# Choice of 16S ribosomal RNA primers impacts urinary microbiota profiling

Vitor Heidrich[1,2], Lilian T. Inoue[1], Paula F. Asprino[1], Fabiana Bettoni[1], Antonio C.H. Mariotti[3], Diogo A. Bastos[4], Denis L.F. Jardim[4], Marco A. Arap[5], Anamaria A. Camargo[1]*

[1]Centro de Oncologia Molecular, Hospital Sírio-Libanês, São Paulo, Brazil
[2]Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, São Paulo, Brazil
[3]Instituto de Ensino e Pesquisa, Hospital Sírio-Libanês, São Paulo, Brazil
[4]Centro de Oncologia, Hospital Sírio-Libanês, São Paulo, Brazil
[5]Departamento de Urologia, Hospital Sírio-Libanês, São Paulo, Brazil
*corresponding author

## Abstract

Accessibility to next-generation sequencing (NGS) technologies has enabled the profiling of microbial communities living in distinct habitats. 16S ribosomal RNA (rRNA) gene sequencing is widely used for microbiota profiling with NGS technologies. Since most used NGS platforms generate short reads, sequencing the full-length 16S rRNA gene is impractical. Therefore, choosing which 16S rRNA hypervariable region to sequence is critical in microbiota profiling studies. All nine 16S rRNA hypervariable regions are taxonomically informative, but due to variability in profiling performance for specific clades, choosing the ideal 16S rRNA hypervariable region will depend on the bacterial composition of the habitat under study. Recently, NGS allowed the identification of microbes in the urinary tract, and urinary microbiota has become an active research area. However, there is no current study evaluating the performance of different 16S rRNA hypervariable regions for urinary microbiota profiling. We collected urine samples from male volunteers and profiled their urinary microbiota by sequencing a panel of six amplicons encompassing all nine 16S rRNA hypervariable regions. After systematically comparing their performance, we show that V1V2 hypervariable regions better assess the taxa commonly present in urine samples and V1V2 amplicon sequencing is more suitable for urinary microbiota profiling. We believe our results will be helpful to guide this crucial methodological choice in future urinary microbiota studies.

**Keywords**: urobiome, urinary microbiota, bladder microbiota, 16S amplicon sequencing, 16S rRNA primers

## Introduction

Urine is not sterile (1). Modified culture protocols and modern sequencing techniques have now enabled the detection of microbes washed out from the whole urogenital tract (2). However, only a small fraction of these microbes is culturable (3), rendering culture-independent sequencing-based methods as the main tool to identify microbes inhabiting the urogenital tract.

Microbial communities colonizing the urinary tract (collectively referred to as the urobiome) are influenced by sex, age, environmental factors and even host genetics (2,4). Most importantly, recent studies have shown that urobiome dysbiosis is linked to several

urological conditions (2), ranging from urinary incontinence (5) to bladder cancer (6). Therefore, a comprehensive and systematic characterization of the urobiome in health and disease is fundamental, and may lead to new prevention, diagnosis and treatment strategies for urological pathologies.

Bacteria are the dominant component of the urobiome and a major technical challenge in DNA-based microbiota studies is the low bacterial biomass of urine samples. The urinary tract contains $<10^5$ colony forming units per milliliter, a number at least a million times lower than that found in feces per gram (7). As a consequence, while gut microbiota DNA-based studies are shifting from 16S ribosomal RNA (rRNA) amplicon sequencing towards shotgun metagenomic sequencing - which is problematic with low amounts of input bacterial DNA (8) -, urinary microbiota profiling still relies on 16S rRNA amplicon sequencing (9).

A critical step in 16S rRNA amplicon sequencing studies is the selection of which 16S rRNA hypervariable regions to sequence. 16S rRNA contains nine hypervariable regions (V1-V9) used to determine taxonomic identity and estimate evolutionary relationships between bacteria. Although all nine hypervariable regions are taxonomically informative, the amount and quality of information retrieved varies per region according to the studied environment. For instance, Fadeev et al. (2021) showed that V4V5 is superior to V3V4 for microbiota profiling of environmental arctic samples (10), and Kameoka et al. (2021) found that V1V2 is more precise than V3V4 for gut microbiota profiling of Japanese individuals (11).

Despite evidence showing that the choice of 16S hypervariable regions in microbiota profiling studies is critical (10–13), no study has systematically compared the performance of different 16S rRNA hypervariable regions for microbial characterization of urine samples. In this work, we compared the performance of different sets of 16S rRNA primers for urinary microbiota profiling. We collected urine samples from male volunteers by transurethral catheterization and used a 16S rRNA sequencing panel encompassing all nine hypervariable regions. We also combined pairs of non-overlapping 16S rRNA amplicons using bioinformatics reconstruction to evaluate their performance. To identify which primer sets and combinations are best suited for urinary microbiota profiling, we evaluated the effect of using different primer sets and combinations on metrics such as taxonomic resolution, taxonomic richness and ambiguity. We show that V1V2 amplicon sequencing is more suitable for urinary microbiota studies. We also observed marginal gains in taxonomic richness when using pairs of amplicons, which may not compensate for the higher costs of sequencing multi-amplicon libraries.

## Materials and Methods

### Sample collection

Twenty-two urine samples were collected from 14 male volunteers between March 2019 and November 2020. Samples were collected by a trained nurse in sterile urine containers during catheterization for BCG instillation in volunteers with non-muscle invasive bladder cancer or for transurethral resection in volunteers with benign prostatic hyperplasia (Table S1). Urine samples were stored at -80 ºC until DNA extraction.

## DNA extraction

Urine samples were thawed at room temperature, and up to 40 ml of urine was used for DNA extraction. Urine samples were centrifuged twice for 15 min at 10 ºC and 3000 g, and the supernatant was discarded sparing 10 ml of urine (containing a pellet). This content was transferred to 15 ml tubes, and centrifugation was repeated (15 min; 10 ºC; 3000 g). Approximately 1 ml of urine (containing the pellet) was resuspended in 3 ml phosphate-buffered saline (PBS) and centrifugation was repeated (15 min; 10 ºC; 3000 g). The supernatant was discarded leaving 1 ml of sample in the tube. Samples and the DNA extraction negative control (1 ml PBS) were processed for DNA extraction using the QIAamp DNA Microbiome kit (Qiagen, Hilden, Germany) following the manufacturer's protocol (*Depletion of Host DNA* protocol).

## Library preparation and sequencing

Twenty-four multi-amplicon libraries were prepared using the QIAseq 16S/ITS Screening Panel kit (Qiagen, Hilden, Germany) as detailed below. These libraries were prepared using 22 urine DNA samples, the DNA extraction negative control and the QIAseq 16S/ITS Smart Control (Qiagen, Hilden, Germany), a synthetic DNA sample used both as positive control for library preparation and sequencing, and as control for the identification of contaminants. DNA concentration was determined using the Qubit dsDNA HS Assay kit and Qubit 2.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA). Next, the fungal taxonomic marker internal transcribed spacer (ITS) and six 16S rRNA amplicons, spanning all nine hypervariable regions (V1V2, V2V3, V3V4, V4V5, V5V7 and V7V9), were amplified by PCR. The ITS region was poorly amplified since we used a DNA extraction protocol which depletes eukaryotic DNA. Sequences originated from the ITS amplicon were therefore discarded. PCR primers and their properties (estimated with OligoCalc (14)) are provided in Table S2. Amplifications were carried out in three independent reactions with primers multiplexed by the manufacturer. For samples in which DNA concentration was ≥0.25 ng/ul, 1 ng of DNA was used as template, and for samples with <0.25 ng/ul, 4 ul of DNA was used. Cycling conditions were: 95 ºC for 2 min; 20 cycles of 95 ºC for 30 s, 50 ºC for 30 s and 72 ºC for 2 min; and 72 ºC for 7 min. PCR products from the same sample were pooled and purified twice using QIAseq beads (Qiagen, Hilden, Germany). Dual-index barcodes and adapters were added to amplified products through a second-round of PCR using the QIAseq 16S/ITS 96-Index I array (Qiagen, Hilden, Germany). Cycling conditions were: 95 ºC for 2 min; 19 cycles of 95 ºC for 30 s, 60 ºC for 30 s and 72 ºC for 2 min; and 72 ºC for 7 min. After an additional purification using QIAseq beads, the presence of target sequences was evaluated with the Agilent Bioanalyzer 2100 System using the Agilent DNA 1000 kit (Santa Clara, CA, USA). Finally, we quantified the libraries using the NEBNext® Library Quant Kit for Illumina (New England Biolabs, Ipswich, MA, USA), size-correcting for the average length reported in the Bioanalyzer report considering a 400-700 bp quantification window. Libraries were normalized to 2 nM and sequenced using the MiSeq Reagent Kit v3 (600-cycle) (Illumina, San Diego, CA, USA) following the 2 x 276 bp paired-end read protocol.

