bioRxiv preprint doi: https://doi.org/10.1101/2020.10.27.358408; this version posted October 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Species level resolution of female bladder microbiota from marker gene surveys

Carter Hoffman, MS[1]; Nazema Y Siddiqui, MD, MHSc[2]; Ian Fields, MD[3]; W. Thomas Gregory, MD[3]; Holly Simon, PhD[4]; Michael A. Mooney, PhD[1]; Alan J. Wolfe, PhD[5]; Lisa Karstens, PhD[1,3]*

Abstract

The human bladder contains bacteria in the absence of infection. Interest in studying these bacteria and their association with bladder conditions is increasing, but the chosen experimental method can limit the resolution of the taxonomy that can be assigned to the bacteria found in the bladder. 16S rRNA gene sequencing is commonly used to identify bacteria, but is typically restricted to genus-level identification. Our primary aim was to determine if accurate species-level identification of bladder bacteria is possible using 16S rRNA gene sequencing. We evaluated the ability of different classification schemes, each consisting of combinations of a 16S rRNA gene variable region, a reference database, and a taxonomic classification algorithm to correctly classify bladder bacteria. We show that species-level identification is possible, and that the reference database chosen is the most important component, followed by the 16S variable region sequenced.

Importance

Species-level information may deepen our understanding of associations between bladder microbiota and bladder conditions, such as lower urinary tract symptoms and urinary tract infections. The capability to identify bacterial species depends on large databases of sequences, algorithms that leverage statistics and available computer hardware, and knowledge of bacterial genetics and classification. Taken together, this is a daunting body of knowledge to become familiar with before the simple question of bacterial identity can be answered. Our results show the choice of taxonomic database and variable region of the 16S rRNA gene sequence makes species level identification possible. We also show this improvement can be achieved through the more careful application of existing methods and use of existing resources.

Author Affiliations:

1. Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR USA
2. Department of Obstetrics and Gynecology, Duke University, Durham, NC USA
3. Department of Obstetrics and Gynecology Oregon Health & Science University, Portland, OR USA
4. AnimalBiome, Oakland, CA USA
5. Department of Microbiology & Immunology, Loyola University Chicago, Maywood, IL, USA

*Corresponding author: karstens@ohsu.edu

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.27.358408; this version posted October 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

The human body provides a wide range of habitats, supporting a variety of microorganisms that include bacteria, archaea, viruses and fungi, collectively known as the human microbiome(1). Recent evidence from sequence-based and enhanced culturing techniques have revealed a population of microbes (bacteria, fungi and viruses) that exist in the bladder, even in the absence of infection(2–7). The discovery of the bladder microbiota (also known as the bladder urobiome) has led researchers to question how these microbes influence the health of the host. Studies have shown that altered bladder urobiome diversity is associated with urgency urinary incontinence (UUI)(4,8), urinary tract infection after instrumentation of the urinary tract(9,10), and is predictive of response to a common UUI drug (11). These studies collectively provide evidence that the bladder urobiome, while previously overlooked, is clinically relevant and warrants further investigation.

To study the relationships between the bacteria found in the human bladder and health of the host, it is necessary to accurately identify bacteria in a rapid and large-scale manner. Reliable methods of determining the bacterial identity of an unknown bacterium include Matrix Assisted Laser Desorption/Ionization-time of flight (MALDI-TOF) analysis or whole genome sequencing (WGS) of purified colonies; both techniques permit species-level identification of bacteria(12). However, culturing specific bacterial species is also time consuming and laborious. This limitation has been circumvented by adopting culture-independent methods of sequencing DNA directly from an environmental sample, such as shotgun metagenomic sequencing and targeted amplicon sequencing, the latter most commonly involving the 16S rRNA gene(13). These culture-independent sequencing methods are an attractive strategy because they can more accurately reveal microbiota diversity by identifying bacteria that are difficult to grow in culture.

Targeted amplicon sequencing is currently the most practical method for identifying bladder bacteria in a large-scale manner. When performing targeted amplicon sequencing, DNA is first extracted from all cells in a sample, including host and bacterial cells. Next, the polymerase chain reaction (PCR) is used to amplify a small segment of the bacterial genome. This segment is then sequenced in a high-throughput manner. Finally, bioinformatics are used to process the resulting sequences and identify the taxonomy of the bacteria. Algorithms compare the short DNA sequences recovered from a sample to known sequences held in a reference database until the closest match is found. In general, longer or more unique strings of sequenced DNA can be used to identify bacteria at a higher level of precision, though sequence length is often limited by the sequencing technology. The 16S rRNA gene is commonly used in amplicon sequencing studies due to its universal presence in bacteria. The 16S rRNA gene conveniently contains multiple "variable regions" with unique strings of sequence that can be used for bacterial identification. A common target is the 4th variable region (V4), as this region has good phylogenetic resolution down to the genus level for many bacteria(14).

When identifying bacteria using targeted amplicon sequencing there are three important components (**Figure 1**). These components are: 1) the identifier, or DNA sequence of the unknown bacterium; 2) a database of DNA sequences annotated with taxonomic information; and 3) a classifier, which is the algorithm that compares the unknown sequence to those in a database until the closest match is found. These components work together as a *classification scheme*. One common classification scheme uses the V4 region from the 16S rRNA gene

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.27.358408; this version posted October 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sequence(15) as the identifier, the Silva database(16), and the Naïve Bayes algorithm(17). A limitation of this particular classification scheme, and many others commonly used, is that the phylogenetic resolution is usually constrained to the genus level. Recently, several new approaches to sequence processing and taxonomy assignment have become available, which may improve resolution to the species level (e.g. amplicon sequence variant algorithms such as DADA2(18), and taxonomic classifiers such as Bayesian Lowest Common Ancestor (BLCA)(19)).

Our primary aim was to determine if species-level identification of bladder bacteria is possible from 16S rRNA gene sequencing studies. To achieve this aim, we used a representative sample of bacteria found in the human female bladder and published by Thomas-White and colleagues(20). This dataset includes bacteria found in the human female bladder that have been cultured, isolated, subjected to whole genome sequencing, and identified using full length sequences of 40 protein-encoding genes.. We used these known DNA sequences to determine which classification schemes would be most useful for future targeted amplicon sequencing studies. We evaluated several variable regions (i.e. potential identifiers), reference databases, and taxonomic classification algorithms for their ability to accurately identify bladder bacteria at the species level.

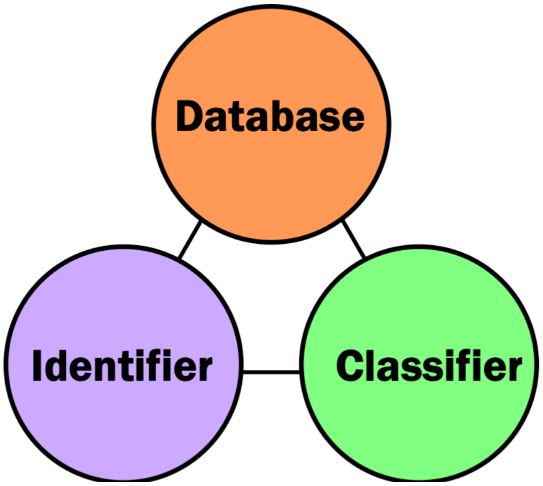


Figure 1. Model of the components that make up a classification scheme to assign taxonomy to unknown sequences. A classification scheme consists of an identifier, a database, and a classifier. The identifiers used in this study are subsequences of the 16S rRNA gene, computationally generated using published primers as coordinates on the gene sequence. These targeted amplicons are the V3, V4, and V6 variable regions of the gene, or span the V1-V3, V2-V3, V3-V4, and V4-V6 variable regions. The databases used in this study are the Greengenes, Silva, and NCBI 16S. The classifiers used in this study are the Naive Bayes and Bayesian Lowest Common Ancestor (BLCA) algorithms. One example of a classification scheme is the V4 region identifier, Silva database, and Naive Bayes classifier. Another example classification scheme is the V6 region identifier, Greengenes database, and BLCA classifier. These two examples are distinct from each other and can have different outcomes when assigning taxonomy.

Results

Representation of bladder bacteria in 16S rRNA gene sequence databases. The Thomas-White genome sequencing dataset consists of 149 bladder bacterial isolates, representing 78 bacterial species from 36 genera(20). There are several databases available for bacterial identification using the 16S rRNA gene(21,22). Of these, Greengenes (v.13_5)(23), Silva (v.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.27.358408; this version posted October 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

132)(24), and NCBI 16S Microbial (v. August 2019) were evaluated due to their widespread use in amplicon sequencing studies and availability of species-level annotation. At the genus level, all but 1 genus (*Globicatella*) were present in the Greengenes database and all genera were present in the Silva and NCBI 16S Microbial databases. At the species level, all 78 bladder bacterial species were present in the Silva and NCBI 16S Microbial databases, whereas only 21 species were present in the Greengenes database.

Information contained in variable regions differs for bladder bacterial species. Sequencing studies frequently focus on a small segment of the 16S rRNA gene that can be rapidly sequenced in a high throughput manner using short read sequencing technology, such as the Illumina HiSeq or MiSeq. To evaluate the performance of different variable regions as identifiers, amplicons were computationally generated from the Thomas-White genome sequencing dataset for the V1-V3, V2-V3, V3-V4, V4-V6, V3, V4, and V6 variable regions using published primers (see **Methods**). These computational amplicons (**Figure 2A**) were used to determine how well the currently available classification schemes can distinguish bladder bacterial species. To assess different classification schemes, we tested multiple permutations of the variable regions listed above with different databases (i.e. Greengenes, Silva, or NCBI 16S Microbial) and different classifiers (i.e. Naive Bayes or BLCA, see **Figure 1**).

To quantify the amount of information contained across variable regions of the 16S rRNA gene among commonly identified bladder bacteria, we performed a sliding window analysis on a multiple sequence alignment (MSA) of all genomes from the Thomas-White dataset. We calculated entropy as a measure of information content along the MSA (**Figure 2B**). As expected, the defined variable regions contained regions of high entropy, suggesting variability across species, whereas variable regions were flanked by conserved regions with low entropy containing sequences that are similar among species. The V1 and V2 regions contained the highest entropy, while V7 and V8 contained the lowest.

Evaluation of classification scheme performance. To evaluate the ability of currently available resources to identify bladder species, we calculated the recall, precision and F-measure for each classification scheme implemented (see **Methods**). Briefly, each resulting taxonomic classification was evaluated as a true match, true non-match, false match or false non-match based on whether the taxonomic classification was correctly assigned or not. Recall refers to the proportion of matches that the classification scheme correctly identified out of all possible matches. Precision refers to the proportion of matches that the classification scheme called correctly out of all classified matches. The F-measure is the equally weighted harmonic mean of recall and precision.

