PLOS ONE

RESEARCH ARTICLE

IM-TORNADO: A Tool for Comparison of 16S Reads from Paired-End Libraries

Patricio Jeraldo[1,2], Krishna Kalari[3], Xianfeng Chen[3], Jaysheel Bhavsar[3], Ashutosh Mangalam[4], Bryan White[2,5], Heidi Nelson[1], Jean-Pierre Kocher[3], Nicholas Chia[1,2,6]*

1. Department of Surgery, Mayo Clinic, Rochester, Minnesota, United States of America, **2.** Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, **3.** Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, United States of America, **4.** Department of Immunology, Mayo Clinic, Rochester, Minnesota, United States of America, **5.** Department of Animal Sciences, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, **6.** Department of Physiology and Biomedical Engineering, Mayo Clinic College of Medicine, Rochester, Minnesota, United States of America

*chia.nicholas@mayo.edu

Citation: Jeraldo P, Kalari K, Chen X, Bhavsar J, Mangalam A, et al. (2014) IM-TORNADO: A Tool for Comparison of 16S Reads from Paired-End Libraries. PLoS ONE 9(12): e114804. doi:10.1371/journal.pone.0114804

Editor: Paul Jaak Janssen, Belgian Nuclear Research Centre SCK•CEN, Belgium

Received: May 1, 2014

Accepted: October 29, 2014

Published: December 15, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All the data is available at Dryad. The data DOI is doi: 10.5061/dryad.fm67n.

Funding: This work was supported by the Mayo Clinic Center for Individualized Medicine Microbiome Program. PJ was funded by the Mayo-Illinois Strategic Alliance for Technology-Based Healthcare. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors would also note here that Prof. Bryan White is member of the Editorial Board of PLOS ONE. This does not alter the authors' adherence to PLOS ONE Editorial policies and criteria.

Abstract

Motivation: 16S rDNA hypervariable tag sequencing has become the *de facto* method for accessing microbial diversity. Illumina paired-end sequencing, which produces two separate reads for each DNA fragment, has become the platform of choice for this application. However, when the two reads do not overlap, existing computational pipelines analyze data from read separately and underutilize the information contained in the paired-end reads.

Results: We created a workflow known as Illinois Mayo Taxon Organization from RNA Dataset Operations (IM-TORNADO) for processing non-overlapping reads while retaining maximal information content. Using synthetic mock datasets, we show that the use of both reads produced answers with greater correlation to those from full length 16S rDNA when looking at taxonomy, phylogeny, and beta-diversity.

Availability and Implementation: IM-TORNADO is freely available at http://sourceforge.net/projects/imtornado and produces BIOM format output for cross compatibility with other pipelines such as QIIME, mothur, and phyloseq.

Introduction

The advent of high-throughput 16S rDNA microbial population sequencing has revolutionized our ability to carry out culture independent surveys [1] and paved the way to understanding the microbiome from environments ranging from the practical [2, 3] to the exotic [4, 5]. The need for bioinformatics has grown hand-

in-hand with the increasing depth and complexities of microbiome analyses [6, 7]. The importance of being able to organize such analyses is epitomized by the popularity of microbiome analysis pipelines such as QIIME [8] and mothur [9], which collectively to date have more than 3000 citations.

Historically, most microbiome analyses were carried out using large amplicons (>500 bp long) sequenced on the 454 pyrosequencing platform. Recently, due to its higher throughput and multiplexing capability, Illumina sequencers have become the preferred platform for microbiome analyses, despite the shorter reads produced (100 to 300 bp long) [10–12]. The Illumina platform is currently used by the microbiome community following two preferred approaches. The first approach consists in designing small amplicons, to force the reads from the same amplicon ends to overlap. These reads can then be assembled into a single read longer than the 250 bp reads, thereby increasing their specificity. The second approach focuses on sequencing the 250 bp ends of the large amplicons originally crafted for the 454 platform. These regions have been previously studied [13, 14] and optimized for capturing phylogenetic or taxonomic information. However, this can lead to reads that do not overlap when sequenced on the Illumina platform [15–18].

The use of non-overlapping paired reads requires a large amount of tracking to maintain the proper linkages while preserving the necessary efficiency expected of 16S rDNA analyses, which typically take advantage of redundancy in the datasets. While cases have been explored previously [16], existing analytical pipelines are only built for analyzing single reads and do not track the linkages between the two non-overlapping reads produced by Illumina, and such tracking can create a "potentially more complicated downstream pipeline" [19]. This can create a conundrum for the researcher who may have to choose drawing conclusions based on a single read [20], or altering their primer design to allow for overlapping reads [11]. These former uses less than the maximum information available and are suboptimal in the accuracy of the analyses they provide while the bioinformatics for the latter has already been solved elsewhere by assembling the paired-end reads [21, 22].

Here, we present an integrated workflow, known as the Illinois-Mayo Taxon Organization from RNA Dataset Operations (IM-TORNADO), for carrying out common microbiome analyses leveraging the information of the paired reads provided by the Illumina sequencers to relate reads belonging to the same amplicon, making the use of these non-overlapping reads accessible to a broader base of users. We use our pipeline to compare the accuracy of results obtained from paired reads versus single reads. We find that with regards to three common operations–making multiple alignment, inferring taxonomy, and phylogeny–utilizing the information contained in both reads leads to better performance than analysis of any single read. Where possible, we build upon existing tools and link to formats for further processing other popular pipelines [8, 9, 23] in order to augment the growing community of computational tools for microbiome analyses. In this article, we will first outline the general methodology of IM-TORNADO and compare the performance of with IM-TORNADO when using

paired reads versus either read individually. We then describe, in detail, the methods employed at every step. Finally, we discuss the significance of IM-TORNADO's improved performance using paired reads and its impact on future experimental designs.

Approach

Paired-end sequencing produces 2 reads, read 1 (R1) and read 2 (R2), each from opposite ends of a DNA fragment. In the case of 16S rDNA hypervariable tag sequencing, these regions are selected by primers targeted to specific regions. We approach the case where R1 and R2 do not overlap and the sequencing of the amplified DNA fragment is incomplete due to the gap between the two reads. Algorithms for taxonomy, multiple alignment, and phylogeny that do not take into account this gap cannot be employed naively and require an organizational framework to properly manage the link between each R1 and R2 read pair.

