Published online 7 November 2013 *Nucleic Acids Research, 2013, Vol. 41, No. 22* ***e205***
doi:10.1093/nar/gkt1070

High-resolution microbial community reconstruction by integrating short reads from multiple 16S rRNA regions

Amnon Amir[1], Amit Zeisel[1], Or Zuk[2,3], Michael Elgart[4], Shay Stern[4], Ohad Shamir[5], Peter J. Turnbaugh[6], Yoav Soen[4] and Noam Shental[7,*]

[1]Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel, [2]Toyota Technological Institute at Chicago, Chicago, IL 60637, USA, [3]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA, [4]Department of Biological Chemistry, Weizmann Institute of Science, Rehovot 76100, Israel, [5]Microsoft Research, Cambridge, MA 02142, USA, [6]FAS Center for Systems Biology, Harvard University, Cambridge, MA 02138, USA and [7]Department of Mathematics and Computer Science, The Open University of Israel, Raanana 43537, Israel

Received February 26, 2013; Revised October 8, 2013; Accepted October 14, 2013

ABSTRACT

The emergence of massively parallel sequencing technology has revolutionized microbial profiling, allowing the unprecedented comparison of microbial diversity across time and space in a wide range of host-associated and environmental ecosystems. Although the high-throughput nature of such methods enables the detection of low-frequency bacteria, these advances come at the cost of sequencing read length, limiting the phylogenetic resolution possible by current methods. Here, we present a generic approach for integrating short reads from large genomic regions, thus enabling phylogenetic resolution far exceeding current methods. The approach is based on a mapping to a statistical model that is later solved as a constrained optimization problem. We demonstrate the utility of this method by analyzing human saliva and *Drosophila* samples, using Illumina single-end sequencing of a 750 bp amplicon of the 16S rRNA gene. Phylogenetic resolution is significantly extended while reducing the number of falsely detected bacteria, as compared with standard single-region Roche 454 Pyrosequencing. Our approach can be seamlessly applied to simultaneous sequencing of multiple genes providing a higher resolution view of the composition and activity of complex microbial communities.

INTRODUCTION

Microorganisms comprise the majority of living organisms on our planet, both in terms of biomass (1) and species diversity [estimated between 10^7 and 10^9 species (2)]. Elucidating the composition of microbial communities is important for understanding ecological systems in nature as well as for pathological scenarios in the clinic.

Owing to the difficulty of culturing microbial species under laboratory conditions (2–4), comprehensive characterization of community structure often relies on conserved marker genes, such as the 16S ribosomal RNA gene (16S). Current methods focus on sequencing a few variable 16S regions flanked by highly conserved domains enabling selective isolation of the relevant regions using 'universal' polymerase chain reaction (PCR) primers (Figure 1A). Millions of 16S sequences have been deposited in the past decade into databases such as Greengenes (5), SILVA (6) and RDP (7), which integrate data from a large number of projects. These, in turn, allow phylogenetic analysis and microbial profiling, namely the identification of bacteria in a sample and their frequency.

The recent introduction of Massively Parallel Sequencing (MPS), also commonly known as Next-Generation Sequencing (8,9), and its combination with DNA barcoding for sample multiplexing (10) have greatly increased the yield of bacterial community analysis [reviewed in (11)]. These improvements have enabled large-scale studies involving hundreds of different individuals (12) or time points (13). In a typical MPS-based experiment, a short

*To whom correspondence should be addressed. Tel: +972 9 7781252; Fax: +972 9 7781615; Email: shental@openu.ac.il

The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

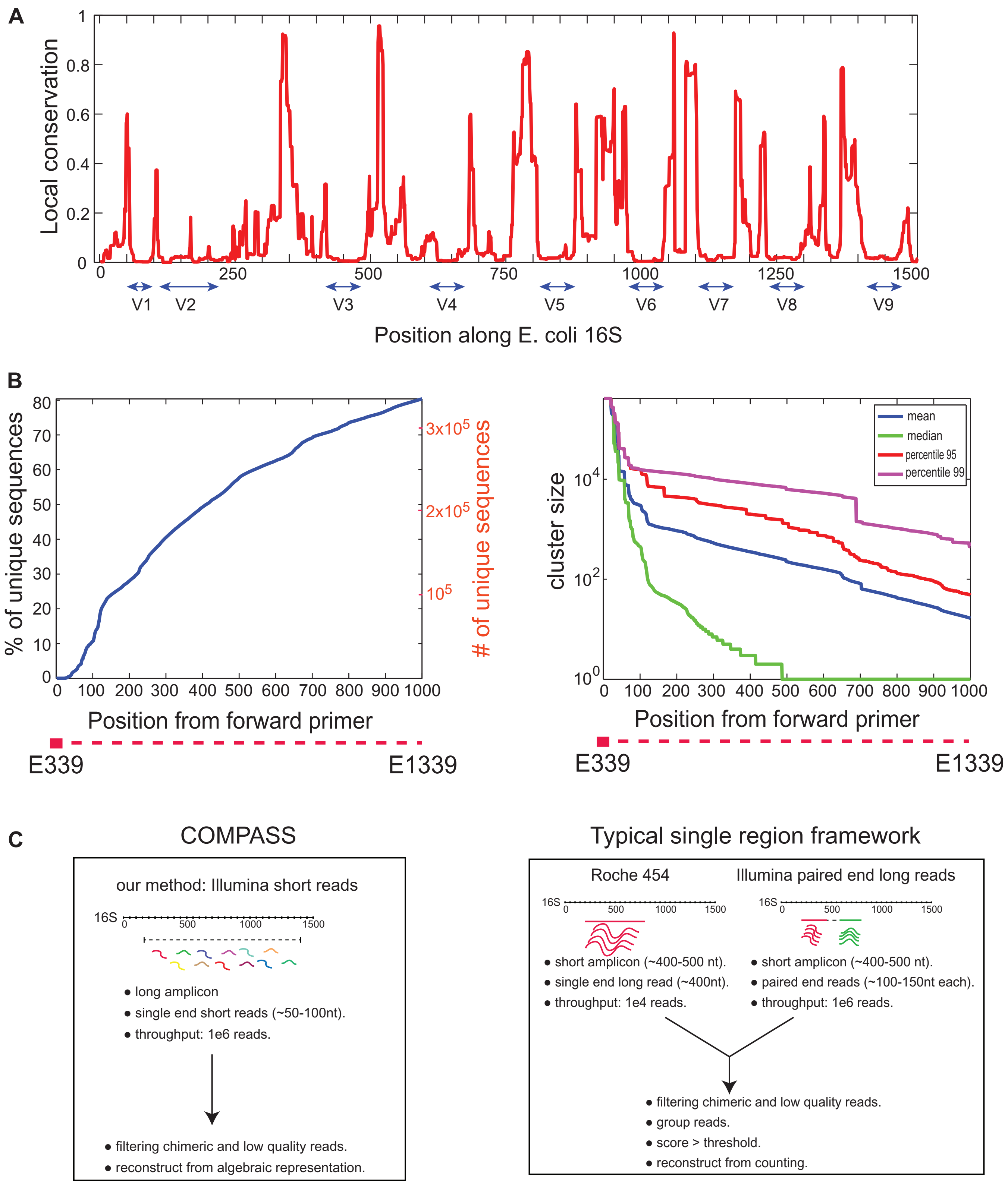

Figure 1. Overview of COMPASS for 16S rRNA microbial profiling. (**A**) Local conservation along the 16S rRNA gene using *E. coli* 16S sequence as reference: The fraction of bacterial database sequences matching the *E. coli* 15-mer sequence is shown as function of the position along the gene. Commonly used variable regions are marked by blue arrows. Data are based on the Greengenes database (5) containing 455 055 sequences. (**B**) Integrating evidence from a larger region increases species resolution. We used the highly conserved sequence E339 to test the gain in resolution achieved by extending the amplicon. We aligned the ~400 000 sequences that contain this primer (out of the full 16S database of ~450 000 sequences), and counted the number of unique sequences as a function of the amplicon length (left). For example, using the first 200 nt results in only ~100 000 unique sequences. We can then count how many (full length) sequences share the same first 200 bp and calculate different statistics, e.g. the average or median groups' sizes etc. Quantiles of the group size distribution versus the region length are presented in the right panel, showing

(continued)

segment covering one or two variable regions of the 16S is amplified and sequenced. The length of the segment is ~300 or 400 bases (out of the entire ~1500 bases of the 16S gene) depending on the use of Illumina paired-end or Roche 454 machines, respectively.

The identity and frequency of bacteria in a sample are determined by assigning reads to known 16S database sequences via sequence homology such as RDP classification (14), or by clustering reads with similarity >97–99% (15). Both database assignment and clustering are aimed at reducing the effect of false-positive sequences due to sample preparation (PCR-based) errors and MPS read errors. We refer to profiling based on MPS of one or two variable regions followed by read mapping or clustering as the 'Single Region Framework' (SRF).

The large number of MPS reads enables an accurate estimate of species abundance and detection of bacteria present at low frequencies, compared with classical Sanger sequencing. However, the reliance on one or two short segments of the 16S limits the phylogenetic resolution (i.e. the ability to distinguish between closely related bacteria), since many different bacteria may share an identical sequence over these short segments. For example, the nonpathogenic *Escherichia fergusonii* str. ATCC 35469 and the uropathogenic *Escherichia coli* str. UTI89 share an almost identical 16S sequence, with <0.5% base pair difference, a change too subtle to be reliably detected by SRF.

Distinguishing between highly similar 16S sequences therefore requires analyzing a large 16S region. Figure 1B demonstrates how the reliance on a partial sequence of the 16S reduces the number of uniquely identifiable sequences, thus leading to underestimation of the true bacterial diversity. This concern is even more important given the fact that the 16S gene itself is highly conserved (16), and may underestimate the total genomic and phylogenetic differences between bacteria (17).

Several methods have been recently suggested for addressing the limitation of phylogenetic resolution (18–20). Fan *et al*. showed that stringent *de novo* assembly of 16S derived reads from Roche 454 metagenomic sequencing overcomes PCR amplification bias and detects bacteria up to the genus level while maintaining a low level of chimera formation. Alternatively, Miller *et al*. suggested EMIRGE, an iterative read mapping approach for short paired-end Illumina reads. By probabilistic mapping of short reads to a candidate set of 16S sequences and applying an expectation maximization algorithm, EMIRGE has been shown to identify sequences present in mixtures containing tens of species with genus level resolution. However, both methods are currently limited to analyzing a contiguous region of DNA. In (20), both EMIRGE and an alternative approach, modQIIME, were applied to analyze a 16S region of length 700–1000 bp. The modQIIME approach uses the taxonomical assignment of each paired-end short fragment read to identify bacteria present in a mixture up to the species level. Recall and precision values of EMIRGE and modQIIME varied both across tested data sets and also depending on whether genus or sequence level profiling was required. An alternative approach for integrating whole genome metagenomic reads is applied by PhylOTU (21) that maps reads into a phylogenetic tree and estimates pairwise distances between OTUs. While this approach enables downstream sample analysis such as unifrac (22), it does not provide direct classification of bacteria in the sample.

In this article, we present Convex Optimization for Microbial Profiling by Aggregating Short Sequence reads (COMPASS)—a method for integrating MPS short reads from a large genomic region or multiple regions, enabling high-resolution microbial profiling (see Figure 1C). Unlike *de novo* assembly methods, we rely on a database of bacterial sequences and assume that sequences of most bacteria in the mixture are represented in this database. Our general formulation of microbial profiling as a linear optimization problem (Figure 2) enables us to expand the analyzed region without the need of longer reads. Each of the millions of reads measured by MPS, together with all 'absent' reads that were not found, set constraints that enable 'zooming in' on the correct species present in the mixture. Theoretical analysis of COMPASS provides sufficient conditions on performance in the limit of infinite number of reads, and bounds on reconstruction error in case of finite reads (23). By design, COMPASS is completely agnostic to chimeric reads, and is robust to read errors. Analyzing a long region reduces the number of false positive sequences since differences among sequences become more significant, hence the probability that read errors would cause false-positive detection is decreased. In addition, while this article presents an application of COMPASS to the 16S case, it can be seamlessly extended to simultaneous sequencing of multiple and not necessarily overlapping regions of a number of genomic regions, provided a rich enough database of known sequences is available.

