Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome

Volodymyr Kuleshov1,2, Chao Jiang2, Wenyu Zhou2, Fereshteh Jahanbani2, Serafim Batzoglou1,3 & Michael Snyder2,3

Identifying bacterial strains in metagenome and microbiome samples using computational analyses of short-read sequences remains a difficult problem. Here, we present an analysis of a human gut microbiome using TruSeq synthetic long reads combined with computational tools for metagenomic long-read assembly, variant calling and haplotyping (Nanoscope and Lens). Our analysis identifies 178 bacterial species, of which 51 were not found using shotgun reads alone. We recover bacterial contigs that comprise multiple operons, including 22 contigs of >1 Mbp. Furthermore, we observe extensive intraspecies variation within microbial strains in the form of haplotypes that span up to hundreds of Kbp. Incorporation of synthetic long-read sequencing technology with standard short-read approaches enables more precise and comprehensive analyses of metagenomic samples.

As yet, only a small fraction of the microbial world has been isolated and studied in the laboratory, and little is known about species that cannot be cultured1. Metagenomics has begun to shed light on this unculturable 'microbial dark matter' by sequencing the DNA of microbial communities directly from the environment2. Metagenomic analyses of soil3, water4 and human microbiome5 samples have already increased our understanding of the microorganisms present in these environments.

Although short-read sequencing technologies have enabled high-throughput metagenomic studies, the limited read lengths, combined with the complexity of microbial samples (hundreds to thousands of species), can make it difficult to accurately identify bacterial strains, recover whole genomes, catalog sample diversity and assess the abundance of species and strains.

Many approaches have been proposed to overcome these problems. Binning de novo assembled contigs using metrics such as tetranucleotide frequencies6, genomic coverage under different DNA extraction methods7 or abundance correlations between samples from multiple individuals8 have all been used to identify species and recover their genomes. However, each approach has limitations: tetranucleotide frequencies may vary within the same species8, closely related bacteria have similar DNA extraction efficiencies7, and establishing abundance correlations among individuals may require a large number of samples8. Metagenomic contigs can also be scaffolded using chromatin-level contact probability maps generated by the high-throughput chromosome conformation capture (Hi-C) technology9; however, Hi-C has high input DNA requirements and its performance has not been assessed on bona fide high-complexity metagenomes. An alternative technology, single-molecule real-time sequencing10, has been used to sequence 16S rRNA amplicons11, but has seen limited application to whole metagenomes.

Recently, TruSeq synthetic long-read sequencing has been developed to increase the effective read length available using the Illumina platform from hundreds to more than ten thousand base pairs. This method uses a modified library-preparation protocol, in which kilobase-long DNA fragments are extracted, diluted, amplified and reassembled from regular short reads12. Synthetic long-read sequencing has been used in metagenomics for assembly validation13 and for studying environmental metagenomes14. Here, we report the first, to our knowledge, study of the human gut microbiome using synthetic long reads.

We present three improvements compared with previous technologies. First, we demonstrate that long reads, in conjunction with Lens, an algorithm developed for this study, reveal extensive haplotype diversity within individual bacterial strains of the same species. This level of resolution was inaccessible using previous technologies, which at best studied microorganisms at the strain level. Whereas earlier studies aimed to distinguish one strain from another, we looked for differences within the strains themselves. Second, we use long reads to assemble de novo hundreds of megabases of genomic sequence in contigs of complete operons and whole bacterial chromosomes. Finally, we accurately determine microbial composition and identify 51 species that were not found with shotgun reads alone.

RESULTS
Sequencing the gut microbiome using synthetic long reads
We applied our long-read sequencing approach to two metagenomic data sets: the human microbiome project’s staggered mock metagenomic community3 (mock metagenome), and a sample from the gut of a healthy male adult individual (human gut metagenome). The mock metagenome is a synthetic community of 20 organisms with known reference genomes (Supplementary Table 1) that is widely used for validation. We generated three TruSeq synthetic long-read libraries

1Department of Computer Science, Stanford University, Stanford, California, USA. 2Department of Genetics, Stanford University School of Medicine, Stanford, California, USA. 3These authors contributed equally to this work. Correspondence should be addressed to V.K. (kuleshov@stanford.edu), S.B. (serafim@cs.stanford.edu) or M.S. (mpsnyder@stanford.edu).

