Med. 2019 Jan 7;216(1):20–40.

4. Manor O, Dai CL, Kornilov SA, Smith B, Price ND, Lovejoy JC, et al. Health and disease markers correlate with gut microbiome composition across thousands of people. Nat Commun. 2020 Oct 15;11(1):5206.

5. Kootte RS, Levin E, Salojärvi J, Smits LP, Hartstra AV, Udayappan SD, et al. Improvement of Insulin Sensitivity after Lean Donor Feces in Metabolic Syndrome Is Driven by Baseline Intestinal Microbiota Composition. Cell Metab. 2017 Oct 3;26(4):611-619.e6.

6. Deo PN, Deshmukh R. Oral microbiome: Unveiling the fundamentals. J Oral Maxillofac Pathol JOMFP. 2019;23(1):122–8.

7. Byrd AL, Belkaid Y, Segre JA. The human skin microbiome. Nat Rev Microbiol. 2018 Mar;16(3):143–55.

8. Li JJ, Yi S, Wei L. Ocular Microbiota and Intraocular Inflammation. Front Immunol [Internet]. 2020 [cited 2022 Nov 23];11. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2020.609765

9. Hilt EE, McKinley K, Pearce MM, Rosenfeld AB, Zilliox MJ, Mueller ER, et al. Urine Is Not Sterile: Use of Enhanced Urine Culture Techniques To Detect Resident Bacterial Flora in the Adult Female Bladder. J Clin Microbiol. 2014 Mar;52(3):871–6

10. Wolfe AJ, Toh E, Shibata N, Rong R, Kenton K, Fitzgerald M, et al. Evidence of uncultivated bacteria in the adult female bladder. J Clin Microbiol. 2012 Apr;50(4):1376–83.

11. Perez-Carrasco V, Soriano-Lerma A, Soriano M, Gutiérrez-Fernández J, Garcia-Salcedo JA. Urinary Microbiome: Yin and Yang of the Urinary Tract. Front Cell Infect Microbiol [Internet].

2021 [cited 2022 Aug 11];11. Available from: https://www.frontiersin.org/articles/10.3389/fcimb.2021.617002

12. Martischang R, François P, Cherkaoui A, Gaïa N, Renzi G, Agostinho A, et al. Epidemiology of ESBL-producing Escherichia coli from repeated prevalence studies over 11 years in a long-term-care facility. Antimicrob Resist Infect Control. 2021 Oct 19;10(1):148.

13. Al Nafeesah A, Al Fakeeh K, Chishti S, Hameed T. E. coli versus Non-E. coli Urinary Tract Infections in Children: A Study from a Large Tertiary Care Center in Saudi Arabia. Int J Pediatr Adolesc Med. 2022 Mar 1;9(1):46–8.

14. Pearce MM, Hilt EE, Rosenfeld AB, Zilliox MJ, Thomas-White K, Fok C, et al. The Female Urinary Microbiome: a Comparison of Women with and without Urgency Urinary Incontinence. mBio. 2014 Jul 8;5(4):e01283-14.

15. Medina M, Castillo-Pino E. An introduction to the epidemiology and burden of urinary tract infections. Ther Adv Urol. 2019 May 2;11:1756287219832172.

16. Bono MJ, Reygaert WC. Urinary Tract Infection. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 [cited 2022 Jun 6]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK470195/

17. Hollenbeak CS, Schilling AL. The attributable cost of catheter-associated urinary tract infections in the United States: A systematic review. Am J Infect Control. 2018 Jul 1;46(7):751–7.

18. Bader MS, Loeb M, Leto D, Brooks AA. Treatment of urinary tract infections in the era of antimicrobial resistance and new antimicrobial agents. Postgrad Med. 2020 Apr 2;132(3):234–50.

19. Paul R. State of the Globe: Rising Antimicrobial Resistance of Pathogens in Urinary Tract Infection. J Glob Infect Dis. 2018;10(3):117–8.

20. Sihra N, Goodman A, Zakri R, Sahai A, Malde S. Nonantibiotic prevention and management of recurrent urinary tract infection. Nat Rev Urol. 2018 Dec;15(12):750–76.

