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

Standard microbial profiling based on 16S rRNA (16S) sequencing suffers from a lack of primer universality, primer biases and often yields low resolution. We introduce 'Comprehensive Small Ribosomal Subunit Mapping and Identification of Communities' (CoSMIC), addressing these challenges, especially in unexplored niches. CoSMIC begins with long-read sequencing of the full-length 16S gene, amplified by generic Locked Nucleic Acid primers over pooled samples, thus augmenting reference databases with novel niche-specific gene sequences. Subsequently, CoSMIC amplifies multiple non-consecutive variable regions along the gene, followed by short-read sequencing of each sample. Data from the different regions are integrated using the SMURF framework, alleviating primer biases and providing de-facto full gene resolution. Using a mock community, CoSMIC identified full-length 16S genes with significantly higher specificity and sensitivity while dramatically increasing resolution compared to standard methods. Evaluating CoSMIC across environmental samples provided higher accuracy and unprecedented resolution while detecting thousands of novel full-length 16S sequences.

Introduction

During the last decade, with the advent of massively parallel sequencing, the study and exploration of microbiomes have 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). Most of these studies rely on sequencing predefined variable regions of the 16S rRNA gene using short reads. However, it has been widely shown that the results of such studies depend on the amplified region due to biases introduced by the specific

53 primers applied(5,6), hence limiting the ability to harmonize findings across studies(7).
54 Moreover, due to short read length, standard methods amplify one or two adjacent
55 variable regions; therefore, many bacteria may share their sequence over the
56 amplicon yet be very different along the rest of the 16S rRNA gene. Such ambiguity in
57 assigning reads to bacteria often results in lower resolution profiling. However, current
58 bioinformatic pipelines do not adequately address either problem, regardless of
59 whether they use amplicon sequence variants (ASVs) or operational taxonomic units
60 (OTUs).
61 To overcome these problems, we have previously presented the Short MUlti-Region
62 Framework (SMURF) approach(8). In SMURF, each sample is amplified by any
63 number of primer pairs, and reads from different regions are computationally combined
64 to provide coherent, high-resolution microbial profiling. For such integration, SMURF
65 relies on available databases of full-length 16S rRNA sequences (e.g., SILVA(9) or
66 Greengenes(10)) containing millions of sequences. These databases are almost
67 comprehensive with respect to human-associated bacteria, while other habitats may
68 be severely underrepresented(11,12), rendering the approach suboptimal for less
69 explored niches.
70 Here, we present Comprehensive 16S rRNA gene Mapping and Identification of
71 Communities (CoSMIC), which builds on SMURF to provide a framework for high-
72 resolution profiling of novel niches. Before applying SMURF in a new niche, we
73 suggest amplifying a pool of samples using generic locked nucleic acid (LNA) primers,
74 which provide an almost universal amplification of full-length 16S rRNA sequences.
75 Sequencing these amplicons by long-read technologies (e.g., PacBio) enriches 16S
76 rRNA gene databases with relevant sequences, thus allowing SMURF to provide
77 efficient and high-resolution profiling of these niches. We provide a workflow for
78 applying CoSMIC detailing the selection of specific subsets of primers to enable cost
79 savings while not compromising resolution. The advantages of CoSMIC over standard
80 methods are demonstrated by sequencing a mock mixture and over a variety of
81 environmental samples.
82 CoSMIC is best suited for profiling novel ecological niches. It provides de facto full-
83 length 16S rRNA resolution at about 10% of the cost of per-sample PacBio
84 sequencing. This cost-effectiveness, combined with CoSMIC's flexibility in primer
85 selection, allows researchers to optimize their experimental designs for both scope
86 and budget. In addition, the full-length 16S rRNA sequences detected in CoSMIC's

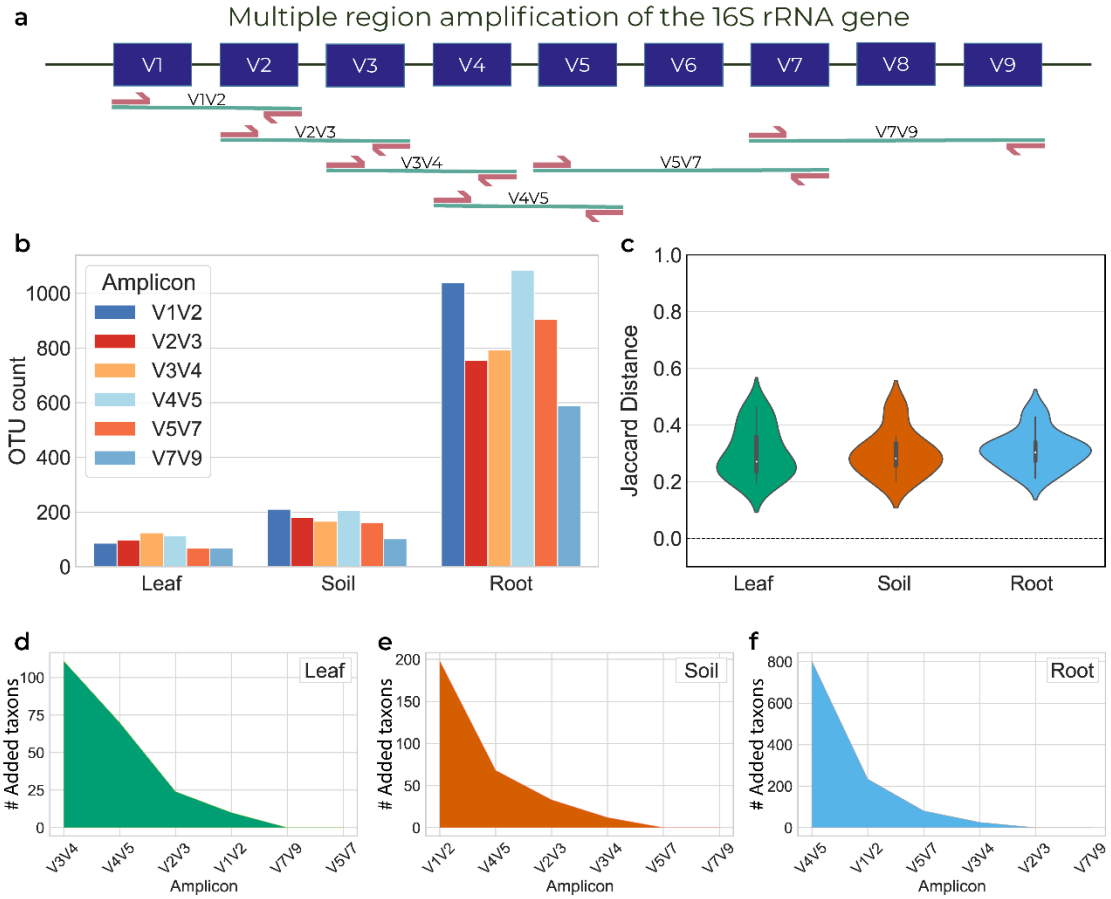
87 initial step may be uploaded to public 16S rRNA databases, thus allowing other
88 researchers to perform even more cost-effective profiling studies over these niches.

89 Results

90 The outcome of microbial profiling highly depends on amplified 91 regions

92 To demonstrate the effect of preselecting a single variable region (formerly shown,
93 e.g., by (5,6)), we performed 16S rRNA profiling of leaves, roots, and their surrounding
94 soil from seven different plants (see Methods). Profiling was conducted using the
95 commercial Qiagen QIAseq 16S/ITS Region Panel(13), which amplifies six regions
96 along the gene (Fig. 1A), resulting in 200K-600K reads per sample (Supplementary
97 Figure 1). We used Qiagen's CLC Genomics Microbiome package, which applies a
98 closed-reference Operational Taxonomic Units (OTU) picking strategy using a
99 similarity threshold of 97% (see Methods). The package independently analyzes each
100 region by assigning taxonomy to OTUs detected in a region and outputs the
101 taxonomies detected across all regions.
102 Fig. 1B illustrates the number of OTUs detected in leaf, soil, and root samples of *Aster*
103 *poliothamnus* using each of the six amplified regions. For example, the same root
104 sample results in 600-1100 OTUs, depending on the amplified region (Fig. 1B, right).
105 Moreover, there is a low overlap in assigned taxonomies detected by different regions,
106 as evident by the beta diversity among all pairs of amplified regions (Fig. 1C). Beta
107 diversity was calculated by a Jaccard distance and is higher than zero for any pair of
108 amplified regions (median Jaccard distances of 0.27, 0.30 and 0.28 for leaf, root, and
109 soil, respectively).
110 To evaluate the contribution of each amplicon to the analysis of a given sample type,
111 we identified the amplicon having the largest number of unique OTUs and sequentially
112 added regions according to their additional unique OTUs with respect to regions
113 selected thus far (Fig.1D-F). The "top" region varied across samples, e.g., the most
114 "informative" regions for leaf, soil, and root in the case of *Aster poliothamnus* were
115 V3V4, V1V2, and V4V5, respectively. In addition, it required four regions to exhaust
116 the number of unique OTUs. Yet, the identity of these regions varied among leaf, soil,

117 and root samples, thus highlighting the inherent biases in region selection.
118 Comparable observations were made when profiling the other plants - *Brandisia*
119 *hancei*, *Lepidium apetalum*, *Lathyrus sativus, Angelica hirsutiflora,* and *Tetraena*
120 *mongolica* (Supplementary Figure 2-6), where the most "informative" amplicon was
121 inconsistent both across plants for a given niche (e.g., leaves across all plants), and
122 over niches for each plant (Supplementary Figure 7-9).