Received 13 December 2014; accepted 23 October 2015; published online 14 December 2015; doi:10.1038/nbt.3416
Figure 1 The Nanoscope pipeline and the Lens algorithm. (a) Nanoscope assembles short and long reads using the Soapdenovo2 and Celera assemblers and merges the results with Minimus2; it then assigns taxonomic labels to contigs with the Fragment Classification Package (FCP) and identifies bacterial strains with Lens; finally, it estimates abundances of detected bacterial species by mapping short reads to contigs and by aggregating the coverage over all contigs assigned to the same species. (b, i) The Lens algorithm identifies heterozygous variants in the assembled genomic contigs. (b, ii) These variants are supported by long reads aligned to the contigs. Each long read originates from a single organism; thus, the variants it supports must belong to the same strain. (b, iii) By connecting reads at their overlapping variants, Lens places the variants into multi-kilobase-long haplotypes associated with bacterial strains. The number of haplotypes is a priori unknown and is inferred from the data.

(2.9 Gbp of sequence, N50 read length of 9.2 kbp; Supplementary Fig. 1) for this data set, in addition to 3.1 Gbp of standard 101-bp paired-end Illumina short-read libraries (Supplementary Table 2). For the human gut metagenome, we generated seven TruSeq synthetic long-read libraries (8.3 Gbp of sequence, N50 read length of 8.6 kbp; Supplementary Fig. 2), and complemented these data with 8.1 Gbp of standard Illumina libraries (Supplementary Table 3). The similar sequencing amounts of short and long reads for both samples helped assess the benefits of using longer read lengths.

We mapped the long reads to the known reference genomes of the mock community using the MUMmer aligner. Accuracy was high, with less than 0.5% of reads misassembled (Supplementary Table 4). However, we observed differences in coverage between short- and long-read technologies, with long reads covering ~15% fewer base pairs overall than short reads (Supplementary Tables 5 and 6 and Supplementary Fig. 3). In particular, four organisms were highly covered (>98%) by short reads but long-read coverage was substantially less thorough (<75%), suggesting that long reads have more sequence bias. Interestingly, six organisms had at least 10% of their genomes covered by long reads but not by short reads, suggesting that the two technologies can be complementary (Supplementary Table 7). We also found that abundance estimates were substantially different between the technologies (sometimes by more than an order of magnitude), indicating that long reads on their own may be insufficient for abundance estimation (Supplementary Fig. 4). Differences in coverage have been previously linked to the long reads’ increased sensitivity to GC content during the PCR amplification step12,15.

We assembled long reads from the human gut metagenome sample using Nanoscope, a new bioinformatics pipeline we created. Nanoscope automates metagenomic assembly, species identification, strain analysis and abundance estimation from a combination of short- and long-read data (Fig. 1 and Supplementary Code). It is available online as a free open-source tool (https://github.com/kuleshov/nanoscope).

Nanoscope starts by invoking the Soapdenovo16 and Celera17 assemblers to independently assemble the short- and long-read libraries, before merging the results using Minimus2 (ref. 18). This method produced contigs for more than 650 Mbp of the human gut metagenome (N50 length of 49 kbp; Table 1); these were longer and more complete than ones assembled from either long reads (600 Mbp of sequence; N50 length of 38 kbp; Celera assembler) or short reads (232 Mbp of sequence; N50 length of 8.6 kbp; Soapdenovo2 assembler) alone. Twenty-two of the contigs we obtained were longer than 1 Mbp (Supplementary Fig. 5), indicating that multiple organisms could be assembled completely or almost completely. The longest contig we recovered was 3.9 Mbp; its length and number of predicted open reading frames (ORFs)19 were comparable to that of a closely related complete bacterial genome (see below). For comparison, long reads by themselves produced 19 contigs longer than 1 Mbp, whereas the longest contig from short reads alone was 410 Kbp. This indicates that synthetic long reads assemble a small number of complete chromosomes in addition to dozens of contigs that are almost an order of magnitude longer than ones obtained from short reads.

We also assessed our assembly strategy using the mock metagenome. Our merged assembly of long and short reads recovered 42 Mbp of sequence (of 83 Mbp total) into contigs with an N50 length of 92 kbp (Table 1 and Supplementary Table 8). This assembly was quite accurate, with approximately one misassembly per 400 kbp on average (Supplementary Table 9), a rate that compares favorably to accuracies reported by previous empirical studies of de novo assembly algorithms20.