21. Rezatofighi SE, Mirzarazi M, Salehi M. Virulence genes and phylogenetic groups of uropathogenic Escherichia coli isolates from patients with urinary tract infection and uninfected control subjects: a case-control study. BMC Infect Dis. 2021 Apr 17;21(1):361.

22. Terlizzi ME, Gribaudo G, Maffei ME. UroPathogenic Escherichia coli (UPEC) Infections: Virulence Factors, Bladder Responses, Antibiotic, and Non-antibiotic Antimicrobial Strategies. Front Microbiol. 2017 Aug 15;8:1566.

23. Thomas-White K, Forster SC, Kumar N, Van Kuiken M, Putonti C, Stares MD, et al. Culturing of female bladder bacteria reveals an interconnected urogenital microbiota. Nat Commun. 2018 Apr 19;9(1):1557.

24. Zalewska-Piątek B, Piątek R. Phage Therapy as a Novel Strategy in the Treatment of Urinary Tract Infections Caused by E. Coli. Antibiotics. 2020 Jun 5;9(6):304.

25. Chapelle C, Gaborit B, Dumont R, Dinh A, Vallée M. Treatment of UTIs Due to Klebsiella pneumoniae Carbapenemase-Producers: How to Use New Antibiotic Drugs? A Narrative Review. Antibiotics. 2021 Nov 1;10(11):1332.

26. Nordmann P, Naas T, Poirel L. Global Spread of Carbapenemase-producing Enterobacteriaceae. Emerg Infect Dis. 2011 Oct;17(10):1791–8.

27. Adu-Oppong B, Thänert R, Wallace MA, Burnham CAD, Dantas G. Substantial overlap between symptomatic and asymptomatic genitourinary microbiota states. Microbiome. 2022 Jan 17;10(1):6.

28. Johnson JA, Delaney LF, Ojha V, Rudraraju M, Hintze KR, Siddiqui NY, et al. Commensal Urinary Lactobacilli Inhibit Major Uropathogens In Vitro With Heterogeneity at Species and Strain Level. Front Cell Infect Microbiol. 2022;12:870603.

29. Woese CR, Fox GE, Zablen L, Uchida T, Bonen L, Pechman K, et al. Conservation of primary structure in 16S ribosomal RNA. Nature. 1975 Mar;254(5495):83–6.

30. Hoffman C, Siddiqui NY, Fields I, Gregory WT, Simon HM, Mooney MA, et al. Species-Level Resolution of Female Bladder Microbiota from 16S rRNA Amplicon Sequencing. mSystems. 6(5):e00518-21.

31. Katiraei S, Anvar Y, Hoving L, Berbée JFP, van Harmelen V, Willems van Dijk K. Evaluation of Full-Length Versus V4-Region 16S rRNA Sequencing for Phylogenetic Analysis of Mouse Intestinal Microbiota After a Dietary Intervention. Curr Microbiol. 2022 Jul 30;79(9):276.

32. Patel S, Ingram C, Scovell J, Link R, Mayer W. The Microbiome and Urolithiasis: Current Advancements and Future Challenges. Curr Urol Rep. 2022 Mar 1;23:1–10.

33. Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun. 2019 Nov 6;10(1):5029.

34. Nguyen NP, Warnow T, Pop M, White B. A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. Npj Biofilms Microbiomes. 2016 Apr 20;2(1):1–8.

35. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 2017 Dec;11(12):2639–43.

36. Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, et al. Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. Methods Ecol Evol. 2013;4(12):1111–9.

37. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 2016 Jul;13(7):581–3.

38. Hon T, Mars K, Young G, Tsai YC, Karalius JW, Landolin JM, et al. Highly accurate long-read HiFi sequencing data for five complex genomes. Sci Data. 2020 Nov 17;7(1):399.

39. Giani AM, Gallo GR, Gianfranceschi L, Formenti G. Long walk to genomics: History and current approaches to genome sequencing and assembly. Comput Struct Biotechnol J. 2019 Nov 17;18:9–19.

40. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 2015 Oct 1;13(5):278–89.

41. Udaondo Z, Sittikankaew K, Uengwetwanit T, Wongsurawat T, Sonthirod C, Jenjaroenpun P, et al. Comparative Analysis of PacBio and Oxford Nanopore Sequencing Technologies for Transcriptomic Landscape Identification of Penaeus monodon. Life. 2021 Aug 23;11(8):862.

