

BRIEF REPORT

Open Access



Nanopore long-read-only metagenomics enables complete and high-quality genome reconstruction from mock and complex metagenomes

Lei Liu, Yu Yang, Yu Deng and Tong Zhang*

Abstract

Background: The accurate and comprehensive analyses of genome-resolved metagenomics largely depend on the reconstruction of reference-quality (complete and high-quality) genomes from diverse microbiomes. Closing gaps in draft genomes have been approaching with the inclusion of Nanopore long reads; however, genome quality improvement requires extensive and time-consuming high-accuracy short-read polishing.

Results: Here, we introduce NanoPhase, an open-source tool to reconstruct reference-quality genomes from complex metagenomes using only Nanopore long reads. Using Kit 9 and Q20+ chemistries, we first evaluated the feasibility of NanoPhase using a ZymoBIOMICS gut microbiome standard (including 21 strains), then sequenced the complex activated sludge microbiome and reconstructed 275 MAGs with median completeness of ~ 90%. As a result, NanoPhase improved the MAG contiguity (median MAG N50: 735 Kb, 44-86X compared to conventional short-read-based methods) while maintaining high accuracy, allowing for a full and accurate investigation of target microbiomes. Additionally, leveraging these high-contiguity reference-quality genomes, we identified 165 prophages within 111 MAGs, with 5 as active prophages, indicating the prophage was a neglected source of genetic diversity within microbial populations and influencer in shaping microbial composition in the activated sludge microbiome.

Conclusions: Our results demonstrated that NanoPhase enables reference-quality genome reconstruction from complex metagenomes directly using only Nanopore long reads. Furthermore, besides the 16S rRNA genes and bio-synthetic gene clusters, the generated high-accuracy and high-contiguity MAGs improved the host identification of critical mobile genetic elements, e.g., prophage, serving as a genomic blueprint to investigate the microbial potential and ecology in the activated sludge ecosystem.

Keywords: Nanopore sequencing, Long-read metagenomics, Activated sludge microbiome, Reference-quality genome reconstruction, NanoPhase, Prophage

Introduction

The surge of shotgun sequencing data and advances in elaborating metagenomics approaches have drastically promoted our understanding of the diversity of microbial life [1]. In particular, genome-resolved metagenomics has been massively employed to unveil the black box of uncultured microbial majority, providing genome-level insights and expanding the tree of life [2] since the first

*Correspondence: zhangt@hku.hk

Environmental Microbiome Engineering and Biotechnology Laboratory, Center for Environmental Engineering Research, Department of Civil Engineering, The University of Hong Kong, Hong Kong SAR, China



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

metagenome-assembled genomes (MAGs) were reconstructed in 2004 [3]. However, even high-quality MAGs could be highly fragmented (worse in complex metagenomes) when assembled using the conventional short-read metagenomics, thus missing crucial genetic information [4]. To bridge the genome gaps, error-prone long reads generated on Nanopore and PacBio platforms were introduced, contributing to the leverage of hybrid (short- and long-read) sequencing strategy and the renaissance of high-contiguity reference-quality genome reconstruction from diverse microbiomes [5–9].

Quite different from Nanopore sequencing, PacBio sequencing could also generate high-accuracy (> 99%) HiFi reads with sacrificing read length and throughput, resulting in much higher sequencing costs in the metagenome project [10]. While the low entry and sequencing cost facilitates Nanopore sequencing accessible for most research labs, allowing a rapid turnaround time [11, 12]. Although Nanopore sequencing has difficulty fully characterizing long homopolymer regions, introducing insertion/deletion errors [13], the continuous improvement of sequencing accuracy, throughput and theoretically unlimited read length empower much more cost- and time-efficient genome reconstruction [14, 15]. To take advantage of the improvement of Nanopore platforms, here, we introduce NanoPhase, an open-source package to reconstruct reference-quality genomes efficiently from complex metagenomes and explore the metabolic potential based on the near-complete genomes.

