Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Biochem Biophys Res Commun. Author manuscript; available in PMC 2017 January 22.

Published in final edited form as:

Biochem Biophys Res Commun. 2016 January 22; 469(4): 967–977. doi:10.1016/j.bbrc.2015.12.083.

Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing

Ravi Ranjan[1,#], **Asha Rani**[1,#], **Ahmed Metwally**[1,2], **Halvor S. McGee**[1], and **David L. Perkins**[1,3]

Ravi Ranjan: ranjan@uic.edu; Asha Rani: asharani@uic.edu; Ahmed Metwally: ametwa2@uic.edu; Halvor S. McGee: hmcgee@uic.edu; David L. Perkins: perkinsd@uic.edu

[1]Department of Medicine, University of Illinois, Chicago, IL 60612 USA

[2]Department of Bioengineering, University of Illinois, Chicago, IL 60612 USA

[3]Department of Surgery, University of Illinois, Chicago, IL 60612 USA

Abstract

The human microbiome has emerged as a major player in regulating human health and disease. Translation studies of the microbiome have the potential to indicate clinical applications such as fecal transplants and probiotics. However, one major issue is accurate identification of microbes constituting the microbiota. Studies of the microbiome have frequently utilized sequencing of the conserved 16S ribosomal RNA (rRNA) gene. We present a comparative study of an alternative approach using shotgun whole genome sequencing (WGS). In the present study, we analyzed the human fecal microbiome compiling a total of 194.1×10^6 reads from a single sample using multiple sequencing methods and platforms. Specifically, after establishing the reproducibility of our methods with extensive multiplexing, we compared: 1) The 16S rRNA amplicon versus the WGS method, 2) the Illumina HiSeq versus MiSeq platforms, 3) the analysis of reads versus *de novo* assembled contigs, and 4) the effect of shorter versus longer reads. Our study demonstrates that shotgun whole genome sequencing has multiple advantages compared with the 16S amplicon method including enhanced detection of bacterial species, increased detection of diversity and increased prediction of genes. In addition, increased length, either due to longer reads or the assembly of contigs, improved the accuracy of species detection.

[*]**Corresponding author:** David Perkins, MD, PhD, University of Illinois at Chicago, Department of Medicine, MC 787, 840 S Wood Street, Suite 1020N CSB, Chicago IL 60612 USA, perkinsd@uic.edu, Phone: 312-413-3382, Fax: 312-355-0499.
[#]These authors contributed equally and considered as co-first authors.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

AUTHOR CONTRIBUTIONS

RR, AR and DLP: designed study; RR: prepared libraries and sequencing, AR and RR performed data analysis, AM: performed bioinformatics pipe line analysis; HSM: assisted with sample collection; RR, AR and DLP wrote the manuscript.

COMPETING FINANCIAL INTERESTS:

The authors have declared that no conflict of interest exists.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Keywords

16S rRNA; Metagenomics; Microbiome; Next-generation sequencing; Amplicon sequencing; Whole genome shotgun sequencing

INTRODUCTION

The human microbiome is important in maintaining health, whereas dysbiosis has been associated with various diseases (e.g., inflammatory bowel disease and coronary artery disease) and conditions (e.g. obesity) [1, 2]. These observations suggest that modulation of the fecal microbiome could become an important therapeutic modality for some diseases. For example, fecal transplants have been shown to alleviate diarrhea caused by *Clostridium difficile* infection and temporarily improve insulin sensitivity [3, 4]. However, a major concern when administering fecal transplants, or even probiotics, is the long-term biological effects of the inoculum on the recipient microbiota. It is essential to precisely identify and enumerate the bacterial species in the inoculum as well as in the recipient microbiome in order to understand the complex interactions among the microbes [5, 6]. The gut microbiome, which has been the most extensively studied of the human microbiomes, is highly diverse and has been shown to include thousands of different bacterial species [7, 8]. The diverse community of bacteria is composed of a small number of abundant species plus a large number of rare or low abundance species [9]. The differential functions of the abundant and rare species remain poorly understood. Thus, to effectively understand the ecology of the fecal microbiome, it is imperative to analyze both the rare and the abundant microbes.

The number of studies investigating the microbiome has exploded since the technological advances in high-throughput sequencing that facilitate culture- and cloning-independent analysis [10]. These technical advances have been paradigm shifting since the majority (>90%) of microbial species cannot be readily cultured using current laboratory culture techniques [11–13]. The most common sequencing approach to analyze the microbiome, which has been used to compile most of the data collated by the Human Microbiome Project (HMP), is amplicon analysis of the 16S ribosomal RNA (rRNA) gene [14, 15]. In this method, a 16S rRNA region is amplified by PCR with primers that recognize highly conserved regions of the gene and sequenced [16]. The limitations of this method are that the annotation is based on putative association of the 16S rRNA gene with a taxa defined as an operational taxonomic unit (OTU). In general, OTUs are analyzed at the phyla or genera level, and can be less precise at the species level. In addition, specific genes are not directly sequenced, but rather predicted based on the OTUs. Due to horizontal gene transfer and the existence of numerous bacterial strains [17–19], the lack of direct gene identification potentially limits understanding of a microbiome.

An alternative approach to the 16S rRNA amplicon sequencing method is whole genome shotgun sequencing (WGS) which uses sequencing with random primers to sequence overlapping regions of a genome. The major advantages of the WGS method are that the taxa can be more accurately defined at the species level. Another important consideration is

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

that the 16S and WGS methods commonly utilize different databases for classification of taxa. However, WGS is more expensive and requires more extensive data analysis [10, 20–22]. In addition, to identify and understand the bacterial genes in a taxa, it may be necessary to sequence a genome with high coverage [20].

In the present study, we analyzed a single human fecal microbiome with a total of 194.1×10^6 reads using multiple methods and platforms. The high number of reads supported an analysis with multiple experimental methods. Specifically, we established the reproducibility of our methods with extensive multiplexing. Also, we investigated four factors of microbiome analysis. We compared: 1) The 16S rRNA amplicon versus the WGS method, 2) the Illumina HiSeq versus MiSeq platforms, 3) the analysis of reads versus de novo assembled contigs, and 4) the effect of shorter versus longer reads. Our results demonstrate important advantages of the WGS sequencing method.

MATERIALS AND METHODS

Subject recruitment and sample collection

Informed consent was obtained from the subject. An adult subject provided self-collected stool. The study was approved by the Institutional Review Board of the University of Illinois at Chicago (Protocol # 2014-0528), and the experimental methods were carried out in accordance with the approved guidelines.

Fecal Metagenomic DNA isolation

A fresh voided stool specimen was processed for total DNA isolation. Approximately 100 mg of stool was transferred to an Eppendorf safe lock tube and processed with a PowerSoil DNA isolation kit (Catalogue # 12888-100, MO BIO Laboratories, Inc) using the manufacture's protocol with slight modifications. For efficient lysis of the microbes, glass beads in 200 μL bead solution plus 200 μL of Phenol/Chloroform/Isoamyl alcohol (25:24:1) pH 7.8–8.2 (Catalogue #327115000, Acros Organics) was added to the sample. The contents were vortexed for 1–2 min and homogenized at speed 10 for 5 min with air-cooling using the Bullet Blender Storm Homogenizer (Catalogue # BBY24M, Next Advance Inc). The contents were centrifuged at 14,000×g, and the lysate was transferred to a sterile tube. The DNA was eluted with 1×TE, pH 8.0, and stored at −80°C. The quality and quantity of the DNA was accessed using a spectrophotometer (NanoPhotometer Pearl, Denville Scientific, Inc), agarose gel electrophoresis, and fluorometer (Qubit® dsDNA High Sensitivity and dsDNA Broad Range assay, Life Technologies Corporation).

Metagenomic Library preparation

For preparation of the libraries for Illumina MiSeq and HiSeq 2000 DNA sequencers, the fecal metagenomic DNA libraries were prepared in two separate batches. For sequencing on MiSeq, approximately 5 μg metagenomic DNA was mechanically sheared to 300 – 600 bp fragments using a Covaris S220 instrument (Covaris, Inc) with the following parameters: temperature 7– 9°C, peak incident power 140 W, duty factor 10%, cycles per burst 200, time 80 s, sample volume 130 μL, and shearing tubes Crimp-cap microTubes with AFA fiber. In total 20 μg of metagenomic DNA was sheared to prepare multiple libraries. The fragmented

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

DNA was analyzed with an Agilent DNA 12000 Kit on the 2100 Bioanalyzer Instrument (Agilent Technologies, Inc). The sheared DNA was processed with a QIAquick PCR purification kit (Qiagen) and eluted in nuclease free water. One microgram fragmented metagenomic DNA was end-repaired and 3′-adenylated, ligated with Illumina adapters, and PCR enriched with Illumina sequencing indexes (barcodes) using the NEBNext Ultra DNA library prep kit for Illumina (Catalogue # E7370L,New England BioLabs Inc). In total, 11 libraries were constructed with unique indexes using the NEBNext Multiplex Oligos for Illumina Set 1 (Catalogue # E7335L, NewEngland BioLabs Inc). For sequencing on a HiSeq 2000 the library was prepared using the same fecal metagenomic DNA using the described procedures. The quality and quantity of all the DNA libraries were analyzed with an Agilent DNA 12000 Kit on the 2100 Bioanalyzer Instrument and Qubit.

