AEM
Journals.ASM.org

Performance Comparison of Illumina and Ion Torrent Next-Generation Sequencing Platforms for 16S rRNA-Based Bacterial Community Profiling

Stephen J. Salipante,[a] Toana Kawashima,[a] Christopher Rosenthal,[a] Daniel R. Hoogestraat,[a] Lisa A. Cummings,[a] Dhruba J. Sengupta,[a] Timothy T. Harkins,[b] Brad T. Cookson,[a,c] Noah G. Hoffman[a]

Department of Laboratory Medicine, University of Washington, Seattle, Washington, USA[a]; Life Technologies, South San Francisco, California, USA[b]; Department of Microbiology, University of Washington, Seattle, Washington, USA[c]

High-throughput sequencing of the taxonomically informative 16S rRNA gene provides a powerful approach for exploring microbial diversity. Here we compare the performances of two common "benchtop" sequencing platforms, Illumina MiSeq and Ion Torrent Personal Genome Machine (PGM), for bacterial community profiling by 16S rRNA (V1-V2) amplicon sequencing. We benchmarked performance by using a 20-organism mock bacterial community and a collection of primary human specimens. We observed comparatively higher error rates with the Ion Torrent platform and report a pattern of premature sequence truncation specific to semiconductor sequencing. Read truncation was dependent on both the directionality of sequencing and the target species, resulting in organism-specific biases in community profiles. We found that these sequencing artifacts could be minimized by using bidirectional amplicon sequencing and an optimized flow order on the Ion Torrent platform. Results of bacterial community profiling performed on the mock community and a collection of 18 human-derived microbiological specimens were generally in good agreement for both platforms; however, in some cases, results differed significantly. Disparities could be attributed to the failure to generate full-length reads for particular organisms on the Ion Torrent platform, organism-dependent differences in sequence error rates affecting classification of certain species, or some combination of these factors. This study demonstrates the potential for differential bias in bacterial community profiles resulting from the choice of sequencing platform alone.

There has been increasing interest in exploring the bacterial communities that populate environmental (1) and biological (2) specimens. One common approach for classifying microbial species present in a sample, rooted in previous cloning-based methods (3), is to PCR amplify a fraction of the taxonomically informative 16S rRNA gene and subject this product to next-generation DNA sequencing, enabling the classification of individual reads to specific taxa. As opposed to shotgun sequencing of genomic DNA extracted from a sample (4), wherein random fragments of bacterial genomes (and, potentially, contaminating DNA from host species or other organisms present) are sequenced and classified, 16S rRNA amplicon sequencing can be targeted specifically against bacteria, does not require the availability of reference genome sequences, and can be employed in cases where only trace amounts or poor-quality bacterial DNA templates are available (5, 6). For these reasons, 16S rRNA amplicon sequencing has found application in a wide range of metagenomic profiling studies, especially in those studies where speed or limited input material is a concern.

General differences among next-generation sequencing platforms that may be used for this purpose, including relative turnaround times, per-base sequencing costs, read lengths, and accuracies, have been discussed elsewhere (7–9). Historically, many 16S rRNA amplicon sequencing experiments were performed by using 454 (Roche) massively parallel pyrosequencing (10), both because it was the first commercially available system and because it later offered the longest read lengths, permitting interrogation of a larger and consequently more informative fraction of the 16S rRNA gene (11). However, this platform is currently being phased out by the manufacturer. At present, the two highest-selling next-generation sequencing technologies are relatively inexpensive "benchtop" sequencers developed by Illumina (12) and Ion Torrent (13), both of which have developed sufficiently long reads as to now permit data generation from a highly informative fraction of the 16S rRNA gene. Accordingly, the Illumina MiSeq and Ion Torrent Personal Genome Machine (PGM) platforms are increasingly being used for 16S rRNA-mediated surveys of bacterially diverse populations (1, 2, 14–19). Although some evaluation of these technologies against 454 sequencing have been performed for various purposes (17, 20), a detailed and direct comparison between the Illumina and Ion Torrent platforms for the specific task of 16S rRNA amplicon sequencing has not yet been described.

Although both the Illumina and Ion Torrent platforms sequence DNA by monitoring the addition of nucleotides during DNA synthesis, they operate on different principles, which could affect their performance in this application. DNA fragments are prepared for Illumina sequencing by isothermic "bridge PCR," which simultaneously amplifies single DNA molecules and cova-

Received 30 July 2014 Accepted 23 September 2014
Published ahead of print 26 September 2014
Editor: F. E. Löffler
Address correspondence to Stephen J. Salipante, stevesal@uw.edu, or Noah G. Hoffman, ngh2@uw.edu.
Supplemental material for this article may be found at http://dx.doi.org/10.1128/AEM.02206-14.

doi:10.1128/AEM.02206-14

lently links amplicons to a solid substrate within a constrained physical location in order to form randomly arrayed "clusters." Clusters are then sequenced through repeated cycles of single-base extension using a mixture of 4 fluorescently labeled, reversible chain terminators (one for each nucleotide); imaging to ascertain the identity of the incorporated base; and chemical cleavage of the terminator to enable further cycles. Illumina sequencing also supports sequencing of templates from both ends (i.e., "paired-end sequencing") (12). In contrast, Ion Torrent sequencing initially prepares templates by using emulsion PCR (12): PCR reagents, primer-coated particles, and a low concentration of template fragments are combined with oil; emulsified to form picoliter-scale microreactions; and subjected to thermal cycling to achieve clonal amplification of single DNA molecules on the surfaces of individual particles. Particles are then deposited into individual, nanowell chambers on a semiconductor sequencing chip (13). Individual nucleotides are cyclically introduced in the presence of DNA polymerase, and successful incorporation of a particular nucleotide is registered by the release of hydrogen ions. Unlike Illumina chemistries, multiple nucleotides may be incorporated during a single sequencing cycle, and it is recognized that errors in quantitating the length of homopolymer repeats are common (7–9).

Here we evaluate the relative performance characteristics of the Illumina and Ion Torrent sequencing platforms for the express purpose of bacterial community profiling by 16S rRNA amplicon sequencing. We examined both a synthetic community of 20 bacterial species and a collection of primary, human-derived microbiological specimens through sequencing of 16S rRNA variable regions 1 and 2 (V1-V2), an ~360-bp span useful for the identification of bacteria to the species level (5) that can be fully covered (albeit by different strategies) on both platforms.

