SPECIES-LEVEL IDENTIFICATION OF THE BACTERIA IN THE HUMAN BLADDER MICROBIOME

By

Carter Hoffman

A THESIS

Presented to the Department of Medical Informatics and Clinical Epidemiology
and the Oregon Health & Science University School of Medicine
in partial fulfillment of
the requirements for the degree of

Master of Science

May 2020

CERTIFICATE OF APPROVAL

This is to certify that the Master's thesis of

Carter Hoffman

"Species-level identification of the bacteria in the human bladder microbiome"

has been approved

Thesis Advisor - Lisa Karstens, PhD

Committee Member - Michael Mooney, PhD

Committee Member - Holly Simon, PhD

Contents

Table of figures

Acknowledgements

I thank Dr. Alan Wolfe and Dr. Krystal Thomas-White for generously sharing their hard work and providing the sequencing results and data which I used in this study. I thank Dr. Christina Zheng for insight into multisequence alignments and assistance with genomic formats. I thank Sean Davin for training me in the benchwork techniques needed to do PCR with the microbiome, and supervising me when he had other things to do. I am especially grateful for the help and patient guidance of my thesis advisory committee, Dr. Michael Moony and Dr. Holly Simon. I have a moment of silence for Dr. Mark Asquith, who passed away in 2019. I also thank my family, who were a never-ending source of encouragement for me to finish this degree. Finally, I thank my mentor Dr. Lisa Karstens, of whom it is absolutely true when I say I could not have done this without her guidance and teaching.

Abstract

The human bladder was long believed to be a sterile environment, except for acute infections. However, evidence from sequenced-based and enhanced culturing techniques have revealed the bladder supports a population of bacteria even in the absence of infection. The discovery of a bladder microbiota naturally leads to the question of how it influences the health of the host, and recent clinical studies collectively provide evidence that understanding changes in the bacterial diversity and abundance of the bladder microbiota is relevant and warrants further investigation.

Bacterial identification depends on comparing the DNA sequence obtained from the collected bacteria with information held in a phylogenetic database, using an algorithm to perform the comparison. Together, these components are called a classification scheme. A common classification scheme is composed of the V4 region of the 16S rRNA gene, the Silva database, and Naive Bayes classifier. Currently, the phylogenetic resolution achieved with this classification scheme is limited to the Genus level, and obscures the true nature of the relation between bacterial species found in the bladder and the host.

The 16S rRNA gene is not the only gene suitable for identifying bacteria in a collected sample. Additional genes that have experienced different selective pressure and have accumulated a different diversity of DNA mutations may provide better taxonomic resolution for the bacteria found in the human bladder. Likewise, there are more classifiers than just Naive Bayes, and more databases than just Silva. Combining these additional resources into new classification schemes can increase the phylogenetic resolution.

This study was done in two parts. The first part computationally compares the phylogenetic resolution that is achieved by combining 4 currently available databases, 2 taxonomic classifiers, and subsequences from 3 genes into 58 classification schemes. The results show that the best overall classification scheme is composed of a custom-built prokaryotic genomic database, a gene subsequence from the RNA polymerase subunit B gene (RpoB), and the Bayesian

Lowest Common Ancestor (BLCA) classifier. The results also show that the V2-V3 region of the 16S rRNA gene had the best performance of any classification scheme that relies on the 16S rRNA gene sequence. The second part of the study was to validate one of the classification schemes by comparing the computational outcome with data generated from targeted amplicon sequencing of bacterial DNA obtained from urine samples. The results from this part of the study show that the number of bacterial species correctly identified *in silico* and the number of bacterial species correctly identified *in vitro* are in good agreement. In summary, these results show that species level classification is possible for the microbiota found in the human bladder with resources that are currently available.

1. Introduction

The human body provides a wide range of habitats, all of which support a variety of bacteria, archaea, viruses and fungi collectively known as the human microbiome(1). Beginning in 2008, The Human Microbiome Project performed a general survey of five body habitats in order to begin to characterize and compare the diversity of the bacteria found in these habitats, both between body sites on individuals and across populations of humans(2). One of the body environments that was left out of the HMP was the human bladder, which has historically been viewed as sterile(3) except for acute infections. However, recent evidence from sequenced-based methods and enhanced culturing techniques have revealed a population of bacteria that exists in the bladder even in the absence of infection(4,5).

The discovery of a bladder microbiota naturally leads to the question of how it influences the health of the host. Recent results have begun to provide evidence that changes in the population diversity of the bladder microbiota are associated with changes in health. Some examples are studies that have identified characteristics in the urinary microbiota that are associated with symptom severity of Urgency Urinary Incontinence (UUI)(6), increased risk of urinary tract infection(7), and response to the common UUI drug treatment solifenacin(8). These results collectively provide evidence that the urinary microbiota is clinically relevant, and warrants further investigation.

In order to correlate the diversity of bacteria found in the human bladder to changes in the health of host, it is necessary to be able to accurately identify bacteria in a rapid and large-scale manner. Currently, the practical method of identifying bacteria found in the bladder is to extract the bacterial DNA obtained from a urine sample and sequence a small portion of a marker gene. An algorithm is used to compare short DNA sequences to sequences held in a reference database until the closest match is found. Then, the taxonomic information from the reference sample with the highest sequence similarity is assigned to the unknown bacterial DNA sample. This method depends on three components: 1) a

gene sequence, 2) an algorithm for comparing sequences, and 3) a taxonomic database. Common choices for these components are: the small subunit of the bacterial ribosome (16S rRNA gene sequence)(9), the Naïve Bayes classifier(10), and the Silva database(11), respectively, but these are not the only choices available.

Taken together, the classification algorithm, the database of known bacterial sequences, and some portion of the genetic sequence obtained from the population of formerly unknown/unidentified bacteria comprise a classification scheme. Each component has nontrivial limitations that have required clever and significant effort to circumvent. As examples, the algorithms that perform the classification have been constrained by the available hardware of computers(10,12); the databases that contain records of agreed upon bacterial taxa have been constrained by the state of bacterial systematics(13); and the length of genetic sequence used to compare sampled DNA to the database records has been constrained by the state of sequencing technology(14).

This project is an attempt to determine if species level identification of the bacterial microbiota found in the human bladder is possible with currently available resources, and if so, find an optimal classification scheme. While the constraints outlined above seem daunting when taken together, I show it is possible to achieve species level identification of bladder bacteria, when the proper consideration of the constraints are is taken.

Section 2 introduces the concepts and tools that are used in identifying bacteria found in the human microbiome and used in this study. Section 2.2 covers how bacteria are placed on the Tree of Life in relation to other organisms and to each other. Section 2.3 covers the molecular biology behind the use of gene sequences as a means to identify bacteria. Section 2.4 and 2.5 covers the computational methods of used for combining databases, classification algorithms, and gene sequences to determine the identity of a bacteria, and how to assess if one computational method is better than another.

This study relies on gene sequence data generated by Thomas-White et. al.(5). In their publication, the authors describe how they successfully cultured, isolated and performed whole genome sequencing of bacteria obtained from urine samples. Section 2.6 introduces this dataset.

The prediction of how well different combinations of databases, classification algorithms, and gene sequences perform to determine the identity of an unknown bacterial sequence is presented in sections 4.1 to 4.6, and an *in vitro* validation of the *in silico* predictions are presented in section 4.7. These results are discussed in section 5.

2. Background

Recognizing the large amount of diversity that exists in the animal and plant kingdoms is not hard, largely due to the fact that many forms of life are visible to the unaided eye. When organizing these life forms into groups that reflect the phylogenetic relationships between them, the visible morphology of the plant or animal resolves a great deal of the work. Additionally, the genotype of the plant or animal is the result of mutation and natural selection of vertically inherited traits from the parent to the offspring. Bacterial identification and classification presents two major challenges in that bacteria are visibly homogenous, and possess the ability to transfer genetic material between among themselves through lateral gene transfer (LGT).

Broadly, there are two concepts that this thesis will deal with. The first is describing the challenges of applying the definition of species to bacteria. While the science of organizing life forms into a hierarchical structure has a long history, bacteria are difficult to classify. The second concept is the method of identifying an unknown bacterium based on a genetic sequence, a method that falls under the field of Information Retrieval. The goal is to take a small amount of descriptive data from an unknown bacterium, such as a short DNA sequence, and use it to find a document that contains more information about the organism, such as where the bacterium fits on the Tree of Life. In general, the

hope is that the short DNA sequence is enough to identify the species of the unknown bacterium.

2.1 The two concepts of this study

2.1.1 Identification and classification

The study of the methods by which a species is recognized and delimited, and the methods by which species are arranged in the form of classification falls under the field of Systematics. In general terms, Systematics is the scientific study of the kinds of diversity of organisms, and all relations among them.

The method of placing an individual in an existing classification paradigm is identification. The classification paradigm contains groups of organisms that are distinct enough to be assigned to a single category, called a taxon. Each category describes the rank or level in a hierarchical classification. For example, organisms that have a backbone, hair, three middle ear bones and produce milk to nurse their young comprise one taxon. The organisms of that taxon can be placed in the category of Mammalia, at the hierarchical rank of Class. The classification schemes in this study attempt to match unknown bacterial DNA sequences to sequences of a known classification in a reference database, and therefore are methods of identification.

2.1.2 Record linkage

The information collected during bacterial classification can be organized, stored and later consulted when presented with a group of characteristics belonging to bacteria of an unknown identity. This aspect of this study falls under Information Retrieval, and more specifically Record Linkage. In short, an identifying characteristic of an unknown bacterium is compared to the stored information held in a reference database. If the information held in a record of the reference database meets pre-determined criteria (e.g., a specific % similarity to the identifying characteristic of the unknown bacteria), then the information of in that reference record is attributed to the unknown bacteria. While this method is

simple in cases where the reference material is small, or when the identifying characteristic is highly distinctive, it quickly becomes a very difficult problem with even moderately sized reference databases. The difficulty is compounded when the identifying characteristic is found to be contained in several records in the reference database. Practically speaking, all record linkage is performed through the use of a computer algorithm specifically designed to do this sort of work.

2.1.3 General terms and glossary

This study uses a wide range of terms, but some of the most commonly used terms describe the parts of record linkage, which will be covered in the next paragraph. A glossary of frequently used terms follows.

The collection of information that pertains to one subject (e.g. a bacterial species) is a *record*. In this study, a record is typically the information held in one FASTA file, and is composed of a unique alphanumeric character string, a taxonomic lineage, and the DNA sequence. A *database* is a collection of records. The *identifier* is information taken from the unknown bacterium that is used to search the database for matching records. In this paper, the identifier is a bacterial DNA sequence obtained through PCR amplification, and subsequently sequenced. During the process of sequencing, each DNA sequence is assigned a unique alphanumeric label. This unique label and the DNA sequence is formatted into a FASTA file and becomes the *query* record. The query record is the compared against the records contained in a reference database. The *classifier* is the algorithm used to do the actual work of comparing query records with reference records. Examples are the Naïve Bayes classifier and the Bayesian Lowest Common Ancestor (BLCA) classifier. The *classification scheme* refers to the specific combination of reference database, DNA sequence used as an identifier, and classification algorithm employed.

Glossary

Broad spectrum primers - degenerate primers that are designed to amplify orthologous DNA from all the species of interest.

Classifier - the algorithm used to do the actual work of comparing query records with reference records. Examples are the Naïve Bayes classifier and the Bayesian Lowest Common Ancestor (BLCA) classifier.

Classification scheme - the specific combination of reference database, DNA sequence used as an identifier, and classification algorithm employed.

Database - a collection of records.

Deterministic record linkage - record linkage that produces two mutually exclusive categories: either the record pair is a match, or the record pair is not a match

Degenerate primers - A primer is called degenerate if any of the positions of the primer can have more than one possible base

Evaluation - Quantifying how much better one classification scheme is compared to another classification scheme.

Expanded quantitative urine culture (EQUC) - Modified urine culture conditions to include the plating of a greater volume of urine, incubation in varied atmospheric conditions, and the use of extended incubation times.

Identifier - the bacterial DNA sequence that is used by a classifier to search a reference database for a match. Examples are the V4 region of the 16S rRNA gene sequence, and the full 16S rRNA gene sequence.

Marker gene – a gene that is used to identify a species of bacteria. The complete gene sequence can be used, or only a shorter subsequence. Examples are the different variable regions of the 16S rRNA gene sequence, most of which have been used to identify bacteria.

Multisequence alignment (MSA) - the algorithmic assembly of three or more sequences such that each residue is matched with its counterpart in every other sequence

Probabilistic record linkage - record linkage that attaches a probability to the comparison as a means to define when to designate a record pair as a match.

Predicted species set - The set of species predicted to be correctly classified by any one of the classification schemes tested *in silico*, and that are also present in the V4 subset.

Predicted paired set - the species/sample pair combination in the V4 subset that exist in the predicted species set of a classification scheme.

Query - the record from the an unknown bacterium that is compared against the records contained in a reference database. In this study, only the unique label and gene sequence used as the identifier is held in this record.

Record - The collection of information that pertains to one subject. In this study, the information is held in one FASTA file, and at a minimum is composed of a unique alphanumeric identifier and DNA sequence.

Record linkage - As applied in this study, Record Linkage is the process by which the genetic information held in two separate records is compared, and a determination of whether or not they are referring to the same species of bacteria is made.

Reference - the record to which a query record is compared against. In this study, the reference records contain the taxonomic lineage in addition to the ID and a more complete sequence of the gene (or genome).

Sliding window analysis - the method by which a list of subsequences are generated by taking successive groups of equal size, in the manner of a window of fixed length sliding across the full sequence.

Targeted amplicon - the span of DNA that is amplified during PCR when using primers that have been designed to uniquely anneal to sites that flank the amplicon.

The Thomas-White dataset - The set of bacterial species cultured from the urine samples, and identified by whole genome sequencing.

V4 subset - The smaller set of species/sample pairs that were subjected to targeted amplicon sequencing within the Thomas-White dataset.

Variable region - regions of a gene sequence that show heterogeneity when compared to analogous regions of the gene across different species.

2.2 Bacterial systematics

In this study, the classification of the unknown bacteria found in the bladder is achieved by correctly identifying the bacterial DNA sequences obtained in that sample. Each element of a classification scheme embodies a surprising amount of conceptual development covering 300 years of natural history, from the debate over how to define a species to the endured compromise of next generation sequencing's low cost, high volume, but short read lengths. While there is a strong desire to just use each element of a classification scheme as found, out of the box, some amount of work is always required before classification results can be interpreted.

Naturally, there are few examples where classification and identification achieve perfectly unambiguous results. Bacteria are just one example. The modern definition of a species is a population of reproductively isolated organisms, defined by their relation to coexisting species who are not interbreeding and not competing for the same resources(15). However, the concept of species has always been a compromise, and even Darwin gave up on a rigorous definition(15). Botanists have provided many examples that have defied any attempt at a comprehensive species definition since the 1700s. While the species definition holds up well for sexually reproducing populations such as bears and pelicans, it certainly falls apart for bacteria.

2.2.1 Polyphasic taxonomy

Bacteria were first described in 1683 by Antony van Leeuwenhoek in a letter to a friend, and while describing the physical characteristics of bacteria remained a method of classification(16), the difficulties of depending on the physical characteristics to classify bacteria were fully realized as early as 1866. Because microbes offered so little in the way of distinguishing characteristics, Ernst Haeckel defined the Protista kingdom by lumping together all the remaining organisms that could not be assigned to any other kingdom(17).

However, bacterial taxonomy was not restricted to just the physical description (morphology), and many other distinguishing characteristics are used to define bacterial taxa. Examples of these characteristics are growth at different temperatures, pH values, salt concentrations and atmospheric conditions(18). Others are the composition of the cell wall, the fatty acids found in the cells, the whole-cell protein analysis by gel electrophoresis, antibiotic resistance and molecules that could be targeted with antibodies (serotyping)(18).

Advancements in Bacterial taxonomy remained slow and arduous until the development of polyphasic taxonomy. With the use of genomic sequences starting in 1987(19), and the combining of genotypic, phenotypic, and phylogenetic qualities in the years later(18), bacterial taxonomy has become a consensus method of classification that has moved beyond the dependence of visual characteristics.

2.2.2 Formal species proposal

The formal taxonomic proposal of a new species is a rigorous and time consuming procedure(20), and if successfully accomplished results in a publication in the *International Journal of Systematic and Evolutionary Microbiology*, the journal of record for novel microbial taxa[www.microbiologyresearch.org]. A protocol for defining a new species is outlined in Rainey et al. (21). The first item, culturing an isolate, is undoubtedly the most difficult. The remainder of the checklist is

- sequence the 16S gene, if not the entire genome
- select a set of known isolates that are similar to the unknown isolate for comparison
- perform a standardized array of phenotypic, chemotaxonomic and genomic tests
- select appropriate names that comply with the *International Code of Nomenclature of Prokaryotes* to describe the new species
- write a proposal justifying the designation of a new species and submit to the IJSEM
- deposit two cultures to two separate international cell banks such as ATCC

Because of the effort involved, the number of published prokaryotic species is extremely small compared to the number of novel bacteria classified through the algorithmic evaluation of gene sequences.

2.2.3 Identification using DNA-DNA hybridization

The gold standard of bacterial identification is the DNA-DNA hybridization method(22). Genomic DNA is sheared into small oligonucleotides and hybridized to a similarly constructed library from a reference bacterium. A high proportion of hybridized DNA indicates the two bacteria are of the same species. While accurate, the process is slow and labor intensive, and it is recognized that additional identification methods which have a high concordance with DNA-DNA hybridization, yet are easier to perform, are needed(23). One of the additional methods that has remained predominant is to use the gene sequence of the fourth variable region (V4) of the small subunit of the bacterial ribosome (16S rRNA gene), because it is estimated to have a high phylogenetic resolution(14). Frequently, this is the only classification scheme implemented.

2.2.4 The ribosome as identifier

Carl Woese and George Fox made a significant advancement in bacterial classification by using the ribosomal RNA (rRNA) gene to organize all known life into three domains. In their 1977 publication, Woese and Fox stated that

determining evolutionary relationships between all living organisms requires measuring the degree of difference between comparable structures. Woese and Fox chose the ribosome, due to the universal presence in all self-replicating systems and ease of isolation. While they initially started their work by comparing enzymatic digestions of ribosomal RNA(24), they moved to comparing rRNA sequences, and later the 16S rRNA gene sequence itself(19).

Most importantly, the gene has alternating conserved and variable regions along the length of its sequence(24). This quality is important when attempting to design one set of primers that will anneal to many different species of bacteria. When designing primers for this thesis, the highly conserved regions in the 16S rRNA gene sequence allowed flexibility in selecting a primer set that would both span a variable region and anneal to many species in the data set. Examples of variable regions in the 16S rRNA gene is shown in Figure 1.

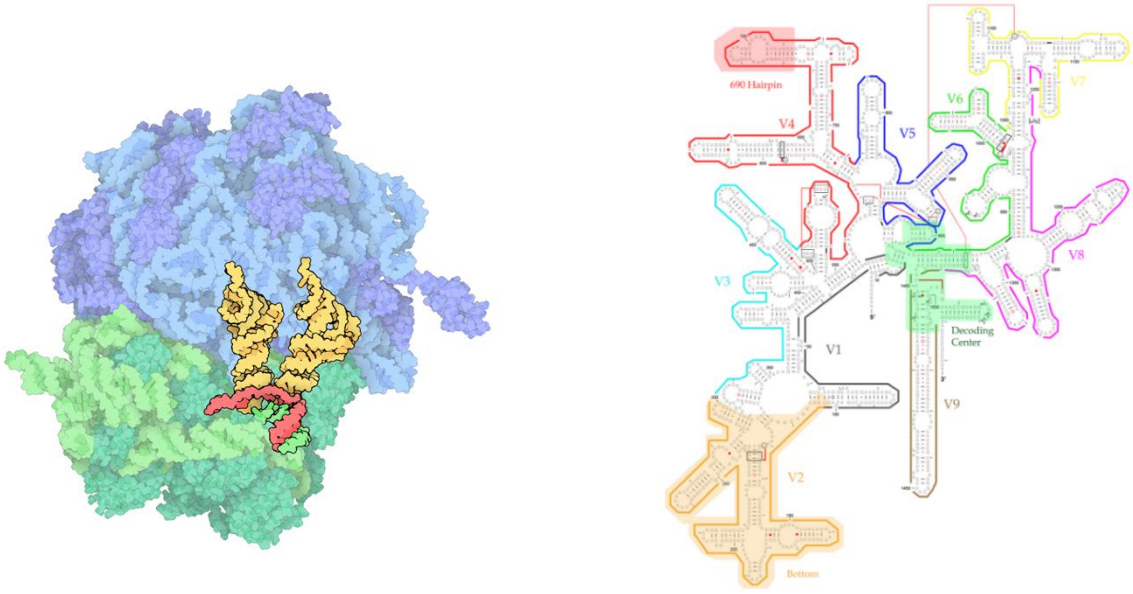

Figure 1: The bacterial ribosome and the fourth variable region of the 16S rRNA gene. (left) The bacterial ribosome. Green portion represents the small subunit (16S) of the complete ribosomal complex. Blue is the large subunit. Two transfer RNAs are shown in yellow, bound a short stretch of mRNA in red. (right) A stem-and-loop diagram of the 16S subunit, with the fourth variable region (V4) shown in red. Picture by David Goodsell doi:10.2210/rcsb_pdb/mom_2010_1. RNA adapted from: Yang et al. BMC Bioinformatics. 2016 Dec;17(1):135.

Homology between 16S rRNA sequences may not always correlate with important character differences that would otherwise define a species. One popular example is described by two *Bacillus* species isolated from river water(25). *B. psychrophilus* was held as a separate species from *B. globisporus* for several

reasons. *B. psychrophilus* could reduce nitrate to nitrite, tolerate higher salinity, consistently fermented different sugars, and was physically smaller than *B. globisporus*. The claim for a separate species was also supported by the low DNA-DNA hybridization values(26). However, if classification were to only rely on the 16S rRNA sequence, these two bacteria would be classified as the same species as the sequence similarity was later found to be 99.8%(27). This example would plot to the blue shaded area of the correlation graph in Figure 2.

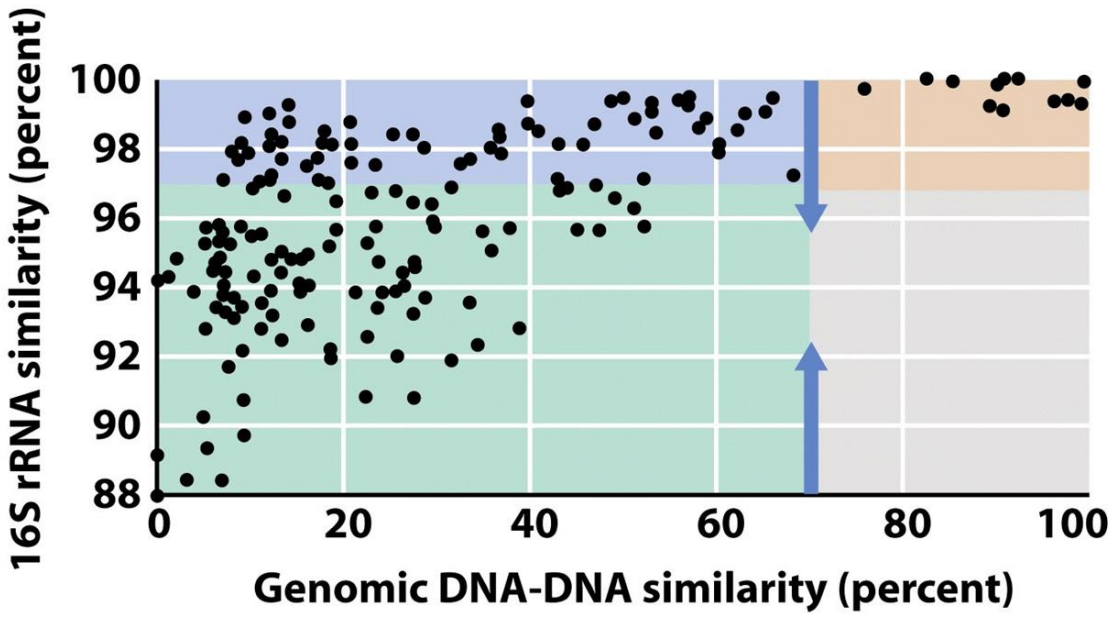

Figure 11-25 Brock Biology of Microorganisms 11/e
© 2006 Pearson Prentice Hall, Inc.

Figure 2: Correlation between DNA-DNA hybridization and the 16S rRNA gene sequence. DNA-DNA similarity above 70% is considered to be from the same species, whereas 16S rRNA similarity above 97% is considered to be from the same species. However, there are many examples of organisms that share high 16S rRNA sequence similarity but whose DNA-DNA similarity is well below 70%.

2.2.5 Lateral gene transfer

Lateral gene transfer (LGT) is a phenomenon demonstrated by bacteria by which they absorb and use genomic information from the environment or other microorganisms, in the form of mobile gene elements. A mobile gene element is any genetic material that is transferred from one bacterium to another in ways that are distinct from the chromosomal duplication during bacterial replication. As a whole, the replication and propagation of mobile gene elements are not

under the control of the host bacteria(28). Instead, they ensure their propagation by either including a replicating origin on the element, or integrating themselves into the host chromosome. Mobile gene elements can be transported between bacteria (pili), and the injection of DNA by bacteriophages(28).

The phenomenon of LGT was first recognized in 1943(29). However, the importance of LGT was not fully appreciated until the mid nineteen-nineties. LGT was viewed as a rare event; so rare that the mutation rate was considered the main driver of bacterial evolution. Recombination events, the shuffling of chunks of DNA as a result of LGT, was viewed as insignificant until David Guttman and Daniel Dykhuizen showed that recombination occurred 50 times more frequently than mutation(30).

LGT allows the size of a bacterial genome to change dramatically. Virulence genes that provide the ability for disease can account for 30% of the genome, as compared between non-pathogenic and pathogenic species(31).

2.2.6 Protein encoding marker genes as classifiers

LGT confounds the conventional paradigm of classification that assumes vertically inherited traits, such as when dealing with wolves or salmon. While there is evidence that almost all genes in the bacterial genome have undergone LGT(32), some genes resist transfer between bacteria, and can be used as phylogenetic markers in the conventional sense. The effectiveness of the gene as a phylogenetic marker can be estimated with the Complexity Hypothesis: the more gene products that are required to assemble into a functional unit, the less likely any of those genes will be successfully transferred between bacterial species. In its original form, the complexity hypothesis was constrained to translational and transcriptional genes(33), but later work has shown that the number of protein to protein interactions (PPI) is a better predictor of resistance to LGT(34). Once sufficient numbers of bacterial genomes representing many species were available, genes were assayed for susceptibility to LGT either with laboratory or computational techniques(32,35), and compiled into a final list of 40 marker genes. Unlike marker genes that code for ribosomal RNA, these 40 genes are

translated into proteins and will be referred to as *protein encoding marker genes*. This study includes the 40 protein encoding marker genes as candidates to identify the bacteria of the bladder microbiome.

2.3 Targeted gene sequencing

Just as the 16S rRNA gene has conserved gene sequences interspersed with regions that show a high amount of variability, the 40 protein encoding marker genes' sequences have conserved and variable regions. While the variable regions for the 16S rRNA gene are well known, the variable regions of the 40 protein encoding marker genes are unknown, and it was necessary to locate them. Identifying the variable regions of a marker gene across many species is achieved by aligning those sequences together. These variable regions can then be targeted by designing PCR primers that anneal in the conserved regions flanking the variable region of interest. The resulting amplicon of the PCR can be sequenced, and phylogenetic study performed on the results.

The steps taken in this study that go from a collection of candidate marker genes to the final PCR primers can be summarized in three steps. The first is to align the orthologous gene sequences of each marker gene found in the Thomas-White dataset in a multi-sequence alignment (MSA). Then, sliding window analysis is performed on the MSA to locate the variable and conserved regions. The final step is an attempt at algorithmically designing primer sets that will anneal to conserved regions flanking one or more variable regions. The goal of designing the primers is not to amplify a variable region of a marker gene in just one of the species found in the Thomas-White dataset, but to amplify the orthologous variable region of a marker gene from all the species in the Thomas-White dataset. The sequences of these orthologues are then used to identify the species of bacteria in the bladder.

2.3.1 Multisequence alignment

Multi-sequence alignment (MSA) is the algorithmic assembly of three or more sequences such that each residue is matched with its counterpart in every other sequence(36). This step is necessary because analogous regions of the marker gene between species will be in different relative positions along the length of their respective sequences. Simultaneously aligning three or more sequences is computationally demanding, and in practice a heuristic method is used instead. A common example of this workaround is progressive alignment. Pairwise global alignments of all sequences in the MSA are scored for percent identity. A distance measure is used to cluster the most similar sequences together, and from this a guide tree is constructed. The guide tree is then used to progressively add sequences to the alignment until the full MSA is complete, as shown in Figure 3.

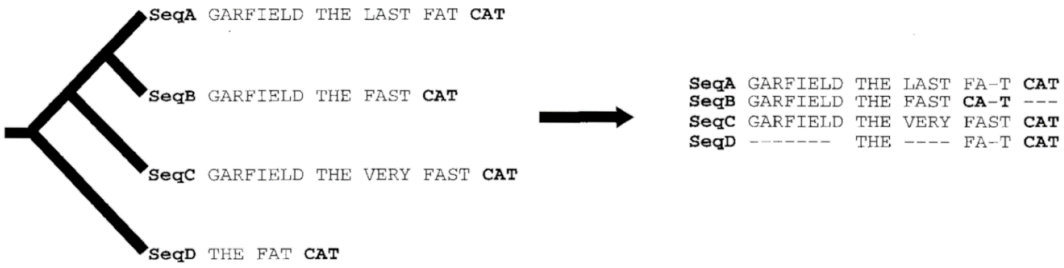

Figure 3: Example of progressive alignment. All possible pairwise alignment combinations are performed for the sequences in a multisequence alignment. Clustering ranks which sequence pairs are the most similar, and a guide tree is constructed based on the similarities. The guide tree is used to construct the final multisequence alignment. Image adapted from Notredame 2000.

Progressive alignment suffers from getting trapped in local minima, and errors introduced early in the MSA are compounded as additional sequences are brought into the alignment. Notredame et al. designed a progressive alignment method called T-coffee that introduces a consistency check(37). Residues that are consistently aligned with each other are encouraged to remain aligned as the larger multi-sequence alignment is assembled, and increases the accuracy of the final MSA(38). This is summarized in Figure 4.

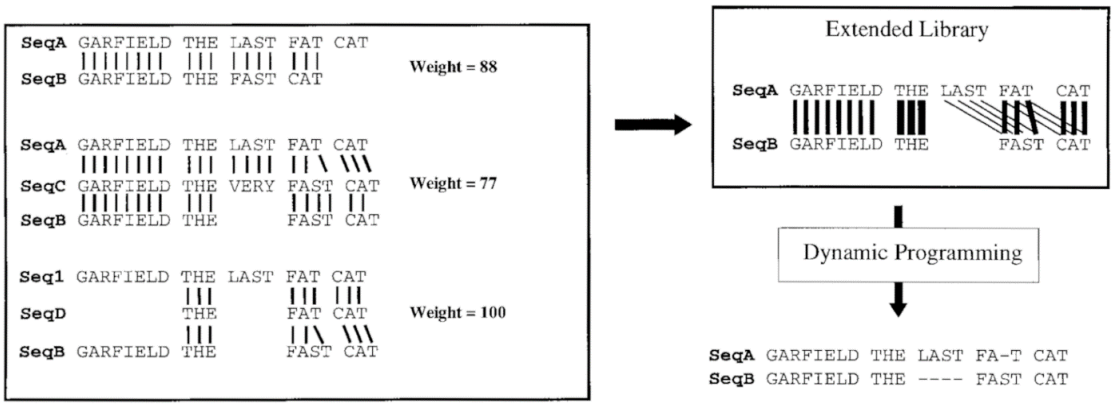

Figure 4: Weighted progressive alignment used by T-coffee. The position in each pairwise alignment is given a weight, and a guide tree is constructed as in progressive alignment. During the construction of the multisequence alignment, the weights ensure that frequently aligned positions are preserved.

2.3.2 Information content of variable regions

Each position in the MSA is a column of nucleotides from the marker gene sequence of each species in the alignment. If the column is comprised of the same nucleotide, there is no information that can be used to identify what species the nucleotide came from. The amount of information is quantified by the Shannon Index, and can be weighted to account for insertions or deletions which impart no information at all. While no single column contains sufficient information for identification, increasingly larger blocks of nucleotides in the MSA will eventually have enough information. Averaging the Shannon Index values of the positions in the block is a way to indicate the amount of useful information of that range. Assessing the average Shannon Index value for successive positions along the MSA, in the manner of sliding a window across the alignment, can be summarized in a graph of high and low areas of information. Peaks in the graph indicate regions of high information and are defined as variable regions, while troughs indicate low information and are defined as conserved regions.

Designing primers that anchor in conserved regions and span across variable regions will have varying degrees of success in assigning taxonomy, as some variable regions will have more information than others along the length of the

marker gene. The amount of taxonomic information in a variable region can be predicted by the height of the peak drawn on the graph that is spanned by the primer set.

2.3.3 Defining variable regions of a MSA

Quantifying the amount of variability in each position along a MSA is done by calculating the Shannon Entropy, defined as

$$S = -\sum_{i} P_i \ln P_i$$

where P is the proportion of each letter i of the alphabet in that position. For the MSA in these examples, the alphabet is the four nucleotides and the gap symbol "-", $i = \{A, T, C, G, -\}$, for a total alphabet of 5 characters.

The minimum Shannon entropy possible is zero, and occurs when there is only one letter of the alphabet present in that position. The maximum Shannon entropy value for any position is reached when all letters in the alphabet are present at that position. For the alphabet i, each letter occurs with a frequency of $\frac{1}{5}$.

Example

Suppose 10 sequences are aligned, and at some position x there are the base pairs $x = (A, A, A, T, A, A, A, -, C, A)$. The value S at that position is then

$$\begin{aligned}
\mathrm{S} &= -\sum_{i=\{A,T,C,G,-\}} P_i \ln P_i \\
&= -\left(P_A \ln P_A + P_T \ln P_T + P_C \ln P_C + P_G \ln P_G + P_{gap} \ln P_{gap}\right) \\
&= -\left(\frac{7}{10}\ln\left(\frac{7}{10}\right) + \frac{1}{10}\ln\left(\frac{1}{10}\right) + \frac{1}{10}\ln\left(\frac{1}{10}\right) + 0 + \frac{1}{10}\ln\left(\frac{1}{10}\right)\right) \\
&= -\Big(\big(.7 \times (-.36)\big) + \big(.1 \times (-2.3)\big) + \big(.1 \times (-2.3)\big) + 0 + \big(.1 \times (-2.3)\big)\Big) \\
&= -\big((-.252) + (-.23) + (-.23) + (-.23)\big) \\
&= .942
\end{aligned}$$

The max value for this example is equal to

$$S_{max} = -\sum_{i=\{A,T,C,G,-\}} P_i \ln P_i$$

$$= -\left(\frac{1}{5}\ln\left(\frac{1}{5}\right) + \frac{1}{5}\ln\left(\frac{1}{5}\right) + \frac{1}{5}\ln\left(\frac{1}{5}\right) + \frac{1}{5}\ln\left(\frac{1}{5}\right) + \frac{1}{5}\ln\left(\frac{1}{5}\right)\right)$$

$$= -\big((-.322) + (-.322) + (-.322) + (-.322) + (-.322)\big)$$

$$= 1.61$$

To make the graphs of the sliding window analysis, each data point is the average of all Shannon entropy values inside the window, as the window slides across the length of the MSA. The optimal size of the window is determined empirically.

2.3.4 Weighted entropy scores

The Shannon Entropy treats gaps in a sequence as information, where in practice gaps are an absence of information. Multisequence alignments can generate many columns of gap characters that indicate insertions or deletions (indels) in the respective sequences that make up the MSA (the *RpoB* gene is a good example), but a graph of the Shannon Entropy will interpret these regions as conserved sequence. This problem can be solved by weighting the entropy scores against gaps, and I used the weighting method described in Valdar 2002(39).

Valdar 2002 starts by defining how conserved a position is by

$$C_{(x)} = \big(1 - t(x)\big) \times \big(1 - g(x)\big)$$

where $g(x)$ is the gap penalty, which will be the proportion of how many gaps are in the column at position x. The entropy is the inverse,

$$E_{(x)} = 1 - C_{(x)}$$

At each position x of the MSA, the number of symbols α are accounted for and the weighted frequency is calculated.

$$t_x = \lambda_t \sum_{\alpha}^{A} p_\alpha \ln p_\alpha$$

where p_α is each symbol in the alphabet A={A,T,C,G,–} at the position p, and λ_t is a scaling constant equal to $\lambda_t = (\ln 5)^{-1}$.

For each position along the MSA, the probability of observing symbol α is the sum weight of all the sequences with symbol α at position x.

$$p_\alpha = \sum_{i \in \{i : s_i(x) = \alpha\}} w_i$$

Example

Suppose there is the following multisequence alignment.

seq	**position**					
	1	2	3	4	5	6
A	A	T	C	-	G	-
B	A	T	T	C	G	-
C	A	-	C	G	G	C

Table 1: Multisequence alignment for weighted entropy example. Rows are sequences, columns are positions in the MSA.

Ranking the amount of information in this alignment should reflect that while positions 1, 2, 5 and 6 are equally noninformative, the presence of gaps in position 2 and 6 should count against them. The weighting should also reflect that there is more information in position 3 than there is in position 4.

Gaps in the example MSA should not be considered as conserved, but as less information to make an assessment. Visually ranking the example sequence by amount of descending information available, and then by ascending heterogeneity (the entropy) is shown in Table 2.

seq	position					
	1	5	3	2	4	6
A	A	G	C	T	-	-
B	A	G	T	T	C	-
C	A	G	C	-	G	C

Table 2: Multisequence alignment, with each position ranked from left to right in order of amount of descending information available and ascending amount of heterogeneity. Position 6 has the least amount of information and the least heterogeneity, while positions 1 and 5 have the most amount of information but least amount of heterogeneity.

Any entropy calculations that weight against gaps should reflect this ordering.

The weight of each sequence is determined by the following:

$$w_i = \frac{1}{L}\sum_{x}^{L}\frac{1}{k_x n_{x_i}}$$

where i is the sequence, L is the length of the sequence (6 in this example), x is the position along the sequence, k_x is the number of unique symbols at position x , and n_{x_i} is the number of times the symbol at position x in sequence i occurs in the whole MSA at position x. Using the MSA in Table 1, for $i = B$, and $x = 2$, the symbol is “T”. In that column, “T” occurs twice. The gap symbol ‘-’ occurs only once.

Moving along the length of sequence B, the weight of the sequence is

$$w_B = \frac{1}{6}\left(\frac{1}{3}+\frac{1}{4}+\frac{1}{2}+\frac{1}{3}+\frac{1}{3}+\frac{1}{4}\right) = .33$$

The other weights for A and C are

$$w_A = \frac{1}{6}\left(\frac{1}{3}+\frac{1}{4}+\frac{1}{4}+\frac{1}{3}+\frac{1}{3}+\frac{1}{4}\right) = .29$$

$$w_C = \frac{1}{6}\left(\frac{1}{3}+\frac{1}{2}+\frac{1}{2}+\frac{1}{3}+\frac{1}{3}+\frac{1}{2}\right) = .416$$

Calculating the entropy of p_1 of the example MSA in Table 1 is comparatively easier. There is only 'A' in that column, so the probability of seeing that symbol in that position is

$$p_1 = w_A + w_B + w_c = 1$$

Then the log value is

$$p_1 \ln p_1 = 1 \ln 1 = 0$$

And t_1 is zero too.

$$t_1 = \lambda_t \times 0 = 0$$

Finally, the gap penalty is $g(1) = 0$, and the entropy of position 1 is.

$$E_1 = 1 - \big((1-0) \times (1-0)\big) = 0$$

The position is completely conserved, so the entropy is zero. For p_2

$$t_2 = (w_A + w_B)\ln(w_A + w_B) + (w_C)\ln(w_C) = -.65$$

The final weighted entropy values for the example are

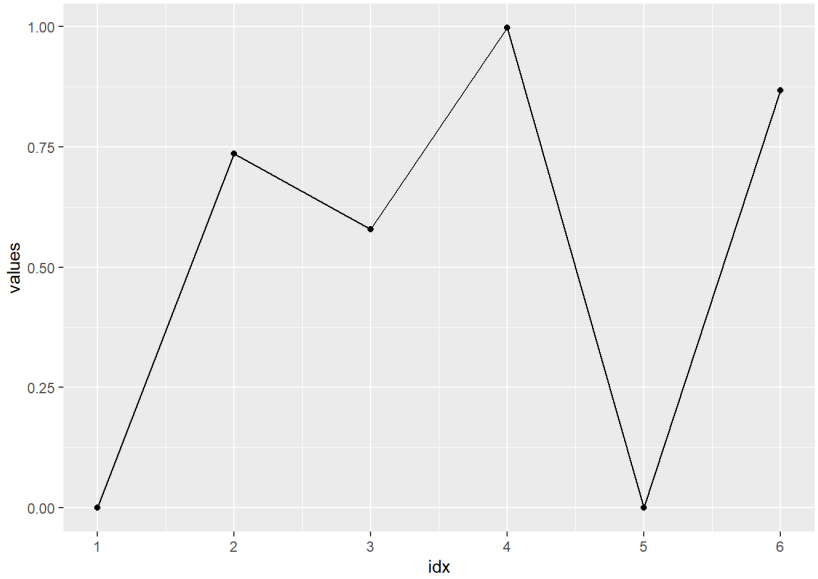

Figure 5: Graph of weighted entropy calculated for each position of the example multisequence alignment. Ranking the positions is done by reading from bottom to top. Positions ranked by weighted entropy are 1&5,3,2,6,4.