Sequencing strategy

The DNA libraries were sequenced using the Illumina MiSeq and HiSeq 2000 platforms. The first batch of eleven DNA libraries was diluted and pooled with an equimolar concentration of each library. The pooled libraries were sequenced following manufacturer's protocol by multiplexing on our MiSeq using the MiSeq® Reagent Kit v2 (300 cycle) and MiSeq® Reagent Kit v3 (600 cycle) for paired-end 151 bases and 301 bases, respectively. The second batch of DNA libraries was sequenced on a HiSeq 2000 using the TruSeq SBS v3 reagent for 100 base paired runs by BGI Americas. Sequencing runs on both platforms included an additional 6 cycles for the 6 base-index.

16S rRNA amplicon library preparation and sequencing

The amplicon libraries were prepared using the NEXTflex 16S V1–V3 Amplicon-Seq kit (Catalog # 4202-02, Bio Scientific Corp) following the manufactures protocol with slight modifications. Metagenomic DNA (50 ng) was used as template for the first PCR (PCR-I) amplification, which amplifies the 16S rRNA region, with the following conditions (initial denaturation at 95°C for 5 min, 10 cycles of denaturation at 95°C for 30 s, annealing at 55°C for 30 s, extension at 72°C for 30 s plus a final extension at 72°C for 5 min). The PCR product was processed using Agencourt AMPure XP Beads (Beckman Coulter, Inc.) and eluted with nuclease free water. Four μL was used as template for the second PCR (PCR-II), which adds the adapter sequences. PCR-II was performed using the following conditions (initial denaturation at 95°C for 5 min, 12 cycles of denaturation at 95°C for 30 s, annealing at 60°C for 30 s, annealing at 72°C at 30 s, plus a final extension at 72°C for 5 min). The final product was processed using the Agencourt AMPure XP Beads and library was analyzed using the Bioanalyzer DNA 1000 (Agilent Technologies). The final amplicon libraries were sequenced on a MiSeq using the v3–600 cycle kit with 301 base paired end chemistry. To check for possible exogenous DNA contamination, samples of TE and water used for the PCR-I and PCR-II step were analyzed on a Bioanalyzer DNA 1000 and Qubit (High Sensitivity Assay) with no detectable DNA. Also, the control libraries generated no detectable sequences.

Data analysis

The sequence reads generated by the 16S rRNA and WGS sequencing methods on MiSeq and HiSeq 2000 sequencers were processed on CLC Genomics workbench (CLC bio). For

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the 16S rRNA amplicon analysis the individual library read files were paired. In addition all the eleven read files for the l6S libraries were combined *in silico* (16S–total) and analyzed for taxonomic annotation using the RDP tool on MG-RAST [23]. For the WGS libraries individual library reads sequenced by v2–300, v3–600 and HiSeq 2000 were paired. In addition, all the reads from v2–300, v3–600, and HiSeq 2000 were combined *in silico* using CLC Genomics workbench, and designated as v2-total, v3-total, v2+v3-total, v2+v3+HiSeq-total. All the files were analyzed for species identification using MG-RAST. For the contig based analysis, the reads of individual libraries (v2, v3, HiSeq) were paired and *de novo* assembled into contigs using *de-novo* assembly algorithm (CLC Genomics workbench). Additionally, the read files of v2-total, v3-total, v2+v3-total, v2+v3+HiSeq-total were combined and *de novo* assembled into contigs. The contigs of the individual libraries and combined reads were annotated for phylogenetic and functional analysis using the automated annotation pipeline at MG-RAST.

Analysis of genome coverage of high, low and rare abundant species

Based on the percentage relative abundance of species, we classified abundance as common (>1.00%), high (>0.50–0.99%), moderate (0.05–0.49%), low (0.01 – 0.49%) and rare (<0.01%). The genome sequences of representative species were downloaded from the NCBI Genome browser. The reads were aligned to the reference genomes using the CLC Genomics workbench, and the percentage of genome coverage was calculated using the formula [total number of bases aligned/genome size (bases) × 100].

Data analysis pipeline

We created a bioinformatics framework pipeline to classify metagenomic sequences at the phylogenetic and functional levels (Fig. S1). The input to the pipeline was raw sequencing reads in fastq format. First, the reads were filtered using the FastX tool (to remove the reads that have a length less than 75 nt for MiSeq v2–300, 151 PE; 150 nt for MiSeq v3–600, 301 PE; and 40 nt for HiSeq 100 PE) or quality for Phred score less than 20. All reads that aligned to the human genome were filtered. Next, reads that had less than 90% identity or *e*-value of $1e^{-3}$ were removed. The filtered high quality reads were assembled into contigs using the MetaVelvet assembler [24]. All parameters were set to default except the minimum contig length of 500 bp and *k-mer* size of 31. A similarity method was utilized to assign each contig to a species, and the assembled contigs were then queried against NCBI NT database. A stringent threshold of 90% identity and *e*-value of $1e^{-3}$ was used to evaluate the alignments, and the best hit was selected. Using blast output, the implemented pipeline returned the relative abundance at each taxonomic level, i.e. kingdom, phylum, class, order, family, genus and species.

Analysis of unique genes

The assembled contigs from the combined reads (v2-total, v3-total, v2+v3-total, v2+v3+HiSeq-total), were used to predict putative genes using the MetaGeneMark [25] tool (Fig. S1). The predicted ORFs were in two formats: nucleotide and amino acid. To identify the unique genes, we utilized the CD-HIT tool to remove shorter ORFs that aligned with another ORF with more than 95% identity and had query coverage of greater than 90 bp of the shorter ORF. This criteria is the same as the parameters used in the MetaHIT project. We

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

aligned the predicted unique ORFs to the current updated gut microbial gene catalog which includes 9.8×10^6 genes [26].

Determining the number of reads to identify abundant species

The v3-total read file (containing ~59 million reads) was randomly sampled for 250, 500, 750, 1,000, 2,000, 3,000, 5,000 and 10,000 reads. Each sample read file was then matched against the NCBI-nt database using our pipeline (Fig. S1).

Sequence Datasets

The datasets supporting the results of this article are available in the MG-RAST repository (http://metagenomics.anl.gov/). The accession numbers are listed in Table S10.

RESULTS

Analysis of amplicon (16S rRNA) versus whole genome shotgun sequencing

In this study, we performed extremely deep sequencing of a fecal sample using different sequencing methods (16S and WGS metagenomic sequencing), sequencing platforms (HiSeq and MiSeq), analysis strategies (reads and *de novo* assembled contigs) and read length (100, 150 and 300 bp) to rigorously determine the optimal methods for microbiome analysis (Fig. 1). A freshly voided sample was processed and high quality (greater than 5 kb) metagenomic DNA was isolated (Fig. S2a,b). To evaluate reproducibility of our methods, the libraries were constructed with 11 fold multiplexing. We prepared 16S rRNA amplicon libraries from the metagenomic DNA using 16S rRNA gene V1–V3 region primers listed by the Human Microbiome Project [27]. The Bioanalyzer tracings of the 11 libraries showed a peak at 650–700 bp without any primer-dimers or detectable adapter contamination confirming high quality of the amplicon libraries (Fig. S2c). To exclude the possibility of contamination of reagents, a control library without metagenomic DNA inserts was prepared following the same methods and no PCR product or DNA inserts were detected (S3a,b). To prepare WGS libraries for the v2–300 and v3–600 MiSeq sequencing, the fecal metagenomic DNA (Fig. S2a,b) was sheared to 300 – 600 bp fragments using Covaris S220 (Fig. S4a). Eleven high quality libraries (S1 – S11) with unique indexes for multiplexed sequencing were prepared (Fig. S4b). Next, to prepare a library for sequencing on HiSeq, the same fecal metagenomic DNA was sheared (Fig. S5a) and a library was prepared. The Bioanalyzer tracing of the library showed a peak at 280–300 bp also without primer-dimers or detectable adapter contamination (Fig. S5b). Taken together, our quality control assessments indicate high quality DNA and metagenomic libraries.

To compare sequencing protocols, we performed 4 sequencing reactions of the libraries utilizing different methods (16S versus WGS), different platforms (HiSeq versus MiSeq), different read lengths (100, 150 and 300 bp paired-end reads) and different data analysis strategies [amplicon, read or contig based analysis] (Table 1). First, with the 16S approach we identified 30.4×10^6 reads totaling 9.1 Gb of sequence. Second, with the HiSeq 2000 using a WGS library sequenced with a TruSeq SBS v3 reagent we generated 67.2×10^6 reads that comprised 6.7 Gb of sequence. Third, with the MiSeq using a library sequenced with a v2–300 kit, we generated a total of 37.5×10^6 reads with 151 read length that totaled 5.8 Gb

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

of sequence. And fourth, with the MiSeq using a library sequenced with a v3–600 we generated a total of 59.0×10^6 reads that totaled 15.6 Gb of sequence. In total we identified 194.1×10^6 reads comprising 39.6 Gb of sequence (Table 1).

