AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

METHODS AND PROTOCOLS
Novel Systems Biology Techniques

Evaluating the Information Content of Shallow Shotgun Metagenomics

Benjamin Hillmann,[a] Gabriel A. Al-Ghalith,[b] Robin R. Shields-Cutler,[c] Qiyun Zhu,[d] Daryl M. Gohl,[e] Kenneth B. Beckman,[e] Rob Knight,[d,f,g] Dan Knights[a,b,c]

[a]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, Minnesota, USA
[b]Bioinformatics and Computational Biology, University of Minnesota, Minneapolis, Minnesota, USA
[c]Biotechnology Institute, University of Minnesota, Minneapolis, Minnesota, USA
[d]Department of Pediatrics, University of California San Diego, San Diego, California, USA
[e]University of Minnesota Genomics Center, Minneapolis, Minnesota, USA
[f]Department of Computer of Science and Engineering, University of California San Diego, San Diego, California, USA
[g]Center for Microbiome Innovation, University of California San Diego, San Diego, California, USA

ABSTRACT Although microbial communities are associated with human, environmental, plant, and animal health, there exists no cost-effective method for precisely characterizing species and genes in such communities. While deep whole-metagenome shotgun (WMS) sequencing provides high taxonomic and functional resolution, it is often prohibitively expensive for large-scale studies. The prevailing alternative, 16S rRNA gene amplicon (16S) sequencing, often does not resolve taxonomy past the genus level and provides only moderately accurate predictions of the functional profile; thus, there is currently no widely accepted approach to affordable, high-resolution, taxonomic, and functional microbiome analysis. To address this technology gap, we evaluated the information content of shallow shotgun sequencing with as low as 0.5 million sequences per sample as an alternative to 16S sequencing for large human microbiome studies. We describe a library preparation protocol enabling shallow shotgun sequencing at approximately the same per-sample cost as 16S sequencing. We analyzed multiple real and simulated biological data sets, including two novel human stool samples with ultradeep sequencing of 2.5 billion sequences per sample, and found that shallow shotgun sequencing recovers more-accurate species-level taxonomic and functional profiles of the human microbiome than 16S sequencing. We discuss the inherent limitations of shallow shotgun sequencing and note that 16S sequencing remains a valuable and important method for taxonomic profiling of novel environments. Although deep WMS sequencing remains the gold standard for high-resolution microbiome analysis, we recommend that researchers consider shallow shotgun sequencing as a useful alternative to 16S sequencing for large-scale human microbiome research studies where WMS sequencing may be cost-prohibitive.

IMPORTANCE A common refrain in recent microbiome-related academic meetings is that the field needs to move away from broad taxonomic surveys using 16S sequencing and toward more powerful longitudinal studies using shotgun sequencing. However, performing deep shotgun sequencing in large longitudinal studies remains prohibitively expensive for all but the most well-funded research labs and consortia, which leads many researchers to choose 16S sequencing for large studies, followed by deep shotgun sequencing on a subset of targeted samples. Here, we show that shallow- or moderate-depth shotgun sequencing may be used by researchers to obtain species-level taxonomic and functional data at approximately the same cost as amplicon sequencing. While shallow shotgun sequencing is not intended to replace

Received 9 May 2018 **Accepted** 10 September 2018 **Published** 13 November 2018

Citation Hillmann B, Al-Ghalith GA, Shields-Cutler RR, Zhu Q, Gohl DM, Beckman KB, Knight R, Knights D. 2018. Evaluating the information content of shallow shotgun metagenomics. mSystems 3:e00069-18. https://doi.org/10.1128/mSystems.00069-18.

Editor John F. Rawls, Duke University School of Medicine

Address correspondence to Dan Knights, dknights@umn.edu.

Shallow shotgun sequencing as an alternative to 16S for human microbiomes

deep shotgun sequencing for strain-level characterization, we recommend that microbiome scientists consider using shallow shotgun sequencing instead of 16S sequencing for large-scale human microbiome studies.

KEYWORDS human microbiome, metagenomics, microbiome, shotgun metagenomics

Despite the close association of microbial communities with many aspects of human, environmental, plant, and animal health (1–4), it is not currently possible to characterize precisely the species and genes present in a microbial community in a cost-effective manner. The microbial communities of human microbiomes are complex, multivariate, and multidimensional, requiring large studies to power novel biomarker discovery and predictive modeling (4, 5). Deep whole-metagenome shotgun (WMS) sequencing can provide highly resolved strain-level taxonomic and functional information but is generally cost-prohibitive for large-scale studies. Many of the largest microbiome studies to date have been performed via 16S rRNA gene amplicon (16S) sequencing, a cost-effective alternative, but 16S sequencing typically provides only genus-level taxonomic assignments (6) and rough estimates of the functional repertoire (7, 8), limiting the amount of information that can be learned from the data. The purpose of this paper is to evaluate shallow shotgun sequencing as a possible cost-effective alternative to 16S sequencing for large-scale biomarker discovery with improved taxonomic resolution and functional accuracy.

A major concern for the use of any microbiome assay is the ability to identify and quantify taxonomic and functional traits from within a complex community. Deep WMS sequencing has a number of advantages over 16S sequencing for microbiome profiling for these purposes in well-characterized environments; for example, deep WMS sequencing of mixed communities, such as the human gut microbiome, has been effective at recovering strain-level polymorphisms and functional traits for abundant strains (9–11). Both shallow and deep shotgun sequencing are also less subject to amplification bias than 16S sequencing because they do not rely on targeted primers to amplify a marker gene (12). However, at the time of writing, WMS sequencing typically costs several times more per sample than 16S sequencing for library preparation and DNA sequencing. In addition to that of the sequencing itself, the cost of WMS sequencing includes labor and reagents for quality control and sequencing library preparation.

Although extensive work has been done to characterize how many 16S sequencing reads are required for quantifying relevant biological signals of different types, the same has not been done for alpha diversity, beta diversity, species profiles, and functional profiles with shotgun metagenomics. One prior study estimated the average coverage of each microbe given its relative abundance in a metagenomic sample at a given depth, concluding that approximately $\sim$7 Gb of pairs of sequencing data were required to achieve $>$20× coverage of each microbe above 1% relative abundance (13). Another study compared 16S and WMS sequencing by rarefying reads from depths of 500 to 100,000 repeatedly, concluding that 16S sequencing has significant primer bias. This study mentioned that 16S sequencing is typically used for dense sampling efforts, including longitudinal studies (14). Another study explored the use of database-free methods for metagenomic analysis and found that approximately 1 Gb of pairs would provide sufficient metagenomic coverage to compare metagenomic samples with such methods (15).