Ranking each position by the increasing amount of entropy gives 1&5, 3, 2, 6, 4. Visually weighting the position as above gave 1&5, 3, 2, 4, 6. The difference in the order between the last two positions can be interpreted as while position 6 appears completely conserved (there's only the one 'C') the two '-' weights against it, and position 4 has both more information and more heterogeneity.

2.3.3 Primer design

The PCR primer is a small oligonucleotide that anneals to a complementary sequence on the DNA template strand. Primers are designed with the intent of annealing to only the target area, and in a specific temperature range. Designing a primer to amplify a targeted region on one DNA template can be done with a text editor and a calculator. Designing a primer that will amplify a targeted region across several different DNA templates, such as a variable region of a marker gene, requires the aid of an algorithm.

The Degenerate Primer Design problem

For an authoritative treatment of the degenerate design problem, I refer the reader to Linhart and Shamir 2002(40). A primer is called *degenerate* if any of the positions of the primer can have more than one possible base. On paper, this is done by using the International Union of Pure and Applied Chemistry (IUPAC) codes shown in Table 3.

base pair	C	T	A	G	W	S	M	K	R	Y	D	V	B	H	N
C	•					•	•			•		•	•	•	•
T		•			•			•		•	•		•	•	•
A			•		•		•		•		•	•		•	•
G				•		•		•	•		•	•	•		•

Table 3: Table of IUPAC designations for ambiguous nucleotides

A naive approach to designing degenerate primers is to align the sequences, pick a window of the right size and make an oligo for each unique sequence in the window. This is not the best way because the degeneracy will always be higher

than necessary, and it ignores the inevitable indels in a MSA. Therefore, the goal of degenerate primer design is to find the oligo with the smallest degeneracy that matches the most number of sequences.

The alphabet to be used is the set $A = \{A, T, C, G\}$. The primer P is the string $P = p_1, p_2, \ldots p_i$, where $p_i \in A$. The length of the primer is denoted by k. The degeneracy of a primer is denoted by d and calculated by

$$d = \prod_{i}^{k} p_i$$

A primer matches a sequence string S if P is a substring of S. The number of sequence strings that a degenerate primer matches is denoted by m.

The problem can be stated in words as: Given a set of n sequence strings and the integers k, d, and m - is there a primer of length k and degeneracy at most d that matches at least m sequence strings?

For real world applications to solve the degenerate primer design problem, each solution is NP-complete(40). However, heuristic methods are available.

DegePrime

To design the degenerate primers for this study, I chose to use the DegePrime(41) program. This algorithm finds a solution to the degenerate primer design problem by bootstrapping the weighted subsequences in a window of the MSA. The unique substrings are randomly sampled accounting for the weights, and checked for coverage. After repeating 100 times, the sampling that performed the best is presented as the degenerate primer.

Degenerate primer synthesis

The method of synthesizing a degenerate primer set for PCR impacts the reaction conditions needed to generate amplicons. In a PCR reaction, the use of a degenerate primer is realized by including all the possible combinations of primer sequence in the reaction volume. Given a primer GTGCCMGC, the primer solution would have an equal mix of GTGCCAGC and GTGCCCGC(42).

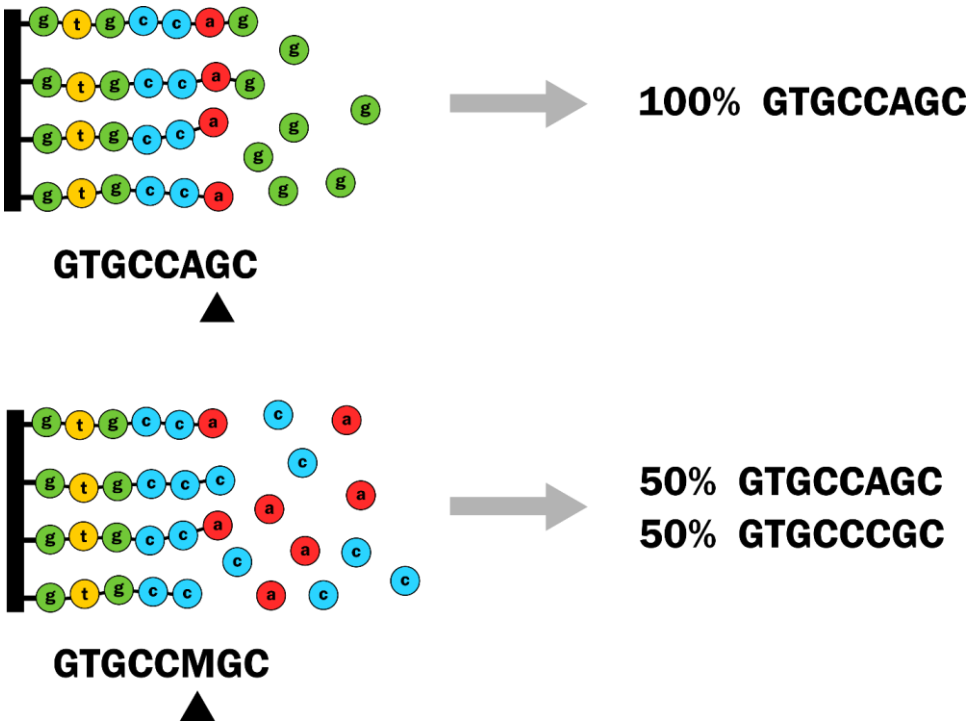

Figure 6: Degenerate primer synthesis. The degenerate nucleotide M is generated by injecting equal amounts of adenosine and cytosine nucleotides into the reaction chamber, resulting in near equal amounts of GTGCCAGC and GTGCCCGC oligonucleotides.

Primer synthesis is a chemical reaction cycle that adds one nucleotide to a growing chain of nucleotides anchored to a support [https://www.idtdna.com/pages/education/decoded/article/oligo-synthesis-why-idt-leads-the-oligo-industry]. Each nucleotide has its 5' carbon blocked by a protection group, and elongation of the oligo chain is a cycle of deblocking, chemical binding with the new nucleotide, stabilization of the bond, and blocking the reactive site on the oligo. To synthesize degenerate primers, several different nucleotides are introduced at the binding stage of the reaction [https://www.idtdna.com/pages/products/custom-dna-rna/mixed-bases]. For

example, if an oligo GTGCCMGC has a degeneracy of two, and the degenerate position is {C,A}, the synthesizer machine will inject an equal amount of dATP and dCTP nucleotides into the reaction chamber, as shown in Figure 6. The final oligos in the chamber are a (near) equal mix of (GTGCCAGC, GTGCCCGC).

When diluting a non-degenerate primer to a stock solution for use in PCR, the molarity is commonly 100μM. When added to the PCR reagent mix, the final molarity is commonly ~1μM per primer. For the example primer GTGCCMGC, the molarity of each degenerate oligo in a common stock dilution is actually 50μM, and much less for primers with higher degeneracy. Optimizing the reaction conditions for PCR requires some additional adjustment to ensure the correct molarity of the degenerate primers.

2.4 The Thomas-White dataset

This thesis relies on a set of identified bacteria found in urine samples collected from women symptomatic for Overactive Bladder (OAB)(43) and a control set of otherwise healthy women. This set of bacteria is then used to evaluate the taxonomic resolution and performance of the 58 classification schemes.

In their publication, Thomas-White et al describe catheterizing and collecting urine samples from women who were symptomatic and asymptomatic for overactive bladder syndrome. Bacteria were cultured from the urine samples using expanded quantitative urine culture (EQUC)(44), isolated, and subject to whole genome sequencing (WGS). In each genomic sequence, the 40 marker gene sequences described in section 2.2.6 were located and used to identify the species of the bacterial isolate.

The set of bacterial species cultured from the urine samples, and identified by whole genome sequencing, will be referred to as the *Thomas-White dataset*. As shown in Figure 7, this set represents 140 sample/species pairs and 78 identified bacterial species from 77 urine samples.

In general, the samples taken from patients with OAB symptoms contain a larger diversity of species than the samples taken from asymptomatic patients, shown in

Figure 8. In the entire dataset, the diversity of bacteria found in the samples ranges from 1 to 23 different species. The majority of samples had less than 5 species.

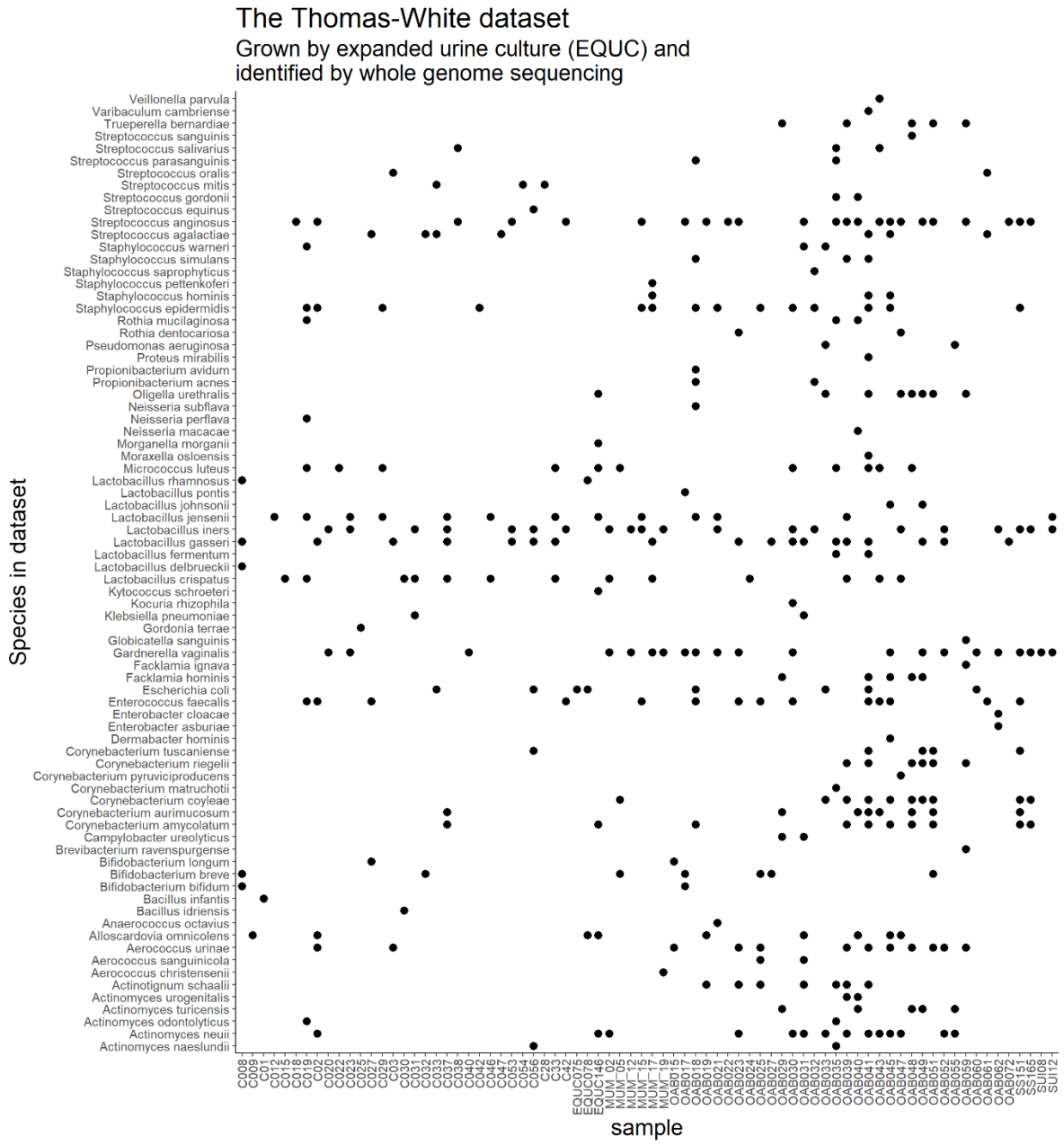

Figure 7: The Thomas-White dataset. Each dot indicates a bacterial species that was identified, for the isolates in each urine sample successfully cultured by the expanded quantitative urine culture (EQUC) method. Sample names that include "OAB" indicate patients that were experiencing overactive bladder symptoms at the time of sampling. Total number of species identified is 79, and the total number of samples is 77.

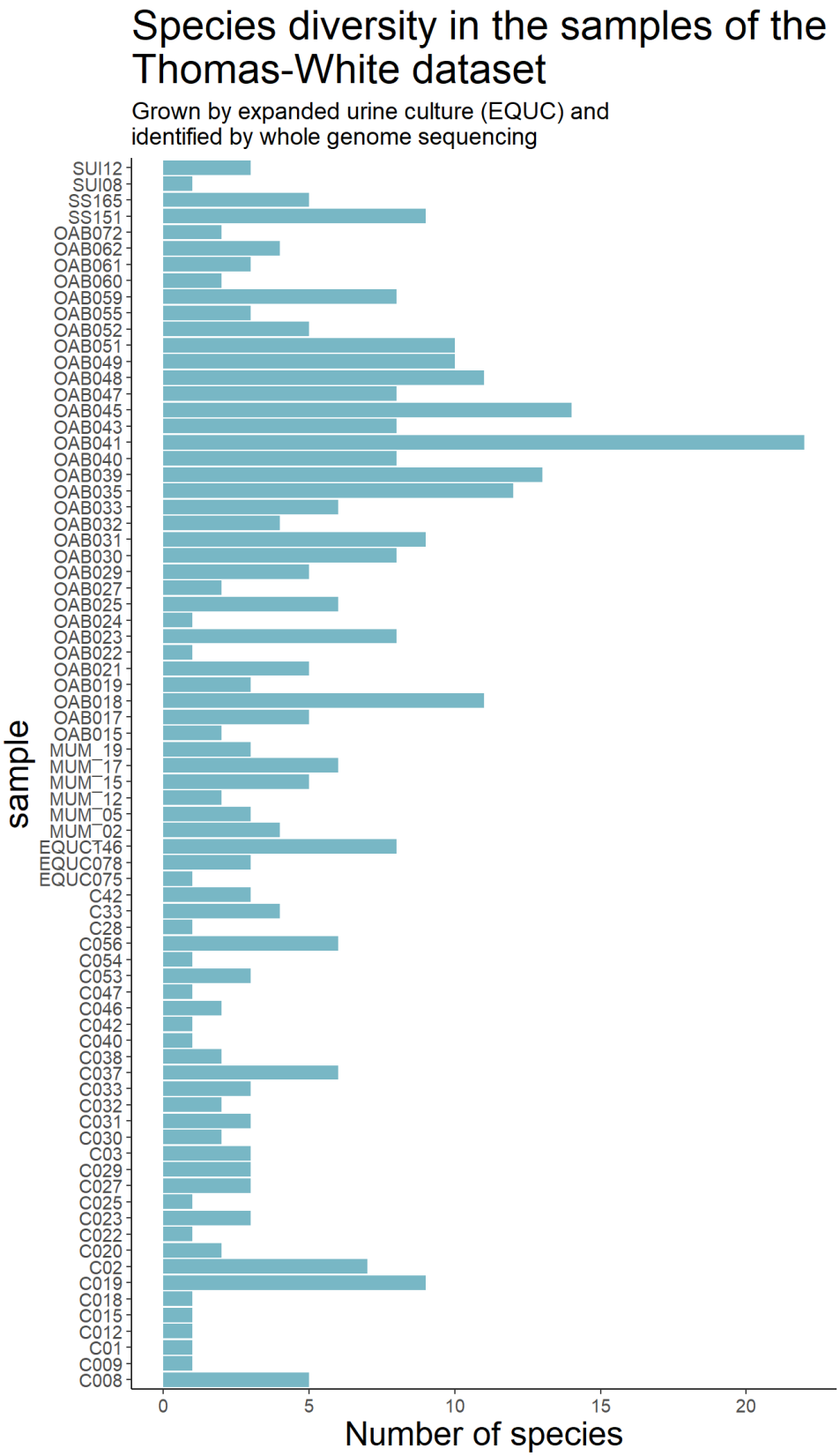

Figure 8: Species diversity of the Thomas-White dataset. Sample names that include "OAB" indicate patients that were experiencing overactive bladder symptoms at the time of sampling. Total number of species identified is 79, and the total number of samples is 77.

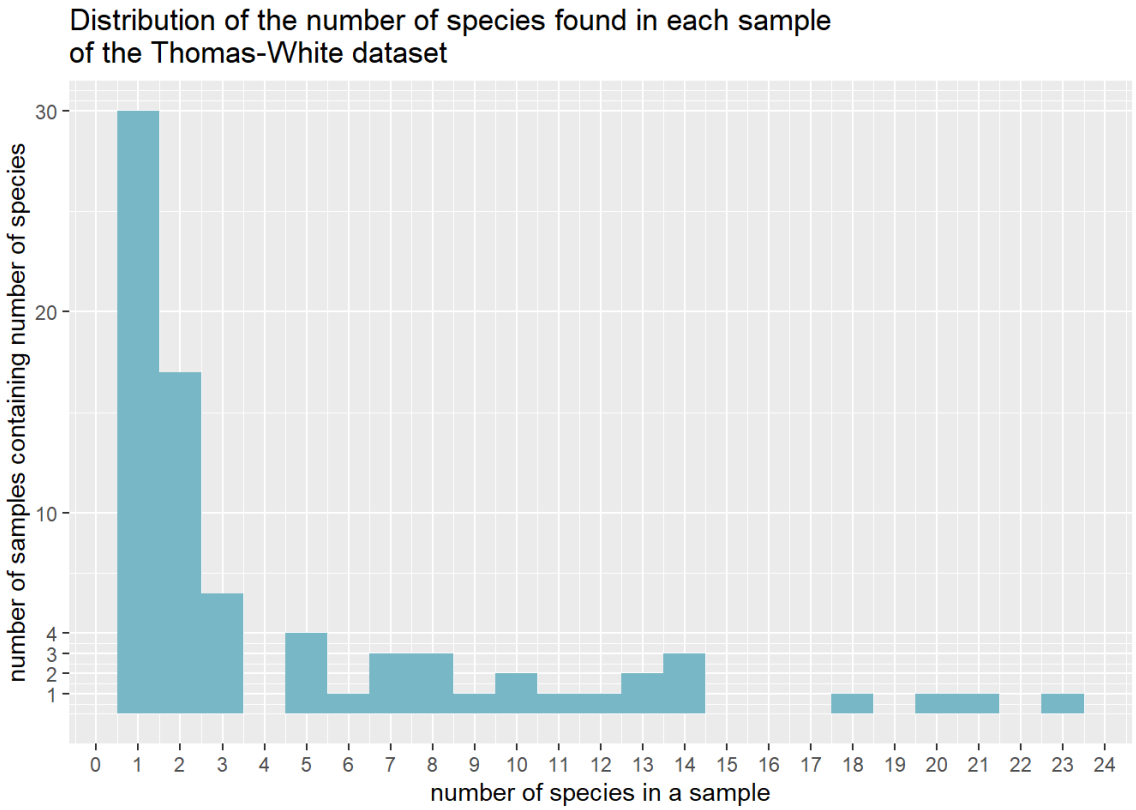

Figure 9: Distribution of the species diversity of the Thomas-White dataset. The majority of samples had 4 or fewer species identified.

Figure 9 shows a surprising 38% of the samples contained only one species. This rarity may be attributed to the difficulty of culturing bacteria from urine samples that contain very low biomass, and a culture-independent method of identification, such as metagenomic sequencing, would be better suited to describe the between-sample diversity. This set of species will be referred to as "singletons", and are listed in Table 4. One item of note is that while Lactobacillus has a large representation of species in the Thomas-White dataset, two of those species, *L. delbrueckii* and *L. pontis*, are singletons.

genus	species
Aerococcus	christensenii
Anaerococcus	octavius
Bacillus	idriensis, infantis
Brevibacterium	ravenspurgense
Corynebacterium	matruchotii, pyruviciproducens
Dermabacter	hominis
Enterobacter	asburiae, cloacae
Facklamia	ignava
Globicatella	sanguinis
Gordonia	terrae
Kocuria	rhizophila
Kytococcus	schroeteri
Lactobacillus	delbrueckii, pontis
Moraxella	osloensis
Morganella	morganii
Neisseria	macacae, perflava, subflava
Propionibacterium	avidum
Proteus	mirabilis
Staphylococcus	pettenkoferi, saprophyticus
Streptococcus	equinus, sanguinis
Varibaculum	cambriense
Veillonella	parvula

Table 4: List of species that were only found once in the entire Thomas-White dataset (singletons).

The species abundance of the Thomas-White dataset are shown in Figure 10. The most commonly found species in the Thomas-White dataset are *Streptococcus anginosus*, *Gardnerella vaginalis*, *Staphylococcus epidermis*, *Enterococcus faecalis*, and the *Lactobacillus* species *L. jensenii*, *L. iners*, *L. gasseri* and *L.*

crispatus. Additional *Lactobacillus* species isolated and identified are *L. rhamnosus*, *L. pontis*, *L. johnsonii*, *L. fermentum* and *L. delbruckii*.

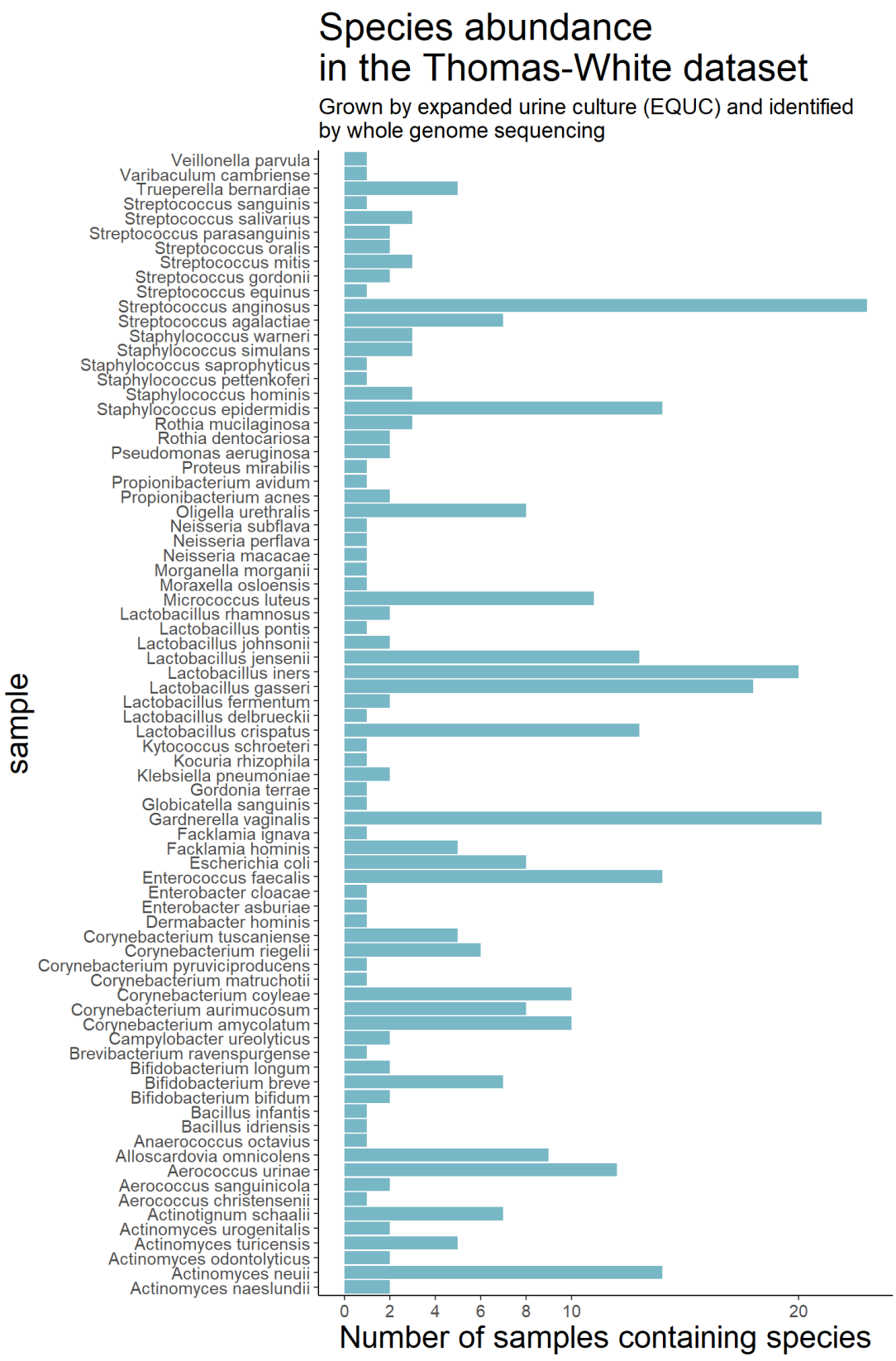

Figure 10: Species abundance of the Thomas-White dataset. Many of the species in this dataset occurred only once. Total number of species identified is 79, and the total number of samples is 77.

2.5 Record linkage

Sections 2.2 to 2.4 have described how the 16S rRNA gene and the 40 protein encoding marker genes are used as the identifier component of a classification scheme. This next sections will review the remaining two components: the classifier and the database. As applied in this study, Record Linkage is the process by which the genetic information held in two separate records is compared, and a determination of whether or not they are referring to the same species of bacteria(45) is made.

In this study, the Naive Bayes and BLCA classification algorithms do the procedure of record linkage, and employ Bayes theorem to calculate the probability of a match. However, these classifiers have a more restricted use of the calculated probabilities than in other classification applications. For both the BLCA and Naive Bayes classifiers, the highest calculated probability (argmax) is designated the match, and the remaining candidates are assigned as non-matches. This will be explained in more detail in section 2.5.2.

A second type of Record Linkage produces two mutually exclusive categories: either the record pair is an exact match, or the record pair is not a match(46). No probabilities are calculated. In this study, exact matching assigns taxonomy by searching the reference database for a sequence that is an exact match to the query, and even one nucleotide difference between the record pairs is sufficient to assign the pair as a non-match.

It is worth mentioning that classification algorithms like BLCA and Naive Bayes are not *predictive*, like log-odds or regression. A predictive model is when an unknown data point is given a probability that the data is a member of possible outcomes, from high probability all the way down to possibly a very small probability. Predicting the next outcome of a coin toss based on previous trails results in a probability assigned for the heads outcome, a probability assigned for the tails outcome, and a very, very small probability that the coin will land on its edge and stick. BLCA and Naïve Bayes are based on a model that is different, in that the probability is only there to assist in assigning a record pair as a match. A

predictive model assumes the characteristics of one of the possible classes is already present in the new data point. A classification model assigns the new data point the characteristics of one of the possible classes, based on how similar the new data is to the classes that are available to choose from. It could be that the values of the probability are so low that a reasonable person would say the new data really should not be assigned to any of the available classes.

For example, suppose a classification scheme is designed for classifying animals into one of two categories, either mammals or fish. Mammal are then defined based on the characteristics of a camel and a giraffe: they have hair, a long neck, produce milk and give birth to live young. Fish are then defined with opposite characteristics of a mammal. When attempting to classify an ostrich, there is no good category to choose from, as the similarity between the ostrich and either category are so small. The classifier algorithm will still classify the ostrich as a mammal, because the bird does have one thing in common with the mammal class (a long neck) and zero things in common with the fish class, whereas reasonable person would say the ostrich should be left unclassified.

2.5.1 Databases used in this study

Greengenes

The Greengenes database(47) was created with the intent to screen sequences for errors generated during PCR (chimeras). Secondary goals were to created standardized fields that were compatible with the bioinformatic suite of analysis programs ARB(48), and taxonomy assignment. Sequences were gathered from the Ribosomal Database Project (RDP) and The International Nucleotide Sequence Database Collaboration (INSDC). Bergey's manual and several experts' opinion were consulted for taxonomic identity. Greengenes predicts phylogeny based on a novel pairwise alignment algorithm to compare a 16S sequence with the seed alignment.

Silva

The Silva database(11) was created as a repository for curated ribosomal sequences from all forms of life. It includes the small subunit (SSU) and large subunit (LSU) sequences of the ribosome. In addition, the database is not limited to full length sequences, but includes much shorter fragments. Sequences are gathered from INSDC, and taxonomy is obtained from Nomenclature Up-to-date and the International Journal of Systematic and Evolutionary Microbiology (IJSEM). Phylogeny is predicted by pairwise sequence comparison to seed alignment through the use of a modified Needleman-Wunsh algorithm.

NCBI

By far, the NCBI taxonomic database(49) is the largest and most comprehensive, with the exception of online sequence aggregators like the All-species Living Tree(50). All sequences are gathered and curated from the INSDC, and taxonomy is obtained from the primary systematic literature. At a minimum, each sequence must map to the species level taxon. Smaller specialty databases are available, such as the 16S SSU database or the representative prokaryotic genomic database.

Other databases considered

Two additional databases were considered, but ultimately rejected from this study. The Ribosomal Database Project offers a curated database of 16S rRNA gene sequences(51). However, this database has no species level taxonomic information, and was rejected. The STIRRUPS database is a database of 16S rRNA gene sequence reference sequences for bacterial taxa likely to be associated with vaginal health(52). However, the reference sequences are limited to the V1-V3 regions of the 16S rRNA gene sequence (about 485 nucleotides) and was rejected.

2.5.2 Classification algorithms used in this study

Classification algorithms that depend on the DNA sequence of bacteria can be grouped into two methods. The first is pairwise alignment. This method is done by comparing the query and reference sequences by global alignment, and noting the match or mismatch at each position. “Exact matching” is the term used when the query is an exact copy of the reference, but many times the sequences differ by some number of nucleotides. At some threshold of difference, it is agreed that the sequences came from different taxa. The second method uses the concept of a “*k*-mer”. This method is done by assessing the frequency that key subsequences of length *k* occur in the query sequence and pooled sequences of the reference. If enough key subsequences are shared between the query and reference, the sequences are agreed to have come from the same taxon. There are several examples of each kind of method, but for this study I chose one algorithm from each that employed a Bayesian approach.

For pairwise alignment, I chose the Bayesian Lowest Common Ancestor (BLCA)(12) algorithm. For the *k*-mer method, I chose the Naive Bayes classifier as implemented by Qiime2(53). In addition, I also use the exact matching function as implemented by DADA2(54).

BLCA

BLCA assigns taxonomy in four steps. The first step is to find a set of closely similar sequences by using the sequence as a BLAST query. Next, Clustal(55) is used to create a multisequence alignment (MSA) of the query sequence and all the hits that pass a threshold requirement. Third, a Bayesian probability that the query sequence could be a subsequence of each sequence returned by the BLAST search is calculated using the pairwise alignment scores. Bayes equation for this application is

$$\Pr(T_i|Q) = \frac{\Pr(Q|T_i)\Pr(T_i)}{\sum_{i=1}^{m} \Pr(Q|T_i)\Pr(T_i)}$$

Where $\Pr(Q|T_i)\Pr(T_i)$ is the likelihood of observing sequence Q if derived from taxon T, and $\sum_{i=1}^{m} Pr(Q|T_i)\Pr(T_i)$ is the sum of all pairwise alignment scores in the MSA. $\Pr(T_i)$ is the model, and the authors state that because the query is equally likely (at first) to be derived from all possible bacteria, they assign the uninformative prior to the model. Because this prior is a constant value, it cancels out of the equation.

Stating the problem in words, the likelihood that a particular taxonomy is the correct one for a query sequence is the proportion of the pairwise alignment score to the maximum possible alignment score, divided by the sum of all pairwise alignment scores of the MSA.

Finally, a confidence score is calculated for the result that reflects how often the classification is changed by random sampling of the pairwise alignment assigned as a match (bootstrapping). Highly similar sequences compared between the reference and query database records will have little to no change in the taxonomy assignment, while more dissimilar sequences will have more frequent changes to the original assignment.

Example

Suppose there are two blast hits, T_1 and T_2, returned for a query sequence Q, and these two hits are pairwise aligned to the query sequence during the construction of a MSA.

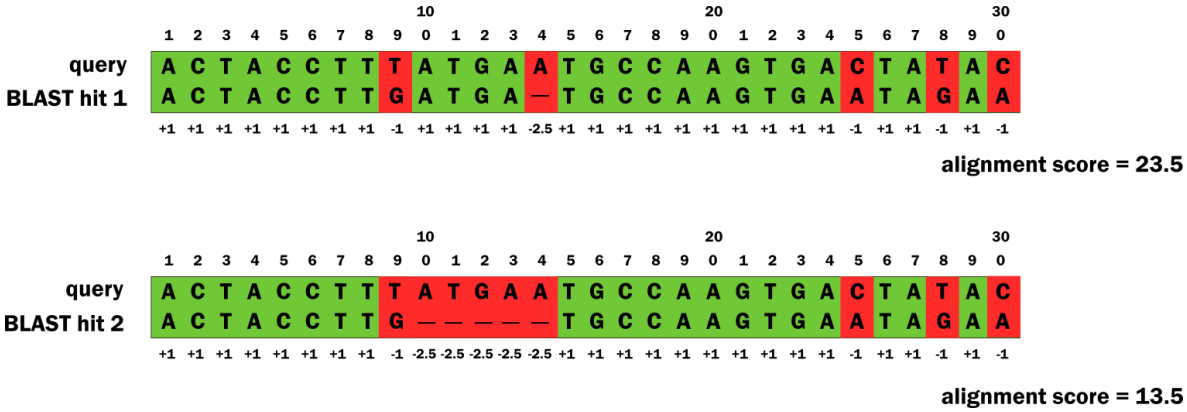

Figure 11: Example of classification by BLCA. The alignment score of the pairwise global alignments between the query sequence and a candidates from the reference database is used to assess the degree of similarity. Many candidates can be aligned and scored. The highest alignment score is designated as the match, and the taxonomy from the reference sequence is given to the query sequence. The position of the global alignment is indicated by the numbers on the top of the alignment, while the score of each position is listed below. Green columns are matches, red columns are mismatches (or insertions or deletions).

In Figure 11, the index of the sequence position is placed above the alignment, and the score for each position below. Each match is worth +1, each mismatch is worth -1, and insertions or deletions (indels) are -2.5. The score of the query aligned to itself is 30.

$$\Pr(T_1|Q) = \frac{\frac{23.5}{30}}{23.5 + 13.5} = .021$$

$$\Pr(T_2|Q) = \frac{\frac{13.5}{30}}{23.5 + 13.5} = .012$$

The taxonomy of the hit with the highest value is saved, in this case T_1. Ties between equally likely hit sequences are decided by randomly choice.

Calculating a confidence score is a bootstrap process. A random sample of each pairwise alignment is taken with replacement. In the code, this is done by randomly choosing an index value of the query sequence, with replacement, from

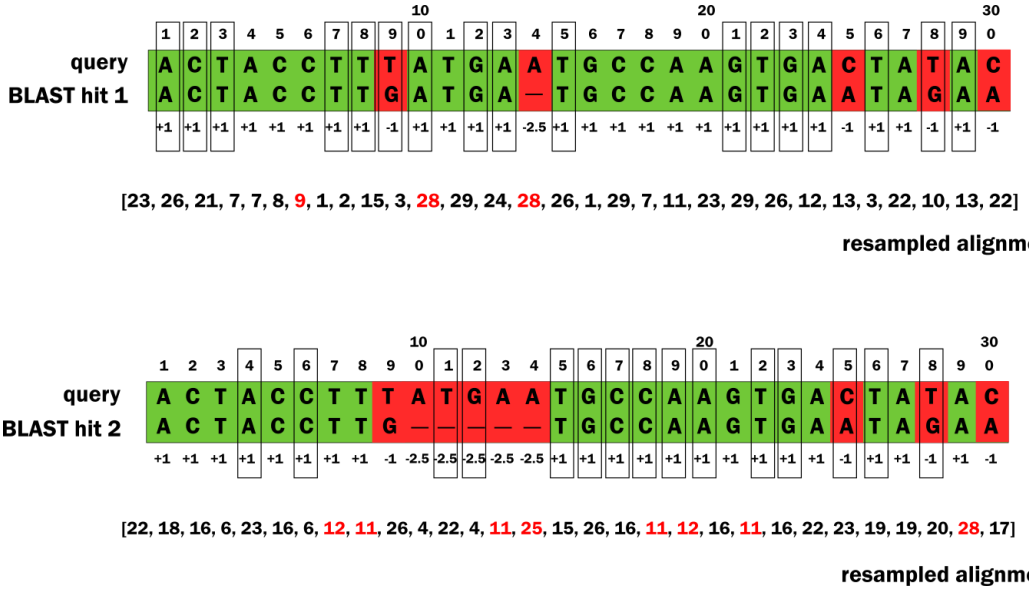

Figure 12: Example of how the confidence score is calculated by BLCA. After all pairwise global alignments are finished, the positions of each alignment are randomly sampled with replacement until the number of samples equals the original alignment. The scores of each position are used to calculate an alignment score, a posterior probability is calculated again, and the argmax of all calculated posterior probabilities is selected. This constitutes one bootstrap iteration. After many iterations, the confidence score is the number of times the original assigned match has occurred.

each pairwise alignment until a list of equal length to the query is reached. This is shown in Figure 12.

The Bayesian probability is calculated for each newly generated pair, and the taxonomy with the highest probability is saved again.

$$\Pr(T_1|Q) = \frac{\frac{27}{30}}{27 + 13} = .023$$

$$\Pr(T_2|Q) = \frac{\frac{13}{30}}{27 + 13} = .011$$

The bootstrap is repeated many times (the default value is 100), and a tally of all the iterations is made. The taxonomy with the highest proportion of all the iterations is assigned to the query. For this example, T1 got the highest tally, and the taxonomy of T1 is assigned to the query. The value of the highest proportion is used to indicate the level of confidence, and in this example the confidence score for T1 is 100%.

Naive Bayes

In Wang et al., the authors recast a solution to the problem of classifying text to that of bacterial classification(10).

Given a DNA sequence from a bacterium of unknown taxonomy. The probability that you are observing the genus G given the sequence S is

$$Pr(G|S) = \frac{Pr(S|G)Pr(G)}{Pr(S)}$$

Where $Pr(G)$ is the prior probability of the sequence S being a member of genus G, and $Pr(S)$ is the probability of observing sequence S from any genus. As in BLCA, the authors state that all genera are equally likely, so $Pr(G)$ cancels out. The probability of observing sequence S from any genus is given as uniform, so $Pr(S)$ cancels out. The "words" used in the classifier are also assumed to be independently and identically distributed (the Bag of Words model), so the Bayes equation is reduced to

$$Pr(G|S) = Pr(S|G)$$

Stating the problem in words, the probability that sequence S derived from genus G is the conditional probability of observing the "words" in a sequence given the genus.

Generating words and the feature space

Subsetting a gene sequence into subsequences of size k (the "words") reduces the size of the feature space needed to calculate the conditional probability. The authors empirically determined a value $k = 8$ to be optimal.

The training sequences and sequences of unknown taxonomy are split into 8-mers in the manner of a window sliding across the sequence. The first 8 nucleotides are the first word $[n_1 : n_8]$, the second nucleotide to the 9th $[n_2 : n_9]$ is the second word, and so on, as shown in Figure 13.

Figure 13: Construction of the Naive Bayes corpus training set. All possible k-mer subsequences are generated from a sequence through a sliding window method.

The total number of unique words generated from the training set is called the corpus by the authors. In general, the corpus is smaller than the feature space and the number words from an unknown sequence will also be smaller than the feature space.

The feature space is the Cartesian product of $\{a, t, c, g\}$ in all 8 positions. This is like an incrementing odometer, where each wheel has "a", "t", "c", and "g" (Figure 14). The result is a dictionary of 65,536 8-mer words with which to create any sequence. Frequencies are calculated for each 8-mer in the dictionary according to how often they appear in the training set. The probability that the sequence S is from genus G is calculated by looking up the frequency value of the words in the unknown sequence and multiplying them all together, for every genus in the training set.

