High-quality genome sequences of uncultured microbes by assembly of read clouds

Alex Bishara1,2,6, Eli L Moss2,6, Mikhail Kolmogorov3, Alma E Parada4, Ziming Weng5, Arend Sidow2,5, Anne E Dekas4, Serafim Batzoglou1 & Ami S Bhatt2

Although shotgun metagenomic sequencing of microbiome samples enables partial reconstruction of strain-level community structure, obtaining high-quality microbial genome drafts without isolation and culture remains difficult. Here, we present an application of read clouds, short-read sequences tagged with long-range information, to microbiome samples. We present Athena, a de novo assembler that uses read clouds to improve metagenomic assemblies. We applied this approach to sequence stool samples from two healthy individuals and compared it with existing short-read and synthetic long-read metagenomic sequencing techniques. Read-cloud metagenomic sequencing and Athena assembly produced the most comprehensive individual genome drafts with high contiguity (>200-kb N50, fewer than ten contigs), even for bacteria with relatively low (20×) raw short-read-sequence coverage. We also sequenced a complex marine-sediment sample and generated 24 intermediate-quality genome drafts (>70% complete, <10% contaminated), nine of which were complete (>90% complete, <5% contaminated). Our approach allows for culture-free generation of high-quality microbial genome drafts by using a single shotgun experiment.

Short-read sequencing and assembly have played an instrumental role in advancing the study of microbial genomes beyond the minority of organisms that have been isolated and cultured. These methods have greatly expanded understanding of the genomic structure and dynamics of complex microbial communities ranging from the human microbiome to environmental communities in the ocean, soil, and beyond. However, the precise gene coding potential and the consequent functional capabilities of organisms within these complex systems remain poorly understood.

Despite large-scale sequencing efforts of cultured isolates, analysis of sequences from diverse environmental samples has revealed that some major novel taxonomic lineages are entirely unrepresented in current reference collections, such as ReSeq. For example, even prevalent clades within heavily sequenced niches, such as Clostridiales and Bacteroides within the human gut, do not currently have a collection of isolate reference genomes that represent organisms observed in metagenomic shotgun sequencing. Thus, methods that accelerate the generation of high-quality genome drafts from shotgun sequencing of microbiome samples are needed.

Metagenomic shotgun sequencing, with the aid of specialized computational techniques, has also been used to generate draft genomes for individual taxa without the use of culture. The computational techniques developed involve dedicated metagenomic assemblers, and metagenome-draft binning based on sequence similarity and coverage-depth covariance. Binning techniques can group assembled sequences into more comprehensive drafts, but these techniques often fail to properly assign sequences that are shared among multiple bacterial strains. Furthermore, sequencing reads produced by existing high-throughput methods are too short to span many types of shared or duplicated sequences, and as a result, regions containing these types of sequences remain unassembled.

In principle, long-read-sequencing approaches can be used to address these issues. Long-read platforms such as Pacific Biosciences' Single Molecule Real Time sequencing approach have been successfully applied to close genomes of cultured isolates and dominant organisms within complex mixtures. However, these single-molecule platforms, in comparison to short reads, have lower throughput and a higher error rate. These single-molecule platforms also typically require higher input-DNA mass (~100 ng), preventing their application to biological samples containing insufficient high-molecular-weight DNA.

Synthetic long-read (SLR) approaches, such as Illumina TruSeq Synthetic Long reads, use long-fragment partitioning and short-read barcoding to obtain virtual long-read sequences, which can in theory be used to improve metagenomic assembly. Deep sequencing applied to a stool sample from a healthy human by using this SLR approach has allowed for the assembly of more contiguous genome sequences from a subset of constituent bacteria. However, SLR sequencing applied to more complex environmental samples, such as soil samples, has not yet resulted in improved genome assemblies. This lack of improvement is probably due to both the higher species richness of these samples and the limited overall throughput of the SLR approach.

1Department of Computer Science, Stanford University, Stanford, California, USA. 2Department of Medicine (Hematology, Blood and Marrow Transplantation) and Department of Genetics, Stanford University, Stanford, California, USA. 3Department of Computer Science and Engineering, University of California San Diego, La Jolla, California, USA. 4Department of Earth System Science, Stanford University, Stanford, California, USA. 5Department of Pathology, Stanford University School of Medicine, Stanford, California, USA. 6These authors contributed equally to this work. Correspondence should be addressed to A.S.B. (asbhatt@stanford.edu) or S.B. (serafim@cs.stanford.edu).
A recent method, introduced by 10x Genomics, streamlines the short-read barcoding process by using more than 1 million droplet partitions to yield uniquely barcoded short-read fragments from one or a few long molecules trapped in each droplet partition. Sequencing of libraries generated with this method yields shallow-coverage groups of barcode-sharing reads, which we will refer to as read clouds (also referred to as linked reads). Although both read-cloud and SLR approaches use long-fragment partitioning, read clouds trade off shallower short-read coverage of each individual long fragment for a larger total number of long fragments sequenced (Supplementary Note 1). This method and similar techniques predate human haplotype phasing, and also in resolving complex structural variations in human genomes. To date, their potential for de novo metagenomic sequence assembly has yet to be explored.

Here, we applied read clouds, generated by the 10x Genomics Chromium platform, to sequence human and marine microbiome samples. We also introduce Athena, an assembler that uses the barcode information from read clouds to produce high-quality genome drafts from a single shotgun sequencing experiment.

