CoSMIC - A hybrid approach for large-scale, high-resolution microbial profiling of novel niches

Maor Knafo*[1], Shahar Rezenman*[1], Tal Idan[1], Michael Elgart[2], Shlomi Dagan[3], Ziv Reich[1], Ruti Kapon[1], Dagan Sade^[1] & Noam Shental^[4]

[1]Department of Biomolecular Sciences, Weizmann Institute of Science Rehovot, Rehovot, Israel.
[2] Metabolomic Discovery Unit, Metabolic Center, Sheba Medical Center.
[3] Fisher Drug Discovery Resource Center, The Rockefeller University, New York USA.
[4]Department of Mathematics and Computer Science, The Open University of Israel.

^Corresponding authors:
Dagan Sade - dagan.sade@weizmann.ac.il &
Noam Shental - shental@openu.ac.il

*****These authors contributed equally to this work

The authors declare they have no competing interests.

CoSMIC - A hybrid approach for large-scale, high-resolution microbial profiling of novel niches

Abstract

Microbiome studies have become integral to a wide range of biological research from ecology to oncology, mainly relying on short-read sequencing of variable regions along the 16S rRNA gene. However, the comprehensiveness of 16S rRNA studies has been persistently challenged due to a lack of primer universality and primer biases, causing differences between a study's results and the underlying bacterial community. Moreover, relying on a small part of the gene often provides low phylogenetic resolution, hampering downstream taxonomy-based analysis and the ability to harmonize results from studies performed using different variable regions.

Here, we introduce a framework called 'Comprehensive Small Ribosomal Subunit Mapping and Identification of Communities' (CoSMIC), effectively addressing these challenges. CoSMIC begins with long-read full-length sequencing of the 16S rRNA gene, using Locked Nucleic Acid primers over pooled samples. This step augments the Small Subunit (SSU) reference database with novel niche-specific SSUs. Subsequently, CoSMIC optimizes a set of primer pairs targeting multiple non-consecutive variable regions along the gene, followed by standard short-read sequencing of each sample. Data from different regions are seamlessly integrated using the SMURF framework, thus alleviating primer-based biases and providing extremely high phylogenetic resolution. We evaluated CoSMIC across plant, root, soil and marine sponge samples, yielding higher profiling accuracy and unprecedented phylogenetic resolution compared to standard methods while detecting ~40,000 novel SSUs.

CoSMIC provides researchers with a robust, customizable, and cost-effective framework addressing the challenges of 16S rRNA sequencing, especially in underexplored habitats, paving the way for data harmonization across microbiome studies.

Introduction

During the last decade, in the advent of massively parallel sequencing, the study and exploration of microbiomes has become an integrative science that spans multiple disciplines and research fields. Medicine, environmental studies, agriculture, and even behavioral studies have associated specific phenotypes with bacteria(1–4). A majority of these studies rely on sequencing predefined variable regions of the 16S rRNA gene using short reads. However, it has been widely shown that results of such studies depend on the amplified region due to biases introduced by the specific primers applied(5,6), and on specific bioinformatic decisions and practices(7–9), hence limiting the ability to harmonize findings across studies(10).

Early microbiome studies used operational taxonomic units (OTUs) either in “closed reference” or “open reference” forms. The former aligns each short read to a database of full-length 16S rRNA sequences (e.g., SILVA(11) or Greengenes(12)), which contain millions of small ribosomal subunits (SSUs). Such alignment implicitly “cleans” sequencing errors and allows harmonizing results from different experiments, using both the SSU itself and its assigned taxonomy, which can also provide ecological analysis at higher taxonomic levels. However, there are two inherent caveats to “closed reference” OTU. First, it implicitly assumes that the SSU database is comprehensive, an assumption that probably holds only for human-associated niches. Second, reads are often aligned to multiple SSUs that share the same short amplicon, and such ambiguity may result in inconsistent taxonomies. To allow database-free assignments researchers then used “open-reference” OTUs, i.e., experimental reads were taken as is, which then led to the development of denoising methods, e.g., Deblur(13) and DADA2(14). The output of such methods, referred to as amplicon sequence variants (ASVs), use error models of Illumina platforms to “denoise” reads, i.e., create a single ASV from different reads that harbor different sequencing errors. It was shown that ASVs can be compared across studies even when using different denoising algorithms(15), allowing a database-free approach. However, such ASV-based cross-study harmonization is not possible when studies amplify different variable regions, thus results still depend on their identity. In addition, the issue of ambiguous assignment to SSUs remains unsolved.

To circumvent the necessity to commit to a specific primer pair, we have previously presented the Sample MUlti-Region Framework (SMURF) approach(8). In SMURF

each sample is amplified by any number of primer pairs, and reads from different regions are computationally combined to provide coherent high-resolution microbial profiling. To perform such integration, SMURF relies on SSU databases, making the approach suboptimal for less explored niches.

Here we present Comprehensive SSU Mapping and Identification of Communities (CoSMIC), which builds on SMURF to provide a framework for high-resolution profiling of novel niches. Prior to applying SMURF in a new niche, we suggest amplifying a pool of samples using locked nucleic acid (LNA) primers which provide an almost universal amplification of full-length 16S rRNA genes. Sequencing these amplicons by long-read technologies (e.g., PacBio) enriches SSU databases with relevant SSUs, thus allowing the use of SMURF to provide efficient and high-resolution profiling of these niches. We provide a workflow for applying CoSMIC detailing the selection of specific subsets of primers to allow cost savings while not compromising phylogenetic resolution. The advantages of CoSMIC over standard methods are demonstrated by sequencing a mock mixture and over a variety of environmental samples as well as marine sponges.

Results

The outcome of microbial profiling highly depends on amplified regions

To demonstrate the effect of preselecting a single variable region, we performed 16S rRNA profiling of leaves, roots, and their surrounding soil, from 7 different plants (see Methods). Profiling was performed using a commercial Qiagen kit (QIAseq 16S/ITS Region Panel (24)) that amplifies six regions along the gene (Fig. 1A; see Methods) resulting in 200K-600K reads per sample (Fig. S10). We then applied Qiagen's accompanying CLC Genomics Microbiome package to analyze results, thus providing a set of Operational Taxonomic Units (OTUs) and their taxonomy for each region and sample.

Fig. 1B illustrates the number of OTUs detected in leaf, soil, and root samples of *Aster poliothamnus* using each of the six amplified regions. For example, the same root sample results in 600-1100 OTUs, depending on the amplified region (Fig. 1B, right). Moreover, there is a low overlap in assigned taxonomies detected by different regions,

as evident by the beta diversity among all pairs of amplified regions (Fig. 1C). Diversity was calculated by a Jaccard distance and is higher than zero for any pair of amplified regions (median Jaccard distances of 0.27, 0.30 and 0.28 for leaf, root, and soil, respectively).

To evaluate the contribution of each amplicon to the analysis of a given sample type, we identified the amplicon as having the largest number of unique OTUs and sequentially added regions according to their additional unique OTUs with respect to regions selected thus far (Fig. 1D-F). The “top” region varied across samples; e.g., the most “informative” regions for leaf, soil, and root in the case of *Aster poliothamnus* were V3V4, V1V2, and V4V5, respectively. In addition, it required four regions to exhaust the number of unique OTUs, yet the identity of these regions varied among leaf, soil, and root samples, thus highlighting the inherent biases in region selection. Comparable observations were made when profiling the other plants - *Brandisia hancei*, *Lepidium apetalum*, *Lathyrus sativus, Angelica hirsutiflora,* and *Tetraena mongolica* (Figs. S1-5), while the most "informative" amplicon was inconsistent both across plants for a given niche (e.g., leaves across all plants), and over niches for each plant (Figs. S6-8).