123

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

133

Combining results from several variable regions partially overcomes problems in profiling

We previously suggested employing the Sample MUlti-Region Framework (SMURF) approach(8) 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 to solve an optimization problem, seeking the gene sequences 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 full-length 16S rRNA genes having an identical sequence over the amplified regions. For example, suppose V1V2 and V7V9 regions are being amplified. In that case, full-length 16S rRNA genes having the same sequence over these two regions are 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 the identification of a *single* full-length gene sequence 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 as could be expected from the added regions; *(b)* relative detection, representing the fraction of identified 16S rRNA sequences out of those identified when all six regions are used (results based on all six regions are the best possible in this case; hence, all subsets are compared to them); and *(c)* Pearson correlation between the frequencies of 16S rRNA sequences detected by a set of regions vs. those detected by the six-region combination.

The average performance across all three parameters improved when more regions were incorporated into the SMURF analysis. The average ambiguity across all 24 samples was 20,000 sequences when using a single region. Yet, by adding only two

more regions to the analysis, ambiguity was reduced to less than 2,000 (Fig. 2A). Secondly, the capacity to accurately detect a substantial proportion of the mutual species increased linearly (Fig. 2B). Since sequences belonging to the same group differ by un-sequenced regions, including each additional region removed sequences that were inaccurately declared by the smaller number of regions. Thirdly, as more amplicons are added, the correlation improves significantly, with many samples achieving a correlation coefficient (r) higher than 0.9 by adding 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, combining 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. 2E, red cross) and leaf (Supplementary Figure 10, red cross).

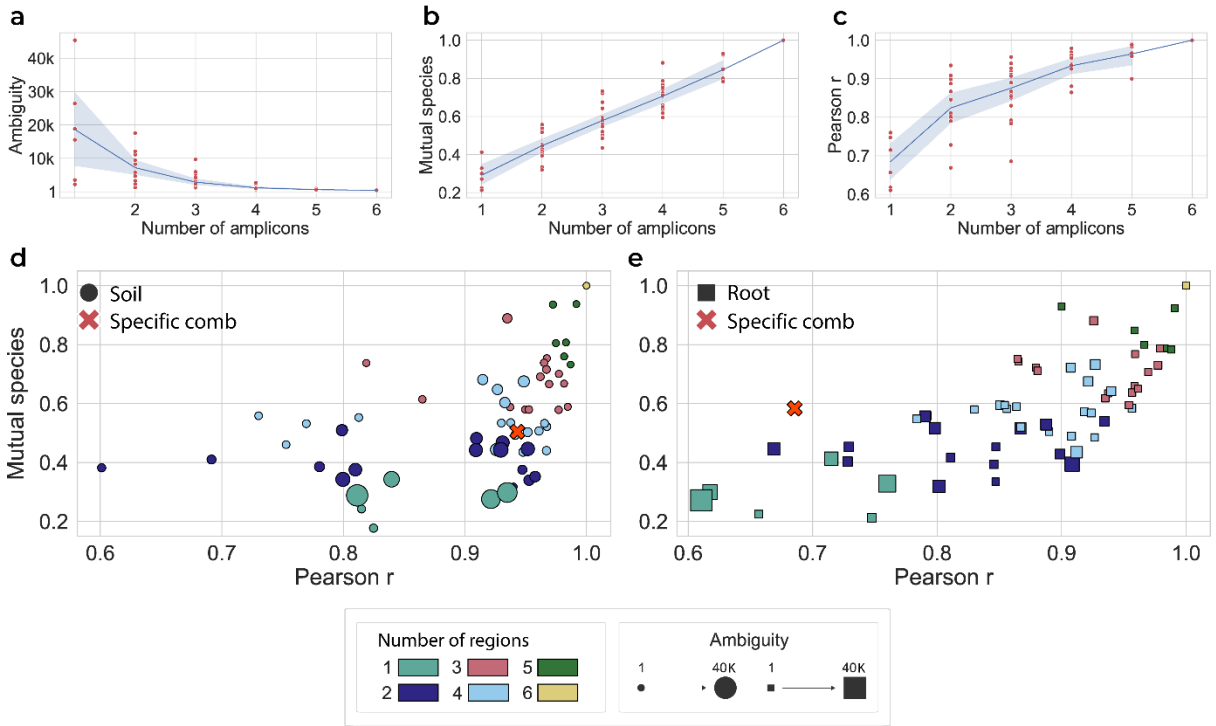


Figure 2. Evaluating the performance of sets of primer pairs across niches. Twenty-four samples corresponding to the leaf, soil, and root of eight plants were profiled by SMURF, using all 63 combinations of the six amplicons. **a.** Ambiguity, i.e., the number of 16S rRNA gene sequences per group as 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 the gold standard and display a fraction of species identified by subsets of the six regions. Shown is the fraction of species whose taxonomy matches the six-region solution. **c.** Pearson correlation between profiling based on all six regions and each set of regions. **d.** A comparison of amplicon sets for soil samples. Each dot corresponds to a specific region combination averaged across 24 samples, 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 six regions (yellow), and six dark green dots correspond to the different groups of five 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 full-length 16S rRNA sequence databases. However, these databases are highly enriched for human-associated bacteria, while other habitats are often poorly covered.

208 To bridge this gap and harness the advantages of SMURF across habitats, we
209 propose a simple preceding step in a microbiome sampling project designated to
210 provide niche-specific full-length 16S rRNA sequences. The procedure entails a PCR
211 step, utilizing a generic pre-designed multiplexed set of locked nucleic acid (LNA)
212 primers aimed at effectively amplifying the whole 16S rRNA gene. LNA primers(14,15)
213 are shorter than regular primers (10-13nt instead of ~20nt), thus allowing an increased
214 universality compared to regular primers while maintaining primer specificity and high
215 Tm (see Methods). We, therefore, propose to pool samples from the relevant niche,
216 apply PCR amplification using the pre-designed LNA primers, and sequence by long-
217 reads. To account for potential chimeras, the resulting reads undergo several quality
218 assurance steps and are clustered at 99% identity to remove redundancy. A
219 representative sequence from each cluster is aligned to the SILVA NR99 database(9),
220 and sequences whose alignment to SILVA is lower than 99% are considered novel
221 (for a detailed description of the abovementioned steps, see Methods). We refer to the
222 database containing SILVA and the novel sequences as “augmented SILVA”.
223 To illustrate this idea and estimate the number of novel sequences that can be
224 detected, we used this approach over pools of samples, each from a different niche:
225 (a) sponges of the species *Spongia officinalis* collected from the Mediterranean coast
226 of France; (b) *Spongia officinalis* collected from the Mediterranean coast of Israel; (c)
227 plants from the Arava desert in Israel (climate type, BWh); and (d) plants from the
228 temperate lowland in Israel (climate type, Csa). These pools were sequenced using
229 PacBio long reads amplicon sequencing(16), resulting in ~200K high-quality, full-
230 length 16S rRNA sequences (Supplementary Figure 11) of average length 1600 bp
231 (Supplementary Figure 12).
232 Figure 3A presents the percentage of novel sequences as a function of the identity
233 threshold when aligning to SILVA. Sponges seem to have significantly higher
234 percentages of novel sequences(17), irrespective of identity threshold. Using the
235 SILVA definition of a novel full-length 16S rRNA sequence, i.e., a threshold of 99%
236 identity (vertical dashed line in Fig. 3A), more than 95% of detected sequences were
237 novel for samples of sponges and about 85% for samples taken from plants. These
238 novel sequences include many whose relative abundance in the pool is high,
239 especially in sponges. Fig. 3B shows histograms of the number of PacBio reads across
240 detected 16S rRNA sequences for the pools of plants and sponges (left and right
241 columns, respectively) when split into known SILVA sequences and novel sequences

