




Species-Level Resolution of Female Bladder Microbiota from 16S rRNA Amplicon Sequencing

 Carter Hoffman,^a Nazema Y. Siddiqui,^b Ian Fields,^c W. Thomas Gregory,^c Holly M. Simon,^d Michael A. Mooney,^a  Alan J. Wolfe,^e  Lisa Karstens^{a,c}

^aDivision of Bioinformatics and Computational Biomedicine, Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, Oregon, USA

^bDivision of Urogynecology and Reconstructive Pelvic Surgery, Department of Obstetrics and Gynecology, Duke University, Durham, North Carolina, USA

^cDivision of Urogynecology, Department of Obstetrics and Gynecology, Oregon Health & Science University, Portland, Oregon, USA

^dAnimalBiome, Oakland, California, USA

^eDepartment of Microbiology & Immunology, Loyola University Chicago, Maywood, Illinois, USA

ABSTRACT The human bladder contains bacteria, even in the absence of infection. Interest in studying these bacteria and their association with bladder conditions is increasing. However, the chosen experimental method can limit the resolution of the taxonomy that can be assigned to the bacteria found in the bladder. 16S rRNA amplicon sequencing is commonly used to identify bacteria in urinary specimens, but it is typically restricted to genus-level identification. Our primary aim here was to determine if accurate species-level identification of bladder bacteria is possible using 16S rRNA amplicon sequencing. We evaluated the ability of different classification schemes, each consisting of combinations of a reference database, a 16S rRNA gene variable region, 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 Accurate species-level identification from culture-independent techniques is of importance for microbial niches that are less well characterized, such as that of the bladder. 16S rRNA amplicon sequencing, a common culture-independent way to identify bacteria, is often critiqued for lacking species-level resolution. Here, we extensively evaluate classification schemes for species-level bacterial annotation of 16S amplicon data from bladder bacteria. Our results show that the proper choice of taxonomic database and variable region of the 16S rRNA gene sequence makes species level identification possible. We also show that this improvement can be achieved through the more careful application of existing methods and resources. Species-level information may deepen our understanding of associations between bacteria in the bladder and bladder conditions such as lower urinary tract symptoms and urinary tract infections.

KEYWORDS 16S sequencing, algorithms, bioinformatics, bladder, microbiome, taxonomic classification, urobiome


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 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 urobiome diversity is associated with urinary tract infections (8–10), neurogenic bladder (7, 11, 12), bladder

Citation Hoffman C, Siddiqui NY, Fields I, Gregory WT, Simon HM, Mooney MA, Wolfe AJ, Karstens L. 2021. Species-level resolution of female bladder microbiota from 16S rRNA amplicon sequencing. *mSystems* 6:e00518-21. <https://doi.org/10.1128/mSystems.00518-21>.

Editor Nicholas Chia, Mayo Clinic

Copyright © 2021 Hoffman et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Lisa Karstens, karstens@ohsu.edu.

 Curious about species level information of the bladder microbiome? A recent study from Hoffman et al. determines components needed for species-level identification from 16S gene amplicon sequencing.

Received 29 April 2021

Accepted 18 August 2021

Published 14 September 2021

cancer (13, 14), and urgency urinary incontinence (UUI) (4, 15). 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 the 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 (16). However, culturing specific bacterial species is 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 (17, 18). 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 (19).

Targeted amplicon sequencing of the 16S rRNA gene is a common method for identifying bladder bacteria in a large-scale manner (20, 21). This is particularly true in urine, where the low microbial biomass (9, 22) often results in smaller bacterial DNA quantities than what is required for shotgun metagenomic sequencing (23). When performing targeted amplicon sequencing, DNA is extracted from all cells in a sample. Next, PCR is used to amplify a small segment of the bacterial genome, typically spanning one or more variable regions of the 16S rRNA gene. This amplified DNA segment is subject to high-throughput sequencing, such as Illumina MiSeq or HiSeq short-read technology. Bioinformatics are used to process the resulting sequences and identify the taxonomy of the bacteria (24).

When identifying bacteria using targeted amplicon sequencing, there are three important components (Fig. 1). These components are (i) a database of DNA sequences annotated with taxonomic information; (ii) the identifier, or DNA sequence, of the unknown bacterium; and (iii) 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 (25). One common classification scheme that can be implemented through the use of the mothur (26) or QIIME2 (27, 28) bioinformatic pipelines uses the Silva database (29), the V4 region from the 16S rRNA gene sequence as the identifier (30), and the naive Bayes algorithm (28) as the classifier. A limitation of this and of many other commonly used classification schemes is that the taxonomic resolution is usually constrained to the genus level. Recently, several new approaches to sequence processing and taxonomy assignment have become available (e.g., amplicon sequence variant algorithms such as DADA2 [31], and taxonomic classifiers such as Bayesian lowest common ancestor [BLCA] [32]), which may improve resolution to the species level.

Our primary aim was to determine if species-level identification of bladder bacteria is possible using 16S rRNA amplicon sequencing studies. To achieve this aim, we used known DNA sequences from a representative sample of bacteria found in the human female bladder, published by Thomas-White and colleagues (33). This data set includes bacterial species that have been isolated from the bladder, cultured, and thoroughly characterized using WGS. We computationally tested multiple classification schemes to see which schemes identified these known DNA sequences at the species level. We evaluated several reference databases, variable regions (i.e., potential identifiers), and taxonomic classification algorithms for their ability to accurately identify bladder bacteria at the species level.

RESULTS

Representation of bladder bacteria in 16S rRNA gene sequence databases. The Thomas-White genome sequencing data set consists of 149 bladder bacterial isolates,

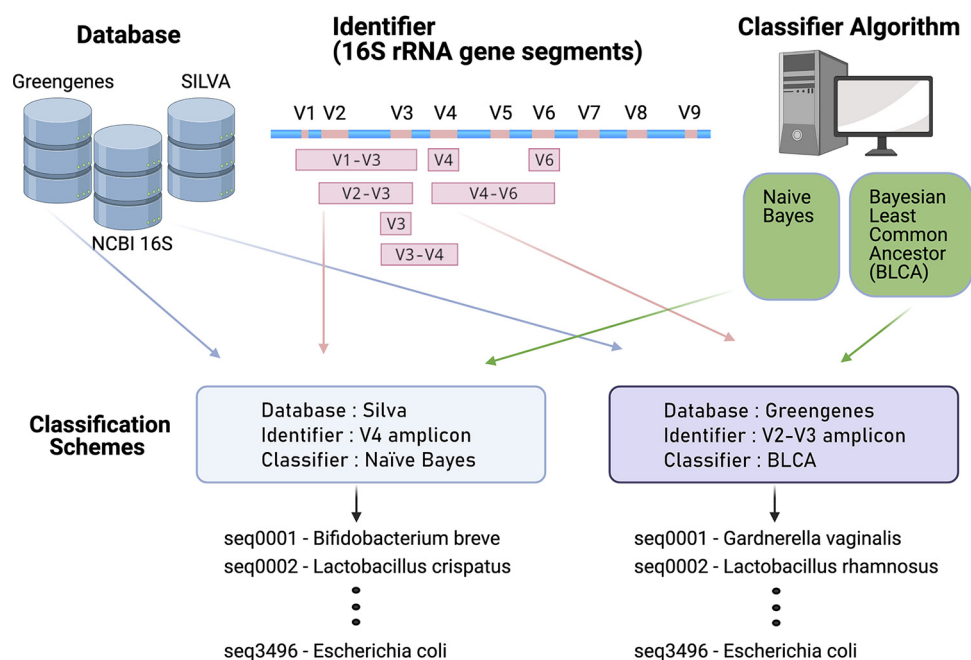


FIG 1 Model of the components that make up a classification scheme to assign taxonomy to unknown sequences. A classification scheme consists of a database, an identifier, and a classifier. The databases used in this study are the Greengenes, Silva, and NCBI 16S databases. The identifiers used in this 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 classifiers used in this study are the naive Bayes and Bayesian lowest common ancestor (BLCA) algorithms. One example of a classification scheme is the Silva database, V4 region identifier, and naive Bayes classifier. Another example of a classification scheme is the Greengenes database, V1-V2 region identifier, and BLCA classifier. These two examples are distinct from each other and can have different outcomes when assigning taxonomy.

representing 78 bacterial species from 36 genera (33). There are several databases available for bacterial identification using the 16S rRNA gene (34, 35). Of these, the Greengenes (v.13_5) (36), Silva (v. 132) (37), and NCBI 16S Microbial (v. August 2019) databases were evaluated due to their widespread use in amplicon sequencing studies and the availability of species-level annotation. At the genus level, all but one genus from the Thomas-White data set (*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 (see Fig. S1 and Table S1 in the supplemental material).