We study COMPASS's performance using extensive simulations, where we create *in silico* mixtures, simulate sequence reads and compare COMPASS's profiling to the 'true' mixture. We test the performance of COMPASS under different conditions, varying parameters such as the mixture diversity (i.e. the number and frequencies of species), number of reads and read length. We show accurate performance of COMPASS under realistic scenarios with resolution far exceeding SRF, simply due to the

Figure 1. Continued
that increasing region length reduces group sizes and enhances resolution. However, even at 1000 bp, only 80% of sequences are unique, while ~20% of the sequences cannot be uniquely resolved. (**C**) Schematic description of two frameworks for 16S sequencing. In the SRF, Roche 454 and Illumina paired-end sequencing are applied to a single variable region of length of ~400 bp. Bacterial profiling is based on counting the unique reads aligned to the specific database sequences. COMPASS integrates short reads originating from a larger amplicon (1200 bp of the 16S in this example). Profiling is performed by mapping to a high-dimensional linear regression problem.

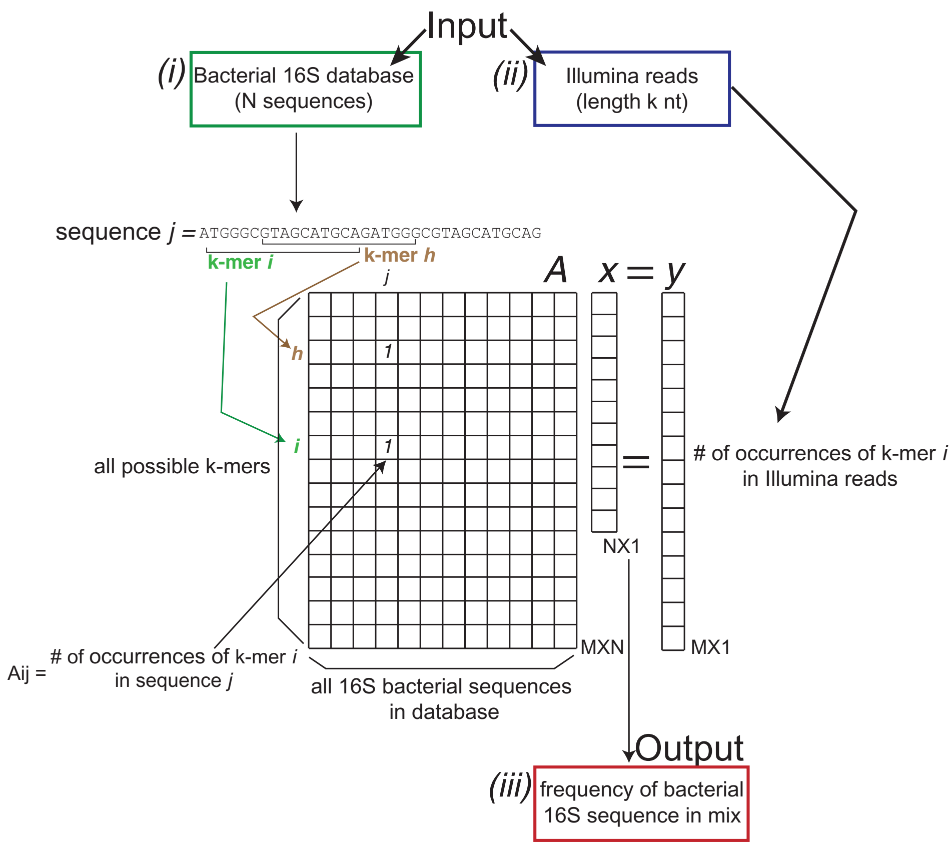

Figure 2. Mathematical formulation of COMPASS. Illustration of COMPASS implementation for the 16S gene. COMPASS receives two inputs: (i) a database of 16S sequences of the relevant amplified region, and (ii) MPS short reads of length k. (iii) COMPASS's output is the identity and frequency of the sample's bacteria. Reconstruction is obtained by solving the set of linear equations Ax = y. Each column in the matrix A corresponds to one of the N sequences in the database. Each one of the M rows of A corresponds to a specific k-mer appearing in the database, where k is the MPS read length. The matrix elements Aij are equal to the number occurrences of k-mer i in sequence j. The reads' vector y counts the number of reads corresponding to each k-mer. The solution vector x (once normalized such that its sum equals 1) represents the frequency of each of the N database bacteria; since M > N, x is obtained as the optimal solution of a regression problem.

extended sequenced region. A comparison with EMIRGE is also provided, displaying the significant advantages of COMPASS.

Next we describe an application of COMPASS to two biological systems: the *Drosophila* gut and eggs, and human saliva. These different biological habitats exhibit different bacterial communities, with different species and complexities. We applied COMPASS with Illumina single-end sequencing of a 750 bp amplicon from the 16S gene. For comparison, the same samples were also sequenced by SRF via Roche 454 pyrosequencing, displaying two advantages of COMPASS over SRF. First, while high concordance is shown between COMPASS and SRF at the latter's finest phylogenetic resolution, COMPASS further increases resolution, allowing accurate higher resolution detection. Second, the number of falsely detected bacteria was greatly reduced compared with SRF, displaying COMPASS's robustness to experimental errors.

Together, our simulations and experimental results suggest that COMPASS may serve as an important tool in analyzing and designing MPS experiments for microbial profiling, and enable more accurate assessment of bacterial diversity and population composition.

MATERIALS AND METHODS

COMPASS—a high-resolution microbial profiling framework

We present COMPASS in its general form and describe its implementation to the 16S case. The COMPASS approach comprises three steps illustrated in Figure 2:

(i) Preparation of a database of sequences from predefined regions.
(ii) DNA Extraction, PCR amplification of the predefined regions, fractionation and sequencing using MPS.
(iii) Computational reconstruction of the identity and frequency of bacteria in the sample.

(i) Preparation of a database of sequences from predefined regions

We use available sequence databases (e.g. RDP, Greengenes, etc.) to compile an ad hoc database of known sequences from regions to be sequenced. These may be multiple (consecutive or nonconsecutive) regions of the 16S or multiple regions from several genes. Since COMPASS can only detect bacterial sequences present in this ad hoc database, the leading principle in choosing the regions is the wealth of the databases and high sequence diversity at these regions. More specifically, since step **(ii)** involves PCR amplification of the predefined regions it should maximize the number of unique sequences over the amplicon. For this reason, the current application of COMPASS to the 16S used highly universal primers predicted to amplify ~330 000 sequences from the original 16S database used (72% of sequences in the original database, described in a following section), corresponding to ~231 000 unique sequences of ~750 bp (covering four variable regions V3–V6, see Figure 1A). In comparison, primer pairs that amplify the V3 region (24) would amplify ~380 000 sequences from the same database, but would result in only 95 000 unique sequences. Also, on the other length extreme, although a single primer pair designed to amplify the whole 16S gene would generate a longer amplicon, it may suffer from the limitation of potentially amplifying a smaller fraction of bacteria (e.g. the primers E8-E1492 (25) amplify <10% of the sequences in the database), and hence, we preferred the >750 bp long region.

(ii) DNA Extraction, PCR amplification of predefined regions and sequencing using MPS

This step follows standard protocols and yields a set of reads that originate from the amplified regions. In the current application of COMPASS, we used Illumina single-end reads of length 100 nt to sequence a 750 bp region. Specific issues regarding amplification, fractionation and Illumina sequencing appear in the experimental procedures section below.

(iii) Computational reconstruction of the identity and frequency of bacteria in the sample

Reconstruction is based on finding 'entries' in the ad hoc database compiled in step **(i)** that best 'explain' sequencing output in step **(ii)**. Since reads are much shorter than the amplified region and contain errors, they often align to multiple sequences in the database. However, each read does provide evidence in support of the existence of the 'correct' bacteria in a probabilistic way. The computational reconstruction step (Figure 2) integrates the statistical evidence from all reads to infer the frequency of each sequence in the database.

For clarity purposes, we first present the naive mathematical formulation ignoring practical issues such as sequencing read errors and computational implementation, and address them later.

Definitions

We use a sequence database S of N bacterial sequences S_j, $j = 1,\dots,N$ [see step **(i)**] and a set of R MPS reads of length k [see step **(ii)**]. Our goal is to use S and the R reads to compute a solution vector $\boldsymbol{x} = (x_1,\dots,x_N)$, where x_j represents the frequency of the j-th database sequence in the mixture.

The matrix A

We create a list of all unique k-mers (i.e. sequences of length k) that appear in S, and record the sequences to which each k-mer belongs. This information can be represented in a matrix $\boldsymbol{A}$, where each row corresponds to a specific k-mer and each column corresponds to one of the N database sequences (see Figure 2). The entry $\boldsymbol{A}_{Ij}$ represents number of occurrences of the i-th k-mer, in sequence j. The number of rows in $\boldsymbol{A}$, denoted M, is bounded by NxL where L is the typical length of the sequences in the database (L need not be equal for all bacteria).

The measurement vector y

Assuming the (unknown) bacteria in the mixture appear in the database and that no read errors occur, each of the R reads correspond to one of the rows in $\boldsymbol{A}$. We, therefore, denote sequencing results by the reads vector $\boldsymbol{y}$, where y_i corresponds to the number of reads of k-mer i. The vector $\boldsymbol{y}$ is sparse, namely most of its entries are zero, yet such zero entries may be informative. Assuming high enough coverage, the absence of reads 'characterizing' specific bacteria means that these bacteria are less likely to be part of the mixture.

Reconstruction

To find $\boldsymbol{x}$, we aim to solve the set of linear equations $\boldsymbol{Ax} = \boldsymbol{y}$ under the constraint that the entries of $\boldsymbol{x}$ are nonnegative, and later normalize $\boldsymbol{x}$ to have sum to 1. Since this is an overdetermined system, namely $M>>N$, we typically cannot satisfy this equation precisely and instead we solve the following regression problem:

$$\min x' \parallel \boldsymbol{A}\boldsymbol{x}' - \boldsymbol{y} \parallel_2 \text{ such that } \boldsymbol{x}'_i \geq 0, i = 1,\dots,N \quad (1)$$

where $\parallel \;\; \parallel_2$ corresponds to the Euclidean norm. This formulation implicitly assumes that reads are sampled uniformly from the relevant amplicon. In all natural scenarios, only a small minority of the N sequences have nonzero frequency, and hence $\boldsymbol{x}$ is a sparse vector.

A short overview of two theoretical results of COMPASS appear in Supplementary Methods S1, and described in detail in (23).

When dealing with a large database S, the matrix $\boldsymbol{A}$ becomes too large to compute and one cannot solve this regression problem in a straightforward manner. For example, if $N \sim 10^6$ and one sequences the whole 16S gene of length ~1500 bp, the number of rows in $\boldsymbol{A}$ is $\sim 10^9$. Another difficulty arises from read errors in which case reads might not be mapped to any row in $\boldsymbol{A}$ and are thus 'lost'. We address these issues in the following two sections.

Divide-and-conquer algorithm allowing scalability

To overcome the computational problems following the huge size of $\boldsymbol{A}$, we developed a divide-and-conquer algorithm. Briefly, we randomly split the database S into blocks of 1000 sequences and solve an optimization problem of reduced size [Equation (1)] in each block separately using CVX, a package for specifying and solving convex programs (26). We then collect all sequences whose frequencies were found to be above a threshold of 0.1% in all blocks (hence promoting a sparse solution); these form a reduced problem S', for which we repeat the procedure. These iterations, which keep reducing the database size, are performed until the number of sequences in the reduced database falls below the block size. We then solve the optimization problem again and normalize the solution to obtain the final sequence frequencies.

