Large-scale 16S gene assembly using metagenomics shotgun sequences

Feng Zeng[1,*], Zicheng Wang[2], Ying Wang[1], Jizhong Zhou[3,4,5] and Ting Chen[6,7,*]

[1] Department of Automation, Xiamen University, Xiamen, Fujian 361005, China, [2] Bioinformatics Division, TNLIST and Department of Automation, Tsinghua University, Beijing 100084, China, [3] Institute for Environmental Genomics and Department of Microbiology and Plant Biology, University of Oklahoma, Norman, OK 73019, USA, [4] State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing 100084, China, [5] Earth Science Division, Lawrence Berkeley National Laboratory, Berkeley, California 94270, USA, [6] Bioinformatics Division, TNLIST, and Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China, and [7] Program in Computational Biology and Bioinformatics, University of Southern California, Los Angeles, CA 90089, USA. *To whom correspondence should be addressed.

Abstract

Motivation

Combining a 16S rRNA (16S) gene database with metagenomic shotgun sequences promises unbiased identification of known and novel microbes.

Results

To achieve this, we herein report reference-based ribosome assembly (RAMBL), a computational pipeline, which integrates taxonomic tree search and Dirichlet process clustering to reconstruct full-length 16S gene sequences from metagenomic sequencing data with high accuracy. By benchmarking against the synthetic and real shotgun sequences, we demonstrated that full-length 16S gene assemblies of RAMBL were a good proxy for known and putative microbes, including Candidate Phyla Radiation. We found that 30–40% of bacteria genera in the terrestrial and intestinal biomes have no closely related genome sequences. We also observed that RAMBL was able to generate a more accurate determination of environmental microbial diversity and yield better disease classification, suggesting that full-length 16S gene assemblies are a powerful alternative to marker gene set and 16S short reads. RAMBL first realizes the access to full-length 16S gene sequences in the near-terabase-scale metagenomic shotgun sequences, which markedly improve metagenomic data analysis and interpretation.

Availability and Implementation

RAMBL is available at https://github.com/homopolymer/RAMBL for academic use.

1 Introduction

Microbial ecology relies on 16S rRNA amplicon sequencing and whole metagenomic shotgun sequencing to explore the taxonomic and phylogenetic composition of previously unknown environmental samples (Franzosa *et al.*, 2015). Accurate determination of microbial taxa and compositional abundance from amplicon sequencing data is challenging (Zhou *et al.*, 2015). This can, in part, be attributed to the fact that PCR primers used for 16S rRNA amplicon sequencing are biased toward certain kinds of microbes and thus cannot fully capture divergent 16S rRNA gene sequences, especially the Candidate Phyla Radiation (CPR) members and uncharacterized archaea that comprise approximately 10% of environmental microbes (Eloe-Fadrosh *et al.*, 2016). In addition, the hyper-variable regions (V1–V9) of 16S rRNA gene evolve at distinct divergence rates (Chakravorty *et al.*, 2007). As a result, amplicon sequences of different hypervariable regions could not reach a consistent characterization of taxonomic composition for a microbial community.

On the other hand, whole metagenomic shotgun sequencing suffers no primer bias and possesses a full characterization for microbial community. It requires reference genome sequences and relies on phylogenetic marker gene set(s) to profile and annotate microbial taxa. For instance, mOTU established a core set of 40 marker genes extracted from 3496 prokaryotic reference genome sequences for taxonomy identification (Sunagawa *et al.*, 2013). MetaPhlan2 started from the genome sequences contributed by the early capillary sequencing and latest metagenomic shotgun sequencing efforts to construct a clade-specific marker gene set, containing one million genes, on average 184 genes per species, for identifying 7500 species (Truong *et al.*, 2015). The more the characterized prokaryotic genome sequences grow, the more archaea and bacteria the metagenomic shotgun sequences could identify. The number of prokaryotic reference genome sequences has been increased by 10-fold in the past few years. However, since metagenomic shotgun sequencing projects (such as the Human Microbiome Project, HMP (The HMP Consortium, 2012) and the Metagenomics of the Human Intestinal Tract, MetaHit (Qin *et al.*, 2010)) were mostly devoted to cataloging the microbes colonized in human body, the speed of characterizing environmental microbes lags far behind that of characterizing human associated microbes in terms of genome sequences. Moreover, more than 99% of microbes, including most underrepresented and uncultivated bacteria, such as CPR members (Sunagawa *et al.*, 2015), still have no closely related reference genome sequences (Sharon and Banfield, 2013). The bias and incompleteness of the reference genome sequence database limit full characterization of taxonomic diversity of microbial communities.

Over the past few decades, millions of 16S rRNA gene sequences have been collected by amplicon sequencing (DeSantis *et al.*, 2006) and *in silico* gene prediction (Cole *et al.*, 2014; Pruesse *et al.*, 2007). An existing strategy for taxonomic profiling involves identifying 16S rRNA sequence reads from shotgun sequencing data and then using them to query a 16S rRNA database

for taxonomic identification, e.g. Parallel-Meta (Su *et al.*, 2014). However, since most metagenomic studies adopt Illumina sequencers for data generation, short reads from 100 to 250 bp cannot differentiate many homologous gene sequences. Therefore, when combined with sequencing errors, short reads could produce an incorrect and biased estimation of taxonomic composition.

Full-length 16S rRNA gene sequences, about 1.5 kb, can delineate a full spectrum of bacteria and archaea (Singer *et al.*, 2016). Unfortunately, current sequencing platforms produce reads that are too short (Illumina), have low quality (PacBio), or have low throughput (Sanger). As a result, high-throughput sequencing of full-length 16S rRNA genes remains unavailable. A worthy alternative is the direct assembly of full-length 16S rRNA gene sequences from shotgun sequencing data. EMIRGE (Miller *et al.*, 2011) and Reago (Yuan *et al.*, 2015), representing reference-based and *de novo* assembly approaches, respectively, have previously attempted this task. EMIRGE assigned sequencing reads to the closest known reference sequences, and reconstructed 16S rRNA gene sequences of a community using a single nucleotide polymorphism (SNP) map and an iterative inference-and-realignment algorithm. The specificity of EMIRGE depends on the accurate assignment of sequencing reads to the *ab origine* 16S rRNA gene reference sequences. However, error-prone short reads and homologous 16S rRNA gene sequences make this task a grand challenge. Reago employed a hidden Markov model (HMM) profile (Nawrocki *et al.*, 2009) trained for small subunit ribosomal genes to extract 16S rRNA gene reads, and utilized overlap graph to assemble full-length 16S rRNA gene sequences. *De novo* assembly can discover new 16S rRNA gene sequences, but cannot differentiate similar strain sequences. In addition, *de novo* assembly algorithms tend to show bias toward high-abundance strains and miss low-abundance strains, restricting the sensitivity.

Therefore, we developed a new computational pipeline termed reference-based ribosome assembly (RAMBL) with markedly improved sensitivity and specificity for 16S rRNA gene assembly. Using synthetic and real benchmark datasets, we showed that RAMBL was able to reconstruct both known and novel full-length 16S rRNA gene sequences from a complex microbial community, which yield a better estimate of taxonomic composition than marker gene set and 16S rRNA short reads. In addition, RAMBL can identify both high-abundance and low-abundance microbes, such as CPR members, with high quality. RAMBL displayed a unique capacity for comprehensive characterization of microbial diversity.

2 Methods

2.1 Overview of RAMBL

Genetic heterogeneity is a predominant factor that declines the accuracy of 16S gene sequence reconstruction. At one side, heterogeneous sequences in a microbiota lower the sensitivity of 16S gene assembly. At the other side,

sequence homology can lead to falsely detect microbes that are not present in the microbiota. To resolve this issue, we proposed a divide-and-conquer approach (Fig. 1a) following the anticipation that reduces genetic heterogeneity as far as possible and avoids eliciting false discoveries in the meantime. Briefly, at the beginning, we separate the entire data into many subgroups at high rank with the aid of a well-established taxonomic tree. The underlying assumption is that taxonomic annotation is correct at high rank. Short reads of a subgroup are assumed to be originated from the same taxon or similar taxa. The subgroups are called as the seeds. To enhance signal intensity for the detection of the seeds, RAMBL merges short reads across multiple samples. Next, for each subgroup, we devise a Bayesian non-parametric model to statistically reconstruct the full-length sequences of similar 16S genes. Technical details are addressed in the following.