MATERIALS AND METHODS

Samples and library preparation. A synthetic mixture of genomic DNA comprising 20 bacterial species (Microbial Mock Community B, catalog number HM-782D) was obtained from BEI Resources (Manassas, VA). This mixture contains DNA from each organism added at equimolar concentrations of 16S rRNA operons at 100,000 copies per organism per microliter (see Table S1 in supplemental material). DNA was extracted from human-derived microbiological specimens by using a High Pure PCR template preparation kit (Roche). Sequencing libraries were prepared by PCR amplification using AmpliTaq DNA polymerase (Applied Biosystems) with 3 mM $MgCl_2$. PCR conditions for construction of all sequencing libraries consisted of 1 cycle of 95°C for 10 min; 28 cycles of 95°C for 30 s, 60°C for 30 s, and 72°C for 1 min 15 s; and then 1 cycle of 72°C for 10 min. PCR products were purified by using 0.7 volumes of AMPure beads (Agencourt), eluted in low Tris-EDTA (TE) buffer, and quantified by using a Qubit dsDNA HS (high sensitivity) kit (Life Technologies).

Primers (PAGE purified; Integrated DNA Technologies) directed against the 16S rRNA V1-V2 region (derivatives of 8F and the reverse complement of 557F [21] incorporating deoxyinosine to improve cross-species binding) were designed to incorporate either Illumina- or Ion Torrent-compatible sequencing adaptors (see Tables S2 to S4 in the supplemental material). For MiSeq library preparation, primer Universal_forward was used in conjunction with a reverse primer incorporating a sample-specific 8-bp barcode sequence. For Ion Torrent sequencing, DNA was amplified in two separate reactions (using either Universal_ion_forward plus a barcoded reverse primer or Universal_ion_reverse plus a barcoded forward primer), in which sequencing adaptors were incorporated on opposite ends of the amplicon, enabling sequencing across the full length of 16S rRNA variable regions 1 and 2 in either orientation. Ion Torrent barcodes were 10- to 12-bp sequences optimized for maximal error correction, average sequence content, and nucleotide flow order (Ion Xpress barcodes; Life Technologies). These independent sequencing orientations were designated "reverse" or "forward," in accordance with the biological orientation of the 16S rRNA gene.

Sequencing. Illumina sequencing was performed by using a MiSeq platform (Illumina) operating Real Time Analysis (RTA) software, version 1.17.28. Paired-end sequencing was done by using custom primers (see Table S2 in the supplemental material) (15) and a 500-cycle sequencing kit (version 2), according to the manufacturer's instructions. Amplicon sequencing was carried out in the presence of either the bacterial whole-genome shotgun sequence or 7% PhiX control (Illumina) to allow proper focusing and matrix calculations. Raw data processing and run demultiplexing were performed by using on-instrument software.

Templating, enrichment, and quantification for Ion Torrent sequencing were performed by using the One-Touch 2 and One-Touch ES systems (Life Technologies) according to the manufacturer's instructions (part number 4479878). Sequencing was performed on an Ion PGM (Life Technologies), using 400-bp sequencing kits (part number 4482002) according to the manufacturer's instructions or using the same reagents with an alternative flow order (TGCTCAGAGTACATCACTGCGATCTCGAG ATG) (see Text S1 in the supplemental material). The default, generic Ion Torrent flow order is designed with a particular efficiency of extension-versus-phase correction effect tradeoff and is optimized for genomes with near-even base usage. This alternative flow order results in more aggressive phase correction, making it better for sequencing of difficult secondary structures or templates with significant bias in base usage, but is less efficient at overall extension (C. C. Lee, personal communication). 314 v2 or 318 v2 chips were used for sequencing of various specimens. Base calling and run demultiplexing were performed by using TorrentServer software, version 3.6.2, with default parameters for the General Sequencing application (-Basecaller–trim-qual-cutoff 15, –trim-qual-window-size 30, –trim-adapter-cutoff 16, with no base recalibration applied).

Data processing. Forward and reverse reads from Ion Torrent sequencing were first processed by discarding reads of <100 bp. Reads were next run length encoded (22), by which each homopolymer was represented by a single nucleotide and the length of the homopolymer tract was recorded. This process optimizes alignments between homopolymer tracts with different lengths, improving the sensitivity for detecting primer sequences by minimizing pairwise alignment differences attributable to disparities in homopolymer runs. Detection of run-length-encoded primer sequences was performed by using the Smith-Waterman alignment algorithm with ssearch36 (23). PCR primer sequences proximal to the direction of sequencing were removed; reads in which proximal primer sequences could not be detected were discarded. These reads subjected to proximal primer trimming were used for the calculation of read length distributions after reversing the run length encoding. Run-length-encoded reads were then further processed to remove PCR primer sequences distal to the direction of sequencing; again, sequences in which primer sequences could not be detected were discarded to ensure that the remaining reads spanned the entire PCR amplicon. Reads from the "reverse" orientation were then reverse complemented. Full-length, primer-trimmed sequences were used to calculate error rates after reversing run length encoding. Finally, for the purposes of data reduction, clustering was performed by using usearch v6.0.307 with an identity threshold of 0.995 using the "-cluster_fast" command, and clusters of <3 sequences were discarded. Clustering was performed for forward and reverse reads independently as well as for forward and reverse reads combined into the same input file.

Illumina sequences were analogously processed, as described in full previously (24), with minor modifications. Briefly, paired-end Illumina reads were self-assembled by using PANDAseq 2.4 (25) to produce sequences spanning the full length of the amplicon with PCR primer sequences removed. Successfully self-assembled reads were used to calculate read length distributions (Fig. 1A) and error rates. For consistency with