The depth necessary for metagenomic sequencing in a particular study depends on the purpose of the study; in many studies, the key goals are to understand which species and functions are present and to identify biomarkers related to experimental groups or outcomes. To address the important question of how many reads are required to capture species-level taxonomic and functional assignments (e.g., KEGG Orthology groups [16]), we analyzed shotgun sequencing data at various depths. We found that shotgun sequencing can produce species and functional profiles at a level of quality similar to that of deep WMS sequencing using few as 0.5 million sequences per sample, as demonstrated on deep whole-genome sequencing (WMS) samples from the Human Microbiome Project (HMP) data set (4), the HMP mock community (12), a diabetes study (17), simulated human gut microbiomes, and two novel human stool samples on which we performed ultradeep sequencing of 2.5 billion sequences per sample.

RESULTS

A comparison between deep WMS sequencing and shallow shotgun sequencing in real and simulated biological data sets demonstrated that shallow shotgun sequencing provides nearly the same accuracy at the species and functional level as deep WMS sequencing for known species and genes in five key aspects of microbiome analysis: (i) beta diversity (Fig. 1B and C); (ii) alpha diversity (Fig. 1E and F); (iii) species composition (Fig. 2A and C); (iv) functional composition (Fig. 2B and D); and (v) clinical biomarker discovery (Fig. 3).

Alpha and beta diversity profiling. We obtained deep WMS sequencing data from the Human Microbiome Project (HMP) (4) and subsampled the data to simulate shallow shotgun sequencing depth across five body subsites representing skin, oral, and gut habitat microbiomes. Surprisingly little sequencing was needed to recover the diversity trends found in deep WMS sequencing data. We annotated the deep WMS sequencing and the shallow shotgun sequencing data using an accelerated version of fully exhaustive gapped Needleman-Wunsch alignment (18, 19) for both taxonomic and functional profiles, against all complete representative bacterial genomes from the reference database RefSeq version number 82 (20). Fully exhaustive alignment allowed us to identify any and all ties for best match for each input sequence according to sequence identity. Species relative abundance profiles were derived by tabulating the number of sequences with at least 80% of the best hits belonging to one species. This is similar to the direct mapping approach used commonly in k-mer-based approaches to shotgun metagenomics taxonomic profiling (21) but with higher sensitivity and recall due to the use of gapped sequence alignment (see the simulated data analysis below). Using the fully exhaustive alignment approach, we found that 20% to 40% of all sequences could be identified as species markers because they were uniquely present in only one species in the database (Fig. 1A). We then bootstrapped these samples repeatedly down to 10, 100, 1,000, 10,000, 100,000, and 1 million sequences per sample and reran the analysis to quantify species-level alpha diversity and beta diversity profiles. In all cases, a depth of 0.5 million sequences was more than sufficient to recover the same alpha- and beta-diversity signals, as with deep WMS sequencing (Fig. 1B and D). We used Procrustes analysis (22) to confirm that the beta diversity matrices based on shallow and deep data were similar (Procrustes test, P value = 0.001).

Species and functional profiles in human stool samples. In order to compare the performances of shallow shotgun sequencing for species and functional profiling with ultradeep sequencing, we obtained ultra-deep WMS sequencing of 2.5 billion sequences per sample on novel human stool samples from two individuals. At the time of writing, to the best of our knowledge, this was the deepest sequencing that had been performed on human stool microbiomes. Using these two novel samples, we measured the species profiles and functional profiles as described above with exhaustive gapped alignment against a full genome database for species and a database of genes annotated with KEGG Orthology groups (KOs) (16). We performed this analysis at the full depth of 2.5 billion sequences and then subsampled to lower depths. Species profiles at 0.5 million sequences per sample had an average correlation of 0.990 with ultradeep WMS sequencing data (Spearman correlation, $n = 112$, $P < 2 \times 10^{-16}$) across the two samples (Fig. 2A and C), and the average KO profile correlation was 0.971 (Spearman correlation, $n = 4{,}394$, $P < 2 \times 10^{-16}$) (Fig. 2B and D). For KO annotation, we used direct gene observation for all but the lowest-abundance genes, for which we augmented the direct KO counts with counts of all KOs contained in observed reference strains in a manner similar to that of Piphillin (8). We weighted the amount of

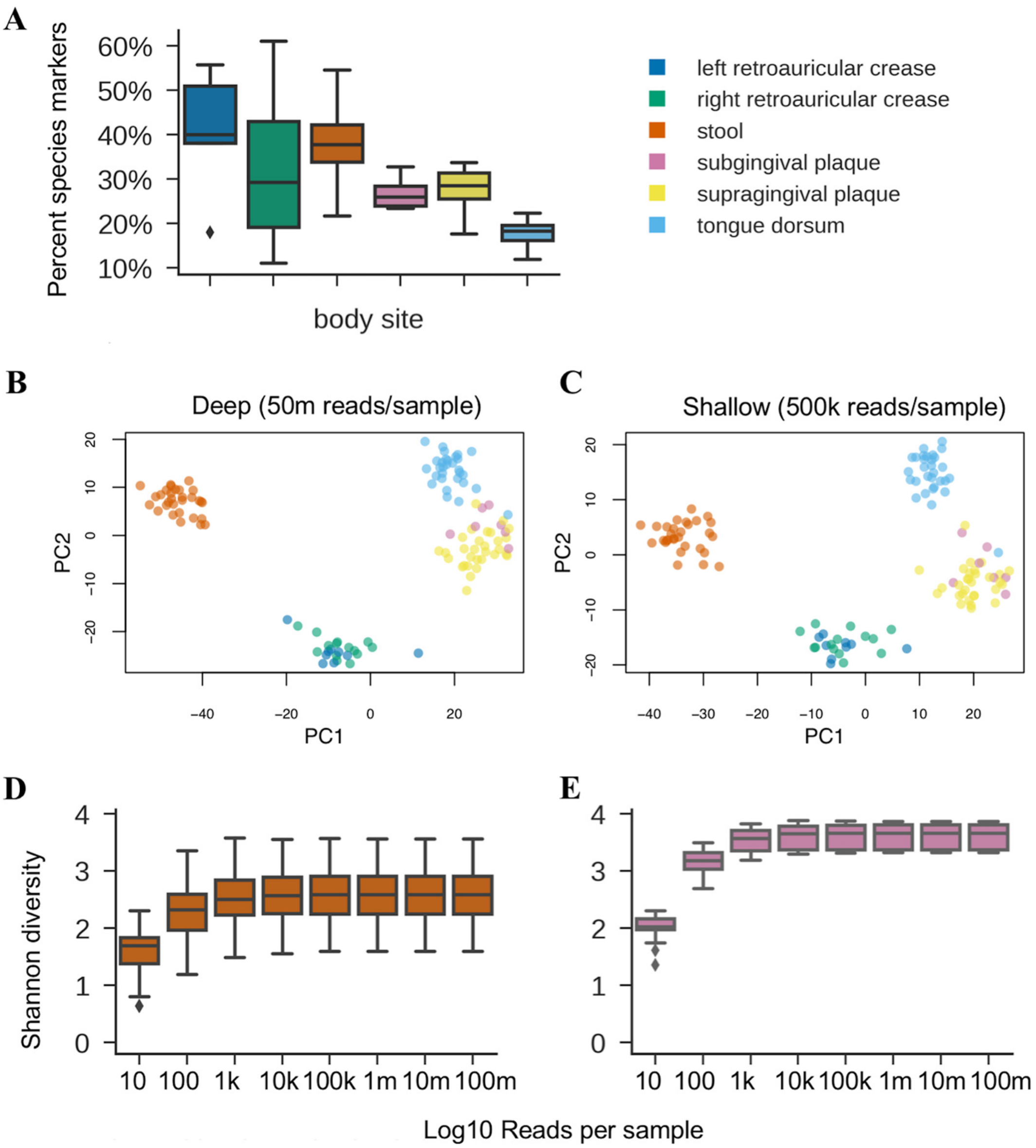

FIG 1 Information content of deep and shallow shotgun sequencing. (A) Percentages of raw shotgun DNA sequences that are unique to one bacterial species across different human body habitats (7 distinct plaque samples, 30 distinct samples from other body sites). (B, C) Principal-coordinate analysis of Bray-Curtis beta diversity using deep (B) and shallow (C) sequencing (sample sizes were as described for panel A). (D, E) Shannon diversity estimates at varied sequencing depths for human stool (D) and subgingival plaque microbiomes (E) (sample sizes were as described for panel B). Boxplots show minimums, first quartiles, medians, second quartiles, and maximums, with outliers beyond 1.5 times the interquartile range plotted individually.

augmentation according to the coefficient of variation of a binomial distribution at the given observed proportion for a given gene, with the augmented gene counts contributing at most 10% of the total counts for a given gene. In practice, this approach affects only the rarest genes with fewer than approximately 10 direct observations and offers a slight improvement in accuracy in Spearman correlation with virtually no change to Pearson correlation (see Materials and Methods). Our observed 97.1% Spearman correlation of the shallow and deep functional profiles is substantially higher than the correlations of functional profiles predicted from 16S sequencing, which typically has 80% to 90% correlation with the directly observed functions (7, 8). We note that we are evaluating the correspondence of the gene profiles using the set of known genes present in the reference database. We did not assemble genes and genomic

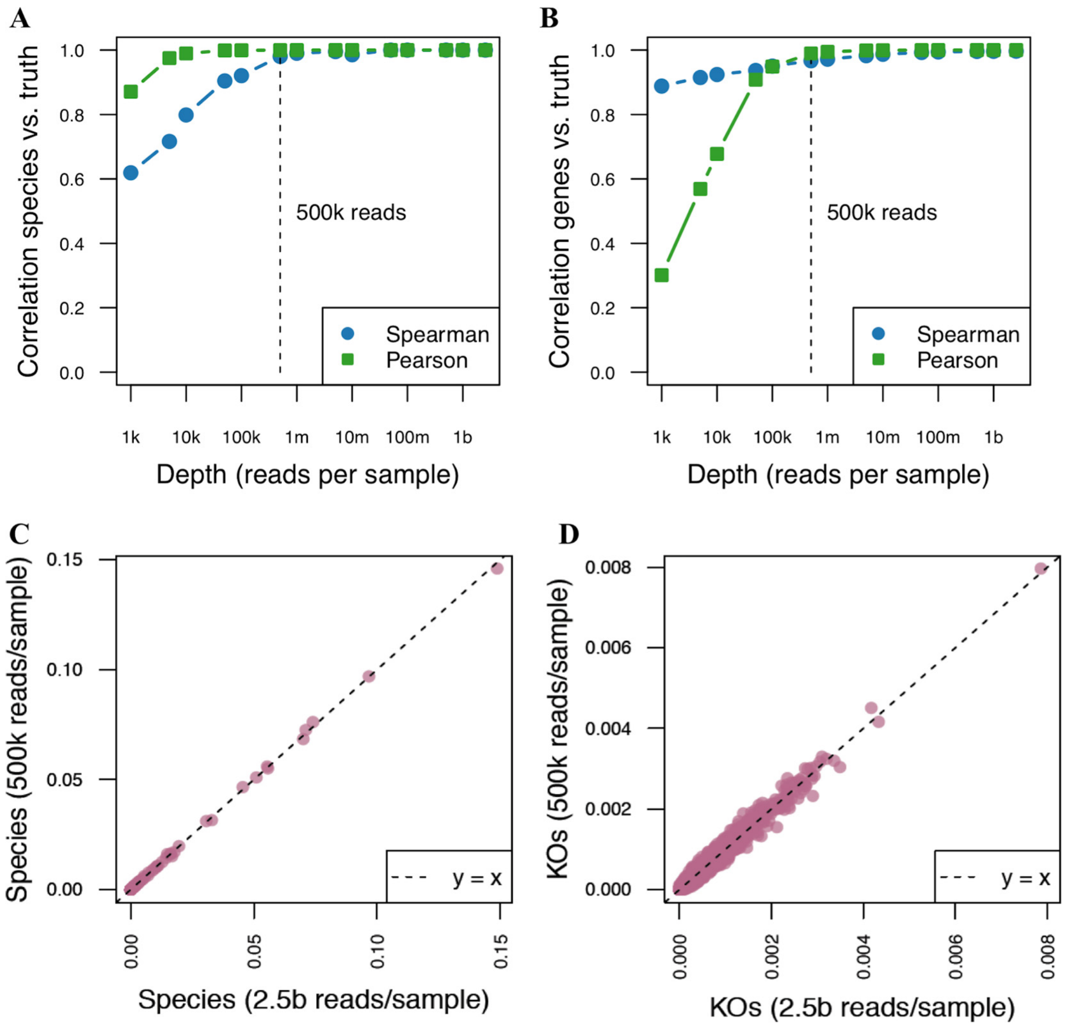

FIG 2 Comparison of species and function profiles with ultradeep sequencing data. (A, B) Correlation with ground truth species (A) and KEGG Orthology group (KO) (B) profile for known genes present in the reference database at different sequencing depths, showing that as few as 0.5 million sequences recover nearly the full species and function profiles (ground truth based on 2.5 billion reads per sample; 4,394 genes and 694 species were used at each subsampling level from the subject 1 ultradeep sequencing sample; comparable results from subject 2 are not shown). Gene and species profiles recovered from the ultradeep data include only direct matches to genes and genomes present in the database; *de novo* assembly of novel genes and contigs from deep data are expected to yield additional uncharacterized gene content and is not possible with shallow shotgun data. (C, D) Scatterplots of species (C) and KOs (D) at 0.5 million versus 2.5 billion reads per sample (we used the same sample size as used for panel A and B above).

regions *de novo* from the deep shotgun data for comparison, as the lack of an ability to assemble new genes and genomes is a fundamental limitation of shallow shotgun sequencing (see Discussion).