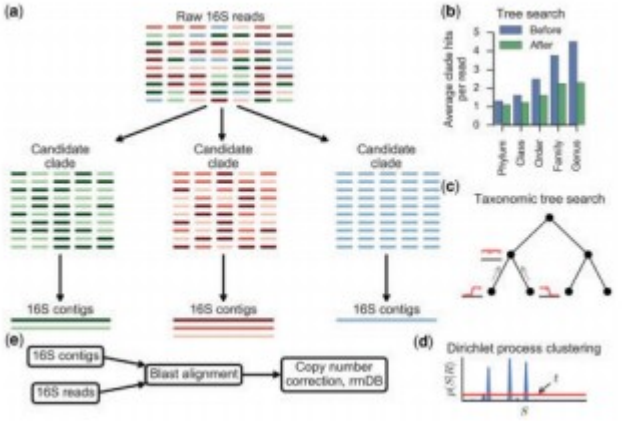

FIGURE 1 The metagenomic 16S assembly pipeline. (**a**) RAMBL consists of two components: taxonomic tree search and Dirichlet process clustering. (**b**) Clade hit analysis for the 10× × dataset of Mock1. Blue and green bars represent the number of clade hits by short reads before and after the taxonomic tree search, respectively. The taxonomic tree search significantly decreases the number of clade hits. P-values of Student's *t*-test for comparison at all taxonomic ranks are less than 0.001. (**c**) The taxonomic tree search sums up the abundance and coverage signals, which are shown as red curves next to the nodes. (**d**) Dirichlet process clustering infers the posterior distribution of the strains. The strains with proportional ratio above threshold *t* are reported. (**e**) The procedure of abundance estimation for the 16S contigs

2.2 Taxonomic tree search and data division

The taxonomic tree search aims to cluster short reads according to their origins. We observe that each short read can be mapped to different 16S rRNA gene sequences. The 10× × Mock1 dataset is an example of this (Fig. 1b). Although short reads are noticeably dispersed at the genus rank, we observe that the number of clade-hits per read gradually decreases as we move up taxonomic ranks (Fig. 1b). This phenomenon leads to a hypothesis that dispersed mapping of short reads could be grouped together if a proper high-rank taxonomic clade was specified. Therefore, we devised a taxonomic tree search algorithm to identify proper clades for the purpose of partitioning reads into subsets.

The taxonomic tree search starts from leaf nodes of a taxonomic tree, e.g. the 16S reference sequences of GreenGenes (DeSantis *et al.*, 2006) (v13.8). The leaf nodes represent 16S rRNA gene reference sequences. Given the read mapping files, we calculate the depth of a reference sequence as the average number of reads observed at a position, and the coverage as the fraction of reference sequences covered by at least one read.

Then, as the algorithm walks up to the root of the taxonomic tree, we sum up the abundance and coverage signals at each taxonomic level (Fig. 1c). The abundance of an internal node is calculated as the sum of the abundances of its offspring nodes. The union of the covered regions of the offspring nodes is used to define the coverage of the internal node. When an internal node is abundant, i.e. depth $\geq\geq 1\times\times$, and its sequence is fully covered, i.e. coverage $\geq\geq 0.9$, we call the subtree under the internal node as a candidate clade. A candidate clade is represented by a subset of short reads of the same taxonomic origin. Short reads of the clades that do not satisfy the coverage and abundance criteria are re-assigned to the closest candidate clades using Bowtie2 (Langmead and Salzberg, 2012).

2.3 The construction of alignment graph

We proposed a data structure named alignment graph to store the short reads as well as their alignments for a subgroup. Alignment graph is an extension of the partial order graph (POG) that was originally proposed to delineate the skeleton of a MSA (Lee *et al.*, 2002). Unlike POG that discards the alignments, alignment graph retains the alignment information. Thus, alignment graph is not only to represent the graph skeleton that guides the generation of all the possible strain sequences, but also to permit the fast establishment of all the MSAs against all the possible strain sequences. In the following, we first describe the procedure of graph construction, and later introduces how to generate the possible strain sequences and establish the strain-specific MSA in linear time.

We used the following procedures to construct an alignment graph for a subgroup.

First, all the reads that are assigned to the subgroup by the data division step are mapped onto the representative reference sequence of the subgroup. The reference sequence of a taxon of the subgroup is selected as the representative if it harnesses the most reads. An aligned read is represented by a 3-tuple as shown in Figure 2a. The first element of the 3-tuple indicates the start position of the alignment on the representative reference sequence. The second element of the 3-tuple is a CIGAR string (Li *et al.*, 2009) that indicates how the read is aligned to the representative exactly. The third element of the 3-tuple is the part of the read sequence that is aligned to the representative.

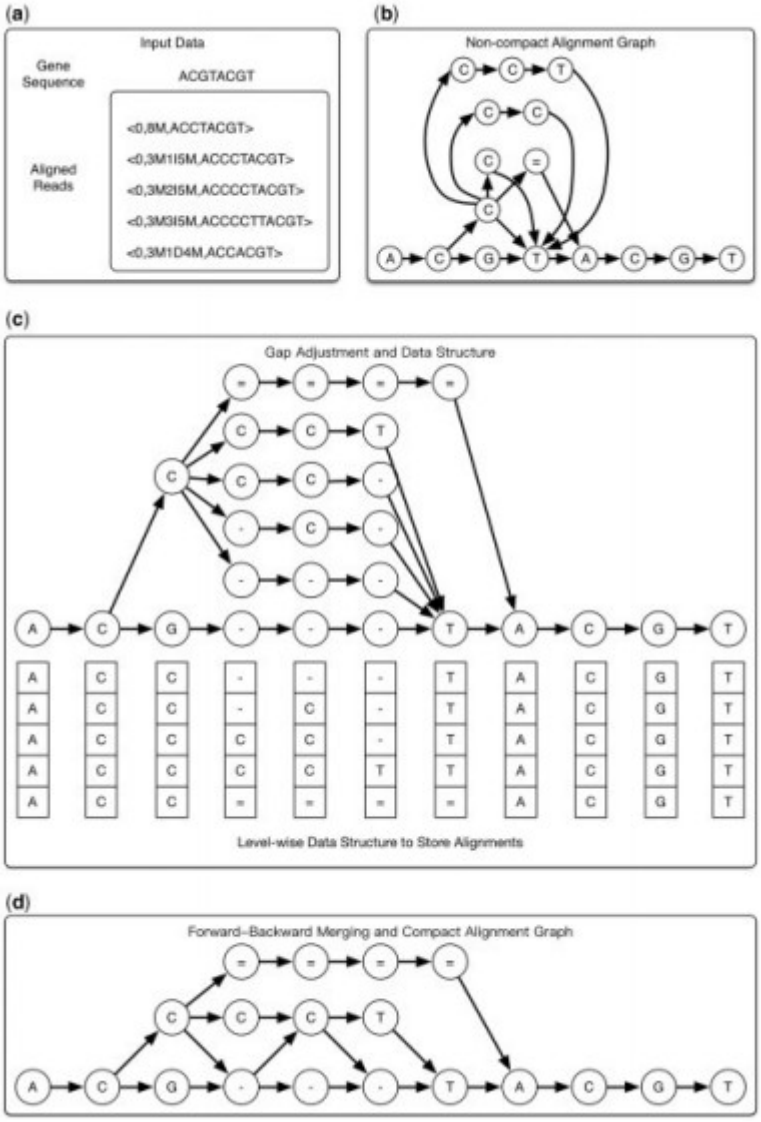

FIGURE 2 Flows to build an alignment graph. (**a**) Input data include a reference gene sequence and the aligned reads. Each read is represented as a 3-tuple, indicating the position where the alignment starts on the gene reference, the CIGAR letter that claims how the read is aligned, and the read sequence. (**b**) The first step is to build a non-compact alignment graph, where a path represents an aligned read or the reference gene sequence. The path on the bottom represents the reference gene sequence. The deletion is explicitly represented as '='. (**c**) The second step is to adjust the gap alignment. A progressive MSA method is utilized to re-compute the alignment of the sequences C, CC and CCT, which are inserted between the reference letters G and T. We explicitly add the letter '-' into the paths that represent the reference gene sequence and the reads that have no insertions at this gap position. Instead of discarding the alignments, we use a level-wise data structure to store the read letters that cover a graph position. (**d**) The third step is to remove the redundant nodes and edges to obtain a compact graph representation. In a forward-and-backward manner, it first walks through the graph from left to right, one level by one level and merges the nodes at a level that have the same letter and the same preceding node. Next, it reverses the direction, walks through the graph from right to left, one level by one level and repeats the node-merging process.