RESULTS Read-cloud sequencing and Athena assembly
We developed the Athena assembler to use long-range information encoded within barcoded short-read sequences. In our approach, we extract long DNA fragments and use the 10x Genomics Chromium platform to obtain barcoded short reads for our samples (Fig. 1a). The resulting short reads are first stripped of their barcodes and assembled with a standard short-read assembler (Online Methods) to obtain an initial assembly of the metagenome in the form of sequence contigs. These seed contigs are then provided to the Athena assembler for further metagenome sequence assembly (Fig. 1b). The same barcoded short reads are mapped back to the seed contigs, and read pairs that span contigs are used to form edges in a scaffold graph. Branches in this scaffold graph correspond to ambiguities encountered by the short-read assembler. At each edge, Athena examines the short-read mappings together with the attached barcodes to propose a simpler subassembly problem of a pooled subset of barcoded reads that can potentially assemble through branches in the scaffold graph (Supplementary Note 2). The selection of this read subset removes most reads considered during the initial assembly while retaining reads that cover the local target sequence, thus isolating the local subassembly problem from the broader metagenome. The much smaller independent subassemblies are performed separately for every edge in the scaffold graph to yield longer, overlapping subassembled contigs that resolve branches in the scaffold graph. The initial seed contigs and intermediary subassembled contigs are then passed as reads to the long-read De Bruijn graph–based assembler Fyfe, which determines how to assemble the target genome from these much longer contigs. The resulting metagenome assembly consists of more comprehensive sequence contigs, which resolve repeats that are too difficult to assemble with short-read techniques alone.

Assembly of a mock metagenome community
As a first validation of our approach, we applied Athena to assemble a read-cloud library of a staggered mixture of genomic DNA from 20 bacterial strains (ATCC MSA-1003, Online Methods). Groups of bacterial strains within the genomic-DNA mixture were present at staggered abundance levels as high as 18% and as low as 0.02% (Supplementary Table 1). The read-cloud library was prepared directly from genomic DNA supplied by the American Type Culture Collection (ATCC) and sequenced on one full lane of an Illumina HiSeq 4000 sequencer, which yielded approximately 74 Gb of raw short-read sequences.

We assembled the read-cloud library of the 20-strain mixture by using Athena and evaluated the overall draft quality against the available closed reference genomes. To compare against conventional short-read assembly, which does not leverage the read-cloud barcode information, we also assembled the raw barcode-stripped read-cloud sequencing data by using a standard short-read assembler (Online Methods). The assembled metagenome drafts of each approach were evaluated with MetaQUAST to assess contiguity, base-error rates, and misassemblies (Supplementary Table 2). The Athena-assembled drafts were significantly more contiguous than short-read-assembled drafts and had a median contig N50 increase of 7.6-fold for organisms with a minimum of 20× raw short-read coverage (0.18% reported DNA fraction; Supplementary Fig. 1). This contiguity was achieved without sacrificing overall accuracy, as compared against conventional short-read assembly. We found Athena assembly to be comparable to short-read assembly on the basis of two important metrics: the base error rate (8.97 versus 10.45 mismatches per 100 kb, respectively) and the total number of misassemblies (67 versus 61, respectively).

We then identified 16S/23S rRNA operons within drafts from both approaches and compared the placement of these repeats (5–7 kb in size) against the available closed reference genomes to ensure correct placement. Conventional short-read assembly was unable to correctly assemble and place a single rRNA operon. In contrast, Athena read-cloud assembly produced 41 copies of the complete rRNA operon across multiple species (Supplementary Table 1). All 41 assembled rRNA operons were correctly assigned to their respective genomes, and only three were determined to be misassembled (Supplementary Note 3).

Sequencing and assembly of the human intestinal microbiome
To test the generalizability of this approach to natural biological samples, we next applied read-cloud sequencing and Athena assembly to stool samples from two healthy human participants, P1 and P2. We used a Puregene DNA-extraction kit after enzymatic cell lysis to extract DNA from sample P1 and a QiaGen DNA-extraction kit after mechanical cell lysis to extract DNA from sample P2. To evaluate performance against alternative metagenomic-sequence-assembly approaches, we also prepared standard Illumina TruSeq short-read and Illumina TruSeq SLR sequencing libraries from extracted DNA. Read-cloud and SLR library preparations both require long DNA fragments, whereas TruSeq library preparation does not. Thus, the extracted DNA to be used in read-cloud and SLR libraries was first subjected to size-selection (Online Methods, Supplementary Table 3). For each stool sample, prepared short-read TruSeq and read-cloud libraries were multiplexed and sequenced with an Illumina HiSeq 4000 sequencer, yielding approximately 40 Gb of raw short-read sequences per library. SLR libraries cannot be multiplexed, so each of the two SLR libraries was given its own full lane for sequencing on a HiSeq 4000 instrument, yielding approximately 102 Gb of raw short-read sequences for each library (Supplementary Table 4).