242 (upper and lower rows, respectively). For sponges, 98% of PacBio sequences are
243 novel, compared to 90% in plants. Moreover, 17 novel sponge sequences appear in
244 more than 1000 reads, potentially corresponding to abundant sequences, compared
245 to zero such sequences in plants (analogously, 214 and 25 sequences have more
246 than 100 reads for sponges and plants, respectively). Hence, LNA amplification and
247 sequencing detect novel sequences, some of potentially significant frequency,
248 especially in the less-explored sponge niche. When subsequently performing short-
249 read sequencing over individual sponge samples and performing SMURF using the
250 augmented SILVA, more than 93% of the relative frequency in each sample was
251 attributed to novel sequences (Fig. 3C, two left columns). Therefore, standard SMURF
252 is inadequate for these sponge samples. In contrast, on average, 90% of our plant
253 samples were attributed to known SILVA sequences (Fig. 3C, two right columns).
254 Another way to evaluate the significance of novel sequences is to calculate the
255 percentage of short reads assigned to each database, thus estimating its "utility" (Fig.
256 3D). More than 82% of short reads in each sponge sample align to augmented SILVA
257 compared to 58% that align to SILVA (Fig. 3D upper panels). We note that each short
258 read may align to SILVA and novel sequences, yet the additional 24% of the reads
259 aligned to only novel sequences "drive" SMURF's optimization to reject SILVA
260 sequences and select novel sequences as part of the profiling solution. The case for
261 plant samples differs as more than 80% of the reads are aligned to SILVA, and novel
262 sequences contribute only an additional 8% and 4% for BWh and Csa, respectively
263 (Fig. 3D lower panels). Consequently, SMURF's solution attributes most of the relative
264 abundance to SILVA sequences.
265 Another observation is that 91% and 97% of the novel sequences for sponges and
266 plants, respectively, appear in less than 10 PacBio reads (Fig. 3B). Also, a rarefaction
267 curve plotted over these pools does not reach a plateau for either sponges or plants
268 (Supplementary Figure 13). Although part of these sequences may correspond to
269 chimeric sequences undetected by the applied algorithms, they may also, in part, be
270 a manifestation of underexplored communities. Hence, additional long-read
271 sequencing may detect novel sequences that may close the remaining gap in
272 unaligned reads.

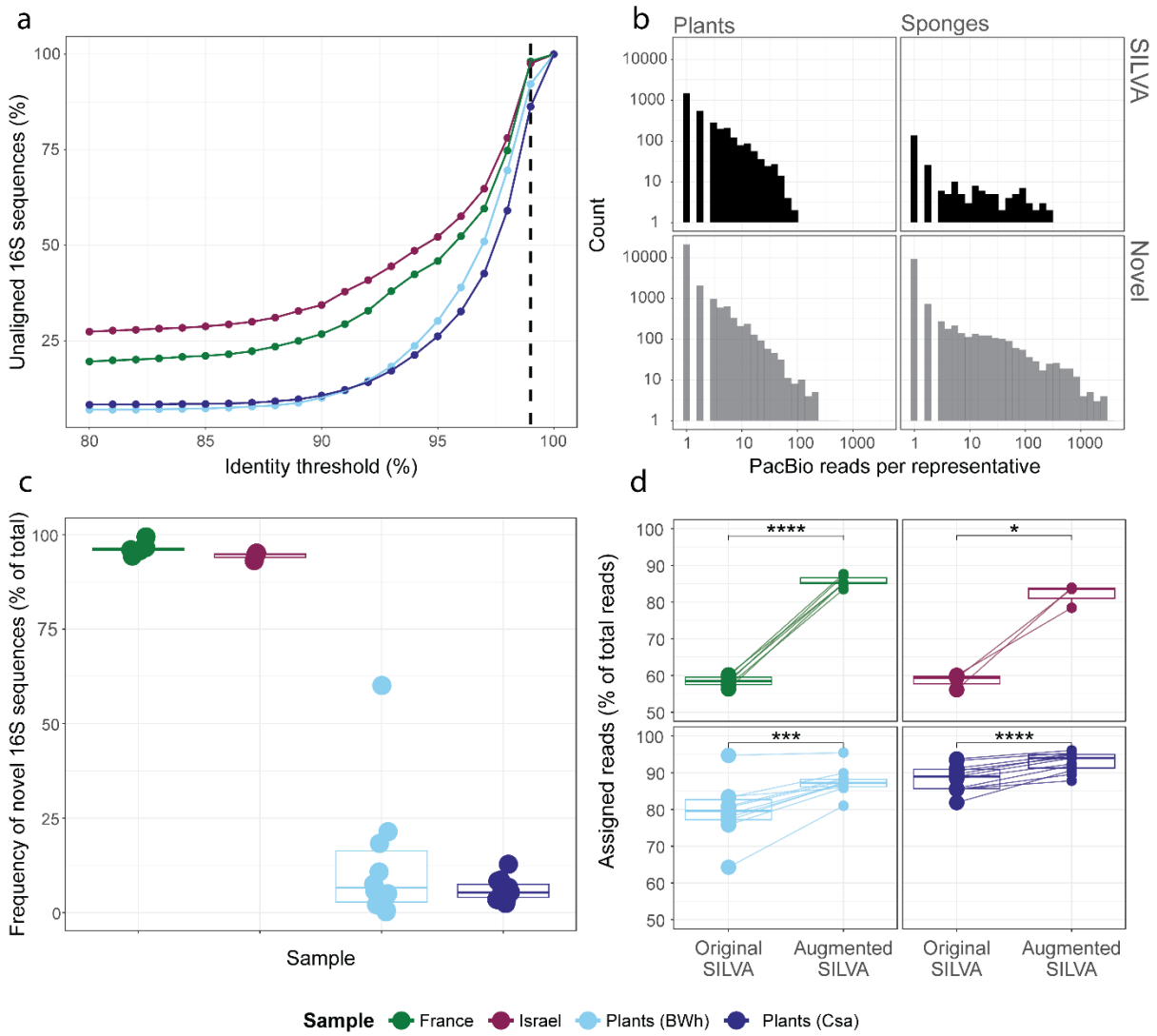


Figure 3. Enriching the database with novel full-length 16S rRNA sequences. We amplified the full-length 16S rRNA gene using pre-designed LNA primers and sequenced the resulting amplicons by PacBio long-read sequencing. Four samples were processed, each containing a pool of samples from a different habitat, 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 PacBio sequences (Plants BWh - 14852, Plants Csa - 14589, France - 6507, Israel - 5468) were aligned to SILVA using the CD-HIT algorithm(30). The percentage of sequences 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 full-length 16S rRNA sequences (SILVA NR99). Colors indicate different biomes sampled. **b.** Histograms of read counts per 16S rRNA sequence for the plants' pool (left panels) and the sponges' pool (right panels). The upper and lower panels show histograms of SILVA and novel sequences, respectively. An identity threshold of 99% was used to call novel sequences with respect to SILVA. **c.** Individual samples from the abovementioned pools were sequenced via short reads and analyzed by SMURF using the augmented database to determine their microbial composition (France – 6 samples, Israel – 3 samples, Plants (BWh) – 10 samples, Plants (Csa) – 10 samples). The total relative frequency assigned by SMURF to novel 16S rRNA sequences per sample (dots) and across types (columns) is shown. **d.** The fraction of reads aligned to either SILVA or augmented SILVA for each

291 sample depicted in C. Each panel corresponds to a different sample type. To evaluate the statistical significance,
292 a paired t-test was performed for each sample type; * P<0.05, *** P<0.001, **** P<0.0001.
293

294 The CoSMIC approach

295 When considering the problems and solutions presented in the former sections, we
296 propose guidelines for large-scale microbiome studies that would provide high-
297 resolution profiling in a new or known niche while accommodating different budget
298 tiers (Fig. 4). We recommend treating each research project as a new ecological niche
299 (orange, right-hand side in the flowchart) and enrich current databases with relevant
300 full-length 16S rRNA gene sequences, i.e., pool multiple samples of different origins
301 of the ecological niche, apply LNA amplification and perform long reads sequencing.
302 Next, using the newly enriched database, we propose to sequence a representative
303 set of samples using the most extensive set of primers available (e.g., all nine Swift
304 pairs or all six Qiagen pairs), apply SMURF, and select the subset of primer pairs that
305 provide sufficient resolution per niche (or select all primer pairs in case costs do not
306 scale considerably with the number of primers). Once these steps are finalized, the
307 project transitions to a “familiar niche” case (purple, left-hand side), which performs
308 routine amplification of the optimal primer set followed by short reads sequencing,
309 subsequently analyzed by augmented-database SMURF.