Second, we create an initial graph using the representative reference sequence, where there is only one path and each node on the path

represents a letter of the representative reference sequence. The path on the bottom line of Figure 2b represents the representative reference sequence. After that, we consider the aligned reads one by one. We scan an aligned read from left to right. We add a node for each letter of the aligned read to the graph. Eventually, this will result in a redundant and non-compact graph, where there are multiple nodes in one position that equally represents the same alignment event, like the letter ‘C’ in the first three lines in Figure 2b. We call this graph as the non-compact graph. On the resultant graph, we explicitly represent the deletion nodes in the graph using the letter ‘=’. Other nodes are called as non-deletion nodes.

Third, we adjust the gap alignment. The mapping program that does not implement a full version of dynamic programming and considers the alignment of short reads one by one cannot result in a consistent alignment for the reads that encounter the insertions and/or deletions (indels). In addition, the random indels that happen during sequencing will confuse the alignment of a gap. As the example in Figure 2b, there are three inserted sequences between the letters G and T on the representative sequence, which are C, CC and CCT. Their alignments are inconsistent. Thus, we pursuit to establish a consistent alignment for the gap. For the insertion, we apply the progressive MSA algorithm to re-align the inserted sequences. A consistent alignment for the sequences inserted between the reference letters G and T is shown in Figure 2c. After the adjustment of the insertion, we re-calculate the graph level for all the non-deletion nodes. Next, we adjust the deletion by simply adding a new deletion to replace the old one. The number of nodes on the new deletion equals to the difference between the level orders of the deletion-starting and deletion-ending nodes. To keep a consistent graph representation, we explicitly represent the insertions on the representative reference sequence.

Although the graph is non-compact, the nodes of the graph can be indexed using a horizontal–vertical coordinator. The horizontal axis indicates the graph level, which is a counterpart of the level counting on a tree. Particularly, the first column is level 1, the second column is level 2 and so on. The vertical axis indicates the order of the variants at a graph level. For example, the letter ‘G’ at the graph level 3 has a horizontal–vertical coordinate of (3,0), where zero indicates that it is a reference allele, and the letter ‘C’ on the top of ‘G’ has a horizontal-vertical coordinate of (3,1).

We introduce an auxiliary data structure $Reads[l]$ to store the aligned reads at the level l by collecting the reads covering the level ll.

Finally, we use a forward–backward algorithm to reduce the redundant nodes to obtain a compact graph. In the forward direction, we traverse the graph from left to right, visit the nodes one level by one level, and merge the nodes at a level having the same letter. After that, we repeat this process but conduct in the reverse direction. Eventually, a compact graph like the one depicted in Figure 2d is obtained.

Alignment graph and the conventional POG are different in two aspects. First, alignment graph explicitly represents the indels, whereas POG adopts an implicit representation for the indels. Second, alignment graph involves an external data structure to store the alignment information. POG discards the alignments.

2.4 The manipulation of alignment graph

During the probabilistic inference described at next section, we use the alignment graph for two purposes. First, we use the alignment graph to generate possible strain sequences. Second, we use the alignment graph to fast establish a MSA of all reads against a possible strain sequence.

Strain sequence generation. We traverse an alignment graph from left to right, and start with a possible strain sequence that accounts for the starting node of the graph. By visiting nodes one level by one level, we can extend the strain sequence by padding the node letter at the next level, when there is no branch. Otherwise, we can expand the set of strain sequences by adding new strain sequences where the variant node letters are padded on the right end of the strain sequences till the previous level per graph branching. Figure 3a depicts all the possible strain sequences at all the graph levels of an exemplar alignment graph. Without the aid of alignment graph, the enumeration of all the possible strain sequences grows exponentially in terms of the number of mutations and indels. The graph-aided generation of strain sequences substantially reduces the enumeration complexity. As shown in Figure 3a, the number of all the possible strains is $1 \times 2 \times 1 \times 3 \times 3 \times 1 = 18$, when all positions are assumed to be mutually independent. However, the use of alignment graph declines the number by 10. This improvement will be noticeable when the graph is complex.

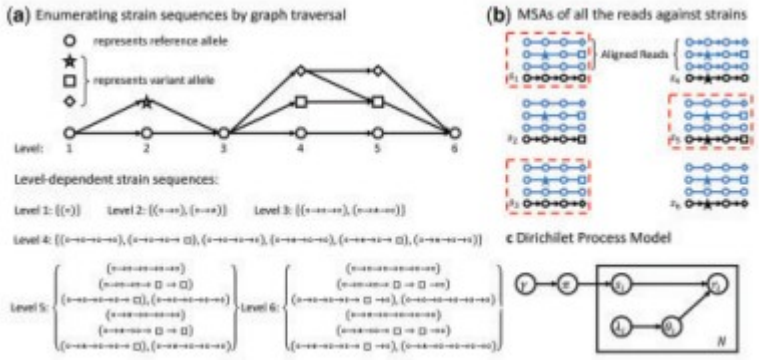

FIGURE 3 Operation on alignment graph. (**a**) Alignment graph is to yield the candidate strain sequences. The graph-aided enumeration avoids the exponential complexity. (**b**) Alignment graph is also to establish a MSA of all the reads against a candidate strain sequence through the constant-time retrieval from the level-wise data structure. All the MSAs of all the candidate strain sequences till the level 4 are listed. (**c**) The Dirichlet process model is to infer the underlying strain sequences that likely yield all the reads till the current level, such as red dashed boxes in (**b**).

Fast MSA establishment. For each possible strain sequence, we can establish the alignments of all the reads against it in linear time. This is achieved by

the constant time retrieval of the alignment parts starting from *Reads*[*1*] to *Reads*[*l*] , where *l* is the current level. Figure 3b depicts the MSAs of all the possible strain sequences at the level 4. This avoids the intensive computation of new MSAs *ab initio*.

2.5 Dirichlet process clustering

Dirichlet process clustering is a machine learning method for the automatic inference of community composition and does not require prior knowledge about the structure of mixture data. It has been widely used in quasi-species sequence assembly for viral population sequencing data (Töpfer *et al.*, 2013) and clonal reconstruction for cancer genomics (Fischer *et al.*, 2014). Therefore, we applied this technique to reconstruct 16S gene sequences for the taxonomic tree-defined read sets.

