nature methods

BRIEF COMMUNICATION

https://doi.org/10.1038/s41592-022-01478-3

Metagenome assembly of high-fidelity long reads with hifiasm-meta

Xiaowen Feng[1,2], Haoyu Cheng[1,2], Daniel Portik[3] and Heng Li[1,2] ✉

De novo assembly of metagenome samples is a common approach to the study of microbial communities. Current metagenome assemblers developed for short sequence reads or noisy long reads were not optimized for accurate long reads. We thus developed hifiasm-meta, a metagenome assembler that exploits the high accuracy of recent data. Evaluated on seven empirical datasets, hifiasm-meta reconstructed tens to hundreds of complete circular bacterial genomes per dataset, consistently outperforming other metagenome assemblers.

A short-read metagenome assembly[1] often results in contigs of tens of kilobases (kb) in length[2], ~1% of a bacterial genome. After years of metagenome sequencing, there were only 62 complete genomes assembled from metagenome samples as of September 2019[3]. Although we can cluster short contigs into metagenome-assembled genomes (MAGs) with binning algorithms[4], binning can be an important source of errors, which complicate or mislead downstream analysis[3]. The limitation of short-read MAGs motivated the development of metaFlye[5], the only published assembler specialized for long-read metagenome assembly. Initially developed for noisy long reads of error rate ~10%, Flye[6], in which metaFlye is based, does not take advantage of PacBio's high-fidelity reads (HiFi) and is suboptimal for single-species HiFi assembly[7]. To leverage the full power of long accurate HiFi reads, we developed hifiasm-meta, extending our earlier work[8] to metagenome samples.

Comparison with the assembly of a single species, metagenome assembly poses several unique challenges[1,9], such as a larger variance in read length distribution in PacBio HiFi data, and high ploidy combined with low coverage in certain haplotypes. We made several major changes in hifiasm-meta to address these challenges. First, hifiasm-meta has an optional read selection step that reduces the coverage of highly abundant strains without losing reads on low abundant strains. Second, during the construction of the assembly graph, hifiasm-meta tries to protect reads in genomes of low coverage, which may be treated as chimeric reads and dropped by the original hifiasm. Third, hifiasm-meta only drops a contained read if other reads exactly overlapping with the read are inferred to come from the same haplotype. This reduces contig breakpoints caused by contained reads[10]. Fourth, after the initial graph construction, hifiasm-meta uses the coverage information to prune unitig overlaps, assuming unitigs from the same strain tend to have similar coverage. It also tries to join unitigs from different haplotypes to patch the remaining assembly gaps. These strategies make hifiasm-meta more robust to features in metagenome datasets.

We first evaluated hifiasm-meta r58-31876a0, metaFlye[5] v.2.9 and HiCanu[7] v.2.2 on two mock communities, ATCC and zymo (Table 1). ATCC consists of 20 distinct species, 15 of which are abundant, at 0.18–18%, and 5 of which are rare, at 0.02% abundance. We were able to reconstruct 13 of the abundant species, each as a complete circular contig, comparable to metaFlye and HiCanu (Supplementary Table 1). All tools assembled *Porphyromonas gingivalis*, at 18% abundance, into two contigs. No assemblers could fully reconstruct the five species of low abundance. We manually checked the read alignment of these species and found their assembly gaps are all caused by insufficient coverage. We would not be able to assemble these species in full with the current data. The zymo dataset features 21 strains of 17 species, including five strains of *Escherichia coli* at 8% abundance each. A challenge of this dataset lies in the phasing of the *E. coli* strains. Hifiasm-meta assembled strain B766 into a complete circular contig, strain B3008 into two contigs and the rest as fragmented contigs. HiCanu assembled both B766 and B3008 into complete circular contigs; metaFlye failed to assemble all five strains as circular contigs. Hifiasm-meta produces a more contiguous assembly for *Methanobrevibacter smithii* at 0.04% abundance (Supplementary Table 1). In general, all three assemblers have comparable accuracy on the two mock community datasets.

We then evaluated the three HiFi metagenome assemblers on real datasets (Table 1). Due to the lack of their true compositions, we used CheckM[11] to measure the completeness and the contamination level of each assembly. We define quality brackets in line with the minimum information requirement[12]. From the sheepA gut sample, hifiasm-meta reconstructed 323 contigs longer than 1 megabase (Mb) (Fig. 1a and Extended Data Fig. 1) totaling 651 Mb in length. Of these, 176 were near-complete according to CheckM (Fig. 1b). Most long contigs that failed to reach this category were due to incompleteness, not due to excessive contamination. Among the 176 near-complete hifiasm-meta contigs, 134 are circular (Fig. 1b), representing a noticeable improvement over HiCanu (71 circular near-complete contigs) and metaFlye (47). We aligned hifiasm-meta, HiCanu and metaFlye assemblies to each other and investigated the similarity between them. We found 86% and 94% of circular, near-complete HiCanu and metaFlye contigs, respectively, are also circular in the hifiasm-meta assembly and are of similar lengths (Supplementary Table 3). The remaining near-complete circular HiCanu and metaFlye contigs are assembled into either one linear contig or two linear contigs by hifiasm-meta. Hifiasm-meta can reconstruct most high-quality contigs found by other assemblers. Furthermore, the mash[13] sequence divergence between hifiasm circular contigs is mostly above 1%, except for four pairs of contigs at 0.62–0.92% divergence. In general, strains of high divergence (more than a couple of percentage points) can be separated into disconnected contigs; a few strains of low divergence tend to be collapsed and represented by mosaic contigs; many strains of mixed divergence may lead to complex assembly subgraphs and are the most challenging to assemble.