In this section, we describe results of an analysis pipeline for preparing non-overlapping reads for analysis as a whole unit, without sacrificing one of the reads in the pair. Our approach, IM-TORNADO, employs a number of smaller scripts that uniformize and merge paired-reads for analysis by widely-used analytical tools, guaranteeing reasonable comparability with existing studies. We test our approach using a synthetic mock dataset that simulates 16S rDNA hypervariable tag sequencing of V3–V5 and V6–V9 regions of 16S rDNA and show that IM-TORNADO allows us to extract meaningful improvements in accuracy from non-overlapping paired read analyses in comparison to single-end read analyses. Details of our methodology are provided in the ''Methods for Pipeline Implementation'' section, and information about operational taxonomic units in these synthetic mock communities is discussed in the S1 Text and S2 Table.

Taxonomy

Determining taxonomy relies on profiling characteristics or alignments of sequences in a reference database. 16S rDNA-based classification tools such as Ribosomal Database Project (RDP) taxonomy classifier use 16S rDNA databases such as Greengenes [24]. Algorithms for determining taxonomy expect an ungapped sequence input, adding a layer of difficulty when utilizing non-overlapping paired-end reads. We joined the paired-end reads using an ambiguous nucleotide character before the data is input into the classifier in order to avoid misinterpretation of the data while still retaining the information from both reads.

Table 1 shows the overall accuracy of taxonomic calls using paired, R1, R2, and full length 16S rDNA. Accuracy was determined by comparison to the original sequence in the synthetic mock dataset, whose taxonomy is pre-determined by the reference database, in this case, Greengenes 13_5. Paired reads perform better than either single end read, and paired reads generally outperform R1 by about the same amount by which full length 16S rDNA outperforms the paired reads. In the

Table 1. Taxonomy comparison.

Library type	Domain	Phylum	Class	Order	Family	Genus	Species
Paired	100.0	99.95	99.91	99.59	98.25	94.32	90.78
R1	100.0	99.91	99.83	99.18	97.41	92.12	87.60
R2	100.0	99.89	99.76	99.10	97.14	87.38	80.47
Full length	100.0	99.99	99.97	99.68	99.88	96.30	94.57

Comparison of accuracy from taxonomy calls using paired, read 1 (R1), read 2 (R2), and full length 16S rDNA. Analyses were carried out using the Ribosomal Database Project taxonomy classifier with the complete Greengenes database. Using Paired read analysis provides more accurate taxonomic classification than either R1 or R2 alone across all taxonomic levels. Full length 16S rDNA was used for comparison purposes.

doi:10.1371/journal.pone.0114804.t001

S1 Text we also discuss the effect of read quality errors and error-trimming on the accuracy of the taxonomy assignment (see also S1 Table).

Multiple Sequence Alignment

Before making a comparison of phylogenies, we must make sure we make a sensible choice for the tool used to create the sequence alignment used for the phylogenies. In the case of high-throughput 16S rDNA sequencing, specialized tools exist to guarantee a good alignment (based on the extensive knowledge of ribosomal RNA) while performing a very fast multiple sequence alignment. Two tools stand out for this task. The Nearest Alignment Space Termination (NAST) algorithm [25] uses a template of pre-aligned sequences which are used to find the closest template reads matching the query read, align to the template, and carefully introduce misalignments to make sure the query matches the template structure. The other tool, Infernal [26] is a more general-purpose RNA toolkit which contains a module which aligns queries to a reference secondary structure. In particular, a set of 16S reads is used by the RDP Pipeline to create a secondary structure model using Infernal, and used to align query reads.

In Fig. 1, we show the result of comparing 100 phylogenetic trees (see Methods), created from the same reads randomly chosen from the Greengenes 13_5 database, but using different sequence aligners. We compared PyNAST [27] as bundled with QIIME version 1.8.0 (relaxing the minimum alignment identity threshold down to 50%, to ensure all queries are processed), and Infernal version 1.1 [28]. We compare the maximum likelihood trees created using FastTree version 2.7.1 [29] by examining the Gamma Log Likelihoods reported for each tree. The trees created using Infernal have significantly higher log likelihoods ($p<0.0001$, Wilcoxon signed ranked test) than the trees created using PyNAST, meaning the multiple sequence alignments are of higher quality. This is in agreement with previous results comparing these two algorithms [30].

Phylogeny

Phylogenetic trees are often used as part of the metric of choice [31] when comparing organismal differences between communities. Using the multiple

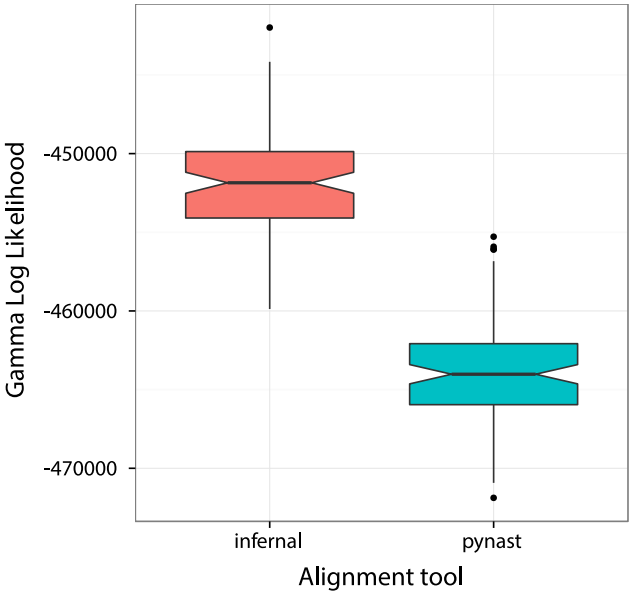

Fig. 1. Comparison of alignment tools. Plot of Gamma Log Likelihoods of 100 trees created from paired reads selected from the Greengenes 13_5 database, and aligned using PyNAST and Infernal version 1.1. Likelihoods in the trees created using Infernal are significantly better than the trees created with PyNAST ($p<0.0001$, Wilcoxon signed ranked test), strongly suggesting that Infernal produces better quality alignments than PyNAST for the same input reads.

doi:10.1371/journal.pone.0114804.g001

alignment scheme from Section ''Multiple Sequence Alignment'' to faithfully incorporate both non-overlapping reads R1R2, we are able to obtain more accurate phylogenies than using either R1 or R2.