Suppose there are *N* short reads, $R = \{r_i\}_{i=1}^{N}$, within a read set. These reads are derived from *K* different strains, $S = \{s_j\}_{j=1}^{K}$. We devise a full probabilistic sequencing model, $\Theta_i = \left\{\theta_{\alpha,\beta}^{i}\right\},\ \alpha,\beta \in \mathcal{A} = \{A, C, G, T, -\},$ for a strain s_j. It formulates the single nucleotide polymorphisms (SNPs), insertions and deletions (indels), which occur during the sequencing of strain s_j. The model parameter $\theta_{\alpha,\beta}^{i}$ represents the probability that a letter α on s_i emits an observation β. The probability of all emission events of α is equal to 1, $\sum_{\beta\in\mathcal{A}}\theta_{\alpha,\beta}^{i} = 1.$

With the above configuration, we can describe the generation of short reads as follows. To generate a short read r_i, a strain s_j is first drawn from the candidate strain set *S*. The distribution over *S* is a multinomial probabilistic model with parameter $\pi = (\pi_1, \pi_2, \ldots, \pi_K), \sum_{j=1}^{K}\pi_j = 1.$ The prior distribution over parameter π is a Dirichlet distribution, *Dirichlet*(γ) , γ=(γ_1,γ_2,...,γ_K) , where γ_j>0 is the pseudocount of the *j*th strain s_j. Next, the sequencer proceeds from the left end to the right end and emits one letter at one position. Suppose the strain letter α is at the current position; then the sequencing model, $\left\{\theta_{\alpha,\beta}^{i}\right\},\ \beta \in \mathcal{A},$ would be specified by drawing from a Dirichlet prior distribution, $Dirichlet(\lambda_{\alpha}^{i})$ where $\lambda_{\alpha}^{i} = \left(\lambda_{\alpha,A}^{i}, \lambda_{\alpha,C}^{i}, \lambda_{\alpha,G}^{i}, \lambda_{\alpha,T}^{i}, \lambda_{\alpha,-}^{i}\right)$ is the pseudocounts of the emission events. Given the strain and sequencing model, a letter β on read r_i is drawn from a multinomial distribution with the probabilistic parameters $\left\{\theta_{\alpha,\beta}^{i}\right\}$. Taken together, the Dirichlet process generation is depicted in Figure 3c and formulated as

$$\pi \sim Dirichlet(\gamma)$$

$$s_j \sim Multinomial(\pi)$$

$$\theta_{\alpha}^{j} \sim Dirichlet(\lambda_{\alpha}^{j})$$

$$\beta \sim Multinomial(\theta_{\alpha}^{j})$$

In this framework, the probability of a read r_i belonging to a strain s_j can be explicitly written by,

$$p(r_i|s_j,\Theta_j) = \prod_{\alpha,\beta\in\mathcal{A}} \theta_{\alpha\beta}^{j,n_{\alpha\beta}}$$

where $n_{\alpha\beta}$ counts the number of the emission event $\alpha\rightarrow\beta$ that occurs during the generation of r_i.

The joint probability of a read r_i belonging to the strain s_j and the parameter Θj can be obtained by

$$p(r_i,\Theta_j|s_j) = \prod_{\alpha\in\mathcal{A}} \frac{\Gamma\left(\sum_{\beta\in\mathcal{A}}\lambda_{\alpha\beta}^{j}\right)}{\prod_{\beta\in\mathcal{A}}\Gamma\left(\lambda_{\alpha\beta}^{j}\right)} \prod_{\beta\in\mathcal{A}} \theta_{\alpha\beta}^{j,n_{\alpha\beta}+\lambda_{\alpha\beta}^{j}-1}$$

By integrating out the parameter Θj, the marginal probability of the strain s_j generating r_i is,

$$\mathrm{p}(r_i|s_j) = \prod_{\alpha\in\mathcal{A}} \frac{\Gamma\left(\sum_{\beta\in\mathcal{A}}\lambda_{\alpha\beta}^{j}\right)}{\prod_{\beta\in\mathcal{A}}\Gamma\left(\lambda_{\alpha\beta}^{j}\right)} \frac{\prod_{\beta\in\mathcal{A}}\Gamma\left(\lambda_{\alpha\beta}^{j}+n_{\alpha\beta}\right)}{\Gamma\left(\sum_{\beta\in\mathcal{A}}\lambda_{\alpha\beta}^{j}+\sum_{\beta\in\mathcal{A}}n_{\alpha\beta}\right)}$$

The posterior probability of Θj, given that r_i has been assigned to s_j, can be explicitly written by

$$\begin{aligned}\mathrm{p}(\Theta_j r_i, s_j) &\propto p(r_i s_j,\Theta_j)p(\Theta_j)\\ &\propto \prod_{\alpha\in\mathcal{A}} \mathrm{Dirichlet}(\lambda_{\alpha}^{j}+n_{\alpha}^{j})\end{aligned}$$

RAMBL uses Gibbs sampling to estimate π and Θ with the aim of maximizing the posterior probability of the strains. In the Gibbs sampling, the conditional posterior probability of the assignment of a read r_i is given by

$$\mathrm{p}(s_j r_1,\ldots,r_{i-1}) = \begin{cases} \dfrac{m_j}{m_j-1+\tau} p(r_i s_j,\Theta_j), & \text{if } s_j \text{ has been populated} \\ \dfrac{\tau}{m_j-1+\tau}\sum_{s_j} p(r_i s_j)p(s_j), & \text{if } s_j \text{ is new strain} \end{cases}$$

where m_j is the number of previous reads that has been assigned to the strain s_j. The update of the parameter Θj is according to the posterior probability that described above.

2.6 Progressive inference

Once an alignment graph for a subgroup is built, we scan the graph from left to right, and conduct the probabilistic inference one level by one level. We call this way that we perform the inference as the progressive Dirichlet process.

Specifically, at the beginning, the candidate strain set has only one element comprising one letter that represents the starting node of the alignment graph under exploitation. The inference at this circumstance is trivial. We move on to the next level of the alignment graph. If the graph branches to indicate variants, we will involve additional strain sequences to expand the candidate set. A new strain sequence is constructed by padding the variant letter of a branching node v to the right end of the old strain sequence of the node u preceding the branch. The weight of the new strain sequence is computed using $\gamma_v = \frac{N_v}{N_u}\gamma_u$, where γ_u is the weight of the old strain sequence, N_u is the number of reads covering u, and N_v is the number of reads covering v. The initial K at current level is specified as the K^* inferred at the previous level. Then, we execute the Dirichlet process clustering to infer the posterior probability for the strain sequences until the current level. After the inference, a strain is discarded if its posterior probability is found to be less than 0.05. In this case, its abundance would be less than 5% within the subgroup. The posterior probability of the strains at the current level is used as the prior probability at the next level. The procedure proceeds toward the right end of the alignment graph. When RAMBL reaches the right end, it outputs the reconstructed 16S gene sequences. Two assembled contigs are merged if sequence identity is higher than 98%.

2.7 Compositional abundance estimation and taxonomic classification

We align short reads to the 16S contigs in order to calculate compositional abundance. The MegaBlast program (Zhang *et al.*, 2000) (v2.2.29) is utilized to search for alignments between short reads and 16S contigs. An alignment is discarded when its identity is less than 0.95 or the *E*-value is higher than 1e−10. After that, each read is assigned to the best-hit contig having the lowest *E*-value. If more than one best-hit contig exists for a read, the read is assigned to best-hit contigs with weight equal to the inverse of the number of the best-hit contigs. Thus, if L equals contig length and l equals read length, then the raw abundance without fixing copy number variation of the contig that has n assigned short reads can be defined as $\frac{n \times l}{L}$.

We use the RDP classifier (Wang *et al.*, 2007) (v2.11) to determine the taxonomic identity of a contig. A taxonomic identification is considered

unreliable and filtered out if the RDP score, i.e. the posterior classification probability, is less than 0.6.

We obtained the copy number of a contig by querying the rrnDB database (Klappenbach, 2001) (v4.4.4) through the RDP-defined taxonomic identity. Suppose the copy number of a contig is c. We refined the raw abundance by dividing the copy number and get the copy number corrected abundance as $\frac{n \times l}{c \times L}$. The abundance of a clade was calculated by summing over the abundances of the contigs within the clade.

2.8 Data analysis for simulation metagenomes

We used two simulation communities, Mock1 and Mock2, for validation. Mock1 consisted of 100 strain sequences (Supplementary Table S1) and three datasets of the simulated Illumina paired-end reads. The three datasets represented coverages of $10\times$, $20\times$ and $30\times$, respectively. Mock2 contained 22 known strain sequences and 2 spiked strain sequences that represented unknown microorganisms (Supplementary Table S2).