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 platforms. To evaluate the performance of different variable regions as identifiers, sequences of the 16S rRNA gene from each species in the Thomas-White data set were downloaded from Silva (see Table S1), and amplicons were computationally generated for the V1-V3, V2-V3, V3-V4, V4-V6, V3, V4, and V6 variable regions using published primers (see Materials and Methods and Fig. 2). These computational amplicons 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 Fig. 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-

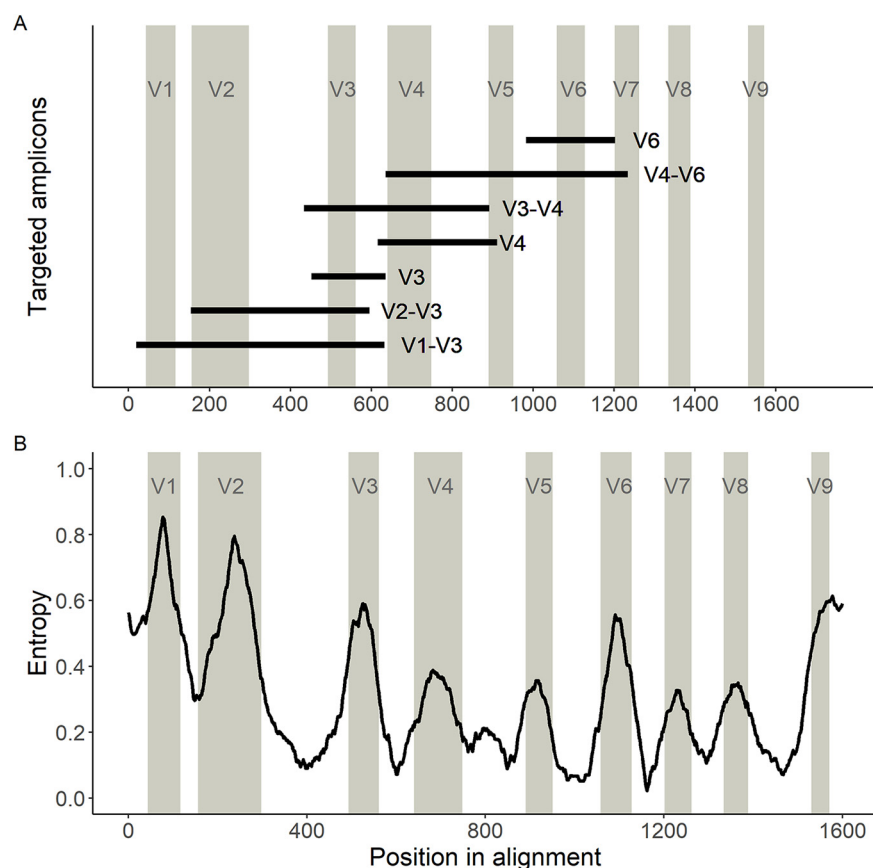


FIG 2 Variable regions of the 16S rRNA gene used in this study. (A) Locations of the primers used in this study on the 16S rRNA gene. Locations of the predicted amplicons are shown as black bars in relation to the multiple-sequence alignment (MSA) of the bacterial species described in Thomas-White et al. (33). Gray columns are the locations of the known variable regions based on the sequence from *Escherichia coli*. (B) Information about 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.

window analysis on a multiple-sequence alignment (MSA) of all 16S rRNA gene sequences from the Thomas-White data set. We calculated entropy as a measure of information content along the MSA (Fig. 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 Materials and Methods). For these calculations, we compared results from computational classification schemes to known bacterial species information from prior WGS studies (33). Each taxonomic classification was evaluated as a true match, true nonmatch, false match, or false nonmatch 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 (38).

In general, classification schemes that use the NCBI 16S Microbial database

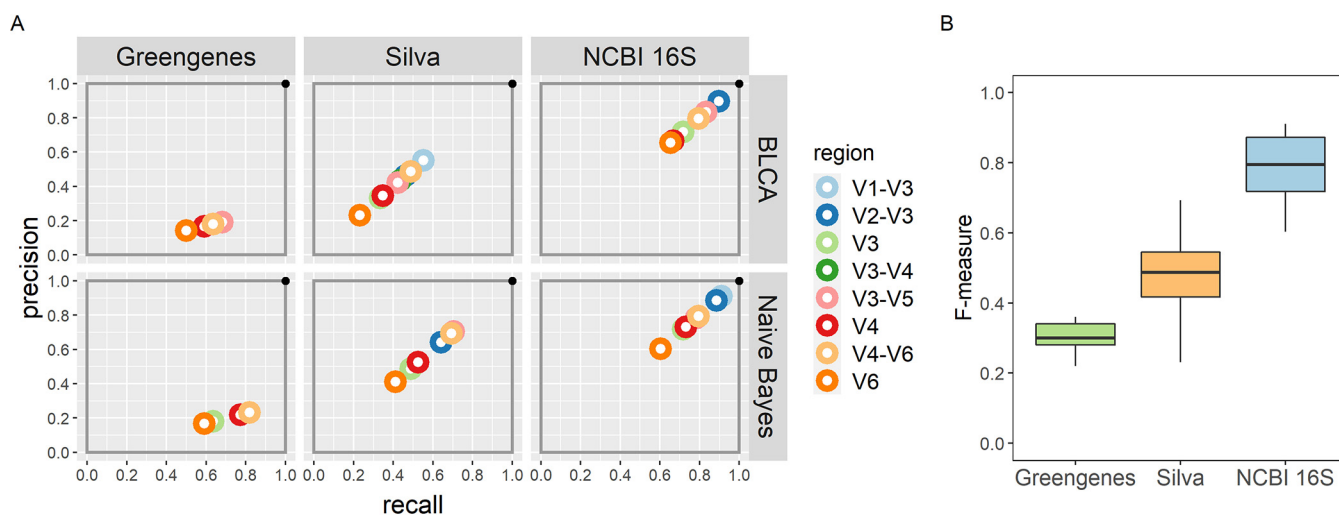


FIG 3 Overall classification scheme evaluation. (A) The performance of each classification scheme is summarized by the precision (y axis) and recall (x axis) for each variable region (color). The best classification scheme would lie in the upper right-hand corner. (B) Box plots depicting the median, quartiles, and range of the F-measure of all classification schemes by database. Overall, classification schemes using the NCBI 16S Microbial database had the highest F-measures, and those using the Greengenes databases had the lowest.

perform the best because they have high recall and precision and consequently the highest F-measures (median, 0.79; range, 0.60 to 0.91), regardless of which classifier is used (Fig. 3). Classification schemes using the Silva database show reduced recall and precision and therefore lower F-measures (median, 0.48; range, 0.23 to 0.69). 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 but low precision, resulting in a low F-measure (median, 0.30; range, 0.22 to 0.36).

While the choice of database appears to have the largest effect on the outcome of a classification scheme, the combinations of identifiers and classifiers with specific databases show differences as well. For example, when using the Silva database, the classifier has an impact on the classification scheme outcome. With the naive Bayes classifier, the identifiers yielding the highest F-measures are the V3-V4 (0.71) and V4-V6 (0.70) targeted amplicons. Using the BLCA classifier, the V1-V3 targeted amplicon has the highest F-measure (0.55).

