bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Implications of error-prone long-read whole-genome shotgun sequencing on characterizing reference microbiomes

Yu Hu[1,*], Li Fang[1,*], Christopher Nicholson[1,2], Kai Wang[1,3, §]

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

2. Department of Biology, University of Pennsylvania, Philadelphia, PA 19104

3. Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA 19104

* These authors contributed equally to this work

§ Correspondence to: wangk@email.chop.edu

Keywords: metagenomics, long-read sequencing, microbiome

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Single-molecule long-read sequencing technologies, such as Nanopore and PacBio, may be particularly relevant for microbiome studies, since they can perform sequencing without PCR amplification or bacteria culture, and the much longer reads may facilitate assignments of operational taxonomic units (OTUs) from genus to species level. However, due to the relatively high per-base error rates (~15%), the application of long-read sequencing on microbiomes remains largely unexplored, and there is a lack of benchmarking study on reference materials to assess their potential utility in microbiome studies. Here we deeply sequenced two human microbiota mock community samples from the Human Microbiome Project (525× coverage on HM-276D with 20 evenly mixed strains, 1068× coverage on HM-277D with 20 unevenly mixed strains). We showed that assembly programs consistently achieved high accuracy (~99%) and completeness (~99%) for bacterial strains with adequate coverage (~99% in 276D and ~72% in 277D). For HM-277D, 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%). Taxonomic binning and profiling were more accurate at higher rank, while performance decreased at the species level. We further compared the results with data generated from the Illumina short-read sequencing and PacBio long-read sequencing. Our results demonstrate the feasibility to characterize complete microbial genomes and populations from error-prone Nanopore sequencing data, but also highlight necessary bioinformatics improvements for future metagenomics tool development. All the data sets on reference microbiomes are made publicly available to facilitate benchmarking studies on metagenomics and the development of novel software tools.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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, hepatocellular and colorectal carcinoma[1-14]. Therefore, accurate profiling of complete genomes and population are 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)[15] and shotgun metagenome sequencing on Illumina platforms[16] 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[17], abundance quantification[18], read assembly[19-23], binning and taxonomic profiling[24-29]. 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 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[30-32]. 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, due to the short read length (150 to 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 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.* [33] 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 while single bacteria chromosome was assembled to more than 50 contigs based on short read data. In recent years, the Oxford Nanopore technologies (ONT) 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 >10kbp, spanning low complexity and repetitive genomic regions, which provides much more continuous assemblies. Subsequently, this approach has become an attractive option in metagenomics sequencing. While the ONT 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[34]. Therefore, a comprehensive evaluation of long-read bioinformatics tools in microbial profiling is needed[35]. Nicholls

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

et al.[36] presented Nanopore sequencing data sets of two mock communities with 10 microbial species from ZymoBIOMICS[37]. They showed the utility of these data sets for future bioinformatics method development for long-read metagenomics. However, publicly available data sets 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 data sets 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 4 long-read assemblers at different depth of coverage. 20 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. [33] sequenced this mock community with PacBio to show the improvement of long-read data in metagenome assembly analysis. Jones *et al.*[5] compared the influence of different NGS platforms on genomic and functional predictions using HM-276D sample. We downloaded these two data sets and compared the performance with Nanopore data. Our results show that Nanopore consistently improved assembly contiguity, and completeness compared to PacBio and Illumina across computational approaches. Next, we sequenced HM-277D

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Mock Community sample with 1068× coverage. HM-277D is unevenly mixed DNA sample of 20 bacterial strains. Kuleshove *et al.*[38] sequenced this sample with Illumina TruSeq synthetic long read technique and showed the improvement in bacterial species identification, genome reconstruction compared to short sequences. Also, Leggett *et al.* [39] 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 bacteria strains with reference genome size 70 Mb in total with 39 chromosomes. 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. As shown in **Figure 1(a)**, read coverage

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

across 20 bacterial strains has good agreement with known abundances. Read depth is relatively homogenous across bacteria 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.

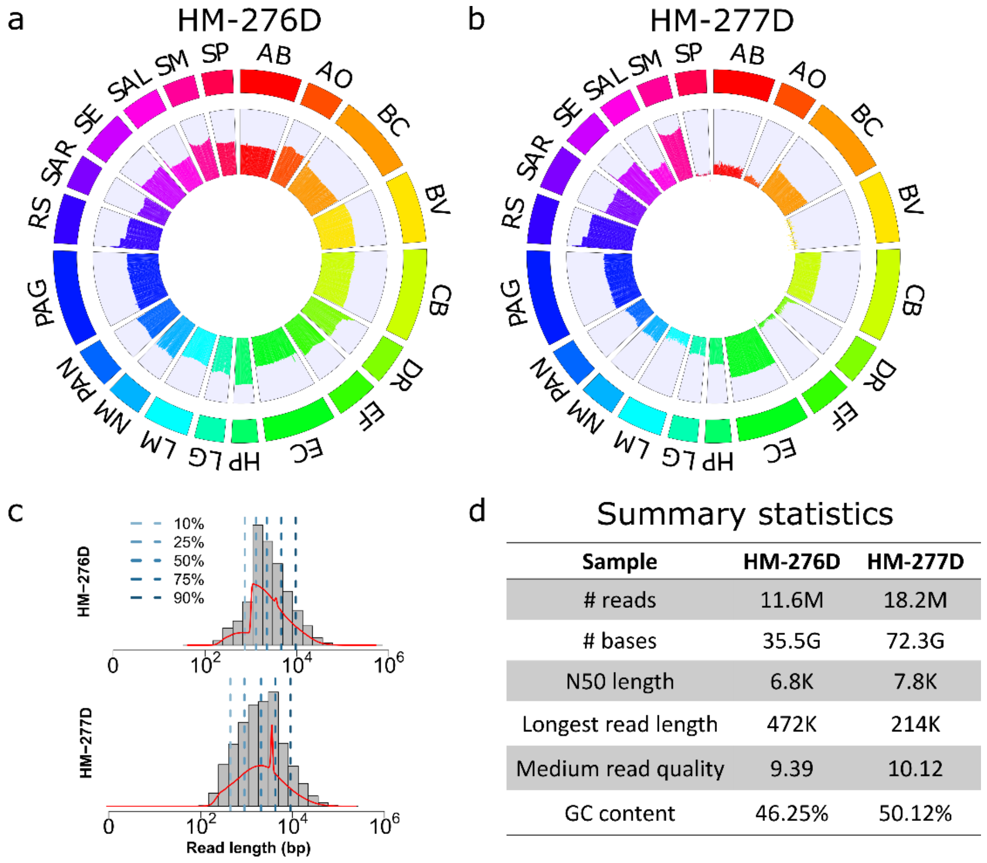
Mapping statistics	HM-276D	HM-277D
# of reads	8,086,684	18,254,839
# of mapped reads	7,640,934	18,110,317
reads unmapped	445,750	144,522
reads MQ0	60,972	103,601
non-primary alignments	287,369	732,671
total length	33,563,573,383	72,312,638,112
bases mapped	32,143,689,158	72,216,146,980
bases mapped (cigar)	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 data set. Sequenced data were mapped against reference genomes of 20 known bacterial strains. Sequences indicates 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 bacteria strains. 18,254,839 reads data set with 72,312,638,112 bases (1068× 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%. As shown in **Figure1(b)**, 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 3 reads covered and 4 strains with sequencing depth less than 10 bp, which makes it more difficult for biological interpretation of this microbial community.



Summary statistics

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%

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 1. Summary of Nanopore Sequencing data from HM-276D and HM-277D microbial communities. (a, 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 to 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 data sets. Blue dashed lines represent different quantiles. Red line represents the density of read length distribution. **(d)** Summary statistics of HM-276D and HM-277D data sets. Each value was calculated by using pycoQC [40] and LongreadQC

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 data set with 525× coverage from HM-276D mock community using 4 assemblers: wtdbg2[19], OPERA-MS[20], Canu[21] and meta-flye[22]. Canu and meta-flye are designed to be capable of handling metagenome data, while 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 67.7Mb genome with 89 contigs. wtdbg2 generated similar results with 64.9

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Mb genome size, 61 contigs and 2.97 Mb N50 length. Assembly metrics of OPERA-MS (67.9 Mb genome size, 4734 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. By mapping all contigs to the reference genomes using MUMMer v3.23, we assessed the accuracy and genome completeness of contigs produced by 4 assemblers. As shown in **Figure 2(a)**, meta-flye achieved the highest genome fraction (99.99%) and 1-to-1 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 95.94% and accuracy 98.73%). Thus, 4 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