We analyzed the multiplexed results of a single fecal sample to establish the technical reproducibility of our methods, which included our analyses of abundance, taxonomy and diversity (Tables S1 and S2). After establishing the reproducibility of our methods, we combined the multiplexed indexes *in silico* to construct the v2-total (37.5×10^6 reads) and v3-total (59.0×10^6 reads) data sets. Next, we combined the v2-total plus v3-total read sequence data sets to form the v2+v3-total (96.5×10^6 reads) data file. And finally, we combined the v2+v3-total with the HiSeq data to construct the v2+v3+HiSeq-total (163.7×10^6 reads) dataset (Table S2). Overall, the combination of the multiplexed data plus the high number of sequence reads facilitated an in depth comparative analysis of the sequencing platforms and methods with our data analysis pipeline (Fig. S1).

Analysis of species abundance

To evaluate the capacity of the different platforms and methods to detect bacterial species, we constructed rarefaction curves for the reads of the 16S, HiSeq, v2-total, v3-total, v2+v3-total and v2+v3+HiSeq-total datasets (Fig. 2a). The striking result was that the species abundance detected by the 16S method was markedly lower than by all of the WGS protocols. At 32.8 million reads, 16S detected 1,800 compared to more than 3,000 species by the WGS method. The different libraries and platforms analyzed with the WGS methods detected similar numbers of species per sequence read; however, the slope of the abundance curve remained positive indicating that additional species were being detected even at 163.7×10^6 total WGS reads. In the combined data in the v2+v3+HiSeq-total dataset, a total of 5,870 unique species were detected. Interestingly, the 21 common species (defined as relative abundance >1%) comprised 34.0% of the cumulative abundance, whereas the 1,222 rare species (defined as relative abundance <0.01%) comprised only 3.3% of the cumulative abundance (Table 2). Thus, we detected a large number of unique species in a single fecal microbiome. In parallel, we asked what is the minimum number of reads necessary to identify abundant species in the sample. We randomly sampled low numbers of reads from the v3-total dataset and constructed an abundance table (Table 3). Interestingly, 100% of the abundant species can be detected with as few as 500 sequence reads. The rarefaction curves for the multiplexed samples showed similar and reproducible results for the individual indexes (Fig. S6a–e and Tables S3, S4, S5). Thus, our data shows that the microbiome contains a small number of abundant species, but a large number of low abundance and rare species.

After performing *de novo* assembly of the WGS metagenome datasets with the CLC Genomics Workbench, we reanalyzed the rarefaction curves for the WGS datasets (Fig. 2b). Interestingly, in this analysis of contigs both the v2-total and v3-total datasets generated with the MiSeq outperformed the HiSeq dataset, In our study both the MiSeq and HiSeq data were paired end reads, but the MiSeq data generated longer reads (150 and 300 versus 100 bp) which may have improved the efficiency of the *de novo* assembly of the contigs, and thus, increased species detection in the MiSeq data.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Analysis of bacterial taxonomy

To investigate the complexity of the microbial community, we determined the relative abundance at the phyla level in our datasets based on the analysis of 16S rRNA amplicons, reads and contigs (Fig. 3a–c). The relative abundance of the major phyla (Firmicutes, Bacteroidetes, Actinobacteria and Proteobacteria) was similar in all datasets sequenced with the WGS methods. However, we did detect significant differences between the 16S rRNA amplicon and WGS results. Specifically, in the 16S analysis we detected an increase in Bacteroidetes to 34% from 14–21% and a decrease in Actinobacteria to 0.4% from 4–7%. The relative abundance of Firmicutes and Proteobacteria was similar. When we analyzed the abundance of phyla in the multiplexed data, we obtained similar and reproducible results for the individual indexes (Fig. S7a–c). To compare the detection of phyla by the 16S rRNA amplicon versus WGS reads, we determined the number of species in each phyla identified by both methods or by only one of the methods (Fig. 4a and Table S6). In summary, both methods detected 70.8% of the Actinobacteria species, but only 8.9% of the Proteobacteria species. Among the Proteobacteria, 1,056 species were detected only by WGS showing increased sensitivity of species detection by WGS.

To drill down on the differences observed at the phyla level, we next analyzed the detection of individual bacterial species by the different methods. First, we analyzed the common species (>1% abundance) detected by 16S rRNA amplicon and by BLAST analysis of reads and contigs (Table 4 and Table S6). The Pearson's correlation coefficient between the abundance determined by 16S rRNA amplicon analysis versus WGS (the v2+v3+HiSeq-total dataset) was 0.6, whereas the correlation among the WGS datasets was >0.9 for all comparisons (Table S7a,b). These results illustrate the similarity among the WGS approaches, and the differences between the 16S and WGS methods.

Similar correlations were detected in an analysis of either reads or contigs of the WGS data. The lower correlation between the 16S and WGS methods was in part due to differences between the classifications of OTUs in the RDP database and species in the NCBI-nt database. First, we suggest that the 16S rRNA method is less sensitive for species detection as shown in the rarefaction plots. Second, the NCBI-nt database includes a greater number of bacterial species. Although some species showed comparable abundance between the 16S and WGS methods, other species (e.g., *Faecalibacterium prausnitzii*) were more abundant in the WGS results (17.0–23.5 versus 7.1), whereas other species were more abundant in the 16S results [e.g., *Bacteroides uniformis* (1.0 versus 3.9) and *Bacteroides stercoris* (1.0 versus 3.8)]. For some sequences, we detected different nomenclature in the RDP versus the 3 NCBI databases (nt, representative genome database and 16S rRNA databases). For example, several species (e.g., *Subdoligranulum variabile, Clostridium saccharolyticum, Bacteroides* sp. 1_1_6, *Ruminococcus* sp. SR1/5, *Clostridium phytofermentans* and *Bacteroides* sp. 4_3_47FAA) were detected by M5NR database but not by 16S analysis of the RDP and NCBI-nt database. In addition, we identified *Blautia* sp. Ser8 in the RDP database; however the same sequence was assigned to uncultured bacterium clone (99%) in NCBI-nt, *Clostridium saccharolyticum* WM1 (93%) in NCBI representative genome database and *Blautia producta* strain JCM (98%) in NCBI 16S ribosomal RNA database. Comparison of a complete list of the unique species identified by the 16S versus WGS

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

methods also showed differences in a read-based versus contig-based analysis (Fig. 4b). First, the read-based analysis detected substantially more unique species (3,875) than the contig-based analysis (2,769). Furthermore, 63% of the species detected by the read-based method, compared with only 48% in the contig-based method, were not detected by the 16S approach.

We also detected substantial quantitative differences in abundance for some species determined by analysis of reads versus contigs. The most striking difference was for *Faecalibacterium prausnitzii* which showed 17.0–23.5% for our WGS read analysis (Table 4), but only 3.6–5.3% for WGS contig analysis (Table S6). Smaller differences were observed for other species. We observed similar results in the multiplexed data (Tables S3, S4, S5). Overall, among the unique bacterial species detected, 43% were detected only in the analysis of reads, whereas 2% were detected only in the analysis of contigs (Fig. 4b). These results could be due to decreased specificity in the read analysis or decreased sensitivity in the contig analysis. Our analysis underscores the importance of understanding the differences among the databases and using the same database for all comparative metagenomic analyses. Thus, both the sequencing and analytical methods showed differences in both detection and abundance of some species. Some of these discrepancies were because the RDP classifier gave a genus-level assignment, whereas NCBI gave only uncultured bacteria-level assignment for the same sequence.

Analysis of diversity

Next, we analyzed the effects of the different sequencing methods on diversity as measured by Shannon diversity, Simpson index and evenness (Fig. 5 and Table S8). All 3 measures of diversity were lower for the 16S than the WGS analyses. Similar results were observed for the individual multiplexed libraries (Figs. S8a,b). Among the datasets prepared by WGS, the contig-based analysis showed increased diversity measured by Shannon diversity, Simpson index and evenness in all datasets compared with the read-based analyses.

Depth of coverage of bacterial genomes

Due to horizontal gene transfer, bacterial genomes of strains even within the same species are potentially diverse. To investigate if the BLAST hits to a reference species mapped to short homologous sequences or with broad regions of a genome, we analyzed the depth and breadth of sequencing coverage for 10 reference genomes with abundance classified as common, high, moderate, low and rare (Table 5). The depth of coverage by sequencing reads ranged from 452 fold in the common abundance to 1.8 fold in the rare abundance species. A representative graphical representation of genome coverage for *Faecalibacterium prausnitzii* shows diffuse coverage for both strands of the genome for both mapped reads and contigs (Fig. 6). For this genome, the mapped reads covered 3,337 features (95.5% of total) and the contigs covered 2,460 features (70.4% of total). To determine if genomes with lower depth of sequencing also had diffuse coverage, we plotted the depth of sequencing on the genomes of the 10 reference species (Figs. S9a–j). These results showed diffuse coverage that, although gaps and spikes were present, was broadly distributed across the genome. For example, the coverage of *Faecalibacterium prausnitzii* L2–6 shows broadly dispersed coverage with few gaps (Fig. 9a). Interestingly, coverage of *Aggregatibacter aphrophilus*

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

NJ8700, which is a rare species with 1.8 fold coverage, also had diffuse coverage with intermittent spikes (Fig. S9j). We also documented genome size for the abundant species which ranged from 2.1–6.8 Mb (Table S9). These results indicate that sequencing coverage of the bacterial genomes was variable but diffuse.