Genus-level community compositions for each of the three sequencing approaches were first assessed with k-mer-based short-read classifications (Fig. 2a,b). Although some less abundant genera differed in their abundance rank, the community composition was largely concordant among all approaches tested (Supplementary Note 4).
To compare the performance of the three sequencing approaches, we applied the appropriate assembly approach to each sequenced library to obtain initial metagenomic drafts. Short-read, read-cloud, and SLR libraries were assembled with a conventional short-read assembler, Athena, and a two-stage assembly process (Online Methods). Despite high amounts of raw short-read sequence for the SLR libraries (~102 Gb per sample for both P1 and P2), the amount of total sequence in the form of virtual long reads was low (0.64 Gb for P1 and 0.55 Gb for P2; Supplementary Table 4). Read-cloud sequencing and assembly resulted in much longer microbial sequence contigs than did both SLR and short-read sequencing and assembly. Nearly 144 Mb of sequence from P1 and 40 Mb of sequence from P2 were assembled with read clouds into contigs with a minimum size of 100 kb, compared with just 68 Mb and 22 Mb with short reads, and 26 Mb and 14 Mb with SLRs (Supplementary Fig. 2). The overall size of the read-cloud metagenome drafts was also larger than that of the SLR metagenome drafts (345 Mb versus 55 Mb in P1 and 229 Mb versus 31 Mb in P2), thus highlighting the benefit of the increased throughput of our approach that allows for assembly of lower-abundance organisms.

Read clouds produce high-quality genomes for individual bacterial species

To assess the ability of each approach to produce genome drafts for constituent bacteria, we binned metagenome draft contigs and used annotations of contigs to obtain genus-level and/or species-level assignments for each resulting bin (Online Methods, Supplementary Fig. 3 and Supplementary Tables 5 and 6). The resulting bins were evaluated as genome drafts by using the presence of lineage-specific single-copy core genes to determine draft completeness and contamination. Using previously described criteria, we refer to a genome bin as a ‘complete’ genome draft if it is >90% complete and <5% contaminated as assessed by checkM. We refer to the subset of complete genome drafts as ‘high quality’, adopting a previously defined standard, if the draft also contains at least 18 tRNA loci and at least one copy each of 5S, 16S, and 23S. We also designate less comprehensive genome bins that are >70% complete and <10% contaminated as ‘intermediate-quality’ genome drafts.

Read-cloud sequencing yielded complete and high-quality genome drafts for bacteria from both samples P1 and P2 (Fig. 2c,d and Supplementary Note 5). Our most contiguous, high-quality
read-cloud draft was for *Bacteroides uniformis* in sample P1, which was contained completely in three contigs of sizes 4.7 Mb, 369 kb, and 25 kb. Several other bacteria from P1 were also well assembled, including *Bifidobacterium longum*, *Escherichia coli*, and *Bacteroides fragilis*. Alignments of input short reads to the assembled genome drafts from each sequenced library of samples P1 and P2 allowed for estimation of the short-read coverage of individual organisms within these libraries (Supplementary Table 6). Read-cloud and short-read libraries showed overall concordance with each other and also discordance with the SLR libraries, in terms of the raw short-read coverage of individual taxa in both samples. All three approaches yielded fewer complete and high-quality genome drafts from sample P2 than from sample P1. Examination of the per-taxon coverage in sample P2 libraries revealed that this sample was largely dominated by a small number of highly abundant taxa, and as a result, libraries of sample P2 contained far fewer well-covered taxa than did libraries of sample P1.

Although read-cloud assembly and binning yielded a single high-quality genome draft that was annotated as *Prevotella copri* in sample P2, the N50 of 103 kb for this read-cloud draft was unexpectedly low, given its 2,836× short-read coverage. Analysis of short reads originating from this genome bin in the read-cloud library illuminated the unusual presence of five high-copy (more than ten copies) genomic elements that probably impeded improvements in assembly by our approach (Supplementary Note 6).

The read-cloud approach was superior to both the short-read and SLR approaches in its ability to generate genome drafts for individual bacterial species (Fig. 3 and Supplementary Fig. 4). The combined results from read-cloud sequencing of samples P1 and P2 yielded a total of 51 intermediate-quality drafts, of which 27 were complete. The short-read approach yielded fewer, with 43 intermediate-quality drafts, of which only 18 were complete. SLR sequencing produced a total of only two intermediate-quality drafts, of which one was complete, despite receiving twice the amount of raw short-read sequencing for each sample (owing to the inability to multiplex SLR libraries). Read clouds produced the most comprehensive drafts, which were also highly contiguous (N50 > 200 kb), with a total of 16, compared with just one each from the short-read and SLR approaches. Read clouds were able to produce complete genome drafts, a large fraction of which were also highly contiguous, with as little as 20× short-read coverage for some bacteria (Fig. 3b,c). The short-read approach also produced multiple complete drafts at low coverage. However, the resulting drafts from short reads were fragmentary compared with the read-cloud drafts, even for bacteria with high short-read coverage. Of all three tested approaches, the read-cloud approach was the only one capable of producing high-quality drafts (Fig. 3d–f).

We next assessed differences among the three approaches in their ability to produce complete drafts for particular taxa (Fig. 4). Read clouds produced by far the most comprehensive and high-quality genome drafts, in which all contigs were clustered into a single bin. In contrast, short-read genomes were most frequently split across two or more bins. For most taxa discovered in samples P1 and P2, read clouds also successfully assembled and binned more genes than either short reads or SLRs.

To assess whether the performance gains of read clouds over short reads would be retained if overall sequencing depth were reduced, we
also evaluated the performance on in silico downsampled data sets of the sequenced mock community sample and a human stool sample. Comparisons of assembly results between the full sequenced data sets and downsampled data sets (8 Gb overall sequencing) revealed that the read-cloud performance gains over short reads were depth dependent and that these gains diminished with lower overall sequencing depth (Supplementary Note 7).