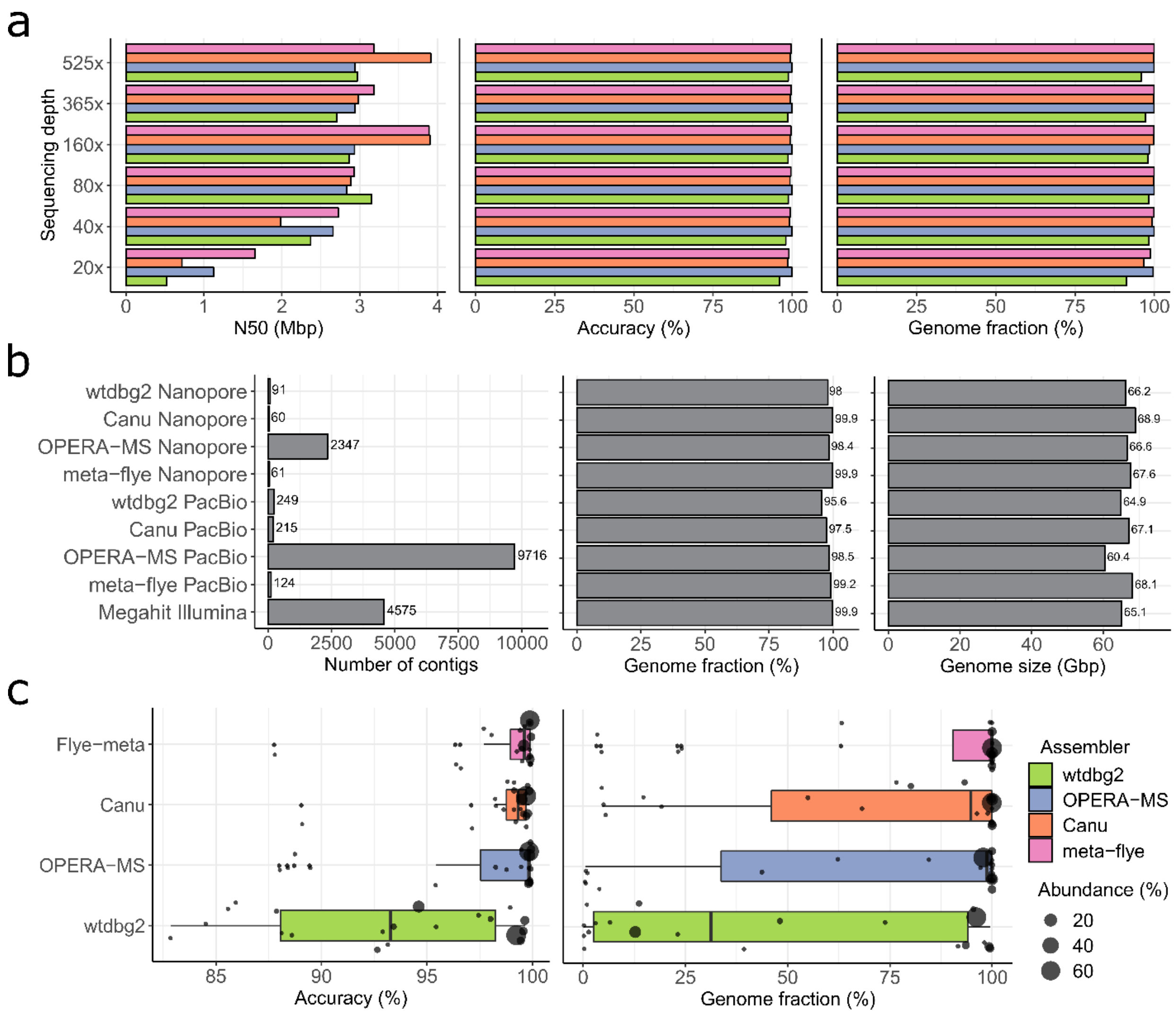


Figure 2. Assembly results for HM-276D and HM-277D data sets. (a) Assembly statistics (N50 length, accuracy and genome fraction) of each assembler at different coverage depths based on HM-276D data set. Colors indicate results from different assemblers (See **Supplementary material** for details in parameter settings). **(b)** Assembly statistics (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 data set was down-sampled to 160× depth of coverage. **(c)** Strain-specific assembly performance of each assembler based on HM-277D data set. Assembly statistics (accuracy and genome fraction) distributions were presented using boxplots with jitter. Radius of each dot indicates the known relative abundance of each bacteria strain from the mock community.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Next, we subsampled 525× data set 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. The assembly results of 4 tools ranges 95.95% to 99.96% in consensus accuracy and 91.26% to 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%). Despite of reduced metrics as sequencing depth becoming 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 520× to 80×. In addition, we examined whether genomes of 20 bacterial strains can be better constructed with Nanopore sequencing technology compared to PacBio and Illumina. As shown in **Figure 2(b)**, 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 PacBio, genome fraction and genome size are 1.56% and 3.1 Mb higher respectively. Assemblies using Illumina sequenced data are 99.9% in accuracy, but with more contigs generated and lower genome size in total compared to 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 1068× data set from HM-277D Mock Community, with unevenly mixed DNA samples of the 20 bacteria strains. Assembly accuracy still remains

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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: 59.7%) and genome sizes (Canu: 50.21 Mb, OPERA-MS: 47.99 Mb, meta-flye: 64.12 Mb, wtdbg2: 41.85 Mb). To assess how strain abundance affects assemblies, we calculated strain-specific genome fraction for each tool in **Figure 2(a)**. Across bacterial strains, meta-flye recovered the highest percentage of genome (median 100%), followed by OPERA-MS (median: 98.75%) and Canu (median 94.78%), while assemblies of wtdbg2 covered only 31.22% (median). 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 belong to two different bacteria species, *Bacteroides vulgatus* (0.19% relative abundance) and *Streptococcus pneumoniae* (0.05% relative abundance), indicating the difficulty in differentiating one bacteria from another with low relative abundance.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

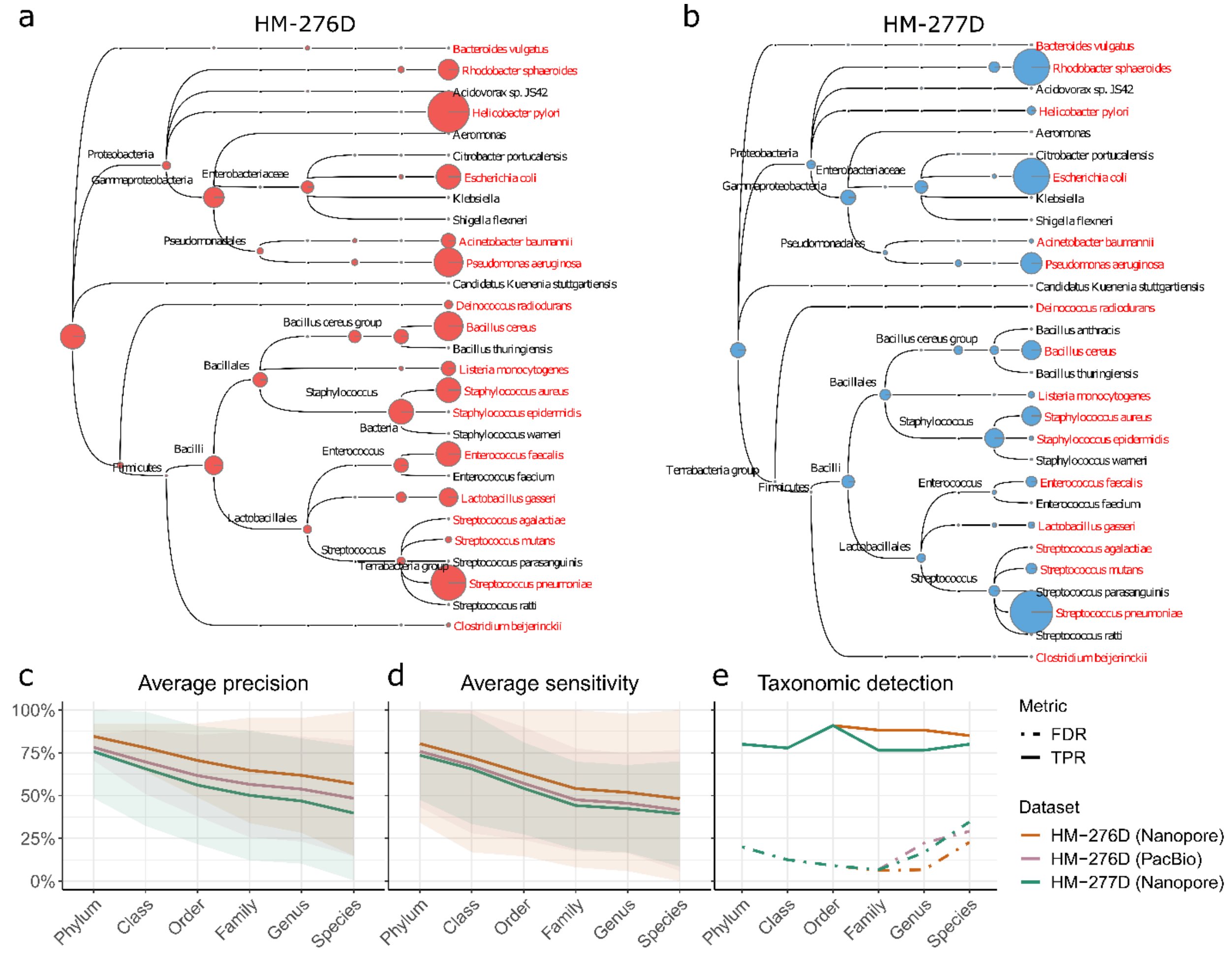


Figure 3. Taxonomic binning results for HM-276D and HM-277D data sets. (a,b) Megan taxonomic tree assignment obtained from HM-276D **(a)** and HM-277D **(b)** Nanopore sequenced data sets. Both data sets were downsampled to 160× depth of coverage. Each read was aligned against NCBI-nr protein reference 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. 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 data sets (indicated by colors). Average **(c)** precision and **(e)** sensitivity 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

identified taxon (reads > 10) at each rank. To make fair comparison, each data set was downsampled to 160× depth of coverage.