Methods

Sample collection, DNA extraction, sequencing, and basecalling of Mock and activated sludge samples

The ZymoBIOMICS gut microbiome standard (catalog number: D6331) includes 21 microbial strains with varying abundances (0.0001–14%) to simulate the human gut microbiome. Microbial composition and reference genomes of the mock community can be accessed at <https://www.zymoresearch.com/products/zymbiomics-gut-microbiome-standard>. The activated sludge (AS) sample was collected on November 4th, 2019, from the aeration tank at the Shatin wastewater treatment plants in Hong Kong. The DNeasy PowerSoil Kit (Qiagen, Hilden, Germany) was used to extract DNA from mock and AS samples following the manufacturer's protocols.

DNA sequencing of Mock and AS samples was on the GridION (Oxford Nanopore Technology, UK) platform using the Ligation Sequencing Kit (SQK-LSK109, the Kit 9 chemistry) on two and five flowcells (R9.4.1), respectively. In addition, another two flowcells (R9.4.1, one flowcell for the mock DNA and the other for the AS DNA) were used to generate long reads using the new Kit 12 (Q20+) chemistry. To evaluate the sequencing

accuracy improvement of the Nanopore platforms, multiple Guppy versions with their best accuracy models were used for basecalling based on generated fast5 files, i.e., Guppy v3.0.3 (hac), Guppy v4.0.11 (hac), Guppy v5.0.16 (sup) and Guppy v6.0.0 (sup), releasing from April 2019 to December 2021. Guppy v5.0.16 (sup) showed the identical basecalling performance with Guppy v6.0.0 (sup), as the latter update was mainly for the Q20+ chemistry. Basecalled Nanopore raw reads were filtered based on the length of 1 Kb and the mean read identity of 80 (QA80), 90 (QA90) and 95 (QA95) using Filtlong (<https://github.com/rwwick/Filtlong>), respectively. The results present in the main text were all based on QA90 reads.

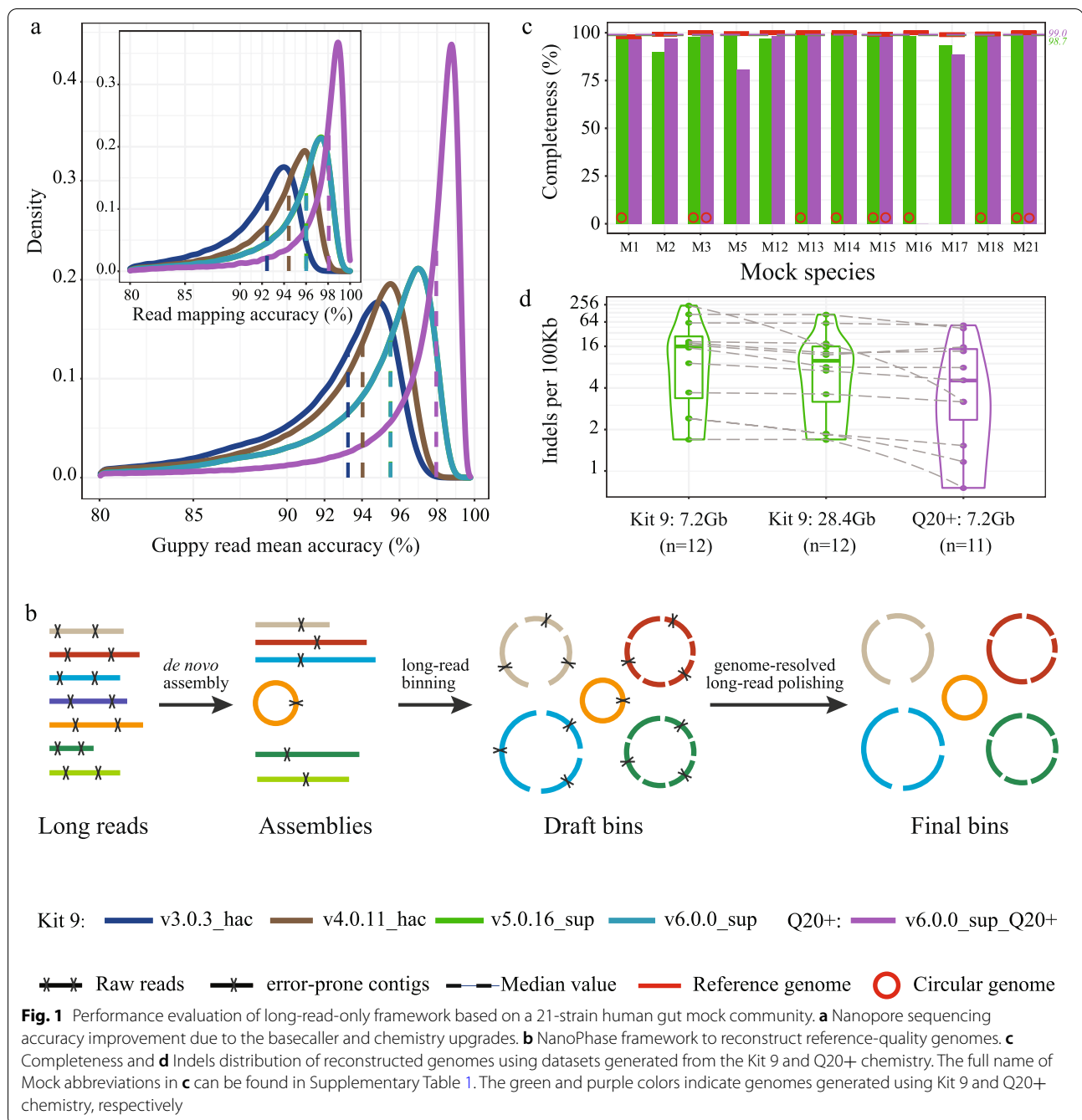
Besides the Nanopore sequencing, the same Mock and AS samples DNA were also delivered to Novogene Company Limited (Beijing, China) for short-read 150bp x 2 paired-end sequencing. In total, we obtained 53.7 Gb short reads for the Mock community and 53.6 Gb for the AS sample.