In general, the classification schemes that use the NCBI 16S Microbial database perform the best (**Figure 3**), with high recall and precision (range 60.3%-91.0% for both classifiers). Those using the Silva database show reduced precision and recall (range 23.1%-70.5% for both measures). Because the Greengenes database is missing many of the bacterial species found in the bladder, it is less precise. As such, classification schemes using the Greengenes database can have good recall values (range 50.0%-81.8%), but the precision values are very low (range 22.0%-36.0%), indicating a large proportion of false matches to the number of true matches.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.27.358408; this version posted October 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

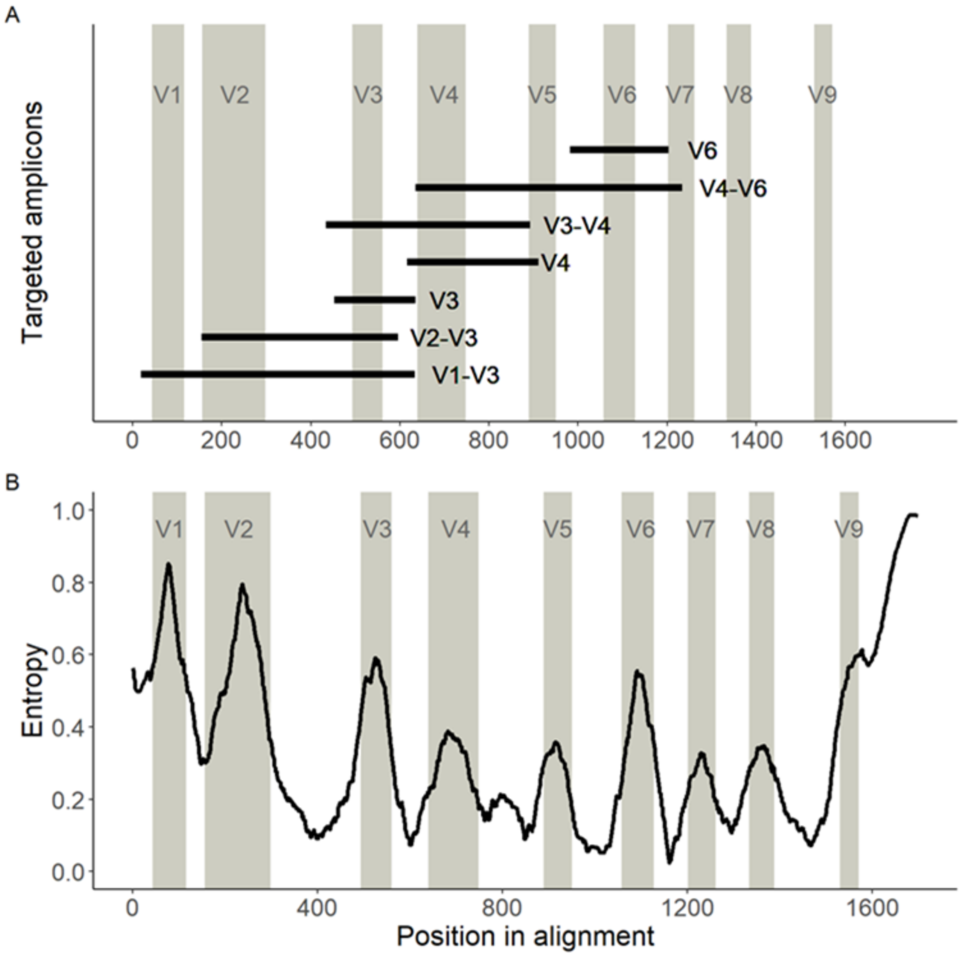


Figure 2. Variable regions of the 16S rRNA gene used in this study. A) Locations of the primers used in this study on the 16S rRNA gene. Locations of the predicted amplicons are shown as black bars in relation to the multi-sequence alignment (MSA) of the bacterial species described in Thomas-White et al. (2018). Gray columns are the locations of the known variable regions based on the sequence from *E. coli*. B) The information of variable regions, measured by entropy from a sliding window analysis of the MSA. Higher entropy indicates that the region has more variability across species, and therefore more information to identify a bacterial species. Lower entropy indicates that the region has little variability (i.e. is conserved) across species and therefore less information to identify a bacterial species.

When different variable regions are used as identifiers with Silva and NCBI 16S databases, there are differences in classification scheme outcomes. Using the Silva database and Naive Bayes classifier, the identifiers yielding the highest recall are the large V3-V4 (69.2%) and V4-V6 (69.2%) targeted amplicons. Using the Silva database and BLCA classifier, the V1-V3 and V4-V6 amplicons have the highest recall (55.1% and 48.7%, respectively). In contrast, the identifiers yielding the highest recall in classification schemes using the NCBI 16S database are the V1-V3 (90.3% on average) and V2-V3 (89.1% on average) targeted amplicons, regardless of the classifier.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.27.358408; this version posted October 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

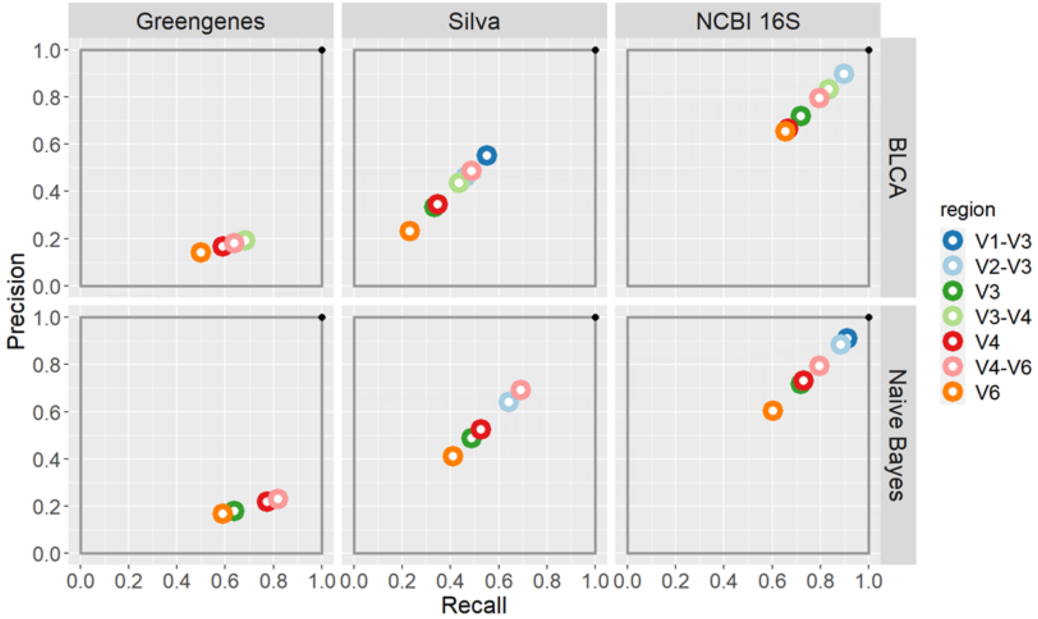


Figure 3: Classification scheme evaluation when ignoring confidence scores. The performance of each classification scheme is summarized by the precision (y axis) and recall (x axis) for each variable region (color). The best classification scheme would lie in the upper right-hand corner. Overall, classification schemes using the NCBI 16S Microbial database performed better than those using the Greengenes or Silva databases.

Confidence scores affect classification.

The BLCA and Naïve Bayes classifiers used in this study will classify an unknown sequence even when the posterior probability for that taxon is very low. To account for this situation, a confidence score is calculated that measures how much the classification changes through random permutation (bootstrapping) and produces a value that reflects the "goodness of fit" of that classification. When lacking any knowledge of how to choose the best confidence score that minimizes the number of errors of a classification scheme (i.e. when a test set is not available), using a predefined confidence score threshold is an option. Here, we evaluated the performance of classification schemes when confidence score thresholds of 50% or 80% were used, such that matches returned with confidence scores less than the threshold were considered non-matches. **Figure 4** shows the effect of increasing the confidence score on the number of true matches returned by each classification scheme.

Almost all classification schemes had a decrease in recall when using a default confidence score of 80% (**Supplemental Figure 1**). This effect is especially marked for the classification schemes that use the Silva database, which shows a 79.3% reduction in recall, on average. Classification schemes that use the NCBI 16S database are unequally affected, for example the V1-V3 identifier shows a slight reduction in recall (7.1% on average), while the V6 identifier shows the largest (43.3% on average). Classification schemes that use the Greengenes database are slightly affected. These reductions in recall are mirrored in all classification schemes when a confidence score of 50% is used as a threshold, but at a smaller magnitude.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.27.358408; this version posted October 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Changes in the precision of the classification schemes are affected the most by the database used (**Supplemental Figure 1**). For the classification schemes that use the NCBI 16S database, precision is generally improved regardless of confidence score, but at unequal amounts. For example, using the 80% threshold, the V1-V3 identifier shows a slight increase of 3.5% on average, while the V6 identifier shows a large 39.7% increase on average. Classification schemes that use the Silva database are unequally affected, with both reduction and gains in precision. A dramatic increase in precision is shown by the classification scheme composed of the Silva database, V4 identifier, and the Naive Bayes classifier. When ignoring a confidence score, this classification scheme has a precision of 52.6%, but shows a 63.1% gain when using a confidence score of 80% as a threshold. In general, precision is reduced when using the BLCA classifier and the Silva database. As with recall, classification schemes that use the Greengenes database show slight changes in precision.

