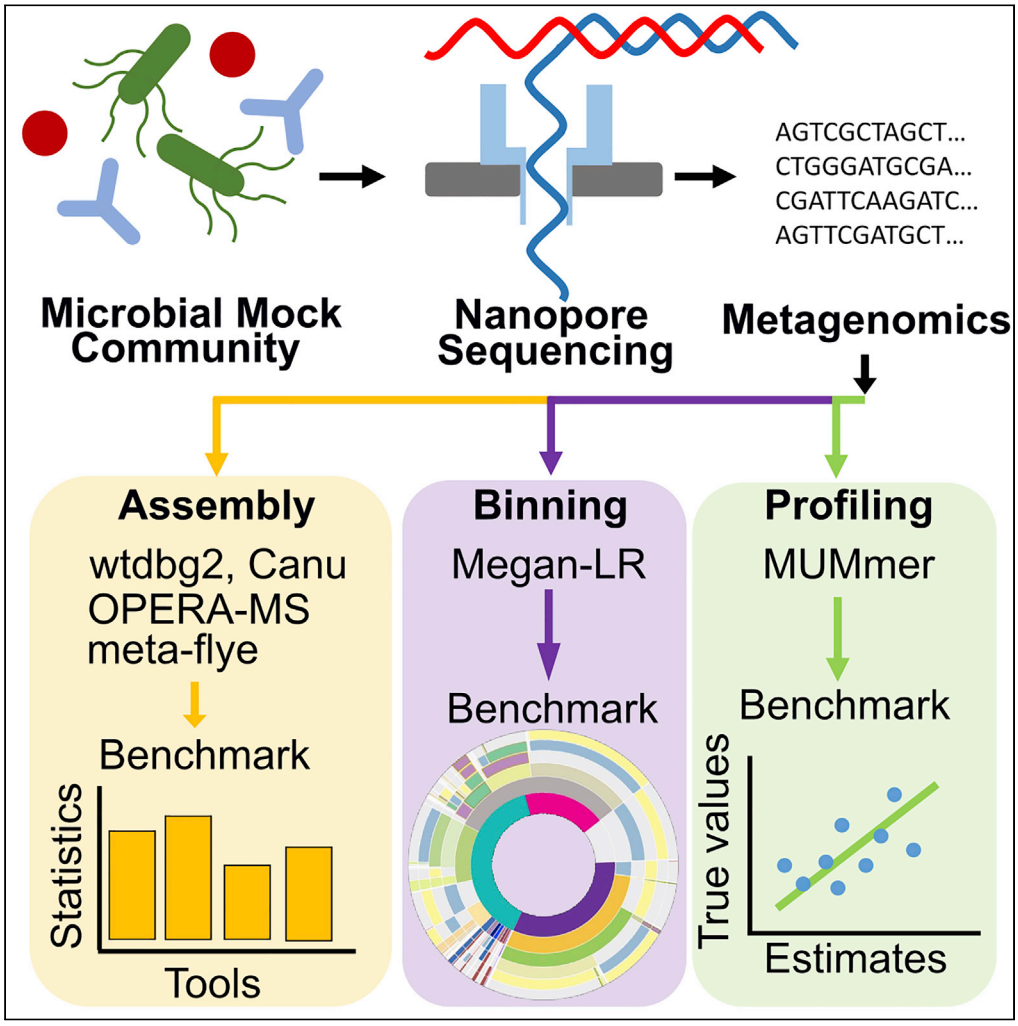
iScience

CellPress
OPEN ACCESS

Article

Implications of Error-Prone Long-Read Whole-Genome Shotgun Sequencing on Characterizing Reference Microbiomes



Yu Hu, Li Fang, Christopher Nicholson, Kai Wang

wangk@email.chop.edu

HIGHLIGHTS

First set of Nanopore long-read data on two NIH designated reference samples

Long-read assemblers achieved high accuracy (~99%) and completeness (~99%)

Nanopore sequencing provides accurate taxonomic profiling and binning

We highlighted challenges in long-read data for metagenomics tool development

Hu et al., iScience *23*, 101223
June 26, 2020
https://doi.org/10.1016/j.isci.2020.101223

iScience

CellPress
OPEN ACCESS

Article

Implications of Error-Prone Long-Read Whole-Genome Shotgun Sequencing on Characterizing Reference Microbiomes

Yu Hu,[1,4] Li Fang,[1,4] Christopher Nicholson,[1,2] and Kai Wang[1,3,5,*]

SUMMARY

Long-read sequencing techniques, such as the Oxford Nanopore Technology, can generate reads that are tens of kilobases in length and are therefore particularly relevant for microbiome studies. However, owing to the higher per-base error rates than typical short-read sequencing, the application of long-read sequencing on microbiomes remains largely unexplored. Here we deeply sequenced two human microbiota mock community samples (HM-276D and HM-277D) from the Human Microbiome Project. We showed that assembly programs consistently achieved high accuracy (~99%) and completeness (~99%) for bacterial strains with adequate coverage. We also found that long-read sequencing provides accurate estimates of species-level abundance (R = 0.94 for 20 bacteria with abundance ranging from 0.005% to 64%). Our results not only demonstrate the feasibility of characterizing complete microbial genomes and populations from error-prone Nanopore sequencing data but also highlight necessary bioinformatics improvements for future metagenomics tool development.

BACKGROUND

The fundamental importance of microbiota as the microbial communities that reside in human body is increasingly recognized. Over the past decade, there have been tremendous amounts of evidence suggesting that microbiota plays a crucial role in human health through modulating the metabolic functions, as well as food energy harvest and storage. Microbiota, especially the gut microbiota, is associated with many chronic diseases such as obesity, diabetes, metabolic syndrome, inflammatory bowel disease (IBD), irritable bowel syndrome (IBS), liver disease, and hepatocellular and colorectal carcinoma (Gill et al., 2006; Lewis et al., 2015; Chehoud et al., 2015; Hooper et al., 2003; Jones et al., 2015; Ley et al., 2006; Liang et al., 2015; Sartor, 2008; Schauber et al., 2003; Turnbaugh et al., 2009; Wang et al., 2016; Wen et al., 2008; Wu et al., 2011; Group et al., 2009). Therefore, accurate profiling of complete genomes and population is crucial to understanding the impact of microbiota on human health. Currently, high-throughput sequencing technologies have been widely used in microbial community characterization. In particular, 16S ribosomal RNA (rRNA) (Janda and Abbott, 2007) and shotgun metagenome sequencing on Illumina platforms (Quince et al., 2017) are two dominant approaches for describing microbiomes. Overall, the high-throughput nature of metagenomics sequencing allows us to interpret microbial community by using computational approaches such as operational taxonomic unit (OTU) identification (Hao and Chen, 2012), abundance quantification (Chen et al., 2017), read assembly (Ruan and Li, 2019; Bertrand et al., 2019; Koren et al., 2017; Kolmogorov et al., 2019; Li et al., 2015), and binning and taxonomic profiling (Gregor et al., 2016; Huson et al., 2016, 2018; Francis et al., 2013; Hong et al., 2014; Byrd et al., 2014). Specifically, 16S rRNA sequencing targets on very specific regions that are highly variable between species, which is much cost-efficient. This is very useful for us to examine and compare the microbiota across a high number of samples in a large-scale project. However, this technique can only identify bacteria but not viruses or fungi, and the low resolution limits its usage in microbiome study below the genus level. As opposed to only the 16S sequences, shotgun metagenome sequencing surveys the whole genomes of all organism in the community (Jovel et al., 2016; Laudadio et al., 2018; Ranjan et al., 2016). It allows us to perform deep investigation of the microbial community as its ability to capture sequences from all organisms.

Despite the theoretical advantage of shotgun metagenome sequencing, owing to the short read length (150–300 nucleotides), metagenomes cannot be fully characterized by next-generation sequencing (NGS) data. In addition, the lack of contextual information has become a barrier for short read to span

[1]Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

[2]Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA

[3]Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

[4]These authors contributed equally

[5]Lead Contact

*Correspondence: wangk@email.chop.edu

https://doi.org/10.1016/j.isci.2020.101223

both intra- and intergenomic repeats, which is crucial for complete *de novo* genome assembly of all dominant species in a microbial community. As a consequence, short-read assemblies remain highly fragmented. In comparison, the use of long-read sequencing has the potential to facilitate the complete and contiguous metagenome assembly. Lee et al. (2014) sequenced a reference mock community sample using PacBio long read and evaluated the metagenome assembly performance. Results showed that single-molecule real-time (SMRT) long-read data offered significantly improved assembly contiguity by spanning many of repetitive regions, whereas single bacterial chromosome was assembled to more than 50 contigs based on short-read data. In recent years, the Oxford Nanopore technologies (ONTs) have offered advantages over traditional short-read NGS technologies in genome study. This single-molecule sequencing platform is able to generate average read length of >10 kbp, spanning low complexity and repetitive genomic regions, which provides much more continuous assemblies. Subsequently, this approach has become an attractive option in metagenomics sequencing. Although the ONTs have great potential, complete and contiguous *de novo* metagenome assembly is still constrained by the high error rate (~15%) of single-molecule long-read sequence data (Sczyrba et al., 2017). Therefore, a comprehensive evaluation of long-read bioinformatics tools in microbial profiling is needed (Mason et al., 2017). Nicholls et al. (2019) presented Nanopore sequencing datasets of two mock communities with 10 microbial species from ZymoBIOMICS (McIntyre et al., 2019). They showed the utility of these datasets for future bioinformatics method development for long-read metagenomics. However, publicly available datasets based other sequencing technologies of these samples are limited as the samples are only commercially available and are not well studied so far by competing approaches. A study to evaluate the advantages of Nanopore sequencing in complete microbial genomes and a comparison over other sequencing technologies is still lacking so far.

In this article, we generated two deeply sequenced Nanopore datasets from new reference samples that are more commonly studied and performed comprehensive analysis to compare microbial community profiling performance with PacBio and Illumina technologies. We first generated 525× coverage data on HM-276D mock community sample from Human Microbiome Project, which is an evenly mixed DNA sample of 20 bacterial strains (each with 5% abundance). We performed *de novo* assembly analysis with four long-read assemblers at different depth of coverage. Twenty bacterial genomes were assembled with high accuracy and genome completeness. This sample also has been well studied by many groups. As mentioned above, Lee et al. (2014) sequenced this mock community with PacBio to show the improvement of long-read data in metagenome assembly analysis. Jones et al.(2015) compared the influence of different NGS platforms on genomic and functional predictions using HM-276D sample. We downloaded these two datasets and compared the performance with Nanopore data. Our results show that Nanopore improved assembly contiguity compared with PacBio and Illumina across computational approaches. Next, we sequenced HM-277D Mock Community sample with 1,068× coverage. HM-277D is unevenly mixed DNA sample of 20 bacterial strains. Kuleshove et al. (2016) sequenced this sample with Illumina TruSeq synthetic long-read technique and showed the improvement in bacterial species identification, genome reconstruction compared with short sequences. Also, Leggett et al., 2020 demonstrated Nanopore metagenomics sequence can be reliably classified using this community. In addition to metagenome assembly, we evaluated taxonomy binning and profiling performance across technologies (Nanopore and PacBio) and samples (HM-276D and HM-277D). High identification and classification accuracy were achieved above the species level. Overall, we demonstrate the technical feasibility to characterize complete microbial genomes and populations from error-prone Nanopore sequencing without any DNA amplification. We also discuss the limitations of current bioinformatics tools, when dealing with error-prone long-read metagenomics sequencing data. All our data are made publicly available, to benefit computational tool development on long-read-based microbial genome assembly for metagenomics studies.

RESULTS

Sequence Data Quality

HM-276D DNA sample includes 20 evenly mixed bacterial strains with reference genome size 70 Mb in total with 39 chromosomes. A total of 11,610,183 reads with 35,578,375,166 bases (525× coverage depth) were generated on the Nanopore GridION platform, with a median length of 1,374 bp. The N50 length is 6,828 bp, and median read quality is 9.39 in Phred scale. By using minimap2, 95% of reads were successfully aligned to reference genomes of 20 bacterial strains with 13.1% error rate (Table 1). As shown in Figure 1A, read coverage across 20 bacterial strains has good agreement with known abundances. Read depth is relatively homogeneous across bacterial strains with 521.9X (sd = 524.7X) in average. Sequencing depth of each strain is at least 150 reads and only 0.03% region is covered by less than 3 reads.

Mapping Statistics	HM-276D	HM-277D
# of reads	8,086,684	18,254,839
# of mapped reads	7,640,934	18,110,317
Reads unmapped	445,750	144,522
Reads MQ0	60,972	103,601
Non-primary alignments	287,369	732,671
Total length	33,563,573,383	72,312,638,112
Bases mapped	32,143,689,158	72,216,146,980
Bases mapped (cigar)	31,156,025,998	70,073,211,829
Mismatches	4,104,593,752	6,925,222,080
Average length	4,150	3,961
Maximum length	472,762	214,792
Average Phred quality per base	13	17

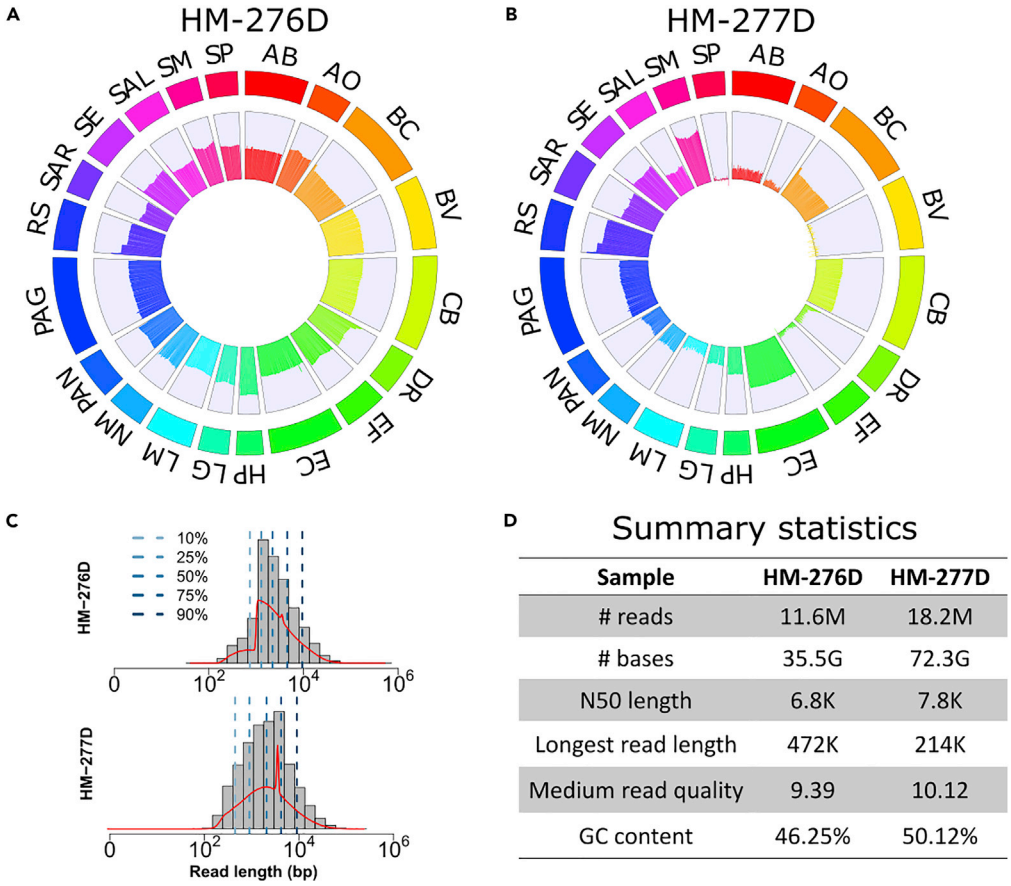
Table 1. Mapping Statistics of HM-276D and HM-277D Sequenced Dataset

Sequenced data were mapped against reference genomes of 20 known bacterial strains. Sequences indicate the number of QC passed reads. Number of mapped and unmapped reads were summarized. MQ0 represents number of mapped reads with MQ = 0. Clipping was ignored when calculating total length, bases mapped. Bases mapped (cigar) provides a more accurate number of mapped bases. Number of mismatches were obtained from NM field of BAM file.