Following the comparison of shallow shotgun sequencing to ultradeep sequencing of real biological samples, we also simulated deep WMS sequencing of complex metagenomes from a reference database to evaluate precision and recall of shotgun sequencing at different depths. Individual sequences were drawn at random from full reference genomes of selected species, with a simulated 5% rate of sequencing error. Three different mixtures of species were selected from the database to match the average species-level composition of HMP samples from stool, oral, and skin body sites, respectively (see Materials and Methods). Precision was defined as the fraction of simulated reads that were correctly assigned to their respective species divided by the total number of reads that mapped to the database. Recall was defined as the fraction of simulated reads that were correctly assigned to their respective species divided by

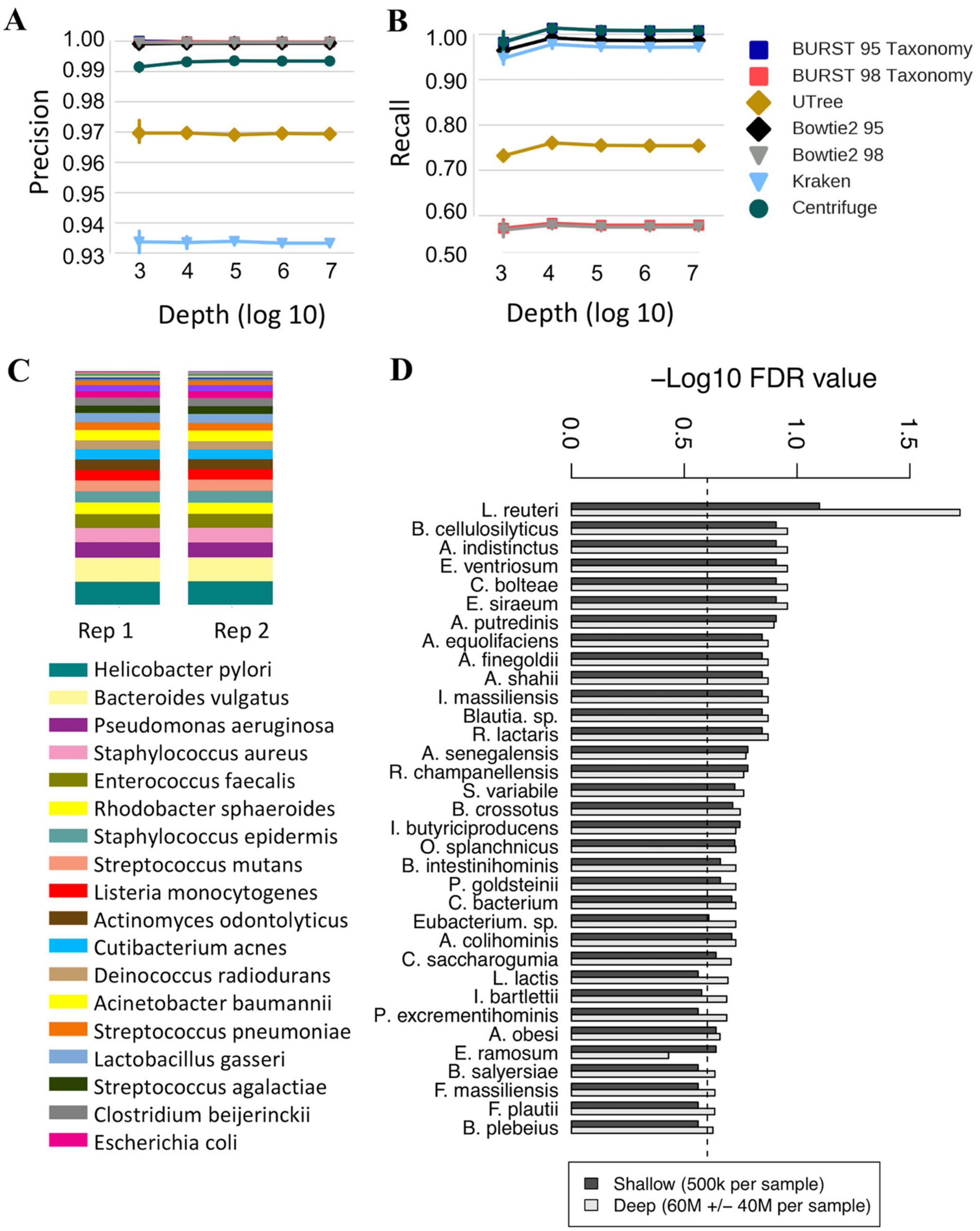

FIG 3 Biomarker discovery using shallow shotgun sequencing. (A, B) Precision, recall for per-read species binning of different metagenomics analysis tools ("95" and "98" refer to the minimum alignment identity threshold used; 5 distinct replicates [rep] were performed per subsampling depth, and error bars show standard deviations). (C) Stacked bar plot of species abundances recovered from HMP mock community shotgun sequencing data. (D) Negative $\log_{10}$ false-discovery rate (FDR)-corrected *P* values using Mann-Whitney U tests for species associated with type 2 diabetes (17), compared between deep and shallow shotgun sequencing (43 healthy patients, 53 patients with type 2 diabetes).

the total number of simulated reads. We found similarly high precision rates of 0.985 to 0.995 when using exhaustive gapped alignment, Bowtie2 (23) at 95% or 98% alignment identity, or Centrifuge (24); k-mer-based methods, including Kraken (21) and an in-house method for comparison (25), had lower precision (Fig. 3A). Recall was considerably higher with 95% identity than with 98% identity alignment with exhaustive gapped alignment or Bowtie2, likely due to the high error rate in the simulated data (Fig. 3B). We also analyzed published shotgun data from the HMP mock community (12), recovering all expected species perfectly as the top 20 taxa, with the

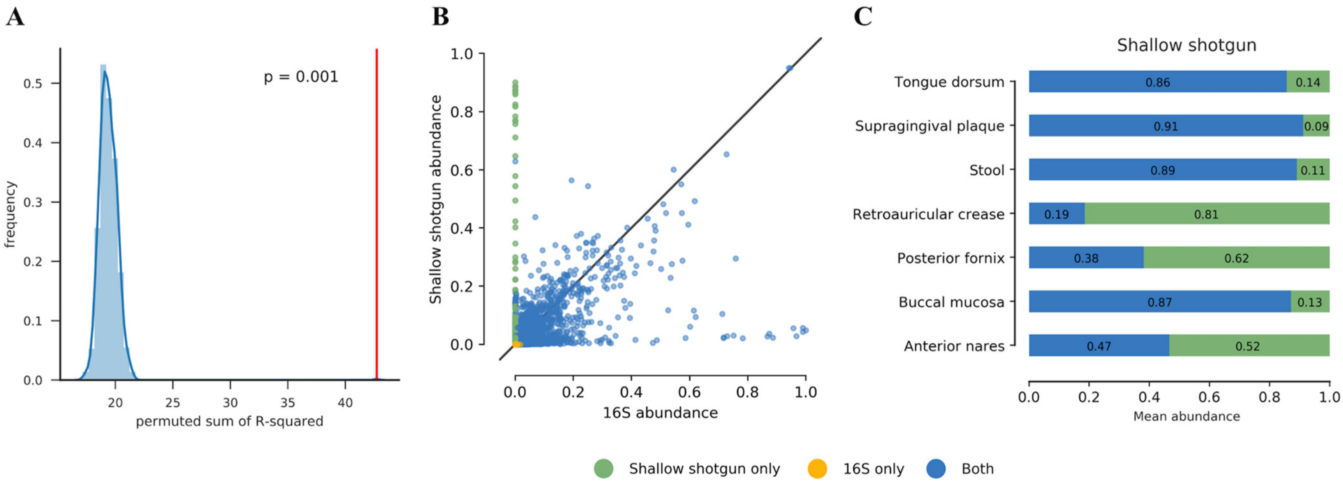

FIG 4 Comparison of 16S sequencing and shallow shotgun recovery of species-level taxa. (A) Histogram of average Pearson correlation (*R*-squared) of species profiles between 16S sequencing and shallow shotgun sequencing from the same HMP sample (*R*-squared = 0.918), compared to the permutation-based null distribution of *R*-squared values for random pairings ($P < 0.001$). (B) Scatterplot of relative abundances of species in shallow shotgun sequencing versus 16S sequencing from the same HMP samples. Species found in only one data type are shown in a different color. (C) Fractions of all observed species, with relative abundance accounted for by species found by 16S sequencing only, shallow shotgun sequencing only, or both.

exception of *Bacillus cereus* which was recovered at the genus level due to highly overlapping species genomes in the genus *Bacillus* (26) (Fig. 3C).

Species-level biomarker discovery. To assess the ability of shallow shotgun sequencing to identify species-level biomarkers in a clinical study, we subsampled deep shotgun sequencing data from a study of stool microbiomes from healthy individuals and individuals with type 2 diabetes (T2D) (17) to 0.5 million sequences per sample. We identified the species significantly associated with T2D in both the deep data and the shallow data using two-sided Mann-Whitney U tests and found high concordance between the *P* values for species down to a 0.0005 relative abundance (rank correlation of lists of differentially abundant species were ordered by raw *P* value; rho = 0.954 across 10 subsampled replicates; $n = 94$; $P < 2 \times 10^{-16}$), indicating that 0.5 million sequences per sample enables discovery of species-level biomarkers with power comparable to deep that of shotgun sequencing down to a relative abundance of approximately 0.0005 (Fig. 3D). Notably, this classification task contained a range of statistical signals ranging from very strong to marginally significant.

Comparison to 16S sequencing species profiles. As noted, 16S sequencing variable-region amplicon sequences often do not resolve taxa below the genus or family level, although some species can be differentiated with 16S sequencing (6). To compare the overall concordance between 16S sequencing species profiles and shallow shotgun sequencing species profiles for pairs from the same sample, we calculated the Pearson correlation *R*-squared value (coefficient of determination) of the 16S sequencing and shallow shotgun sequencing species profiles in paired samples from the HMP (see Materials and Methods; Table S1) and found that the average *R*-squared was 0.918. This demonstrated high overall concordance between 16S sequencing and shallow shotgun sequencing species profiles within a subject. We then permuted the pairing of the 16S sequencing and shallow shotgun sequencing profiles and repeated the average *R*-squared calculation to obtain a null distribution, showing that the *R*-squared value between the true pairs of samples was better in all cases than the randomly assigned pairs (Monte Carlo permutation test, $P < 0.001$) (Fig. 4A). To compare the contributions to total relative abundance of observed species between 16S sequencing and shallow shotgun sequencing profiles, we merged the species-level taxonomic profiles for paired 16S sequencing and shallow shotgun sequencing analyses and measured the fraction of species attributed to 16S sequencing only, shallow shotgun sequencing only, or both. We found that there were many species observed only in the shallow shotgun sequencing data, with some observed at high levels of abundance (Fig. 4B and C), indicating that the 16S sequencing identified a subset of the dominant taxa at the species level. A higher concordance was observed between genus-level profiles from 16S sequencing and shallow shotgun sequencing than between species-level profiles from 16S sequencing and shallow shotgun sequencing (Fig. S1). We confirmed that these results were similar when we compared the 16S sequencing data to deep shotgun metagenomics of the HMP data (Fig. S1).

DISCUSSION

In this work, we evaluated the information content of shallow shotgun sequencing as a potential alternative to 16S sequencing in certain situations. We found that surprisingly few shotgun metagenomic sequences are needed to obtain more reliable species and gene group profiles than 16S sequencing at approximately the same cost as 16S sequencing. We also compared shallow shotgun sequencing to deep shotgun sequencing of a number of biological data sets, including samples from the HMP, a published deep shotgun diabetes study, and simulated and mock communities, and found that we could recover similar trends in alpha and beta diversity, species profiles, and species biomarker discovery down to a 0.05% relative abundance with as few as 0.5 million sequences per sample. We then analyzed two human stool samples with new ultradeep shotgun sequencing data at 2.5 billion reads per sample, the deepest sequencing coverage yet obtained of any microbiome to our knowledge. We found that shallow sequencing recovers 97% to 99% of the correlated species and KEGG (16) Orthology group (KO) profiles found by ultradeep sequencing. Although we obtained shallow shotgun sequencing on the data described in this paper by subsampling originally deep shotgun sequencing data, we expect subsampling to provide an unbiased representation of the data that would be observed in actual shallow shotgun sequencing. We also found, using HMP samples with paired 16S sequencing and shotgun data, that shallow shotgun sequencing is superior to 16S sequencing for recovery and annotation of species.

We did not attempt to perform an exhaustive comparison of different sequence annotation tools, as that was outside the scope of our investigation. Instead, we selected several tools representing different approaches to database search for comparison, including exhaustive semiglobal gapped alignment (18, 19), heuristic gapped alignment using the Burrows-Wheeler transformation (23), and k-mer-based searching (21, 25). Using simulated metagenomic data, we found that tools using gapped alignment obtained higher precision and recall than tools using k-mer-based mapping. This result was expected, as k-mer mapping requires exact matches of fixed-size k-mers, whereas gapped alignment allows insertion of gaps at random to maximize overall sequence identity. In our simulated data, fully exhaustive end-to-end gapped alignment with a minimum threshold of 95% identity using an accelerated version of Needleman-Wunsch alignment (18, 19) performed best in terms of recall. Several methods were approximately tied for highest precision. A potential advantage of gapped alignment over k-mer mapping is that current alignment-based tools report the genomic coordinates of each match, allowing estimation of strain-level coverage, which may be useful for future work into novel algorithms that use strain-level coverage to further improve precision and recall for rare species.

We note a number of important limitations to shallow shotgun sequencing. Shallow shotgun sequencing, as with deep shotgun sequencing, may not be a viable replacement for 16S sequencing when characterizing blood or biopsy specimen microbiomes, where there is likely to be more host DNA contamination and relatively low bacterial biomass. Shallow shotgun sequencing does not allow *de novo* assembly of genes and genomes and, thus, relies on whole-genome reference databases and will require expansion of reference genomes to cover novel environments. When analyzing poorly characterized environments, researchers may consider combining 16S sequencing for identification of novel taxonomic groups with shallow shotgun sequencing for functional profiling. We have not attempted to compare deep or shallow shotgun sequencing with 16S sequencing in environments with a low representation of strains in the reference database, such as marine or soil samples. In these cases, it is likely that shallow shotgun sequencing will still reveal useful functional profiles due to the homology of some observed sequences to known genes and species, but we expect 16S sequencing to provide superior profiling of novel taxa due to the lack of available representative genomes covering endemic species, as has been observed for freshwater samples (27).

Shallow shotgun sequencing is not meant to be a replacement for deep WMS sequencing for strain-level resolution or tracking polymorphisms in strains. Shallow shotgun sequencing cannot be used for novel gene and genome assembly. Our analysis has compared the performance of shallow shotgun sequencing metagenomics to those of 16S sequencing and deep shotgun metagenomics for known genes and genomes only; *de novo* assembly of genes and genomes directly from deep shotgun metagenomic data is likely to reveal novel gene content not readily identifiable from shallow shotgun sequencing data. For many of the metrics that we examined, a depth of 0.5 million sequences per sample was sufficient, but deeper sequencing is warranted for detection of rare species below a relative abundance of approximately 0.0005. We chose 0.5 million for evaluation here because it is the highest depth at which the sequencing cost is still less than approximately half of the total cost of generating data, and yet it is a depth at which one still obtains reasonable sensitivity for species-level recovery; however, we do encourage readers to increase depth toward 1 million or 2 million reads when budget allows, as this will continue to increase sensitivity for detecting rare species. Still, we found that depths of as low as 1,000 reads per sample may be sufficient for some purposes, such as assessing differences in alpha diversity and beta diversity. Therefore, the minimum depth required is dependent on the experimental hypothesis under consideration. In addition, a general concern with any taxonomic annotation from shotgun metagenomic data is that the boundaries of traditional species taxonomic labels do not necessarily reflect consistent entities at the genomic level when accounting for horizontal gene transfer and inaccurate database annotations. These concerns can be alleviated to some extent using deep shotgun sequencing and metagenomic assembly (10), coabundance clustering (11), or proximity-based assembly (28), and *de novo* identification of strains from complex metagenomes remains an active area of research.

We found that shallow sequencing of human stool microbiomes provides high-quality species and functional profiles of human microbiome samples for little more than the cost of 16S amplicon sequencing when a miniaturized library preparation protocol was used (see Materials and Methods). We have made available the gene and genome databases that we used together with a convenient Python-based wrapper script that allows users to compare several existing tools for performing both taxonomic and functional annotation (see Materials and Methods). Shallow shotgun sequencing has a number of important limitations and is not intended to replace deep whole-genome shotgun sequencing for strain-level analysis or novel gene and genome assembly. Nonetheless, shallow shotgun analysis provides considerably more accurate functional profiles and more precise taxonomic resolution than 16S amplicon sequencing for human microbiome studies. Thus, shallow shotgun sequencing is a viable alternative to 16S sequencing for researchers performing large-scale human microbiome studies where deep shotgun sequencing may not be possible.

MATERIALS AND METHODS

Alignment algorithms. Alignment was performed using several existing tools and algorithms, including Bowtie2 (23), Centrifuge (24), Kraken (21), an in-house k-mer-based aligner for comparison with Kraken (25), and an accelerated adaptation of Needleman-Wunsch alignment for exhaustive gapped semiglobal alignment (18, 19).

Shotgun species profiling. After sequences were trimmed until the quality score was above 20 and trimmed sequences shorter than 80 bases or with average quality score of less than 30 were discarded, query reads were mapped with several different alignment tools against representative and reference genomes from the RefSeq database, version 82 (20), using a 95% identity threshold (also compared to 98% for precision and recall evaluation on the simulated data). A read that mapped to a single reference genome is labeled with the NCBI taxonomic annotation. All reads that mapped to multiple reference genomes are labeled as the last common ancestor (LCA) of each label according to the NCBI taxonomy, and only species-level assignments are retained. We use a confidence-adjusted LCA that requires at least 80% of all tied best matches to agree for species annotation. All source codes can be found at the GitHub repository (https://github.com/knights-lab/SHOGUN). Additional analysis code used to generate figures and run tests for this paper can also be found in GitHub (https://github.com/knights-lab/analysis_SHOGUN).

Shotgun functional profiling. Functional profiling was obtained using KEGG Orthology group (KO) (16) annotations for RefSeq-derived genes (20) from directly observed exhaustive gapped alignments in ultradeep WMS sequencing. To improve the accuracy of the direct KO profiles for low-abundance genes, the KO profiles were separately predicted from reference genomes and the predicted profiles were used to augment the estimates of low-abundance KOs. Specifically, we identified those query reads with a 100% match to exactly one reference genome and predicted the entire KO profile of that genome to be present in the sample, which is similar to a previously published approach (8). This is similar to the PICRUSt algorithm for amplicon sequencing data (7) but without the intermediate steps of clustering short-read amplicons and identifying closely related reference genomes. The final KO profiles reported by SHOGUN are a weighted average of predicted and directly observed KO profiles. The predicted KO counts are weighted between 0.0 and 0.1 by a linear function of the coefficient of variation of the count for a given KO, estimated from the size of the binomial confidence interval for the observed count of a given KO divided by the count of that KO. The direct KO profiles receive the remainder of the weight, such that the direct KO profiles receive at least 90% of the weight for all genes and the predicted KO profiles are trusted only for the lowest-abundance genes for which the expected variance in observed count is high. 16S sequences were aligned to Greengenes version 13_8 (29) at 98% identity with exhaustive gapped alignment (18, 19). Where a query sequence aligned equally well to multiple reference sequences, the taxonomic assignment was made using the last common ancestor conserved across at least 80% of the set of references.

Human microbiome project data. We obtained deep WMS sequencing data from the Human Microbiome Project (HMP) (4) and subsampled the data to simulated shallow shotgun sequencing depth. We annotated the deep WMS sequencing data using fully exhaustive gapped alignment for both taxonomy and functional profiles against all complete representative bacterial genomes from the reference database RefSeq version number 82 (20). We then rarefied these samples repeatedly to 1,000, 10,000, 100,000, 1 million, and 10 million sequences per sample and ran the SHOGUN pipeline to quantify species and gene profiles. The list of HMP WMS sequencing and corresponding 16S sequencing samples used are provided in Tables S1 and S2, respectively. The HMP mock community data are from runs SRR2726671 and SRR2726672 from NCBI accession number SRX1342165 (12).

Simulated human metagenomes. The body sites analyzed from the HMP1 project were first grouped according to the broad stool, skin, and oral body sites. We calculated the average relative abundances of all samples within each group. The 100 most abundant species for each group were used for simulating communities. The reads were simulated from a randomly selected strain belonging to each of those most abundant species according to the average proportion of that species in the respective body site group using the tool dWMSim (30). The reads were simulated with default settings for Illumina single-end sequencing machines with a 5% mutation rate where 2% of mutations are indels and with a maximum of 10 ambiguous bases per query sequence.

Sequencing library preparation. Shotgun DNA sequencing was performed on the Illumina HiSeq platform. DNA was extracted using the Qiagen DNeasy PowerSoil kit and was quantified using the Quant-iT PicoGreen dsDNA assay (Thermo Fisher). DNA sequencing libraries were prepared using one-quarter-scale NexteraXT reactions (Illumina). The resulting DNA libraries were denatured with NaOH, diluted to 8 pM in Illumina's HT1 buffer, and spiked with 1% PhiX, and a HiSeq 1× 100-cycle v3 kit (Illumina) was used to sequence samples. For the ultradeep shotgun sequencing, 64 separate libraries were prepared as described above but using full Nextera reaction mixtures from a homogenized stool sample and were multiplexed on a HiSeq 3000 high-output run, using an entire run per sample.

Ethics statement. Volunteers contributing stool samples for the two ultradeep WMS sequencing analyses were recruited as part of research protocol number 150275 approved by the University of California San Diego Institutional Review Board. Research was conducted in accordance with the Helsinki Declaration. Informed consent was obtained from all subjects recruited into the study.

Data availability. The data for the ultradeep WMS sequencing have been deposited in the European Nucleotide Archive with the accession code PRJEB24152.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/mSystems.00069-18.

FIG S1, PDF file, 0.1 MB.

TABLE S1, TXT file, 0.02 MB.

TABLE S2, TXT file, 0.4 MB.

ACKNOWLEDGMENTS

We are grateful to the University of Minnesota Genomics Center for assistance with implementing shallow shotgun DNA sequencing.

The research leading to these results has received funding from NIH grant R01AI121383 (to D.K.), grant P01DK078669, and in-kind support from Illumina (to R.K.).

We declare a conflict of interest. D. Knights, K. B. Beckman, and D. M. Gohl serve as CEO, COO, and CSO, respectively, and hold equity in CoreBiome, a company involved in the commercialization of microbiome analysis. The University of Minnesota also has financial interests in CoreBiome under the terms of a license agreement with CoreBiome. These interests have been reviewed and managed by the University of Minnesota in accordance with its conflict-of-interest policies.

B.H., G.A.A., and D.K. designed the algorithms. B.H., G.A.A., D.K., R.R.S-C., and Q.Z. performed data analysis. D.M.G. and K.B.B. designed the wet-lab methods. B.H., D.K., and R.K. wrote the manuscript with contributions from G.A.A., R.R.S.-C., D.M.G., and K.B.B.

