High-resolution single-molecule long-fragment rRNA gene amplicon sequencing of bacterial and eukaryotic microbial communities

Fang, Chao; Sun, Xiaohuan; Fan, Fei; Zhang, Xiaowei; Wang, Ou; Zheng, Haotian; Peng, Zhuobing; Luo, Xiaoqing; Chen, Ao; Zhang, Wenwei; Drmanac, Radoje; Peters, Brock A.; Song, Zewei; Kristiansen, Karsten

Published in:
Cell reports methods

DOI (link to publication from Publisher):
10.1016/j.crmeth.2023.100437

Publication date:
2023

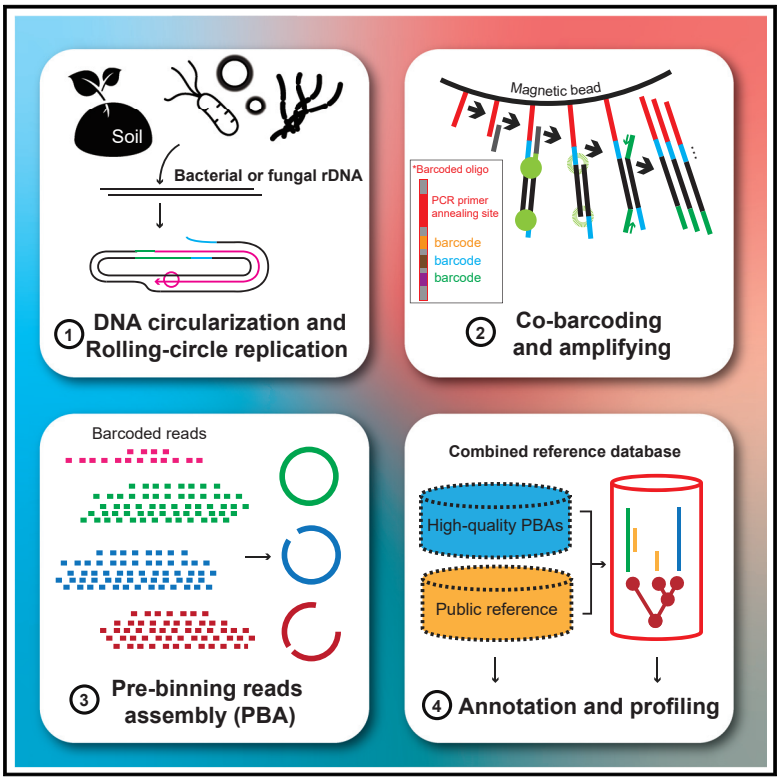
Document Version
Publisher's PDF, also known as Version of record

Cell Reports
Methods

Report

High-resolution single-molecule long-fragment rRNA gene amplicon sequencing of bacterial and eukaryotic microbial communities

Graphical abstract



Highlights

- Co-barcoding long fragment method for sequencing of bacterial and fungal rDNA
- Uses second generation sequencing data to assemble nearly full-length rDNA
- Has high identification accuracy and is reproducible for complex microbial communities
- Is high-throughput and cost-effective compared to similar methods

Authors

Chao Fang, Xiaohuan Sun, Fei Fan, ..., Brock A. Peters, Zewei Song, Karsten Kristiansen

Correspondence

bpeters@mgi-tech.com (B.A.P.), songzewei@genomics.cn (Z.S.), kk@bio.ku.dk (K.K.)

In brief

Fang et al. present a co-barcoding long fragment reads method based on second-generation sequencing to obtain nearly full-length rDNA sequences from bacteria and fungi. Benchmarking using synthetic consortia of bacteria and fungi demonstrates high accuracy, and analyses of soil demonstrate high taxonomic resolution for complex communities.

Fang et al., 2023, Cell Reports Methods *3*, 100437
March 27, 2023
https://doi.org/10.1016/j.crmeth.2023.100437

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Report

High-resolution single-molecule long-fragment rRNA gene amplicon sequencing of bacterial and eukaryotic microbial communities

Chao Fang,[1,2,10] Xiaohuan Sun,[1,10] Fei Fan,[1,10] Xiaowei Zhang,[3,10] Ou Wang,[1,10] Haotian Zheng,[1,4,5] Zhuobing Peng,[1,5] Xiaoqing Luo,[1,6] Ao Chen,[1] Wenwei Zhang,[1] Radoje Drmanac,[7,8] Brock A. Peters,[7,8,*] Zewei Song,[1,*] and Karsten Kristiansen[1,2,9,11,*]

[1]BGI-Shenzhen, Shenzhen 518083, China
[2]Laboratory of Genomics and Molecular Biomedicine, Department of Biology, University of Copenhagen, 2100 Copenhagen, Denmark
[3]Department of Obstetrics and Gynecology, Peking University Shenzhen Hospital, Shenzhen, 518036, China
[4]Section of Microbiology, University of Copenhagen, 2100 Copenhagen, Denmark
[5]BGI Education Center, University of Chinese Academy of Sciences, Shenzhen, 518083, China
[6]State Key Laboratory of Biocontrol and Guangdong Provincial Key Laboratory of Plant Resources, School of Life Sciences, Sun Yat-Sen University, Guangzhou, 510275, China
[7]Advanced Genomics Technology Lab, Complete Genomics Inc., 2904 Orchard Parkway, San Jose, CA 95134, USA
[8]MGI, BGI-Shenzhen, Shenzhen 518083, China
[9]PREDICT, Center for Molecular Prediction of Inflammatory Bowel Disease, Faculty of Medicine, Aalborg University, 2450 Copenhagen, Denmark
[10]These authors contributed equally
[11]Lead contact
*Correspondence: bpeters@mgi-tech.com (B.A.P.), songzewei@genomics.cn (Z.S.), kk@bio.ku.dk (K.K.)
https://doi.org/10.1016/j.crmeth.2023.100437

MOTIVATION Current metagenomics studies are often limited by a lack of accurate taxonomic identification related to the species and strain level. This situation is more severe for complex environment samples. Moreover, obtaining full-length rDNA for taxonomic identification is still a challenge using cost-efficient shotgun sequencing and expensive using single-molecule sequencing. Co-barcoding of shotgun sequencing reads provides additional information to assemble full-length rDNA at the single molecule level and is also high-throughput and cost-efficient.

SUMMARY

Sequencing of hypervariable regions as well as internal transcribed spacer regions of ribosomal RNA genes (rDNA) is broadly used to identify bacteria and fungi, but taxonomic and phylogenetic resolution is hampered by insufficient sequencing length using high throughput, cost-efficient second-generation sequencing. We developed a method to obtain nearly full-length rDNA by assembling single DNA molecules combining DNA co-barcoding with single-tube long fragment read technology and second-generation sequencing. Benchmarking was performed using mock bacterial and fungal communities as well as two forest soil samples. All mock species rDNA were successfully recovered with identities above 99.5% compared to the reference sequences. From the soil samples we obtained good coverage with identification of more than 20,000 unknown species, as well as high abundance correlation between replicates. This approach provides a cost-effective method for obtaining extensive and accurate information on complex environmental microbial communities.

INTRODUCTION

Currently, there are two main approaches for sequencing rRNA genes in complex microbial environmental samples: second-generation technologies such as sequencing by synthesis and DNA nanoball sequencing (DNBseq) and third-generation technologies such as nanopore and single molecule real-time sequencing. Second-generation sequencing suffers from relatively short read lengths (less than 300 bases) making it difficult to sequence the entire rRNA gene. Recently, co-barcoding technologies combined with second-generation sequencing and *de novo* assembly have enabled recovery of nearly full-length

rRNA genes, but the quantification reproducibility using complex samples is unknown.[1,2] Recently, third-generation technologies using unique molecular identifiers (UMI) or circular consensus sequence (CCS) technologies have been developed, providing much longer read lengths and enabling full coverage of the rRNA genes with decent accuracy, but they are still hampered by relatively high cost and low throughput.[3–8] As such, there is still a great demand for high-throughput and cost-effective approaches to support the deciphering of complex microbial communities. In this study, we developed a strategy whereby a combination of rolling-circle replication (RCR) and DNA co-barcoding[9] techniques allows sequence information of long amplicons to be obtained by a high-throughput, single-molecule level *de novo* assembly approach to achieve species resolution and highly accurate profiles in an efficient and cost-effective manner using a short read sequencing platform. To demonstrate the applicability of this method, we performed benchmarking using two mock communities of various bacterial and fungal species. We also applied the protocol to field soil samples as a demonstration for the discovery of unknown species and were able to recover rRNA sequences even longer than those found in current reference databases.

RESULTS

The basic idea behind the present protocol is to assemble a single DNA molecule by using DNA co-barcoding technology enabled by the single-tube long fragment read (stLFR) method. Using stLFR, it is possible to label DNA sub-fragments from a single long DNA molecule with the same barcode. Importantly, in a single stLFR library there are approximately 3.6 billion different barcodes, allowing each long DNA molecule in a sample to be labeled by a unique barcode.[10] One shortcoming of the current stLFR method is that sequence coverage of each long molecule is less than 1X. As a result, using the De Bruijn graph (DBG) algorithm, it is usually impossible to achieve full assembly of DNA fragments from a single barcode. To overcome this lack of co-barcoded sequence coverage, we applied a modified version of the stLFR technology, named single tube complete-coverage long fragment read (stcLFR). This approach involves an initial process of RCR, which performs a linear amplification whereby tandem copies of a single DNA molecule are joined head to tail in one long single-stranded concatemer. This method also avoids cumulative replication errors because the same DNA fragment is used as the template for replication. After conversion to double-stranded DNA, a transposon containing a capture sequence is inserted at regular intervals of approximately 500 base pairs in the RCR amplified products. After insertion of the transposon, the DNA is captured by barcodes containing oligonucleotides anchored to the surface of a micron-sized magnetic bead, whereupon the DNA is fragmented and ligated to these oligonucleotides (Figure 1A). After PCR on the beads, the DNA fragments with unique barcodes, corresponding to multiple copies of each single molecule, are subjected to DNBseq. By decoding barcoded sequences from the sequencing dataset, copy depth can be estimated in each barcoded binning group. Bins with depth >5X are then used for parallel *de novo* assembly (Figure 1B). Successfully assembled contigs are termed pre-binning assemblies (PBAs) and used for subsequent analysis.