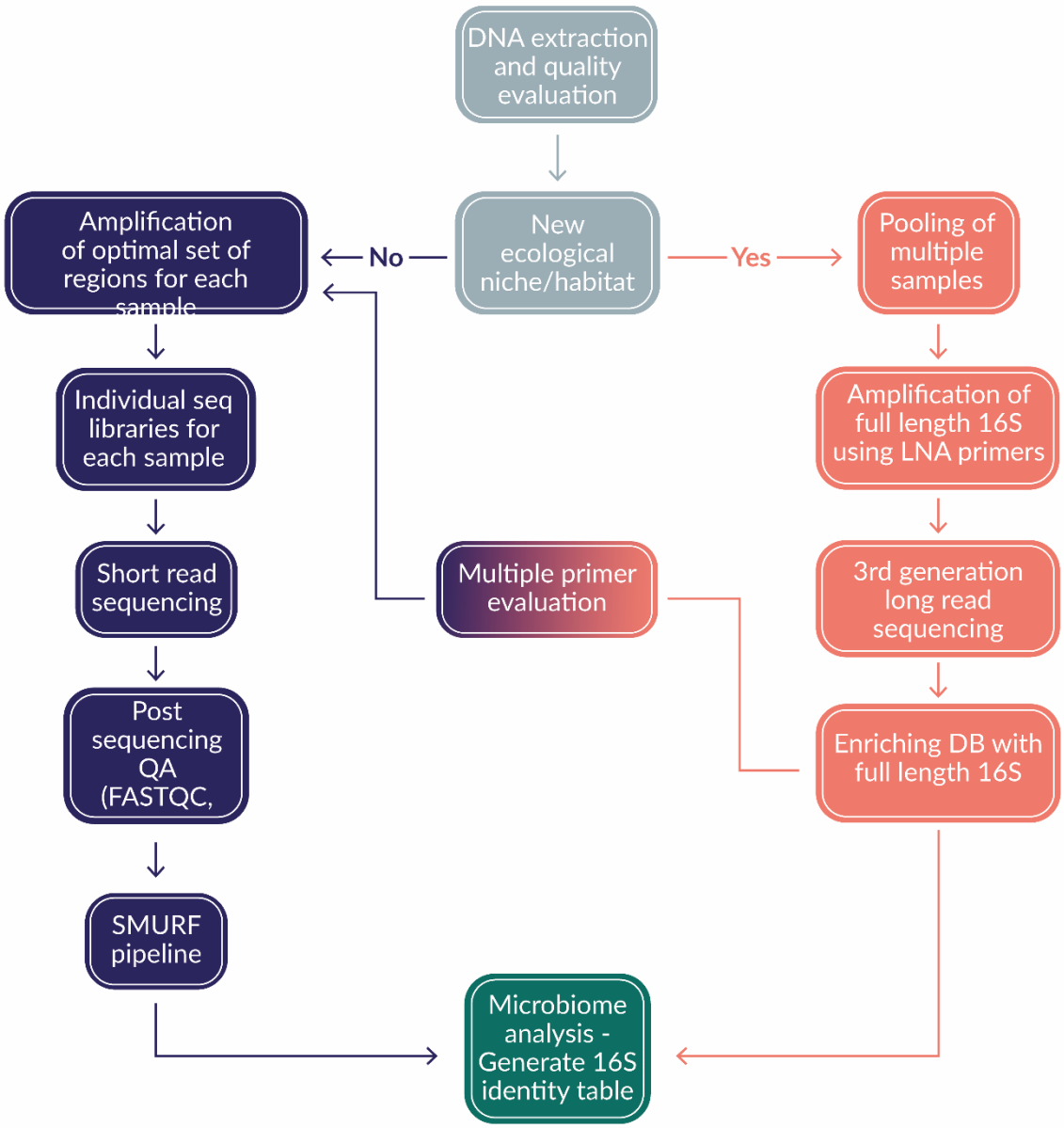


Figure 4. The suggested CoSMIC flowchart. Gray - Perform standard DNA extraction from environmental samples. **Orange -** Establish a niche-relevant full-length 16S rRNA gene database by LNA-based amplification and long-read sequencing. **Orange-purple -** Test multiple primer sets to find an optimal 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.

318 # Experimental evaluation of CoSMIC

319 To evaluate the advantages of CoSMIC, we sampled *Hymelaea hirsuta*, *Solanum*
320 *lycopersicum,* and *Triticum aestivum* at their leaves, roots, and surrounding soil. In
321 addition, we created a mock community of 12 known species. To construct our ground
322 truth, we performed standard shotgun metagenomic sequencing of each of the ten
323 samples and used phyloFlash(18) to extract 16S rRNA sequences and their relative
324 frequencies in each sample (see Methods). We refer to these sequences as ground
325 truth (gt16S), and their relative frequencies are called ground truth frequencies. An
326 optimal performance corresponds to detecting the exact full-length sequences of
327 gt16S while adding no other sequence. Note that each species in the community may
328 contain several different 16S rRNA sequences, thus the task is to detect all these
329 sequences.

330 As the first step in CoSMIC, we pooled DNA from all ten samples and used our LNA
331 primers to amplify the full-length16S rRNA gene. We sequenced the amplicons using
332 synthetic long reads technology (Loop Genomics) to achieve full-length 16S rRNA
333 gene sequences. Novel sequences, with respect to the SILVA NR99 standard, were
334 added to SILVA to create our reference database.

335 Each sample was then sequenced using Illumina short reads using the Swift pipeline,
336 which amplifies several regions along the 16S rRNA gene (see Methods section for
337 detailed information about obtained reads for Loop, Swift, and shotgun
338 metagenomics). To apply SMURF, we used our reference database while considering
339 nine amplicons spanning the 16S rRNA gene based on the abovementioned primers.
340 For comparison, we provide a standard analysis provided by Swift utilizing the
341 QIIME2(19) package to generate tables of operational taxonomic units (OTUs) or
342 amplicon sequence variants (ASVs) utilizing DADA2(20), each based on either all
343 amplified regions or only the widely used V3V4 amplicon.

344 Comparison to the ground truth of each sample was based on four criteria: (a)
345 *Alignment identity* - the average alignment score between detected 16S rRNA
346 sequences and gt16S; (b) *Explained frequency* - the sum of ground truth relative
347 frequencies for gt16S that were also detected by each method; (c) *Erroneous 16S*
348 *rRNA genes sequences* - the number of detected sequences that were not present in
349 gt16S; and (d) *Ambiguity* - the number of sequences in the reference database that
350 match detected sequences by each method (for details regarding a-d see Methods).

351 CoSMIC outperforms traditional methods on all parameters (Fig 5).
352 *Alignment identity* - Mean and overall alignment identities are higher for CoSMIC
353 across all samples (Fig. 5A). The distribution varies according to the tissue sampled,
354 with mock, root, and soil consistently producing better scores than leaf tissue. This is
355 likely related to the low number of bacteria detected due to chloroplasts occupying
356 most reads. While the difference between analysis methods is evident across all
357 tissues, the most striking example is the mock community scores. Swift-based
358 analysis, based on all regions or only on V3V4, had a median alignment identity of
359 about 90% even for this simple, well-defined mixture, compared to an almost 100%
360 alignment identity for CoSMIC.
361 *Explained frequency* - Significant improvements are also apparent in the accuracy of
362 community composition. Fig. 5B presents the total explained frequency as a function
363 of the identity threshold averaged per sample type. Using CoSMIC, more sequences
364 are accurately reported, accounting for a higher proportion of the ground-truth sample.
365 The difference between the methods is particularly evident when only high-quality
366 alignments are considered, resulting in almost zero frequency explained by traditional
367 methods.
368 In three of the four cases, such reduction also seems to affect CoSMIC, yet to a lesser
369 degree. To understand this phenomenon, we considered the mock mixture whose
370 species composition is known. The significant drop in explained frequency occurs at
371 an identity threshold of 98.4% and mainly originates from a mismatch between an *E.*
372 *coli* sequence in gt16S and the detected CoSMIC sequence. Aligning Illumina short
373 reads of the Swift pipeline to both sequences showed a perfect match to the detected
374 CoSMIC sequence, while the gt16S *E. coli* sequence contained 12 SNPs relative to
375 the reads. Therefore, the actual value of CoSMIC's explained frequency at identity
376 thresholds higher than 99% is at least 88%. This indicates that phyloFlash-derived
377 16S rRNA sequences may not be entirely accurate, demonstrating the inherent
378 problems in achieving highly accurate assembly of 16S rRNA sequences from shotgun
379 metagenomics. This problem is not restricted to phyloFlash and also characterizes
380 other state-of-the-art assemblers(21,22). To alleviate this problem yet provide an
381 estimate solely based on shotgun results, we used MarkerMAG(23) to improve the
382 gt16S sequences (see Methods), yielding an explained frequency of 89% for identity
383 thresholds higher than 99% (Fig. 5B left panel, grey dashed line). This also indicates