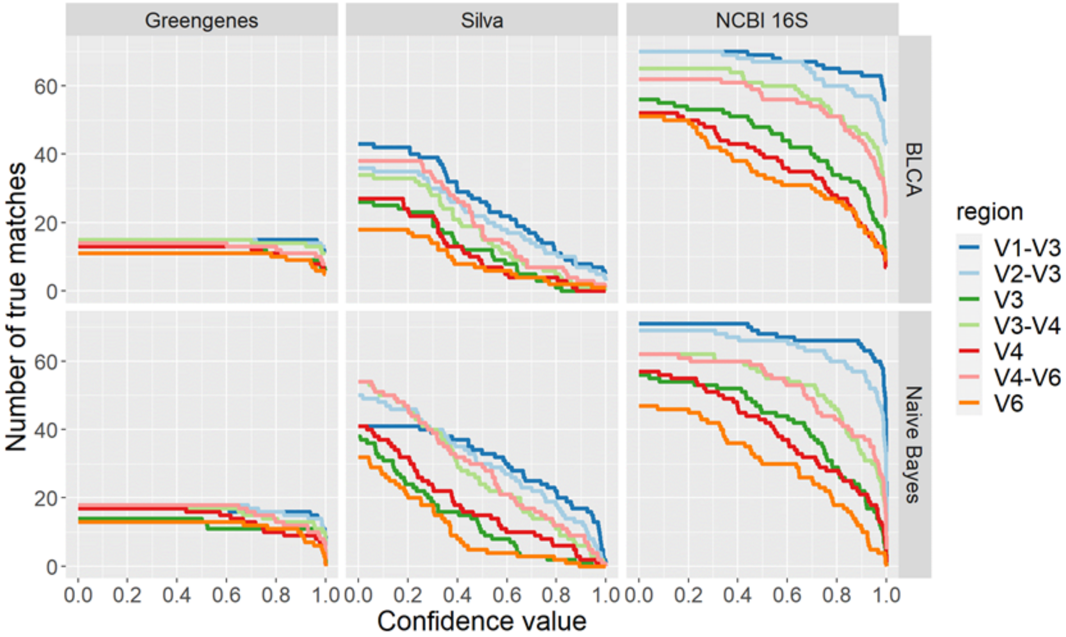


Figure 4: The number of true matches returned for each classification scheme across all confidence score values. As the confidence score value is increased, the number of true matches dramatically decreases, especially for schemes using the Silva database.

The overall changes in how these classification schemes perform when using a 50% or 80% confidence score can be summarized by comparing the F-measure values shown in **Supplemental Figure 2**. In almost every classification scheme, the F-measure value decreases when a threshold is used, indicating a larger proportion of false matches and false non-matches to the number of true matches. The classification schemes that use the Silva database clearly demonstrate this effect, which show a 66.9% reduction in F-measure values on average. The classification schemes that use the NCBI 16S database show slight decreases in the F-measure values, with the exceptions of those that use the V3, V4 and V6 regions as identifiers. Those classification schemes show a large 27.2% reduction in F-measure values on average. Finally, the classification schemes that use the Greengenes database have slight changes in their F-measure values, regardless of using a threshold or not.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.27.358408; this version posted October 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Amplicons spanning more than one variable region identify a higher number of bladder bacterial species. Amplicons spanning more than one variable region identified more unique bladder bacteria at the species level than amplicons spanning a single variable region. For example, with the commonly used V4 variable region and Naïve Bayes classifier, 21.8% of bladder bacteria are correctly identified with the Greengenes database, whereas 52.6% are identified with the Silva database and 73.1% with the NCBI 16S database (**Figure 5**). However, with the NCBI database, when using amplicons spanning more than one variable region, such as the V1-V3 region, 91.0% of bacteria are correctly identified at the species level.

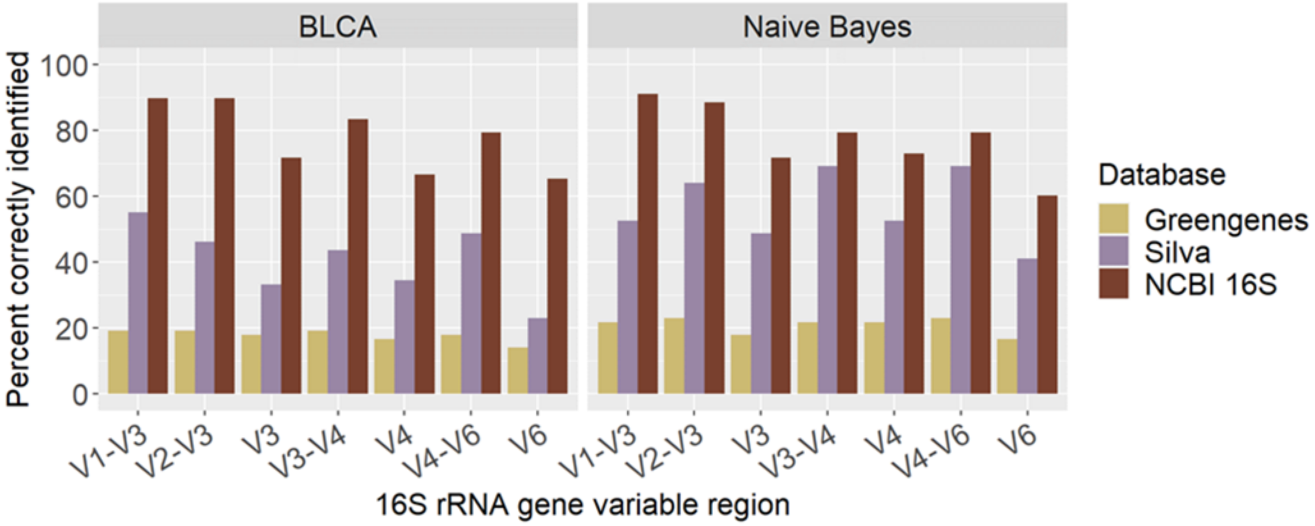


Figure 5: Percent of bladder bacteria correctly identified for each classification scheme. With the commonly used V4 variable region and BLCA classifier, 17% of bladder bacteria are correctly identified using the Greengenes database, compared with 35% correctly identified using the Silva database and 67% using the NCBI 16S database. A similar trend is seen with the Naïve Bayes classifier. Using other variable regions can lead to improved species-level resolution to a maximum number of 91% correctly identified.

Species identified depends on choice of database and variable region. While the results thus far have focused on summarizing overall performance of classification schemes for identifying bladder bacteria at the species level, we also sought to determine which classification schemes could be used to identify specific bacteria (**Table 1, Supplemental Figure 3**). Although the NCBI database contains the largest representation of bladder species, some species were not identified with certain variable regions, if at all. For example, *Lactobacillus* species were overall best represented within the NCBI database, with 8 out of 9 species being identified with the V1-V3 and V2-V3 variable regions (**Figure 6**). However, the other variable regions only identified between 4 and 6 *Lactobacillus* species when using the NCBI database. Interestingly, *Lactobacillus crispatus* was identifiable with the Silva and NCBI databases when using the V4-V6 regions, but only with the NCBI database using the V1-V3 and V2-V3 regions, and only the Silva database when using the V4 and V6 regions independently. *Lactobacillus iners* was not correctly identified from our dataset with any classification scheme.

Additionally, we found that there were important discrepancies for bacteria that are thought to play a role in bladder health and disease (**Supplemental Figure 3**). Several bladder species, such as *Gardnerella vaginalis*, were only detected with the NCBI and Silva databases. *Staphylococcus*

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.27.358408; this version posted October 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

species were poorly identified with the V4 region but were distinguishable with all other regions. *Streptococcus* and *Corynebacterium* species were best identified with NCBI. *Escherichia coli* is not well represented in any of the databases, and was only detected with the V4 region and the NCBI database.

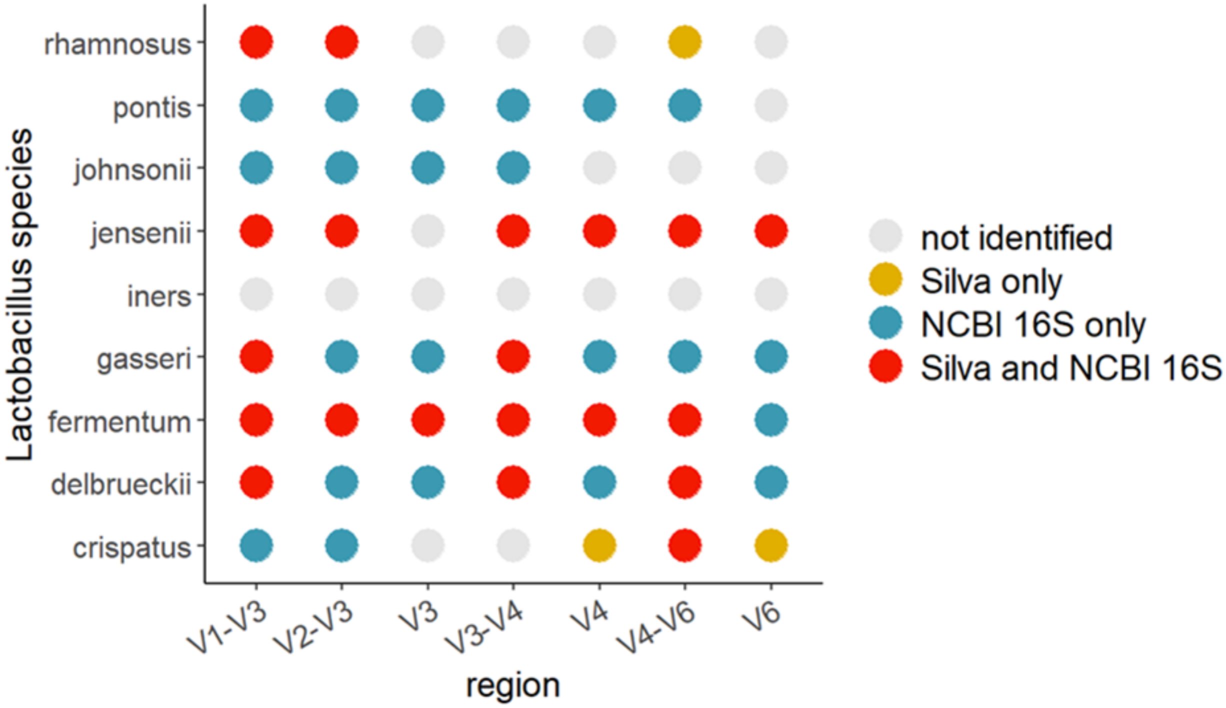
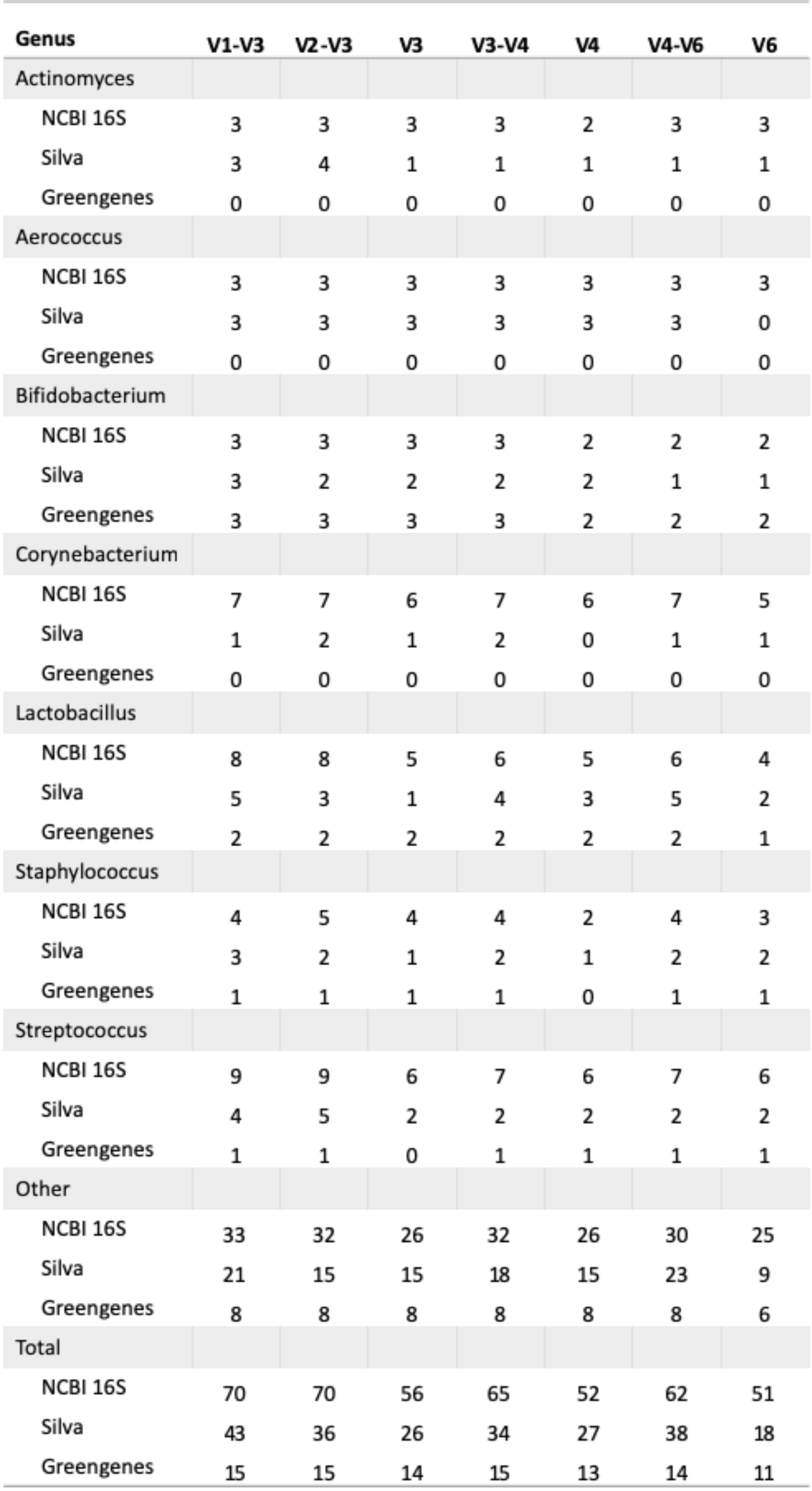