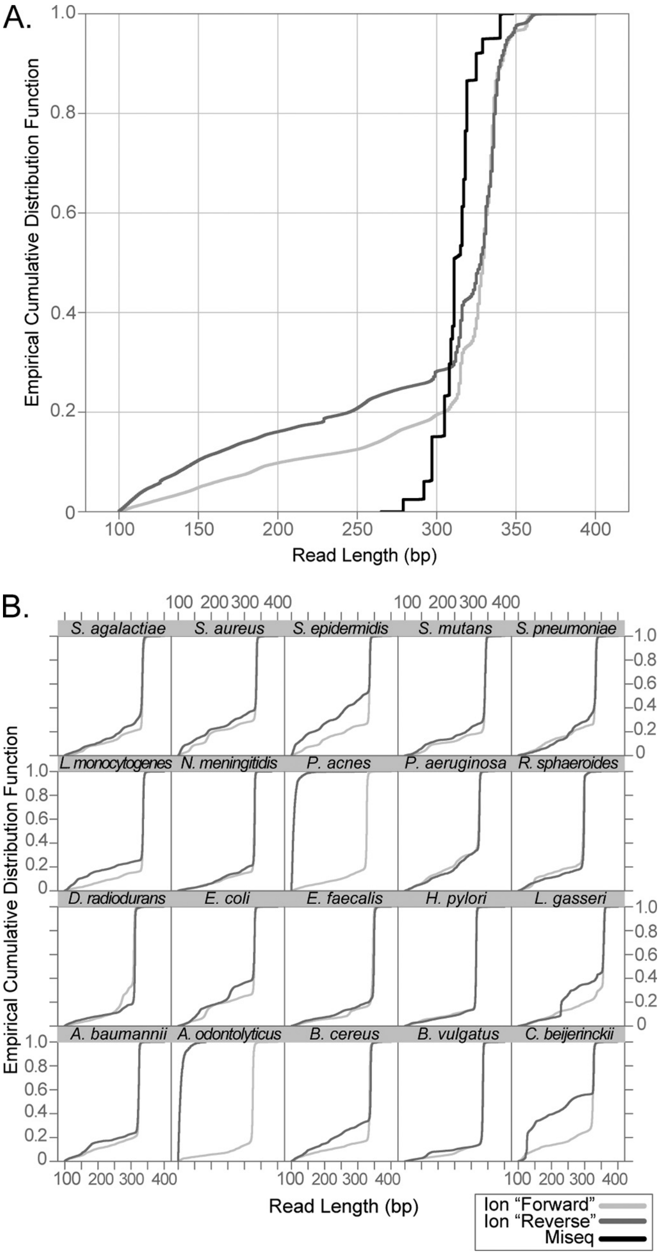

FIG 1 16S rRNA read lengths in MiSeq and Ion Torrent sequence data. (A) Comparison of absolute sequence lengths for assembled paired-end MiSeq reads and unidirectional reads from the Ion Torrent PGM platform derived from a mock-community mixture of 20 bacterial DNAs. Read length is plotted as a cumulative distribution function. A function for which all sequences were uniformly full length would be represented as a vertical line at ~320 bp (exact size is dependent on the species). (B) Ion Torrent sequence length for individual organisms, displayed as described above for panel A. The key applies to both panels. *L. gasseri*, *Lactobacillus gasseri*; *D. radiodurans*, *Deinococcus radiodurans*; *E. faecalis*, *Enterococcus faecalis*; *B. cereus*, *Bacillus cereus*; *C. beijerinckii*, *Clostridium beijerinckii*.

Ion Torrent read processing, the remaining sequences were run length encoded and clustered by using usearch, as described above.

Error rate calculations. To tally and categorize errors observed among reads generated from the mock-community specimen, we first constructed a database of reference sequences representing each 16S rRNA gene allele from the whole-genome assemblies of each organism in the mixture, as indicated in the product literature. Reference sequences and accompanying annotations are provided in Tables S5 and S6 in the supplemental material. Each sequence read was classified as one of the 20 species represented in the mock community on the basis of the highest-scoring pairwise alignment to the library of reference sequences. To calculate read length distributions, reads ≥100 bp in length that aligned successfully with the appropriate proximal PCR primer relative to the direction of sequencing were further cropped to 400 bp to remove artifactual sequences extending beyond the distal PCR primer and then aligned with the library of reference sequences for the mock community by using ssearch36 with default gap opening and extension parameters to permit assignment of each read to one of the 20 species represented in the mock community (Fig. 1).

For error rate calculations, pairwise alignments were first performed by aligning full-length, primer-trimmed (i.e., both forward and reverse primer sequences were located and removed), and run-length-encoded reads against the 20-organism reference library by using ssearch36 with a gap extension penalty of 3 and a gap-open penalty of 8. These provisions result in an alignment model that heavily favors the correct alignment of homopolymer tracts of differing lengths, because it eliminates nucleotide tract length from consideration in the alignment model. Errors were identified in pairwise alignments of each experimentally generated sequence relative to the highest-scoring reference sequence. Details of the procedures for primer trimming, pairwise alignments of run-length-encoded reads, error tallies, and classifications were described previously (5).

Separately, in order to enable a more general description of errors observed and to statistically assess comparative differences, the error rate was modeled for raw Ion Torrent reads in forward and reverse orientations and for individual self-assembled MiSeq reads by using a negative binomial regression with a generalized estimation equation incorporating robust variance based on an exchangeable working correlation matrix, to account for multiple observations across error types per base length (26). Reads classified as either *Actinomyces odontolyticus* or *Propionibacterium acnes* were excluded from the model due to the very low numbers of reads representing these species in the Ion Torrent data, in the reverse orientation. Potential two-way and three-way interactions between the three covariates (sequencing platform/orientation, organism, and error type) were examined and were adjusted for in the final full multivariable model due to statistical significance. All expected estimates were expressed as error rates, and *P* values were 2 sided with a significance level at an α value of 10. These analyses were carried out by using the geem function provided by the geeM package for R, version 3.0.2 (27).

Classification of reads from clinical specimens. Consensus sequences were classified by sequence identity as described previously (5, 24). Briefly, experimentally generated sequence data ("query" sequences) were compared to a library of representative reference sequences downloaded from the Ribosomal Database Project (RDP 10, update 32) by using BLAST. Classification of each query sequence was performed by assigning one or more species-level taxonomic names represented among reference sequences with ≥99% identity and 95% alignment coverage. Query sequences with no BLAST hits meeting the 99% identity threshold were identified with the label "≤99%." Ambiguity among closely related species was signified by assigning a compound name (e.g., *Streptococcus constellatus*/*S. intermedius*). To simplify comparison among specimens, query sequences that were classified as any combination of certain closely related organisms were renamed as follows: (i) any member of the genus *Enterobacter*, *Escherichia*, or *Shigella* was renamed *Enterobacteriaceae*; (ii) any combination of *Streptococcus mitis*, *S. oralis*, *S. pneumoniae*, or *S. pseudopneumoniae* was renamed *Streptococcus mitis* group, and any combination of *Streptococcus salivarius*, *S. vestibularis*, or *S. thermophiles* was renamed *Streptococcus salivarius* group; (iii) any combination of *Staphylococcus capitis*, *S. caprae*, *S. epidermidis*, or *S. saccharolyticus* was renamed *Staphylococcus epidermidis* group. The relative abundance of organisms within a specimen was calculated by weighting each denoised read or cluster centroid according to the number of reads included in the corresponding cluster. Reads were downsampled so that specimens had the same number of reads across platforms, with a maximum of 100,000 reads allotted per specimen.

Data availability. Sequence reads for this project are available from the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under study accession number SRP040453.

RESULTS

Sequencing. Sequencing libraries for both platforms were prepared by PCR amplification using forward primers incorporating (5′-to-3′) Ion Torrent or Illumina sequencing adaptors, sample-specific barcodes, and a "universal" template-specific sequence (28) as well as reverse primers similarly incorporating (5′-to-3′) Ion Torrent or Illumina sequencing adaptors and a universal bacterial sequence.