Alignments of read-cloud genomes against closed reference genomes

Comparisons of our high-quality drafts against available closed reference genomes showed both cases in which genome structure was largely maintained and cases in which large structural rearrangements were apparent (Fig. 5). Both Dialister invisus and Eubacterium eligens were present and assembled into high-quality genome drafts in both samples P1 and P2. Alignments of both D. invisus drafts from samples P1 and P2 illustrated large-scale rearrangement with respect to the available reference genome. Inspection of these reference alignments indicated that the D. invisus strains generated by the read clouds in each sample were largely structurally divergent from each other as well. Interestingly, the draft recovered for E. eligens from sample P2 was structurally similar to the reference genome, whereas the draft recovered from sample P1 displayed two large-scale inversions. Despite structural concordance in most our assembled drafts to the available reference genomes, all of them deviated substantially from the available references in sequence identity for alignable bases as well as the total number of bases that were unalignable (Supplementary Table 7). The median nucleotide sequence identity was 98.5%, and the median fraction of reference-unaligned bases in each draft was 15.7%.

For the organisms assembled into high-quality drafts by using read clouds, alignments of the corresponding SLR and short-read drafts illustrate the fragmentary nature of the drafts recovered by these two approaches. Organisms that were not present at sufficiently high abundance within each of the samples received only sparse virtual long-read coverage in the SLR libraries, such that further sequence assembly of these virtual long reads into sequence contigs was generally not possible. Although the short-read approach did not have the same throughput limitation, it was nonetheless capable of producing only fragmentary genome drafts. The read-cloud approach was the
only one capable of producing high-quality and highly contiguous genome drafts \textit{de novo} from the studied human stool samples.

**Assembly of a marine-sediment microbial community**

To test the ability of read clouds to generate genome drafts from samples that are generally regarded as more complex than human stool microbiomes, we applied read-cloud sequencing and Athena to deep-sea marine sediment obtained approximately 115 km off the coast near San Francisco, California, USA. DNA was extracted from this sample through a combination of mechanical bead-beating and chemical lysis, and subjected to a size-selection to enrich for long DNA fragments (Online Methods). A read-cloud library was prepared and sequenced on one full lane and a quarter lane on an Illumina HiSeq 4000 flow cell, yielding approximately 72 Gb of raw short-read sequences (Supplementary Table 4). To successfully assemble this sample, which was substantially more complex than our human stool samples, we applied a specialized short-read assembler designed for use with large and complex metagenomes (Online Methods). Modifications were also made to Athena to successfully assemble the sequencing data by using the read-cloud barcode information (Online Methods).
Figure 5 Comparisons of representative read-cloud genome drafts to reference genomes, and corresponding short-read and SLR drafts. Dot-plot alignments between read-cloud drafts (y axis) and the closest available reference genome (x axis) are shown. For each dot plot, a given color corresponds to the alignment of a single contig in the read-cloud draft against the available reference. Large-scale structural concordance and differences including inversions are visually apparent. Alignments of SLR and short-read drafts to the read-cloud drafts for each taxon are also shown. In all cases, read-cloud drafts were the most contiguous. For each approach, contigs belonging to the largest genome bin for a particular taxon are represented with a darker color, and the rest of the contigs in other bins are represented with a lighter color.
Figure 6 Comparison of marine-sediment genome drafts generated by read-cloud sequencing with standard short-read versus Athena assembly. Athena read-cloud assembly (gold) consistently produced more genome drafts than standard short-read assembly (blue), with genome bins assessed as genome drafts under various quality criteria. Athena read-cloud assembly allowed for significantly more 16S rRNA (16S) taxonomic sequences to be assigned to genome drafts than short-read assembly. (a–c) The number of intermediate-quality (>70% completeness and <10% contamination) genome drafts (a) intermediate-quality genome drafts with assembled 16S rRNA sequences (b), and high-quality genome drafts with assembled 16S rRNA sequences (c) with a minimum short-read coverage depth are shown.

The short-read-assembled metagenome was 5.3 Gb, as compared with just 574 Mb from the combined metagenomes of the human stool samples, suggesting a much higher species richness in our marine-sediment sample (Supplementary Note 8). Athena read-cloud assembly produced more large sequence contigs (351 Mb versus 135 Mb in contigs >10 kb; Supplementary Fig. 5) and 16S rRNA sequences (130 versus 23) than did short-read assembly alone.

We next assessed the ability of each assembly approach to produce genome drafts from the marine microbiome (Online Methods and Supplementary Table 8). Read-cloud sequencing and Athena assembly consistently produced more genome drafts than short-read assembly alone (Fig. 6). Athena assembly produced nine complete genome drafts, of which eight were also high quality. Short-read assembly was unable to produce a single complete or high-quality draft. Athena produced 49 intermediate-quality genome drafts, of which 24 also contained assembled 16S RNA sequences. Short-read assembly produced 28 intermediate-quality genome drafts, of which only four contained 16S rRNA sequences. Alignments of input short reads to the assembled genome drafts from the read-cloud library of the marine-sediment sample allowed for estimation of short-read coverage of individual organisms within this sample (Supplementary Table 9). Higher-quality drafts tended to be better covered within our sequenced sample: high-quality genome bins and intermediate-quality genome bins had median coverage of 27× and 13× respectively.