Figure 6: The ability of classification schemes to distinguish between different Lactobacillus species. Results shown for classification schemes using the BLCA classifier. Classification schemes using the NCBI 16S database have the most coverage, regardless of variable region chosen. The Greengenes database is not shown since it only classified two species (*L. pontis* and *L. delbrueckii*).

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.27.358408; this version posted October 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1. Number of species identified with each database by variable region (BLCA).



Genus	V1-V3	V2-V3	V3	V3-V4	V4	V4-V6	V6
Actinomyces							
NCBI 16S	3	3	3	3	2	3	3
Silva	3	4	1	1	1	1	1
Greengenes	0	0	0	0	0	0	0
Aerococcus							
NCBI 16S	3	3	3	3	3	3	3
Silva	3	3	3	3	3	3	0
Greengenes	0	0	0	0	0	0	0
Bifidobacterium							
NCBI 16S	3	3	3	3	2	2	2
Silva	3	2	2	2	2	1	1
Greengenes	3	3	3	3	2	2	2
Corynebacterium							
NCBI 16S	7	7	6	7	6	7	5
Silva	1	2	1	2	0	1	1
Greengenes	0	0	0	0	0	0	0
Lactobacillus							
NCBI 16S	8	8	5	6	5	6	4
Silva	5	3	1	4	3	5	2
Greengenes	2	2	2	2	2	2	1
Staphylococcus							
NCBI 16S	4	5	4	4	2	4	3
Silva	3	2	1	2	1	2	2
Greengenes	1	1	1	1	0	1	1
Streptococcus							
NCBI 16S	9	9	6	7	6	7	6
Silva	4	5	2	2	2	2	2
Greengenes	1	1	0	1	1	1	1
Other							
NCBI 16S	33	32	26	32	26	30	25
Silva	21	15	15	18	15	23	9
Greengenes	8	8	8	8	8	8	6
Total							
NCBI 16S	70	70	56	65	52	62	51
Silva	43	36	26	34	27	38	18
Greengenes	15	15	14	15	13	14	11

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.27.358408; this version posted October 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Validation of computational findings on V4 amplicon data. To evaluate the performance of our computational findings on actual data, we acquired targeted amplicon sequencing data from 24 urine samples. These urine samples were a subset of those that were used to derive cultures in the Thomas-White dataset and thus should contain the same bacteria. Sequencing data were generated as part of two other studies using Illumina sequencing of the V4 region of the 16S rRNA gene(4,11). We reprocessed the raw sequencing data (see **Methods**) and performed taxonomic classification to assess the performance of our computational findings. Since 16S rRNA gene sequencing will detect many more bacteria than those identified even with enhanced culture, we restricted the evaluation to only the bacteria that grew in culture from a given sample. We used accuracy to assess the number of predicted matches that were correctly identified in the V4 dataset, using classification schemes composed of the V4 identifier, each of three databases, and the BLCA classifier (**Figure 7**). All databases had good accuracy with high proportions of accurate identifications at the species level (80% for NCBI 16S, 86% for Silva, and 88% for Greengenes). Accuracy was reduced when the default confidence score of 80% was applied (61% for NCBI 16S, 67% for Silva, and 86% for Greengenes). The default confidence score of 50% reduced the accuracy of two of the classification schemes (76% for NCBI 16S and 80% for Silva). We also evaluated classification schemes with the Naive Bayes classifier and found similar results (**Supplemental Figures 4 and 5**)

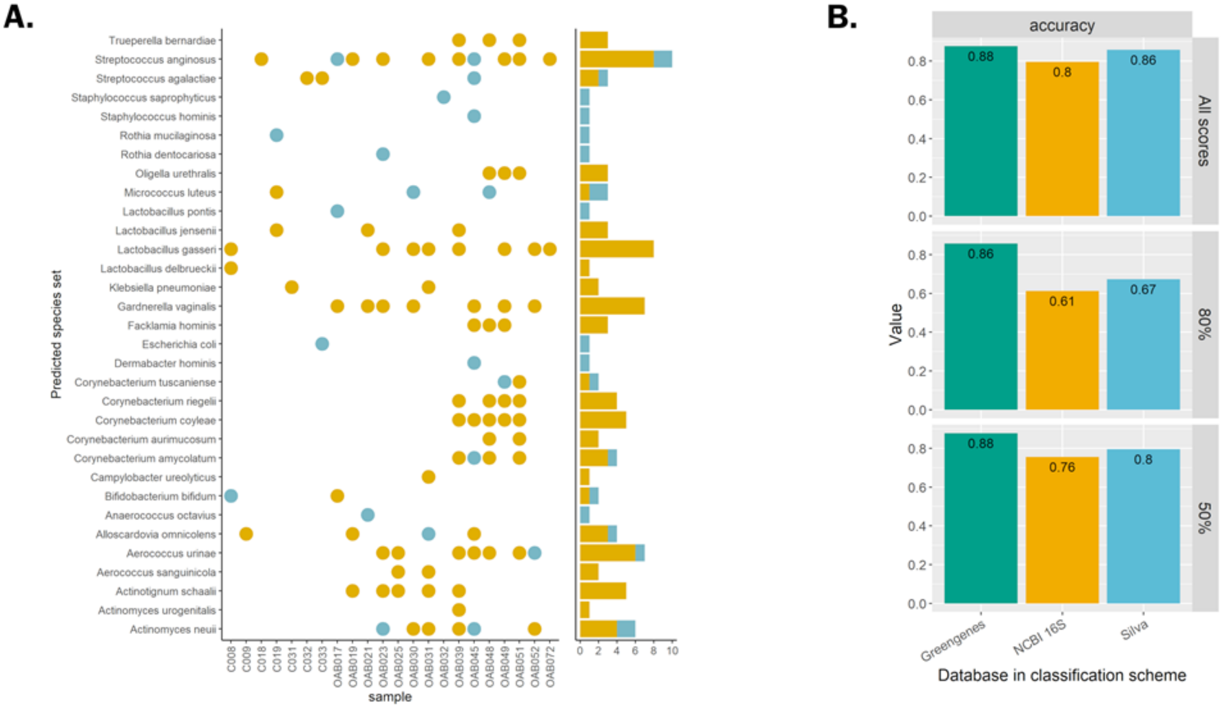


Figure 7. Taxonomic classification of the V4 validation dataset. A) Results when using a classification scheme including the V4 identifier, NCBI 16S database, and BLCA classifier. Blue dots represent species identified in cultured isolates, but not identified in targeted amplicon sequencing using this classification scheme. Yellow dots represent the species that were present in cultured isolates and successfully identified by the classification scheme. B) Summary of accuracy for classification schemes that use the V4 rRNA identifier, BLCA classifier, and the three databases (Greengenes, Silva and NCBI 16S). Rows show accuracy results when ignoring confidence scores, and when using confidence scores of 50% or 80% as thresholds.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.27.358408; this version posted October 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discussion

Our study demonstrates that it is possible to gain higher resolution results at the species level with existing resources when performing targeted amplicon sequencing of urinary specimens. Though higher resolution is possible, it requires a carefully chosen classification scheme. Within the classification scheme, the reference database strongly influences the identification of bacteria at the species level. Overall, we found the NCBI 16S database performs the best, whereas the Greengenes database performs the worst, primarily because it does not currently contain representatives of bladder bacteria. The identifier, or 16S rRNA variable region that is chosen, can also influence the types of bacterial species that are identified. The choice of classifier did not drastically affect the identification of species and thus is less critical within the classification scheme.

The largest limitation of any reference database is that the number of records of accurately classified bacteria is dwarfed by the number and diversity of unidentified sequences obtained through metagenomic sequencing of environmental samples. Because of the considerable amount of work required to construct and maintain databases, they will undoubtedly incompletely represent existing bacteria.