Illumina sequencing was performed by using paired-end 250-bp reads; that is, 250 bp of sequence data was generated from each end of the approximately 320- to 350-bp V1-V2 amplicons. At the time when the experiments were performed, this was the longest read length offered by Illumina (subsequent improvements in sequencing chemistries have increased this range to 300 bp at present), and although the length of these individual reads does not permit the entire V1-V2 amplicon to be fully sequenced from each direction independently, the entire target can be covered by using paired, bidirectional reads with an expected central overlap of ~150 to 180 bp.

In contrast, Ion Torrent chemistries presently offer longer continuous sequencing reads (up to 400 bp) that are able to fully span the 16S rRNA V1-V2 target region, but paired-end sequencing is not supported as a standard protocol on this platform. We performed full-length sequencing of the V1-V2 target region in both orientations separately (in the "forward" direction, from the 5′ to the 3′ end of the gene, or in the "reverse" direction, from the 3′ to the 5′ end), using independently prepared sequencing libraries that incorporated adaptors at the appropriate end of the fragment. In subsequent analyses, these reads from different orientations were considered either individually or in combination (full-length reads combined into the same data set, without self-assembly).

Organism-specific read truncation by Ion Torrent. We initially generated 16S rRNA amplicon sequence data from a defined bacterial community encompassing 20 species distributed throughout the eubacterial tree of life and for which 16S rRNA reference sequences are available (29). We used a formulation of this community containing equimolar concentrations of rRNA operons from each species. A total of 715,306 reads were obtained from the Illumina platform for this specimen, 675,175 (94.38%) of which were successfully self-assembled into full-length sequences. For the Ion Torrent platform, 1,928,511 reads were obtained from the forward orientation, and 1,782,317 reads were obtained from the reverse orientation; 684,003 (35.5%) and 567,362 (31.8%) reads, respectively, passed our quality filters for primer trimming. We assigned species names to each read according to its highest-scoring pairwise alignment against each 16S rRNA gene allele represented among the panel of bacteria included in the mock community and evaluated the cumulative distribution of read lengths observed for each species (Fig. 1A).

Using the Ion Torrent platform, we observed marked differences among the length distributions of reads representing different species (Fig. 1B), and for certain organisms (most notably *S. epidermidis*, *P. acnes*, and *A. odontolyticus*), there were substantial differences in read length distributions obtained with sequencing in opposite orientations. We inactivated all on-instrument quality filtering and quality control read truncation and found that this decreased the overall quality of reads without changing the observed distribution of read lengths (data not shown). Use of the standard sequencing flow order, rather than the flow order optimized for abnormal secondary structures and/or base composition, exaggerated the degree of read truncation (not shown). We therefore elected to use the optimized flow order and default on-instrument data processing in all subsequent experiments (see Text S1 in the supplemental material).

In general, these observations suggest both organism- and orientation-dependent biases contributing to premature read truncation. For most species, longer read lengths were obtained with sequencing in the forward orientation. The majority of reads from *A. odontolyticus* and *P. acnes* did not extend beyond 100 bp in the forward orientation.

Per-read error rates. We next estimated the error rate, which is one variable determining the specificity with which reads can be classified (5). We tabulated the number of errors observed with respect to the reference sequences of each component organism empirically (see Table S7 in the supplemental material) and also calculated per-nucleotide error frequencies using a multiple-regression model (Fig. 2A; see also Table S8 in the supplemental material). Consistent with previous reports (7–9), we found that the error rate of Ion Torrent sequencing (averages of 1.5 and 1.4 errors per 100 bases for read sequences from the forward and reverse directions, respectively) was higher than that for equivalent Illumina libraries (average of 0.9 errors per 100 bases).

Error rates were also found to differ among templates from different organisms sequenced by using the same platform (Fig. 2B; see also Table S9 in the supplemental material). For some bacteria (*Escherichia coli* and *Bacteroides vulgatus*), the observed per-read error rates were more consistent across platforms and across read orientations. Comparing other organisms, the error rates of sequence data varied substantially for both platforms (for example, *Listeria monocytogenes* and *Acinetobacter baumannii*) and depending on the orientation of sequencing using the Ion Torrent platform (for example, *Pseudomonas aeruginosa*, *Streptococcus agalactiae*, *Helicobacter pylori*, and *Rhodobacter sphaeroides*). For most organisms, Illumina data had a higher fraction of reads with perfect quality than did the Ion Torrent data, although for *H. pylori*, the population of Ion Torrent forward reads contained slightly fewer errors overall than did the Illumina data (see Table S9 in the supplemental material).

16S rRNA amplicon resequencing of a mock bacterial community. We next compared the abilities of each platform to capture the microbially diverse populations within the defined mock-community sample by inferring the relative representation of each species for both raw and processed (primer-trimmed) reads (Fig. 3). This specimen was prepared to contain an equal representation of each organism based on 16S rRNA gene read counts. Some deviation from the expected abundance of each organism was observed for both platforms, likely reflecting both errors in the formulation of the mock-community DNA mixture and biases introduced during PCR amplification (11, 30, 31). To assess the degree of inherent interlibrary variability, we generated three technical replicates for each library type (MiSeq and Ion Torrent separately in the forward and reverse orientations) using the mock community template, with each replicate incorporating different barcode

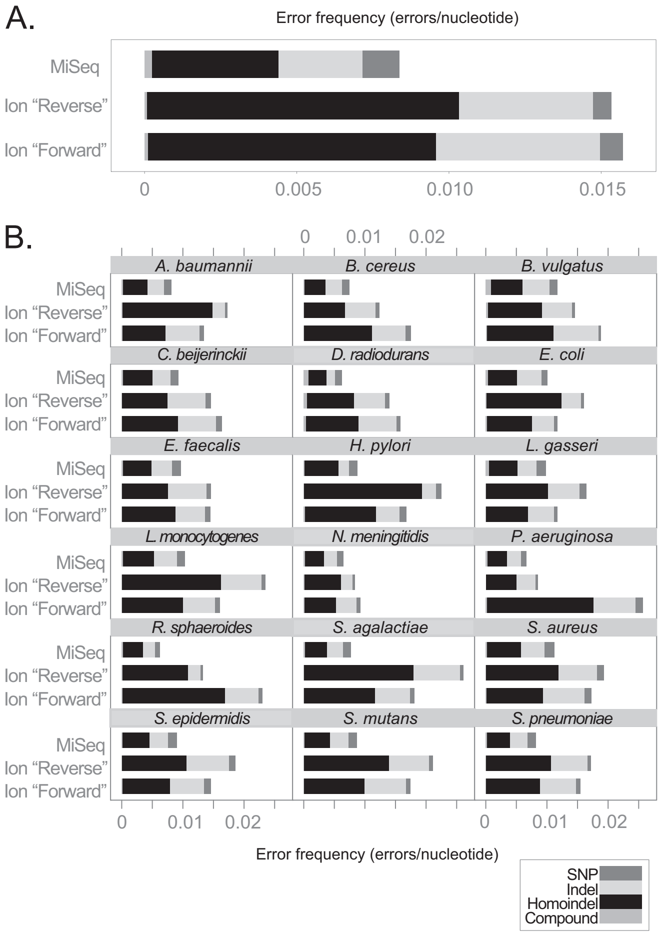

FIG 2 Modeled error rates and error types. (A) Comparison of error rates and error types for assembled paired-end MiSeq reads and unidirectional reads from the Ion Torrent PGM platform derived from a mock-community mixture of 20 bacterial DNAs. Errors for single nucleotide substitutions, homopolymer indels (homoindels), other indels outside homopolymer tracts, and compound errors (event involving two or more categories) are shown separately. (B) Error rates and error types for both platforms stratified by organism. Two organisms for which insufficient reads were obtained for statistical analysis are not displayed. The key applies to both panels A and B. SNP, single nucleotide polymorphism.

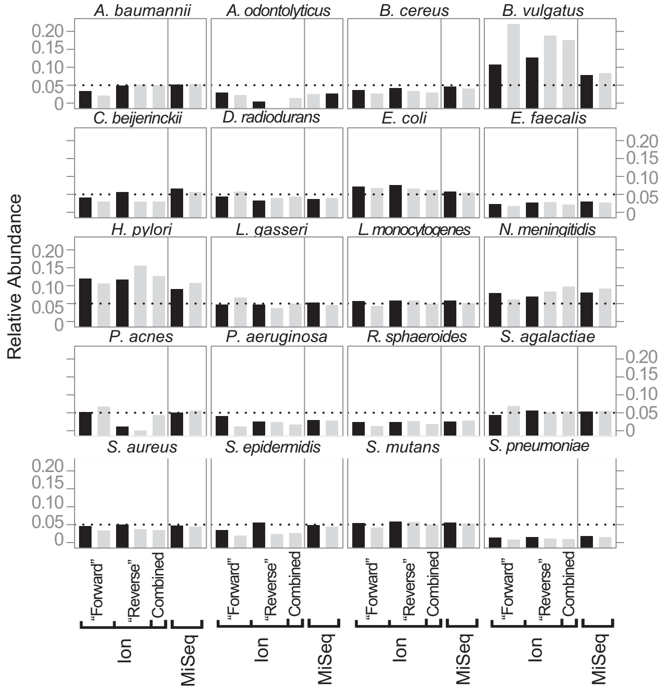

FIG 3 Relative abundance of bacterial species in a 20-organism mock community. The relative read abundance for each organism is indicated with respect to sequencing platform and read orientation (for Ion Torrent data) and for raw and processed reads (light shading and dark shading, respectively). The expected relative abundance of species is indicated by a horizontal dashed line.

sequences. We found minimal variability among replicates (see Fig. S1 in the supplemental material), consistent with independent reports addressing the experimental variability of amplicon-based 16S rRNA surveys (32–36).

The relative abundance of most organisms was generally in good agreement with predicted values and was in most cases relatively consistent among platforms, with a few exceptions. The inferred abundances of *B. vulgatus* and *H. pylori* were substantially increased across all platforms. Notably, representation of *A. odontolyticus* was decreased in reverse-orientation Ion Torrent raw reads and absent in the resultant processed reads. To a lesser degree, the same phenomenon was observed for *P. acnes* with reverse-orientation reads and *P. aeruginosa* with forward-orientation reads from the Ion Torrent platform.

Community profiling of human-derived specimens. Finally, to assess performance using naturally occurring bacterial populations, we explored the microbiological diversity of a panel of 18 human-derived specimens using both platforms (Fig. 4; see also Table S10 in the supplemental material), downsampling the total number of reads for each specimen to be equal across platforms. To meter against organism- and orientation-specific sequencing artifacts from the Ion Torrent data, reads from opposite orientations were combined in equal numbers for this platform. The relative abundances of organisms as determined by both platforms were similar in 5 out of 18 cases (cases S02, S04, S15, S16, and S18), differing in calculated abundance by a maximum of ~10%. In the majority of cases, this 10% tolerance was exceeded for one or more organisms (13 cases [cases S01, S03, S05, S06, S07, S08, S09, S10, S11, S12, S13, S14, and S17]). Disparate cases were marked by a higher rate of failure to classify sequences at the species level from the Ion Torrent reads (cases S01, S03, S06, S07, S08, S09, S10, S11, and S17) and, in a subset of those cases, a relative underrepresentation of one or more organisms by the Ion Torrent platform compared to MiSeq data (cases S01, S03, S08, and S10). *P. acnes* was present in 4 such specimens (cases S01, S08, S10, and S17). *S. constellatus*/*S. intermedius* (case S09), *Morococcus cerebrosus*/*Neisseria macacae* (case S03), and *Moraxella nonliquefaciens* (case S03) were identified by MiSeq but were not detected at any level by using Ion Torrent sequencing. In two cases, a *Staphylococcus* species (case S17) or a *Streptococcus* species (case S01) was identified by Ion Torrent sequencing but was not detected by MiSeq. For one specimen (case S15), both platforms failed to produce BLAST hits that met our identity threshold for species-level classification, indicating the presence of one or more species not represented in our sequence database.