384 that most of the improvement in gt16S is attributed to the errors in the abovementioned
385 *E. coli* sequence.
386 *Erroneous 16S rRNA sequences* - The percentage of erroneous sequences, i.e., those
387 reported yet not present in the ground truth sample, is significantly lower for CoSMIC
388 than the other methods (Fig. 5C). Moreover, in three of the four cases, the number of
389 erroneous sequences using standard methods grows linearly with the identity
390 threshold, while CoSMIC displays an increase starting from a threshold of 95%. To
391 understand this phenomenon in the case of the mock mixture, we used the
392 MarkerMAG-improved sequences as ground truth, yielding a reduction of 9.5% (Fig.
393 5C left panel, grey dashed line). Although the percentage of erroneous sequences at
394 a high identity threshold is still significant, we note that since MarkerMAG-based
395 sequences explain 89% of the mock mixture sequences at an identity threshold of
396 99%, the erroneous sequences correspond to only ~11% of the population. Moreover,
397 the absolute number of these erroneous sequences is merely 12 (Supplementary
398 Figure 14). In addition, these errors may also be another manifestation of the
399 limitations of shotgun-based ground truth. Although more than 60 million reads were
400 allocated to shotgun sequencing of the mock mixture (Supplementary Figure 15), only
401 ~88,000 were assigned to 16S rRNA sequences. Hence, the ground truth may be
402 missing some sequences, and the actual erroneous sequences at 99% identity may
403 be lower than 12.
404 *Ambiguity* - Fig. 5D presents a histogram of the ambiguity across all 16S rRNA
405 sequences detected in the ten samples. Most OTUs or ASVs detected by Swift
406 correspond to hundreds or thousands of the database's 16S rRNA sequences, while
407 CoSMIC detects a *single* sequence, yielding a more precise, actionable result.
408 Figure 5E combines all criteria using radar plots for each method, clearly showing
409 CoSMIC's improved performance, especially in addressing ambiguity (Amb) and
410 explained frequency at high alignment identity scores (Cov-99). CoSMIC outperforms
411 standard pipelines regardless of methodology (OTU or ASV) and whether multi-region
412 or single-region analysis is used. While standard methods can identify species (Id) at
413 low sequence resolution (Cov-80), they often result in many erroneous species (FP)
414 not present in the sample and have very high ambiguity (Amb). As expected, adding
415 more regions to the standard analysis does little to improve results since these
416 methods do not combine inputs from the different regions.
417

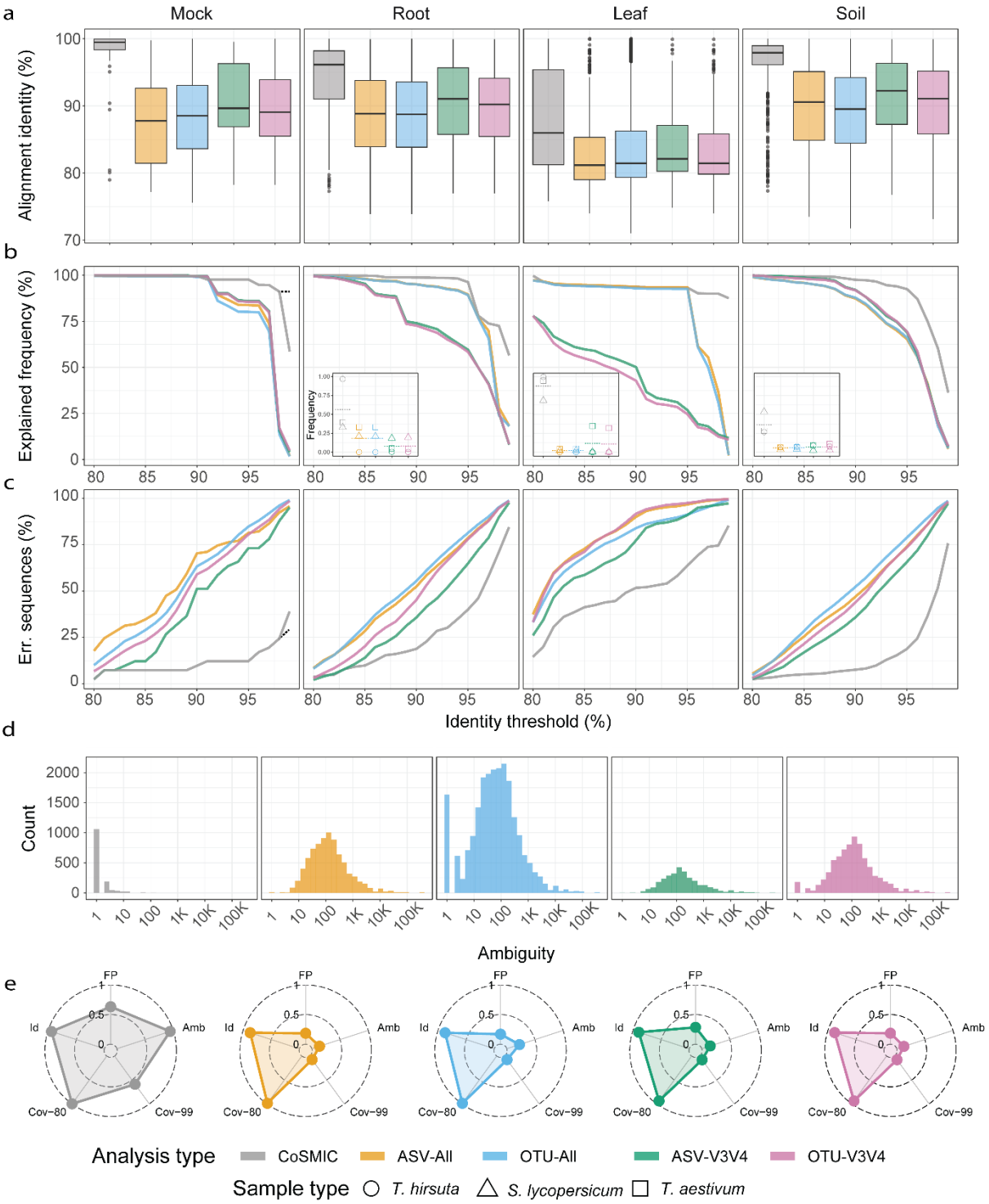


Figure 5. Experimental evaluation of standard methods vs. CoSMIC. Ten samples were sequenced - leaf, soil, and root samples of *Hymelaea hirsuta*, *Solanum lycopersicum,* and *Triticum aestivum*, and a mock mixture of 12 known species (see Methods). Each sample underwent standard shotgun metagenomic sequencing, from which we detected its full-length 16S rRNA sequences to be considered ground truth. In addition, as part of CoSMIC, we pooled extracted DNA from all ten samples, amplified by LNA primers, and sequenced via synthetic long reads (Loop Genomics) to create full-length 16S rRNA sequences 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 CoSMIC using

428 nine primer pairs and by a standard QIIME2 pipeline. The latter used either nine primer pairs (All) or the V3V4
429 region only (V3V4) and was performed using OTUs or by ASV calling. Hence, four QIIME2-based results are shown
430 – OTU-All, OTU-V3V4, ASV-All, and ASV-V3V4. **a.** Alignment identity, i.e., the average alignment score between
431 detected 16S rRNA sequences and ground truth 16S rRNA genes (gt16S) detected by shotgun metagenomics.
432 Boxplots present the distribution of alignment identities for each tissue and analysis method. Leaf, soil, and root
433 panels correspond to an average of over three samples (one from each plant). A Mann-Whitney U test was
434 performed for all comparisons, which resulted in very low p-values, with the highest being 0.0064. **b.** The total
435 frequency of correctly detected gt16S by each method as a function identity threshold. The dotted horizontal line
436 in the left panel (mock mixture) marks the results if we improve the ground truth sequences by MarkerMag. Insets
437 present the total frequency achieved per sample at a cutoff of 99%; the dotted horizontal line is their average. The
438 y-axis, shown for the leftmost inset, is relevant for all insets. **c.** Percentage of erroneous 16S rRNA sequences, i.e.,
439 those identified by an analysis method yet not detected by shotgun metagenomics. The percentage is a function
440 of the identity threshold, where alignments below the threshold are considered erroneous. The dotted line marks
441 the MarkerMag-based improvement for the mock community using CoSMIC for up to a 99% identity threshold. **d.**
442 Histogram of the number of taxa matching an identification by each analysis type. **e.** Radar plot combining all
443 samples over several criteria (normalized on a 0-1 scale, where large values correspond to better performance):
444 **Id,** mean alignment identity for each analysis method as described in A; **Cov-80,** mean total ground truth frequency
445 at 80% identity threshold as described in B; **Cov-99,** the same for 99% identity threshold; **FP,** One minus the mean
446 false positive rate as described in C; **Amb,** mean ambiguity as described in d.

447 Discussion

448 The diverse applications of microbiome studies have underscored the need for robust
449 and standardized methodologies(24). However, primer selection and arbitrary analysis
450 decisions made by standard methods lead to lower profiling resolution (Figs. 5A, D)
451 and fall short due to the omission of species (Fig. 5B) or the inclusion of species that
452 are not present in the sample (Fig. 5C).