For species level taxonomy assignments, the reference database must contain species-level information. In other words, if species of bacteria are expected in a sample, it must be verified that the database contains those species. For example, we found that the Greengenes database does not currently contain many bacterial species that are found in the human bladder. In contrast, the NCBI 16S Microbial and Silva databases had representation of all species that were identified from prior studies of bladder bacteria. Thus, the latter two databases are better choices for evaluating bacterial species from the bladder.

While the databases reviewed in this study do have species-level information associated with the records, additional work was needed before species-level identification could be achieved with the Naïve Bayes classifier. This classifier requires a database that has undergone the "training" steps that convert the DNA sequences to the calculated frequencies that each *k*-mer occurs in a taxon. For available classification algorithms like the RDP classifier(25) and QIIME2(17), the training is only currently done to reliably identify bacteria to the genus level. For this study, it was necessary to train the Silva and NCBI 16S databases to the species level for use with the Naive Bayes classifier. While training the reference databases did take significant computational effort, once completed it was used repeatedly.

The classifiers used in this study are examples of two different strategies designed to overcome the common challenges of searching an extremely large dataset in order to find matching pairs of query sequences and reference records. While these two approaches are different in concept, we did not find significant differences in their performance for species-level classification of bladder bacteria.

BLCA is an example of sequence comparison by pairwise alignment. The strength of this method is due to the fact that the similarities between two DNA samples are directly compared. This is the most effective way to compare the characteristics of a sample to those that define a taxon; however, until recent advances in computer technology, it remained impractical because

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.27.358408; this version posted October 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of the computational burden. The Naive Bayes classifier is an example of a *k*-mer-based classification approach, and was designed to circumvent the computational challenges that are faced with use of a pairwise alignment classifier. However, there are limitations when using Naive Bayes for species-level identification. The first limitation arises from the database training process. If one taxon has more training examples than another, Naive Bayes generates unfavorable weights for the decision boundary(26). The second limitation is that all features (i.e. the *k*-mers generated from the DNA sequences) are assumed to be independent, and weights for taxa with strong dependencies among the associated *k*-mers are larger than those taxa with weakly dependent *k*-mers(26).

Finally, as shown by both the computational and V4 validation results, the use of the 50% or 80% confidence score thresholds significantly reduced the recall and accuracy of the classification schemes. Precision increased in several cases, for example with classification schemes that use the NCBI 16S database or those that use the Silva database and Naïve Bayes classifier, but at the cost of severely decreasing the number of species identified. These results show that the default settings of 50% or 80% are restrictive, and limit the ability to detect bladder species, especially when using the Silva and Greengenes databases. This could be resolved through the use of a comparative data set to find the confidence score values yielding optimal performance of these classification schemes.

Affordable sequencing of large-scale data is presently done with short read sequencing technology, such as Illumina MiSeq. This is currently limited to sequencing reads up to 300 nucleotides in length. Until full-length 16S rRNA gene sequencing can be achieved affordably on a large scale (such as with Oxford Nanopore and PacBio technologies), choosing the optimal region of the 16S rRNA gene for identification purposes remains a significant part of the experimental design. Thus, the variable regions that are used as identifiers require some consideration.

Our findings show that use of the V2-V3, and V1-V3 regions of the 16S rRNA gene allowed for the correct identification of the most bladder bacterial species when combined with the NCBI 16S database and either classifier. In general, amplicons that span more than one variable region perform better than those that contain single variable regions. This is likely due to the increased information available with longer reads. It is important to note that longer reads can also have limitations, which are discussed in more detail below. While shorter variable regions, such as the V4 region, did not perform as well as longer amplicons, they were able to identify many bladder bacteria at the species level (52 out of 78). These shorter amplicons are widely used with Illumina sequencing and may be valid, depending on the study design and level of precision desired. However, other variable regions may be explored for practical application, or when more detailed information is desired.

By taking a computational approach to evaluating classification schemes that are capable of identifying bladder bacterial species, we were able to thoroughly assess the ability of classification schemes to identify known bacterial species. However, there are several practical limitations of amplicon sequencing that were not captured in this approach.

Targeted amplicons are generated by priming the polymerase chain reaction with specially designed oligonucleotides (PCR primers). The challenge of PCR primer design is to identify a

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.27.358408; this version posted October 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 becomes very difficult when considering the 16S rRNA gene sequence of many species. We used published primers to create computational amplicons, but this may not reflect the actual experimental efficiency. Finally, quality control with DNA sequence processing must be conducted before classification is performed. Targeted amplicon sequencing generates a large number of overlapping reads and provides the data for 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 reads that are complementary before attempting to classify them. If the sequence reads do not overlap, loss of phylogenetic information occurs in the gaps and impacts the accuracy of identification, which may occur for longer amplicons, such as the V1-V3 and V4-V6 regions.

In our study, we identified the V1-V3 region of the 16S rRNA gene as having the greatest taxonomic resolution for the bacteria that are found in the bladder. This may be attributed to the high occurrence of insertions and deletions (indels) in the conserved regions between the first three variable regions across the bacteria in the Thomas-White dataset. Designing one degenerate primer set that would amplify the entire dataset may not be possible for this region. A future research direction could be to stratify the Thomas-White dataset into smaller, more closely related phylogenetic groups for more specific primer design.

Conclusion

Species level taxonomy assignment will greatly benefit studies focused on the urobiome and its relationship to bladder health and disease. Our results show that it is possible to reliably classify bladder bacterial species using targeted amplicon sequencing of the 16S rRNA gene variable regions with existing classification algorithms and databases. We determined that the most important component of the classification scheme is the database used, and that the NCBI database allows for best identification of bladder species. Our validation with V4 amplicon data demonstrates that the predicted computational outcomes are a good approximation for how a classification scheme will perform on real data. The knowledge that a majority of the predicted matches reflect reality is encouraging. It can be expected that the alternate variable regions covered in this study, such as the V2-V3 region of the 16S rRNA gene, would have similar outcomes.

Importantly, we found that no single variable region gives 100% coverage of all bladder bacteria species. Thus, the choice of variable region may significantly affect the results of a given study. One approach to resolve this could be to use multiple amplicon sequencing or long read sequencing technology. These technologies are emerging and may prove to be beneficial for the urobiome community. Furthermore, no database has 100% coverage across a variable region. This could be resolved by using more than one database for classification, though this approach is complicated by differences in databases in terms of formatting, as well as conflicting classifications. Both of these components are important for planning experimental and computational aspects of urobiome studies, and should be considered when comparing results across studies.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.27.358408; this version posted October 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Material and Methods

Code resources. All scripts that were written for this project can be found in the GitHub repository (https://github.com/lakarstens/BladderBacteriaSpecies). All scripts sourced from this repository are referred to as "custom."

The Thomas-White dataset. The 78 species of bladder bacteria used in this study were identified by culturing 149 urine samples and performing whole-genome sequencing, as described in Thomas-White et al.(20). This set of identified species served as the basis for our computational analysis and is referred to as the Thomas-White dataset. For each species identified, the 16S rRNA gene sequence of the corresponding type strain was downloaded from the Silva v132 release (https://www.arb-silva.de/) on 4/27/2019. 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(27), and is the agreed bacterial organism to which the taxonomic name refers. Sequences were searched using the "[T]" filter setting, and sequences longer than 1450 nt with alignment and pintail quality scores greater than 95% were selected. For the species that had no hits, the taxonomic synonym (see below) was used as the search query, if available. One unidentified *Corynebacterium* species had no type strain available, and was excluded from the analysis.

The V4 validation dataset. Targeted amplicon sequences from 24 urine samples, using the V4 region of the 16S rRNA gene sequence, is referred to as the V4 validation dataset. These 24 urine samples originated from a subsample of the women whose samples comprised the Thomas-White dataset. Sequencing data were generated as part of two other published studies using Illumina sequencing of the V4 region of the 16S rRNA gene(4,11). The raw sequence reads were processed with DADA2 version 1.14.1(18) to generate amplicon sequence variants (ASVs). The ASVs were subjected to taxonomic classification with the BLCA algorithm.

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(28) (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.

Databases. 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 (https://github.com/qunfengdong/BLCA/). For use with the Qiime2 package, the FASTA file was reformatted to work with Qiime2 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.

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.27.358408; this version posted October 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

"makeblastdb" utility provided in the Blast+ suite. The taxonomy file that was required by BLCA was generated with the custom "write_taxonomy.py" script. For use with the Qiime2 package, the FASTA file was reformatted to work with Qiime2 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" Qiime2 script.

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 included with BLCA. For use with the Qiime2 package, a FASTA file was extracted from the bundled BLCA database using "blastdbcmd" utility provided in the Blast+ suite, and reformatted to work with Qiime2 using the custom "write_qiime_train_db.py" script. The database was trained to work with the Naive Bayes classifier with the provided "fit-classifier-naive-bayes" script included in Qiime2.

Presence of Thomas-White species in 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 a FASTA file (if needed) using the "blastdbcmd" utility included in the Blast+ suite. The FASTA file was then searched using regular expressions and the Linux command-line program *grep* for a match of each species in the dataset. The commands were implemented using the custom "species_in_db.bash" script. The presence or absence of each species was recorded.

Multisequence alignment. The 16S gene sequences from the Thomas-White dataset were formed into a multi-sequence alignment using the T-coffee program(29). T-coffee version 12.00.7fb08c2 was downloaded from (http://tcoffee.org/Packages/Stable/Latest/) on 4/5/2019. Alignments were performed using the default settings.

Sliding window analysis. Comparing the 16S rRNA gene sequences of the species in the Thomas-White dataset reveals regions of conserved sequence and regions of variability. The degree that variable regions of species differ from each other can aid the identification of each species; therefore, quantifying the amount of variability of a region across a set of species is important.

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. Quantifying the amount of variability along a MSA is achieved by combining SWA with calculating the Shannon Entropy contained in each column framed by the window.