42. Balvočiūtė M, Huson DH. SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? BMC Genomics. 2017 Mar 14;18(2):114.

43. Almeida A, Mitchell AL, Tarkowska A, Finn RD. Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. GigaScience. 2018 May 1;7(5):giy054.

44. Dixit K, Davray D, Chaudhari D, Kadam P, Kshirsagar R, Shouche Y, et al. Benchmarking of 16S rRNA gene databases using known strain sequences. Bioinformation. 2021 Mar 31;17(3):377–91.

45. RDP Release 11 -- Sequence Analysis Tools [Internet]. [cited 2022 Nov 11]. Available from: http://rdp.cme.msu.edu/index.jsp

46. Documentation [Internet]. [cited 2022 Nov 11]. Available from: https://www.arb-silva.de/documentation/

47. Somerville V, Lutz S, Schmid M, Frei D, Moser A, Irmler S, et al. Long-read based de novo assembly of low-complexity metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system. BMC Microbiol. 2019 Jun 25;19(1):1–18.

48. Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem Biophys Res Commun. 2016 Jan 22;469(4):967–77.

49. Sharpton TJ. An introduction to the analysis of shotgun metagenomic data. Front Plant Sci. 2014 Jun 16;5:209.

50. Staden R. A strategy of DNA sequencing employing computer programs. Nucleic Acids Res. 1979 Jun 11;6(7):2601–10.

51. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014 Mar 3;15(3):R46.

52. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019 Nov 28;20(1):257.

53. Lu J, Salzberg SL. Ultrafast and accurate 16S rRNA microbial community analysis using Kraken 2. Microbiome. 2020 Aug 28;8(1):124.

54. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res. 2007 Mar;17(3):377.

55. Dilthey AT, Jain C, Koren S, Phillippy AM. Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps. Nat Commun. 2019 Jul 11;10(1):3066.

56. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, et al. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell. 2019 Jan 24;176(3):649-662.e20.

57. Anyansi C, Straub TJ, Manson AL, Earl AM, Abeel T. Computational Methods for Strain-Level Microbial Detection in Colony and Metagenome Sequencing Data. Front Microbiol [Internet]. 2020 [cited 2022 Aug 2];11. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2020.01925

58. Quince C, Nurk S, Raguideau S, James R, Soyer OS, Summers JK, et al. STRONG: metagenomics strain resolution on assembly graphs. Genome Biol. 2021 Jul 26;22(1):214.

59. Bajic P, Van Kuiken ME, Burge BK, Kirshenbaum EJ, Joyce CJ, Wolfe AJ, et al. Male Bladder Microbiome Relates to Lower Urinary Tract Symptoms. Eur Urol Focus. 2020 Mar 15;6(2):376–82.

60. Price TK, Lin H, Gao X, Thomas-White KJ, Hilt EE, Mueller ER, et al. Bladder bacterial diversity differs in continent and incontinent women: a cross-sectional study. Am J Obstet Gynecol. 2020 Nov;223(5):729.e1-729.e10.

61. Al Mohajer M, Darouiche RO. Staphylococcus aureus Bacteriuria: Source, Clinical Relevance, and Management. Curr Infect Dis Rep. 2012 Dec;14(6):601–6.

62. Selim S, Faried OA, Almuhayawi MS, Saleh FM, Sharaf M, El Nahhas N, et al. Incidence of Vancomycin-Resistant Staphylococcus aureus Strains among Patients with Urinary Tract Infections. Antibiot Basel Switz. 2022 Mar 18;11(3):408.

63. Walker JN, Flores-Mireles AL, Pinkner CL, Schreiber HL, Joens MS, Park AM, et al. Catheterization alters bladder ecology to potentiate Staphylococcus aureus infection of the urinary tract. Proc Natl Acad Sci U S A. 2017 Oct 10;114(41):E8721–30.

64. Atassi F, Pho Viet Ahn DL, Lievin-Le Moal V. Diverse Expression of Antimicrobial Activities Against Bacterial Vaginosis and Urinary Tract Infection Pathogens by Cervicovaginal Microbiota Strains of Lactobacillus gasseri and Lactobacillus crispatus. Front Microbiol. 2019;10:2900.