We used the $10\times$ dataset of Mock1 to analyze the clade hits of short reads. We used Bowtie2 (Langmead and Salzberg, 2012) (v2.2.4, with local option) to map short reads of the $10\times$ dataset to the 16S reference sequences of GreenGenes. We counted the number of clade hits by short reads derived from the same strain. After the taxonomic tree search, we remapped short reads to the representatives of the candidate clades. We counted the number of clade hits again for comparison. The comparison results were summarized in Figure 1b and evaluated by Student's *t*-test.

To evaluate the recovery of strain sequences, we used Blast to align 16S contigs to the mock strain reference sequences. We discarded alignments whose sequence identities were below 95%. A strain was considered to be recovered if a 16S contig hit the strain with sequence identity above 95%. Therefore, we defined the sensitivity of strain recovery as the fraction of the mock strain sequences that could be found by 16S contigs. We also defined the specificity of strain recovery as the fraction of 16S contigs that could be used to recover the mock strain sequences. The F_1 measurement, $F_1 = \frac{2 \times \text{sensitivity} \times \text{specificity}}{\text{sensitivity} + \text{specificity}}$, was used to summarize the sensitivity and specificity.

To evaluate the recovery of microbial taxa, we used the RDP classifier to determine the taxonomic identities of the 16S contigs. Sensitivity was determined as the fraction of the mock taxa that could be recovered by the 16S contigs. Specificity was estimated by the fraction of taxa in the mock community defined by the 16S contigs.

2.9 Metagenomic data of soil biomes and Chinese gut

Raw sequencing reads of the 16 soil biomes were downloaded from the MG-RAST server: hot deserts (4477805.3, 4477872.3 and 4477873.3), cold

deserts (4477803.3, 4477900.3 to 4477904.3) and green soils (4477804.3, 4477807.3, 4477874.3 to 4477877.3 and 4477899.3). On average, there were 7 042 164 100 bp Illumina reads per sample. It was noted that only first segments of paired-end reads were available on the MG-RAST server. A total of 2 289 307 895 Illumina paired-end reads of the Chinese gut microbiota were downloaded from the Short Read Archive (SRA, accession number SRA045646) of the NCBI server. Each segment of a paired-end read was 75 bp at length.

2.10 Data analysis for cross-biome soil metagenomes

We ran RAMBL, EMIRGE, MetaPhlan2 and Parallel-Meta with the default parameters. We did not test Reago or mOTU because they failed to process the soil metagenomic datasets.

To perform the principal coordinate analysis (PCoA), we ran MUSCLE (Edgar, 2004) (v3.8.31, using default parameters) to compute the multiple sequence alignment (MSA) for the 16S contigs. Next, we input the MSA to FastTree (Price *et al.*, 2009) (v2.1.9, using default parameters) and built the taxonomic tree for the 16S contigs. Following that, we used the beta_diversity.py script (with the -m weighted_unifrac option) of the QIIME package (Caporaso *et al.*, 2010) to compute the weighted UniFrac distance matrix. With the distance matrix, we used the principal_coordinates.py script of QIIME to perform PCoA analysis. For MetaPhlan2 and Parallel-Meta, we ran beta_diversity.py with the -m bray_curtis option to compute the Bray–Curtis distance matrix because these two methods did not provide the representative sequences of the identified taxa for taxonomic tree construction.

2.11 T2D classification

We used the scikit-learn package (http://scikit-learn.org) to carry out the analysis. The support vector machine-recursive feature extraction (SVM-RFE) method was used to select the discriminative 16S gene assemblies. The linear SVM classifier was used for the T2D classification. We partitioned the samples into 10 folds to perform the cross validation. Among the partitioned samples, 7 folds were used to train the SVM classifier, and 3 folds were used to test the classification accuracy. The cross validation procedure was replicated 1000 times.

2.12 Computational resource

We conducted all the experiments using one node of the Tsinghua BigData cluster, where the CPU is Intel® Xeon® E5-2680 and memory size is 512GB. The CPU clocks of RAMBL on the 16S gene reconstruction of the soil and T2D microbiome were 18 and 389 min, respectively, using 20 threads.

3 Results

3.1 Improvement of full-length 16S gene assembly

To assess the accuracy of RAMBL, we created a mock community consisting of 100 strain sequences (termed Mock1, Supplementary Table S1). The strain abundance levels were simulated through a stick-breaking process (Paisley *et al.*, 2010), producing values ranging from 0.02% to 5.79%. We used a simulator called Mason (Holtgrewe, 2010) to generate three Illumina paired-end datasets with read length of 100 bp, insert size of 300 ± 30 30 bp (mean ± s.d.) and mean sequencing depths of 10×, 20× and 30× (Supplementary Table S1) for evaluation.

The average sequence identity of reconstructed 16S rRNA assemblies for the 10× × simulated dataset was 99.5%, compared to the ground truth reference sequences (Supplementary Fig. S1a). These results indicated that RAMBL could reconstruct 16S rRNA gene sequences nearly identical to the reference sequences which is a significant improvement over that of EMIRGE (v0.60, average sequence identity of 98.5%) and Reago (v1.1, average sequence identity of 99.1%). In addition, the lengths of the reconstructed 16S rRNA assemblies by RAMBL were closer to 1.5 kb with median deviation of 5 bp (Supplementary Fig. S1b), while those by EMIRGE were shorter, with median deviation of −12 bp, and those by Reago were longer, with median deviation of 124 bp. For all three simulated datasets, RAMBL outperformed the other two methods, as measured by F1F1 (Supplementary Fig. S1c), with higher sensitivity and accuracy. RAMBL could also accurately identify low-abundance taxa. For the 10× dataset, RAMBL could recover all simulated phyla, whereas EMIRGE and Reago missed many microbes with abundance below 0.01 (Supplementary Fig. S1e). The speed of RAMBL is more than 300 assemblies per hour (Supplementary Fig. S1d).

3.2 Accurate identification of both known and novel microbes

RAMBL can detect sequences from potentially novel microorganisms. To verify this, we created another mock community (termed Mock2, Supplementary Table S2) comprised of 22 known microbial genome sequences and 2 spike-in (novel) microbial strains, *Aminiphilus circumscriptus* DSM 16581 (GenBank: GCA_000526375.1) and *Arsenophonus endosymbiont* str. Hangzhou of *Nilaparvata lugens* (GenBank: GCA_000757905.1). The Illumina paired-end reads were simulated using Mason with an even sequencing depth (15×) for the 24 strains. We evaluated three 16S rRNA reconstruction methods, including RAMBL, EMIRGE and Reago, along with two genome query methods, including MetaPhlan2 (v2.2.0) and mOTU (v1.3), and one 16S rRNA read-based method, Parallel-Meta (v3.1.0). All programs were run using default parameters.

The 16S rRNA reconstruction methods demonstrated distinct features from those of genome query methods (Supplementary Fig. S1f and g). For example, both EMIRGE and Reago succeeded in identifying all of the spiked-in microbes, but failed to identify a number of known microbes; the sensitivity of EMIRGE was 71.4%, while that of Reago was 76.2%. On the other hand, MetaPhlan2 and mOTU, both genome query methods, recovered

all the known microbes, but missed the spiked-in microbes. In comparison, RAMBL detected all microorganisms correctly and estimated microbial abundance accurately with cosine similarity to the ground truth at 97.9% (Supplementary Fig. S1g). By assembling 16S rRNA reads to full-length gene sequences, we can significantly improve specificity, as demonstrated in Supplementary Figure S1g.