The minimum Shannon entropy occurs when all nucleotides in a position (column) of the MSA are the same. The maximum occurs when all possible nucleotides in the MSA are present at that position. However, the Shannon Entropy treats gaps in a sequence as relevant, where in practice gaps reflect an absence of useful information. Multisequence alignments can generate many columns of gap characters due to insertions or deletions (indels) in the respective sequences that make up the MSA. A consequence of treating gaps as relevant is the Shannon Entropy will interpret these indel regions as conserved sequence. This limitation was overcome by weighting the entropy scores against gaps(30). The locations of known variable regions of the 16S gene sequence were validated, and the relative amount of variability was quantified, using the custom "weighted_ent.py" script.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.27.358408; this version posted October 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Primers. Amplicons were computationally generated from the Thomas-White genome sequencing dataset for the V1-V3, V2-V3, V3-V4, V4-V6, and V4 variable regions using published primers and the V3 and V6 regions using designed primers. The primer sequences used, listed in order of amplicon spanning variable region(s), forward primer name and sequence, reverse primer name and sequence are: **V1-V3**: A17F 5'-GTT TGA TCC TGG CTC AG-3', 515R 5'-TTA CCG CGG CMG CSG GCA-3'(31,32). **V2-V3**: 16S_BV2f 5'-AGT GGC GGA CGG GTG AGT AA-3', HDA-2 5'-GTA TTA CCG CGG CTG CTG GCA C-3'(33,34). **V3:** v3_579F 5'-THT TSS RCA ATG GRS GVA-3', v3_779R 5'-GKN SCR AGC STT RHY CGG-3'. **V3-V4:** V3F 5'-CCT ACG GGA GGC AGC AG-3', V4R 5'-GGA CTA CHV GGG TWT CTA AT-3'(35). **V4**: F515 5'-GTG CCA GCM GCC GCG GTA A-3', R806 5'-CCT GAT GHV CCC AWA GAT TA-3'(36). **V4-V6:** 519F 5'-GTG CCA GCT GCC GCG GTA ATA-3', 1114R 5'-GGG GTT GCG CTC GTT GC-3'(32). **V6:** v6_1183F 5'-CCG CCT GGG GAS TAC GVH-3', v6_1410R 5'-AGT CCC RYA ACG AGC GCA-3'. Degenerate primer design was employed to generate primer sets for the V3 and V6 regions of the 16S rRNA gene that would anneal to as many species in the Thomas-White dataset as possible with DegePrime(37) (https://github.com/EnvGen/DEGEPRIME.git). DegePrime has the option to ignore columns of a MSA if the number of "-" characters exceed a user-defined threshold. The MSAs were preprocessed with this threshold set to .01. The main script of DegePrime was run using a degeneracy setting of 4096 and a window length of 18.

Extracting computational amplicons. For each primer set, 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 (accession number 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 output as a multi-record FASTA formatted file.

Taxonomic classifiers. Taxonomic classification was performed with Bayesian lowest common ancestor (BLCA) and Naïve Bayes classifiers. BLCA(19) was cloned from the GitHub repository https://github.com/qunfengdong/BLCA.git. For the 16S variable regions, the BLCA was run using default settings but pointing to the selected reference database, either Greengenes, Silva, or NCBI 16S. The Naïve Bayes classifier as implemented by Qiime2(17) was used with the Greengenes, Silva, and NCBI 16S databases and a confidence setting of 0, 50, and 80, but otherwise default settings.

Evaluating computational results. To evaluate the taxonomic classification results of each classification scheme on the computational amplicon dataset, the custom "new_taxonomy_results_2020-3-14.Rmd" file was used. These scripts compare the results of each record pair (each comparison between the query sequence and sequence held in the reference database) from the classification scheme to the known identify of the query sequence from the Thomas-White dataset. All record pairs that were assigned a match by the classification schemes were evaluated according to the following definitions (**Figure 8**):

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.27.358408; this version posted October 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

False non-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 false non-match.

True non-match - All records in the reference database that were not in the Thomas-White dataset. While records assigned to this category were not used in evaluating the classification schemes in this manuscript, the definition is still included for completeness.

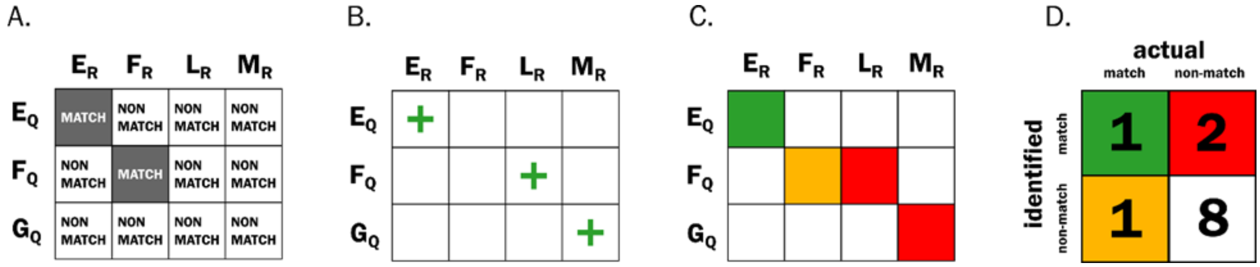


Figure 8. Example of classification evaluation used in this study. Suppose there is a classification scheme comprising a set of query sequences (the rows E,F,G) and the set of reference sequences (the columns E,F,L,M) held in a reference database. In this example, the number of reference records is greater than the query records, and the reference is missing a corresponding G record from the query set. **A**) If the query and reference record letters are the same, then they are designated as a **match**. If they are different they are designated as a **non-match**. **B**) Next, the classifier is allowed to assign record pairs as matches or non-matches for all query sequences, represented as green plus signs for matches and blank cells as non-matches. Some results are correct, and some are not. Note that despite the lack of a matching record in the reference database, the classifier still designated the (G:M) pair as a match. **C**) Using the definitions for assigning the classifications to the confusion matrix, there is one **true match** (green square), two **false matches** (red squares), one **false non-match** (yellow square), and 8 **true non-matches** (white squares). D) The cell values of the confusion matrix are then filled out, and performance measurements can be calculated. For this classification scheme, the precision is 1/(1+2)=.33, recall is 1/(1+1)=.5, and the F-measure is (2*.33*.5)/(.33+.5)=.40.

Performance measures. Recall, precision and the F-measure were used to evaluate the performance of each classification scheme implemented. Recall refers to the proportion of matches that the classification scheme correctly identified (true matches) out of all possible matches (true matches plus false non-matches). Precision refers to the proportion of matches that the classification scheme called correctly (true matches) out of all classified matches (true matches plus false matches). The F-measure is the equally weighted harmonic mean of recall and precision. For this study, we chose to maximize recall over precision, because the number of true matches impacts the subsequent work on diversity measures, such as species richness and evenness(38).

Evaluating V4 validation results. The species of bacteria in the V4 seqeuncing data 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. To determine the expected bacterial species in each sample, the results of the whole genome sequencing on the isolates cultured from the corresponding subject was used. For each

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.27.358408; this version posted October 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

classification scheme, accuracy was calculated by enumerating the number of species identified by WGS that were also identified by the V4 16S targeted amplicon sequencing using the custom "real_world_data_2020-4-17.Rmd" file.

The results of the V4 validation set were evaluated according to the following definitions (**Figure 9**):

True match - All matches from the computational classification scheme that were correctly identified by V4 16S targeted amplicon sequencing
False match - All species identified by V4 16S targeted amplicon sequencing that were not identified by the computational classification scheme
False non-match - All matches from the computational classification scheme that were not identified by V4 16S targeted amplicon sequencing
True non-match - All species that were not identified by either the computational classification schemes or the V4 16S targeted amplicon sequencing

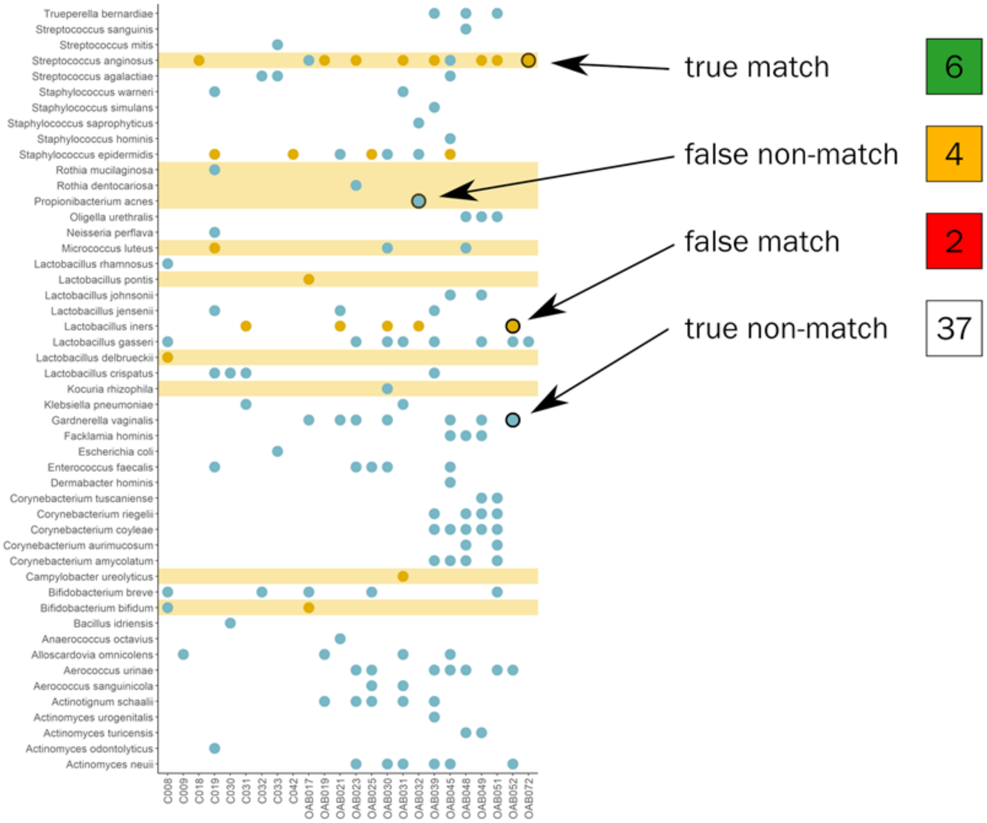


Figure 9: Definitions of how the classification scheme outcomes are assigned to the cells of the confusion matrix for the V4 validation results. This example shows the classification scheme composed of the Greengenes database, BLCA classifier, and the V4 region of the 16S rRNA gene as the identifier. When the Thomas-White dataset is subsetted by the 24 samples that underwent targeted amplicon sequencing, a smaller set of 49 species remains. The light yellow rows indicate the species correctly identified by the computational classification scheme. 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 targeted amplicon sequencing. Yellow dots in light yellow rows are true matches, when found elsewhere they are false matches. Blue dots in the light yellow rows are false non-matches, when found elsewhere they are true non-matches.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.27.358408; this version posted October 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Data Availability.

This project used previously acquired publicly available data.[20] All code that was written for this project can be found in the GitHub repository: https://github.com/lakarstens/BladderBacteriaSpecies.

Acknowledgements.

This work was supported by the National Institutes of Health (NIH): NIDDK award number K01 DK116706 (LK); NIA award numbers R03 AG060082 and P30AG028716 (NS). The content of the manuscript is solely our responsibility and does not represent the official views of the NIH or any other funding agency

Author contributions.