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 3 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 (**Figure 3(a,b)**). Overall, 4.22 Mb (0.087%) from Nanopore data of HM-276D sample were mis-assigned, while 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 bacteria 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 3(c)**, HM-276D (Nanopore) has the highest precision, which are all above 60% from phylum to genus. HM-277D (Nanopore)

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

followed, with all above 50%, while HM-276D (PacBio) has the lowest average precision due to predicted small false positive bins at the species level. Sensitivity has similar pattern (**Figure 3(d)**). HM-276D (Nanopore) still appears to the best data set for read classification than other two and the difference in accuracy between these 3 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 scenarios 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 bacteria 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, while 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, while identity with less 10 supporting reads was marked as absence. We considered two other metrics to quantify the detection accuracy, true positive rate (TPR) and false discover 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 3(e). TPR were consistent across 3 data sets from phylum to order level (90%-77%). Below the order level, PacBio (HM-276D) and Nanopore (HM-277D) are 22% lower compared to Nanopore (HM-276D) (92%-87%). From phylum to family level, FDRs were controlled under 15% for all 3 data sets. However, at the genus level, more than 20% of detections are false for PacBio (HM-276D) and Nanopore (HM-277D) while 6% for Nanopore (HM-276). All 3 scenarios have inflated FDR (>20%) at the species level. Across data sets, 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 data sets across ranks. However, accurately predicting taxonomic profiles at the species level still remains challenging due to many false predicted taxonomic identities with 10 to 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 data set at strain level. First, we examined the ability in identifying these 13 extremely rare strains based on annotated target genes. To explore the sensitivity of strain detection using this data set, we mapped raw sequenced reads to reference genomes of the 20 bacterial strains with Minimap2. Then, for each strain-specific gene, the average coverage were estimated by summing up read depth across all exonic region, normalized for gene length. In addition, exon coverage fractions were

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 **Figure 4(a)**. Detection rates and average coverage among all genes largely keep high in abundant strains (>1%), ranging from 96.4 bp to 4207.6 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 RefSeq annotation. As shown in **Figure 4(a)**, though *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. 14 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 to the gold

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

standard. Relative abundance was obtained by normalizing total read coverage with chromosome length for each taxon at different ranks. As shown in **Figure 4(b)**, 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, while 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 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

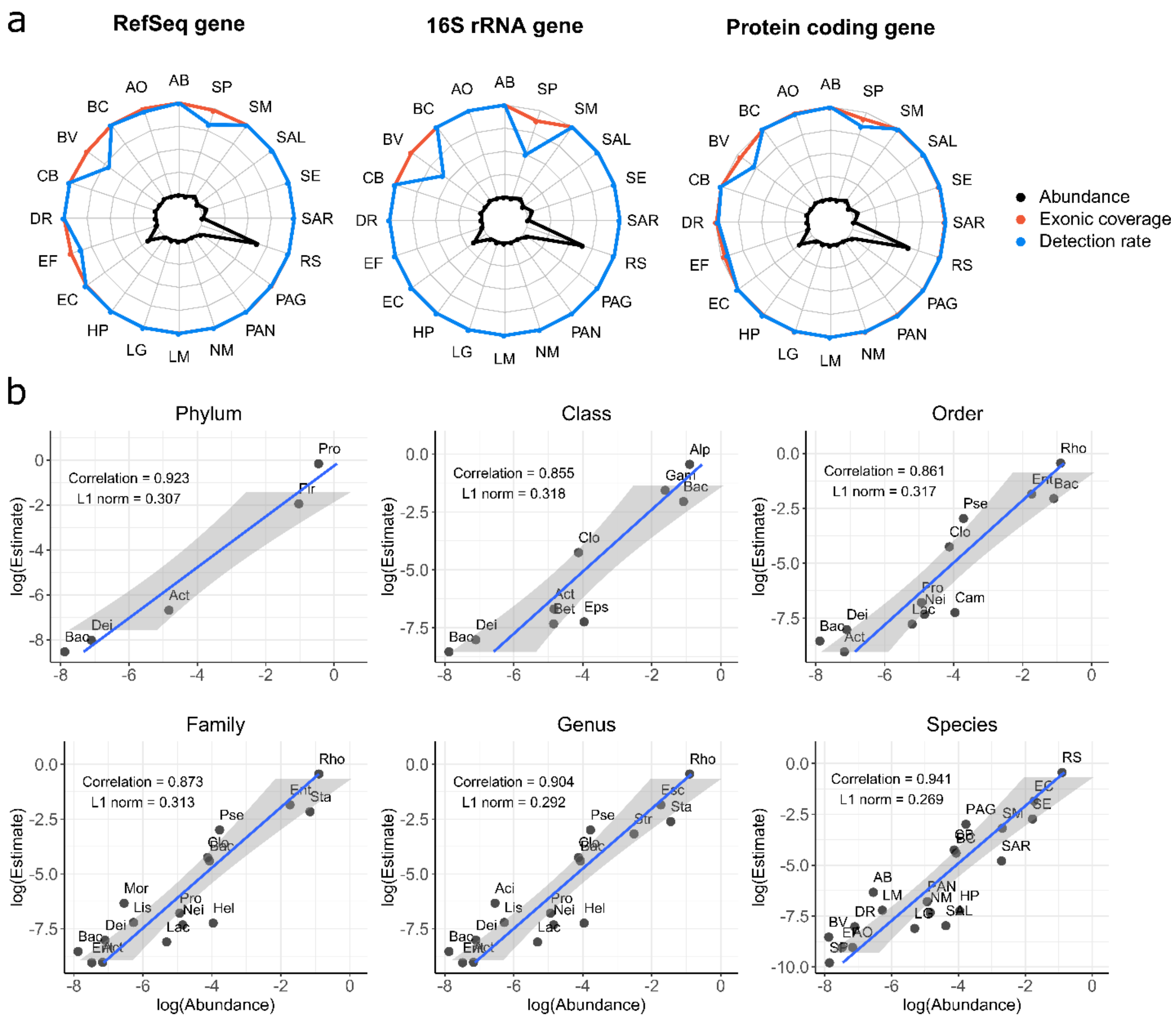


Figure 4. Taxonomic profiling results for HM-277D data sets. (a) Gene identification performance of 20 bacterial strains. 3 gene sets (RefSeq, 16S rRNA, protein coding) were evaluated. Colors indicate different metrics (exonic coverage and detection rate). Exonic coverage (orange) is the percentage of exonic region covered by at least 1 read out 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)** Bacteria abundance estimation. Scatter plots 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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).

Discussion

Complete genome assembly and population profiling are critical for the interpretation of microbial community diversity. However, a benchmarking long-read data set 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

sequencing data and bioinformatics tools in characterizing microbial community. Assemblers consistently achieved high accuracy and completeness for nontrivial bacteria 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 that 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 is always a major challenge for genome assembly. In the future, we anticipate

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 to 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 to R9. 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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, 4 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 4 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 remain high for all 4 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 while meta-flye,

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 while performance decreased at species and strain level.

In summary, our results demonstrate the feasibility to characterize 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 data sets on reference microbiomes are made publicly available to facilitate benchmarking studies on metagenomics and the development of novel software tools.

Methods and materials

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).

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 Canu and meta-flye are designed to be capable to handle metagenome while wdtbg2 and OPERA-MS 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[5] and HM-277D data sets[38]. 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[33] 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. In specific, we specified estimated genome size as 70M, where the parameters are “-x ont -g 70m –t 20” for wtdbg2, “genomeSize=70M useGrid=True” for Canu, and

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

"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. 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-

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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[25] 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[26], 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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.

Acknowledgements

This work was supported by CHOP Research Institute 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.

Competing interests

The authors declare no conflict of interest.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

1. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE: **Metagenomic analysis of the human distal gut microbiome.** *Science* 2006, **312:**1355-1359.

2. Lewis JD, Chen EZ, Baldassano RN, Otley AR, Griffiths AM, Lee D, Bittinger K, Bailey A, Friedman ES, Hoffmann C, et al: **Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease.** *Cell Host Microbe* 2015, **18:**489-500.

3. Chehoud C, Albenberg LG, Judge C, Hoffmann C, Grunberg S, Bittinger K, Baldassano RN, Lewis JD, Bushman FD, Wu GD: **Fungal Signature in the Gut Microbiota of Pediatric Patients With Inflammatory Bowel Disease.** *Inflamm Bowel Dis* 2015, **21:**1948-1956.

4. Hooper LV, Stappenbeck TS, Hong CV, Gordon JI: **Angiogenins: a new class of microbicidal proteins involved in innate immunity.** *Nat Immunol* 2003, **4:**269-273.

5. Jones MB, Highlander SK, Anderson EL, Li W, Dayrit M, Klitgord N, Fabani MM, Seguritan V, Green J, Pride DT, et al: **Library preparation methodology can influence genomic and functional predictions in human microbiome research.** *Proc Natl Acad Sci U S A* 2015, **112:**14024-14029.

6. Ley RE, Turnbaugh PJ, Klein S, Gordon JI: **Microbial ecology: human gut microbes associated with obesity.** *Nature* 2006, **444:**1022-1023.

7. Liang X, Bittinger K, Li X, Abernethy DR, Bushman FD, FitzGerald GA: **Bidirectional interactions between indomethacin and the murine intestinal microbiota.** *Elife* 2015, **4:**e08973.

8. Sartor RB: **Microbial influences in inflammatory bowel diseases.** *Gastroenterology* 2008, **134:**577-594.

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

10. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, et al: **A core gut microbiome in obese and lean twins.** *Nature* 2009, **457:**480-484.

11. Wang F, Kaplan JL, Gold BD, Bhasin MK, Ward NL, Kellermayer R, Kirschner BS, Heyman MB, Dowd SE, Cox SB, et al: **Detecting Microbial Dysbiosis Associated with Pediatric Crohn Disease Despite the High Variability of the Gut Microbiota.** *Cell Rep* 2016, **14:**945-955.

12. Wen L, Ley RE, Volchkov PY, Stranges PB, Avanesyan L, Stonebraker AC, Hu C, Wong FS, Szot GL, Bluestone JA, et al: **Innate immunity and intestinal microbiota in the development of Type 1 diabetes.** *Nature* 2008, **455:**1109-1113.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

13. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R, et al: **Linking long-term dietary patterns with gut microbial enterotypes.** *Science* 2011, **334:**105-108.

14. Group NHW, Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Bonazzi V, McEwen JE, Wetterstrand KA, et al: **The NIH Human Microbiome Project.** *Genome Res* 2009, **19:**2317-2323.

15. Janda JM, Abbott SL: **16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls.** *J Clin Microbiol* 2007, **45:**2761-2764.

16. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N: **Shotgun metagenomics, from sampling to analysis.** *Nat Biotechnol* 2017, **35:**833-844.

17. Hao X, Chen T: **OTU analysis using metagenomic shotgun sequencing data.** *PLoS One* 2012, **7:**e49785.

18. Chen EZ, Bushman FD, Li H: **A Model-Based Approach For Species Abundance Quantification Based On Shotgun Metagenomic Data.** *Stat Biosci* 2017, **9:**13-27.

19. Ruan J, Li H: **Fast and accurate long-read assembly with wtdbg2.** *Nat Methods* 2019.

20. Bertrand D, Shaw J, Kalathiyappan M, Ng AHQ, Kumar MS, Li C, Dvornicic M, Soldo JP, Koh JY, Tong C, et al: **Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes.** *Nat Biotechnol* 2019, **37:**937-944.

21. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM: **Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation.** *Genome Res* 2017, **27:**722-736.

22. Kolmogorov M, Yuan J, Lin Y, Pevzner PA: **Assembly of long, error-prone reads using repeat graphs.** *Nat Biotechnol* 2019, **37:**540-546.

23. Li D, Liu CM, Luo R, Sadakane K, Lam TW: **MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph.** *Bioinformatics* 2015, **31:**1674-1676.

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

26. Huson DH, Beier S, Flade I, Gorska A, El-Hadidi M, Mitra S, Ruscheweyh HJ, Tappu R: **MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data.** *PLoS Comput Biol* 2016, **12:**e1004957.

27. Francis OE, Bendall M, Manimaran S, Hong C, Clement NL, Castro-Nallar E, Snell Q, Schaalje GB, Clement MJ, Crandall KA, Johnson WE: **Pathoscope: species identification and strain attribution with unassembled sequencing data.** *Genome Res* 2013, **23:**1721-1729.

28. Hong C, Manimaran S, Shen Y, Perez-Rogers JF, Byrd AL, Castro-Nallar E, Crandall KA, Johnson WE: **PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples.** *Microbiome* 2014, **2:**33.

29. Byrd AL, Perez-Rogers JF, Manimaran S, Castro-Nallar E, Toma I, McCaffrey T, Siegel M, Benson G, Crandall KA, Johnson WE: **Clinical PathoScope: rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data.** *BMC Bioinformatics* 2014, **15:**262.

30. Jovel J, Patterson J, Wang W, Hotte N, O'Keefe S, Mitchel T, Perry T, Kao D, Mason AL, Madsen KL, Wong GK: **Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics.** *Front Microbiol* 2016, **7:**459.

31. Laudadio I, Fulci V, Palone F, Stronati L, Cucchiara S, Carissimi C: **Quantitative Assessment of Shotgun Metagenomics and 16S rDNA Amplicon Sequencing in the Study of Human Gut Microbiome.** *OMICS* 2018, **22:**248-254.

32. Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL: **Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing.** *Biochem Biophys Res Commun* 2016, **469:**967-977.

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

34. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Droge J, Gregor I, Majda S, Fiedler J, Dahms E, et al: **Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software.** *Nat Methods* 2017, **14:**1063-1071.

35. Mason CE, Afshinnekoo E, Tighe S, Wu S, Levy S: **International Standards for Genomes, Transcriptomes, and Metagenomes.** *J Biomol Tech* 2017, **28:**8-18.

36. Nicholls SM, Quick JC, Tang S, Loman NJ: **Ultra-deep, long-read nanopore sequencing of mock microbial community standards.** *Gigascience* 2019, **8**.

37. McIntyre ABR, Alexander N, Grigorev K, Bezdan D, Sichtig H, Chiu CY, Mason CE: **Single-molecule sequencing detection of N6-methyladenine in microbial reference materials.** *Nat Commun* 2019, **10:**579.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

39. Leggett RM, Alcon-Giner C, Heavens D, Caim S, Brook TC, Kujawska M, Martin S, Hoyles L, Clarke P, Hall LJ: **Rapid profiling of the preterm infant gut microbiota using nanopore sequencing aids pathogen diagnostics.** *bioRxiv* 2018**:**180406.

40. Leger A, Leonardi T: **pycoQC, interactive quality control for Oxford Nanopore Sequencing.** *The Journal of Open Source Software* 2019, **4:**1236.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Material for Evaluation of single-molecule long-read whole-genome shotgun sequencing on characterizing reference microbiomes

Yu Hu*, Li Fang*, Christopher Nicholson, Kai Wang[§]

[§] Correspondence to: wangk@email.chop.edu

*** These authors contributed equally to this work**

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Tables

Tools	Depth	N50 length	Accuracy (%)	Coverage fraction (%)	NG50 length	# contigs	# long contigs	Longest contig	Genome size
Canu	20x	717267	98.5	96.8	616530	298	254	2612567	65503873
	40x	1987236	99.07	99.29	1975600	132	112	6286130	67676017
	80x	2886059	99.24	99.86	2731942	62	57	6316623	68735511
	160x	3901381	99.27	99.93	3901381	60	52	6299115	68879111
	365x	2983818	99.28	99.83	2983818	64	58	6292103	68964121
	480x	3911963	99.4	99.81	3911963	83	65	6359094	69425747
OPERA-MS	20x	1122204	99.83	99.71	1122204	5117	201	6324007	67168904
	40x	2657727	99.96	99.99	2657727	1695	81	5220208	67629371
	80x	2835709	99.96	99.99	2732545	1921	74	4636570	67632885
	160x	2933262	99.95	98.45	2792941	2347	65	6255842	66580943
	365x	2938016	99.91	99.98	2938016	4734	64	6255878	67858470
	480x	2938019	99.92	99.98	2938019	4732	63	6255756	67892051
wtdbg2	20x	519021	95.95	91.26	400703	443	363	3338270	61551400
	40x	2371130	97.94	98.4	2253156	175	124	6222827	66248572
	80x	3152360	98.73	98.34	2920496	122	80	6230107	66026593
	160x	2863759	98.7	98.08	2863759	91	69	6242719	66161138
	365x	2706888	98.66	97.33	2706888	90	73	6251621	65543654
	480x	2968720	98.73	95.94	2942294	61	53	8884115	64898035
meta-flye	20x	1653589	98.96	98.76	1547909	223	206	5630982	66808399
	40x	2725547	99.43	99.97	2653197	64	52	6274273	67627825
	80x	2930772	99.52	99.99	2930772	59	43	6251934	67630110
	160x	3888260	99.54	99.97	3180529	61	39	6252579	67595608
	365x	3181836	99.62	99.98	2934283	88	44	6245780	67727067
	480x	3181822	99.62	99.99	2934277	89	43	6245565	67700317

Supplementary Table 1. Comprehensive assembly statistics on HM-276D using Canu, OPERA-MS, wtdbg2 and meta-flye.

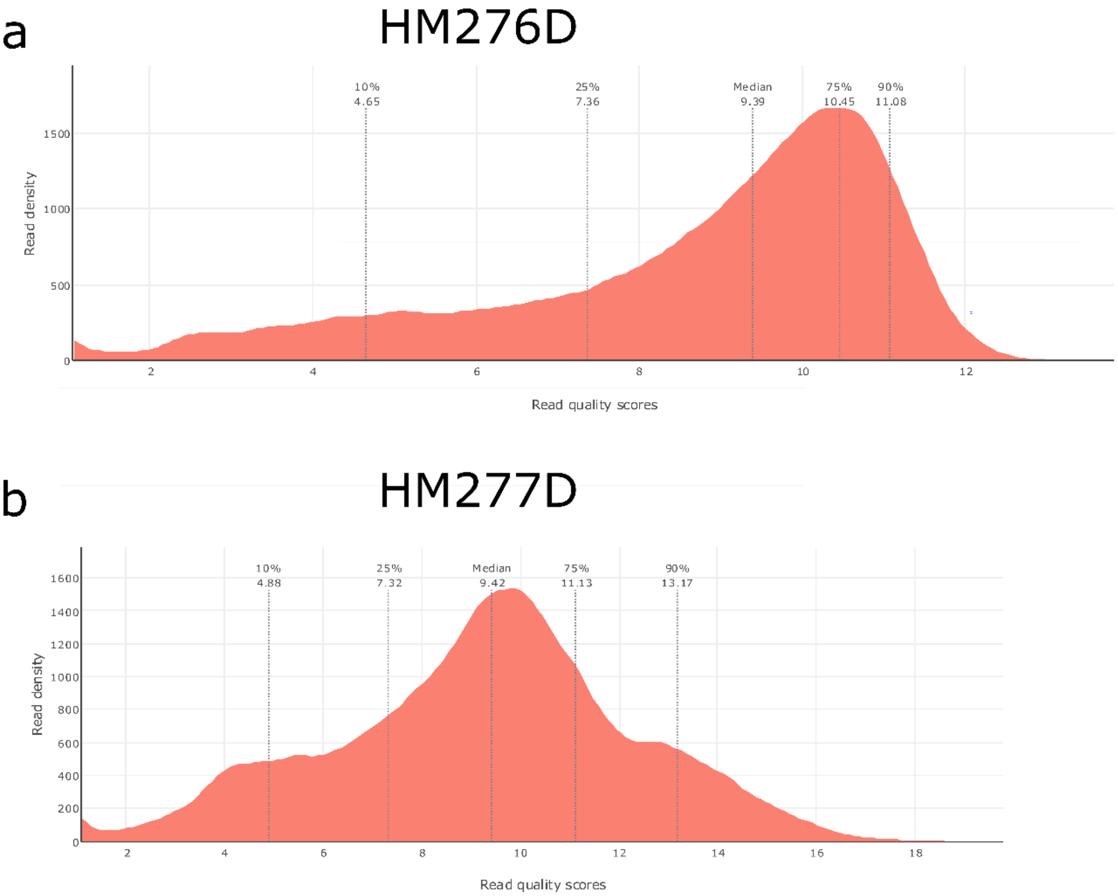
bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

Supplementary Table 2. Species-specific gene coverage summary of HM-277D data set. 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.