**Read processing**

Paired-end reads were library demultiplexed and adapters were removed in the Illumina BaseSpace Sequence Hub. Each library was amplicon demultiplexed using cutadapt (v3.4) (15), generating two FASTQ files (with forward or reverse reads) for every library-amplicon combination. FASTQ files from the same amplicon were grouped in QIIME 2 artifacts and processed as independent datasets (hereinafter referred to as amplicon-specific datasets) using QIIME 2 (16).

Using DADA2 (17) (*q2-dada2* QIIME 2 plugin), reads were filtered based on default quality criteria, denoised and truncated (at the first instance of median quality score <30) to remove low quality bases at 3' ends. Next, paired-end reads were merged using DADA2 to produce amplicon sequence variants (ASVs). Finally, chimeric ASVs were filtered using VSEARCH (18) (*q2-vsearch* QIIME 2 plugin) and the SILVA database (v138) (19) as reference.

**Taxonomic assignment, nomenclature homogenization and contaminant removal**

Custom slices of the SILVA database (v138) for each amplicon were generated using RESCRIPt (20) (*q2-rescript* QIIME 2 plugin). Low-quality reference sequences were removed, identical reference sequences were dereplicated and 16S rRNA hypervariable regions were selected using primer sequences from the first-round of PCR as target sequences. Only selected regions within a reasonable length-range (100-600 nt) were kept in the final amplicon-specific databases. To achieve a more accurate taxonomic assignment for each amplicon-specific dataset (21), amplicon-specific taxonomic classifiers trained in amplicon-specific databases were built using the *q2-feature-classifier* QIIME 2 plugin (22). Finally, taxonomic assignment of ASVs was performed using amplicon-specific databases and classifiers.

Assigned taxonomies often contain incomplete information or generic proxies, especially at species level. To homogenize taxonomic nomenclature and to prevent inflation of taxonomic richness at species level, we replaced missing data, generic proxies (terms including "_sp.", "uncultured", "metagenome", or "human_gut") and ambiguous taxonomic entries (e.g., "phylum: Bacteroidota|Proteobacteria)" by the lowest taxonomic level with complete nomenclature and the corresponding taxon (e.g. "(…) genus: Streptococcus; species: uncultured_bacterium" is replaced by "(…) genus: Streptococcus; species: Genus_Streptococcus").

Next, we filtered non-bacterial and bacterial contaminants using taxonomic and abundance information. Non-bacterial contaminants were filtered by removing ASVs classified as not being from bacterial origin (taxonomy assigned to mitochondria, chloroplast or unassigned kingdom). Bacterial contaminants were identified using the R package *decontam* (23). Briefly, using the DNA extraction negative control and QIAseq 16S/ITS Smart Control libraries as controls for contaminants, we tested whether each ASV was a contaminant by combining frequency and prevalence *decontam* methods. Due to the limited number of DNA extraction negative control libraries, there was limited statistical power to identify contaminants exclusively from abundance data. Therefore, we evaluated manually if potentially contaminant ASVs ($P < 0.25$) had been previously described as belonging to human microbiotas by searching the taxon associated with such ASVs at PubMed (search in

May 2021). Potentially contaminant ASVs whose taxonomy had not been previously described in urine (namely, *Pelomonas,* which is a known laboratory contaminant (24), *Mycoplasma wenyonii* and *Candidatus Obscuribacter* ASVs) were considered true bacterial contaminants and were removed from all amplicon-specific datasets.

**Sidle-reconstruction of amplicons combinations**

The Short MUltiple Reads Framework (SMURF) algorithm (25) as implemented in Sidle (SMURF Implementation Done to acceLerate Efficiency) (26) was used to reconstruct datasets combining all six 16S rRNA amplicons. The *q2-sidle* QIIME 2 plugin was used (as described below) with amplicon-specific datasets after contaminants removal.

For ASVs in each amplicon-specific dataset to have a consistent length (as demanded by SMURF algorithm), ASVs were truncated at 300 nt. Reference sequences in the amplicon-specific databases generated previously were also truncated at 300 nt. For each truncated amplicon-specific database, regional k-mers were aligned (with 5 nt maximum mismatch) and a reconstructed database incorporating all amplicon-specific databases was built. Next, we reconstructed the abundance (0 minimum number of counts) and the taxonomic table incorporating all amplicon-specific datasets. Finally, we removed all libraries classified as defective and homogenized taxonomic nomenclature as previously described. Sidle-reconstructed datasets combining pairs of amplicons were built through an analogous pipeline.

**Microbiota analyses**

Amplicon-specific datasets were normalized prior to diversity analyses by Scaling with Ranked Subsampling (27) using the R package *SRS* (28). The number of reads of the library with the lowest number of reads per dataset was used as normalization cutoffs. The normalized amplicon-specific datasets were used to compute taxonomic and ASV richness (where richness is defined as the number of different observed features per dataset), and Faith's phylogenetic diversity index (29) using the R package *picante* (30). Compositional dissimilarity between samples (beta-diversity) was estimated using either Bray-Curtis or Jaccard (31) indices using the R package *phyloseq* (32).

ASVs were aligned using the R package *DECIPHER* (33) to calculate the entropy per nucleotide for each dataset, and the entropy score was calculated using the R package *Bios2cor* (34).

Genera intersections between datasets were determined using the R package *UpSetR* (35). Taxonomic trees were generated using the R package *metacoder* (36) employing the Reingold-Tilford layout. Only the 32 most abundant taxa were shown when plotting taxa relative abundances (based on minimum relative abundance in at least one sample, which is adjusted for each plot).

Ambiguity was estimated using the abundance output tables generated using Sidle. In these tables, the number of potential 16S rRNA source sequences for each feature is provided. For each dataset, ambiguity was calculated as the sum of the log of the number of potential 16S rRNA source sequences for each feature in the abundance table over the total number of features in the abundance table. To calculate the ambiguity for amplicon-specific datasets (not generated by Sidle), Sidle abundance tables for each amplicon were built as described in the previous section.

The full bioinformatics pipeline and R scripts (37) used for plotting (mainly with the R package *ggplot2* (38)) are available at https://github.com/vitorheidrich/urine-16S-analyses.

## Results

### Sequencing output and taxonomic resolution

We were able to amplify target sequences from 18 out of the 22 (82%) urine samples. In total we generated 20 amplicon libraries spanning all 16S rRNA hypervariable regions for urinary microbiota profiling (18 libraries from urine samples and libraries for the DNA extraction negative control and the QIAseq 16S/ITS Smart Control). A total of 13,638,685 reads were generated from urine sample libraries (median per library: 609,902; range: 383,982-1,489,196), out of which ~59% were short unspecific reads not associated with any of the amplicons of interest (the read length distribution of each amplicon-specific dataset is shown in Figure S1). After amplicon demultiplexing, each amplicon-specific dataset was analyzed in parallel. The total number of reads generated for each amplicon-specific dataset varied between 668,509 (V3V4) and 1,674,525 (V4V5) (Figure 1A; Table S3). After read filtering and removal of contaminants (see Methods), amplicon-specific datasets showed on average a 28% decrease in the number of reads (Figure 1A; Table S3). The number of reads removed at each step in our bioinformatics pipeline is detailed in Table S3.