453 To address these limitations, we formerly proposed applying sets of primers from
454 different variable regions and computationally combining their results via the SMURF
455 approach(8). To do so, SMURF requires a comprehensive database of full-length 16S
456 rRNA sequences. Yet, such comprehensiveness holds mainly for human-associated
457 samples, and SMURF's performance is sub-optimal for other niches (e.g., Fig. 3C-D).
458 CoSMIC aims to bridge this gap in a cost-effective way by introducing a preliminary
459 step of proactively enriching the database with niche-relevant full-length 16S rRNA
460 sequences. We pool samples from the novel niche and apply amplification by our
461 generic LNA-based (almost) universal primers followed by long-read sequencing.
462 Subsequently, one can apply short-read sequencing over each sample and apply
463 SMURF using the augmented database to achieve high-resolution, more accurate

464 profiling results (Fig. 3C-D, Fig. 5). Applying CoSMIC is more important in less
465 explored environments, e.g., marine sponges, where close to 100% of the reads align
466 to novel 16S rRNA gene sequences detected by CoSMIC (Fig. 3C). Therefore, we
467 hope CoSMIC will be adopted by researchers seeking efficient and accurate large-
468 scale profiling of less explored niches.

469 To our surprise, CoSMIC also outperformed standard methods when profiling a mock
470 community of known species. The average alignment score between 16S rRNA
471 sequences detected by CoSMIC and the ground truth sequences was 97.2% (Fig. 5A
472 left panel), while the score of standard methods was lower by ~8%. This corresponds
473 to a ~120bp difference along the gene, thus potentially leading to a lower resolution.
474 Also observed was an overestimation of the diversity of 16S rRNA sequences over the
475 mock mixture, evident in both ASV or OTU-based standard analysis (Fig. 5C left
476 panel). Hence, CoSMIC provides a more robust estimation of this important ecological
477 property.

478 Overall costs per sample in CoSMIC (including all expenses - from DNA extraction to
479 those related to short- and long-read sequencing) currently range between $25 to $30,
480 depending on the number of primer pairs applied. Hence, overall costs only marginally
481 depend on the number of primer pairs since the expensive stages are DNA extraction
482 and other sample preparation steps, which do not depend on the number of applied
483 primer pairs. Also, amplification is often applied in multiplex; hence, sample
484 preparation and labor may be comparable to using a single primer pair. Moreover,
485 large-scale projects may benefit from a 10-15% reduction in costs associated with the
486 number of applied primer pairs. Therefore, we advise researchers to estimate the
487 minimal subset of primer pairs that provide adequate results for their samples (e.g.,
488 Fig. 2).

489 The straightforward approach of using our generic LNA primers and applying PacBio
490 long-read sequencing to individual samples rather than pools of samples is possible
491 and would achieve equivalent results to CoSMIC. However, its costs become
492 prohibitive for large projects. For example, the estimated cost for PacBio sequencing
493 is about $300 per sample for tens of samples and even higher costs for larger cohorts
494 due to the price of additional indices. Hence, CoSMIC provides a means of
495 circumventing such high costs while providing analogous results.

496 Researchers may also turn to shotgun metagenomics techniques to retrieve 16S rRNA
497 genes, yet the aims of shotgun sequencing differ from mere profiling, and their costs
498 are much higher. Moreover, state-of-the-art assemblers often fail to provide highly
499 accurate 16S rRNA gene sequences of a sample's organisms (e.g., the *E. coli*
500 sequence in our mock mixture of Fig. 5). The gene's conserved regions across species
501 and often its almost identical copies within the same organism hampers accurate
502 assembly. As a result, relying solely on metagenomics-based approaches for 16S
503 rRNA profiling may provide less accurate results than CoSMIC.

504 While the CoSMIC approach represents an advancement in microbiome analysis, it
505 requires careful quality assurance by researchers. In CoSMIC, a novel niche refers to
506 cases where a significant fraction of the reads cannot be matched to the available
507 database of full-length 16S rRNA sequences. Hence, researchers are expected to
508 monitor this fraction to ensure comprehensive results and re-apply the LNA step upon
509 encountering new communities.

510 Adding irrelevant or incorrect full-length 16S rRNA sequences to an ad-hoc CoSMIC
511 database does not affect its reconstruction results. For example, thousands of 16S
512 rRNA gene sequences that appeared as single PacBio reads and were part of the
513 augmented database did not participate in any of CoSMIC's reconstructions. Also,
514 chimeric reads that end up in an augmented database cannot be part of CoSMIC's
515 reconstructions since there will not be a set of short reads that supports such an
516 erroneous sequence. However, as performed here, we advise researchers to apply
517 chimera-detection (e.g., intrinsic algorithms by PacBio, VSEARCH
518 uchime_denovo(25), etc.) prior to adding sequences to the ad-hoc database to keep
519 the database "clear" of irrelevant sequences. Moreover, when compiling a set of novel
520 sequences to be added to a public general-purpose 16S rRNA database, or a niche-
521 specific pool of sequences, researchers should only consider sequences detected by
522 SMURF in at least one sample. By this, databases would only contain "validated"
523 sequences, i.e., those that passed both chimera-detection algorithms and were
524 detected by SMURF. This would potentially allow researchers of the same niche to
525 directly apply the short-read part of CoSMIC and skip long-read sequencing once an
526 adequate "coverage" of the environment has been achieved.

527 Comparisons between CoSMIC and other methods were based on detecting the
528 correct bacteria yet ignoring their relative frequencies. This omission is due to the fact
529 that frequency estimates are not robust enough in all primer-based methods, including
530 CoSMIC. These inaccuracies mainly result from noise and biases in PCR
531 amplification, yet they also stem from different copy numbers of the 16S rRNA
532 genes or different sequences of the gene in the same cell. To overcome this problem
533 and allow accurate estimates of relative frequencies, we aim to utilize unique
534 molecular identifiers (UMIs) as part of the CoSMIC sample preparation, similar to our
535 former quantification method(26).

536 In conclusion, employing CoSMIC will circumvent the problems caused by standard
537 profiling methods, allow harmonizing results across studies from the same niche or
538 from different niches, and potentially aid in inferring species functions and biological
539 importance. We hope that this framework, allowing low-cost, high-resolution microbial
540 profiling of any niche, would provide a modular and optimized solution for the evolving
541 landscape of microbiome research.

542

Methods

Sample collection and preservation

Overall, collected samples consisted of 33 individual 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 (Supplementary Table 1). Leaf, root, and soil samples were collected in the field and kept in a nucleic acid protection buffer(27). Sponge samples were collected and immediately 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 of soil samples, and an approximately 1cm x 0.5cm piece from the sponges using the ZymoBIOMICS DNA Kit (Cat. D4300), following 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.

The following subsections describe the specific samples and all steps per processing pipeline (the order of the presentation corresponds to the Results sections).

The Qiagen processing pipeline as used in Figures 1 and 2

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.

570 **Library preparation and sequencing**

571 Libraries for deep sequencing were prepared using the QIAseq 16S/ITS Region Panel
572 protocol (Cat. 333842), which amplifies the V1V2, V2V3, V3V4, V4V5, V5V7, and
573 V7V9 regions (primer sequences may be available upon request from Qiagen).
574 Sequencing was performed on an Illumina MiSeq using 250|8|8|250 cycles.

575 **Qiagen-based analysis in Figure 1**

576 The resulting FASTQ files were analyzed using the QIAGEN CLC Genomics
577 Workbench software with the CLC Microbial Genomics module, which performs
578 closed-reference Operational Taxonomic Unit (OTU) analysis. More specifically, the
579 software preprocesses the raw sequencing data, aligns the reads against a reference
580 database (Greengenes(10) clustered at 99%), and clusters the sequences into OTUs
581 based on a similarity threshold of 97%. It then assigns taxonomy to each OTU and
582 generates an OTU table containing the frequency of each OTU in each sample.

583 **Combinatorial estimation of sets of variable regions in Figure 2**

584 The same FASTQ files were provided to SMURF which was independently applied to
585 the 63 combinations of the six amplified regions (Fig. 2). The performance of each
586 combination was evaluated by: (a) ambiguity - the number of sequences in each
587 SMURF-detected group, i.e., in each set of 16S rRNA genes having the same
588 sequence over the amplified regions; (b) mutual species - the proportion of correctly
589 identified species, i.e., the number of headers that appear in both groups (six regions-
590 full, and region combination-partial) this parameter would be 1 if the partial
591 combination has all the headers and 0 if they share none with the full six-region
592 analysis; and (c) Pearson correlation between the correctly identified mutual species.
593 A Jupyter Notebook tutorial is available (GitHub NoamShental/CoSMIC).