DISCUSSION

We present a novel approach using read clouds to generate de novo genome drafts from microbiome samples with the use of a single shotgun sequencing experiment. Application of our approach across diverse samples should provide high-quality genome drafts across the microbial tree of life, thereby increasing the comprehensiveness of reference collections without a need for laborious isolation and culture. Our work represents a step toward enabling fine-grained comparative genomics for microorganisms within complex communities.

We anticipate that our read-cloud sequencing approach will benefit from future improvements in both DNA-extraction techniques and long-fragment barcoding approaches. Our approach currently requires a relatively high input-DNA mass, because the application of size-selection after existing mechanical lysis techniques results in substantial loss. Improvements to DNA extraction that better preserve high-molecular-weight DNA across all constituent bacteria should enhance the usability of this and other approaches. Although our approach produced highly contiguous drafts for many taxa present in our human microbiome samples, the genome draft for a highly abundant Prevotella copri strain was notably fragmented. We found that this strain contained several high-copy genomic repeat elements that probably complicated the correct resolution of local genomic structure during subassembly in Athena. The current 10x Genomics Chromium method groups several (approximately ten) long fragments per barcode. Improvements that allow for only a single long fragment per partition would greatly decrease the complexity of each subassembly task within Athena and potentially allow read clouds to better assemble organisms with these high-copy repeats.

Further development of binning methods that take advantage of the read-cloud barcode information should allow for recovery of even more individual microbial genome drafts from the communities presented. Our current approach to produce individual genome drafts leveraged both our Athena assembler to improve metagenomic contig assembly and existing binning tools that were designed for use with conventional short-read-assembly techniques. These binning tools cluster contigs into groups with similar nucleotide composition (for example, tetramer frequencies) and coverage depth. Although application of these tools worked well for our improved metagenome-draft contigs, they were unable to properly deconvolve several members of some genera in our stool microbiome samples, such as Bacteroides and Fecalibacterium, and probably members of many less characterized genera within our marine-sediment samples. Multiple species belonging to each of these genera are probably present at similar abundance and have similar nucleotide compositions, such that the current metrics do not allow contigs from these taxa to be correctly separated into individual draft genomes. Read clouds have the potential to solve this issue. Pairs of sequences sharing many barcodes are indicative of sequences originating from the same input-DNA fragments, which should then be binned together. Binning approaches that aim to incorporate this linkage information will probably provide a stronger signal that can further disentangle closely related taxa within complex metagenomic samples.
Among the methods evaluated, our read-cloud approach was the only one capable of generating complete and high-quality genome drafts for the marine-sediment sample. Read clouds also generated more intermediate-quality genome drafts, nearly half of which included the 16S rRNA gene. The added ability to link genomic sequences with 16S rRNA sequence provides an opportunity to improve functional characterization of the vast number of environmental samples for which taxonomic composition (i.e., 16S rRNA data sets) but not functional characterization (i.e., metagenomic data) is readily available. Extensions of binning approaches to use the linkage information present in read clouds will probably allow for the generation of far more comprehensive bins from these complex samples. Further applications of our read-cloud approach to diverse environmental samples, especially those in which isolation and culture have been limited, should help illuminate the vast microbial life that is currently unknown.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The authors thank E. Tkachenko for assistance in preparing TruSeq libraries and M. Snyder and members of the laboratory of A.S.B. for helpful feedback. The authors also thank H. Xu at Illumina for sharing read-cloud sequencing data of ATCC 20 for the mock metagenome. This work was supported by NCIC90148420, the Amy Stelzer Manasevit Award from the National Marrow Donor Program, and a Damon Runyon Clinical Investigator Award to A.S.B. E.L.M. was supported by National Science Foundation Graduate Research Fellowship DGE-114747. A.B. was supported by the Stanford Genome Training Program (SGTP; NIH/NHGRI) and a Training Grant of the Joint Initiative for Metrology in Biology (JMB; NIST). A.E.D. and the marine-sample collection and extraction were supported by National Science Foundation grant OCE-1634297. A.E.P. was supported by a Center for Dark Energy Biosphere Investigations Postdoctoral Fellowship. Access to shared computer resources was supported in part by NIH P30 CA124453 via the Stanford Cancer Institute Shared Resource Genomics Bioinformatics Service Center.

AUTHOR CONTRIBUTIONS

A.B., E.L.M., A.S.B. and S.B. conceived the study. Z.W. performed read-cloud library generation from E. coli. E.L.M. extracted DNA and prepared SLR sequencing libraries. E.L.M. performed PCR validation and Sanger sequencing. A.B. and S.B. conceived the assembly approach. A.B. implemented the Athena assembler. M.K. modified the Flye assembler for use with Athena. A.E.P. and A.E.D. collected the marine-sediment sample, extracted DNA from the marine-sediment sample and assisted in the analysis of these samples. A.B., A.S.B. and E.L.M. carried out all analyses, wrote the manuscript, and generated figures. All authors commented on the manuscript.