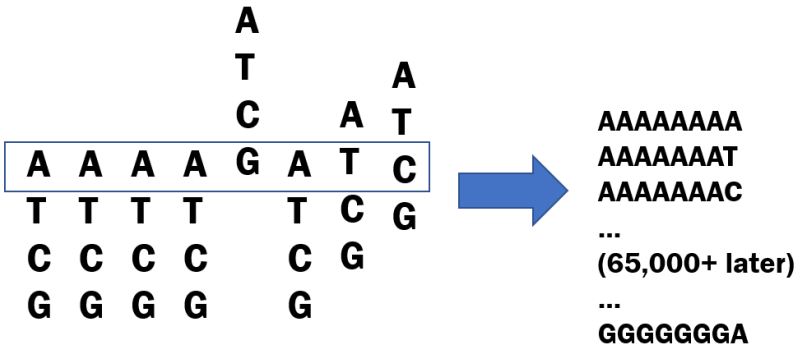

Figure 14: Construction of the Naive Bayes feature space. All possible words are generated for length k=8 in the manner of an incrementing odometer.

Calculating $\boldsymbol{Pr(S|G)}$

The following definitions will be used. $S = \{s_1, s_2, \ldots s_n\}$ is the set of all sequences to classify and $G = \{g_1, g_2, \ldots g_j\}$ is the set of all genera in the training set. Let $G = \{g_1, g_2, \ldots g_j\}$ represent the words in the feature space, and let $V = \{v_1, v_2, \ldots v_i\}$ represent the words generated from an unknown sequence.

corpus words in the feature space

We start with calculating the frequency of the "words" and work up. From the corpus comprising N sequences, $n(w_i)$ is defined as the number of sequences containing the word w_i . To calculate the conditional probabilities for each of the words in the feature space, make a note of how many sequences contain that word. Some words will be absent from the corpus. While it is expected that there may be no information available for that word, the actual likelihood is not *zero*. It will be some small number that relates to the number of sequences. The authors used a variant of Laplace's Rule of Succession(56) to calculate the expected likelihood estimate P_i.

$$P_i = \frac{n(w_i) + \frac{1}{2}}{N + 1}$$

conditional probabilities per genera

The same steps are taken to calculate the expected likelihood of words for each genus in the training set. Let $m(w_i)$ represent the number of sequences in each genus *G* that contain the word w_i, where the total number of sequences in the genus is *M*. The conditional probability $Pr(w_i|G)$ is then

$$Pr(w_i|G) = \frac{m(w_i) + P_i}{M + 1}$$

The value of P_i is used because if $m(w_i)$ happens to be zero, it can be expected that the value would fall back to the likelihood of the word occurring in the feature space, based on the training set.

joint probability

Finally, the probability that a sequence *S* is a member of the genus *G* is found by looking up the frequencies of the words v_i that compose *S* in the feature space dictionary and multiplying them together, for each genus.

$$Pr(S|G) = \prod_{i=1}^{n} Pr(v_i|G)$$

$Pr(S|G)$ is then substituted into the Naive Bayes classifier.

$$Pr(G|S) = \underset{(g_j \in \text{possible genera})}{\text{argmax}} \left(\prod_{i=1}^{n} Pr(v_i|g_j) \right)$$

Example

Suppose there are 2 genera G, each having 2 species S. Each species has a highly improbable sequence of 9 nucleotides, of which all but one is adenosine (A).

G1:S1 -> AAAAAAAAA

G1:S2 -> AAAAAAAAT

G2:S3 -> AAAAAAAAC

G2:S4 -> AAAAAAAAG

This list of sequences is the training corpus, and represents the database that query sequences will be compared against. The following steps are shown in Table 5. After generating the feature space, the number of times the words in the training corpus occurred in the feature space was counted (column 2). The frequency P_i is shown in column 3. Next, the number of times the training corpus words occur in each genus is counted (columns 4 & 5), and $Pr(w_i|G)$ is calculated (columns 6&7).

words in feature space	counts in corpus	P_i	G1 counts	G2 counts	G1 $Pr(w_i\|T)$	G2 $Pr(w_i\|T)$
aaaaaaaa	5	1.1	3	2	1.367	1.033
aaaaaaat	1	0.3	1	0	0.433	0.100
aaaaaaac	1	0.3	0	1	0.100	0.433
aaaaaaag	1	0.3	0	1	0.100	0.433
all others	0	.1	0	0	.033	.033

Table 5: Construction of the Naive Bayes lookup table. First column: each word of the feature space. Second column: the number of times the particular word from the feature space occurs in the corpus. Third column: the frequency of each feature space word. Fourth and Fifth columns: the number of times the feature space word occurs in each genera. Sixth and seventh columns: the frequency of the feature space words occurring in each genera.

Now, given the unknown bacterial sequence AAAAAAAT, what is $Pr(S|G)$? After generating the words from the query sequence, the frequency that each word occurs in each genera is looked up and multiplied together, as shown in Table 6.

query words	G1 $Pr(S\|T)$	G2 $Pr(S\|T)$
aaaaaaaa	1.367	1.033
aaaaaaat	0.433	0.100
final probability	0.592	0.103

Table 6: Calculating the probability that a query sequence belongs to either genera. The frequency that each word from the query sequence occurs in each genera is looked up, and all values from that genera are multiplied together. The genus with the highest value is the most likely source of the query sequence.

The highest posterior probability that the query sequence is from Genus1 is .592, while Genus2 is .103. Therefore, the query sequence is most likely from Genus1.

Confidence level

During the training step above, each sequence had a full complement of 8-mer words generated. The confidence of the assignment can be done by bootstrapping the independent features of a data set, in a similar way to BLCA. For a sequence, the independent features of the dataset are the non-overlapping 8mer "words". So, a subset of $\frac{1}{8}$th of the total "words" are randomly chosen. In Figure 15, an example of non-overlapping words are the black bars under the sequence.

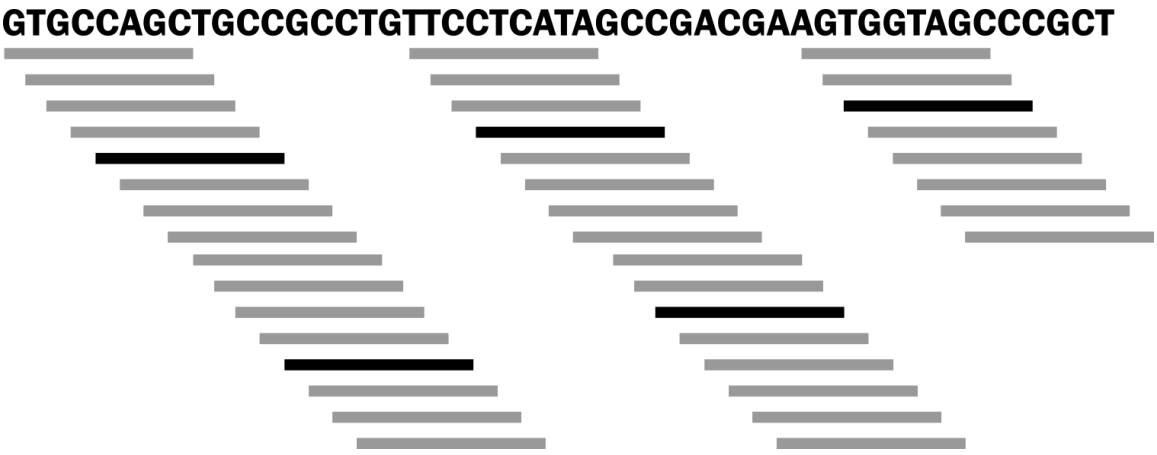

Figure 15: Example of how the confidence score is calculated by Naive Bayes. In a similar way to BLCA, non-overlapping 8mer segments (the black bars) are sampled from the query sequence and used to calculate the posterior probability from the lookup table. This constitutes one bootstrap iteration. After many iterations, the confidence score is the number of times the original assigned match has occurred.

Exact matching

Exact Matching is an attempt to compensate for the known ambiguity of using the 16S rRNA gene sequence as a taxonomic marker gene(27,57). Exact Matching is a deterministic classification method of assigning taxonomy done by searching the reference database for a sequence that is an exact match to the query. A common processing step done after sequencing but before classification is to cluster the reads into Operational Taxonomic Units (OTU) as a means to overcome sequencing errors and gene sequence variation. Classification is then done on the consensus sequences of the OTUs. OTUs have several limitations that reduce their utility, the largest being that all sequences in one cluster are delimited by a consensus distance metric(58,59). A recent advancement is to

remove sequencing errors by statistical methods based on the premise that true sequence diversity will be observed more than the errors introduced by the sequencing process, and yields Amplicon Sequence Variants (ASV)(60). As both of these methods attempt to accommodate rather than reduce the diversity found in an environmental DNA sample, finding an exact match in a reference database is limited by the difficulty of defining a species with standing in nomenclature.

2.5.3 When data is missing or removed

Missing data is a common problem. The obvious impact of missing data is that there are less data available for analysis, and possibly so much data is missing that analysis is precluded entirely. A more insidious impact of missing data is that the results of analysis are biased. In this study, missing data is defined as when a species in the Thomas-White dataset has no representative sequence in the query or reference records of a classification scheme. For the query records, an example would be not all query records have sequences for all proposed marker genes. For the reference records, an example would be that the database does not include one or more of the species in the Thomas-White dataset. In addition, the requirements for performing the Exact Matching classification necessitated that query records be removed, and these removed records are considered missing data as well.

Missing data can be ignored if specific remedies to compensate for the missing data are not needed, either because the missing data will not bias the results or due to the design of the analysis(61). For example, while a random sample of a finite population does not include a large portion of that population, the missing data can be ignored because the process of random sampling is designed to account for the missing data. The design of this study does not compensate for missing data, but small amounts of missing data will not bias the results. Furthermore, it is possible to verify if all the species in the Thomas-White dataset are present in the query and reference records. Therefore, a procedure was used to determine the extent of missing data in the query and reference components of

each classification scheme, and if that classification scheme could be used based on that determination.

The flowchart in Figure 16 shows the steps in determining how a classification scheme is assessed for missing data. Based on recommendation(61), if the query or reference records in a classification scheme were missing more than 10% of the species in the Thomas-White dataset they were removed from the study. This equates to 78 x .1 = 8 or more species, rounding up. Otherwise, the missing data were either ignored and the classification scheme was used as normal, or an attempt was made to replace the missing data.

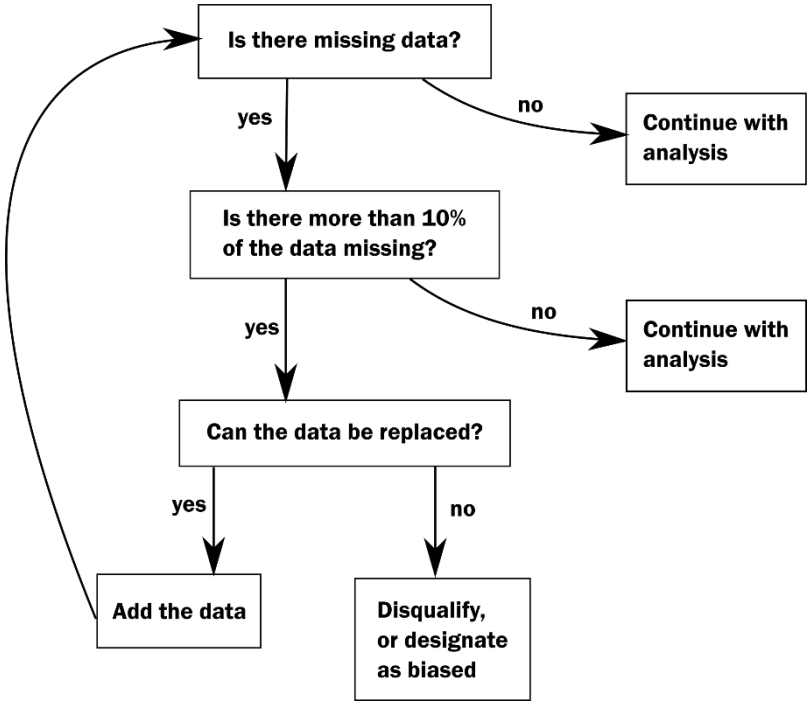

Figure 16: Flowchart for handing missing data. If more than 10% of the data were missing, then the data is either replaced if possible, designated as biased, or disqualified from further analysis.

A simple example of how this procedure was used is illustrated by removal of the RDP database from this study. This database does not include any species-level classification, and was excluded from this study. A more complex example is the approach taken with the NCBI representative prokaryotic database. This database was missing 15% of the species in the Thomas-White dataset, but it was possible to add enough records to fill out all but one of the missing species. The

largest amount of work involved the classification schemes that used Exact Matching. This classification scheme required that the query sequences did not include any ambiguous nucleotides such as N (representing A,T,C or G) or W (representing T or A). The query records in this classification scheme were the set of sequences that would be amplified during PCR for each primer set (the targeted amplicon), and some of those sequences contained ambiguous nucleotides. If more than 7 sequences were removed, that targeted amplicon was labeled as biased. For this reason, only the V3 and V6 regions of the 16S gene were designated as unbiased results.

It needs to be emphasized that the point of this study was not to create fully functioning classification schemes from all possible resources, but to assess if available resources allow species level identification. The Greengenes database lacks about 75% of the species in the Thomas-White dataset, but was still included in this study due to its continued use in classification schemes and possessing some representation of species-level taxonomy.

2.6 Evaluation

The results of how a classification scheme performs Record Linkage will vary depending on the particular combination of the scheme components, and some combinations will be more effective than others. Quantifying how much better one classification scheme is than another is the role of evaluation. While it is easy to recognize that the choice of identifier to use in a classification scheme is important, it must be reiterated that evaluation involves all three components – the identifier, the classifier, and the database.

2.6.1 The confusion matrix

The Naive Bayes and BLCA classifiers are designed to designate one and only one reference record as a match to the query record, and by omission the remaining record pairs are designated as non-matches. As a result there are two linkage errors that can occur. A false match occurs when a record pair is designated a match when in fact they are not, while a missed match occurs when a record pair is designated as a non-match when in fact they are(62). Tabulating the correct

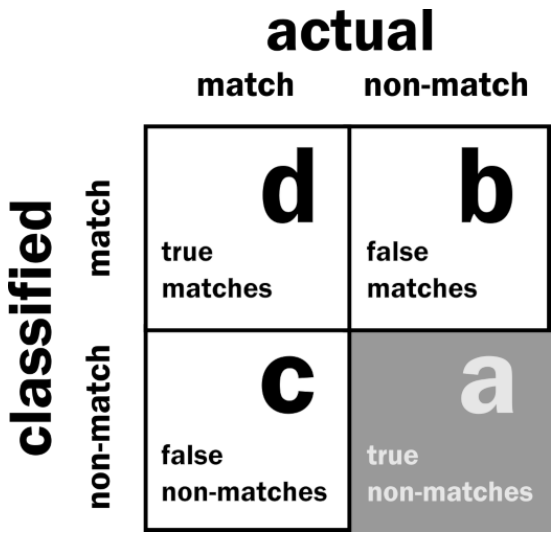

Figure 17: The confusion matrix. Evaluating the results of Record Linkage only use true matches, false matches and false non-matches (missed matches).

designations and errors yields the confusion matrix, shown in Figure 17. All evaluation measurements use proportions of the cells in this matrix to assess the effectiveness of the classification scheme.

Linkage errors can be identified and mitigated by using a known data set as a test case(63). The characteristics of matched and unmatched record pairs can be compared, the linkage errors can be quantified, and changes noted when the classification procedure is modified. Test cases for record linkage can be difficult to obtain, especially in the context of the human bladder. The Thomas-White dataset is a fortunate development in this respect, because it is an initial census of the bacteria in this environment.

Example

Suppose there is a classification scheme composed of a probabilistic classifier, a set of query sequences $\{E, F, G\}$ and the set of sequences $\{E, F, L, M\}$ held in a reference

database. In this example, the number of reference records is greater than the query records, and the reference is missing a corresponding G record in the query set.

If the query and reference record letters are the same, then they are designated as a match. Starting with the query sequence E, all possible pairwise comparisons between the records in the reference database with the query are performed, which will be written as $\{E_q: E_r, E_q: F_r, E_q: L_r, E_q: M_r\}$. We can summarize the outcomes of these comparisons as $\{$*match, non-match, non-match, non-match, non-match*$\}$. After all query records are compared to the reference records, the outcomes can be written in a matrix defined as the classification space (Figure 18). The size of this classification space is the number of all pairwise comparisons of the query and reference records, or 3x4=12. As we know all the outcomes in the classification space, this matrix is the test case.

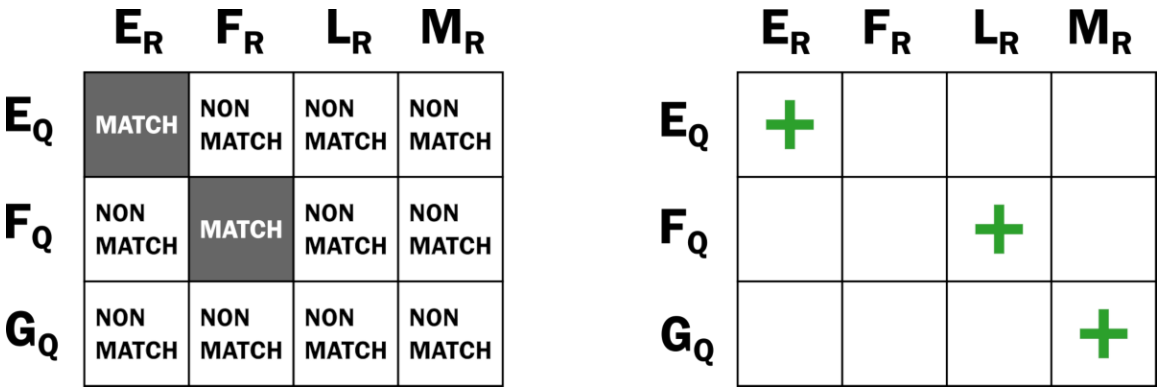

Figure 18: The classification space for the known test set (left) and the results of a classifier (right). Classification results are a green cross for matches and blank cells for non-matches.

Next, suppose we allow the classifier to assign record pairs as matches or non-matches for this classification space, represented as green plus signs for matches and blank cells as non-matches. Some results are correct, and some are not.

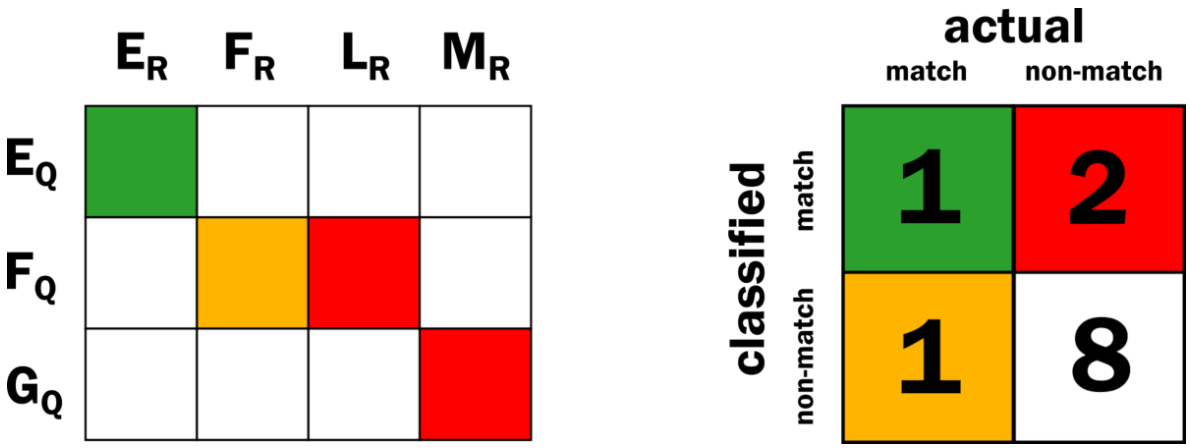

Figure 19: Evaluation of the classifier results (left). Green cells are true matches, red cells are false matches, yellow cells are missed matches, and white cells are true non-matches. The confusion matrix (right) is shown with tabulated results for the classification space on the left.

The correct and incorrect designations are tabulated into the confusion matrix and evaluation can begin (Figure 19). There are a few characteristics of this example that need to be pointed out that result from this test case. The first is that despite the lack of a matching record in the reference database, the classifier still designated the $G_q : M_r$ pair as a match. Both the Naive Bayes and BLCA classifiers choose the record pair with the highest posterior probability to designate as a match, even if the magnitude of the probability is very small. A consequence of the way these classifiers designate matches is that the number of false matches will always be higher when using a database that is lacking one or more of the query species. Therefore, it is important to verify that the reference database has a sufficient number of corresponding sequences found in the test case, before incorporating that reference in a classification scheme with actual data. In the absence of a test case, using a database that reflects the current state of bacterial taxonomy is the only way to mitigate linkage errors. The second characteristic is that the number of true non-matches is much larger than any other cell in the confusion matrix, and using evaluation measurements that incorporate true non-matches in the context of record linkage leads to problems that will be described in detail below.

2.6.2 Evaluation measurements

There are many kinds of measurements available to evaluate the results held in a confusion matrix, and all basic measures are the proportion of one cell to the sum of the cell's row or column. For this project, it is desired that the classifier returns the most true matches possible, while minimizing the number of missed matches and false matches. The definitions in this section use the lettering scheme in Figure 17.

The first evaluation measurement is Recall, which is the number of matches that the classifier got right (true matches, or cell D), out of all the matches that were listed in the test case (true matches, or cell D + missed matches, or cell C). It is defined as

$$\text{Recall} = \frac{d}{d + c}$$

The next is evaluation measurement Precision, which is the number of matches the classifier got right (true matches, or cell D), out of all the record pairs that the classifier thought were matches (true matches, or cell D + false matches, or cell B). It is defined as

$$\text{Precision} = \frac{d}{d + b}$$

In this study, the values of Precision and Recall are used by themselves and as an average, described in further detail below.

Another measure, called the False Positive Rate (FPR), shows the proportion of the number of matches the classifier got wrong (false matches, or cell B), out of all the record pairs that were not matches (false matches, or cell B + true non-matches, or cell A). In a similar way as Recall and Precision, the False Positive Rate is defined as

$$\text{False Positive Rate} = \frac{b}{b + a}$$

Finally, a popular measure is the Accuracy, which shows the proportion of the number of correctly assigned record pair comparisons to all comparisons. It is defined as

$$\text{Accuracy} = \frac{d + a}{a + b + c + d}$$

Recall, Precision and the FPR are similarly constructed fractions, and behave in the same way. Figure 20 shows this behavior for Recall, but is accurate for Precision and the FPR. For descriptive purposes we will focus on Recall. In this graph, the value of true matches is held constant, and the number of missed matches ranges over many values. The vertical line is where the number of missed matches is equal to the number of true matches. As the number of missed matches exceeds the number of true matches, the value for Recall asymptotically approaches zero. As the number of missed matches becomes smaller than the number of true matches and finally reaches a value of zero (an ideal case), the value for Recall approaches 1. When the number of missed matches is less than the number of true matches, small reductions of missed matches results in large gains in Recall. In contrast, when the number of missed matches is greater than the number of true matches, small reductions only result in small gains in Recall. This is also true for Precision and the FPR.

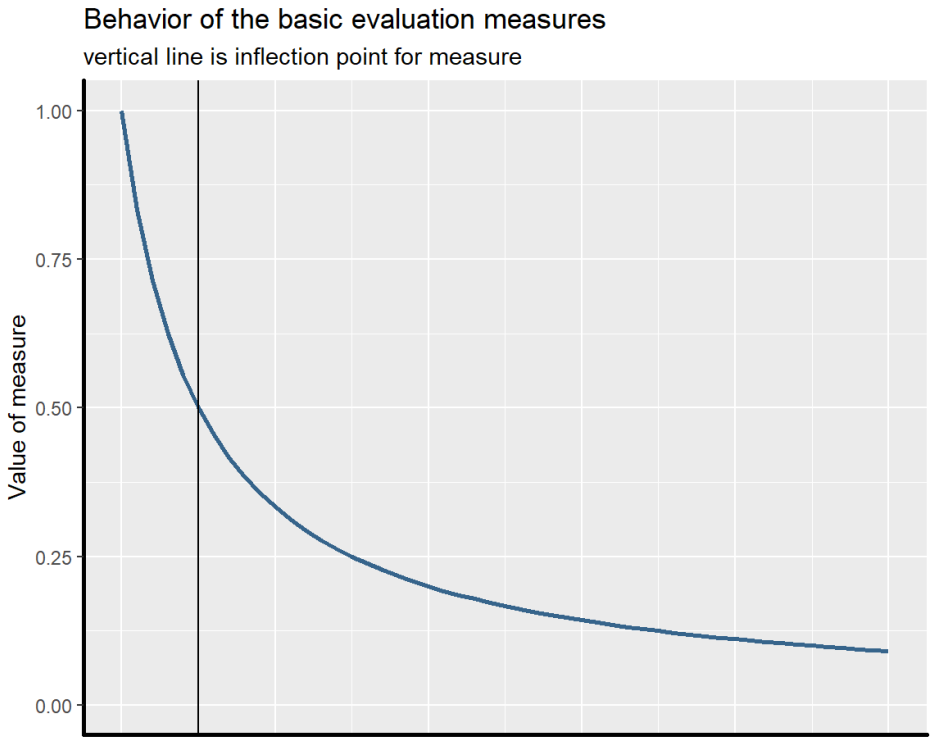

Figure 20: General behavior of Recall, Precision and the False Positive Rate (FPR). Description will be of Recall. Y-axis is the value of Recall, while the x-axis maps the increasing value of Missed Matches. The vertical line is when the value of Missed Matches equals the number of True Matches. This description holds for the remaing performance measures as well. For Precision, the vertical line is when the number of False Matches equals the number of True Matches, and the x-axis is increasing False Matches. For the FPR, the vertical line is where the number of true non-matches is equal to the number of false matches, and the x-axis is increasing True Non-matches.

In sum, the most desirable classifier will have the highest value for Recall and Precision. Comparing two or more classification schemes by graphing these two measures is difficult, and a third measure defined as the harmonic mean between Recall and Precision is used, called *F*.

$$F = \frac{2PR}{P+R}$$

There are several ways to write this measure, depending on how it is calculated, but here it will be written as the F-measure.

The Precision, Recall, FPR, and F-measure for the example in Figure 19 are:

- Recall = 1/2
- Precision = 1/3
- F-measure = 2/5
- FPR = 1/5

- Accuracy = 3/4

2.6.3 When the classification space is large

The number of records in a taxonomic database is easily in the hundreds of thousands. How does this affect a classification scheme? An estimate of the classification space that occurs in this study is roughly 16 million. Real world sequencing data from the Illumina platform can be in the tens of thousands, resulting in a classification space of roughly 2 billion. Waiting for a computer to complete this number of comparisons is prohibitive.

The solution to the first problem is to do some kind of low level comparison that will sift through all the records of the reference database and return a smaller (but still sizable) subset of the entire database, called *blocking*. Both classifiers in this study use blocking, but employ different methods.

Example

For the example shown in Figure 21, the letters for reference and query records are omitted. Gray boxes indicate actual matches in the test case, while white boxes indicate non-matches. There are more records in the reference database than there are queries, and not all queries have a corresponding match in the reference. The classification space on the right shows which record pairs are considered a match by the classifier, and the size

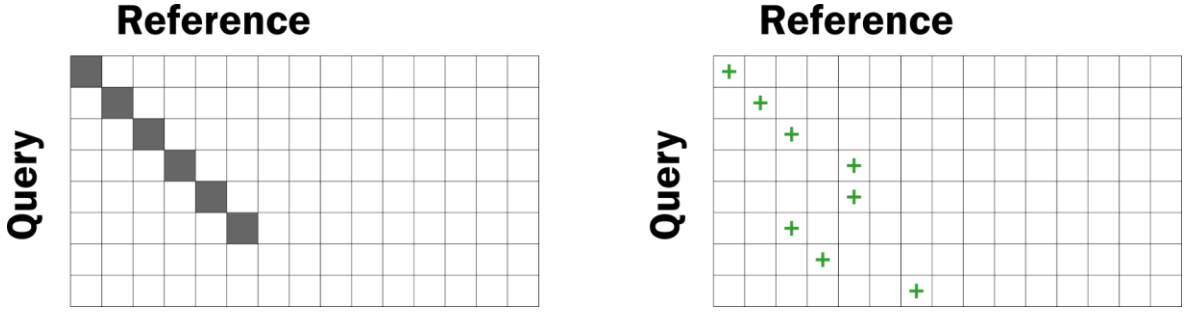

Figure 21: (left) Classification space of the test case for when the reference is larger than the query set. Gray cells indicate a match while white cells indicate non-matches. The results of a classifier for this test case is shown on the right.

of the classification space is 120.

Evaluating the classification scheme is shown in Figure 22. There are two record pairs that are missed matches, four record pairs that are false matches, four record pairs that are true matches, and the remaining 110 record pairs are true non-matches.

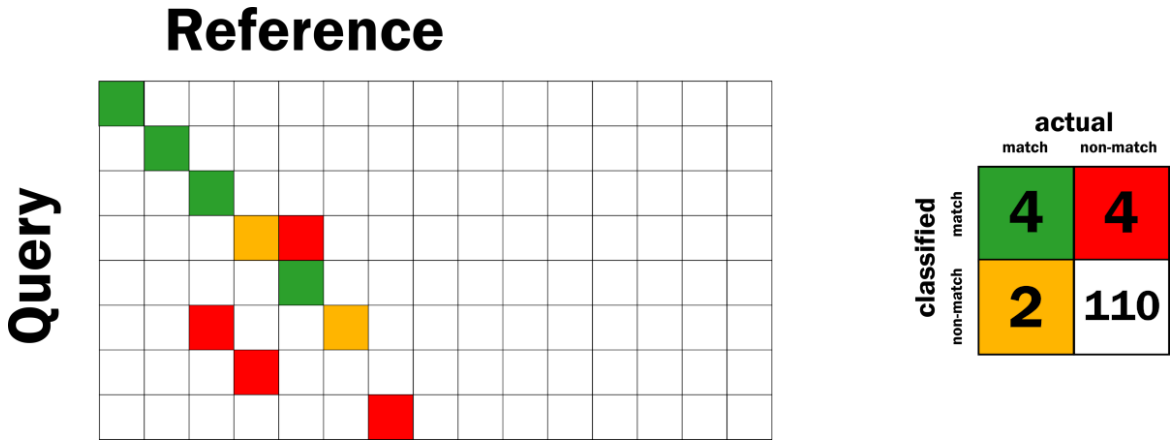

Figure 22: Evaluation of the classification results from Figure 21. Four of the record pairs that were called a match were correct and evaluate as True Matches (green boxes), and four of the record pairs that were called a match were incorrect and evaluate as False Matches (red boxes). The two correct record:reference pairs that were not called a match are evaluated as Missed Matches (yellow boxes). The remaining white boxes are True Non-matches.The tabulated results are shown in the confusion matrix on the right.

The evaluation measures are:

- Recall = 4/6 = 2/3
- Precision = 4/8 = 1/2
- F-measure = 4/7
- FPR = 4/114 = 2/57
- Accuracy = 114/120 = 57/60

It is clear that most of the classification space is true non-matches, and blocking would reduce the number of computations needed without changing the outcome of the evaluation. One result of blocking is shown in Figure 23.

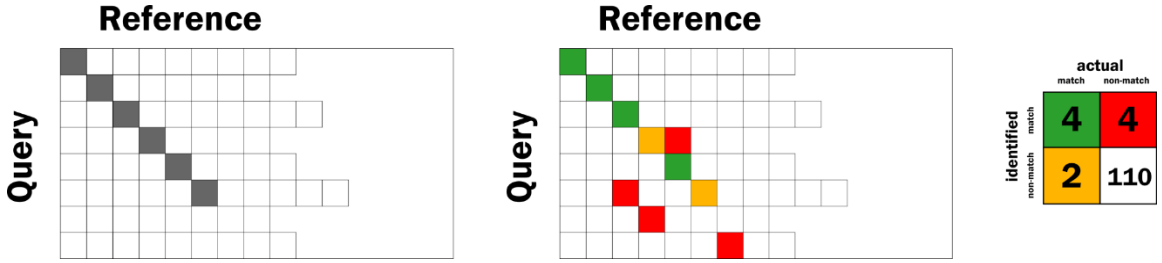

Figure 23: Blocking example for the classification space shown in Figure 21. When the classification space is large, a way to reduce the computational time is to perform a low level, fast comparison that disregards as many obvious True Non-matches as possible. The remaining record pairs in the classification space can then be worked on by the classifier.

The ragged edge of the rows is to show that the blocking method will return a different number of records based on similarity to the query. This classification space is still

considered to be 120, but now only 72 record pair comparisons need to be done by the classifier.

However, blocking assumes that all the records that are not selected to go on to the classification step are true non-matches, and this assumption can be wrong. The case where the blocking step returns zero reference records is another assumption that no reference records will match the query. BLCA uses the Blast+ suite of tools to perform blocking, and in the event where all BLAST scores are below a threshold, the query is assigned as 'unclassified'. The Naive Bayes classifier reduces the classification space during the calculation of the conditional probability that a word occurs in a taxon.

Note what is happening to the Accuracy. This classification scheme managed to make an equal number of correct and incorrect assignments, and this is reflected in the lousy value for Precision. Yet the accuracy is still 57/60 = ~92%, due to the large number of true non-matches. When working with Record Linkage, Accuracy is one evaluation measurement that is not informative.

2.6.4 Confidence score

The mammal vs fish classification example in section 2.4.2 describes how the classifier will still classify a query even though the posterior probability that the query came from that taxon is very low. How do Naive Bayes and BLCA account for this situation? They include a step where the classification is checked, and produce a value that reflects the "goodness of fit" of that classification. This is the confidence score, and through an unfortunate overloading of terminology, it has nothing to do with a confidence *interval*. The confidence value has more in common with a power test, or leave one out validation. As explained above, the pairwise alignments are randomly permutated and classification assigned again through a bootstrap process. The method measures how much the query classification changes through random permutation. If the classification outcome was due to a record pair that are highly similar, the confidence score will be high. Otherwise, the confidence score will be low. The classification space in Figure 25 shows the previous setup with the addition of the confidence scores for each classification result.

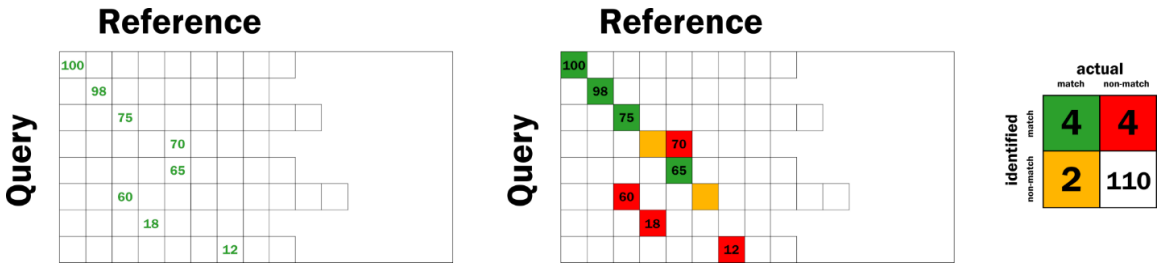

Figure 24: Inclusion of confidence scores. The classification outcomes are the same as the previous figures, but each match is assigned a confidence score. These classifications are evaluated as before, but now the classifications have an associated bootstrap value where the same outcome was observed after random sampling

There is a property of the confidence score that needs to be highlighted. One of the red squares has a confidence score of 70, while one of the green squares has a confidence score of 65. The confidence score reflects how similar the query record is to the reference records, and in general a high confidence score does correspond with a true match.

However, there are cases where the confidence score does not correspond to a correct classification. One case is when the reference records are neither highly similar or highly dissimilar to the query record, but somewhere in the middle. In this situation, it is possible for the bootstrap process to generate moderately high confidence scores for record pairs that are a false match. Another example is when the database has no corresponding record for the query, but the identifier is still highly similar. The point of using variable regions as record identifiers is that the accumulated variation is as different as possible between species, but even variable regions can share a respectively high amount of similarity. A good example is the nearly identical 16S rRNA gene sequence found in *B. psychrophilus* and *B. globisporus*.

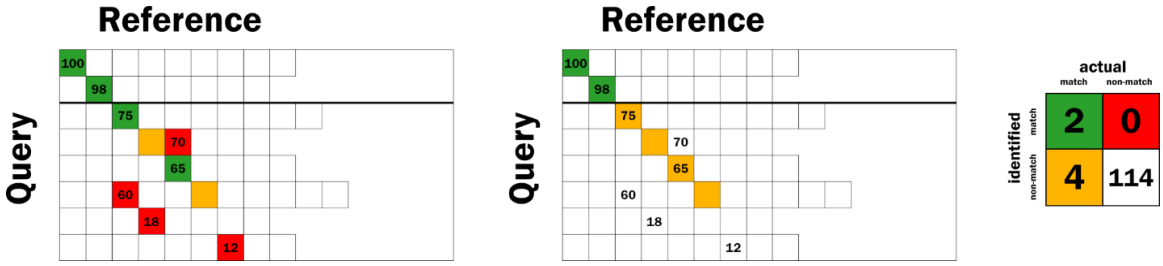

Figure 25: The addition of a confidence score as a threshold for the classification space shown in Figure 24. Confidence scores of each classification are shown as numbers, and the horizonal line is a confidence score of 80%.

The heavy black line in Figure 25 is the default cutoff that Naive Bayes classifiers like the RDP classifier, DADA2 and Qiime set for their confidence scores (although it can be easily changed). If all the record pairs below that cutoff are rejected, how does that affect the evaluation? In effect, we are saying that the classifier is only allowed to designate a match for the first two rows of record pairs. So, all classifications above the cutoff remain the same. All true matches below the cutoff are now assigned as missed matches. All false matches below the cutoff are now assigned as true non-matches as the result of a kind of double negative operation. The final outcomes are shown on the right hand side of the picture above, and the evaluation measures are now compared to the previous values in Table 7.

	80%	No threshold
Recall	.5	.667
Precision	1	.5
F	.667	.572
FPR	0	.035

Table 7: Evaluation measure values for the classification space shown in Figure 25

Recall has gotten worse, Precision is absolutely fabulous, and the F measure is worse (naturally). These conditions have also created a situation in which there are no false positives, and therefore the false positive rate is zero. The largest disappointment is that the number of actually correct classifications is cut in half. Reduction in false matches is to be pursued, as long as a loss in useable data can be avoided. This is still applicable to a sequencing run where the number of useable sequencing reads that have confidence scores above the default threshold numbers in the thousands, because the reduction in true matches does not have any reassurance that the remaining data necessarily has a higher chance of a correct assignments, and a smaller dataset may bias the results.

However, the confidence score is still useful. Low confidence scores still indicate that the classification scheme is making the best of a miserable choice of outcomes, but how to find out which score is best? After visually inspecting the evaluation of the classification space, there is a setting that minimizes the false matches and maximizes the number of true matches, somewhere between 60 and 65, shown in Figure 26.

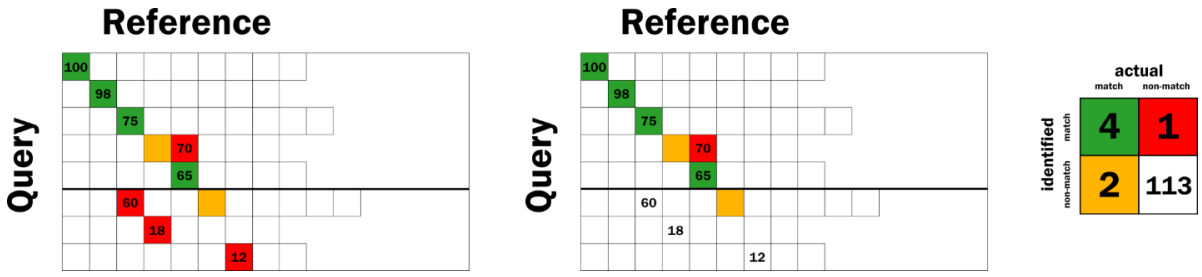

Figure 26: Optimal confidence score that minimized the number of false matches and missed matches. Horizontal line is a confidence score of 63%

The calculated evaluation measures are now shown in Table 8. Precision and the F-measure are now as high as they can be, and the FPR is very low.

	80%	63%	No threshold
Recall	.5	.667	.667
Precision	1	.8	.5
F	.667	.727	.572
FPR	0	.009	.035

Table 8: Evaluation measure values for confidence scores of 0, 63% and 80% used as thresholds.

2.6.5 ROC curves and record linkage

For large datasets, visual inspection of the classification space is not practical, but it can be done through code. Evaluating the classification outcomes at all confidence scores can be computed and the evaluation measurements graphed.

Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1

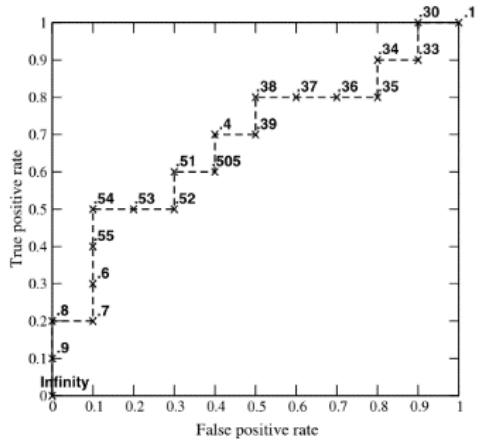

Figure 27: ROC curve and associated data from Fawcett 2006. The results of a binary classifier for each data point are listed in the "score" column. The known class for each data point is listed in the "class" column as either "n" or "p". Incrementing a threshold value places the data points in either a "n" class or "p" class depending if the score is less than or greater than the threshold value, e.g. a threshold of .2 places data point #20 in the "n" class and data points 1-19 in the "p" class. A confusion matrix is constructed for the results of each threshold value, and the true positive rate (y-axis) is plotted against the false positive rate (x-axis).

One such graph is a Receiver Operator Characteristic (ROC) graph shown in Figure 27, which plot the values of Recall vs the FPR. This graph is from Fawcett 2006(64), and shows the classification results of a classifier using a predictive model. The data are arranged by Score from low to high, which indicates the probability that the data point is of class P or class N. Starting at zero, a threshold is incremented from low to high, and those data points above the threshold are classified as class P, otherwise as class N. It is important to note that the classification is done after the threshold is set. At each step the Recall and FPR is calculated and graphed. The optimal threshold to use is the one where the value of Recall is the highest and the FPR is the lowest, which maps to the upper left corner.

However, an ROC curve is not suitable for evaluating Record Linkage. In the examples above, the FPR is very small due to the large number of true non-matches in the denominator of the FPR equation. Large values of the true non-matches push the FPR values far to the left of the ROC graph. Like accuracy, the ROC curve is an uninformative evaluation measure for record linkage.

2.6.6 Using the F-measure for best performance

The best evaluation measure for record linkage is the F measure, as it only depends on the values of Recall and Precision. Because it's the harmonic mean between the two, if a classification scheme has a high value of both Recall and Precision, it is reflected in the high value of the F measure. As shown in the graph below, the F-measure is a good indicator of the performance of a classification scheme.

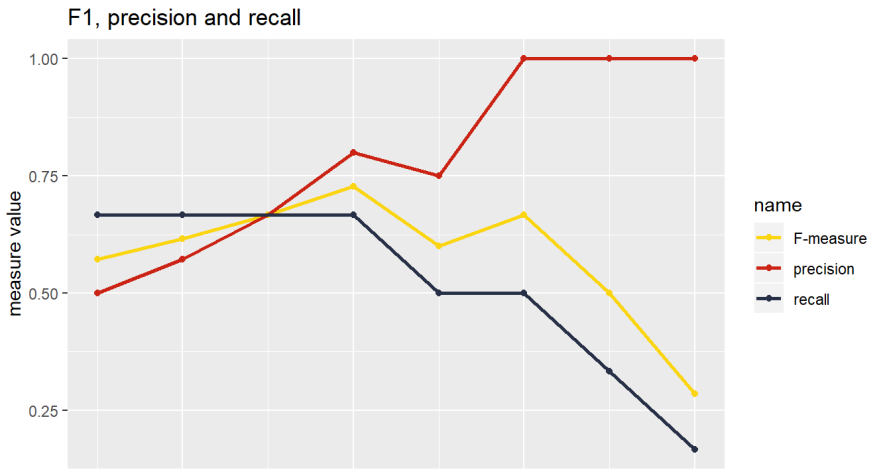

Figure 28: Values of the F-measure, precision and recall using data from Figure 26. As the F-measure is the harmonic mean between precision and recall, the values will not always plot exactly between the two.

Better still, different classification schemes can be compared using the F-measure. One way to determine the confidence score that yields the best performance is to graph the confidence score against the calculated F-measure, like in Figure 29.

In this example the highest F-measure is achieved when a confidence score of 63% is used as a threshold, just as when found by visual inspection. However, this method only works when the outcomes of the classification space are known beforehand, as with a test case. For a real world application, this could be accomplished by including a mock community in the sequencing run. Since the identity of each bacteria in the mock community is known, the optimal confidence score for that classification scheme can be found and applied to the actual sequencing data.

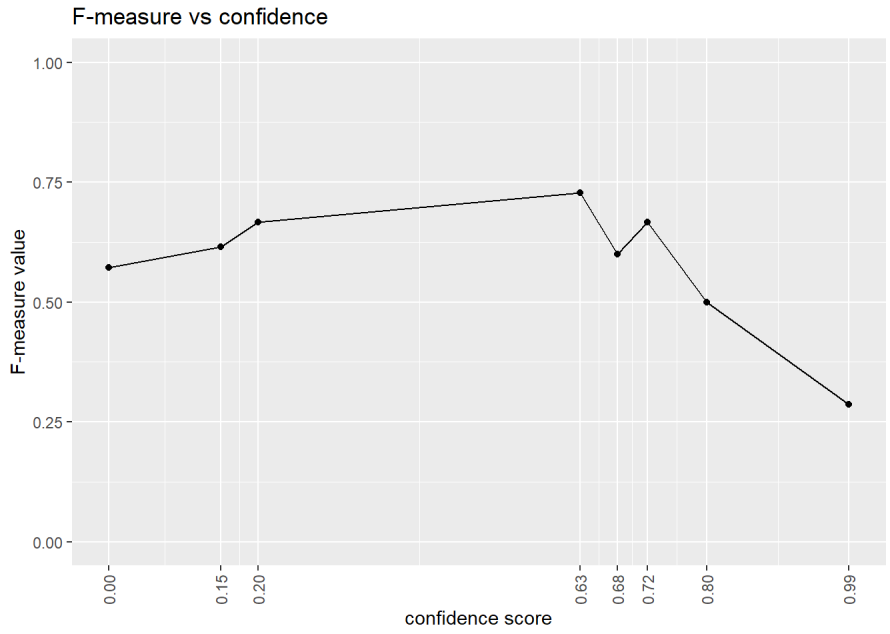

Figure 29: F-measure vs confidence score using the data from Figure 26. The maximum value of the F-measure is at a confidence score of 63% is used as a threshold, which agrees with the visual inspection of the data.

A note about the F-measure

For the data in this project, the goal is to show which classification scheme has the most true matches, lowest number of false matches, and lowest number of missed matches. The F-measure is the best way to incorporate these requirements, and be able to compare all the schemes with each other. However, the F-measure remains a measurement of proportions, and high ranking F-measure values can still be derived from record counts that are otherwise useless. Consider these two examples. Two classification schemes are used to assign taxonomy to a data set of unknown sequences. Classification scheme A yields 63 true matches, 14 false matches, and 7 False Non-matches. The F-measure is then .857, which indicates a reasonably good performance.

$d = 63, b = 14, c = 7$

$$\text{F-measure} = \frac{2d}{2d + c + b}$$

$$= \frac{(2 \times 63)}{(2 \times 63) + 7 + 14}$$

$$= \frac{126}{141} = .857$$

On the other hand, Classification scheme B yields 9 true matches, 2 false matches, and one False Non-match. Compared to the first classification scheme, 9 sequences with accurately assigned taxonomy is almost not worth the trouble to use, but the F-measure values of the two schemes are the same.

$d = 9, b = 2, c = 1$

$$\text{F-measure} = \frac{2d}{2d + c + b}$$

$$= \frac{(2 \times 9)}{(2 \times 9) + 1 + 2}$$

$$= \frac{18}{21} = .857$$

2.6.7 Ranking classification schemes

The way to show which classification schemes have the best performance, and at the same time the greatest number of true matches, is to graph the F-measure vs the Recall of each classification scheme. Since I'm only interested in when the classification scheme is performing at its best, I'm going to plot the number of true matches where the confidence score yields the highest F-measure value. The result is a scatterplot, where the better classifications schemes will be close to the upper right corner. This is shown in Figure 30. The classification scheme represented by the red dot is considered to be better than the classification scheme represented by the black dot, because the performance and the number of true matches is larger and the distance to the upper right corner is shorter. In the event of a tie, the largest number of true matches followed by the highest F-measure decides the winner.

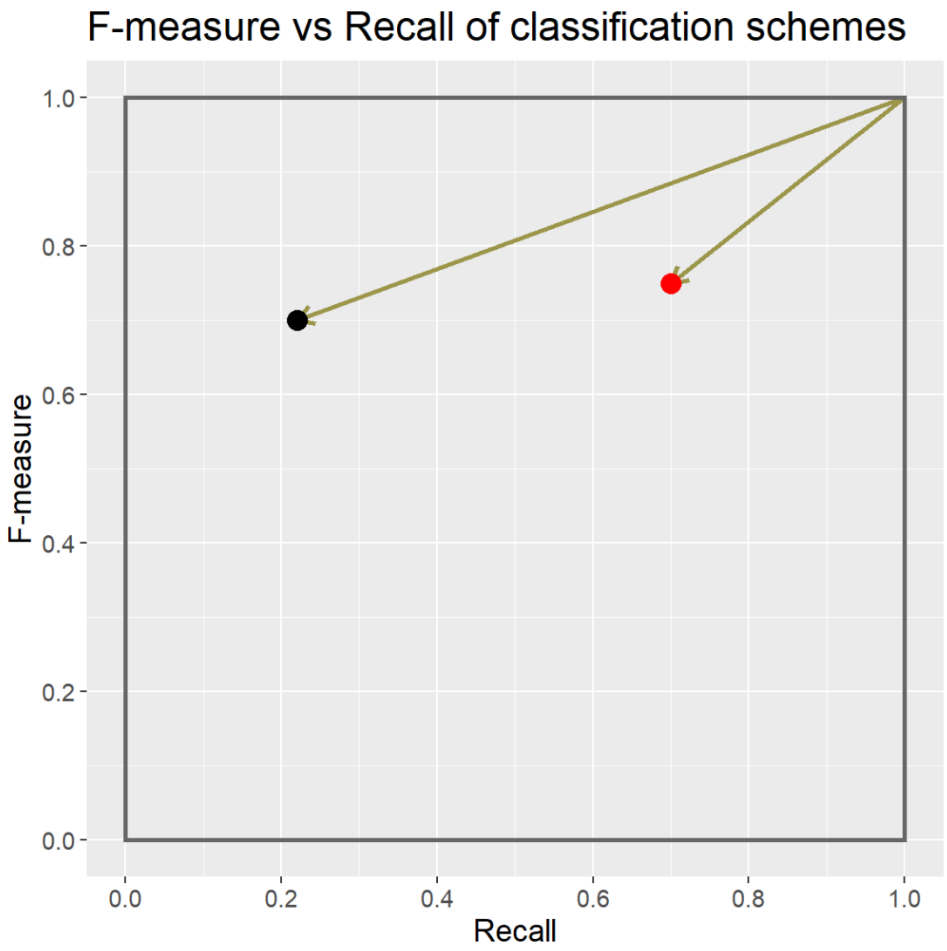

Figure 30: Graphing method to compare the classification schemes used in this study. F-measure vs Recall for all classification schemes are graphed, and those that plot closest to the upper right corner (red point) outperform classification schemes that plot further away (black point).

3. Methods

3.1 Code resources

All scripts that were written for this project can be found in the GitHub repository https://github.com/lakarstens/-Hoffman_MS_Thesis_Bladder_Species_2020

3.2 Data

3.2.1 The Thomas-White dataset

Culturing and identification

As described in their paper, Thomas-White et al obtained urine samples from 77 female subjects with and without symptoms of overactive bladder. Using Expanded Quantitative Urine Culture (EQUC)(44) Thomas-White was able to expose collected urine from symptomatic and non-symptomatic participants to a wide range of food sources, culturing environments and incubation time. This method was in stark contrast to the standard clinical method of urine culture which was designed for rapid turnaround and favored pathogenic bacteria. Once the genomic sequencing of the isolates were completed, Thomas-White used a novel technique for identifying the bacteria based on the sequences of the protein encoding marker genes.

This novel technique is described in Mende 2011(65). To begin, the authors collected as many complete bacterial genomic sequences as were available at the time, roughly about 3000 genomes. For each genome, the 40 protein encoding marker genes were located and compiled. The 16S rRNA gene for each bacterial species was also used for this study, but those sequences were obtained from the SILVA database as described above.. For each marker gene, pairwise global alignments between all the possible pairs of species were done and the percent dissimilarity was calculated. Next, each pairwise dissimilarity value was weighted by gene length, and averaged across all marker genes. This is analogous to generating a dissimilarity matrix for each marker gene, each cell representing one

permutation of all genome alignments, then stacking the 40 matrices on top of each other and averaging each cell. The final supermatrix is all the possible pairs of genomes, and each cell holds the weighted average of all 40 marker genes. When this supermatrix was done, they used average cluster distance (UPGMA) to cluster cells together until all were merged, and then compared the resulting hierarchical cluster to the known Tree of Life. Any discrepancies were verified by referring to published literature. Finally, the authors created the web based tool specI that will search submitted sequences against their genomic database. If there is at least a 97% identity between the submitted sequence and one of the sequences kept in the specI database, a match is declared and the submission is labeled as the corresponding species.

After culturing clonal populations of the bacteria from their urine samples, Thomas-White et al assembled each genome de novo and annotated them to identify the 40 protein encoding gene sequences. Then, for each genome, the protein encoding genes were concatenated end to end in a specific order and submitted to the online tool specI to identify the species.

V4 targeted amplicon sequencing

Targeted amplicon sequences of a set of those urine samples, using the V4 region of the 16S rRNA gene sequence, was generously provided by Dr. Alan Wolfe. The raw sequence reads were processed with DADA2 version 1.14.1 to generate amplicon sequence variants (ASVs).

Marker gene sequences

For the 16S ribosomal rRNA gene sequences, I used the type strains for each species identified by Thomas-White. The gene sequence of a type strain is the sequence of the cultured isolate that was subject to the metabolic, genotypic and phenotypic evaluations taken to define the bacterial species(21). It is the agreed bacterial organism to which the taxonomic name is referring. 16S gene sequences were downloaded from the Silva v132 release on 4/27/2019, using the “[T]” filter setting, and selecting for sequences longer than 1450nt with alignment and pintail quality scores greater than 95%. For the species that had no hits, I used

the synonym, if available. The species identified as *Corynebacterium sp* had no type strain available, and was excluded from the query set. For the protein encoding gene sequences, I used the sequence data graciously provided by Dr. Krystal Thomas-White.

3.2.2 Databases

Greengenes

The Greengenes database version 13_5 was downloaded on 9/23/19 from http://greengenes.secondgenome.com/?prefix=downloads/greengenes_database/gg_13_5/. For use with BLCA, the database was processed using the provided "1.subset_db_gg.py" script. For use with the Qiime package, the FASTA file was reformatted to work with Qiime using the custom "write_qiime_train_db.py" script, and trained to work with the Naive Bayes classifier with the provided "fit-classifier-naive-bayes" script.

Silva

The Silva database version 132 was downloaded on 9/14/19 from https://www.arb-silva.de/no_cache/download/archive/release_132/Exports/ as a fasta formatted file. The fasta file was compiled into a database that could be used with BLCA by using the "makeblastdb" utility from the Blast+ suite. The taxonomy file that was required by BLCA was generated with the custom "write_taxonomy.py" script. For use with the Qiime package, the FASTA file was reformatted to work with Qiime using the custom "write_qiime_train_db.py" script, and trained to work with the Naive Bayes classifier with the provided "fit-classifier-naive-bayes" script.

NCBI

16SMicrobial

The 16SMicrobial database is bundled with the BLCA package, but is available from ftp://ftp.ncbi.nlm.nih.gov/blast/db/. For use with BLCA, the database was processed using the provided “1.subset_db_acc.py” script. For use with the Qiime

package, a FASTA file was extracted from the bundled BLCA database using "blastdbcmd", and reformatted to work with Qiime using the custom "write_qiime_train_db.py" script, and trained to work with the Naive Bayes classifier with the provided "fit-classifier-naive-bayes" script.

Prokaryotic genomes

A custom prokaryotic database comprising full genomes was constructed. To begin, the file ref_prok_rep_genomes.XX.tar.gz (where XX are replaced with the numbers 00-05) was downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/db/. Next, the records contained in the database and Thomas-White dataset were compared, and any species present in the Thomas-White dataset that were missing from the genomic database were downloaded from NCBI. The genomic sequence for *Bacillus idriensis* was unavailable. The genomic assemblies for *Corynebacterium amycolatum, Anaerococcus octavius, Kytococcus schroeteri, Neisseria perflava, Neisseria subflava, Streptococcus oralis*, and *Cutibacterium acnes* were compiled into a new database using "makeblastdb" from the Blast+ suite. Finally, a new alias file was created with "blastdb_aliastool" in order for the missing records and the downloaded prokaryotic database to function as one virtual database that could be used by BLCA.

Presence of Thomas-White species in the databases

To verify that all species from the Thomas-White dataset were present in the databases used in this study, each database was first converted to FASTA files (if needed) using the Blast+ utility "blastdbcmd". The FASTA files were then searched with a "grep" expression for a match of the fasta header, for each species in the dataset using the custom "species_in_db.bash" script. The presence or absence of each species was recorded.

3.2.3 Duplicate sequences in the Thomas-White dataset

The sequence data of the protein encoding genes from the Thomas-White group included all 149 isolates obtained from patient samples. As there were only 79

species identified in this data set, there are many duplicate sequences. Neighbor joining trees were generated from the multisequence alignments of Ffh and RpoB by using UGENE(66). Clusters of sequences were evaluated for distance from each other and grouped into three categories. Group one were sequences of the same species that were separated by a distance of zero. Group two were sequences of the same species that were separated by no more than a distance of .03. Group three were sequences that were split across more than one cluster, and this group was further divided by the number of sequences in each cluster.

A representative species was chosen at random from groups 1 and 2. For example, *E. coli* has 4 duplicates. In Ffh, all *E. coli* duplicates were in the same cluster and separated by a distance of zero, while in RpoB the sequences were in the same cluster but separated by a distance of no more than .006. Therefore one of those duplicates was chosen at random. All species of group 3 that had one sequence in each cluster was disqualified due to the lack of any knowledge of which sequence to use. For example, *Streptococcus mitis* had 5 duplicates, and each sequence was split into 5 different clusters of unrelated species, and therefore was disqualified. Species in group 3 that had two or more sequences split into two or more clusters were chosen based on the cluster with the largest number of shared sequences between Ffh and RpoB. For example, *Aerococcus urinae* has 5 duplicates, and the largest cluster that held species in both genes contained 2 sequences, therefore one of those two sequences was chosen at random.

In total, 4 bacterial species were disqualified (*Alloscardovia omnicolens, Staphylococcus warneri, Streptococcus mitis*, and *Streptococcus salivarius*). As this was under the actionable number of missing data, this query set of protein encoding gene sequences needed no more modifications. Finally, in order to allow a fair comparison between the protein encoding genes and the 16S rRNA gene when using the custom genomic database, the sequences of those 4 species were removed from the 16S targeted amplicon query set. The evaluations between the targeted amplicons of the 16S rRNA gene sequence remained unchanged, and included all 78 species.

3.3 Variable regions

3.3.1 Multisequence alignment

The 16S gene sequences from the Thomas-White dataset and all protein encoding marker gene sequences obtained from the Thomas-White group were formed into respective multi-sequence alignments using the T-coffee program. T-coffee version 12.00.7fb08c2 was downloaded from http://tcoffee.org/Packages/Stable/Latest/ on 4/5/2019 and installed on OHSU's Advanced Computing Cluster as per instructions. Alignments were performed using the default settings.

3.3.2 Sliding window analysis

Sliding window analysis (SWA) is the method by which a list of subsequences are generated by taking successive groups of equal size, in the manner of a window of fixed length sliding across the full sequence. Variable regions of the 16S gene sequence and all protein encoding marker genes were identified through sliding window analysis using the custom "weighted_ent.py" script. The identified variable regions were used to inform which set of degenerate primers (described in the next section) would be optimal.

3.3.3 Primer design

Degenerate primers were designed for each protein encoding marker gene and the 16S gene with DegePrime. The program was cloned from the GitHub repository [https://github.com/EnvGen/DEGEPRIME.git] on 4/23/19 and installed as per instructions. DegePrime has the option to ignore columns of a multisequence alignment if the number of “-“ characters exceed a user-defined threshold. The multisequence alignments were preprocessed with this threshold set to .01, a setting that does not ignore any column. The main script of DegePrime was run using a degeneracy setting of 4096 and a window length of 18.

Broad spectrum primers were attempted for all 40 protein encoding marker genes and the 16S rRNA gene. Of the protein encoding genes, *RpoB* and *Ffh* were the only ones that fulfilled the requirement of maximum coverage and minimal degeneracy.

All annealing temperatures were calculated using OligoCalc v3.27 (http://biotools.nubic.northwestern.edu/OligoCalc.html). Min temp & max temp are calculated by nearest neighbor method, Na_{min} and Na_{max} are calculated by salt adjusted method.

3.3.4 Extracting amplicons

For each successfully designed primer set, and primer set obtained from the literature, the DNA sequence bracketed by the forward and reverse primers was extracted from the multisequence alignment. Coordinates of the MSA were identified by searching the *E. coli* sequence (EU014689.1.1541) included in the MSA for a match to the forward and reverse primer sequences, and then mapping those position to the MSA of the Thomas-White dataset. This procedure was done using the custom "extract_16s_vr.py" script and written in the FASTA format as a multirecord file.

3.4 Classification

3.4.1 BLCA

BLCA was cloned from the GitHub repository https://github.com/qunfengdong/BLCA.git and installed as per instructions. For the 16S variable regions, the program was run using default settings but pointing to the correct reference database, either Greengenes, Silva, or NCBI 16S. For the protein encoding gene variable regions, the program was run on default settings and pointing to the custom database build.

3.4.2 Qiime

Qiime2 was installed in a conda environment. The program Miniconda was downloaded (https://docs.conda.io/en/latest/miniconda.html#linux-installers) to a Linux system and installed with the command “bash Miniconda3-latest-Linux-x86_64.sh”. Then Qiime2 was installed with the commands “wget https://data.qiime2.org/distro/core/qiime2-2020.2-py36-linux-conda.yml” and “conda env create -n qiime2-2020.2 --file qiime2-2020.2-py36-linux-conda.yml”

Qiime2 was used with the Greengenes, Silva, and NCBI 16S databases and a confidence setting of 0, but otherwise default settings.

3.4.3 Exact Matching

Exact Matching was performed on the targeted amplicons of the 16S rRNA gene sequence using the DADA2 package. As Exact Matching requires a pair of identical sequences to score as a match, the 16S gene query sequences in the Thomas-White dataset needed to be screened for ambiguous nucleotides. Any amplicon dataset that exceeded 7 sequences with ambiguous nucleotides were disqualified from evaluation. At the end, all targeted amplicon datasets except v3 and v6 were disqualified.

Exact matching was performed using the addSpecies() function of the DADA2 package(54). Unlike assignSpecies (), the addSpecies() function takes the results of the function assignTaxonomy() as input. The assignTaxonomy() function is DADA2's implementation of the Naive Bayes classifier and classifies reads to the genus level. The species rank is then assigned to those results that exactly match a reference record in the provided 'silva_nr_v132_train_set.fa.gz' database.

3.5 Synonyms of species

Species names have changed in response to advances in bacterial systematics. All currently known species synonyms were downloaded from the Prokaryotic Nomenclature Up-to-Date(67) (PNU) website on 1/5/2020. PNU includes information down to the strain level, but these entries were consolidated to the species level. For example, entries like *Enterobacter cloacae* and *Enterobacter*

cloacae dissolvens are treated as synonyms of *Enterobacter cloacae*. Classification results were then checked for synonyms using the custom "validate_match_batch.py" script.

3.6 Evaluation

3.6.1 *in silico* evaluation

BLCA and Naive Bayes classification evaluation of the 16S rRNA gene targeted amplicons and protein encoding amplicons were performed using the "new_taxonomy_results_2020-3-14.Rmd" file.

Exact Matching classification evaluation of the 16S rRNA gene targeted amplicons and protein encoding amplicons were performed using the "exact_match_taxonomy_results_2020-3-16.Rmd" file.

The V4 targeted amplicon sequence reads provided by the Wolfe group was evaluated using the "real_world_data_2020-4-17.Rmd" file.

All record pairs that were assigned a match by the classification schemes were evaluated according to the following definitions:

True match - All record pairs assigned as a match that have identical genus and species labels

False match - All record pairs assigned as a match that did not have identical genus and species labels

Missed match - If a record representing a species in the Thomas-White dataset was present in the database, but was not assigned as a match, the record was evaluated as a missed match

True non-match - All records in the reference database that were not in the Thomas-White dataset

3.6.2 16S V4 targeted amplicon sequencing validation

Standard targeted amplicon sequencing using the V4 region of the 16S rRNA gene was performed on 24 of the 77 urine samples from the Thomas-White dataset. The species of bacteria in these results were identified using classification schemes composed of the V4 sequencing results as the identifier,

BLCA classifier, and the Greengenes, Silva, and NCBI 16S microbial databases. The results of the whole genome sequencing on the isolates cultured from each of the urine samples was a list of species identified as present in each sample. There is no way to correlate an ASV taken from the culture-independent method of V4 targeted amplicon sequencing to the isolates whose genomes were sequenced. Therefore, validating the *in silico* with the *in vitro* results will be restricted to the comparison of the list of species produced by WGS of the isolates from a sample, and the list of species produced from identifying the ASVs produced from targeted amplicon sequencing.

Thomas-White et al. used the full sequence of 40 protein encoding genes to identify the sequences obtained by whole genome sequencing, and this method produces more reliable identification than targeted amplicon sequencing. It can be expected that the urine samples contain at least the species identified by that method. Targeted amplicon sequencing will return a much larger list of species identified in each sample, but it can be expected that those results will include the species identified by whole genome sequencing.

For each classification scheme, validating the results will be done by enumerating the number of species identified by WGS that were identified by the V4 16S targeted amplicon sequencing. In other words, validation will be done by comparing the Recall of each classification scheme. Referring back to Figure 17, the species of bacteria identified using WGS as the identifier represents the columns of the confusion matrix (the "actual"), and the species identified using targeted amplicon sequencing as the identifier represents the rows (the "classified").

Defining how the evaluation is performed

Figure 32 shows the distribution of the species found in each sample on which V4 16S targeted amplicon sequencing was preformed, and will be referred to as the *V4 validation set*. Some species are found in more than one sample, and some are only represented once. Each *in silico* classification scheme that uses the 16S

V4 region correctly identified a set of species, and those correctly identified species that are contained in the V4 validation set will be called the *predicted matches* of that classification scheme. The cells in the confusion matrix will be filled out according to the following definitions:

True match - All predicted matches of the *in silico* classification scheme identified by V4 16S targeted amplicon sequencing

False match - All species identified by V4 16S targeted amplicon sequencing that are not predicted matches of the *in silico* classification scheme

Missed match - Predicted matches that were not identified by V4 16S targeted amplicon sequencing

True non-match - All species that were not predicted matches, and were not identified by V4 16S targeted amplicon sequencing

Figure 31 is a graphical representation of the definitions above, using the *in silico* results of the classification scheme composed of the V4 16S region, BLCA classifier and Greengenes database as an example. There are a few additional notes about this evaluation. Some species are represented in more than one sample. If any one sample in a predicted match is correctly identified by V4 16S targeted amplicon sequencing, the whole row is counted as a true match. Likewise for false matches. On the other hand, if none of the samples of a predicted match are identified by V4 16S targeted amplicon sequencing, the row is counted as a missed match, and likewise for true non-matches.

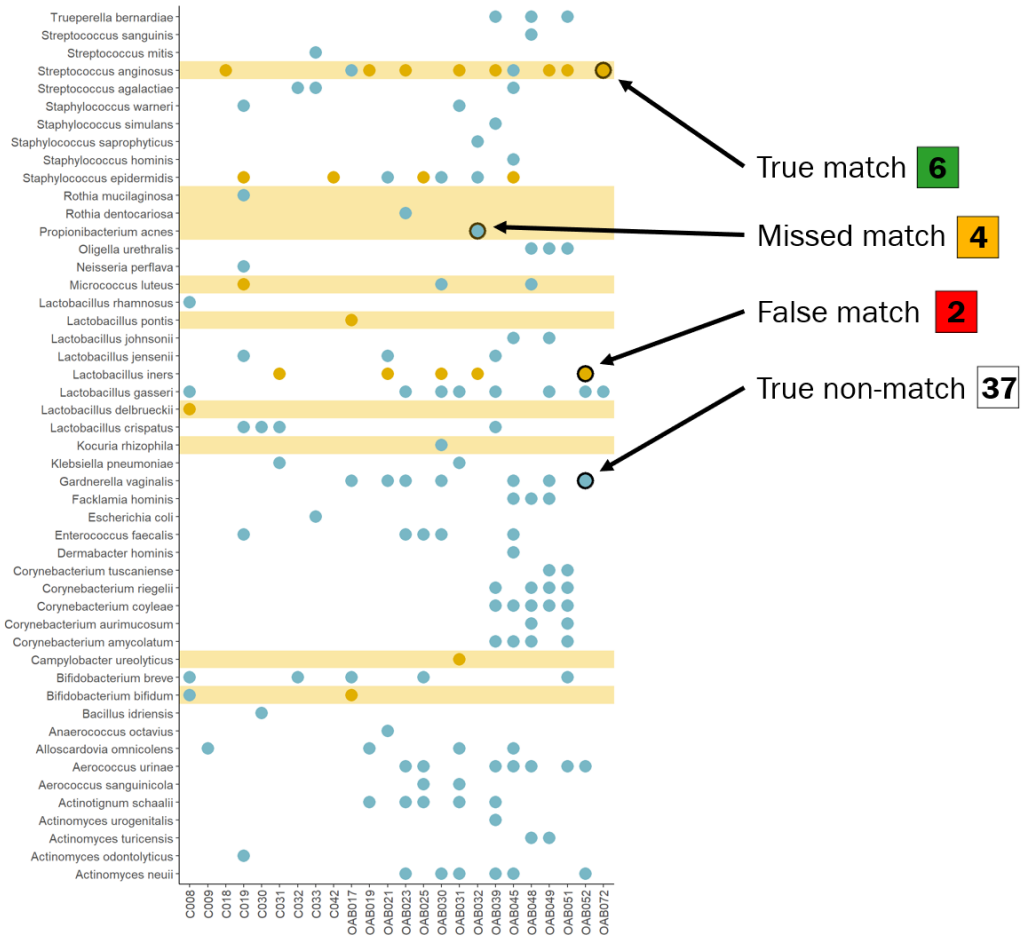

Figure 31: Definitions of how the classification scheme outcomes are assigned to the cells of the confusion matrix. This example describes the classification scheme composed of the Greengenes database, BLCA classifier, and the V4 region of the 16S rRNA gene as the identifier. Blue dots represent species identified in the collected samples by whole genome sequencing after expanded urine culturing and isolation. Yellow dots indicate the species were identified in those samples by V4 16S targeted amplicon sequencing. Yellow rows indicate the species correctly identified by the in silico methods. Yellow dots in yellow rows are True Matches, otherwise they are False Matches. Blue dots in yellow rows are Missed Matches, otherwise they are True Non-matches.

Because Figure 32 is large, and this validation is primarily concerned with Recall, the discussion of the evaluation results will only be accompanied by graphs of the predicted matches of the classification schemes plotted against the samples in which those species are found. However, the full results are shown in Figure 56.

Figure 32: The V4 validation set used in the in vitro results. When the Thomas-White dataset is subsetted by the 24 samples that underwent targeted amplicon sequencing, a smaller set of species/sample pairs remains out of all the possible pairings. The total number of species in this set is 49, and contains 106 species/sample pairs.

3.7 Whole genome and targeted amplicon sequencing comparison

The bacterial species identified in urine samples from symptomatic and non-symptomatic subjects through whole genome sequencing was obtained from the supplementary information of White et al 2018.

Targeted amplicon sequences of those same urine samples using the V4 region of the 16S rRNA gene sequence was generously provided by Alan Wolfe. The raw sequence reads were processed with DADA2 version 1.14.1 to generate amplicon sequence variants (ASVs). The ASVs were classified with BLCA using the Silva, Greengenes, and NCBI 16S databases as described in this paper.

3.8 PCR optimization

3.8.1 Reaction conditions

Because degenerate primers are a mix of all possible oligonucleotides for the designated primer sequence, the working molarity of the PCR reaction solution needs to be calculated. The molarity of a non-degenerate primer used in a PCR reaction is calculated starting from the stock solution.

$$\left(\frac{10 \times 10^{-6}\text{mol stock oligo solution}}{\text{L}}\right) \times \left(\frac{6.022 \times 10^{23}\text{molecules}}{\text{mol}}\right) \times (1 \times 10^{-6}\text{L primer})$$

$$= 6.022 \times 10^{12} \text{ molecules of oligo}$$

From this value, the final molarity in a 50µL reaction is

$$(6.022 \times 10^{12} \text{ total molecules of oligo}) \times \left(\frac{\text{mol}}{6.022 \times 10^{23}\text{molecules}}\right) \times \left(\frac{1}{50 \times 10^{-6} \text{ L reaction volume}}\right)$$

$= 2 \times 10^{-7}\text{M}$

$=$ 200nM of plain non-degenerate primer

For degenerate primers (Table 9), the numbers of each different oligo in the solution was assumed to be equal.

Name	Degeneracy	Gene target
v3_579F	288	16S rRNA
v3_779R	768	16S rRNA
v6_1183F	18	16S rRNA
v6_1410R	4	16S rRNA
541_811F	3072	Ffh
541_995R	3072	Ffh
85_3124F	1536	RpoB
85_3274R	2048	RpoB

Table 9: Degeneracy of the primers designed in this study.

Using the reverse primer of *RpoB* as an example, divide the number of oligos in solution by 2048 to get the molecules of each oligo.

$$(6.022 \times 10^{12} \text{ total molecules of oligo}) \times \frac{1}{2048 \text{ different RpoB oligos}}$$

$$= 2.9 \times 10^{9} \text{ molecules of each degenerate RpoB oligo}$$

The molarity of the RpoB reverse primer in the reaction volume is:

$$(2.9 \times 10^{9} \text{ molecules RpoB oligos}) \times \left(\frac{\text{mol}}{6.022 \times 10^{23} \text{molecules}}\right) \times \left(\frac{1}{50 \times 10^{-6} \text{ L reaction volume}}\right)$$

$$= 9.6 \times 10^{-11} \text{M}$$

$$\approx 0.1\text{nM of degenerate RpoB reverse primer}$$

This molarity is about a thousand times less than found in a common non-degenerate primer PCR reaction.

Because the V6 reaction amplified something to the point that could see it on a gel, and corresponded to a reaction molarity of 10μM, this value was used as a minimum molarity for all other degenerate primers used.

In general, the volume of degenerate primer needed is found by

$$(x \times 10^{-6}\text{L}) \times \left(\frac{10 \times 10^{-6}\text{mol stock oligo solution}}{\text{L}}\right) \times \left(\frac{1}{\text{primer degeneracy}}\right) \times \left(\frac{1}{50 \times 10^{-6}\text{ L reaction volume}}\right) > \frac{10 \times 10^{-9}\text{mol}}{\text{L}}$$

$$(x \times 10^{-6}\text{L}) > \left(\frac{10 \times 10^{-9}\text{mol}}{\text{L}}\right) \times (50 \times 10^{-6}\text{ L}) \times \text{primer degeneracy} \times \left(\frac{\text{L}}{10 \times 10^{-6}\text{mol stock oligo solution}}\right)$$

$$(x\mu\text{L}) > (5 \times 10^{-8}\text{L}) \times \text{primer degeneracy}$$

Table 10 lists the reaction volumes that were used for PCR

	RpoB	Ffh	V3	V6
forward primer	11μL of 100μM	16μL of 100μM	5μL of 100μM	1μL of 10μM
reverse primer	11μL of 100μM	16μL of 100μM	5μL of 100μM	1μL of 10μM
polymerase	0.25μL	0.25μL	0.25μL	0.25μL
DNA	1μL	1μL	1μL	1μL
5x buffer	10μL	10μL	10μL	10μL
MgCl2	3μL	3μL	3μL	3μL
dNTP	1μL	1μL	1μL	1μL
water	12.75μL	2.75μL	24.75μL	32.75μL

Table 10: Reagent volumes used in the PCRs to generate the amplicons from RpoB, Ffh and the 16S rRNA gene.

4. Results

4.1 Targeted gene sequences

4.1.1 Known variable regions of the 16S gene

To verify that the variable regions of the 16S gene sequence identified by SWA are actually capturing the previously published locations, the coordinates published in Chakravorty 2007(68) were consulted. These coordinates were found on the *E. coli* sequence included in the MSA of the Thomas-White dataset, and mapped to the weighted entropy graph of the MSA, shown in Table 11.

The results of the SWA in Table 11 show that the predicted variable regions match the positions of the known variable regions, and the highest peaks are located in the first 500 nucleotides of the multisequence alignment. This reflects the large amount of information available to discriminate between the different species.

variable region	start	stop	MSA start	MSA stop
V1	69	99	108	159
V2	137	242	198	323
V3	433	497	517	585
V4	576	682	664	772
V5	822	879	914	975
V6	986	1043	1083	1152
V7	1117	1173	1226	1283
V8	1243	1294	1352	1407
V9	1435	1465	1548	1583

Table 11: Coordinates of the 16S rRNA gene sequence variable regions projected onto the multisequence alignment of the Thomas-White dataset. "Start" and "stop" columns show variable region coordinates as described in Chakravorty 2007. "MSA start" and "MSA stop" columns show corresponding variable regions in the multisequence alignment.

Primer sets designed to amplify this region of the 16S gene are predicted to perform better than those targeted elsewhere in the sequence.

4.1.2 Variable regions of the protein encoding genes

Once the variable regions of the 16S gene sequence identified by SWA was verified to be accurate, the variable regions of the 40 protein encoding genes were calculated. Because *Ffh* and *RpoB* were the only genes to have successfully designed degenerate primers, only those SWA graphs are included (Figure 33 and Figure 34).

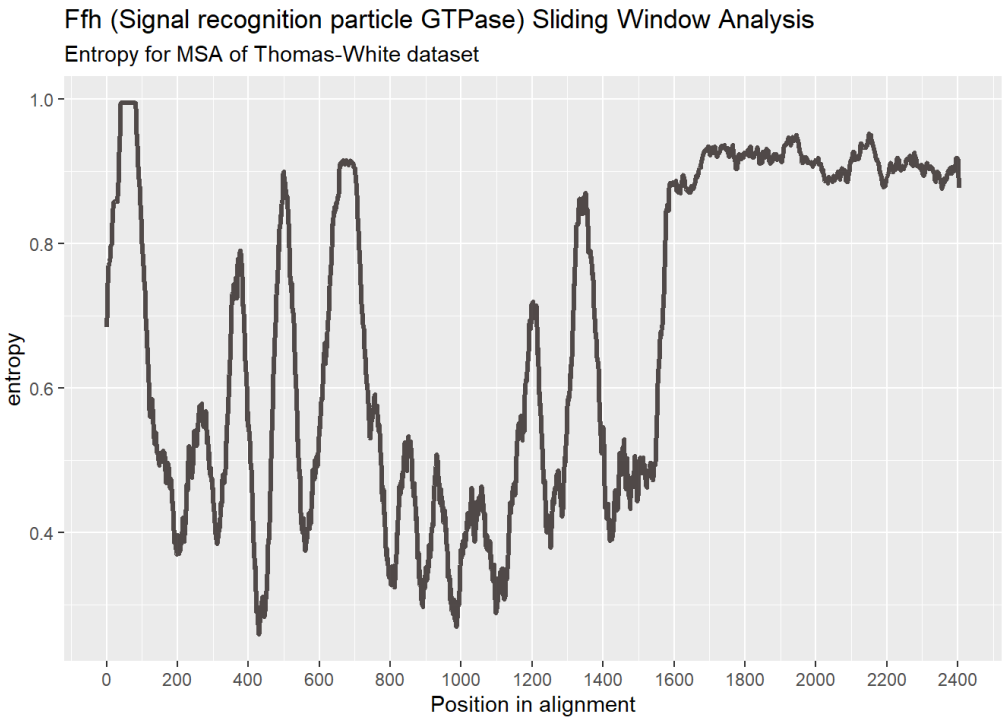

Figure 33: Predicted variable regions for the Ffh genes found in the Thomas-White dataset after sliding window analysis. The high entropy values after position 1600 indicate frequent insertions and deletions in the sequences that make up the multisequence alignment.