REFERENCES

1. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, Navas-Molina JA, Janssen S, Kopylova E, Vázquez-Baeza Y, González A, Morton JT, Mirarab S, Zech Xu Z, Jiang L, Haroon MF, Kanbar J, Zhu Q, Jin Song S, Kosciolek T, Bokulich NA, Lefler J, Brislawn CJ, Humphrey G, Owens SM, Hampton-Marcell J, Berg-Lyons D, McKenzie V, Fierer N, Fuhrman JA, Clauset A, Stevens RL, Shade A, Pollard KS, Goodwin KD, Jansson JK, Gilbert JA, Knight R, Earth Microbiome Project Consortium. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551:457–463. https://doi.org/10.1038/nature24621.]
2. Singh BK, Bardgett RD, Smith P, Reay DS. 2010. Microorganisms and climate change: terrestrial feedbacks and mitigation options. Nat Rev Microbiol 8:779–790. https://doi.org/10.1038/nrmicro2439.
3. Fitzpatrick CR, Copeland J, Wang PW, Guttman DS, Kotanen PM, Johnson MTJ. 2018. Assembly and ecological function of the root microbiome across angiosperm plant species. Proc Natl Acad Sci U S A 115: E1157–E1165. https://doi.org/10.1073/pnas.1717617115.
4. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger J, Chinwalla A, Creasy H, Earl A, FitzGerald M, Fulton R, Giglio M, Hallsworth-Pepin K, Lobos E, Madupu R, Magrini V, Martin J, Mitreva M, Muzny D, Sodergren E, Versalovic J, Wollam A, Worley K, Wortman J, Young S, Zeng Q, Aagaard K, Abolude O, Allen-Vercoe E, Alm E, Alvarado L, Andersen G, Anderson S, Appelbaum E, Arachchi H, Armitage G, Arze C, Ayvaz T, Baker C, Begg L, Belachew T, Bhonagiri V, Bihan M, Blaser M, Bloom T, Bonazzi V, Brooks J, Buck G, Buhay C, Busam D, Campbell J, et al. 2012. Structure, function and diversity of the healthy human microbiome. Nature 486:207–214. https://doi.org/10.1038/nature11234.
5. Knights D, Costello EK, Knight R. 2011. Supervised classification of human microbiota. FEMS Microbiol Rev 35:343–359. https://doi.org/10.1111/j.1574-6976.2010.00251.x.
6. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, Whitman WB, Euzéby J, Amann R, Rosselló-Móra R. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol 12:635–645. https://doi.org/10.1038/nrmicro3330.
7. Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R, Beiko RG, Huttenhower C. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol 31: 814–821. https://doi.org/10.1038/nbt.2676.
8. Iwai S, Weinmaier T, Schmidt BL, Albertson DG, Poloso NJ, Dabbagh K, DeSantis TZ. 2016. Piphillin: improved prediction of metagenomic content by direct inference from human microbiomes. PLoS One 11: e0166104. https://doi.org/10.1371/journal.pone.0166104.
9. Luo C, Knight R, Siljander H, Knip M, Xavier RJ, Gevers D. 2015. ConStrains identifies microbial strains in metagenomic datasets. Nat Biotechnol 33:1045–1052. https://doi.org/10.1038/nbt.3319.
10. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. 2014. Binning metagenomic contigs by coverage and composition. Nat Methods 11:1144–1146. https://doi.org/10.1038/nmeth.3103.
11. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, Plichta DR, Gautier L, Pedersen AG, Le Chatelier E, Pelletier E, Bonde I, Nielsen T, Manichanh C, Arumugam M, Batto J-M, Quintanilha Dos Santos MB, Blom N, Borruel N, Burgdorf KS, Boumezbeur F, Casellas F, Doré J, Dworzynski P, Guarner F, Hansen T, Hildebrand F, Kaas RS, Kennedy S, Kristiansen K, Kultima JR, Léonard P, Levenez F, Lund O, Moumen B, Le Paslier D, Pons N, Pedersen O, Prifti E, Qin J, Raes J, Sørensen S, Tap J, Tims S, Ussery DW, Yamada T, MetaHIT Consortium P, Renault P, Sicheritz-Ponten T, Bork P, Wang J, Brunak S, Ehrlich SD. 2014. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol 32:822–828. https://doi.org/10.1038/nbt.2939.
12. Jones MB, Highlander SK, Anderson EL, Li W, Dayrit M, Klitgord N, Fabani MM, Seguritan V, Green J, Pride DT, Yooseph S, Biggs W, Nelson KE, Venter JC. 2015. Library preparation methodology can influence genomic and functional predictions in human microbiome research. Proc Natl Acad Sci U S A 112:14024–14029. https://doi.org/10.1073/pnas.1519288112.
13. Ni J, Yan Q, Yu Y. 2013. How much metagenomic sequencing is enough to achieve a given goal? Sci Rep 3:1968. https://doi.org/10.1038/srep01968.
14. Jovel J, Patterson J, Wang W, Hotte N, O'Keefe S, Mitchel T, Perry T, Kao D, Mason AL, Madsen KL, Wong GK-S. 2016. Characterization of the gut microbiome using 16S or shotgun metagenomics. Front Microbiol 7:459. https://doi.org/10.3389/fmicb.2016.00459.
15. Rodriguez-R LM, Konstantinidis KT. 2014. Estimating coverage in metagenomic data sets and why it matters. ISME J 8:2349–2351. https://doi.org/10.1038/ismej.2014.76.
16. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. 2012. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 40:D109–D114. https://doi.org/10.1093/nar/gkr988.
17. Karlsson F, Tremaroli V, Nielsen J, Backhed F. 2013. Assessing the human gut microbiota in metabolic diseases. Diabetes 62:3341–3349. https://doi.org/10.2337/db13-0844.
18. Al-Ghalith GA, Knights D. 2017. BURST enables optimal exhaustive DNA alignment for big data. https://github.com/knights-lab/burst.
19. Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453. https://doi.org/10.1016/0022-2836(70)90057-4.
20. Tatusova T, Ciufo S, Fedorov B, O'Neill K, Tolstoy I. 2014. RefSeq microbial genomes database: new representation and annotation strategy. Nucleic Acids Res 42:D553–D559. https://doi.org/10.1093/nar/gkt1274.
21. Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15:R46. https://doi.org/10.1186/gb-2014-15-3-r46.
22. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H. 2013. Package 'vegan.' Community Ecol Packag version 2. https://cran.r-project.org/web/packages/vegan/index.html.
23. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923.
24. Kim D, Song L, Breitwieser FP, Salzberg SL. 2016. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res 26: 1721–1729. https://doi.org/10.1101/gr.210641.116.
25. Al-Ghalith GA, Knights D. 2017. Faster and lower-memory metagenomic profiling with UTree. https://github.com/knights-lab/utree.
26. Radnedge L, Agron PG, Hill KK, Jackson PJ, Ticknor LO, Keim P, Andersen

GL. 2003. Genome differences that distinguish Bacillus anthracis from Bacillus cereus and Bacillus thuringiensis. Appl Environ Microbiol 69: 2755–2764. https://doi.org/10.1128/AEM.69.5.2755-2764.2003.

27. Tessler M, Neumann JS, Afshinnekoo E, Pineda M, Hersch R, Velho LFM, Segovia BT, Lansac-Toha FA, Lemke M, DeSalle R, Mason CE, Brugler MR. 2017. Large-scale differences in microbial biodiversity discovery between 16S sequencing amplicon and shotgun sequencing. Sci Rep 7:6589. https://doi.org/10.1038/s41598-017-06665-3.

28. Burton JN, Liachko I, Dunham MJ, Shendure J. 2014. Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. G3 (Bethesda) 4:1339–1346. https://doi.org/10.1534/g3.114.011825.

29. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J 6:610–618. https://doi.org/10.1038/ismej.2011.139.

30. Homer N. 2010. DWGSIM. https://github.com/nh13/DWGSIM.