594 Enriching the database with novel 16S rRNA gene sequences

595 **Samples for pooled LNA amplification and PacBio long-read sequencing in**
596 **Figure 3**

597 Four pools were created, each by combining extracted DNA from relevant samples:
598 *Spongia officinalis France pool*: Samples from seven different individuals were
599 collected along the coast of Marseille, France.
600 *Spongia officinalis Israel pool*: Samples from three different individuals were collected
601 along the coast of Haifa, Israel.

602 *Arava desert Israel pool*: Leaf, root and surrounding soil samples from the following
603 Arava desert species were pooled: *Moringa oleifera, Pulicaria incisa, Ochradenus*
604 *baccatus, Anabasis articulata, Salsola vermiculata, Commiphora gileadensis,*
605 *Ochradenus baccatus, Fagonia mollis, Zilla spinosa,* and *Astersicus gravelonis*.
606 *Temperate lowlands Israel pool*: Leaf, root, and surrounding soil samples from the
607 following lowlands region species were collected near Rehovot, Israel: *Retama*
608 *raetam, Ferula communis, Vicia villosa, Heterotheca subaxillaris, Sarcopoterium*
609 *spinosum, Prasium majus* and *Withania somnifera*.
610

611 **Samples for pooled LNA amplification and Loop Genomics synthetic long reads**
612 **sequencing in Figure 5**

613 A pool was created for LNA amplification and Loop Genomics synthetic long-read
614 sequencing(28) by combining 11 different libraries: a mixture of the 24 samples
615 prepared using the Qiagen QIAseq kit; one library per each of the nine samples
616 prepared using the Swift 16S+ITS PANEL kit; and one library for the mock community.
617

618 **Designing LNA primers for full-length 16S rRNA sequences**

619 Using a refined version of the Greengenes database containing 1.4M full-length 16S
620 rRNA gene sequences(10) we sought short primer pairs (of length 10-13nt) that
621 maximize the number of amplified genes. Considering only perfect matches while
622 ignoring thermodynamic considerations(14), we identified three candidate pairs that
623 potentially amplify 1.2M out of the 1.4M sequences (Table 1). Candidate pairs were
624 provided to Qiagen to replace specific nucleotides with locked nucleic acids, resulting
625 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

626

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

631

632 **Sample preparation**

633 The LNA primer set was applied to augment the database with novel full-length 16S
634 rRNA gene sequences to be sequenced by either PacBio or Loop Genomics pipelines.
635 The forward and reverse primers were combined into two separate mixes at a stock
636 concentration of 20μM.
637 Samples were individually amplified at an input concentration of 10-50ng. Each
638 sample was amplified using the following volumes: 25μl Kapa HiFi HotStart ReadyMix
639 (cat. KK2602), 1.25μl forward primer pool, and 1.25μl reverse primer pool (each to a
640 final concentration of 0.5μM), 50ng template, and DDW up to 50μl. PCR plan used:
641 95°C - 3 minutes, 2 cycles: 98°C - 20 seconds, 55°C - 15 seconds, 72°C - 1 minute;
642 24 cycles: 98°C - 20 seconds, 58°C - 15 seconds, 72°C - 1 minute; 72°C - 5 minutes,
643 hold at 4°C. Following amplification, the products were cleaned using Ampure XP
644 beads (Cat. A63881) at 0.6X concentration. The relevant samples from each niche
645 were pooled at 100 ng/μl each (for a full list of samples for each pool, see
646 Supplementary Table 1).

647

648 **PacBio sequencing**

649 16S rRNA full-length amplicons, obtained from the LNA amplification of the *Spongia*
650 *officinalis* microbial community (France and Israel) and of microbial communities in
651 plants from Israel's lowland and Arava, were used as a starting template for the
652 SMRTbell Express template Prep Kit 2.0 kit (Product PN. 100-938-900). Sequencing
653 was done on the PacBio Sequel sequencer. An initial analysis step of demultiplexing
654 and filtering was performed using the SMRT(28) analysis software. The software also
655 detects and discards chimeric reads (reads are eliminated if two SMRTbell adaptors
656 are sequenced).

657 **Loop Genomics sequencing**

658 Samples used to validate the CoSMIC method (Fig. 5) were sequenced using Loop
659 Genomics(29). 16S rRNA full-length amplicons were used as a template for the
660 LoopSeq PCR Amplicon Kit. Sequencing was done on the Illumina NovaSeq
661 sequencer using the following cycles 151|8|8|151, resulting in 2.5-17 Mbp reads per
662 sample (Supplementary Figure 16). Analysis was performed on the company's remote
663 server, yielding fasta files containing assembled long reads.

664

665 **Filtering potential chimeras**

666 Although the SMRT analysis software already filters out potential chimeric reads, and
667 to support chimera detection for Loop Genomics reads, the uchime_denovo of
668 VSEARCH(25) was applied to detect and remove chimeric sequences. While no
669 chimeric reads were found among full-length 16S rRNA sequences that were to be
670 "augmented" to our database, we acknowledge the possibility that such sequences
671 may still be present. The formation of chimeric sequences during 16S rRNA
672 amplification, especially for full-length genes, has been well-documented in the
673 literature (29). Combining chimera-detection algorithms and the fact that CoSMIC's
674 profiling is robust to chimeric sequences provides significant assurances regarding the
675 validity of detected sequences.

676

677 **Calling novel 16S rRNA genes following the sequencing of LNA amplicons**

678 Candidate 16S rRNA gene sequences were clustered within each experiment using
679 CD-HIT(30) at 99% identity to remove redundant sequences. Non-redundant
680 sequences were compared to the SILVA database using the cd-hit-est-2d function to
681 determine their alignment to SILVA at different identity levels (Fig. 3A). Script for this
682 analysis appears in the GitHub repository "NoamShental/CoSMIC". This script was
683 used for sequences of both PacBio and Loop Genomics.

684 The Swift processing pipeline as used in Figure 5

685 **Samples**

686 Leaf, soil, and root samples of *Hymelaea hirsuta, Solanum lycopersicum,* and *Triticum*
687 *aestivum* were collected.

688 A mock community was formed using 12 species obtained from environmental isolates
689 - *Acidovorax citrulli*, *Bacillus subtilis*, *Paenibacillus dendritiformis*, *Clavibacter*
690 *michiganensis*, *Escherichia coli*, *Zymomonas mobilis*, *Proteus mirabilis*,
691 *Pseudomonas syringae*, *Xanthomonas campestris*, *Agrobacterium tumefaciens*,
692 *Sphingomonas yanoikuyae*, and *Bacillus pumilus*. Species identification was
693 performed by Sanger sequencing using E8f-939r as primers.

694 **Library preparation and sequencing**

695 Libraries for deep sequencing were prepared with the Swift AMPLICON™ 16S+ITS
696 PANEL (Cat. AL-51696), using five forward and four reverse primers pooled in a single
697 reaction (Table 2)(31). Sequencing was performed on the Illumina NovaSeq
698 sequencer with the following cycles: 151|8|8|151, yielding 2.5-15 Mbp reads per
699 sample (Fig. S15 lower panel). Base calling and demultiplexing were performed using
700 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	

701 **Table 2. Primers used in the Swift AMPLICON™ 16S+ITS PANEL.** Primers were applied to amplify the variable
702 regions along the 16S rRNA gene as described in the kit's protocol.

703 **Data analysis**

704 *Standard analysis*: Microbial profiling was performed using the standard Swift analysis
705 scripts to create Amplicon Sequence Variant (ASV) or OTU tables. These tables were
706 based on the QIIME2(19) package, where paired R1 and R2 reads were filtered and
707 analyzed using R's DADA2(20) package to account for sequencing errors.
708 *CoSMIC analysis*: Each sample was demultiplexed to its variable regions using
709 Cutadapt(32). In principle, since five forward primers were mixed with four reverse
710 primers, SMURF can be applied to up to 20 specific regions. We selected nine
711 potential amplicons (Table 3) while ignoring long amplicons (e.g., the pair V1_f and
712 V9_r was omitted) since the number of such reads was negligible. Paired R1 and R2
713 reads were filtered, analyzed using R's DADA2 package to account for sequencing
714 errors, and submitted to SMURF.

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 applied in SMURF-based reconstruction. Multiple amplicon types are generated in a multiplex amplification of the Swift nine primer pairs. SMURF analysis considered nine forward-reverse pairs of adjacent regions. For example, two amplicon types starting with V3_f were considered, i.e., those that end with V4_r or V5_r.

Generating ground truth 16S rRNA sequences in Figure 5

Samples

The same samples used for the Swift processing pipeline were used to generate ground truth data.

Library preparation and 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. S15 upper panel).