Since this divide-and-conquer approach provides an approximation of the correct solution, low-frequency sequences in the mixture may not be selected for the next iteration due to the random partitioning. To increase robustness with respect to detection of such low-frequency bacteria, random splitting is repeated 10 times when size of the database S' is <150 000 sequences. Sequences whose average frequency over these 10 iterations is >0.1% are kept.

The threshold of 0.1% sets a bound on the lowest frequency that is still considered 'interesting'. In case lower frequencies are of interest, this threshold can be reduced together with increasing the block size. For example, in the 'Results' section, we describe a case where the threshold was set to 0.025% and the block size to 4000, so as to collect low abundant bacteria. Setting a lower threshold and increasing the block size increases run time as discussed in the Section Time and memory requirements

Incorporating read errors using COMPASS

Reads may contain errors of several types, depending on sequencing technology. Such errors may hamper profiling of COMPASS in two ways. First, reads that 'originate' from one bacterium and are subject to errors may seem like 'legitimate' k-mers of a different bacterium. Second, in the more probable scenario, errors cause reads not to be mapped to any row in $\boldsymbol{A}$ at all, resulting in loss of information. The former case cannot be recovered since such reads appear errorless. However, since the total number of reads is large, read errors that happen to create 'legitimate' k-mers are negligible. In the more common scenario, some reads simply do not match any k-mer and we need to devise a way to make use of them (note that chimeric reads are always automatically discarded as they simply never match any row in $\boldsymbol{A}$).

Assuming the sequencing device error profile is known, it is straightforward to integrate read errors into the set of linear equations (1), by adding rows to $\boldsymbol{A}$ corresponding to possible additional k-mers. This, however, greatly increases (the already huge) size of $\boldsymbol{A}$, hence further increasing COMPASS's computational complexity. We therefore applied the following approximation: whenever a read is not mapped to any k-mer we map it to its nearest possible rows in $\boldsymbol{A}$, namely all k-mers with Hamming distance 1, and modify the vector $\boldsymbol{y}$ accordingly. This approximation effectively uses the information obtained from reads with up to a single base substitution error, which would otherwise be lost. The effect of applying this procedure is shown in Supplementary Figure S8. Given the limited effect of read errors, correction for unmapped reads may be ignored, thus greatly decreasing running times, without significant loss in performance.

Measuring performance using *in silico* simulations

Reconstruction quality depends both on correct sequence prediction and on an accurate frequency estimate. Several measures of similarity between microbial profiles based on species similarity (22) or combined species similarity and frequency (27,28) have been previously suggested. However, we opted for an alternative measure that provides a more intuitive score while combing sequence and frequency similarities. We defined 'weighted recall' and 'weighted precision', which quantify false-positive and false-negative predictions, respectively. Intuitively, 'weighted recall' corresponds to the probability of correctly reconstructing a randomly selected sequence from the original mixture, with respect to both sequence and frequency. Similarly 'weighted precision' corresponds to the probability that a randomly selected sequence from the reconstructed set is correct in its sequence and frequency.

The first step in calculating weighted recall and precision is to go over all reconstructed bacteria and find their most similar bacteria, in terms of Hamming distance, in the list of (known) mixture bacteria. We then mark each such pair of bacteria as 'valid' if two conditions hold: (i) both sequences should match completely; (ii) the relative difference between the estimated frequency and the true simulated frequency should be <20%, or the absolute value of their difference should be <0.002. Based on these definitions, weighted precision (recall) is calculated by summing the frequencies of reconstructed (simulated) bacteria that are 'valid'. Since 'valid' reconstruction is declared only for zero mismatches between the correct and predicted bacteria, we refer to this case as the 'MM 0%' case. We also calculated weighted recall and weighted precision when up to 2% mismatches are allowed, and refer to this case as 'MM 2%'.

Comparison with other methods using *in silico* simulations

COMPASS was compared with single region sequencing of the 16S variable region V4 and to EMIRGE (18) using *in silico* simulations (see Supplementary Methods S2 for a description).

16S-V4 comparison

In silico PCR was applied to select all sequences from the Greengenes database (see details in Section Database and *in silico* PCR) that are amplified by the 16S-V4 universal primers F515 and R806 (29), resulting in 133,173 unique sequences holding 352,983 amplified sequences. The average region length was 292 bp, and we truncated all sequences to a length of 280 bp (deleting the 98 sequences whose length was shorter). The 16S-V4 method was compared with COMPASS in our toy mixture example and in our wide scale *in silico* simulations. In the toy mixture example, we provided the phylogenetic resolution given by the 16S-V4, namely the number of 16S sequences that share the same 280 bp of the 16S-V4. In our *in silico* simulations, we estimated weighted precision and recall of 16S-V4 results in the following way. For each bacterium in the simulated mixture, we located the group of bacteria that share the same sequence over V4 (the group can be of size 1 in cases the V4 region uniquely identifies the original bacterium). We then chose the 'reconstructed' sequence by randomly selecting a bacterium from each group of each of the 'true' bacteria. We calculated weighted precision and recall based on the sequence similarity criterion [criterion (i)] described above for COMPASS, while ignoring the frequency criterion [criterion (ii)]. Therefore, reconstruction results correspond to an infinite number of errorless reads of length 280 bp.

EMIRGE comparison

We used the EMIRGE amplicon software package for single-end reads (emirge_amplicon.py, emirge_rename_fasta.py) following the guidelines specified at the EMIRGE Web site (https://github.com/csmiller/EMIRGE). The same bacterial mixtures tested for COMPASS were used to create fasta files with reads that were supplied to EMIRGE. The only difference from COMPASS regarding the input was that we did not include read errors in the case of EMIRGE. The EMIRGE output, namely the list of reconstructed bacteria and their frequencies, were then used to calculate weighted precision and recall in the same way as for COMPASS. In addition, we calculated weighted precision and recall while considering the reconstructed sequences [criterion (i)] and ignoring EMIRGE reconstructed frequencies [criterion (ii)].

Time and memory requirements

Profiling time depends on the number of blocks that need to be processed, which mostly depends on the database size. Since COMPASS is completely parallelizable, we get a linear speedup in the number of cores, e.g. running using 10 cores would take half the time as running with 5 cores. Memory requirements per core are ~3.5 GB (either with or without applying read error correction). We estimated profiling times using a Linux machine with 6 cores having 37 GB of RAM in total. Computing time increased significantly when read error correction was applied. Detailed comparisons of COMPASS and EMIRGE with respect to time and memory usage appear in Supplementary Results SR.4.

COMPASS without correcting for read errors

A block of 1000 bacteria is solved in $\sim 6 \pm 1$ s, where ~4 s are needed to create the matrix $\boldsymbol{A}$ for these 1000 bacteria, and 2 s for solving the optimization problem over the block (numbers refer to a single core of the Linux machine mentioned above, running read length of 100 nt). The total profiling time was 1–2 h, using the above 6 cores Linux machine and the Greengenes database. Time was independent of the number of reads, it slightly increased with increasing number of bacteria and when decreasing the read length (the only case where profiling took 3.5 h was for read length of 35 nt).

COMPASS with correcting for read errors

Correcting for read errors significantly increase COMPASS run times. Our default conditions (a mixture of 200 bacteria sequenced by 10^6 reads of length 100 nt) takes ~11 h to profile. Increasing read length to 200 nt or increasing the number of bacteria to 1000 increased running time to 24 and 35 h, respectively. Given the limited effect of read errors, the correction procedure for unmapped reads may be ignored, without significant loss in performance (see Supplementary Figure S8). This would dramatically reduce running times. However, in case sequencing results contain a large number of read errors and correction is required, one can simply use a larger number of CPUs to reduce run time.

COMPASS—effect of block size

Detecting low abundance bacteria requires setting a lower COMPASS threshold while increasing the block size. When using a block of 4000 bacteria, it takes ~17 s to build the matrix, while solving the optimization problem takes ~3 s (memory did not vary significantly). Hence, in case detection of lower abundance bacteria is needed, it may be beneficial to apply faster k-mer counting algorithms such as JELLYFISH (30).

Database and *in silico* PCR

The 16S sequences were downloaded from Greengenes (current_prokMSA_unaligned.fasta, version dated 2010), containing 455 055 unique 16S sequences. We aimed at selecting primers that would maximize the number of unique sequences over the amplicon. *In silico* PCR was used to search for sequences that could possibly be amplified by a given pair of primers, by identifying perfect match of the primers within database sequences and calculating the product's length. Several combinations of known primers [e.g. (24)] were tested and selected primers' sequences and amplicon length histogram appear in Supplementary Figure S9. The reduced number of unique sequences with respect to the full 16S database results from the fact that not all sequences have the primers' binding region and, more importantly, that many potentially amplified sequences share the same sequence over the specific amplicon (i.e. phylogenetic resolution is limited). The primers selected for Illumina sequencing produce an amplicon of ~750 bp (covering variable regions V3–V6) and correspond to 231 299 unique sequences. The primers used for the Roche 454 sequencing produce an amplicon of ~450 bp (covering variable regions V3–V4) and correspond to 176 674 unique sequences. We used the same forward primer for Illumina and Roche 454 to allow further comparison between SRF and COMPASS. We refer to the Illumina case as the '750 database', while the Roche 454 database is termed '350 database' since analysis of the Roche 454 data used only the first 350 bp to decrease the effect of read errors.

Experimental procedures

Drosophila samples

Larvae (L1 and L2). h-Gal4 males flies were crossed to UAS-*neoGFP* females for 3 days [described in (31)], the progenies of this cross (*h-Gal4*;UAS-*neoGFP* flies) were developed in standard 25 × 95 mm *Drosophila* vials (cat# 51–0500, Biologix) containing 15 ml of fly food (http://flystocks.bio.indiana.edu/Fly_Work/media-recipes/).

Sample L1 was supplemented with G418 (400 µl/ml, GIBCO). Ten guts of third instar larvae were dissected from each sample and pooled together. DNA was extracted using a chemagic DNA bacteria Kit (Chemagen).

Eggs (E1 and E2). About 300 *h-Gal4*;UAS-*neoGFP* adult flies that were developed in a food with or without G418 (E1 and E2, respectively), and were allowed to lay eggs for 2 h on agar plates. These 0–2 h-old embryos were collected and DNA was extracted from ~200 eggs of each sample as above.

Human saliva samples
1 ml of saliva was collected from one female and one male adult (34 years) in two consecutive days. DNA was extracted as above.

Single-region Roche 454 sequencing

Sequencing
Samples were amplified by PCR (E339–789), all amplicon products from different samples were mixed in equal concentrations, purified and then sequenced by a Roche titanium 454 machine. The Illumina amplicon, of length 750 nt, includes the Roche 454 amplicon as its first 450 nt. Preprocessing the reads included truncating reads to a length of 350 nt and keeping only those for which the Phred quality score was >25 in at least 80% of their bases. After preprocessing, the number of reads per sample ranged from $\sim 10^4$ to 3×10^4.

Analysis
Analysis was performed by two methods. First, reads were mapped to the '350 database' using BLAST, marking reads for which similarity was >98%. Bacterial abundance was estimated by the fraction of reads that were mapped to each sequence in the database, and thus we refer to this method as '454-BLAST'. In the second analysis method, we applied MG-RAST (32). Apart from strain-level classification, which is too coarse and does not allow for direct comparison with COMPASS, MG-RAST provides for each read the best hit in the database according on its internal pipeline (MG-RAST also enables selecting Greengenes as its baseline database). Hence, bacterial abundance was estimated by the fraction of reads that were mapped to the best hit. There were cases in which MG-RAST similarity search provided multiple same-scoring hits to the same read. However, in all of these cases, these multiple classifications still corresponded to the same entry in the '350 database', and hence, no classification ambiguities occurred.