Using only long reads resulted in much shorter contigs (33 Mbp of sequence; N50 length of 43 kbp), suggesting that for low-complexity metagenomes, combining short- and long-read technologies may substantially improve assembly quality. We assembled one contig longer than 1 Mbp and three contigs longer than 500 Kbp, all of which were assembled using long reads alone, suggesting that short reads mainly improved assembly completeness rather than contiguity.

One of the potential uses of long contigs is the recovery of complete bacterial operons (clusters of functionally related genes

| Table 1 Assembly of the human gut metagenome |
|---------------------------------------------|
| Short | Long | Joint |
|-------|------|-------|
| Number of contigs | 92,247 | 24,199 | 34,786 |
| Largest contig (Mbp) | 0.63 | 3.94 | 3.94 |
| Total length (Mbp) | 233 | 610 | 656 |
| N50 (Kbp) | 8.7 | 37.3 | 49.2 |
| Number of predicted genes | 274,600 | 523,358 | 552,680 |

Short- and long-read libraries were assembled with the Soapdenovo2 and Celera assemblers, respectively. The results were merged using Minimus2 to produce a joint assembly. Long reads assemble into significantly longer contigs that contain about twice as many genes.
that are transcribed together). By using the known positions of operons in the reference genomes of the species in the mock metagenome, we confirmed that operon recovery is feasible with a combined assembly of short and long reads (Supplementary Table 10).

Our combined assembly recovered 4,500 operons, which represents more than half of all the known operons in the mock metagenome and twice the number that can be obtained using short reads alone. Interestingly, long and short reads by themselves reconstructed only about 2,500 operons each, and thus many were assembled from only one data set (Supplementary Figs. 6 and 7).

We attribute this discrepancy to differences in coverage between short and long reads.

In particular, long reads enabled us to recover long flagellar operons present in *E. coli* (Supplementary Table 11); these operons are clinically relevant as flagella contribute to pathogenicity. We assembled complete sequences of 11 flagellar operons from three bacterial species (*E. coli*, *Rhodobacter sphaeroides*, *Pseudomonas aeruginosa*; Supplementary Table 12), which comprised half of the known flagellar operons in the mock metagenome. We also recovered multiple flagellar operons from the gut metagenome (Supplementary Table 13). For example, a 2.3-Mbp contig belonging to the genus Acinetobacter was found to contain ten flagellar operons, the longest of which contained 11 genes.

### Identification of substrains in the gut microbiome

We assessed variation among the bacterial strains whose genomes we assembled using Lens, a new tool that we created (Table 2). In brief, we mapped long reads to assembled contigs using the Burrows-Wheeler aligner (BWA); at many positions, read and contig sequences differed, which we interpreted as variation among recently diverged organisms of the same strain. We used Lens to determine and phase single-nucleotide variants (SNVs) and short indels based on this alignment; Lens

| Genomic variants | 202,574 |
|------------------|--------|
| Genomic regions harboring haplotypes | 2,204 |
| Number of haplotypes | 5,024 |
| N50 region length (bp) | 18,985 |
| Max region length (bp) | 112,271 |
| Fraction of regions intersecting a gene | 95% |
| Fraction of genes intersecting a region | 4.4% |

The human gut microbiome sample contains hundreds of thousands of variants, which are distributed across 2,204 genomic regions of up to 112 kbp in length. A region is defined as a maximal set of variants that can be phased by Lens using long reads.

![Table 2](image)

The variation within bacterial strains identified by the Lens algorithm

We assessed variation among the bacterial strains whose genomes we assembled using Lens, a new tool that we created (Table 2). In brief, we mapped long reads to assembled contigs using the Burrows-Wheeler aligner (BWA); at many positions, read and contig sequences differed, which we interpreted as variation among recently diverged organisms of the same strain. We used Lens to determine and phase single-nucleotide variants (SNVs) and short indels based on this alignment; Lens

![Figure 2](image)

Long reads aligned to assembled metagenomic contigs reveal extensive variation within bacterial strains. (a) Fragment of a 110-kbp long region within a metagenomic contig belonging to the species *Odoribacter splanchnicus*; the region harbors numerous strain variants that can be assembled into bacterial haplotypes. (b) Left: fragment of a bacterial region containing 32 genomic variants that assemble into four bacterial haplotypes. Right: the haplotypes placed in an evolutionary tree satisfying perfect phylogeny; for simplicity, we visualize this tree over 4 of the 32 positions in the region (upper left corner).
performed these tasks by means of algorithms we wrote that do not make any assumptions on either the read length or the ploidy of the organism (Supplementary Methods and Supplementary Code; https://github.com/kuleshov/lens).