Despite an overall balanced relative abundance of reads for each amplicon-specific dataset (Figure 1B), some libraries presented a disproportionate number of reads for a particular amplicon (Figure 1C). Specifically, libraries #2 and #3 showed a high proportion (>⅓) of V1V2 and V2V3 reads, respectively. We also noted that, despite the high median total number of reads generated for each library (204,433), the extremes varied by orders of magnitude (from 5,366 to 504,852 reads), so that the library with the lowest number of reads (#1) had less than 1,000 reads in 4 out of 6 amplicon-specific datasets. These disparities lead us to remove libraries #1, #2 and #3 from further analyses to prevent the introduction of bias due to low-quality libraries. Finally, we confirmed that the remaining libraries achieved satisfactory sequencing depth by calculating the Good's coverage (39) (~100% for all samples) and drawing rarefaction curves (Figure S2) for each amplicon-specific dataset.

Within these refined datasets, virtually all sequences in V1V2, V2V3 and V3V4 datasets received a taxonomic assignment up to genus level (Figure 1D). On the other hand, V4V5 and V5V7 showed a marked decrease in the percentage of assigned sequences at genus level, suggesting a lack of taxonomic resolution for relatively abundant taxa. Taxonomic assignment up to species level was rarely achieved, with V1V2 (19.7%) and V2V3 (21.8%) datasets showing the highest percentage of sequences assigned up to species level.

In summary, our results indicate that the protocol used herein is suitable for urinary microbiota characterization, providing enough sequencing depth to assess several amplicons simultaneously. We also confirmed that 16S rRNA hypervariable regions sequencing of urine samples can provide reliable taxonomic information up to genus level. However, taxonomic resolution varies along the 16S rRNA hypervariable regions, with V1V2 and V2V3 achieving the highest taxonomic resolution when considering genus and species levels together.

## Richness across 16S rRNA amplicon-specific datasets

Next, we evaluated ASV and taxonomic (phylum to species level) richness for each amplicon-specific dataset (Figure 2A-B). V1V2 and V3V4 datasets showed the highest ASV richness, while V4V5 and V7V9 presented a markedly lower ASV richness (Figure 2A). There was no correlation between ASV richness per dataset and the median ASV length per dataset (Spearman $\rho$ = -0.37, P = 0.47). The ASV length distribution of each amplicon-specific dataset is shown in Figure S3. At phylum and class level, all amplicons showed a remarkably similar richness (Figure 2B), with 6 phyla and 10 classes observed for all datasets, except for the V3V4 dataset (7 phyla and 11 classes). At lower taxonomic levels, differences between amplicons emerged, with the V1V2 dataset showing consistently the highest taxonomic richness from order to species level (Figure 2B). There was no correlation between taxonomic richness per dataset and the median ASV length per dataset (Table S4).

As expected from its higher ASV richness, V1V2 showed the highest taxonomic richness at genus level. However, we noticed that ASV richness did not always translate into taxonomic richness. For instance, V3V4 goes from the 2nd to the 4th position when richness was assessed at genus level instead of ASV level, suggesting that part of its ASVs correspond to ASVs phylogenetically close to other ASVs observed in the dataset, which do not contribute to increase taxonomic richness. Indeed, V3V4 ASVs showed a much lower phylogenetic diversity compared to V1V2 ASVs (Figure 2C). In fact, there is a decreasing trend in phylogenetic diversity along the 16S rRNA hypervariable regions, which is in line with the sequence variability (entropy) observed for each amplicon-specific dataset (Figure 2D). There was no correlation between ASV phylogenetic diversity and the median ASV length (Spearman $\rho$ = -0.09, P = 0.92).

Together, our results indicate that V1V2 is the most informative 16S rRNA amplicon in terms of taxonomic richness and phylogenetic diversity for urinary microbiota characterization.

## Taxonomic composition across 16S rRNA amplicon-specific datasets

The phyla Actinobacteriota, Bacteroidota, Firmicutes, Fusobacteriota and Proteobacteria were detected in all amplicon-specific datasets. However, some phyla were detected exclusively in a subset of them (Figure S4A). Therefore, we analyzed how taxa detection varied across amplicon-specific datasets at genus level. The full picture of the genera detected in each amplicon-specific dataset is provided in Figure S4B. Taxonomic trees depicting the contribution of each taxon (tree nodes) to the genera detected in each dataset are provided in Figure S5.

When evaluating the intersection of genera present in each amplicon-specific dataset (Figure 3A), we see that 27 genera were detected in all amplicon-specific datasets. The next larger subgroup, composed of 15 genera, comprises genera detected exclusively in the V1V2 dataset. Noteworthy, the V1V2 dataset is the only amplicon-specific dataset without exclusively undetected genera. All other datasets also presented “exclusive” genera, which summed up to 31 genera.

Due to such substantial differences, we aimed to assess how the choice of amplicon affects taxonomic profiles. To do so, we used beta-diversity analysis to evaluate whether the

taxonomic composition of a given sample is similar to itself irrespectively of the amplicon used for characterization (Figure 3B). For all taxonomic levels, using either Bray-Curtis or Jaccard beta-diversity indices, the compositional dissimilarities within samples (same sample profiled with different amplicons) are significantly lower than between samples, suggesting that the choice of amplicons will marginally impact the overall taxonomic compositions, especially at higher taxonomic levels.

The robustness of the taxonomic profile obtained irrespectively of the amplicon of choice can be further contemplated by the similar genera relative abundance profile (averaged over all samples) obtained for each amplicon-specific dataset (Figure 3C). In Figure 3C, there is an apparently disparate average taxonomic composition for V4V5 and V5V7 datasets. However, this is mainly due to loss of taxonomic resolution for some taxa, with ASVs otherwise classified as genera *Variovorax* and *Klebsiella* being only resolved up to family level (Comamonadaceae and Enterobacteriaceae, respectively) in these datasets. This loss of taxonomic resolution is also observed for *Halomonas* ASVs, which were classified as so in V4V5, V5V7 and V7V9 datasets, but as “Family_Halomonadaceae” in the remaining ones. This phenomenon is even more evident when evaluating taxa relative abundances per sample for each dataset at different taxonomic levels (Figure S6), with examples of higher taxonomic resolution at species level (e.g. for *Staphylococcus sp.* in V1V2 and V2V3 datasets).

Despite small variations in taxonomic resolution across amplicon-specific datasets for specific taxa, the overall taxonomic composition of urinary samples is similar independently of the amplicon of choice. Still, each amplicon is able to capture a different subset of the taxa, with V1V2 providing the highest number of exclusively detected genera. These results are in line with the higher genus richness observed for the V1V2 dataset and indicate that V1V2 better captures the actual microbiota composition of urinary samples.

**Comparison with Sidle-reconstructed datasets**

We next evaluated how the taxonomic richness and composition differ when considering a single amplicon-specific dataset vs. the Sidle-reconstructed taxa abundance, considering the information for all amplicon-specific datasets simultaneously. This “full” dataset can then serve as a compiled reference for urinary microbiota analysis. We also used Sidle to reconstruct taxa abundances for the following pairs of non-overlapping amplicons: V1V2-V4V5, V1V2-V5V7, V1V2-V7V9, V2V3-V5V7, V2V3-V7V9 and V4V5-V7V9.

As expected, there is a considerable gain in richness in the full dataset, mainly at species level, with 3.9x more species observed in the full dataset when compared to amplicon-specific datasets (Figure 4A). The use of pairs of amplicons also increases richness, but to a lower extent (Figure S7A), with V2V3-V7V9 combination providing the greatest increase in richness at species level (1.9x). This result can be explained by a more complete taxonomic assignment being achieved for a greater proportion of sequences in the full dataset (Figure S7B). Indeed, better taxonomic resolution observed for the Sidle-reconstructed datasets is due to the lower ambiguity (see Methods) in taxonomic assignment (Figure 4B). Noteworthy, the V1V2 dataset shows the lowest ambiguity when comparing only single amplicon-specific datasets.