HM-277D DNA sample includes 20 unevenly mixed bacterial strains. A total of 18,254,839 reads dataset with 72,312,638,112 bases (1,068× coverage depth) were generated, leading to 2,065 bp in median read length with 10.12 median read quality. The N50 length is 7,857 bp; 99.2% of QC-passed reads were mapped to the reference genome and the error rate was 9.8% (Table 1). As shown in Figure 1B, read distribution is more heterogeneous across strains due to unevenly mixed samples. The average coverage is 988.8 reads with standard deviation 1941.6 bp. This leads to 1.6% of region with less than three reads covered and four strains with sequencing depth less than 10 bp, which makes it more difficult for biological interpretation of this microbial community (Figures 1C and 1D).

De Novo Assembly of HM-276D Mock Community

To assess the ability of Nanopore sequencing in profiling microbial community, we first conducted a *de novo* assembly of dataset with 525× coverage from HM-276D mock community using four assemblers: wtdbg2 (Ruan and Li, 2019), OPERA-MS (Bertrand et al., 2019), Canu (Koren et al., 2017), and meta-flye (Kolmogorov et al., 2019). OPERA-MS and meta-flye are designed to be capable of handling metagenome data, whereas wtdbg2 and canu are broadly used for haploid or diploid genomes. Overall, the results show promise for the characterization of microbial genomes using long-read sequencing data. Canu produced the largest assembly of 69.5 Mb (99.3% of the benchmark data), including 83 contigs with contig N50 length of 3.91 Mb. meta-flye assembled 67.7 Mb genome with 89 contigs. wtdbg2 generated similar results with 64.9 Mb genome size, 61 contigs, and 2.97 Mb N50 length. Assembly metrics of OPERA-MS (67.9 Mb genome size, 4,734 contigs with contig N50 length of 2.94 Mb) are similar with Canu and wtdbg2, whereas much more contigs were generated because OPERA-MS utilizes both long and short sequencing reads for assembly. In addition, for aligned blocks, meta-flye yielded the highest NA50 with 1.71Mb in length compared with other assemblers (wtdbg2: 1.2Mb, OPERA-MS: 1.21Mb, Canu: 1.4Mb). Furthermore, by mapping all contigs to the reference genomes using MUMMer v3.23, we assessed the accuracy and genome completeness of contigs produced by four assemblers. As shown in Figure 2A, meta-flye achieved the highest genome fraction (99.99%) and one-to-one identity percentage (99.62%), followed by OPERA-MS (genome fraction: 99.98% and accuracy 99.92%), Canu (genome fraction 99.81% and accuracy 99.4%), and wtdbg2 (genome fraction 96.02% and accuracy 98.60%). Moreover, we evaluated aligned blocks for each method based on NA50 length. As shown in Table S1, meta-flye achieved the highest NA50 with 1.71 Mb in length compared with other assemblers (wtdbg2: 1.2Mb, OPERA-MS: 1.21Mb,



Sample	HM-276D	HM-277D
# reads	11.6M	18.2M
# bases	35.5G	72.3G
N50 length	6.8K	7.8K
Longest read length	472K	214K
Medium read quality	9.39	10.12
GC content	46.25%	50.12%

Figure 1. Summary of Nanopore Sequencing Data from HM-276D and HM-277D Microbial Communities

(A–C) (A and B) Circos plots of read coverage across whole genome of 20 bacterial strains from (A) HM-276D and (B) HM-277D. Each chromosome was divided into bins with 5,000 bp width. Average read coverage was calculated within each bin and converted to log scale to facilitate viewing and comparing between bacterial strains. AB, *Acinetobacter baumannii*; AO, *Actinomyces odontolyticus*; BC, *Bacillus cereus*; BV, *Bacteroides vulgatus*; CB, *Clostridium beijerinckii*; DR, *Deinococcus radiodurans*; DF, *Enterococcus faecalis*; EC, *Escherichia coli*; HP, *Helicobacter pylori*; LG, *Lactobacillus gasseri*; LM, *Listeria monocytogenes*; NM, *Neisseria meningitides*; PAN, *Propionibacterium acnes*; PAG, *Pseudomonas aeruginosa*; RS, *Rhodobacter sphaeroides*; SAR, *Staphylococcus aureus*; SE, *Staphylococcus epidermidis*; SAL, *Streptococcus agalactiae*; SM, *Streptococcus mutans*; SP, *Streptococcus pneumonia*. (C) Read length distribution of HM-276D and HM-277D datasets. Blue dashed lines represent different quantiles. Red line represents the density of read length distribution.

(D) Summary statistics of HM-276D and HM-277D datasets. Each value was calculated by using pycoQC (Leger and Leonardi, 2019) and LongreadQC. Real-time statistics are shown in Figures S1–S5.

Canu: 1.4Mb). Overall, four tools generated results with similar good quality in term of contiguity, accuracy, and completeness using long-read data with evenly mixed samples at 525× coverage depth.

Next, we subsampled 525× dataset to 365× (70%), 160× (30%), 80× (15%), 40× (7.5%), and 20× (3.75%) to examine the effect of sequencing depths on *de novo* assembly (Figure 2A, Table S1). The assembly results of four tools ranges 95.95%–99.96% in consensus accuracy and 91.26%–99.99% in genome fraction. In specific, OPERA-MS outperforms others with the highest and most consistent metrics for completeness and accuracy across different sequencing depths because its metagenomics design substantially improves the robustness to low sequencing depth, where genome fractions are 99.68% in average (sd = 0.61%) and consensus identities are 99.92% in average (sd = 0.05%). In spite of reduced metrics as the sequencing depth becomes lower, meta-flye and Canu still recovered at least 96.8% genomes with 98.5% accuracy. Notably, wtdbg2 improved the assembly metrics with coverage depth reduced from 365× to 80×. In addition, we examined whether genomes of 20 bacterial strains can be better constructed with Nanopore

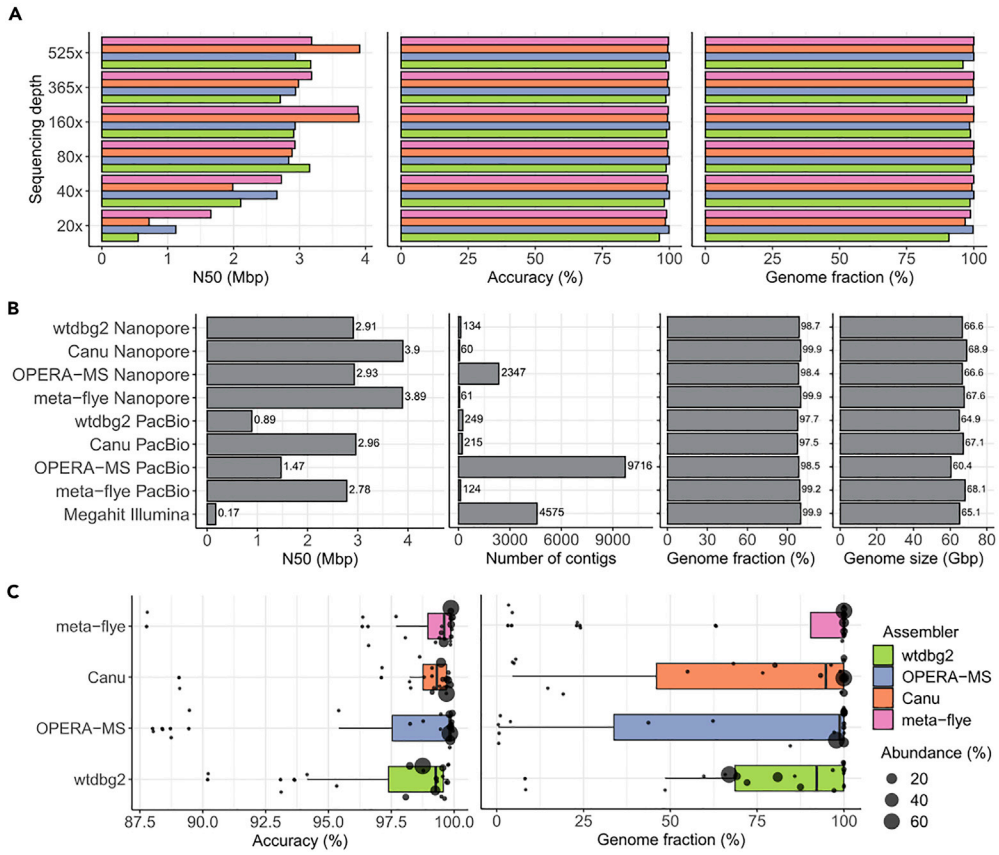


Figure 2. Assembly Results for HM-276D and HM-277D Datasets

(A) Assembly statistics (N50 length, accuracy, and genome fraction) of each assembler at different coverage depths based on HM-276D dataset. Colors indicate results from different assemblers (see Supplemental Information for details in parameter settings).

(B) Assembly statistics (N50 length, number of contigs, genome fraction, and genome size) of each assembler based on HM-276D sample sequenced by different technologies (Nanopore, PacBio, Illumina). To make fair comparison, each dataset was downsampled to 160× depth of coverage.

(C) Strain-specific assembly performance of each assembler based on HM-277D dataset. Assembly statistics (accuracy and genome fraction) distributions were presented using boxplots with jitter. Radius of each dot indicates the known relative abundance of each bacterial strain from the mock community.

sequencing technology compared with PacBio and Illumina. As shown in Figure 2B, assemblers using Nanopore sequenced data outperforms other two technologies. With the same assembler, on average, the number of contigs of Nanopore is ~30% lower than that of PacBio; genome fraction and genome size are 1.56% and 3.1 Mb higher, respectively. To understand the reason, we compared read length characteristics between these two datasets. For N50 length, ONT (7,350 bp) is 15% longer than PacBio (6,357 bp), and for N05 length, ONT (35.9 kbp) is even 159% longer than PacBio (13.8 kbp). This indicates that read length is the main advantage for ONT. Therefore, as shown in Figure 2B, N50 length of ONT (13.3 Mbp) is 68% longer than PacBio (8.1 Mbp). Assemblies using Illumina sequenced data have 99.9% accuracy, but more contigs generated and lower genome size in total compared with Nanopore.

De Novo Assembly of HM-277D Mock Community

To evaluate the metagenome reconstruction in a more realistic setting, we carried out another *de novo* assembly of 1,068× dataset from HM-277D Mock Community, with unevenly mixed DNA samples of the 20 bacterial strains (Figure S6). Assembly accuracy still remains high, ranging from 97.78% to 99.75% across tools. However, not surprisingly, genome fractions and genome sizes of all methods are substantially lower than even community. This is because 13 bacterial strains have extremely low abundances (<1%) in this unevenly mixed samples, leading to reduced genome coverage fractions (Canu: 71.68%, OPERA-MS: 71.25%,

meta-flye: 91.57%, wtdbg2: 82.95%) and genome sizes (Canu: 50.21 Mb, OPERA-MS: 47.99 Mb, meta-flye: 64.12 Mb, wtdbg2: 61.75 Mb). To assess how strain abundance affects assemblies, we calculated strain-specific genome fraction for each tool. Across bacterial strains, meta-flye recovered the highest percentage of genome (median 100%), followed by OPERA-MS (median: 98.75%), Canu (median 94.78%), and wtdbg2 (median: 91.66%) (Figure 2C). For bacteria with relative abundance higher than 0.2%, least 99.99% of reference genome can be covered by assembly contigs (meta-flye), with identity consensus reaching to 99.93%. These results suggest that bacterial strain with nontrivial abundance can be accurately assembled with Nanopore sequenced data. Overall, we observed that meta-flye returned assemblies for 20 bacterial strains with the best performance in completeness and accuracy. Metric for each strain is correlated with abundance of the corresponding bacteria. Some strains were proved hard to assemble for all assemblers due to extremely low relative abundance. For example, 13.6% of region of *Enterococcus faecalis* (0.011% relative abundance) were covered by 0 or 1 read and 56.1% covered by less than 3 reads, leading to 4.47% genome fraction for meta-flye. Moreover, there were 2 contigs that belong to two different bacterial species, *Bacteroides vulgatus* (0.19% relative abundance) and *Streptococcus pneumoniae* (0.05% relative abundance), indicating the difficulty in differentiating one bacterium from another with low relative abundance.

Taxon Binning and Identification

Metagenome assemblers construct contigs with variable length to recover original genome of each bacteria from microbial community. Subsequently, another major challenge in studying the identity and diversity of this community member is to classify sequenced reads or contigs correctly according to their taxonomic origins. Here we investigated the taxonomic binning performance based on three scenarios of long-read sequencing data, HM-276D (Nanopore, PacBio) and HM-277D (Nanopore) at 160× depth of coverage, using a state-of-art taxonomic binner Megan-LR. First, all long reads were aligned to NCBI-nr database. Then, we used Megan-LR with interval-union LCA algorithm to assign ~2 million aligned reads (~4.6 Mb bases) to taxonomic nodes (Figures 3A, 3B, and Figures S7–S10). Overall, 4.22 Mb (0.087%) from Nanopore data of HM-276D sample were mis-assigned, whereas 4.37 Mb (0.075%) and 4.66 Mb (0.141%) for Nanopore data of HM-277D and PacBio data of HM-276D, respectively. Specifically, we evaluated the recovery of taxon bins at different ranks. We considered two metrics to quantify the read assignment accuracy, average precision, and sensitivity of 20 bacterial strains. For each taxonomic bin, we obtained precision by calculating the percentage of reads correctly classified out of all binned reads. Sensitivity is the percentage of correctly assigned reads out of all reads originally from the bin. As shown in Figure 3C, HM-276D (Nanopore) has the highest precision, which are all above 60% from phylum to genus. HM-277D (Nanopore) followed, with all above 50%, whereas HM-276D (PacBio) has the lowest average precision due to predicted small false-positive bins at the species level. Sensitivity has a similar pattern (Figure 3D). HM-276D (Nanopore) still appears to be the best dataset for read classification than the other two, and the difference in accuracy between these three scenarios is similar across ranks. Nanopore is ~8% higher than PacBio and HM-276D is 10% higher than HM-277D. To evaluate the stability of read assignment accuracy, we calculated 95% confidence interval of precision and sensitivity for each scenario at each rank. Not surprisingly, confidence bands are narrower at higher rank, indicating that more taxon recovery accuracy can be reached. Owing to unevenly mixed bacterial strains, sensitivity is much more variable for HM-277D than other HM-276D. Overall, these results demonstrated the advantage of long-read data in accurate taxon recovery above the family level, whereas binning accuracy and stability were relatively at the species level.