Comparing databases to assign taxonomy to metagenome sequences

In our analyses we used four different databases to assign taxonomy to the sequences. When discrepancies among the databases occurred, discrepant sequences were compared among the databases. For example, we observed that *Blautia* sp. Ser8 was only detected in the 16S amplicon library, whereas *Clostridium saccharolyticum* WM1 was absent in the 16S amplicon library. To analyze a discrepancy we submitted the reference query sequences, for example *Blautia* sp. Ser8 (S001745585; GU124472, reference sequence obtained from RDP), and *Clostridium saccharolyticum* WM1 (S002287206; CP002109) to the online RDP classifier at http://rdp.cme.msu.edu/classifier/classifier.jsp [28]. We parsed the detailed RDP match results using taxon assignments with ≥95% bootstrap support and saved the assignment detail for each RDP match. Also, we submitted the reference query sequence *Blautia* sp. Ser8 and *Clostridium saccharolyticum* WM1 to the online NCBI databases at (http://www.ncbi.nlm.nih.gov). NCBI BLAST computed a pairwise alignment between the query and the database sequences using default parameters. BLAST results for each sequence from each database were saved for comparison (Fig. S10). In general, comparison at the phylum level using the MG-RAST M5NR and NCBI-nt database revealed no differences in relative abundance of major phyla, however detection of unclassified bacteria was higher (~ 4 %) in NCBI-nt database (Fig. S11).

Gene detection

To determine the putative number of genes encoded by fecal metagenome, we predicted the number of genes in the *de novo* assembled contigs in each dataset with the MetaGeneMark algorithm (Table 6). After identifying the predicted genes, we selected all unique genes with the CD-HIT algorithm. Next, we aligned the putative genes with tblastx to the amino acid sequences in the NCBI database. Alignment with the amino acid sequences increased the confidence that the predicted genes could be translated and functional. The v2+v3+HiSeq-total dataset had the largest number of unique predicted genes (811,933), followed by the v3-total (783,887), HiSeq dataset (733,705) and the v2-total dataset (422,174). Thus, the HiSeq dataset detected the lowest number of genes. Although the HiSeq data generated the most identified reads (67.2×10^6 versus 59.0×10^6), the MiSeq v3-total dataset generated the largest number of total bases (15.6 Gb versus 6.7 Gb) due to the longer reads produced by the v3 technology. Interestingly, the longer MiSeq reads actually resulted in fewer total contigs, but the N50 length was substantially longer. These results suggest that longer reads can improve the efficiency of the assembly process resulting in the identification of a larger number of predicted genes.

DISCUSSION

In this study we performed extremely deep sequencing (194.1×10^6 reads) of a single sample using multiple approaches to evaluate parameters that can affect sequencing results and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

analytical interpretation to determine optimal methods. First, we compared the results of 16S versus WGS sequencing. The 16S amplicon approach has been the most commonly employed method to analyze bacterial microbiomes and has several important advantages: 1) it is cost effective, 2) data analysis can be performed by established pipelines, and 3) there is a large body of archived data for reference. However, our work demonstrated multiple substantial advantages of the WGS approach. As shown in a rarefaction plot, WGS identified significantly more bacterial species per read than the 16S method (Fig. 2a). For example, with $>32\times10^6$ reads, the WGS method identified approximately twice as many species as the 16S method (4100 versus 2050 species). It is noteworthy that studies using the 16S method commonly generate approximately 10,000 reads[29]. Since the slope of the 16S rarefaction curve remains positive at 32×10^6 reads, it is likely that the lower number of reads commonly used does not fully detect the richness of a microbiome. In terms of diversity, 3 different metrics (Shannon diversity, Simpson index and evenness) all showed greater diversity with the WGS approach (see Fig. 6 and Table S8). In addition, the WGS, but not the 16S, approach can identify organisms in additional kingdoms including viruses, fungi and protozoa.

Another consideration is the role of the reference database in the interpretation of the 16S versus WGS approaches. The 16S method assigns OTUs based on the 16S amplicon which is used to predict classification of taxa. The classifications are most effective at the phyla and, to a lesser extent, genera levels, but often lack accuracy at the species level. Due to this limitation, the RDP classifier frequently assigns a 16S amplicon sequence to a genus without specifying a species. In contrast, the WGS approach can confidently assign classifications for many sequences at the species level. Thus, our analyses were occasionally qualitatively and quantitatively divergent. For example, 2,441 species detected by the WGS reads were not detected by the 16S method. Also, identical reference genomes occasionally had different classifications. For example, the *Blautia* sp. Ser8 genome in the RDP database was annotated as an uncultured bacterium clone (99%) in NCBI-nt, *Clostridium saccharolyticum* WM1 (93%) in NCBI representative genome database and *Blautia producta* strain JCM (98%) in NCBI-16S ribosomal RNA database (Fig. S10). Thus, the development of more complete and universal reference databases will be essential to enhance the accuracy of species classifications.

Using the WGS method, we performed a direct comparison of the Illumina HiSeq 2000 using the TruSeq SBS v3 reagent with the MiSeq v2–300 and v3–600 reagents. As expected, the HiSeq run generated the most identified reads (67.2×10^6) compared with 59.0×10^6 and 37.5×10^6 for the v3–600 and v2–300 MiSeq runs, respectively (Table 1). However, due to longer reads, the v3–600 MiSeq run generated the highest yield of 18.0 Gb. In the analysis of the rarefaction curves based on the number of sequencing reads, all of the WGS curves were approximately overlaid (Fig. 2a). However, when we performed the same analysis using contigs rather than reads, both the v2–300 and v3–600 MiSeq results outperformed the HiSeq data on a per read basis and also an absolute basis (Fig. 2b). All reads were paired-end, however, the v2–300 and v3–600 reads were longer, a factor that likely improved the efficiency of the *de novo* assembly. Interestingly, the v2–300 reads were only 50 bp longer than the HiSeq reads (151 versus 100 bp, respectively). Since the rarefaction curves using contigs from the v2–300 (read length 151) and v3–600 (read length 301) dataset were

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

similar, it is possible that increases from 100 to 150 supports critical increases in effectiveness of *de novo* assembly. Length also may be a critical factor in the differences between the read and contig comparisons with the 16S amplicons (Fig. 2). In this comparison, the read analysis of the v2+v3+HiSeq-total dataset identified 2,459 species not detected by the 16S amplicon analysis, whereas the contig analysis of the same v2+v3+HiSeq-total dataset identified only 1,353 species not detected by the 16S amplicon analysis. One interpretation of these results is that the shorter reads generate more false positives than the longer contigs. In summary, based on the v2–300 versus v3–600 and the read versus contig comparisons, it is apparent that read length exerts a major effect on data analysis.

A major limitation of the 16S amplicon method is that it sequences only a single region of the bacterial genome, whereas the WGS method can sequence broad regions of the genome. To directly assess the breadth and depth of genome coverage with the WGS method, we analyzed coverage of 10 reference genomes with common, high, moderate, low and rare abundance. Average coverage ranged from 452.3 fold for the high abundance *Faecalibacterium prausnitzii* L2–6 to 1.8 fold for the rare *Aggregatibacter aphrophilus* NJ8700 (Table 5). For all species analyzed, we observed occasional spikes and gaps, but overall diffuse coverage throughout the genomes (Figs. S9a–j). Thus, with adequate sequencing depth the WGS method can generate both depth and breadth of coverage of the bacterial genomes even including bacteria of rare abundance. A major strength of the WGS method is the capacity to identify specific genes in the microbiota. In this study, we used the MetaGeneMark algorithm to identify predicted genes in our contig data and then used tblastx against the NCBI database to confirm that the predicted genes encoded amino acid sequences. In this analysis, the v2+v3+HiSeq-total identified 811,933 genes, whereas the HiSeq and v2-total identified 733,705 and 422,174, respectively (Table 6). These results support the conclusion that longer reads enhance gene prediction possibly due to more efficient *de novo* assembly of contigs. Although the HiSeq has been reported to generate fewer errors, this advantage was apparently neutralized in our analysis by the longer reads produced by the MiSeq. In summary, our study demonstrates that WGS has multiple advantages compared with the 16S rRNA amplicon method including enhanced detection of bacterial species, increased detection of diversity and increased prediction of genes. In addition, increased length, either due to longer reads or the assembly of contigs, improved the accuracy of species detection.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by RO1 HL081663 and RO1 AI053878 to DLP.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

REFERENCES

1. Pflughoeft KJ, Versalovic J. Human microbiome in health and disease. Annu. Rev. Pathol. 2012; 7:99–122. [PubMed: 21910623]
2. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. Nat. Rev. Genet. 2012; 13:260–270. [PubMed: 22411464]
3. Kelly CP. Fecal microbiota transplantation — An old therapy comes of age. N. Engl. J. Med. 2013; 368:474–475. [PubMed: 23323865]
4. Vrieze A, Van Nood E, Holleman F, Salojarvi J, Kootte RS, Bartelsman JF, Dallinga-Thie GM, Ackermans MT, Serlie MJ, Oozeer R, Derrien M, Druesne A, Van Hylckama Vlieg JE, Bloks VW, Groen AK, Heilig HG, Zoetendal EG, Stroes ES, de Vos WM, Hoekstra JB, Nieuwdorp M. Transfer of intestinal microbiota from lean donors increases insulin sensitivity in individuals with metabolic syndrome. Gastroenterology. 2012; 143:913–916. e917. [PubMed: 22728514]
5. Ratner M. Fecal transplantation poses dilemma for FDA. Nat. Biotechnol. 2014; 32:401–402. [PubMed: 24811495]
6. Vrieze A, de Groot PF, Kootte RS, Knaapen M, van Nood E, Nieuwdorp M. Fecal transplant: a safe and sustainable clinical therapy for restoring intestinal microbial balance in human disease? Best Pract. Res. Clin. Gastroenterol. 2013; 27:127–137. [PubMed: 23768558]
7. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Caporaso JG, Lozupone CA, Lauber C, Clemente JC, Knights D, Knight R, Gordon JI. Human gut microbiome viewed across age and geography. Nature. 2012; 486:222–227. [PubMed: 22699611]
8. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE. Metagenomic analysis of the human distal gut microbiome. Science. 2006; 312:1355–1359. [PubMed: 16741115]
9. Lynch MD, Neufeld JD. Ecology and exploration of the rare biosphere. Nat Rev Microbiol. 2015; 13:217–229. [PubMed: 25730701]
10. Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D, Knight R. Experimental and analytical tools for studying the human microbiome. Nat Rev Genet. 2012; 13:47–58. [PubMed: 22179717]
11. Rappe MS, Giovannoni SJ. The uncultured microbial majority. Annu Rev Microbiol. 2003; 57:369–394. [PubMed: 14527284]
12. Stewart EJ. Growing Unculturable Bacteria. Journal of Bacteriology. 2012; 194:4151–4160. [PubMed: 22661685]
13. Nichols D, Cahoon N, Trakhtenberg EM, Pham L, Mehta A, Belanger A, Kanigan T, Lewis K, Epstein SS. Use of ichip for high-throughput in situ cultivation of "uncultivable" microbial species. Appl Environ Microbiol. 2010; 76:2445–2450. [PubMed: 20173072]
14. C. Human Microbiome Project. Structure, function and diversity of the healthy human microbiome. Nature. 2012; 486:207–214. [PubMed: 22699609]
15. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Dore J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, Meta HITC, Bork P, Ehrlich SD, Wang J. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010; 464:59–65. [PubMed: 20203603]
16. Sanschagrin S, Yergeau E. Next-generation Sequencing of 16S Ribosomal RNA Gene Amplicons. 2014:e51709.
17. Poretsky R, Rodriguez RL, Luo C, Tsementzi D, Konstantinidis KT. Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. PLoS One. 2014; 9:e93827. [PubMed: 24714158]
18. Konstantinidis KT, Tiedje JM. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. Curr. Opin. Microbiol. 2007; 10:504–509. [PubMed: 17923431]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

19. Konstantinidis, K.; Stackebrandt, E. Defining Taxonomic Ranks. In: Rosenberg, E.; DeLong, E.; Lory, S.; Stackebrandt, E.; Thompson, F., editors. The Prokaryotes. Berlin Heidelberg: Springer; 2013. p. 229-254.

20. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014; 15:121–132. [PubMed: 24434847]

21. Luo C, Rodriguez RL, Konstantinidis KT. A user's guide to quantitative and comparative analysis of metagenomic datasets. Methods Enzymol. 2013; 531:525–547. [PubMed: 24060135]

22. Luo C, Rodriguez RL, Konstantinidis KT. MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. Nucleic Acids Res. 2014; 42:e73. [PubMed: 24589583]

23. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics. 2008; 9:386. [PubMed: 18803844]

24. Namiki T, Hachiya T, Tanaka H, Sakakibara Y. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res. 2012

25. Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. Nucleic Acids Res. 2010; 38:e132. [PubMed: 20403810]

26. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, Arumugam M, Kultima JR, Prifti E, Nielsen T, Juncker AS, Manichanh C, Chen B, Zhang W, Levenez F, Wang J, Xu X, Xiao L, Liang S, Zhang D, Zhang Z, Chen W, Zhao H, Al-Aama JY, Edris S, Yang H, Wang J, Hansen T, Nielsen HB, Brunak S, Kristiansen K, Guarner F, Pedersen O, Dore J, Ehrlich SD, Meta HITC, Bork P, Wang J, Meta HITC. An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol. 2014; 32:834–841. [PubMed: 24997786]

27. C. Human Microbiome Project. A framework for human microbiome research. Nature. 2012; 486:215–221. [PubMed: 22699610]

28. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. 2014; 42:D633–D642. [PubMed: 24288368]

29. Kuczynski J, Costello EK, Nemergut DR, Zaneveld J, Lauber CL, Knights D, Koren O, Fierer N, Kelley ST, Ley RE, Gordon JI, Knight R. Direct sequencing of the human microbiome readily reveals community differences. Genome Biol. 2010; 11:210. [PubMed: 20441597]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

HIGHLIGHTS

- The human microbiome has emerged as a major player in regulating human health and disease.
- Accurate identification of microbes constituting the microbiota is a major challenge.
- We report a comparative study of an alternative approach using shotgun whole genome sequencing (WGS) compared to 16S ribosomal RNA amplicon sequencing.
- WGS have multiple advantages with enhanced detection of bacterial species with high accuracy, increased detection of diversity and prediction of genes.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

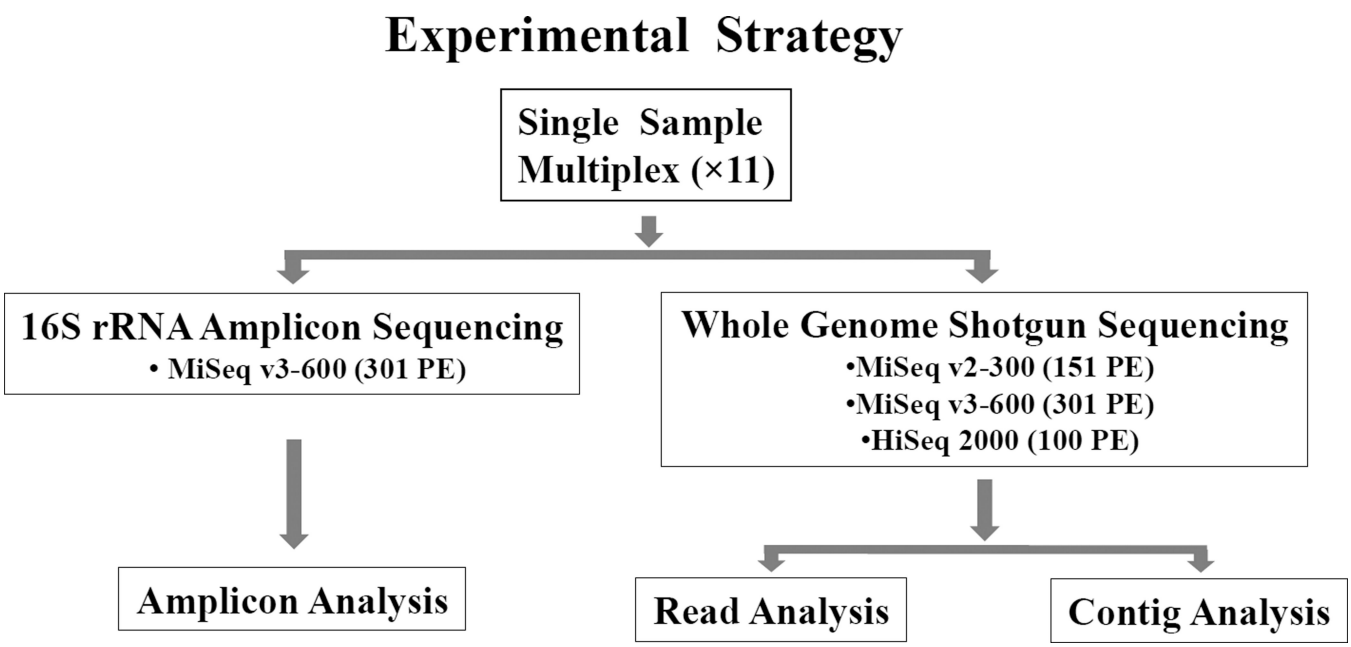


Figure 1. Experimental strategy to compare sequencing methods, platforms and data analysis

Experimental design for 16S rRNA amplicon and WGS sequencing for a single fecal sample multiplexed in 11 libraries is shown. 16S amplicon sequencing was performed using MiSeq v3–600 and WGS sequencing was performed using MiSeq v2–300, MiSeq v3–600 and HiSeq 2000. The 16S data was analyzed using OTU based amplicon approach and the WGS read and contig data were analyzed using the MG-RAST M5NR and NCBI nt database.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