We examine the quality of a given tree generated from paired-end sequencing data by looking at how well it agrees with a tree generated using the full length 16S rDNA from our synthetic mock dataset. Greater correlation between the branch distances indicates more similarity between the trees. A tree that agrees well with its full length counterpart captures the relationships between samples. Fig. 2 shows the results from 100 comparisons between full length 16S rDNA phylogeny and its short read counterparts generated using paired reads, R1, or R2. As shown, paired read analysis preserves significantly more of the phylogenetic tree structure than either R1 or R2 alone.

β-diversity

Similarities and differences between microbial communities, or β-diversity, are often calculated from a phylogenetic tree. Here, we investigate the effect of using paired reads R1R2 versus single reads R1 or R2 on the fidelity of β-diversity.

We examine the effect of using paired versus single-end reads in determining β-diversity between 100 randomly constructed communities. In order to assess

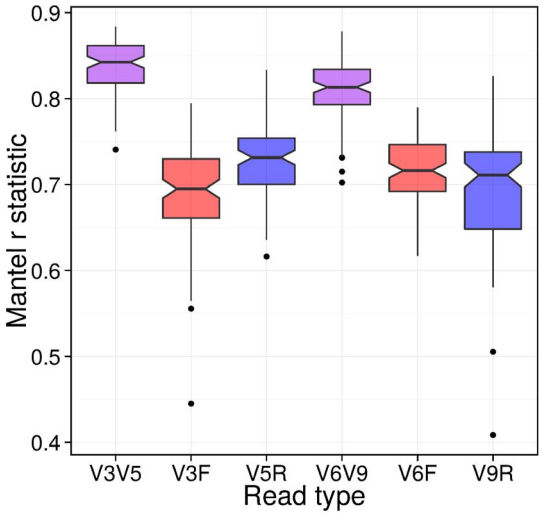

Fig. 2. Comparison of phylogenetic trees between libraries. Plot of a Mantel correlation test comparing cophenetic distance matrices calculated from phylogenetic trees created using paired, R1 and R2 (for both the V3–V5 and V6–V9 primer pairs) versus the distance matrix created from the corresponding full-length 16S trees. A higher correlation value means the trees are more closely related to the full-length trees. Here, the paired trees are significantly closer to the full-length trees than the R1 and R2 trees ($p<0.0001$, Wilcoxon signed ranked test, using 100 synthetic mock communities), strongly suggesting that combining the use of paired reads leads to phylogenies closer to what is obtained from full-length reads, even when the chosen primers create non-overlapping reads.

doi:10.1371/journal.pone.0114804.g002

accuracy, we use the corresponding full length 16S rDNA sequences from the reference database used to construct the synthetic mock dataset. Fig. 3 shows a comparison between the β-diversity distance matrices generated from paired, R1 and R2 reads. Paired reads retain a significantly greater amount of information than either single-end read alone.

Use Case

The purpose of this use case test was to create a synthetic dataset that contains similar diversity and relatedness of organisms that might be found in a real dataset. For our example case, we chose to mimic the characteristics of a stool dataset (included with the validation datasets) by using closed reference mapping to full-length 16S rDNA from Greengenes 13_5 (see Methods). These full-length 16S rDNA reference reads then allowed us to the reproduce multiple artificial datasets using different primer regions what we could then compare against full length 16S rDNA reads. We then tested the different read libraries by computing their respective unweighted UniFrac matrices and compared the results of full-length to paired, R1 and R2 libraries using a Mantel correlation test. In Table 2, we show the results of the Mantel correlation test. We see that the paired reads have a higher correlation to the full-length reads than the single reads. This is concordant with the results shown in Fig. 3, which shows we expect to capture the same improvements when using real data.

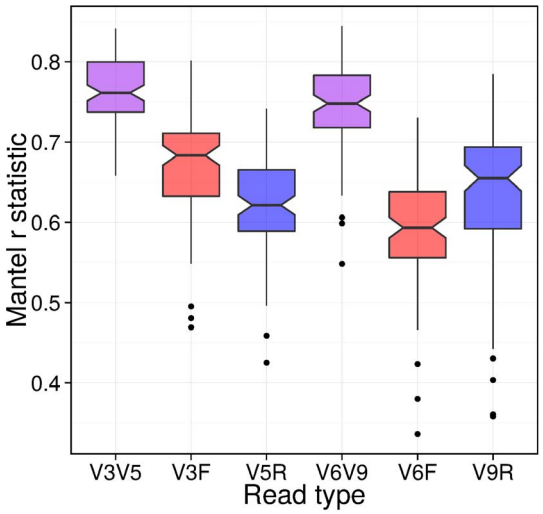

Fig. 3. Comparison of β-diversity between libraries. Plot of a Mantel correlation test comparing unweighted UniFrac distance matrices created using synthetic mock communities from paired, R1 and R2 reads (for both the V3–V5 and V6–V9 pairs) versus the distance matrix created from the corresponding full-length 16S synthetic mock communities. A higher correlation value means the distance matrices, and hence their β-diversity, are more closely related to the full-length communities. Here, the communities from paired reads are significantly closer to the full-length communities than the R1 and R2 communities ($p<0.0001$, Wilcoxon signed ranked test, using 100 synthetic mock communities), strongly suggesting that combining the use of paired reads leads to results closer to what is obtained from full-length reads, even when the chosen primers create non-overlapping reads.

doi:10.1371/journal.pone.0114804.g003

Methods for Pipeline Implementation

In the following, we describe the steps performed by the IM-TORNADO pipeline to clean, merge, track, pick OTU, infer taxonomy and calculate phylogeny for a project. A graphical outline of the pipeline is presented in Fig. 4

Input Data

The input for the IM-TORNADO pipeline is a set of demultiplexed fastq-formatted files. A metadata file describes the samples and other data the end user wants to be included in the output BIOM-formatted file.