Nanopore long-read-only framework to reconstruct genomes

Figure 1b illustrates the brief framework of NanoPhase. In detail, metaFlye (v2.9-b1768) [16] was used to assemble filtered Nanopore long reads under the option firstly “--nano-hq -i 5 -g 4m” to generate assemblies. Then MetaBAT2 [17] and MaxBin2 [18] integrated with the coverage information were adopted to reconstruct two candidate genome sets, followed by the bin refinement step of MetaWRAP (v1.3.2) [19] to generate draft bins. Finally, long reads were mapped to the above draft bins using minimap2 (v2.21-r1071; map-ont) [20] with at least 90% identity and 90% coverage, producing draft-bin-based clusters. The long-read cluster was used to polish the draft bins individually with two rounds of Racon (v1.4.22) [21] and three rounds of medaka (v1.4.3; <https://github.com/nanoporetech/medaka>) to generate high-accuracy final bins. GNU Parallel [22] was used to speed up the analysis for parallel computation. This genome-resolved polishing would significantly improve the genome reconstruction efficiency, reducing polishing time and computational requirement, especially for deeply sequenced, complex metagenomes. However, we suggest including one round of Racon and one round of medaka polishing in the future, because additional rounds of long-read polishing were found to have deteriorated the genome quality slightly. Only MAGs with the completeness of above 50% and contamination of below 10% were retained for the downstream analysis.

Short-read polishing and genome accuracy evaluation

The short-read dataset was also divided into different clusters by mapping short reads to polished final bins



using minimap2 (-sr) with at least 90% identity and 90% coverage. Finally, the above final bins were polished using the short-read clusters using multiply rounds of Pilon [23]. The first round of Pilon polishing corrects the most errors, although additional rounds of Pilon polishing could further resolve the remaining errors but had little impact on the genome quality improvement.

Mapping read accuracy was estimated by mapping Nanopore reads to the mock reference genomes with

blastn (v2.5.0) [24]. Misassembly rate and Indels of reconstructed MAGs were evaluated using QUAST (v5.0.2) [25] in the mock dataset. Completeness and contamination of reconstructed MAGs were appraised by CheckM (v1.0.80) [26]. True and observed homopolymer length distribution was computed by Counterr (<https://github.com/dayzerodx/counterr>). As no ground-truth genome standards for the AS metagenome, IDEEL [27], the fraction of predicted full-length proteins in each

MAG, was used as an indirect indicator for evaluating genome accuracy with dependencies Prodigal (v2.6.3) [28] and Diamond (v2.0.13) [29]. Full-length proteins were defined when their length was more than 95% of the best-hit known protein [30] in the UniProt/TrEMBL [31] database (release 2021_04). Genome features, including rRNA operons and tRNA, were identified by Prokka (v1.14.6) [32]. Taxonomic assignment of MAGs was classified with GTDB-Tk (v1.6.0) [33] based on the GTDB R06-RS202 [34].

Prophage and active prophage identification

Prophage sequences within reconstructed MAGs were determined and extracted using VIBRANT (v1.2.1) [35] with default parameters. Then, the prophage sequences with lengths more than 1Kb and their corresponding coordinates were used to estimate prophage activity using PropagAtE (v1.1.0) [36] by providing short reads.

Results and discussion

The primary concern in the application of Nanopore sequencing is the error rate, and the median raw read accuracy for the R9.4 (the current most widely used version) was below 90% in 2019 [7]. However, based on the sequencing dataset of the mock community in this study, the median value of “Guppy read mean accuracy” has been substantially improved by the basecaller upgrades and newly developed chemistry, achieving 95.5% (Kit 9) and 98.0% (Q20+) basecalled by Guppy v6.0.0 (the sup model) (Fig. 1a). In addition, the density profile between “Guppy read mean accuracy” and “Read mapping accuracy” suggested that Nanopore read quality scores predicted by Guppy correlated well to the empirical read accuracy estimated from read-to-reference alignments, and some sequences quality was even underestimated (Fig. 1a and Supplementary Figure S1).

To facilitate rapid genome reconstruction, we proposed NanoPhase, a package to generate MAGs from a single long-read dataset (Fig. 1b). NanoPhase is designed to detangle the complex dataset into different clusters of draft bins and achieve genome-resolved efficient polishing. Totally, 28.4 Gb (N50: 5.9 Kb, two flowcells) and 7.2 Gb (N50: 5.4 Kb, one flowcell) were generated from the mock community using the Kit 9 and Q20+ chemistry, respectively. As expected, bacterial and archaeal genomes with sequencing coverage of $< 5\times$ cannot be reconstructed, and only one *E. coli* MAG was resolved to represent five closely related strains due to a very high average nucleotide identity (98.3–99.4%). Thus, 12 MAGs with median completeness of 98.7% were reconstructed from the Kit 9 dataset, and the Q20+ chemistry demonstrated slightly better performance, recovering 11 MAGs with median completeness of 99.0% (Fig. 1c). MAGs from

both datasets were very close to reference genomes, benefiting from the read-accuracy and homopolymer resolution improvement (Supplementary Figure S2), which was also supported by low Indels errors (Fig. 1d) and high IDEEL [37] scores (Supplementary Figure S3). Notably, 8 (75%) MAGs were assembled into circular, complete genomes in the Kit 9 dataset, more than those generated from the Q20+ dataset (3), mainly due to a much higher sequencing coverage (~ 4 -fold).

We next evaluated the genome reconstruction performance of NanoPhase to resolve a complex AS sample harboring thousands of microbial species. Both Kit 9 and Q20+ chemistries were used for noisy long-read sequencing on five and one flowcells, generating 85.3 (Kit 9 dataset, N50: 6.8 Kb) and 9.4 Gb (Q20+ dataset, N50: 6.5 Kb), respectively. In addition, we observed that filtration of sequencing reads with $> 90\%$ accuracy (QA90) is suitable for genome reconstruction in the complex microbiome, balancing yield and accuracy and generating more reference-quality MAGs (Supplementary Table 2).