In addition to assigning sequence fragments (reads or contigs) to taxon bins, we recognized the importance of accurate determination of taxonomic identity presence or absence from microbial community. Therefore, we continued to investigate the performance of taxonomic identity prediction between data from HM-276D (Nanopore, PacBio) and HM-277D (Nanopore). For taxon prediction, we defined that the species is significantly present in the community when at least 10 reads were assigned to it, whereas identity with less than 10 supporting reads was marked as absence. We considered two other metrics to quantify the detection accuracy, true-positive rate (TPR), and false discovery rate (FDR), where TPR is the percentage of correctly predicted taxonomic identities out of known existing taxon and FDR is the percentage of incorrectly predicted taxonomic identities out of all predicted taxon. TPR and FDR were calculated at different ranks in Figure 3E. TPRs were consistent across three datasets from phylum to order level (90%–77%). Below the order level, PacBio (HM-276D) and Nanopore (HM-277D) are 22% lower compared with Nanopore (HM-276D) (92%–87%). From phylum to family level, FDRs were controlled under 15% for all three datasets.

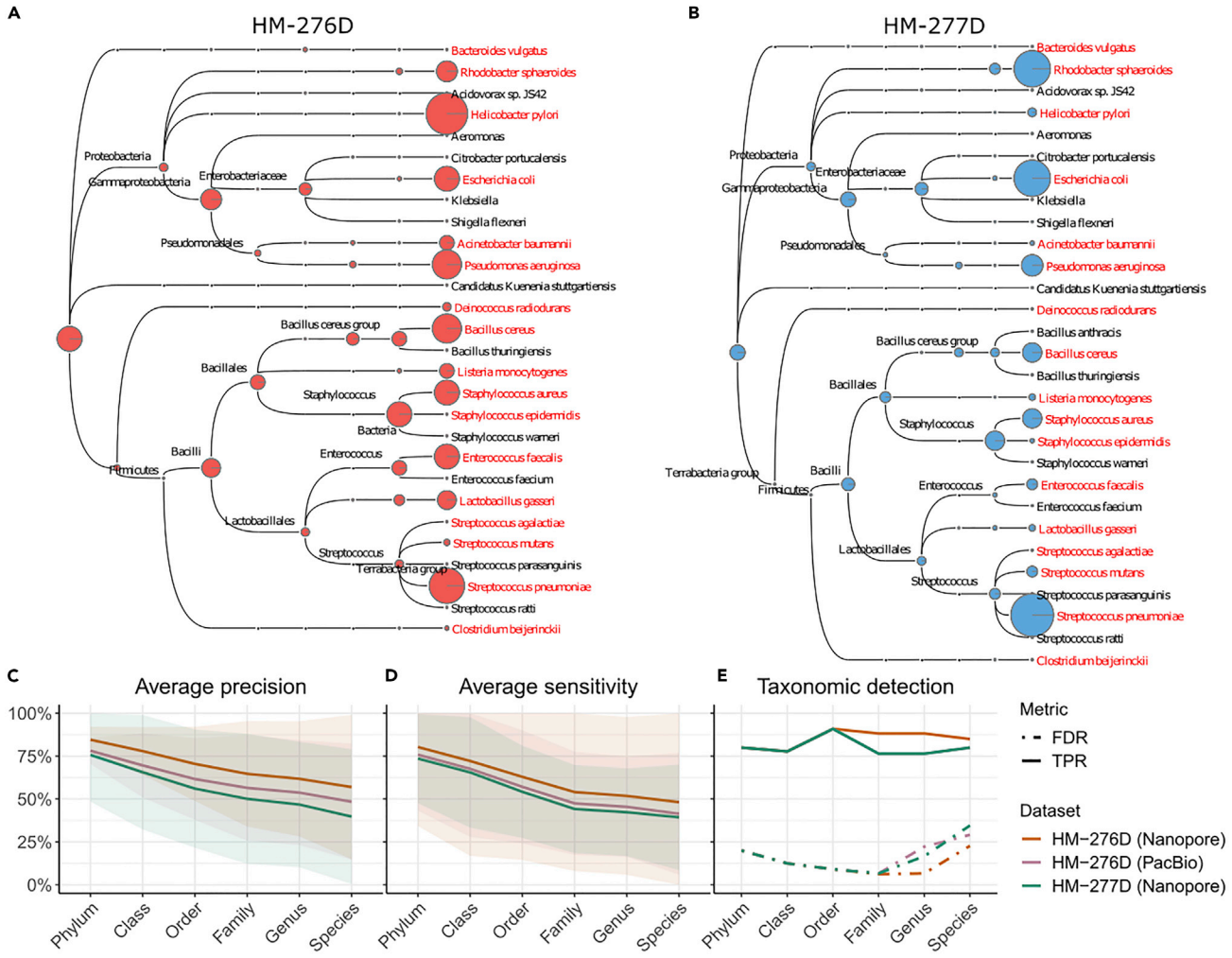


Figure 3. Taxonomic Binning Results for HM-276D and HM-277D Datasets

(A and B) Megan taxonomic tree assignment obtained from HM-276D (A) and HM-277D (B) Nanopore sequenced datasets. Both datasets were downsampled to 160× depth of coverage. Each read was aligned against NCBI-nr protein reference database, then binned and visualized using Megan-LR. Megan taxonomic tree showing bacterial taxa identified and their corresponding abundances across taxonomic rank. The radius of circle represents the number of reads assigned for each taxon. Bacterial strains highlighted in red represent true organisms in the mock community.

(C–E) Taxonomic binning and identification performance metrics across ranks based on different datasets (indicated by colors). Average precision (C), average sensitivity (D), and their 95% CIs were calculated based on metrics from different taxon at each rank. (E) Taxonomic detection accuracy metrics, true-positive rate (solid), and false-positive rate (dashed), were calculated based on identified taxon (reads >10) at each rank. To make fair comparison, each dataset was downsampled to 160× depth of coverage.

However, at the genus level, more than 20% of detections are false for PacBio (HM-276D) and Nanopore (HM-277D), whereas it was 6% for Nanopore (HM-276). All three scenarios have inflated FDR (>20%) at the species level. Across datasets, there was drastic increase in FDR between phylum to family level and below family level, 10% ± 3% and 21% ± 5%. Similar to binning results, Nanopore data of HM-276D still consistently performed better than other two datasets across ranks. However, accurately predicting taxonomic profiles at the species level still remains challenging owing to many false predicted taxonomic identities with 10–100 reads assigned incorrectly.

Strain Profiling

Despite the challenges in assembly and binning of HM-277D microbial community even at the species level, especially for low abundance bacteria (relative abundance <1%), the golden standard profile of this mock community still allows us to evaluate other unique advantages of this deeply sequenced dataset at strain level. First, we examined the ability in identifying these 13 extremely rare strains based on

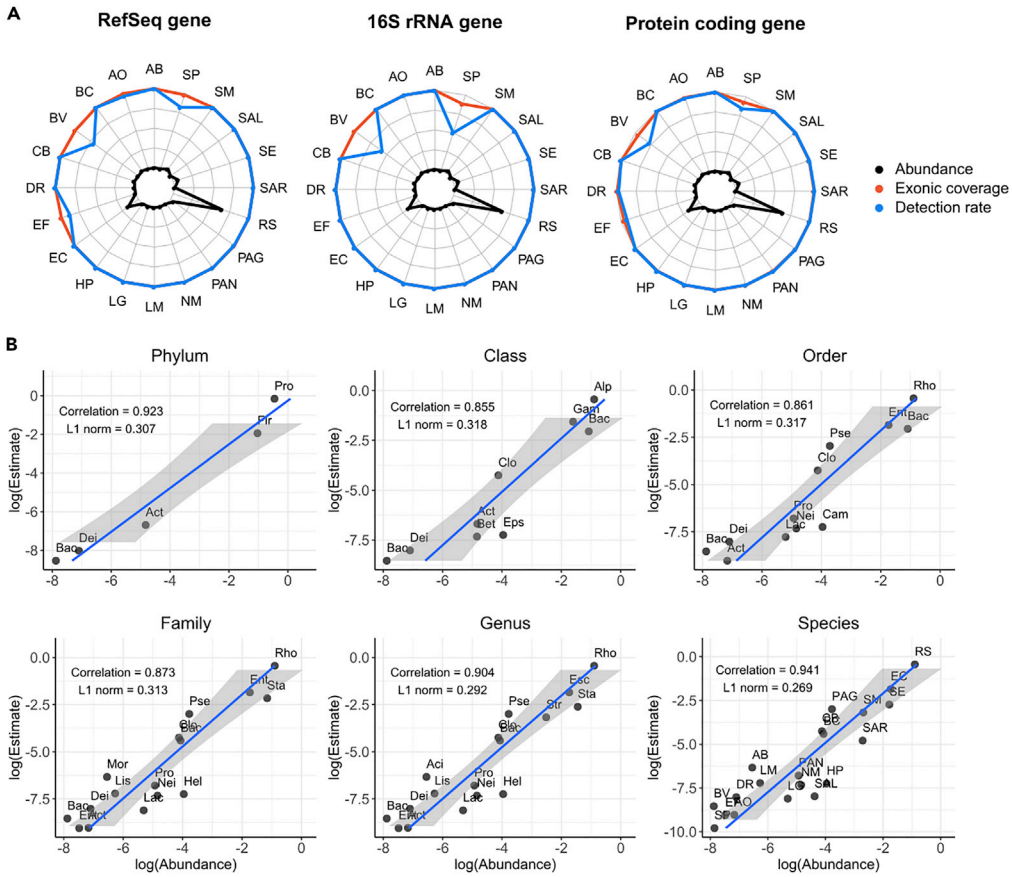


Figure 4. Taxonomic Profiling Results for HM-277D Datasets

(A) Gene identification performance of 20 bacterial strains. Three gene sets (RefSeq, 16S rRNA, protein coding) were evaluated. Colors indicate different metrics (exonic coverage and detection rate). Exonic coverage (orange) is the percentage of exonic region covered by at least one readout of all exons. Detection rate (blue) is the percentage of genes with coverage depth >1 and exonic coverage >50% out of all genes. Gold standard abundance of each strain was indicated in black.

(B) Bacterial abundance estimation. Scatterplots abundance estimates versus gold standard abundances from HM-277D mock community across taxonomic ranks. Abundances were converted to log scale to facilitate viewing. Pearson correlation and L1 norm were utilized to quantify the performance. Estimates consistently share a good agreement with gold standard across ranks with correlation >0.85 and L1 norm <0.32. Abbreviations for bacterial name above the species level are listed below. Phylum level: Actinobacteria, Bacteroidetes (Bac), Deinococcus-Thermus (Dei), Firmicutes (Fir), Proteobacteria (Pro); Class level: Actinobacteria (Act), Alphaproteobacteria (Alp), Bacilli (Bac), Bacteroidia (Bact), Betaproteobacteria (Bet), Clostridiales (Clo), Deinococcus (Dei), Epsilonproteobacteria (Eps), Gammaproteobacteria (Gam); Order level: Actinomycetales (Act), Bacillales (Bac), Bacteroidales (Bact), Campylobacterales (Cam), Clostridiales (Clo), Deinococcales (Dei), Enterobacteriales (Ent), Lactobacillales (Lac), Neisseriaceae (Nei), Propionibacteriaceae (Pro), Pseudomonadales (Pse), Rhodobacterales (Rho); Family level: Actinomycetaceae (Act), Bacillaceae (Bac), Bacteroidaceae (Bact), Clostridiaceae (Clo), Deinococcaceae (Dei), Enterobacteriaceae (Ent), Enterococcaceae (Ent), Helicobacteraceae (Hel), Lactobacillaceae (Lac), Listeriaceae (Lis), Moraxellaceae (Mor), Neisseriaceae (Nei), Propionibacteriaceae (Pro), Pseudomonadaceae (Pse), Rhodobacteraceae (Rho), Staphylococcaceae (Sta); Genus level: Acinetobacter (Act), Actinomyces (Act), Bacillus (Bac), Bacteroides (Bact), Clostridium (Clo), Deinococcus (Dei), Enterococcus (Ent), Escherichia (Esc), Helicobacter (Hel), Lactobacillus (Lac), Listeria (Lis), Neisseria (Nei), Propionibacterium (Pro), Pseudomonas (Pse), Rhodobacter (Rho), Staphylococcus (Sta), Streptococcus (Str).

annotated target genes. To explore the sensitivity of strain detection using this dataset, we mapped raw sequenced reads to reference genomes of the 20 bacterial strains with Minimap2. Then, for each strain-specific gene, the average coverage was estimated by summing up read depth across all exonic region, normalized for gene length. In addition, exon coverage fractions were calculated. We required a gene with average coverage greater than 1 and exon coverage fraction greater 50% simultaneously in order to be declared as a detected gene. The results are shown in Figures 4A, S11, and S12. Detection rates

and average coverage among all genes largely keep high in abundant strains (>1%), ranging from 96.4 bp to 4,207.6 bp, as well as most of rare strains (<1%). Most of bacterial strains except for *Bacteroides vulgatus* (69.1%) and *Streptococcus pneumoniae* (81.7%) have achieved at least 97% gene detection rate.

Next, we recognized that 16S rRNA genes are most commonly used as gene marker for bacteria identification; we further selected them out for each strain based on RefSeq annotation. As shown in Figure 4A, although *Bacteroides vulgatus* and *Streptococcus pneumoniae* still have about 50% of 16S rRNA genes undetected by raw sequenced reads, 18 strains have 100% detection rates and exon coverage fraction with 434.77 bp coverage in average, which demonstrates the feasibility of identifying rare strain (<1%) in microbial community with long-read sequencing data. Additionally, read coverage of protein coding genes for 20 bacterial strains was summarized, which shows similar results. Fourteen strains have average coverage above 100 bp and gene detection rates for 18 strains have reached to 99%, indicating the presence of bacterial strains in the sample.

To understand the composition, diversity, and spatial dynamics of microbial communities, we continued to evaluate the bacterial abundance estimation accuracy based on Nanopore data. We determined two abundance metrics to measure the accuracy, Pearson correlation, and L1 norm. These two metrics assess how well Nanopore sequenced reads can reconstruct the bacterial abundances in comparison with the gold standard. Relative abundance was obtained by normalizing total read coverage with chromosome length for each taxon at different ranks. As shown in Figure 4B, abundance estimates at the species level agrees well with the known relative abundances from the mock community. However, abundance estimation at higher ranks appears to be more challenging, as correlation coefficient ranges from 0.87 to 0.85 and L1 norm is above 0.3 from class to family level, whereas two metrics improved with Pearson correlation >0.9 and L1 < 0.29 when rank is below the family level. Poor abundance estimation at class or family level may be due to the presence of extremely rare bacterial strains in the HM-277D sample, as read coverages were simply summed up between species belonging to the same family or class without accounting for abundance heterogeneity.

DISCUSSION

Complete genome assembly and population profiling are critical for the interpretation of microbial community diversity. However, a benchmarking long-read dataset with consistent evaluation metrics is still lacking, which has hindered our understanding of long-read sequence data in metagenome assembly. In this study, we deeply sequenced HM-276D and HM-277D samples to assess the performance of error-prone Nanopore sequencing data and bioinformatics tools in characterizing microbial community. Assemblers consistently achieved high accuracy and completeness for nontrivial bacterial strains, and genome binners performed well at above the genus level. Furthermore, by targeting on marker genes, we were able to identify rare strains with extremely low abundance in microbial community. Overall, our results have demonstrated the technical feasibility to characterize complete microbial genomes and populations from Nanopore sequencing data with metagenomic software.