Once again, the overall taxonomic composition is similar between datasets at genus level (Figure 4C). However, we see cases in which identification at species level was only possible in Sidle-reconstructed datasets (e.g., *Klebsiella pneumoniae* was identified in the full dataset and in most of the pairs of non-overlapping amplicons combinations) (Figure S7C). The taxa relative abundance per sample for the Sidle-reconstructed datasets at different taxonomic levels is provided in Figure S8.

Overall, the combination of amplicons through Sidle increases the taxonomic resolution achievable from 16S rRNA amplicon sequencing. However, the increase of combining pairs of amplicons is modest compared to the full reconstruction using all 16S hypervariable regions, which increases up to 4-fold the number of species detected. Still, this has limited impact in the taxonomic compositions, as evaluated by comparison with the taxonomic profiles generated by single amplicons. Once again, V1V2 stands out as the least ambiguous amplicon for urinary microbiota characterization.

**V1V2 taxonomic composition**

Due to the great number of taxa identified in the V1V2 dataset, we next investigated whether these taxa are commonly associated with the urogenital microbiota. In our cohort, six phyla were detected using V1V2 amplicon sequencing: Proteobacteria (72.4% of the sequences), Firmicutes (11.3%), Actinobacteriota (9.6%), Fusobacteriota (4.5%), Bacteroidota (2.3%) and Campilobacterota (<0.1%). All of these phyla have been previously reported in studies using catheterized urine samples (40–42). Only one of such studies reported the overall phyla abundance. The top-three most abundant phyla in Mansour et al. (2020) were Firmicutes, Proteobacteria and Actinobacteriota (42). However, their cohort included females, which are known to have a Firmicutes-enriched urogenital microbiota due to the high abundance of lactobacilli (43). In fact, a study with voided urine specimens from male bladder cancer patients found the same top-three most abundant phyla as described in this study (6).

Next, we examined the 15 genera detected exclusively in the V1V2 dataset. The average relative abundance of these genera varied between <0.001% (*Alkalibacterium* and *Jeotgalibaca*) and 2.9% (*Comamonas*), summing up to ~4% of the bacterial microbiota exclusively detected by V1V2 16S amplicon sequencing (Table S5). Due to the overall low relative abundance of these genera, we excluded the possibility of them being contaminants by searching the literature for the presence of these genera in urine samples. Briefly, 12 out of the 15 (80%) genera exclusively detected in the V1V2 dataset have been previously detected in human samples, and 10 out of 12 (83%) have been associated with urinary infections or detected in urogenital microbiota (Table S5). The three genera that were not previously detected in human microbiotas (*Alkalibacterium*, *Chromohalobacter*, *Salipaludibacillus*) sum up to only <0.01% average relative abundance in the V1V2 dataset. They have been described mainly as environmental high salt tolerant bacteria (44–46), indicating they may indeed represent undetected contamination or taxonomic misclassifications.

Finally, we compared our results with 16S amplicon sequencing-based microbiota studies using catheterized urine samples. In Forster et al. (2020), the urinary microbiota from 34 children with neuropathic bladder was characterized by V4 amplicon sequencing (47). More than 75% of the samples were dominated (relative abundance >30%) by family

Enterobacteriaceae members, but the genera involved in this phenomenon could not be determined due to limited taxonomic resolution. We also observed dominance by Enterobacteriaceae members in this cohort (samples #8.1 and #8.2; Figure S6), but because all nine Enterobacteriaceae ASVs in the V1V2 dataset were classified up to genus level (either to *Klebsiella* or *Escherichia-Shigella*), we were able to determine that *Klebsiella sp.* were responsible for this phenomenon. Noteworthy, in V4V5 and V5V7 datasets, family Enterobacteriaceae ASVs could not be classified up to genus level (Figure S6), recapitulating the limited taxonomic resolution for family Enterobacteriaceae observed in the aforementioned study.

Together, these data corroborate that V1V2 amplicon sequencing can provide reliable and richer taxonomic information for microbiota profiling of catheterized urine samples.

## Discussion

Many studies have compared the performance of different sets of 16S rRNA primers for microbiota profiling in different environments (10–13). These studies consistently demonstrated that the choice of the 16S rRNA primer set can significantly influence the analysis of microbiota diversity and composition. Similar studies for urinary microbiota profiling are lacking. As reviewed by Cumpanas et al. (2020) (9), out of 38 urobiome studies, 17 evaluated the V4 and 4 evaluated the V3V4 16S rRNA hypervariable regions. This is probably because these amplicons are commonly used in 16S rRNA amplicon sequencing commercial kits. It is also worth mentioning that some of the early seminal studies were based on V1V3 amplicon sequencing using the Roche 454 platform (2), which allows longer reads. Therefore, up to now library preparation kits and sequencing platforms have heavily influenced the choice of 16S rRNA hypervariable regions used in urinary microbiota profiling studies. Consequently, studies that provide evidence for a more informed choice are urgent.

In this study, we tested the performance of six different 16S rRNA primer sets, spanning all nine hypervariable regions, for microbiota profiling of 22 urine samples collected from male volunteers by transurethral catheterization. We show that V1V2 amplicon sequencing is more suitable for urinary microbiota profiling. We found that V1V2 provides the greatest taxonomic and ASV richness, which translates into a higher number of exclusively detected genera. This result is likely attributed to V1V2 having a higher taxonomic resolution for assessing the taxa commonly present in human urine samples.

We also evaluated combinations of pairs of non-overlapping amplicons, from which we observed only marginal gains in taxonomic richness in comparison with single amplicons. Combining all six amplicons leads to a substantial increase in taxonomic richness at species level, but with little impact on the overall taxonomic compositions, indicating these gains are largely due to low-abundant taxa. Therefore, they may not compensate for the higher costs of sequencing multi-amplicon libraries. Moreover, as amplicon combinations cannot be reconstructed as single sequences, the eventual equivocal association between amplicons may have caused inflation of taxonomic richness by false-positive taxa in Sidle-reconstructed datasets.

We observed huge discrepancies between amplicon-specific datasets when evaluating bacterial compositions by taxa relative abundances. This is mainly because some amplicons presented lower taxonomic resolution for profiling specific clades. Low taxonomic resolution may impact community-wide metrics and preclude the identification of

associations between taxa and covariates. Furthermore, low taxonomic resolution may also drastically impact beta-diversity metrics that do not take phylogenetic information into account (e.g., Bray-Curtis).

Amplicon-specific datasets also differed in the set of taxa detected. V1V2 profiling minimized the number of undetected genera, but because all other datasets possessed exclusively detected genera, we conclude that missing a fraction of the urine bacterial richness is inevitable with 16S rRNA amplicon sequencing. Still, low relative abundance taxa drive these observed differences so that analyses will not be harshly influenced by this limitation, except when evaluating beta-diversity with metrics that do not take bacterial evenness into account (e.g., Jaccard).

In this study, removal of contaminants was a key step, since laboratory and reagent contaminants disproportionately affect the microbiota profiling of low bacterial biomass samples (7). However, the method used for contaminant removal has limitations. Since we had low statistical power to detect contaminants exclusively using sequencing data, we validated our findings using information available in the literature. This is questionable because most urobiome studies available did not use strategies to control for contaminants (9), therefore a previous description of a taxon in human urobiomes does not imply it is a true urinary tract-resident microbe. On the other hand, some lists of known reagent and laboratory contaminants are available in the literature (e.g., (24)), but many of the taxa included in such lists are known to be present in human microbiotas. Obviously, these disputes are more frequent when studying less characterized environments. For instance, the genus *Variovorax*, which dominated a few samples in our study, is described as a contaminant by Salter et al. (2014) (24). At the same time, in a contaminant-controlled study, a *Variovorax* strain was identified in the urethra of a non-chlamydial non-gonococcal urethritis patient (48).

Another important limitation of urobiome studies is the lack of information on what are the true microbial members of the human urobiome. Urine samples were not included within the Human Microbiome Project, and due to the lack of an external reference, we focused on comparisons between amplicon-specific datasets. In addition, to partially mitigate the lack of a reference, we also compared amplicon-specific datasets to a bioinformatic reconstruction of the microbial community present in the urine samples using the full set of 16S hypervariable regions. Although further studies with experimentally validated references will be needed to confirm our findings, the results presented can guide methodological decisions in future urobiome studies.