The ZymoBIOMICS mock community, which contains eight bacterial species (see STAR Methods), was used to test the ability of stcLFR to properly assemble the rDNA amplicons. A region covering 4.5 kb from the ribosomal small subunit (SSU) gene (515Fng forward primer) to the ribosomal large subunit (LSU) gene (TW13 reverse primer) was amplified for sequencing. From 187,942,995 raw reads generated from 3 replicates, 186,580,349 passed quality control as high-quality reads. Among these, 158,291,516 harbored a barcode, comprising in total 26,512,829 unique barcodes. Among those with valid barcoded read bins, 148,814 bins with a coverage >5X were used for *de novo* assembly. With a relatively small number of reads and simple genomic content, the assembly proceeded faster and was more complete compared to general metagenomic assembly. We finally retrieved 146,509 assembled PBAs with a centroid length distribution close to 800 bp (Figure S1A). Most of them covered either the SSU (34.15%) or the LSU (38.58%). Only 25.23% of PBAs covered both the SSU and the LSU (Figure S1B). The assembly quality was high, with 9.72% of the aligned PBAs exhibiting 100% identity to the reference genomes, and more than 60% of the PBAs achieved 99.5% or higher identity with the reference genomes. By contrast, only 1.78% of the PBAs could not be aligned to the reference genomes (Figure 2A). Details of numbers obtained in each step are summarized in Table S1.

As the ZYMO mock community is widely used as a reference sample for benchmarking, we also collected public datasets on sequencing of 16S rRNA and 23S rRNA genes using Illumina HiSeq2500, HiSeq4000, PacBio, and Nanopore platforms. Substitutions and insertions error rates in the 16S rRNA and 23S rRNA genes were computed by using clean reads mapping to the reference sequences (Figure S2). For substitutions, we observed that the stcLFR reads exhibited slightly higher error rates than those observed for HiSeq4000 reads but lower error rates than what was observed for the other sequencing platforms. For deletions and insertions, the error rate was similar to that of the HiSeq2500 reads but higher than those observed for HiSeq4000, PacBio, and Nanopore reads. For the stcLFR PBAs, we observed that the error rates for deletions and insertions were comparable to those observed for the HiSeq2500 platform but higher that those determined using the HiSeq4000, PacBio, and Nanopore platforms, For substitutions, the stcLFR PBAs exhibited slightly higher error rates than those observed for the HiSeq4000 platform but lower error rates than those observed using the HiSeq2500, PacBio, and Nanopore platforms (Figure S2).

For tests on fungi, a mock community comprising seven common fungal species was used to assess the classification resolution at the species level (see STAR Methods). Primers spanning a region ~2.5 kb covered ITS1 and ITS2, and parts of the flanking SSU and LSU rRNA genes were used. Following the DNA library preparation and sequencing protocols used for bacterial species, we generated 190,208,124 high-quality reads and decoded 21,215,624 barcoded bins, with 270,794 having a coverage >5X. In total, 263,416 PBAs were generated with a size centered on 2.3 kb, which is close to the value expected from the design (Figure S1C). Exceeding the results obtained for the

CellPress
OPEN ACCESS

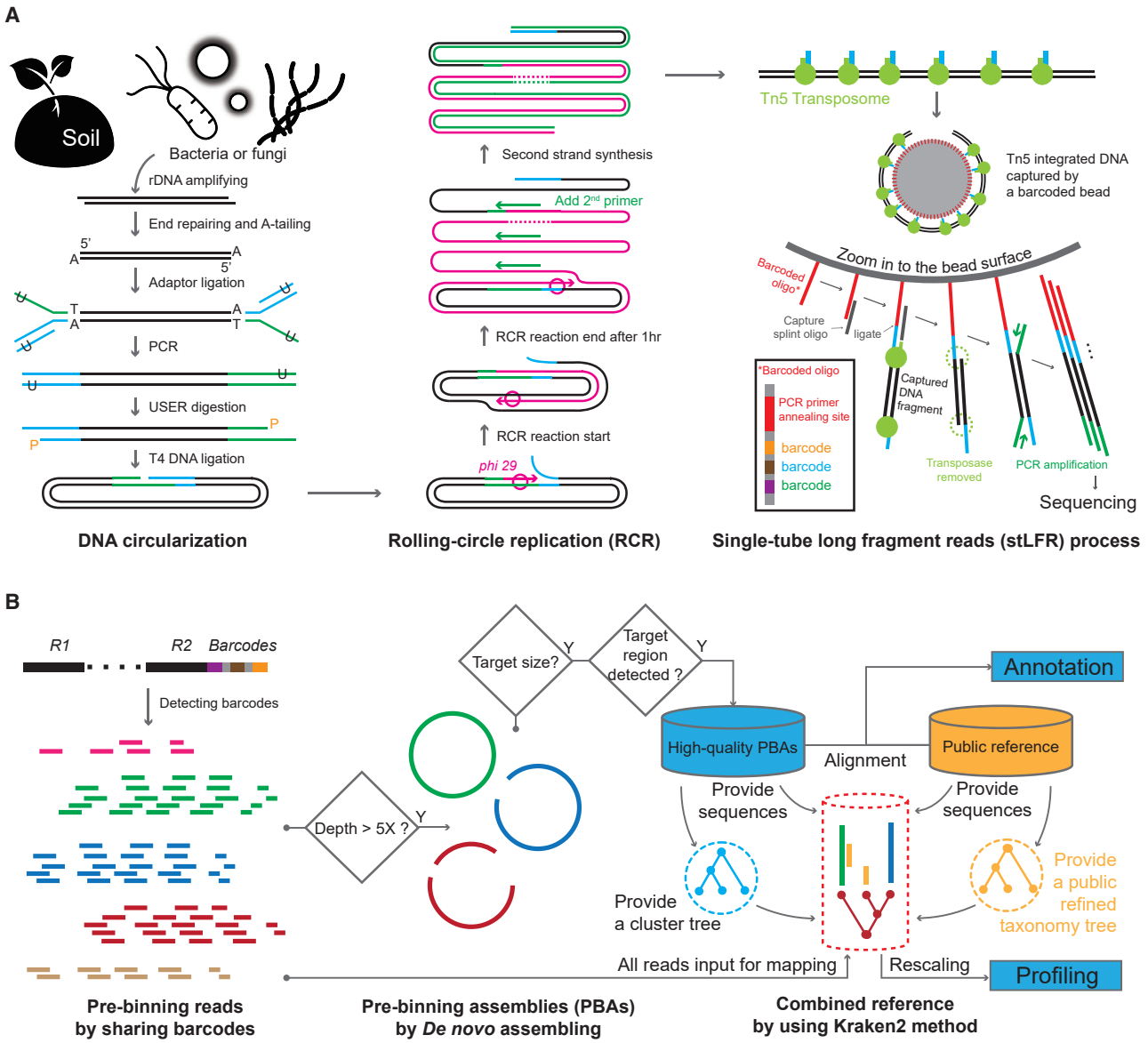


Figure 1. Sequencing framework

(A) *In vitro* processes of rolling-circle replication (RCR) and stcLFR. DNA from bacteria and fungi was extracted and amplified using designed rDNA primers. Amplicons were turned into long concatemers with circularization and RCR. The products of RCR were labeled with unique barcodes by integrating transposons and hybridizing onto 30 million different clonal barcoded beads in a single tube. After PCR, these sheared short sub-fragments were subjected to DNA sequencing.

(B) *In silico* processes of assembly, classification, and quantification. Reads with attached barcodes were decoded during the quality-control process. Pre-binning was done by grouping reads sharing the same barcode. Subsequently, *de novo* assembly was performed for each bin independently and in parallel. The pre-binning assemblies (PBAs) were then selected by target size range, and rDNA subunits or internal transcribed spacer (ITS) regions were detected by Barnnap and ITSx software. The open reference databases SILVA and UNITE were used for taxonomic classification. Kraken2 was used to generate a database for profiling by mapping all barcode-detected reads.

bacterial samples, 12.54% of the fungal PBAs exhibited a 100% alignment to the reference genomes, and more than 72.4% of the PBAs achieved 99.5% or higher identity to the reference genomes. Only 0.43% of the PBAs could not be aligned to the reference genomes (Figure 2B). 76.8% of the PBAs covered the entire region from the SSU to the LSU, and the ITS region was included in 83.7% of the PBAs (Figure S1D).