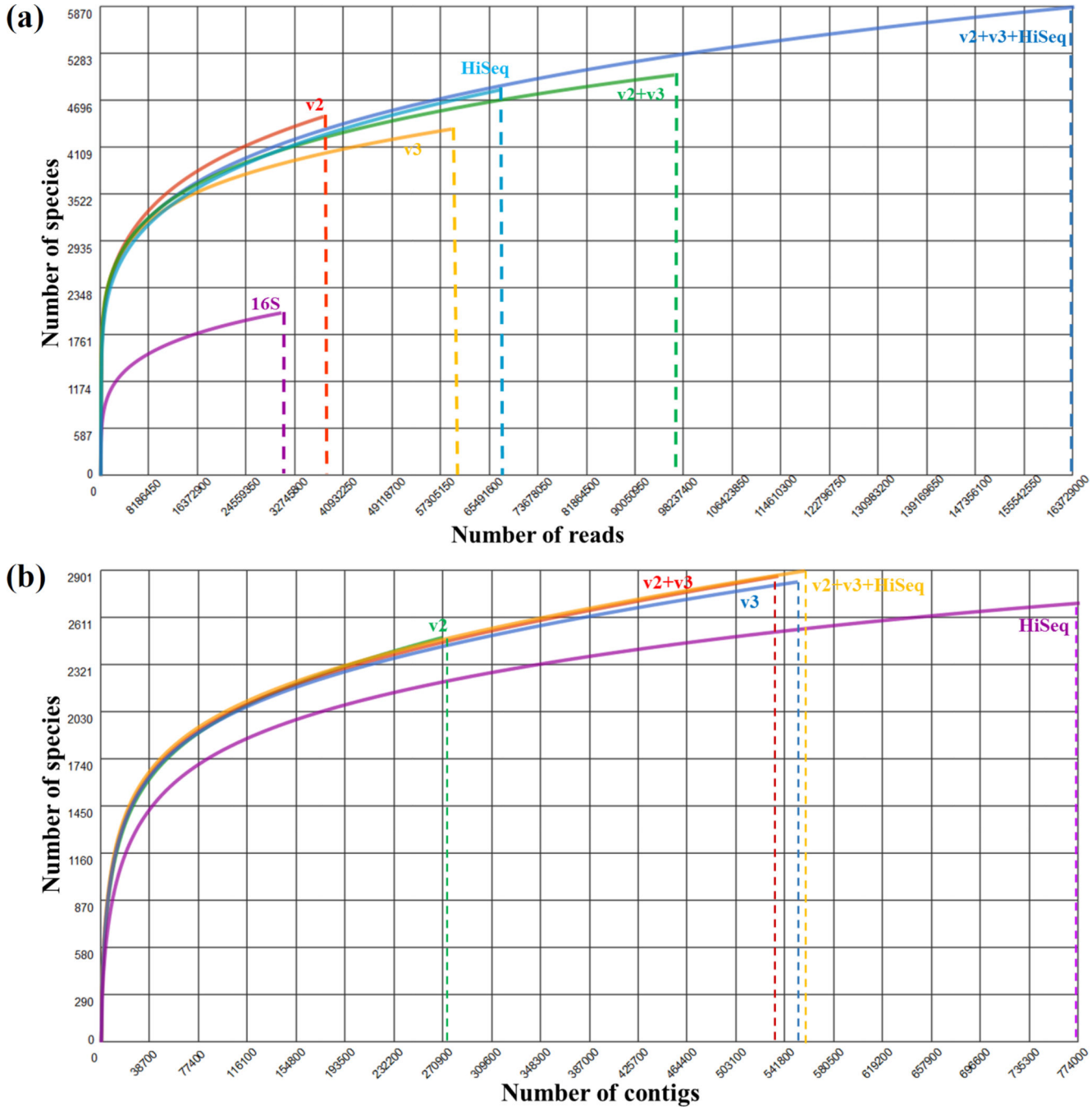


Figure 2. Rarefaction curves of combined reads and contigs

Rarefaction curve for 16S amplicon, HiSeq, v2, v3, v2+v3 and v2+v3+HiSeq data using a read-based analysis (a) and for HiSeq, v2-total, v3-total, v2+v3-total and v2+v3+HiSeq-total data using a contig-based analysis (b). Graph shows total number reads or contigs (x-axis) and total number of species identified (y-axis). Vertical dashed lines mark the number of reads or contigs detected for each dataset.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

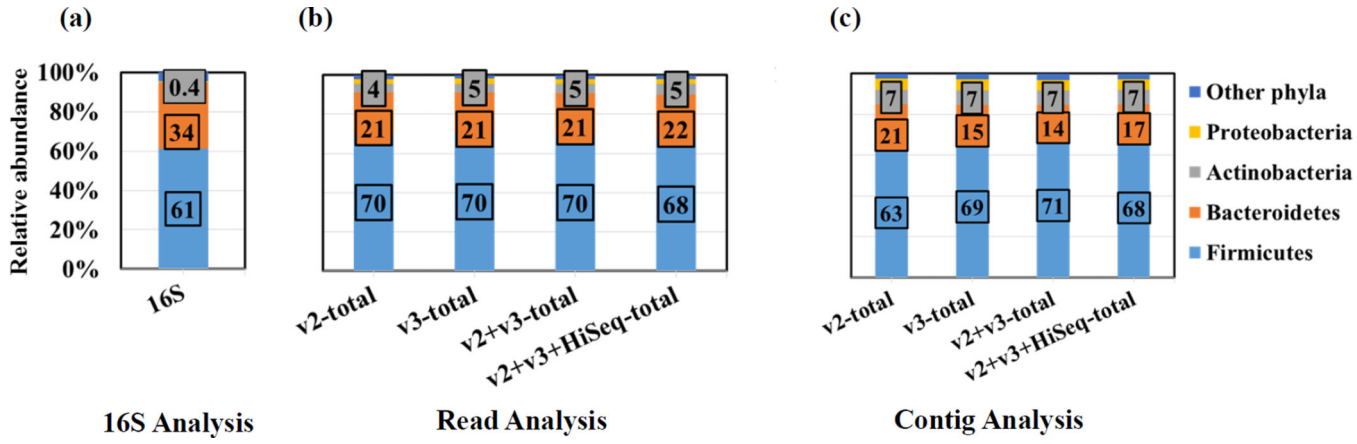


Figure 3. Relative abundance of bacterial phyla

Stacked bar graph of relative abundance of bacterial phyla identified in 16S amplicon based analysis (a), read-based analysis of WGS data (b) and contig-based analysis of WGS data (c) in the v2-total, v3-total, v2+v3-total and v2+v3+HiSeq-total datasets. Relative abundance (y-axis) of the dominant bacterial phyla includes Firmicutes, Bacteroidetes, Actinobacteria and Proteobacteria. The “other phyla” for 16S amplicon analysis contains 19 non-abundant phyla and unclassified bacteria representing <5% of total abundance. The “other phyla” for the WGS analysis contains 27 non-abundant phyla and unclassified bacteria representing <2% of total abundance.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

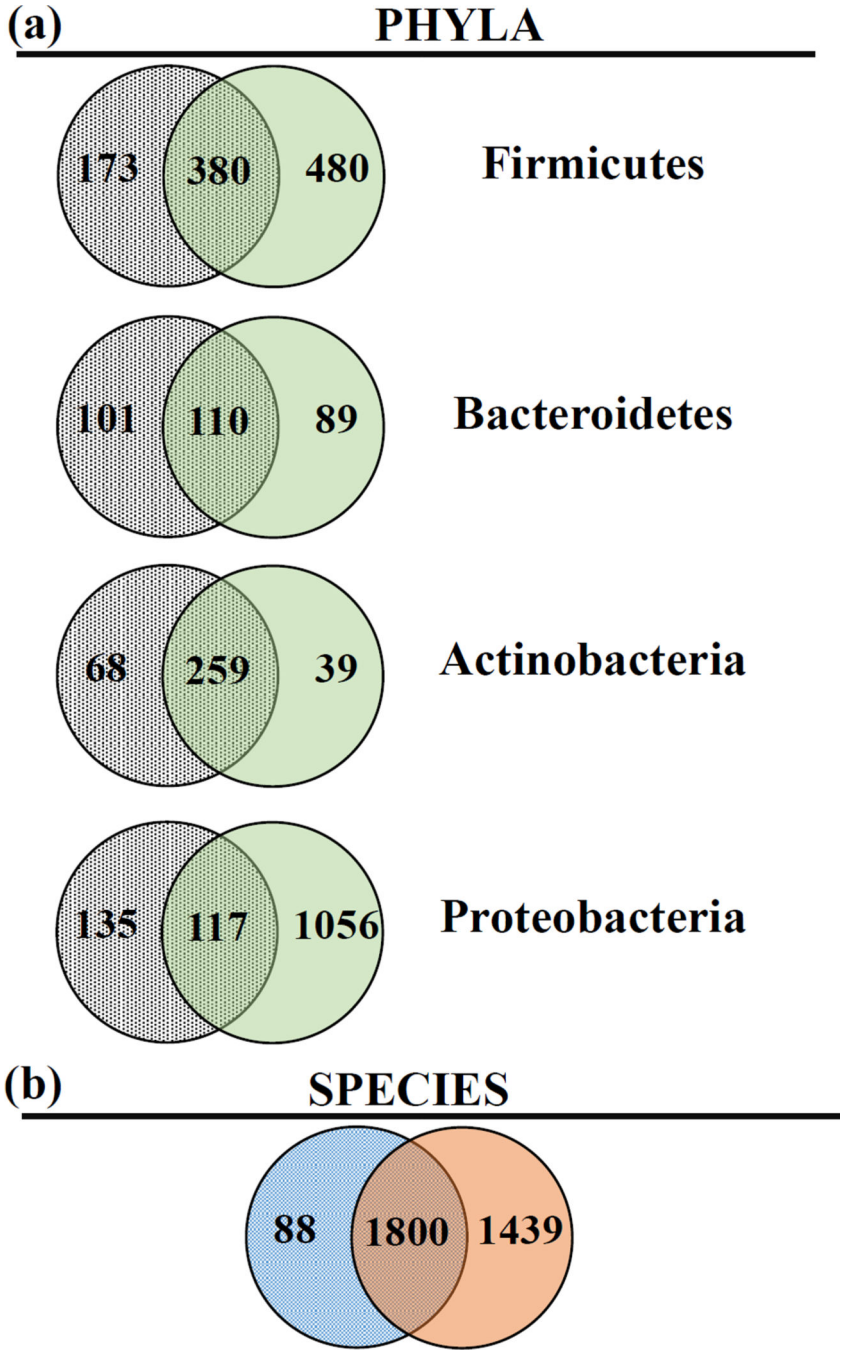


Figure 4. Comparison of taxa identified by different sequencing and analysis methods