Because genitalia and the urinary tract contain distinct bacterial communities (49), an important variable in urobiome studies is the choice of the sampling method (50). Many urobiome studies evaluate voided urine samples (9), which may contain bacteria from the urethra and genital skin, such that voided urine samples represent the whole urogenital tract. In this study, we evaluated urine samples collected via transurethral catheterization, which reduces the presence of distal urinary tract contaminants compared to voided urine (51). This sampling method, similarly to suprapubic aspiration, allows the specific characterization of the urinary bladder microbiota (52). Even though this was a fundamental consideration to avoid cross-site contamination, further studies will be necessary to evaluate whether our results extend to voided urine specimens. Likewise, since we included only male volunteers in this study, further studies including samples from females are desired to test whether our results can be extrapolated to the female urobiome.

In conclusion, similarly to other reports of primer bias in microbiota studies, we provided strong evidence that V1V2 is the most suitable 16S rRNA amplicon for the characterization of catheterized urine samples microbiotas. To our knowledge, this is the first study to address this question by systematically analyzing all 16S hypervariable regions. This is true not only for catheter-derived urine samples, but actually for any kind of urine sample. We believe that our results might help other researchers make an informed decision about which 16S rRNA hypervariable regions to use for urobiome analysis.

### Data Availability Statement

Raw sequencing data was deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB49145 (https://ebi.ac.uk/ena/browser/view/PRJEB49145).

### Ethics Statement

This study was approved by the Ethics Committee of Hospital Sírio-Libanês (#HSL 2018-72). All volunteers provided informed consent to participate before sample collection.

### Author Contributions

Conceptualization and study design: VH, AM, DB, DJ, MA, and AC. Volunteer recruitment and clinical evaluation: AM, DB, DJ, and MA. Samples preparation and sequencing: VH and LI. Bioinformatics and statistical analyses: VH. Writing original draft: VH and AC. Reviewing and editing the manuscript: PA, FB, and AM. Supervision: AC. All authors contributed to the article and approved the submitted version.

### Funding

VH was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, process no. 13996-0/2018).

### Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## References

1. Wolfe AJ, Brubaker L. “Sterile Urine” and the Presence of Bacteria. *Eur Urol* (2015) 68:173–174. doi: 10.1016/j.eururo.2015.02.041
2. Perez-Carrasco V, Soriano-Lerma A, Soriano M, Gutiérrez-Fernández J, Garcia-Salcedo JA. Urinary Microbiome: Yin and Yang of the Urinary Tract. *Front Cell Infect Microbiol* (2021) 11:617002. doi: 10.3389/fcimb.2021.617002
3. Hilt EE, McKinley K, Pearce MM, Rosenfeld AB, Zilliox MJ, Mueller ER, Brubaker L, Gai X, Wolfe AJ, Schreckenberger PC. Urine Is Not Sterile: Use of Enhanced Urine Culture Techniques To Detect Resident Bacterial Flora in the Adult Female Bladder. *J Clin Microbiol* (2014) 52:871–876. doi: 10.1128/JCM.02876-13

4. Adebayo AS, Ackermann G, Bowyer RCE, Wells PM, Humphreys G, Knight R, Spector TD, Steves CJ. The Urinary Tract Microbiome in Older Women Exhibits Host Genetic and Environmental Influences. *Cell Host Microbe* (2020) 28:298-305.e3. doi: 10.1016/j.chom.2020.06.022

5. Pearce MM, Hilt EE, Rosenfeld AB, Zilliox MJ, Thomas-White K, Fok C, Kliethermes S, Schreckenberger PC, Brubaker L, Gai X, et al. The Female Urinary Microbiome: a Comparison of Women with and without Urgency Urinary Incontinence. *mBio* (2014) doi: 10.1128/mBio.01283-14

6. Wu P, Zhang G, Zhao J, Chen J, Chen Y, Huang W, Zhong J, Zeng J. Profiling the Urinary Microbiota in Male Patients With Bladder Cancer in China. *Front Cell Infect Microbiol* (2018) 8:167. doi: 10.3389/fcimb.2018.00167

7. Karstens L, Asquith M, Caruso V, Rosenbaum JT, Fair DA, Braun J, Gregory WT, Nardos R, McWeeney SK. Community profiling of the urinary microbiota: considerations for low-biomass samples. *Nat Rev Urol* (2018) 15:735–749. doi: 10.1038/s41585-018-0104-z

8. Pereira-Marques J, Hout A, Ferreira RM, Weber M, Pinto-Ribeiro I, van Doorn L-J, Knetsch CW, Figueiredo C. Impact of Host DNA and Sequencing Depth on the Taxonomic Resolution of Whole Metagenome Sequencing for Microbiome Analysis. *Front Microbiol* (2019) 10:1277. doi: 10.3389/fmicb.2019.01277

9. Cumpanas AA, Bratu OG, Bardan RT, Ferician OC, Cumpanas AD, Horhat FG, Licker M, Pricop C, Cretu OM. Urinary Microbiota—Are We Ready for Prime Time? A Literature Review of Study Methods' Critical Steps in Avoiding Contamination and Minimizing Biased Results. *Diagnostics* (2020) 10:343. doi: 10.3390/diagnostics10060343

10. Fadeev E, Cardozo-Mino MG, Rapp JZ, Bienhold C, Salter I, Salman-Carvalho V, Molari M, Tegetmeyer HE, Buttigieg PL, Boetius A. Comparison of Two 16S rRNA Primers (V3–V4 and V4–V5) for Studies of Arctic Microbial Communities. *Front Microbiol* (2021) 12:637526. doi: 10.3389/fmicb.2021.637526

11. Kameoka S, Motooka D, Watanabe S, Kubo R, Jung N, Midorikawa Y, Shinozaki NO, Sawai Y, Takeda AK, Nakamura S. Benchmark of 16S rRNA gene amplicon sequencing using Japanese gut microbiome data from the V1–V2 and V3–V4 primer sets. *BMC Genomics* (2021) 22:527. doi: 10.1186/s12864-021-07746-4

12. Sirichoat A, Sankuntaw N, Engchanil C, Buppasiri P, Faksri K, Namwat W, Chantratita W, Lulitanond V. Comparison of different hypervariable regions of 16S rRNA for taxonomic profiling of vaginal microbiota using next-generation sequencing. *Arch Microbiol* (2021) 203:1159–1166. doi: 10.1007/s00203-020-02114-4

13. Cabral DJ, Wurster JI, Flokas ME, Alevizakos M, Zabat M, Korry BJ, Rowan AD, Sano WH, Andreatos N, Ducharme RB, et al. The salivary microbiome is consistent between subjects and resistant to impacts of short-term hospitalization. *Sci Rep* (2017) 7:11040. doi: 10.1038/s41598-017-11427-2

14. Kibbe WA. OligoCalc: an online oligonucleotide properties calculator. *Nucleic Acids Res* (2007) 35:W43–W46. doi: 10.1093/nar/gkm234

15. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* (2011) 17:10–12. doi: 10.14806/ej.17.1.200

16. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* (2019) 37:852–857. doi: 10.1038/s41587-019-0209-9

17. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* (2016) 13:581–583. doi: 10.1038/nmeth.3869

18. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* (2016) 4:e2584. doi: 10.7717/peerj.2584

19. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* (2013) 41:D590–D596. doi: 10.1093/nar/gks1219

20. Robeson MS, O'Rourke DR, Kaehler BD, Ziemski M, Dillon MR, Foster JT, Bokulich NA. RESCRIPt: Reproducible sequence taxonomy reference database management for the masses. (2020)2020.10.05.326504. doi: 10.1101/2020.10.05.326504

21. Werner JJ, Koren O, Hugenholtz P, DeSantis TZ, Walters WA, Caporaso JG, Angenent LT, Knight R, Ley RE. Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. *ISME J* (2012) 6:94–103. doi: 10.1038/ismej.2011.82

22. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, Huttley GA, Gregory Caporaso J. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* (2018) 6:90. doi: 10.1186/s40168-018-0470-z

23. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* (2018) 6:226. doi: 10.1186/s40168-018-0605-2

24. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* (2014) 12:87. doi: 10.1186/s12915-014-0087-z

25. Fuks G, Elgart M, Amir A, Zeisel A, Turnbaugh PJ, Soen Y, Shental N. Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling. *Microbiome* (2018) 6:17. doi: 10.1186/s40168-017-0396-x

26. Debelius JW, Robeson M, Hugerth LW, Boulund F, Ye W, Engstrand L. A comparison of approaches to scaffolding multiple regions along the 16S rRNA gene for improved resolution. (2021)2021.03.23.436606. doi: 10.1101/2021.03.23.436606

27. Beule L, Karlovsky P. Improved normalization of species count data in ecology by scaling with ranked subsampling (SRS): application to microbial communities. *PeerJ* (2020) 8:e9593. doi: 10.7717/peerj.9593

28. Heidrich V, Karlovsky P, Beule L. "SRS" R Package and "q2-srs" QIIME 2 Plugin: Normalization of Microbiome Data Using Scaling with Ranked Subsampling (SRS). *Appl Sci* (2021) 11:11473. doi: 10.3390/app112311473

29. Faith DP. Conservation evaluation and phylogenetic diversity. *Biol Conserv* (1992) 61:1–10. doi: 10.1016/0006-3207(92)91201-3

30. Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* (2010) 26:1463–1464. doi: 10.1093/bioinformatics/btq166

31. Jaccard P. Distribution de la flore alpine dans le Bassin des Dranses et dans quelques régions voisines. *Bull Société Vaudoise Sci Nat* (1901) 37:241–272. doi: 10.5169/seals-266440

32. McMurdie PJ, Holmes S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLOS ONE* (2013) 8:e61217. doi: 10.1371/journal.pone.0061217

33. Wright ES. Using DECIPHER v2.0 to analyze Big Biological Sequence Data in R. *R J* (2016) 8:352–359. doi: 10.32614/RJ-2016-025

34. Taddese B, Garnier A, Deniaud M, Henrion D, Chabbert M. Bios2cor: an R package integrating dynamic and evolutionary correlations to identify functionally important residues in proteins. *Bioinformatics* (2021) 37:2483–2484. doi: 10.1093/bioinformatics/btab002 10.1002/9781118655252.ch13

35. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* (2017) 33:2938–2940. doi: 10.1093/bioinformatics/btx364

36. Foster ZSL, Sharpton TJ, Grünwald NJ. Metacoder: An R package for visualization and

manipulation of community taxonomic diversity data. *PLOS Comput Biol* (2017) 13:e1005404. doi: 10.1371/journal.pcbi.1005404

37. R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria* (2021). url: https://www.R-project.org/

38. Valero-Mora PM. ggplot2: Elegant Graphics for Data Analysis. *J Stat Softw Book Rev* (2010) 35:1–3. doi: 10.18637/jss.v035.b01

39. Good IJ. The Population Frequencies of Species and the Estimation of Population Parameters. *Biometrika* (1953) 40:237–264. doi: 10.2307/2333344

40. Oresta B, Braga D, Lazzeri M, Frego N, Saita A, Faccani C, Fasulo V, Colombo P, Guazzoni G, Hurle R, et al. The Microbiome of Catheter Collected Urine in Males with Bladder Cancer According to Disease Stage. *J Urol* (2021) 205:86–93. doi: 10.1097/JU.0000000000001336

41. Hussein AA, Elsayed AS, Durrani M, Jing Z, Iqbal U, Gomez EC, Singh PK, Liu S, Smith G, Tang L, et al. Investigating the association between the urinary microbiome and bladder cancer: An exploratory study. *Urol Oncol Semin Orig Investig* (2021) 39:370.e9-370.e19. doi: 10.1016/j.urolonc.2020.12.011

42. Mansour B, Monyók Á, Makra N, Gajdács M, Vadnay I, Ligeti B, Juhász J, Szabó D, Ostorházi E. Bladder cancer-related microbiota: examining differences in urine and tissue samples. *Sci Rep* (2020) 10:11042. doi: 10.1038/s41598-020-67443-2

43. Pearce MM, Zilliox MJ, Rosenfeld AB, Thomas-White KJ, Richter HE, Nager CW, Visco AG, Nygaard IE, Barber MD, Schaffer J, et al. The female urinary microbiome in urgency urinary incontinence. *Am J Obstet Gynecol* (2015) 213:347.e1-347.e11. doi: 10.1016/j.ajog.2015.07.009

44. Yumoto I, Hirota K, Nakajima K. “The genus Alkalibacterium”, *Lactic Acid Bacteria: Biodiversity and Taxonomy*. John Wiley & Sons, Ltd (2014). p. 147–158 doi: 10.1002/9781118655252.ch13

45. Ventosa A, Gutierrez MC, Garcia MT, Ruiz-Berraquero FY 1989. Classification of “Chromobacterium marismortui” in a New Genus, Chromohalobacter gen. nov., as Chromohalobacter marismortui comb. nov., nom. rev. *Int J Syst Evol Microbiol* 39:382–386. doi: 10.1099/00207713-39-4-382

46. Sultanpuram VR, Mothe T 2016. Salipaludibacillus aurantiacus gen. nov., sp. nov. a novel alkali tolerant bacterium, reclassification of Bacillus agaradhaerens as Salipaludibacillus agaradhaerens comb. nov. and Bacillus neizhouensis as Salipaludibacillus neizhouensis comb. nov. *Int J Syst Evol Microbiol* 66:2747–2753. doi: 10.1099/ijsem.0.001117

47. Forster CS, Panchapakesan K, Stroud C, Banerjee P, Gordish-Dressman H, Hsieh MH. A cross-sectional analysis of the urine microbiome of children with neuropathic bladders. *J Pediatr Urol* (2020) 16:593.e1-593.e8. doi: 10.1016/j.jpurol.2020.02.005

48. Riemersma WA, Schee CJC van der, Meijden WI van der, Verbrugh HA, Belkum A van. Microbial Population Diversity in the Urethras of Healthy Males and Males Suffering from Nonchlamydial, Nongonococcal Urethritis. *J Clin Microbiol* (2003) doi: 10.1128/JCM.41.5.1977-1986.2003

49. Gottschick C, Deng Z-L, Vital M, Masur C, Abels C, Pieper DH, Wagner-Döbler I. The urinary microbiota of men and women and its changes in women during bacterial vaginosis and antibiotic treatment. *Microbiome* (2017) 5:99. doi: 10.1186/s40168-017-0305-3

50. Brubaker L, Gourdine J-PF, Siddiqui NY, Holland A, Halverson T, Limeria R, Pride D, Ackerman L, Forster CS, Jacobs KM, et al. Forming Consensus To Advance Urobiome Research. *mSystems* (2021) doi: 10.1128/mSystems.01371-20

51. Dong Q, Nelson DE, Toh E, Diao L, Gao X, Fortenberry JD, Pol BVD. The Microbial Communities in Male First Catch Urine Are Highly Similar to Those in Paired Urethral Swab Specimens. *PLOS ONE* (2011) 6:e19709. doi: 10.1371/journal.pone.0019709

52. Wolfe AJ, Brubaker L. Urobiome updates: advances in urinary microbiome research. *Nat Rev Urol* (2019) 16:73–74. doi: 10.1038/s41585-018-0127-5