3.3 Reanalysis of 16 soil metagenomes

We applied RAMBL to reanalyze the soil metagenomic data collected from 16 spatial locations, from Antarctica to Argentina, representing the typical desert and green (nondesert) soil biomes of the Americas (Fierer *et al.*, 2012). RAMBL assembled 192 16S gene contigs (Supplementary Table S3), and most of the contigs (86.5%) were assigned to 104 microbial genera with high confidence levels (RDP score > 0.8, Supplementary Fig. S2). Using ChimeraSlayer (Haas *et al.*, 2011), we found that two contigs were chimeric assemblies, which correspond to a low chimera rate of ~1%. The assembled contigs were highly dissimilar and captured diverse microbial species/strains for high-resolution taxonomic profiling (99.7% of intra-genus sequence similarities were less than 0.98, Supplementary Fig. S3). RAMBL recovered more accurate soil microbiota with the assembled 16S contigs; 12 out of the 17 phyla found by RAMBL were confirmed by previous 16S amplicon sequencing data (Fig. 4a). For comparison, MetaPhlan2 missed half of these 12 phyla; EMIRGE failed to recover Verrucomicrobia, Planctomycetes and Nitrospirae; Parallel-Meta did not detect Thaumarchaeota. RAMBL did not report any false positive for the non-CPR phyla, while all other methods exhibited high false positive rates (MetaPhlan2 14.3%, EMIRGE 30.8%, Parallel-Meta 26.7%). The other five phyla identified by RAMBL were the CPR members that were not found by the 16S amplicon sequences (Supplementary Fig. S4). Although the CPR phyla were present at low abundance, one of them, Armatimonadetes (formerly OP10), exhibited a positive correlation with nitrogen content of desert soils (Supplementary Fig. S5) and contributed to the separation of desert and green soil biomes (Supplementary Fig. 6), which was not reported previously.

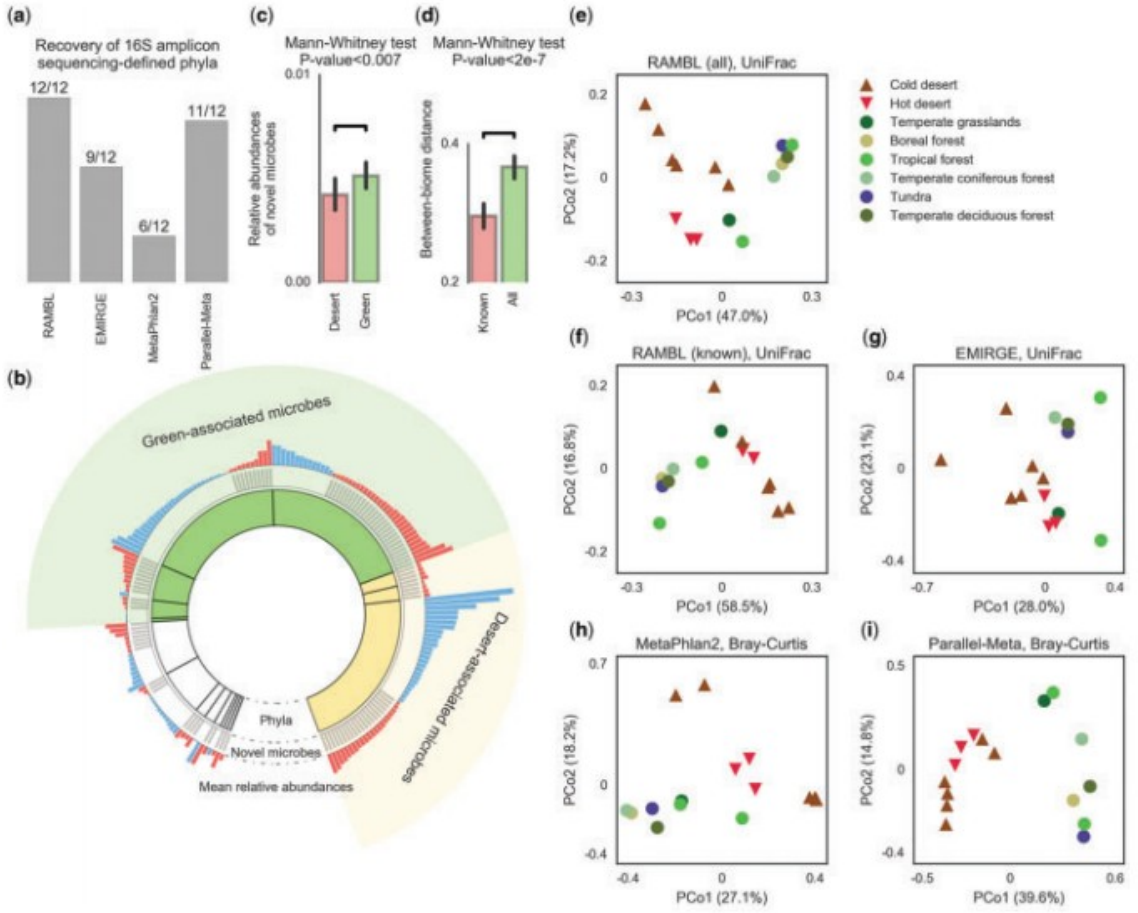

FIGURE 4 RAMBL yields better results for the soil metagenomes. (**a**) Taxonomic profiling results of the four methods in comparison with previous 16S amplicon sequencing. (**b**) The distribution and abundance of novel microbes. An extensive number of assembled contigs highlighted in gray (middle circle) were not aligned to close representatives in the NCBI database, representing novel microbes. Novel microbes were in high abundance (outer circle, highlighted in red), and mainly distributed among the desert-associated phyla (inner circle, highlighted in yellow) and green-associated phyla (inner circle, highlighted in green). In anticlockwise direction, the desert-associated phyla in anticlockwise direction were Actinobacteria, Chloroflexi and Armatimonadetes; the green-associated phyla were Acidobacteria, Proteobacteria, Verrucomicrobia, Planctomycetes and Firmicutes. (**c**) The comparison of novel microbes in the abundance between the desert and green soil biomes. (**d**) The comparison of the between-biome distances (weighed UniFrac measurement) based on known microbes and all microbes. (**e**) The principal coordinates analysis (PCoA) plot of RAMBL based on all microbes. (**f**) The PCoA plot of RAMBL based on known microbes. (**g–i**) The PCoA plots of EMIRGE, MetaPhlan2 and Parallel-Meta

By aligning the RAMBL assemblies against representative microbial genomes in the National Center for Biotechnology Information (NCBI) database, we found that 99 16S gene contigs among the assemblies, at an abundance of 45.7 ± 13.9% (mean ± SD), had no close representatives in the NCBI database (sequence identity > 90%), indicating novel microbes. These novel contigs were distributed among 15 phyla, excluding Firmicutes and Nitrospirae. Among 104 genera that were found, the novel contigs were from 42 genera (40.4%), of which 41 (97.6%) were completely novel, displaying no closely related genomes in the NCBI database. As shown in Figure 4a,

MetaPhlan2 achieved a high false negative rate of 50%; it was attributed to that 66.1% of the contigs of the missing phyla were novel, and that the rest of the contigs of the missing phyla showed a low sequence identity (averagely 92%) to the closest representative microbial genomes (Supplementary Table S3).

Among the aforementioned 99 novel contigs, 80 (80.8%) belonged to the phyla relevant to the classification of desert and green soil biomes, i.e. Acidobacteria and Actinobacteria (Fig. 4b and Supplementary Fig. S6). The non-desert-associated novel genera comprised 51.5% of the uncharacterized microbial community, and they were more abundant than the desert-associated novel genera (Fig. 4c), demonstrating a high prevalence of underexplored microbes in green soil biomes. Many novel microbes in the fertile and moist green soil samples, such as the Acidobacteria genera Gp2 and Gp5 and the Proteobacteria genus Pedomicrobium, had a preference for high organic carbon and nitrogen contents (Supplementary Figs S7 and S8). But in the dry and unproductive desert soils, novel microbes favored high pH levels, including the Actinobacteria genus Iamia and the Chloroflexi genus Sphaerobacter (Supplementary Fig. S8). These novel microbes were important to the characterization of the inter-biome variability. When only known microbes were considered, the sample distance between different biomes decreased significantly from 0.367 ± 0.075 (mean $\pm$ SD) to 0.296 ± 0.082 (mean $\pm$ SD), as shown in Figure 4d. In particular, the novel microbes contributed the most to the separation of the hot and cold desert soil samples (Fig. 4e and f). Although the associated ecological functionalities remained unknown, novel microbes that were found in the terrestrial samples were extensive, and substantially accounted for the underexplored environmental diversity.