We note that, despite the feasibility to characterize complete microbial genomes from long-read sequencing data, there are still challenges to be resolved in our study. Even for evenly mixed samples, the best performing assembler meta-flye achieve 99.99% consensus accuracy. However, as the reference genomes contains 70 Mb, 0.04% error rate has led to 28 kbp of mismatches. These erroneous bases could be due to sequencing errors in low-quality read, a major drawback of long-read sequence data and base modification, which may complicate the genome assembly. To prevent these errors, a sequencer with unbiased and methylation-aware base caller is in need. (We also acknowledge that some of the mismatches may be due to natural differences between reference microbiome samples and the reference genomes that were used.) In addition, there is still room for further improvement in assembly completeness by using longer reads or better designed assemblers to account for long repeats in genomes. In our study, we assembled long-read sequenced data from 20 bacterial strains across species. However, the performance at strain-level still remains unknown as closely related genomes are always a major challenge for genome assembly. In the future, we anticipate that more mock microbial community will be released with bacteria at strain level for benchmarking study.

By evaluating the performance of bioinformatics tools across different technologies, we found that third-generation sequencing generally facilitates the complete characterization of complex bacterial genomes by overcoming many limitations of second-generation sequencing. The short read length has limited the

ability of Illumina sequencing in genome interpretation. For example, the length of repetitive genomic region is larger than a single read. As a consequence, intra- and intergenomic diversities are unlikely to be captured by short sequencing data. This issue has been resolved by long-read sequencing technologies (ONT and PacBio), which is able to span low complexity and repetitive regions by providing sequence reads with at least 10 kb in length. While generating data with much higher error rate than PacBio, ONT has become a promising platform in many applications, especially for studies requiring large amounts of data. This is because ONT provides longer reads (up to 900 kb in length) with higher throughput compared with PacBio (10–15 kb in length). Moreover, ONT is currently more affordable with lower per-base cost of data generation, which is a key factor in long-read sequencing studies. Overall, the application of these two major long-read sequencing platforms in metagenomics analysis of complex communities is still restricted by higher error rate. This problem could be addressed with improvement of consensus sequences. Recently, newly released R10 chip from ONT has longer base-contacting constriction in the pore, which improves the homopolymer resolution as compared with R9 and improved per-base error rates. Similarly, the HiFi protocol from PacBio can provide Sanger-quality accuracy (>99%) with reduced read length, which are still much longer than short-read sequencing for assembly of complex genomes (Wenger et al., 2019).This can lead to metagenome assembly with higher accuracy and completeness, as well as more accurate OTU identification. Future metagenomics studies are expected to be changed dramatically by this approach. For example, strain UA159 and NN2025 under species *Streptococcus mutans* only share 8% common regions, which can be uniquely assigned. We then found that 20% of ONT reads can cover the unique region of these two strains, respectively, which is infeasible for short reads. Therefore, with better quality of long-read data, this approach may allow us to identify bacteria of interest directly at strain level instead of performing binning analysis in the future.

In addition to illustrating the advantages brought by long-read sequence data, we also assessed the performance of four *de novo* assembly algorithms and a long-read genome binner. The bioinformatics challenges to interpret rich information from complex microbial community include high error rates and low throughput for long-read sequencing, fragmented nature for short-read sequencing, and large CPU hours requirement. For evenly mixed (each with 5% abundance) HM-276D mock community, four tools consistently achieved high accuracy and completeness. No single assembler significantly outperforms others. By subsampling data to less coverage depths, not surprisingly, we found that the corresponding metrics for four tools decreased. In terms of speed, wtdbg2 is tens of times faster than other tools. For the unevenly mixed mock community HM-277D, assembly accuracy still remains high for all four tools (~97%–98%). Genome fraction was reduced because 13 rare bacterial strains (<1%) were poorly assembled. Hybrid-assembler OPERA-MS, which combines the advantages from long- and short-read technologies, shows more robust performance to bacterial strains with extremely low abundance than other tools. However, it produced much more contigs with less contiguity, whereas meta-flye, Canu, and wtdbg2 returned single contig for 18, 15, and 17 strains respectively. Furthermore, taxonomic binning results show that Megan-LR performs well when genomes are not closely related. Taxon bins were reconstructed with acceptable accuracy down to the genus level, whereas performance decreased at species and strain levels.

In summary, our results not only demonstrate the feasibility of characterizing complete microbial genomes and populations from error-prone Nanopore sequencing data but also highlight necessary bioinformatics improvements for future metagenomics tool development to handle specific challenges in error-prone long-read sequencing data. We believe that future metagenomics studies will benefit from this approach to assemble complete microbial genomes, while maintaining the theoretical ability to detect DNA methylations and base modifications, infer repetitive elements and structural variants, and achieve strain-level resolution within microbial communities. All the datasets on reference microbiomes are made publicly available to facilitate benchmarking studies on metagenomics and the development of novel software tools.

Limitations of the Study

In this study, we note that there is still room for further improvement in assembly completeness using long reads. Also, the performance of binning analysis using long read at strain-level still remains unknown.

Resource Availability

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Kai Wang (wangk@email.chop.edu).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

The Oxford Nanopore sequencing data that support the findings of this study have been deposited in the BioProject database at http://www.ncbi.nlm.nih.gov/bioproject/630658 (reference number: **PRJNA630658**).

The PacBio data used in this study were generated from the PacBio RS II sequencer. The data were downloaded from the following URL: https://github.com/PacificBiosciences/DevNet/wiki/Human_Microbiome_Project_MockB_Shotgun.

The Illumina paired-end data for HM-276D were downloaded from NCBI SRA database with accession numbers **SRR2726671** and **SRR2726672**.

The Illumina TruSeq synthetic long-read data for HM-277D were downloaded from NCBI SRA database with accession number **SRR2822457**.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.101223.

ACKNOWLEDGMENTS

This work was supported by Children's Hospital of Philadelphia Research Institute (United States) and NIH/NIGMS (United States) grant GM132713 to K.W.. The following reagent was obtained through BEI Resources, NIAID, NIH as part of the Human Microbiome Project: Genomic DNA from Microbial Mock Community B (Staggered, High Concentration), v5.2H, for Whole Genome Shotgun Sequencing, HM-277D.The following reagent was obtained through BEI Resources, NIAID, NIH as part of the Human Microbiome Project: Genomic DNA from Microbial Mock Community B (Even, High Concentration), v5.1H, for Whole Genome Shotgun Sequencing, HM-276D.

AUTHOR CONTRIBUTIONS

Y.H. performed data analysis and wrote the manuscript. L.F. designed the study, performed long-read sequencing, and analyzed the data. C.N. performed long-read sequencing. K.W. designed the study, supervised the study, and wrote the manuscript. All authors read, revised, and approved the manuscript.

DECLARATION OF INTERESTS

The authors declare no conflict of interest.

Received: March 4, 2020
Revised: May 9, 2020
Accepted: May 28, 2020
Published: June 26, 2020

REFERENCES

Bertrand, D., Shaw, J., Kalathiyappan, M., Ng, A.H.Q., Kumar, M.S., Li, C., Dvornicic, M., Soldo, J.P., Koh, J.Y., Tong, C., et al. (2019). Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. Nat. Biotechnol. *37*, 937–944.

Byrd, A.L., Perez-Rogers, J.F., Manimaran, S., Castro-Nallar, E., Toma, I., McCaffrey, T., Siegel, M., Benson, G., Crandall, K.A., and Johnson, W.E. (2014). Clinical PathoScope: rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data. BMC Bioinformatics *15*, 262.

Chehoud, C., Albenberg, L.G., Judge, C., Hoffmann, C., Grunberg, S., Bittinger, K., Baldassano, R.N., Lewis, J.D., Bushman, F.D., and Wu, G.D. (2015). Fungal signature in the gut microbiota of pediatric patients with inflammatory bowel disease. Inflamm. Bowel Dis. *21*, 1948–1956.

Chen, E.Z., Bushman, F.D., and Li, H. (2017). A model-based approach for species abundance quantification based on shotgun metagenomic data. Stat. Biosci. *9*, 13–27.

Francis, O.E., Bendall, M., Manimaran, S., Hong, C., Clement, N.L., Castro-Nallar, E., Snell, Q., Schaalje, G.B., Clement, M.J., Crandall, K.A., and Johnson, W.E. (2013). Pathoscope: species identification and strain attribution with unassembled sequencing data. Genome Res. *23*, 1721–1729.

Gill, S.R., Pop, M., Deboy, R.T., Eckburg, P.B., Turnbaugh, P.J., Samuel, B.S., Gordon, J.I., Relman, D.A., Fraser-Liggett, C.M., and Nelson, K.E. (2006). Metagenomic analysis of the human distal gut microbiome. Science *312*, 1355–1359.

Gregor, I., Droge, J., Schirmer, M., Quince, C., and McHardy, A.C. (2016). PhyloPythiaS+: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. PeerJ *4*, e1603.

Group, N.H.W., Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J.A., Bonazzi, V., McEwen, J.E., Wetterstrand, K.A., et al. (2009). The NIH human microbiome project. Genome Res. *19*, 2317–2323.

Hao, X., and Chen, T. (2012). OTU analysis using metagenomic shotgun sequencing data. PLoS One *7*, e49785.

Hong, C., Manimaran, S., Shen, Y., Perez-Rogers, J.F., Byrd, A.L., Castro-Nallar, E., Crandall, K.A., and Johnson, W.E. (2014). PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. Microbiome *2*, 33.

Hooper, L.V., Stappenbeck, T.S., Hong, C.V., and Gordon, J.I. (2003). Angiogenins: a new class of microbicidal proteins involved in innate immunity. Nat. Immunol. *4*, 269–273.

Huson, D.H., Albrecht, B., Bagci, C., Bessarab, I., Gorska, A., Jolic, D., and Williams, R.B.H. (2018). MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. Biol. Direct *13*, 6.

Huson, D.H., Beier, S., Flade, I., Gorska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H.J., and Tappu, R. (2016). MEGAN community edition - interactive exploration and analysis of large-scale microbiome sequencing data. PLoS Comput. Biol. *12*, e1004957.

Janda, J.M., and Abbott, S.L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J. Clin. Microbiol. *45*, 2761–2764.

Jones, M.B., Highlander, S.K., Anderson, E.L., Li, W., Dayrit, M., Klitgord, N., Fabani, M.M., Seguritan, V., Green, J., Pride, D.T., et al. (2015). Library preparation methodology can influence genomic and functional predictions in human microbiome research. Proc. Natl. Acad. Sci. U S A *112*, 14024–14029.

Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T., Perry, T., Kao, D., Mason, A.L., Madsen, K.L., and Wong, G.K. (2016). Characterization of the gut microbiome using 16S or shotgun metagenomics. Front. Microbiol. *7*, 459.

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P.A. (2019). Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. *37*, 540–546.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. *27*, 722–736.

Kuleshov, V., Jiang, C., Zhou, W., Jahanbani, F., Batzoglou, S., and Snyder, M. (2016). Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. Nat. Biotechnol. *34*, 64–69.

Laudadio, I., Fulci, V., Palone, F., Stronati, L., Cucchiara, S., and Carissimi, C. (2018). Quantitative assessment of shotgun metagenomics and 16S rDNA amplicon sequencing in the study of human gut microbiome. Omics *22*, 248–254.

Lee, C.H., Bowman, B., and Hall, R. Developments in PacBio® metagenome sequencing: Shotgun whole genomes and full-length 16S. International Plant and Animal Genome Conference Asia, 2014.

Leger, A., and Leonardi, T. (2019). pycoQC, interactive quality control for Oxford Nanopore Sequencing. J. Open Source Softw. *4*, 1236.

Leggett, R.M., Alcon-Giner, C., Heavens, D., Caim, S., Brook, T.C., Kujawska, M., Martin, S., Hoyles, L., Clarke, P., and Hall, L.J. (2020). Rapid profiling of the preterm infant gut microbiota using nanopore sequencing aids pathogen diagnostics. Nat. Microbiol. *5*, 430–442.

Lewis, J.D., Chen, E.Z., Baldassano, R.N., Otley, A.R., Griffiths, A.M., Lee, D., Bittinger, K., Bailey, A., Friedman, E.S., Hoffmann, C., et al. (2015). Inflammation, antibiotics, and diet as environmental stressors of the gut microbiome in pediatric Crohn's disease. Cell Host Microbe *18*, 489–500.

Ley, R.E., Turnbaugh, P.J., Klein, S., and Gordon, J.I. (2006). Microbial ecology: human gut microbes associated with obesity. Nature *444*, 1022–1023.

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.

Liang, X., Bittinger, K., Li, X., Abernethy, D.R., Bushman, F.D., and Fitzgerald, G.A. (2015). Bidirectional interactions between indomethacin and the murine intestinal microbiota. Elife *4*, e08973.

Mason, C.E., Afshinnekoo, E., Tighe, S., Wu, S., and Levy, S. (2017). International standards for genomes, transcriptomes, and metagenomes. J. Biomol. Tech. *28*, 8–18.

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.

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.

Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. Nat. Biotechnol. *35*, 833–844.

Ranjan, R., Rani, A., Metwally, A., McGee, H.S., and Perkins, D.L. (2016). Analysis of the microbiome: advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem. Biophys. Res. Commun. *469*, 967–977.

Ruan, J., and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. Nat. Methods *17*, 155–158.

Sartor, R.B. (2008). Microbial influences in inflammatory bowel diseases. Gastroenterology *134*, 577–594.

Schauber, J., Svanholm, C., Termen, S., Iffland, K., Menzel, T., Scheppach, W., Melcher, R., Agerberth, B., Luhrs, H., and Gudmundsson, G.H. (2003). Expression of the cathelicidin LL-37 is modulated by short chain fatty acids in colonocytes: relevance of signalling pathways. Gut *52*, 735–741.

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Droge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., et al. (2017). Critical assessment of metagenome interpretation-a benchmark of metagenomics software. Nat. Methods *14*, 1063–1071.

Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Duncan, A., Ley, R.E., Sogin, M.L., Jones, W.J., Roe, B.A., Affourtit, J.P., et al. (2009). A core gut microbiome in obese and lean twins. Nature *457*, 480–484.

Wang, F., Kaplan, J.L., Gold, B.D., Bhasin, M.K., Ward, N.L., Kellermayer, R., Kirschner, B.S., Heyman, M.B., Dowd, S.E., Cox, S.B., et al. (2016). Detecting microbial dysbiosis associated with pediatric Crohn disease despite the high variability of the gut microbiota. Cell Rep. *14*, 945–955.

Wen, L., Ley, R.E., Volchkov, P.Y., Stranges, P.B., Avanesyan, L., Stonebraker, A.C., Hu, C., Wong, F.S., Szot, G.L., Bluestone, J.A., et al. (2008). Innate immunity and intestinal microbiota in the development of Type 1 diabetes. Nature *455*, 1109–1113.

Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat. Biotechnol. *37*, 1155–1162.