CH and LK conceived and designed experiments, and wrote the manuscript. CH performed analyses, LK reviewed analyses. AJW supplied the data. LK, CH, MM, HS, TG, IF, NS, and AJW interpreted results and revised the manuscript.

References

1. Lederberg J, McCray AT. 'Ome Sweet 'Omics-- A Genealogical Treasury of Words. The Scientist. 2001 Apr 2;15(7):2.

2. 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.

3. Khasriya R, Sathiananthamoorthy S, Ismail S, Kelsey M, Wilson M, Rohn JL, et al. Spectrum of Bacterial Colonization Associated with Urothelial Cells from Patients with Chronic Lower Urinary Tract Symptoms. J Clin Microbiol. 2013 Jul 1;51(7):2054–62.

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

5. Fouts DE. Next Generation Sequencing to Define Prokaryotic and Fungal Diversity in the Bovine Rumen. PLOS ONE. 2012;7(11).

6. Price TK, Dune T, Hilt EE, Thomas-White KJ, Kliethermes S, Brincat C, et al. The Clinical Urine Culture: Enhanced Techniques Improve Detection of Clinically Relevant Microorganisms. Forbes BA, editor. J Clin Microbiol. 2016 May;54(5):1216–22.

7. Ackerman AL, Underhill DM. The mycobiome of the human urinary tract: potential roles for fungi in urology. Ann Transl Med. 2017 Jan;5:31–31.

8. 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. 2016 Jul 27;6.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.27.358408; this version posted October 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

9. 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.

10. Thomas-White KJ, Gao X, Lin H, Fok CS, Ghanayem K, Mueller ER, et al. Urinary microbes and postoperative urinary tract infection risk in urogynecologic surgical patients. Int Urogynecology J. 2018 Dec;29(12):1797–805.

11. 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.

12. Kliem M, Sauer S. The essence on mass spectrometry based microbial diagnostics. Curr Opin Microbiol. 2012 Jun;15(3):397–402.

13. Pace NR. A Molecular View of Microbial Diversity and the Biosphere. Science. 1997 May 2;276(5313):734–40.

14. Yang B, Wang Y, Qian P-Y. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. BMC Bioinformatics. 2016 Dec;17(1):135.

15. Hugenholtz P, Skarshewski A, Parks DH. Genome-Based Microbial Taxonomy Coming of Age. Cold Spring Harb Perspect Biol. 2016 Jun;8(6).

16. 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.

17. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019 Aug;37(8):852–7.

18. 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.

19. 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.

20. 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.

21. Breitwieser FP, Lu J, Salzberg SL. A review of methods and databases for metagenomic classification and assembly. Brief Bioinform. 2019 Jul 19;20(4):1125–36.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.27.358408; this version posted October 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

23. 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.

24. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2012 Nov 27;41(D1):D590–6.

25. 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.

26. Rennie JDM, Shih L, Teevan J, Karger DR. Tackling the Poor Assumptions of Naive Bayes Text Classifiers. 2003;8.

27. Rainey FA. How to Describe New Species of Prokaryotes. In: Methods in Microbiology. Elsevier; 2011. p. 7–14.

28. 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.

29. 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.

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

31. Kumar PS, Griffen AL, Moeschberger ML, Leys EJ. Identification of Candidate Periodontal Pathogens and Beneficial Species by Quantitative 16S Clonal Analysis. J Clin Microbiol. 2005 Aug 1;43(8):3944–55.

32. Kumar P, Brooker M, Dowd S, Camerlengo T. Target Region Selection Is a Critical Determinant of Community Fingerprints Generated by 16S Pyrosequencing. PLOS ONE. 2011 Jun;6(6).

33. 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.

34. Walter J, Tannock GW, Tilsala-Timisjarvi A, Rodtong S, Loach DM, Munro K, et al. Detection and Identification of Gastrointestinal Lactobacillus Species by Using Denaturing Gradient Gel Electrophoresis and Species-Specific PCR Primers. Appl Environ Microbiol. 2000 Jan 1;66(1):297–303.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.27.358408; this version posted October 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

35. 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.

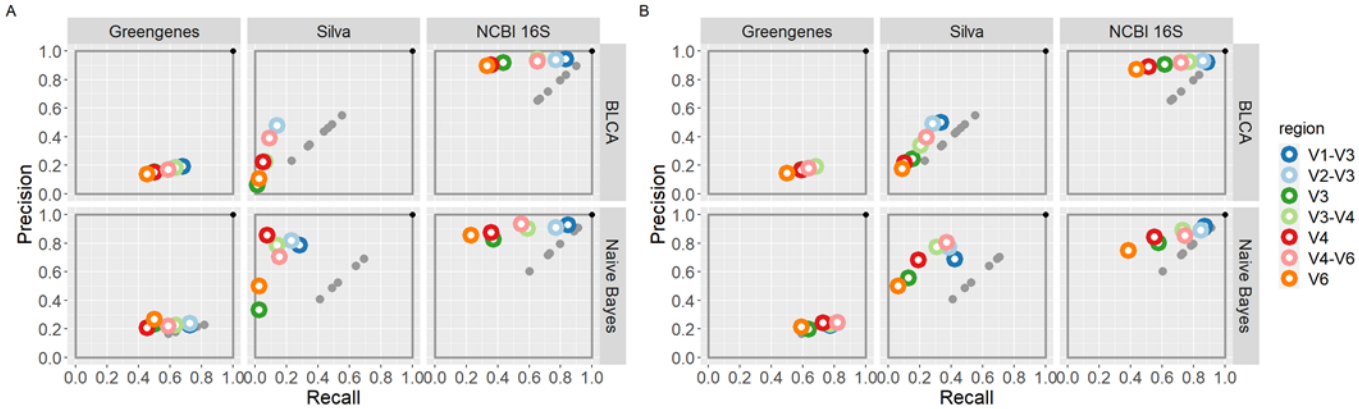
36. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci. 2011 Mar 15;108(Supplement_1):4516–22.

37. 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.

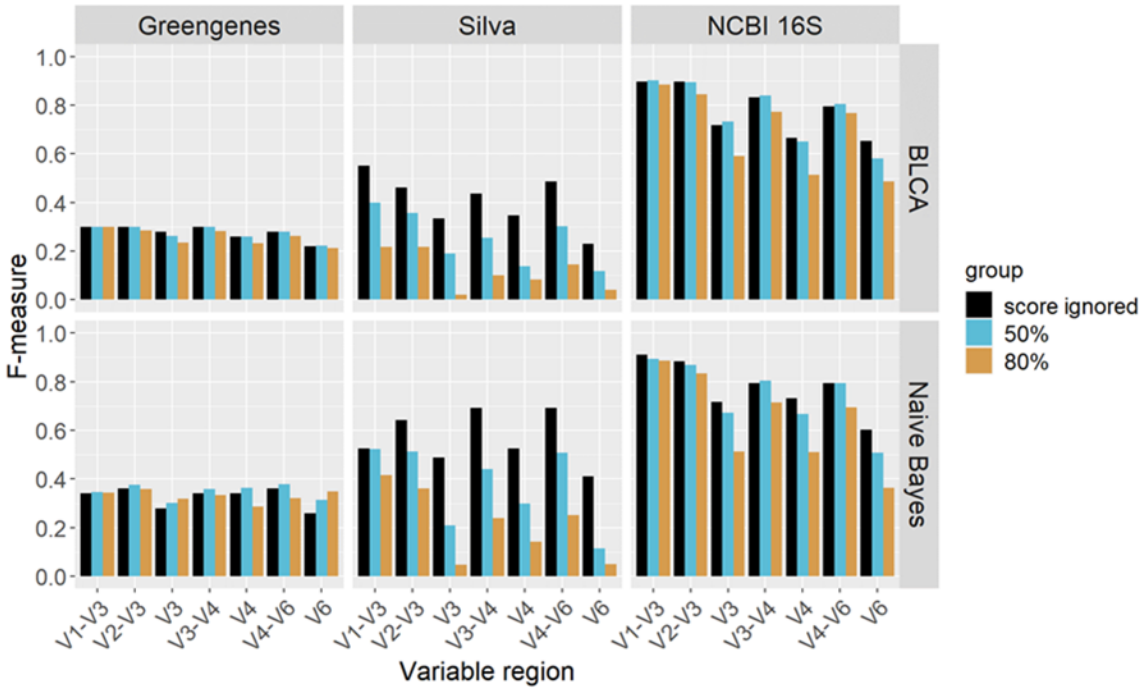
38. Morris EK, Caruso T, Buscot F, Fischer M, Hancock C, Maier TS, et al. Choosing and using diversity indices: insights for ecological applications from the German Biodiversity Exploratories. Ecol Evol. 2014 Sep;4(18):3514–24.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.27.358408; this version posted October 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplemental Figures