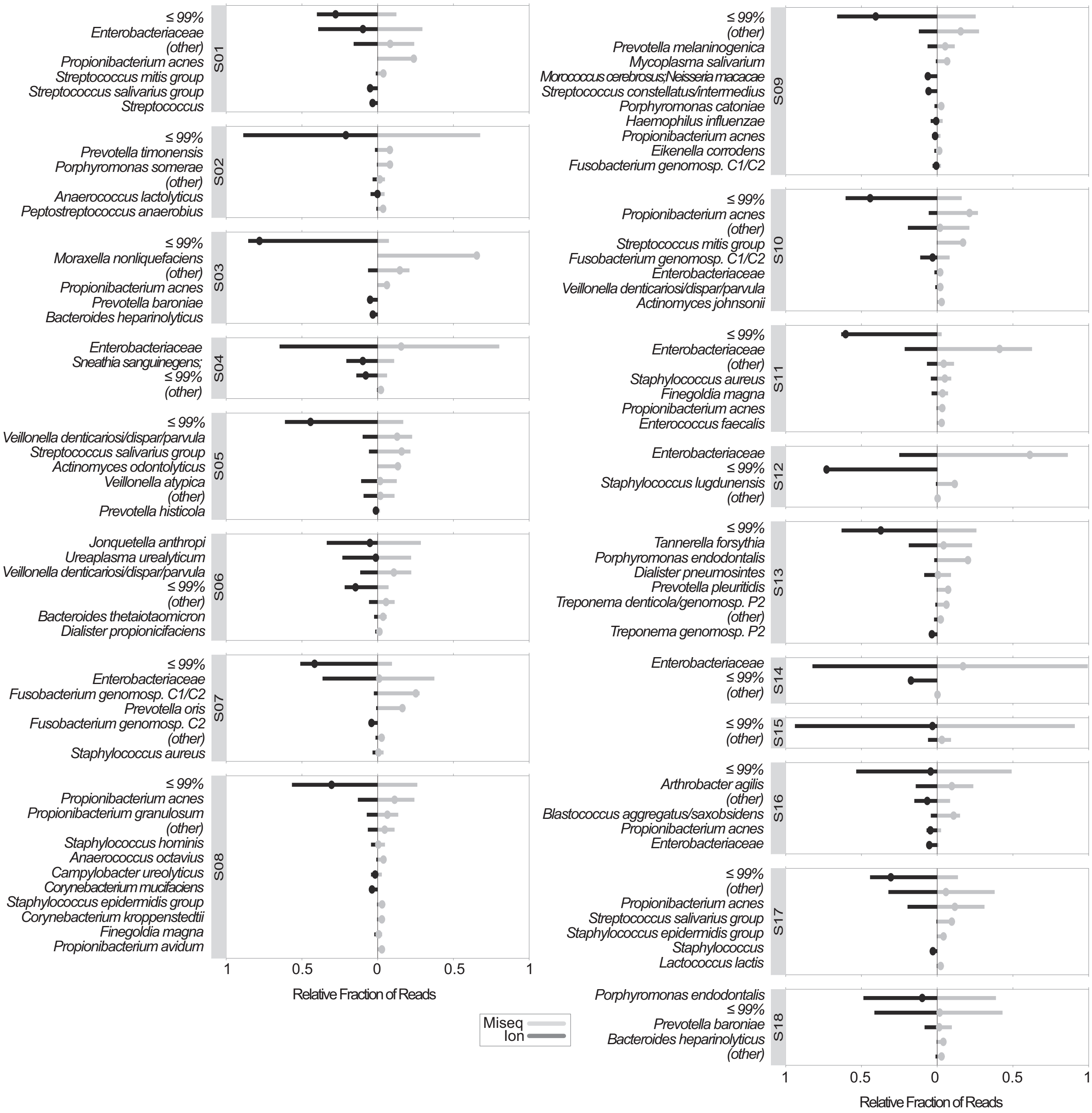

FIG 4 Bacterial communities inferred by MiSeq or Ion Torrent sequencing. The bacterial community structures of 18 human-derived polymicrobial specimens determined by using each sequencing platform were compared. The relative abundance of reads corresponding to each specified organism is indicated by a shaded bar, extending to the left of the midpoint for Ion Torrent data and to the right for MiSeq data. A circle indicates a difference in percent abundances between methods, falling at the midpoint when no difference is observed or to the extreme end of a bar when a species is detected exclusively on one platform. Reads for which more than one species met criteria for classification are labeled with all possible classifications; classification to different species within a genus are separated by slashes, and classifications combining different genera are separated by semicolons. Sequences failing to meet species-level classification criteria are labeled "≤99%." Individual organisms comprising <2.5% of the overall population are aggregated into the "other" category.

DISCUSSION

With the exception of mandatory differences in platform-specific sequencing adaptors, which necessarily represent an uncontrolled variable, we kept all aspects of library production identical when we prepared sequencing templates for these studies. Furthermore, we employed a bidirectional sequencing strategy on both platforms, using self-assembled paired-end sequencing for the Illumina MiSeq platform and full-length sequencing of the target from both orientations on the Ion Torrent PGM platform (albeit combining those reads without self-assembly). When we examined data generated under these matched experimental conditions, we noted several important differences in the performance characteristics of the two platforms as applied to 16S rRNA amplicon sequencing.

Notably, we observed that a fraction of 16S rRNA sequence reads were prematurely truncated in PGM data, although virtually all MiSeq reads were full length (Fig. 1). Read truncation was highly dependent on both the orientation in which sequencing was performed and the specific organism from which the sequence was derived. Several conclusions can be drawn regarding the nature of this phenomenon. Because the Ion Torrent sequencing described in this work was performed on two different instruments, the phenomenon does not appear to be unique to a particular sequencer. It is similarly not specific to the 16S rRNA primer set used, as the forward and reverse sequencing reactions on the Ion Torrent platform utilized the same core primer sequences yet yielded different patterns of read truncation. Given the fact that DNA fragments must retain primer binding sites at opposing ends of the molecule, either primer binding sites during PCR amplification or adaptor sequences during templating, the PCR process can be ruled out as the source of read truncation, both at the initial amplification stage and during emulsion PCR prior to sequencing. We conclude that the sequencing stage can be implicated as the source of this bias, which is supported by our observation that the use of an optimized flow order somewhat mitigated the read truncation effect. The underlying source of this artifact is unclear, but the directionality of the effect suggests that local sequence attributes such as homopolymer tracts, local GC content, or local nucleotide composition are responsible, rather than global properties such as a fragment's overall GC content. However, we examined affected reference sequences with respect to these features and failed to identify any consistent or unifying sequence property that preceded read truncation or that was unique to sequences demonstrating premature truncation (see Fig. S2 to S4 in the supplemental material). This suggests that the underlying causes of this phenomenon are complex and may involve properties outside the nucleotide sequence itself, such as secondary structure.

Whatever the origin of the bias, this finding has at least three major practical implications for users of Ion Torrent sequencing platforms. First, given that the removal of reads that are not full length is an initial data-filtering step for many 16S rRNA amplicon analysis pipelines (2, 5), our findings suggest that this requirement might need to be relaxed when Ion Torrent data are examined in order to avoid categorical exclusion of certain species. Such a strategy comes with potential tradeoffs in data quality, and the partial sequence fragments included may not contain enough taxonomic information for robust classification. Second, it should be recognized that under some conditions, 16S rRNA amplicon fragments of the expected length may not be detectable for some species on the Ion Torrent platform, independently of whether PCR products can be generated from those organisms. Among the 20 organisms specifically surveyed in this study, *A. odontolyticus* and *P. acnes* strains were strongly affected by this problem when sequenced in the reverse orientation. It should be safe to assume that this phenomenon would extend to at least a few additional species or genera not represented in the mock community, for example, as suggested by the failure to detect *M. nonliquefaciens* in one experimental sample (Fig. 4). Third, different community profiles can be obtained when interrogating a 16S rRNA gene target from the two possible orientations. We have found that the use of an optimized flow order and a combination of sequencing data from both orientations (bidirectionally) on the Ion Torrent platform partially mitigates the effects of organism- and orientation-specific read truncations. We therefore recommend both of these strategies for 16S rRNA amplicon sequencing studies performed on the Ion Torrent platform.