To reconstruct MAGs from non-circular contigs, we applied the MetaBAT2 binning algorithm[4] to each assembly. Not optimized for

[1]Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA. [2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [3]Pacific Biosciences, Menlo Park, CA, USA. ✉e-mail: hli@ds.dfci.harvard.edu

Table 1 | Evaluated metagenome datasets

Sample	Accession	Number of bases (Gb)	N50 read length (kb)	Median read QV	Sample description
ATCC	SRR11606871	59.2	12.0	36	Mock community ATCC MSA-1003 (ref. [17])
Zymo	SRR13128014	18.0	10.6	40	Mock community ZymoBIOMICS D6331
SheepA	SRR10963010	51.9	14.3	25	Sheep gut microbiome[5]
SheepB	SRR14289618	206.4	11.8	NA	Sheep gut microbiome[14]
Chicken	SRR15214153	33.6	17.6	30	Chicken gut microbiome
Sludge	ERR7015089	15.3	15.4	32	Anaerobic digester sludge
HumanO1	SRR15275213	18.5	11.4	40	Human gut from a pool of four omnivore samples
HumanO2	SRR15275212	15.5	10.3	41	Human gut from a pool of four omnivore samples
HumanV1	SRR15275211	18.8	11.0	39	Human gut from a pool of four vegan samples
HumanV2	SRR15275210	15.2	9.6	40	Human gut from a pool of four vegan samples

The N50 read length is the length of the shortest read at 50% of the total number of read bases. The quality value (QV) of a read is $-10\log_{10}e$, where e is the expected sequencing error rate of the read, assuming accurate base quality. No base quality is available for the sheepB dataset.

long-read assemblies, MetaBAT2 may mistakenly group different strains of the same species into one MAG and even group two circular contigs occasionally. Such MAGs would be considered to be contaminated by CheckM. To improve binning, we separated circular contigs into individual bins. In the end, we identified more than 110 non-circular MAGs of medium or higher quality from each sheepA assembly (Fig. 1b). Hifiasm-meta still finds more quality MAGs in total.

We applied hifiasm-meta to the larger sheepB dataset[14] (Table 1) and obtained 379 near-complete MAGs and 279 circular contigs. Bickhart et al.[14] assembled the combined sheepA and sheepB datasets with metaFlye and clustered contigs into MAGs using additional Hi-C data. They reported 44 circular contigs and 428 near-complete MAGs evaluated by DAS Tool. For a direct comparison, we ran CheckM on their assembly and identified 241 near-complete MAGs instead. Hifiasm-meta produced a more contiguous assembly with HiFi data alone.

For the chicken and the four human gut metagenomes (Table 1), hifiasm-meta consistently produced more circular contigs and more total MAGs than both HiCanu and metaFlye (Fig. 1b). Hifiasm-meta and metaFlye have comparable performance on the sludge dataset, both being better than HiCanu. All assemblers produced fewer MAGs in comparison to the sheepA gut sample. To see how much this is caused by the higher data volume of sheepA, we randomly sampled sheepB, which represents the same specimen, but was sequenced in SequelII and had similar read length distribution to that of humanO1, to ~18 gigabases (Gb) of sequences, comparable to the size of humanO1 and sludge. On the downsampled dataset, we could assemble 70 circular contigs, much more than the number of circular contigs in humanO1 and sludge. This suggests that data volume does affect the assembly quality, but that the more contiguous sheepA assembly is probably more related to the composition of the sample.

Among the four human gut datasets, two were collected from omnivore donors and the other two from vegan donors. Each dataset represents a pool of four individuals (Table 1). We further pooled the four datasets together and co-assembled them. With read names reported in the final hifiasm-meta assembly, we can identify the composition of each contig on the basis of the sources of reads. We found that the great majority of contigs of ≥1 Mb in size, and almost all ≥1 Mb circular contigs, are either omnivore specific or vegan specific (Fig. 1c), whereas the two omnivore samples are well mixed in long omnivore-specific contigs, as is the also the case with the two vegan samples. We see more reads coming from the humanO1 and humanV1 datasets, probably because humanO1 and humanV1 have more and longer reads than the other two human gut samples.