Lens found extensive intraspecies variation in almost every bacterial species in the human gut metagenome (Fig. 2 and Supplementary Figs. 8 and 9), which in total contained more than 200,000 variants (Supplementary Methods). Lens assembled these variants into 5,024 haplotypes distributed across about 2,204 genomic regions with an N50 length of 19 kbp (Supplementary Fig. 10); each region contained on average 3.93 bacterial haplotypes. More than 95% of regions overlapped with ORFs, and the longest region we found spanned 112 Kbp and 242 variants and contained four distinct haplotypes.

We observed that significantly fewer variants were present in essential genes compared to nonessential genes, as expected from evolutionary pressure (Supplementary Table 14; P < 0.02). We repeated the same analysis on the mock metagenome and also uncovered a small number of genomic variants (few were expected, as the sample is synthetic; see Supplementary Table 15); as in the gut metagenome, significantly fewer variants were found in essential genes (Supplementary Table 14; P < 1 × 10−3). We used the variant annotation package SNPEff23 to predict the deleteriousness of each mutation in the genome of E. coli (Supplementary Table 16), and found that most variants had low to moderate effects. Only six variants had high effect, and they were all found in nonessential genes rhsB, ydfK, icd and perR. These observations suggest that the variants Lens uncovered are not attributable to noise.

To evaluate the correctness of the phased haplotypes, we determined whether they satisfied perfect phylogeny24. A tree over haplotypes satisfies perfect phylogeny if all strains evolved from a common ancestor, and if during this process, each position mutated once, at most. Although this criterion is not applicable to distantly diverged species, it is useful when organisms are separated by short evolutionary distances, as in the case of bacterial subspecies. In the human gut metagenome, most (85%) of the genomic regions harboring at least four haplotypes satisfied perfect phylogeny (for three haplotypes or less, it is useful when organisms are separated by short evolutionary distances).

We obtain similar results to alternative techniques that used hundreds of pooled samples (Nielsen et al.8) or potentially inaccurate binning approaches (Albertsen et al.7, Iverson et al.6). We also analyze strains at the resolution of individual variants and haplotypes rather than strains or species. WGS, whole genome sequencing; N/A, not available.

### Table 3. Selected metagenomic sequencing approaches compared

| Sample type | Our method | Nielsen et al.8 | Albertsen et al.7 | Iverson et al.6 | Sharon et al.14 |
|-------------|------------|-----------------|-------------------|-----------------|-----------------|
| No. of samples | Gut microbiome | Gut microbiome | Environmental | Environmental | Environmental |
| Seq. platform | TruSeq SLR + Illumina WGS | Illumina WGS | Illumina WGS | SOLID mate-pairs Sanger sequencing | 3 independent |
| Seq. amount | 8 Gbp (long reads) ×40 (subassembly) | 4.5 Gbp/sample | 86 Gbp | 59 Gbp | 1.5 Gbp (long reads) ×40 (subassembly) |
| Analysis type | De novo assembly; phasing | Correlation across multiple samples | DNA extraction efficiency binning | Tetranucleotide binning | De novo assembly |
| Resolution | Individual SNV | Strain | Species with diff. GC content | Family | Strain |
| Longest scaffold | 3.9 Mbp | 733 Kbp | 3.6 Mbp | 2.2 Mbp | <20 Kbp |
| Scaffolding | 49 Kbp | 39 Kbp | 4.1 Kbp overall ~100 Kbp for top species | 6.8 Kbp | 8.2 Kbp |
| Bases assembled | 656 Mbp | 45 Mbp (genes) 35 Gbp (total) | 423 Mbp | 300 Mbp | 500 Mbp/sample |
| No. variants | 200 K | N/A | N/A | N/A | N/A |
| No. haplotypes | 5 K | N/A | N/A | N/A | N/A |

We note that the contigs of Nielsen et al. are also clustered into unordered sets belonging to the same species.
the extent to which it is violated by estimating the number of positions that can be excluded to make perfect phylogeny hold.

We were able to place more than 92% of all gut metagenome regions in perfect phylogeny by excluding at most 5% of positions within each region (Supplementary Figs. 10 and 11). These observations support the hypothesis that the variants we found correspond to distinct bacterial strains that have evolved from one another. Our approach is the first, to our knowledge, to achieve strain resolution and offers a snapshot of how strains evolve in vivo.