65. Atassi F, Servin AL. Individual and co-operative roles of lactic acid and hydrogen peroxide in the killing activity of enteric strain Lactobacillus johnsonii NCC933 and vaginal strain Lactobacillus gasseri KS120.1 against enteric, uropathogenic and vaginosis-associated pathogens. FEMS Microbiol Lett. 2010 Mar;304(1):29–38.

66. Kalyoussef S, Nieves E, Dinerman E, Carpenter C, Shankar V, Oh J, et al. Lactobacillus proteins are associated with the bactericidal activity against E. coli of female genital tract secretions. PloS One. 2012;7(11):e49506.

67. Adewale BA. Will long-read sequencing technologies replace short-read sequencing technologies in the next 10 years? Afr J Lab Med. 2020 Nov 26;9(1):1340.

68. Ugarcina Perovic S, Ksiezarek M, Rocha J, Cappelli EA, Sousa M, Ribeiro TG, et al. Urinary Microbiome of Reproductive-Age Asymptomatic European Women. Microbiol Spectr. 2022 Nov 16;0(0):e01308-22.

69. Callahan BJ, Wong J, Heiner C, Oh S, Theriot CM, Gulati AS, et al. High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution [Internet]. Bioinformatics; 2018 Aug [cited 2022 Aug 2]. Available from: http://biorxiv.org/lookup/doi/10.1101/392332

70. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013 Jan;41(Database issue):D590–6.

71. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome. 2018 Dec 17;6(1):226.

72. Grenié M, Denelle P, Tucker CM, Munoz F, Violle C. funrar: An R package to characterize functional rarity. Divers Distrib. 2017;23(12):1365–71.

73. Wickham H. Reshaping Data with the reshape Package. J Stat Softw. 2007 Nov 13;21:1–20.

74. ggpubr: Publication Ready Plots - Articles - STHDA [Internet]. [cited 2022 Aug 29]. Available from: http://www.sthda.com/english/articles/24-ggpubr-publication-ready-plots/

75. Valero-Mora PM. ggplot2: Elegant Graphics for Data Analysis. J Stat Softw. 2010 Jul 30;35:1–3.

76. Callahan BJ, Wong J, Heiner C, Oh S, Theriot CM, Gulati AS, et al. High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. Nucleic Acids Res. 2019 Oct 10;47(18):e103.

77. Wolfe AJ, Brubaker L. Urobiome Updates: Advances in Urinary Microbiome Research. Nat Rev Urol. 2019 Feb;16(2):73–4.

78. Hu D, Fuller NR, Caterson ID, Holmes AJ, Reeves PR. Single-gene long-read sequencing illuminates Escherichia coli strain dynamics in the human intestinal microbiome. Cell Rep. 2022 Jan 11;38(2):110239.

79. Jiménez-Guerra G, Lara-Oya A, Martínez-Egea I, Navarro-Marí JM, Gutiérrez-Fernández J. Urinary tract infection by Aerococcus sanguinicola. An emerging opportunistic pathogen. Rev Clínica Esp Engl Ed. 2018 Oct 1;218(7):351–5.

80. Narten M, Rosin N, Schobert M, Tielen P. Susceptibility of Pseudomonas aeruginosa urinary tract isolates and influence of urinary tract conditions on antibiotic tolerance. Curr Microbiol. 2012 Jan;64(1):7–16.

81. Senneby E, Eriksson B, Fagerholm E, Rasmussen M. Bacteremia with Aerococcus sanguinicola: Case Series with Characterization of Virulence Properties. Open Forum Infect Dis. 2014 May 23;1(1):ofu025.

82. Carlstein C, Marie Søes L, Jørgen Christensen J. Aerococcus christensenii as Part of Severe Polymicrobial Chorioamnionitis in a Pregnant Woman. Open Microbiol J. 2016 Mar 10;10:27–31.