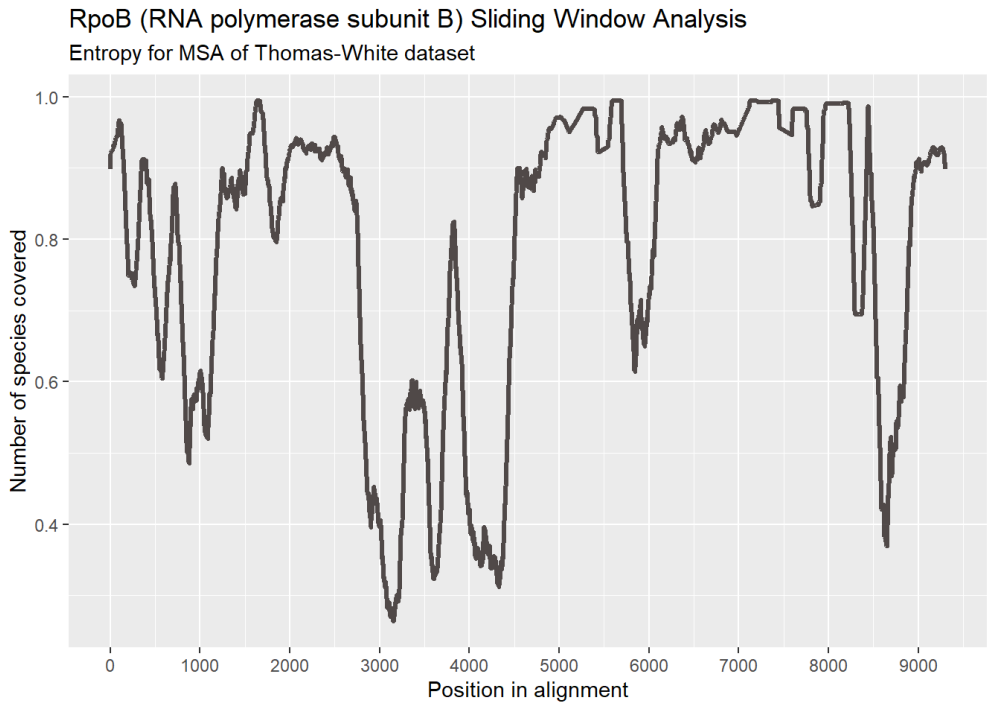

Figure 34: Predicted variable regions for the RpoB genes found in the Thomas-White dataset after sliding window analysis. The high entropy values surrounding the region from 3000 to 4500 indicate frequent insertions and deletions in the sequences that make up the multisequence alignment.

4.1.3 Primers

primers from literature

After a literature search, the primers shown in Table 12 were used to generate the *in silico* 16S rRNA gene amplicons from the Thomas-White dataset.

var_region	primers	publication	notes
V1-V3	A17F, 515R	Int Urogynecol J. 2017 May;28(5):711–20	possible typo at pos 9
V2-V3	16S_BV2f, 16S_BV3r	Sci Data. 2019 Mar;6(1):190007	reverse primer sequence is exactly the same as BV2f, and isn't a palindrome. Typo?
V3-V4	V3F, V4R	Sci Rep. 2018 Dec;8(1):9678	
V4-V6	515F, 1114R	Int Urogynecol J. 2017 May;28(5):711–20	
V3-V4	MiCSQ_343FL, MiCSQ_806R	Sci Data. 2019 Mar;6(1):190007	no matches to my MSA, assuming numbers are position of E. coli 16S sequence
V4	F515, R806	Mar 15;108(Supplement_1):4516–22	

Table 12: Descriptions of the primers found in the literature used in this study.

Designed primers

16S gene

In addition, two sets of primers for the 16S rRNA gene sequence that are optimal for the Illumina sequencing platform were successfully designed, shown in Table 13. One set spans the V3 region, and the other spans the V6 region, as shown in Figure 35.

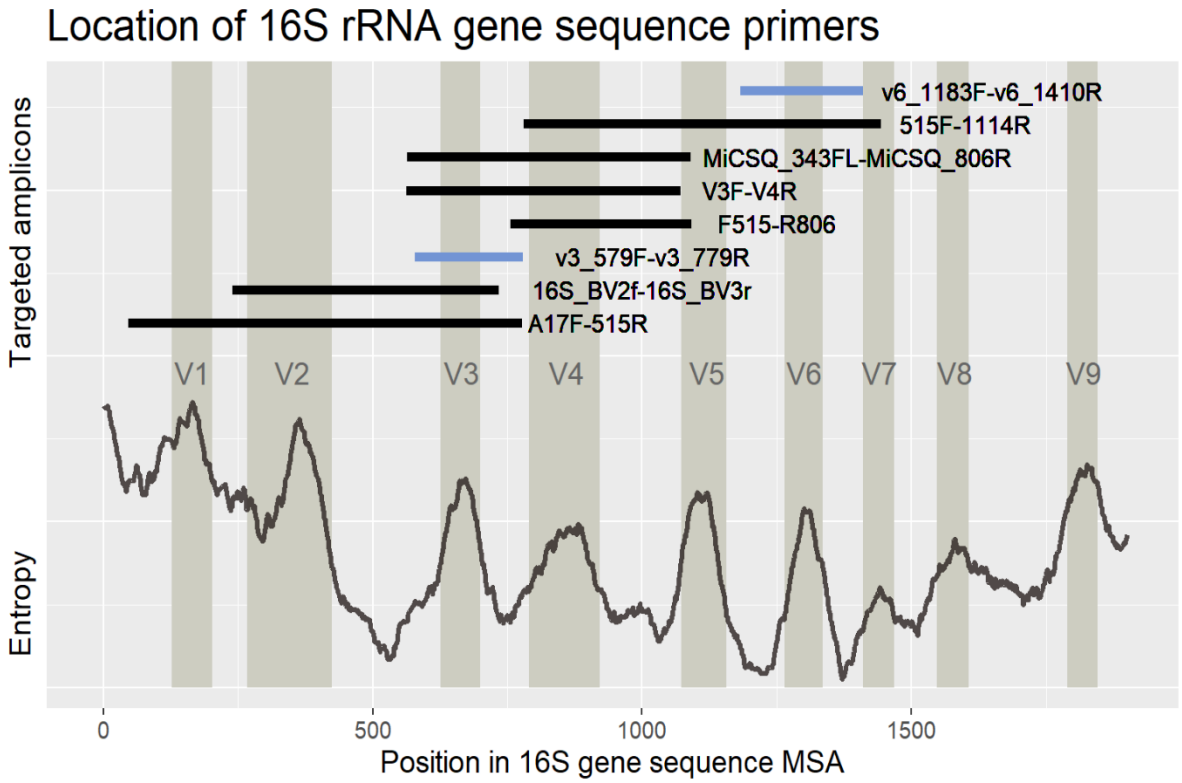

Figure 35: Locations of the primers used in this study on the 16S rRNA gene. Locations of the predicted amplicons are shown on the top of the graph, and the variable regions determined by sliding window analysis is shown by the entropy mapped on the bottom of the graph. Gray columns are the locations of the known variable regions based on the sequence from E. coli.

name	degeneracy	coverage	sequence	min temp	max temp
v3_579F	288	79/79	THTTSSRCAATGGRSGVA	30.89	54.57
v3_779R	768	79/79	GKNSCRAGCSTTRHYCGG	33.95	57.23
v6_1183F	18	79/79	CCGCCTGGGGASTACGVH	52.89	56.79
v6_1410R	4	79/79	AGTCCCRYAACGAGCGCA	53.95	56.89

Table 13: Primers designed in this study for the V3 and V6 variable regions, using the 16S rRNA gene sequence found in the Thomas-White dataset.

RpoB

There were several primer set candidates for this gene. The best primer set, listed in Table 14, was located just outside a variable region and is shown in Figure 36.

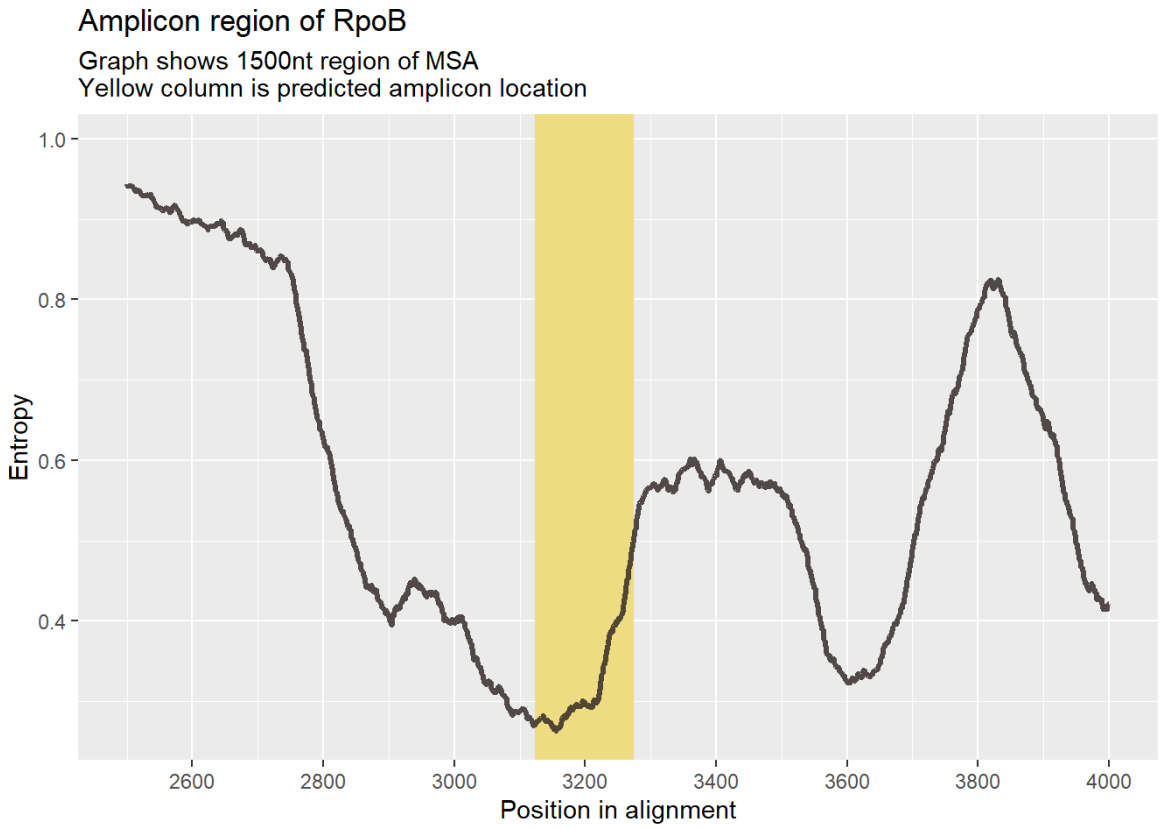

Figure 36: Location of the RpoB primers designed in this study. Graph is cropped from the full length MSA to show the entropy of the 1500 nucleotides surrounding the predicted amplicon.

name	sequence	coverage	min temp	max temp	degeneracy	Na min	Na max
85_3124F	TTYATGGAYCWVNMNAAY	79/79	23.35	46.36	1536	40.3	56.3
85_3274R	CCNGARGGNYMNAAYATY	79/79	27.90	51.58	2048	44.8	62.9

Table 14: Primers designed in this study for the RpoB gene.

Ffh

One primer set was successfully designed, listed in Table 15, and is predicted to span two variable regions shown in Figure 37.

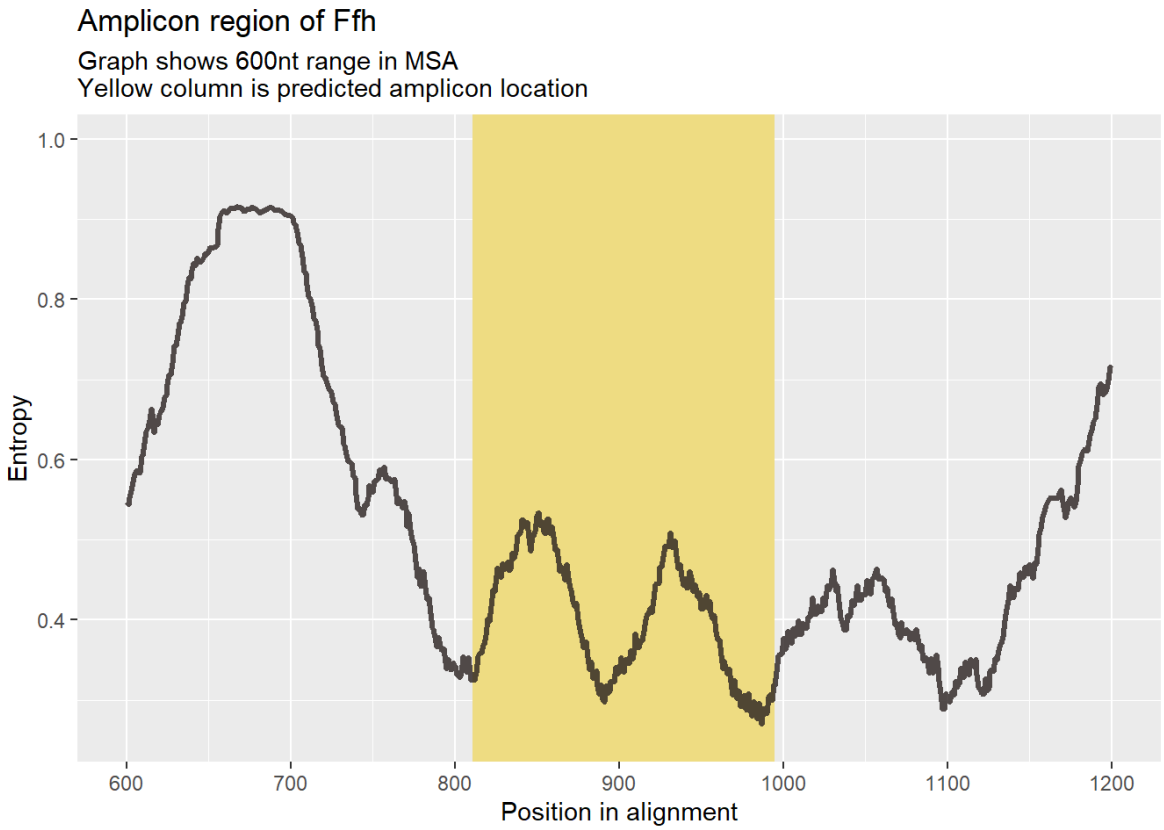

Figure 37: Location of the Ffh primers designed in this study. Graph is cropped from the full length MSA to show the entropy of the 600 nucleotides surrounding the predicted amplicon.

name	sequence	coverage	min temp	max temp	degeneracy	Na min	Na max
541_811F	BGAYACNGCNGGNCGNYT	78/79	36.32	58.85	3072	51.4	67.4
541_995R	GAYGGNGAYDCNCGNGGN	76/79	38.29	60.93	3072	53.8	69.8

Table 15: Primers designed in this study for the Ffh gene.

4.2 Records present in databases

When evaluating the performance of a classification scheme, it is important that the databases used meet a minimum number of species in the set of query sequences. Figure 38 shows which species are available in each of the databases used in this study. The Silva and NCBI 16S databases have a complete representation, while Greengenes and the NCBI genomic database did not. In contrast, the genera in Figure 39 shows that the Silva, NCBI 16S and NCBI genomic databases have a complete representation of the Thomas-White dataset. Greengenes database is the exception, and is missing the Globicella genus.

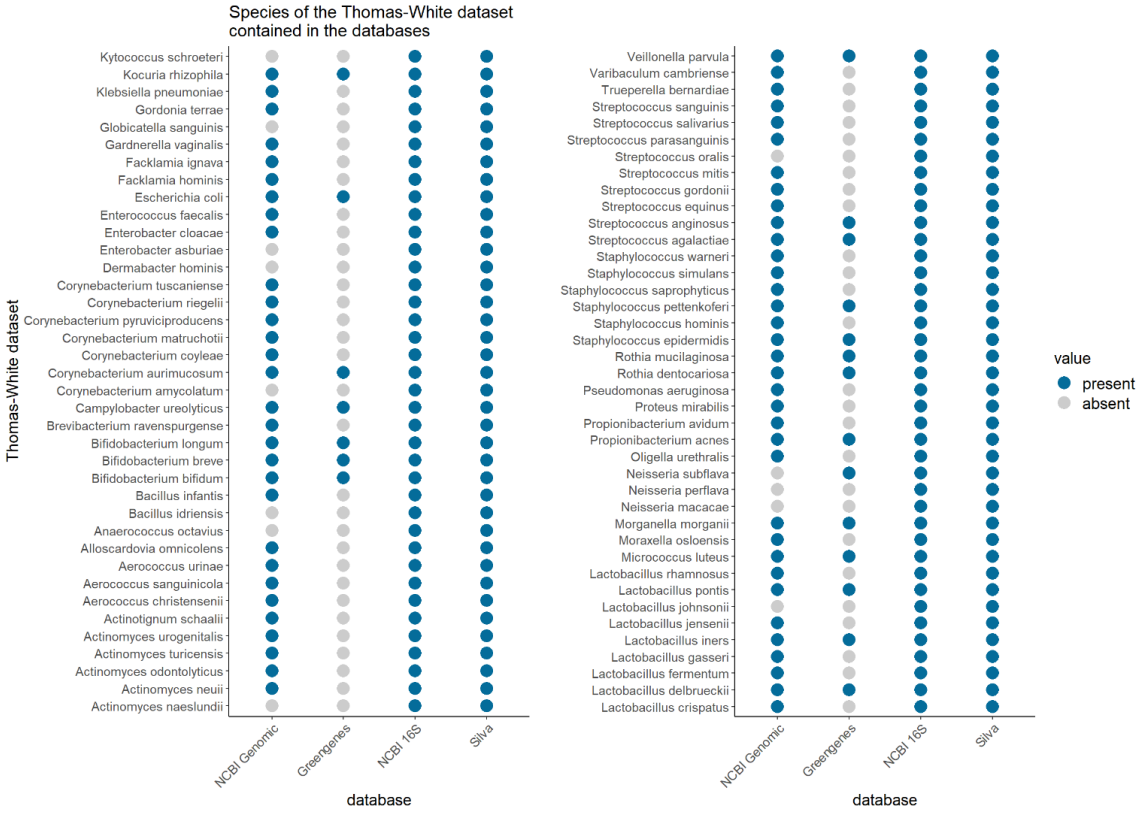

Figure 38: Presence of species from the Thomas-White dataset in each of the databases used in this study. All species of the dataset are present in the Silva and NCBI 16S databases, but many are missing from the Greengenes database. All missing species from the NCBI 16S genomic database were added while building the custom genomic database, with the exception of Bacillus idriensis.

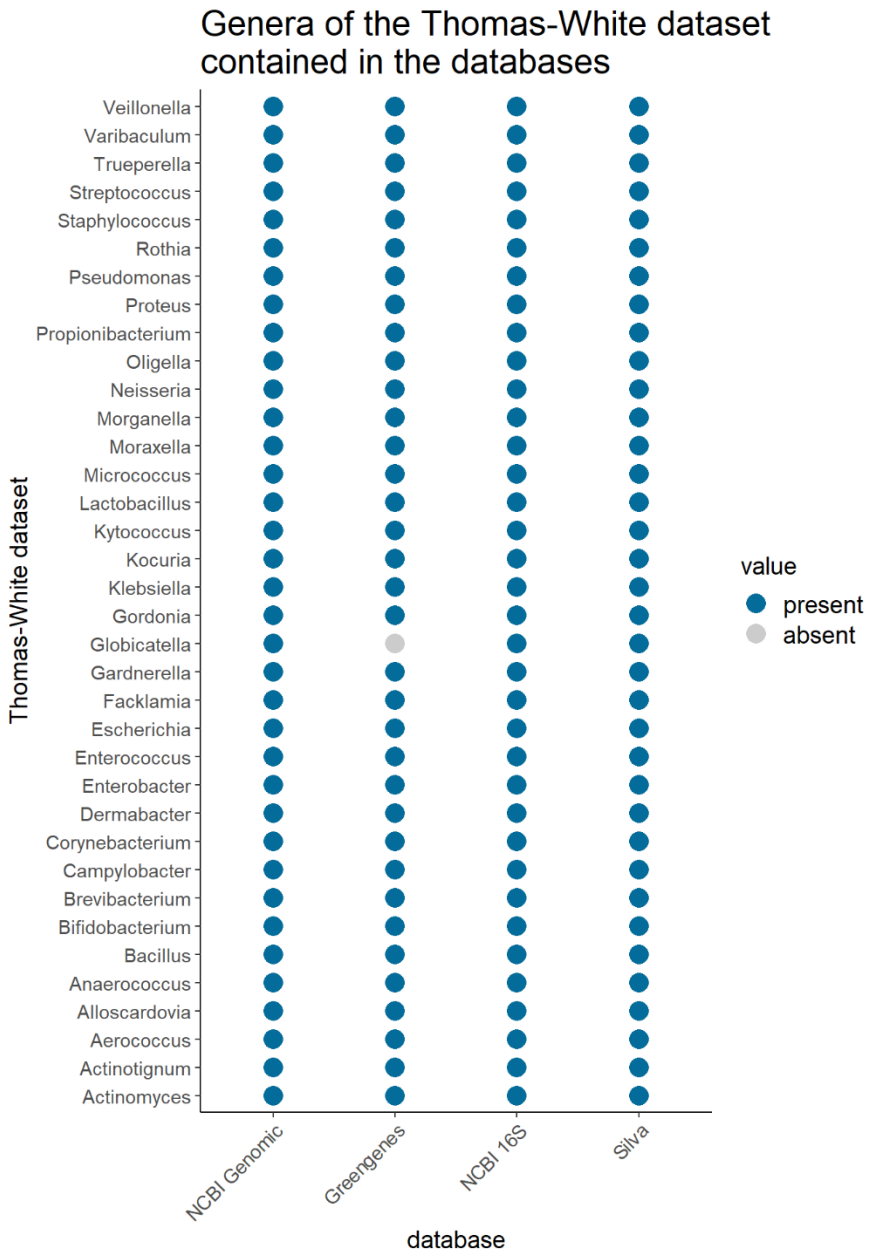

Figure 39: Presence of genera from the Thomas-White dataset in each of the databases used in this study. All genera are present in the databases, with the exception of the absence of Globicatella from the Greengenes database.

4.3 16S targeted amplicon classification schemes

4.3.1 Total number of true matches

The graphs in this section plot the performance (the F-measure) of the classifications schemes. However, because the F-measure is a proportion, it is still informative to look at the total number of true matches that are returned for each classification scheme when a choice of confidence score is made, because the total number of true matches returned by a classification scheme represents the actual data that can be used for further analysis. The graphs that show the relative performance of the classification schemes will be plotted using the F-measure versus the Recall. This will show how the classification schemes minimize the number of false matches and missed matches, and how many true matches are achieved relative to the total number of possible correct identifications.

Figure 40 shows the actual number of true matches for each value of the confidence score. In general, as the confidence score increases, the number of true matches decreases. Compared to the other databases, Greengenes has very little species-level records available, which is reflected in the small number of matches to the Thomas-White dataset. The number of true matches using the Silva database decreases gradually with an increase in confidence score. For the NCBI 16S database, there is a range of decreasing values for each variable region. V6 shows a steady decline as the confidence score increases, and the V1-V3 region shows a steep decrease after a confidence score of 0.75.

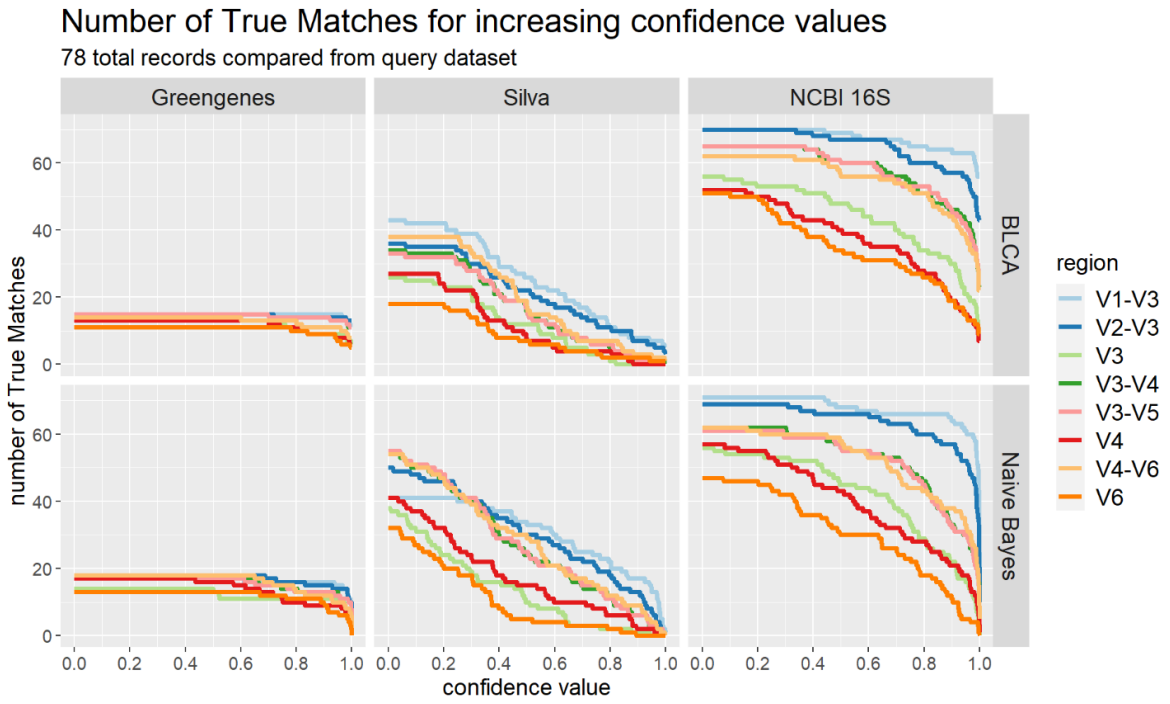

Figure 40: Number of True Matches returned for each classification scheme across all confidence score values.

The BLCA classifier assigned a range of confidence scores from 0 to 100, as shown by the endpoints of the lines terminating above the x-axis. This indicates that for some of the classification schemes, there were taxonomic assignments that had no other candidates arise during the bootstrap process. In contrast, the Naive Bayes classifier did not assign a confidence score of 100 to any of the results, as shown by the number of true matches plunging to zero for all variable regions.

4.3.2 Recall at best performance

While the BLCA and Naive Bayes classifiers have the option of using a confidence score to filter their results, it is worthwhile to see how the classification schemes evaluate when no filtering is done. Figure 41 shows the evaluation of the classification schemes when the confidence score is ignored. The results show that the choice of database has a dramatic effect on the performance of the classification scheme.

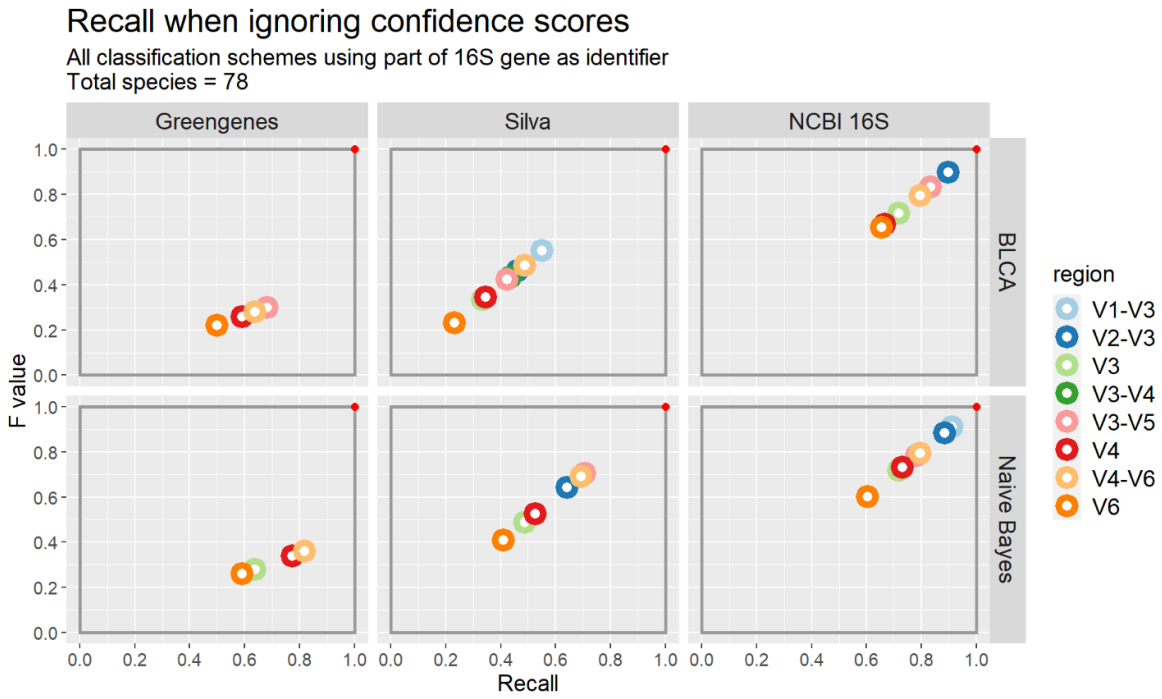

Figure 41: Classification results when confidence scores are ignored. The classification schemes that use the Silva and NCBI 16S database lie on a 45-degree slope, while the classification schemes that use the Greengenes database lie on a shallower slope. This may be because there are fewer species-level records in this database compared to the others.

In general, the classification schemes that use the NCBI 16S database perform well, while Silva has significantly reduced performance and Recall. Despite the lack of a full complement of species in the Greengenes database, this database does have comparable Recall to the classification schemes that use the Silva database, although the performance is well below Silva in comparison. Most classification schemes also tend to lie on a straight line terminating on the upper right hand corner. This is not surprising, because Recall is used as one of the

arguments in calculating the F-measure, and the evaluation of these classification schemes reflect the same number of species allocated in different numbers to the various parts of the confusion matrix. The exceptions are the classification schemes that use the Greengenes database. This database is lacking many of the species-level taxonomy found in the other databases. The lack of a complete representation of the Thomas-White dataset will be addressed below.

Some of these classification schemes in Figure 41 have Recall and F-measure values that are very close to others, and are difficult to see. The results for the classification schemes excluding the Greengenes database are listed in Table 16 and Table 17.

region	classifier	recall	fmeasure
V3-V5	Naive Bayes	0.705	0.705
V3-V4	Naive Bayes	0.692	0.692
V4-V6	Naive Bayes	0.692	0.692
V2-V3	Naive Bayes	0.641	0.641
V1-V3	BLCA	0.551	0.551
V1-V3	Naive Bayes	0.526	0.526
V4	Naive Bayes	0.526	0.526
V4-V6	BLCA	0.487	0.487
V3	Naive Bayes	0.487	0.487
V2-V3	BLCA	0.462	0.462
V3-V4	BLCA	0.436	0.436
V3-V5	BLCA	0.423	0.423
V6	Naive Bayes	0.410	0.410
V4	BLCA	0.346	0.346
V3	BLCA	0.333	0.333
V6	BLCA	0.231	0.231

Table 16: results of the classification schemes that use the Silva database.

There is a large difference between the outcomes of the identifiers of the classifications schemes that use the Silva and NCBI 16S database. When used with Silva and the Naive Bayes classifier, the identifiers that yield the highest Recall are the large V3-V5, V3-V4, and V4-V6 targeted amplicons (Figure 16). When using the BLCA classifier, the V1-V3 and V4-V6 amplicons have the highest

Recall. In contrast, the identifiers that yield the highest Recall in classification schemes that use the NCBI 16S database are the V1-V3 and V2-V3 targeted amplicons, regardless of classifier (Figure 17).

region	classifier	recall	fmeasure
V1-V3	Naive Bayes	0.910	0.910
V1-V3	BLCA	0.897	0.897
V2-V3	BLCA	0.897	0.897
V2-V3	Naive Bayes	0.885	0.885
V3-V4	BLCA	0.833	0.833
V3-V5	BLCA	0.833	0.833
V4-V6	BLCA	0.795	0.795
V3-V4	Naive Bayes	0.795	0.795
V4-V6	Naive Bayes	0.795	0.795
V3-V5	Naive Bayes	0.782	0.782
V4	Naive Bayes	0.731	0.731
V3	BLCA	0.718	0.718
V3	Naive Bayes	0.718	0.718
V4	BLCA	0.667	0.667
V6	BLCA	0.654	0.654
V6	Naive Bayes	0.603	0.603

Table 17: Results of the classification schemes that use the NCBI 16S database, arranged by decreasing values of Recall.

Peak performance of a classification scheme will minimize the number of missed matches and false matches while maximizing the number of true matches. However, the maximum number of true matches may be much lower than when ignoring the confidence score altogther. The effect on the classification schemes' Recall value when using the confidence score that yeilds the best performance is shown in Figure 42. These classification schemes are expected to show increased performance, but there may be some reduction in Recall due to the effect of when true matches that have confidence scores below the chosen threshold are evaluated as missed matches.

The results show the performance of all classification schemes slightly increase or remain the same, with the largest gains in performance shown by classification schemes that use the Greengenes database and Naive Bayes classifier.

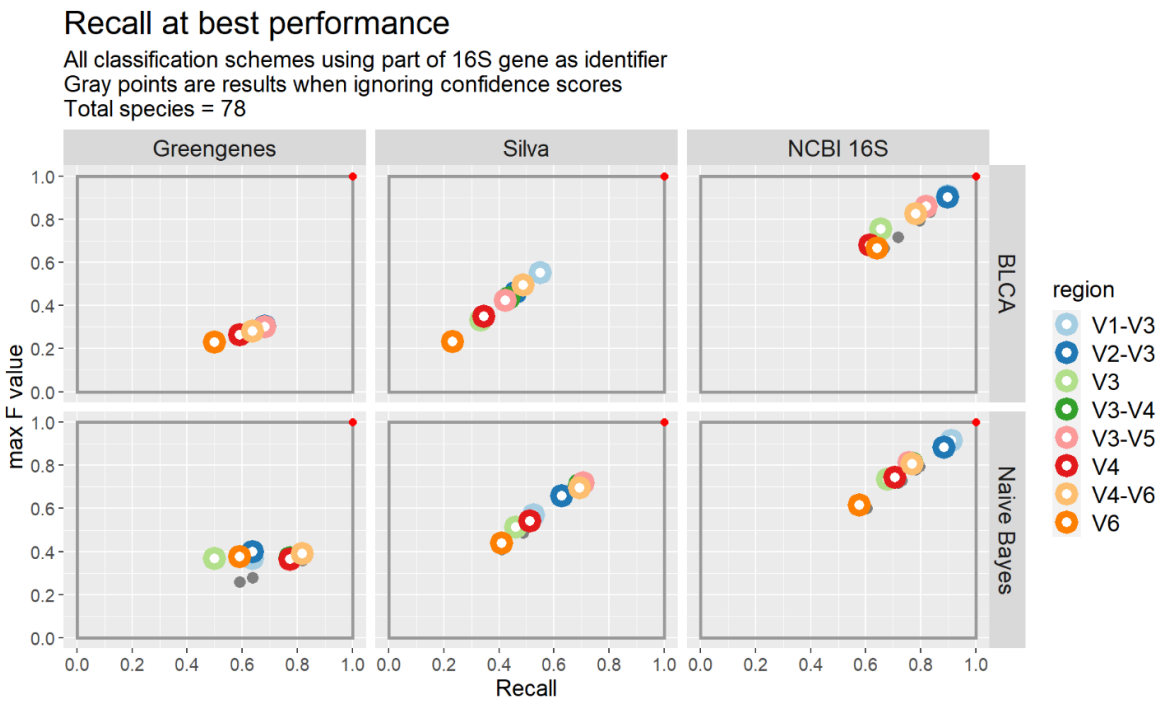

Figure 42: The percent of the Thomas-White dataset returned as true matches, at the best performance of each classification scheme. The classification schemes that use the Greengenes database slightly better performance.

Adjusting for Greengenes

Clearly, Greengenes is unsuited for species-level identification of this dataset. However, adjusting the other two databases to reflect the smaller set of species that are available in the Greengenes database may give a more accurate comparison of the performance and recall between the classification schemes used in this study. To do this comparison, the Silva and NCBI 16S databases were filtered to remove the species that were not contained in the Greengenes database (a set of 33 species), and the same evaluation steps were performed as before.

These results are shown in Figure 43. Overall, there is marked improvement in the performance and recall in the classification schemes that use the Greengenes database, to the point where those schemes are comparable to those that use the

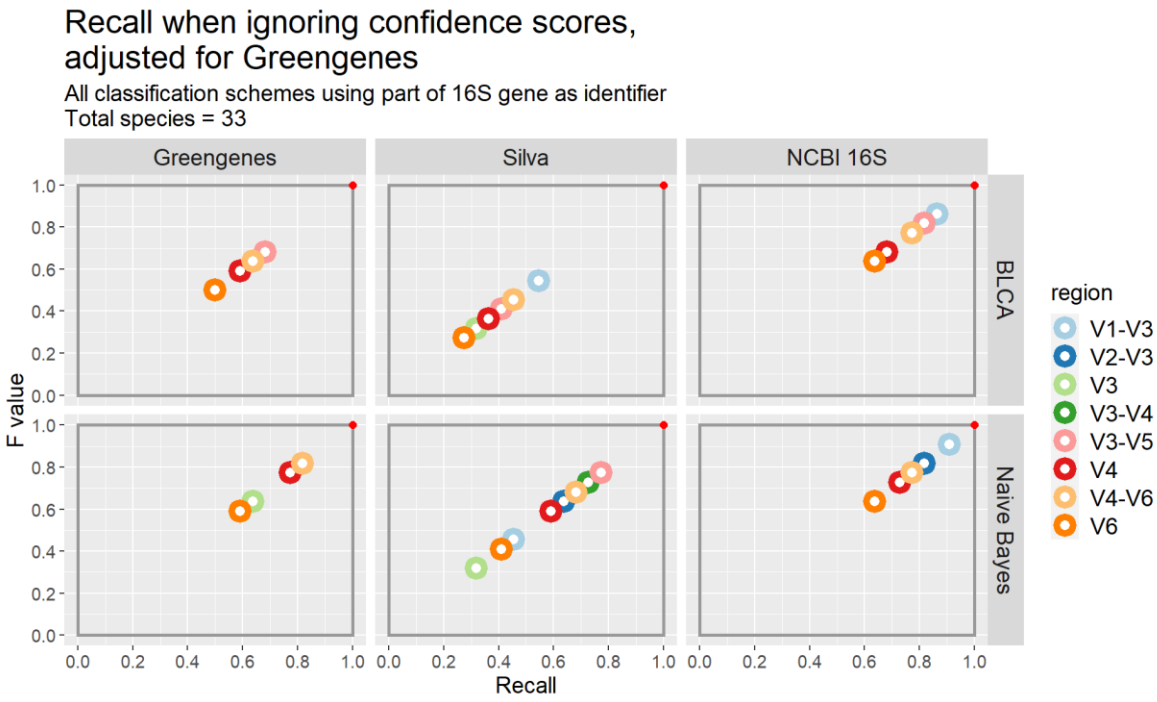

Figure 43: Performance and Recall for all classification schemes. The species in the Silva and NCBI 16S databases were adjusted to only match those found in the Greengenes database.

Silva database. The classification schemes that use the BLCA classifier and the Silva database are now the lowest performing and have the smallest values of Recall of the dataset. The classification schemes that use the NCBI 16S database still show the highest performance and values of Recall in general, and the V1-V3 and V1-V2 targeted amplicons are still the predominant identifiers.

4.3.3 Recall and performance at predefined confidence scores

When lacking any knowledge of how to choose the best confidence score that maximize the performance of a classification scheme, falling back on a predefined confidence score is an option. Two such confidence score values are 80% and 50%, and the next set of results show how these values affects the performance and number of true matches of each classification scheme.