Confidence scores affect classification. The BLCA and naive 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. This confidence score measures how much the classification changes through random permutation (bootstrapping), and the value reflects the “goodness of fit” of that classification (39). When lacking any knowledge of how to choose the best confidence score that minimizes the number of errors of a given classification scheme, using a predefined confidence score threshold is an option (40, 41). We evaluated the performance of classification schemes against our test set when confidence score thresholds were used. For this analysis, matches that returned with confidence scores less than the designated confidence threshold were considered nonmatches. 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% (Fig. S2A and Fig. 4). This effect is especially marked for the classification schemes that use the Silva database, which shows an average 79.3% reduction in recall with higher confidence scores (Fig. 4A). Classification schemes that use the NCBI 16S database are unequally affected by increasing confidence scores; 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 only slightly

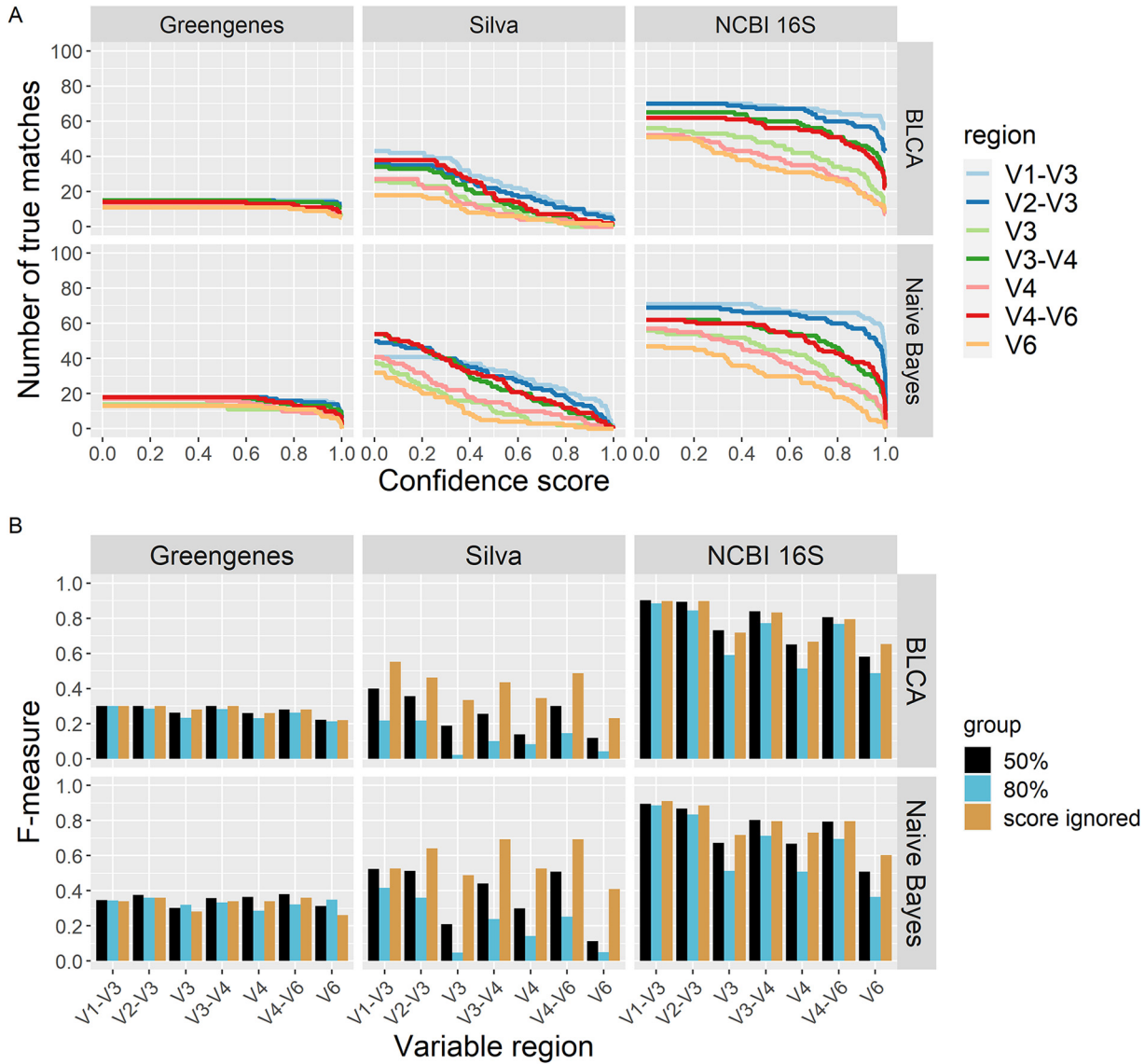


FIG 4 The effect of confidence score on classification scheme outcomes. (A) Number of true matches returned for each classification scheme across all confidence score values. For all classification schemes, as the confidence score value is increased the number of true matches decreases. For Greengenes and NCBI 16S databases, true matches dramatically decrease with confidence scores higher than 80%. With schemes using the Silva database, true matches steadily decrease. (B) F-measure values for all classification schemes. Values shown are for confidence scores of 50% and 80% and when confidence scores are ignored. Classification schemes that use the Silva database show the greatest change in F-measure values when different thresholds are used.

affected, with different confidence scores resulting in smaller-magnitude differences in recall.

Changes in the precision of classification schemes are mostly affected by the database used (Fig. S2), although confidence scores can further affect precision. For the classification schemes that use the NCBI 16S database, precision is generally improved with increasing confidence score, but at unequal amounts. For example, using the 80% threshold, the V1-V3 identifier shows a slight average increase of 3.5%, while the V6 identifier shows a large average increase of 39.7% (Fig. S2A). Classification schemes that use the Greengenes database show only minimal changes in precision. In contrast, classification schemes using the Silva database are differentially affected, with both reduction and gains in precision depending on the identifier and classifier combinations that are chosen. In general, for the Silva database, precision is reduced when using any 16S rRNA identifier with the BLCA classifier. However, a dramatic increase in

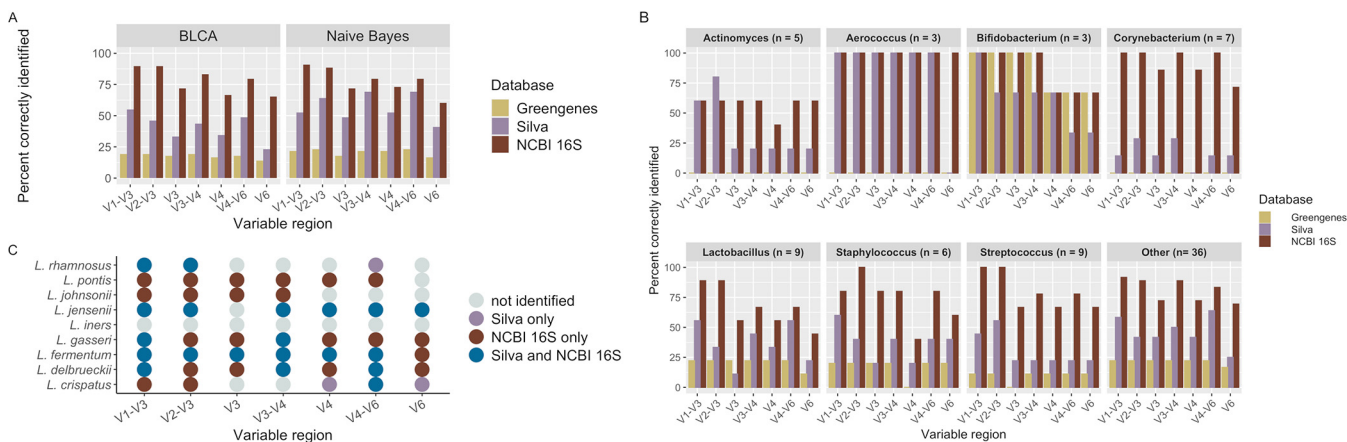


FIG 5 Percentage of bladder bacteria correctly identified for each classification scheme. (A) Overall results by classifier. With the commonly used V4 variable region and BLCA classifier, 17% of bladder bacteria are correctly identified using the Greengenes database, compared with 35% correctly identified using the Silva database and 67% using the NCBI 16S database. A similar trend is seen with the naive Bayes classifier. Using other variable regions can lead to improved species-level resolution to a maximum of 91% correctly identified. (B) Percentage of bladder bacterial species identified with each database by variable region (BLCA). Genera with fewer than 3 species are grouped as “other.” (C) Identification of *Lactobacillus* species depends on the variable region and database used, with the NCBI 16S database having the most coverage, regardless of the variable region chosen. The Greengenes database is not shown, since it only classified two species (*Lactobacillus pontis* and *Lactobacillus delbrueckii*).