COMPEting INTERESTS

S.B. is an employee of and owns stock in Illumina. Shotgun sequencing products developed, marketed and/or sold by Illumina were used in this work.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Schloss, P.D. & Handelsman, J. Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. Genome Biol. 6, 229 (2005).
2. Turnbaugh, P.J. et al. An obesity-associated gut microbiota with increased capacity for energy harvest. Nature 444, 1027–1031 (2006).
3. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature 486, 207–214 (2012).
4. Lloyd-Price, J. et al. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature 550, 61–66 (2017).
5. Kashlan, N. et al. Single-cell genomics reveals hundreds of coexisting subpopulations in mixed Parachlorella. Science 344, 416–420 (2014).
6. Baker, B.J., Lazar, C.S., Teske, A.P. & Dick, G.J. Genomic resolution of linkages in carbon, nitrogen, and sulfur cycling among widespread estuary sediment bacteria. Microbiology 3, 14 (2015).
7. Eyice, O. et al. SIP metagenomics identifies uncultivated Methylphosphatase as dimethylsulphide degrading bacteria in soil and lake sediment. ISME J. 9, 2336–2348 (2015).
8. He, Y. et al. Genomic and enzymatic evidence for acetogenesis among multiple lineages of the archaeal phylum Bathycyanaeota widespread in marine sediments. Nat. Microbiol. 1, 16035 (2016).
9. Brown, C.T. et al. Unusual biology across a group comprising more than 15% of domain bacteria. Nature 523, 208–211 (2015).
10. Rieger, L.A. et al. A new view of the tree of life. Nat. Microbiol. 1, 16048 (2016).
11. DeLeary, N.A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44, D733–D745 (2016).
12. Peing, Y., Leung, H.C.M., Yiu, S.M. & Chin, F.Y.L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28, 1420–1428 (2012).
13. Namiki, T., Hachiya, T., Tanaka, H. & Sakakibara, Y. MetaVelvet: an extension of Velvet assembler de novo assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res. 40, e158 (2012).
14. Cleary, B. et al. Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. Nat. Biotechnol. 33, 1053–1060 (2015).
15. Wu, Y.-W., Tang, Y.-H., Tringe, S.G., Simmons, B.A. & Singer, S.W. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiology 2, 26 (2014).
16. Kang, D.D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 3, e1165 (2015).
17. Nielsen, H.B. et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat. Biotechnol. 32, 822–828 (2014).
18. Alineberg, J. et al. Binning metagenomic contigs by coverage and composition. Nat. Methods 11, 1144–1146 (2014).
19. Popic, V., Kuleshov, V., Snyder, M. & Batzoglou, S. GATTACA: lightweight metagenomic binning with compact indexing of kmer counts and minhash-based panel selection. Preprint at https://www.biorxiv.org/content/early/2017/04/26/130997 (2017).
20. Koren, S. et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat. Biotechnol. 30, 693–700 (2012).
21. Chin, C.-S. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods 10, 563–569 (2013).
22. Loman, N.J., Quick, J. & Simpson, J.T. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat. Methods 12, 733–735 (2015).
23. Leonard, M.T. et al. The methylome of the gut microbiome: disparate Dam methylation patterns in intestinal Bacteroides dorei. Front. Microbiol. 5, 361 (2014).
24. Voskoboynik, A. et al. The genome sequence of the colonial chordate, Botryllus schlosseri. elife 2, e00569 (2013).
25. Kuleshov, V. et al. Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. Nat. Biotechnol. 34, 64–69 (2016).
26. Sharon, I. et al. Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. Genome Res. 25, 534–543 (2015).
27. White, R.A. III et al. Molecule long-read sequencing facilitates assembly and genomic binning from complex soil metagenomes. mSystems 1, e00045–16 (2016).
28. Zheng, G.Y.X. et al. Haplotyping genome and cancer genomes with high-throughput linked-read sequencing. Nat. Biotechnol. 34, 303–311 (2016).
29. Bishara, A. et al. Read clouds uncover variation in complex regions of the human genome. Genome Res. 25, 1570–1580 (2015).
30. Peters, B.A. et al. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. Nature 487, 190–195 (2012).
31. Kitzman, J.O. et al. Haplotyping-resolved genome sequencing of a Gujarati Indian individual. Nat. Biotechnol. 29, 59–63 (2011).
32. Amini, S. et al. Haplotype-resolved whole-genome sequencing by contig-preserving transposition and combinatorial indexing. Nat. Genet. 46, 1343–1349 (2014).
33. Spies, N. et al. Genome-wide reconstruction of complex structural variants using read clouds. Nat. Methods 14, 915–920 (2017).
34. Lin, Y. et al. Assembly of long error-prone reads using de Bruijn graphs. Proc. Natl. Acad. Sci. USA 113, E8396–E8405 (2016).
35. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. Assembly of long error-prone reads using repeat graphs. Preprint at https://www.biorxiv.org/content/early/2018/01/12/ 247149 (2018).
36. Mikheenko, A., Saveliev, V. & Gurevich, A. MetaQUAST: evaluation of metagenome assemblies. Bioinformatics 32, 1088–1090 (2016).
37. 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).
38. 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).

NATURE BIOTECHNOLOGY VOLUME 36 NUMBER 11 NOVEMBER 2018 1075

ARTICLES

Supplementary Note 2

DNA preparation. For the ATCC 20 mock metagenome sample, DNA from ATCC 20 Strain Staggered Mix Genomic Material was used directly without size-selection for the mock metagenome. A single read-cloud library was prepared for 10x Genomics Chromium sequencing according to the manufacturer's standard protocol.