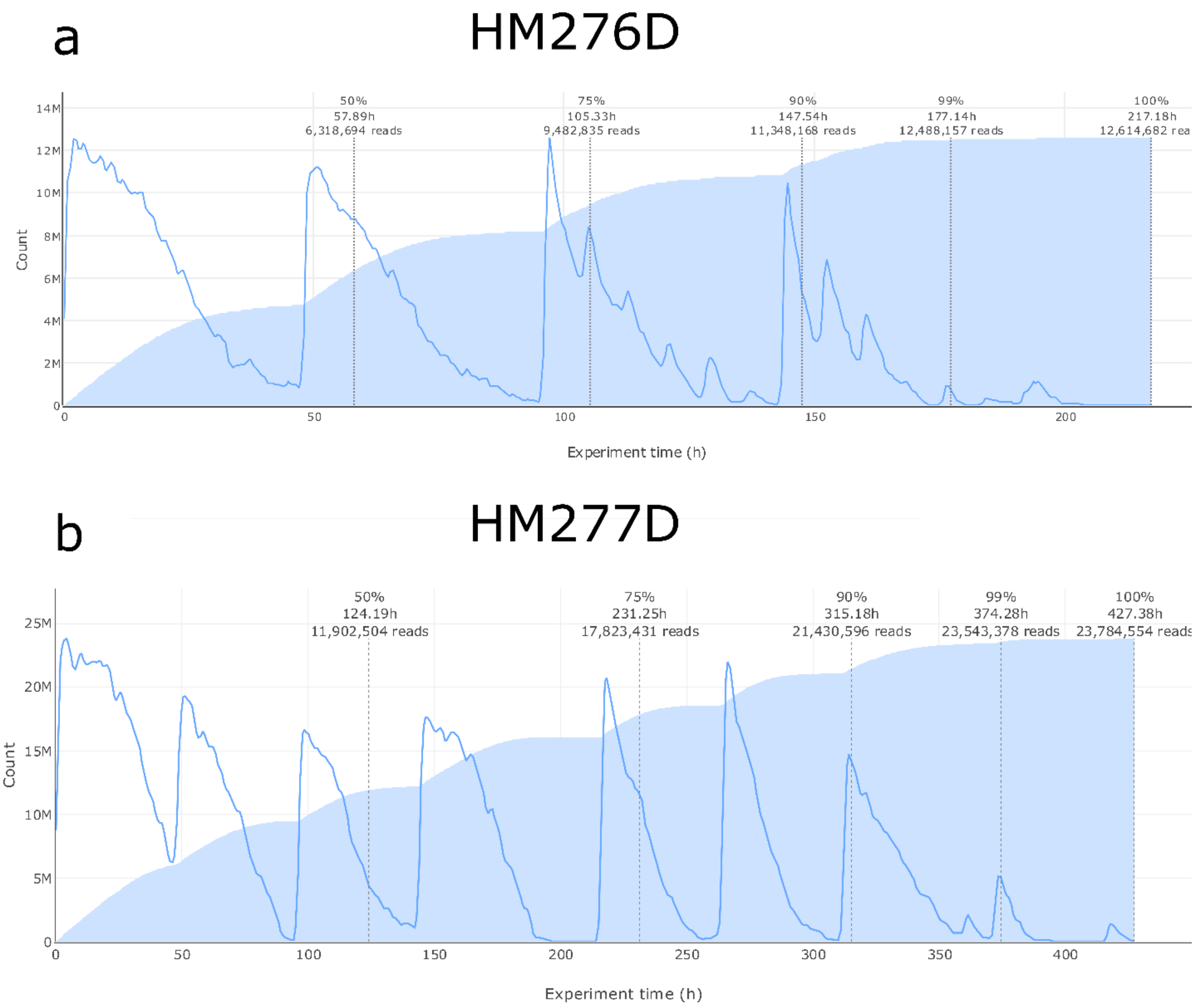
bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Figures



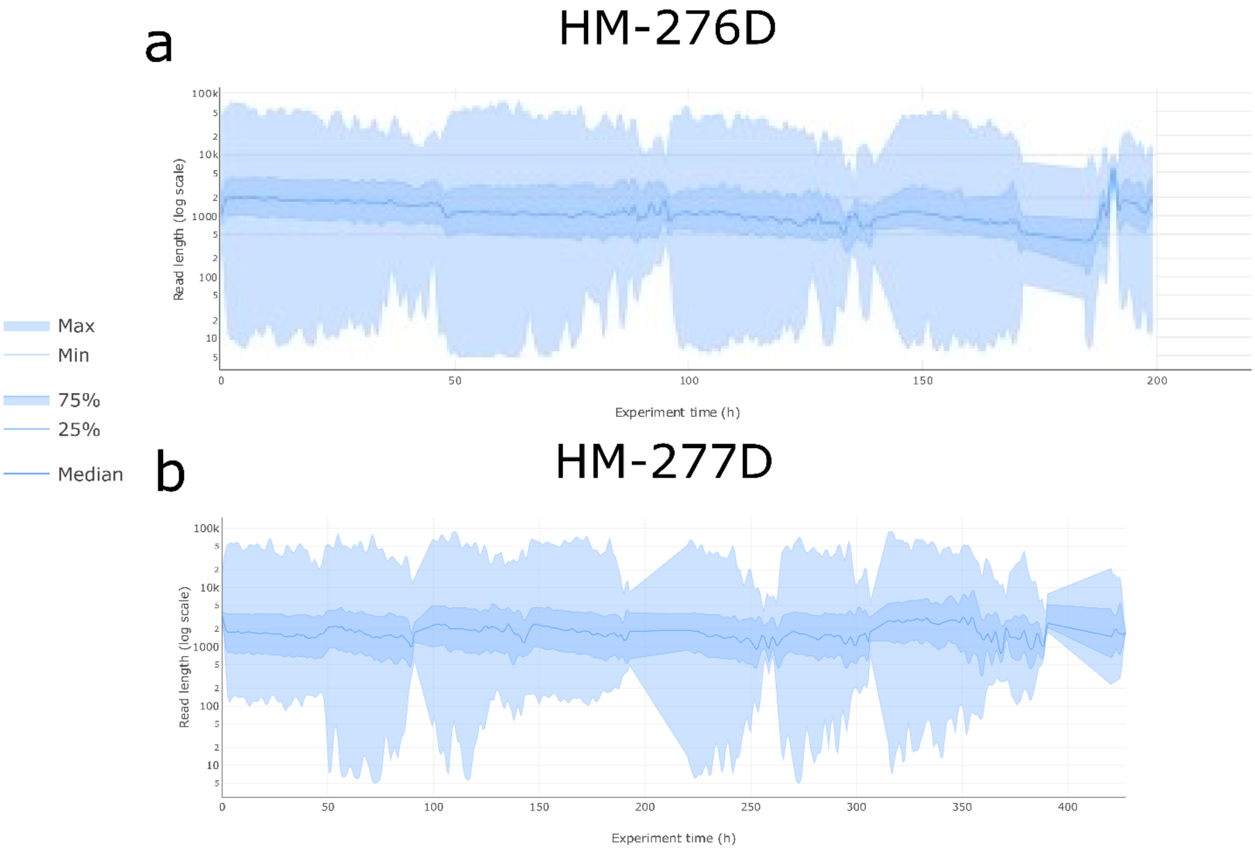
Supplementary Figure 1. Read quality of Nanopore sequencing data. Read quality of sequenced data sets, HM-276D **(a)** and HM-277D **(b)**, were summarized using PycoQC respectively. Dashed lines indicate different quantiles (10%, 25%, 50%, 75%, 90%).

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



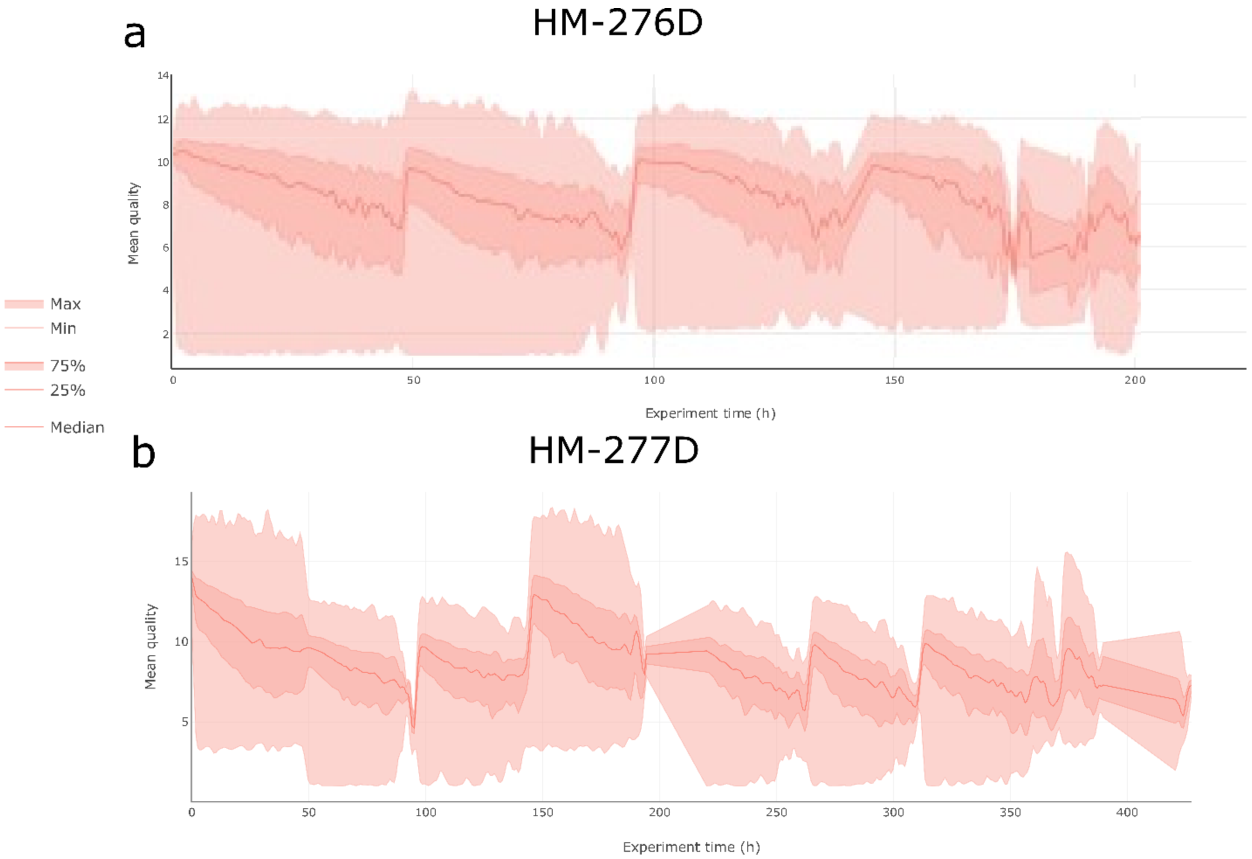
Supplementary Figure 2. Read output over experiment of Nanopore sequencing data. Number of output reads over experiment time for sequenced data sets, HM-276D **(a)** and HM-277D **(b)**, were summarized using PycoQC. Blue line indicates output velocity at specific time. Shaded area represents cumulative read output over experiment time.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



Supplementary Figure 3. Read length over experiment of Nanopore sequencing data. Read length in log scale over experiment time for sequenced data sets, HM-276D **(a)** and HM-277D **(b)**, were summarized using PycoQC.

