Microbiome

CrossMark

Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling

Garold Fuks[1†], Michael Elgart[2†], Amnon Amir[3], Amit Zeisel[4], Peter J. Turnbaugh[5], Yoav Soen[2] and Noam Shental[6*]

Abstract

Background: Most of our knowledge about the remarkable microbial diversity on Earth comes from sequencing the 16S rRNA gene. The use of next-generation sequencing methods has increased sample number and sequencing depth, but the read length of the most widely used sequencing platforms today is quite short, requiring the researcher to choose a subset of the gene to sequence (typically 16–33% of the total length). Thus, many bacteria may share the same amplified region, and the resolution of profiling is inherently limited. Platforms that offer ultra-long read lengths, whole genome shotgun sequencing approaches, and computational frameworks formerly suggested by us and by others all allow different ways to circumvent this problem yet suffer various shortcomings. There is a need for a simple and low-cost 16S rRNA gene-based profiling approach that harnesses the short read length to provide a much larger coverage of the gene to allow for high resolution, even in harsh conditions of low bacterial biomass and fragmented DNA.

Results: This manuscript suggests Short MUltiple Regions Framework (SMURF), a method to combine sequencing results from different PCR-amplified regions to provide one coherent profiling. The de facto amplicon length is the total length of all amplified regions, thus providing much higher resolution compared to current techniques. Computationally, the method solves a convex optimization problem that allows extremely fast reconstruction and requires only moderate memory. We demonstrate the increase in resolution by in silico simulations and by profiling two mock mixtures and real-world biological samples. Reanalyzing a mock mixture from the Human Microbiome Project achieved about twofold improvement in resolution when combing two independent regions. Using a custom set of six primer pairs spanning about 1200 bp (80%) of the 16S rRNA gene, we were able to achieve ~ 100-fold improvement in resolution compared to a single region, over a mock mixture of common human gut bacterial isolates. Finally, the profiling of a *Drosophila melanogaster* microbiome using the set of six primer pairs provided a ~ 100-fold increase in resolution and thus enabling efficient downstream analysis.

Conclusions: SMURF enables the identification of near full-length 16S rRNA gene sequences in microbial communities, having resolution superior compared to current techniques. It may be applied to standard sample preparation protocols with very little modifications. SMURF also paves the way to high-resolution profiling of low-biomass and fragmented DNA, e.g., in the case of formalin-fixed and paraffin-embedded samples, fossil-derived DNA, or DNA exposed to other degrading conditions. The approach is not restricted to combining amplicons of the 16S rRNA gene and may be applied to any set of amplicons, e.g., in multilocus sequence typing (MLST).

Keywords: Microbial profiling, Microbiome, 16S rRNA gene, High resolution

---

* Correspondence: shental@openu.ac.il
†Equal contributors
6Department of Computer Science, The Open University of Israel, 43107
Ra'anana, Israel
Full list of author information is available at the end of the article

Background

Bacteria are the most diverse domain on our planet [1], playing a crucial role in numerous ecosystems as well as an inseparable role in human health and disease. With the advent of next-generation sequencing technologies, molecular identification has replaced traditional culturing, allowing a window into the diversity of microbial communities. Most of these studies are based on sequencing of the 16S rRNA gene, which has several highly conserved regions interleaved with variable regions. Conserved regions serve as "anchors" for designing polymerase chain reaction (PCR) primers while sequencing the variable regions identifies the bacteria.

Such PCR primers are required to be "universal," i.e., amplify most bacterial 16S rRNA gene sequences, while maximizing the phylogenetic resolution, namely, the ability to distinguish among different bacteria based on the amplified sequence (for a comparison between different regions see [2] and references therein). However, the main design criterion for many current 16S rRNA gene primers is to increase the universality under the constraint of amplifying a region that matches the read length available by the sequencing platform (e.g., V3-V4 primers for Illumina paired-end sequencing [3–6]). Hence, the resolution is limited since many bacteria may share the same short amplified sequence.

High-resolution identification of bacteria in mixtures may prove highly useful in subsequent analysis, especially when follow-up studies require isolation and monoculture experiments. The basic aim is, therefore, to allow an exact identification of each bacterium in a mixture, or at least minimize the *ambiguity*, i.e., the number of bacteria that share the same 16S rRNA gene "footprint."

Using short reads over a long amplicon to increase resolution

The main reason for matching the amplicon length to the read length is technical and not biological. In recent years, several computational methods that allow higher resolution profiling based on an amplicon longer than the read length have been proposed [7–12]. These methods receive short reads that originate from one long amplified region and solve some optimization problem that seeks a combination of bacteria that best explains the set of reads in the experiment. Different methods vary in their fine details, representation, applied cost function, and optimization method. This approach, which we term the "single long-region framework," suffers from several drawbacks:

(i) *Non-standard 16S rRNA gene sample preparation*: Prior to sequencing, a long amplicon requires shearing (e.g., by sonication) followed by size selection and adaptors' ligation. Apart from being quite labor-intensive and time-consuming, both size
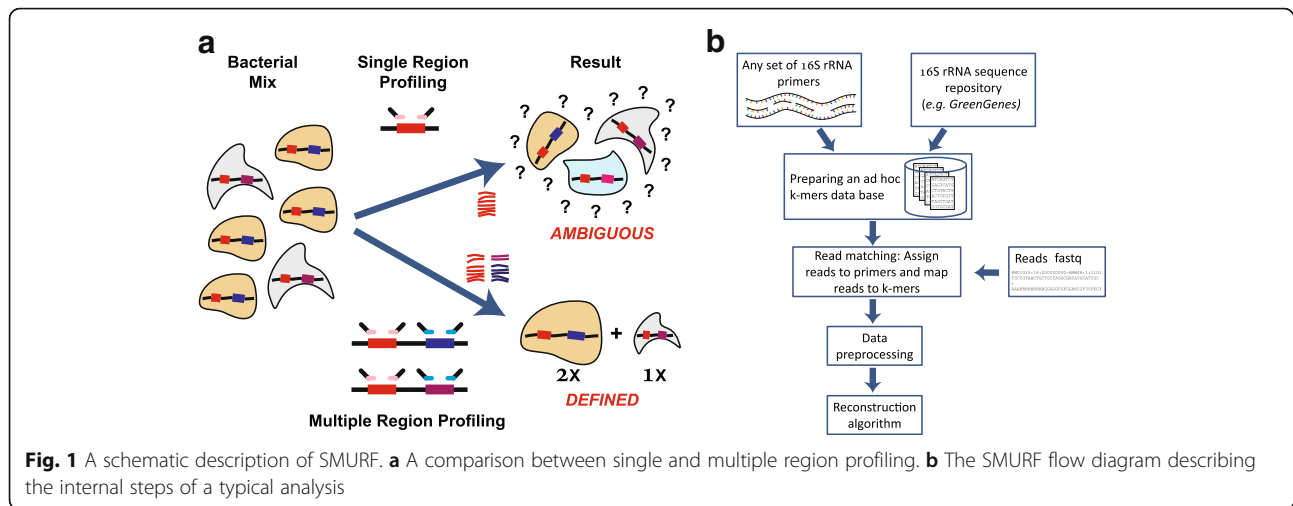
selection and ligation steps cause significant material loss that could be detrimental when detecting low-frequency bacteria in complicated biological mixtures. An additional PCR step is required to compensate for such material loss, hence additional PCR bias introduced.

(ii) *Primers are far from being universal*: Current primers for amplifying the whole 16S rRNA gene (e.g., 8F and 1492R [13]) significantly improve resolution, compared to sequencing a short region; however, they are highly non-universal. To increase universality, some mix of primer pairs is required.

(iii) *Sequencing fragmented DNA* may not be feasible: Increasing resolution via sequencing a long amplicon implicitly assumes that available DNA molecules are long enough. Although this is often the case, there are conditions in which molecules are fragmented, e.g., when DNA is extracted from formalin-fixed and paraffin-embedded (FFPE) blocks, fossil-derived DNA, or DNA exposed to harsh environments (radiation, reducing agents, etc.). An average length of a DNA molecule in this case may be shorter than needed for amplifying a long region, and thus, amplification may not be possible. Moreover, sometimes, even standard primers may not be suitable since the molecules' lengths are too short (if the average molecule length is about 200 bp then also custom V3-V4 primers may not be applied).