Omnivore and vegan samples are also well separated among co-assembled MAGs, although omnivore- and vegan-specific MAGs are mixed in the phylogenetic tree (Fig. 1d); in this tree, 20 genera consist of three or more MAGs, 17 of which contain both omnivore- and vegan-specific MAGs. This suggests hifiasm-meta assembly is better at untangling subtle composition differences. Also, notably, a clade of seven circular contigs (in the north-eastern direction in Fig. 1d) only has 75–79% CheckM completeness, but they all have 5S, 16S and 23S ribosomal RNA genes and ≥18 transfer RNA genes (Supplementary Table 4). Two of them were assembled by HiCanu into circular contigs of near-identical length, so they are less likely to be truncated by misassembly. We speculate that this clade may be under-represented in CheckM.

With regard to performance, hifiasm-meta took ~18 hours over 48 CPU threads to assemble the sheepA and the chicken datasets, and took ~3 hours for the human gut samples (Supplementary Table 3). On these datasets, it is as fast as metaFlye and is consistently faster than HiCanu by a few folds. Hifiasm-meta tends to use more memory than metaFlye and HiCanu, consuming ~200 Gb RAM for the sheepA and chicken gut samples. Hifiasm-meta assembled the largest sheepB dataset in 8.9 days and used 724 Gb RAM at the peak.

In the era of short-read sequencing, metagenome assembly was rarely considered as a method to reconstruct full genomes[3]. This view has been changed by recent progress in long-read assembly[5,14–16]. Optimized for long, accurate HiFi reads, hifiasm-meta moves metagenome assembly even further. It possibly assembles more circular MAGs from one deeply sequenced sample, without manual intervention, than all circular MAGs published previously. Such high-quality metagenome assemblies may fundamentally change the practice in metagenome analysis and shed light on the biological and biomedical implications of microbial communities.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of