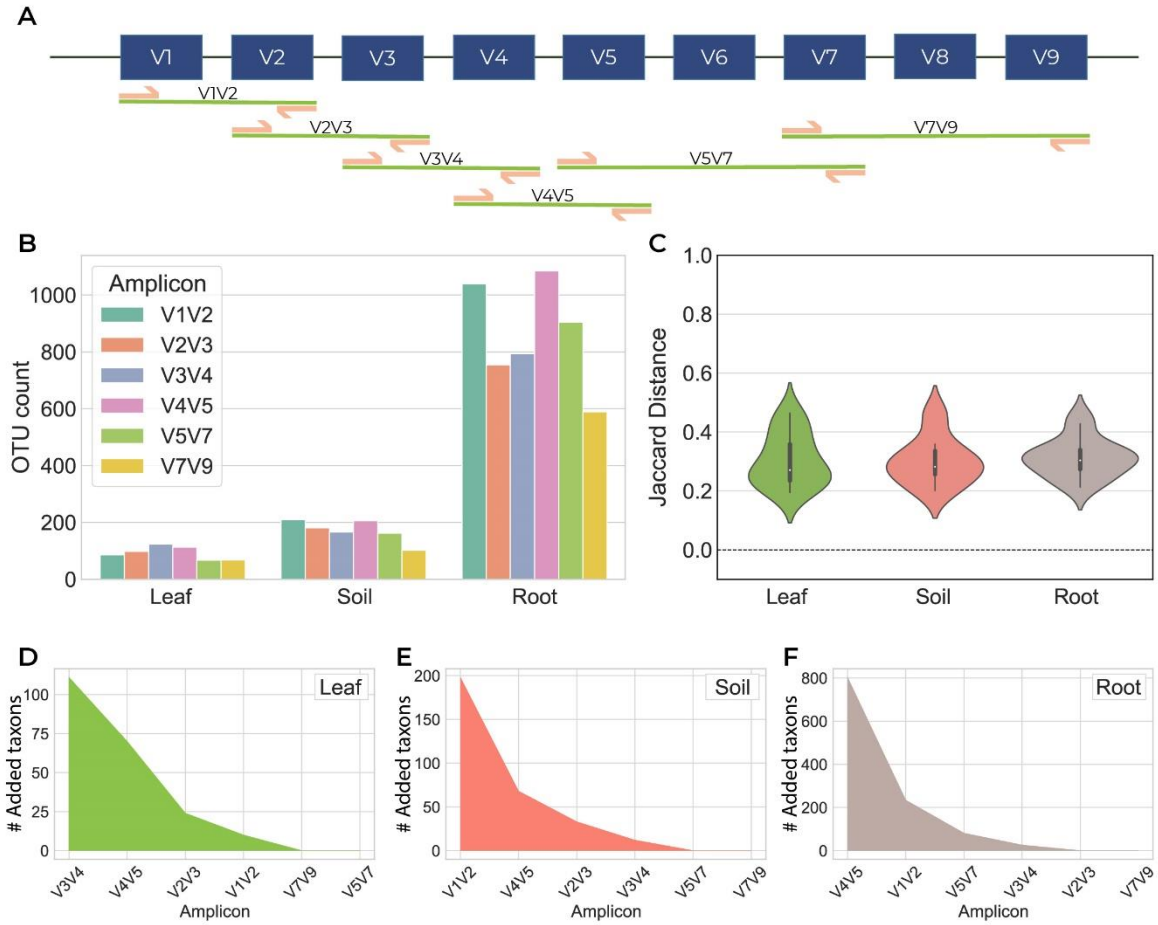

Figure 1. Microbial profiling of *Aster altaicus* samples depends on the applied amplicon. Leaf, soil, and root samples of *Aster altaicus* were profiled by the Qiagen kit that provides six amplicons along the 16S rRNA gene and subsequently analyzes their reads using the CLC Genomics platform. **A.** The six amplicons of the Qiagen kit, each amplifying two or three adjacent variable regions. Taken together, these amplicons cover all variable regions. **B.** The number of OTUs detected for soil, leaf, and root samples of *Aster altaicus* for each of the six amplicons. **C.** Distribution of Jaccard distances of among pairs of amplicons. In each pair, e.g., V1V2 vs. V3V4, a Jaccard distance is calculated between the lists of taxons detected by CLC Genomics in each region. **D-F.** Decay graph of the number of unique taxons identified by the amplicons analyzed. Amplicons are ordered by the number of unique taxons a region contributes to the former set of regions. D-Leaf, E-Soil, and F-Root.

Combining results from several variable regions partially overcomes problems in profiling

We previously suggested employing the Sample MUlti-Region Framework (SMURF) approach to avoid issues related to pre-selecting a single region. In SMURF, any combination of regions is amplified, and their sequencing results are computationally combined, resulting in a single coherent profiling solution. SMURF relies on a database of full-length 16S rRNA sequences, i.e., SSUs, to solve an optimization

problem, seeking the SSUs and their relative frequencies that have given rise to the set of reads from the relevant regions. The output of SMURF is a set of “groups” and their relative frequencies, where each group contains SSUs having an identical sequence over the amplified regions. For example, if V1V2 and V7V9 regions are being amplified, SSUs having the same sequence over these two regions are, therefore, indistinguishable and, if detected by SMURF, would be part of the same group. Apart from circumventing the issue of selecting a single region, with its associated inherent complications, SMURF provides higher profiling resolution. The larger the number of regions, the smaller the size of a characteristic group, typically allowing to identify a *single* SSU that unambiguously appears in the sample.

We applied SMURF to analyze all eight plants across leaf, soil, and root samples. Every sample underwent analysis through all 63 region combinations of the amplified six regions. For example, an analysis was performed using solely V1V2, by V1V2 combined with V3V4, followed by V1V2 combined with V3V4 and V7V9, and so forth. Our assessment of each region's contribution was based on three parameters: *(a)* ambiguity, i.e., the average group size detected by SMURF. The lower the ambiguity, the higher the resolution; *(b)* detection, representing the percentage of correctly identified SSUs when the full six-region combination is used as a benchmark*; (c)* Pearson correlation between the frequencies of SSUs detected by a set of regions vs. those detected by the full six-region combination.

The average performance across all three parameters improves when more regions are incorporated into the SMURF analysis. The average ambiguity across all 24 samples was 20,000 SSUs when using a single region, yet by adding only two more regions to the analysis, ambiguity was less than 2,000 (Fig. 2A). Secondly, the capacity to accurately detect a substantial proportion of the population increased linearly (Fig. 2B). Each additional amplicon contributed to the correct identifications, eliminating unique SSUs that were inaccurately declared by single regions, then subsequently dismissed with the inclusion of more comprehensive information. Thirdly, as more amplicons are added, the correlation improves significantly, with many samples achieving a correlation coefficient (r) higher than 0.9 with the addition of just two or three regions (Fig. 2C). Figs. 2A-C displays a significant variance in performance for a given number of regions, depending on their identity. Assuming that costs scale with the number of amplified regions, the optimal selection of primers may be crucial. Such selection can be tailored to suit the specific aims and budget. One

can identify an optimal selection by visualizing mutual taxonomies and Pearson correlation along the horizontal and vertical axes, respectively, and ambiguity designated by the marker size (Fig. 2D-E). For example, a combination of three primer pairs, V1V2, V2V3, and V4V5, yields better results in soil samples than some sets of four primers (Fig. 2D, red cross vs. all other pink circles). However, the same combination performs poorly for root (Fig. 2D-E, red cross) and leaf (Fig. S9, red cross) samples.

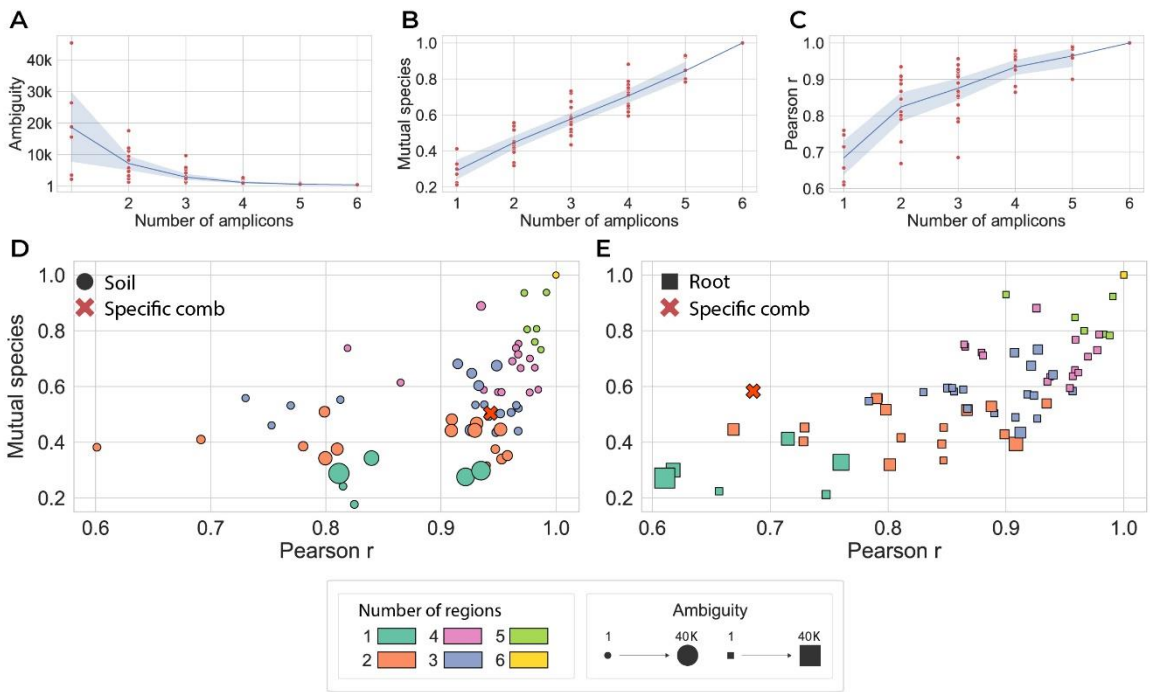

Figure 2. Evaluating the performance of sets of primer pairs across niches. 24 samples corresponding to leaf, soil, and root of eight plants were profiled by SMURF, using all 63 combinations of the six amplicons. **A.** Ambiguity, i.e., the number of SSUs per group as detected by SMURF, as a function of the number of amplicons. Each dot corresponds to the average ambiguity in one of the 63 distinct cases. The blue line and gray background correspond to the average and standard deviation of the ambiguity across combinations of a specific number of regions. **B.** We consider SMURF's solution based on all six regions as the gold standard and display a fraction of species identified by subsets of the six regions. Shown is the percent of species whose taxonomy matches the six-region solution. **C.** Pearson correlation between profiling based on all six regions and each individual set of regions. **D.** A comparison of amplicon sets for soil samples. Each dot corresponds to a specific region combination averaged across samples (soil or root), where the marker size scales with the ambiguity and axes correspond to Pearson correlation (horizontal) and mutual species (vertical). The marker's color indicates the number of amplicons used for SMURF analysis. Hence, a single marker corresponds to 6 regions (yellow), and 6 green dots correspond to the different groups of 5 regions. The marker "x" corresponds to a specific set of 3 regions comprising V1V2, V2V3, and V4V5. **E.** The same for root samples.

Enriching the database using long-read sequencing

Although SMURF provides improved profiling, its results depend on the comprehensiveness of the available databases of SSU sequences. However, most databases are highly enriched for, e.g., human-associated bacteria, while other habitats may be severely underrepresented.

To bridge this gap and harness the advantages of SMURF across habitats, we propose a simple preceding step in a microbiome sampling project designated to provide niche-specific full-length SSUs. The procedure entails a PCR step, utilizing a multiplex set of locked nucleic acid (LNA) primers aimed at effectively amplifying the whole 16S rRNA gene. LNA primers(16,17) are shorter than regular primers (10-13nt instead of ~20nt), thus allowing an increased universality compared to regular primers while maintaining primer specificity and high Tm (see Methods). We, therefore, propose pooling samples from the relevant niche, applying PCR amplification using the designed LNA primers, and sequencing to create full-length SSUs that potentially enrich existing databases with novel sequences.

To illustrate this idea and estimate the number of novel SSUs that can be detected, we used this approach over pools of samples, each from a different niche: (a) plants from the Arava desert in Israel; (b) plants from the temperate lowland in Israel; (c) sponges of the species *Spongia officinalis* collected from the Mediterranean coast of France; and (d) *Spongia officinalis* collected from the Mediterranean coast of Israel. These pools were sequenced using PacBio long reads amplicon sequencing(18), resulting in ~200K high-quality SSUs (Fig. S11) of average length 1600 bp (Fig. S12). Fig. 3A presents the percentage of novel SSUs as a function of the identity threshold to SILVA SSUs. In general, sponges seem to have significantly higher percentages of novel SSUs(19), irrespective of threshold. Using the SILVA definition of a novel SSU, i.e., a threshold of 99% identity (SILVA NR99 shown as a vertical dashed line in Fig. 3A), more than 95% of detected SSUs were novel for samples of *Spongia officinalis* and about 85% for samples taken from plants.

To remove redundancy in our initial pools, we used the SILVA database standard threshold for their NR99 database(11), and clustered our samples with 99% identity. This resulted in a table where each cluster represents a single SSU. Fig. 3B shows the histograms of cluster sizes, i.e., the number of PacBio reads matching this SSU for one pool of samples derived from *S. officinalis*. Upper and lower panels correspond

to SILVA alignments lower or higher than 97%, respectively. Overall, the distribution of read counts is very similar between known SSUs and novel SSUs. The tail of the distribution shows highly abundant SSUs (e.g., having more than 1000 reads) that are not represented in SILVA, suggesting that the underrepresentation problem is not limited to rare species. Furthermore, 96% of the novel SSUs have fewer than 10 reads (3758 of 3880), indicating more could be detected with deeper sequencing. The same effect was observed for other pools, as evident from rarefaction curves in Fig. 3C.

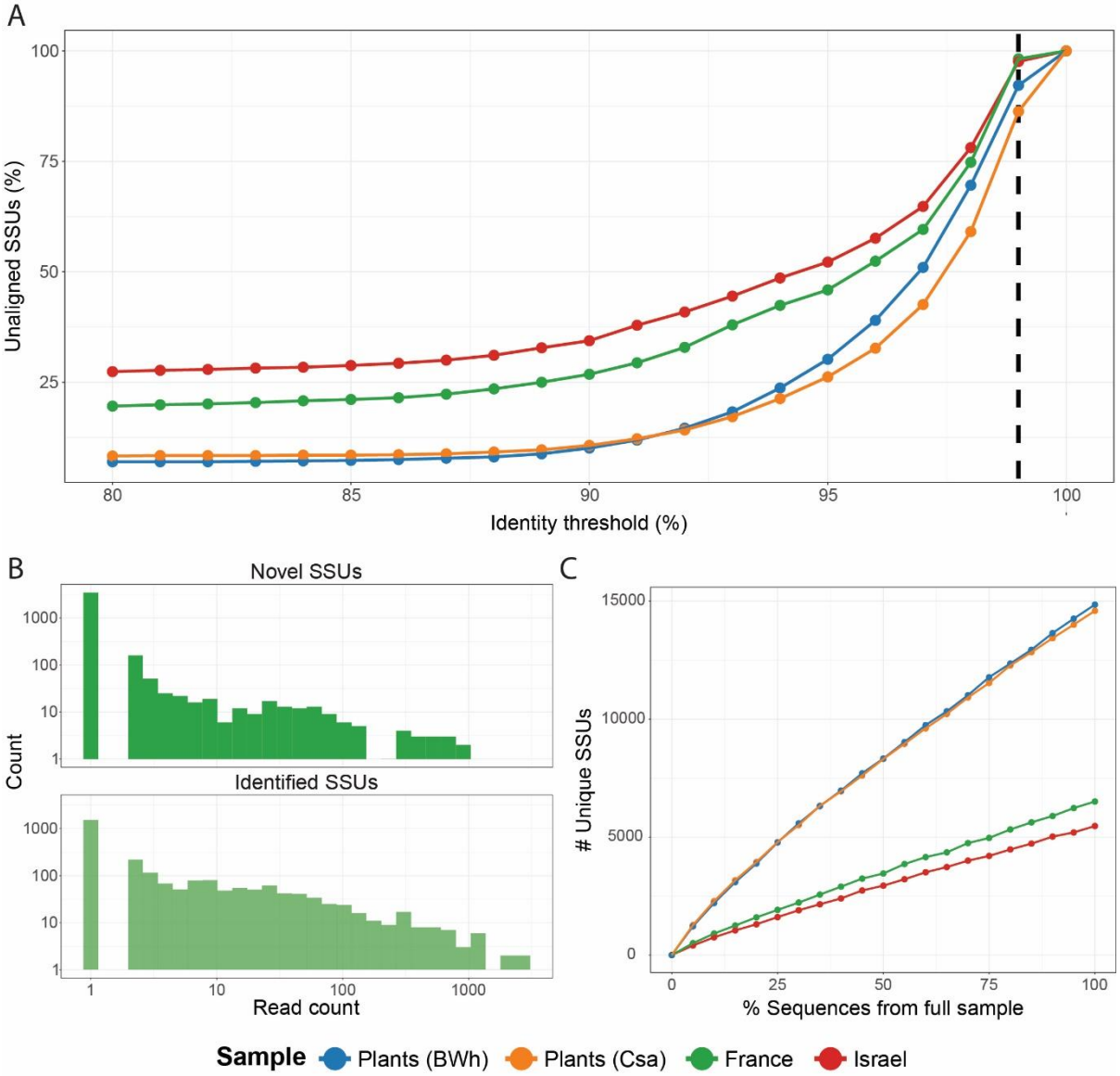

Figure 3. Enriching the database with novel full-length SSUs. We amplified the full-length 16S rRNA gene using designed LNA primers and sequenced the resulting amplicons by PacBio long-read sequencing. Four samples were processed, each containing a pool of samples from a different habitat, i.e., plants from the Arava desert in Israel, characterized by a climate type BWh, according to Koppen climate classification; plants from the lowland region in Israel, characterized by a climate type Csa, and pools of samples from the marine sponge *Spongia officinalis* collected in the Mediterranean Sea along the coast of either France or Israel. **A**. All non-redundant SSUs (Plants BWh - 14852, Plants Csa - 14589, France - 6507, Israel - 5468) from the PacBio

sequencing were aligned to the SILVA database using the CD-HIT algorithm(20). The percentage of SSUs that did not appear in SILVA is shown as a function of the identity threshold. The vertical dashed line corresponds to SILVA's standard threshold for calling novel SSUs (SILVA NR99). Colors indicate different biomes sampled. **B.** Histogram of read counts per SSU for the France sponge pool of samples, for novel SSUs (upper panel), and for known SILVA SSUs (lower panel), using a 97% identity threshold. **C.** Rarefaction curves presenting the number of novel SSUs identified (y-axis) based on the percentage of reads sampled (x-axis), for each of the sequenced pools.

The CoSMIC approach

When considering the problems and solutions presented in the former sections, we propose guidelines for large-scale microbiome studies that would provide high-resolution profiling in a new or known niche while accommodating for different budget tiers (Fig. 4). We recommend treating each research project as a new ecological niche (orange, right-hand side in the flowchart) and enrich current SSU databases with relevant sequences, i.e., pool multiple samples of different origins of the ecological niche, apply LNA amplification and perform long reads sequencing. Next, using the newly enriched database, we propose to sequence a representative set of samples using the largest set of primers available (e.g., all 10 Swift pairs or all 6 Qiagen pairs), apply SMURF, and select the subset of primer pairs that provide sufficient resolution per niche (or select all primer pairs in case costs do not scale considerably with the number of primers). Once these steps are finalized, the project transitions to a "familiar niche" case (purple, left-hand side), which performs routine amplification of the optimal primer set followed by short reads sequencing, subsequently analyzed by augmented-database SMURF.

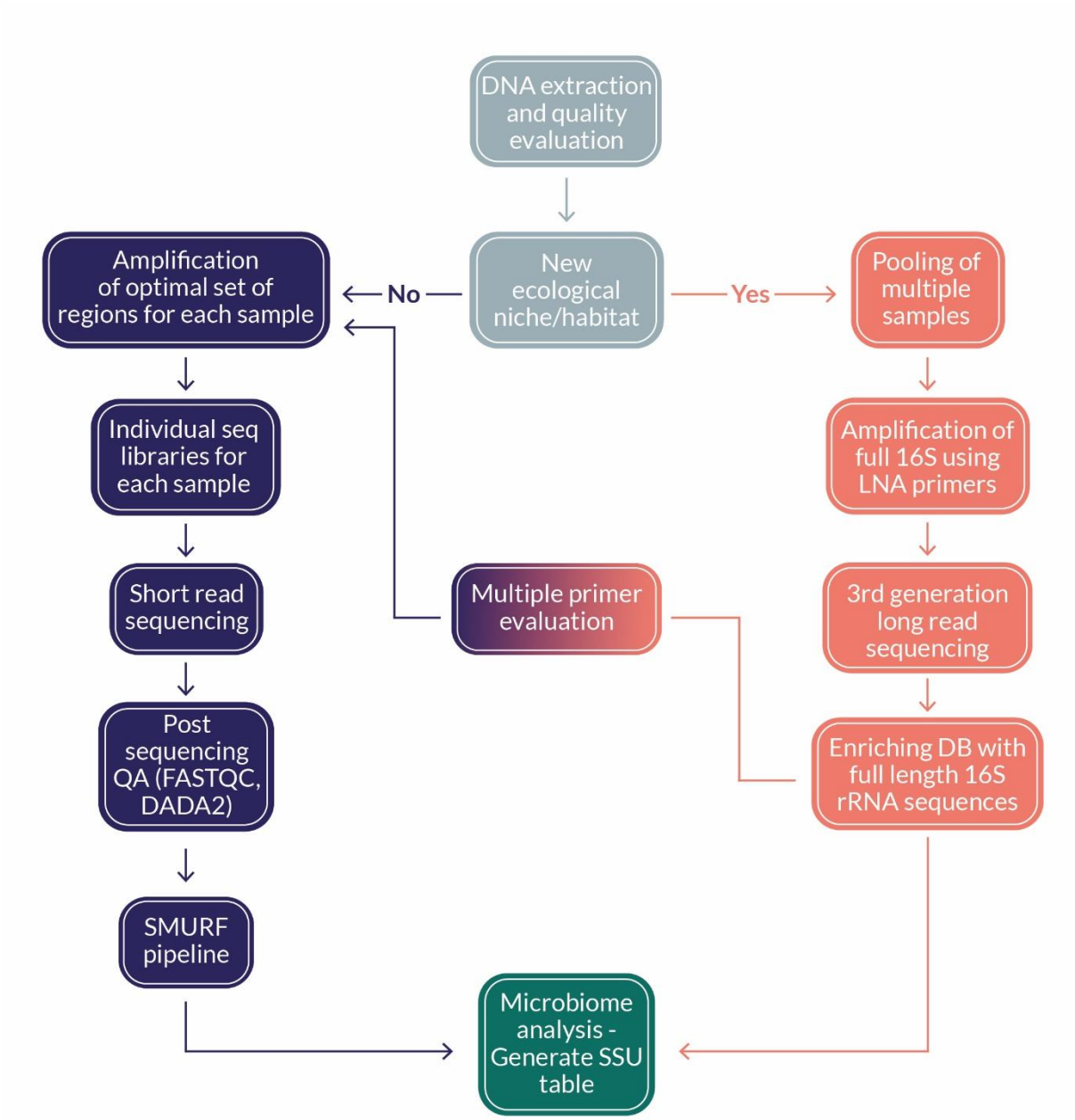

Figure 4. The CoSMIC suggested flowchart. Gray - Perform standard DNA extraction from environmental samples. **Orange -** Establish a relevant SSU database by LNA-based amplification and long-read sequencing. **Orange-purple -** Test multiple primer sets to find an optimal primer pair combination. **Purple -** Perform high throughput profiling in a large-scale project. Samples are amplified using the optimal primer combination and analyzed by SMURF based on the augmented database.

Experimental evaluation of CoSMIC

To evaluate the advantages of CoSMIC, we sampled *Hymelaea hirsuta*, *Solanum lycopersicum,* and *Triticum aestivum*, at their leaves, roots, and surrounding soil. In addition, we created a mock community of 12 known species. To construct our ground

truth microbial communities, we performed standard shotgun metagenomic sequencing of each of the ten samples and extracted SSUs and their relative frequencies in each sample (see Methods). We refer to these SSUs as ground truth SSUs (gtSSUs), and their relative frequencies are called ground truth frequencies.

As a first step in CoSMIC, we pooled all ten samples and used our LNA primers to amplify full-length SSU sequences, then sequenced them using synthetic long reads technology (Loop genomics(21)) to achieve full-length SSU sequences. Novel SSUs (using the SILVA NR99 standard) were added to the SILVA database. We refer to this database as “augmented SILVA”.

Each sample was then sequenced by Illumina short reads using the Swift pipeline, which amplifies several regions along the 16S rRNA gene (see Methods section and detailed information about obtained reads in supplementary for Loop, SWIFT and metagenomics S13-14). To apply SMURF, we considered 10 amplicons spanning the 16S gene, based on the abovementioned primers, and applied the augmented SILVA database. For comparison, we provide the standard analysis provided by Swift, based on either all primers or on the V3V4 amplicon.

The comparison was based on four criteria: (a) *Alignment identity* - the average alignment score between detected SSUs and gtSSUs; (b) *Ambiguity* - the number of SSUs in the augmented SILVA database that match detected SSUs by each method; (c) *Explained frequency* - the sum of ground truth relative frequencies for gtSSUs that were also detected by each method; (d) *Erroneous SSUs* - the number of detected SSUs that were not present in gtSSU (for details regarding a-d see Methods).

CoSMIC outperforms traditional methods on all parameters (Fig 5).

Alignment identity: Mean and overall alignment identities are higher for CoSMIC across all samples (Fig. 5A). The distribution varies according to the tissue sampled, with mock, root, and soil consistently producing better scores compared to leaf tissue. This is possibly related to the low number of bacteria detected due to chloroplasts occupying most of the reads. While the difference between analysis methods is evident across all tissues, the most striking example is the mock community scores. Swift-based analysis, based on all regions or on V3V4, had a median alignment identity of about 90% even for this simple, well-defined mixture, compared to an almost 100% alignment identity for CoSMIC.

Ambiguity - Fig. 5B presents a histogram of the ambiguity across all SSUs detected in the 10 samples. In most cases each detected OTU by Swift corresponds to hundreds

or thousands of the database's SSUs, while the CoSMIC detects a *single* SSU, yielding a more precise, actionable result.

Explained frequency - Significant improvements are also apparent in the accuracy of community composition. Fig. 5C presents the total explained frequency as a function of the identity threshold averaged per sample type across plants. Using CoSMIC, more species are accurately reported, accounting for a higher proportion of the ground-truth sample. The difference between the methods is particularly evident when only high-quality alignments are considered, resulting in almost zero frequency explained by traditional methods.

Erroneous SSUs - The percentage of erroneous SSUs, i.e., those that are reported yet do not exist in the ground truth sample, is significantly lower for CoSMIC compared to the other methods (Fig. 5D).

Fig. 5E shows all criteria using a radar plot for each method, clearly showing the improved performance by CoSMIC, especially when addressing ambiguity (Amb) and explained frequency at high alignment identity scores (Cov-99).

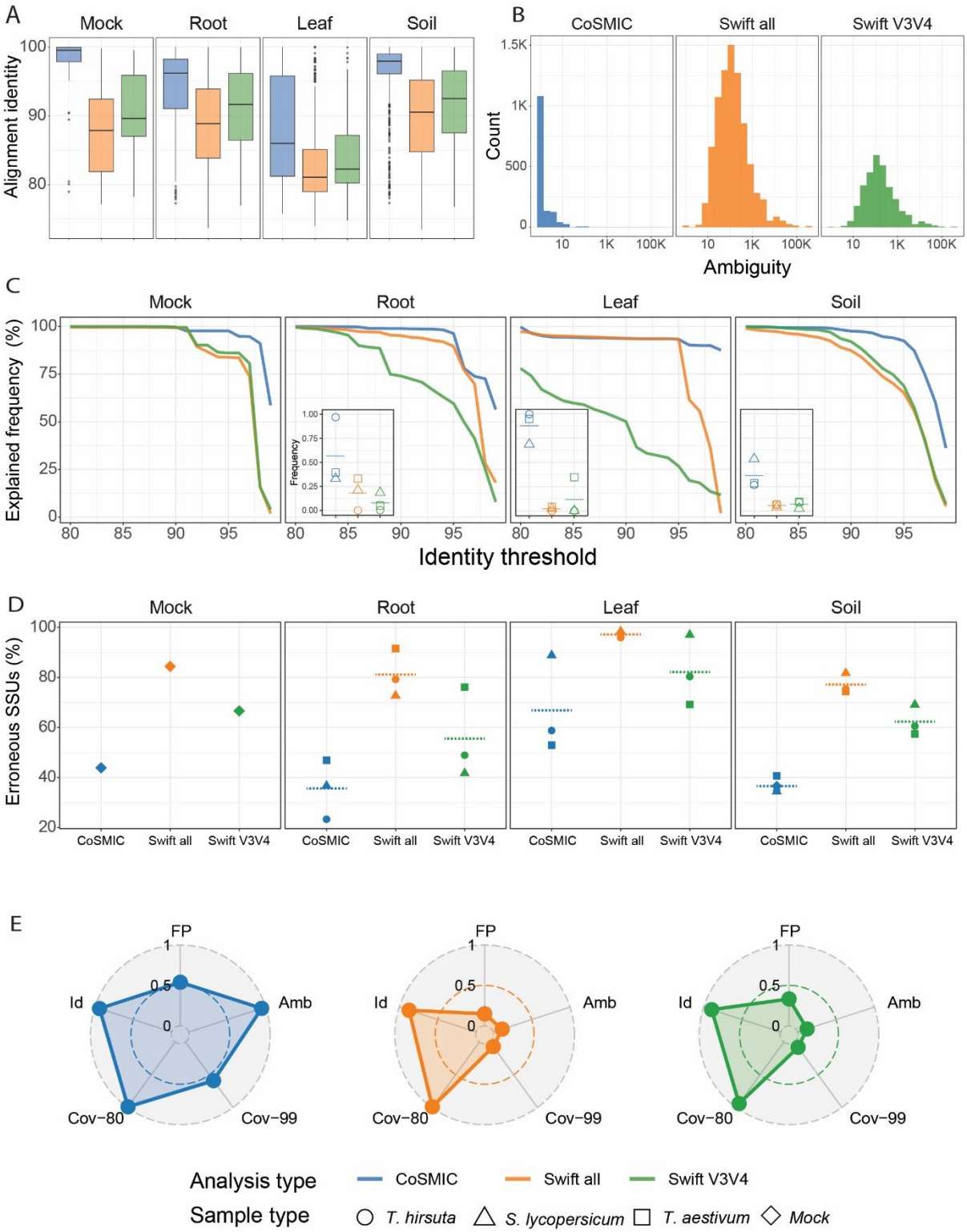

Figure 5. Experimental evaluation of standard methods vs. CoSMIC. Ten samples were sequenced - leaf, soil, and root samples of *Thymelaea hirsuta*, *Solanum lycopersicum,* and *Triticum aestivum*, and a mock mixture of 12 known species (see Methods). Each sample underwent standard shotgun metagenomic sequencing, from which we detected its full-length 16S rRNA sequences to be considered ground-truth. In addition, as part of CoSMIC we first pooled extracted DNA from all 10 samples, amplified by LNA primers, and sequenced via synthetic long reads (Loop genomics[22]) to create SSUs that were added to the SILVA database. Standard amplicon sequencing was