For the healthy-volunteer stool samples, DNA was extracted from the stool of P1 with a Qiagen Genta Puregene Yeast/Bacteria kit according to the manufacturer's standard protocol with two modifications: a chilling step at −80 °C for 5 min before DNA precipitation, and DNA precipitation with a 14,000g, 20-min centrifugation at 4 °C. DNA was extracted from the stool of P2 with a Qiagen QIAmp Stool Mini Kit according to the manufacturer's standard protocol, which was modified with an additional step after addition of buffer ASL. The additional step comprised seven cycles of alternating 30-s periods of beating with zirconia beads in a Minibeadbeater (Biospec Products) and chilling on ice. DNA concentration was measured with Qubit fluorometric quantification (measured concentrations in Supplementary Table 3).

DNA that was to be used for to 10x Chromium preparation was size-selected with a BluePippin instrument targeting the size range of 10–50 kb, the maximum yielding measurable output. DNA for the SLR library preparation was size-selected with the BluePippin instrument targeting the size range of 8–12 kb, per the manufacturer's recommended protocol. DNA for TruSeq conventional short-read library preparation was not size-selected. Libraries were prepared for sequencing with the 10x Genomics Chromium system, Illumina TruSeq SLR kit, or Illumina TruSeq Nano kit according to the respective manufacturer's standard protocol. Library fragment size was quantified with an Agilent 2100 Bioanalyzer instrument (Agilent Technologies) with a High Sensitivity DNA kit.

For the marine-sediment sample, DNA was extracted with an RNasey PowerSoil DNA elution kit (Qiagen, cat. no. 12867-25) in combination with an RNasey PowerSoil Total RNA kit (Qiagen, cat. no. 12866-25). The protocol was modified from the manufacturer's instructions to include a bead-beating step of 5.5 m/s for 2 × 45 s with a FastPrep-24 instrument (MP Biomedicals, cat. no. 116005500). DNA was eluted in 100 μl DNAse- and RNase-free water and stored at −80 °C until further processing. DNA was then size-selected with the BluePippin instrument targeting the size range of 10–50 kb (the maximum yielding measurable output), and a library was prepared for sequencing with the 10x Genomics Chromium system, according to the manufacturer's standard protocol.

Sequencing. Chromium libraries from the mock metagenome, stool samples from healthy individuals, and ocean sediment were sequenced with 2 × 151-bp sequencing on an Illumina HiSeq 4000 instrument. The stool samples were allocated a half lane each. The marine sediment was allocated a quarter lane and a full lane. The mock metagenome was allocated one lane (total gigabase coverage in Supplementary Table 4). The resulting sequences were demultiplexed and barcoded with the 10x Longranger v2.1.3 mkfastq tool to generate raw reads, then subjected to quality control.

DNA from the stool samples was prepared for sequencing with an Illumina TruSeq library prep kit according to the manufacturer's standard protocol and subjected to 2 × 101-bp sequencing on an Illumina HiSeq 4000 instrument. Each library was allocated a half lane of sequence coverage (total gigabase coverage in Supplementary Table 4). Raw reads were then subjected to quality control (described below).

DNA from the stool samples was prepared for sequencing with an Illumina TruSeq Synthetic Long-read library prep kit according to the manufacturer's standard protocol. Because these libraries use the sample barcode to identify the 384 molecular partitions, samples cannot be multiplexed. Thus, each library was necessarily allocated one full lane of 2 × 151-bp coverage on an Illumina HiSeq 4000 instrument (total gigabase coverage in Supplementary Table 4). Raw reads were then subjected to quality control (described below).

After sequencing, all libraries were trimmed with cutadapt39 v1.8.1 with a minimum length of 60 bp and minimum terminal base score of 20 (with the exception of the ATCC mock metagenome reads, which were trimmed with a minimum trimmed read length of 80 bp and minimum terminal base score of 35, as well as 8 bp removed from the 5′ end and 15 bp removed from the 3′ end, owing to low read quality). Reads were synced, and orphans (reads whose pair mates were filtered out) were placed in a separate single-ended fastq file with an in-house script.

Assembly of the mock metagenome and human stool samples. Data from read-cloud 10x Genomics Chromium and short-read TruSeq libraries were assembled with MetaSPAdes v3.11.1 (ref. 40) with default parameters. For read-cloud libraries, MetaSPAdes assembled seed contigs were then assembled with Athena (Supplementary Note 2).

Synthetic long reads were assembled with a two-stage process: (i) synthetic long reads were assembled from trimmed sequencing reads with TruSPAdes41 v3.11.1 with default parameters, and (ii) these assembled synthetic long reads were then further assembled into contigs with CANU v1.5 (ref. 42) with the following parameters: errorRate = 0.06, genomeSize = 45.00m, contigFilter = “2 2000 1.0 1.0 2.0”, stopOnReadQuality = false.

Assembly of marine-sediment sample. Data from the marine-sediment read-cloud library were assembled with MEGAHIT v1.1.2 (ref. 43) with default parameters. MEGAHIT short-read–assembled contigs were then used as seed contigs and assembled with Athena (Supplementary Note 2).

To make Athena assembly tractable on complex metagenomes, Athena was modified to perform subassembly for only well-covered seed contigs with a minimum short-read–sequence coverage of 20×. MEGAHIT contigs excluded from Athena assembly were then mapped back to the initial Athena draft, and each of these contigs was included in the final output if more than 2,000 bases did not align to the initial draft.