For assessing the accuracy of quantification, we mapped all decoded reads to the assembled complete rRNA gene sequences. Mapped reads sharing the same barcode represent a single DNA molecule and as such were counted once. We determined the difference between the observed and the calculated relative abundance of each species to estimate the error. For analyses of the mock bacterial community, we

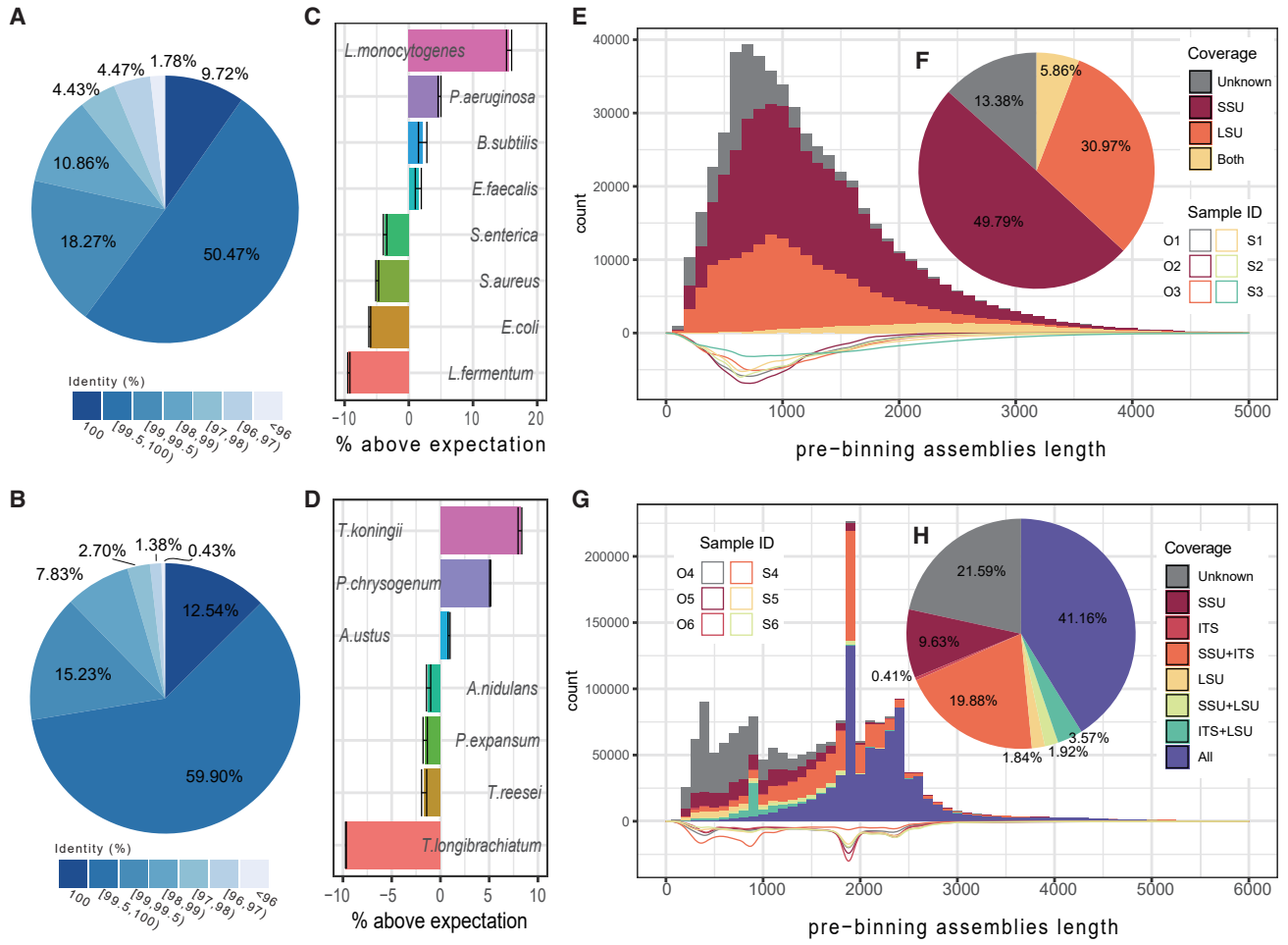


Figure 2. Long fragments restore performance

(A) Identity of alignments of mock bacterial rDNA PBAs to reference sequences.

(B) Identity of alignments of mock fungal rDNA PBAs to reference sequences.

(C) Difference between the observed and the calculated abundances of bacterial species in the mock community.

(D) Difference between the observed and the calculated abundances of fungal species in the mock community.

(E) rDNA sequences assembled from soil bacterial DNA. Subunits (16S/SSU and 23S/LSU) were detected by alignment to the SILVA and UNITE database or predicted by barrnap. The length distribution of each sample is plotted on the negative part of the y axis.

(F) Pie plot of percentages of subunits detected.

(G) rDNA sequences assembled from soil fungal DNA. Percentage of covered 18S/SSU, 28S/LSU, and ITS is indicated by different colors.

(H) Pie plot of percentages of subunits and ITS regions detected.

compared the relative abundance of each taxonomy to the theoretical expectation, revealing that the differences ranged from −9.3% to 15.6%, with a standard deviation of 7.9% (Figure 2C). For analyses of the fungal mock community, we estimated the error to range from −9.7% to 8.2%, with an error standard deviation of 5.6% (Figure 2D). We estimated that the variation in the relative abundance of each species, as defined by the variance of error, was as low as 0.32% for bacterial species and 0.15% for fungi (Figure 2C). Though we only obtained a low percentage of amplicons covering the entire bacterial rRNA gene region as designed, we successfully recovered long fragments of bacterial 16S and 23S rRNA genes, as well as fungal fragments covering the entire ITS region and parts of the flanking 18S and 28S rRNA genes.

From both bacterial and fungal mock samples, the identity between PBAs and the corresponding reference sequences exceeded 99% in most cases, a percentage higher than the 97% that often is used for amplicon-based species identification. In addition, the observed low variation in the relative abundance of each species also demonstrated the consistency of the protocol, an important requirement for comparative analyses using high-resolution profiles.

We had limited success in retrieving the entire ~4 kb sequence of bacterial rDNA regions and observed a gap in the coverage of the bacterial rDNA genes, but we still successfully recovered long fragments of bacterial 16S and 23S rRNA genes. The gap harbors transfer RNA (tRNA) genes (Figure S3A), which may be a target for cleavage by the Tn5 transposase, the enzyme used

for fragmentation during sequencing library preparation.[11–14] Although cleavage by the Tn5 transposase exhibits limited sequence bias, target preferences might still exist and cause the failure of effectively retrieving fragments covering the entire rDNA gene region.[22–24] In eukaryotes, tRNA genes are generally located outside the SSU and LSU regions, not in between, which may at least in part explain why the eukaryotic amplicons survived but the bacterial amplicons were split.

From the reference alignment benchmarking, we observed that when a PBA exceeded its designed size, part of the sequences could be aligned to reference sequences belonging to different species. We consider such PBAs as chimeric PBAs, which were subsequently filtered out (Figure S3B, red lines). Since we successfully enriched for the designed size of fungal rDNA sequences, this trend was prominent when the length of PBAs exceeded 2.3 kb for fungal samples. In addition, for bacterial PBAs with length <500 bp, we also observed a higher probability for achieving the exact same identity and bit score by multiple reference sequences representing different species, leading to an ambiguous annotation (Figure S3B, solid blue line). This trend became more severe in fungal PBAs with sizes <2000 bp (Figure S3B, solid blue line). These findings might point to one limitation of using short sequences to distinguish taxonomies at the species level, especially for fungi. However, coverage of multiple regions greatly eliminated the ambiguous annotation (Figure S3C).

To further explore the abundance and reproducibility of taxonomies in real environmental samples, we sequenced three replicates of two natural soil samples to identify bacterial rRNA genes and fungal ITS regions using the same primer sets and protocols used for the mock samples. We obtained 632,574,076 bacterial reads and 1,006,826,680 fungal reads with valid barcodes. For bacterial amplicons, we successfully generated 544,433 PBAs from 724,038 candidate bins (Figure 2E). As we observed using mock communities, only a small fraction (5.86%) of the bacterial PBAs covered both the SSU and the LSU regions. Most bacterial PBAs only covered the SSU (49.79%) or the LSU (30.97), whereas 13.38% did not match any currently known rDNA region (Figure 2F). For fungal amplicons, 1,807,779 PBAs from 1,845,429 high-coverage bins were generated with a size distribution that peaked at 2.3kb and 1.8kb (Figure 2G). A PBA size of 2.3kb was consistent with the primer design, showing full coverage of the flanking region of the SSU region, the ITSs, and the flanking region of the LSU region. By contrast, half of the PBAs with sizes that peaked around 1.8kb lacked the LSU region. We detected ITS regions from 65.05% of the PBAs (Figure 2H). This percentage increased when the size exceeded 1 kb.

We also tried to profile the bacterial and fungal community at different rank levels. Because soil samples contain very complex microbial communities, we chose Kraken2 to handle the profiling task. To build a Kraken2 database, we first used PBAs to generate cluster trees for bacteria and eukaryotes (mostly fungi). Since we retrieved individual rDNA subunit assemblies for bacteria, we selected PBAs harboring SSU sequences for clustering. For eukaryotes, and especially fungi, the ITS region provided the highest level of discrimination, and accordingly we selected PBAs covering ITS1 and ITS2 for clustering. The operational taxonomic unit (OTU) clusters were constructed using a set of identity thresholds enabling classification of taxa at multiple taxonomy levels, from domain to species. These thresholds were determined by pairwise global alignments of ribosomal gene subunit sequences including SSU and LSU from the SILVA database and ITS from the UNITE database (Figure S4), setting 97% identity as the threshold for genus association and 99% for species association for bacteria, and 95% and 97% identity as the thresholds for genus and species association, respectively, for fungi. Additional thresholds were the same for bacteria and fungi (see STAR Methods). The trees generated using PBAs were then merged with taxonomy trees based on the SILVA and the UNITE databases (Figure 3A). In this process, clades were merged with taxonomies when the representing PBAs achieved sequence identities greater than the rank's threshold. The combined tree contained three categories of branches. One type was represented by the public SILVA and UNITE sequences (PUBs, green in Figure 3, defined as exclusively PUBs), one represented by sequences only present in PBAs (blue in Figure 3, exclusively PBAs), and one represented by both the PBA and the PUB sequences (red in Figure 3, the shared PBAs and PUBs). The merged taxonomy tree and combined sequences were then used to build the database.