performed by a commercially available kit by Swift, which uses primers spanning the 16S rRNA gene to amplify several variable regions (see Methods). Analysis was then performed by either standard Swift pipeline using 10 primer pairs (Swift all), by the V3V4 region (Swift V3V4), and by CoSMIC. **A.** Alignment identity, i.e., the average alignment score between detected SSUs and SSUs detected by shotgun metagenomics. Boxplots present the distribution of alignment identities for each tissue and for each analysis method. Leaf, soil, and root panels correspond to an average of over 3 samples (one from each plant). A Mann-Whitney U test was performed for all comparisons, which resulted in very low p-values, with the highest being 0.0064. **B.** Distribution of the number of taxa matching an identification by each analysis type. **C.** Total frequency of ground truth detected by each method as a function identity cutoff. The insets present the total frequency achieved per sample at a cutoff of 99%, while the dotted horizontal line is their average. **D.** Percentage of erroneous SSUs, i.e., those that were identified by an analysis method yet were not detected by shotgun metagenomics. Each marker represents a specific plant, while each facet corresponds to a tissue. The dotted line marks the mean for each tissue and method. **E.** Radar plot combining all samples over several criteria (normalized on a 0-1 scale, where large values correspond to better performance): **FP,** mean false positive rate as described in D; **Amb,** mean ambiguity as described in B; **Cov-99,** mean total frequency ground truth coverage at 99% identity; **Cov-80,** mean ground truth coverage at 80% identity cutoff as described in C; **Id,** mean identity for each analysis method as described in A.

Discussion

The diverse applications of microbiome studies have underscored the need for robust and standardized methodologies(22). However, primer selection and arbitrary analysis decisions lead to lower profiling resolution (Figs. 5A-B), as well as fall short due to the omission of species (Fig. 5C) or the inclusion of species that are not present in the sample (Fig. 5D). To address these limitations, we propose to circumvent the issue of primer selection by applying sets of primers from different variable regions and computationally combining them via the SMURF approach to achieve extremely high phylogenetic resolution profiling. When profiling novel niches, we suggest a preliminary step that pools samples from the novel niche and applies amplification by our LNA-based (almost) universal primers followed by long-read sequencing. Subsequently, one can optimize the primer set for the novel niche and perform low-cost and high-resolution large-scale profiling of multiple samples from the relevant niche. These concepts were incorporated into the CoSMIC pipeline (Fig. 4) and then shown to provide significant improvements across all tested parameters over a new niche, i.e., reducing ambiguity, improving alignment quality, and providing a more accurate and reliable picture of the microbial community (Fig. 5).

While the CoSMIC approach represents an advancement in microbiome analysis, it requires careful quality assurance by researchers. A novel niche in terms of CoSMIC

refers to cases where a significant fraction of the reads cannot be matched to the available SSU database (e.g., SILVA). Hence, researchers are expected to monitor this fraction to ensure comprehensive results and re-apply the LNA step upon encountering new populations.

Comparisons between CoSMIC and other methods were based on the detection of the correct bacteria yet ignored their relative frequencies. The reason for this omission is that frequency estimates are not robust enough in all primer-based methods, including CoSMIC. The main reason is the noisy and biased nature of PCR amplification, yet it also stems from different copy numbers of the 16S rRNA gene(7,23) or intra-individual SSU variations(24). To overcome this problem and allow accurate estimates of relative frequencies we aim to utilize unique molecular identifiers (UMIs) as we previously used in other quantification methods(25) as part of the CoSMIC sample preparations.

In conclusion, we believe that employing CoSMIC will circumvent the problems caused by OTU or ASV tables, allow harmonizing results across studies from the same niche or from different niches, and infer species functions and biological importance. Moving microbiome research beyond mere taxonomic bookkeeping and toward functional and biologically relevant attributes requires rigorous research practices and standards. CoSMIC offers such a framework, allowing a modular and optimized solution for the evolving landscape of microbiome research.

Methods

Samples

Overall, we collected 33 individual samples of plants and their surroundings, one mock community sample, two sample pools derived from multiple plants and their surroundings (collected in distinct locations in Israel), two sample pools derived from sponges collected in Israel and France, and another pool composed of all 33 individual samples used in the study (Table S1). The following subsection describes the samples in each experiment.

Qiagen-processed samples

Leaf, soil and root samples of the following plants were collected: *Aster poliothamnus*, *Brandisia hancei*, *Tetraena mongolica*, *Lathyrus sativus* and *Angelica hirsutiflora*. In addition, two *Lepidium apetalum* plants were collected. In total, 24 samples were processed.

Samples for pooled LNA amplification and PacBio long read sequencing

Four pools were created, each by combining extracted DNA from relevant samples (extraction procedure appears below):

Spongia officinalis France pool: Samples from seven different individuals were collected along the coast of Marseille, France.

Spongia officinalis Israel pool: Samples from three different individuals were collected along the coast of Haifa, Israel.

Arava desert Israel pool: Leaf, root and surrounding soil samples from the following Arava desert species were pooled - *Moringa oleifera, Pulicaria incisa, Ochradenus baccatus, Anabasis articulata, Salsola Vermiculata, Commiphora gileadensis, Ochradenus baccatus, Fagonia mollis, Zilla spinosa,* and *Astersicus gravelonis*.

Temperate lowlands Israel pool: Leaf, root and surrounding soil samples from the following lowlands region species were collected near Rehovot, Israel - *Retama raetam, Ferula communis, Vicia villosa, Heterotheca subaxillaris, Sarcopoterium spinosum, Prasium majus* and *Withania somnifera*.

Swift-processed samples

Leaf, soil and root samples of the following plants were collected: *Hymelaea hirsuta, Solanum lycopersicum* and *Triticum aestivum*.

Swift-processed samples - mock community

We created a mock community of 12 species derived from environmental isolates - *Acidovorax citrulli*, *Bacillus subtilis*, *Paenibacillus dendritiformis*, *Clavibacter michiganensis*, *Escherichia coli*, *Zymomonas mobilis*, *Proteus mirabilis*, *Pseudomonas syringae*, *Xanthomonas campestris*, *Agrobacterium tumefaciens*, *Sphingomonas yanoikuyae*, and *Bacillus pumilus*. Species identification was performed by Sanger sequencing using E8f-939r as primers.

Samples for pooled LNA amplification and Loop Genomics synthetic long reads sequencing

A pool was created for LNA amplification and Loop Genomics synthetic long read sequencing by combining 11 different libraries: A mixture of the 24 samples that were prepared using the Qiagen QIAseq kit; one library per each of the 9 samples that were prepared using the Swift 16S+ITS PANEL kit; and one library for the mock community.

Sample Collection and Preservation

Plant samples were collected in the field and kept in a nucleic acid protection buffer(26). Sponge samples were sampled and preserved in ethanol absolute. Upon arrival at the lab, samples were kept at -80°C until DNA extraction.

DNA Extraction

Genomic DNA was extracted from 100mg of the root and leaf samples, 250mg soil samples, and an approximately 1cm x 0.5cm size piece from the sponges using ZymoBIOMICS DNA Kit (Cat. D4300) based on the manufacturer's protocol.

For the mock community, DNA was extracted from each isolate using GenElute Bacterial Genomic DNA Kits (Cat. NA2110-1KT), and 20ng from each species was used to create a mock community. Pools were created by adding 50ng from each amplicon post LNA amplification and cleaning up to generate a 500ng pool.

Library Preparation and Sequencing

The following sections describe sample preparation and sequencing procedures used for each method.

Qiagen-Processed Samples

A total of 24 samples were processed. Libraries for deep sequencing were prepared using the QIAseq 16S/ITS Region Panel protocol (Cat. 333842), which amplifies the V1V2, V2V3, V3V4, V4V5, V5V7, and V7V9 regions (primer sequences may be available upon request from Qiagen). Sequencing was performed on an Illumina MiSeq using 250|8|8|250 cycles, and the resulting fastq files were analyzed using Qiagen Genomics Workbench software with the CLC Microbial Genomics module.

PacBio Sequencing

16S rRNA full-length amplicons, obtained from the LNA amplification of the *Spongia officinalis* microbial community (France and Israel) and of microbial communities in plants from Israel's lowland and Arava, were used as a starting template for the SMRTbell Express template Prep Kit 2.0 kit (Product PN. 100-938-900). Sequencing was done on the PacBio Sequel sequencer. An initial analysis step of demultiplexing, filtering, and removal of chimeric reads was performed using the SMRT analysis software.

Loop Genomics Sequencing

Samples used for validation of CoSMIC method (Fig. 5) were sequenced using Loop Genomics. 16S rRNA full-length amplicons were used as a template for the LoopSeq PCR Amplicon Kit. Sequencing was done on the Illumina NovaSeq sequencer using the following cycles 151|8|8|151, resulting in 2.5-17 Mbp reads per sample (Fig. S13). Analysis was performed on the company's remote server, yielding fasta files containing assembled long reads.

Shotgun metagenomics sequencing

Extracted DNA from each sample was used with the Nextera DNA Flex Library Prep (Cat. 20018704) to prepare the libraries for metagenomic sequencing. Libraries were sequenced on the Illumina NovaSeq sequencer using the following cycles 151|8|8|151 resulting in 20-80 Mbp reads per sample (Fig. S13 upper panel).

SWIFT-Processed Samples

Libraries for deep sequencing were prepared with the SWIFT AMPLICON™ 16S+ITS PANEL (Cat. AL-51696), using 5 and 4 forward and reverse primers, respectively, pooled in a single reaction (Table 1). Sequencing was performed on the Illumina NovaSeq sequencer with the following cycles: 151|8|8|151 yielding 2.5-15 Mbp reads per sample (Fig. S13 lower panel). Base calling and demultiplexing were performed using Illumina's bcl2fastq command line interface.

Forward primers	Reverse primers
V1_f: GAGTTTGATCMTGGCTCAG	V4_r: CTACCAGGGTATCTAATCC
V3_f: CCTACGGGAGGCAGCAG	V5_r: CCGTCAATTCMTTTGAGTTT
V4_f: GCCAGCAGCCGCGGTAA	V8_r: GACGGGCGGTGTGTACAA
V6_f: ATGGCTGTCGTCAGCT	V9_r: TACCTTGTTACGACTT
V7_f: GYAACGAGCGCAACCC	

Table 1. Primers used in the SWIFT AMPLICON™ 16S+ITS PANEL. Primers applied to amplify the variable regions along the 16S gene as described in the kit's protocol.

Enriching the database with novel SSUs.

Designing LNA primers for full-length 16S rRNA sequences

Using a refined version on the Greengenes database containing 1.4M SSUs[3] we sought short primer pairs (of length 10-13nt) that maximize the number of amplified SSUs. Considering only perfect matches while ignoring thermodynamic considerations, we identified three candidate pairs that potentially amplify 1.2M out the 1.4M SSUs (Table 2). Candidate pairs were provided to Qiagen that replaced specific nucleotides by locked nucleic acids resulting in our LNA primer set.

A.

primer	sequence	Qiagen catalog number
fwd_pair1 (LNA-F-52)	CCACATGCAAG	YCO0190926 (339406)
rev_pair1 (LNA-R-1382)	GGAACGTATTCAC	YCO0190932 (339406)
fwd_pair2 (LNA-F-32)	AACGCTGGCG	YCO0190928 (339406)
rev_pair2 (LNA-R-1376)	CCTATTCACCG	YCO0190929 (339406)
fwd_pair3 (LNA-F-57)	CCGCAAGTCG	YCO0190930 (339406)
rev_pair3 (LNA-R-1406)	GACGGGCGGT	YCO0190931 (339406)

B.

	pair1	pair2	pair3
pair1	1008809	700830	570143
pair2	700830	751633	379021
pair3	570143	379021	712836

Table 2. LNA primers. A. Three primer pairs (pairs 1-3) and their location along the *E. coli* 16S rRNA. Primers are available via the Qiagen catalog number. **B.** The number of amplified SSUs by each pair and by their intersection. About 360K are amplified by all three pairs; 570K are amplified by two pairs; and 250K SSUs are amplified by just a single pair.

Sample preparation

To augment the database with novel SSUs we applied the LNA primer set. The three forward primers and the three reverse primers were combined into two separate mixes at a stock concentration of 20µM.

Samples were individually amplified in an input concentration of 10-50ng. Each sample was amplified using the following volumes: 25µl Kapa HiFi HotStart ReadyMix (cat. KK2602), 1.25µl forward primer pool, and 1.25µl reverse primer pool (each to a final concentration of 0.5µM), 50ng template, and DDW up to 50µl. PCR plan used: 95°C - 3 minutes, 2 cycles: 98°C - 20 seconds, 55°C - 15 seconds, 72°C - 1 minute; 24 cycles: 98°C - 20 seconds, 58°C - 15 seconds, 72°C - 1 minute; 72°C - 5 minutes, hold at 4°C. Following amplification, the products were cleaned using Ampure XP beads (Cat. A63881) at 0.6X concentration. The relevant samples from each niche were pooled at 130ng each (for a full list of samples for each pool, see Table S1).

Data Analysis

Combinatorial estimation of sets of variable regions

SMURF was independently applied to the 63 combinations of the 6 amplified regions produced by the QIAseq 16S/ITS Region Panel. The performance of each combination was evaluated by: (a) ambiguity - the number of SSUs in each SMURF-detected group, i.e., in each set of SSUs having the same sequence over the amplified regions; (b) mutual species - the proportion of correctly identified species, i.e., the number of headers that appear in both groups (six regions-full, and region combination-partial) this parameter would be 1 if the partial combination has all the headers and 0 if they share none with the full six-region analysis; and (c) Pearson correlation between the correctly identified mutual species. A Jupyter Notebook tutorial is available (GitHub rezenman/CoSMIC).

Defining novel SSUs following the sequencing of LNA amplicons

To evaluate the similarity to the SILVA database, candidate SSUs were clustered within each experiment using CD-HIT(20) at 99% identity to remove redundant SSUs. Non-redundant sequences were compared to the SILVA database using the cd-hit-est-2d function to determine the fraction of SSUs that are unmatched to SILVA at different identity levels (Fig. 3A). Script for this analysis appears in the GitHub repository “rezenman/CoSMIC”.

SWIFT reads analysis

Standard analysis: Microbial profiling was performed by the standard Swift analysis scripts, which are based on the Qiime2[19] package.