Quality Filtering

Demultiplexed sequence files are subject to quality filtering using Trimmomatic [32] version 0.30, with a hard cutoff of PHRED score Q3 for 5′ and 3′ ends of the reads (parameters LEADING: 3 and TRAILING: 3), trimming of the 3′ end with a moving average score of Q15, with a window size of 4 bases (parameter SLIDINGWINDOW: 4:15), and removing any remaining reads shorter than 75% [10] of the original read length (for example, parameter MINLEN: 112 for 150 bp long reads, MINLEN: 187 for 250 bp long reads or MINLEN: 225 for 300 bp long reads). Reads with any ambiguous base calls are discarded.

Table 2. Library comparison in a realistic sample.

Library type	Mantel *r* statistic
Paired	0.931
R1	0.910
R2	0.813

Mantel *r* statistic comparing the unweighted UniFrac matrices for different short-read 16S libraries against a full-length 16S library, based on a real paired-end libary sequenced from stool samples. The paired reads have a higher correlation to the full-length library than any of the other single read libraries.

doi:10.1371/journal.pone.0114804.t002

Merging of Reads, Length Trimming and Concatenation

Surviving read pairs are grouped into two files, one for "read 1" (R1) and one for "read 2" (R2) sequences. We keep track of the sample of origin for each read, for later use in writing out the final OTU observation table. We trim the read lengths down to a specified cutoff. For R1, we keep the whole read; for R2, we trim down to around 88% of the maximum read length (e.g. 200 bp long for 250 bp long reads), given the overall lower sequencing quality of read 2. This uniformizes the length of both reads to satisfy the global alignabiltiy requirement of UPARSE [33]. At this stage, we keep all reads shorter than this cutoff. A third file is created by concatenating the matching trimmed reads from both reads 1 and read 2. Only reads with matching pairs in both read files are considered. This file will be used for further processing. Finally, a fourth file is created by stitching matching reads with an ambiguous nucleotide character "N" between read 1 and read 2. This file is to be used for taxonomy assignment, where a k-mer based approach will ignore k-mers containing ambiguous bases during the classification steps.

Find OTU Representatives

The reads from the "Merging of Reads, Length Trimming and Concatenation" section are dereplicated, building clusters of reads with 100% similarity using mothur, and annotated with cluster size. To ensure the use of high quality reads when finding the OTU representatives, reads shorter than the cutoff length are discarded, as well as singleton reads. Reads are sorted by cluster size and processed using USEARCH to find *de novo* the OTU representatives using the UPARSE algorithm. This step also removes *de novo* chimeric reads, resulting in a set of OTU representatives of very high sequence quality [33].

Taxonomy Assignment

Out of many possible ways of assigning taxonomy for pared-end 16S reads, we have decided to use k-mer-based methods, since, besides their speed and accuracy, they allow us to query reads with gaps in them (marked as unknown nucleotides) without loss of accuracy. In particular, we use mothur's implementation [9] of the Ribosomal Database Project's naive Bayesian classifier [34]. For this step, we classify the reads against the Greengenes 13_5 database, or in the case of the paired

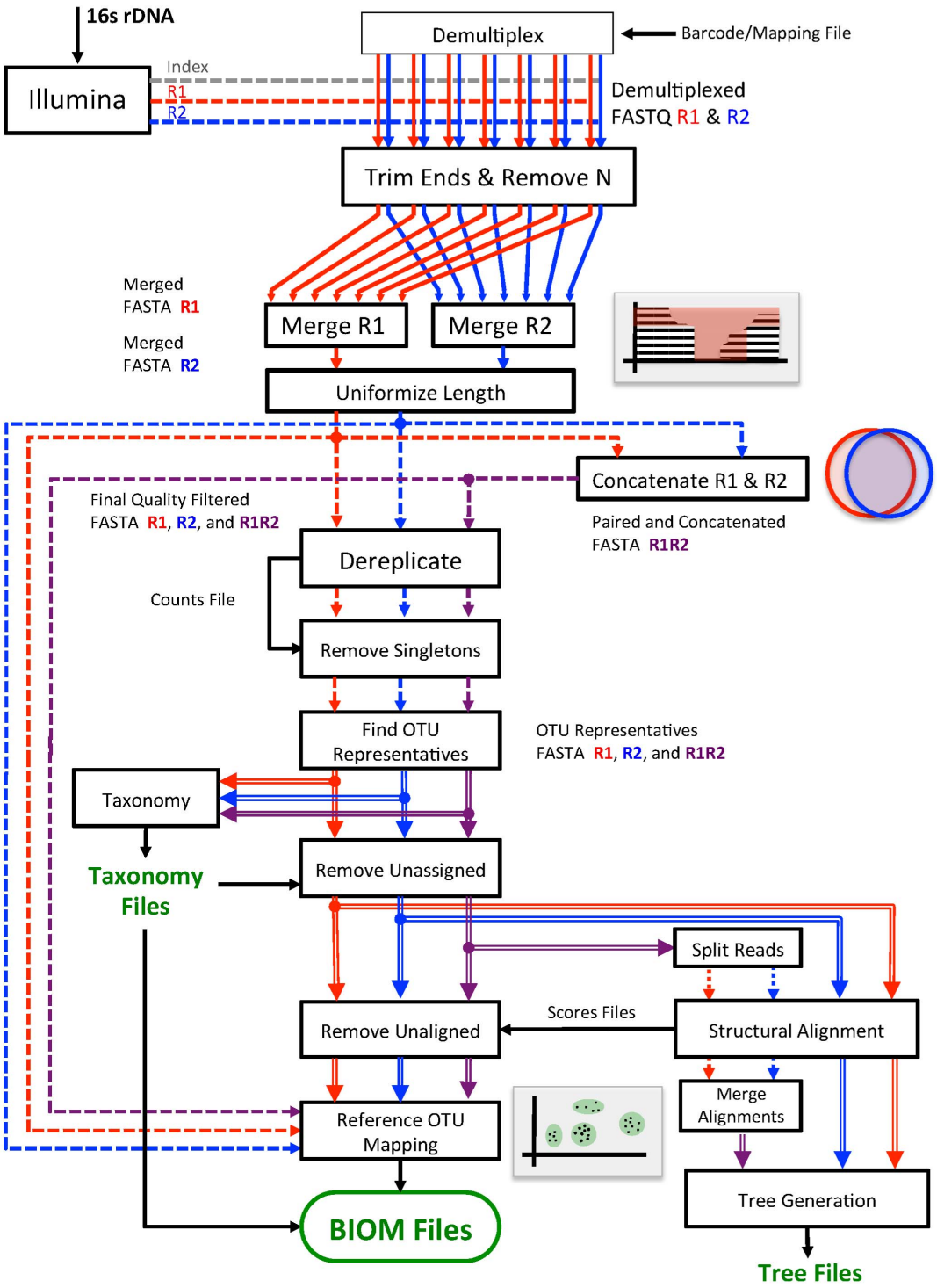

Fig. 4. IM-TORNADO pipeline workflow. Schematic of the IM-TORNADO pipeline workflow.