## Figures

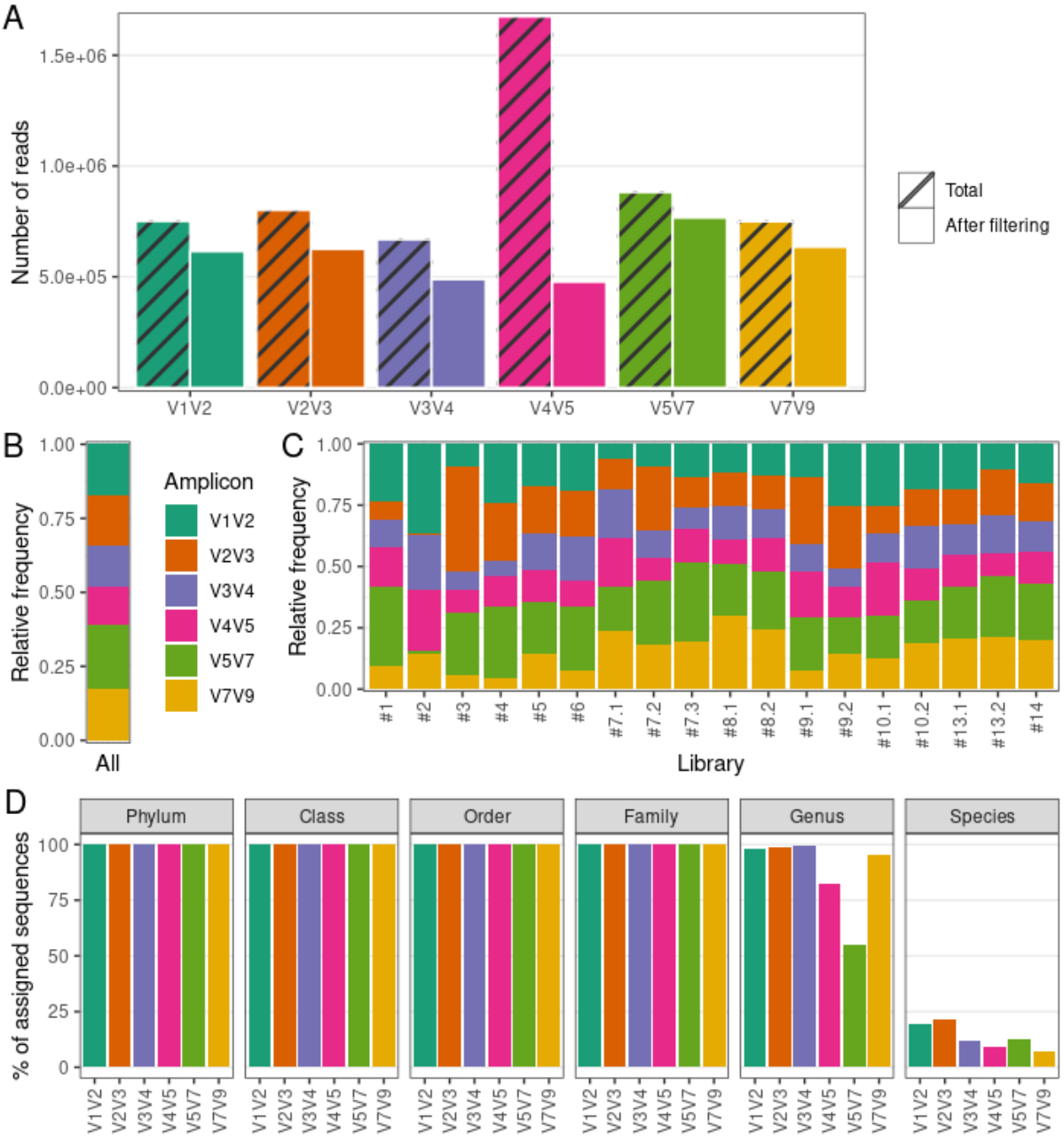


**Figure 1: Sequencing output and taxonomic resolution for each 16S rRNA amplicon-specific dataset.** (A) Number of reads generated and retained after filtering steps for each amplicon-specific dataset. (B) Relative frequency of reads retained after filtering steps averaged over all libraries for each amplicon-specific dataset. (C) Relative frequency of reads retained after filtering steps per library for each amplicon-specific dataset. (D) Percentage of sequences with assigned taxonomy (per taxonomic level) for each amplicon-specific dataset.

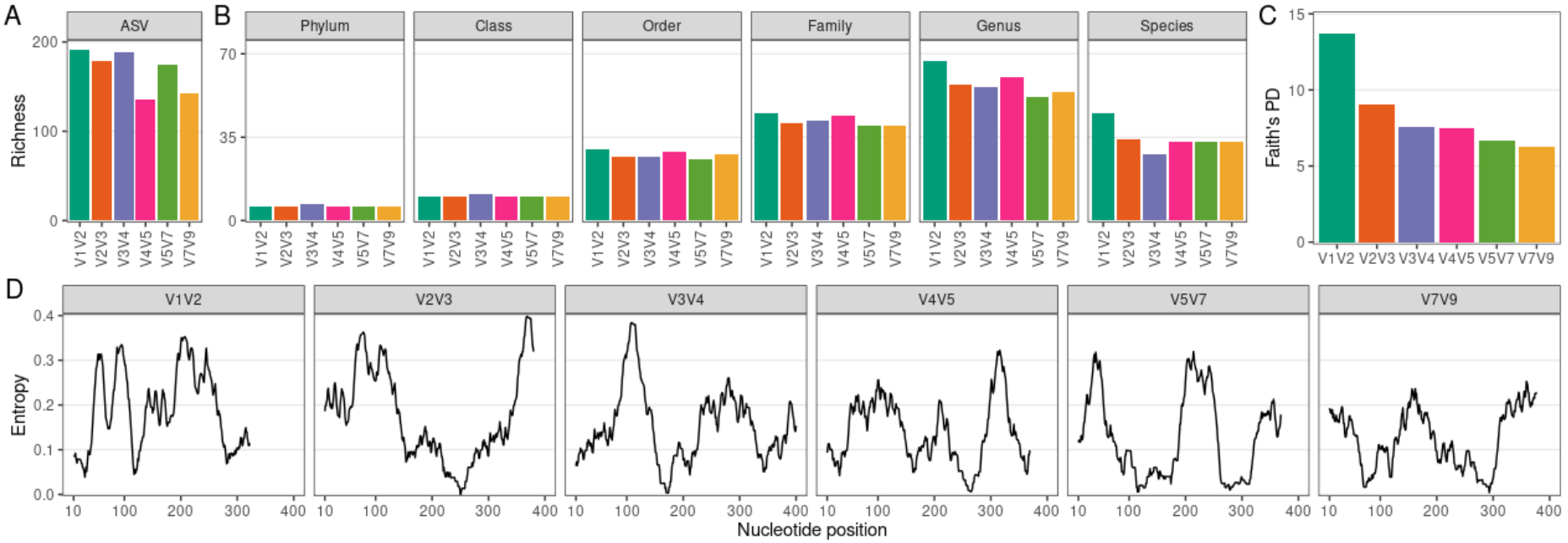


**Figure 2: Richness and phylogenetic diversity across 16S rRNA amplicon-specific datasets.** (A) Amplicon sequence variant (ASV) richness per amplicon-specific dataset. (B) Taxonomic richness (phylum to species level) per amplicon-specific dataset. (C) Faith's Phylogenetic Diversity (PD) across amplicon-specific datasets. (D) Sequence variability (entropy) along ASVs nucleotide positions (20-nucleotides rolling average) for each amplicon-specific dataset. Only nucleotide positions up to the median ASV size per amplicon-specific dataset are considered.