CoSMIC analysis: Each sample was demultiplexed to its variable regions using cutadapt. In principle, since 5 forward primers were mixed with 4 reverse primers, SMURF can be applied to up to 20 specific regions. We selected 10 potential amplicons (Table 3) while ignoring long amplicons (e.g., the pair V1_f and V9_r was omitted) since the number of such reads is negligible.

Reads that matched the relevant primer pairs (i.e., their R1 and R2 prefixes matched these primer pairs) were filtered and analyzed using R’s DADA2[11] package to account for sequencing inaccuracies and error-corrected reads were submitted to SMURF. Scripts and usage instructions appear in the GitHub repository rezenman/CoSMIC.

Forward primer	Reverse primer
V1_f	V4_r
V1_f	V5_r
V3_f	V4_r
V3_f	V5_r
V4_f	V4_r
V4_f	V5_r
V6_f	V8_r
V6_f	V9_r
V7_f	V8_r
V7_f	V9_r

Table 3. SWIFT primer pairs considered in SMURF. Multiple amplicon types are generated in a multiplex amplification of the SWIFT 9 primer pairs. SMURF analysis considered 10 forward-reverse pairs of adjacent regions. For example, two amplicon types starting with V1_f were considered, i.e., those that end with V4_r or V5_r.

Generating ground truth for benchmarking CoSMIC to other methods

Metagenomic reads were analyzed using PhyloFlash(27) together with the Loop-genomics long reads that were used as trusted contigs. The resulting community composition served as the ground truth for comparing existing analysis pipelines (Swift all and Swift V3V4) to our suggested CoSMIC pipeline. For each sample, reads aligned to the SILVA SSU database (after removing LSU sequences) were kept. Those reads were aligned to one of the following: (a) SSUs extracted from the metaSPAdes(28) assembly of each sample. (b) Full-length SSU sequences obtained by Loop Genomics sequencing of the LNA amplicons; or (c) SILVA database of SSUs, for the reads that were unassigned to the two previous sources. The complete list of SSU served as the ground truth for CoSMIC evaluation purposes.

Calculating ambiguity, alignment identity, explained frequency & erroneous SSUs

The output of SMURF is a set of groups, where each group contains SSUs whose sequence is identical over the amplified regions yet vary over the rest of the gene. We created an analogous definition of a group for standard analysis based on the SILVA database, in which we extracted all relevant SSUs that share the same taxonomy as

decided by the Qiime2 pipeline. This definition is equivalent to SMURF's groups as all SSUs sharing the same taxonomy are indistinguishable based on the reads provided. These groups were used for the comparisons below:

Ambiguity: To calculate ambiguity (Fig. 5B), we counted the number of indistinguishable (equally likely to be the relevant species) matches to the SILVA database for each identification. For example, if a hit to the SILVA database was to the genus level, we defined ambiguity as the number of SILVA SSUs for that genus. To resolve ambiguity for later comparison to the gtSSU, for each identification, we clustered all possible hits to the database using cd-hit and used the most frequent seed as a representative.

Alignment identity: To evaluate the method's performance against our gtSSU, we aligned all seeds produced by each method to its corresponding gtSSU list using blast, and its score serves as the alignment identity.

Explained frequency: The relative proportion originating from the correctly aligned SSUs for a given alignment score.

Erroneous SSUs: Percent of declared SSUs not present in the gtSSU.

Bibliography

1. Hamdi C, Balloi A, Essanaa J, Crotti E, Gonella E, Raddadi N, et al. Gut microbiome dysbiosis and honeybee health. Journal of Applied Entomology. 2011 Aug;135(7):524–33.

2. Johnson KVA, Foster KR. Why does the microbiome affect behaviour? Nat Rev Microbiol. 2018 Oct;16(10):647–55.

3. Nejman D, Livyatan I, Fuks G, Gavert N, Zwang Y, Geller LT, et al. The human tumor microbiome is composed of tumor type-specific intracellular bacteria. Science. 2020 May 29;368(6494):973–80.

4. Livyatan I, Straussman R. A spatial perspective on bacteria in tumours. Nature. 2022 Nov;611(7937):674–5.

5. Parada AE, Needham DM, Fuhrman JA. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. Environ Microbiol. 2016 May;18(5):1403–14.

6. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res. 2013 Jan;41(1):e1.

7. Fouhy F, Clooney AG, Stanton C, Claesson MJ, Cotter PD. 16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform. BMC Microbiology. 2016 Jun 24;16(1):123.

8. Fuks G, Elgart M, Amir A, Zeisel A, Turnbaugh PJ, Soen Y, et al. Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling. Microbiome. 2018 Jan 26;6(1):17.

9. Odom AR, Faits T, Castro-Nallar E, Crandall KA, Johnson WE. Metagenomic profiling pipelines improve taxonomic classification for 16S amplicon sequencing data. Sci Rep. 2023 Aug 26;13(1):13957.

10. Schloss PD. Amplicon Sequence Variants Artificially Split Bacterial Genomes into Separate Clusters. mSphere. 2021 Aug 25;6(4):e0019121.

11. SILVA ribosomal RNA gene database project: improved data processing and web-based tools | Nucleic Acids Research | Oxford Academic [Internet]. [cited 2023 Jun 27]. Available from: https://academic.oup.com/nar/article/41/D1/D590/1069277

12. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB - PMC [Internet]. [cited 2023 Sep 3]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1489311/

13. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, et al. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. mSystems. 2017 Mar 7;2(2):10.1128/msystems.00191-16.

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

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

16. Kurreck J, Wyszko E, Gillen C, Erdmann VA. Design of antisense oligonucleotides stabilized by locked nucleic acids. Nucleic Acids Res. 2002 May 1;30(9):1911–8.

17. Owczarzy R, You Y, Moreira BG, Manthey JA, Huang L, Behlke MA, et al. Effects of sodium ions on DNA duplex oligomers: improved predictions of melting temperatures. Biochemistry. 2004 Mar 30;43(12):3537–54.

18. Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019 Oct;37(10):1155–62.

19. Thomas T, Moitinho-Silva L, Lurgi M, Björk JR, Easson C, Astudillo-García C, et al. Diversity, structure and convergent evolution of the global sponge microbiome. Nat Commun. 2016 Jun 16;7(1):11870.

20. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012 Dec;28(23):3150–2.

21. Callahan BJ, Grinevich D, Thakur S, Balamotis MA, Yehezkel TB. Ultra-accurate microbial amplicon sequencing with synthetic long reads. Microbiome. 2021 Jun 5;9(1):130.

22. Thomas T, Gilbert J, Meyer F. Metagenomics - a guide from sampling to data analysis. Microb Informatics Exp. 2012 Feb 9;2(1):3.

23. Starke R, Pylro VS, Morais DK. 16S rRNA Gene Copy Number Normalization Does Not Provide More Reliable Conclusions in Metataxonomic Surveys. Microb Ecol. 2021;81(2):535–9.

24. Bradshaw MJ, Aime MC, Rokas A, Maust A, Moparthi S, Jellings K, et al. Extensive intragenomic variation in the internal transcribed spacer region of fungi. iScience. 2023 Jul 10;26(8):107317.

25. gUMI-BEAR, a modular, unsupervised population barcoding method to track variants and evolution at high resolution | PLOS ONE [Internet]. [cited 2023 Jun 14]. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0286696

26. Camacho-Sanchez M, Burraco P, Gomez-Mestre I, Leonard JA. Preservation of RNA and DNA from mammal samples under field conditions. Mol Ecol Resour. 2013 Jul;13(4):663–73.

27. Gruber-Vodicka HR, Seah BKB, Pruesse E. phyloFlash: Rapid Small-Subunit rRNA Profiling and Targeted Assembly from Metagenomes. mSystems. 2020 Oct 27;5(5):10.1128/msystems.00920-20.

28. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017 May;27(5):824–34.