Assessing strain abundance with long reads

Nanoscope uses the FCP software package to assign taxonomic labels to assembled contigs. FCP determines labels using either a homology-based approach (based on the lowest common ancestor, or LCA, algorithm), or a composition-based approach (a naive Bayes classifier trained on k-mer frequencies). In principle, longer contigs should be easier to label because they contain more species-specific sequences, and they should map with less ambiguity to known reference genomes.

In the human gut metagenome, 61.4% of contigs assembled from long reads could be labeled using the LCA method, compared with 46.5% of contigs derived from short reads (Supplementary Table 17). Similarly, 89.8% of contigs assembled using long reads could be labeled by the naive Bayes method, compared to 11.0% of contigs assembled using short reads. To assess the accuracy of these assignments, we used the mock metagenome. LCA assigned contigs with 100% accuracy on both long and short reads, whereas the naive Bayes classifier had accuracies of 99% and 98% on contigs obtained from long and short reads, respectively (Supplementary Tables 18–21).

By examining the taxonomic labels assigned to the longest contigs in the gut metagenome, we were able to identify which bacteria could be assembled completely or almost completely (Supplementary Table 22). Five of the ten longest contigs belonged to the genus Bacteroides, and the longest 3.9-Mbp contig was from the genus Odoribacter. Many contigs—including one measuring 2.3 Mbp—could not be assigned accurate labels; these contigs correspond to either unknown species, or to species whose genomes have not yet been fully assembled (Supplementary Tables 23 and 24). Notably, the 2.3-Mbp contig did not match any known reference genome by more than 3.2 Kbp; however, we found that contigs from a fragmented draft assembly of a species from the genus Acinetobacter mapped to the 2.3-Mbp contig completely (Supplementary Fig. 12). This suggests that the genome we recovered of that bacterium was of higher quality than one in a previous study that used metagenomics samples from 396 different individuals. We observed similar mapping results for other unclassified contigs as well.

Finally, we determined the abundance of each bacterium by mapping short reads to gut metagenome contigs and then using the taxonomic labels above to propagate the resulting coverage estimates to each identified bacterium (Supplementary Methods). We found 178 species in the human gut metagenome and these species varied greatly in their abundance: some comprised as much as 5% of the metagenome, and others as little as 0.02% (Fig. 3 and Supplementary Table 25). Interestingly, different species were recovered by short and by long reads: short reads helped find two relatively high-abundance bacteria, whereas long reads uncovered 51 species (mostly of low abundance) that were missed by short reads (given the same amount of sequencing). Moreover, by combining both short and long reads, 58 additional low-abundance bacteria could be identified. Finally, we found that on the mock metagenome, abundance estimates were highly concordant with those obtained from mapping short reads directly to the 20 known reference genomes ($r^2 = 0.97$; Supplementary Fig. 13 and Supplementary Table 26). This serves as an indicator of the accuracy of our approach.

DISCUSSION

Including synthetic long reads in metagenomic analyses significantly improves the delineation of complex metagenomic samples compared with using short-read sequences on their own. Although synthetic long reads have sequencing biases that affect coverage in specific genomic regions, these biases can be overcome using a small amount of additional shotgun sequencing. The resulting approach offers three main advantages over existing methods. First, we recover long bacterial contigs (up to megabases in length) that span operons that could not be recovered by short-read sequencing alone. Second, analysis of long reads produces kilobase-long haplotypes that can reveal evolutionary trends in microbial communities. Finally, longer sequencing read lengths enable identification of bacteria at abundances as low as 0.02% that are undetected by short reads.

Our approach is not without limitations. Synthetic long-read technologies rely on a dilution step that attempts to reduce the number of copies of each repeat to at most one per well; genomes of bacteria at high abundances of 10% or more may not be sufficiently diluted, and several repeat copies may remain in a single well, preventing the subassembly of their long reads. This may explain why some gut metagenome bacteria at ~5% abundance were not identified by long reads alone. A related issue is the presence of short tandem repeats, which may also prevent subassembly. A third shortcoming of our approach is its reliance on a PCR step, which could introduce bias (Supplementary Figs. 3 and 4) and error. However validation using the mock metagenome and comparison with short shotgun reads indicates that this error does not exceed 2–3% (Supplementary Table 27).

Despite these limitations, our approach can shed light on the configuration of complete operons, and facilitate the identification of pathogenic strains (especially in a mixed population). For instance, we assembled multiple kilobase-long flagellar operons that affect motility and thus play a role during infection. Furthermore, the substrain resolution enabled by our methods could assist in understanding how evolution shaped strains over time in situ (Supplementary Fig. 14). Finally, the ability to identify low-abundance bacteria will help reveal the complete composition of environmental samples and discover new species.