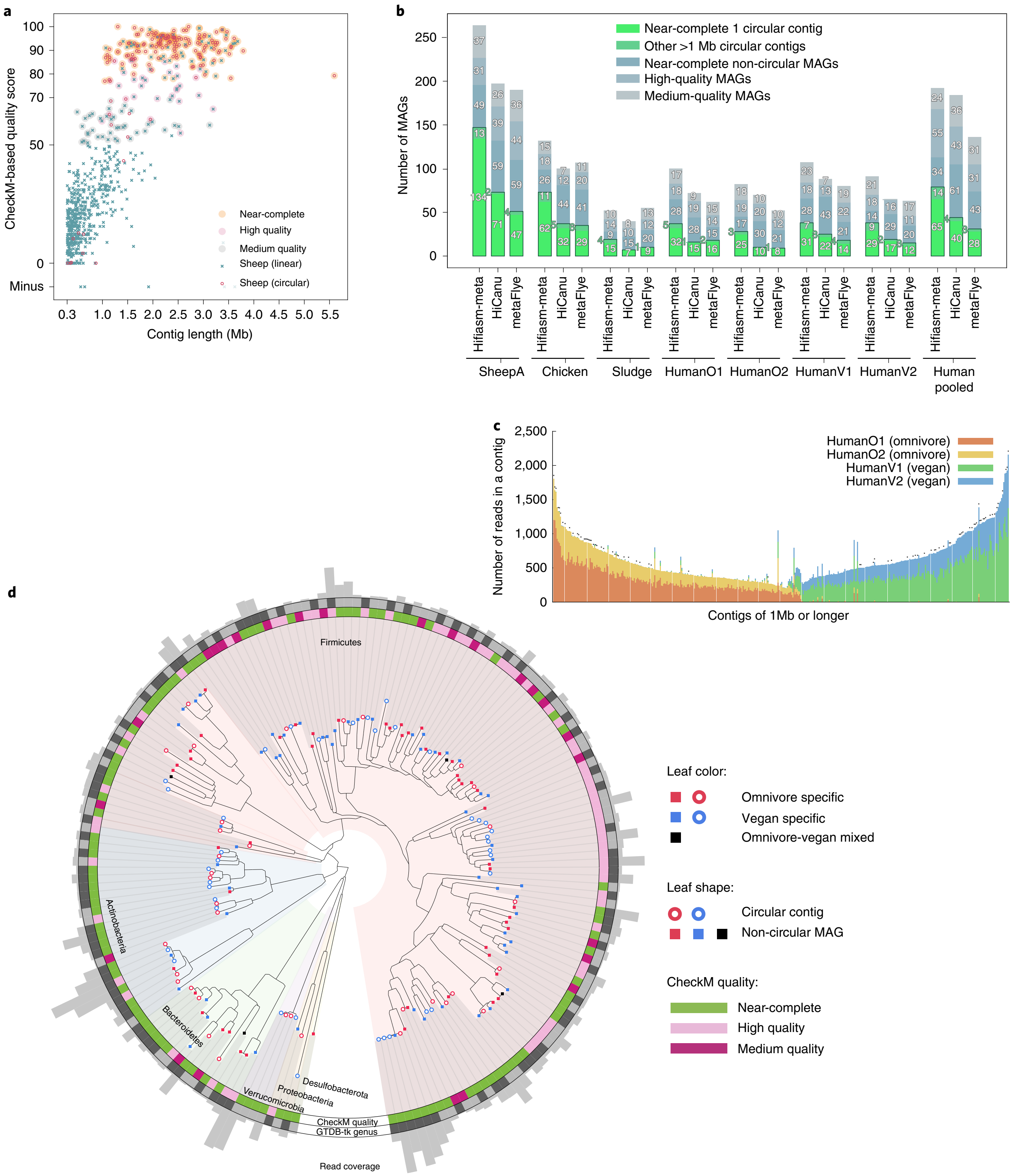


Fig. 1 | Metagenome assemblies of empirical datasets. a, Quality score of contigs longer than 300 kb from the hifiasm-meta assembly of sheepA. The quality score of a MAG is defined as 'completeness – 5 × contamination' on the basis of CheckM reports. **b**, CheckM evaluation. A MAG is 'near-complete' if its CheckM completeness is ≥90% and its contamination level ≤5%, is 'high quality' if completeness ≥70% and contamination ≤10% or is 'medium quality' if its quality score is ≥50%. 'HumanPooled' represents the co-assembly of all four human gut samples. **c**, Composition of long contigs in the hifiasm-meta co-assembly of four human gut samples. Each bar represents a contig of ≥1 Mb in length. It gives the number of reads used in the contig. Each color corresponds to a human gut sample. A cross at the top of a bar indicates the contig being circular. **d**, Phylogeny of human gut MAGs from the co-assembly. A colored clade corresponds to a phylum inferred by GTDB-Tk[18,19]. A MAG is omnivore/vegan specific if 90% of reads in the MAG come from omnivore/vegan samples, respectively. Two adjacent leaves coming in the same genus have the same shade on the 'genus' outer ring.

author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-022-01478-3.

Received: 15 October 2021; Accepted: 28 March 2022;
Published online: 09 May 2022

References

1. Lapidus, A. L. & Korobeynikov, A. I. Metagenomic data assembly—the way of decoding unknown microorganisms. *Front. Microbiol.* **12**, 613791 (2021).
2. Almeida, A. et al. A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).
3. Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M. & Banfield, J. F. Accurate and complete genomes from metagenomes. *Genome Res.* **30**, 315–333 (2020).
4. Kang, D. D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
5. Kolmogorov, M. et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110 (2020).
6. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
7. Nurk, S. et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305 (2020).
8. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
9. Cao, C. et al. Reconstruction of microbial haplotypes by integration of statistical and physical linkage in scaffolding. *Mol. Biol. Evol.* **38**, 2660–2672 (2021).
10. Hui, J., Shomorony, I., Ramchandran, K. & Courtade, T. A. Overlap-based genome assembly from variable-length reads. In *IEEE International Symposium on Information Theory, ISIT 2016* 1018–1022 (IEEE, 2016).
11. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
12. Bowers, R. M. et al. Minimum information about a single amplified genome (misag) and a metagenome-assembled genome (mimag) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
13. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using minhash. *Genome Biol.* **17**, 132 (2016).
14. Bickhart, D. M. et al. Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-021-01130-z (2022).
15. Moss, E. L., Maghini, D. G. & Bhatt, A. S. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat. Biotechnol.* **38**, 701–707 (2020).
16. Vicedomini, R., Quince, C., Darling, A. E. & Chikhi, R. Strainberry: automated strain separation in low-complexity metagenomes using long reads. *Nat. Commun.* **12**, 4485 (2021).
17. Hon, T. et al. Highly accurate long-read HiFi sequencing data for five complex genomes. *Sci. Data* **7**, 399 (2020).
18. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* **36**, 1925–1927 (2019).
19. Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C. & Segata, N. Compact graphical representation of phylogenetic data and metadata with graphlan. *PeerJ* **3**, e1029 (2015).

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

Methods

Overview of the hifiasm-meta algorithm. The hifiasm-meta workflow consists of optional read selection, sequencing error correction, read overlapping, string graph construction and graph cleaning. The error correction and read overlapping steps are largely identical to the original hifiasm. We added optional read selection and revamped the rest of the steps.