doi:10.1371/journal.pone.0114804.g004

reads, we look up the OTU representatives in the read library built for this purpose in the ''Merging of Reads, Length Trimming and Concatenation'' section and classify those reads. After classification, we remove fully unclassified reads from the OTU representatives, as they are presumed contaminants (in most cases, these are phiX reads leftover from sequencing).

Structural Alignment

After taxonomy assignment, reads R1 and R2 are aligned using Infernal version 1.1 [28], using a secondary structure model created using the Ribosomal Database Project's (RDP) [35] recommended training set. As per RDP recommendations, options for the aligner are set at ''-g –sub –notrunc'' (with the ''–g'' and ''–notrunc'' options added as a requirement in version 1.1 of Infernal). Structural alignments produces better multiple sequence alignments of 16S rDNA short reads amplicons (see Fig. 1) than other template-based methods [30]. For the paired reads, the reads are split into their R1 and R2 components by looking up the corresponding reads in the R1 and R2 files, and then aligned separately using Infernal. This is necessary because Infernal does not expect a large gap in between the two components, otherwise alignment will fail. After alignment, the scores are analyzed and all alignments with negative scores are removed, as well as the corresponding unaligned reads. These negative scores indicate reads that aligned very poorly to the secondary structure model. For the paired alignments, after removing these low scoring reads, the R1 and R2 components are concatenated back together.

Mapping to OTU Representatives

The reads output in the ''Merging of Reads, Length Trimming and Concatenation'' section are then mapped against their corresponding set of unaligned, clean OTU representatives using USEARCH at 97% sequence identity. The resulting table is then parsed to obtain the different OTU counts per sample, and finally combined with taxonomy information and other metadata into a BIOM-formatted [36] file using the biom-format python package. The unmapped reads are presumed to be either chimeric reads, contaminants or singletons distant to any reported OTU.

Generation of Phylogenetic Trees

In order to obtain the evolutionary relationships between the members of the microbial communities, the clean alignments obtained in the ''Structural Alignment'' section are used to create a phylogeny. Empty columns are removed from the alignments of the OTU representatives, and the trees are calculated using FastTree version 2.1.7 [29], using options ''-nt -gtr -gamma.''

Methods for Validation Using Synthetic Communities

To asses the performance and accuracy of our choice of primers and tools when compared to full-length 16S reads, we used 100 synthetic communities of 20 samples each, for two pairs of primers targeting regions V3–V5 and V6–V9. We ran a simplified version of our pipeline and compared β-diversity metrics and phylogeny metrics to quantify differences with the equivalent result obtained from the corresponding full-length 16S reads.

Input Data

To create our synthetic communities, we started with the complete 16S Greengenes 13_5 database. Using the PrimerProspector tool [37] we selected artificial paired amplicons matching forward primer 357F (CCTACGGGA-GGCAGCAG) and reverse primer 926R (CCGTCAATTCMTTTRAGT) for the V3-V5 primer pair, each of length 250 bp. For the V6-V9 primer pair, we used forward primer 968F (AACGCGAAGAACCTTAC) and reverse primer 1492R (CGGTTACCTTGTTACGACTT), to obtain amplicons also of length 250 bp. For the full length reads, we selected artificial amplicons matching primers 27F (AGAGTTTGATCMTGGCTCAG) and 1492R (CGGTTACCTTGTTACGACTT). We kept reads whose IDs were found both in the paired-end and full-length libraries, as well as having no ambiguous nucleotide characters in their reads. The total number of reads surviving this selection was 142,610 reads for the V3–V5 library, and 143,085 reads for the V6–V9 library. No artificial quality scores were created, hence the reads will be assumed to be perfect.

Synthetic Communities

For each of the primer pairs used for validation, we created 100 synthetic communities, each with 20 samples. Each sample has 2,500 unique IDs randomly picked from the selected read set from the ''Input Data'' section. For each of the IDs we created a set of clone reads whose size was chosen from a random distribution resembling a log-series model of ecological communities [38]. The selected IDs with their respective multiplicities were used to select the corresponding reads from the master full-length and paired-end libraries to create three different sets of communities with identical IDs: full-length, R1 and R2.

Merging of Reads and Concatenation for Validation

As in the ''Merging of Reads, Length Trimming and Concatenation'' section. the R1 and R2 libraries were grouped into a single file for each read. Read lengths in both library R1 and R2 were left unmodified. A third library was created by concatenating the corresponding reads from both R1 and R2, created the paired library.

Finding OTU Representatives for Validation

The three libraries from the ''Merging of Reads and Concatenation for Validation'' section and the full-length library are dereplicated, building clusters of reads with 100% similarity using mothur, and annotated with cluster size. To ensure the use of high quality reads when finding the OTU representatives, singleton reads are discarded. No inspection of read lengths is necessary because, in the case of libraries R1, R2 and paired, the reads are of uniform length by construction, and in the case of the full-length library, the reads are globally alignable by construction (from primers at both ends of the reads). Reads are sorted by cluster size and processed using USEARCH to find *de novo* the OTU representatives using the UPARSE algorithm. This step also removes *de novo* putative chimeric reads, resulting in a set of OTU representatives of very high sequence quality [33].