It is well documented that data from the Ion Torrent platform exhibit a higher rate of sequencing errors than data from the Illumina platform (7–9), and our study was consistent with this conclusion, although the absolute difference in error rates between the two platforms is not great (Fig. 2). We additionally note that different, organism-specific biases in error rates were observed for both platforms. Library preparation for this study was done by PCR amplification using a robust yet error-prone polymerase lacking proofreading function, and the calculated error rates for both platforms incorporate those introduced by library preparation as well as sequencing. Thus, although it is instructive to perform relative comparisons of error rates between the two platforms, the values presented here should not be interpreted as the absolute error rates for either sequencing technology.

In a direct comparison of species compositions of a mock community consisting of an equimolar mixture of bacterial 16S rRNA gene templates, results were largely consistent when a bidirectional sequencing strategy was employed for the Ion Torrent PGM platform (Fig. 3). Some bias in the relative representation of bacterial species from heterogeneous mixtures is expected from PCR-based sequencing library preparations due to factors including differences in primer mismatches and relative GC contents (11, 30, 31), some of which will be specific to both the particular 16S rRNA region targeted and the profile of organisms present in a sample. Indeed, neither platform produced a precisely equal representation of bacterial species, likely reflecting these underlying issues impacting library preparation as well as potentially real differences in organism abundance introduced unintentionally during formulation of the mock community. In most cases, we observed consistent results for both platforms, with similar patterns of over- and underrepresentation of specific species. For some organisms, there were substantial differences in the relative abundances inferred from raw versus processed Ion Torrent reads in some orientations (notably, *A. odontolyticus*, *P. acnes*, and *P. aeruginosa*). This disparity in *P. aeruginosa* reads likely reflects a pronounced degradation of terminal sequence quality and a subsequent loss of reads during trimming of the distal primer sequences in the forward orientation (Fig. 2B). An underrepresentation of the other two organisms (*A. odontolyticus* and *P. acnes*) was apparent in the reverse orientation for both raw and processed reads and is attributable to premature read truncation (Fig. 1B).

We analyzed human-derived microbial specimens in order to explore functional differences between the two platforms when applied to specimens of human origin (Fig. 4). Although the true composition of these samples cannot be known, in general, we found reasonable concordance between the results obtained by using the MiSeq and Ion Torrent PGM sequencing platforms. Nevertheless, for several samples, the abundance of one or more organisms detected by one platform was significantly different from that detected by the other. It is important to note that, in this analysis, significantly biased detection of even a single organism will significantly distort the overall population profile of a specimen, as the relative abundance of the remaining organisms will be artifactually increased or decreased if one species is significantly underrepresented or overrepresented, respectively. Indeed, specimens with major population-level differences in comparisons of the two platforms can be parsimoniously explained by global differences in the ability to classify reads at the species level or by differences in the inferred abundance of one or a few specific bacteria. In the majority of cases, disparities among the platforms can be explained by the reduced ability of the Ion Torrent platform to detect specific organisms due to premature read truncation: *P. acnes*, known to be problematic in this regard, was implicated in 4 of the 6 discordant cases, and the failure of Ion Torrent sequencing to detect *M. nonliquefaciens*, *Morococcus cerebrosus-Neisseria macacae*, and *S. constellatus/S. intermedius* may similarly reflect this issue.

The choice of the 16S rRNA variable region or regions selected for analysis, the library construction procedures used, and the taxonomic distribution of microorganisms expected in a sample are inevitably study dependent. Thus, although the principles established in this work are broadly applicable, the practical impact of the observations that we report may be greater or less depending on a specific project's experimental objectives, design, and selected target gene(s). As the premature read truncation observed for Ion Torrent sequence reads could extend to other 16S rRNA variable regions and perhaps other marker genes, analogous studies should be considered to evaluate Ion Torrent amplicon resequencing, and next-generation sequencing technologies in general, for specific applications. Similarly, next-generation sequencing technologies are developing at a rapid pace, and as such, it will be important to periodically assess performance characteristics of sequencing platforms in light of the latest read lengths and sequencing chemistries as they become available.

ACKNOWLEDGMENTS

This work was supported in part by the National Center for Advancing Translational Sciences of the National Institutes of Health (grant number UL1TR000423) and in part by Life Technologies.