Optional downsampling of input reads. If read selection is enabled, hifiasm-meta will first make a crude guess of whether there are too many alignments to be performed for the whole read set. This is done by examining anchors and is alignment free. We proceed to do the selection if two-thirds of reads have more than 300 target reads. We start with an empty hash table, which will record *k*-mer counts, and go through reads in batches of 2,000. In a batch, for each read encountered, we collect its canonical *k*-mers and query the hash table for their occurrence. Three percentiles 3%, 5% and 10%, are checked against the corresponding thresholds of 10, 50 and 50, respectively. If any percentile is lower than the given threshold, the read is kept. The rationale is that we would like to keep a read when it has some rare *k*-mers, that is when discarding it will lead to loss of information. Note that the 'rare *k*-mers' here are not necessarily rare globally, and the read selection result might change if the inputs are shuffled. We assume that the input is not particularly sorted. After all reads in the batch have been processed, we update the *k*-mer counting hash table with them (*k*-mers of discarded reads are also counted). The termination criterion of the read selection is either that the total number of reads being kept has exceeded the desired count, or that all reads have been processed.

Modified chimera detection. Before graph construction, the original hifiasm regards a read to be chimeric and discards it if a middle part of the read is not covered by other reads. A read from a genome of low abundance may have such an uncovered region due to statistical fluctuation. Hifiasm-meta disables the heuristic if both ends of the read overlap with five or fewer other reads. This extra threshold improves the contiguity of genomes of low abundance.

Treatment of contained reads. The standard procedure to construct a string graph discards a read contained in a longer read. This may lead to an assembly gap if the contained read and the longer read actually reside on different haplotypes[10]. The original hifiasm patches such gaps by rescuing contained reads after graph construction. Hifiasm-meta tries to resolve the issue before graph construction instead. It retains a contained read if other reads exactly overlapping with the read are inferred to come from different haplotypes. In other words, hifiasm-meta only drops a contained read if there are no other similar haplotypes around it. This strategy often retains extra contained reads that are actually redundant. These extra reads usually lead to bubble-like subgraphs and are later removed by the bubble popping algorithm in the original hifiasm.

Changes to graph cleaning. At the graph construction stage, the original hifiasm-meta rejects overlaps between unitigs inferred to come from different haplotypes. Hifiasm-meta may do this to patch remaining assembly gaps. Hifiasm-meta also uses the unitig coverage to prune overlaps. Suppose unitig A overlaps unitig B and C in the same orientation. Such a bifurcation is an ambiguity in the assembly graph. Let $r_{AB} = \min\{\mathrm{cov}(A), \mathrm{cov}(B)\}$, where cov(A) is the coverage of A. r_{AC} is defined similarly. Hifiasm-meta drops the overlap between A and C if $r_{AB} > 0.7$ and $r_{AC} < 0.7$. This strategy is only applied to unitigs longer than 100 kb as it is difficult to accurately estimate coverage for short unitigs. In addition, attempting to resolve short unitigs would not greatly improve the assembly quality in our testing.

Assembly of metagenome datasets. We evaluated hifiasm-meta r58, HiCanu v.2.2 and metaFlye v.2.9 all with 48 CPU threads. We used 'hifiasm-meta reads.fa' for the assembly of empirical gut samples and used 'hifiasm-meta --force-rs reads.fa' to enable read selection for the two mock community datasets. We ran HiCanu with 'canu maxInputCoverage=1000 genomeSize=100m batMemory=200 -pacbio-hifi reads.fa'. We tried to increase the 'genomeSize' parameter to 1000m for sheepA and got identical results. We ran metaFlye with 'flye --pacbio-hifi reads.fa --plasmids --meta'. Hifiasm-meta and metaFlye report assembly time and peak memory usage. We used a script (https://gist.github.com/xfengnefx/d4abf19de8ebae9cc8ccd56e9898604d) to check /proc/ID/status files to measure the performance of HiCanu. For general file manipulations, we used seqtk (https://github.com/lh3/seqtk, 1.3-r107-dirty), readfq.py (https://github.com/lh3/readfq, 7c04ce7), GNU Parallel[20] and SAMtools[21].

Metagenome binning. We used MetaBAT2 for initial binning and then post-process MetaBAT2 results to get final MAGs. We aligned raw reads to an assembly with 'minimap2 -ak19 -w10 -I10G -g5k -r2k --lj-min-ratio 0.5 -A2 -B5 -O5,56 -E4,1 -z400,50 contigs.fa reads.fa' [22], calculated the depth with 'jgi_summa_rsize_bam_contig_depths --outputDepth depth.txt input.bam' and ran MetaBAT2 with 'metabat2 --seed 1 -i contigs.fa -a depth.txt'. We tried different random seeds or '-s 500000', and got similar results. We only applied MetaBAT2 to the primary hifiasm-meta and HiCanu assemblies, as including alternative assemblies led to worse binning. After MetaBAT2 binning, we separate circular contigs of 1 Mb or longer into a separate MAG if it is binned together with other contigs.