The results that we obtained with our sequencing and assembly strategy are comparable to those obtained using previously published methods that required hundreds of human subjects, multiple DNA extraction methods or tetranucleotide binning with a mix of Sanger and mate-pair sequencing (Table 3). Our method requires only a single metagenomic sample, does not involve binning and is superior at identifying bacterial strains. It is also feasible that long reads might work in tandem with previous approaches, as longer contigs will increase the accuracy of their statistical components. Similarly, longer contigs might improve the accuracy of scaffolding techniques such as ones based on Hi-C. Finally, our approach is most closely related to single-molecule real-time sequencing (SMRT), an alternative long-read technology. We assembled a publicly available data set for the mock metagenome (Supplementary Tables 28 and 29 and Supplementary Methods, “comparison to alternative methods”) using the MHAP assembly strategy. This produced long contigs with fewer misassemblies than those that we obtained from synthetic long reads; however, the SMRT contigs also had a 5x higher indel rate and only 88% of SNVs called using Lens in these contigs could be confirmed with short reads (compared to 99% for synthetic long reads). It seems that SMRT reads produce excellent draft metagenomic assemblies, but their high error rates make it difficult to identify and phase variants in the metagenome.
Finally, synthetic long reads were recently used by Sharon et al.\textsuperscript{14} to analyze a soil metagenome sample. By focusing on a set of marker genes, they showed that the community comprised a combination of closely related strains and rare species. Our work demonstrates that long reads can produce much longer contigs than those described by Sharon et al.\textsuperscript{14} (Table 3 and Supplementary Tables 30 and 31); in addition, we didn't need to use a marker gene set, our analysis pipeline can find additional species, including a phylum that marker genes did not reveal on the Sharon et al.\textsuperscript{14} environmental data set (see Supplementary Tables 32 and 33 and Supplementary Methods, “comparison to alternative methods” for full discussion).

In conclusion, we show that the human gut microbiome is more complex than previously thought, particularly in terms of subspecies diversity. The rapid evolution of bacterial strains at the subspecies level could affect human physiology\textsuperscript{29}. Armed with more complete inventories of microorganisms, it might be possible to strengthen associations between human and bacterial phenotypes.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. SRA: SRP065223.

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

ACKNOWLEDGMENTS

This work was supported by US National Institutes of Health/National Human Genome Research Institute (NIH/NHGRI) grant T32 HG000044. V.K. was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) post-graduate fellowship. We thank Illumina, Inc. for their assistance in sample preparation.