Employing the Kit 9 dataset, 275 MAGs were reconstructed with the median completeness, contig count and coverage of 89.5%, 9 and 17X (Fig. 2a), representing 46.9% of the microbial community. Furthermore, the median N50 of these MAGs was 735 Kb, about 44- or 86-fold improvement compared to the short-read methods (Supplementary Table 3), demonstrating that genome gaps were remarkably closed by long reads. Of these MAGs, 94 MAGs with median coverage of 28X were classified as high-quality, fitting the stringent criteria of including full-length rRNA operons [38]. Notably, circular genomes were also recovered.

Compared to short-read-based methods, NanoPhase also reconstructed highly accurate genomes from the complex sample, from the aspect of completeness (median value of 89.5%; Fig. 2b) and IDEEL (median value of 0.58; Fig. 2c). As expected, the Q20+ chemistry performed better with higher IDEEL (median value of 0.64; Fig. 2c) but only generated half of the flowcell (Kit 9 chemistry) throughput. Therefore, the yield limitation has increased the sequencing cost of the Q20+ chemistry in genome reconstruction at present. In addition, short-read-based Pilon polishing of the MAGs did not considerably improve the genome accuracy (Fig. 2b, c), suggesting that it is not an essential step in the future, particularly given the continuous improvement of Nanopore sequencing.

Besides the improved reconstruction of 16S rRNA genes and prediction of secondary metabolites potential discussed in our previous study [7], some mobile genetic elements identification that depended heavily on bacterial isolates could also benefit from these high-contiguity reference-quality genomes, e.g., prophage. As a major