RAMBL also yields more accurate relative abundance estimations for the environmental microbes, which are key to the correct interpretation of metagenomic data. The results of ββ-diversity analysis of all 16 soil samples based on RAMBL, EMIRGE, MetaPhlan2 and Parallel-Meta were compared, as shown in Figure 4e and g–i. The results obtained by RAMBL give better separation of the three different biomes than that achieved by the other methods. In addition, we observe a higher percentage of explanation of the data variations at the first two PCoA components by RAMBL than that by the other methods. Accurate abundance estimation, combined with full taxonomic profiling, makes RAMBL the most accurate tool for diversity estimation and community profiling of environmental samples.

3.4 Reanalysis of 145 human gut metagenomes

We applied RAMBL to characterize the gut microbiota in a Chinese population, which consists of 74 healthy individuals and 71 T2D individuals (Qin *et al.*, 2012). The 896 diverse 16S gene sequences were reconstructed from the metagenomic datasets (with an average intra-genus sequence identity of 0.893, Supplementary Fig. S9). The chimera rate was 13.7%

determined by ChimeraSlayer. Based on these recovered 16S rRNA gene sequences, RAMBL revealed an uncharted microbial community in the gut of Chinese individuals (Fig. 5a and Supplementary Table S4). The gut microbiota was composed of 11 bacterial and archaeal phyla, including 91.3 ± 9.7% (mean ± SD) Firmicutes and Bacteroidetes, and 148 genera (with an average RDP score of 0.94, Supplementary Fig. S10). Novel microbes were abundant in the Chinese gut (18.7 ± 7.6%, mean ± SD), and represented 49 genera (33.1%), of which 37 genera (75.5%) had no close related genome sequences in NCBI database. We found 27 genera significantly correlated with the health and diabetes (Spearman correlation analysis followed by Benjamini–Hochberg correction for multiple tests, *P*-value < 0.05). As shown in Figure 5b, 10 genera (37%) were novel, which was defined as the fraction of novel contigs within a genus. These novel genera included seven health-associated genera Oribacterium, Faecalibacterium, Butyricicoccus, Prevotella, Lachnospiracea_incertae_sedis, Clostridium XIVa and XIVb and three diabetes-associated genera Anaerovibrio, Anaerovorax and Erysipelotrichacaea_incertae_sedis. The health-associated novel genera are known to supply methane, acetate and butyrate (Vital *et al.*, 2015), which play a crucial role in the increasing of insulin sensitivity and the fermentation of polysaccharide and fatty-acid-producing sugar. These novel genera are important for maintaining host health. Two diabetes-associated novel genera, Anaerovibrio and Anaerovorax, are putative bacteria fermenting glycerol, taurine, glucose and putrescine, which are elevated in diabetic individuals. Our finding that Erysipelotrichacaea_incertae_sedis was increased in diabetic individuals was also consistent with the previous observation of this bacterial genus having close relevance to metabolic disorders (Kaakoush, 2015).

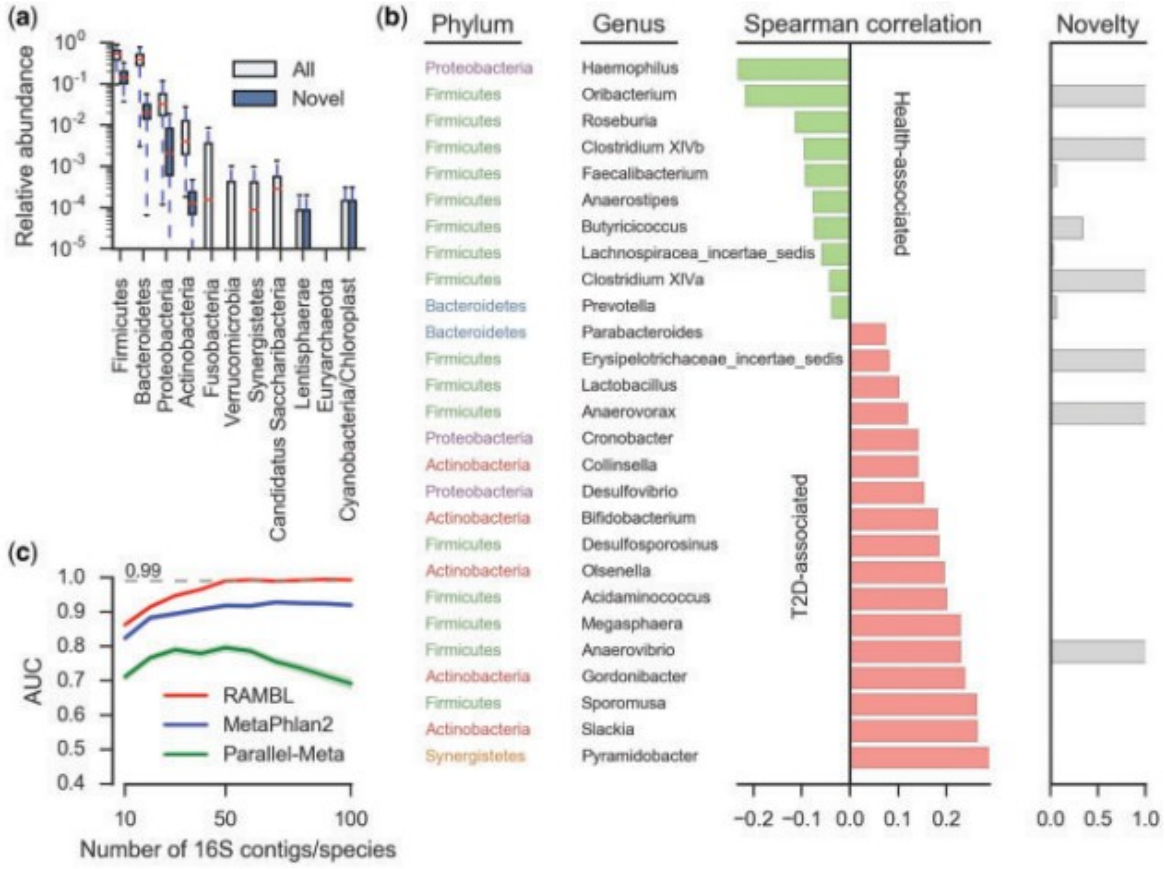

FIGURE 5 Microbial landscape in the gut of a Chinese population. (**a**) The abundance distributions of all microbes and novel microbes. (**b**) The correlation analysis between microbial taxa and individual health (Spearman correlation and Benjamini–Hochberg multiple testing adjustment, *P*-value < 0.05). Novelty is the proportion of novel microbes (contigs) within a taxon. (**c**) The AUC scores of RAMBL, MetaPhlan2 and Parallel-Meta for T2D classification. The red, blue and green regions indicate the 95% confidence intervals

The HMP has identified a list of 'most wanted' OTUs that represent novel species that have never been sequenced in the western population (Fodor *et al.*, 2012). We compared 226 novel 16S contigs that were found in the Chinese population with the HMP 'most wanted' OTU sequences. Out of 119 'most wanted' OTUs, 56 were close to our contigs with sequence identity >90% and MegaBlast *E*-value <1e-30. Near half of the HMP 'most wanted' taxa did not present in the Chinese population. This implies that the unknown microbiomes of different populations are divergent. Our 16S contigs can serve as a complement of the HMP 'most wanted' taxa list to indicate novel species across populations.

Compared with human gut microbial gene catalogs, marker genes and 16S rRNA gene short reads, the 16S assemblies were the better biomarkers for the T2D classification. Using only 10 discriminative 16S assemblies and a trained support vector machine (SVM) classifier, RAMBL achieved the area under the receiver operating characteristic curve (AUC) score 0.86 (Fig. 5c), higher than the AUC score of 0.81 obtained by the previously reported gene

catalog-based classification (Qin *et al.*, 2012) that used 50 gene markers and leave-one-out cross-validation (LOOCV). RAMBL achieved higher accuracy with much fewer marker features. We demonstrated that our classification was also more accurate than the classifications obtained by MetaPhlan2 and Parallel-Meta. As shown in Figure 5c, our classification obtained a relative increase of 7% in the AUC score compared to that of the MetaPhlan2 classification for a wide range of discriminative 16S assemblies and species. When 50 or more discriminative 16S gene assemblies were used, our classification achieved an AUC score of 0.99. In comparison, the AUC score of MetaPhlan2 was 0.92–0.93 when the same number of species was selected for classification. The best AUC score of Parallel-Meta was 0.796 when 50 OTUs determined by 16S rRNA short reads were used. These results reveal that the 16S assemblies serve as good diagnostic markers for diabetes.