Evaluating assemblies of mock metagenome libraries. To evaluate the quality of assemblies, we mapped contigs with 'minimap2 -cxasm20' to the reference genomes, and inspected structural changes in the alignment. Out of 22 circular hifiasm-meta contigs assembled from the two mock communities, 21 are consistent with the reference, except for the assembly of *Streptococcus mutans*. For this genome, hifiasm-meta introduced a 20 kb deletion that is supported by a small fraction of reads in alignment, suggesting this is a real but rare allele in the community.

In the contig-to-reference alignment, we observed up to several thousand mismatches and gaps per genome (Supplementary Table 1). The number of small differences is much lower between HiFi assemblies. For example, for *Neisseria meningitidis*, there are 6,090 small differences between the hifiasm-meta contig and the reference genome, but there are only two small base-pair differences between the hifiasm-meta and HiCanu contigs. We suspect most of these 6,090 differences may be consensus errors in the reference genome.

Evaluating metagenome assemblies. We ran CheckM v.1.1.3 to measure the completeness and the contamination level of MAGs. The command line is 'checkm lineage_wf -x fa input/ wd/; checkm qa -o 2 wd/lineage.ms .'. We also tried DAS Tool[23] for evaluation on the sheepA dataset. DAS Tool is more optimistic, identifying 22% more near-complete MAGs in comparison to CheckM. As CheckM is more often used for evaluation, we only applied CheckM to all assemblies. For sheepB, additionally, yak QV was used to evaluate contig correctness (Extended Data Fig. 2).

We used GTDB-Tk v.1.3.0 and its database version r95 for phylogenetic placement with command line 'gtdbtk classify_wf'. We annotated the tree and used GraPhlAn for visualization.

We used INFERNAL[24] to identify rRNAs and tRNAs from contigs. The command line is 'cmscan –cut_ga –rfam –nohmmonly –fmt 2 –tblout cmscan.table Rfam.cm in.fa'. Entries with *E*-value larger than 0.01 were dropped. A total of 733 of the 738 long circular contigs assembled by hifiasm-meta in this manuscript were RNA complete (Supplementary Table 5), that is having at least one full-length copy for all three types of rRNAs, and at least 18 full-length copies of tRNAs.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

HiFi data were obtained from NCBI Sequence Read Archive (SRA) with accession numbers shown in Table 1. All generated assemblies and underlying data for the figures are available at https://zenodo.org/record/6330282. ZymoBIOMICS mock reference genomes were downloaded from https://s3.amazonaws.com/zymo-files/BioPool/D6331.refseq.zip. The list of reference genomes in the ATCC mock community is available at https://www.atcc.org/products/msa-1003. CheckM database: https://data.ace.uq.edu.au/public/CheckM_databases/checkm_data_2015_01_16.tar.gz. GTDB-Tk database: https://data.ace.uq.edu.au/public/gtdb/data/releases/release95/95.0/auxillary_files/. Source data are provided with this paper.

Code availability

Hifiasm-meta is available at https://github.com/xfengnefx/hifiasm-meta.

References

20. Tange, O. GNU Parallel - the command-line power tool. *The USENIX Magazine* **36**, 42–47 (2011).
21. Li, H. et al. The sequence alignment/map format and samtools. *Bioinformatics* **25**, 2078–2079 (2009).
22. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
23. Sieber, C. M. K. et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).
24. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).

Acknowledgements

We thank W. Fan from Agricultural Genomics Institute, Shenzhen, China, for sharing the chicken gut dataset. This study was supported by US National Institutes of Health (grant R01HG010040 and U01HG010971, to H.L.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

X.F. and H.L. conceived the project, designed the algorithm and wrote the manuscript. X.F. implemented the algorithm and evaluated the metagenome assemblies. H.C. helped with the algorithm implementation. D.P. helped with assembly evaluations. All authors helped with the data analysis and revised the manuscript.

Competing interests

D.P. is an employee of Pacific Biosciences. H.L. is a consultant of Integrated DNA Technologies and on the Scientific Advisory Boards of Sentieon and Innozeen. The remaining authors declare no competing interests.