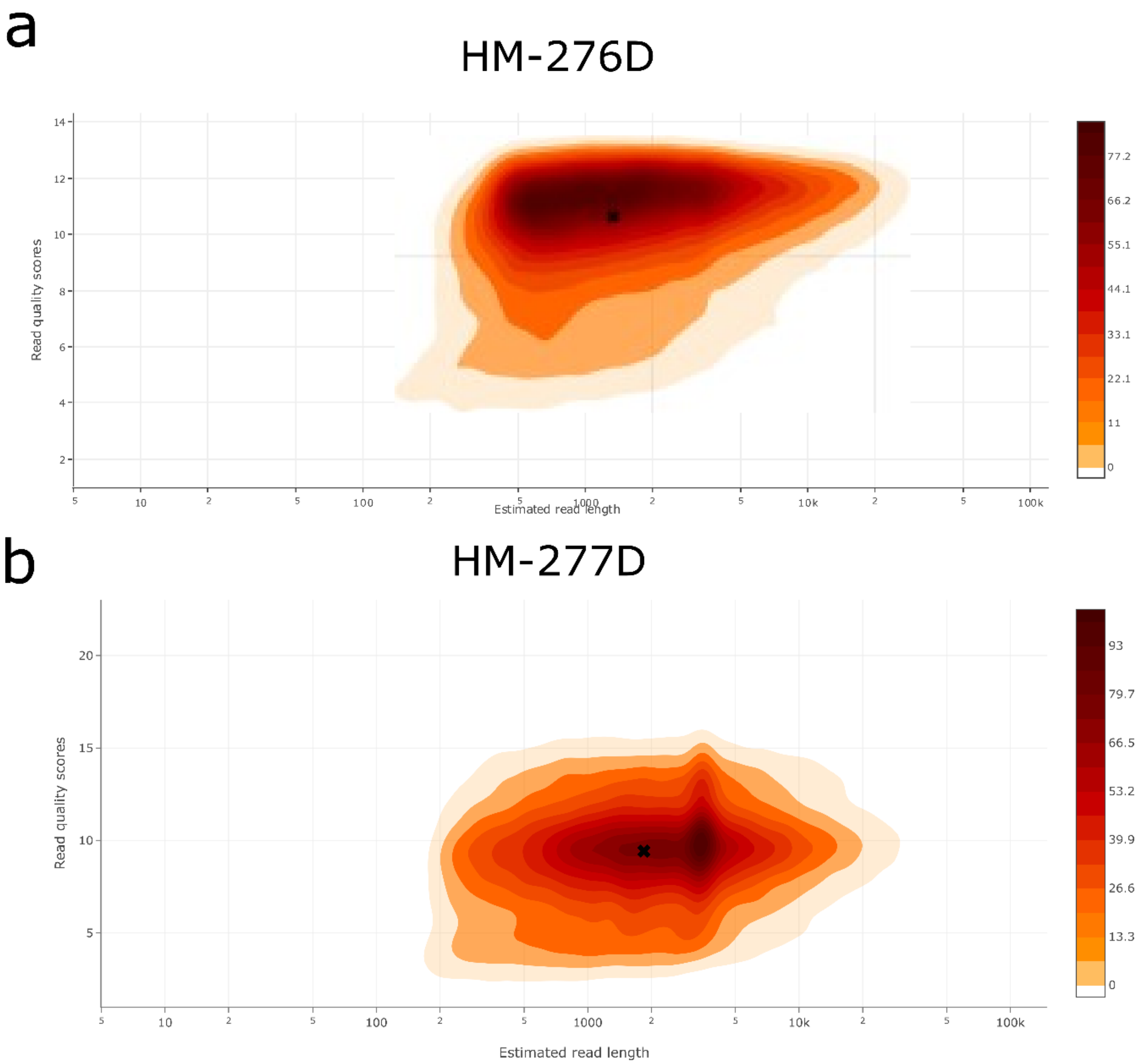
bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



784

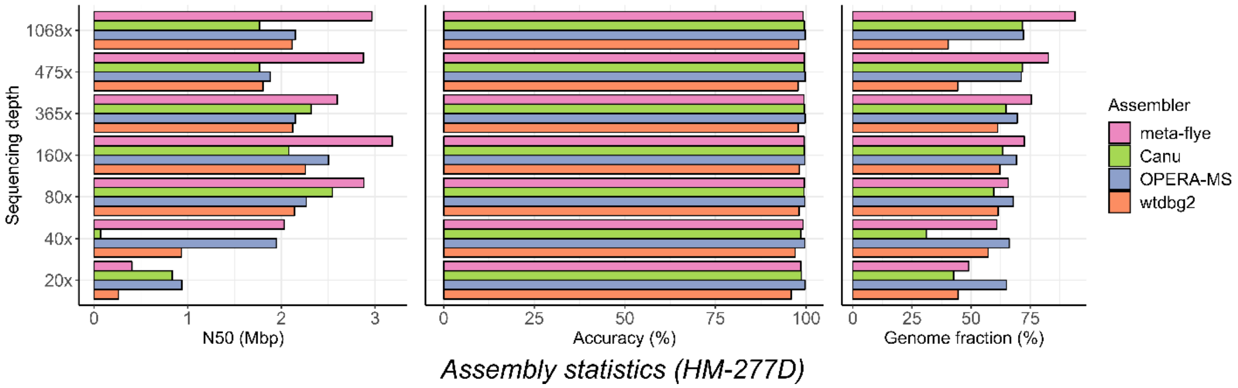
785 **Supplementary Figure 4. Read quality over experiment of Nanopore sequencing**
786 **data.** Mean read quality over experiment time for sequenced data sets, HM-276D **(a)**
787 and HM-277D **(b)**, were summarized using PycoQC.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



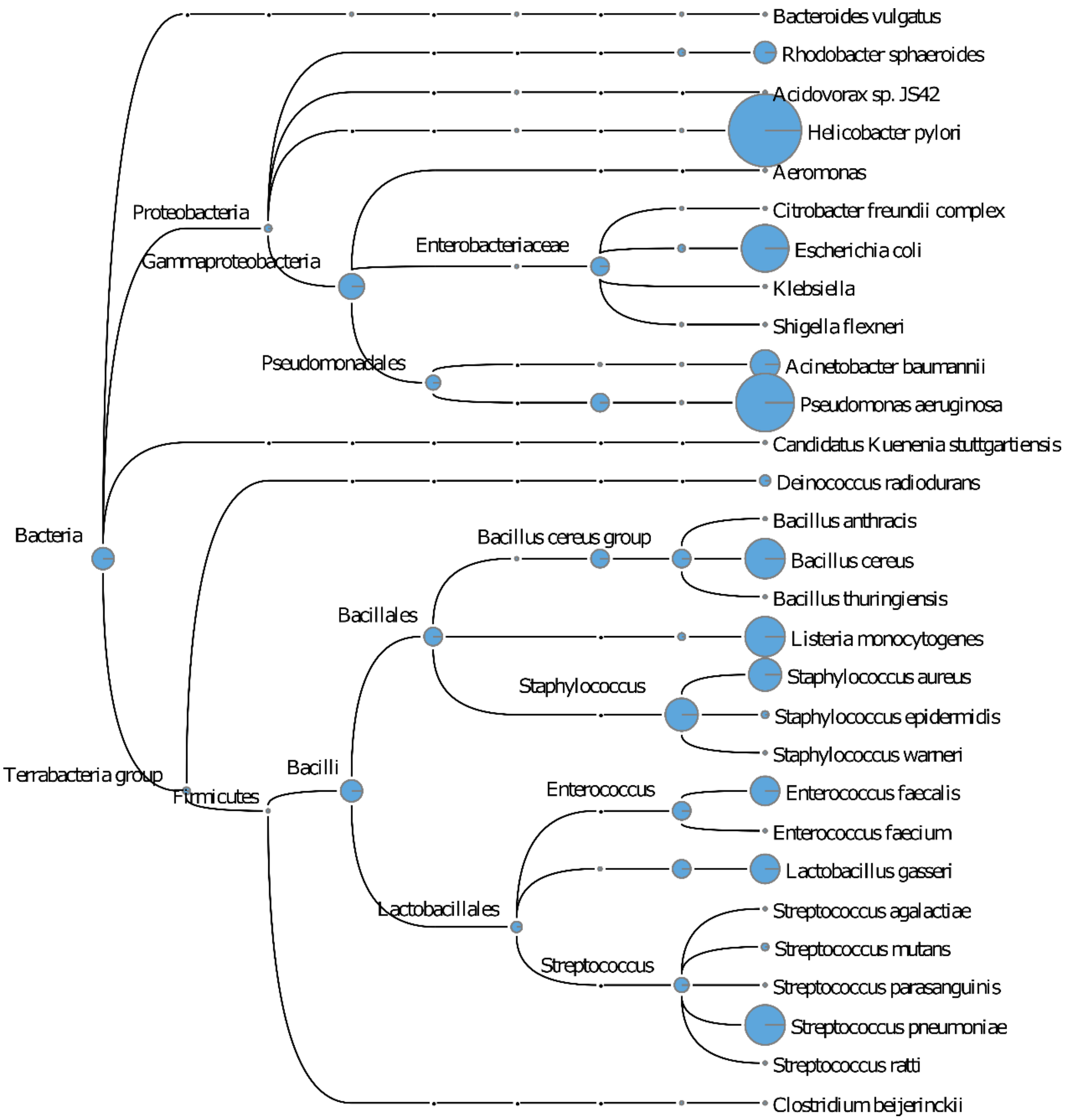
Supplementary Figure 5. Read quality score vs estimated read length. Nanopore read distribution of read length and quality score for sequenced data sets, HM-276D **(a)** and HM-277D **(b)**, were summarized using PycoQC. Color indicates read density.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