Assembly classification, genome draft binning, and gene identification. For each approach, raw short reads were aligned to assembled contigs with BWA v0.7.10 (ref. 44) to generate contig–coverage profiles. Contigs were then binned with Metabat v2.12.1 (ref. 16) to form genome drafts. Bins were evaluated with Metaquast v4.6.0 (ref. 36) for assembly size and contiguity, CheckM v1.0.7 (ref. 37) for completeness and contamination as genome drafts, Prokka v1.12 (ref. 45) for gene content, Aragorn v1.2.36 (ref. 46) to count tRNA sequences, and Barrnap v0.7 (ref. 47) to count 5S, 16S, and 23S ribosomal RNA loci. We defined an intermediate-quality genome as one with >70% completeness and <10% contamination. We adopted previously described standards defining a high-quality genome as one containing at least 18 tRNA loci; at least one copy each of 5S, 16S, and 23S; >90% completeness; and <5% contamination48.

Individual contigs from all assemblies were assigned taxonomic classifications with Kraken v0.10.6 (ref. 48) with a custom database constructed from the RefSeq and GenBank41–48 bacterial-genome collections. Each genome draft was assigned a species-level label if >60% of total bases within the draft shared a species-level classification. Otherwise, drafts were assigned the majority genus-level label.

Code availability. The Athena assembler together with a demonstration data set can be found at https://github.com/abihara/athena_meta/. Version 1.1 was used to generate the results in this paper. The demonstration data set contains a subset of the read clouds from the ATCC 20 mock metagenome, for which assembly with Athena yields the full Lactobacillus gasseri genome in
two sequence contigs. The binning, annotation, and evaluation workflow can be found at https://github.com/elimoss/metagenomics_workflows/.

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

**Data availability.** The data sets generated during the current study are available in the NCBI Sequence Read Archive under Bioproject accession PRJNA380276. 10x read barcodes have been encoded as sample barcodes and must be reformatted as molecular barcodes for use with Athena.

39. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
40. Bankevich, A. et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
41. Bankevich, A. & Pevzner, P.A. TruSPAdes: barcode assembly of TruSeq synthetic long reads. *Nat. Methods* **13**, 248–250 (2016).
42. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
43. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHit: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Brujin graph. *Bioinformatics* **31**, 1674–1676 (2015).
44. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
45. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
46. Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* **32**, 11–16 (2004).
47. Seemann, T. barrnap. Github https://github.com/tseemann/barrnap/ (2018).
48. Wood, D.E. & Salzberg, S.L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
49. Benson, D.A. et al. GenBank. *Nucleic Acids Res.* **41**, D36–D42 (2013).
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 Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

- **n/a** Confirmed

- An exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- An indication of 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 statistics 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
- Clearly defined error bars
- State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on [statistics for biologists](#) may be useful.

Software and code

Policy information about **availability of computer code**

**Data collection**

No software was used in the process of data collection.

**Data analysis**

Read cloud data were processed with 10X Genomics Longranger v2.1.3. Read cloud, synthetic long read and short read data were subjected to quality control using cutadapt v1.8.1. Data were assembled with MetaSPAdes v3.11.1, TruSPAdes 3.11.1, CANU v1.5, MEGAHIT v1.1.2, and Athena, the focus of the present manuscript. The Athena read cloud assembler is available at [www.github.com/abishara/athena_meta](http://www.github.com/abishara/athena_meta). Assemblies were further analyzed with BWA v0.7.10, Metabat v2.12.3, Metaquast v4.6.0, CheckM v1.0.7, Prokka v1.12, Aragorn v1.2.36, Barrnap v0.7, and Kraken v0.10.6.

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/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.
Field-specific reporting

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

Life sciences

Study design

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

- **Sample size**: No sample size calculation was performed. Four independent samples were chosen to evaluate the technical contributions presented in the manuscript. We selected four samples for the technical development and evaluation of this method to meet or exceed the precedence of prior publications that have developed methods for improved genome assembly (Kuleshov et al, NBT, 2015). The samples selected represent a spectrum of species richness that spans nearly two orders of magnitude and which reflects what is observed in most human stool samples and some environmental samples.

- **Data exclusions**: No data were excluded from analysis.

- **Replication**: Both human samples were sequenced in the presence and absence of size selection, as well as with DNA extraction by two alternative approaches. All samples yielded compositionally concordant results.

- **Randomization**: Randomization was not relevant to our study as there were no groups to which participants could be assigned.

- **Blinding**: Blinding was not relevant to our study as there were no differing selection or exclusion criteria between sampled participants.

Materials & experimental systems

Policy information about availability of materials

- **n/a**: Involved in the study
- **Unique materials**: Antikaryotic cell lines, Research animals, Human research participants

**Unique materials**

- **Obtaining unique materials**: The mock mixture DNA is available from ATCC (product MSA-1003). The ocean sediment sample and human stool samples are unique materials and can be made available through a material transfer agreement with the corresponding author.

**Human research participants**

Policy information about studies involving human research participants

- **n/a**: Involved in the study
- **Population characteristics**: Samples were obtained from human volunteers that have been consented under a Stanford IRB approved protocol.
| Method-specific reporting |
|--------------------------|
| n/a Involved in the study |
| ✖️ ChIP-seq               |
| ✖️ Flow cytometry         |
| ✖️ Magnetic resonance imaging |