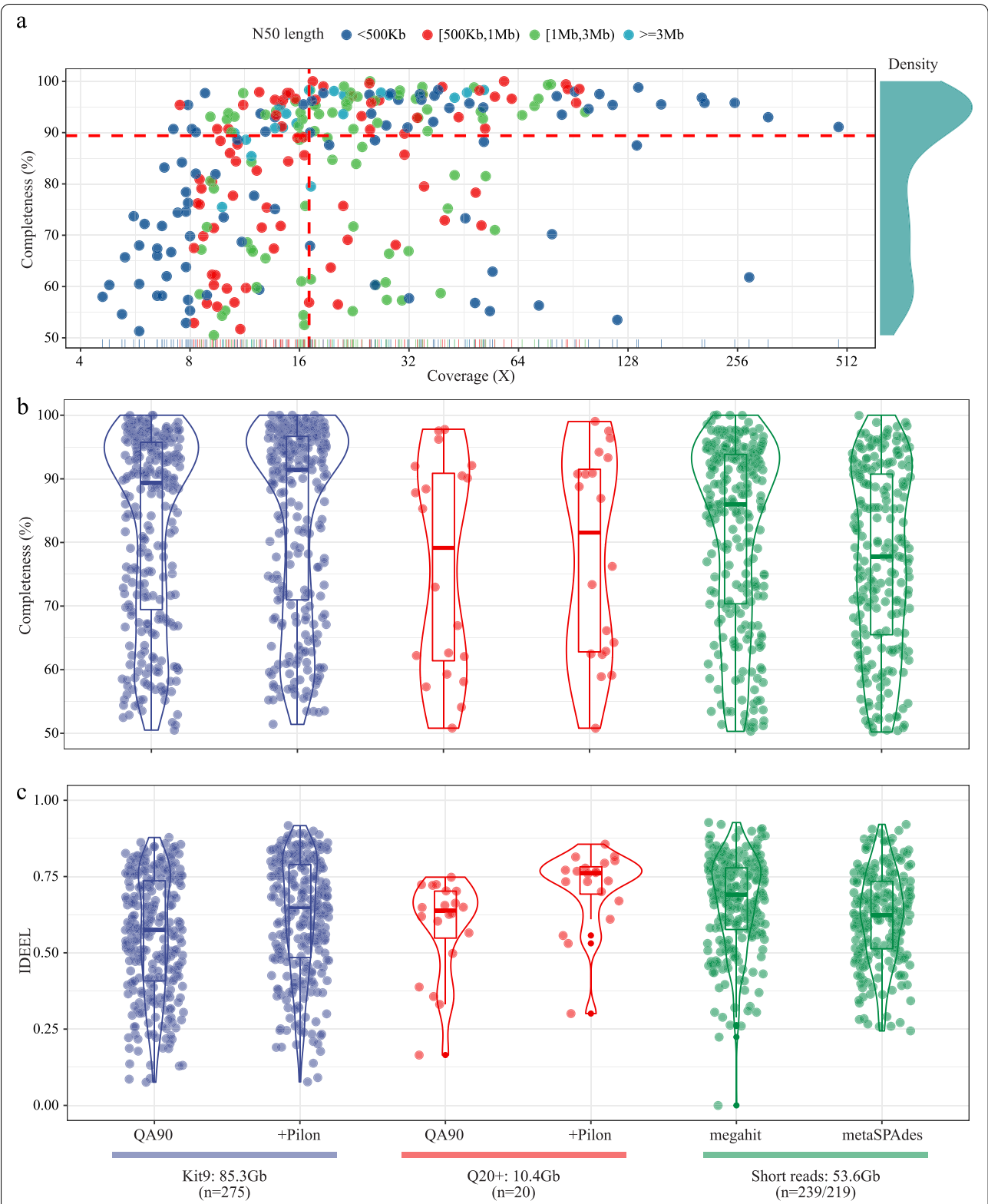


Fig. 2 Reference-quality genome reconstruction from a complex activated sludge metagenome. **a** Genome quality distribution of reconstructed MAGs and their corresponding coverage using the Kit 9 dataset. **b** Distribution of the completeness and **c** IDEEL scores in different genome reconstruction strategies. “QA90” means read accuracy was above 90%. “+Pilon” means MAGs were polished with one round of Pilon correction. “megahit” and “metaSPAdes” indicate MAGs were reconstructed using short-read-based methods. The reconstructed genome numbers were presented in the bracket

contributor to the diversity of bacterial gene repertoires, the relationship between bacteria and prophage is multifaceted, i.e., increasing bacterial fitness while at the risk of future lysis. In total, 165 prophages were identified within 111 MAGs with a median length of 14.3 Kb, and 1 MAG even possesses 5 different prophage sequences. The widespread prophages in recovered MAGs inferred that prophage was a critical factor in the evolution of microbial genomes via horizontal gene transfer between bacterial populations. Interestingly, no virulence factor and antibiotic resistance gene (ARG) were observed within recovered prophage sequences, suggesting that prophage mediated virulence factor and ARG transfer play a minor role in the studied activated sludge microbiome. In addition, the most prevalent accessory gene of prophage was methyltransferases, with a putative role in protecting prophages from the host immune system, which were found in 24 prophage sequences. Notably, most prophages remain dormant, transmitting vertically along with bacterial replication. However, 5 of them were determined as active prophages, indicating their host populations are undergoing prophage induction. Therefore, the active prophage lysis events may have altered the microbial community composition after sampling, suggesting the overlooked role of prophage in shaping the microbial community.

Conclusions

Reference-quality genome reconstruction from complex metagenomes benefited significantly from long reads, promoting critical insights into “complete metagenomics” [39]. We efficiently reconstructed highly accurate and reference-quality genomes from mock and complex AS metagenomes by NanoPhase. Furthermore, the CheckM-based completeness of genomes generated from Nanopore long reads was even higher than those MAGs from conventional short reads, indicating that the error rates (mainly the Indels rate) in Nanopore sequencing reads were low enough to barely affect the genome quality at the protein (amino acid) level.

In addition, the superiority of NanoPhase was also supported by the high IDEEL score, which would further expand its promising application in genome-centric studies, providing a near-complete genomic blueprint and benefiting a finely detailed overview of diverse ecosystems. Currently, NanoPhase also supports genome reconstruction from bacterial isolates and antibiotic resistance genes (ARGs) identification from reconstructed MAGs based on the structured ARG database [40]. However, it cannot distinguish different microbial strains from metagenomics datasets, which would be the next step in the field.