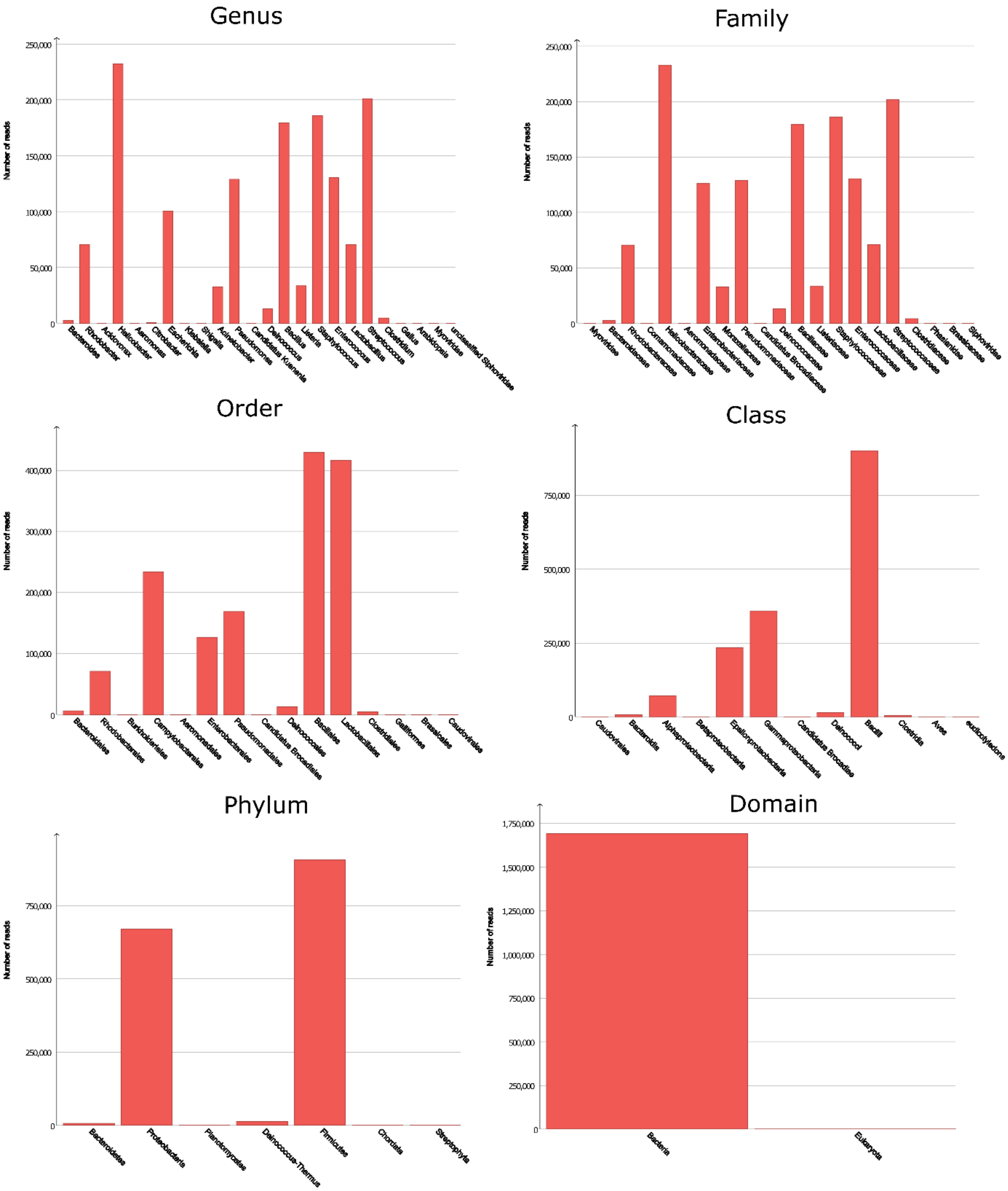
Supplementary Figure 6. Assembly performance on HM-277D data set. Assembly statistics (N50 length, accuracy and genome fraction) of each assembler at different coverage depths based on HM-277D data set. Colors indicate results from different assemblers (Canu, OPERA-MS, wtdbg2, meta-flye). Assembly accuracy remains high compared to HM-276D, ranging around ~99% across tools. N50 lengths and genome fractions of all methods are substantially lower than even community.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



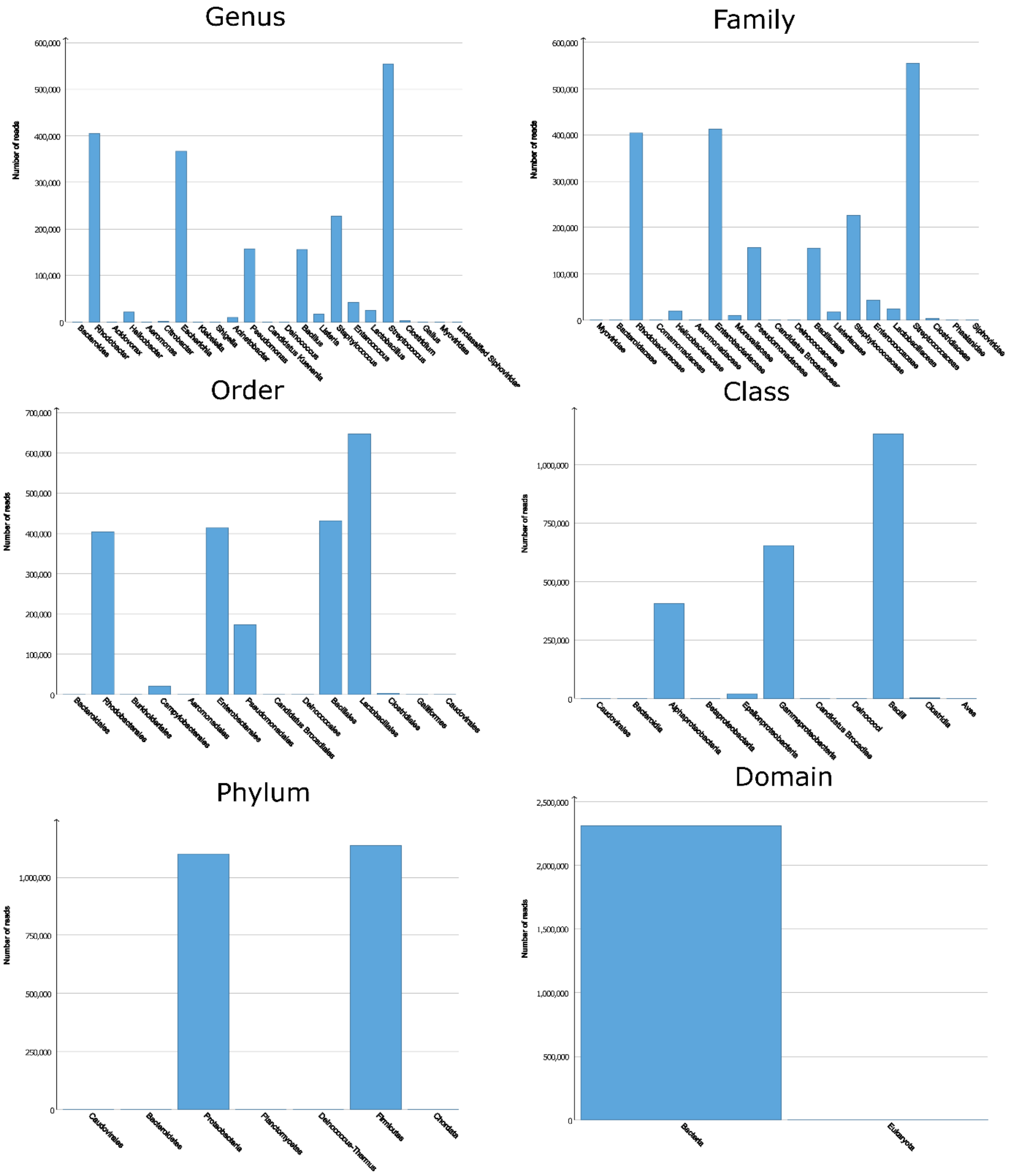
Supplementary Figure 7. Megan taxonomic tree assignment obtained from HM-276 PacBio sequenced data set. HM-276D PacBio data set was subsampled to 160× depth of coverage. Each read was aligned against NCBI-nr protein reference data base, then binned and visualized using Megan-LR. Megan taxonomic tree showing bacteria taxa identified and their corresponding abundances across taxonomic rank. The radius of circle represents the number of reads assigned for each taxa.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