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 this increases to 85.7% when a confidence score threshold of 80% is used (Fig. S2).

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 Fig. 4B. In almost every classification scheme, the F-measure value decreases when a threshold is used. The classification schemes that use the Silva database clearly demonstrate this effect and show a 66.9% reduction in F-measure values, on average, when higher confidence scores, such as 80%, are used. 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 (0.78 to 6.78%) in their F-measure values, regardless of whether a threshold is used or not.

Amplicons spanning more than one variable region identify a larger 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 naive 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 (Fig. 5A). In contrast, 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 when using the NCBI 16S database.

Species identified depends on choice of database and variable region. While the results thus far have focused on summarizing the 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 (Fig. 5B and Fig. S3). Although the NCBI 16S database contains the greatest representation of bladder species, some species were not identified using certain variable regions, if at all. For example, *Lactobacillus* species, which are commonly found in the female bladder and vagina (33, 42, 43), were overall best represented within the NCBI 16S database, with 8 out of 9 species being identified with the V1-V3 and V2-V3 variable regions (Fig. 5C). However, the other variable regions only identified between 4 and 6 *Lactobacillus* species when using the NCBI 16S database. Interestingly, *Lactobacillus*

TABLE 1 Cohen's kappa values demonstrate moderate agreement of the species identified through computational and experimental amplicon data sets

Scheme	Cohen's kappa
V4-Silva-BLCA	0.70
V4-Greengenes-BLCA	0.59
V4-NCBI 16S-BLCA	0.58
V4-Silva-naive Bayes	0.47
V4-Greengenes-naive Bayes	0.65
V4-NCBI 16S-naive Bayes	0.50

crispatus and *Lactobacillus rhamnosus* were only detectable with specific combinations of variable regions and databases, and *Lactobacillus iners* was not correctly identified from our data set 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 (44) (Fig. S3). Several bladder species, such as *Gardnerella vaginalis*, were only detected with the NCBI 16S and Silva databases. *Staphylococcus* species were poorly identified with the V4 region but were distinguishable with all other regions. *Streptococcus* and *Corynebacterium* species were best identified with the NCBI 16S database. *Escherichia coli*, a common uropathogen (45), is not well represented in any of the databases and was only detected with the V4 region and the NCBI 16S database.

Validation of computational findings on V4 amplicon data. To evaluate the performance of our computational findings on experimental 16S rRNA amplicon 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 data set 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, 46). We reprocessed the raw sequencing data (see Materials and Methods) and performed taxonomic classification to assess the performance of our computational findings. Since 16S rRNA amplicon 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 (see Fig. S4). We used Cohen's kappa to assess the agreement between species assignment in the V4 computational data set to the experimental V4 data set (see Materials and Methods). All classification schemes had moderate agreement (Table 1).

DISCUSSION

Our study demonstrates that it is possible to attain results at the species level with existing resources when performing targeted amplicon sequencing of urinary specimens. Although species-level taxonomic resolution is possible, it requires a carefully chosen classification scheme. Within the classification scheme, we found that the classifier chosen did not drastically affect the identification of species, but the database and identifier do. Importantly, we found that the reference database strongly influences the identification of bacteria at the species level. Overall, using our approach, the NCBI 16S database performed the best, whereas the Greengenes database performed the worst.

In our study, we compared several combinations of classification scheme components in the context of the female bladder microbiome to identify bacterial species. This builds on several prior studies comparing databases and classifiers (47–49), identifiers and databases (50–52), classifiers and identifiers (53, 54), or individual components of the classification scheme for application in microbiome studies. However, to our knowledge, there have been no studies that systematically compared combinations of all the components of a classification scheme (classifiers, databases, and identifiers) for species-level assignment from 16S amplicon experiments. This is an

advancement over previous studies in that we were able to comprehensively show that certain combinations of identifier, classifier, and database perform better than others, particularly on bladder bacteria.

While database selection is recognized as an important decision in taxonomy assignment for amplicon sequencing studies, few studies have compared the effect of different databases on taxonomy assignment (47, 49, 55, 56), and none have focused on bladder bacterial species. Park and Won recently compared database performance on mock microbial communities and similarly found that the Silva database outperformed the Greengenes database (55). Several studies have also demonstrated that niche-specific databases perform the best for species-level identification (57), an approach that has been taken in research on many microbiome communities, such as the vaginal (58), oral (59), and gut (60) microbiomes.

For species-level taxonomy assignments, the reference database must contain species-level information of the bacteria to be identified. 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. Even though both of these databases contain species of interest, the NCBI 16S database consistently outperformed the Silva database on the computational data set.

The 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. While these two approaches are different in concept, we did not find significant differences in their performance for species-level classification of bladder bacteria on the computational data set. 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 a highly effective approach; however, until recent advances in computer technology, it remained impractical because 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. It is a commonly used classifier, implemented in Ribosomal Database Project (RDP), QIIME2, mothur, and DADA2 bioinformatics workflows. 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, leading to preferential classification of taxa with more training examples (61). The second limitation is that all features (i.e., the *k*-mers generated from the DNA sequences) are assumed to be independent, an assumption that is violated with the palindromic structures found in mRNA. Naive Bayes will favor classifications into taxa containing dependent *k*-mers over independent *k*-mers (61). A third limitation was that the available naive Bayes classifiers (such as those available through the RDP classifier [62] and QIIME2 [28]) only had genus-level information for the Silva and NCBI 16S databases. To use these for species-level assignment, it was necessary to retrain the classifier to use the specified databases at the species level. This training converts the DNA sequences to the calculated frequency that each *k*-mer occurs in a taxon.

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 of up to 300 nucleotides in length. Until full-length 16S rRNA gene sequencing can be achieved affordably on a large scale (e.g., 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 have experimental limitations (63), such as difficulties combining partial reads if portions of those reads do not overlap; these limitations were not assessed in our study. 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 if the NCBI 16S database was used (52 out of 78). These shorter amplicons are widely used with Illumina sequencing and are valid, depending on the study design and level of precision desired.

This study adds to a growing body of literature informing and improving experimental design for urobiome research. Several studies have now informed on best practices (64) and how method of sample collection (2, 65), sample handling and storage (66), sample chemical composition (67), and DNA extraction protocol (67, 68) can influence downstream results and interpretations. By taking a computational approach to evaluating classification schemes, we were able to thoroughly assess the ability of various combinations of databases, identifiers, and classifiers to identify known bladder bacterial species. However, there are several practical limitations of amplicon sequencing that were not captured in this approach. We used published primers to create computational amplicons, but this may not reflect the efficiencies of primer binding in actual experiments. Additionally, targeted amplicon sequencing generates a large number of overlapping reads and provides the data for methods to correct for errors introduced in the lab by the polymerase enzyme and through sequencing. Our method did not account for the noise that is found in experimental data, which may decrease the ability of the classification schemes to differentiate between bacterial species with similar sequences. Furthermore, our results may not extend to bladder samples that contain novel bacterial species that are not well represented in the existing databases.

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 among those that we tested, the NCBI 16S database allows for the best identification of bladder species. Our validation with V4 amplicon data demonstrates that the predicted computational outcomes are a reasonable approximation for how a classification scheme will perform on experimental data. 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, and we recommend that experimental validation be performed.

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, although this approach is complicated by differences in databases in terms of formatting, as well as conflicting classifications (69). Both of these components are important for planning experimental and computational aspects of urobiome studies and should be considered when comparing results across studies.