REFERENCES

1. **Whiteley AS, Jenkins S, Waite I, Kresoje N, Payne H, Mullan B, Allcock R, O'Donnell A.** 2012. Microbial 16S rRNA Ion Tag and community metagenome sequencing using the Ion Torrent (PGM) platform. J. Microbiol. Methods **91:**80–88. http://dx.doi.org/10.1016/j.mimet.2012.07.008.
2. **Junemann S, Prior K, Szczepanowski R, Harks I, Ehmke B, Goesmann A, Stoye J, Harmsen D.** 2012. Bacterial community shift in treated periodontitis patients revealed by ion torrent 16S rRNA gene amplicon sequencing. PLoS One **7:**e41606. http://dx.doi.org/10.1371/journal.pone.0041606.
3. **Schmidt TM, DeLong EF, Pace NR.** 1991. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. J. Bacteriol. **173:**4371–4378.
4. **Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE.** 2006. Metagenomic analysis of the human distal gut microbiome. Science **312:**1355–1359. http://dx.doi.org/10.1126/science.1124234.
5. **Salipante SJ, Sengupta DJ, Rosenthal C, Costa G, Spangler J, Sims EH, Jacobs MA, Miller SI, Hoogestraat DR, Cookson BT, McCoy C, Matsen FA, Shendure J, Lee CC, Harkins TT, Hoffman NG.** 2013. Rapid 16S rRNA next-generation sequencing of polymicrobial clinical samples for diagnosis of complex bacterial infections. PLoS One **8:**e65226. http://dx.doi.org/10.1371/journal.pone.0065226.
6. **Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, Coelho LP, Arumugam M, Tap J, Nielsen HB, Rasmussen S, Brunak S, Pedersen O, Guarner F, de Vos WM, Wang J, Li J, Dore J, Ehrlich SD, Stamatakis A, Bork P.** 2013. Metagenomic species profiling using universal phylogenetic marker genes. Nat. Methods **10:**1196–1199. http://dx.doi.org/10.1038/nmeth.2693.
7. **Lam HY, Clark MJ, Chen R, Natsoulis G, O'Huallachain M, Dewey FE, Habegger L, Ashley EA, Gerstein MB, Butte AJ, Ji HP, Snyder M.** 2012. Performance comparison of whole-genome sequencing platforms. Nat. Biotechnol. **30:**78–82. http://dx.doi.org/10.1038/nbt.2065.
8. **Quail M, Smith ME, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y.** 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics **13:**341. http://dx.doi.org/10.1186/1471-2164-13-341.
9. **Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ.** 2012. Performance comparison of benchtop high-throughput sequencing platforms. Nat. Biotechnol. **30:**434–439. http://dx.doi.org/10.1038/nbt.2198.
10. **Humblot C, Guyot JP.** 2009. Pyrosequencing of tagged 16S rRNA gene amplicons for rapid deciphering of the microbiomes of fermented foods such as pearl millet slurries. Appl. Environ. Microbiol. **75:**4354–4361. http://dx.doi.org/10.1128/AEM.00451-09.
11. **Claesson MJ, Wang Q, O'Sullivan O, Greene-Diniz R, Cole JR, Ross RP, O'Toole PW.** 2010. Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. Nucleic Acids Res. **38:**e200. http://dx.doi.org/10.1093/nar/gkq873.
12. **Shendure J, Ji H.** 2008. Next-generation DNA sequencing. Nat. Biotechnol. **26:**1135–1145. http://dx.doi.org/10.1038/nbt1486.
13. **Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, Hoon J, Simons JF, Marran D, Myers JW, Davidson JF, Branting A, Nobile JR, Puc BP, Light D, Clark TA, Huber M, Branciforte JT, Stoner IB, Cawley SE, Lyons M, Fu Y, Homer N, Sedova M, Miao X, Reed B, Sabina J, Feierstein E, Schorn M, Alanjary M, Dimalanta E, Dressman D, Kasinskas R, Sokolsky T, Fidanza JA, Namsaraev E, McKernan KJ, Williams A, Roth GT, Bustillo J.** 2011. An integrated semiconductor device enabling non-optical genome sequencing. Nature **475:**348–352. http://dx.doi.org/10.1038/nature10242.
14. **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. http://dx.doi.org/10.1128/AEM.02772-10.
15. **Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R.** 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc. Natl. Acad. Sci. U. S. A. **108**(Suppl 1)**:**4516–4522. http://dx.doi.org/10.1073/pnas.1000080107.
16. **Milani C, Hevia A, Foroni E, Duranti S, Turroni F, Lugli GA, Sanchez B, Martin R, Gueimonde M, van Sinderen D, Margolles A, Ventura M.** 2013. Assessing the fecal microbiota: an optimized ion torrent 16S rRNA gene-based analysis protocol. PLoS One **8:**e68739. http://dx.doi.org/10.1371/journal.pone.0068739.
17. **Yergeau E, Lawrence JR, Sanschagrin S, Waiser MJ, Korber DR, Greer CW.** 2012. Next-generation sequencing of microbial communities in the Athabasca River and its tributaries in relation to oil sands mining activities. Appl. Environ. Microbiol. **78:**7626–7637. http://dx.doi.org/10.1128/AEM.02036-12.
18. **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. http://dx.doi.org/10.1128/AEM.01043-13.
19. **Fadrosh DW, Ma B, Gajer P, Sengamalay N, Ott S, Brotman RM, Ravel J.** 2014. An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. Microbiome **2:**6. http://dx.doi.org/10.1186/2049-2618-2-6.
20. **Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT.** 2012. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. PLoS One **7:**e30087. http://dx.doi.org/10.1371/journal.pone.0030087.
21. **Turner S, Pryer KM, Miao VP, Palmer JD.** 1999. Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. J. Eukaryot. Microbiol. **46:**327–338. http://dx.doi.org/10.1111/j.1550-7408.1999.tb04612.x.

22. **Reeder J, Knight R.** 2010. Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. Nat. Methods **7:**668–669. http://dx.doi.org/10.1038/nmeth0910-668b.

23. **Pearson WR.** 2000. Flexible sequence similarity searching with the FASTA3 program package. Methods Mol. Biol. **132:**185–219. http://dx.doi.org/10.1385/1-59259-192-2:185.

24. **Salipante SJ, Hoogestraat DR, Abbott AN, Sengupta DJ, Cummings LA, Butler-Wu SM, Stephens K, Cookson BT, Hoffman NG.** 2014. Coinfection of Fusobacterium nucleatum and Actinomyces israelii in mastoiditis diagnosed by next-generation DNA sequencing. J. Clin. Microbiol. **52:** 1789–1792. http://dx.doi.org/10.1128/JCM.03133-13.

25. **Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD.** 2012. PANDAseq: paired-end assembler for Illumina sequences. BMC Bioinformatics **13:**31. http://dx.doi.org/10.1186/1471-2105-13-31.

26. **Kung-Yee L, Zeger S.** 1986. Longitudinal data analysis using generalized linear model. Biometrika **73:**1322.

27. **R Core Team.** 2013. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

28. **Baker GC, Smith JJ, Cowan DA.** 2003. Review and re-analysis of domain-specific 16S primers. J. Microbiol. Methods **55:**541–555. http://dx.doi.org/10.1016/j.mimet.2003.08.009.

29. **Martin J, Sykes S, Young S, Kota K, Sanka R, Sheth N, Orvis J, Sodergren E, Wang Z, Weinstock GM, Mitreva M.** 2012. Optimizing read mapping to reference genomes to determine composition and species prevalence in microbial communities. PLoS One **7:**e36427. http://dx.doi.org/10.1371/journal.pone.0036427.

30. **Suzuki MT, Giovannoni SJ.** 1996. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Appl. Environ. Microbiol. **62:**625–630.

31. **Hongoh Y, Yuzawa H, Ohkuma M, Kudo T.** 2003. Evaluation of primers and PCR conditions for the analysis of 16S rRNA genes from a natural environment. FEMS Microbiol. Lett. **221:**299–304. http://dx.doi.org/10.1016/S0378-1097(03)00218-0.

32. **Brown SP, Callaham MA, Jr, Oliver AK, Jumpponen A.** 2013. Deep Ion Torrent sequencing identifies soil fungal community shifts after frequent prescribed fires in a southeastern US forest ecosystem. FEMS Microbiol. Ecol. **86:**557–566. http://dx.doi.org/10.1111/1574-6941.12181.

33. **Stearns JC, Lynch MD, Senadheera DB, Tenenbaum HC, Goldberg MB, Cvitkovitch DG, Croitoru K, Moreno-Hagelsieb G, Neufeld JD.** 2011. Bacterial biogeography of the human digestive tract. Sci. Rep. **1:**170. http://dx.doi.org/10.1038/srep00170.

34. **Nelson MC, Morrison HG, Benjamino J, Grim SL, Graf J.** 2014. Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys. PLoS One **9:**e94249. http://dx.doi.org/10.1371/journal.pone.0094249.

35. **Poretsky R, Rodriguez RL, Luo C, Tsementzi D, Konstantinidis KT.** 2014. Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. PLoS One **9:**e93827. http://dx.doi.org/10.1371/journal.pone.0093827.

36. **Ibarbalz FM, Perez MV, Figuerola EL, Erijman L.** 2014. The bias associated with amplicon sequencing does not affect the quantitative assessment of bacterial community dynamics. PLoS One **9:**e99722. http://dx.doi.org/10.1371/journal.pone.0099722.