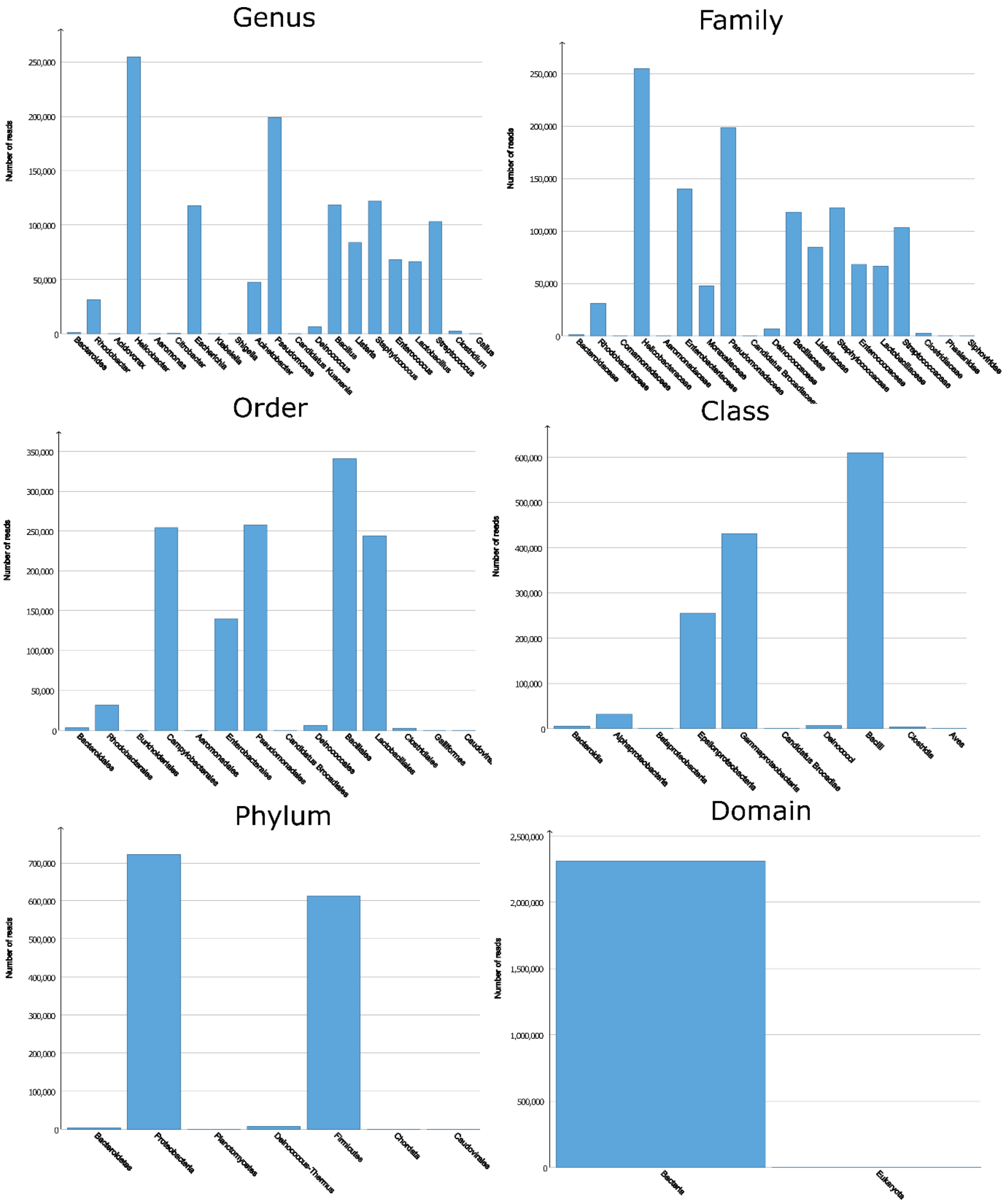
Supplementary Figure 8. Megan taxonomic read distribution at different ranks obtained from HM-276 Nanopore sequenced data set. HM-276D Nanopore data set was subsampled to 160× depth of coverage. Each read was aligned against NCBI-nr protein reference data base, then binned and visualized using Megan-LR.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



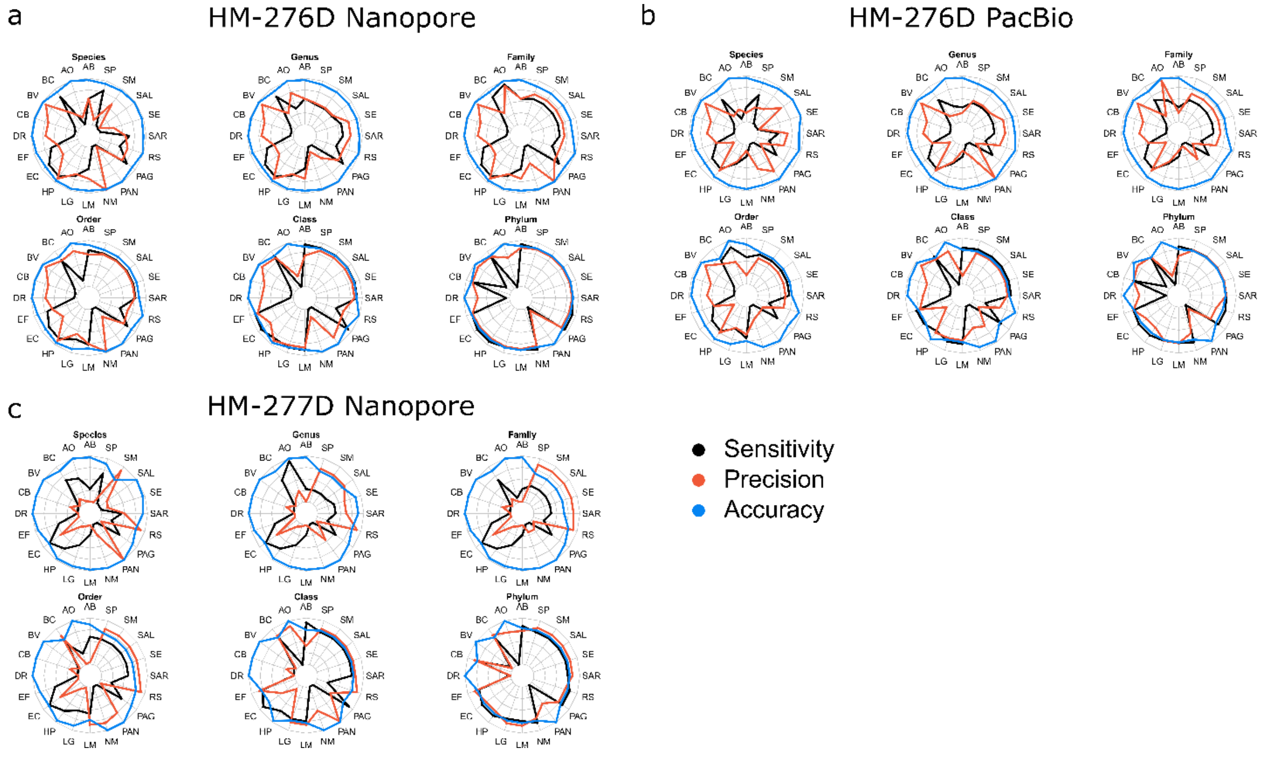
Supplementary Figure 9. Megan taxonomic read distribution at different ranks obtained from HM-277 Nanopore sequenced data set. HM-277D Nanopore data set was subsampled to 160× depth of coverage. Each read was aligned against NCBI-nr protein reference data base, then binned and visualized using Megan-LR.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



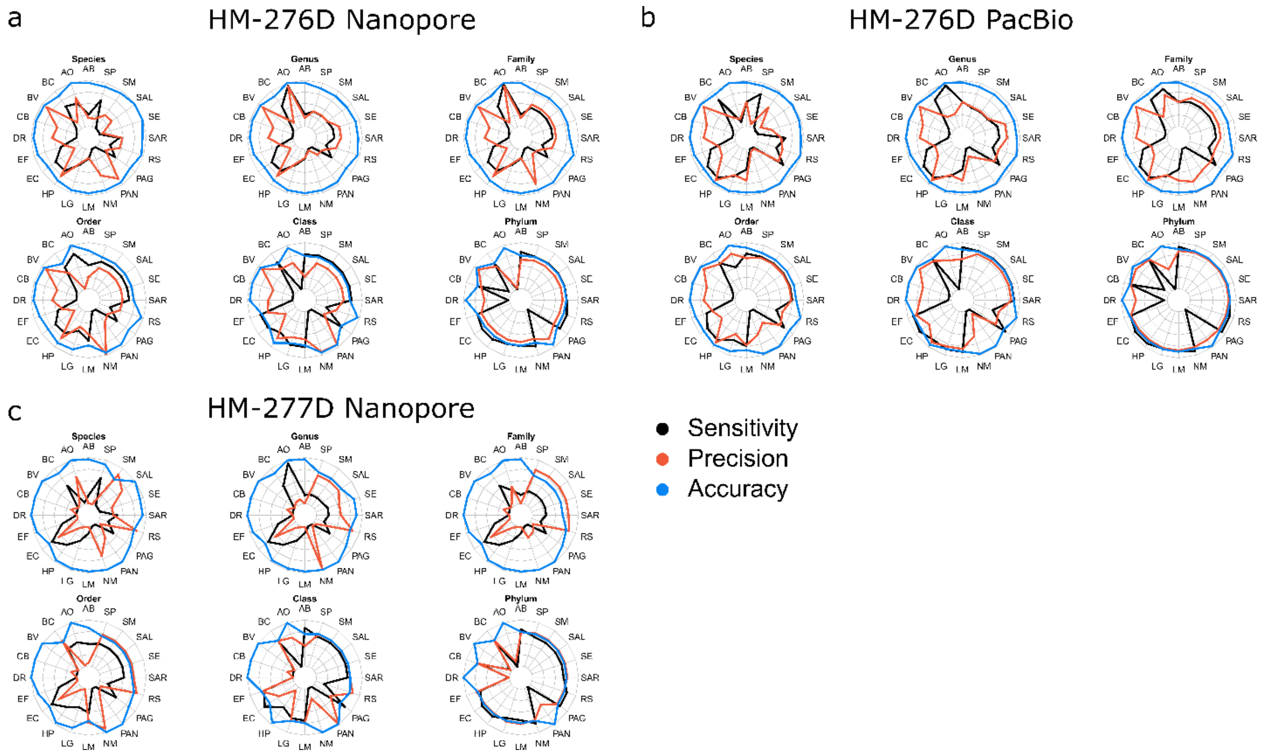
Supplementary Figure 10. Megan taxonomic read distribution at different ranks obtained from HM-276 PacBio sequenced data set. HM-276D PacBio data set was subsampled to 160× depth of coverage. Each read was aligned against NCBI-nr protein reference data base, then binned and visualized using Megan-LR.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



Supplementary Figure 11. Strain-specific read assignment performance comparison across sequencing technologies. 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*.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.05.978866; this version posted March 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



Supplementary Figure 12. Strain-specific base pair assignment performance comparison across sequencing technologies. 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*.