Short multiple-regions framework

Here, we propose a different approach that allows high-resolution profiling. The method is based on independent PCR amplification and sequencing of several regions (actually any number of regions) along the 16S rRNA gene, which are computationally combined to provide a joint estimate of the microbial community composition (Fig. 1). We term this method "Short MUltiple Regions Framework (SMURF)." Since several regions are amplified, the de facto amplicon is large allowing for a high resolution similar to the single long-region framework, while having the following advantages:

(i) *Availability of custom primers and standard sample preparation*: SMURF can be applied to any combination of common primer pairs (e.g., V1-V3, V3-V5, V4) without the need for designing custom long-amplicon primers. Also, DNA shearing, size selection, and ligation steps needed in the long-amplicon case are not required. Basically, the approach allows high resolution without the need for modifying experimental procedures

Fig. 1 A schematic description of SMURF. **a** A comparison between single and multiple region profiling. **b** The SMURF flow diagram describing the internal steps of a typical analysis

(namely, library preparation is independently performed for each primer pair, as in standard protocols).

(ii) *"Mix-and-match" primers to increase primers' universality and resolution*: Increasing universality and resolution in SMURF amounts to amplifying additional regions until the required universality and resolution are achieved. Rather than aiming to improve universality [14] or searching for the single and optimal set of primers [2], SMURF allows to combine any set of "sub-optimal" primers to achieve both superior universality and resolution compared to each of these primers applied by it. Results would evidently depend on the properties of the chosen set of primers, and therefore, a wise selection would yield more accurate profiling.

(iii) *SMURF allows sequencing fragmented DNA*: The "building blocks," i.e., the amplicons, are short and thus allow amplification also in case of highly fragmented DNA.

(iv) *Multiple regions tend to "average" PCR bias*: PCR amplification may have a different efficiency for different bacteria [15, 16], which results in considerable differences in profiling depending on primer choice [17–20]. Moreover, Gohl et al. [21] have shown that the *same* primer pair may yield different profiling results depending on the specific library preparation protocol. Since SMURF reconstruction is performed based on several amplified regions, and assuming that biases are independent in every region, reconstruction tends to average the effects of such biases.

(v) *Algorithmic advantages*: The mathematical formulation of SMURF is analogous to the one formerly suggested for the long-region methods [7, 12]. However, there are several computational advantages of SMURF that stem from combining

short regions. First, long-region methods require aligning reads to reference genomes before estimating bacterial frequencies. This time-consuming and error-prone step (e.g., via Bowtie [22] or BWA [23]) is not needed in SMURF since each read is readily associated with its primer and then matched to the corresponding region's k-mer database (a k-mer in this context is a bacterial sequence from the amplified region whose length matches the read length). Second, data representation in this case is much more efficient. The number of k-mers representing a bacterium is orders of magnitude smaller than for the long-region framework (a bacterium is represented by one or two k-mers per region as opposed to the long-region framework, where the number of k-mers is equal to the region's length). Consequently, optimization does not require handling huge matrices, which together with fixed-point iteration formula results in an extremely fast reconstruction. The third computational advantage is that the SMURF cost function is convex, and thus, convergence to a global optimum is guaranteed.

The manuscript presents the improvement in resolution achieved by SMURF and demonstrates the advantages of detecting near full-length 16S rRNA gene sequences. The work is divided into several parts: (i) First, using in silico simulations, we show a marked increase in resolution as a function of the number of amplified regions; (ii) second, the correctness of the method is assessed and verified via analysis of several mock mixtures. In the first experiment, a mock mixture was sequenced using a new set of six primer pairs each amplifying about 200 bp of the 16S rRNA gene. The experiment allowed us to confirm the increase in resolution as a function of the number of applied primer pairs. Next, we re-analyzed a mock mixture from the Human Microbiome Project combining reads originating

from two regions (V1-V3 and V3-V6). Apart from correct reconstruction and improvement in resolution when combining two regions, this experiment displayed the applicability of SMURF to available data acquired via standard protocols; finally, (iii) a real-world experiment characterizing a dynamic change occurring in the complex *D. melanogaster* microbiome was analyzed. We demonstrate the correctness of SMURF's results and describe how high-resolution profiling was essential for efficient downstream analyses.

Methods
Steps in SMURF are detailed (Fig. 1b), followed by a description of simulations and experiments performed (a Matlab code of SMURF is available at https://github.com/NoamShental/SMURF).

Database preparation
Database
The basic database used was the Greengenes (GG) 16S rRNA gene database (May 2013 version) [24] (other databases, e.g., SILVA [25], can also be used as described in the SMURF software package). Sequences containing *ambiguity* were duplicated by all possible sequence values (up to 64 combinations per sequence; sequences that contained more than three ambiguous nucleotides were discarded). Consequently, the basic database consisted of 1,402,801 unique 16S rRNA gene sequences. *Preparing an ad hoc, protocol-specific database*: Given the sequencer's read length, k, and each region's primer pair, a list of all possible k-mers in each region was generated based on the basic database. A bacterium was termed "amplified" in a specific region if each of the forward and reverse primers had at most two mismatches with the bacterial sequence. Data was presented using a database matrix, M, where M_{hj} is the probability of observing the k-mer h in bacterium j, $M_{hj} = \Pr(kmer = h| bacterium = j)$. The matrix M holds the information of all regions (i.e., for all J bacteria and H possible k-mers). In (the rare) case that the same k-mer appears in two regions, M holds a separate line for each.

For paired-end sequencing, each bacterial sequence may contribute a single k-mer per region (the k-mer is a concatenation of two bacterial sequences of length k that begin from the two ends of the amplicon). Hence, $M_{hj} = 1/R_j$, where R_j is the number of regions amplified for bacterium j, which also equals the number of k-mers originating from bacterium j. For single-end sequencing, there are two possible k-mers per amplified region (one from each amplicon ends), i.e., $M_{hj} = 0.5/R_j$. Hence, the maximal matrix size is the number of bacterial sequences in the database multiplied by twice the number of regions, which is orders of magnitude smaller than for single long-region framework, which greatly simplifies algorithm complexity and run time

(e.g., the analogous matrix M in COMPASS [9] contains ~ 100-fold more rows).

Read matching
Reads were assigned to a region by the primer sequence, and hence, "alignment" amounts to counting the number of mismatches between a read and the list of k-mers in that region (compared to applying complicated and costly alignment algorithms such as Bowtie required in the long region approach). A read is termed "matched" to a k-mer if there were at most two mismatches between the read and the k-mer after excluding the primer sequence. Following alignment of read i, it was assigned a probability that it originated from a k-mer h, E_{ih}. Assuming a constant error probability per nucleotide p_e, and independence between different locations along the read, then,

$$E_{ih} = \Pr(read = i|kmer = h) = \left(\frac{p_e}{3}\right)^{n_e}(1-p_e)^{(L-n_e)}$$

where n_e is the number of mismatches between the k-mer and the read (results were almost independent of the selected value of p_e (Additional file 1: Figure S4), and thus $p_e = 0.005$ was used [26]).

Data preprocessing
Two filters were applied:

Filtering low frequency reads: Reads that appeared less than 10^{-4} times the number of reads in a region were discarded. The filter sets the lowest detectable frequency (0.01%) while significantly decreasing false positive detections resulting from reads errors. The threshold value, 10^{-4}, may be lowered if the number of reads per region is high enough, and bacteria whose frequencies are lower than 0.01% are sought. To decrease false positive detections in such case, a read de-noising algorithm [27, 28] may be applied prior to SMURF.
Filtering candidate bacteria: In case no reads were aligned to a bacterium in a region for which there was a perfect match between the bacterial sequence and the primers, this bacterium was not considered in subsequent analysis.

Reconstruction algorithm
For reconstruction, we followed the maximum likelihood probabilistic framework suggested in [7, 12].