80% confidence score

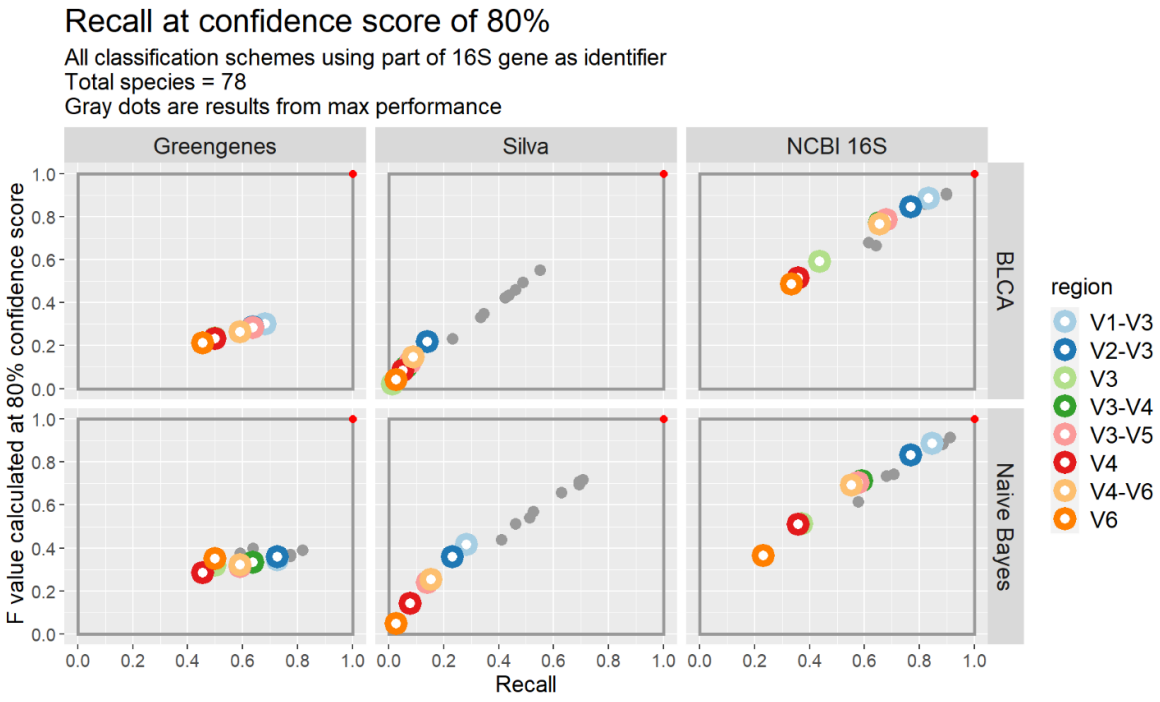

Figure 44: The percent of the Thomas-White dataset returned as true matches, using a confidence score of 80%. The classification schemes that include the Silva databases have a very low proportion of true matches returned.

Contrary to what is a reasonable expectation, almost all classification schemes suffered from using a default confidence score of 80%. As shown in Figure 44, this effect is especially marked for the classification schemes that use the Silva database. Classification schemes that use the Greengenes database are slightly affected, and it appears that the schemes that use the BLCA classifier with this database are hardly affected at all. Both the performance and Recall are severely reduced when using the Silva database, indicating that the number of missed matches and false matches are in greater proportion to the number of true matches. The classification schemes that use the NCBI 16S database have been unequally affected by the high threshold. While the V6 and V4 targeted amplicons have large losses in performance and Recall, the V1-V3 and V2-V3 targeted amplicons show slight decreases.

50% confidence score

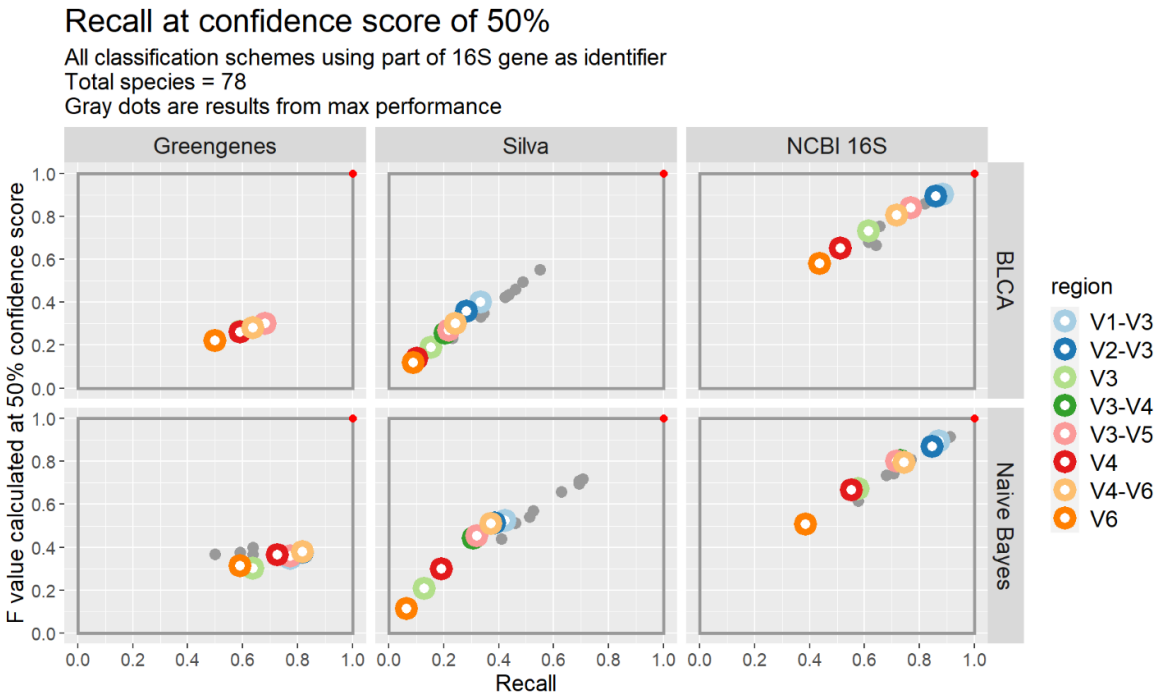

Figure 45: The percent of the Thomas-White dataset returned as true matches, using a confidence score of 50%. Many of the classification schemes still return a low proportion of true matches compared with the confidence score associated with peak performance.

As shown in Figure 45, using a confidence score of 50% is much the same as the 80% confidence score, but the performance of the classification schemes are slightly better than using a higher confidence score. The classification schemes using the NCBI 16S database have formed a tighter group, and the results using the Silva database have shown moderate improvement over the results using the 80% confidence score. Somewhat surprising result is that the classification schemes using the Greengenes database and Naive Bayes show an improvement in Recall beyond the values at peak performance, but show no gains in performance.

Comparing the number of true matches

In Figure 46, the number of true matches are compared for all classification schemes at the default confidence scores and the confidence score that yields peak performance. By far, the impact of the confidence score is highest when

using the Silva database. In general, the classification schemes that use the NCBI 16S database for either default confidence score tend to show a steady decrease in true matches compared to peak performance. The Greengenes database are largely unaffected by choice of confidence score. Overall, the classification schemes that use the NCBI 16S database return the most True matches of any other classification scheme, followed by the Silva database, and finally the Greengenes database.

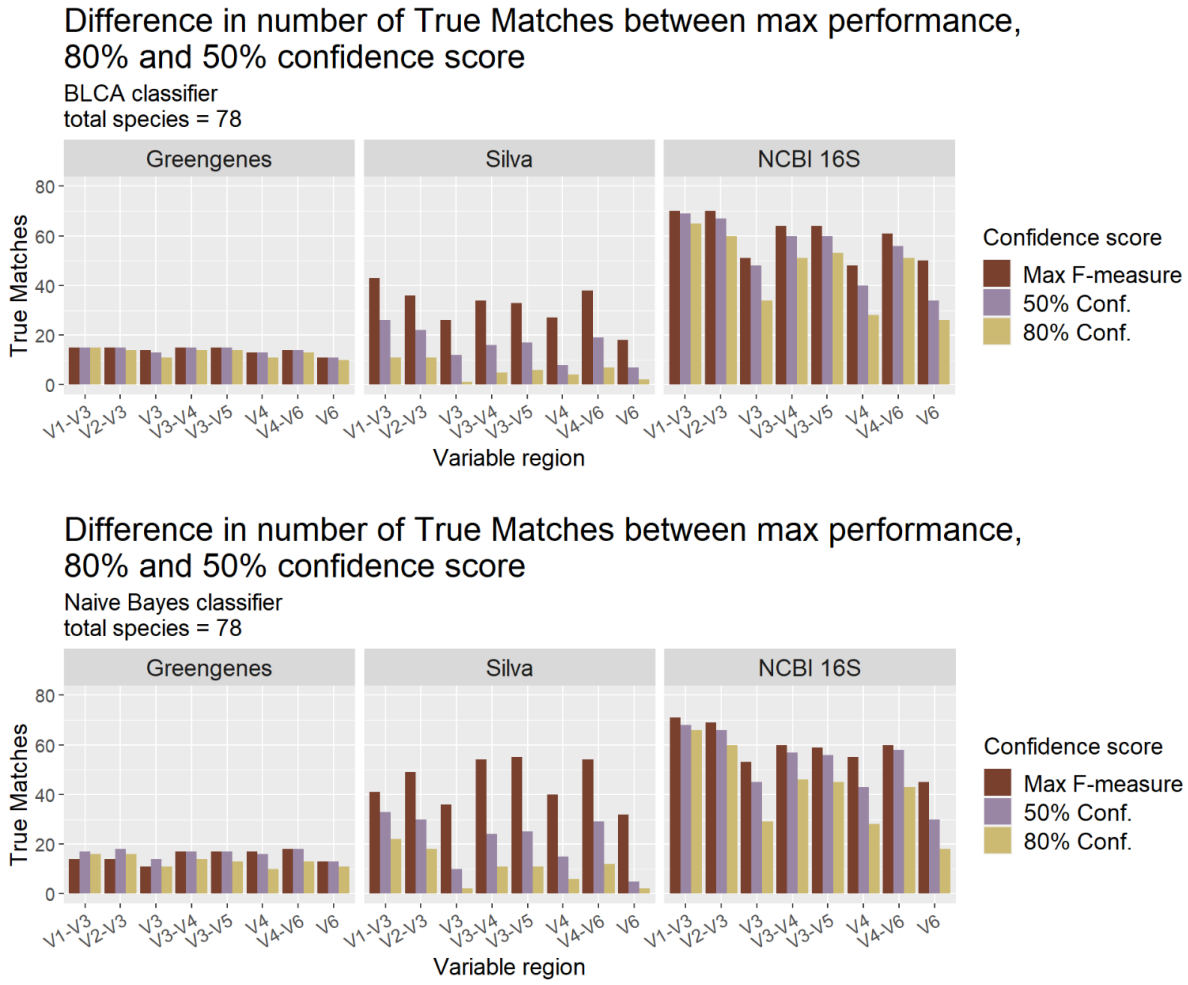

Figure 46: Comparison of the number of True Matches returned at the two default confidence score values, for each of the classification schemes. "Max F1" indicates the confidence score associated with the highest F-measure score for the classification scheme. (Top) Number of True Matches when using the BLCA classifier. (Bottom) Number of True Matches when using the Naive Bayes classifier.

4.4 Protein encoding and 16S targeted amplicon classification schemes

The results of the previous section shows that using the NCBI 16S database and targeted amplicons that include the V1 to V3 regions as identifiers are the best classification schemes when using the ribosomal small subunit gene as a marker. To compare this gene with the additional 40 protein coding genes, a database that includes all of these sequences is required. This requirement is be met by using the custom database build described in this study, which contains genomic sequences of bacteria and has a representation of almost all the species in the Thomas-White dataset.

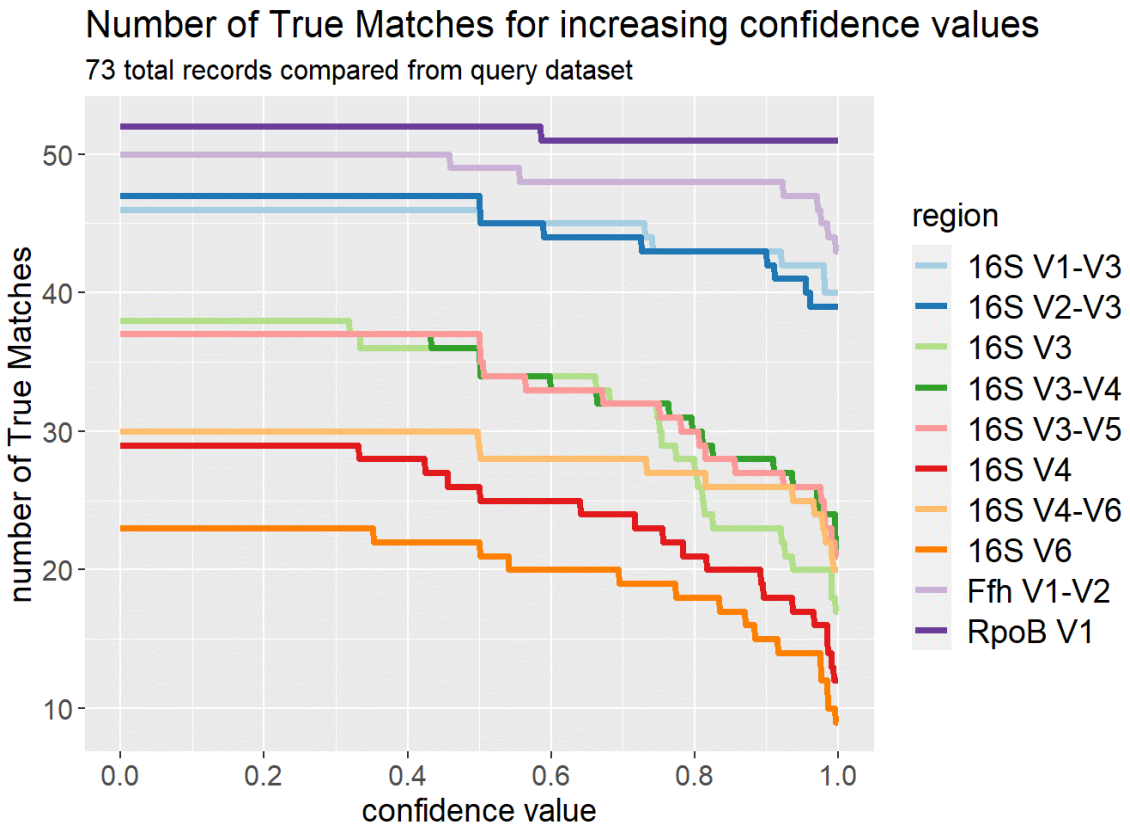

Figure 47: Number of true matches returned for increasing confidence values, using the custom database with all primer sets and the BLCA classifier. The protein encoding genes Ffh (light purple) and RpoB (dark purple) return more true matches than any variable region of the 16S rRNA gene.

Training a genomic database for use with the Naive Bayes classifier is computationally prohibitive, and this section compares the classification schemes that are only composed of the BLCA classifier and the custom database.

4.4.1 Classification scheme performance

Figure 47 shows that in all classification schemes, the two protein encoding genes returned more true matches than any identifier derived from the 16S rRNA gene.

Figure 48: Performance and Recall for all classification schemes that use variable regions from the 16S rRNA gene, Ffh and RpoB as identifiers. (left) Recall and performance of classification schemes when the confidence value is ignored. (Right) Recall and performance of classification schemes that use a confidence score value which yeilds the maximum F-measure. The protein-encoding marker genes return more true matches than any of the 16S rRNA variable regions used as identifiers.

The right hand side of Figure 48 shows the Recall and maximum performance for all identifiers used in this study. Overall, the protein encoding genes had higher F-measures and higher Recall than any other identifier, which indicates that these identifiers have the lowest proportion of false matches and missed matches to the number of true matches. For the identifiers derived from the 16S rRNA gene, the V2-V3 region had the best performance and highest recall of the remaining identifiers. These results also show that identifiers that use any of the

variable regions between V1 to V3 of the 16S rRNA gene do better than identifiers that include the V4 to V7 variable regions.

4.5 Exact matching

Exact matching searches a reference database for a corresponding record that has the exact same sequence as the query. As stated in the methods section, many of the sequences comprising the Thomas-White dataset contained ambiguous nucleotides, and those targeted amplicons in which the number of removed sequences exceeded 7 species was labeled as *biased*. What remained was the V3 and V6 targeted amplicons, and are labeled as *unbiased*. As the classification schemes that include the V3 amplicon are consistently in the mid-range ranking, and those that include the V6 are consistently in the last rank, these provide a good range of evaluated performance and Recall values to compare against.

Figure 49 shows how may true matches are returned as confidence scores are increased. At confidence score of zero, Exact Matching returns about the same number of true matches as BLCA, but fewer than Naïve Bayes. The number of

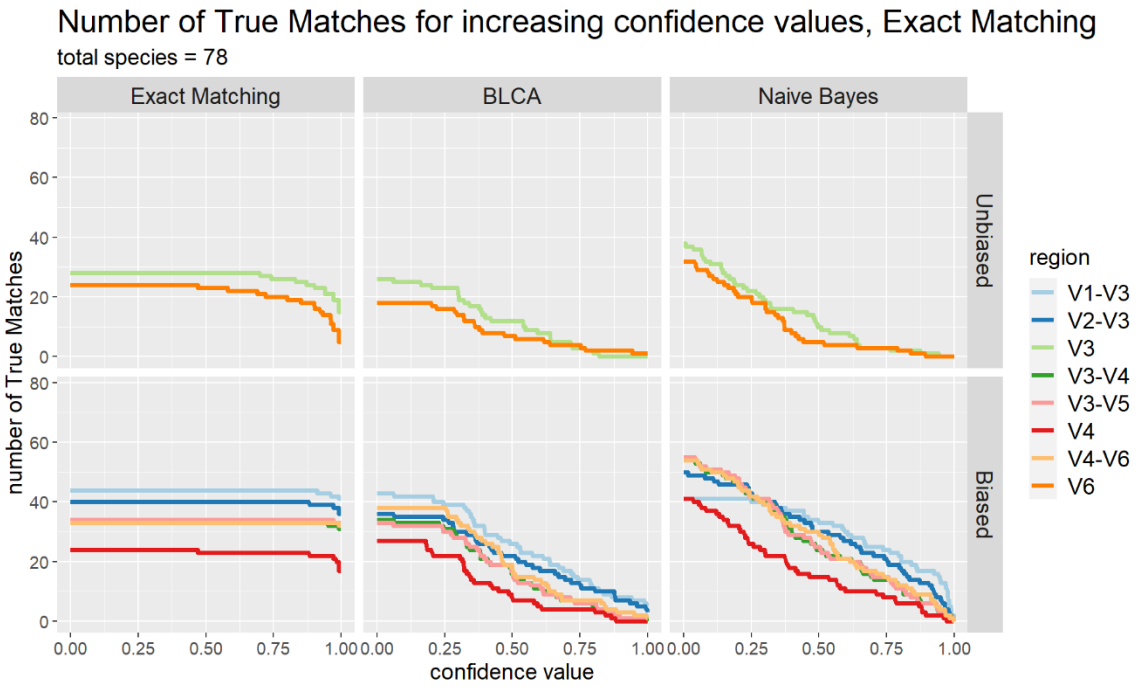

Figure 49: Number of true matches returned for increasing values of confidence score using Exact Matching, BLCA and Naive Bayes as the classifier. The top row are the two targeted amplicons that were below the missing data threshold, and are considered unbiased results. Bottom row are the classification schemes that had exceed the missing data threshold, and are considered biased results.

true matches returned by Exact Matching tends to hold as confidence scores are increased, until a quick reduction after a value of about 75. BLCA and Naïve Bayes both show a constant decrease as confidence scores increase.

Figure 50 shows the number of true matches returned at the maximum F-measure value. In general, all classification schemes have similar values for Recall, but Exact Matching and Naïve Bayes have higher F-measures. These results may be contrary to what would be expected from a classifier that is described as *exact*, but Recall is not the performance measurement that shows the best character of the method. While the true matches are assigned by the classifier at what is practically a 100% confidence score, the classifier can still assign an match to a record that has no useful taxonomic information, which is then evaluated as a false match.

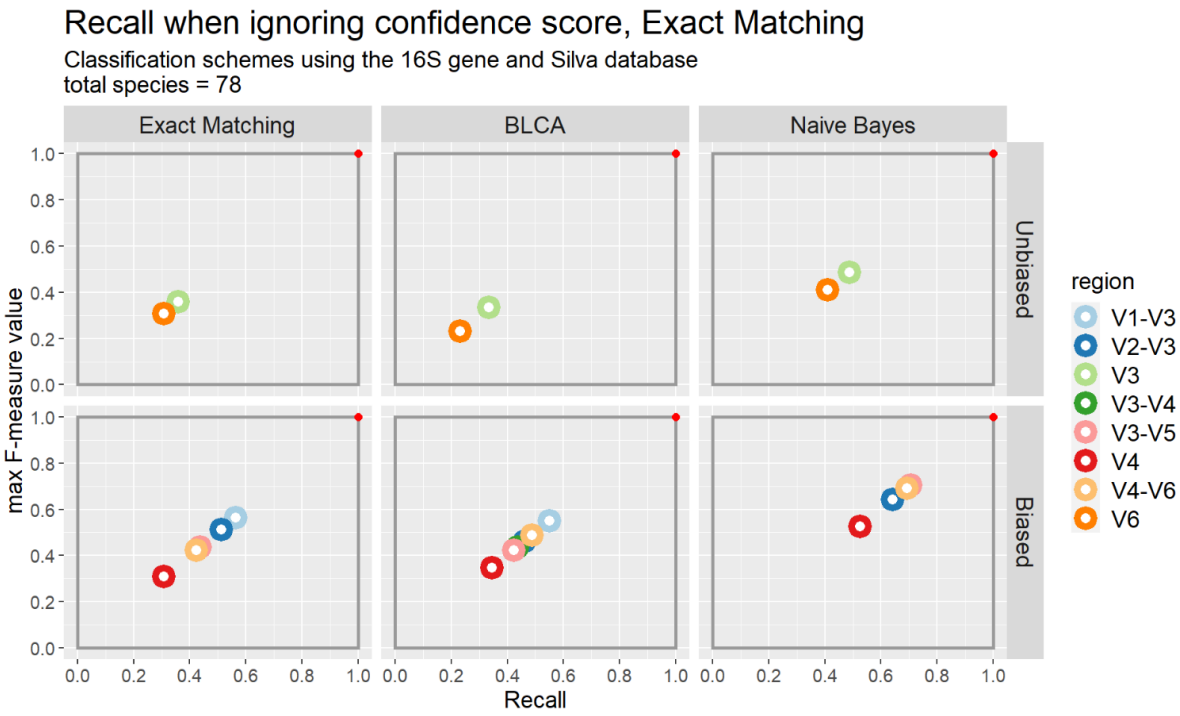

Figure 50: Number of true matches returned at the best performance of each classification scheme using Exact Matching, BLCA and Naive Bayes as the classifier. The top row are the two targeted amplicons that were below the missing data threshold, and are considered unbiased results. Bottom row are the classification schemes that had exceed the missing data threshold, and are considered biased results.

At maximum performance, the otherwise false matches are evaluated as missed matches. Recall is then dominated by a large number of missed matches, but the remaining set of classified matches have very few false matches. Precision is the right performance measurement for the Exact Match classifier, and this can be seen in Figure 51. It should be noted that the F-measure is still brought down by the low Recall, and the moderate value of the F-measure in Figure 51 reflects this relationship.

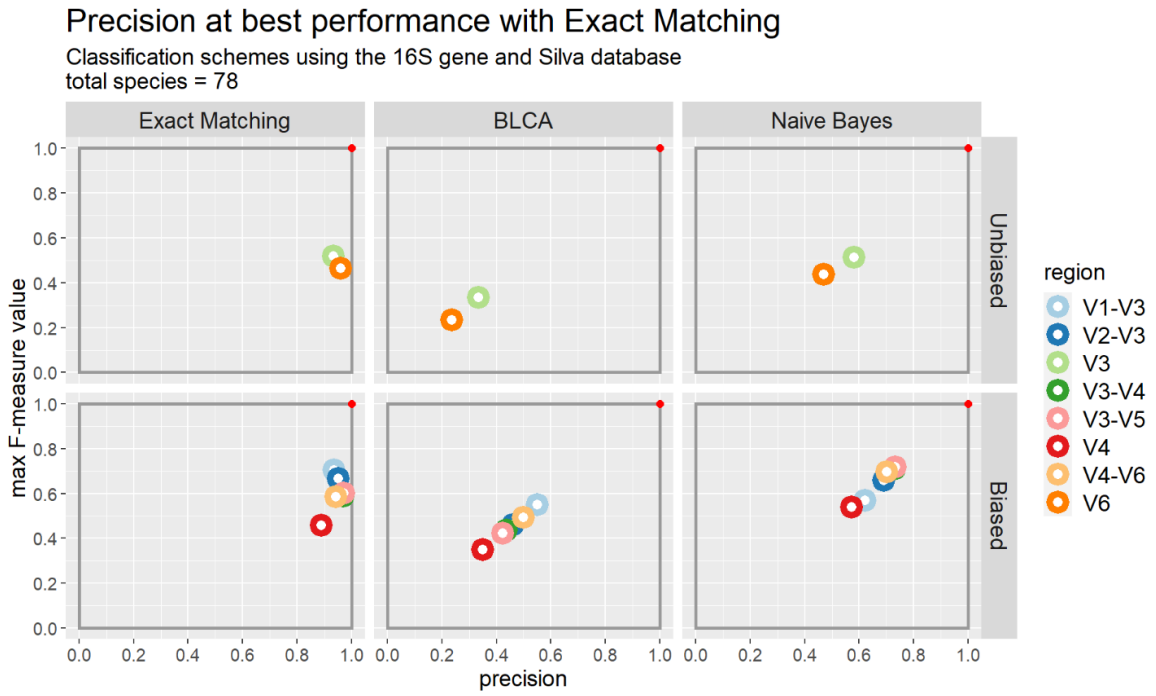

Figure 51: Precision at highest performance when using Exact Matching, BLCA and Naive Bayes. The Exact Matching classifier achieves very high values for precision due to its demand that two compared records are exactly the same before assigning the pair as a match. However, the classification schemes that use exact matching are dominated by a high number of missed matches. The top row are the two targeted amplicons that were below the missing data threshold, and are considered unbiased results. Bottom row are the classification schemes that had exceed the missing data threshold, and are considered biased results.

80% and 50% confidence scores

One strength of the Exact Match classifier is shown in Figure 52 and Figure 53. The confidence scores for the assigned true matches are extremely high, and using a default threshold of 50% or 80% does not affect those classification schemes that depend on Exact Matching. In contrast, the other unbiased classification schemes that use Naive Bayes or BLCA are still heavily impacted by those thresholds.

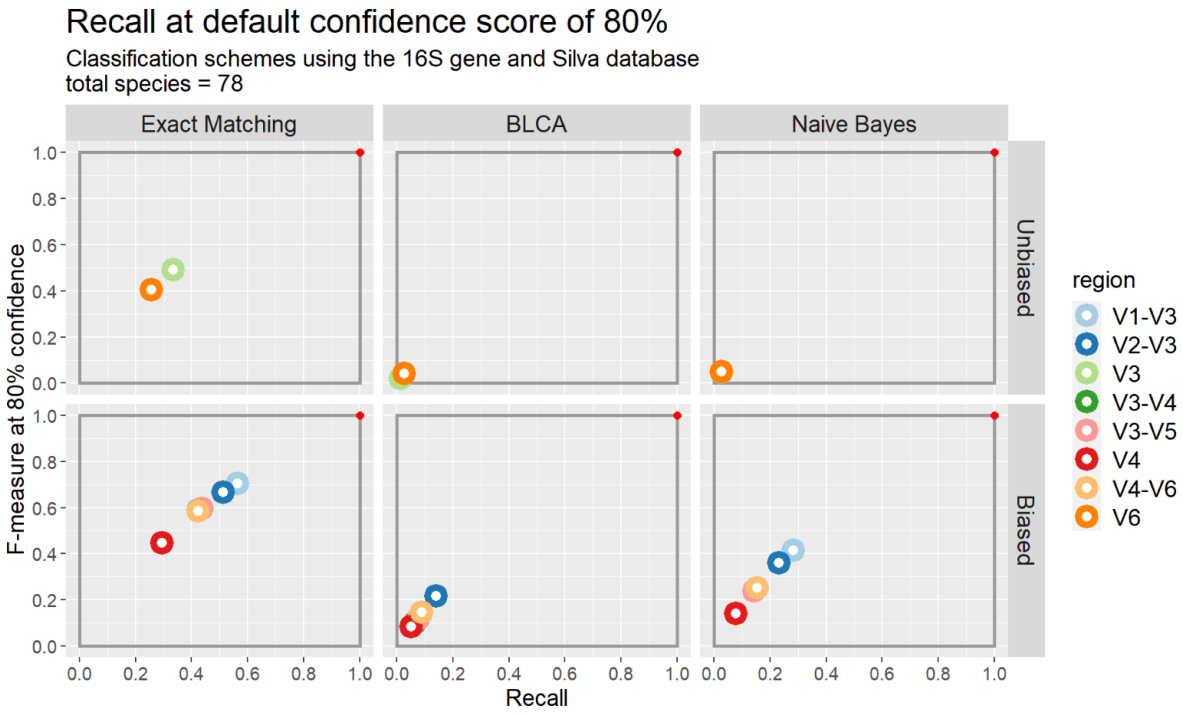

Figure 52: Number of true matches returned using a confidence score of 50% as a threshold, for each classification scheme using Exact Matching, BLCA and Naïve Bayes as the classifier. The top row are the two targeted amplicons that were below the missing data threshold, and are considered unbiased results. Bottom row are the classification schemes that had exceed the missing data threshold, and are considered biased results.

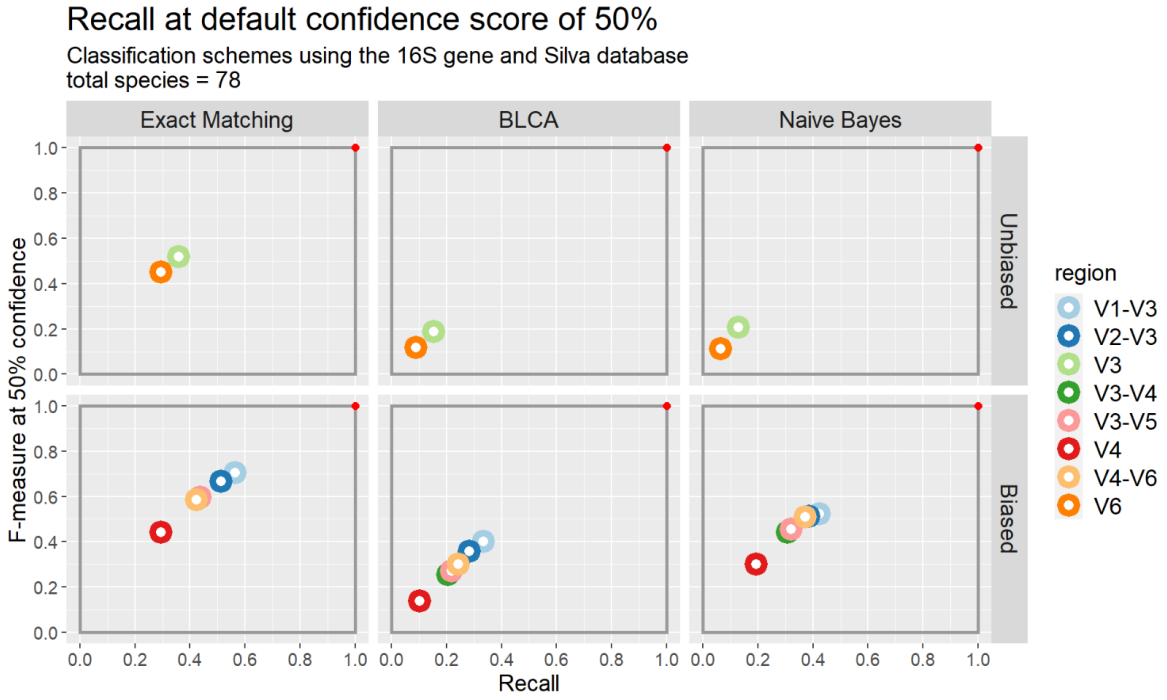

Figure 53: Number of true matches returned using a confidence score of 80% as a threshold, for each classification scheme using Exact Matching, BLCA and Naïve Bayes as the classifier. The top row are the two targeted amplicons that were below the missing data threshold, and are considered unbiased results. Bottom row are the classification schemes that had exceed the missing data threshold, and are considered biased results.

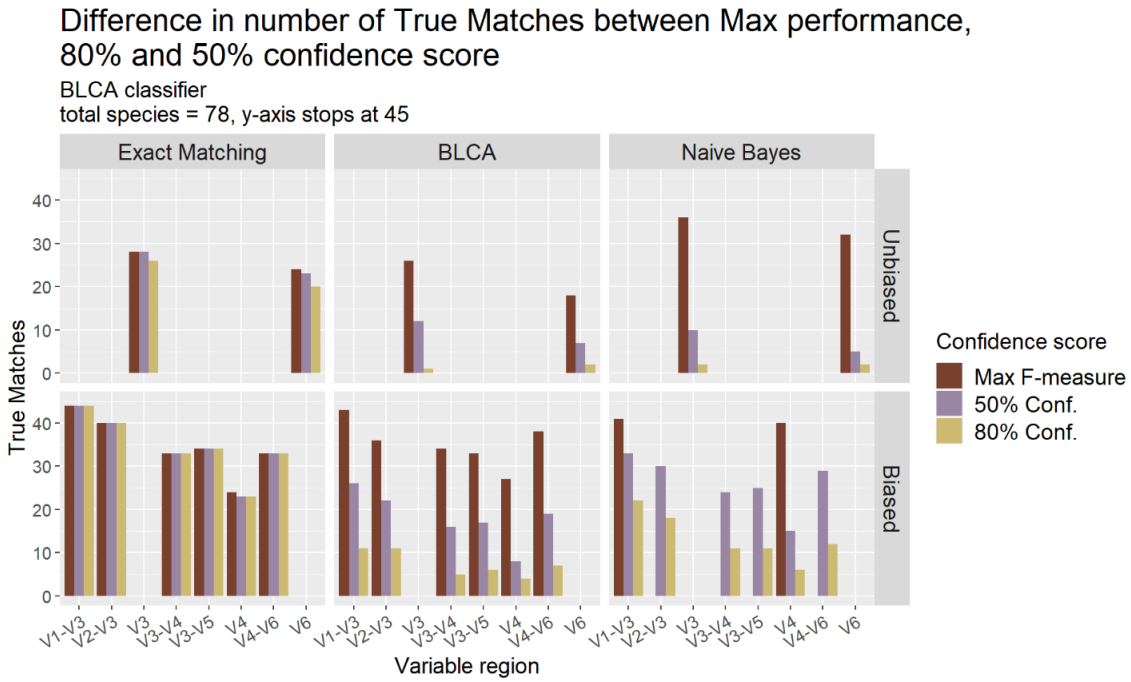

Figure 54: The number of True Matches returned by the Exact Matching classifier, compared to the classification schemes that use BLCA and Naive Bayes. The top row are the two targeted amplicons that were below the missing data threshold, and are considered unbiased results. Bottom row are the classification schemes that had exceed the missing data threshold, and are considered biased results.

While Exact Matching has very good Precision, the total number of true matches returned by classification schemes remains the goal. Figure 54 shows the classification schemes that use Naive Bayes still return more true matches than those that depend on Exact Matching, even when the identifiers are arguably not the best when compared to the V1-V3 or V4-V6 identifiers in the preceding sections.

4.6 Amplification of broad spectrum primers

4.6.1 PCR primers

Of the four primer sets designed for this study, only the primers designed for *RpoB* failed to generate an amplicon in any PCR. The degeneracy of a primer is calculated by multiplying the number of possible nucleotides at each position in the oligo. While all primers used in this study had some degeneracy, they ranged in values between 0 and 24. These are slightly degenerate primers. For the broad-spectrum primers designed to anneal to as many of the 79 species identified by Thomas-White as possible, the degeneracy was much higher. The V6 primer was of comparable degeneracy to the published primers, but V3 had a value of 768, and the calculated degeneracy of the protein encoding gene primers ranged in the thousands. The high degeneracy necessitated additional work to ensure the molarity of the reaction solutions was comparable to typical PCR conditions. It was disappointing that the RpoB PCR consistently failed, because this gene has been designated as a suitable supplement to DNA-DNA hybridization for bacterial identification(69).

4.6.2 PCR amplification

Optimization of the reaction conditions for PCR began with using DNA extracted from fecal samples, because these samples contain a higher diversity of bacteria at a much greater biomass than bacterial DNA extracted from urine samples.

Once amplicons could be reliably amplified from fecal samples, template DNA obtained from urine samples was attempted.

The V3, V6, and Ffh reactions yielded correctly sized amplicons from fecal samples. Ffh and V6 were able to yield amplicons of the predicted size (290nt and 300nt respectively) from one urine sample. The V3 primers were unsuccessful at generating a visible amplicon from urine samples. Figure 55 shows the results of the PCR.

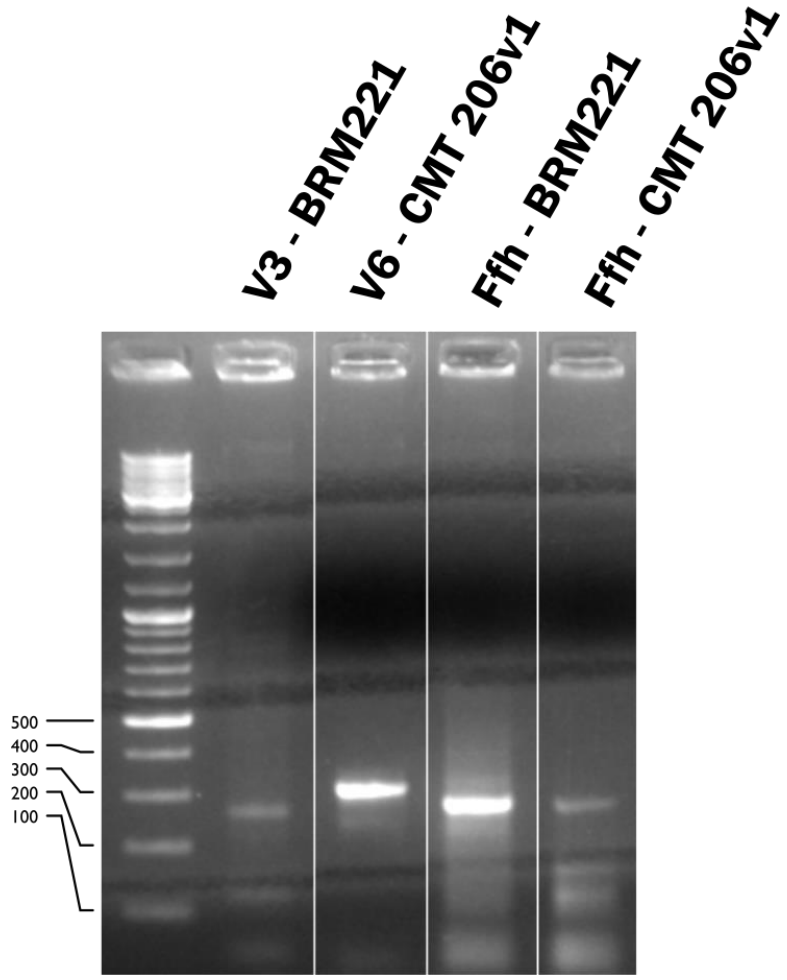

Figure 55: Amplicons of the successful primers designed in this study. V3, V6 are the primers for the 16S rRNA gene sequence. Primers are (v3_579F, v3_779R) and (v6_1183F, v6_1410R) respectively. Ffh is the primer set (541_811F, 541_995R). BRM is prokaryotic DNA extracted from a fecal sample. CMT is prokaryotic DNA extracted from a urine sample. No amplicon was successfully produced by the 16S V3 primers from urine extracted DNA.

4.7 Validation using V4 16S rRNA targeted amplicon sequencing

Figure 56 shows the results of how the three classification schemes identified the bacteria from the V4 validation set. As shown in the *in silico* results, there is a wide range in the number of species each classification scheme is predicted to correctly identify. Table 18 lists the bacteria that were not predicted to be correctly identified by any of the classification schemes that includes the V4 16S

genus	species
Actinomyces	odontolyticus, turicensis
Bifidobacterium	breve
Enterococcus	faecalis
Lactobacillus	iners, johnsonii, rhamnosus
Neisseria	perflava
Staphylococcus	epidermidis, simulans, warneri
Streptococcus	mitis, sanguinis

Table 18: The 13 species in the Thomas-White dataset that were not correctly classified by any classification scheme in silico.

rRNA gene identifier. Of these bacteria, *Streptococcus sanguinis* and *Neisseria perflava* are singletons.