Although sufficient coverage ($\sim 30\times$) is vital for the reference-quality genome reconstruction and challenging for most microbial populations in many natural complex communities, we expected that the extraction of ultra-high molecular weight (UHMW) DNA and methylation profiles might further reduce the requirement of sequencing depth and resolve strain-level genome reconstruction. Furthermore, integrating the Nanopore adaptive sampling strategy [41, 42], ultra-low-cost and accurate reference-quality genomes reconstruction to represent a near-complete microbial community is accessible from complex metagenomes. Ultimately, these reference-quality genomes are valuable harnesses for investigating microbial population potential, interaction, and evolution in the studied microbiomes.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s40168-022-01415-8>.

Additional file 1: Table S1. Mock ID and strain taxonomic information. **Table S2.** Assembly and genome reconstruction quality statistics generated from the Kit 9 dataset of the activated sludge microbiome by different cutoffs of read-accuracy filtering. **Table S3.** Genome reconstruction comparison between Nanopore long reads (Kit 9) and Illumina short reads. **Figure S1.** Guppy predicted mean read quality scores with different versions versus read mapping accuracy as measured by alignment to the reference genome. Nanopore raw reads were basecalled by a, Guppy v3.0.3 (hac), b, Guppy v4.0.11 (hac), c, Guppy v5.0.16/6.0.0 (sup) and d, Guppy v6.0.0 with the Q20+ chemistry (sup) then 10K reads were subsampled for the comparison. The orange dashed line indicates the perfect correlation between the two quantities. **Figure S2.** The homopolymer identification improvement of Nanopore raw reads due to the basecaller and chemistry upgrades. Nanopore raw reads were basecalled by a, Guppy v3.0.3 (hac), b, Guppy v4.0.11 (hac), c, Guppy v5.0.16/6.0.0 (sup) and d, Guppy v6.0.0 with the Q20+ chemistry (sup), then one million reads were subsampled for computing homopolymer identification by comparing the basecalled reads to the reference genome using Counterr. The black dashed line indicates the true homopolymer length. **Figure S3.** The IDEEL distribution of reconstructed MAGs from the Kit 9 and Q20+ datasets. The dashed line represents the IDEEL score of reference genomes. Genomes generated from the Q20+ chemistry showed better IDEEL scores.

Acknowledgements

L.L. acknowledges the University of Hong Kong for the postdoctoral research fellowship. Y.Y. would like to thank The University of Hong Kong for the postgraduate scholarship. In addition, we would also like to thank the tremendous support from the Information Technology Services (ITS) at The University of Hong Kong for bioinformatics computing. Technical assistance from Ms. Vicky Fung and Ms. Lilian Y L Chan is greatly appreciated.

Code availability

NanoPhase source code is available on the GitHub page: <https://github.com/Hydro3639/NanoPhase>

Authors' contributions

T.Z. supervised the work. L.L. and T.Z. conceived the project and implemented the analysis pipeline. Y.Y. generated the nanopore sequence data for the ZymoBIOMICS gut microbiome standard and activated sludge samples. L.L., Y.Y., and Y.D. did the data analysis. L.L. and T.Z. wrote the manuscript. All authors read, discussed, and approved the final manuscript.

Funding

This work was substantially supported by the Theme-based Research Scheme funded by the University Grants Committee of Hong Kong, China (Project No. T21-705/20-N).

Availability of data and materials

The Nanopore long-read ZymoBIOMICS gut microbiome standard dataset has been deposited in the NCBI database with BioProject accession number PRJNA804004. In addition, the activated sludge Nanopore long reads are available at NCBI with BioProject accession number PRJNA803959. The reconstructed 275 MAGs using only Nanopore long reads have been deposited in Figshare, allowing bulk download under DOI 10.6084/m9.figshare.20654796.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 7 June 2022 Accepted: 7 November 2022