Structural Alignment for Validation

Libraries R1, R2 and full-length are aligned using Infernal version 1.1 [28], using a secondary structure model created using the Ribosomal Database Project's (RDP) [35] recommended training set. As per RDP recommendations, options for the aligner are set at ''-g –sub –notrunc'' (with the ''–g'' and ''–notrunc'' options added as a requirement in version 1.1 of Infernal). For the paired reads, the reads are split into their R1 and R2 components by looking up the corresponding reads in the R1 and R2 libraries, and then aligned separately using Infernal. This is necessary because Infernal does not expect a large gap in between the two components, otherwise alignment will fail. Alignment scores are not inspected since the reads are free of contaminants by construction.

Mapping to OTU Representatives for Validation

The reads output in the ''Merging of Reads and Concatenation for Validation'' section are then mapped against their corresponding set of unaligned, clean OTU representatives using USEARCH at 97% sequence identity. The resulting table is then parsed to obtain the different OTU counts per sample, and finally combined with taxonomy information and other metadata into a BIOM-formatted [36] file using the biom-format python package. The unmapped reads are presumed to be either chimeric reads or singletons distant to any reported OTU.

Generation of Phylogenetic Trees for Validation

In order to obtain the evolutionary relationships between the members of the microbial communities, the clean alignments obtained in the ''Structural Alignment for Validation'' section are used to create a phylogeny. Empty columns are removed from the alignments of the OTU representatives, and the trees are calculated using FastTree [29]. These trees will be used for comparisons of β-diversity.

These trees are not suitable for direct comparison of the phylogenetic structure between the different libraries, since the OTU picking procedure may choose different reads as representatives, as well as a different number of OTUs. To solve this, we took as a reference the trees created for the full-length reads, created the equivalent libraries with R1, R2 and paired reads and re-run the alignment and calculate the new trees. This way, the trees will have the same number of leaves, same leaf IDs and hence directly comparable.

Comparison of β-Diversity Between the Libraries for Validation

For each of the 100 synthetic communities, we took the resulting BIOM files and phylogenetic trees (for all four libraries, R1, R2, paired and full-length), and calculated the unweighted UniFrac distance matrix [31] between the 20 samples of each community, using QIIME 1.7.0 [8]. We compared the distance matrices of the libraries R1, R2 and paired to the corresponding full-length distance matrix using a Mantel correlation test, as implemented by the R package vegan [39] (see Fig. 3), and assessed the significance of the results using a signed, ranked Wilcoxon test as implemented in R.

Comparison of Phylogenies Between the Different Libraries for Validation

For each of the 100 synthetic communities, we took the full-length tree and the equivalent and comparable trees created using libraries R1, R2 and paired, and calculated the cophenetic distance matrix (which measures the distances between leaves of the phylogenetic tree through branch lengths), as implemented in the R package ape [40]. The distance matrices were compared using a Mantel correlation test, as implemented in the R package vegan [39] (see Fig. 2), and assessed the significance of the results using a signed, ranked Wilcoxon test as implemented in R.

Data Availability

All the data created for the synthetic communities analysis, including reads, resulting OTU tables and scripts, are available for download from the Dryad repository. The data DOI is doi: 10.5061/dryad.fm67n.

Methods for Validation Using a Sample-Modeled Synthetic Mock Community

As another validation of our approach, we used an existing set of read libraries from stool samples to create a synthetic mock community based on real gastrointestinal microbiome data, for a more realistic case testing.

Ethics Statement and Sample Collection

Subjects were consented under IRB #11-002697, which was approved and reviewed by the Mayo Clinic Institutional Review Board. Written consent was provided by all subjects. A total of 20 multiple sclerosis patients were consented under Mayo Clinic IRB #11-002697 and stool samples were collected either in clinic or using a specimen mailer tube. Upon arrival, samples were frozen at −80℃ until they were ready to be processed.

DNA Library Preparation

DNA extraction was carried out according to the Manual of Procedures on the Human Microbiome Project website http://www.hmpdacc.org using physical and chemical lysis with a FastPrep-24 (MP Biomedicals, Santa Ana, CA) and PowerSoil Extraction Kit (MoBio, Carlsbad, CA). Amplification targeted V3 and V5 regions, whose performance has been previously characterized [41, 42], using primers 357F (AATGATACGGCGACCACCGAGATCTACACTATGGTAATTG-TCCTACGGGAGGCAGCAG) and 926R (CAAGCAGAAGACGGCATACGAGA-T-NNNNNNNNNNN-AGTCAGTCAGCCCCGTCAATTCMTTTRAGT) with barcodes 55–76 from Caporaso *et al.* [10]. PCR was run through 34 cycles of 98℃ for 15 seconds, 70℃ for 20 seconds, 72℃ for 15 seconds, in accordance to the temperatures from the polymerase hotstart mix for Kapa Hi-Fi (Kapa Biosystems, Boston, MA). Electrophoresis was then carried out using a 2200 TapeStation (Agilent Technologies, Santa Clara, CA). After verifying there were no other large bands besides the amplicon, purification was carried out using magnetic beads. Concentrations for each sample were then diluted to 10 nM and pooled for sequencing. Amplicons were then sequenced on a MiSeq sequencer (Illumina, San Diego, CA) using a 500 cycle kit and custom read1 (TATGGTAATTGTCCT-ACGGGAGGCAGCAG), read2 (AGTCAGTCAGCCCCGTCAATTCMTTTR-AGT), and index (ACTYAAAKGAATTGACGGGGCTGACTGACT) sequencing primers.

Sample Modeling

To create the synthetic full-length 16S samples, we mapped five of the short-read libraries onto the full Greengenes 13_5 database using USEARCH's ''-usearch_global'' option [43], selecting the database reads that matched at 97% sequence identity. Using PrimerProspector [37] we extracted 250 bp long V3 and V5 paired reads from the matches, and also created an almost-full-length 16S library by trimming their read lengths down to 1300 bp long.

OTU Representatives, Alignment and Phylogeny

As in the previous synthetic library processing, we dereplicated each library, sorted by cluster size, discarded singleton clusters and picked OTUs at 97% sequence identity using USEARCH [33]. Then the libraries were mapped onto the OTU

representatives using USEARCH to obtain the OTU sizes. We multiple sequence aligned the OTU representatives to a model secondary structure of the 16S gene using Infernal 1.1 [28] and created a phylogenetic tree using FastTree 2.1.7 [29]. All of the above information was consolidated into BIOM files using IM-TORNADO scripts.

Comparison of β-Diveristy

For each of the BIOM files correspoonding to each library (R1, R2, paired and full-length), we calculated the corresponding unweighted UniFrac distance matrix using QIIME 1.7.0 [8]. We compared the matrices of the three short libraries (R1, R2 and paired) against the full-length UniFrac matrix using a Mantel correlation test, as implemented in the R package vegan [39].

Data Availiability

All the data created for the sample synthetic mock community, including original reads and synthetic reads, resulting OTU tables and scripts, are available for download at the Dryad repository. The data DOI is doi: 10.5061/dryad.fm67n.

Discussion

We have shown that the use of paired reads improves performance over R1 or R2 alone. Moreover, significance tests revealed that in all cases R1 and R2 together almost always provided a better reflection of the results from full length 16S rDNA. By itself, this fact is unsurprising since R1 or R2 both use only a subset of the information available to paired reads, which are made up of both R1 and R2 together. On the other hand, what may seem more surprising is that paired read approaches are not utilized as the norm. However, the overall gains from using paired reads is moderate when considering the enormous amount of book-keeping and tracking involved (shown in Fig. 4). It is clear that the travails of making maximal use of paired reads outweighs the benefit for many microbiome researchers. It is our hope that with the introduction and public availability of IM-TORNADO that microbiome researchers will be able to overcome the logistical barrier to use of non-overlapping paired-end 16S rDNA reads.

It is our view that the most valuable tools are also the most widely usable. Now more than ever, community standards within the field of microbiome research are becoming increasingly important as the field (and the data) continues to grow at a rapid pace. One of the important design principles of IM-TORNADO is our adherence to already widespread analyses in microbiome studies. This allows results from IM-TORNADO to be readily compared to results from pipelines such as QIIME or mothur [8, 9]. Furthermore, we utilize the BIOM format popularized by QIIME which allows for our results to be manipulated by existing tools that accept the BIOM format, such as mothur and phyloseq [44]. The scripts and tools

for running IM-TORNADO are available at http://sourceforge.net/projects/imtornado.

Finally, as an important consideration in future studies, the ability to use non-overlapping paired end 16S rDNA reads means added flexibility to experimental design. This comes in direct contrast to overlapping paired read designs which require primers to be adjacent to each other. While overlapping primer design has potential advantages (see Supplemental Materials) and could in principle use nearly the same amount of information as non-overlapping reads, with a small number of bases given to assemble the reads, primer choice ideally would vary between studies due to emphasis on different taxa in particular environments. Creating a means for the more holistic analysis of non-overlapping paired reads allows for greater flexibility in primer choice, allowing researchers to target a larger variety of regions to maximize sensitivity and specificity for a particular environment without using only adjacent regions or resorting to single end read analysis. This applies not only to 16S applications. For example, studies using the 18S rDNA gene [45] will benefit from a paired-end sequencing approach. And these benefits extend to any other structurally alignable RNA molecule. Being able to utilize non-overlapping reads also ameliorates issues arising from poor sequencing quality and shorter than expected read lengths after quality trimming by enabling analysis of data from lower quality sequencing runs.

Conclusions

We implemented a workflow for computing taxonomy, OTUs, and phylogeny from non-overlapping 16S rDNA paired-end reads. As a quality metric for our particular implementation, we demonstrate the differences between the use of single versus non-overlapping paired-end reads. This is meant to address another possible methodology for data analysis in studies that utilize primer designs where there are non-overlapping paired-end reads [12, 15].

As shown by the results of our tests on synthetic mock datasets in Tables 1 and 2, as well as Figs. 2 and 3, including both reads in the computational analysis of taxonomic classification, phylogeny, and β-diversity produces results that have greater agreement with those from full length 16S rDNA then when using either single-end read alone. Given the implied increase in accuracy when analyzing environmental data, it is evident that paired-end analysis should be used when possible in these cases.

Supporting Information

S1 Table. Taxonomy comparison. Accuracy of taxonomy assignments for different read lengths, with errors and read trimming. The table shows the effect of base errors in the accuracy of the taxonomy assignment process (Paired, with errors, average 15.6 errors per read pair, standard deviation 3.7). It also shows the improvement in accuracy after applying the read trimming step pf the pipeline

(Paired, trimmed, with errors, average 6.0 errors per read pair, standard deviation 2.9, average read pair length 438.7 bp, standard deviation 28 bp).
doi:10.1371/journal.pone.0114804.s001 (PDF)

S2 Table. Comparison of OTU counts for synthetic mock communities. Comparison of OTU counts at 97% sequence identity, averaged for the 100 synthetic mock communities used for validation. From the table we can observe the trend of Full length reads showing comparable or higher OTU counts than the short read types, and among these short reads, R1 shows the higher OTU counts.
doi:10.1371/journal.pone.0114804.s002 (PDF)

S1 Text. Supplemental Information for "IM-TORNADO: A Tool for Comparison of 16S Reads from Paired-End Libraries".
doi:10.1371/journal.pone.0114804.s003 (PDF)

S1 Code. Source code for the IM-TORNADO pipeline, version 2.0.0. This and subsequent versions of the pipeline are available in the SourceForge repository at http://sourceforge.net/projects/imtornado.
doi:10.1371/journal.pone.0114804.s004 (GZ)

Acknowledgments

The authors thank Bruce Eckloff and Álvaro Hernández for discussion and sequencing support. We also thank Melissa Cregger, Andrés Gómez, Marina Walther-Antonio and Michael Mundy for testing prototypes of this workflow, Mat Wiepert for IT support, and Jay Doughty, and Janet Yao for useful discussions.

Author Contributions

Conceived and designed the experiments: PJ NC. Performed the experiments: PJ NC. Analyzed the data: PJ KK XC JB. Contributed reagents/materials/analysis tools: AM BW HN JK. Wrote the paper: PJ JK NC.