Additional information

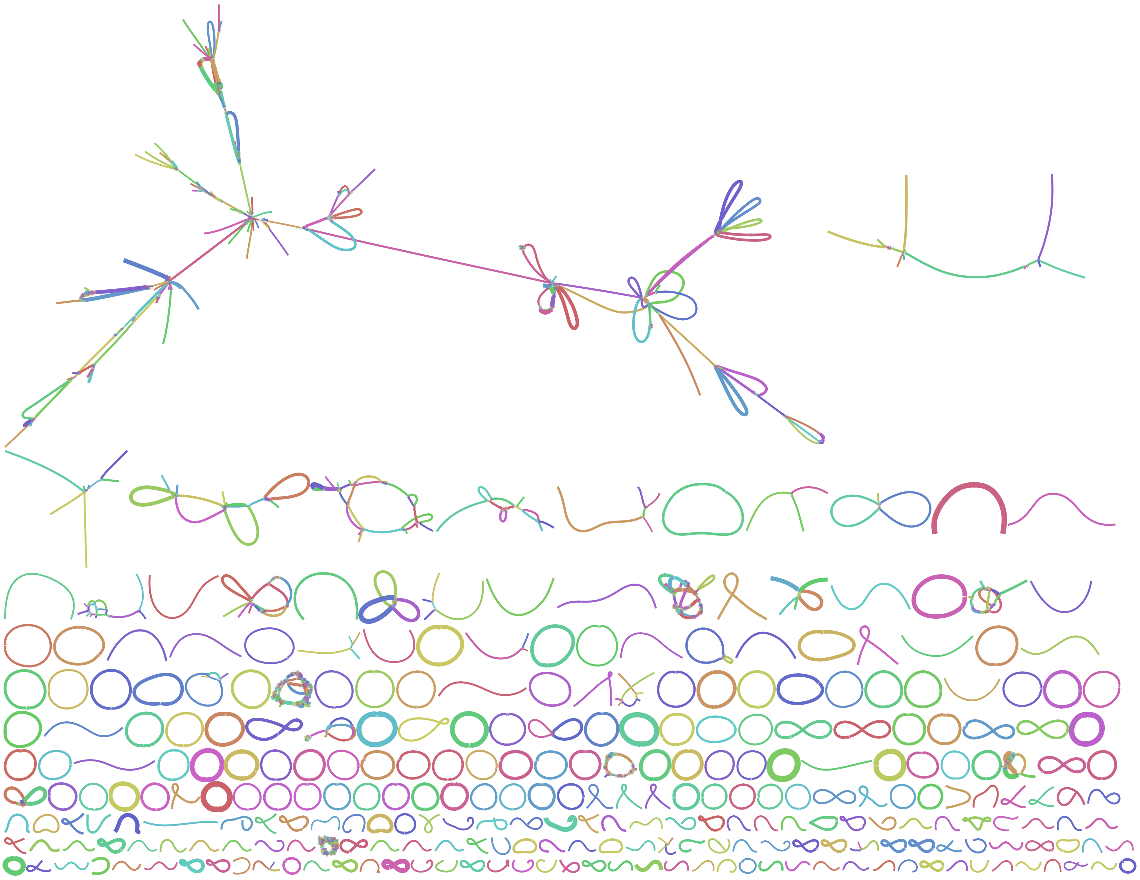
Extended data is available for this paper at https://doi.org/10.1038/s41592-022-01478-3.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-022-01478-3.

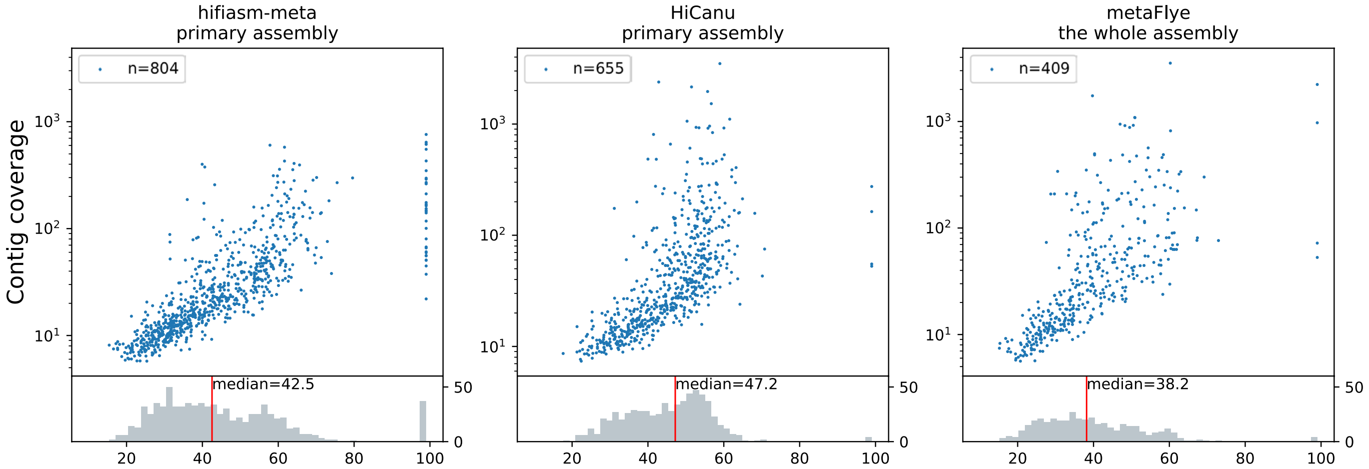
Correspondence and requests for materials should be addressed to Heng Li.