Wu, G.D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.Y., Keilbaugh, S.A., Bewtra, M., Knights, D., Walters, W.A., Knight, R., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. Science *334*, 105–108.

iScience, Volume *23*

Supplemental Information

Implications of Error-Prone Long-Read Whole-Genome Shotgun Sequencing on Characterizing Reference Microbiomes

Yu Hu, Li Fang, Christopher Nicholson, and Kai Wang

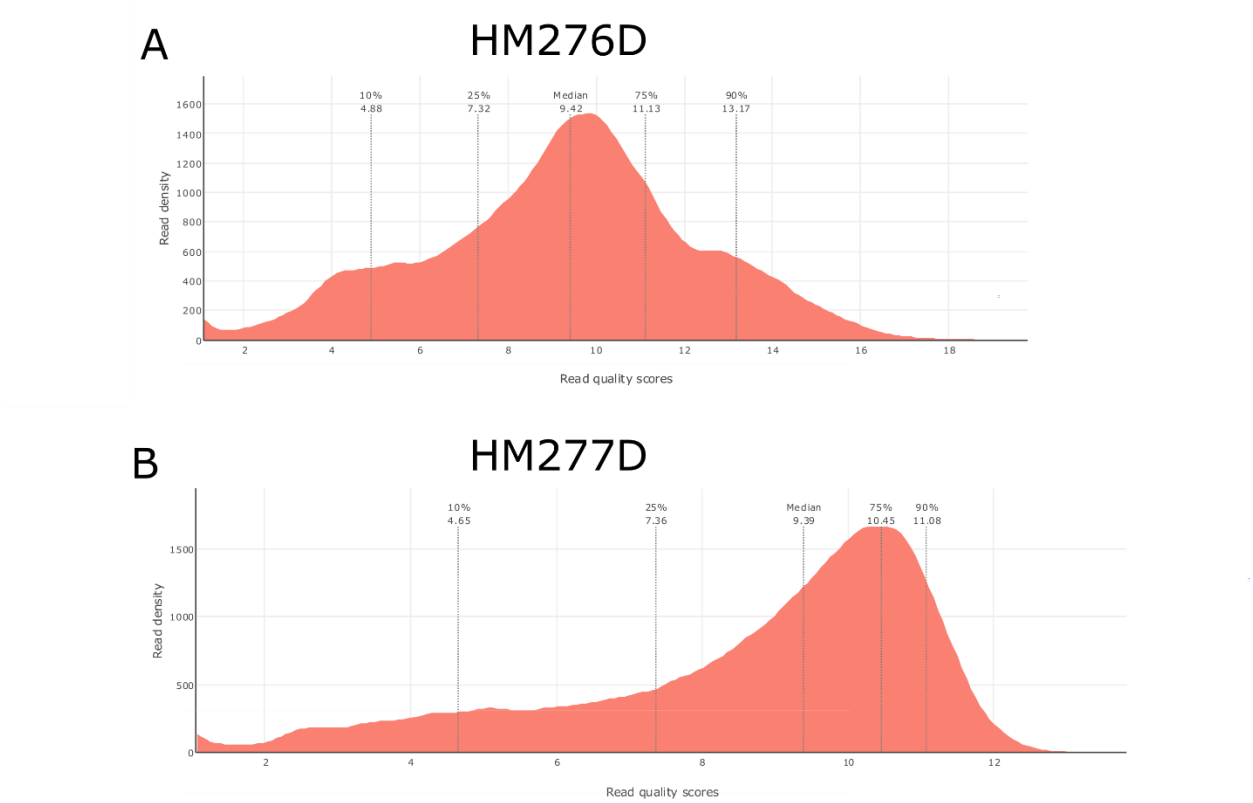


Figure S1. Read quality of Nanopore sequencing data, Related to Table 1. Read quality of sequenced data sets, HM-276D **(A)** and HM-277D **(B)**, were summarized using PycoQC respectively. Dashed lines indicate different quantiles (10%, 25%, 50%, 75%, 90%).

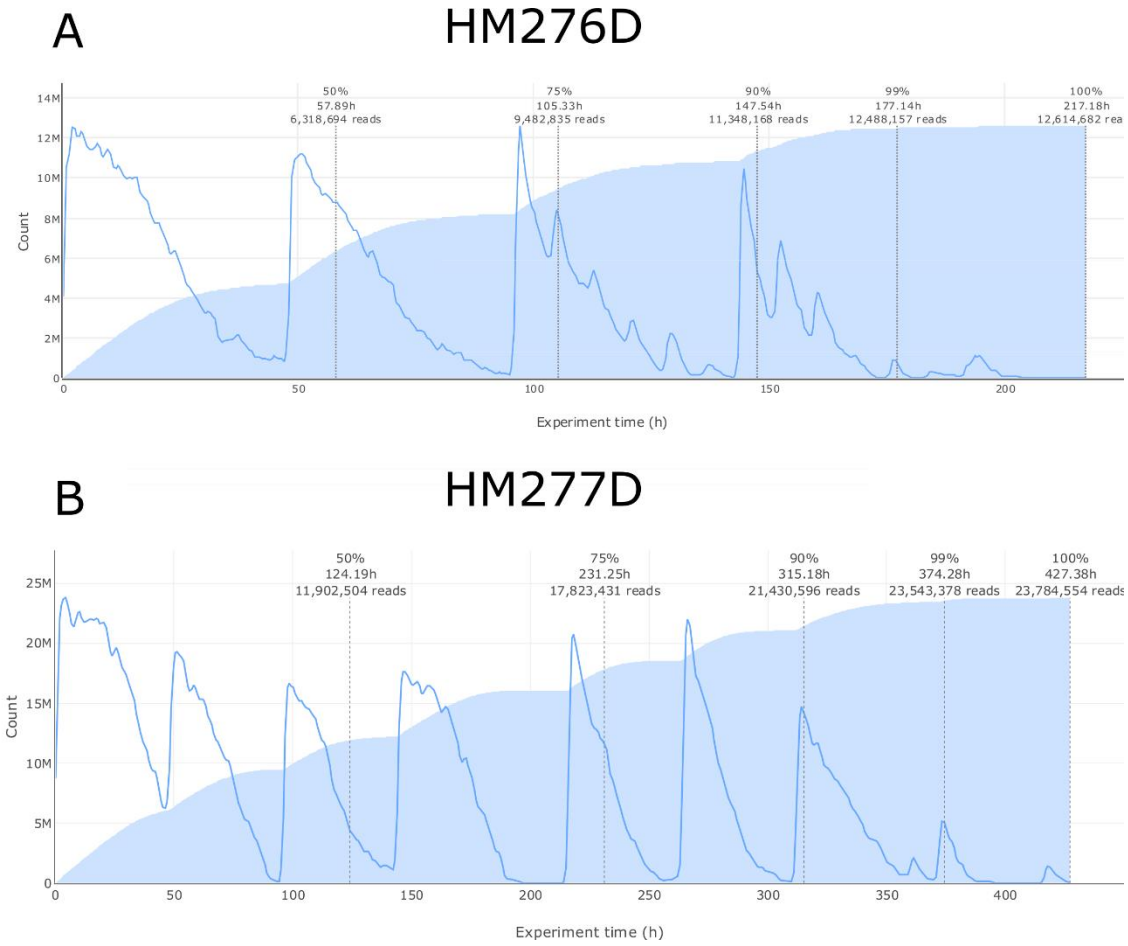


Figure S2. Read output over experiment of Nanopore sequencing data, Related to Table 1. Number of output reads over experiment time for sequenced data sets, HM-276D **(A)** and HM-277D **(B)**, were summarized using PycoQC. Blue line indicates output velocity at specific time. Shaded area represents cumulative read output over experiment time.

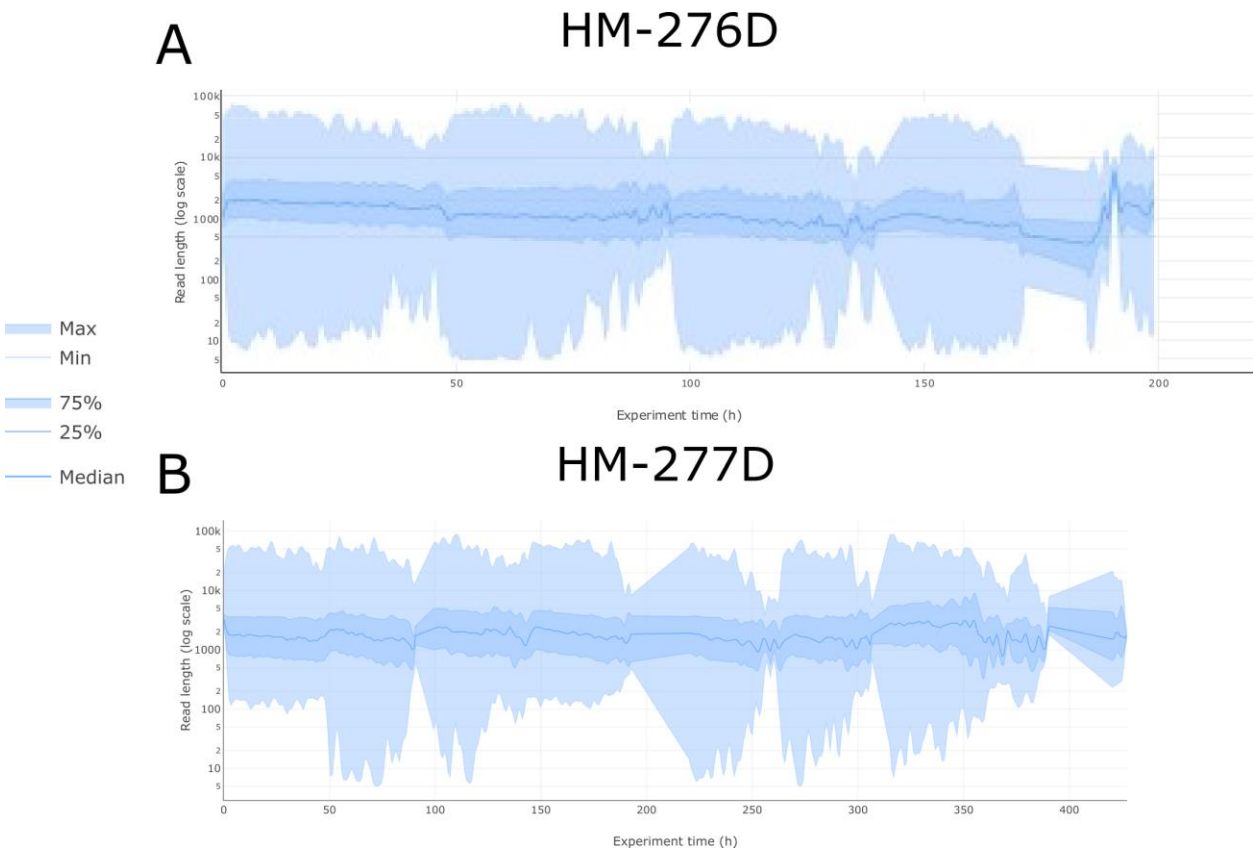


Figure S3. Read length over experiment of Nanopore sequencing data, Related to Table 1. Read length in log scale over experiment time for sequenced data sets, HM-276D **(A)** and HM-277D **(B)**, were summarized using PycoQC.

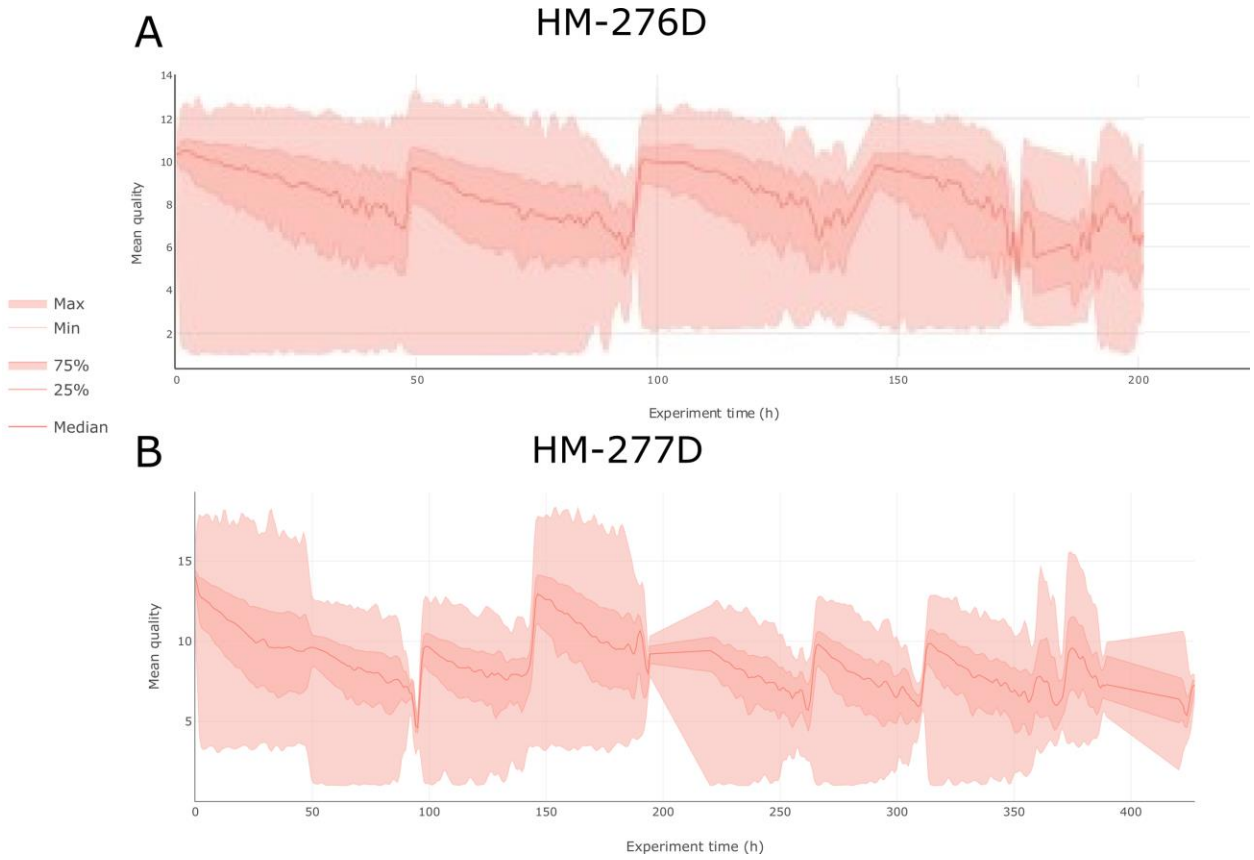


Figure S4. Read quality over experiment of Nanopore sequencing data, Related to Table 1. Mean read quality over experiment time for sequenced data sets, HM-276D **(A)** and HM-277D **(B)**, were summarized using PycoQC.

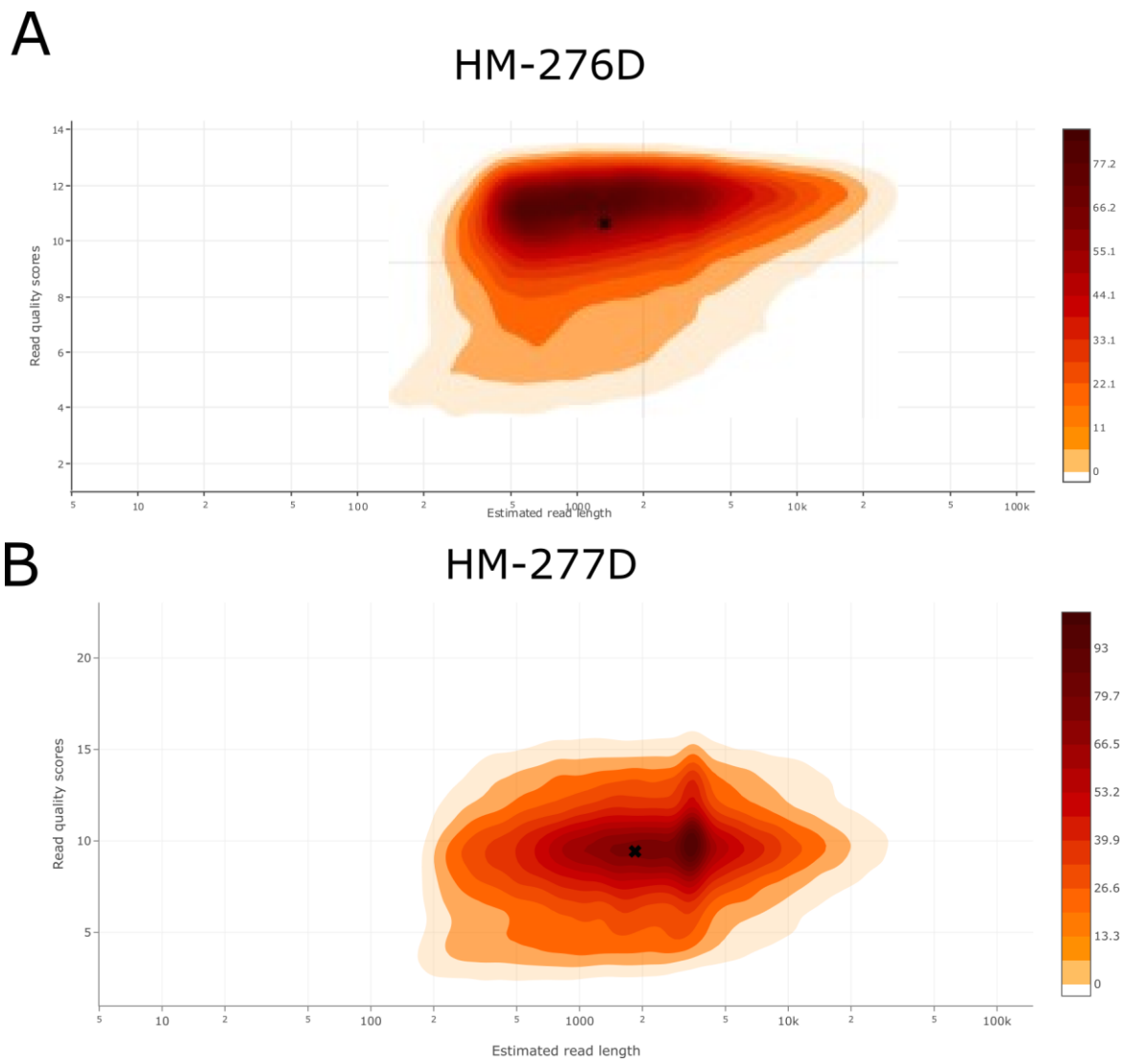


Figure S5. Read quality score vs estimated read length, Related to Table 1. Nanopore read distribution of read length and quality score for sequenced data sets, HM-276D **(A)** and HM-277D **(B)**, were summarized using PycoQC. Color indicates read density.

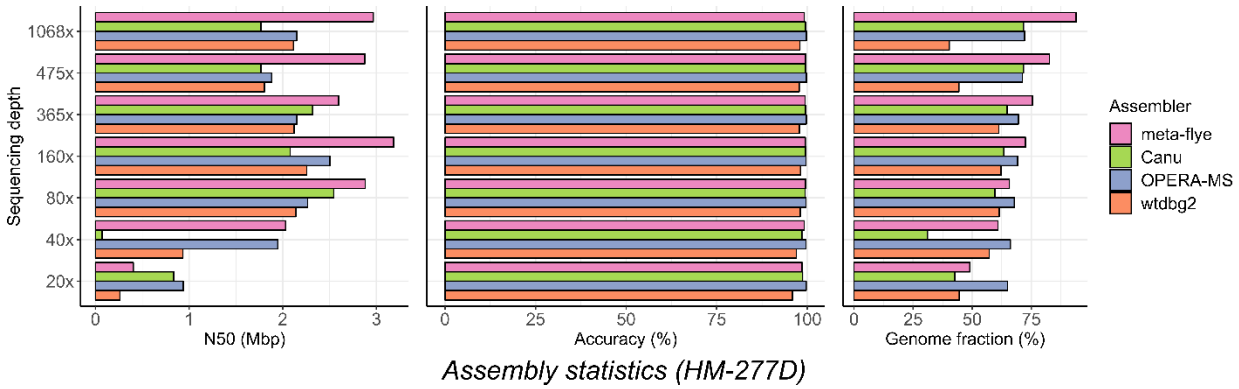


Figure S6. Assembly performance on HM-277D data set, Related to Figure 2. Assembly statistics (N50 length, accuracy and genome fraction) of each assembler at different coverage depths based on HM-277D data set. Colors indicate results from different assemblers (Canu, OPERA-MS, wtdbg2, meta-flye). Assembly accuracy remains high compared to HM-276D, ranging around ~99% across tools. N50 lengths and genome fractions of all methods are substantially lower than the even community.

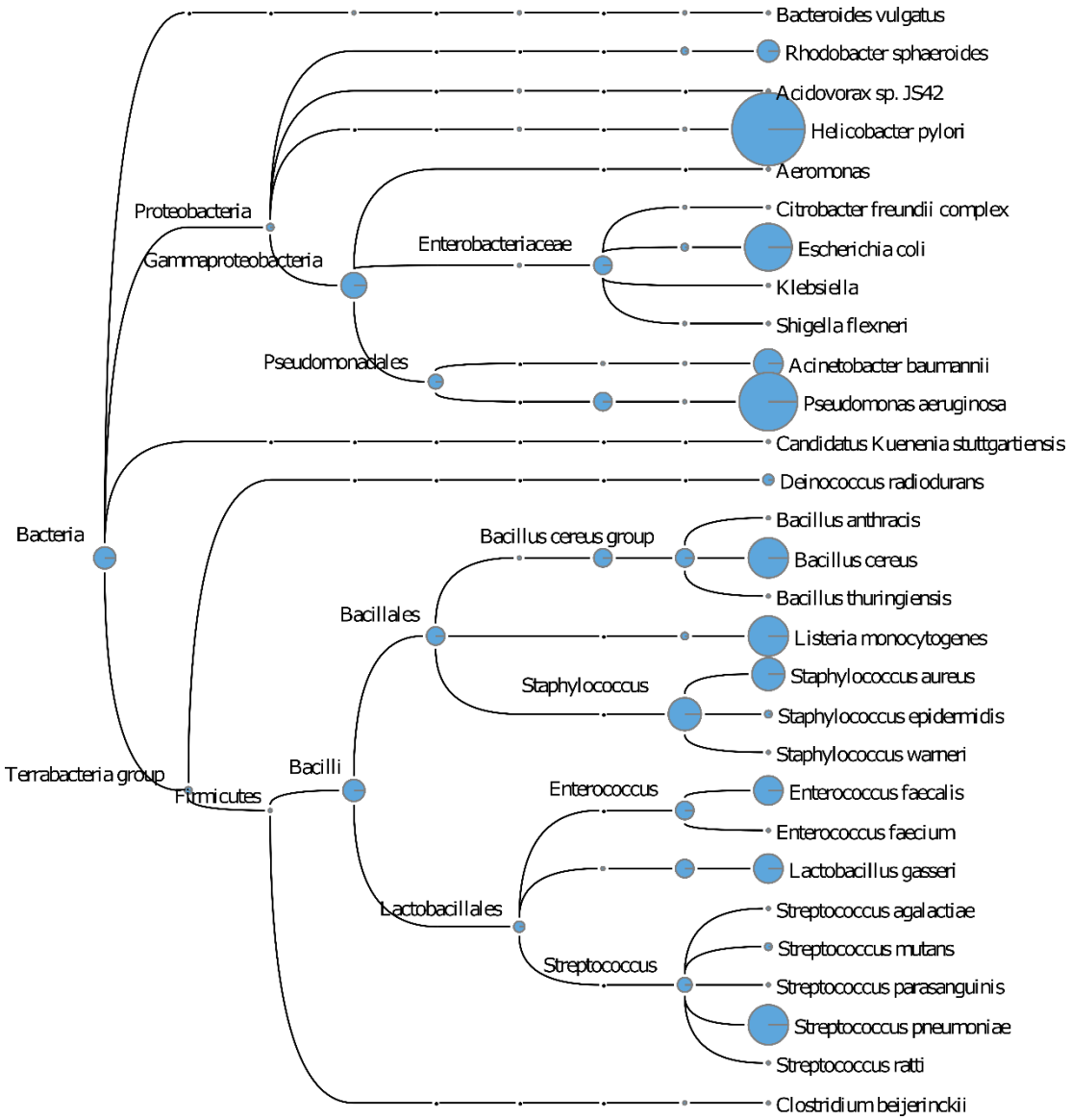


Figure S7. Megan taxonomic tree assignment obtained from HM-276D PacBio sequenced data set, Related to Figure 3. HM-276D PacBio data set was subsampled to 160× depth of coverage. Each read was aligned against NCBI-nr protein reference data base, then binned and visualized using Megan-LR. Megan taxonomic tree showing bacteria taxa identified and their corresponding abundances across taxonomic rank. The radius of circle represents the number of reads assigned for each taxa.

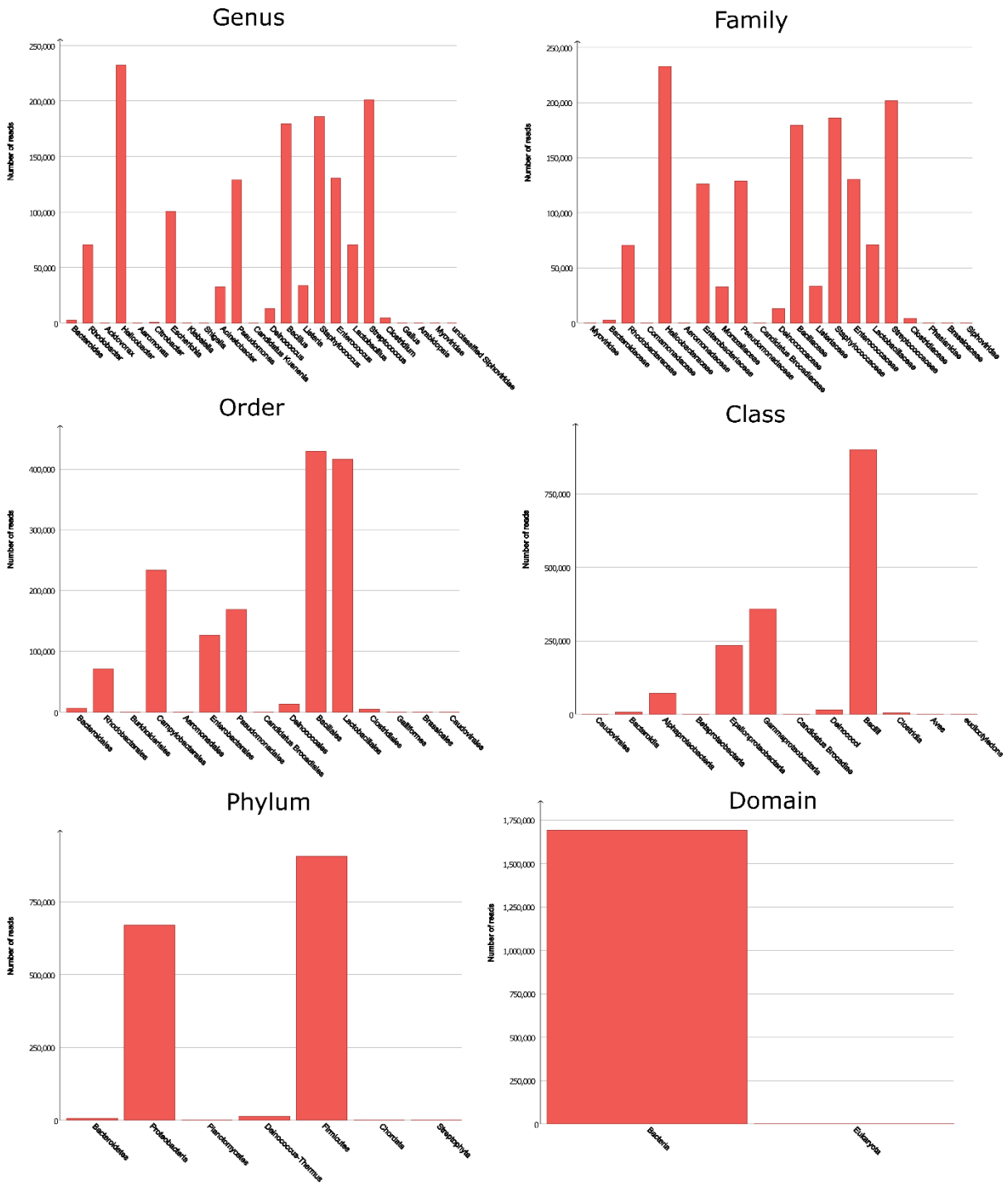


Figure S8. Megan taxonomic read distribution at different ranks obtained from HM-276D Nanopore sequenced data set, Related to Figure 3. HM-276D Nanopore data set was subsampled to 160× depth of coverage. Each read was aligned against NCBI-nr protein reference data base, then binned and visualized using Megan-LR.

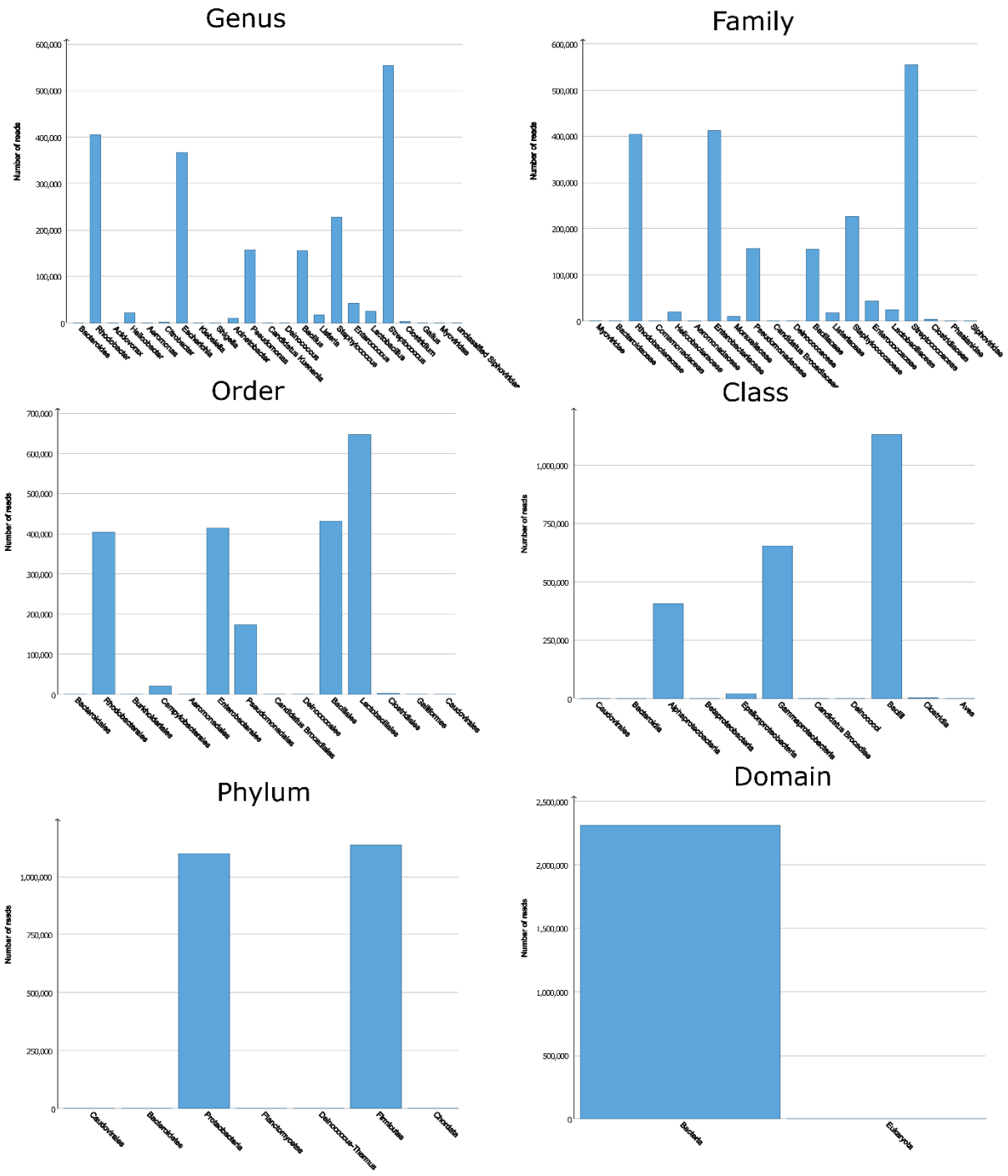


Figure S9. Megan taxonomic read distribution at different ranks obtained from HM-277D Nanopore sequenced data set, Related to Figure 3. HM-277D Nanopore data set was subsampled to 160× depth of coverage. Each read was aligned against NCBI-nr protein reference data base, then binned and visualized using Megan-LR.

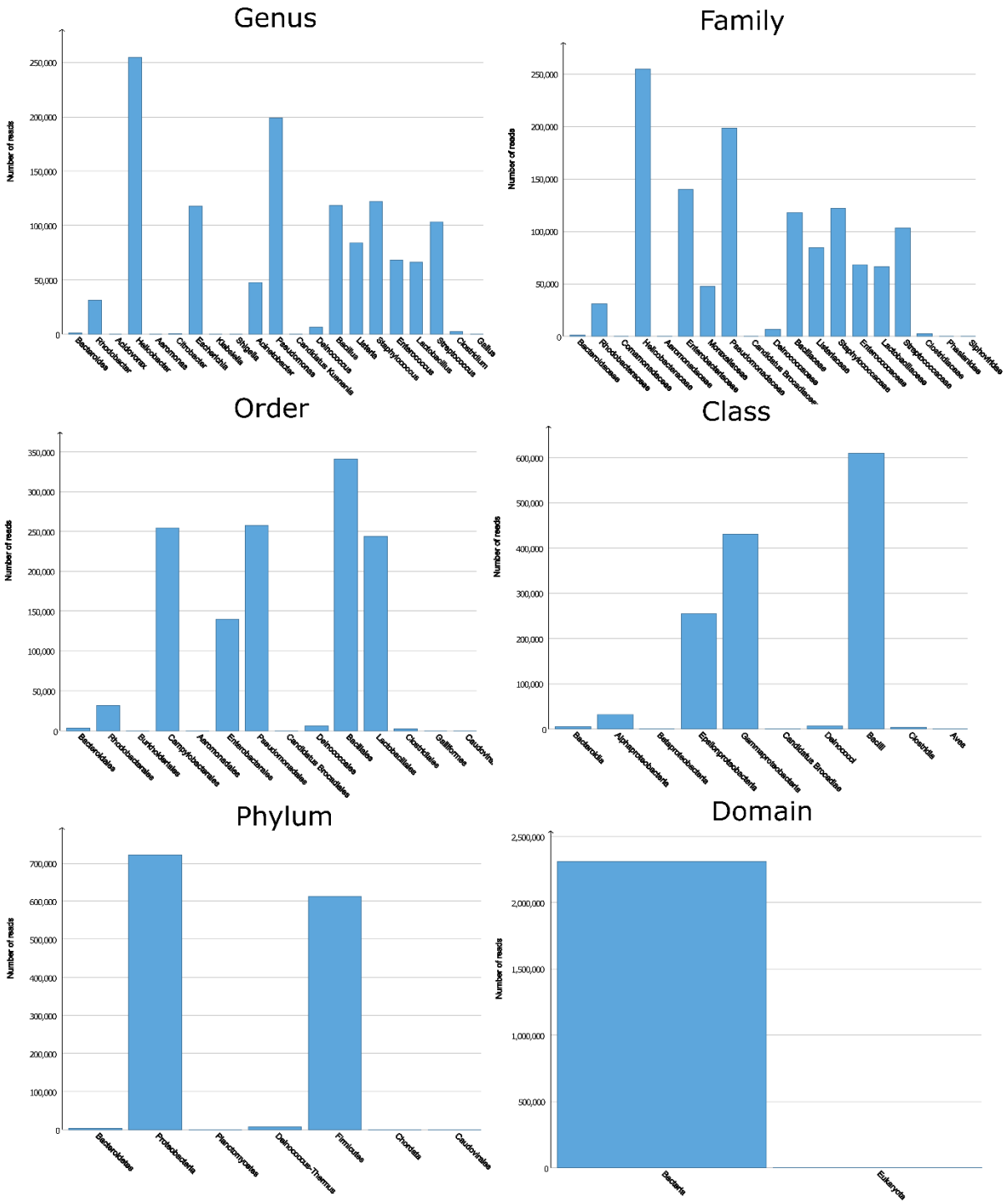


Figure S10. Megan taxonomic read distribution at different ranks obtained from HM-276D PacBio sequenced data set, Related to Figure 3. HM-276D PacBio data set was subsampled to 160× depth of coverage. Each read was aligned against NCBI-nr protein reference data base, then binned and visualized using Megan-LR.

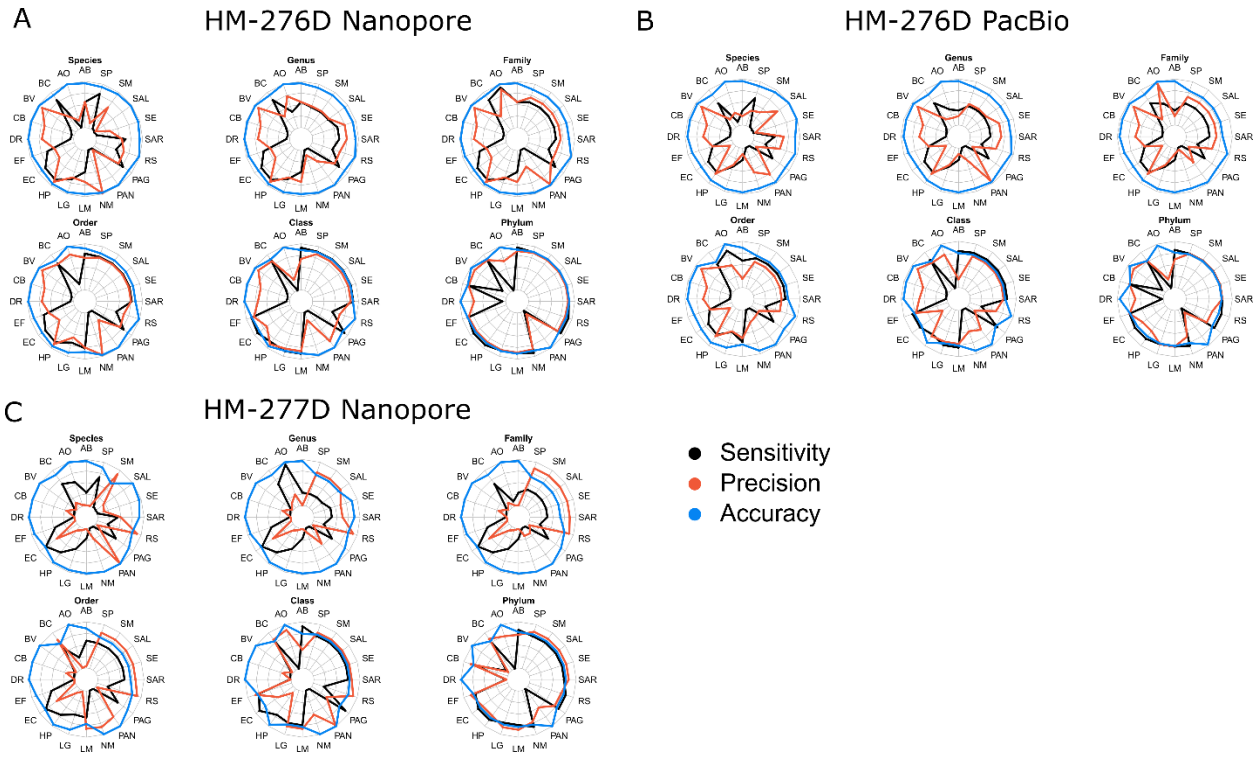


Figure S11. Strain-specific read assignment performance comparison across sequencing technologies, Related to Figure 4. Read assignment accuracy statistics for each bacterial strain were summarized based on datasets: HM-276D Nanopore **(A)**, HM-276D PacBio **(B)** and HM-277D Nanopore **(C)** across ranks. Colors indicates different metrics: sensitivity, precision and accuracy. Taxon were accurately recovered above the family level. HM-276D Nanopore outperformed other two data sets. AB, *Acinetobacter baumannii*; AO, *Actinomyces odontolyticus*; BC, *Bacillus cereus*; BV, *Bacteroides vulgatus*; CB, *Clostridium beijerinckii*; DR, *Deinococcus radiodurans*; DF, *Enterococcus faecalis*; EC, *Escherichia coli*; HP, *Helicobacter pylori*; LG, *Lactobacillus gasseri*; LM, *Listeria monocytogenes*; NM, *Neisseria meningitides*; PAN, *Propionibacterium acnes*; PAG, *Pseudomonas aeruginosa*; RS, *Rhodobacter sphaeroides*; SAR, *Staphylococcus aureus*; SE, *Staphylococcus epidermidis*; SAL, *Streptococcus agalactiae*; SM, *Streptococcus mutans*; SP, *Streptococcus pneumonia*.

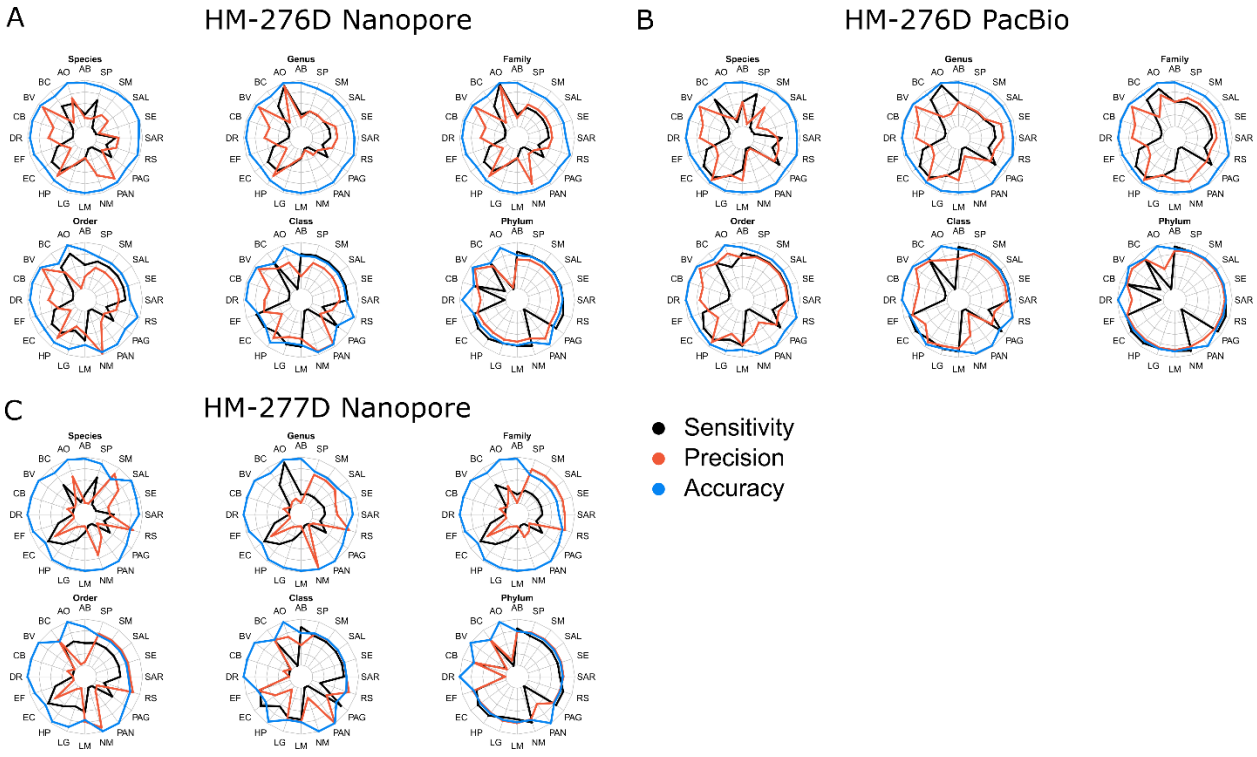


Figure S12. Strain-specific base pair assignment performance comparison across sequencing technologies, Related to Figure 4. Read base assignment accuracy statistics for each bacterial strain were summarized based on datasets: HM-276D Nanopore **(A)**, HM-276D PacBio **(B)** and HM-277D Nanopore **(C)** across ranks. Colors indicates different metrics: sensitivity, precision and accuracy. PacBio performed better than Nanopore data above the family level because of lower error rate. AB, *Acinetobacter baumannii*; AO, *Actinomyces odontolyticus*; BC, *Bacillus cereus*; BV, *Bacteroides vulgatus*; CB, *Clostridium beijerinckii*; DR, *Deinococcus radiodurans*; DF, *Enterococcus faecalis*; EC, *Escherichia coli*; HP, *Helicobacter pylori*; LG, *Lactobacillus gasseri*; LM, *Listeria monocytogenes*; NM, *Neisseria meningitides*; PAN, *Propionibacterium acnes*; PAG, *Pseudomonas aeruginosa*; RS, *Rhodobacter sphaeroides*; SAR, *Staphylococcus aureus*; SE, *Staphylococcus epidermidis*; SAL, *Streptococcus agalactiae*; SM, *Streptococcus mutans*; SP, *Streptococcus pneumonia*.

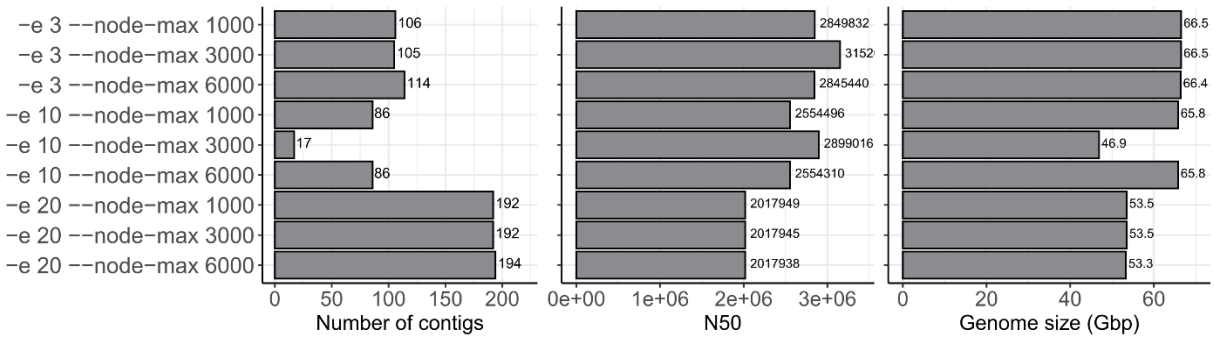


Figure S13. Assembly results for wtdbg2 based on HM-276D data sets, Related to Figure 2. Barplots indicate assembly statistics (Number of contigs, N50 length and genome size). Each row represents a run of wtdbg2 with parameters “-e” and “--node-max”.

Tools	Depth	N50 length	NG50 length	NA50 length	NGA50 length	Accuracy (%)	Fraction (%)	# contigs	# long contigs	Longest contig	Genome size	CPU time (min)
Canu	**20x**	717267	616530	450953	413727	98.5	96.8	298	254	2612567	65503873	98
Canu	**40x**	1987236	1975600	893966	893966	99.07	99.29	132	112	6286130	67676017	250
Canu	**80x**	2886059	2731942	1284219	1284219	99.24	99.86	62	57	6316623	68735511	678
Canu	**160x**	3901381	3901381	1302124	1506778	99.27	99.93	60	52	6299115	68879111	1537
Canu	**365x**	2983818	2983818	1219953	1219953	99.28	99.83	64	58	6292103	68964121	2651
Canu	**480x**	3911963	3911963	1400993	1400993	99.4	99.81	83	65	6359094	69425747	3858
OPERA-MS	**20x**	1122204	1122204	450323	489324	99.83	99.71	5117	201	6324007	67168904	61
OPERA-MS	**40x**	2657727	2657727	1210958	1210958	99.96	99.99	1695	81	5220208	67629371	102
OPERA-MS	**80x**	2835709	2732545	1189226	1189226	99.96	99.99	1921	74	4636570	67632885	186
OPERA-MS	**160x**	2933262	2792941	1273171	1273171	99.95	98.45	2347	65	6255842	66580943	382
OPERA-MS	**365x**	2938016	2938016	1298425	1298425	99.91	99.98	4734	64	6255878	67858470	856
OPERA-MS	**480x**	2938019	2938019	1213537	1298125	99.92	99.98	4732	63	6255756	67892051	1238
wtdbg2	**20x**	552100	415407	243026	208370	96.22	90.73	439	367	3542441	60910472	3
wtdbg2	**40x**	2106610	2057746	751004	751004	98.1	98.58	143	110	6265362	66286613	7
wtdbg2	**80x**	3152112	2920474	1331021	1331021	98.79	98.97	105	66	6229939	66511311	13
wtdbg2	**160x**	2910424	2910424	1286826	1286826	98.95	98.75	134	76	6215258	66615029	13
wtdbg2	**365x**	2706821	2706821	1278483	1265683	98.66	97.34	90	73	6251621	65544245	19
wtdbg2	**480x**	3168384	2922530	1201346	1184259	98.73	95.95	201	119	6210210	65977641	23
meta-flye	**20x**	1653589	1547909	558652	534341	98.96	98.76	223	206	5630982	66808399	57
meta-flye	**40x**	2725547	2653197	1209147	1209147	99.43	99.97	64	52	6274273	67627825	86
meta-flye	**80x**	2930772	2930772	1588493	1588493	99.52	99.99	59	43	6251934	67630110	140
meta-flye	**160x**	3888260	3180529	1622636	1622636	99.54	99.97	61	39	6252579	67595608	372
meta-flye	**365x**	3181836	2934283	1315358	1315358	99.62	99.98	88	44	6245780	67727067	603
meta-flye	**480x**	3181822	2934277	1718698	1718698	99.62	99.99	89	43	6245565	67700317	756

Table S1. Comprehensive assembly statistics on HM-276D using Canu, OPERA-MS, wtdbg2 and meta-flye, Related to Figure 2.

Species	Abundance	RefSeq gene		16S rRNA gene		Protein coding gene	
		average coverage (#bases)	Significantly detected gene	average coverage (#bases)	Significantly detected gene	average coverage (#bases)	Significantly detected gene
Acinetobacter baumannii	0.18%	9.83	94	9.50	6	9.86	3,817
Actinomyces odontolyticus	0.01%	4.27	56	3.10	2	4.65	1,999
Bacillus cereus	1.22%	100.51	138	94.04	12	102.33	5,675
Bacteroides vulgatus	0.02%	2.32	65	1.77	4	2.39	3,067
Clostridium beijerinckii	1.43%	96.40	143	78.49	14	97.42	5,149
Deinococcus radiodurans	0.03%	4.94	57	5.19	3	4.86	3,060
Enterococcus faecalis	0.01%	2.76	53	3.81	2	3.37	2,497
Escherichia coli	15.75%	1,032.93	179	1,003.79	7	1,060.46	4,341
Helicobacter pylori	0.07%	113.13	43	117.15	2	114.16	1,444
Lactobacillus gasseri	0.03%	27.95	96	24.06	6	28.97	1,783
Listeria monocytogenes	0.07%	10.74	184	8.92	6	11.42	2,864
Neisseria meningitides	0.07%	42.67	71	28.53	4	47.85	1,926
Propionibacterium acnes	0.11%	41.60	58	38.75	3	43.02	2,506
Pseudomonas aeruginosa	5.01%	141.55	105	160.86	4	137.90	5,572
Rhodobacter sphaeroides	64.44%	2,219.40	67	1,993.22	3	2,438.52	4,279
Staphylococcus aureus	0.83%	323.26	79	289.00	5	404.68	2,982
Staphylococcus epidermidis	6.52%	976.37	76	1,117.10	5	1,002.43	2,472
Streptococcus agalactiae	0.03%	72.99	101	70.16	7	75.54	2,127
Streptococcus mutans	4.15%	4,207.60	80	3,598.02	5	3,818.93	1,953
Streptococcus pneumoniae	0.01%	1.91	58	1.30	2	2.39	1,868

Table S2. Species-specific gene coverage summary of HM-277D data set, Related to Figure 4. Gene coverage statistics were summarized for 3 different gene sets: all Refseq genes, 16S rRNA genes and protein coding genes. Average coverage = number of bases mapped to the exonic region / length of exonic region. Gene is noted as

significantly detected when 50% exonic region is covered by at least 1 read and average coverage > 1.

Transparent Methods

Oxford nanopore sequencing of HM-276D and HM-277D

DNA samples of HM-276D and HM-277D were ordered from BEI Resources. Concentration of DNA was assessed using the dsDNA HS assay on a Qubit fluorometer (Thermo Fisher).

For library preparation, 1.0 μg DNA was used as the input DNA of each library. The library was prepared using the ligation sequencing protocol (SQK-LSK109) from ONT. Concretely, end repair, dA-tailing and DNA repair was performed using NEBNext Ultra II End Repair/dA-tailing Module (catalog No. E7546) and NEBNext FFPE Repair Mix (M6630). In all, 3.5 μl Ultra II End-prep reaction buffer, 3 μl Ultra II End-prep enzyme mix, 3.5 μl NEBNext FFPE DNA Repair Buffer and 2 μl NEBNext FFPE DNA Repair Mix were added to the input DNA. The total volume was adjusted to 60 μl by adding nuclease-free water (NFW). The mixture was incubated at 20 °C for 5 min and 65 °C for 5 min. A 1 × volume (60 μl) AMPure XP clean-up was performed and the DNA was eluted in 61 μl NFW. One microliter of the eluted dA-tailed DNA was quantified using the Qubit fluorometer. A total of 0.7 μg DNA should be retained if the process is successful.

Adaptor ligation was performed using the following steps. Five microliter Adaptor Mix (ONT, SQK-LSK109 Kit), 25 μl Ligation Buffer (ONT, SQK-LSK109 Kit) and 10 μl NEBNext Quick T4 DNA Ligase (NEB, catalog No. E6056) were added to the 60 μl dA-tailed DNA from the previous step. The mixture was incubated at room temperature for 10 min. The adaptor-ligated DNA was cleaned up using 40 μl AMPure XP beads. The mixture of DNA and AMPure XP beads was incubated for 5 min at room temperature and the pellet was washed twice by 250 μl Long Fragment Buffer (ONT, SQK-LSK109). The purified-ligated DNA was resuspended in 15 μl Elution Buffer (ONT, SQK-LSK109). A 1-μl aliquot was quantified by fluorometry (Qubit) to ensure ≥ 400 ng DNA was retained.

The final library was prepared by mixing 37.5 µl Sequencing Buffer (ONT, SQK-LSK109), 25.5 µl Loading Beads (ONT, SQK-LSK109), and 12 µl purified-ligated DNA. The library was loaded to R9.4 flow cells (FLO-MIN106, ONT) according to the manufacturer's guidelines. GridION sequencing was performed using default settings for the R9.4 flow cell and SQK-LSK109 library preparation kit. The sequencing was controlled and monitored using the MinKNOW software developed by ONT.

Metagenome assembly

Genome assemblies of the 20-mixed bacteria from HM-276D and MH-277D mock communities were conducted using 4 existing assemblers based on generated long-read sequencing reads. These 4 dedicated long-read assemblers we used are wtdbg2 (v2.4), OPERA-MS, Canu (v1.8) and meta-flye, where OPERA-MS and meta-flye are designed to be capable to handle metagenome while wdtbg2 and Canu are for broadly application. To evaluate the impact of coverage depth in genome assembly, in addition to 525× (HM-276D) and 1068× (HM-277D), we subsampled 5 data sets with 365×, 160×, 80×, 40× and 20× coverages for these two mock communities. In addition to long-read data, OPERA-MS requires short reads to improve the assembly accuracy. Hence, we downloaded Illumina sequenced HM-276D(Jones et al., 2015) and HM-277D data sets(Kuleshov et al., 2016). Similarly, these short-read data were also subsampled with depths 160×, 80×, 40× and 20×, which were provided to OPERA-MS in corresponding data set analysis. We also analyzed a PacBio data set(Lee et al., 2014) of HM-276D sample using wtdbg2, OPERA-MS, Canu and meta-flye to compare assembly performance across sequencing technologies. For comparison fairness, we applied consistent configuration settings for each tool across different coverage depths. For wtdbg2, we first tuned parameters on "-e", "--node-max". Assemblies were conducted under different parameter values (-e: 3, 10, 20) (--node-max: 1000, 3000, 6000). Based on the genome contiguity and completeness results in **Fig S13**, we specified parameter "-e 3 –node-max" for wtdbg2. For other tools, we set estimated genome size as 70M, where the parameters are "genomeSize=70M useGrid=True" for Canu, and "CONTIG_LEN_THR 500, CONTIG_EDGE_LEN 80, CONTIG_WINDOW_LEN 340, KMER_SIZE 60, LONG_READ_MAPPER minimap2" for OPERA-MS, "-t 40 -g 70m -o ./ --meta" for meta-flye. 40 contig output files were obtained

(2 mock community samples, 6 depths of coverage, 4 assembly tools) for further evaluation.

Metagenome assembly evaluation

Assembled genomes produced by each tool based on different samples and coverage depths were evaluated with metrics related to contiguity, genome completeness and accuracy. To assess the assembly contiguity, we first used our script to calculate the widely-used statistic N50, which is the shortest contig needed to cover at least 50% of the assembly. In addition, other related statistics, such as number of contigs, number of long contigs (>10kb), longest contigs and total assembly size, were collected from the FASTA output file of each assembler. Furthermore, we summarized NG50 for each method by replacing the assembly size with estimated genome size. This quantity represents the shortest contig needed to cover 50% of the genome. Instead of contigs, we also evaluated the performance based on aligned blocks for each method by using QUAST(Gurevich et al., 2013) to calculate NA50 and NGA50, which represent shortest aligned block to cover 50% of the assembly and genome respectively. Based on these metrics, the contiguity of assemblies was comprehensively evaluated. Next, we downloaded the reference genome FASTA files of all 20 bacteria from NCBI database to measure the concordance between the references and assemblies. First, assemblies were mapped to the reference genomes using Mummer v3.23 with parameters “-maxmatch -c 100 -p nucmer”. Then, by comparing all contigs mapped onto the reference using dandiff, assembly accuracy was calculated using 1-to-1 alignment identity, which is the correctly matched base-pair percentage of contigs uniquely mapped to the reference genome (1-mismatch%). In addition, to assess the assembly completeness, we calculated the percentage of genome covered by the contigs. In real case, instead of evenly mixed in HM-276D mock community, bacterial strains are non-uniformly distributed, where some are likely to share extremely low abundance. Therefore, we evaluated the impact of the genomic DNA abundance on genome assembly. For the unevenly mixed HM-277D mock community samples, we calculated the abundance for each bacterial strain by normalizing the concentration with related reference genome size. The relationship between abundances

and assessment metrics was displayed using scatter plots. For each plot, linearity was measured based on Spearman correlation using R v3.3.3.

Taxonomic binning analysis

Taxon bins of the 20-mixed bacteria from two mock communities were recovered using taxonomic binner Megan-LR(Huson et al., 2018) with 3 long-read sequencing data sets: HM-276D (Nanopore, PacBio) and HM-277D (Nanopore) at 160× depth of coverage. We first aligned all reads against NCBI-nr protein reference database using LAST with parameters "-P 100 -F15". Next, output MAF files were converted to DAA format in smaller size. Then, we meganized the DAA files using MEGAN(Huson et al., 2016), which allows us to interactively visualize and explore these taxonomic results. To evaluate the taxonomic binning performance, we first counted the number of reads and bases which were correctly assigned to each taxon from the mock microbial community. We determined the metrics (precision, sensitivity, true positive rate and false positive rate). Precision and sensitivity assess how accuracy each read is classified across different sequencing technologies. Precision is the percentage of reads assigned correctly to the corresponding taxa out of all reads. Sensitivity is the percentage of correct reads out of reads assigned to the particular taxa. Next, we use true positive rate (TPR) and false discover rate (FDR) to assess the accuracy in taxonomic detection across sequencing technologies. TPR is the percentage of correctly detected taxon out of known taxon from the microbial community. FDR is the percentage of correctly detected taxon out of all detected taxon. All metrics are defined at each taxonomic rank.

Supplemental References

GUREVICH, A., SAVELIEV, V., VYAHHI, N. & TESLER, G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics,* 29**,** 1072-5.

HUSON, D. H., ALBRECHT, B., BAGCI, C., BESSARAB, I., GORSKA, A., JOLIC, D. & WILLIAMS, R. B. H. 2018. MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. *Biol Direct,* 13**,** 6.

HUSON, D. H., BEIER, S., FLADE, I., GORSKA, A., EL-HADIDI, M., MITRA, S., RUSCHEWEYH, H. J. & TAPPU, R. 2016. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Comput Biol,* 12**,** e1004957.

JONES, M. B., HIGHLANDER, S. K., ANDERSON, E. L., LI, W., DAYRIT, M., KLITGORD, N., FABANI, M. M., SEGURITAN, V., GREEN, J., PRIDE, D. T., YOOSEPH, S., BIGGS, W., NELSON, K. E. & VENTER, J. C. 2015. Library preparation methodology can influence genomic and functional predictions in human microbiome research. *Proc Natl Acad Sci U S A,* 112**,** 14024-9.

KULESHOV, V., JIANG, C., ZHOU, W., JAHANBANI, F., BATZOGLOU, S. & SNYDER, M. 2016. Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. *Nat Biotechnol,* 34**,** 64-9.

LEE, C. H., BOWMAN, B. & HALL, R. Developments in PacBio® metagenome sequencing: Shotgun whole genomes and full-length 16S. International Plant and Animal Genome Conference Asia, 2014.