Given M_{hj} and E_{ih}, using the law of total probability, the probability of a read i given a bacterium j, is given by $Q_{ij} = \Pr(read = i|bacterium = j) = \sum_{h=1}^{H} E_{ih}M_{hj}$. The probability of observing the read is $p_i = \sum_{j=1}^{J} Q_{ij}\pi_j$, where π_j is a probability of selecting a read belonging to bacterium j. The likelihood of the N reads is given by

$\Pr(y) = \prod_{i=1}^{N} p_i$, which is a convex function of the unknown read proportion vector, π. Hence, using the expectation-maximization algorithm [29], an iterative procedure for the estimation of π was derived.

$$\pi_j^{(t+1)} = \pi_j^{(t)} \frac{1}{N} \sum_{i=1}^{N} \frac{Q_{ij}}{\sum_{j=1}^{J} Q_{ij} \pi_j^{(t)}}$$

The above procedure is similar to the one used in [12], while assuming constant error probability, which results in a convex cost function and a highly efficient fixed point iteration formula.

The frequency of bacterium j, x_j, was estimated by normalizing π_j by the number of k-mers originating from this bacterium j, which was simply the number of regions, R_j, amplified in bacterium j,

$$x_j = \frac{\frac{\pi_j}{R_j}}{\sum_{j=1}^{J} \frac{\pi_j}{R_j}}$$

Resolution, ambiguity, and the definition of a group

SMURF results in a list of reconstructed bacteria (i.e., full-length 16S rRNA gene sequences) and their frequencies. Some of these sequences may be indistinguishable by SMURF since they share the same sequence over all amplified regions. A set of reconstructed sequences that shares the same "footprint" over the amplified regions was defined as a "*group*."

A bacterium may also correspond to several reconstructed *groups*, e.g., due to several distinct 16S rRNA gene operons or because of profiling errors. Hence, in order to measure resolution, we define "*ambiguity*" as the effective number of 16S rRNA gene sequences that were "assigned" to a bacterium. In case a bacterium corresponds to a single *group*, *ambiguity* is in fact the *group's* size. To quantify *ambiguity*, we used the exponent of the Shannon entropy [30, 31]. Although *ambiguity* is actually a standard alpha diversity measure, we decided to use this term to emphasize the fact that lower values, which correspond to higher resolution, are preferred.

Resolution may not necessarily be associated with changes in taxonomy. In many cases, even a single region allows species-level identification; however, subspecies may not be decided. Hence, studies that require a unique identification of full-length 16S rRNA gene sequences would benefit from applying SMURF.

In silico simulations

Performance of SMURF was first evaluated by in silico simulations. Bacterial communities of 100 bacteria were randomly selected from the database, where their frequency followed a power law distribution ($1/x$). Profiling such a mixture is challenging since most bacteria have very low frequencies (half of the bacteria in the mixture have a cumulative frequency of 13%). The set of six primer pairs, later used in our experimental mock mixture, was applied in our simulations (the total amplicon length was about $6 \cdot 200$bp). The simulated number of reads over all regions was 200,000, where the sequencing error model was adopted from [9]. Results are based on 1000 simulated communities.

Effect of the number of regions

Each simulated mixture was reconstructed six times, starting from a single region and adding one region at a time (the order of regions was set so as to maximize the number of unique groups in the ad hoc database at each step). The total number of reads was kept the same (200,000) in all cases, i.e., the number of reads per region decreased when increasing the number of regions.

Simulation performance measures

The performance was measured by an adaptation of weighted precision and weighted recall defined in [9], quantifying the true discovery rates (i.e., one minus false discovery rate (1-FDR)) and true positive rate, respectively. Sequences in the simulated mixture were compared to sequences in the reconstructed mixture (comparison was performed over the whole 16S rRNA gene sequence). Whenever a perfect match occurred between a simulated bacterium m and reconstructed bacterium r, the indicator functions $I_m = 1$ and $I_r = 1$ were assigned. Even a single mismatch between r and m was considered as an error.

Weighted recall, i.e., the probability that a bacterium belongs to the reconstructed mixture given it appeared in the simulated mixture was $\sum_{m=1}^{100} f_m \cdot I_m$, where f_m was the frequency of a bacterium in the simulated mixture. Weighted precision, i.e., the probability that a bacterium belongs to the simulated mixture given it appeared in the reconstructed mixture, $\sum_{r=1}^{N_R} x_r \cdot I_r$, where x_r is the frequency of bacterium r in reconstructed mixture, and N_R is the total number of reconstructed bacteria.

Experimental profiling of a mock mixture
Primers

To test the method, a set of six primer pairs was designed (Table 1). Each primer pair was selected to maximize the number of bacteria amplified in a region while satisfying some desired primers' characteristics (e.g., GC content, melting temperature). Amplicon length was about 200–250 bp.

Mock mixture-sample preparation

Genomic DNA from ten bacterial strains, *Akkermansia muciniphila* ATCC BAA-835, *Lactobacillus casei* ATCC 334, *Bacteroides caccae* ATCC 43185, *Bacteroides fragilis* ATCC 25285, *Bacteroides ovatus* ATCC 8483T, *Eggerthella*

Table 1 The set of six primer pairs used in this study. The number of Greengenes sequences (out of the total 1,402,801 sequences) that have a perfect match with each primer pair. The primers' sequences and their median start locations along the forward strand are shown. The set of primers was used for profiling our mock mixture, the *D. melanogaster* samples, and for in silico simulations

Region	# amplified	Forward primer		Reverse primer	
		Sequence	Position	Sequence	Position
1	196415	5′-TGGCGGACGGGTGAGTAA-3′	74	5′- CTGCTGCCTCCCGTAGGA-3′	315
2	1122118	5′-TCCTACGGGAGGCAGCAG-3′	316	5′- TATTACCGCGGCTGCTGG-3′	484
3	660912	5′-CAGCAGCCGCGGTAATAC-3′	486	5′- CGCATTTCACCGCTACAC-3′	650
4	660342	5′-AGGATTAGATACCCTGGT-3′	752	5′- GAATTAAACCACATGCTC-3′	911
5	591604	5′-GCACAAGCGGTGGAGCAT-3′	901	5′- CGCTCGTTGCGGGACTTA-3′	1057
6	783882	5′-AGGAAGGTGGGGATGACG-3′	1143	5′- CCCGGGAACGTATTCACC-3′	1336

lenta DSM2243, *Eggerthella lenta* FAA 1-3-56 (both *Eggerthella lenta* strains have the same 16S rRNA gene sequence), *Escherichia coli* BL21, *Eubacterium rectale* ATCC 33656, and *Lactobacillus plantarum* ATCC 8014, was isolated using the UltraClean® Microbial DNA Isolation Kit (MO BIO Laboratories, Inc.), quantified using the Qubit™dsDNA BR Assay Kit (ThermoFisher Scientific), and mixed in equal proportions.

PCR amplification of six predefined regions was performed using the primers described above, and their products were mixed. Samples were then cleaned on column (Promega, Fitchburg, WI), concentration was measured using NanoDrop (NanoDrop Technologies, Wilmington, DE), and diluted to 50 ng/ul. Subsequently, the sample went through standard Illumina library preparation and sequenced on an Illumina HiSeq2000 sequencer using paired-end 100 bp reads (the sample was sequenced together with many other samples). The number of reads per region varied between 800,000 to about 2,000,000.

Analysis
Preprocessing reads
The barcodes (7 bp) and the primers (18 bp) were removed from the reads resulting in 76 bp long reads assigned to each region. Reads were discarded in three cases (i) a Phred score of less than 30 in more than 25% of nucleotides, (ii) more than three nucleotides having a Phred score of less than 10, or (iii) when containing one or more ambiguous base calling (e.g., "*N*").

Postprocessing SMURF results
All reconstructed 16S rRNA gene sequences were classified via the Ribosomal Database Project (RDP) Sequence Match engine [32], and results were combined (i.e., frequencies added) for bacteria sharing the same RDP assigned species. Combined *groups* having a total frequency lower than 0.1% were discarded, and the relative frequencies of the remaining *groups* were renormalized. Since both *Eggerthella lenta* strains share the same 16S rRNA gene sequence, their combined frequency was shown.