AUTHOR CONTRIBUTIONS

S.B. and M.S. conceived the study. W.Z. and F.J. performed library preparation. V.K. developed the Nanoscope pipeline and the Lens algorithm. V.K. and C.J. performed computational analyses. V.K., C.J., S.B. and M.S. wrote the paper. S.B. and M.S. supervised the study.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Rinke, C. et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature \textbf{499}, 431–437 (2013).
2. Thomas, T., Gilbert, J. & Meyer, F. Metagenomics - a guide from sampling to data analysis. Microb. Inform. Exp. \textbf{2}, 3 (2012).
3. Daniel, R. The metagenomics of soil. Nat. Rev. Microbiol. \textbf{3}, 470–478 (2005).
4. Venter, J.C. et al. Environmental genome shotgun sequencing of the Sargasso Sea. Science \textbf{304}, 66–74 (2004).
5. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature \textbf{486}, 207–214 (2012).
6. Iverson, V. et al. Unangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. Science \textbf{335}, 587–590 (2012).
7. Albertsen, M. et al. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat. Biotechnol. \textbf{31}, 533–538 (2013).
8. Nissen, T.B. et al. MetaHIT Consortium. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat. Biotechnol. \textbf{32}, 822–828 (2014).
9. Nenov, J.N., Liachko, I., Dunham, M.J. & Shendure, J. Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. G3 (Bethesda) \textbf{4}, 1339–1346 (2014).
10. Ed, J. et al. Real-time DNA sequencing from single polymerase molecules. Science \textbf{323}, 133–138 (2009).
11. Fichtot, E.B. & Norman, R.S. Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform. Microbiome \textbf{1}, 10 (2013).
12. Kuleshov, V. et al. Whole-genome haplotyping using long reads and statistical methods. Nat. Biotechnol. \textbf{32}, 261–266 (2014).
13. Di Rienzo, S.C. et al. The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. eLife \textbf{2}, e01102 (2013).
14. Sharon, I. et al. Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. Genome Res. \textbf{25}, 534–543 (2015).
15. McCoy, R.C. et al. Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. PLoS One \textbf{9}, e106689 (2014).
16. Li, R. et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. \textbf{20}, 265–272 (2010).
17. Meyers, E.W. et al. A whole-genome assembly of Drosophila. Science \textbf{287}, 2196–2204 (2000).
18. Sommer, D.D., Delcher, A.L., Salzberg, S.L. & Pop, M. Minimus: a fast, lightweight genome assembler. BMC Bioinformatics \textbf{8}, 64 (2007).
19. Zhu, W., Lomsadze, A. & Borodovsky, M. Ab initio gene identification in metagenomic sequences. Nucleic Acids Res. \textbf{38}, e132 (2010).
20. Blagon, J. et al. GAGE-B: an evaluation of genome assemblers for bacterial organisms. Bioinformatics \textbf{29}, 1718–1725 (2013).
21. Mao, F., Dam, P., Chou, J., Olman, V. & Xu, Y. DOOR: a database for prokaryotic operons. Nucleic Acids Res. \textbf{37}, D459–D463 (2009).
22. Chen, W.H., Minguez, P., Lercher, M.J. & Bork, P. OGEE: an online gene essentiality database. Nucleic Acids Res. \textbf{40}, D901–D906 (2012).
23. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) \textbf{6}, 80–92 (2012).
24. Gusfield, D. Efficient algorithms for inferring evolutionary trees. Networks \textbf{21}, 19–28 (1991).
25. Parks, D.H., MacDonald, N.J. & Beiko, R.G. Classifying short genomic fragments from novel lineages using composition and homology. BMC Bioinformatics \textbf{12}, 328 (2011).
26. Almeberg, J. et al. Binning metagenomic contigs by coverage and composition. Nat. Methods \textbf{11}, 1144–1146 (2014).
27. Burton, J.N. et al. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat. Biotechnol. \textbf{31}, 1119–1125 (2013).
28. Berlin, K. et al. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat. Biotechnol. \textbf{33}, 623–630 (2015).
29. Lieberman, T.D. et al. Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. Nat. Genet. \textbf{43}, 1275–1280 (2011).
ONLINE METHODS

Metagenomic sample preparation. Mock microbial DNA, HM-277D Staggered v5.2H, was obtained from BEResourses. Gut microbiome DNA was isolated from the frozen feces of a healthy man using PowerSoil DNA Isolation Kit (MO BIO Laboratories, Inc.). Both DNA samples were sequenced using Illumina TruSeq synthetic long-read technique (three and seven libraries for mock and gut microbiome samples, respectively) and the standard shotgun technique, with each library sequenced on one full lane of HiSeq. All libraries were prepared according to the manufacturer’s standard protocol. Shotgun sequence reads for both the mock and the gut metagenomic samples were subsampled at random to produce subsampled libraries containing the same number of base pairs as the in TruSeq synthetic long-read libraries. The results were assembled on the Illumina Basespace platform, according to the standard protocol.

We used Ovation Ultralow DR Multiplex Systems 1–8 (0330-32, NuGEN Technologies, Inc.) for whole genome library preparation. Briefly, 100 ng of intact gDNA was diluted into 120 μl of 1× low EDTA-TE buffer and transferred to Covaris snap-cap microtube and fragmented to 300 bp following Covaris recommended settings. Fragmented DNA was purified using Agencourt RNAClean XP bead, provided by NuGen Library preparation kit. The sheared DNA was then subjected to end repair and adaptor ligation. Adaptor ligated libraries were purified with Agencourt RNAClean XP bead and amplified using 18 PCR cycle of 94 °C for 30 s, 60 °C for 30 s and 72 °C for 1 min. Agencourt RNAClean XP bead was used for amplified library purification and libraries fragment distribution was validated on Bioanalyzer DNA Chip 100.

Overview of the Nanoscope pipeline. In order to facilitate the analysis of synthetic long-read data in the context of metagenomics, we have developed a bioinformatics pipeline called Nanoscope (Fig. 1). Nanoscope takes as input a set of long-read libraries together with optional (but highly recommended) short-read libraries. It then performs a four-stage analysis of these data that includes de novo assembly, variant calling and haplotyping, taxonomic identification and abundance estimation.