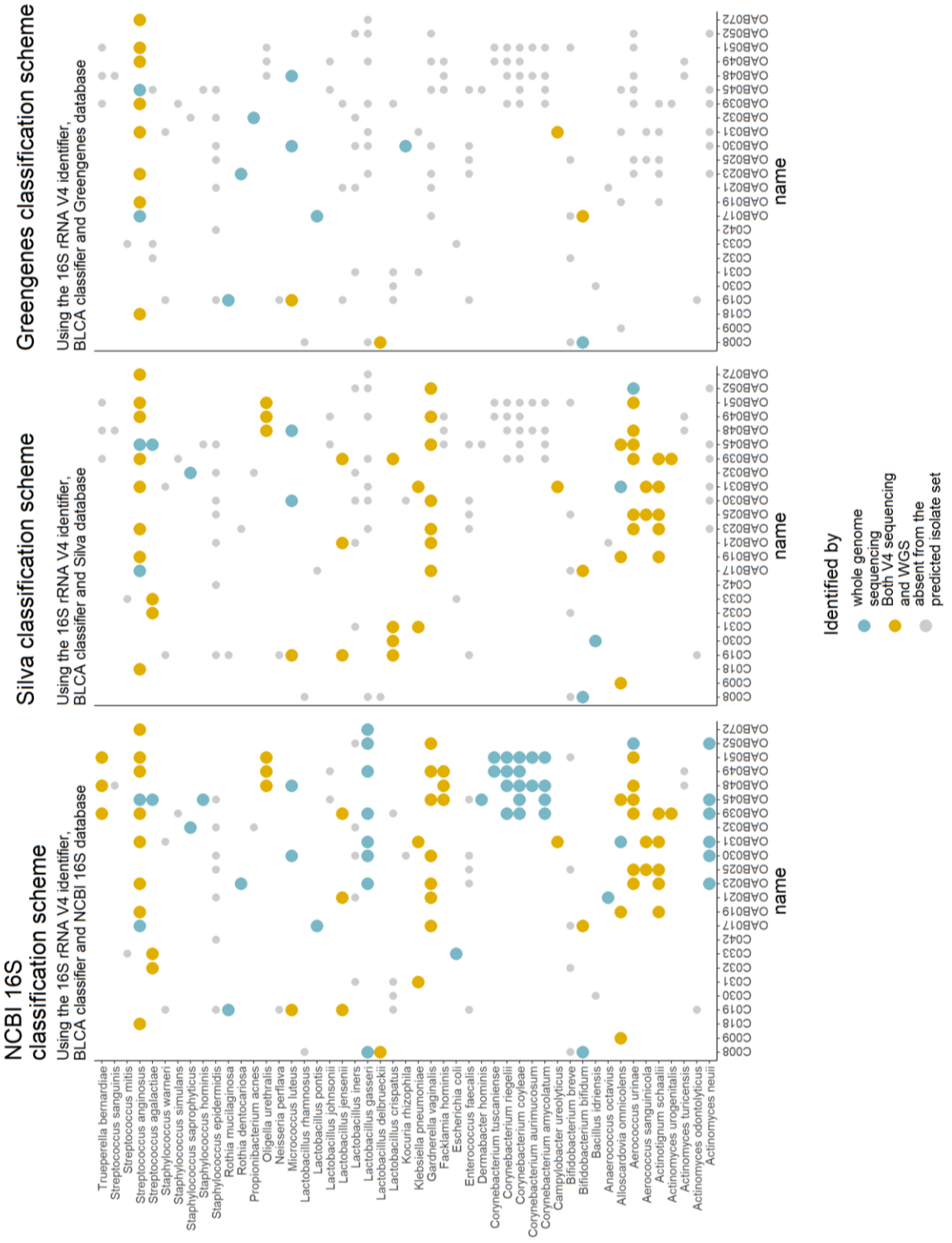

Figure 56: The identification results of each classification scheme composed of the V4 16S rRNA identifier, BLCA classifier, and the Silva, Greengenes, and NCBI 16S databases. Each panel includes the V4 validation dataset as shown in Figure 32. Gray dots represent the species/sample pair that was not included in each of the classification schemes. Blue dots represent the species/sample pair that was present in the predicted isolate set, but was not identified by the classification scheme. Yellow dots represent the species/sample pair that was successfully identified by the classification scheme.

4.7.1 Recall of the V4 validation

Figure 57 shows the number of species each classification scheme correctly identified out of the total number of predicted matches, when using confidence score values of 0, 50% and 80% as thresholds. These numbers are used in calculating the Recall of each classification scheme (see Figure 59), but it is informative to show the actual numbers of true matches and missed matches.

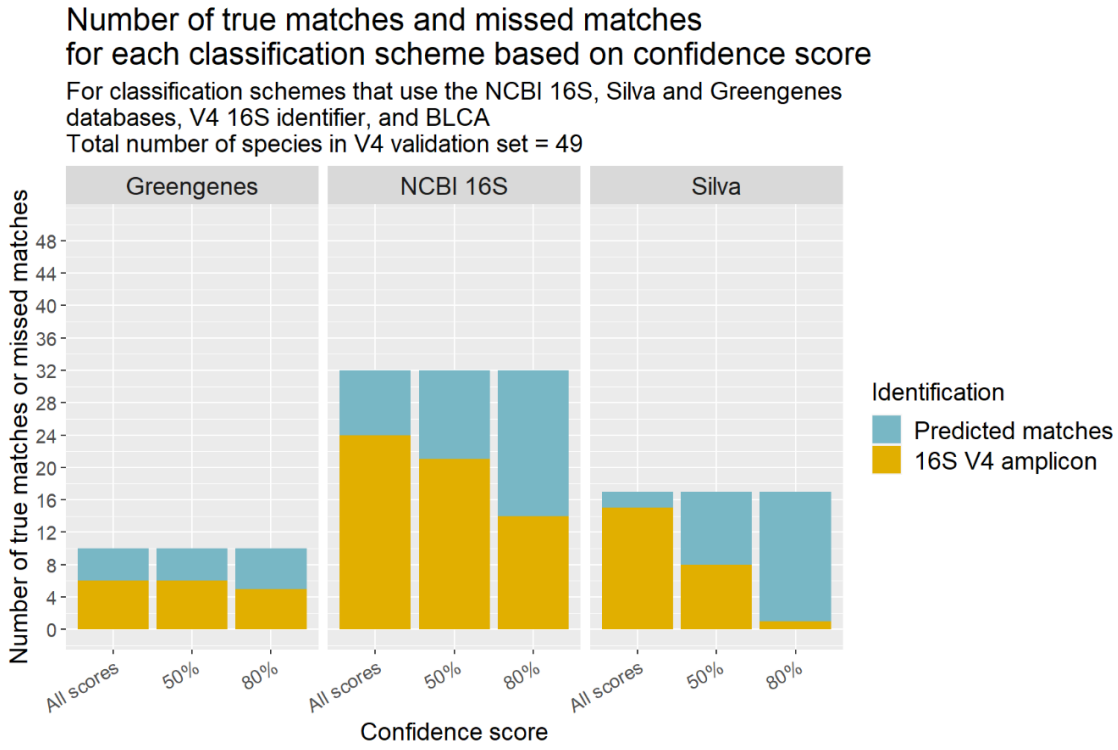

Figure 57: Summary of the data shown in Figure 56, for all predicted species sets of the classification schemes and when using a confidence score to filter the classification results. The number of species in each of the predicted matches is shown by the full height of each column. Blue segments indicate the number of species that were not identified by the classification scheme (missed matches). Yellow segments indicate the number of species that were successfully identified by the classification scheme (true matches). For each classification scheme, the results of filtering the classifications are shown for when the confidence score is ignored (all scores), and confidence scores of 50% and 80%.

The NCBI 16S classification scheme correctly identified 24 of the 32 species of its predicted matches when the confidence score is not used. This is 9 more species than the Silva classification scheme, and 18 more than when using Greengenes.

Using the default confidence score settings of 50% or 80% did not significantly change the proportion of correctly identified species of the Greengenes

classification scheme. In contrast, the NCBI and Silva classification schemes were impacted by the confidence score. The most extreme effect is shown in the Silva classification scheme, as the number of true matches is reduced to one when using an 80% confidence score.

4.7.2 Evaluations

It is possible to do some basic performance evaluation of the classification schemes, such as accuracy, precision and recall. The definitions of the cells in the confusion matrix are described in the Methods section but is also summarized here in Figure 58.

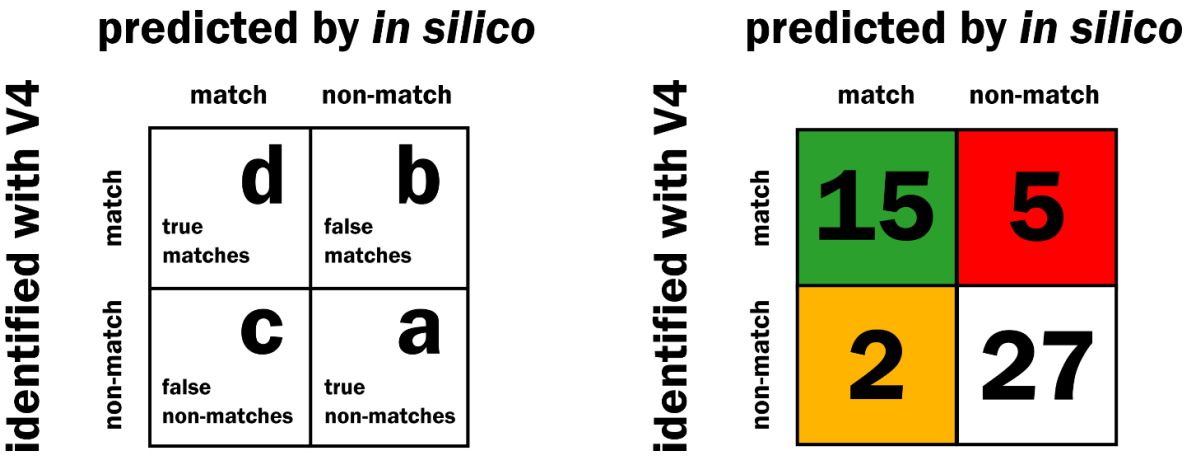

Figure 58: (Left) The confusion matrix for evaluating the results of a classification scheme on the V4 subset. Columns are the correct species identification (match) and incorrect species identification (non-match results) as predicted by the in silico methods. Rows are the match and non-match results of the classification schemes that use the V4 rRNA gene identifier. (Right) An example of how the confusion matrix is filled out for the results of the Silva classification scheme. This classification scheme identified 15 of the 17 species predicted by the in silico methods, and failed to identify 2. The scheme also correctly identified 5 of the 32 species this scheme was not predicted to identify.

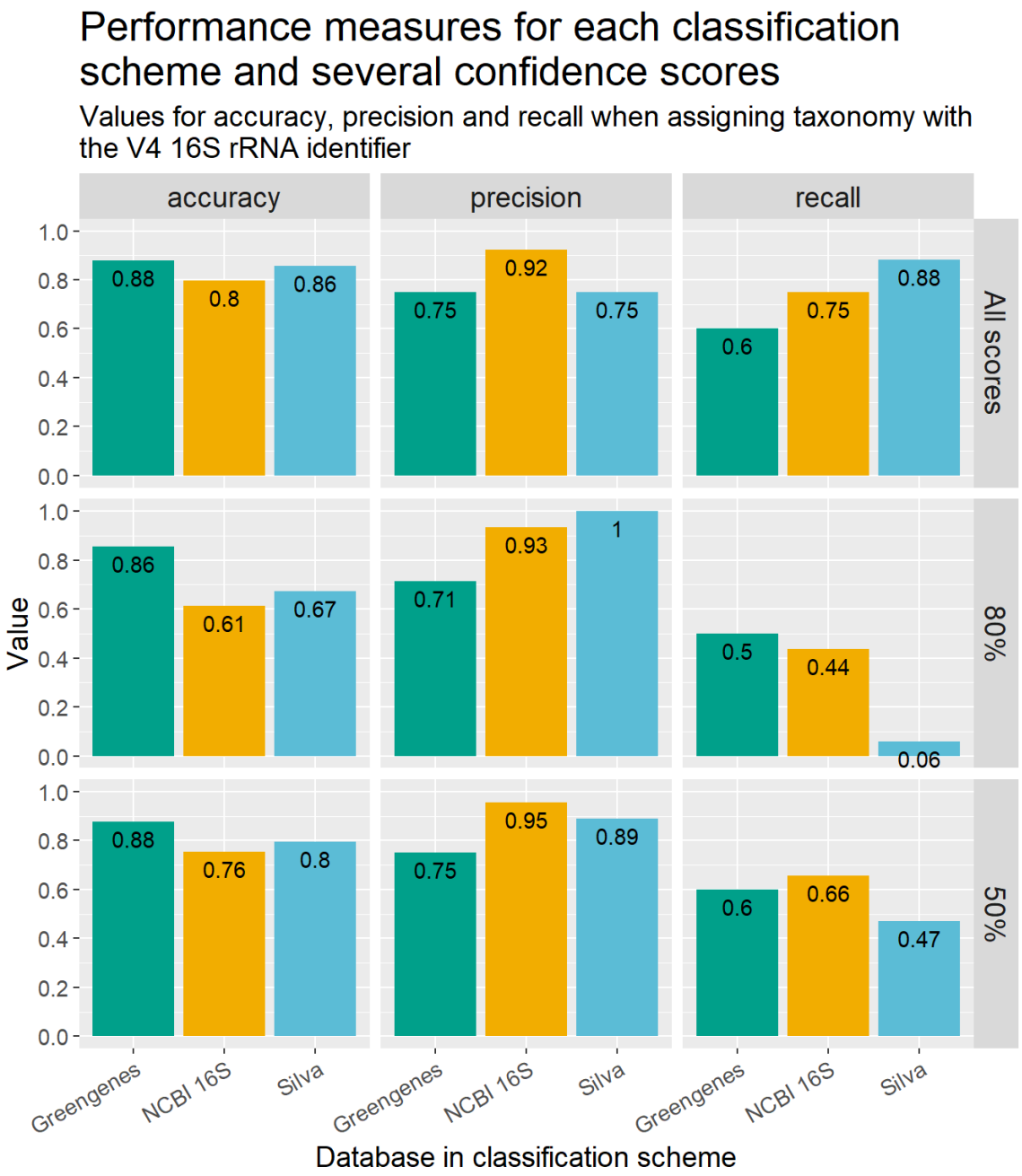

Figure 59: Performance measurements for each classification that uses the V4 rRNA identifier, BLCA classifier, and the Greengenes, Silva and NCBI 16S databases. Rows show the effect on the performance measures when filtering the results by three confidence scores (All scores, 50%, 80%). Columns are values for Accuracy, Precision and Recall.

The performance measures for each classification scheme show very different responses when applying the default confidence scores, as shown in Figure 59. All performance measurements here range from zero to one, so results will be

expressed in percentages for convenience. The Greengenes classification scheme shows very little change in the Accuracy, Precision or Recall values, regardless of whether or not a confidence score is used. While the performance measures for this classification scheme can be good, such as 88% Accuracy, it should be remembered that the predicted Matches for this scheme is only 10 species. The Silva and NCBI 16S databases have a larger number of 17 and 32 species in their respective predicted match sets.

Accuracy reflects the proportion of how many classifications are correctly placed as true matches or true non-matches out of all classifications. missed matches and false matches reduce the Accuracy of a classification scheme. The Silva and NCBI 16S databases both have a good Accuracy (80% for NCBI 16S and 86% for Silva) when ignoring the confidence score. These high Accuracy values show that the predicted Matches for each of these classification schemes are a reasonable expectation of what species can be identified when using the V4 region of the 16S rRNA gene as an identifier. When using a default confidence score of 80%, classification schemes that use the NCBI 16S and Silva database lose about 30% of their Accuracy (53% and 62% respectively). Accuracy is moderately decreased when using a 50% confidence score.

Precision measures the proportion of true matches to all predicted matches, and indicates how well a classification scheme avoids making false matches. The NCBI 16S classification scheme does very well best across all confidence score thresholds, and achieved a high value of 95% when using a 50% threshold. In contrast, the Silva classification scheme has a Precision value of 75% when ignoring confidence scores, and achieves a maximum value of 100% when using a default threshold of 80%.

Recall measures the proportion of true matches to the total possible matches, and indicates how well a classification scheme can classify all the possible matches available. In the same way that is shown in the *in silico* results, the values of Recall for the Silva classification scheme in this validation set are severely affected by the default confidence scores, and plummet to an extremely low Recall value of 6% when using a threshold of 80%. The NCBI 16S classification

scheme has a good Recall score of 75% when ignoring the confidence scores, and show a moderate decrease when using a 50% threshold, but is strongly decreased to 44% when using a threshold of 80%.

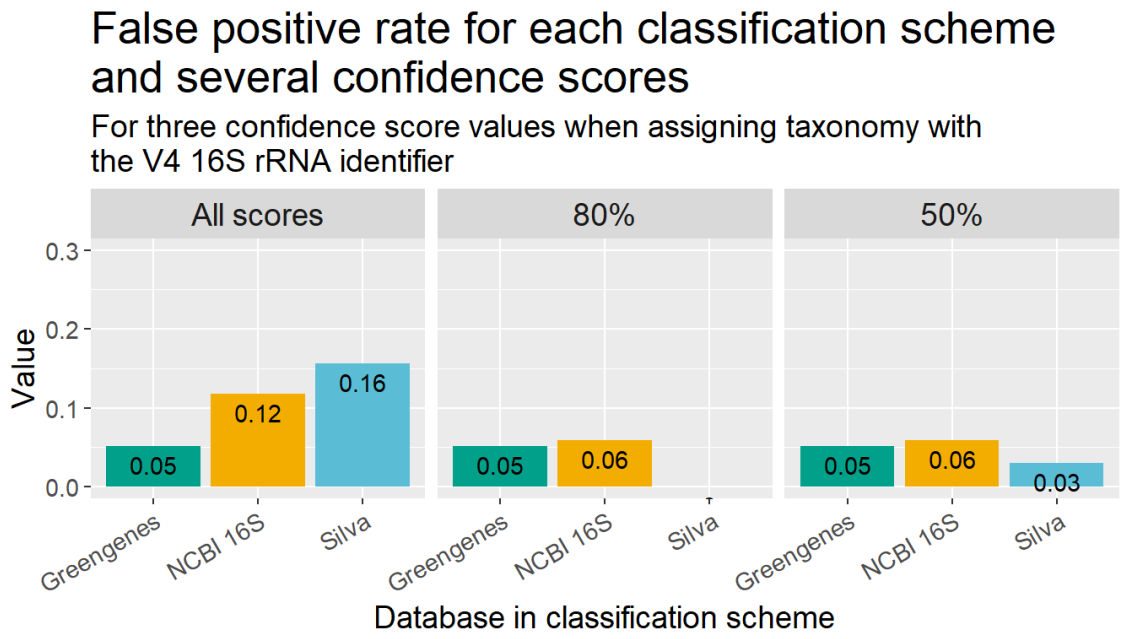

Figure 60: The False Positive Rate (FPR) for each classification that uses the V4 rRNA identifier, BLCA classifier, and the Greengenes, Silva and NCBI 16S databases. In this evaluation, the FPR indicates the number of species that were not predicted to be correctly identified by the classification scheme, but in fact were. Larger values indicate the change in proportion of true matches over all predicted matches.

Under this evaluation, the false positive rate (FPR) has a happier connotation, shown in Figure 60. Unlike a standard evaluation where a larger number of false matches is a detriment to the performance of a classification scheme, in this case it indicates the number of species outside the predicted matches that are correctly identified in addition to what is expected. However, when higher values of a confidence score are used to filter the results, it indicates the number of otherwise correct matches that are removed from the final results and classified as missed matches. This effect can be clearly seen with the Silva classification scheme at confidence score thresholds of 50% and 80%. In a similar manner as the results in Figure 59, the Greengenes classification scheme shows no effect between ignoring the confidence score and using the default thresholds.

4.7.3 Recall of predicted matches

Silva predicted matches

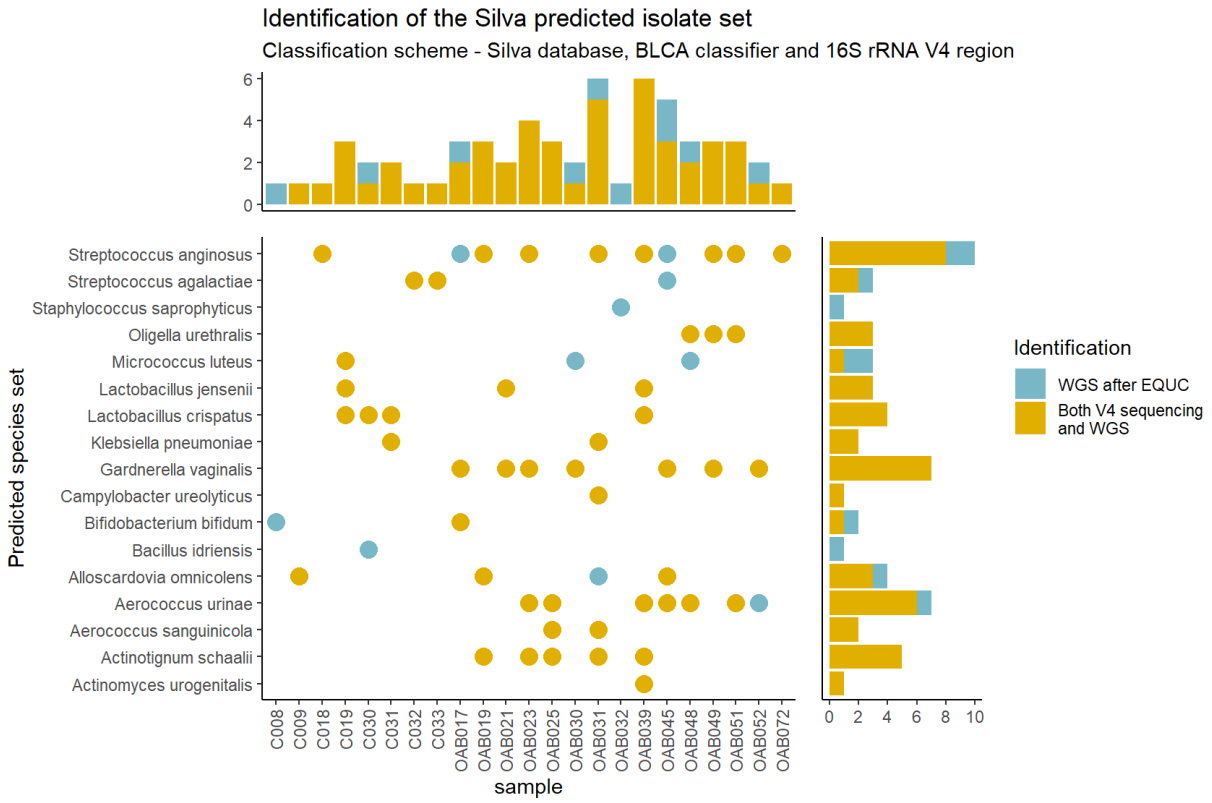

Figure 61: Predicted matches for the Silva classification scheme. This graph shows the true matches and missed matches that were present for each species/sample pair of the V4 subset that were members of the Silva predicted species set. Blue indicates species/sample pairs that were not identified by the classification scheme, and yellow indicates those that were. Sample diversity is graphed plotted on the top of the scatterplot, and species abundance is plotted on the right of the scatterplot.

Compared to the other two predicted matches, the Silva predicted matches is of moderate size, shown in Figure 61. The two species that were not classified were *S. saprophyticus* and *B. idriensis*, both of which are singletons. Two of the Lactobacillus species of interest to the urobiome, *L. jensenii* and *L. crispatus*, were present in this dataset and were correctly identified.

Greengenes predicted matches

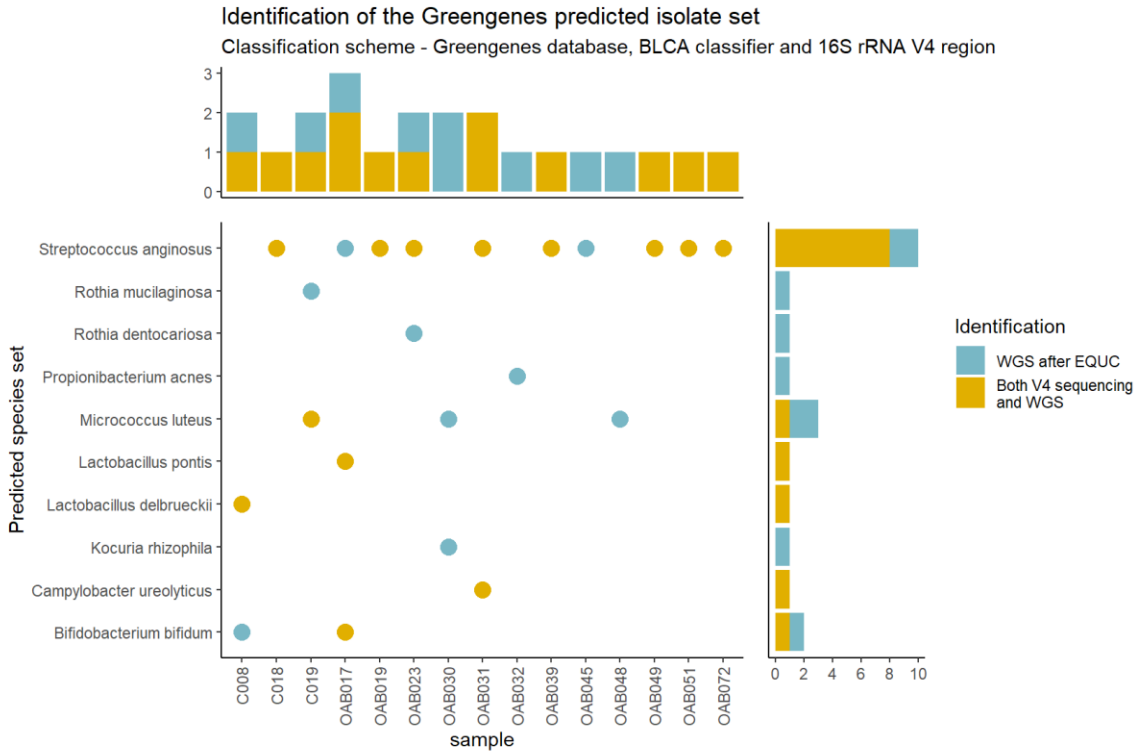

Figure 62: Predicted matches for the Greengenes classification scheme. This graph shows the true matches and missed matches that were present for each species/sample pair of the V4 subset that were members of the Greengenes predicted species set. Blue indicates species/sample pairs that were not identified by the classification scheme, and yellow indicates those that were. Sample diversity is graphed on the top of the scatterplot, and species abundance is plotted on the right of the scatterplot.

The Greengenes predicted matches is the smallest of the three datasets, shown in Figure 62, but the Greengenes classification scheme was able to identify 60% of the species in the set. There were two Lactobacillus species in this set, *L. luteus* and *L. pontis*, and both were classified correctly. One of the species missed by this classification scheme was the singleton *Kocuria rhizophila*.

NCBI 16S predicted matches

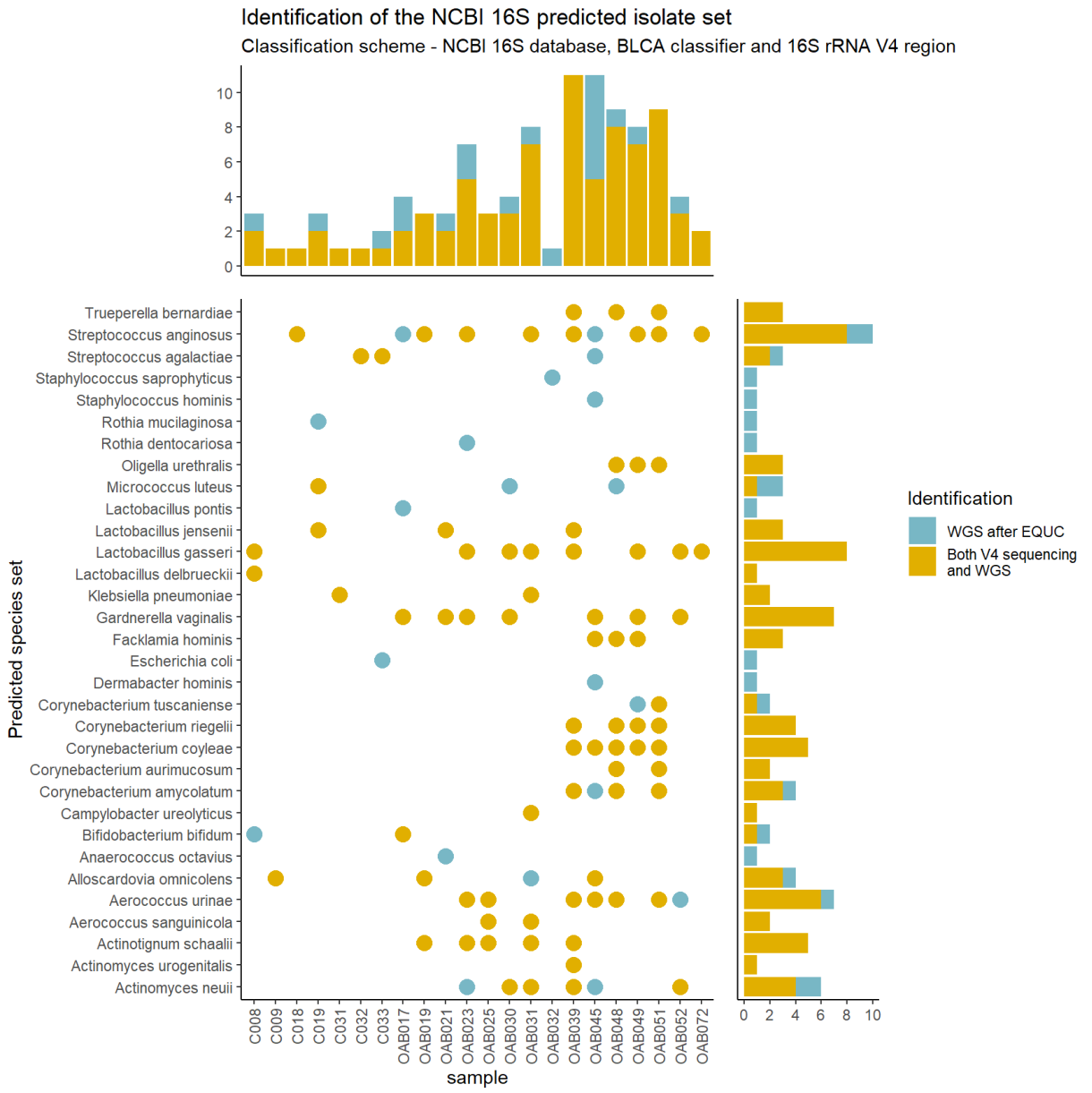

Figure 63: Predicted matches for the NCBI 16S classification scheme. This graph shows the true matches and missed matches that were present for each species/sample pair of the V4 subset that were members of the NCBI 16S predicted species set. Blue indicates species/sample pairs that were not identified by the classification scheme, and yellow indicates those that were. Sample diversity is graphed on the top of the scatterplot, and species abundance is plotted on the right of the scatterplot.

The NCBI 16S predicted matches is the largest of the three predicted datasets, shown in Figure 63, and contained the largest number of Lactobacillus species (*L.*

pontis, *L. jensenii*, *L. gasseri*, and *L. delbruecki*). Of these Lactobacillus species, the NCBI classification scheme failed to identify the singleton *L. pontis*. Also of note is this classification scheme failed to identify the common bacteria *Escherichia coli*, although this bacteria was only found in the NCBI 16S predicted matches.

5. Discussion

With this study, I attempted to determine if species level identification can be achieved by combining several databases, classifiers and subsequences of three marker genes. For the 16S rRNA gene sequence, the results show that the V2-V3 region allows the identification of the most species out of the Thomas-White dataset compared to other variable regions in the gene sequence, when used in combination with the BLCA classifier and NCBI 16S rRNA gene sequence database. When compared with the protein encoding genes Ffh and RpoB, the results show that RpoB identified the most species in the Thomas-White dataset than any other subsequence of any other marker gene, when combined with the BLCA classifier and a custom genomic database. In addition, a broad-spectrum primer set for Ffh was successfully shown to generate an amplicon from bacterial DNA extracted from a urine sample. One classification scheme was validated by comparing the number of species correctly identified *in silico* with sequencing results from *in vitro* data. The results show this classification scheme allows the identification of 74% of the predicted species. In summary, these results show that species level classification is possible for the microbiota found in the human female bladder with resources that are currently available.

5.1 Limitations

The transition from the predicted successful classification schemes and actual data generated in a laboratory remains challenging. In this section, I consider some of those challenges.

5.1.1 Classifier limitations

The BLCA and Naïve Bayes classifiers used in this study are examples of two different strategies, designed to overcome the common challenges of searching an extremely large data set in order to find matching pairs of query sequences and reference records. The limitations of a chosen classifier is further compounded by the dependence on the other two components of the classification scheme. Classifier algorithms will not perform at their best when the database is poorly constructed, or if the query sequence lacks sufficient variation between the sampled species.

The Exact Matching classifier is a stringent method of identification, and when few matches are returned there are only two strategies to consider, either accept the results or attempt to find a better identifier to compare the query and reference records(46).

One serious limitation of the Exact Match classifier is that it is very rare that sequences obtained from environmental samples are exactly the same as the sequences held in a reference database. There are two reasons for this. The first is that exact matching does not account for the variation found in an environmental sample, and will always underrepresent diversity found in the real world until it is updated. The second is that sequence errors introduced by PCR and the sequencing platform will always be present.

BLCA is an example of a classification method that uses sequence comparison by pairwise alignment. The strengths of this method are that the similarities between two DNA samples are directly compared, as found in the bacterial cell. This is the desired way to compare the characteristics of a sample to those that define a taxon, but until recent advances in computer technology it remained unattainable.

The Naive Bayes classifier is an example of a *k*-mer based classification. Its strength is that large query and reference databases can be searched for matches quickly with low computational effort, and is more robust against taxa that contain a large amount of variation. While building the corpus of *k*-mer

frequencies from a reference database can take significant computational effort, once completed the index can be used repeatedly.

The largest limitation of a *k*-mer classifier is that all identifying characteristics of a taxon are treated as a 'bag of words' model, and all positional information held in the query and reference sequences are lost. Just as two sentences can have the same words but differ in meaning, two unrelated species can have the same *k*-mer frequencies for a genetic sample but differ in the actual linear DNA sequence.

A second limitation is how the conditional probabilities are calculated. Each successive step of calculating the frequencies depends on counting how many times a word occurs in the corpus or taxon. When classifying a taxon at the genus level taxon, there are usually many species of each genera included in a database, and each of those species can have more than one occurrence of a word because the sequences are long. The difference is when attempting to classify species. Frequently there are very few representative sequences for a species in the databases described here. Because the conditional probabilities are calculated using very few occurrences, small variations in the query sequence result in drastically different classification. For these reasons, and despite the demonstrated high performance of the Naïve Bayes classifier for identifying species in this study, this classifier is not recommended for species level identification until there are more representatives for each species available in databases.

Finally, the confidence scores calculated for both classifiers can be confounded by the formatting of the reference database records. Candidate records retrieved from the reference database may have different information contained in the names of the records, even though the records contain the exact same sequence. Neither classifier is equipped to make the distinction between identical sequences contained in the records but have different labels, and this results in a low confidence score. I do not consider this to be a problem of the classifier *per se*, because conserved regions of gene sequences should be very similar by definition.

However, the multiple copies of the 16S rRNA gene sequence can affect the classification of easily identifiable bacterial species such as *E. coli*.

A simple example is demonstrated using mock databases containing only the 16S gene copies found in *E. coli* and applying the BLCA classifier (although both BLCA and Naïve Bayes calculate the confidence score in the same manner). The *E. coli* genome has 7 *rrn* copies, which are identical except for one nucleotide in the V1 region, and one 12 nucleotide insertion on a single gene copy as can be seen in Figure 64.

Figure 64: Multisequence alignment of the seven 16S rRNA gene copies found in E. coli. The only differences between the sequences are the 13 nucleotides shown in this part of the alignment. Visualization done with UGENE.

BLCA is then run under two conditions, using an identical query sequence found in all seven *rrn* records. In the first condition, the records are labeled with the same ID, while the second condition uses a mock database where that all *rrn* records are labeled with a unique ID.

In the first condition, BLCA correctly identifies the query sequence as E. coli, and the confidence scores for all ranks are an unambiguous 100% as shown in Table 19.

E.Coli *rrn* copy 1, IDs identical

level	taxon	confidence
superkingdom	Bacteria	100.0
phylum	Proteobacteria	100.0
class	Gammaproteobacter	100.0
order	Enterobacteriales	100.0
family	Enterobacteriaceae	100.0
genus	Escherichia-Shigella	100.0
species	Escherichia coli	100.0

Table 19: Classification result for one of the seven 16S rRNA gene sequence as described in the text. Because the labels used for each of the records in the reference database are the same, the confidence scores for each taxonomic rank is 100%.

When BLCA performs bootstrapping to calculate the confidence score, ties are broken by random choice. Faced with seven records containing the same

E.Coli *rrn* copy 1, IDs unique

level	taxon	confidence
superkingdom	Bacteria	14.28
phylum	Proteobacteria	14.28
class	Gammaproteobacteria	14.28
order	Enterobacteriales	14.28
family	Enterobacteriaceae	14.28
genus	Escherichia-Shigella	14.28
species	Escherichia coli	14.28

Table 20: Classification result for one of the seven 16S rRNA gene sequence as described in the text. Because the labels used for each of the records in the reference database are different, the confidence scores for each taxonomic rank reflects how often the tiebreaking step is selecting one of the seven equally probable candidates. The resulting confidence score is always 1 out of 7, or 14%.

sequences but different unique identifiers, the assigned taxonomy is given a confidence score equal to $\frac{1}{7}$, or ~14%.

To reiterate, BLCA has correctly matched all seven reference sequences to the query sequence. Moreover, these are *exact* matches between query and reference. However, due to the way these classification algorithms calculate their confidence scores, the results are likely viewed with suspicion.

5.1.2 Database limitations

A popular Naive Bayes classifier and reference database combination is the RDP classifier and associated RDP database. The RDP database does not include the species level taxon, and is a good example of how the information in database records is critical for species level taxon assignment. The curation, frequency of update and quality of information is an important consideration in choosing which database to use.

Even then, none of the databases used in this study could be considered a perfect match of the requirements. The Greengenes database has not been updated since 2014 [https://greengenes.secondgenome.com/?prefix=downloads/greengenes_database/gg_13_5/], and taxonomic practices have changed since then. The Silva database includes many records sourced from metagenomic sequencing of environmental samples that were not identified by polyphasic taxonomy. During the course of this study, species-level taxonomy was frequently observed to be labeled “uncultured bacterium”. NCBI provided very little documentation describing the characteristics of either database.

5.1.3 Targeted amplicon limitation

There are several practical limitations of the targeted amplicon used, and these can be grouped under sequencing technology, the amount of phylogenetic resolution provided by the targeted amplicon, and the reaction conditions of PCR.

Resolution

The 16S gene sequence was one of the first genes identified by Woese and Fox as ideal for phylogenetic study(24). Ever since their initial publication, this gene has become a cornerstone in phylogenetic research. The compromises of using a much smaller variable region for classification have been explained. An additional drawback of the bacteria in the Thomas-White dataset is there is an average of 5 copies of the 16S rRNA gene present (Figure 65). These multiple copies confound the assessment of population diversity.