By this approach, we mapped all barcoded reads to the combined database, retrieving 60,942 bacterial OTUs at the species level of which 1,078 OTUs were supported by both the PBA sequences and PUB sequences, 18,403 could not be annotated, representing unknown species, and the remaining 41,461 were annotated by publicly available sequences. At the genus level, 6,765 OTUs represented PBAs with no annotation. Using 72% identity as the threshold for association with a phylum, we identified 241 OTUs of which 19 were unknown. For the eukaryotic communities, we retrieved 37,355 fungal OTUs at the species level of which 1,592 OTUs were supported by both the PBA sequences and PUB sequences, 1,817 could not be annotated, representing unknown species, and the remaining 33,946 OTUs were annotated by PUB sequences. At the genus level, 1,266 OTUs represented PBAs with no annotation. Using 72% identity as the threshold for association with a phylum, we identified 244 OTUs, all of which were annotated (data of sample S are shown, Figure 3B, left panel).

We found that species-level OTUs supported by both PBA and PUB sequences represented only 1.8% of all OTUs; the relative abundance of these OTUs combined corresponded to about 12% of the relative abundances of all OTUs. At the genus level, the relative abundance of OTUs supported by both the PBA and PUB sequences was even higher (62%), and at the phylum level the relative abundance of the OTUs supported by both PBA and PUB sequences reached close to 100%. This trend was more obvious for fungal OTUs, where 4.2% of OTUs supported by both the PBAs and PUBs contributed 79% of relative abundance at the species level (Figure 3B, right panel). This indicated that OTUs supported by PBAs (whether annotated or unknown) were abundant and dominated the microbial community. Compared with the fungal community, more dominant taxonomies may represent novel species.

Further analysis showed that the three categories of branches in the combined tree exhibited distinct differences

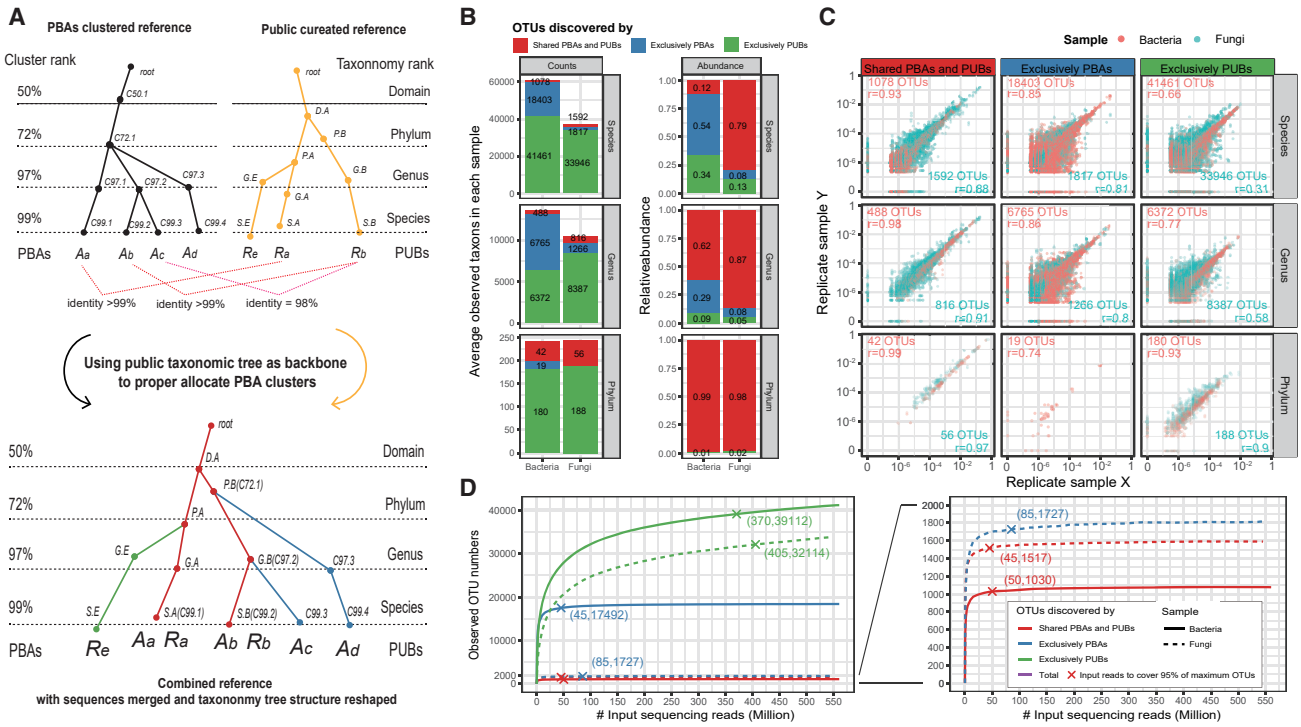


Figure 3. Identification and performance of a combined Kraken2 profiling strategy

(A) The pre binning assembly (PBA) cluster tree was generated by a series of clustering of PBAs at default identity cutoff for each level of taxonomy. The identity cutoffs were estimated by pairwise alignments of annotated SSU, LSU, and ITS sequences from public databases. The public taxonomy tree was mainly based on the SILVA SSU taxonomy tree and supplemented with new taxonomies from the SILVA LSU and UNITE databases. A combination of PBAs and public reference sequences (PUBs) was established based on the identities between the PBA and PUB sequences. The combined taxonomy tree contains three branch categories of sequences: 1) sequences shared by PBAs and the PUBs (red), 2) sequences unique to PBAs with no taxonomy association (blue), and 3) reference sequences from public taxonomy tree not shared by PBA (green).

(B)The performance of the three branches of the taxonomic trees. Left: the number of OTUs shown at the species, genus, and phylum levels, visualized as bar plots. Right, relative abundances. The colors indicate OTUs belonging to PUBs (green) or PBAs with (red) or without (blue) annotation.

(C) Correlation of relative abundance between duplicates. At each level of taxonomy, the relative abundance observed in each duplicate is colored light red for bacteria and light green for fungi. The Spearman correlation values for bacteria and fungi are indicated.

(D) Rarefaction curves of observed bacterial and fungal OTUs at the species level using the different databases. One bacterial sample (solid line) and one fungal sample (dash line) with three replicates and sequenced to more than 500 million read pairs were analyzed. On each curve, a cross marks the number of reads needed to cover 95% of the total counts.

in terms of reproducibility between replicates. OTUs supported by both PBA and PUB sequences exhibited high correlation between replicates (r = 0.93 for bacterial communities and r = 0.88 for eukaryotic communities; Spearman's correlation, Figure 3C). Higher correlations were observed at the genus and phylum levels. By contrast, the correlation between replicates of OTUs supported exclusively by PBA was lower, and the correlation decreased significantly when the abundance became lower than 10^{-4}, both for bacteria and eukaryotes. Some of the unannotated bacterial phyla taxa still varied, indicating that they might not represent an independent phylum. For the taxa supported by public references, though well organized by taxonomic information, the species taxa performed worse than those supported by unannotated PBAs. However, the performance was improved dramatically at the phylum levels. These dominant annotated OTUs supported both by PBAs and PUBs and highly abundant (>10^{-4}) unknown OTUs exhibited high reproducibility for quantification comparison tasks at the species and genus levels.

The rarefaction curves also showed a better performance in terms of the number of reads needed for sufficient coverage of OTUs from the shared PBAs and PUBs sequences. To retrieve the majority of taxa (here defined as 95% of maximum OTUs), the publicly available database required the highest number of read-pairs, 410 million read-pairs for bacteria and 425 million read-pairs for fungi. To assess exclusively PBA OTUs, this requirement was reduced to 110 million and 150 million for bacteria and eukaryotes, respectively. The shared PBAs and PUBs representing OTUs required the lowest number of reads, only 40 million bacterial read-pairs and 55 million eukaryotic read-pairs, respectively.