References

1. **Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, et al.** (1985) Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc Natl Acad Sci USA 82: 6955–6959.
2. **Gevers D, Knight R, Petrosino JF, Huang K, McGuire AL, et al.** (2012) The Human Microbiome Project: A Community Resource for the Healthy Human Microbiome. PLOS Biol 10: e1001377.
3. **Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, et al.** (2010) Forensic identification using skin bacterial communities. Proc Natl Acad Sci USA 107: 6477–6481.
4. **Di Rienzi SC, Sharon I, Wrighton KC, Koren O, Hug LA, et al.** (2013) The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. eLife 2: e01102.
5. **Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, et al.** (2013) Insights into the phylogeny and coding potential of microbial dark matter. Nature 499: 431–437.

6. **Hamady M, Knight R** (2009) Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. Genome Res 19: 1141–1152.

7. **Gevers D, Pop M, Schloss PD, Huttenhower C** (2012) Bioinformatics for the human microbiome project. PLOS Comput Biol 8: e1002779.

8. **Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, et al.** (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7: 335–336.

9. **Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, et al.** (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75: 7537–7541.

10. **Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, et al.** (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J 6: 1621–1624.

11. **Degnan PH, Ochman H** (2012) Illumina-based analysis of microbial community diversity. ISME J 6: 183–194.

12. **Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, et al.** (2012) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat Methods 10: 57–59.

13. **Liu Z, DeSantis TZ, Andersen GL, Knight R** (2008) Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. Nucleic Acids Res 36: e120.

14. **Soergel DAW, Dey N, Knight R, Brenner SE** (2012) Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. ISME J 6: 1440–1444.

15. **Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, et al.** (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci USA 108: 4516–4522.

16. **Werner JJ, Zhou D, Caporaso JG, Knight R, Angenent LT** (2012) Comparison of Illumina paired-end and single-direction sequencing for microbial 16S rRNA gene amplicon surveys. ISME J 6: 1273–1276.

17. **Walther-António MRS, Jeraldo P, Berg Miller ME, Yeoman CJ, Nelson KE, et al.** (2014) Pregnancy's Stronghold on the Vaginal Microbiome. PLOS ONE 9: e98514.

18. **Kang SS, Jeraldo PR, Kurti A, Miller MEB, Cook MD, et al.** (2014) Diet and exercise orthogonally alter the gut microbiome and reveal independent associations with anxiety and cognition. Mol Neurodegener 9: 36.

19. **Miller CS, Handley KM, Wrighton KC, Frischkorn KR, Thomas BC, et al.** (2013) Short-read assembly of full-length 16S amplicons reveals bacterial diversity in subsurface sediments. PLOS ONE 8: e56018.

20. **Bartram AK, Lynch MD, Stearns JC, Moreno-Hagelsieb G, Neufeld JD** (2011) Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end illumina reads. Appl Environ Microbiol 77: 3846–3852.

21. **Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD** (2012) PANDAseq: paired-end assembler for Illumina sequences. BMC Bioinformatics 13: 31.

22. **Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD** (2013) Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol 79: 5112–5120.

23. **Robertson CE, Harris JK, Wagner BD, Granger D, Browne K, et al.** (2013) Explicet: graphical user interface software for metadata-driven management, analysis and visualization of microbiome data. Bioinformatics 29: 3100–3101.

24. **DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, et al.** (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72: 5069–5072.

25. **DeSantis TZ, Hugenholtz P, Keller K, Brodie EL, Larsen N, et al.** (2006) NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. Nucleic Acids Res 34: W394–W399.

26. **Nawrocki EP, Kolbe DL, Eddy SR** (2009) Infernal 1.0: inference of RNA alignments. Bioinformatics 25: 1335–1337.

27. **Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, et al.** (2010) PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics 26: 266–267.

28. **Nawrocki EP, Eddy SR** (2013) Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29: 2933–2935.

29. **Price MN, Dehal PS, Arkin AP** (2010) FastTree 2 – approximately maximum-likelihood trees for large alignments. PLOS ONE 5: e9490.

30. **Sipos M, Jeraldo P, Chia N, Qu A, Dhillon AS, et al.** (2010) Robust computational analysis of rRNA hypervariable tag datasets. PLOS ONE 5: e15220.

31. **Lozupone CA, Hamady M, Kelley ST, Knight R** (2007) Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. Appl Environ Microbiol 73: 1576–1585.

32. **Bolger AM, Lohse M, Usadel B** (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120.

33. **Edgar RC** (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods 10: 996–998.

34. **Wang Q, Garrity GM, Tiedje JM, Cole JR** (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73: 5261–5267.

35. **Cole JR, Wang Q, Cardenas E, Fish J, Chai B, et al.** (2009) The ribosomal database project: improved alignments and new tools for rrna analysis. Nucleic Acids Res 37: D141–D145.

36. **McDonald D, Clemente JC, Kuczynski J, Rideout J, Stombaugh J, et al.** (2012) The Biological Observation Matrix (BIOM) format or: how i learned to stop worrying and love the ome-ome. GigaScience 1: 7.

37. **Walters WA, Caporaso JG, Lauber CL, Berg-Lyons D, Fierer N, et al.** (2011) PrimerProspector: de novo design and taxonomic analysis of barcoded polymerase chain reaction primers. Bioinformatics 27: 1159–1161.

38. **Fisher RA, Corbet AS, Williams CB** (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. J Anim Ecol 12: pp.42–58.

39. **Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, et al.** (2013) vegan: Community Ecology Package. URL http://CRAN.R-project.org/package=vegan. R package version 2.0–9.

40. **Paradis E, Claude J, Strimmer K** (2004) APE: Analyses of phylogenetics and evolution in R language. Bioinformatics 20: 289–290.

41. **Yu Z, Morrison M** (2004) Comparisons of different hypervariable regions of rrs genes for use in fingerprinting of microbial communities by PCR-denaturing gradient gel electrophoresis. Appl Environ Microbiol 70: 4800–4806.

42. **Jumpstart Consortium Human Microbiome Project Data Generation Working Group** (2012) Evaluation of 16S rDNA-based community profiling for human microbiome research. PLOS ONE 7: e39315.

43. **Edgar RC** (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26: 2460–2461.

44. **McMurdie PJ, Holmes S** (2013) phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLOS ONE 8: e61217.

45. **Lie AAY, Liu Z, Hu SK, Jones AC, Kim DY, et al.** (2014) Investigating microbial eukaryotic diversity from a global census: insights from a comparison of pyrotag and full-length sequences of 18S rRNA genes. Appl Environ Microbiol 80: 4363–4373.