Out of the 50 16S assemblies that well classified the disease status, 14 16S assemblies represented the novel microbes. These novel microbes were mostly from the phylum Firmicutes, including 1 Blautia microbe, 1 Clostridium IV microbe, 6 Clostridium XIVa microbes, 1 Clostridium XVIII microbe, 1 Faecalibacterium microbe, 1 Flavonifractor microbe, 1 Lachnospiracea_incertae_sedis microbe and 1 Ruminococcus microbe. A novel Bacteroidetes microbe Barnesiella also contributed to the disease classification. Except for the Clostridium IV microbe that positively correlated with the disease, all the other microbes were the health-associated microbes. Although it is well known that the health microbiome is more divergent than the disease microbiome, we unraveled that the health microbiome possesses more novel microbes that play an important role for the maintenance of the individual health.

4 Discussion

Identifying a full spectrum of microbes is critical to the interpretation of microbial diversity, but remains unachievable for metagenomic shotgun sequencing data because over 99% of microbes are uncharacterized in terms of genome sequences. We offer RAMBL, a scalable pipeline to assemble short and error-prone 16S rRNA sequencing reads to full-length high-quality 16S gene sequences, maximizing taxonomic identification from metagenomic shotgun sequences. To the best of our knowledge, RAMBL is the first tool that realizes the assembly of full-length 16S rRNA gene sequences for very large metagenomic datasets, as demonstrated by the soil and T2D datasets that had 11 gigabases and 359 gigabases of shotgun sequencing reads, respectively. Our work suggests that full-length 16S gene assemblies are superior to marker gene set and 16S short reads, because they can identify both known and novel genera, and accurately quantify them to a wide range of abundance levels.

We observed that RAMBL generated few chimeric assemblies in the soil (1%) and gut (13.7%) datasets. Results of chimera checking indicate full-length 16S gene assemblies of RAMBL are of high accuracy. In comparison, 91.7%

of the EMIRGE soil contigs could not be aligned to known 16S rRNA reference sequences, and thus were invalid for chimera checking. Since both EMIRGE and Reago failed to assemble 16S gene sequences for the T2D data, we could not determine the chimera rates of them. We attribute the low chimera rate for the soil data to the fact that the soil biomes harness diverse microbes (Fierer and Jackson, 2006). The gut microbiota, in the contrary, is abundant of closely related strains (Schloissnig *et al.*, 2013). This implies that RAMBL would suffer a higher risk of the chimeric assembly when a community harnesses a higher proportion of similar strains. We hope to resolve this issue in the future.

Binning of metagenomic contigs is a widely adopted method to identify potential novel microbial sequences of a community (Nielsen *et al.*, 2014), but in general, the objective is not very clearly defined beyond the binning itself. In contrast, full-length 16S gene assemblies provide a crystalline depiction of a community, of which uncharacterized and novel genera can be accurately determined. Overall, full-length 16S gene assemblies open the door to the uncharacterized microbial community, and make possible the future investigation of genetic and metabolic functionalities of these novel microbes.

Funding

This work has been supported by the National Natural Science Foundation of China (Nos: 61503314, 61561146396 and 61203282) and the Tsinghua TNLIST Big Data Grant.

References

Caporaso,J.G. et al. (2010) QIIME allows analysis of high-throughput community sequencing data. Nat. Methods, 7, 335–336.

Chakravorty,S. et al. (2007) A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. J. Microbiol. Methods, 69, 330–339.

Cole,J.R. et al. (2014) Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res., 42, D633–D642.

DeSantis,T.Z. et al. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol., 72, 5069–5072.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res., 32, 1792–1797.

Eloe-Fadrosh,E.A. et al. (2016) Metagenomics uncovers gaps in ampliconbased detection of microbial diversity. Nat. Microbiol., 15032.

Fierer,N. et al. (2012) Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. Proc. Natl. Acad. Sci. USA, 109, 21390–21395.

Fierer,N. and Jackson,R.B. (2006) The diversity and biogeography of soil bacterial communities. Proc. Natl. Acad. Sci. USA, 103, 626–631.

Fischer,A. et al. (2014) High-definition reconstruction of clonal composition in cancer. Cell Rep., 7, 1740–1752.

Fodor,A.A. et al. (2012) The ‘Most Wanted’ taxa from the human microbiome for whole genome sequencing. PLoS One, 7, e41294.

Franzosa,E. a. et al. (2015) Sequencing and beyond: integrating molecular ‘omics’ for microbial community profiling. Nat. Rev. Microbiol., 13, 360–372.

Haas,B.J. et al. (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res., 21, 494–504.

Holtgrewe,M. (2010) Mason—a read simulator for second generation sequencing data. Technical Report, FU Berlin.

Kaakoush,N.O. (2015) Insights into the role of erysipelotrichaceae in the human host. Front. Cell. Infect. Microbiol., 5, 84.

Klappenbach,J.A. (2001) rrndb: the ribosomal RNA operon copy number database. Nucleic Acids Res., 29, 181–184.

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. Nat. Methods, 9, 357–359.

Lee,C. et al. (2002) Multiple sequence alignment using partial order graphs. Bioinformatics, 18, 452–464.

Li,H. et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25, 2078–2079.

Miller,C.S. et al. (2011) EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. Genome Biol., 12, R44.

Nawrocki,E.P. et al. (2009) Infernal 1.0: inference of RNA alignments. Bioinformatics, 25, 1335–1337.

Nielsen,H.B. et al. (2014) Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat. Biotechnol., 32, 822–828.

Paisley,J. et al. (2010) A Stick-Breaking Construction of the Beta Process. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 847–854.

Price,M.N. et al. (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol., 26, 1641–1650.

Pruesse,E. et al. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res., 35, 7188–7196.

Qin,J. et al. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature, 464, 59–65.

Qin,J. et al. (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature, 490, 55–60.

Schloissnig,S. et al. (2013) Genomic variation landscape of the human gut microbiome. Nature, 493, 45–50.

Sharon,I. and Banfield,J.F. (2013) Genomes from Metagenomics. Science (80-.), 342, 1057–1058.

Singer,E. et al. (2016) High-resolution phylogenetic microbial community profiling. ISME J., 10, 2020–2032.

Su,X. et al. (2014) Parallel-META 2.0: enhanced metagenomic data analysis with functional annotation, high performance computing and advanced visualization. PLoS One, 9, e89323.

Sunagawa,S. et al. (2013) Metagenomic species profiling using universal phylogenetic marker genes. Nat. Methods, 10, 1196–1199.

Sunagawa,S. et al. (2015) Ocean plankton. Structure and function of the global ocean microbiome. Science, 348, 1261359.

Consortium The HMP,. (2012) A framework for human microbiome research. Nature, 486, 215–221.

Topfer,A. et al. (2013) Probabilistic inference of viral quasispecies subject to recombination. J. Comput. Biol., 20, 113–123.

Truong,D.T. et al. (2015) MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat. Methods, 12, 902–903.

Vital,M. et al. (2015) Diet is a major factor governing the fecal butyrateproducing community structure across Mammalia, Aves and Reptilia. ISME J., 9, 832–843.

Wang,Q. et al. (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol., 73, 5261–5267.

Yuan,C. et al. (2015) Reconstructing 16S rRNA genes in metagenomic data. Bioinformatics, 31, i35–i43.

Zhang,Z. et al. (2000) A greedy algorithm for aligning DNA sequences. J. Comput. Biol., 7, 203–214.

Zhou,J. et al. (2015) High-throughput metagenomic technologies for complex microbial community analysis: open and closed formats. MBio, 6, e02288–14.