Number of species identified in the four predominant phyla identified in the 16S rRNA amplicon (grey) and in the WGS v2+v3+HiSeq-total read (green) datasets (a). The union of species for the Firmicutes (37%), Bacteroidetes (37%), Actinobacteria (32%) and Proteobacteria (9%) is shown in the overlap. A comparison of total species detection using a contig-based analysis (blue) versus a read-based analysis (orange) shows overlap in species detection of 54% (b).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

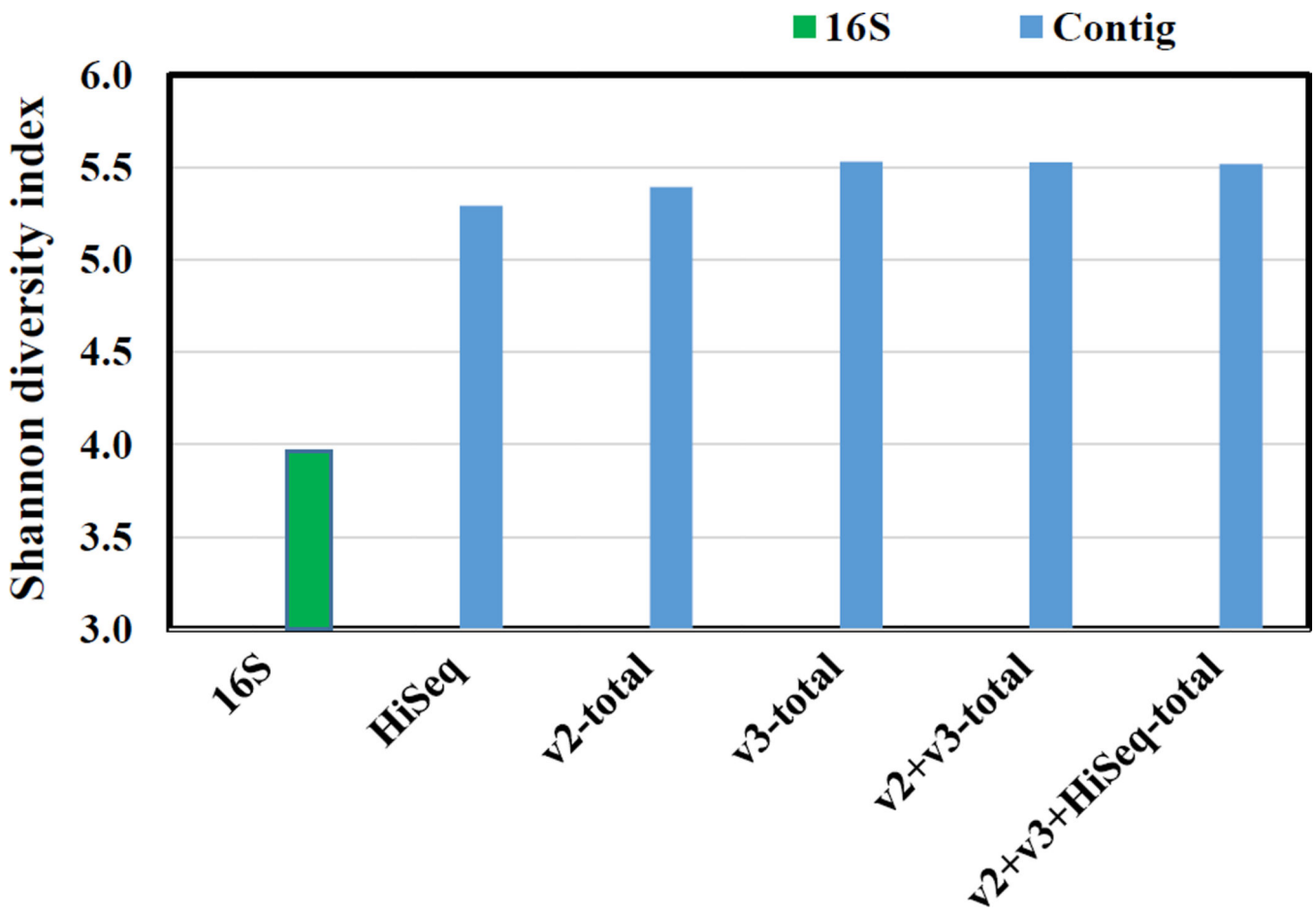
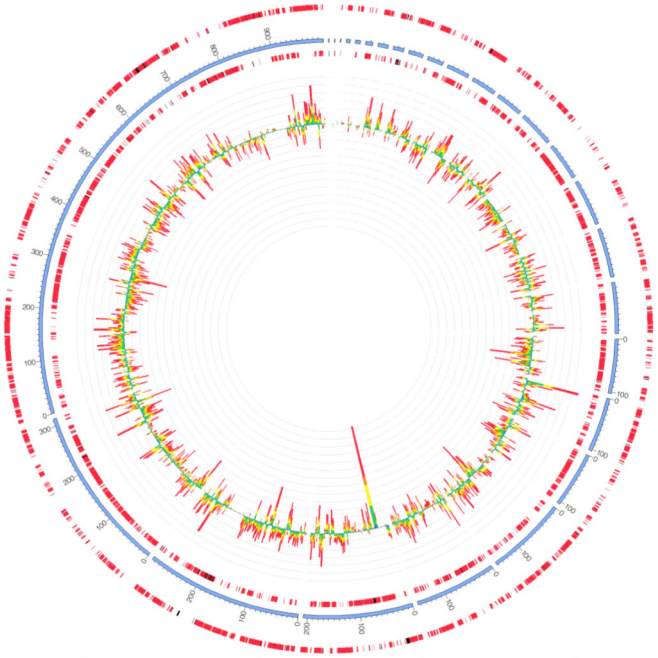


Figure 5. Greater diversity detected with the WGS than 16S method

Bar chart of Shannon diversity index calculated at species level from 16S, HiSeq, v2-total, v3-total, v2+v3-total and v2+v3+HiSeq-total datasets. The diversity of the WGS datasets was analyzed on de novo assembled contigs. Read based and contig based methods showed consistent and reproducible diversity index values among the samples.

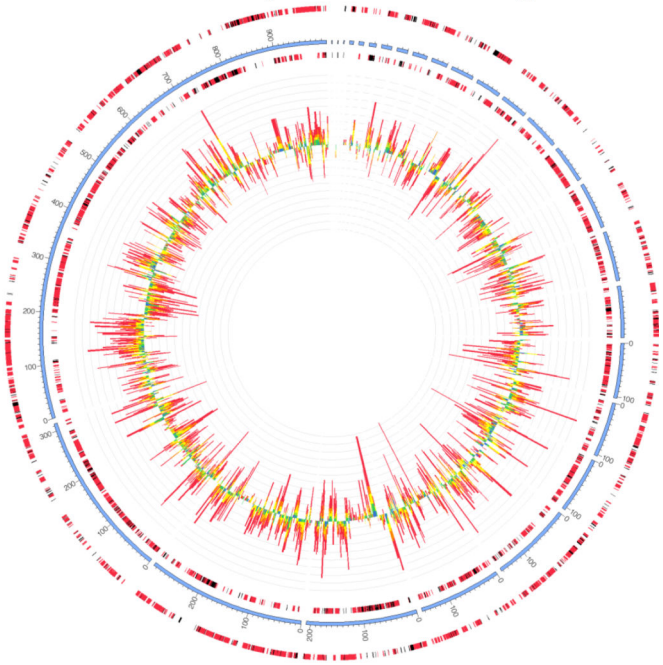
Author Manuscript Author Manuscript Author Manuscript Author Manuscript

v2+v3+HiSeq-total reads mapped on *Faecalibacterium prausnitzii* reference genome



Hits distribution by *e*-value exponent range	Summary of mapped hits
Features Mapped	6579946
Features Covered	3337
Total Features	3493
Contigs Shown	26
Total Contigs	27

v2+v3+HiSeq-total contigs mapped on *Faecalibacterium prausnitzii* reference genome



Hits distribution by *e*-value exponent range	Summary of mapped hits
Features Mapped	10038
Features Covered	2460
Total Features	3493
Contigs Shown	26
Total Contigs	27

Figure 6. Comparison of coverage of a representative genome using read-based versus contig-based analysis

Genome recruitment plots of a representative reference genome, *Faecalibacterium prausnitzii*, using a read-based (left) versus a contig-based (right) analysis of the v2+v3+HiSeq-total dataset. Circular plots were created using the MG-RAST genome recruitment tool using a maximum *e*-value of $1e^{-3}$ and a $\log_2$ abundance scale. The leading and lagging strands are represented by the outer and inner most rings, separated by the blue ring, which indicates the position within the genome. Metagenomic features are depicted as bar graphs inside the genome. The greater height represent more mapped features and the *e*-value exponent is color-coded as blue (−3 to −5), green (−5 to −10), yellow (−10 to −20), orange (−20 to −30) and red (less than −30).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Sequencing statistics.