Evaluating the effect of the number of regions
The sample was reconstructed six times, starting from a single region and adding one region at a time in the same order as described in the simulations section.

Estimating resolution
The resolution was separately estimated for each bacterium in the mock mixture, using the exponent of the Shannon entropy [30, 31]. For each bacterium in the mock mixture, all groups whose RDP classification matched the bacterium were considered in this *ambiguity* calculation.

Re-analyzing Human Microbiome Project mock mixture
To further demonstrate the advantages of SMURF, we examined a microbial mock community used to validate the protocols of the Human Microbiome Project (HMP) [33]. This "even mock mixture" sample (https://www.ncbi.nlm.nih.gov/sra/SRX020130) containing 21 known bacterial strains was independently profiled over regions V1-V3 (SRR42565 and SRR42567) and V6-V9 (SRR42570 and SRR42571). Reads from the two experimental repeats for each region were pooled together for reconstruction. The total number of reads that passed a quality filter was 56,000 and 27,000 for V1-V3 and V6-V9, respectively. SMURF was used to combine results of these two regions to display the increase in resolution. Reads were trimmed to length 170 nt. Preprocessing the reads and postprocessing SMURF results were performed as described for the experimental mock mixture. To display the increase in resolution, reconstruction was performed based on each region separately and for combining the two regions.

Profiling of a *Drosophila* microbiome
Drosophila stocks
A *hairy-GAL4* line was obtained from the Bloomington Stock Center.

Extracting bacterial DNA
Flies were reared either on standard medium or medium containing 400 μg/ml of the G418 toxin (Sigma). For each experiment, seven to ten third stage larvae were collected,

their gut dissected, and total bacterial DNA was extracted via the "chemagic DNA Bacteria" kit (Chemagen).

Sample preparation and sequencing
PCR was performed using the set of six primer pairs described. Sample preparation was performed similarly to our mock mixture described above and sequenced on the same lane. The number of reads per region varied between ~ 300,000 to ~ 1,800,000.

Isolation and growth of commensal bacteria
Ten male flies were shaken in 1 ml of PBS buffer at room temperature for 30 min. One hundred milliliter of this fluid were serially diluted, spread on yeast peptone dextrose (YPD) agar plates, and colonies were grown at room temperature for 3 days. Several colonies were picked, underwent three additional rounds of isolation, grown overnight at 30 °C in liquid YPD, and stocked in 35% glycerol at – 80 °C.

Sanger sequencing of colonies
The 16S rRNA gene was PCR-amplified from bacterial DNA using universal primers 8F (AGAGTTTG ATCCTGGCTCAG) and 1492R (GGTTACCTTGTT ACGACTT) [13]. The PCR product was cloned into PGMT plasmid, sequenced, and compared to SMURF predicted sequences.
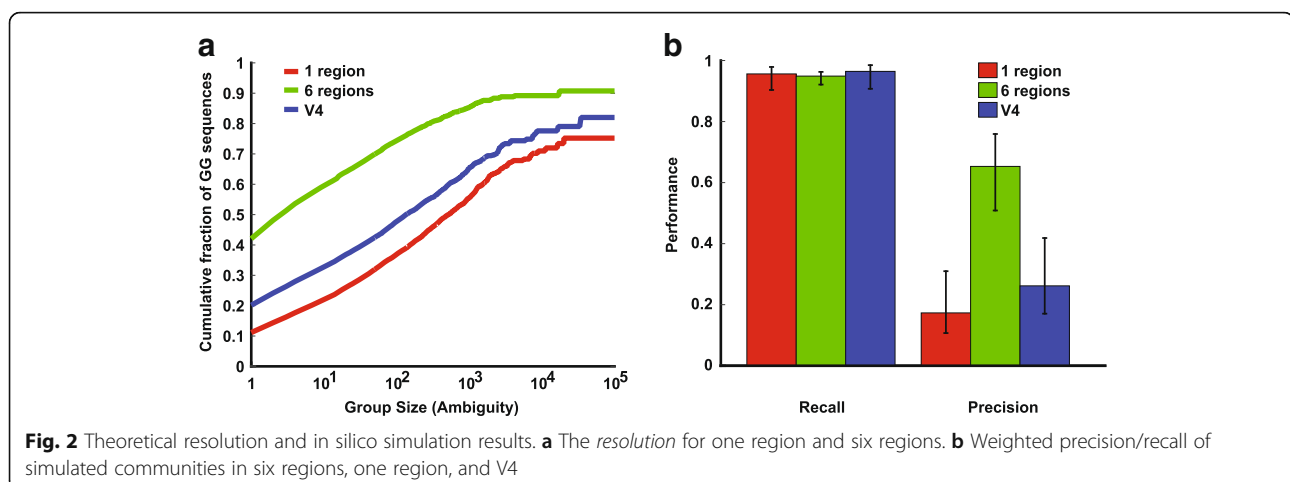
Results
Using the set of six primers pairs, the theoretical increase in resolution as a function of the number of regions is first presented. Second, SMURF's results of both in silico simulations and an experimental mock mixture using this set of primer pairs are shown. Third, we display the effect of applying SMURF to re-analyze an HMP mock mixture. Finally, we present the advantages of increasing resolution in detecting specific bacterial

strains in a real-world example of *D. melanogaster* microbiome undergoing a change.

Theoretical effect of combining short regions
To assess the potential improvement in resolution when combining several short regions, we calculated the *group* size of each bacterial sequence in the Greengenes (GG) database using our six primer pairs. Namely, the number of GG sequences that share the same sequence over the relevant regions and are thus indistinguishable were calculated (the *group's* size coincides with the *ambiguity* in this case and hence measures the resolution by which a bacterium may be identified). This procedure was repeated six times, starting from a single region and adding one region at a time as described in the "Methods" section. For comparison, data for variable region 4 (V4) is also shown (using standard forward and reverse primers: 5′-GTG CCAGCMGCCGCGGTAA-3′, 5′-GGACTACHVGGG TWTCTAAT-3′).

Figure 2a shows the fraction of GG sequences that belong to a *group* of up to a certain size, for one and six regions and for the V4 region (results for two to five regions were omitted for clarity and appear in Additional file 1: Figure S1). For example, 13% of the GG sequences belong to a *group* of size one, i.e., were uniquely identifiable using a single region. Using V4, the fraction of uniquely identifiable sequences was 21%, while for six regions, the fraction was doubled to about 40%, which displays the potential increase in resolution using a larger number of regions. The same phenomenon occurred also for sequences that were not uniquely identifiable and belong to larger *groups*—much more sequences belonged to *smaller groups* when using six regions compared to V4, corresponding to a significant increase in resolution. Although our single region provided inferior resolution compared to V4 (which uses a set of 20 oligonucleotides instead of just a single pair in our case), using merely two regions was sufficient for better performance than V4 (Additional file 1: Figure S1).

Fig. 2 Theoretical resolution and in silico simulation results. **a** The *resolution* for one region and six regions. **b** Weighted precision/recall of simulated communities in six regions, one region, and V4

In silico simulations

Figure 2b shows reconstruction performance when simulating 200,000 reads from a mixture of 100 bacteria randomly selected from the GG database, having a power law frequency distribution. Weighted precision and weighted recall were plotted for one and six regions and for using V4 primers (results for two to five regions were omitted for clarity and appear in Additional file 1: Figure S2).

Weighted recall (i.e., the chances that a bacterium from the simulated mixture appeared in the reconstruction) was very high in all cases. The 200,000 reads were divided among regions, and hence simulating six regions used six-fold less reads per region, resulting in slightly lower performance detecting the low-frequency sequences. These differences became more pronounced when simulating a lower number of reads (Additional file 1: Figure S2). Hence, a certain minimal number of reads per region was required in order to maintain high recall performance.

Weighted precision (i.e., the chances that a bacterium from the reconstructed community appeared in the original mixture) greatly improved when using six regions compared to a single region and to V4. Improvement in precision was mainly achieved by a reduction in the *groups*' sizes, due to the longer effective amplicon. The proba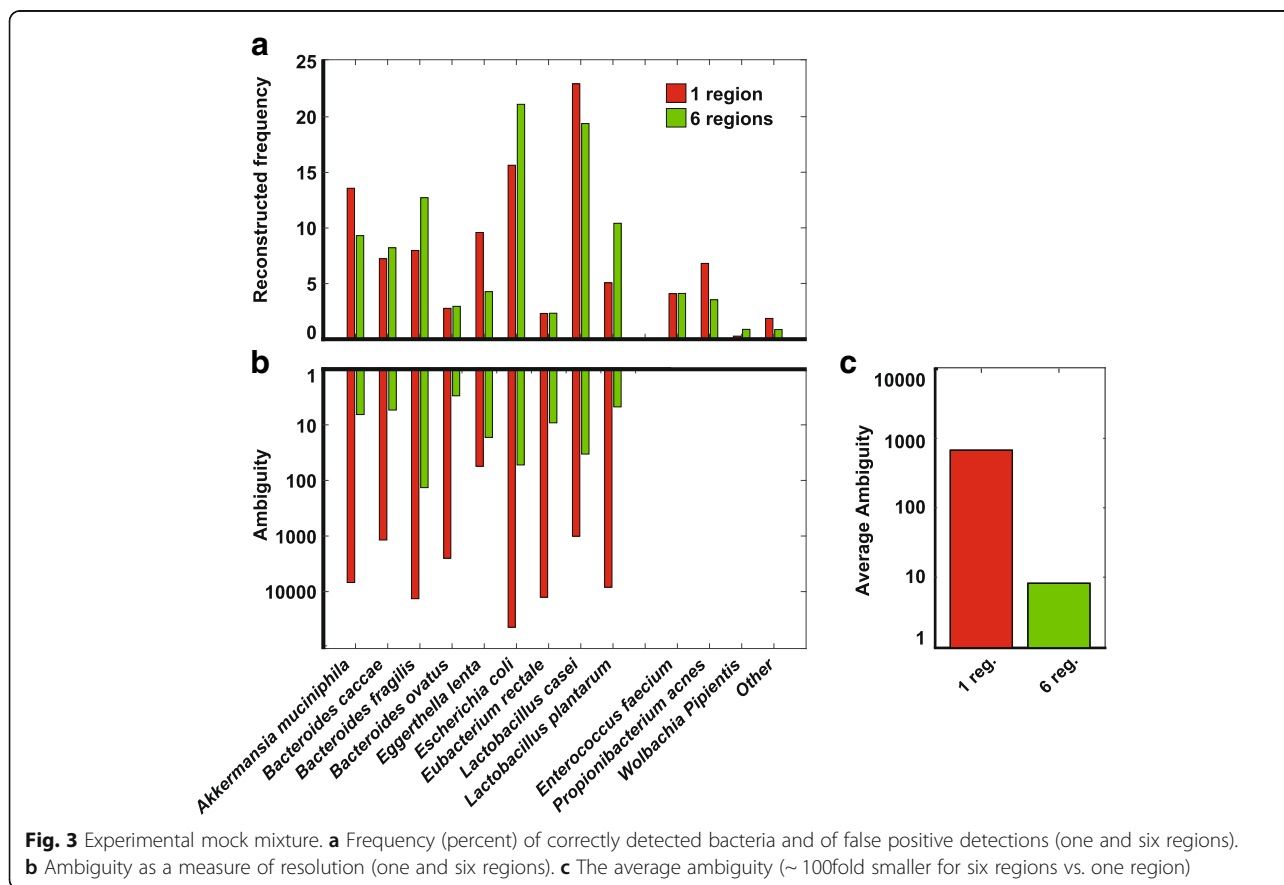bility of correctly detecting a bacterium which is uniquely identifiable, i.e., belongs to a *group* of size one, was about 39% when applying six regions, while being 13 and 21% for a single region and V4, respectively (similar results appeared in the former section describing the theoretical effects).

Since results may vary depending on the selected regions, performance was also tested over all singleton regions as well as over pairs of regions, triplets etc. (Additional file 1: Figure S3). Results varied significantly among singleton regions, yet performance was consistently better and more homogeneous when combining regions (e.g., variability in recall/precision among triplets was smaller than for pairs). Ambiguity also monotonically decreased with increasing the number of regions. However, unlike precision and recall, ambiguity highly depended on the specific selected regions and displayed more than a twofold difference between the worst and best combinations.

Experimental mock mixture
Reconstruction accuracy

Figure 3a shows SMURF profiling of our experimental mock mixture, based on one and six regions (results for two to five regions were omitted for clarity and appear in Additional file 1: Figure S5). All bacteria were detected, i.e., there were no false negative detections (100% recall),

Fig. 3 Experimental mock mixture. **a** Frequency (percent) of correctly detected bacteria and of false positive detections (one and six regions). **b** Ambiguity as a measure of resolution (one and six regions). **c** The average ambiguity (~ 100fold smaller for six regions vs. one region)

independent of the number of regions applied. There were several false positive detections, mainly *Enterococcus faecium* and *Propionibacterium acnes* which are known contaminations from our lab, and *Wolbachia pipientis* which was a contamination originating from *Drosophila melanogaster* samples prepared and sequenced in conjunction. The remaining false positive detections, which may also result from contaminations, sum to about 1%, i.e., precision was 99%. For some bacteria, frequencies varied between one and six regions (Additional file 1: Figure S5). For some bacteria, the reconstructed frequency was different from the expected 10% frequency, even after adjusting for the number of 16S rRNA gene copies (Additional file 1). These differences can be attributed to PCR amplification biases; specifically, the lower frequency estimates were related to primers mismatches (e.g., *Eubacterium rectale*, Additional file 1: Table S1). Results were almost identical when using the SILVA database instead of the Greengenes database (Additional file 1: Figure S6).

Resolution as function of the number of regions

Figure 3b shows the *ambiguity* of species detected in one and six regions (results for two to five regions appear in Additional file 1: Figure S5). *Ambiguity*, measured by the exponent of Shannon's entropy, measures the effective number of 16S rRNA gene sequences that were associated with a specific bacterium. For example, the *ambiguity* of *E. coli BL21* was 5384 based on one region and significantly dropped to 25 when using six regions. Hence, if we were to follow up on the exact *E. coli* sub-strain, we would have had 200-fold less hypotheses to check, as we show later on in the "Profiling a Drosophila microbiome" section. Ambiguity was smaller than 100 for all bacteria, and was lower than ten for six of them, when using the six regions protocol. For comparison, using a single region, *ambiguity* was larger than 100 in nine out of the ten bacteria, while being larger than 1000 for a five of them. Overall, the increase in resolution using six regions was ~ 100-fold better compared to a single region (Fig. 3c).

Re-analyzing an HMP mock mixture
Reconstruction accuracy

Figure 4a shows SMURF's profiling results of an HMP mock mixture containing 21 species (SRX 020130). All species were detected in both regions, apart from *Methano-brevibacter* which neither regions amplified [34]. Although mixed so as to create an equal number of 16S rRNA gene copies per bacterium, some species were nearly undetectable (e.g., *E. coli* and *Helicobacter pylori*) while others were significantly overrepresented (e.g., *Staphylococcus aureus*). As in our mock mixture, these biases result either from PCR amplification biases [35] or from specific library preparation steps [21].

Combining regions increases resolution

As in former cases, the major effect of combining regions was related to increasing resolution. When reconstructing the community jointly from two regions, the resolution improved compared to both single regions for 11 bacteria out of 20 detected in the mixture (Fig. 4b). For example, *ambiguity* for *H. pylori* using regions V1-V3 and V6-V9 was 137 and 76, respectively. However, combing regions resulted in a significantly smaller *ambiguity*, of seven *H. pylori* 16S rRNA gene sequences, which included the correct sequence. Similarly, for *E. coli*, resolution increased from an *ambiguity* of 1892 and 1705 in V1-V3 and V6-V9, respectively, to 602 for the combined solution in which the correct sequence was included. For eight other bacteria, resolution of the combined solution improved with respect to one of the regions but was inferior to the other. This is due to sequencing errors and PCR amplification biases that promote inclusion of sequences with similar amplicon or sequences amplified only in one of the regions to account for the differences in read counts between the regions (see Additional file 1 for an illustrating example). For the bacterium *Actinomyces odontolyti-cus*, *ambiguity* was equal (1) for either regions and for their combination, making it invisible on the logarithmic scale. In total, the average *ambiguity* for the V1-V3 and V6-V9 regions was 261 and 216, respectively, while when combining the two regions *ambiguity* reduced to 131 sequences per species, corresponding to a ~ 2-fold increase in resolution (Fig. 4c).

Profiling a *Drosophila* microbiome

We previously showed that toxic stress in specific tissues of fly larvae can induce developmental phenotypes that persist for multiple generations in non-exposed offspring. An exposure to a G418 toxin led to a delay in larval development, induced the expression of *neoGFP*, and modified the adult morphology [36]. The former two phenotypes persisted in subsequent generations of non-exposed offspring.
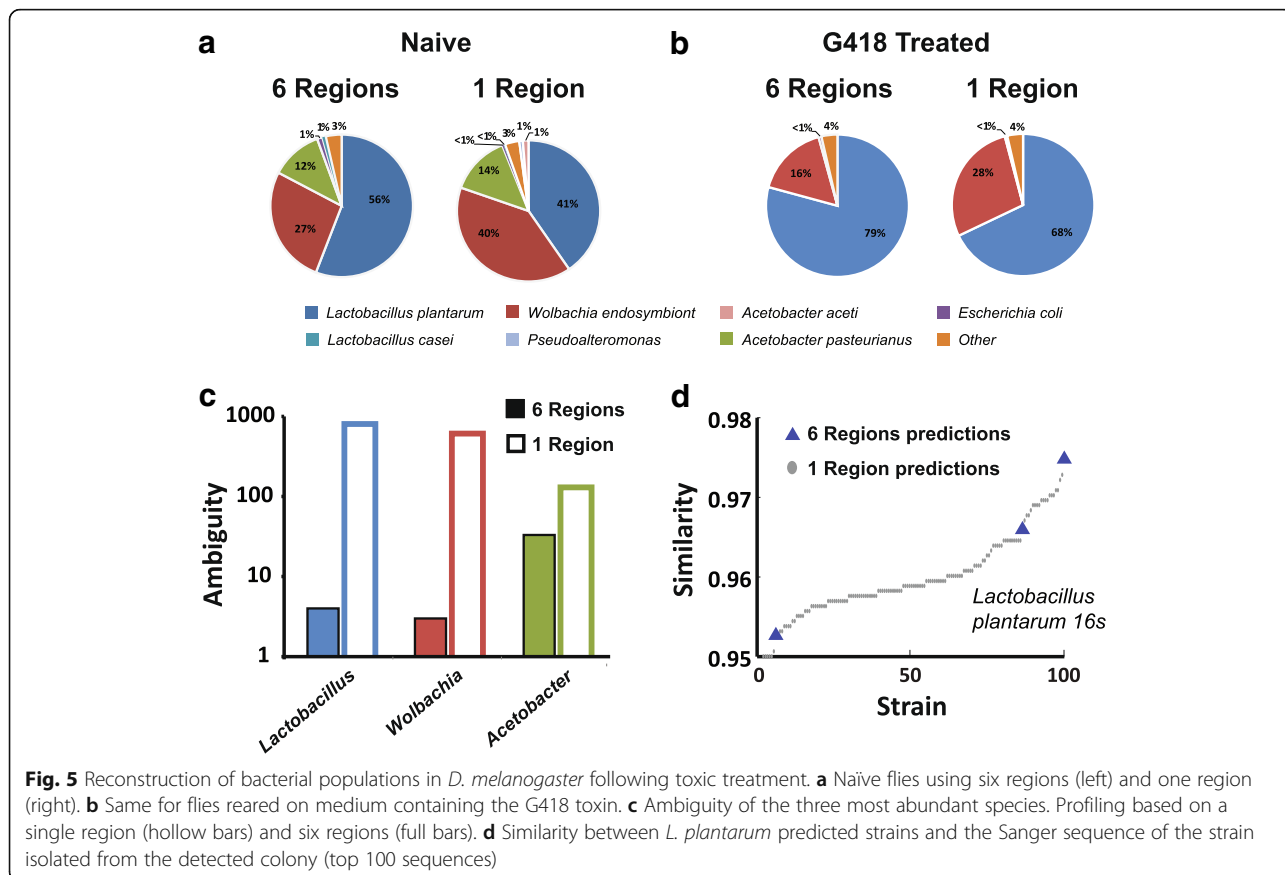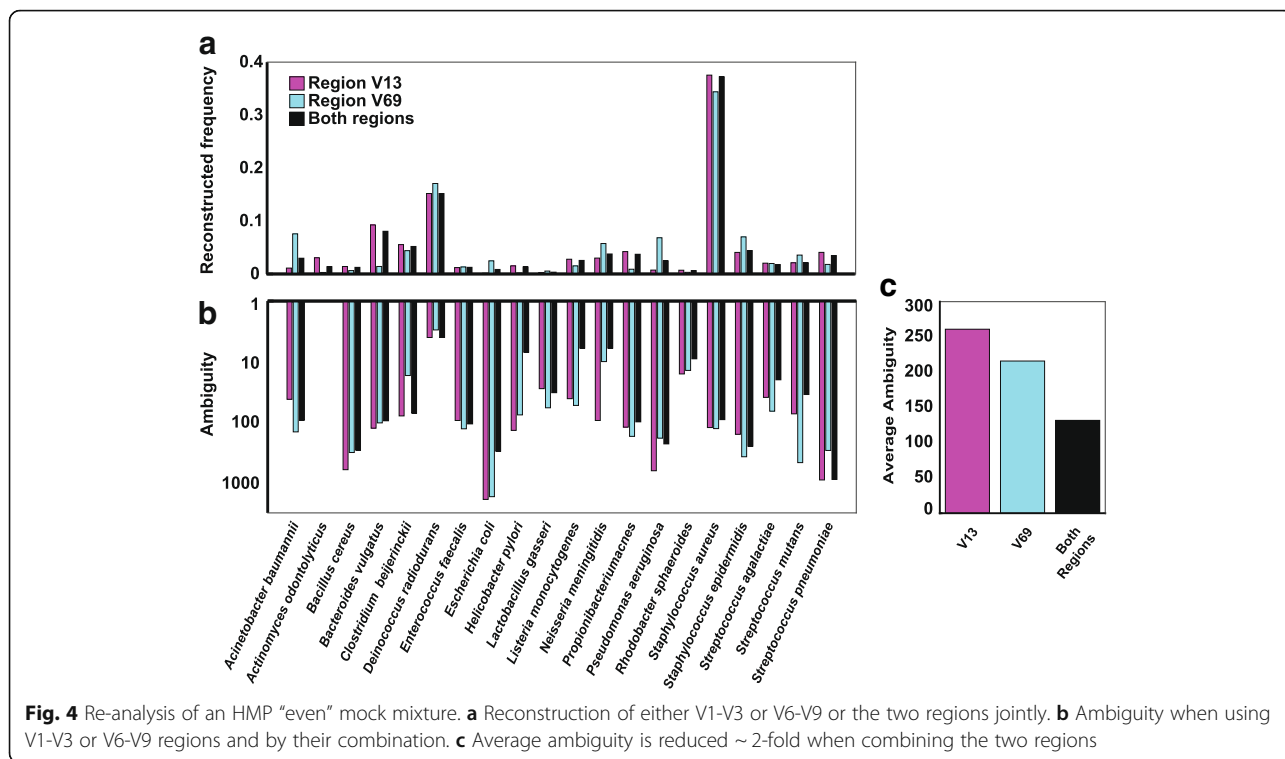
Since G418 is a broad spectrum antibiotic, this experiment sought to investigate the impact of the G418 toxin on the *Drosophila* microbial composition and its implications on the inheritance of induced responses (this section presents unpublished data from our project [37]).

Reconstruction accuracy

Profiling results using one and six regions showed roughly the same microbial composition, namely that G418 lead to a depletion of *Acetobacter* species and to an increase in the frequency of *Lactobacilli* (Fig. 5a, b and [37]).

SMURF enables highly efficient downstream analysis

To check whether the observed change in the host's microbiome played a role in the induction or inheritance of the reported phenotypes, we needed to functionally test

Fig. 4 Re-analysis of an HMP "even" mock mixture. **a** Reconstruction of either V1-V3 or V6-V9 or the two regions jointly. **b** Ambiguity when using V1-V3 or V6-V9 regions and by their combination. **c** Average ambiguity is reduced ~ 2-fold when combining the two regions

Fig. 5 Reconstruction of bacterial populations in *D. melanogaster* following toxic treatment. **a** Naïve flies using six regions (left) and one region (right). **b** Same for flies reared on medium containing the G418 toxin. **c** Ambiguity of the three most abundant species. Profiling based on a single region (hollow bars) and six regions (full bars). **d** Similarity between *L. plantarum* predicted strains and the Sanger sequence of the strain isolated from the detected colony (top 100 sequences)

whether these phenotypes were induced upon the removal of each of the strains present in naïve flies, and whether the inheritance of the phenotypes was interrupted upon reintroduction of these strains to G418 treated flies. Hence, exact strain isolation and identification were required. To achieve that, a set of 16S rRNA gene sequences predicted by SMURF (i.e., hypotheses regarding exact strain identities) was used to design strain-specific PCR/qPCR primers [37], later applied to colonies of individually isolated bacteria in order to identify the various strains. This approach was feasible due to the low *ambiguity* of SMURF's predicted strains, i.e., 4, 3, and 33 for *Lactobacillus*, *Wolbachia*, and *Acetobacter*, respectively (Fig. 5c, solid). Using a single region, there were 887 possible strains for *Lactobacillus*, 610 for *Wolbachia*, and 130 for *Acetobacter* (Fig. 5c, hollow). Hence, based on a single region, such an approach for detecting specific colonies would have been impossible since hundreds of primers should have been used. In such a case, one would have to Sanger sequence each of the hundreds of colonies which would have been slow, laborious, and quite expensive.

To assess the correctness of SMURF's predictions, we Sanger sequenced the full 16S rRNA gene of bacteria from individual colonies identified by the abovementioned PCR primers to verify their identity. As an example of one such colony, the overlap between the Sanger sequence of a strain of *Lactobacillus* and the predictions based on one and six regions is presented (Fig. 5d, only the top 100 predictions for the single region were shown). One of the sequences indeed correctly identified the strain, using both one and six regions, while the rest were wrong. Figure 5d emphasizes a tremendous advantage of the multiple region approach whose number of false positive detections was orders of magnitude smaller, thus allowing identification of the specific strain present in the sample quickly and efficiently.

Finally, we used the different isolated bacteria to functionally test the strains and to uncover two strains having an interesting biological activity [37]. The G418 toxin led to a heritable selective depletion of commensal *Acetobacter*s and explained the heritable developmental delay. Reintroducing *Acetobacter*s, detected by SMURF, prevented this inheritance.

Discussion

The two main approaches for inferring the composition of an unknown mixture of bacteria are either 16S rRNA gene sequencing or whole genome shotgun sequencing. Although sequencing whole genomes provides much more information and does not require bias-prone PCR amplification, sequencing the 16S rRNA gene has several advantages. First, most computational methods for analyzing sequencing results of either the 16S rRNA gene or of whole genome sequencing rely on a database of sequences. Although current databases of bacterial genomes are growing fast, they still contain about an order of magnitude less sequences than 16S rRNA gene databases. Consequently, a species may be identified when sequencing the 16S rRNA gene while being missed when analyzing shotgun sequencing results since its genome has not yet been deposited in the databases. A second potential advantage of 16S rRNA gene-based methods is related to profiling samples having a low bacterial biomass which makes it imperative to amplify bacteria-related sequences and renders shotgun methods highly inefficient (e.g., sequencing without prior amplification would result in an overwhelming number reads originating from the host's DNA and almost no bacterial sequences). Third, 16S rRNA gene sequencing allows for multiplexing of far more samples per lane compared to shotgun sequencing and thus yields a much lower cost per sample. Hence, we view 16S rRNA gene-based profiling as a complementary approach to whole genome sequencing, where each approach should be chosen according to the experimental setting.

In this work, we focused on 16S rRNA gene sequencing and sought to improve upon the state of the art in the field. Sequencing the 16S rRNA gene faces three main challenges that eventually hamper the accuracy of profiling and its reproducibility. The first challenge is limited resolution of current primer pairs, i.e., many bacteria share the same sequence over the amplicon, and thus, the exact identity of bacteria in the community cannot be inferred. The second challenge is "partial" primer universality, i.e., different primers pairs amplify certain subsets of all bacterial 16S rRNA gene sequences [18–20], and hence, primer choice governs profiling results. Primers may also be biased towards amplifying certain genera because of the sequence composition of the database used for their design [38]. The third challenge in 16S rRNA gene profiling is PCR-related errors and biases, which may result from several sources, e.g., primers' mismatches [35], adding sample indices (i.e., 'barcodes') to the primer sequence [21, 39], or other library preparation steps [21].

In this work, we present a novel analysis framework termed SMURF that directly addresses the first two problems via computationally combining multiple 16S rRNA gene short amplicon sequencing data and thus improving upon current best approaches. The last issue of sample preparation biases is only indirectly addressed in this work. Although combining results from different regions tends to average the effects of such biases and errors, it is certainly beneficial to carefully choose the set of primers (Additional file 1: Figure S3) and to modify library preparation procedures so as to reduce such errors and biases [21].

The set of six primer pairs used in this study should be considered a first trial that served for presenting the

advantages of combining regions. Since accurate profiling evidently depends on the properties of the chosen set of primers, it would be interesting to seek an optimal set. Also, it should be noted that accuracy and resolution are limited by the number of sequencing reads. Hence, in order to assure accurate profiling of low-frequency bacteria, a sufficient number of reads should be allocated to each region. The latter also dictates the number of samples that may be multiplexed per lane.

Conclusions

Short MUltiple Regions Framework (SMURF) represents a new approach in bacterial community reconstruction. A significant improvement in resolution is achieved as has been displayed theoretically, using in silico simulations and by analyzing two mock mixtures and real-life samples.

In many cases, microbial profiling at the genus or even species level may be achieved using standard primers (e.g., V4 primers) and does not require combining regions. However, in cases where downstream analyses are required, increasing resolution via SMURF, i.e., reducing the reported number of full-length 16S rRNA gene sequences per profiled bacterium, is crucial as demonstrated by our *D. melanogaster* experiments.

SMURF can be applied to standard sample preparation with very little modifications at the PCR stage (i.e., sample preparation is independently performed using each primer pair) and allow re-analysis of samples that were profiled using several primer pairs. Apart from increased resolution, the effects of PCR biases will tend to be "averaged" and hence be less pronounced. Also, the basic algorithm which is an adaptation of [12] allows fast convergence and avoids the necessity to align reads, and hence, the computational overhead in applying SMURF is small.

SMURF paves the way for high-resolution profiling of fragmented DNA in low biomass samples, e.g., FFPE blocks, which are abundant and relatively easy to acquire (compared to snap frozen samples); hence, SMURF may be highly beneficial in, e.g., cancer microbiome studies. Additionally, it would be beneficial to optimize a set of primer pairs so as to support a multiplexed PCR reaction. Although such optimization requires significant one-time labor, it would eventually save precious material and allow highly efficient profiling.

SMURF is not restricted to amplicons of the 16S rRNA gene and can be seamlessly adapted to combine any set of amplicons, e.g., in multilocus sequence typing (MLST). Such profiling may be required when the 16S rRNA gene is not sufficient to provide the required resolution.