Nanoscope starts by invoking the Soapdenovo16 and Celera17 assemblers to independently assemble the short and long-read libraries, before merging the results using Minimus2 (ref. 18). In the next step, it invokes a variant calling and phasing algorithm called Lens to analyze the assembled contigs for strain variation. Lens reveals hundreds to thousands of sites where individual bacteria of the same strain differ from each other and then phases these variants into bacterial haplotypes. A typical contig might harbor more than a dozen different strain haplotypes, each of which may contain thousands of sequence variants. Variants and haplotypes are determined using a simple model (see the section on Lens below).

Finally, Nanoscope invokes the FCP software package25 to assign taxonomic labels to assembled contigs and to estimate bacterial abundances. The latter task is done by mapping short reads to assembled contigs and by aggregating the coverage over all contigs assigned to the same species. Computing abundances from short reads avoids certain biases inherent to synthetic long reads; this task is done by mapping short reads to assembled contigs and by aggregating the coverage over all contigs assigned to the same species. The Lens haplotyper and variant caller. Lens is a new variant calling and phasing tool specialized for metagenomes and synthetic long reads. It is based on algorithms that, unlike previous approaches30,31, do not make any assumptions on the length of sequencing reads or the ploidy of the organism; we have found that these factors may confuse existing bacterial variant callers and lead to suboptimal results (Supplementary Methods, “comparison to alternative methods”). At a high level, Lens does two things: starting from an alignment of long reads to assembled contigs (or to bacterial reference genomes), it first determines positions at which the reads and the reference differ; these positions are indicative of multiple closely related strains of the same bacterium. Then, Lens phases these variants into long haplotypes, each haplotype being defined in this context as a set of variants that co-occur within the same bacterial strain.

The Lens haplotyper leverages the fact that each long read originates from a single organism, and therefore all variants within a read must belong to the same strain. By connecting reads at their overlapping variants, Lens places the variants into multi-kilobase-long haplotypes in a process that is reminiscent of single-individual haplotyping (SIH) techniques32. In our setting, the number of true haplotypes is an unknown parameter that may be greater than two, making the phasing problem considerably more difficult. Although there exist well-known phasing algorithms for polyploid genomes (e.g., plants or cancer genomes), they all assume a fixed, known ploidy33,37, with the notable exception of some recent methods developed while this paper was under review34,35. Lens on the other hand infers the ploidy directly from the data. More precisely, Lens assembles haplotypes using an approximate greedy procedure (Supplementary Methods, “the Lens haplotyper and variant caller”); this choice is in part due to the fact that the SIH problem (of which ours is a generalization) is computationally intractable36. In brief, Lens sorts aligned reads from left to right and in turn uses each read to either extend an existing haplotype or to form a new one, depending on the read-haplotype overlap and on the cost of forming a new cluster (both are tunable parameters for the algorithm). Our high-level approach may in principle have applications outside metagenomics, such as in cancer genome phasing.

30. Walker, B.J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9, e112963 (2014).
31. Nijkamp, J.F., Pop, M., Reinders, M.J.T. & de Ridder, D. Exploring variation-aware contig graphs for (comparative) metagenomics using MaryGold. Bioinformatics 29, 2826–2834 (2013).
32. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29, 1072–1075 (2013).
33. Treangen, T.J. et al. MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. Genome Biol. 14, R2 (2013).
34. Schloss, P.D. et al. Introducing methur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75, 7537–7541 (2009).
35. Duitama, J. et al. Fasmid-based whole genome haplotype calling of a HapMap trio child: evaluation of Single Individual Haplotyping techniques. Nucleic Acids Res. 40, 2041–2053 (2012).
36. Berger, E., Yorukoglu, D., Peng, J. & Berger, B. HapTree: a novel Bayesian framework for single individual polyploidy using NGS data. PLoS Comput. Biol. 10, e1003502 (2014).
37. Aguiar, D. & Istrail, S. Haplo-type assembly in polyploid genomes and identical by descent shared tracts. Bioinformatics 29, i352–i360 (2013).
38. Niklas, N. et al. cFinder: definition and quantification of multiple haplotypes in a mixed sample. BMC Res. Notes 8, 422 (2015).
39. Pulido-Tamayo, S. et al. Frequency-based haplotype reconstruction from deep sequencing data of bacterial populations. Nucleic Acids Res. 43, e105 (2015).
40. Gusfield, D. Inference of haplotypes from samples of diploid populations: complexity and algorithms. J. Comput. Biol. 8, 305–323 (2001).