Method	Cycles	Reads identified	Yield (Gb)
16S amplicon	2×301	30,378,368	9.1
HiSeq 2000	2×100	67,229,282	6.7
MiSeq v2–300	2×151	37,482,018	5.8
MiSeq v3–600	2×301	59,018,428	18.0
Total		194,108,096	39.6

Sequencing data for the 16S rRNA amplicon, HiSeq 2000, MiSeq v2–300 and MiSeq v3–600 methods. Sequencing method, number of cycles per run, sequence length, total number of reads identified and total yield (Gb) is shown from each run. A total of ~194 million reads with ~ 40 Gb sequence data were generated. Gb (Giga bases), 2×(Paired-end sequencing chemistry).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Cumulative abundance of high, low and rare abundance species.

Abundance category	Relative abundance (%)	Number of species	Total abundance (%)
Common	>1.00	21	34.0
High	0.50 to 0.99	35	23.1
Moderate	0.05 – 0.49	213	31.7
Low	0.01 to 0.05	397	7.8
Rare	<0.01	1222	3.3

Number of species identified in v2+v3+HiSeq-total (contig based) were binned into common, high, moderate, low and rare abundance based on their percent relative abundance.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Biochem Biophys Res Commun. Author manuscript; available in PMC 2017 January 22.

Table 3

Detection of abundant species with low numbers of randomly sampled reads.

Abundant species	Number of randomly sampled reads							
	250	**500**	**750**	**1,000**	**2,000**	**3,000**	**5,000**	**10,000**
Faecalibacterium prausnitzii	32.7	40.7	39.7	40.0	38.0	39.3	36.3	38.0
Eubacterium rectale	8.8	8.6	6.1	8.2	8.2	7.2	9.3	7.7
Roseburia intestinalis	4.4	3.8	5.4	4.3	5.4	5.9	5.9	5.7
Bacteroides vulgatus	5.3	5.7	3.8	5.5	5.7	6.3	4.4	5.1
Unclassified bacteria	8.0	2.9	5.1	4.6	3.7	5.2	4.2	4.3
Bacteroides xylanisolvens	2.7	2.4	1.9	3.9	3.2	3.6	3.5	3.8
Ruminococcus torques	3.5	5.3	5.1	3.6	2.8	3.7	3.8	4.2
Ruminococcus sp. SR1/5	0.9	2.4	2.9	2.7	2.8	3.5	2.7	3.5
Bacteroides dorei	10.6	3.8	1.6	2.7	4.1	1.6	3.7	3.1
Alistipes finegoldii	1.8	2.9	2.2	2.7	2.2	1.7	2.1	2.0
Bifidobacterium longum	0.0	1.4	0.6	2.4	2.2	1.7	1.6	1.4
Roseburia hominis	1.8	1.4	2.9	2.9	1.5	0.7	1.9	1.4
butyrate-producing bacterium SSC/2	1.8	1.0	1.3	1.4	1.6	1.8	1.5	1.6
Parabacteroides distasonis	0.9	1.0	0.0	1.4	0.7	1.1	1.7	1.5
Bacteroides thetaiotaomicron	0.9	3.8	1.0	0.7	2.0	1.7	1.2	1.3
Ruminococcus obeum	1.8	1.0	3.2	1.9	0.9	2.2	1.2	1.3
Bacteroides fragilis	0.0	1.4	0.6	1.0	1.1	0.9	1.2	1.1
butyrate-producing bacterium SS3/4	1.8	1.0	1.3	0.5	1.2	1.2	2.7	1.2
Eubacterium eligens	1.8	1.0	2.9	1.9	0.9	0.4	1.2	1.5
Coprococcus catus	0.9	1.4	1.0	1.0	1.1	1.1	1.1	1.1
Total abundant species identified	18	20	19	20	20	20	20	20

Abundant species were defined as ≥ 1% relative abundance in the v3-total dataset. To determine the sensitivity of species detection with low numbers of reads, 250–10,000 reads were randomly sampled from the v3-total dataset and blast hits were determined in the NCBI nt database and quantified by relative abundance.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 4

Identification of abundant species using 16S, HiSeq, and MiSeq data (v2-total, v3-total, v2+v3-total, v2+v3+HiSeq-total).

Rank	Phylum	Species	16S	HiSeq	v2-total	v3-total	v2+v3-total	v2+v3+HiSeq-total
1	Firmicutes	*Faecalibacterium prausnitzii*	7.1	23.5	18.4	18.3	17.0	17.3
2	Firmicutes	*Eubacterium rectale*	8.0	5.1	6.1	6.4	6.2	5.8
3	Firmicutes	*Roseburia intestinalis*	2.2	2.2	2.8	2.9	2.8	2.7
4	Firmicutes	*Ruminococcus* sp. 5_1_39BFAA	0.0	2.0	2.6	2.5	2.6	2.5
5	Bacteroidetes	*Bacteroides vulgatus*	7.5	1.9	2.0	2.0	2.0	2.0
6	Firmicutes	*Subdoligranulum variabile*	0.0	2.1	1.5	1.6	1.6	1.8
7	Firmicutes	*Eubacterium eligens*	1.3	1.4	1.7	1.6	1.7	1.6
8	Bacteroidetes	*Alistipes putredinis*	1.0	1.9	1.4	1.4	1.4	1.6
9	Bacteroidetes	*Bacteroides fragilis*	0.7	1.5	1.5	1.5	1.5	1.5
10	Firmicutes	*Clostridium saccharolyticum*	0.0	1.1	1.4	1.5	1.5	1.4
11	Firmicutes	*Ruminococcus obeum*	1.4	1.0	1.3	1.3	1.3	1.3
12	Bacteroidetes	*Bacteroides helcogenes*	0.1	1.2	1.2	1.3	1.3	1.3
13	Firmicutes	*Roseburia inulinivorans*	0.1	1.0	1.3	1.4	1.3	1.2
14	Actinobacteria	*Collinsella aerofaciens*	0.0	1.7	0.9	0.9	0.9	1.2
15	Bacteroidetes	*Bacteroides* sp. 1_1_ 6	0.0	1.2	1.2	1.2	1.2	1.2
16	Firmicutes	*Ruminococcus* sp. SR1/5	0.0	1.0	1.2	1.2	1.2	1.2
17	Firmicutes	*Clostridium phytofermentans*	0.0	0.9	1.1	1.2	1.2	1.1
18	Actinobacteria	*Bifidobacterium longum*	0.0	1.5	0.9	0.9	0.9	1.1
19	Bacteroidetes	*Bacteroides uniformis*	3.9	1.1	1.0	1.1	1.1	1.1
20	Bacteroidetes	*Bacteroides thetaiotaomicron*	1.3	1.0	1.1	1.0	1.1	1.1

The abundant bacterial species (defined as ≥ 1% abundance in the v2+v3+HiSeq-total dataset) were analyzed with a read based analysis using the M5NR database at MG-RAST server.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 5

Genome coverage of representative genomes of high, low and rare abundance detected in the v2+v3+HiSeq-total dataset.

Abundance category	Reference species	Genome length	Number of bases mapped	Coverage
Common	*Faecalibacterium prausnitzii* L2–6	3,321,367	1,502,163,364	452.3
Common	*Eubacterium rectale* M104/1	3,698,419	847,530,632	229.2
High	*Ethanoligenens harbinense* YUAN-3	3,008,576	295,813,386	98.3
High	*Parabacteroides distasonis* ATCC 8503	4,811,379	340,577,118	70.8
Moderate	*Atopobium parvulum* DSM 20469	1,543,805	29,714,000	19.2
Moderate	*Caldanaerobacter subterraneus subsp. tengcongensis* MB4	2,689,445	41,320,021	15.4
Low	*Listeria innocua* FSL J1–023	2,914,007	15,862,353	5.4
Low	*Geobacter sulfurreducens* PCA	3,814,128	28,127,049	7.4
Rare	*Acidiphilium cryptum* JF-5	3,389,227	21,305,814	6.3
Rare	*Aggregatibacter aphrophilus* NJ8700	2,313,035	4,118,299	1.8

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 6

Putative genes predicted by the MetaGeneMark algorithm.

Library	Contigs	Predicted genes	Tblastx hits
HiSeq-total	270,227	435,828	422,174
v2-total	551,174	822,724	783,887
v3-total	774,445	782,181	733,705
v2+v3-total	535,729	820,535	783,773
v2+v3+HiSeq-total	556,831	857,426	811,933