Published online: 02 December 2022

References

- Castelle CJ, Banfield JF. Major new microbial groups expand diversity and alter our understanding of the tree of life. *Cell*. 2018;172(6):1181–97.
- Hug LA, et al. A new view of the tree of life. *Nat Microbiol*. 2016;1(5):1–6.
- Tyson GW, et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*. 2004;428(6978):37–43.
- Liu L, et al. High-quality bacterial genomes of a partial-nitrification/anammox system by an iterative hybrid assembly method. *Microbiome*. 2020;8(1):1–17.
- Bertrand D, et al. Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nat Biotechnol*. 2019;37(8):937–44.
- Singleton CM, et al. Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. *Nat Commun*. 2021;12(1):1–13.
- Liu L, et al. Charting the complexity of the activated sludge microbiome through a hybrid sequencing strategy. *Microbiome*. 2021;9(1):1–15.
- Deng Yu, et al. Microbiome assembly for sulfonamide subsistence and the transfer of genetic determinants. *The ISME Journal* 15.10. 2021:2817–29.
- Jin H, et al. Hybrid, ultra-deep metagenomic sequencing enables genomic and functional characterization of low-abundance species in the human gut microbiome. *Gut Microbes*. 2022;14(1):2021790.
- Bickhart Derek M, et al. Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. *Nature biotechnology* 40.5. 2022:711–9.
- Yang Y, et al. Rapid absolute quantification of pathogens and ARGs by nanopore sequencing. *Sci Total Environ*. 2022;809:152190.
- Gorzynski JE, et al. Ultrarapid nanopore genome sequencing in a critical care setting. *New England Journal of Medicine* 386.7. 2022:700–2.
- Delahaye C, Nicolas J. Sequencing DNA with nanopores: troubles and biases. *PLoS One*. 2021;16(10):e0257521.
- Wang Y, et al. Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol*. 2021;39(11):1348–65.
- Sereika M, et al. Oxford Nanopore R10. 4 long-read sequencing enables near-perfect bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *bioRxiv*; 2021.
- Kolmogorov M, et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods*. 2020;17(11):1103–10.
- Kang DD, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*. 2019;7:e7359.
- Wu Y-W, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*. 2016;32(4):605–7.
- Uritskiy GV, DiRuggiero J, Taylor J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome*. 2018;6(1):158.
- Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34(18):3094–100.
- Vaser R, et al. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res*. 2017;27(5):737–46.
- Grabherr MG, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011;29(7):644–52.
- Walker BJ, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*. 2014;9(11):e112963.
- Ye J, McGinnis S, Madden TL. BLAST: improvements for better sequence analysis. *Nucleic Acids Res*. 2006;34(suppl_2):W6–9.
- Gurevich A, et al. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;29(8):1072–5.
- Parks DH, et al. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*. 2015;25(7):1043–55.
- Watson M, Warr A. Errors in long-read assemblies can critically affect protein prediction. *Nat Biotechnol*. 2019;37(2):124–6.
- Hyatt D, et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform*. 2010;11(1):1–11.
- Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. 2015;12(1):59–60.
- Wick RR, et al. Tricycler: consensus long-read assemblies for bacterial genomes. *Genome Biol*. 2021;22(1):1–17.
- Consortium, U. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*. 2019;47(D1):D506–15.
- Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30(14):2068–9.
- Chaumeil Pierre-Alain, et al. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. 2020:1925–7.
- Parks DH, et al. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat Biotechnol*. 2020;38(9):1079–86.
- Kieft K, Zhou Z, Anantharaman K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome*. 2020;8(1):1–23.
- Kieft K, Anantharaman K. Deciphering active prophages from metagenomes: bioRxiv; 2021.
- Stewart RD, et al. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nat Biotechnol*. 2019;37(8):953–61.
- Bowers RM, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol*. 2017;35(8):725–31.
- Marx V. Long road to long-read assembly. *Nat Methods*. 2021;18(2):125–9.
- Yin X, et al. ARGs-OAP v2.0 with an expanded SARG database and Hidden Markov Models for enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes. *Bioinformatics*. 2018;34(13):2263–70.
- Martin S, et al. Nanopore adaptive sampling: a tool for enrichment of low abundance species in metagenomic samples. *bioRxiv*; 2021.
- Kovaka S, et al. Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. *Nat Biotechnol*. 2021;39(4):431–41.

Publisher's Note

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