We finally analyzed the evolutionary phylogeny of bacterial and eukaryotic fully assembled OTUs. We failed to obtain both SSU and LSU of bacterial rDNA, and only 5,574 nearly fully covered SSU PBAs were aligned for maximum likelihood (ML) model fitting (Figure 4A). Most of the clades were clustered well at the phylum rank, but more than half of the clades at the genus rank and nearly all clades at the species rank lack

CellPress
OPEN ACCESS

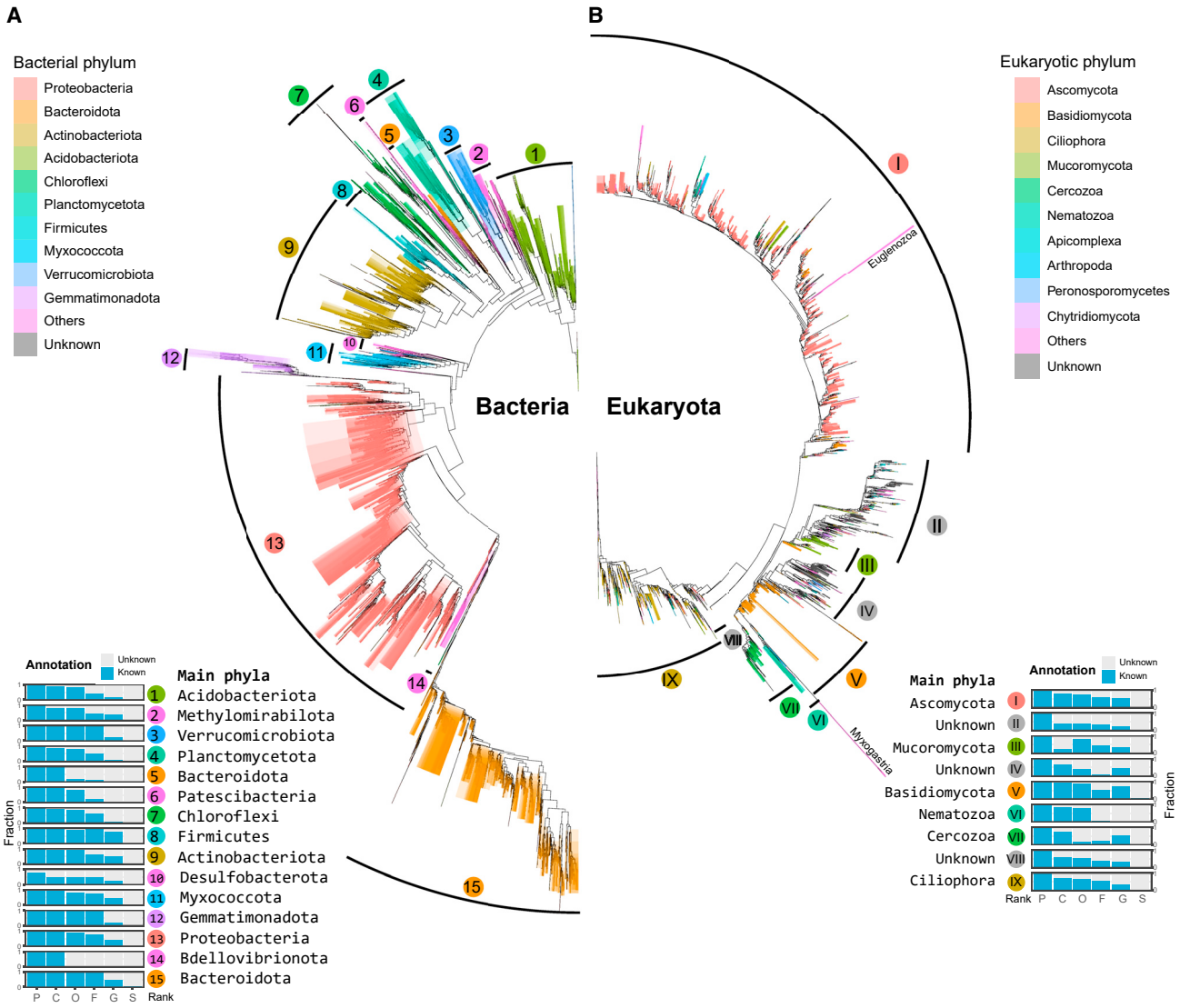


Figure 4. Phylogenetic tree based on bacterial SSU rDNA and eukaryotic SSU-LSU rDNA long sequences

(A) After filtering of too-short assemblies and potential chimeras, 5,574 bacterial PBAs covering nearly the full region of SSU rDNA were aligned and fitted using a GTR model with 100 bootstraps. Branches representing 15 clades are colored according to phyla. Based on current public databases, the fractions of known taxonomic rank annotation at the levels of (P)hylum, (C)lass, (O)rder, (F)amily, (G)enus, and (S)pecies are shown in the bar plot at the left-bottom corner.

(B) After filtering of too-short assemblies and potential chimeras, 3,031 fully assembled eukaryotic PBAs were aligned. SSU and LSU regions were directly joint for GTR model fitting with 100 bootstraps. According to the annotations, we clustered nine clades even though many of the clusters presented a mixture of different phyla. The fractions of known taxonomic rank annotation at the levels of (P)hylum, (C)lass, (O)rder, (F)amily, (G)enus, and (S)pecies are shown in the bar plot in the right-bottom corner.

annotation. These clades may represent novel taxonomies, but these findings may also reflect that current identification mostly is based on 97% similarity as a threshold, which seems to be insufficient for classification of long amplicons. For eukaryotes, 3,031 PBAs covering both SSU and LSU were aligned for ML model fitting (Figure 4B). Compared with the bacterial tree, major groups of the eukaryotic tree are mixed with ambiguous phyla. In addition, unknown phyla were found in groups II, IV, and VIII. Thus, for both bacterial and fungal long OTUs, nearly half of the OTUs at the genus level and almost all at the species level are unknown. This indicates that current references still might be insufficient for species level identification, warranting more studies especially on the eukaryotic microbiome.

DISCUSSION

Here we describe a cost-efficient method to sequence and assemble nearly full-length rDNA sequences by combining DNA co-barcoding with stLFR technology and second-generation sequencing. Recently, Karst and co-workers[3] published an alternative approach for long-read amplicon sequencing. In

that work, redesigned UMIs and single-molecular sequencing were used for obtaining long-read high-accuracy amplicon sequences using Nanopore or PacBio sequencing. In their work, they applied the PacBio UMI method to generate 253,089 high-quality, full-length bacterial rRNA operon sequences from 70 human fecal samples. According to their estimation, the cost was about US$396 per sample. For comparison, to obtain 50 million short reads per sample using the stcLFR method, each sample could recover 19,522 bacterial PBAs and close to 30,000 OTUs, or 2,547 fungal PBAs and about 20,000 OTUs, for a cost of US$70 per sample. While their work demonstrated an elegant approach, our approach relying on second-generation sequencing enables a more cost-effective high-throughput sequencing coupled with robust and reproducible quantification. The generation of long amplicons in a single study enables high taxonomic resolution of even very complex microbial communities as those existing in the rumen of ruminants[15] and the soil.[16,17] In addition, we observed an increase in the occurrence of chimeras and assembly gaps between the bacterial 16S and 23S sequences, where tRNA tandem genes are located. Based on this observation, regions with complex structures, like tRNA cloverleaf structures, should be avoided when designing target amplicons, as such structures may negatively impact co-barcoding continuity and randomness. Long amplicon sequences are also of great value for refining current reference databases. In summary, our approach provided a 99% identity at the species level, pointing to a high-throughput strategy to expand current rRNA gene databases by including long marker sequences and potential novel taxonomies, as well as a comparable accurate quantification profiling strategy. This approach provides a cost-effective method for obtaining extensive and accurate information on environmental complex microbial communities.

Limitations of the study

In this study, we selected specific primers to examine bacteria and fungi from soil samples. Primer-biases were not evaluated in this stage, which should be considered in generalized studies. For taxonomic identification, an algorithm was designed to merge the SILVA and UNITE databases, but a perfect merger is not possible. Scientific taxon names and ranks will need further validation and continued updating.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Bacterial and fungal community mock
  - Fungal microbial community mock
- METHOD DETAILS
  - Sample collection and DNA extraction
  - rDNA long fragments amplification
  - Circularization and rolling-circle replication
  - Single-tube long fragment read barcoded DNA library construction
  - Sequencing and decoding long fragments
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Single-molecular pre-binning and *de novo* assembly strategy
  - Taxonomic annotation of pre-binning assemblies
  - Default taxonomic rank threshold determination
  - Cluster tree generation
  - Relative abundance computation of each rank
  - Phylogenetic tree building

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2023.100437.

ACKNOWLEDGMENTS

This work was partially supported by a grant to Ou Wang from the National Natural Science Foundation of China (32001054). The samples were obtained from Dr. Yue-hua Hu, who is funded by the West Light Foundation of the Chinese Academy of Sciences.

AUTHOR CONTRIBUTIONS

F.F, O.W., and B.P. conceived the wet lab method. X.S., F.F., and O.W performed wet lab experiments and BGISEQ500 sequencing. C.F. and Z.S. conceived the bioinformatics method. C.F developed the software pipeline and performed data analysis as well as visualization. X.S. and X.Z. interpreted the microbial characteristics of data. X.S. performed the phylogenetic analysis. H.Z., Z.P., and X.L. assisted in performing the data. C.F., X.S., X.Z., O.W., B.P., Z.S., and K.K. wrote the manuscript. All authors participated in discussions and contributed to the revision of the manuscript. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

R.D. and B.P. were employed by the company Complete Genomics Inc., US.

Received: August 26, 2022
Revised: January 28, 2023
Accepted: March 1, 2023
Published: March 27, 2023

REFERENCES

1. Bishara, A., Moss, E.L., Kolmogorov, M., Parada, A.E., Weng, Z., Sidow, A., Dekas, A.E., Batzoglou, S., and Bhatt, A.S. (2018). High-quality genome sequences of uncultured microbes by assembly of read clouds. Nat. Biotechnol. *36*, 1067–1075. https://doi.org/10.1038/nbt.4266.
2. Karst, S.M., Dueholm, M.S., McIlroy, S.J., Kirkegaard, R.H., Nielsen, P.H., and Albertsen, M. (2018). Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias. Nat. Biotechnol. *36*, 190–195. https://doi.org/10.1038/nbt.4045.
3. Karst, S.M., Ziels, R.M., Kirkegaard, R.H., Sørensen, E.A., McDonald, D., Zhu, Q., Knight, R., and Albertsen, M. (2021). High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. Nat. Methods *18*, 165–169. https://doi.org/10.1038/s41592-020-01041-y.