83. Tindall BJ, Sutton G, Garrity GM. Enterobacter aerogenes Hormaeche and Edwards 1960 (Approved Lists 1980) and Klebsiella mobilis Bascomb et al. 1971 (Approved Lists 1980) share the same nomenclatural type (ATCC 13048) on the Approved Lists and are homotypic synonyms, with consequences for the name Klebsiella mobilis Bascomb et al. 1971 (Approved Lists 1980). Int J Syst Evol Microbiol. 2017 Feb;67(2):502–4.

84. Wesevich A, Sutton G, Ruffin F, Park LP, Fouts DE, Fowler VG, et al. Newly Named Klebsiella aerogenes (formerly Enterobacter aerogenes) Is Associated with Poor Clinical Outcomes Relative to Other Enterobacter Species in Patients with Bloodstream Infection. J Clin Microbiol. 2020 Aug 24;58(9):e00582-20.

85. Álvarez-Artero E, Campo-Nuñez A, García-García I, García-Bravo M, Cores-Calvo O, Galindo-Pérez I, et al. Urinary tract infection caused by Enterococcus spp.: Risk factors and mortality. An observational study. Rev Clin Esp. 2021;221(7):375–83.

86. Rampersaud R, Planet PJ, Randis TM, Kulkarni R, Aguilar JL, Lehrer RI, et al. Inerolysin, a cholesterol-dependent cytolysin produced by Lactobacillus iners. J Bacteriol. 2011 Mar;193(5):1034–41.

87. Srinivasan S, Hoffman NG, Morgan MT, Matsen FA, Fiedler TL, Hall RW, et al. Bacterial communities in women with bacterial vaginosis: high resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. PloS One. 2012;7(6):e37818.

88. Yoo JJ, Song JS, Kim WB, Yun J, Shin HB, Jang MA, et al. Gardnerella vaginalis in Recurrent Urinary Tract Infection Is Associated with Dysbiosis of the Bladder Microbiome. J Clin Med. 2022 Apr 20;11(9):2295.

89. Garretto A, Miller-Ensminger T, Ene A, Merchant Z, Shah A, Gerodias A, et al. Genomic Survey of E. coli From the Bladders of Women With and Without Lower Urinary Tract Symptoms. Front Microbiol [Internet]. 2020 [cited 2022 Apr 12];11. Available from: https://www.frontiersin.org/article/10.3389/fmicb.2020.02094

90. Ryan MP, Pembroke JT, Adley CC. Ralstonia pickettii: a persistent Gram-negative nosocomial infectious organism. J Hosp Infect. 2006 Mar 1;62(3):278–84.

91. Nurjadi D, Boutin S, Schmidt K, Ahmels M, Hasche D. Identification and Elimination of the Clinically Relevant Multi-Resistant Environmental Bacteria Ralstonia insidiosa in Primary Cell Culture. Microorganisms. 2020 Oct 17;8(10):1599.

92. Moustafa A, Li W, Singh H, Moncera KJ, Torralba MG, Yu Y, et al. Microbial metagenome of urinary tract infection. Sci Rep. 2018 Mar 12;8(1):4333.

93. Khan S, Sistla S, Dhodapkar R, Parija SC. Fatal Delftia acidovorans infection in an immunocompetent patient with empyema. Asian Pac J Trop Biomed. 2012 Nov;2(11):923–4.

94. Preiswerk B, Ullrich S, Speich R, Bloemberg GV, Hombach M 2011. Human infection with Delftia tsuruhatensis isolated from a central venous catheter. J Med Microbiol. 60(2):246–8.

95. Bilgin H, Sarmis A, Tigen E, Soyletir G, Mulazimoglu L. Delftia acidovorans: A rare pathogen in immunocompetent and immunocompromised patients. Can J Infect Dis Med Microbiol. 2015;26(5):277–9.

96. Ranc A, Dubourg G, Fournier PE, Raoult D, Fenollar F. Delftia tsuruhatensis, an Emergent Opportunistic Healthcare-Associated Pathogen. Emerg Infect Dis. 2018 Mar;24(3):594–6.

97. Kam SK, Lee WS, Ou TY, Teng SO, Chen FL. Delftia acidovorans Bacteremia Associated with Ascending Urinary Tract Infections Proved by Molecular Method. J Exp Clin Med. 2012 Jun 1;4(3):180–2.