MATERIALS AND METHODS

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

The Thomas-White data set. The 78 species of bladder bacteria used in this study were identified by culturing 149 bacterial isolates from 77 urine samples and performing whole-genome sequencing, as described in Thomas-White et al. (33). This set of identified species served as the basis for our

computational analysis and is referred to as the Thomas-White data set. 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 (see Table S1 in the supplemental material). A type strain is defined by the sequence of the cultured isolate that was subject to the metabolic, genotypic, and phenotypic evaluations taken to define the bacterial species (70) and is the agreed bacterial organism to which the taxonomic name refers. Sequences were searched using the “[T]” filter setting, and sequences longer than 1,450 nucleotides (nt) with alignment and pntail quality scores greater than 95% were selected. For the species that had no hits, the taxonomic synonym (see below, “Synonyms of species”) 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 data set. Targeted amplicon sequences from 24 urine samples, using the V4 region of the 16S rRNA gene sequence, are referred to as the V4 validation data set. These 24 urine samples originated from a subsample of the women whose samples comprised the Thomas-White data set. Sequencing data were generated as part of two other published studies using Illumina sequencing of the V4 region of the 16S rRNA gene (4, 46). The raw sequence reads were processed with DADA2 v. 1.14.1 (31) 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 (PNU) (71) website on 5 January 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* subsp. *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 v. 13_5 was downloaded on 23 September 2019 from (https://greengenes.secondgenome.com/?prefix=downloads/greengenes_database/gg_13_5/). The Silva database v. 132 was downloaded on 14 September 2019 from https://www.arb-silva.de/no_cache/download/archive/release_132/Exports/ as a FASTA format file. The NCBI 16S Microbial database is bundled with the BLCA package and is also available from <ftp://ftp.ncbi.nlm.nih.gov/blast/db/>. For use with BLCA, each database was processed with provided scripts (“1_subset_db_gg.py” from BLCA for Greengenes; “makeblastdb utility” from the BLAST+ suite combined with custom “write_taxonomy.py” for Silva, and “1_subset_db_acc.py” from BLCA for NCBI 16S Microbial). For use with naive Bayes, the FASTA file for each database was reformatted to work with QIIME2 using the custom “write_qiime_train_db.py” script. The naive Bayes classifier was then trained with the “fit-classifier-naive-bayes” script QIIME2 script.

Presence of Thomas-White species in databases. To verify that all species from the Thomas-White data set 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 data set. 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 data set were formed into a multisequence alignment using the T-coffee program (72). T-coffee v. 12.00.7fb08c2 was downloaded from <http://tcoffee.org/Packages/Stable/Latest/> on 5 April 2019. Alignments were performed using the default settings.

Sliding-window analysis. Comparison of the 16S rRNA gene sequences of the species in the Thomas-White data set reveals regions of conserved sequence and regions of variability. The degree to which 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 is 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 an MSA is achieved by combining SWA with calculation of 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 calculation 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 that the Shannon entropy will interpret these indel regions as conserved sequence. This limitation was overcome by weighting the entropy scores against gaps (73). 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.

Primers. Amplicons were computationally generated from the Thomas-White genome sequencing data set for the V1-V3, V2-V3, V3-V4, V4-V6, and V4 variable regions using published primers and for the V3 and V6 regions using designed primers. The primer sequences used, listed in order of amplicon spanning the variable region(s), forward primer name and sequence, and reverse primer name and sequence are as follows: V1-V3, A17F 5'-GTT TGA TCC TGG CTC AG-3', 515R 5'-TTA CCG CGG CMG CSG GCA-3' (74, 75); V2-V3, 16S_BV2f 5'-AGT GGC GGA CCG GTG AGT AA-3', HDA-2 5'-GTA TTA CCG CGG CTG CTG GCA C-3' (76, 77); 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' (50); V4, F515 5'-GTG CCA GCM GCC GCG GTA A-3', R806 5'-CCT GAT GHV CCC AWA GAT TA-3' (18); V4-V6, 519F 5'-GTG CCA GCT GCC GCG GTA ATA-3', 1114R 5'-GGG GTT GCG CTC GTT GC-3' (75); andV6, v6_1183F 5'-

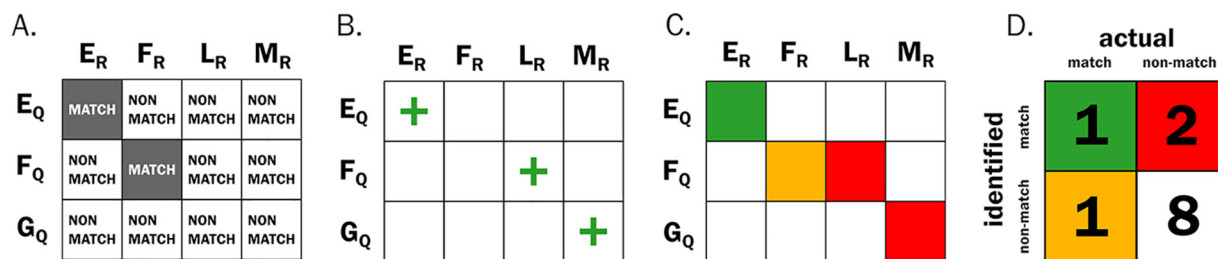


FIG 6 Example of classification evaluation used in this study. Suppose there is a classification scheme comprising a set of query sequences (rows E, F, and G) and the set of reference sequences (columns E, F, L, and M) held in a reference database. In this example, the number of reference records is 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 a match. If they are different, they are designated a nonmatch. (B) Next, the classifier is allowed to assign record pairs as matches or nonmatches for all query sequences, represented as green plus signs for matches and blank cells as nonmatches. 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 nonmatch (yellow square), and 8 true nonmatches (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 $1/(1 + 2) = 0.33$, recall is $1/(1 + 1) = 0.5$, and the F-measure is $(2 \times 0.33 \times 0.5)/(0.33 + 0.5) = 0.40$.

CCG CCT GGG GAS TAC GVH-3', v6_1410R 5'-AGT CCC RYA ACG AGC GCA-3'. Degenerate primer design with DegePrime (78) (<https://github.com/EnvGen/DEGEPRIME.git>) 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 data set as possible. DegePrime has the option to ignore columns of a MSA if the number of "-" characters exceeds 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 (GenBank accession number [EU014689.1](https://www.ncbi.nlm.nih.gov/nuccore/EU014689.1)) 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 data set. This procedure was done using the custom "extract_16s_vr.py" script and output as a multirecord FASTA formatted file.

Taxonomic classifiers. Taxonomic classification was performed with Bayesian lowest common ancestor (BLCA) and naive Bayes classifiers. BLCA (32) was cloned from the GitHub repository <https://github.com/qunfengdong/BLCA.git>. For the 16S variable regions, the BLCA classifier was run using default settings but pointing to the selected reference database, either Greengenes, Silva, or NCBI 16S. The naive Bayes classifier as implemented by QIIME2 (28) was used with the Greengenes, Silva, and NCBI 16S databases and confidence settings 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 data set, 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 the sequence held in the reference database) from the classification scheme to the known identity of the query sequence from the Thomas-White data set. All record pairs that were assigned a match by the classification schemes were evaluated according to the following definitions (Fig. 6):

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.

False nonmatch: if a record representing a species in the Thomas-White data set was present in the database, but was not assigned as a match, the record was evaluated as a false nonmatch.

True nonmatch: all records in the reference database that were not in the Thomas-White data set. While records assigned to this category were not used in evaluating the classification schemes in this study, the definition is still included for completeness.

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 nonmatches). 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 (79).

Evaluating V4 validation results. To determine the expected bacterial species in each sample, the results of the whole-genome sequencing of the isolates cultured from the corresponding subject was used. The species of bacteria in the V4 sequencing data were identified using classification schemes composed of (i) the Greengenes, Silva, and NCBI 16S microbial databases, (ii) the V4 sequencing results as the identifier, and (iii) the BLCA and naive Bayes classifiers. For each classification scheme, the species identification from the computational and validation data sets was scored as a "1" if correct, and "0" otherwise. This resulted in a table for each classification scheme consisting of bacterial species as the rows and the scores from the computational or validation outcomes as columns. Cohen's kappa was then

calculated for each table using the “kappa2” function from the *irr* R package (<https://CRAN.R-project.org/package=irr>).