4. Nicholls, S.M., Quick, J.C., Tang, S., and Loman, N.J. (2019). Ultra-deep, long-read nanopore sequencing of mock microbial community standards. GigaScience *8*, giz043. https://doi.org/10.1093/gigascience/giz043.

5. Callahan, B.J., Wong, J., Heiner, C., Oh, S., Theriot, C.M., Gulati, A.S., McGill, S.K., and Dougherty, M.K. (2019). High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. Nucleic Acids Res. *47*, e103. https://doi.org/10.1093/nar/gkz569.

6. Wagner, J., Coupland, P., Browne, H.P., Lawley, T.D., Francis, S.C., and Parkhill, J. (2016). Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene classification. BMC Microbiol. *16*, 274. https://doi.org/10.1186/s12866-016-0891-4.

7. Deamer, D., Akeson, M., and Branton, D. (2016). Three decades of nanopore sequencing. Nat. Biotechnol. *34*, 518–524. https://doi.org/10.1038/nbt.3423.

8. Benítez-Páez, A., Portune, K.J., and Sanz, Y. (2016). Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer. GigaScience *5*, 4. https://doi.org/10.1186/s13742-016-0111-z.

9. Peters, B.A., Liu, J., and Drmanac, R. (2014). Co-barcoded sequence reads from long DNA fragments: a cost-effective solution for "perfect genome" sequencing. Front. Genet. *5*, 466. https://doi.org/10.3389/fgene.2014.00466.

10. Wang, O., Chin, R., Cheng, X., Wu, M.K.Y., Mao, Q., Tang, J., Sun, Y., Anderson, E., Lam, H.K., Chen, D., et al. (2019). Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. Genome Res. *29*, 798–808. https://doi.org/10.1101/gr.245126.118.

11. Adey, W.H., Kangas, P.C., and Mulbry, W. (2011). Algal turf scrubbing: cleaning surface waters with solar energy while producing a biofuel. Bioscience *61*, 434–441. https://doi.org/10.1525/bio.2011.61.6.5.

12. Picelli, S., Björklund, A.K., Reinius, B., Sagasser, S., Winberg, G., and Sandberg, R. (2014). Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. Genome Res. *24*, 2033–2040.

13. Hennig, B.P., Velten, L., Racke, I., Tu, C.S., Thoms, M., Rybin, V., Besir, H., Remans, K., and Steinmetz, L.M. (2018). Large-scale low-cost NGS library preparation using a robust Tn5 purification and tagmentation protocol. G3 *8*, 79–89.

14. Wang, Y., Zhang, Y., Jin, H., Deng, Z., Li, Z., Mai, Y., Li, G., and He, H. (2018). A practical random mutagenesis system for Ralstonia solanacearum strains causing bacterial wilt of Pogostemon cablin using Tn5 transposon. World J. Microbiol. Biotechnol. *35*, 7.

15. Stewart, R.D., Auffret, M.D., Warr, A., Walker, A.W., Roehe, R., and Watson, M. (2019). Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. Nat. Biotechnol. *37*, 953–961. https://doi.org/10.1038/s41587-019-0202-3.

16. Ramirez, K.S., Knight, C.G., de Hollander, M., Brearley, F.Q., Constantinides, B., Cotton, A., Creer, S., Crowther, T.W., Davison, J., Delgado-Baquerizo, M., et al. (2018). Detecting macroecological patterns in bacterial communities across independent studies of global soils. Nat. Microbiol. *3*, 189–196. https://doi.org/10.1038/s41564-017-0062-x.

17. Tedersoo, L., Sánchez-Ramírez, S., Kõljalg, U., Bahram, M., Döring, M., Schigel, D., May, T., Ryberg, M., and Abarenkov, K. (2018). High-level classification of the Fungi and a tool for evolutionary ecological analyses. Fungal Divers. *90*, 135–159. https://doi.org/10.1007/s13225-018-0401-0.

18. Guo, X., Chen, F., Gao, F., Li, L., Liu, K., You, L., Hua, C., Yang, F., Liu, W., Peng, C., et al. (2020). CNSA: a data repository for archiving omics data. Database *2020*, ebaaa055. https://doi.org/10.1093/database/baaa055.

19. Chen, F.Z., You, L.J., Yang, F., Wang, L.N., Guo, X.Q., Gao, F., Hua, C., Tan, C., Fang, L., Shan, R.Q., et al. (2020). CNGBdb: China national GeneBank DataBase. Yi Chuan *42*, 799–809. https://doi.org/10.16288/j.yczz.20-080.

20. McIntyre, A.B.R., Alexander, N., Grigorev, K., Bezdan, D., Sichtig, H., Chiu, C.Y., and Mason, C.E. (2019). Single-molecule sequencing detection of N6-methyladenine in microbial reference materials. Nat. Commun. *10*, 579. https://doi.org/10.1038/s41467-019-08289-9.

21. Sun, X., Hu, Y.-H., Wang, J., Fang, C., Li, J., Han, M., Wei, X., Zheng, H., Luo, X., Jia, Y., et al. (2021). Efficient and stable metabarcoding sequencing data using a DNBSEQ-G400 sequencer validated by comprehensive community analyses. Gigabyte *2021*, 1–15. https://doi.org/10.46471/gigabyte.16.

22. Dong, Z., Zhao, X., Li, Q., Yang, Z., Xi, Y., Alexeev, A., Shen, H., Wang, O., Ruan, J., Ren, H., et al. (2019). Development of coupling controlled polymerizations by adapter-ligation in mate-pair sequencing for detection of various genomic variants in one single assay. DNA Res. *26*, 313–325. https://doi.org/10.1093/dnares/dsz011.

23. Volden, R., Palmer, T., Byrne, A., Cole, C., Schmitz, R.J., Green, R.E., and Vollmers, C. (2018). Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. Proc. Natl. Acad. Sci. USA *115*, 9726–9731.

24. Adams, M., McBroome, J., Maurer, N., Pepper-Tunick, E., Saremi, N.F., Green, R.E., Vollmers, C., and Corbett-Detig, R.B. (2020). One fly–one genome: chromosome-scale genome assembly of a single outbred Drosophila melanogaster. Nucleic Acids Res. *48*, e75.

25. Wang, O., Chin, R., Cheng, X., Wu, M.K.Y., Mao, Q., Tang, J., Sun, Y., Anderson, E., Lam, H.K., Chen, D., et al. (2019). Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. Genome Res. *29*, 798–808. https://doi.org/10.1101/gr.245126.118.

26. Cheng, X., Wu, M., Chin, R., Lam, H., Chen, D., Wang, L., Fan, F., Zou, Y., Chen, A., Zhang, W., et al. (2018). A simple bead-based method for generating cost-effective co-barcoded sequence reads. Protocol Exchange. https://doi.org/10.1038/protex.2018.116.

27. Fang, C., Zhong, H., Lin, Y., Chen, B., Han, M., Ren, H., Lu, H., Luber, J.M., Xia, M., Li, W., et al. (2018). Assessment of the cPAS-based BGISEQ-500 platform for metagenomic sequencing. GigaScience *7*, 1–8. https://doi.org/10.1093/gigascience/gix133.

28. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

29. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. *17*, 132. https://doi.org/10.1186/s13059-016-0997-x.

30. Li, D., Liu, C.M., Luo, R., Sadakane, K., and Lam, T.W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics *31*, 1674–1676. https://doi.org/10.1093/bioinformatics/btv033.

31. Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F.O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. *41*, D590–D596. https://doi.org/10.1093/nar/gks1219.

32. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410. https://doi.org/10.1016/s0022-2836(05)80360-2.

33. Nilsson, R.H., Larsson, K.-H., Taylor, A.F.S., Bengtsson-Palme, J., Jeppesen, T.S., Schigel, D., Kennedy, P., Picard, K., Glöckner, F.O., Tedersoo, L., et al. (2019). The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. Nucleic Acids Res. *47*, D259–D264.

34. Abarenkov, K.Z., Allan; Piirmann, T., Pöhönen, R., Ivanov, F., and Nilsson, R.H.; Kõljalg (2020). UNITE General FASTA Release for Eukaryotes 2Version 04.02.2020 (UNITE Community). https://doi.org/10.15156/BIO/786370.

35. Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. PeerJ *4*, e2584. https://doi.org/10.7717/peerj.2584.

36. Wood, D.E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. Genome Biol. *20*, 257. https://doi.org/10.1186/s13059-019-1891-0.

37. Bengtsson-Palme, J., Ryberg, M., Hartmann, M., Branco, S., Wang, Z., Godhe, A., De Wit, P., Sánchez-García, M., Ebersberger, I., de Sousa, F., et al. (2013). Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. Methods Ecol. Evol. https://doi.org/10.1111/2041-210x.12073.

38. Nakamura, T., Yamada, K.D., Tomii, K., and Katoh, K. (2018). Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics *34*, 2490–2492. https://doi.org/10.1093/bioinformatics/bty121.

39. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973. https://doi.org/10.1093/bioinformatics/btp348.

40. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033.

41. Yu, G., Smith, D.K., Zhu, H., Guan, Y., and Lam, T.T. (2017). ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. *8*, 28–36. https://doi.org/10.1111/2041-210x.12628.