Additional file

Additional file 1: Supplementary Results. (DOCX 642 kb)

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Departments of Physics of Complex Systems, Weizmann Institute of Science, 7610001 Rehovot, Israel. [2]Department of Biomolecular Sciences, Weizmann Institute of Science, 7610001 Rehovot, Israel. [3]Department of Pediatrics, University of California San Diego, La Jolla, CA 92093, USA. [4]Division of Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, 10 Karolinska Institutet, S-171 77 Stockholm, Sweden. [5]Department of Microbiology and Immunology, University of California San Francisco, San Francisco, CA 94143, USA. [6]Department of Computer Science, The Open University of Israel, 43107 Ra'anana, Israel.

References
1. Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: the unseen majority. Proc Natl Acad Sci U S A. 1998;95:6578–83.
2. Barb JJ, Oler AJ, Kim HS, Chalmers N, Wallen GR, Cashion A, Munson PJ, Ames NJ. Development of an analysis pipeline characterizing multiple hypervariable regions of 16S rRNA using mock samples. PLoS One. 2016;11(2):e0148047.
3. Suez J, Korem T, Zeevi D, Zilberman-Schapira G, Thaiss CA, Maza O, Israeli D, Zmora N, Gilad S, Weinberger A. Artificial sweeteners induce glucose intolerance by altering the gut microbiota. Nature. 2014;514:181–6.
4. Statnikov A, Alekseyenko AV, Li Z, Henaff M, Perez-Perez GI, Blaser MJ, Aliferis CF. Microbiomic signatures of psoriasis: feasibility and methodology comparison. Sci Rep. 2013;3:2620.
5. Zhang T, Shao M-F, Ye L. 454 pyrosequencing reveals bacterial diversity of activated sludge from 14 sewage treatment plants. ISME J. 2012:1137–47.
6. Fahlén A, Engstrand L, Baker BS, Powles A, Fry L. Comparison of bacterial microbiota in skin biopsies from normal and psoriatic skin. Arch Dermatol Res. 2012;304:15–22.
7. Miller CS, Baker BJ, Thomas BC, Singer SW, Banfield JF. EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. Genome Biol. 2011;12:R44.

8. Meinicke P, Asshauer KP, Lingner T. Mixture models for analysis of the taxonomic composition of metagenomes. Bioinformatics. 2011;27:1618–24.

9. Amir A, Zeisel A, Zuk O, Elgart M, Stern S, Shamir O, Turnbaugh PJ, Soen Y, Shental N. High-resolution microbial community reconstruction by integrating short reads from multiple 16S rRNA regions. Nucleic Acids Res. 2013;41(22):e205.

10. Koslicki D, Foucart S, Rosen G. Quikr: a method for rapid reconstruction of bacterial communities via compressive sensing. Bioinformatics. 2013;29:2096–102.

11. Chatterjee S, Koslicki D, Dong S, Innocenti N, Cheng L, Lan Y, Vehkaperä M, Skoglund M, Rasmussen LK, Aurell E, Corander J. SEK: Sparsity exploiting k-mer-based estimation of bacterial community composition. Bioinformatics. 2014;30(17):2423–31.

12. Jiang H, An L, Lin SM, Feng G, Qiu Y. A statistical framework for accurate taxonomic assignment of metagenomic sequencing reads. PLoS One. 2012; 7(10):e46450.

13. Weisburg WG, Barns SM, Pelletier DA, Lane DJ. 16S ribosomal DNA amplification for phylogenetic study. J Bacteriol. 1991;173:697–703.

14. Wang S, Sun B, Tu J, Lu Z. Improving the microbial community reconstruction at the genus level by multiple 16S rRNA regions. J Theor Biol. 2016;398:1–8.

15. Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF. PCR-induced sequence artifacts and bias: insights from comparison of two 16s rRNA clone libraries constructed from the same sample. Appl Environ Microbiol. 2005;71:8966–9.

16. Aird D, Chen W-S, Ross M, Connolly K, Meldrim J, Russ C, Fisher S, Jaffe D, Nusbaum C, Gnirke A. Analyzing and minimizing bias in Illumina sequencing libraries. Genome Biol. 2010;11(Suppl 1):P3.

17. Fouhy F, Clooney AG, Stanton C, Claesson MJ, Cotter PD, Shendure J, Ji H, Glenn T, Mardis E, Shokralla S, Spall J, Gibson J, Hajibabaei M, Weinstock G, Raoult D, Henrissat B, Walker A, Martin J, Scott P, Parkhill J, Flint H, Scott K, Yu Z, Morrison M, Nelson M, Morrison H, Benjamino J, Grim S, Graf J, McOrist A, et al. 16S rRNA gene sequencing of mock microbial populations—impact of DNA extraction method, primer choice and sequencing platform. BMC Microbiol. 2016;16:123.

18. Tremblay J, Singh K, Fern A, Kirton ES, He S, Woyke T, Lee J, Chen F, Dangl JL, Tringe SG. Primer and platform effects on 16S rRNA tag sequencing. Front Microbiol. 2015;6:771.

19. Gibson J, Shokralla S, Porter TM, King I, van Konynenburg S, Janzen DH, Hallwachs W, Hajibabaei M, Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, DM MG, Marsh T, Garrity GM, Tiedje JM, Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glöckner FO, Illumina, van Elsas JD, Speksnijder AJ, van Overbeek LS, et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. PLoS One. 2013;6:1–8.

20. Hiergeist A, Reischl U. Multicenter quality assessment of 16S ribosomal DNA-sequencing for microbiome analyses reveals high inter-center variability. Int J Med Microbiol. 2016;306:334–42.

21. Gohl, Daryl M., Vangay, Pajau, Garbe, John, MacLean, Allison, Hauge, Adam, Becker, Aaron, Gould, Trevor J., Clayton, Jonathan B., Johnson, Timothy J., Hunter, Ryan, Knights, Dan, Beckman KB. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. Nat Biotechnol. 2016;34:942–9. doi:https://doi.org/10.1038/nbt.3601.

22. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.

23. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.

24. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006;72:5069–72.

25. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41(D1):D590–6.

26. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB. Characterizing and measuring bias in sequence data. Genome Biol. 2013;14:R51.

27. Callahan BJ, Mcmurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13:581.

28. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R. Deblur rapidly resolves single-nucleotide community sequence patterns. mSystems. 2017;2(2):e00191–16.

29. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B. 1977;39:1–38.

30. Jost L. Entropy and diversity. Oikos. 2006:363–75.

31. Tucker CM, Cadotte MW, Carvalho SB, Davies TJ, Ferrier S, Fritz SA, Grenyer R, Helmus MR, Jin LS, Mooers AO, Pavoine S, Purschke O, Redding DW, Rosauer DF, Winter M, Mazel F. A guide to phylogenetic metrics for conservation, community ecology and macroecology. Biol Rev. 2016;92:698–715. doi:https://doi.org/10.1111/brv.12252.

32. Cole JR, Chai B, Farris RJ, Wang Q, Kulam SA, McGarrell DM, Garrity GM, Tiedje JM. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. Nucleic Acids Res. 2005;33(Issue suppl_1):D294–6.

33. Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, Ciulla D, Tabbaa D, Highlander SK, Sodergren E, Methé B, DeSantis TZ, Petrosino JF, Knight R, Birren BW. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res. 2011;21:494–504.

34. Group JCHMPDGW. Evaluation of 16S rDNA-based community profiling for human microbiome research. PLoS One. 2012;7:e39315.

35. Schloss PD, Gevers D, Westcott SL. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PLoS One. 2011;6:e27310.

36. Stern S, Fridmann-Sirkis Y, Braun E, Soen Y. Epigenetically heritable alteration of fly development in response to toxic challenge. Cell Rep. 2012; 1:528–42.

37. Fridmann-Sirkis Y, Stern S, Elgart M, Galili M, Zeisel A, Shental N, Soen Y. Delayed development induced by toxicity to the host can be inherited by a bacterial-dependent, transgenerational effect. Front Genet. 2014;5:27.

38. Mao D-P, Zhou Q, Chen C-Y, Quan Z-X. Coverage evaluation of universal bacterial primers using the metagenomic datasets. BMC Microbiology. 2012;12:66

39. O'donnell JL, Kelly RP, Lowell NC, Port JA: Indexed PCR primers induce template—specific bias in large-scale DNA sequencing studies. PLoS One. 2016;11(3):e0148698.