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 at <https://github.com/KarstensLab/BladderBacteriaSpecies>.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, PDF file, 0.4 MB.

FIG S2, PDF file, 0.4 MB.

FIG S3, TIFF file, 17.5 MB

FIG S4, PDF file, 0.3 MB.

TABLE S1, DOCX file, 0.02 MB.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health (NIH) by NIDDK award number K01DK116706 (to L.K.); NCAT award number UL1TR002369 (to L.K.); and NIA award numbers R03AG060082 and P30AG028716 (to N.Y.S.).

The contents of the article are solely our responsibility and do not represent the official views of the NIH or of any other funding agency.

C.H. and L.K. conceived and designed experiments and wrote the manuscript. C.H. performed analyses and L.K. reviewed analyses. A.J.W. supplied the data. L.K., C.H., M.A.M., H.M.S., W.T.G., I.F., N.Y.S., and A.J.W. interpreted results and revised the manuscript.

We acknowledge our late colleague Mark Asquith, Ph.D., who passed away in June of 2019 shortly after the start of this project. Dr. Asquith was a dedicated Assistant Professor in the Department of Medicine at Oregon Health & Science University, with a passion for research pertaining to the human microbiome. Initial discussions with Dr. Asquith about species-level identification of bacteria from 16S rRNA gene sequencing studies helped to shape this project. We also thank the anonymous reviewers who provided critiques that helped strengthen the manuscript.

REFERENCES

- Lederberg J, McCray AT. 2001. ‘Ome sweet ‘omics—a genealogical treasury of words. *Scientist* 15:2.
- Wolfe AJ, Toh E, Shibata N, Rong R, Kenton K, FitzGerald M, Mueller ER, Schreckenberger P, Dong Q, Nelson DE, Brubaker L. 2012. Evidence of uncultivated bacteria in the adult female bladder. *J Clin Microbiol* 50: 1376–1383. <https://doi.org/10.1128/JCM.05852-11>.
- Khasriya R, Sathiananthamoorthy S, Ismail S, Kelsey M, Wilson M, Rohn JL, Malone-Lee J. 2013. Spectrum of bacterial colonization associated with urothelial cells from patients with chronic lower urinary tract symptoms. *J Clin Microbiol* 51:2054–2062. <https://doi.org/10.1128/JCM.03314-12>.
- Pearce MM, Hilt EE, Rosenfeld AB, Zilliox MJ, Thomas-White K, Fok C, Kliethermes S, Schreckenberger PC, Brubaker L, Gai X, Wolfe AJ. 2014. The female urinary microbiome: a comparison of women with and without urgency urinary incontinence. *mBio* 5:e01283-14. <https://doi.org/10.1128/mBio.01283-14>.
- Price TK, Dune T, Hilt EE, Thomas-White KJ, Kliethermes S, Brincat C, Brubaker L, Wolfe AJ, Mueller ER, Schreckenberger PC. 2016. The clinical urine culture: enhanced techniques improve detection of clinically relevant microorganisms. *J Clin Microbiol* 54:1216–1222. <https://doi.org/10.1128/JCM.00044-16>.
- Ackerman AL, Underhill DM. 2017. The mycobiome of the human urinary tract: potential roles for fungi in urology. *Ann Transl Med* 5:31–31. <https://doi.org/10.21037/atm.2016.12.69>.
- Fouts DE, Pieper R, Szpakowski S, Pohl H, Knoblach S, Suh M-J, Huang S-T, Ljungberg I, Sprague BM, Lucas SK, Torralba M, Nelson KE, Groah SL. 2012. Integrated next-generation sequencing of 16S rDNA and metaproteomics differentiate the healthy urine microbiome from asymptomatic bacteriuria in neuropathic bladder associated with spinal cord injury. *J Transl Med* 10:174. <https://doi.org/10.1186/1479-5876-10-174>.
- Thomas-White KJ, Gao X, Lin H, Fok CS, Ghanayem K, Mueller ER, Dong Q, Brubaker L, Wolfe AJ. 2018. Urinary microbes and postoperative urinary tract infection risk in urogynecologic surgical patients. *Int Urogynecol J* 29:1797–1805. <https://doi.org/10.1007/s00192-018-3767-3>.
- Moustafa A, Li W, Singh H, Moncera KJ, Torralba MG, Yu Y, Manuel O, Biggs W, Venter JC, Nelson KE, Pieper R, Telenti A. 2018. Microbial metagenome of urinary tract infection. *Sci Rep* 8:4333. <https://doi.org/10.1038/s41598-018-22660-8>.
- Neugent ML, Hulyalkar NV, Nguyen VH, Zimmern PE, De Nisco NJ. 2020. Advances in understanding the human urinary microbiome and its potential role in urinary tract infection. *mBio* 11:e00218-20. 1–15 <https://doi.org/10.1128/mBio.00218-20>.
- Groah SL, Pérez-Losada M, Caldovic L, Ljungberg IH, Sprague BM, Castro-Nallar E, Chandel NJ, Hsieh MH, Pohl HG. 2016. Redefining healthy urine: a cross-sectional exploratory metagenomic study of people with and without bladder dysfunction. *J Urol* 196:579–587. <https://doi.org/10.1016/j.juro.2016.01.088>.
- Bajic P, Wolfe AJ, Gupta GN. 2019. The Urinary microbiome: implications in bladder cancer pathogenesis and therapeutics. *Urology* 126:10–15. <https://doi.org/10.1016/j.urology.2018.12.034>.
- Zeng J, Zhang G, Chen C, Li K, Wen Y, Zhao J, Wu P. 2020. Alterations in urobiome in patients with bladder cancer and implications for clinical outcome: a single-institution study. *Front Cell Infect Microbiol* 10:555508. <https://doi.org/10.3389/fcimb.2020.555508>.
- Bučević Popović V, Šitum M, Chow CET, Chan LS, Roje B, Terzić J. 2018. The urinary microbiome associated with bladder cancer. *Sci Rep* 8:12157. <https://doi.org/10.1038/s41598-018-29054-w>.
- Karstens L, Asquith M, Davin S, Stauffer P, Fair D, Gregory WT, Rosenbaum JT, McWeeney SK, Nardos R. 2016. Does the urinary microbiome play a role in urgency urinary incontinence and its severity? *Front Cell Infect Microbiol* 6:78. <https://doi.org/10.3389/fcimb.2016.00078>.