42. Team, R. (2020). RStudio: Integrated Development Environment for R. RStudio (PBC).

43. Team, R.C. (2020). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
Pseudomonas aeruginosa	ZymoBIOMICS® Microbial Community DNA Standard	Cat#D6305
Escherichia coli	ZymoBIOMICS® Microbial Community DNA Standard	Cat#D6305
Salmonella enterica	ZymoBIOMICS® Microbial Community DNA Standard	Cat#D6305
Lactobacillus fermentum	ZymoBIOMICS® Microbial Community DNA Standard	Cat#D6305
Enterococcus faecalis	ZymoBIOMICS® Microbial Community DNA Standard	Cat#D6305
Staphylococcus aureus	ZymoBIOMICS® Microbial Community DNA Standard	Cat#D6305
Listeria monocytogenes	ZymoBIOMICS® Microbial Community DNA Standard	Cat#D6305
Bacillus subtilis	ZymoBIOMICS® Microbial Community DNA Standard	Cat#D6305
Saccharomyces cerevisiae	ZymoBIOMICS® Microbial Community DNA Standard	Cat#D6305
Cryptococcus neoformans	ZymoBIOMICS® Microbial Community DNA Standard	Cat#D6305
Aspergillus ustus	*BeNa Culture Collection (BNCC)*	Cat#BNCC144426
Trichoderma koningii	*BeNa Culture Collection (BNCC)*	Cat#BNCC144774
Penicillium expansum	*BeNa Culture Collection (BNCC)*	Cat#BNCC146144
Aspergillus nidulans	*BeNa Culture Collection (BNCC)*	Cat#BNCC336164
Penicillium chrysogenum	*BeNa Culture Collection (BNCC)*	Cat#BNCC336234
Trichoderma reesei	*BeNa Culture Collection (BNCC)*	Cat#BNCC341839
Trichoderma longibrachiatum	*BeNa Culture Collection (BNCC)*	Cat#BNCC336352
Biological samples		
Soil microbiome	Nanbanhe tropical rainforest, Yunnan, China	21.612 N, 101.574 E
Chemicals, peptides, and recombinant proteins		
Kapa Hifi DNA polymerase	Roche	Cat#07958838001
EX Taq DNA polymerase	Takara Bio	Cat#RR01CM
QIAquick PCR Purification Kit	QIAGEN	Cat#28104
T4 DNA ligase	MGI	Cat#1000004279
PNK buffer	NEB	Cat#B0201S
SPRI beads purification reagent (AMPure XP Reagent)	Beckman Coulter Life Sciences	Cat#A63882
TE buffer	Thermo Fisher	Cat#AM9849
Pfu Turbo Cx	Agilent Technologies, Inc.	Cat#600414
USER enzyme	NEB	Cat#M5505S
TA buffer	Teknova	Cat#T0380
Plasmid-Safe™ DNase	Lucigen	Cat#E3110K
Phi29	MGI	Cat#1000007887
Bst 2.0 polymerase	NEB	Cat#M0538M
MGIEasy stLFR Library Prep Kit	MGI	Cat#1000005622

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
stcLFR DNA sequencing data	China National GeneBank DataBase (CNGBdb)	CNGBdb: https://doi.org/10.26036/CNP0001509
Hiseq4000 data for *E. coli*	NCBI	SRR7415647
PacBio RSII data for *E. coli*	NCBI	SRR7498044
WGBS of *E. coli*	NCBI	SRR8137545
PacBio RSII sequencing of *E. coli* K12	NCBI	SRR8154667
PacBio RSII sequencing of *E. coli* K12	NCBI	SRR8154668
PacBio RSII sequencing of *E. coli* K12	NCBI	SRR8154669
R9 MinION sequencing of *E. coli* K12, lambda phage, and mouse: Mason Lab	NCBI	SRR8154670
R9 MinION sequencing of *E. coli* K12, lambda phage, and mouse: Mason Lab	NCBI	SRR8154671
WGBS of *E. coli* K12	NCBI	SRR8154672
Oligonucleotides		
16S rDNA universal primer 27F	China National GeneBank (CNGB)	AGAGTTTGATCATGGCTCAG
23S rDNA universal primer 23S-2850R	China National GeneBank (CNGB)	CTTAGATGCCTTCAGCRVTTATC
18S rDNA universal primer SSU515Fngs	China National GeneBank (CNGB)	GCCAGCAACCGCGGTAA
28S rDNA universal primer TW13	China National GeneBank (CNGB)	GGTCCGTGTTTCAAGACG
Software and algorithms		
Metabbq	Zendo	GitHub: https://doi.org/10.5281/zenodo.7671268
Other		
Code Notes for reproduction of analyses	Zendo	GitHub: https://doi.org/10.5281/zenodo.7671262

RESOURCE AVAILABILITY

Lead contact

Further information and requests about resources and reagents should be directed to and will be fulfilled by the lead contact, Prof. Karsten Kristiansen (kk@bio.ku.dk).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Metagenomic sequencing data have been deposited into China National GeneBank Sequence Archive (CNSA)[18] of China National GeneBank DataBase.[19] Accession numbers are listed in the key resources table. Public dataset used for error type comparison are collected from the work by Alexa and co-workers.[20] The accession numbers are listed in the key resource table.
- All original codes have been deposited at Zenodo and are publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this work is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Bacterial and fungal community mock

We used ZymoBIOMICS Microbial Community DNA Standard (D6305) as a microbial mock community. This mock standard contains pooled DNA extracted from eight cultured bacteria and two cultured yeast strains. The species name, NRRL accession No. and % final composition are listed below.

- *Listeria monocytogenes*, B-33116, 12%;
- *Pseudomonas aeruginosa*, B-3509, 12%;

- *Bacillus subtilis*, B-354, 12%;
- *Escherichia coli*, B-1109, 12%;
- *Salmonella enterica*, B-4212, 12%;
- *Lactobacillus fermentum*, B-1840, 12%;
- *Enterococcus faecalis*, B-537, 12%;
- *Staphylococcus aureus*, B-41012, 12%;
- *Saccharomyces cerevisiae* (Yeast), Y-567, 2%;
- *Cryptococcus neoformans* (Yeast), Y-2534, 2%.

Fungal microbial community mock

We use an in-house prepared fungal mock community. This mock sample contains pooled DNA extracted from seven individually cultured fungi. The species name, BeNa Culture Collection accession No. and the final % composition are listed below.

- *Aspergillus ustus*, BNCC144426, 14.29%;
- *P. aeruginosa*, No. BNCC144774, 14.29%;
- *Penicillium expansum*, No. BNCC146144, 14.29%;
- *Aspergillus nidulans*, No. BNCC336164, 14.29%;
- *Penicillium chrysogenum*, No. BNCC336234, 14.29%;
- *Trichoderma reesei*, No. BNCC341839, 14.29%;
- *Trichoderma longibrachiatum*, No. BNCC336352, 14.29%.

METHOD DETAILS

Sample collection and DNA extraction

The mock bacterial DNA standard, ZymoBIOMICS Microbial Community Standard (D6305), was purchased from ZymoBIOMICS. Fungi strains were purchased from China General Microbiological Culture Collection Center. The fungal mock DNA sample consisted of equally mixed genomic DNA of 7 fungi, including *A. ustus* (No. BNCC144426), *Trichoderma koningii* (No. BNCC144774), *P. expansum* (No. BNCC146144), *A. nidulans* (No. BNCC336164), *P. chrysogenum* (No. BNCC336234), *T. reesei* (No. BNCC341839) and *T. longibrachiatum* (No. BNCC336352). The sequence of the rDNA region for each strain was rechecked by Sanger sequencing. Soil samples were randomly picked from plots in the tropical rainforest of the Nabanhe National Nature Reserve (21.612 N, 101.574 E) in Yunnan, China in 2017. Topsoil cores of a depth 0–10 cm were collected by hammering a ring knife (10 cm in diameter) into the soil at a regular grid of points every 50 m. Before sampling, all litter and loose debris above the sample points were removed from the forest floor.[21] Soil samples were immediately stored at −80°C after collection, and DNA extraction was performed within two months. DNA extraction of all samples was performed using the PowerSoilDNA Isolation Kit (Mobio) according to manufacturer's instructions. DNA concentration was measured by Qubit flex fluorometer (Invitrogen).

rDNA long fragments amplification

We initially amplified the region of bacterial 16S-23S rRNA gene and fungi full length ITS region with Kapa Hifi DNA polymerase (Roche) and EX Taq DNA polymerase (Takara Bio). PCR products failed to be generated using the Kapa Hifi DNA polymerase, probably due to its high fidelity and the fact that hybridization of the used primers to the conserved target regions may result in mismatches of several nucleotides. By contrast, EX Taq DNA polymerase (Takara Bio), which also has proofreading activity, successfully amplified the rDNA targets, and was therefore used for full length rDNA amplification. For PCR amplification we used 5′ phosphorylated primers. The primer sequences are listed below. For bacteria, the forward primer was 27F (AGAGTTTGATCATGGCTCAG) and the reverse primer was our in-house designed 23S-2850R (CTTAGATGCCTTCAGCRVTTATC). For fungi, highly universal eukaryotic primers for high eukaryotic and fungal taxonomic coverage were selected based on a previous study.[17] The forward primer was SSU515Fngs (GCCAGCAACCGCGGTAA) and the reverse primer was TW13 (GGTCCGTGTTTCAAGACG). The conditions for PCR were as follows: 1 μL of diluted template DNA, 1 μL of forward primer (10 M), 1 μL of reverse primer (10 M), 15.5 μL of nuclease-free water, 5 μL of 10X Ex Taq Buffer, 1 μL of dNTPs Mixture (2.5 mM each), and 0.5 μL of EX Taq Polymerase. We amplified samples using the following cycling conditions: 95°C for 5 min; 30 cycles of 95°C for 30s, 55°C for 30s, and 72°C for 3 min; and then a final extension at 72°C for 10 min. The amplified long amplicons were purified using QIAquick PCR Purification Kit (Qiagen).