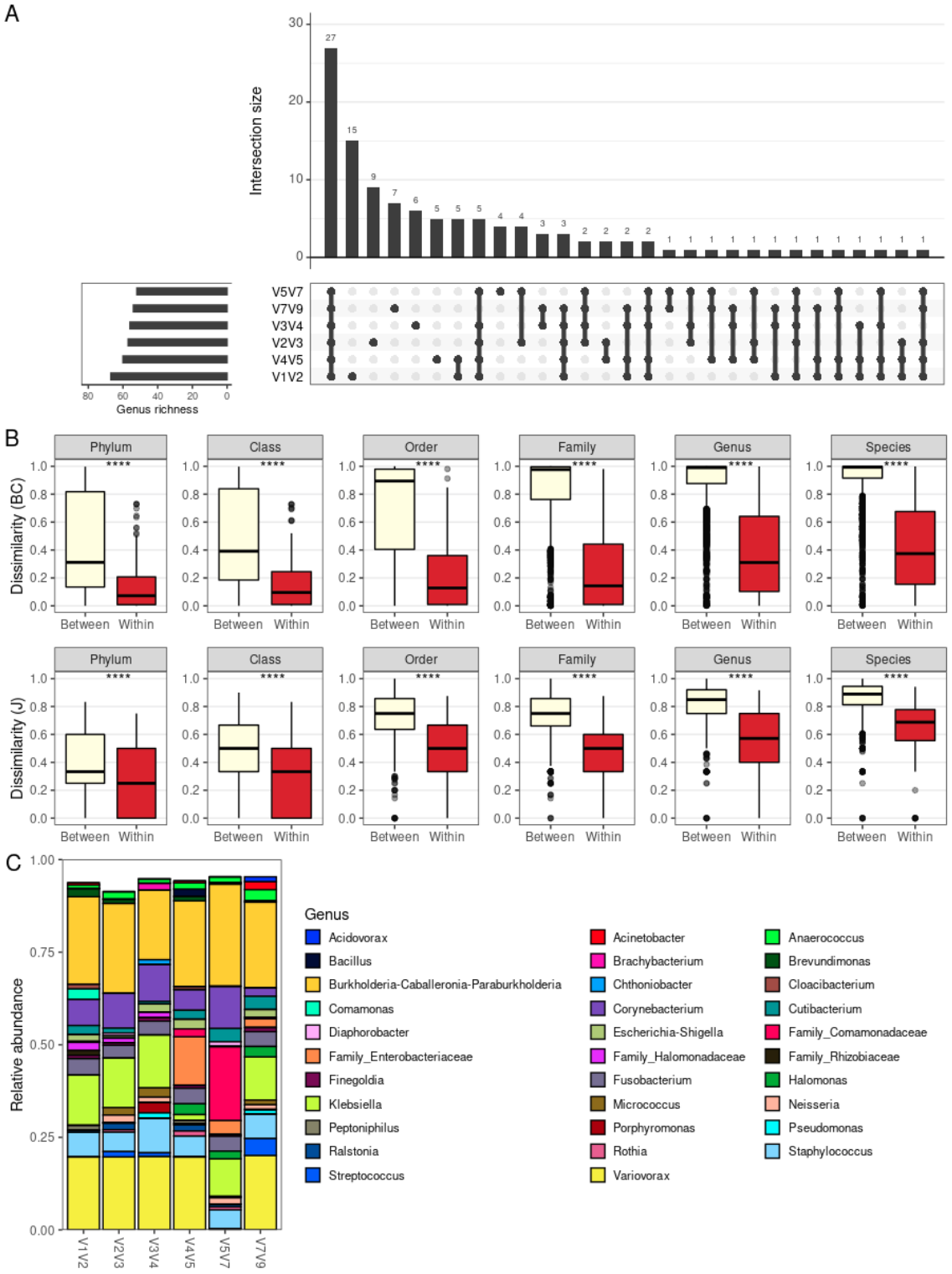


**Figure 3: Taxonomic composition across 16S rRNA amplicon-specific datasets.** (A) Barplot depicting intersections between the genera detected in each amplicon-specific dataset. Total rIchness at genus level is shown in the lower-left subplot. (B) Boxplot comparing dissimilarities between different libraries and within the same libraries as profiled with different amplicons. Dissimilarity metrics considered are Bray-Curtis (BC) and Jaccard (J). Statistical significance was evaluated by the Mann-Whitney U test. The boxes highlight the median value and cover the 25th and 75th percentiles, with whiskers extending to the more extreme value within 1.5 times the length of the box. (C) Average genera relative abundance per amplicon-specific dataset. ***, P < 0.001.

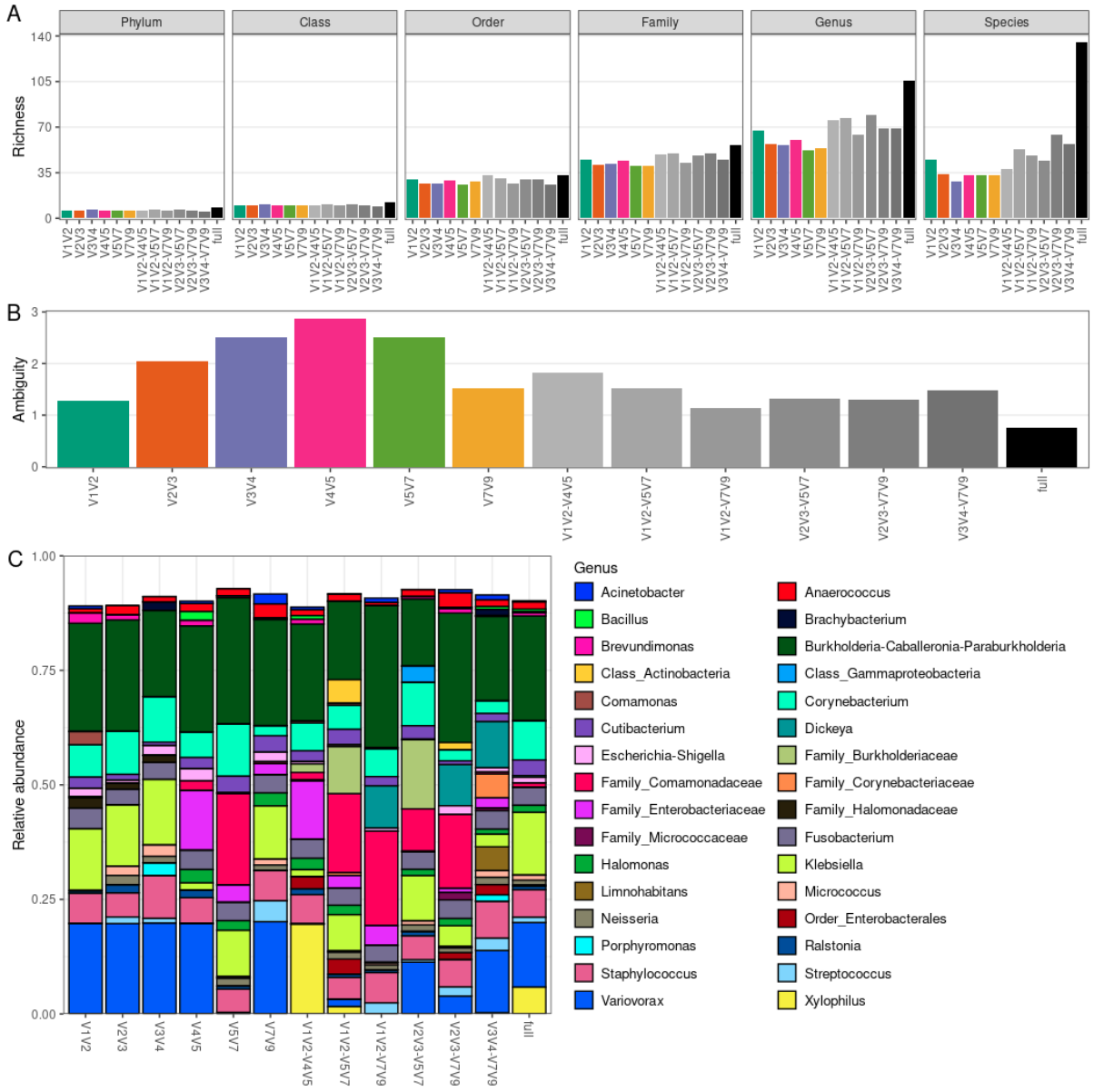


**Figure 4: Richness and taxonomic composition of Sidle-reconstructed datasets.** (A) Taxonomic richness (phylum to species level) per amplicon-specific or Sidle-reconstructed dataset. (B) Ambiguity in taxonomic assignment per amplicon-specific or Sidle-reconstructed dataset. (C) Average genera relative abundance per amplicon-specific or Sidle-reconstructed dataset.