16. Kliem M, Sauer S. 2012. The essence on mass spectrometry based microbial diagnostics. *Curr Opin Microbiol* 15:397–402. <https://doi.org/10.1016/j.mib.2012.02.006>.
17. Pace NR. 1997. A molecular view of microbial diversity and the biosphere. *Science* 276:734–740. <https://doi.org/10.1126/science.276.5313.734>.
18. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R. 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci U S A* 108:4516–4522. <https://doi.org/10.1073/pnas.1000080107>.
19. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Cresy HH, Earl AM, FitzGerald MG, Fulton RS, Giglio MG, Hallsworth-Pepin K, Lobos EA, Madupu R, Magrini V, Martin JC, Mitreva M, Muzny DM, Sodergren EJ, Versalovic J, Wollam AM, Worley KC, Wortman JR, Young SK, Zeng Q, Aagaard KM, Abolude OO, Allen-Vercocoe E, Alm EJ, Alvarado L, Andersen GL, Anderson S, Appelbaum E, Arachchi HM, Armitage G, Arze CA, Ayvaz T, Baker CC, Begg L, Belachew T, Bhonagiri V, Bihan M, Blaser MJ, Bloom T, Bonazzi V, Paul Brooks J, Buck GA, Buhay CJ, Busam DA, Campbell JL, Canon SR, Cantarel BL, Chain PSG, Chen I-MA, Chen L, Chhibba S, Chu K, Ciulla DM, Clemente JC, The Human Microbiome Project Consortium, et al. 2012. Structure, function and diversity of the healthy human microbiome. *Nature* 486:207–214. <https://doi.org/10.1038/nature11234>.
20. Karstens L, Asquith M, Caruso V, Rosenbaum JT, Fair DA, Braun J, Gregory WT, Nardos R, McWeeney SK. 2018. Community profiling of the urinary microbiota: considerations for low-biomass samples. *Nat Rev Urol* 15: 735–749. <https://doi.org/10.1038/s41585-018-0104-z>.
21. Aragón IM, Herrera-Imbroda B, Queipo-Ortuño MI, Castillo E, Del Moral JS-G, Gómez-Millán J, Yucel G, Lara MF. 2018. The urinary tract microbiome in health and disease. *Eur Urol Focus* 4:128–138. <https://doi.org/10.1016/j.euf.2016.11.001>.
22. Adebayo AS, Ackermann G, Bowyer RCE, Wells PM, Humphreys G, Knight R, Spector TD, Steves CJ. 2020. The Urinary tract microbiome in older women exhibits host genetic and environmental influences. *Cell Host Microbe* 28:298–305.e3. <https://doi.org/10.1016/j.chom.2020.06.022>.
23. Oechslin CP, Lenz N, Liechti N, Rytter S, Agyeman P, Bruggmann R, Leib SL, Beuret CM. 2018. Limited correlation of shotgun metagenomics following host depletion and routine diagnostics for viruses and bacteria in low concentrated surrogate and clinical samples. *Front Cell Infect Microbiol* 8: 375. <https://doi.org/10.3389/fcimb.2018.00375>.
24. Prodan A, Tremaroli V, Brolin H, Zwinderman AH, Nieuwdorp M, Levin E. 2020. Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PLoS One* 15:e0227434. <https://doi.org/10.1371/journal.pone.0227434>.
25. Manning C, Raghavan P, Schuetze H. 2009. Introduction to information retrieval. Cambridge University Press, Cambridge, United Kingdom.
26. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75:7537–7541. <https://doi.org/10.1128/AEM.01541-09>.
27. Navas-Molina JA, Peralta-Sánchez JM, González A, McMurdie PJ, Vázquez-Baeza Y, Xu Z, Ursell LK, Lauber C, Zhou H, Song SJ, Huntley J, Ackermann GL, Berg-Lyons D, Holmes S, Caporaso JG, Knight R. 2013. Advancing our understanding of the human microbiome using QIIME. *Methods Enzymol* 531:371–444. <https://doi.org/10.1016/B978-0-12-407863-5.00019-8>.
28. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, et al. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* 37:852–857. <https://doi.org/10.1038/s41587-019-0209-9>.
29. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glockner FO. 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 35:7188–7196. <https://doi.org/10.1093/nar/gkm864>.
30. Hugenholtz P, Skarshewski A, Parks DH. 2016. Genome-based microbial taxonomy coming of Age. *Cold Spring Harb Perspect Biol* 8:a018085. <https://doi.org/10.1101/cshperspect.a018085>.
31. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods* 13:581–583. <https://doi.org/10.1038/nmeth.3869>.
32. Gao X, Lin H, Revanna K, Dong Q. 2017. A Bayesian taxonomic classification method for 16S rRNA gene sequences with improved species-level accuracy. *BMC Bioinformatics* 18:247. <https://doi.org/10.1186/s12859-017-1670-4>.
33. Thomas-White K, Forster SC, Kumar N, Van Kuiken M, Putonti C, Stares MD, Hilt EE, Price TK, Wolfe AJ, Lawley TD. 2018. Culturing of female bladder bacteria reveals an interconnected urogenital microbiota. *Nat Commun* 9:1557. <https://doi.org/10.1038/s41467-018-03968-5>.
34. Breitwieser FP, Lu J, Salzberg SL. 2019. A review of methods and databases for metagenomic classification and assembly. *Brief Bioinform* 20: 1125–1136. <https://doi.org/10.1093/bib/bbx120>.
35. Hugerth LW, Andersson AF. 2017. Analysing microbial community composition through amplicon sequencing: from sampling to hypothesis testing. *Front Microbiol* 8:1561. <https://doi.org/10.3389/fmicb.2017.01561>.
36. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. 2006. Greengenes, a Chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 72:5069–5072. <https://doi.org/10.1128/AEM.03006-05>.
37. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 41: D590–D596. <https://doi.org/10.1093/nar/gks1219>.
38. Hand D, Christen P. 2018. A note on using the F-measure for evaluating record linkage algorithms. *Stat Comput* 28:539–547. <https://doi.org/10.1007/s11222-017-9746-6>.
39. Efron B, Tibshirani RJ. 1993. An introduction to the bootstrap. Springer US, Boston, MA.
40. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, Huttley GA, Gregory Caporaso J. 2018. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* 6:90. <https://doi.org/10.1186/s40168-018-0470-z>.
41. Claesson MJ, O'Sullivan O, Wang Q, Nikkila J, Marchesi JR, Smidt H, de Vos WM, Ross RP, O'Toole PW. 2009. Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. *PLoS One* 4:e6669. <https://doi.org/10.1371/journal.pone.0006669>.
42. Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SSK, McCulle SL, Karlebach S, Gorle R, Russell J, Tacket CO, Brotman RM, Davis CC, Ault K, Peralta L, Forney LJ. 2011. Vaginal microbiome of reproductive-age women. *Proc Natl Acad Sci U S A* 108:4680–4687. <https://doi.org/10.1073/pnas.1002611107>.
43. Martinez RM, Hulten KG, Bui U, Clarridge JE. 2014. Molecular analysis and clinical significance of *Lactobacillus* spp. recovered from clinical specimens presumptively associated with disease. *J Clin Microbiol* 52:30–36. <https://doi.org/10.1128/JCM.02072-13>.
44. Mueller ER, Wolfe AJ, Brubaker L. 2017. Female urinary microbiota. *Curr Opin Urol* 27:282–286. <https://doi.org/10.1097/MOU.0000000000000396>.
45. Flores-Mireles AL, Walker JN, Caparon M, Hultgren SJ. 2015. Urinary tract infections: epidemiology, mechanisms of infection and treatment options. *Nat Rev Microbiol* 13:269–284. <https://doi.org/10.1038/nrmicro3432>.
46. Thomas-White KJ, Hilt EE, Fok C, Pearce MM, Mueller ER, Kliethermes S, Jacobs K, Zilliox MJ, Brincat C, Price TK, Kuffel G, Schreckenberger P, Gai X, Brubaker L, Wolfe AJ. 2016. Incontinence medication response relates to the female urinary microbiota. *Int Urogynecol J* 27:723–733. <https://doi.org/10.1007/s00192-015-2847-x>.
47. Henderson G, Yilmaz P, Kumar S, Forster RJ, Kelly WJ, Leahy SC, Guan LL, Janssen PH. 2019. Improved taxonomic assignment of rumen bacterial 16S rRNA sequences using a revised SILVA taxonomic framework. *PeerJ* 7: e6496. <https://doi.org/10.7717/peerj.6496>.
48. Price T, Hilt E, Thomas-White K, Mueller E, Wolfe A, Brubaker L. 2020. The urobiome of continent adult women: a cross-sectional study. *BJOG* 127: 193–201. <https://doi.org/10.1111/1471-0528.15920>.
49. Gwak H-J, Rho M. 2020. Data-driven modeling for species-level taxonomic assignment from 16S rRNA: application to human microbiomes. *Front Microbiol* 11:570825. <https://doi.org/10.3389/fmicb.2020.570825>.
50. Graspeuntner S, Loeper N, Künzel S, Baines JF, Rupp J. 2018. Selection of validated hypervariable regions is crucial in 16S-based microbiota studies of the female genital tract. *Sci Rep* 8:9678. <https://doi.org/10.1038/s41598-018-27757-8>.
51. Conlan S, Kong HH, Segre JA. 2012. Species-level analysis of DNA Sequence data from the NIH Human Microbiome Project. *PLoS One* 7:e47075. <https://doi.org/10.1371/journal.pone.0047075>.