COMPASS using Illumina sequencing

Sequencing
DNA was PCR amplified using 16S rRNA universal primers whose product was ~750 bp long for most bacteria in the database (see Supplementary Figure S9). PCR was performed in 96-well plate such that each sample had eight reactions that were mixed together immediately following PCR. Samples were then cleaned on column (Promega, Fitchburg, WI) and concentration was measured using NanoDrop (NanoDrop Technologies, Wilmington, DE). All samples were diluted to 50 ng/µl and a volume of 100 µl. Samples were then sonicated (Bioruptor, Diagenode, Philadelphia, PA) between 80 and 100 cycles (30/30 s on/off), resulting in length distribution in the range of 100–300 bp (see Supplementary Figure S10). Subsequently, samples went through standard Illumina library preparation including barcode ligation, and were sequenced on a single lane of an Illumina HiSeq 2000 sequencer using 100 nt reads.

Analysis
Reads for which the Phred quality score was >30 in at least 80% of their bases were kept, resulting in 2×10^6–2×10^7 reads per sample. Since coverage along the amplicon was not uniform, we first normalized the number reads before applying COMPASS (see normalization issues below).

Data normalization and a COMPASS Algorithm modification for Illumina reads

According to our experimental protocol, Illumina data showed several biases, including unequal nucleotide coverage along the sequence. Different coverage is observed at different positions along the sequence due to nonuniform fractionation of the 16S DNA and systematic Illumina biases in amplifying different DNA regions, even when sequencing a single bacterium. For example, for reads originating from both the forward and reverse strands in *Drosophila* sample L1, Supplementary Figure S11A shows high coverage variability as a function of position along the 16S sequence (the mean per-nucleotide coverage varies ~1.8-fold across the sequenced region). We refer to this as 'global' coverage-variability pattern. In addition, high 'local' variability is observed, namely coverage varies significantly even between adjacent locations. This contradicts the COMPASS implicit assumption that the number of reads of each k-mer of a specific bacterium linearly depend on its frequency. Violation of this assumption results in increased noise and potential degradation in performance. To deal with unequal coverage, we performed both preprocessing and postprocessing steps described below.

Preprocessing reads
We performed two normalization steps addressing local and global variability patterns on the raw Illumina reads before applying the COMPASS algorithm.

1. Illumina reads preprocessing overcoming local coverage bias
We split each read to shorter overlapping reads. More specifically, a read of length 100 nt was converted into 11 contiguous reads of length 90, which were mapped to 11 starting points, hence averaging out coverage variability on the single nucleotide level. This normalization step comes at the expense of slightly shorter reads, but this small decrease is predicted to hardly impact COMPASS performance (see Figure 3 panel B). Supplementary Figure S11B presents the read coverage following local noise normalization in sample Supplementary Figure S1.

2. Illumina reads preprocessing overcoming global coverage bias
Following averaging out of the local bias, long-range position-dependant variability in read coverage is still

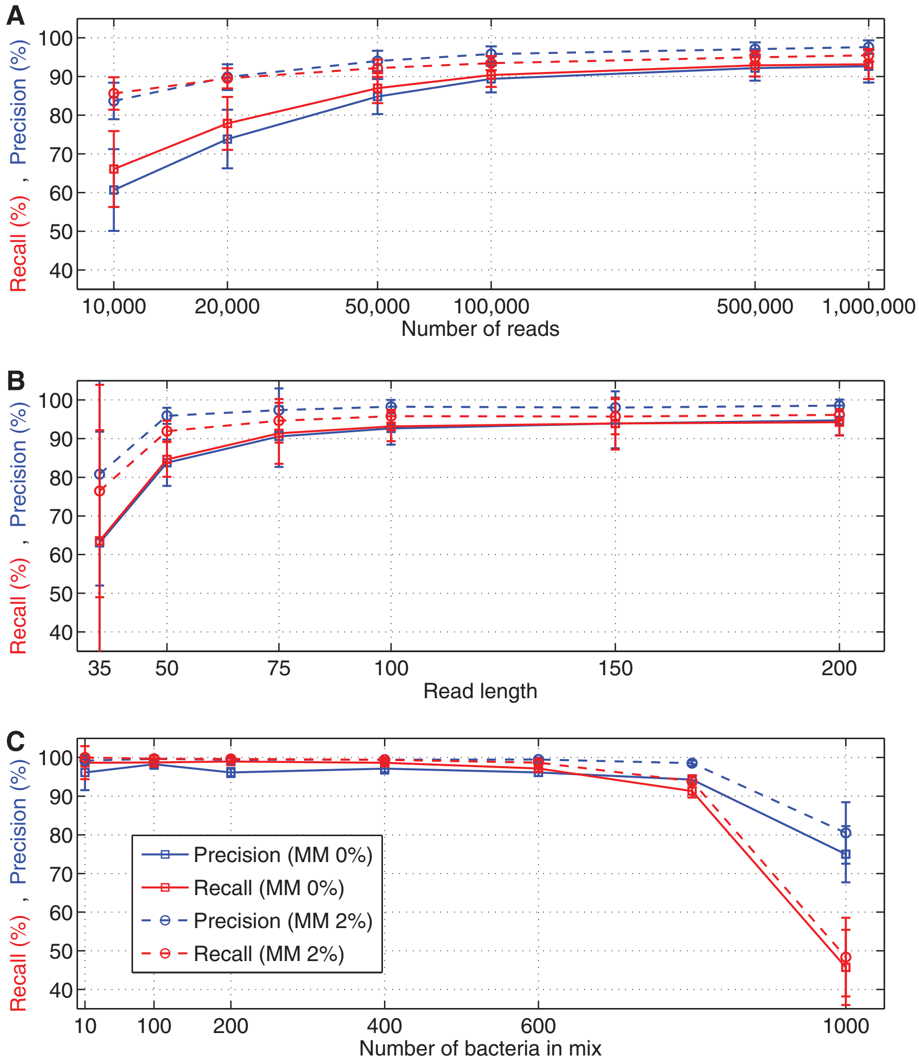

Figure 3. Extensive simulations—performance as a function of several parameters Panels (**A–C**) present weighted recall and precision as a function of different variables. Unless stated otherwise, parameters are as follows: each simulation contained 200 bacteria, randomly selected from the (full length) 16S database, with relative frequencies following a power law distribution (1/x). The read length was 100 nt, and 106 reads were simulated, and were subject to read errors. The blue and red lines denote weighted recall and precision, respectively; a solid line refers to the case in which complete sequence identity is required, while a dashed line refers to the case where up to 2% differences in sequence are acceptable. Frequencies can differ by <20% or the absolute value of their difference should be <0.002. Data present average and one standard deviation over 100 simulations of each scenario. (A) Effect of number of reads, with read number changed from 106 down to 104, and other parameters as above. (**B**) Effect of read length, with read length varying between 35 and 200, and other parameters as above. (C) Effect of number of bacteria in the original mixture. Here for each number of bacteria n, all bacterial frequencies are set to 1/n. Other parameters are as above.

apparent (Supplementary Figure S11B). To reduce this effect, the global coverage profile was independently calculated for forward and reverse reads by taking the mean coverage over a sliding window of length 20 nt. Each read was mapped to the vector **y**, although its original weight was divided by the value in Supplementary Figure S11B at the read's average location. The resulting coverage appears in Supplementary Figure S11C.

Postprocessing—modifying the COMPASS optimization problem

Although preprocessing greatly reduced the position-dependent coverage variability, coverage is still nonuniform as evident from the outliers in Supplementary Figure S11C. To further reduce their effect, we modified the COMPASS algorithm. The original COMPASS algorithm minimizes the L_2 norm of the solution. While being computationally feasible, the L_2 norm is sensitive to unequal coverage, thus may introduce false positives. In case of low-coverage, L_2 norm tends to 'divide' the reads among several low-frequency sequences (i.e. database entries), rather than keeping a single high-frequency sequence. To reduce this effect, we performed a postprocessing step using the list of COMPASS-inferred sequences. The sequences found in the last COMPASS iteration were solved again while changing the optimization criterion to minimizing the L_1 norm:

$$\min_{x'} \| \tilde{\boldsymbol{A}}\boldsymbol{x}' - \boldsymbol{y} \|_1 \text{ such that } \boldsymbol{x}'_i \geq 0, i = 1, \ldots, N$$

where $\tilde{A}$ is the matrix based on the COMPASS-inferred sequences. Since the number of such sequences is small, the reads' average location in Supplementary Figure S11B can be directly incorporated into the matrix rather than normalizing the read vector **y**. This L_1 minimization postprocessing step prunes the results and reduces the number of false positives in the list of inferred sequences.

Sanger sequencing and validation

Primers used

Using *in silico* PCR we designed *Wolbachia*- and *Acetobacter*-specific primers. The *Wolbachia* primers (forward: TGATCAGCCACACTGGAACT, reverse: A AGTCCCCAGCATTACCTGA) potentially amplify 98 bacteria having a product size of ~800 bp. Three 454-BLAST predicted bacterial groups (*EU137480.1*, *EU137491.1* and *EU137473.1* in Figure 4) were not amplified by these Sanger primers. As for *Acetobacter*, the primers (forward: GAGCTGCATTTGATACGTGC, reverse: CACTGTCACCGCCATTGTAG) amplify 60 bacteria having a product of ~600 bp.

Analyzing the Sanger chromatogram

We used an automatic peak detection algorithm to locate and evaluate peaks in the chromatogram. A minor peak was declared when its value was >3% of the corresponding major peak (see Supplementary Figure S6 for a cartoon). Since the primers were designed to amplify highly similar bacteria, the number of multi-peak

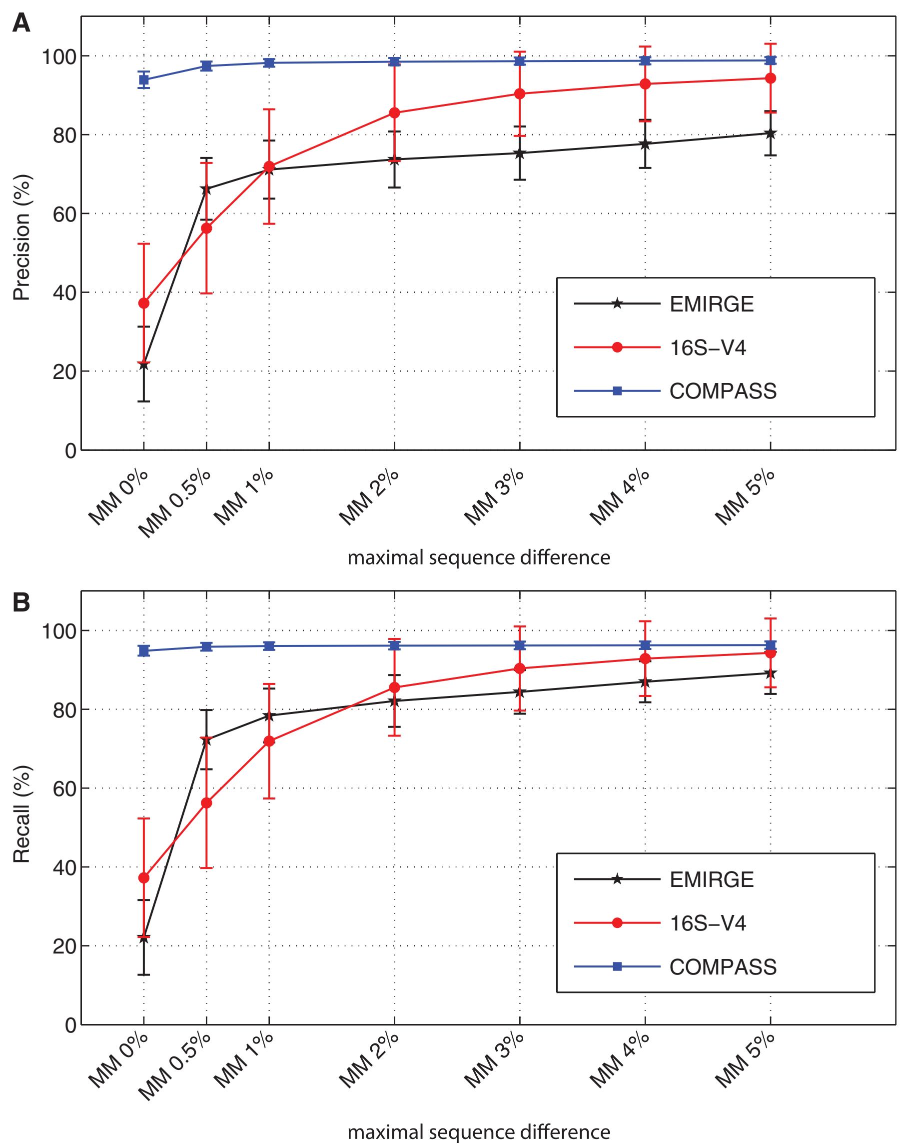

Figure 4. Comparison between algorithms. Comparison between COMPASS, EMIRGE and 16S-V4 sequencing. Weighted precision and recall (panels A and B, respectively) as a function of the maximal allowed sequence mismatch. Simulated mixtures correspond to the default setting. Black, red and blue lines display the mean and standard deviation of 100 simulations using EMIRGE, 16S-V4 and COMPASS, respectively, performed over the same mixtures. Weighted precision and recall for COMPASS also enforce the frequency criterion, as opposed to the case of EMIRGE and 16S-V4. The EMIRGE simulations did not contain read errors. The 16-V4 simulations were performed over an infinite number of errorless reads.

locations is limited, and in most cases only a single minor peak was detected. The length of the Sanger sequence and its starting point were manually set.

Mismatch detection

We created a list of all bacterial sequences that are potentially amplified by the corresponding primer pair, and locally aligned each of them to the Sanger sequence. A 'Sanger-validated' sequence was declared whenever the bacterial sequence matched the nucleotide corresponding to either major or minor peaks along the chromatogram. We counted three numbers for each bacterial sequence *s*—the total number of mismatches between *s* and the Sanger sequence, and the number of mismatches over the intersection between the Sanger sequence and the Roche 454 or Illumina amplicons. Supplementary Figure S7A–D presents the mismatch locations for all potentially amplified bacteria in the *Wolbachia* an *Acetobacter* cases.

Data and software

An implementation of COMPASS is available at https://github.com/NoamShental/COMPASS. Roche 454 data for *Drosophila* and saliva experiments can be downloaded from the MG-RAST Web site at http://metagenomics.anl.gov/linkin.cgi?project = 4932. Profiling results files for 454-BLAST, MG-RAST and COMPASS appear as Supplementary Data sets S1–S5. In addition MG-RAST read classification files appear as Supplementary Data sets S6–S9. Illumina data for *Drosophila* and saliva experiments can be downloaded from the MG-RAST Web site http://metagenomics.anl.gov/linkin.cgi?project = 5237. Read data were subject to preprocessing overcoming local coverage bias (step 1).

RESULTS

Simulation of a toy mixture

To demonstrate the added value of COMPASS in providing improved phylogenetic resolution and accurate frequency estimates we first constructed a simulated *in silico* toy mixture. The mixture comprised 10 bacteria of the human gut model community bacteria used by Faith *et al.* (33), to which we added another bacterium whose 16S sequence is only 11 bp different from one of these bacteria. Frequencies assigned to each sequence were taken from the measurements of Walker *et al.* (34).

We simulated 10^6 short Illumina-like reads of length 100 nt (with read errors) covering the full length 16S gene and applied COMPASS. To demonstrate the improvement in phylogenetic resolution compared with SRF, we also provide the results based on the 16S variable region V4 (referred to as 16S-V4) using an infinite number of errorless reads. Table 1 presents the names of simulated bacteria and their frequencies together with COMPASS and 16S-V4 results.

COMPASS

Table 1 displays the names of COMPASS-reconstructed sequences and their frequencies, when the equality symbol corresponds to a correct reconstruction, i.e. zero

Table 1. Simulated toy mixture results

Toy mixture		COMPASS		16S-V4		
Sequence name	Simulated frequency (%)	COMPASS reconstructed sequences	COMPASS frequency (%)	Number of sequences sharing the same 16S-V4	Median similarity over whole 16S	Minimal similarity over whole 16S
Eubacterium rectale ATCC 33656	14.17	=	12.73	1321	0.89	0.78
Collinsella aerofaciens ATCC 25986	11.71	=	11.53	356	0.89	0.74
Blautia hydrogenotrophica DSM 10507	11.58	=	11.15	5	0.89	0.88
Desulfovibrio piger GOR1	10.21	=	9.84	6	0.88	0.88
Clostridium symbiosum ATCC 14940	9.44	=	9.63	1	1	1
Escherichia coli str. K-12 substr. MG1655	8.14	=	7.21	1539	0.88	0.71
Marvinbryantia formatexigens DSM 14469	7.88	=	8.14	2	0.95	0.95
Bacteroides ovatus ATCC 8483	7.88	=	7.6	100	0.88	0.86
Bacteroides thetaiotaomicron VPI-5482	6.73	=	6.45	369	0.88	0.8
Bacteroides caccae ATCC 43185	6.35	=	5.73	488	0.89	0.82
Bacteroides sp. str. D1	5.9	=	5.51	100	0.89	0.87
		Human fecal clone JTU_G_10_09	0.1			
		Metagenomic gut microbiome healthy human stool clone EB35	0.13			
		Pervasive effects antibiotic on human gut microbiota deep sequencing fecal clone B3_148	0.11			
		Human fecal clone SJTU_G_07_07	0.11			
		Human fecal clone 014B-H5	0.12			

Toy mixture sequences and their frequency appear on the left panel. The middle panel corresponds to COMPASS reconstructed sequence names and their frequencies. The equality sign ('=') stands for an exact match, namely complete identity between the full 16S gene sequence of simulated and reconstructed bacteria. The right panel displays the number of Greengenes sequences that share the same V4 region for each simulated bacterium. We also provide the median and mean similarities between the simulated sequence and the other sequences that share its V4 region. The mixture's sequences appear in Supplementary Table S3 and in Data set 10.

mismatches between the full 16S gene sequences of the simulated and reconstructed bacteria. COMPASS successfully identified the 11 bacteria in the mixture as the top 11 most frequent sequences while providing highly accurate frequency estimates. This includes correct detection of the two highly similar bacteria *Bacteroides ovatus* ATCC 8483 and *Bacteroides* sp. str. D1 that differ by 11 bp along their 16S genes. In addition, COMPASS detected five false-positive bacteria whose total predicted frequency by COMPASS was <0.6%.

16S-V4

Table 1 presents the number of Greengenes sequences that share the same V4 sequence as the simulated sequence, and the median and minimal sequence similarities between their full 16S sequences and the simulated sequence. A larger number of such sequences and low similarity values correspond to lower phylogenetic resolution. For 10 out of 11 bacteria, 16S-V4 displays lower phylogenetic resolution than COMPASS. Namely, while COMPASS identified the exact full 16S gene sequence of each of these bacteria, the 16S-V4 can only provide a larger 'group' of equally likely sequences whose similarity to the correct sequence ranges between 0.71 and 0.95 even when using an infinite number of errorless reads. For example, *E. coli* str. K-12 sub strain MG1655 belongs to a 'group' of 1539 bacteria having an identical V4 region. The median (minimal) sequence similarity along the 16S gene between the correct sequence and the other 1538 sequences in the group is ~0.88 (0.71) displaying reduced phylogenetic resolution. Therefore, e.g. 16S-V4 can not differentiate between *E. coli* str. K-12 substrain MG1655 and *E. coli* NMU-ST2 although their similarity is only ~79%.

In addition, the short sequenced region results in incorrect frequency estimates. For example, 16S-V4 identifies *Bacteroides ovatus* ATCC 8483 and *Bacteroides* sp. str. D1, which share the same V4 region, as a single low-resolution 'bacterium' having frequency of ~13% instead of the correct 2 bacteria detected by COMPASS. Their simulated frequencies were ~7.9 and 5.9%, with COMPASS-predicted frequencies being close at 7.6 and 5.5%, respectively.

COMPASS shows high accuracy in various simulated experimental scenarios

We defined a default simulation setting and explored the performance of COMPASS when modifying each one of its parameters. The default setting included a mixture of 200 bacteria randomly selected from the (full length) 16S database, with relative frequencies conforming to a power law distribution ($1/\times$). The number of reads was set to 10^6 and the read length was chosen to be 100 nt. All reads were subject to Illumina-specific sequencing errors (see Supplementary Methods for the error model). Figure 3 displays the performance using the default settings while modifying the number of reads (panel A), the read length (panel B) or the number of bacteria in the mixture (panel C).

Accurate profiling amounts to correct identification of the bacteria in the mixture, together with their frequency. We aimed at providing a biologically relevant measure that would combine sequence and frequency predictions in an intuitive way. We defined 'weighted recall' and 'weighted precision', which quantify false-positive and false-negative predictions, respectively. Briefly, 'weighted recall' of 95% means that there is 0.95 probability that a bacterium randomly selected from the simulated mixture is identified by COMPASS with zero mismatches and also its inferred frequency is close enough to its simulated frequency (weighted precision quantifies false negatives in a similar way, see 'Materials and Methods' section). To allow for less stringent criterion, we also display performance in case that up to 2% (i.e. ~30 nt of the ~1500 bases of the 16S) mismatch in sequence is allowed and refer to this as the 'MM 2%' case, as opposed to the former 'MM 0%' case. Each point in the results presents the average and standard deviation based on simulating 100 realizations of bacterial mixtures. For example, the default setting displays a level of 93% in both weighted recall and precision for 'MM 0%', while for 'MM 2%' weighted recall is 95% and weighted precision is 98%.

Effect of varying the number of reads

Recall and precision >90% were achieved even when reducing the number of reads to 100 000 and 50 000 for MM 0% and the MM 2%, respectively. This demonstrates high fidelity reconstruction of the bacterial community with low numbers of reads, thus enabling analysis of many samples per lane via barcoding (i.e. multiplexing).

Effect of varying read length

Increasing read length beyond 100 nt or shortening it down to 50–75 nt had almost no effect on performance. Thus, current read lengths of single-end sequencing suffice for good reconstruction, and advantages of longer or pair-end reads or applying assembly methods to increase the effective read length are less significant. Read lengths <50 nt are less informative, as reads potentially appear in many different bacteria, resulting in lower quality profiling.

Effect of varying the number of bacteria in a mixture

To analyze the dependence of COMPASS's performance on the complexity of the bacterial mixture, we replaced the power law distribution by a fixed frequency of 1/n (same frequency for all species within the sample) and varied the number of bacteria, n. This allows bacterial frequencies to be higher than COMPASS's threshold of 0.1% for up to $n = 1000$ (whereas in a power law distribution, most frequencies will be lower than this threshold for $n > 200$). Reconstruction error rate was almost constant for up to $n = 800$ bacteria, but increased substantially for larger numbers due to the predefined threshold in COMPASS for the frequency of 'interesting' bacteria (0.1%, see 'Materials and Methods' section). The error at a level for 1000 bacteria increased mainly due to increase in weighted recall, while weighted precision retains a level of 80%. This indicates that even in this case, COMPASS still almost exclusively finds correct bacteria, although the number of missed bacteria grows. The minimal frequency threshold (0.1%) was chosen to enable faster running times, and can be changed in case detection of low frequencies is required.

To display detection of low abundant bacteria, we replaced the 0.1% threshold value by 0.025% and used blocks of 4000 bacteria instead of 1000 bacteria in the divide and conquer step (see Supplementary Results S1). Repeating the case of $n = 1000$ bacteria in panel C resulted in ~90 and 97% weighted recall and precision for MM 0 and MM 2%, respectively. Results appear in Supplementary Table S1. The table also presents a second scenario of simulating 500 bacteria having a power law distribution (1/x), in which case the frequencies of >350 bacteria fall below 0.1% and the minimal frequency is ~0.03%. Using the abovementioned parameters resulted in ~90 and 97% weighted recall and precision for MM 0% and MM 2%, respectively. Changing the parameters increased simulation run times of each divide and conquer block by a factor of ~4 (see 'Materials and Methods' section).

Comparison with other methods

We compared COMPASS with EMIRGE (18) and with SRF using 16S-V4. Weighted precision and recall were estimated for 16S-V4 and EMIRGE using the default simulation conditions (mixtures of 200 bacteria having relative frequencies conforming to a power law distribution 'sequenced' using 10^6 single-end reads of length 100 nt). Figure 4 presents weighted precision and recall of the three algorithms as a function of the required phylogenetic resolution, namely the maximal allowed difference between the 'correct' and estimated sequences (i.e. from MM 0% to MM 5%).

Weighted precision and recall in COMPASS were calculated as in the former section, i.e. the inferred sequences and their frequencies should be close enough to the 'correct' sequences and their frequencies. When calculating weighted precision and recall for EMIRGE and 16S-V4, we, however, did not enforce the frequencies criterion, thus improving the measured performance (in addition, we used an infinite number of errorless reads for 16S-V4, and no read errors in the case of EMIRGE). COMPASS's results display higher recall and precision than both EMIRGE and 16S-V4, especially when high accuracy is required.

This gain in performance is also apparent in other scenarios. We used EMIRGE to profile the same mixtures used in Figure 3 (although ignoring read errors). The EMIRGE results, with and without enforcing the frequencies criterion, appear in Supplementary Figures S1 and S2, respectively (A comparison with 16S-V4 is less relevant since we consider an infinite number of reads and assume that read length covers the whole region). The lower performance of EMIRGE, with respect to COMPASS probably corresponds to the fact that EMIRGE was optimized for other purposes, for example, using paired-end reads. Additionally, COMPASS and 16S-V4 have an inherent advantage over EMIRGE under the conditions mentioned above, since all simulated sequences were drawn from the Greengenes database later used by the algorithms. EMIRGE, which does not directly rely on a specific database, tends to find novel species and therefore performance deteriorates. However, EMIRGE will probably have a significant advantage in environments that contain many unknown bacteria.

Experimental profiling of Fly and Human samples: COMPASS versus SRF

We applied COMPASS to analyze the microbiota of two biological systems: The first included four samples of the larva and eggs of *Drosophila melanogaster* (referred as L1 and L2 for larva and E1 and E2 for eggs), and the second included four samples of human saliva of two individuals taken on two consecutive days (referred as H1 and H2 for person #1 and H3 and H4 for person #2).

Since ground truth is not available for these experimental mixtures, we compared COMPASS using Illumina sequencing with single-region Roche 454 Pyrosequencing over the same samples. Following DNA extraction, we PCR amplified and sequenced a 450 bp region by Roche 454 Pyrosequencing as an example of SRF. COMPASS was based on a 750 bp region that was PCR amplified and subjected to sonication before sequencing by an Illumina HiSeq2000 using 100 nt single-end reads. The 750 bp amplicon included the 450 bp region (aligned with the start of the sequence), thus supporting comparison of the two methods. Sequencing results were analyzed using COMPASS for Illumina sequencing, with several modifications applied to the basic COMPASS algorithm to overcome experimental biases as described in the 'Materials and Methods' section. For the SRF method, the Roche 454 reads were truncated to the first 350 nt to reduce the effect of read errors, and then analyzed by two methods. The first method mapped each to the database using BLAST, and thus we refer to its results as '454-BLAST'. In the second method, reads were analyzed by MG-RAST (32).

The fly and human samples display two different levels of microbial complexity: while *Drosophila* samples contained a rather low number of bacterial strains (~20 species with frequency >1%), human saliva samples were more diverse (~80 bacterial species with frequency >1%). These differences in complexity are consistent with previous works (12,35). Results of the *Drosophila* experiments presented in Figure 5 compare COMPASS and 454-BLAST. The analogues comparison between MG-RAST and COMPASS appears in Supplementary Figure S3 and displays similar results (see Supplementary Results S2). Results summary of the three methods appear in Table 2. Human saliva samples appear in the Supplementary Figure S5.

Concordance between COMPASS and SRF at the 454-BLAST finest resolution

To quantify the similarity between COMPASS and 454-BLAST, we display the inferred sequences by the two methods in Figure 5. Phylogenetic trees of the sequences found by each method are shown on the top and left parts of the figure, for 454-BLAST and COMPASS, respectively. Each tree leaf is termed a 'group' as it may represent

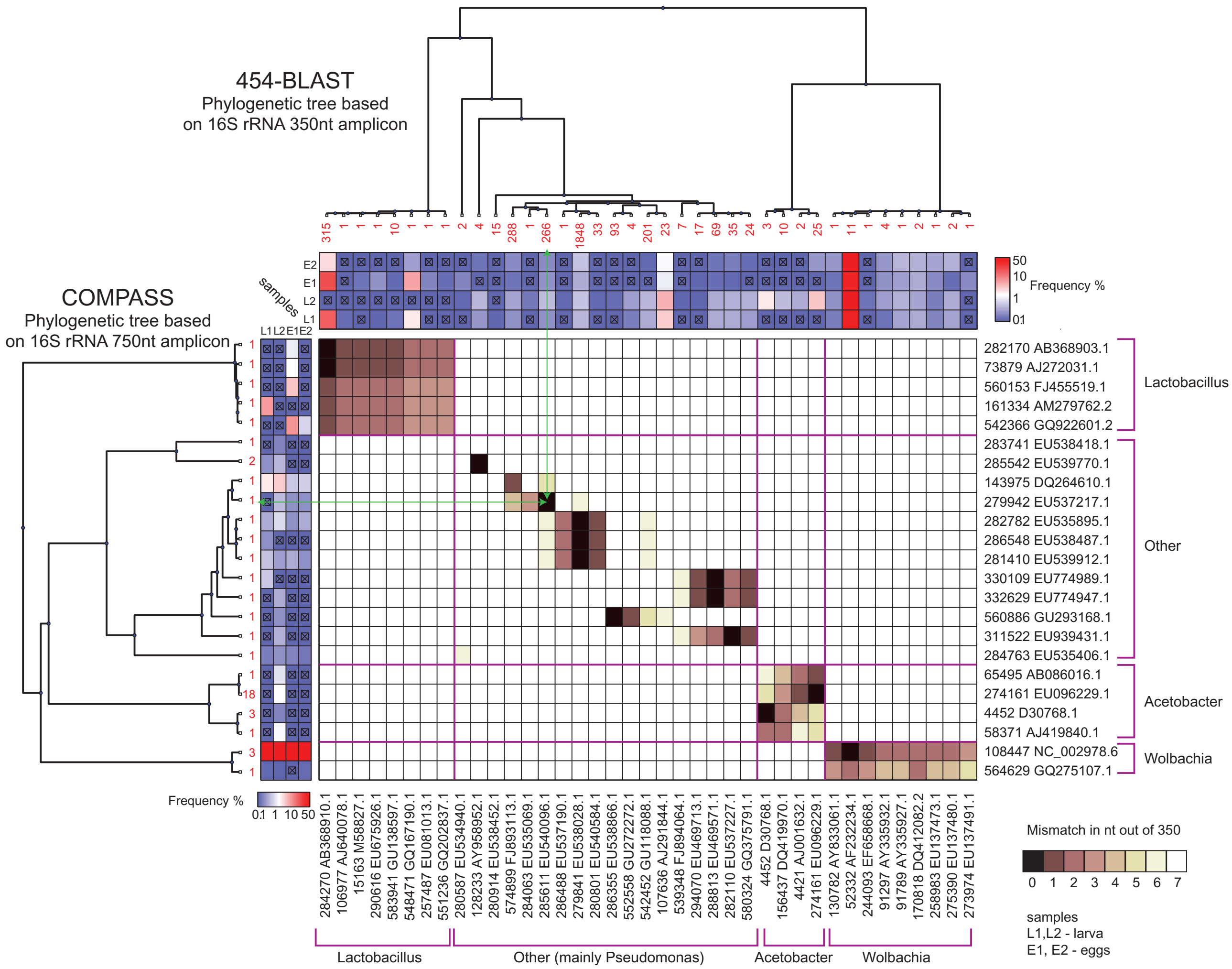

Figure 5. Reconstruction of *Drosophila* samples—COMPASS and 454-BLAST. Experimental results comparing 454-BLAST SRF and Illumina-based COMPASS framework on four *Drosophila* samples, L1, L2, E1 and E2. The Roche 454 amplicon matches the first 450 bp of the 750 bp Illumina amplicon. To decrease the effect of read errors, the 454-BLAST approach was based on the first 350 nt. COMPASS was based on the 750 bp amplicon. On the left we show a phylogenetic tree based on sequences inferred by COMPASS (with frequency >0.1%), and a similar tree based on the 454-BLAST is shown on top (for a description of phylogenetic tree building see Supplementary Methods S3). Database accession names are shown on the left and below, respectively (further details appear in Supplementary Data sets S1–S2). Heatmaps along each tree display the inferred frequency of each sequence in each sample (frequencies <0.1% are marked by 'x') and the displayed numbers correspond to the number of sequences from the full 16S database that are identical on the relevant amplicon, thus representing resolution. The central matrix displays the similarity between sequences found by 454-BLAST and COMPASS, calculated over the shared 350 bp long sequence. Complete identity in shown in black, while seven or more mismatches appear as white. Four main clusters were found—*Lactobacillus*, *Acetobacter*, *Wolbachia* and 'Other' that were mostly Pseudomonas. An example of the improved resolution obtained by COMPASS is highlighted by two green lines pointing to a sequence found by both methods. The 454-BLAST solution corresponds to 266 sequences that share the 350 bp amplicon, while COMPASS allows selecting a single sequence.

several 16S sequences that share the same sequence over the 350 bp or 750 bp amplicon. The central matrix compares the results of the two methods, showing the Hamming distance between pairs of sequences (calculated over the overlapping region). The comparison shows a high agreement between the two methods, as evident from the matrix, detecting the same four main clusters of bacteria (*Wolbachia*, *Lactobacillus*, *Acetobacter* and 'Other' which comprised mostly *Pseudomonas*). A closer look into the matrix shows that 22 out of the 23 sequences found by COMPASS had counterparts found by 454-BLAST: 14 out of 23 sequences found by COMPASS had identical counterparts in the 454-BLAST method (namely the overlapping 350 bp region matches completely), while the other eight sequences differed by up to 4 nt. The COMPASS sequence that was not found by 454-BLAST was, however, found by MG-RAST together with all other COMPASS sequences; 454-BLAST found 39 bacteria, many of which are false positives, as shown in the next section. A single bacterium found by 454-BLAST, which had frequency of ~0.1%, was later validated and not found by COMPASS.

Table 2 summarizes the total abundance of each cluster as inferred by the three methods for each sample. Although absolute frequencies differ, probably owing to biases caused by different primers used, the rank order is consistent between methods—both rank order among the different clusters within each sample, and also for each

Table 2. A summary of the frequencies in each sample and each method for each of the four 'clusters'

	Lactobacillus			Other		
Method⇒ Sample⇓	COMPASS over 750	454-BLAST over 350 (%)	MG-RAST over 350 (%)	COMPASS over 750 (%)	454-BLAST over 350 (%)	MG-RAST over 350 (%)
L1	7.6%	21.1	22.3	4.6	6.5	9.0
L2	0%	0	0	7.3	10.2	15.1
E1	16.0%	33.6	34.7	1.7	2.5	3.7
E2	0.8%	3.1	3.9	1.8	3.2	5.3

	Acetobacter			*Wolbachia*		
Method⇒ Sample⇓	COMPASS over 750 (%)	454-BLAST over 350 (%)	MG-RAST over 350 (%)	COMPASS over 750 (%)	454-BLAST over 350 (%)	MG-RAST over 350 (%)
L1	0	0	0	87.5	68.2	65.3
L2	4.5	7.3	8.9	88.0	78.2	73.0
E1	0	0	0	82.1	61.3	59.7
E2	0	0.3	0.4	97.1	90.6	88.0

Values correspond to a summation of the relevant bacterial frequencies for each sample. Rank order is preserved over the 4 clusters within the same sample, and over the same cluster between samples.

cluster across samples. For example, both 454-BLAST, MG-RAST and COMPASS find that the *Lactobacillus* cluster has highest abundance in sample E1, with L1 next, then E2, and completely vanishes in L2.

COMPASS displays increased phylogenetic resolution

Each black entry in the matrix in Figure 5, which corresponds to a complete match between the methods, also manifests COMPASS's increased phylogenetic resolution compared with SRF. The numbers next to each leaf represent the number of sequences in the full 16S database that share the same 350 or 750 nt amplicon for SRF and COMPASS, respectively. The lower these numbers are, the higher the phylogenetic resolution provided. As shown, sequences found by COMPASS correspond to a smaller number of full 16S sequences compared with 454-BLAST. For example, sequence *EU537217.1* corresponds to 266 sequences sharing an identical 350 nt region as found by 454-BLAST, while it is uniquely identified in COMPASS (follow green arrows in Figure 5). In all 14 complete matches between COMPASS and 454-BLAST, COMPASS achieved higher resolution.

COMPASS displays less false-positive detections than SRF

In several cases, COMPASS detected a single sequence while 454-BLAST detected several highly similar bacteria. For example, in the case of the major *Wolbachia* strain detected (AF232234.1), 454-BLAST detected eight additional highly similar bacteria, which were all experimentally ruled out as described in the next section. The low number of falsely detected bacteria is a direct outcome of the longer amplicon, as the number of differences between the correct sequence and other highly similar bacteria increases with the length of the amplicon. Hence, the probability that specific read errors would divert COMPASS from the correct bacteria is highly reduced.

We performed a similar comparison for human saliva samples (see Supplementary Figure S5) showing good concordance between the two methods (e.g. 58 out of 82 sequences found by COMPASS had identical counterparts in 454-BLAST).

Validation via Sanger sequencing

We aimed at testing the predictions set by COMPASS, 454-BLAST and MG-RAST, validating the higher resolution provided by COMPASS's predictions and the reduction in the number of falsely detected bacteria. Validation was performed for the *Acetobacter* and *Wolbachia* genera over samples L1 and L2 (see Figure 6).

Overview of Sanger-based validation

We designed two specific primer pairs predicted to amplify long regions in most sequences of the *Acetobacter* and *Wolbachia* genera (See 'Materials and Methods' section). Following PCR amplification using each of these primers, the resulting product was sequenced using Sanger sequencing, thus allowing efficient detection of low abundance bacteria in these genera (as opposed to the more labor-intensive cloning and sequencing approach). Since different bacteria are present in the mixture, each chromatogram is expected to display more than a single peak at some locations, as demonstrated in Supplementary Figure S6 (the primers were designed to amplify highly similar bacteria, and therefore the number of such locations is limited). Therefore, we collected all possible base calls for each location along the Sanger sequence and compared it with sequences found by 454-BLAST, MG-RAST and COMPASS. A predicted sequence was termed 'Sanger validated' if its sequence matched the base calls along the Sanger sequence. A prediction is termed a false positive in case a sequence's nucleotide differs from the Sanger base calls in at least one location. For completeness, we used *in silico* PCR to check all sequences, which were neither detected by COMPASS nor by 454-BLAST/

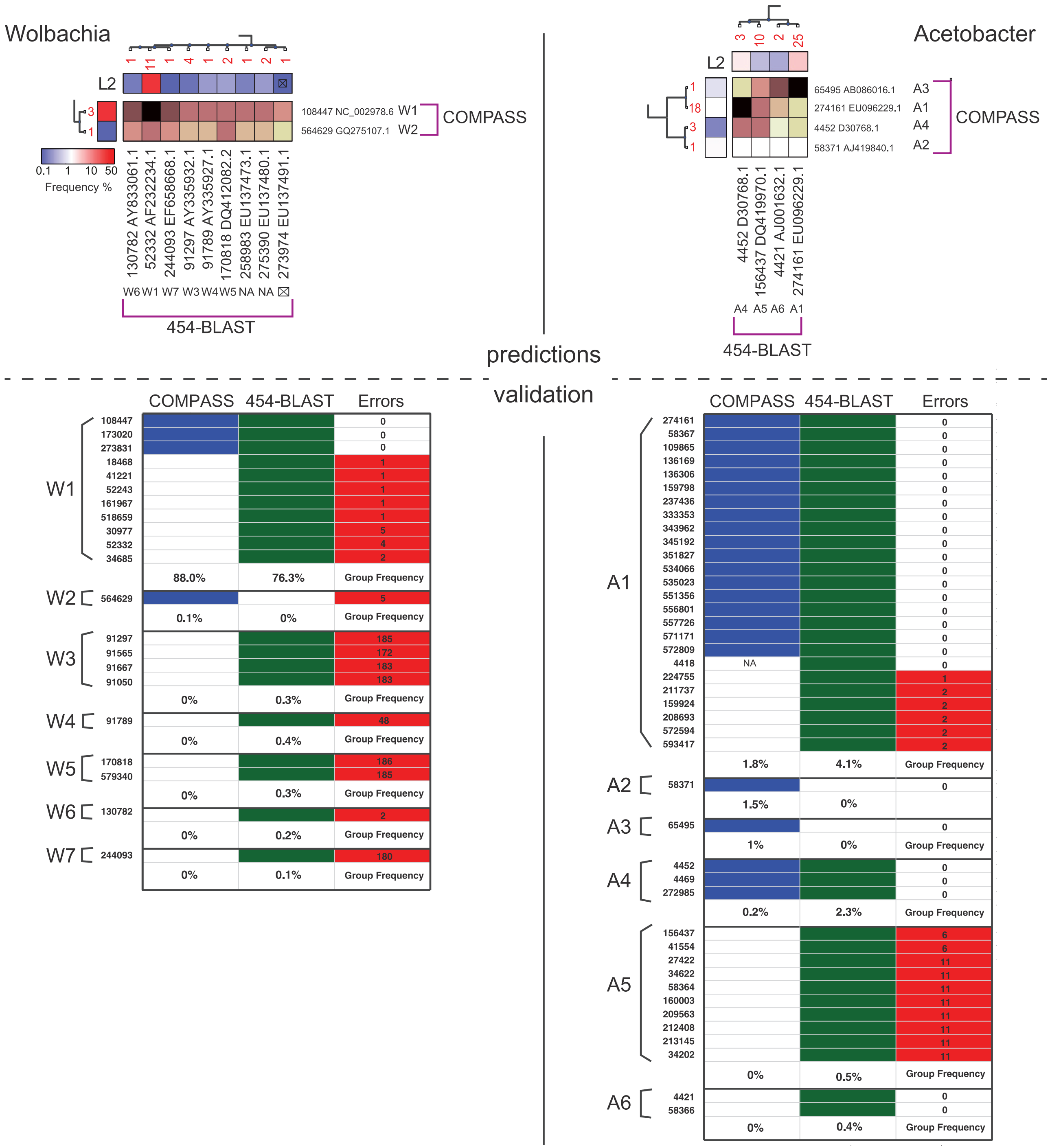

Figure 6. Validation via Sanger sequencing—COMPASS and 454-BLAST. The left and right sides in the upper part of the figure display a zoom in to the Wolbachia and Acetobacter regions of Figure 5 for sample L2, indicating the predicted groups by 454-BLAST and COMPASS on the columns and rows, respectively. Validation by Sanger sequencing is shown in the lower part of the figure. Each row corresponds to a sequence found by either COMPASS (marked blue) or by 454-BLAST (marked green). The last column presents the number of mismatches between the bacterium and the Sanger sequence, when combining results of forward and reverse strands. Cases of nonzero mismatches indicate a wrong prediction and are marked red. Bacteria are grouped according to Figure 5, for both COMPASS and 454-BLAST predictions, where the first bacterium in each group matches the representative bacterium in each group. Groups for which the COMPASS bacteria are a subset of 454-BLAST bacteria appear as a single group, where only a subset of the bacteria is marked green.

MG-RAST, but may have potentially been amplified by the relevant primer pair. None of these sequences were 'Sanger-validated', and hence no 'false negatives' occurred for both *Acetobacter* and *Wolbachia*. A cartoon of the procedure appears in Supplementary Figure S6.

Summary of results displays reduction in falsely detected bacteria

Results for sample L2 are summarized in Table 3 (results for L1 were similar, and thus are omitted), displaying the number of 'Sanger validated' groups (each group contains

Table 3. Sanger validation results summary: The total number of 'Sanger-validated' groups and bacteria, together with those predicted by COMPASS, 454-BLAST and MG-RAST for *Wolbachia* and *Acetobacter*

Family	Number of Sanger-validated groups	Number of bacteria in Sanger-validated groups	Method	Number of predicted groups	Number of bacteria in predicted groups	Number of false-positive groups	Number of false-positive bacteria	Total nonvalidated frequency (%)
Acetobacter	5	26	COMPASS over 750	4	23	0	0	0
			454-BLAST over 350	4	40	1	16	0.5
			MG-RAST over 350	3	38	1	16	4.9
Wolbachia	1	3	COMPASS over 750	2	4	1	1	0.1
			454-BLAST over 350	6	20	5	17	1.3
			MG-RAST over 350	5	15	4	12	5.4

Also shown is the number of false-positive groups and bacteria, and the total nonvalidated frequency, namely, the sum of inferred frequencies of groups that had nonzero mismatches with Sanger sequencing. The total nonvalidated frequency in the case of MG-RAST is higher owing to a single group to which the method incorrectly assigns relatively high frequency for both *Acetobacter* and *Wolbachia* (see Supplementary Figure S4).

bacteria sharing the same sequence over the 350 bp or 750 bp amplicon). Also presented are the corresponding numbers of groups predicted by COMPASS, 454-BLAST and MG-RAST, and the number of false-positive bacteria. The same information appears also for individual bacteria within these groups. For *Acetobacter*, COMPASS had zero falsely detected bacteria, compared with 16 in 454-BLAST/MG-RAST, and for *Wolbachia*, 454-BLAST had 17 false-positive detections and MG-RAST had 12, compared with a single case in COMPASS (whose predicted frequency was 0.1%).

Detailed results validate increased resolution of COMPASS predictions

Groups predicted by COMPASS and 454-BLAST appear in Figure 6, while the analogues comparison between MG-RAST and COMPASS appears in Figure S4 (see Supplementary Results S3).

Validation results for each bacterium within these groups appear in the left and right parts of Figure 6 for *Wolbachia* and *Acetobacter*, respectively. The upper part of the figure presents the bacteria predicted by COMPASS and 454-BLAST, while the lower part presents validation of these predictions. Bacteria predicted by COMPASS and 454-BLAST are marked by green and blue, respectively. Groups for which the COMPASS bacteria are a subset of 454-BLAST bacteria appear as a single group, while only predicted bacteria are marked green. The last column in each table presents Sanger validation results, and the number of mismatches between the relevant bacterium sequence and the chromatogram (See 'Materials and Methods' section).

Wolbachia (Figure 6 left)

The dominant high abundance group of sequences (W1) contained 11 bacteria in the 454-BLAST method. Out of these, only the three bacteria that were found by COMPASS were 'Sanger validated', while the remaining eight sequences were false positives. This demonstrates the correctness of COMPASS predictions together with its improved phylogenetic resolution, namely instead of 11 indistinguishable bacteria in the 454-BLAST case, COMPASS correctly ruled out 8 bacteria. The chromatogram showed that mismatches between these 8 bacterial sequences and the Sanger sequence occurred outside the Roche 454 amplicon (See Supplementary Figure S7), indicating that lower resolution of 454-BLAST is an inherent result of its short amplicon.

In addition, COMPASS predicted a low abundance (0.1%) bacterium (W2) that was found to be incorrect. The 454-BLAST predicted the existence of five groups, W3–W7 (total frequency of 1.3%), in addition to group W1, which were all false positives. In both cases, these incorrect predictions are probably due to read errors that were mapped to bacterial sequences in the database, since mismatches were found within the Illumina/Roche 454–amplified regions (See Supplementary Figure S7). However, 454-BLAST found 17 incorrect bacteria, with much higher relative abundance, compared with the single false positive in COMPASS, suggesting that 454-BLAST is more prone to such read errors than COMPASS.