98. Jumpstart Consortium Human Microbiome Project Data Generation Working Group. Evaluation of 16S rDNA-based community profiling for human microbiome research. PloS One. 2012;7(6):e39315.

99. Siddiqui NY, Ma L, Brubaker L, Mao J, Hoffman C, Dahl EM, et al. Updating Urinary Microbiome Analyses to Enhance Biologic Interpretation. Front Cell Infect Microbiol [Internet]. 2022 [cited 2022 Nov 9];12. Available from: https://www.frontiersin.org/articles/10.3389/fcimb.2022.789439

100. Brubaker L, Gourdine JPF, Siddiqui NY, Holland A, Halverson T, Limeria R, et al. Forming Consensus To Advance Urobiome Research. mSystems. 2021 Aug 31;6(4):e0137120.

101. Thomas-White KJ, Hilt EE, Fok C, Pearce MM, Mueller ER, Kliethermes S, et al. Incontinence medication response relates to the female urinary microbiota. Int Urogynecology J. 2016 May;27(5):723–33.

102. Edgar RC. High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny [Internet]. bioRxiv; 2022 [cited 2022 Nov 8]. p. 2021.06.20.449169. Available from: https://www.biorxiv.org/content/10.1101/2021.06.20.449169v2

103. Durazzi F, Sala C, Castellani G, Manfreda G, Remondini D, De Cesare A. Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota. Sci Rep. 2021 Feb 4;11(1):3030.

104. Sanabria AM, Janice J, Hjerde E, Simonsen GS, Hanssen AM. Shotgun-metagenomics based prediction of antibiotic resistance and virulence determinants in Staphylococcus aureus from periprosthetic tissue on blood culture bottles. Sci Rep. 2021 Oct 21;11(1):20848.

105. Janes VA, Matamoros S, Munk P, Clausen PTLC, Koekkoek SM, Koster LAM, et al. Metagenomic DNA sequencing for semi-quantitative pathogen detection from urine: a prospective, laboratory-based, proof-of-concept study. Lancet Microbe. 2022 Aug 1;3(8):e588–97.

106. Keegan KP, Glass EM, Meyer F. MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. Methods Mol Biol Clifton NJ.

2016;1399:207–33.

107. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Mar 4;9(4):357–9.

108. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. CONCOCT: Clustering cONtigs on COverage and ComposiTion [Internet]. arXiv; 2013 [cited 2022 Nov 14]. Available from: http://arxiv.org/abs/1312.4038

109. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403–10.

110. Mueller ER, Wolfe AJ, Brubaker L. Female urinary microbiota. Curr Opin Urol. 2017 May;27(3):282–6.

111. Cameron D, Jiang JH, Hassan K, Elbourne L, Tuck K, Paulsen I, et al. Insights on virulence from the complete genome of Staphylococcus capitis. Front Microbiol [Internet]. 2015 [cited 2022 Dec 2];6. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2015.00980

112. Schaffer JN, Pearson MM. Proteus mirabilis and Urinary Tract Infections. Microbiol Spectr. 2015 Oct;3(5):10.1128/microbiolspec.UTI-0017–2013.

113. Hochstedler BR, Burnett L, Price TK, Jung C, Wolfe AJ, Brubaker L. Urinary microbiota of women with recurrent urinary tract infection: collection and culture methods. Int Urogynecology J. 2022 Mar;33(3):563–70.

VITA

Delaney Sauer is from Warrensburg, Missouri. She attended Loyola University Chicago for her bachelor's degree in Bioinformatics. While at Loyola, she participated in the Interdisciplinary Honors program and the First Year Research Experience (FYRE). In her junior year, she was accepted into the first cohort of students in the Loyola Adventures in Urobiome Data (LAUD) program. There, she was introduced to the urinary microbiome niche of microbial genetics and her thesis advisor, Dr. Catherine Putonti. After presenting at the NIDDK (National Institute of Diabetes and Digestive and Kidney Diseases) conference at the conclusion of LAUD, she joined the Putonti Lab at Loyola University Chicago and began work on her thesis. After a successful thesis defense, Sauer accepted a job with the Translational Data Services team at Precision for Medicine, a company aiding in pharmaceutical clinical trials.