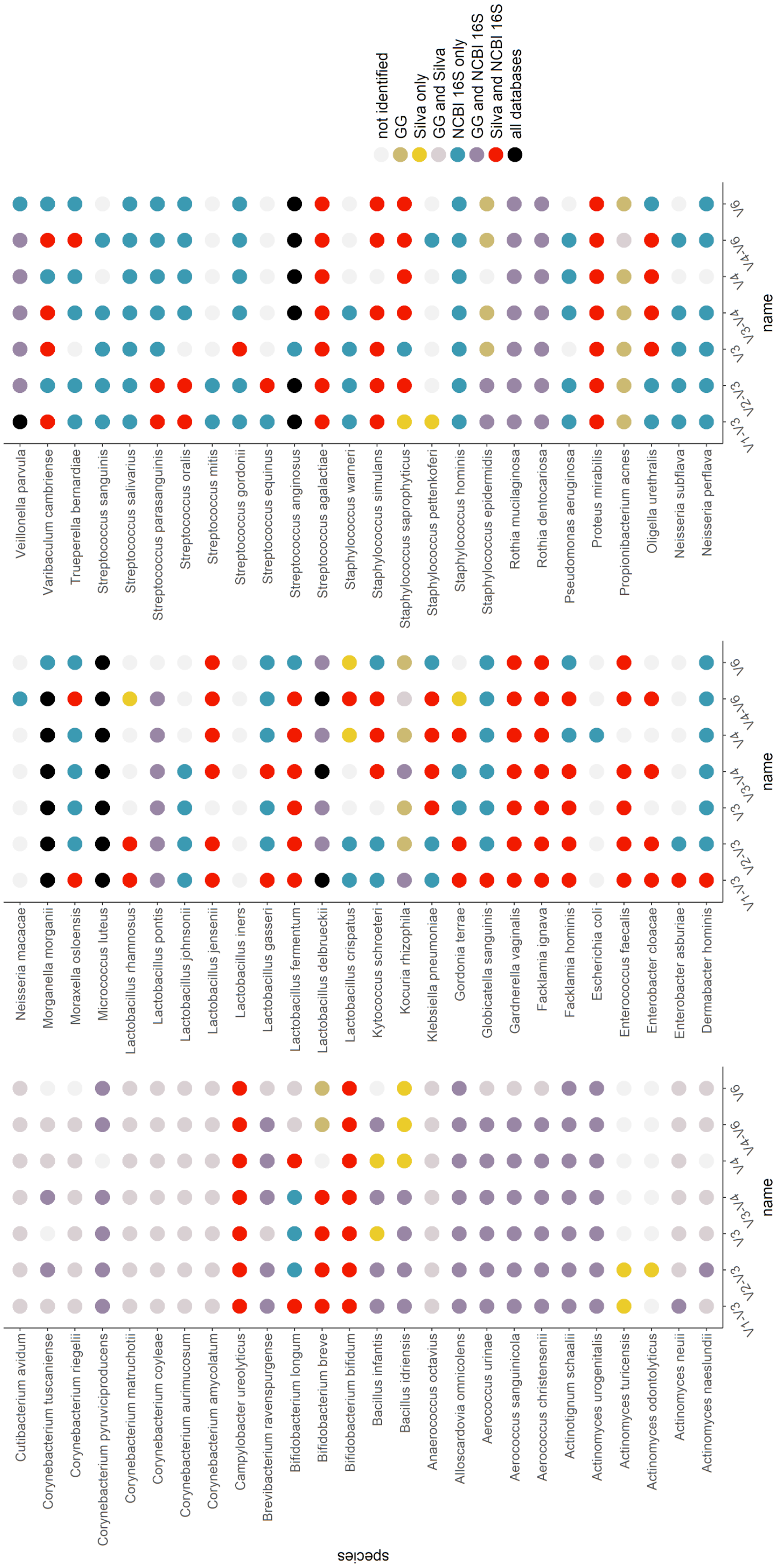


Supplemental Figure 1. Classification scheme precision and recall when using a confidence scores of (A) 80% and (B) 50% as a threshold. The schemes that use the Silva database have very low recall compared to when the confidence score is ignored (gray dots), whereas schemes that use Greengenes and NCBI 16S are not as affected.



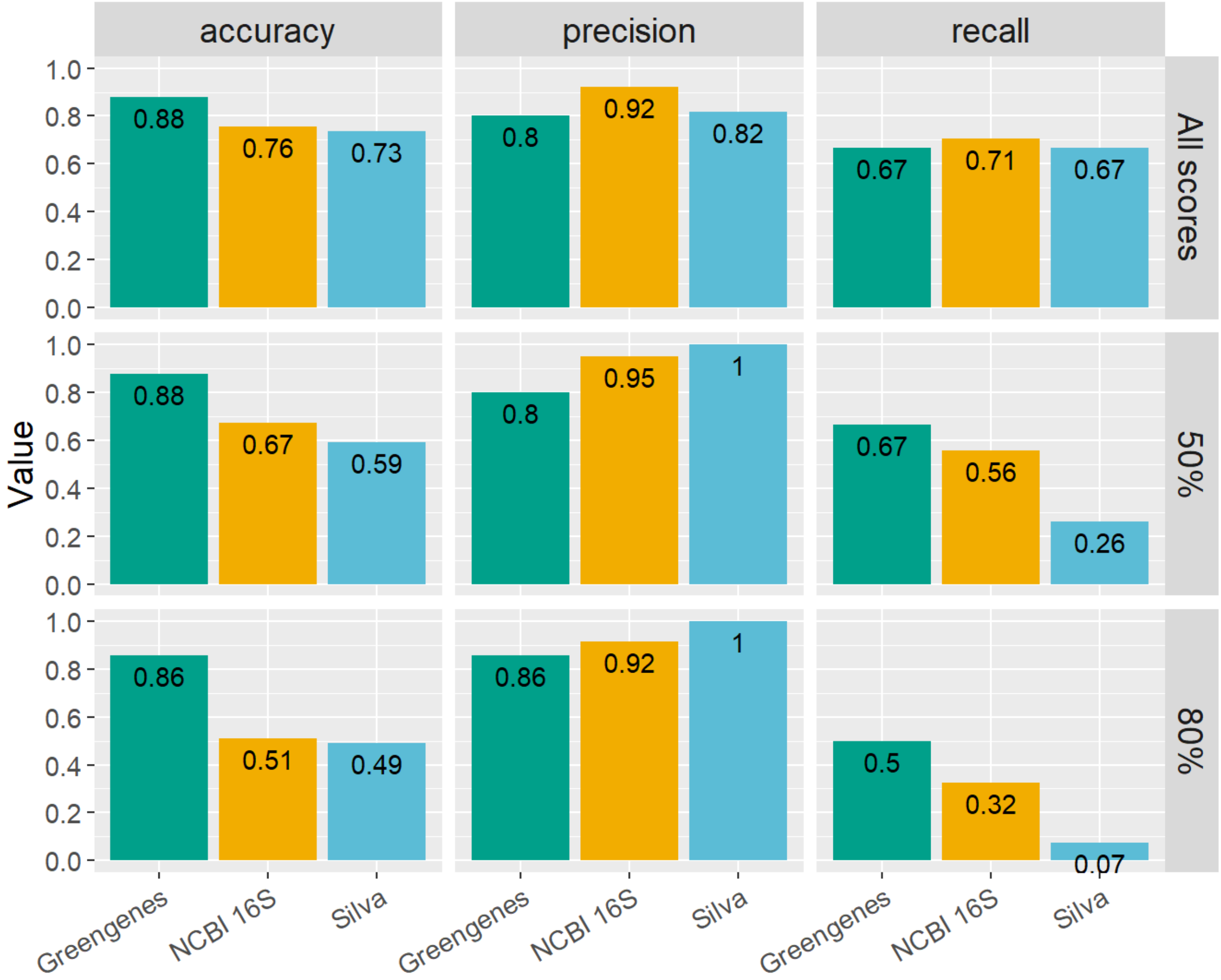
Supplemental Figure 2. F-measure values for all classification schemes. Values shown are for confidence scores of 50%, 80%, and when confidence scores are ignored.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.27.358408; this version posted October 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



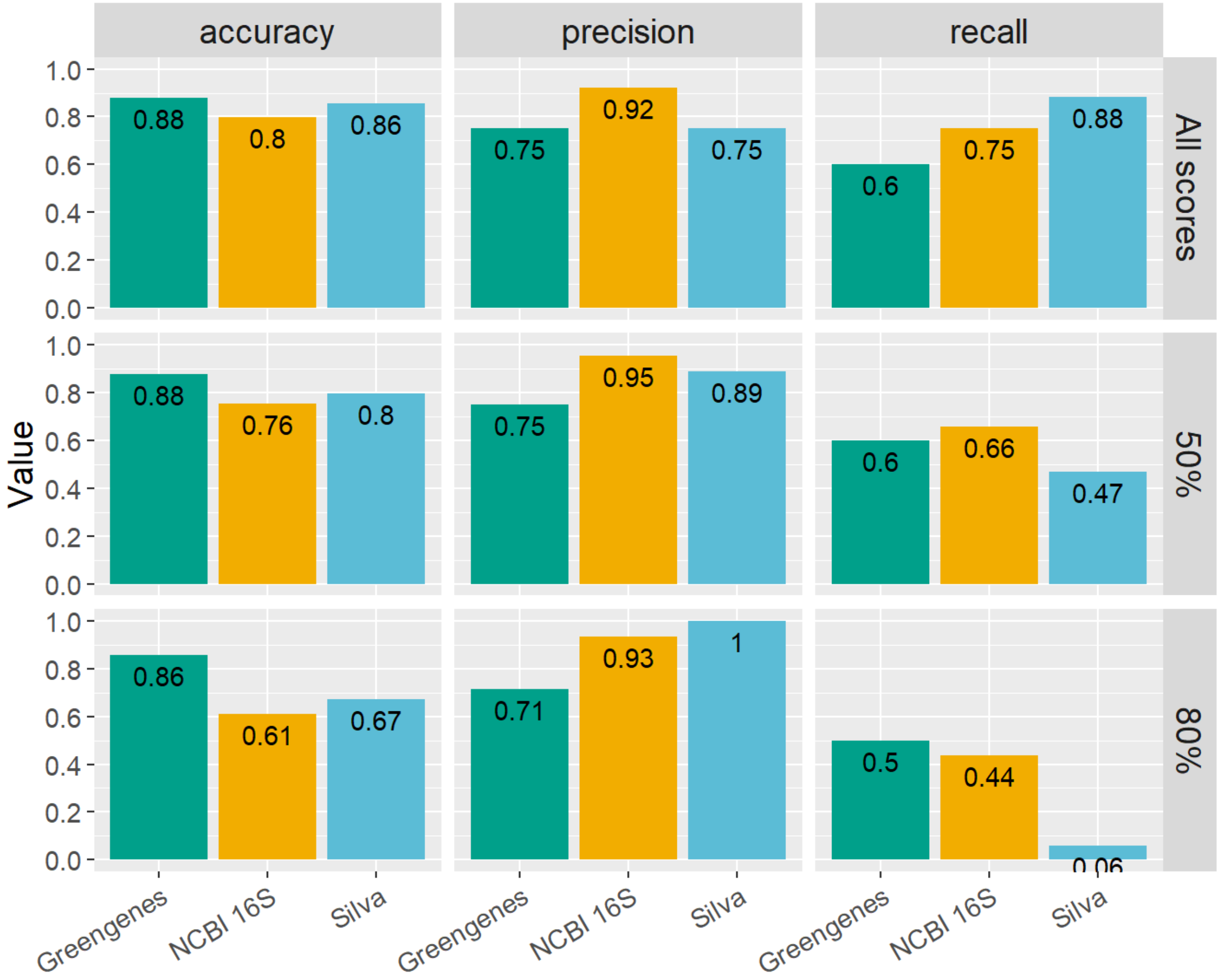
701 **Supplemental Figure 3.** Bladder bacterial species identified by database, variable region, and
702 BLCA classifier.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.27.358408; this version posted October 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplemental Figure 4. Values for accuracy, precision and recall (columns) when assigning taxonomy with the V4 identifier, Naive Bayes classifier and all databases. Rows are values when ignoring confidence scores, and when using confidence scores of 50% or 80% as thresholds.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.27.358408; this version posted October 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplemental Figure 5. Values for accuracy, precision and recall (columns) when assigning taxonomy with the V4 identifier, BLCA classifier and all databases. Rows are values when ignoring confidence scores, and when using confidence scores of 50% or 80% as thresholds.