Data analysis – generating ground truth 16S rRNA sequences

Ground truth sequences were extracted using phyloFlash(18), which employs a multi-step approach to identify full-length 16S rRNA sequences from metagenomic data. First, phyloFlash assembles the metagenomic reads using metaSPAdes(33) to generate metagenome-assembled genomes (MAGs). Next, it aligns the metagenomic reads to the SILVA database using a low identity threshold and retaining reads that align to 16S rRNA gene sequences. These reads are then aligned to three reference sets: (a) Trusted contigs, i.e., full-length 16S rRNA gene sequences obtained by either Loop Genomics sequencing or PacBio; b) 16S rRNA genes extracted from the

737 metaSPAdes assembly of each sample using Barrnap(34); and (c) The SILVA
738 database of 16S rRNA genes, for reads that were unassigned to the two previous
739 sources.
740 This alignment step allows phyloFlash to determine the composition of the bacterial
741 community in the sample. The list of full-length 16S rRNA sequences from the trusted
742 contigs, assembled MAGs, and SILVA database alignments served as the ground
743 truth 16S rRNA sequences (gt16S).

744 **Data analysis – improving candidate 16S rRNA sequences by MarkerMAG**
745 To improve the ground truth sequences for the mock mixture further, MarkerMAG(23)
746 was applied, which uses a set of marker genes to identify and validate the presence
747 of specific taxons in metagenomic data. The *matam_16s* script in MarkerMAG, which
748 is based on the SILVA database, was utilized to generate a list of 16S rRNA marker
749 sequences for the taxons present in our sample. Next, the marker sequences were
750 linked to contigs assembled by metaSPAdes using the *link* script in MarkerMAG. This
751 step allowed us to validate the marker list by confirming that the marker sequences
752 were linked to the assembled contigs.
753 To evaluate alignment identity and the erroneous frequency, the validated marker
754 sequences of the mock mixture were compared against the output of CoSMIC.

755 Calculating alignment identity, explained frequency, erroneous 756 16S rRNA sequences, and ambiguity in Figure 5

757 The output of SMURF is a set of groups, where each group includes 16S rRNA genes
758 whose sequence is identical over the amplified regions yet varies over the rest of the
759 gene. An analogous definition of a group for standard analysis was created based on
760 the SILVA database, in which we extracted all relevant 16S rRNA gene sequences
761 that share the same taxonomy set by the QIIME2 pipeline. For example, if QIIME2
762 provides a genus-level taxonomy for a certain ASV or OTU, its “group” comprises all
763 SILVA sequences of this genus. This definition is equivalent to SMURF’s groups as
764 all 16S rRNA sequences sharing the same taxonomy are indistinguishable based on
765 the reads provided.
766 Groups of QIIME2 or CoSMIC were used for the comparisons below:
767 **Ambiguity (Fig. 5D):** Ambiguity is the number of full-length sequences in a group. Fig.
768 5D displays a histogram of group sizes for both CoSMIC and QIIME2-based analyses.

769 The other measures in Fig. 5 required processing the groups. The sequences in each
770 group were clustered using CD-HIT, and the most frequent sequence was selected as
771 the group's representative.
772 **Alignment identity (Fig 5A):** Each representative sequence was aligned to its
773 corresponding gt16S list using Blast, and the gt16S sequence with the highest
774 alignment was selected. The unweighted average of Blast's identity score served as
775 the alignment identity.
776 **Explained frequency (Fig. 5B):** Overall relative frequencies of groups whose
777 representative 16S rRNA gene sequences aligned to gt16S for a given alignment
778 score.

779 **Erroneous 16S rRNA gene sequences (Fig. 5C):** Percentage of 16S rRNA
780 sequences that did not appear in gt16S out of the total number of sequences detected
781 by each method.
782

783 The CoSMIC software package

784 The software package (REF) provides a comprehensive set of scripts facilitating all
785 steps of CoSMIC. The package includes scripts augmenting an existing database by
786 full-length 16S rRNA sequences obtained from long-read sequencing (e.g., PacBio).
787 It then provides code that formats the augmented database in a SMURF-like format
788 per the specific primers' set to be used in the short-read sequencing experiments. The
789 package includes scripts for denoising and demultiplexing paired-end Illumina short
790 reads using the DADA2 algorithm and running SMURF. The package supports
791 SMURF profiling based on any subset of the original set of primers.

792 **Hardware requirements**
793 The CoSMIC package was optimized for efficiency, requiring a 20-core machine with
794 16GB of RAM per core, a specification within reach for most academic settings. With
795 adjustments, our processes can run on standard PCs, albeit more slowly, making our
796 pipeline accessible and practical for a wide range of research environments without
797 the need for exceptional computational resources.
798 Documentation for code and processing requirements is provided at the GitHub
799 repository.

References

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

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. 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

10. 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/

11. Prodan A, Tremaroli V, Brolin H, Zwinderman AH, Nieuwdorp M, Levin E. Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. PLoS One. 2020;15(1):e0227434.

12. Castelle CJ, Banfield JF. Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. Cell. 2018 Mar 8;172(6):1181–97.

13. 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.

838 14. Kurreck J, Wyszko E, Gillen C, Erdmann VA. Design of antisense
839 oligonucleotides stabilized by locked nucleic acids. Nucleic Acids Res. 2002 May
840 1;30(9):1911–8.

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

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

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

850 18. Gruber-Vodicka HR, Seah BKB, Pruesse E. phyloFlash: Rapid Small-Subunit
851 rRNA Profiling and Targeted Assembly from Metagenomes. mSystems. 2020 Oct
852 27;5(5):10.1128/msystems.00920-20.

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

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

859 21. Pericard P, Dufresne Y, Couderc L, Blanquart S, Touzet H. MATAM:
860 reconstruction of phylogenetic marker genes from short sequencing reads in
861 metagenomes. Bioinformatics. 2018 Feb 15;34(4):585–91.

862 22. Yuan C, Lei J, Cole J, Sun Y. Reconstructing 16S rRNA genes in metagenomic
863 data. Bioinformatics. 2015 Jun 15;31(12):i35–43.

864 23. Song W, Zhang S, Thomas T. MarkerMAG: linking metagenome-assembled
865 genomes (MAGs) with 16S rRNA marker genes using paired-end short reads.
866 Bioinformatics. 2022 Aug 2;38(15):3684–8.

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

869 25. Edgar RC. UCHIME2: improved chimera prediction for amplicon sequencing
870 [Internet]. bioRxiv; 2016 [cited 2024 May 21]. p. 074252. Available from:
871 https://www.biorxiv.org/content/10.1101/074252v1

872 26. Rezenman S, Knafo M, Tsigalnitski I, Barad S, Jona G, Levi D, et al. gUMI-
873 BEAR, a modular, unsupervised population barcoding method to track variants
874 and evolution at high resolution. PLOS ONE. 2023 Jun 7;18(6):e0286696.

27. 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.

28. Picelli S, Faridani OR, Björklund ÅK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc. 2014 Jan;9(1):171–81.

29. 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.

30. 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.

31. GitHub - swiftbiosciences/16S-SNAPP-py3: 16S SNAPP on Python 3 [Internet]. [cited 2024 Feb 14]. Available from: https://github.com/swiftbiosciences/16S-SNAPP-py3

32. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011 May 2;17(1):10–2.

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

34. Seemann T. tseemann/barrnap [Internet]. 2024 [cited 2024 Jun 6]. Available from: https://github.com/tseemann/barrnap

Declarations

Ethics approval and Consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

Raw reads used in this study can be obtained from the NIH SRA database BioProject ID:PRJNA1172999 -https://www.ncbi.nlm.nih.gov/bioproject/1172999. Code and supporting data are available at https://github.com/NoamShental/CoSMIC

Acknowledgments

Shira Holand for scientific illustrations and graphic design; Micha Ilan for the collections of *Spongis officinallis* from Israel; Theirry Perez and Marie Grenier for the collections of *Spongis officinallis* from France. David Pilzer for PacBio sequencing support. Jonathan Nutkiewitz and Elad Jacobi-Carmeli for helpful thoughts and discussions. Funding was provided by Sustainability and Energy Research Initiative by TI; The Yotam project and the Weizmann Institute sustainability and energy research initiative to DS; Combined Weizmann-Abroad Postdoctoral Grant for Advancing Women in Science and the Emergency Fellowships for Israeli Postdocs Affected by COVID-19 to TI. The Open University of Israel internal grant 10024 to NS.

Authors' contributions

MK, SR, DS, and NS conceived and initiated the project; TI and DS collected most of the samples; MK, SR, and DS designed and carried out the experiments; MK, SR, and NS wrote the data analysis code and the software package; ME, SD, RK, and ZR were involved in the work design and have substantively revised it. MK, SR, TI, and NS prepared the manuscript.

All authors have approved the submitted version (and any substantially modified version that involves the author's contribution to the study), read and approved the final manuscript, and have agreed both to be personally accountable for the author's contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature.

Competing interests

The authors declare that they have no competing interests.