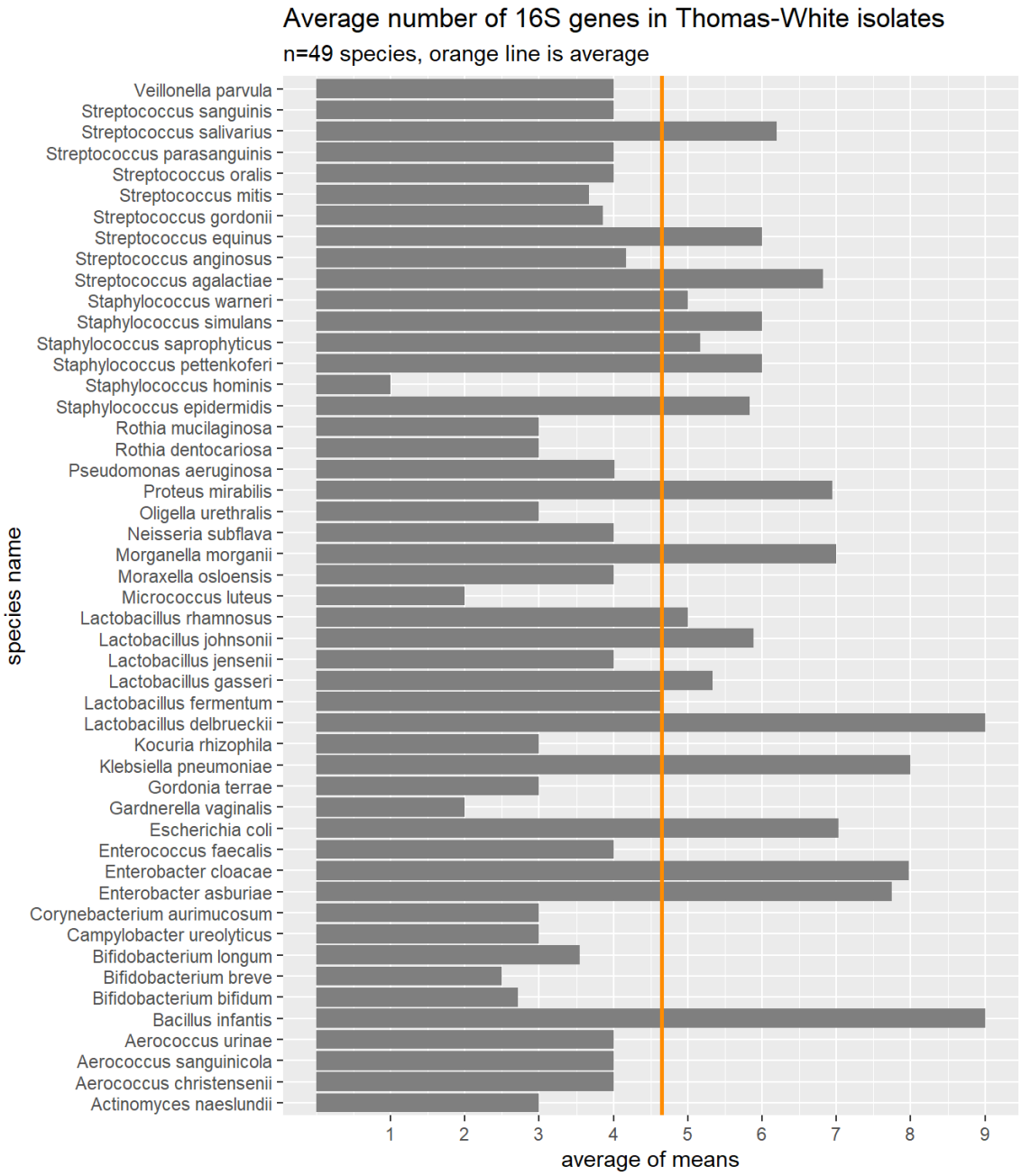

Figure 65: Number of 16S rRNA genes found in the species of the Thomas-White dataset. Data taken from the rrnDB website. Gene copy information was not available for all species.

One possible reason that has been presented for a bacterium containing multiple copies of the gene is a shorter response time in which to take advantage of available resources(70).

Sequencing

Affordable sequencing of large scale data is presently done on the Illumina MiSeq platform, currently limited to sequencing reads of 300 nucleotides. Choosing which 300 nucleotides of the 16S gene to sequence is challenging, and there is a large body of literature that describes the benefits and limitations of all parts of the gene sequence. Until the time where third generation sequencing is widespread, and full length gene sequences can be achieved on a large scale, the choice of which 300 nucleotides of the 16S gene is suitable for the research question will remain a significant part of the experimental design.

PCR

The amplicons yielded from a PCR are generated by priming the reaction with specially designed oligonucleotides. The challenge of PCR primer design is to identify a sequence of nucleotides that will anneal to only one location on the template DNA. Finding suitable annealing sites that flank the variable region of interest can be easy enough to do for one species, but becomes very difficult when considering the gene sequences of many species. Clearly, this design is predicated on having some knowledge of what species to expect in a sample, and for this reason the V4 region of the 16S gene has been used as the way to initially explore the microbiota diversity of a sample to the Family and Genus level taxons(14).

Designing a primer set that will anneal to the target DNA sequence found in more than one species is accomplished by synthesizing a specific oligonucleotide for the specific DNA template sequence found in each species. The pool of oligonucleotides are collectively called a degenerate primer. All the primers in this study are at least slightly degenerate. Graspeutner's(71) primers are a modification of Caporaso's(72) primers towards higher degeneracy, and the V3 and V6 primers designed in this study are even more degenerate. The drawback of this strategy is that non-specific binding becomes more prevalent, and the chemistry and thermocycling settings must by optimized. In the event where no suitable degenerate primer can be designed for all the expected species, designing

degenerate primers for subsets of the expected species can be done, but increases the amount of work downstream.

Finally, the processing work of error correction and accounting for sequence variants must be done on the sequencing results before classification is performed. Targeted amplicon sequencing generates a large number of overlapping reads and provides the data for statistical methods to correct for errors introduced by the polymerase enzyme. After error correction, similar reads are aggregated into operational taxonomic units or amplicon sequence variants. The last step is to attempt to merge those reads that can be shown to be complementary before attempting to classify. In all the steps here, reads that overlap by a large margin are more desirable than otherwise, due to the benefit that denoising and dereplicating steps gain from the increased information provided. All sequencing reads require low quality base calls at the end of the read to be discarded, and if the amplicons generated by PCR are larger than the 300 nucleotides that can be sequenced, either the overlapping nucleotides are removed through trimming or the nucleotides of the interior of the amplicon are never sequenced. In both cases, valuable phylogenetic information is lost.

5.2 Optimal marker gene

Based on the Recall results, and with respect to the species of bacteria found in the bladder, the protein encoding genes *RpoB* and *Ffh* return more true matches than the 16S rRNA gene. These genes appear to be well suited for the bladder microbiome. However, these results are based on the evaluation of one moderately sized custom-built database. In comparison, the 16S rRNA gene has a long history of use as a phylogenetic marker, with the use of the ribosomal RNA sequence dating from the 1970s(24). This gene has also been used in many environmental studies covering very different habitats, from seawater(73) to mussels living on whale carcasses(74) and of course the human microbiome. The diversity found in the collected and annotated 16S rRNA sequences to date certainly aids in the classification of bacteria found in novel environments.

More study is needed to determine if the protein encoding marker genes listed in this study have the same representation of diverse environments and species of bacteria. It can be expected that because this collection of genes were only just described in 2006(75), they would be slow to be incorporated in phylogenetic studies when there is a much more established alternative. It can be argued that as the interactions of human microbiota are more carefully studied, and the demand for higher taxonomic resolution becomes more frequent, the availability of reliable, additional marker genes will become more important. One way to achieve this goal is to include the *Ffh* and *RpoB* genes in the same workflow as the 16S rRNA gene.

5.3 Optimal variable region

Primers

The amplicons used in this study are a best case scenario, but allow the optimization of parameters that are under far more control than variables presented in a bacterial population sampled from the bladder. Designing primers that are predicted to anneal to as many of the bacterial species in the Thomas-White dataset as possible has been one of the goals of this study, but it is understood that there could be many more unknown species in the bladder. This study has shown that while including the V1 to V3 region of the 16S gene in a targeted amplicon has the highest success at identifying the species in the Thomas-White set, attempting to produce a working broad-spectrum primer set that anneals anywhere within that span of the 16S gene sequence has failed.

I was successful in designing a broad-spectrum primer set for the protein encoding genes *Ffh* and *RpoB*, and these outperformed any primer set from the 16S gene *in silico*. PCR using *Ffh* primers was successful in yielding amplicons from gut and urine samples, but the *RpoB* PCR failed to yield visible amplicons. The primers for *Ffh* show promise, and (I think) further optimization is warranted for inclusion in standard sequencing workflow.

Failure to find a solution to the degenerate primer design problem for a data set is not uncommon, and a reasonable next step is *stratification*, subsetting the original set of species into several smaller groups. These groups are determined by estimating the phylogenetic relatedness and splitting on common ancestors that bridge phylogenetic groups. In this sense, the primers found in the literature search are stratified primers. While they are not the optimal solution, they are a practical one. Further work is warranted on the V1-V3 primer set described by Komesu(76) and the V2-V3 described by Bukin(77).

Variable regions

For the classification scheme that used the variable regions from *Ffh*, *RpoB* and the 16S rRNA gene as identifiers, custom-built database, and the BLCA classifier, the variable regions with the highest Recall were clear. *RpoB* V1 was the best overall, followed by *Ffh* V1-V2, and then the V2-V3, and V1-V3 region of the 16S rRNA gene. In general, amplicons that span more than one variable region have a higher Recall and perform better than those that contain single variable regions.

A consistently poor Recall value was observed for the V4 region of the 16S rRNA gene, and was second only to the V6 region in terms of low Recall values. It appears that this variable region is not suited for assigning taxonomy to the bacteria of the bladder microbiome. However, amplicons that include variable regions adjacent to the V4 region do improve both the performance and Recall of those classification schemes. One possible explanation is that there is more phylogenetic information content for the classification scheme to work with.

For the remaining classification schemes (excluding those that used the Exact Matching classifier), the results do not show a clear winner. The classification schemes that use the NCBI 16S database have the best Recall when they use the V1-V3 region of the 16S rRNA gene, followed by the V2-V3, regardless of the classifier used. Classification schemes that use the Silva database have the best Recall when using the V3-V5 region, followed by the V3-V4 and then the V4-V6. These highest Recall values occurred when using the Naive Bayes classifier.

The reason why these identifiers yielded different results is unclear. It is possible the lengths of the amplicons has some effect. On average, the amplicons with the highest Recall were 477 nucleotides long, compared to 223 nucleotides for the amplicons that contain only one variable region. However, the longer amplicons consistently had higher Recall values than the shorter amplicons, and amplicon length was inconsistent with the results across the classification schemes. While it was clear that the variable regions do affect the Recall of the classification schemes, the differences in Recall do not appear to be fully explained by the identifiers.

5.4 Optimal classifier

When comparing the BLCA classifier against the Naive Bayes classifier, both algorithms appear to do equally well. As an example, the results of the classifiers when using the NCBI 16S database and the V1-V3 region of the 16S rRNA gene are equally matched. However, there was a lot of preparation that went into the identifiers and databases before the classifiers were used, and this preparation is the most likely reason for the comparable performance and Recall of these classification schemes.

The classification schemes that used the Exact Matching classifier were hindered by the requirement that there be no ambiguous nucleotides in the query sequence. For the two unbiased classification schemes that could be compared to BLCA and Naive Bayes, Exact Matching did no better than either of them. As a practical option for species level identification, this classifier is not suitable. On the other hand, this classifier is *very* precise, and in the event that high precision is needed, this classifier is well suited for the task.

The choice between using the Naive Bayes classifier or BLCA for species level identification is partly based on the necessary preparation, and partly based on conceptual reasons. Training a very large database for use with Naive Bayes, such as the genomic database used in this study, was prohibitive. As reference databases will only get larger as new species are discovered and characterized,

the computational effort and time required to transform DNA sequences into the conditional frequencies of 8-mers found in the taxa of the corpus will only get larger. In contrast, BLCA has the advantage of being able to use any database made of FASTA-formatted records. Naive Bayes was designed to address the computer hardware limitations of the time and circumvent the problem of pairwise alignment. However, much of the work done in this study using BLCA was done on a modern laptop computer, and the workflow was later moved to OHSU's Advanced Computing Cluster only for convenience. For these reasons, and the conceptual limitations of Naive Bayes outlined in section 5.1, BLCA is best suited for species level identification.

5.5 Optimal database and effects

Effects

The choice of database has the largest affect on the performance and Recall of the classification schemes in this study. Besides the Greengenes database, the Silva database shows a marked influence when ignoring the confidence scores, and strongly reduces the performance and Recall when using either default threshold. Why this effect is so predominant in the Silva database is an obvious question.

One explanation is that the Silva database contains a large number of ambiguous species-level labels. While the Silva database contains records of the 16S rRNA gene for all species in the Thomas-White dataset, it also contains a large number of records that will be called "ambiguous" species. For example, searching the SILVA_132_SSURef_Nr99 database for the taxonomic string "Lactobacillus;uncultured bacterium" (representing the genus and species ranks) using the command-line program *grep* returned 1499 records. A search using the string "Lactobacillus delbrueckii" returned 109 records.

During evaluation of the classification schemes using the Silva database, the taxonomic label of any record pair assigned as a match was evaluated as a false match if the genus and species labels were not identical. Examples are record pairs like "Lactobacillus iners:Lactobacillus delbrueckii" and "Escherichia

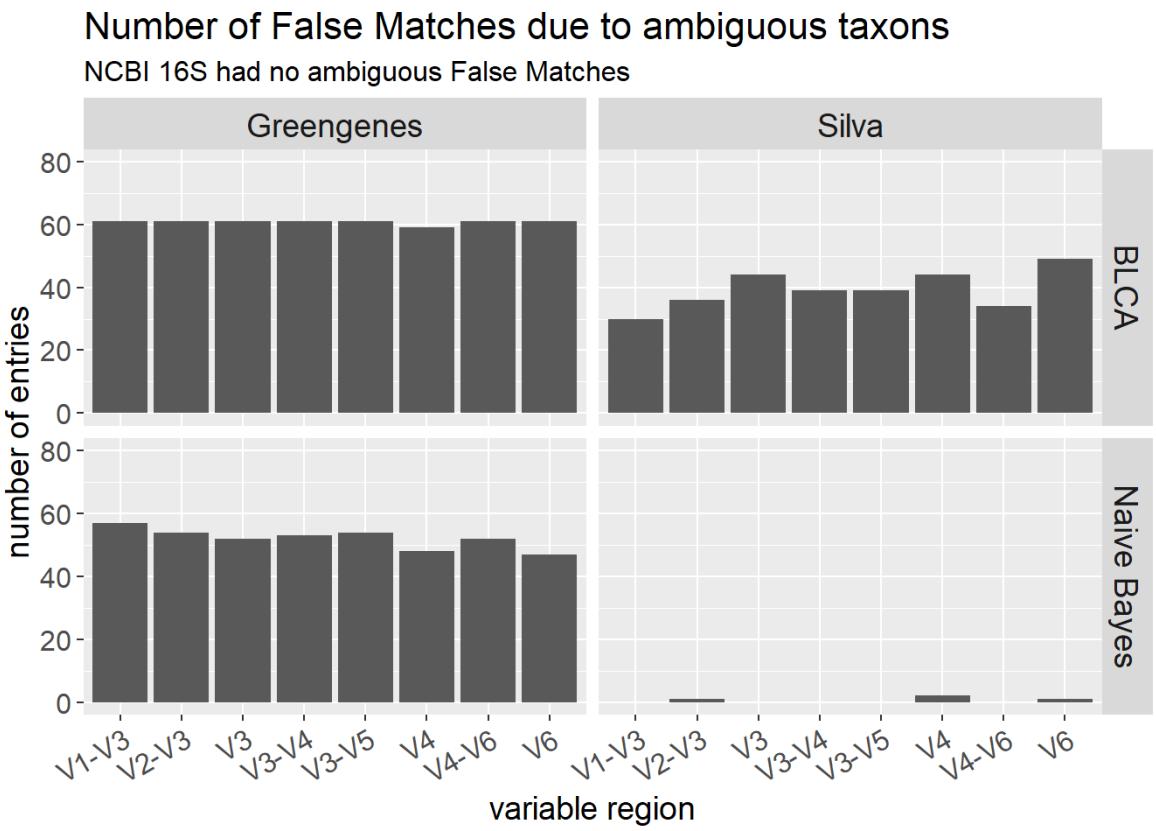

Figure 66: Number of False Matches due to ambiguous species in the Silva database. The high number of False Matches in the Greengenes database is due to the lack of a large number of species-level labels. The Silva database was trained for species level taxonomy for use with the Naïve Bayes classifier, and the small number of ambiguous False Matches is most likely due to the grouping of ambiguous taxa into separate categories when calculating the k-mer frequencies.

coli:uncultured bacterium". Graphing the number of ambiguous species from all classification schemes that used the 16S rRNA identifier is shown in Figure 66.

As expected, the Greengenes database has many false matches due to ambiguous species because of the lack of annotation for that taxonomic rank. The surprising result is the number of false matches when using the Silva database. There are far more false matches due to ambiguous species in the classification schemes that use the BLCA classifier compared to the schemes that use Naive Bayes. A further

surprise is that the NCBI 16S database had zero false matches due to ambiguous species.

If there are more than one candidate records with equally high posterior probabilities (in the event that two different records have identical sequences, as described in section 5.1), BLCA will break the tie by randomly choosing one of the equally possible candidates. It is possible that the large number of ambiguous species have highly similar sequences to those of the identifiers, and BLCA may assign a match to the identifier and reference record with an ambiguous species, generating a false match. This may explain the lower Recall and performance of

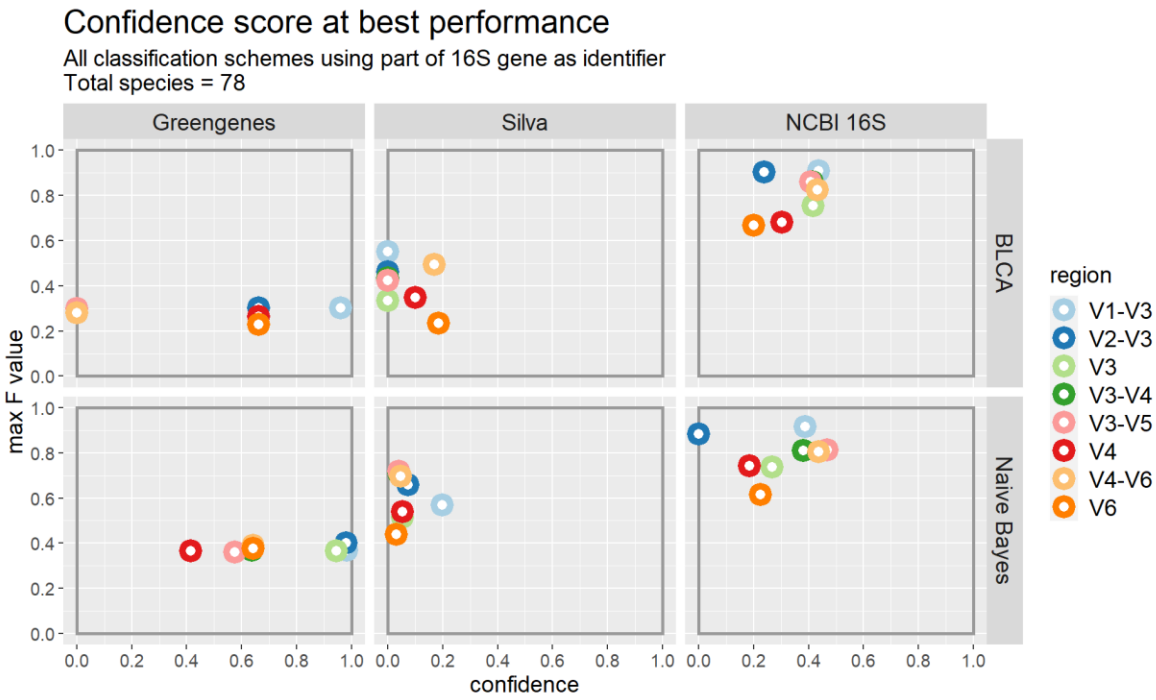

Figure 67: Confidence scores that yield the highest performance for each classification scheme. Many of the confidence scores are roughly 60% or less, but the classification schemes that use the Silva database are no more than 20%.

those classification schemes compared to ones that use the NCBI 16S database. However, this reasoning does not explain the lower Recall and performance for the classification schemes that use Naive Bayes and the Silva database.

Figure 67 shows the confidence scores at which the maximum F-measure value was achieved. A notable character associated with the Silva database is the uniformly low value of the confidence scores for all classification schemes. This

may explain the drastic reduction of the performance and Recall when using one of the default thresholds. Any confidence score larger than about 20% will begin to reassign true matches as missed matches and reduce both the values of Recall and the F-measure.

Optimal database

The best overall classification scheme uses the custom genomic database. This is not a surprising result, because each record contains all the marker genes reviewed in this study and all but one species of the Thomas-White dataset were represented. However, it is limited to only some 200,000 records. At a total file size of 6Gb, it is a much larger file than the other databases which contain more diversity.

The number of bacterial databases dedicated to *Ffh* and *Rpob* is scant. A literature search for a dedicated *RpoB* database yielded a general-purpose database Polbase(78), which collects existing information from public resources on RNA polymerases of all life forms. A database for *Ffh* exists(79), but the website has not been updated since 2010 [https://rth.dk/resources/rnp/SRPDB/srprna.html]. It would be worthwhile to build a large scale database of the *Ffh* gene sequence. Until a large and diverse *Ffh* database is constructed, it is recommended that the NCBI genomic database be used with the *Ffh* marker gene. As the information on Polbase is not specific to bacteria, it is also recommended that the NCBI genomic database be used for *RpoB*.

Classification schemes that included the NCBI 16S database performed the best with the 16S gene amplicons. As this database is curated and active, and it is recommended that it be used with both the Naive Bayes and BLCA classifiers.

5.6 Validation of predicted results

These results show that the predicted *in silico* outcomes are a good approximation for how a classification scheme will perform *in vitro*, at least for the schemes that use the V4 region of the 16S rRNA gene sequence as an

identifier. These results also show that it is possible to reliably classify targeted amplicons from an environmental sample down to the species level using existing classification algorithms, databases, and variable regions of the 16S rRNA gene.

The majority of the classification schemes' predicted matches were identified *in vitro*. While these results only pertain to using the V4 region of the 16S gene as an identifier and the BLCA classifier, they are still a promising result. While the *in silico* results show that the V4 region is not the best identifier to use in the context of the bladder microbiome, the knowledge that a majority of the identifications produced reflect reality is encouraging. It can be expected that the alternate identifiers covered in this study, such as the V1-V3 region of the 16S rRNA gene or the V1-V2 region of *Ffh*, would have similar outcomes.

As shown by the *in silico* results, the use of the 50% or 80% default confidence scores does not increase the Recall of the classification schemes, but Precision is improved. However, the effect of using one of the default confidence scores on the number of true matches on the Silva classification scheme can easily be described as extreme.

The best application of the confidence score is to exclude those classifications in which the bootstrapping values are low due to the low similarity of the query to the reference records. Finding the optimal confidence score that minimizes the missed and false matches while maximizing the number of true matches can be accomplished, as long as there is some comparative data set to tune the parameters. One example is to include a mock bacterial community in the experiment design. As the inclusion of a mock community is recommended for identifying and removing results due to bacterial contamination(80) it can also be used to determine the optimal confidence score. In the absence of a comparative data set, these results show that the default settings of 50% or 80% are too high, and it is recommended that the confidence score be ignored.

5.7 Determining the optimal confidence score

The most common classification scheme for identifying bacteria uses the Silva database, the V4 region, and Naive Bayes with a default confidence score of 80%. This study has shown that this default does not result in a high classification scheme performance, and therefore does not identify the highest number of true matches for further analysis. I now present a method to determine the optimal confidence score. These steps are quickly done by consulting a table of calculated values, but are shown here visually for simplicity.

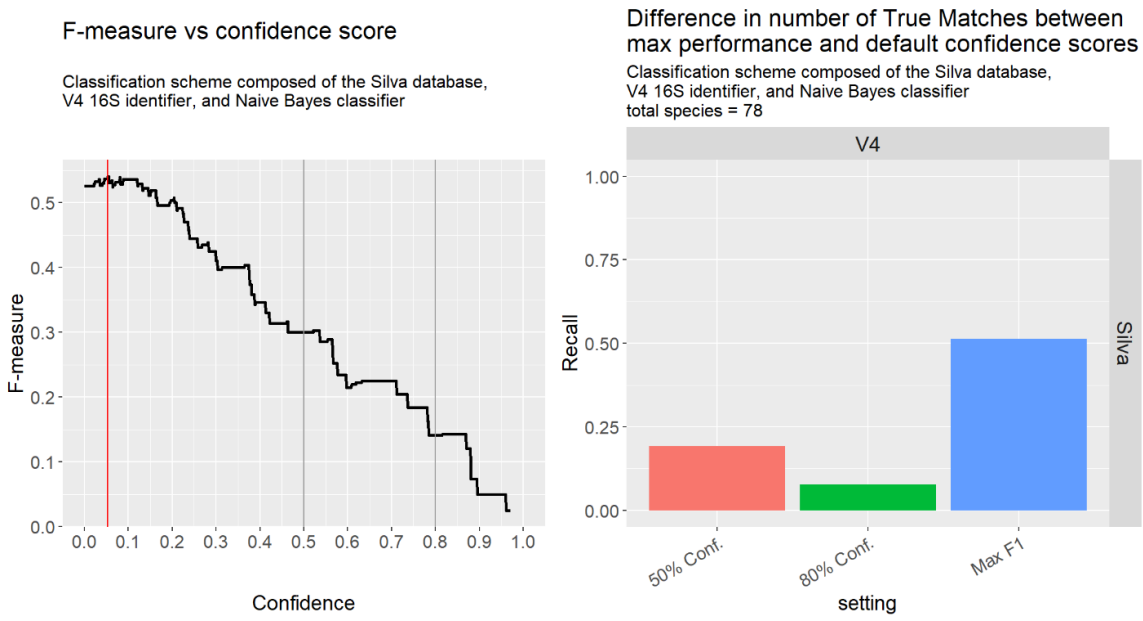

Figure 68: (Left) Finding the optimal confidence score by graphing the F-measure values for all confidence scores. Example uses the common classification scheme of the Naïve Bayes classifier, V4 region of the 16S gene, and the Silva database. The highest F-measure value is achieved when using a confidence score of 5.3%. This value yields the most true matches (right), and minimized the number of false matches and missed matches. Grey lines indicate the default 50% and 80% confidence score thresholds.

Calculating the F-measure for all possible confidence score values, such as ranging over the confidence scores from zero to 1 in increments of .01 as shown in Figure 68, easily shows that the default confidence scores of 50 and 80 yield much smaller numbers of true matches.

6. Conclusion

This study presents *in silico* and *in vitro* evidence that species level taxon assignment is possible with the current resources available. The application of these resources for future work is clear. For the research community interested in urinary tract infections, urinary health, and the entangled vaginal microbiome, the focus has moved past information that stops at the genus level.

One application of species level resolution is developing therapeutic treatments that do not involve antibiotics. Of all patients that were prescribed antibiotics that were not hospitalized (ambulatory care), 9% of those cases were for a urinary tract infection(82). The high occurrence of this disease makes it a likely arena for the emergence of antibiotic resistance. Treating recurrent UTI through the antimicrobial competition and interference provided by probiotic inoculation has been shown to be an effective countermeasure(38).

Understanding the interactions between species in the bladder microbiota is also of importance. *Lactobacillus* is a dominant genus in both the bladder and vagina. However, the species of this genus that are found in these microbiota interact in different ways. The proportion of *Lactobacillus* in the bladder and vagina microbiomes has been found to decrease with age(83), but two species in particular are of interest. *L. gasseri* and *L. crispatus* secrete the antimicrobial metabolites lactic acid and hydrogen peroxide, and are responsible for maintaining a low pH environment(84). In a study of urgency urinary incontinence, 91 genera of bacteria were found only in the conditions of symptomatic UUI, five of which have previously been implicated in UUI(6). Identifying the individual species of those genera will advance the understanding of the role of those bacteria in UUI.

In each of the examples above, determining which of the species in the identified genera contribute to dysbiosis, and which are merely commensal, can only be done with methods that are able to identify species. Hopefully, the results of this study will have application to those efforts.

7. References

1. Lederberg J, McCray AT. 'Ome Sweet 'Omics-- A Genealogical Treasury of Words. :2.

2. The Human Microbiome Project Consortium, Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, et al. Structure, function and diversity of the healthy human microbiome. Nature. 2012 Jun;486(7402):207–14.

3. Thomas-White K, Brady M, Wolfe AJ, Mueller ER. The bladder is not sterile: History and current discoveries on the urinary microbiome. Curr Bladder Dysfunct Rep. 2016 Mar;11(1):18–24.

4. 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 1;50(4):1376–83.

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

6. Karstens L, Asquith M, Davin S, Stauffer P, Fair D, Gregory WT, et al. Does the Urinary Microbiome Play a Role in Urgency Urinary Incontinence and Its Severity? Front Cell Infect Microbiol [Internet]. 2016 Jul 27 [cited 2019 May 29];6. Available from: http://journal.frontiersin.org/Article/10.3389/fcimb.2016.00078/abstract

7. Pearce MM, Zilliox MJ, Rosenfeld AB, Thomas-White KJ, Richter HE, Nager CW, et al. The female urinary microbiome in urgency urinary incontinence. Am J Obstet Gynecol. 2015 Sep;213(3):347.e1-347.e11.

8. 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.

9. Hugenholtz P, Skarshewski A, Parks DH. Genome-Based Microbial Taxonomy Coming of Age. Cold Spring Harb Perspect Biol [Internet]. 2016 Jun [cited 2018 Aug 8];8(6). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4888819/

10. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Appl Environ Microbiol. 2007 Aug 15;73(16):5261–7.

11. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal

RNA sequence data compatible with ARB. Nucleic Acids Res. 2007 Nov 14;35(21):7188–96.

12. Gao X, Lin H, Revanna K, Dong Q. A Bayesian taxonomic classification method for 16S rRNA gene sequences with improved species-level accuracy. BMC Bioinformatics. 2017 Dec;18(1):247.

13. Balvočiūtė M, Huson DH. SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? BMC Genomics. 2017 Mar;18(S2):114.

14. Yang B, Wang Y, Qian P-Y. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. BMC Bioinformatics [Internet]. 2016 Dec [cited 2018 Sep 8];17(1). Available from: http://www.biomedcentral.com/1471-2105/17/135

15. Mayer E. The growth of biological thought: Diversity, evolution, and inheritance [Internet]. 11th ed. Vol. 5. Cambridge, MA: The Belknap Press of Harvard University Press; 2000 [cited 2018 Aug 7]. 974 p. Available from: http://linkinghub.elsevier.com/retrieve/pii/0162309584900384

16. Leifson E. Bacterial Taxonomy: a Critique. BACTERIOL REV. 1966;30:10.

17. van Niel CB. vanNiel_century_of_progress.odt. In: A Century of Progress in the Natural Sciences 1853-1953. San Francisco: Calif. Acad. Sci.; 1955. p. 87–114.

18. Vandamme P, Pot B, Gillis M, Vos PD, Kersters K, Swings J. Polyphasic Taxonomy, a Consensus Approach to Bacterial Systematics. MICROBIOL REV. 1996;60:32.

19. Woese C. Bacterial Evolution. Microbiol Rev. 1987;51(2):221–71.

20. Oren A, Garrity GM. Then and now: a systematic review of the systematics of prokaryotes in the last 80 years. Antonie Van Leeuwenhoek. 2014 Jul;106(1):43–56.

21. Rainey FA. How to Describe New Species of Prokaryotes. In: Methods in Microbiology [Internet]. Elsevier; 2011 [cited 2019 Nov 4]. p. 7–14. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780123877307000024

22. Goris J, Klappenbach JA, Vandamme P, Coenye T, Konstantinidis KT, Tiedje JM. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol. 2007 Jan 1;57(1):81–91.

23. Stackebrandt E. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int J Syst Evol Microbiol. 2002 May 1;52(3):1043–7.

24. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proc Natl Acad Sci. 1977 Nov 1;74(11):5088–90.

25. Larkin JM, Stokes JL. Taxonomy of Psychrophilic Strains of Bacillus. J Bacteriol. 1967;94(4):7.

26. Nakamura LK. Bacillus psychrophilus sp. nov. norn. rev. Int J Syst Bacteriol. 1984;34(2):3.

27. Fox GE, Wisotzkey JD, Jurtshuk P. How Close Is Close: 16S rRNA Sequence Identity May Not Be Sufficient To Guarantee Species Identity. Int J Syst Bacteriol. 1992 Jan 1;42(1):166–70.

28. Hall JPJ, Brockhurst MA, Harrison E. Sampling the mobile gene pool: innovation via horizontal gene transfer in bacteria. Philos Trans R Soc B Biol Sci. 2017 Dec 5;372(1735):20160424.

29. Avery OT, MacLeod CM, McCarty M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. J Exp Med. 1944 Feb 1;79(2):137–58.

30. Guttman D, Dykhuizen D. Clonal divergence in Escherichia coli as a result of recombination, not mutation. Science. 1994 Nov 25;266(5189):1380–3.

31. Koonin EV, Wolf YI. Evolution of microbes and viruses: a paradigm shift in evolutionary biology? Front Cell Infect Microbiol. 2012 Sep 13;2:15.

32. Creevey CJ, Doerks T, Fitzpatrick DA, Raes J, Bork P. Universally Distributed Single-Copy Genes Indicate a Constant Rate of Horizontal Transfer. Liberles D, editor. PLoS ONE. 2011 Aug 5;6(8):e22099.

33. Jain R, Rivera MC, Lake JA. Horizontal gene transfer among genomes: The complexity hypothesis. Proc Natl Acad Sci. 1999 Mar 30;96(7):3801–6.

34. Cohen O, Gophna U, Pupko T. The Complexity Hypothesis Revisited: Connectivity Rather Than Function Constitutes a Barrier to Horizontal Gene Transfer. Mol Biol Evol. 2011 Apr 1;28(4):1481–9.

35. Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, Rubin EM. Genome-Wide Experimental Determination of Barriers to Horizontal Gene Transfer. Science. 2007 Nov 30;318(5855):1449–52.

36. Feng D-F, Doolittle RF. Progressive sequence alignment as a prerequisitet to correct phylogenetic trees. J Mol Evol. 1987 Aug;25(4):351–60.

37. Notredame C, Higgins DG, Heringa J. T-coffee: a novel method for fast and accurate multiple sequence alignment 1 1Edited by J. Thornton. J Mol Biol. 2000 Sep;302(1):205–17.

38. Edgar RC, Batzoglou S. Multiple sequence alignment. Curr Opin Struct Biol. 2006 Jun;16(3):368–73.

39. Valdar WSJ. Scoring residue conservation. Proteins Struct Funct Bioinforma. 2002;48(2):227–41.

40. Linhart C, Shamir R. The degenerate primer design problem. Bioinformatics. 2002 Jul 1;18(Suppl 1):S172–81.

41. Hugerth LW, Wefer HA, Lundin S, Jakobsson HE, Lindberg M, Rodin S, et al. DegePrime, a Program for Degenerate Primer Design for Broad-Taxonomic-Range PCR in Microbial Ecology Studies. Löffler FE, editor. Appl Environ Microbiol. 2014 Aug 15;80(16):5116–23.

42. Roy S, Caruthers M. Synthesis of DNA/RNA and Their Analogs via Phosphoramidite and H-Phosphonate Chemistries. Molecules. 2013 Nov 18;18(11):14268–84.

43. Drake MJ, Morris N, Apostolidis A, Rahnama'i MS, Marchesi JR. The urinary microbiome and its contribution to lower urinary tract symptoms; ICI-RS 2015. Neurourol Urodyn. 2017 Apr 1;36(4):850–3.

44. 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.

45. Newcombe HB, Kennedy JM. Record linkage: making maximum use of the discriminating power of identifying information. Commun ACM. 1962 Nov 1;5(11):563–6.

46. Sayers A, Ben-Shlomo Y, Blom AW, Steele F. Probabilistic record linkage. Int J Epidemiol. 2016 Jun;45(3):954–64.

47. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Appl Env Microbiol. 2006 Jul 1;72(7):5069–72.

48. Ludwig W. ARB: a software environment for sequence data. Nucleic Acids Res. 2004 Feb 23;32(4):1363–71.

49. Federhen S. The NCBI Taxonomy database. Nucleic Acids Res. 2012 Jan 1;40(D1):D136–43.

50. Yarza P, Munoz R. The All-Species Living Tree Project. In: Methods in Microbiology [Internet]. Elsevier; 2014 [cited 2020 Jan 22]. p. 45–59. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0580951714000129

51. Maidak B. The Ribosomal Database Project (RDP). Nucleic Acids Res. 1996 Jan 1;24(1):82–5.

52. Fettweis JM, Serrano MG, Sheth NU, Mayer CM, Glascock AL, Brooks JP, et al. Species-level classification of the vaginal microbiome. 2012;9.

53. Navas-Molina JA, Peralta-Sánchez JM, González A, McMurdie PJ, Vázquez-Baeza Y, Xu Z, et al. Advancing our understanding of the human microbiome using QIIME. Methods Enzymol. 2013;531:371–444.

54. 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.

55. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2014 Apr 16;7(1):539–539.

56. Zabell SL. Symmetry and its discontents [Internet]. 1st ed. Cambridge, UK: Cambridge University press; 2005 [cited 2019 Oct 2]. 293 p. Available from: https://www.mobt3ath.com/uplode/book/book-41229.pdf

57. Edgar RC. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. Valencia A, editor. Bioinformatics. 2018 Jul 15;34(14):2371–5.

58. Schloss PD, Handelsman J. Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. Genome Biol. 2005;4.

59. Stackebrandt E, Goebel BM. Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. Int J Syst Evol Microbiol. 1994 Oct 1;44(4):846–9.

60. Callahan BJ. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. :5.

61. Hair JF, editor. Multivariate data analysis. 7. ed., Pearson new internat. ed. Harlow: Pearson; 2014. 734 p. (Pearson custom library).

62. Fellegi I, Sunter A. A theory for record linkage. J Am Stat Assoc. 1969 Dec;64(328):1183–210.

63. Harron KL, Doidge JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, et al. A guide to evaluating linkage quality for the analysis of linked data. Int J Epidemiol. 2017 Oct 1;46(5):1699–710.

64. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006 Jun;27(8):861–74.

65. Mende DR, Sunagawa S, Zeller G, Bork P. Accurate and universal delineation of prokaryotic species. Nat Methods. 2013 Sep;10(9):881–4.

66. Okonechnikov K, Golosova O, Fursov M. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics. 2012 Apr;28(8):1166–7.

67. Parte AC. LPSN – List of Prokaryotic names with Standing in Nomenclature (bacterio.net), 20 years on. Int J Syst Evol Microbiol. 2018 Jun 1;68(6):1825–9.

68. Chakravorty S, Helb D, Burday M, Connell N, Alland D. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. J Microbiol Methods. 2007 May;69(2):330–9.

69. Adekambi T, Shinnick TM, Raoult D, Drancourt M. Complete rpoB gene sequencing as a suitable supplement to DNA-DNA hybridization for bacterial species and genus delineation. Int J Syst Evol Microbiol. 2008 Aug 1;58(8):1807–14.

70. Klappenbach JA, Dunbar JM, Schmidt TM. rRNA Operon Copy Number Reflects Ecological Strategies of Bacteria. Appl Environ Microbiol. 2000 Apr 1;66(4):1328–33.

71. Graspeuntner S, Loeper N, Künzel S, Baines JF, Rupp J. Selection of validated hypervariable regions is crucial in 16S-based microbiota studies of the female genital tract. Sci Rep. 2018 Dec;8(1):9678.

72. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci. 2011 Mar 15;108(Supplement_1):4516–22.

73. Venter JC. Environmental Genome Shotgun Sequencing of the Sargasso Sea. Science. 2004 Apr 2;304(5667):66–74.

74. Fujiwara Y, Kawato M, Noda C, Kinoshita G, Yamanaka T, Fujita Y, et al. Extracellular and Mixotrophic Symbiosis in the Whale-Fall Mussel Adipicola pacifica: A Trend in Evolution from Extra- to Intracellular Symbiosis. Goldstien SJ, editor. PLoS ONE. 2010 Jul 27;5(7):e11808.

75. Ciccarelli FD, Doerks T, Mering C von, Creevey CJ, Snel B, Bork P. Toward Automatic Reconstruction of a Highly Resolved Tree of Life. Science. 2006 Mar 3;311(5765):1283–7.

76. Komesu YM, Richter HE, Dinwiddie DL, Siddiqui NY, Sung VW, Lukacz ES, et al. Methodology for a vaginal and urinary microbiome study in women with mixed urinary incontinence. Int Urogynecology J. 2017 May;28(5):711–20.

77. Bukin YuS, Galachyants YuP, Morozov IV, Bukin SV, Zakharenko AS, Zemskaya TI. The effect of 16S rRNA region choice on bacterial community metabarcoding results. Sci Data. 2019 Mar;6(1):190007.

78. Langhorst BW, Nichols NM. Database of DNA Polymerases. Curr Protoc Mol Biol [Internet]. 2012 Jul [cited 2020 Apr 26];99(1). Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/0471142727.mb0325s99

79. Andersen ES. The tmRDB and SRPDB resources. Nucleic Acids Res. 2006 Jan 1;34(90001):D163–8.

80. 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;6(1):226.

81. Karstens L, Asquith M, Davin S, Fair D, Gregory WT, Wolfe AJ, et al. Controlling for Contaminants in Low-Biomass 16S rRNA Gene Sequencing Experiments. Gilbert JA, editor. mSystems. 2019 Jun 4;4(4):e00290-19, /msystems/4/4/mSys.00290-19.atom.

82. Shapiro DJ, Hicks LA, Pavia AT, Hersh AL. Antibiotic prescribing for adults in ambulatory care in the USA, 2007–09. J Antimicrob Chemother. 2014 Jan;69(1):234–40.

83. Komesu YM, Dinwiddie DL, Richter HE, Lukacz ES, Sung VW, Siddiqui NY, et al. Defining the relationship between vaginal and urinary microbiomes. Am J Obstet Gynecol. 2019 Aug;S0002937819310105.

84. 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 Dec 20;10:2900.