52. Van Der Pol WJ, Kumar R, Morrow CD, Blanchard EE, Taylor CM, Martin DH, Lefkowitz EJ, Muzny CA. 2019. *In silico* and experimental evaluation of primer sets for species-level resolution of the vaginal microbiota using 16S ribosomal RNA gene sequencing. *J Infect Dis* 219:305–314. <https://doi.org/10.1093/infdis/jiy508>.
53. Guo F, Ju F, Cai L, Zhang T. 2013. Taxonomic precision of different hyper-variable regions of 16S rRNA gene and annotation methods for functional bacterial groups in biological wastewater treatment. *PLoS One* 8:e76185. <https://doi.org/10.1371/journal.pone.0076185>.
54. Wasserstein RL, Lazar NA. 2016. The ASA's statement on *p*-values: context, process, and purpose. *Am Stat* 70:129–133. <https://doi.org/10.1080/00031305.2016.1154108>.
55. Park S-C, Won S. 2018. Evaluation of 16S rRNA databases for taxonomic assignments using a mock community. *Genomics Inform* 16:e24. <https://doi.org/10.5808/GI.2018.16.4.e24>.
56. Sierra MA, Li Q, Pushalkar S, Paul B, Sandoval TA, Kamer AR, Corby P, Guo Y, Ruff RR, Alekseyenko AV, Li X, Saxena D. 2020. The influences of bioinformatics tools and reference databases in analyzing the human oral microbial community. *Genes* 11:878. <https://doi.org/10.3390/genes11080878>.
57. Myer PR, McDanel TG, Kuehn LA, Dedonder KD, Apley MD, Capik SF, Lubbers BV, Harhay GP, Harhay DM, Keele JW, Henniger MT, Clemmons BA, Smith TPL. 2020. Classification of 16S rRNA reads is improved using a niche-specific database constructed by near-full length sequencing. *PLoS One* 15:e0235498. <https://doi.org/10.1371/journal.pone.0235498>.
58. Fettweis JM, Serrano MG, Sheth NU, Mayer CM, Glascock AL, Brooks JP, Jefferson KK, Buck GA, Vaginal Microbiome Consortium (additional members). 2012. Species-level classification of the vaginal microbiome. *BMC Genomics* 13:S17. <https://doi.org/10.1186/1471-2164-13-S8-S17>.
59. Chen T, Yu W-H, Izard J, Baranova OV, Lakshmanan A, Dewhirst FE. 2010. The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database (Oxford)* 2010:baq013. <https://doi.org/10.1093/database/baq013>.
60. Ritari J, Salojärvi J, Lahti L, de Vos WM. 2015. Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database. *BMC Genomics* 16:1056. <https://doi.org/10.1186/s12864-015-2265-y>.
61. Rennie JDM, Shih L, Teevan J, Karger DR. 2003. Tackling the poor assumptions of naive Bayes text classifiers, p 8. In *Proceedings of the Twentieth International Conference on Machine Learning*. AAAI Press, Washington, DC.
62. Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73:5261–5267. <https://doi.org/10.1128/AEM.00062-07>.
63. Abellan-Schneyder I, Machado MS, Reitmeier S, Sommer A, Sewald Z, Baumbach J, List M, Neuhaus K. 2021. Primer, pipelines, parameters: issues in 16S rRNA gene sequencing mSphere 6:e01202-20. <https://doi.org/10.1128/mSphere.01202-20>.
64. Brubaker L, Gouridine J-PF, Siddiqui NY, Holland A, Halverson T, Limeria R, Pride D, Ackerman L, Forster CS, Jacobs KM, Thomas-White KJ, Putonti C, Dong Q, Weinstein M, Lukacz ES, Karstens L, Wolfe AJ. 2021. Forming consensus to advance urobiome research. *mSystems* 6:e0137120. <https://doi.org/10.1128/mSystems.01371-20>.
65. Pohl HG, Groah SL, Pérez-Losada M, Ljungberg I, Sprague BM, Chandan N, Caldovic L, Hsieh M. 2020. The urine microbiome of healthy men and women differs by urine collection method. *Int Neurourol J* 24:41–51. <https://doi.org/10.5213/inj.1938244.122>.
66. Jung CE, Chopyk J, Shin JH, Lukacz ES, Brubaker L, Schwanemann LK, Knight R, Wolfe AJ, Pride DT. 2019. Benchmarking urine storage and collection conditions for evaluating the female urinary microbiome. *Sci Rep* 9:13409. <https://doi.org/10.1038/s41598-019-49823-5>.
67. Munch MM, Chambers LC, Manhart LE, Domogala D, Lopez A, Fredricks DN, Srinivasan S. 2019. Optimizing bacterial DNA extraction in urine. *PLoS One* 14:e0222962. <https://doi.org/10.1371/journal.pone.0222962>.
68. Ackerman AL, Anger JT, Khaliq MU, Ackerman JE, Tang J, Kim J, Underhill DM, Freeman MR, the NIH Multidisciplinary Approach to the Study of Chronic Pelvic Pain (MAPP). 2019. Optimization of DNA extraction from human urinary samples for mycobiome community profiling. *PLoS One* 14:e0210306. <https://doi.org/10.1371/journal.pone.0210306>.
69. Balvočiūtė M, Huson DH. 2017. SILVA, RDP, Greengenes, NCBI and OTT—how do these taxonomies compare? *BMC Genomics* 18:114. <https://doi.org/10.1186/s12864-017-3501-4>.
70. Rainey FA. 2011. How to describe new species of prokaryotes, p 7–14. In *Rainey F, Oren A (ed), Taxonomy of prokaryotes. Methods in microbiology*, vol 38. Elsevier, Amsterdam, The Netherlands.
71. Parte AC. 2018. LPSN—List of Prokaryotic names with Standing in Nomenclature (bacterio.net), 20 years on. *Int J Syst Evol Microbiol* 68:1825–1829. <https://doi.org/10.1099/ijsem.0.002786>.
72. Notredame C, Higgins DG, Heringa J. 2000. T-coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 302:205–217. <https://doi.org/10.1006/jmbi.2000.4042>.
73. Valdar WSJ. 2002. Scoring residue conservation. *Proteins Struct Funct Bioinforma* 48:227–241. <https://doi.org/10.1002/prot.10146>.
74. Kumar PS, Griffen AL, Moeschberger ML, Leys EJ. 2005. Identification of candidate periodontal pathogens and beneficial species by quantitative 16S clonal analysis. *J Clin Microbiol* 43:3944–3955. <https://doi.org/10.1128/JCM.43.8.3944-3955.2005>.
75. Kumar P, Brooker M, Dowd S, Camerlengo T. 2011. Target region selection is a critical determinant of community fingerprints generated by 16S pyrosequencing. *PLoS One* 6:e20956. <https://doi.org/10.1371/journal.pone.0020956>.
76. Bukin YS, Galachyants YP, Morozov IV, Bukin SV, Zakharenko AS, Zemskaya TI. 2019. The effect of 16S rRNA region choice on bacterial community metabarcoding results. *Sci Data* 6:190007. <https://doi.org/10.1038/sdata.2019.7>.
77. Walter J, Tannock GW, Tilsala-Timisjarvi A, Rodtong S, Loach DM, Munro K, Alatossava T. 2000. Detection and identification of gastrointestinal *Lactobacillus* species by using denaturing gradient gel electrophoresis and species-specific PCR primers. *Appl Environ Microbiol* 66:297–303. <https://doi.org/10.1128/AEM.66.1.297-303.2000>.
78. Hugerth LW, Wefer HA, Lundin S, Jakobsson HE, Lindberg M, Rodin S, Engstrand L, Andersson AF. 2014. DegePrime, a program for degenerate primer design for broad-taxonomic-range PCR in microbial ecology studies. *Appl Environ Microbiol* 80:5116–5123. <https://doi.org/10.1128/AEM.01403-14>.
79. Morris EK, Caruso T, Buscot F, Fischer M, Hancock C, Maier TS, Meiners T, Müller C, Obermaier E, Prati D, Socher SA, Sonnemann I, Wäschke N, Wubet T, Wurst S, Rillig MC. 2014. Choosing and using diversity indices: insights for ecological applications from the German Biodiversity Exploratories. *Ecol Evol* 4:3514–3524. <https://doi.org/10.1002/ece3.1155>.