Circularization and rolling-circle replication

To enable a highly efficient circularization of long molecules, a double stranded circularization method was applied. 100 ng amplicon were first subjected to end-repair and A-tailing according to the protocol described previously.[22] Then the product was incubated with 8 pmol of adapter and 3000 units T4 DNA ligase (MGI, 1,000,004,279) in 80 μL of 1X PNK buffer (NEB, B0201S) with extra 1 mM ATP and 7.5% PEG-8000 at room temperature for 1 h, followed by a 0.5X SPRI beads purification (Beckman, A63882) and elution with 20 μL of TE buffer (Thermo Fisher, AM9849). Next polymerase extension with 1 pmol of primer containing uracil was carried out, by adding prior heat-activated 200 units Pfu Turbo Cx (Agilent Technologies, Inc., 600,414) in 50 μL of 1X PfuCx buffer

at 72°C for 20 min, followed by a 1.5X SPRI beads purification and elution with 30 μL of TE buffer. Sticky ends were created by injecting 20 units of USER enzyme (NEB, M5505S) and incubating with 50 μL of 1X TA buffer (Teknova, T0380) at 37°C for 1 h. T4 DNA ligase-mediated circularization was performed in 150 μL of 1X TA buffer with extra 1 mM ATP. All linearized DNA were removed with 0.4 units Plasmid-Safe DNase (Lucigen, E3110K) followed by a 1X SPRI beads purification.

The double stranded circular DNA from the last step was designed with a 3nt gap on one strand acting as the initial extending site for RCR. The RCR reaction was carried out by incubation with 5 units Phi29 (MGI, 1,000,007,887) in 21 μL of 1X Phi buffer at 30°C for 1 h. In this step the original long amplicons were transformed into long concatemers with multiple copies of each long molecule. Next, unlike other studies,[23,24] we applied 60 units warm-start Bst 2.0 polymerase (NEB, M0538M) and primer extension with specific sequences (CGCTGATAAGGTCGCCATGCCTCTCAGTAC) to generate the second strand of the RCR product ready for standard stLFR library preparation.

Single-tube long fragment read barcoded DNA library construction

Extended amplicons produced after RCR were labeled by unique barcodes with MGIEasy stLFR Library Prep Kit (MGI, 1,000,005,622). Briefly, indexed transposons were inserted into 1ng of double strand RCR products from different samples, followed by hybridization of the transposon integrated DNA onto clonally barcoded beads. After capture, the sub-fragments of each transposon inserted DNA molecule were ligated to the barcode oligo. Then the excessive oligos were removed by exonuclease digestion and the second adapter was ligated to the 3′OH recessive end using branch ligation generating a product ready for PCR amplification. This method and a detailed protocol were previously described by Wang et al.[25] and Cheng et al.,[26] respectively.

Sequencing and decoding long fragments

DNA libraries were sequenced using the pair-end 100 bp mode on the BGISEQ-500 platform.[27] During sequencing, the barcode part was sequenced first and attached to the tail of read2. The barcode detection and sequence read quality control were managed by the fastp software.[28] This tool was initially designed for traditional NGS data QC, with a parallel function to speed up the process. We added a barcode detection module to it. In the detection process, each 10-base barcode sequence (three 10-base sequences make up a full barcode) was scanned against our available barcode list, both by forward and reverse strand, within 1 base mismatch tolerance. In addition, a module to perform the BGISEQ platform specific quality filtering and trimming was also implemented as described previously.[27] Reads were sorted by barcode and stored in a fastq format. More than 85% of reads were associated with a valid barcode. Less than 1% of the detected barcodes were recovered after barcode error correction (data not shown).

QUANTIFICATION AND STATISTICAL ANALYSIS

Single-molecular pre-binning and *de novo* assembly strategy

An estimation of kmer coverage was initially performed to ensure that each bead had enough coverage for an independent *de novo* assembly. The kmer coverage was determined by the formula:

$$cov_{kmer} = \frac{n_{reads} \times L_{read} - k + 1}{n_{kmer}}$$

Where n_{reads} means reads number belonging to a bead; L_{read} means the length of each read, which here is 100bp; k means the length of kmer, which is 31; n_{kmer} means the unique number of kmers from n_{reads}, here calculated by mash version 2.1.1.[29]

According to the coverage distribution, we only allowed beads with $cov_{kmer} \geq 5$ for the assembly process. Each bead's assembly was performed by megahit version 1.1.2,[30] with parameters *–k-min 21 –k-step 20 –prune-level 0 –min-count 1*.

Since amplicons were head to tail adjoined by the RCR adaptor, it is possible to assemble a circular sequence with the RCR adapter. In this case, a script to clip the RCR adapter from contigs and linearize the sequence was used.

Taxonomic annotation of pre-binning assemblies

For taxonomic assignment, PBAs were initially aligned to SILVA version 138 SSU and LSU refseq[31] by blastn,[32] separately. For fungal PBAs, UNITE[33,34] (released on Apr 2, 2020) was used as reference of ITS region. Alignment results were then summarized to provide the most probable taxonomy for each PBA according to the summary of alignments of the SSU, ITS and LSU regions, measured by sequence identity, bit score, and length coverage. PBAs with ambiguous annotation pointing to multiple different taxonomies with the same score were eliminated. Chimeric PBAs with several fragments uniquely assigned to different taxonomies were also discarded.

Default taxonomic rank threshold determination

The pairwise global alignments were performed for the SSU, LSU and ITS region sequences, separately. The SSU sequences were collected from SILVA version 138.1 SSU refseq.[31] The LSU sequences were collected from SILVA version 138.1 LSU refseq.[31] The ITS sequences were collected from UNITE[33,34] (released on Apr 2, 2020). For each region, the alignments were initiated from its top rank (species or subspecies). For each taxonomy, if multiple associated sequences existed, a pairwise global alignment was performed by vsearch v2.14.1[35] to calculate sequences identities between each two sequences, with the following command:

vsearch –allpairs_global <belonging sequence file> –acceptall –uc -

As the computing time increased exponentially with an increased number of sequences, we limited the number of sequences to no more than 100 from a given taxa to ensure computation could be finished within an acceptable time. After completion of assignment to all ranks, the sequence identity distribution of the median value of each taxon was used for visualization. For each rank, the peak position was selected manually as the estimated threshold for this rank.

Cluster tree generation

For each rank, taxonomy assignment of a given taxon was determined and organized for Kraken version 2.0.8-beta[36] database generation. Bacterial PBAs ranging from 500 bp to 1500 bp and fungal PBAs ranged from 1700 bp to 2500bp were picked. PBAs used to process clusters satisfied the following criteria: 1) target rDNA region(s) detected; 2) size of length in defined range; 3) no ambiguous annotation; 4) no chimera detected. For bacterial PBAs, barrnap version 0.9 (https://github.com/tseemann/barrnap) was used for rDNA regions detection, and the picked length ranged from 500 bp to 1500bp. For fungal PBAs, barnnap version 0.9 and ITSx version 1.0.11[37] were both used for rDNA regions detection, and the picked length ranged from 1700 bp to 2500 bp. Any PBA with partial segments aligned with conflicting annotations was discarded. Retained PBAs were then clustered by vsearch with a series of identities at 0.4, 0.5, 0.66, 0.72 (for phylum), 0.77, 0.83, 0.89, 0.92, 0.95, 0.97(for genus), 0.98, 0.99 (for species), 0.995, 0.999 and 1, according to the taxonomies similarity centroids calculated from all annotated sequences from the SILVA and the UNITE databases (See Figure S4). Each cluster with a certain identity was regarded as a clade of a taxonomy tree if the member PBAs could not provide a unified rank annotation. Note that the cutoff of rank above species varied greatly among different groups. The cutoff values were only assigned for clades without any useful information, otherwise they were determined by the clade members' classifications. Finally, at species rank, singleton OTUs without annotation were discarded in order to reduce the false discovery rate.

Relative abundance computation of each rank

The rank abundance was calculated by Kraken2, which uses kmer alignments to determine the position of each read. We then re-scaled it to barcode unit whereby reads sharing the same barcode represent a single DNA molecule. The generated results are compatible with Kraken2 so each rank's profile can be classified directly.

Phylogenetic tree building

For bacterial SSU rRNA gene sequences, barrnap version 0.9 (https://github.com/tseemann/barrnap) was used to secure quality of the PBAs sequences with no chimeras. Fungal ITS regions were predicted by ITSx version 1.0.11[37] which also enabled determination of the boundary of the SSU and the LSU. The SSU and LSU sequence were then joined together for heterogeneity rate computation. Both bacterial SSU and fungal joint SSU+LSU sequences were multiple aligned by mafft v7.407[38] and terminal gaps trimmed by trimAl v1.4.rev22.[39] RAxML version 8.2.1[40] was then employed to perform 100 boostraps General Time Reversible model of nucleotide substitution under the Gamma model of rate heterogeneity, with accommodated searches incorporated (GTRCAT). The processes were executed on a 40-core node of the Danish National Supercomputer for Life Sciences (Computerome 2.0) with following parameters:

raxmlHPC-HYBRID-AVX -f a -p 12,345 -x 12345 -T 40 -# 100 -m GTRCAT -s <input.aligned.trimed.fasta>

The results of the best-scoring ML tree with support values were imported and visualized by ggtree package[41] in Rstudio v1.3.1073[42] IDE with R 4.0.3.[43]