Acetobacter (Figure 6 right)

All bacteria in all groups predicted by COMPASS were validated, ranging from abundance of 1.8% to as low as 0.2%. Two of these groups (A1 and A4) were also predicted by 454-BLAST. However, group A1 contained 25 bacteria out of which the 18 bacteria shared by the COMPASS method were validated and other bacteria were found to be incorrect, which is another manifestation of increased resolution provided by COMPASS (an additional bacterium, *X71863.1*, which is part of this group was not amplified by the Illumina primers, hence was not predicted by COMPASS).

A single group (A6) was found by 454-BLAST with frequency 0.4% and did not appear in COMPASS's predictions. The A5 group found by the 454-BLAST, with frequency of 0.5% and including 10 bacteria was a false positive.

In summary, Sanger validation showed the correctness of COMPASS's predictions proving higher phylogenetic resolution together while reducing the number of false-positively detected bacteria.

DISCUSSION

The presented COMPASS approach provides means for high-resolution microbial profiling. COMPASS bridges the gap between the short read limitation of MPS and the necessity to sequence large regions to allow unique identification of bacterial sequences. The large analyzed region also provides increased robustness to experimental errors, e.g. chimeric reads and read errors, resulting in more accurate predictions. In addition, theoretical analysis provides guarantees and bound on performance.

We presented an application of COMPASS based on the 16S rRNA gene using Illumina single-end sequencing of eight samples and showed its extremely high concordance with the SRF 454-BLAST method, while providing an additional increased resolution and dramatically reducing falsely detected bacteria, as further validated via Sanger sequencing. Such accurate predictions make COMPASS suitable in scenarios in which high-resolution profiling is important.

Our extensive simulations have shown that COMPASS can cope with reads as short as 50 nt and a rather small number of reads to provide accurate high-resolution profiling, which sets the ground for large-scale and low-cost multiplexed profiling of many samples.

A basic assumption underlying COMPASS is that most bacteria in the mixture are represented in the sequence database. In case mixture bacteria do not appear in the COMPASS working database, for example, due to mutations in already known sequences or due to poorly studied ecosystems, COMPASS would provide the closest possible sequences in the database. In that respect, COMPASS is more adequate for profiling mixtures of rather well-studied environments, where changes in species identities and composition may have biological and/or clinical importance, and less for analyzing environments that mainly contain unknown bacteria.

A natural application for COMPASS is analyzing RNA-seq data from bacterial populations. Since ribosomal RNAs constitute a large fraction of total RNA produced, in many experimental cases a large fraction of the reads originate from the 16S gene, even without prior PCR amplification. Since 16S RNA-seq reads originate from random locations along the 16S, typical SRF methods can only use a small fraction of the reads originating from a single region. In contrast, COMPASS integrates all available reads and produces a coherent snapshot of bacterial metabolic activity. Unlike *de novo* assembly based methods, COMPASS can also seamlessly integrate reads from noncontiguous regions or multiple genes, thus can also be applied to databases such as MLST (36) or Ribosomal MLST (37) to further increase phylogenetic resolution. Such multigene integration of information will remain important also in case the future MPS read length would be larger. An application of COMPASS to whole-genome sequencing is in principle possible, and would become increasingly beneficial as more and more bacteria are sequenced.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online, including [38,39].

ACKNOWLEDGEMENTS

A.A., A.Z., O.Z. and N.S. would like to thank Miki Ben-Yosef, Zohar Pasternak and Edouard Jurkevitch for fruitful discussions, and Eytan Domany for his encouragement and support. N.S. thanks Ido Karlan and Rafi Sadowsky for their assistance.

FUNDING

The Open University of Israel grant [IDD-12/02 to N.S.]; NIH [P50 GM068763 to P.J.T.]. Funding for open access charge: The Open University of Israel internal grant.

Conflict of interest statement. None declared.

REFERENCES

1. Whitman,W.B., Coleman,D.C. and Wiebe,W.J. (1998) Prokaryotes: the unseen majority. *Proc. Natl Acad. Sci. USA*, **95**, 6578–6583.
2. Curtis,T.P., Sloan,W.T. and Scannell,J.W. (2002) Estimating prokaryotic diversity and its limits. *Proc. Natl Acad. Sci. USA*, **99**, 10494–10499.
3. Rappe,M.S. and Giovannoni,S.J. (2003) The uncultured microbial majority. *Annu. Rev. Microbiol.*, **57**, 369–394.
4. Oliver,J.D. (2005) The viable but nonculturable state in bacteria. *J. Microbiol.*, **43**, 93–100.
5. DeSantis,T.Z., Dubosarskiy,I., Murray,S.R. and Andersen,G.L. (2003) Comprehensive aligned sequence construction for automated design of effective probes (CASCADE-P) using 16S rDNA. *Bioinformatics*, **19**, 1461–1468.
6. Pruesse,E., Quast,C., Knittel,K., Fuchs,B.M., Ludwig,W., Peplies,J. and Glockner,F.O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.*, **35**, 7188–7196.
7. Cole,J.R., Wang,Q., Cardenas,E., Fish,J., Chai,B., Farris,R.J., Kulam-Syed-Mohideen,A.S., McGarrell,D.M., Marsh,T., Garrity,G.M. *et al.* (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.*, **37**, D141–D145.
8. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.-J., Chen,Z. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
9. Bennett,S.T., Barnes,C., Cox,A., Davies,L. and Brown,C. (2005) Toward the 1,000 dollars human genome. *Pharmacogenomics*, **6**, 373–382.
10. Hamady,M., Walker,J.J., Harris,J.K., Gold,N.J. and Knight,R. (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat. Methods*, **5**, 235–237.
11. Hamady,M. and Knight,R. (2009) Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res.*, **19**, 1141–1152.
12. The Human Microbiome Project Consortium. (2012) Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207–214.
13. Caporaso,J.G., Lauber,C.L., Costello,E.K., Berg-Lyons,D., Gonzalez,A., Stombaugh,J., Knights,D., Gajer,P., Ravel,J., Fierer,N. *et al.* (2011) Moving pictures of the human microbiome. *Genome Biol.*, **12**, R50.
14. Wang,Q., Garrity,G.M., Tiedje,J.M. and Cole,J.R. (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into

the new bacterial taxonomy. *Appl. Environ. Microb.*, **73**, 5261–5267.
15. Kuczynski,J., Lauber,C.L., Walters,W.A., Parfrey,L.W., Clemente,J.C., Gevers,D. and Knight,R. (2012) Experimental and analytical tools for studying the human microbiome. *Nat. Rev. Genet.*, **13**, 47–58.
16. Schleifer,K.H. (2009) Classification of Bacteria and Archaea: past, present and future. *Syst. Appl. Microbiol.*, **32**, 533–542.
17. Fox,G.E., Wisotzkey,J.D. and Jurtshuk,P. (1992) How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *Int. J. Syst. Bacteriol.*, **42**, 166–170.
18. Miller,C.S., Baker,B.J., Thomas,B.C., Singer,S.W. and Banfield,J.F. (2011) EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Genome Biol.*, **12**, R44.
19. Fan,L., McElroy,K. and Thomas,T. (2012) Reconstruction of ribosomal RNA genes from metagenomic data. *PLoS One*, **7**, e39948.
20. Ong,S.H., Kukkillaya,V.U., Wilm,A., Lay,C., Ho,E.X.P., Low,L., Hibberd,M.L. and Nagarajan,N. (2013) Species identification and profiling of complex microbial communities using shotgun illumina sequencing of 16S rRNA amplicon sequences. *PLoS One*, **8**, e60811.
21. Sharpton,T.J., Riesenfeld,S.J., Kembel,S.W., Ladau,J., O'Dwyer,J.P., Green,J.L., Eisen,J.A. and Pollard,K.S. (2011) PhylOTU: a high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data. *PLoS Comput. Biol.*, **7**, e1001061.
22. Lozupone,C. and Knight,R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.*, **71**, 8228–8235.
23. Zuk,O., Amir,A., Zeisel,A., Shamir,O. and Shental,N. (2013) Accurate profiling of microbial communities from massively parallel sequencing using convex optimization. In: Kurland,O., Lewenstein,M. and Porat,E. (eds), *SPIRE 2013*. Jerusalem, Israel, pp. 279–297.
24. Wang,Y. and Qian,P.Y. (2009) Conservative fagments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. *PLos One*, **4**, e7401.
25. Weisburg,W.G., Barns,S.M., Pelletier,D.A. and Lane,D.J. (1991) 16S ribosomal DNA amplification for phylogenetic study. *J. Bacteriol.*, **173**, 697–703.
26. Grant,M. and Boyd,S. (2008) Graph implementations for nonsmooth convex programs. In: Blondel,V., Boyd,S. and Kimura,H. (eds), *Recent Advances in Learning and Control*. Lecture Notes in Control and Information Sciences, Zurich, pp.95–110.
27. Lozupone,C.A., Hamady,M., Kelley,S.T. and Knight,R. (2007) Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.*, **73**, 1576–1585.
28. Fukuyama,J., McMurdie,P.J., Dethlefsen,L., Relman,D.A. and Holmes,S. (2012) Comparisons of distance methods for combining covariates and abundances in microbiome studies. *Pac. Symp. Biocomput.*, **2012**, 213–224.
29. Caporaso,J.G., Lauber,C.L., Walters,W.A., Berg-Lyons,D., Lozupone,C.A., Turnbaugh,P.J., Fierer,N. and Knight,R. (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl Acad. Sci. USA*, **108**, 4516–4522.
30. Marcais,G. and Kingsford,C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–770.
31. Stern,S., Fridmann-Sirkis,Y., Braun,E. and Soen,Y. (2012) Epigenetically heritable alteration of fly development in response to toxic challenge. *Cell Rep.*, **1**, 528–542.
32. Meyer,F., Paarmann,D., D'Souza,M., Olson,R., Glass,E.M., Kubal,M., Paczian,T., Rodriguez,A., Stevens,R., Wilke,A. *et al.* (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.
33. Faith,J.J., McNulty,N.P., Rey,F.E. and Gordon,J.I. (2011) Predicting a human gut microbiota's response to diet in gnotobiotic mice. *Science*, **333**, 101–104.
34. Walker,A.W., Ince,J., Duncan,S.H., Webster,L.M., Holtrop,G., Ze,X.L., Brown,D., Stares,M.D., Scott,P., Bergerat,A. *et al.* (2011) Dominant and diet-responsive groups of bacteria within the human colonic microbiota. *ISME J.*, **5**, 220–230.
35. Chandler,J.A., Lang,J.M., Bhatnagar,S., Eisen,J.A. and Kopp,A. (2011) Bacterial communities of diverse *Drosophila* species: ecological context of a host-microbe model system. *PLoS Genet.*, **7**, e1002272.
36. Maiden,M.C., Bygraves,J.A., Feil,E., Morelli,G., Russell,J.E., Urwin,R., Zhang,Q., Zhou,J., Zurth,K., Caugant,D.A. *et al.* (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl Acad. Sci. USA*, **95**, 3140–3145.
37. Jolley,K.A., Bliss,C.M., Bennett,J.S., Bratcher,H.B., Brehony,C., Colles,F.M., Wimalarathna,H., Harrison,O.B., Sheppard,S.K., Cody,A.J. *et al.* (2012) Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology*, **158**, 1005–1015.
38. Kao,W.-C., Stevens,K. and Song,Y.S. (2009) BayesCall: a model-based base-calling algorithm for high-throughput short-read sequencing. *Genome Res.*, **19**, 1884–1895.
39. Kao,W.-C. and Song,Y.S. (2011) naiveBayesCall: an efficient model-based base-calling algorithm for high-throughput sequencing. *J. Comput. Biol.*, **18**, 365–377.