Peer review information *Nature Methods* thanks Mads Albertsen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Lei Tang was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 1 | The hifiasm-meta assembly graph of the sheepA dataset. Short disconnected contigs are not shown.



Extended Data Fig. 2 | Yak QV score correlated with contig coverage. Plots showing >1Mb contigs in sheepB assemblies. Contig coverage was estimated by jgi_summarize_bam_contig_depths from metabat2, and alignment was done with minimap2 -a -k 19 -w 10 -I 10G -g 5000 -r 2000 –lj-min-ratio 0.5 -A 2 -B 5 -O 5,56 -E 4,1 -z 400,50. Hifiasm-meta assembled 37 contigs without k-mer errors. HiCanu and metaFlye each assembled 4 contigs without k-mer errors.

nature research

Corresponding author(s): Heng Li

Last updated by author(s): Mar 10, 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☒	□	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	□	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	□	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	□	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	□	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software were used for data collection.
Data analysis	The manuscript described hifiasm-meta, a fork of the hifiasm assembler. Hifiasm-meta is at https://github.com/xfengnefx/hifiasm-meta and hifiasm is at https://github.com/chhylp123/hifiasm . Hifiasm-meta r58-31876a0 was used. We used the following softwares which are all openly available: - assembler: Flye(2.9-b1774, this is metaflye), canu(v2.2, this is hiCanu) - general file manipulation: seqtk (1.3-r107-dirty), readfq.py (https://github.com/lh3/readfq), GNU parallel (20210422 'Ever Given') - evaluation: minimap2 (2.18-r1028-dirty), mash (2.2), prodigal (v2.6.3), INFERNAL/cmsearch (1.1.4), checkM (v1.1.3), MetaBAT2 (version 2:2.15 Bioconda), viralVerify (https://github.com/ablab/viralVerify/ commit 5f811e), DAS Tool (1.1.3-0 bioconda), GTDB-Tk (v1.3.0), yak (https://github.com/lh3/yak commit 4bdd51d). - visualization: graphlan (commit c8b8a3), bandage (0.8.1). We used the following custom script to log the peak memory of hicanu: - pstrace.py: https://gist.github.com/xfengnefx/d4abf19de8ebae9cc8ccd56e9898604d

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Assemblies, their binning and evaluated completeness/contamination generated in this study are publicly available on Zenodo: https://zenodo.org/record/6330282
Sequencing data and their companion data (if any) used i n this study are all publicly available. The accession IDs:
- ATCC MSA-1003 mock community: SRR11606871; reference assemblies and other information from the manufacture can be found at https://www.atcc.org/products/msa-1003 . In the relevant evaluations, we substituted the references of Candida albican and Saccharomyces cerevisiae with RefSeq ASM18296v3 and ASM308665v1, respectively, in favor of less fragmentation of the references.
- ZymoBIOMICS D6331 mock community: SRR13128014; reference assemblies and other information from the manufacture can be found at https://files.zymoresearch.com/protocols/_d6331_zymobiomics_gut_microbiome_standard.pdf
- sheepA: SRR10963010
- sheepB: SRR14289618
- chicken: SRR15214153
- sludge: ERR7015089
- humanO1: SRR15275213
- humanO2: SRR15275212
- humanV1: SRR15275211
- humanV2: SRR15275210
- humanPooled is simply humanO1, humanO2, humanV1 and human V2 combined.

Companion datasets for evaluation tools:
- CheckM used https://data.ace.uq.edu.au/public/CheckM_databases/checkm_data_2015_01_16.tar.gz
- ViralVerify's HMMER reference was downloaded from https://figshare.com/s/f897d463b31a35ad7bf0 on 2020/12/14.
- INFERNAL used Rfam databse v14.6: http://ftp.ebi.ac.uk/pub/databases/Rfam/14.6/

Data restrictions:
- ATCC mock's manufacture requires an account to be created before downloading their reference assemblies. This is free of charge.
- ATCC mock's manufacture prohibits redistribution of their reference assemblies. Please refer to their data policy for details.

Figure data:
All figures have underlying data available.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Not applicable: no sample-size calculation was involved in this study. We did not do statistical tests or power analyses. For the number of datasets, all public PacBio Hifi WGS metagenome datasets with reasonable coverage available for analysis at the time of writing were included.
Data exclusions	No data were excluded from analysis.
Replication	Not applicable since this study involves no wet lab experiments and no opportunistic methods.
Randomization	Not applicable since this study have no experimental or control groups. Sequencing data were assembled by all assemblers.
Blinding	Not applicable since this study does not involve statistic analysis and data acquisition. Knowing which assembler produced which assembly has no influence on the evaluation, as the process is deterministic and carried out by well-defined pipelines (MetaBAT2, checkM, GTDB-tk, INFERNAL etc) that take no advantage of knowing the origin of the input data.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging