

2017

## Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing

Michael Tessler  
*American Museum of Natural History*

Johannes S. Neumann  
*Stiftung Tierärztliche Hochschule Hannover*

Ebrahim Afshinnekoo  
*New York Medical College*

Michael Pineda  
*Weill Cornell Medicine*

Rebecca Hersch  
*American Museum of Natural History*

*See next page for additional authors*

[How does access to this work benefit you? Let us know!](#)

More information about this work at: [https://academicworks.cuny.edu/ny\\_pubs/357](https://academicworks.cuny.edu/ny_pubs/357)


Discover additional works at: <https://academicworks.cuny.edu>

---

**Authors**

Michael Tessler, Johannes S. Neumann, Ebrahim Afshinnekoo, Michael Pineda, Rebecca Hersch, Luiz Felipe M. Velgo, Bianca T. Segovia, Fabio A. Lansac-Toha, Michael Lemke, Rob DeSalle, Christopher E. Mason, and Mercer R. Brugler

# SCIENTIFIC REPORTS



OPEN

## Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing

Michael Tessler<sup>1,2</sup>, Johannes S. Neumann<sup>3</sup>, Ebrahim Afshinnekoo<sup>4,5,6</sup>, Michael Pineda<sup>4,5</sup>, Rebecca Hersch<sup>1</sup>, Luiz Felipe M. Velho<sup>7,8</sup>, Bianca T. Segovia<sup>7</sup>, Fabio A. Lansac-Toha<sup>7</sup>, Michael Lemke<sup>9</sup>, Rob DeSalle<sup>1</sup>, Christopher E. Mason<sup>4,5,10</sup> & Mercer R. Brugler<sup>1,11</sup>

Modern metagenomic environmental DNA studies are almost completely reliant on next-generation sequencing, making evaluations of these methods critical. We compare two next-generation sequencing techniques – amplicon and shotgun – on water samples across four of Brazil's major river floodplain systems (Amazon, Araguaia, Paraná, and Pantanal). Less than 50% of phyla identified via amplicon sequencing were recovered from shotgun sequencing, clearly challenging the dogma that mid-depth shotgun recovers more diversity than amplicon-based approaches. Amplicon sequencing also revealed ~27% more families. Overall the amplicon data were more robust across both biodiversity and community ecology analyses at different taxonomic scales. Our work doubles the sampling size in similar environmental studies, and novelly integrates environmental data (e.g., pH, temperature, nutrients) from each site, revealing divergent correlations depending on which data are used. While myriad variants on NGS techniques and bioinformatic pipelines are available, our results point to core differences that have not been highlighted in any studies to date. Given the low number of taxa identified when coupling shotgun data with clade-based taxonomic algorithms, previous studies that quantified biodiversity using such bioinformatic tools should be viewed cautiously or re-analyzed. Nonetheless, shotgun has complementary advantages that should be weighed when designing projects.

With the advent of next-generation sequencing (NGS), studies on DNA from environmental samples (environmental DNA or eDNA) have flourished. It is well known that inferences made from these studies can vary with the field, lab, and analytic techniques utilized<sup>1,2</sup>. There are two principal ways that comparisons can be made when assessing the impact of NGS approaches on eDNA studies. The first entails comparison of sequencing platforms, such as 454 Roche vs. Illumina MiSeq using the same amplicon sequencing approach. The second compares sequencing approaches, the primary techniques being amplicon (sequencing all amplified products from a single gene; e.g., 16S) and shotgun (random sequencing across entire genomes). Several studies have performed such comparisons, with foci ranging from humans to studies of water and soil (Table 1).

The results of prior comparative studies regarding eDNA sequencing vary (Table 1). When Sanger methods are compared to 454 and SOLiD, these approaches perform comparably<sup>3,4</sup>. Illumina and 454 platforms also behave similarly<sup>2,5-7</sup>. In contrast, for amplicon strategies, higher error rates are found with the Ion Torrent due

<sup>1</sup>Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, NY, 10024, USA.

<sup>2</sup>Richard Gilder Graduate School, American Museum of Natural History, New York, NY, 10024, USA. <sup>3</sup>Stiftung Tierärztliche Hochschule Hannover, ITZ - Ecology and Evolution, Bünteweg 17d, D-30559, Hannover, Germany.

<sup>4</sup>Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, 10021, USA. <sup>5</sup>The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, New York, NY, 10021, USA. <sup>6</sup>School of Medicine, New York Medical College, Valhalla, NY, 10595, USA. <sup>7</sup>Universidade Estadual de Maringá, Nupelia/PEA, Maringá, Parana, Brazil. <sup>8</sup>Unicesumar/Instituto Cesiumar de Ciência, Tecnologia e Inovação (ICETI), Maringá, Parana, Brazil. <sup>9</sup>Biology Department, University of Illinois Springfield, Springfield, IL, 62703, USA. <sup>10</sup>The Feil Family Brain and Mind Research Institute (BMRI), New York, NY, 10021, USA. <sup>11</sup>Biological Sciences Department, NYC College of Technology, City University of New York, Brooklyn, NY, 11201, USA. Correspondence and requests for materials should be addressed to M.T. (email: [mtessler@amnh.org](mailto:mtessler@amnh.org)) or C.E.M. (email: [chm2042@med.cornell.edu](mailto:chm2042@med.cornell.edu)) or M.R.B. (email: [mbrugler@citytech.cuny.edu](mailto:mbrugler@citytech.cuny.edu))

Strategy	Platform	Sample #	Target	Comment	Reference
A	Illumina HiSeq, MiSeq	24	Bacteria: Soil, human, and canine stool, mouth and skin	The results were very similar across lanes, read directions, and platforms ( $P < 0.0001$ ), and also comparable to results obtained with the older GA-Ix. Increased sequencing depth did not provide additional information on beta diversity, but helped detect rare species. The HiSeq platform was recommended for large projects that aim at minimizing sequencing cost, while the MiSeq platform can give faster results for monitoring or preliminary studies.	Caporaso, J. G. <i>et al.</i> (2012) <sup>19</sup>
A	454, Illumina MiSeq	10	Eukaryotes (microbial): Soil	The two NGS approaches were extremely similar in the results they provided, especially for abundant amplicons.	Mahé, F. <i>et al.</i> (2015) <sup>5</sup>
A	454, Illumina MiSeq	7	Bacteria: Human stool, mouse, cow, leech, termites, sewage, mock	Reference-based operational taxonomic unit (OTU) clustering alone introduced biases compared to de novo clustering, preventing certain taxa from being observed. Low levels of dataset contamination were observed with Illumina sequencing. This cost-effective alternative to 454 was best when the same template primers, read merging, chimera checking, control libraries, and alternating indices between runs were applied.	Nelson, M. C. <i>et al.</i> (2014) <sup>7</sup>
A	Ion Torrent, Illumina MiSeq	3	Bacteria: Soil	The UniFrac distances between samples sequenced on both Illumina MiSeq and Ion Torrent were significantly correlated. "Differences between sequence technologies can be adjusted by adopting the correct pipeline of analysis". The Q scores generated by different platforms were not directly comparable.	Pyro, V. S. <i>et al.</i> (2014) <sup>9</sup>
A	Ion Torrent, Illumina MiSeq	19	Bacteria: Human-derived, mock	The Ion Torrent platform had comparatively higher error rates and a pattern of premature sequence truncation specific to semiconductor sequencing. This led to organism- and direction-dependent biases provoking underrepresentation or failed identification of species.	Salipante, S. J. <i>et al.</i> (2014) <sup>8</sup>
A	454, Ion Torrent	17	Bacteria and Archaea: River sediments & oil sands tailings ponds	454 and Ion Torrent allowed for highly similar relative abundance estimates for major taxa and almost identical community structure patterns. Emulsion PCR limited amplicon size, which resulted in different forward primers being used. Apart from the following primer bias, "the 454 and Ion Torrent data sets were almost interchangeable, and both would have yielded the exact same ecological conclusions". These ecological conclusions were based on physiochemical sediment data like clay and naphthenic acid values.	Yergeau, E. <i>et al.</i> (2012) <sup>37</sup>
A C	454 Sanger	6	Bacteria: Human dentition	454 resulted in significantly higher coverage estimates than the clonal analysis and provided a higher chance of finding rare species. Pyrosequencing, however, also significantly underestimated the relative abundance of Actinobacteria compared to culture.	Schulze-Schweifing, K. (2014) <sup>4</sup>
A C W	454, SOLiD Sanger SOLiD	1	Bacteria: Human stool	Sanger, 454, and SOLiD amplicon sequencing provided results comparable to the result based on SOLiD shotgun sequencing for overall community composition, but WGS sequencing allowed better identification of species.	Mitra, S. <i>et al.</i> (2013) <sup>3</sup>
A W ->W*	454, Illumina MiSeq Illumina GA-II, HiSeq	15	Bacteria: Soil	The small subunit (SSU) extracted from the shotgun approach yielded higher diversity estimates than straight amplicon methods, both taxonomy- and OTU-based (mainly due to primer bias and chimeras in amplicon sequencing). On the other hand, samples were clustered in similar ways using the two approaches. Another advantage of shotgun sequencing was that it allowed the calculation of the fungus/bacteria ratio, which is an important measure of soil health. The large subunit (LSU) rRNA gene provided even better phylogenetic resolution than SSU.	Guo, J. <i>et al.</i> (2015) <sup>2</sup>
A W	Illumina MiSeq Illumina HiSeq	1 each	Bacteria: Hot spring water thermophiles	Amplicon and shotgun sequencing allowed for comparable phylum detection, but shotgun sequencing found more. The 16S rarefaction curve indicated that a fraction of the species diversity remains to be discovered. Complete functional groups were missed by this approach, like thermophile denitrifying bacteria.	Chan, C. S. <i>et al.</i> (2015) <sup>13</sup>
A W	Ion Torrent, Illumina MiSeq Ion Torrent, Illumina MiSeq, HiSeq	6	Bacteria: Human stool	Changing sequencing methods and informatics approaches to binning sequences to taxa had the greatest impact on variance in the analysis – greater than the difference in between samples. Compared to amplicon sequencing, WGS approaches increased the information gained and reduced biases, but had their own issues mainly related to sequencing depth and read length. While HiSeq offered a much greater sequencing depth that allowed the detection of rare species, the high species count might have been inflated due to misalignments of short reads. At the same time, it performed worst in predicting genes. Ion Torrent generally showed an intermediate performance.	Clooney, A. G. <i>et al.</i> (2016) <sup>14</sup>
A W	Illumina HiSeq Illumina GA-II	16	Bacteria: Soils (deserts, tundra, forests)	The two methods yielded nearly identical estimates of the overall differences in soil bacterial community diversity and composition. The study showed clear limitations of shotgun sequencing depth, that only 13–23% of reads could be annotated, and many of these were misannotated. Still, "for certain questions, shallower sequencing of many samples may be more useful than deeper sequencing of fewer samples".	Fierer, N. <i>et al.</i> (2012) <sup>20</sup>

Continued

Strategy	Platform	Sample #	Target	Comment	Reference
A & W	Illumina MiSeq	16	Bacteria: Kefir, human stool, mouse stool, mock mix	Shotgun metagenomics offered a greater potential for identification of strains, which still remained unsatisfactory. It also allowed increased taxonomic and functional resolution, as well as the discovery of new genomes and genes.	Jovel, J. <i>et al.</i> (2016) <sup>15</sup>
A W	454, Illumina MiSeq 454, Illumina HiSeq	4 to 10, depending on comparison	Bacteria: Marine plankton	Metagenomic approaches were reported to have an advantage over amplicon approaches. They rendered more truthful community richness and evenness estimates by avoiding PCR biases, and provided additional functional information. While both platforms “presented a good agreement by recovering taxa from the same evolutionary groups” when comparing metagenomic shotgun sequencing, many more unique genera were recovered with Illumina than with 454 sequencing. This was partly due to better detection of rare taxa.	Logares, R. <i>et al.</i> (2014) <sup>24</sup>
A W	454 Illumina GA-II	4	Bacteria: Freshwater	Taxonomic composition of each 16S rRNA gene library was generally similar to its corresponding metagenome at the phylum level. At the genus level, however, there was a large amount of variation between the 16S rRNA sequences and the metagenomic contigs, which had a tenfold resolution and sensitivity for genus diversity.	Poretzky, R. <i>et al.</i> (2014) <sup>10</sup>
A W	Illumina MiSeq Illumina HiSeq, MiSeq	1	Bacteria: Human stool	Whole genome sequencing approaches “enhanced detection of bacterial species, increased detection of diversity and increased prediction of genes”. The MiSeq platform provided better de novo contig assembly and species detection with its longer reads.	Ranjan, R. <i>et al.</i> (2016) <sup>11</sup>
A W	454 Illumina GAIIx	51	Bacteria: Human Microbiome Project, vaginal microbiomes	The developers of the Metagenomic Phylogenetic Analysis tool MetaPhlAn showed that it was advantageous to comparable tools. They further underlined the advantages of analyzing taxonomically specific marker genes selected from WGS data (~4% of genes) over amplicon approaches, by “enabling efficient, high-resolution taxonomic profiling”. Yet, while they reported better statistical support for metagenomic sequencing (~10 <sup>8</sup> as compared to <10 <sup>4</sup> reads/sample), the advantages were not evident from their data, as the results for relative abundances of genera were “remarkably similar in all clusters”, and they did not include species level results for amplicon data.	Segata, N. <i>et al.</i> (2012) <sup>30</sup>
A W	454 454, Illumina HiSeq	3	Bacteria and Archaea: Synthetic communities of 64 sequenced species	“Both Illumina and 454 metagenomic data outperformed amplicon sequencing in quantifying the community composition, but the outcome was dependent on analysis parameters and platform.” Metagenomic sequencing outperformed most SSU rRNA gene primer sets, with V13 recovering the best accuracy. Archaea had distinct biases to Bacteria.	Shakya, M. <i>et al.</i> (2013) <sup>38</sup>
A: SSU, LSU, ITS W	Illumina MiSeq Illumina HiSeq	14	Fungi: Soil	The metagenomic shotgun and amplicon approaches performed similarly for identification of most fungal classes. WGS was far inferior in detecting OTUs and identifying species than the amplicon approach using internal transcribed spacers (ITS) as an amplicon target. This was largely due to low (0.005% of DNA) and uneven recovery of fungal rDNA sequences, and lacking fungal data in the reference databases. This “identification bias” was very difficult to quantify or compare among studies.	Tedersoo, I. <i>et al.</i> (2015) <sup>12</sup>
W	454, Illumina GA-II (HiSeq)	1	Bacteria: Freshwater planktonic community	The two platforms performed similarly as 90% of the microbial taxa from the two methods overlapped and the abundance of taxa as determined by the two approaches was highly correlated (R <sup>2</sup> = 0.9). While Illumina recovered longer & more accurate contigs and 14% more complete genes; pyrosequencing might be superior for resolving sequences with repetitive structures or palindromes, and for metagenomic studies based on unassembled reads. Illumina HiSeq seemed to perform similarly to GA-II.	Luo, C. <i>et al.</i> (2012) <sup>6</sup>
W	PacBio RS, Ion Torrent, Illumina GA-IIx, HiSeq, MiSeq	4 genomes	Bacteria: 4 species	Pacific Biosciences RS needed far more DNA, but may be useful for studies focused on de novo sequencing, alternative splicing or epigenetics. It featured read lengths an order of magnitude higher than the other platforms (average: 1500 bases) and insert sizes of up to 10 kb. This read length combined with a very high raw error rate of 13% led to 0% of reads being error-free (75% and 15% for Illumina and Ion Torrent, respectively), which complicated single nucleotide polymorphism (SNP) calling. The errors were evenly distributed, though, while Illumina had higher error rates after long homopolymer tracts and the GGC motif. Ion Torrent failed at sequencing homopolymer tracts, had strand-specific errors, and severe coverage bias for AT-rich genomes.	Quail, M. A. <i>et al.</i> (2012) <sup>39</sup>

**Table 1.** Summary of studies comparing different NGS sequencing strategies and sequencing platforms. In the Strategy column abbreviations are A = 16S amplicon, C = clonal amplification, W = WGS shotgun, and W\* = WGS where SSU sequences are extracted and used.





There are over 50 phyla represented in the whole genome database, which currently contains over 83,000 fully sequenced prokaryotic genomes. These genomes show the upper extent of species representation that would be available for searching in any of the currently available classification programs. We examined the distribution of phyla found in our samples for both the amplicon and shotgun approaches in the context of these fully sequenced genomes. Supplementary Figure 1 shows the phyla available in the National Institutes of Health (NIH) genome database, while highlighting those found using either amplicon or shotgun sequencing that are also in this database. Of the 20 phyla we identified in the amplicon based study, 16 have phyla members with whole genomes sequenced in the NIH database. The four phyla identified using the amplicon approach that do not have whole genome sequences in the NIH database are Aminicenantes, Latescibacteria, Parcubacteria, and Saccharibacteria. All nine shotgun-identified phyla have representatives with whole genomes sequenced (17 phyla have data available for MetaPhlAn).

We next compared the overall composition of phyla in the 49 samples for shotgun and amplicon-derived sequences to global datasets of lake bacteria (Fig. 2). We found strong congruence between the amplicon results and those from all prior amplicon research across the globe. While there is overlap in some of the major phyla in lake systems, the shotgun approach detects different proportions of these phyla that are dissimilar to known freshwater systems. Specifically, the shotgun approach detects higher proportions of Proteobacteria and Cyanobacteria than the amplicon approach conducted here, the Newton study<sup>17</sup>, and our prior global amplicon-based comparison<sup>16</sup>, with the exception being somewhat similar levels of Proteobacteria in our prior global amplicon comparisons.

**Impact on Ecological Inference.** In order to assess the impact of the two sequencing strategies on ecological inferences, we compared both datasets using a variety of standard comparisons used in community ecology, focusing on taxonomic richness, taxonomic abundance, and community composition. Our simplest comparison for sequencing strategy – box and whisker plots of taxonomic richness across each of the river floodplain systems – revealed clear differences (Fig. 3). Each river floodplain system had lower taxonomic richness from shotgun sequencing, which corresponds with the overall richness findings mentioned above. However, more notable is that in the amplicon results, the Pantanal stands out based on taxonomic richness. This pattern is not recovered with shotgun sequencing. In fact, shotgun sequencing at the family level hints at the Paraná being slightly richer.

Heatmaps show the abundance of taxa at each site to be more homogenous in shotgun sequences (Fig. 4). This is partially a reflection of fewer taxa being found with this method, as noted in the comparisons above. However, Cyanobacteria and Proteobacteria at the phylum level particularly drive this pattern, as is further reflected by the accompanying cluster diagrams.

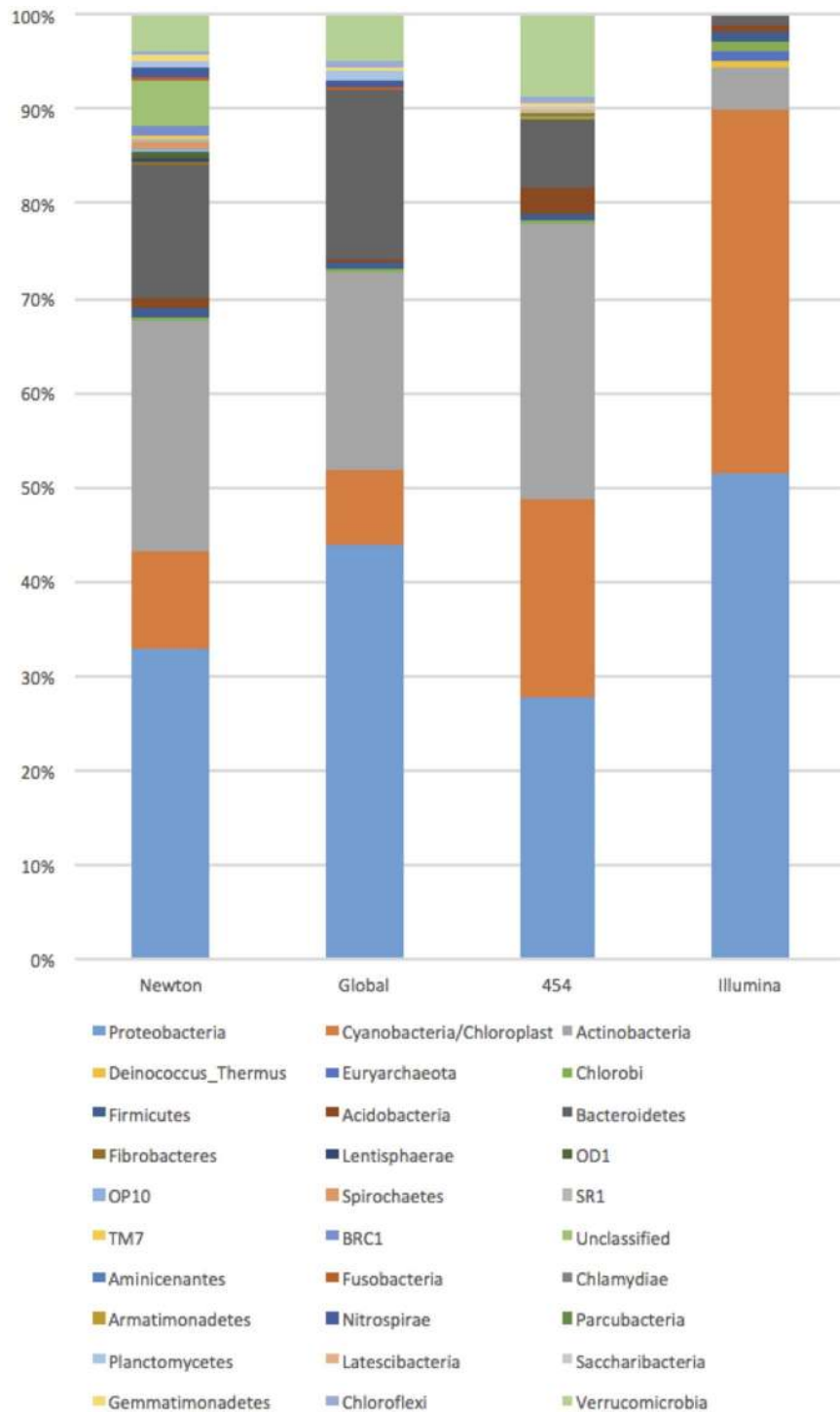
Nonmetric multidimensional scaling (NMDS) analyses for both the amplicon and shotgun approach do not result in any of the four river floodplain systems in the study being particularly distinct (Fig. 5). The environmental variables significantly corresponding with the ordinations are only somewhat similar between amplicon and shotgun approaches both in terms of which variables are significant and how they align with the ordination (Supplementary Table 1). Notably there are many more environmental correlates in the shotgun dataset, while variables significant for amplicon-generated sequences represent a subset of those found with shotgun.

Figure 6 depicts Procrustes tests of the NMDS ordinations produced for the two data sets: significant correlation of the sequencing strategies was found, but it was weak given their use of the same extracted DNA. The finer scale family level comparisons are more similar, despite the actual taxa named for each sequencing strategy at the family level having poor overlap.

To get a further sense of the quality of the datasets and to compare the strength of their correspondence to one another, we compared each dataset (amplicon or shotgun) to themselves using the phylum and family level identifiers. This showed much stronger congruences for the amplicon dataset than the shotgun dataset. The correlation was only slightly stronger for the shotgun comparison than it was for the family level comparison of shotgun vs. amplicon, whereas the amplicon comparisons at the phylum and family level were about twice as strong as the comparisons between sequencing methods.

**Quality Assessments of Analyses.** Our comparisons verifying the quality of our data showed our sampling was thorough. Following QC of the 454 GS Junior generated sequences, 346,042 reads were moved downstream. This number is only a small fraction of the reads generated by the Illumina HiSeq 2500, which, after QC, was ~575 M reads (averaging 12 M pairs of 125 × 125 bp reads per sample). Despite the discrepancies in read count, on a per site basis, it was clear that rarefaction curves reached their asymptotes consistently, indicating that read depth was likely sufficient for both methods, given these taxonomic classifiers (Fig. 7A). The asymptotes are higher and more consistent in taxon richness for the amplicon data. For example, the shotgun data for both taxonomic levels revealed approximately one third of the taxon richness found from amplicons, corroborating our comparisons of taxon richness above.

These rarefaction results are further borne out by the species accumulation curves, where both methods at each taxonomic level have generally reached their asymptote (Fig. 7B). The amplicon data reached full asymptotes with around 10 sites, showing that the method has robust taxonomic sampling even for small numbers of sites. In contrast, the shotgun asymptotes never fully level out, indicating that a large number of sites would be necessary to have a robust taxonomic sample. This is further indicated by our estimates of true taxon richness, which found that, while still lower than amplicon richness, true taxon richness is notably higher with shotgun than could be found with the total sites used in this study (Supplementary Table 2); also note the predictions for shotgun data have a high degree of uncertainty.

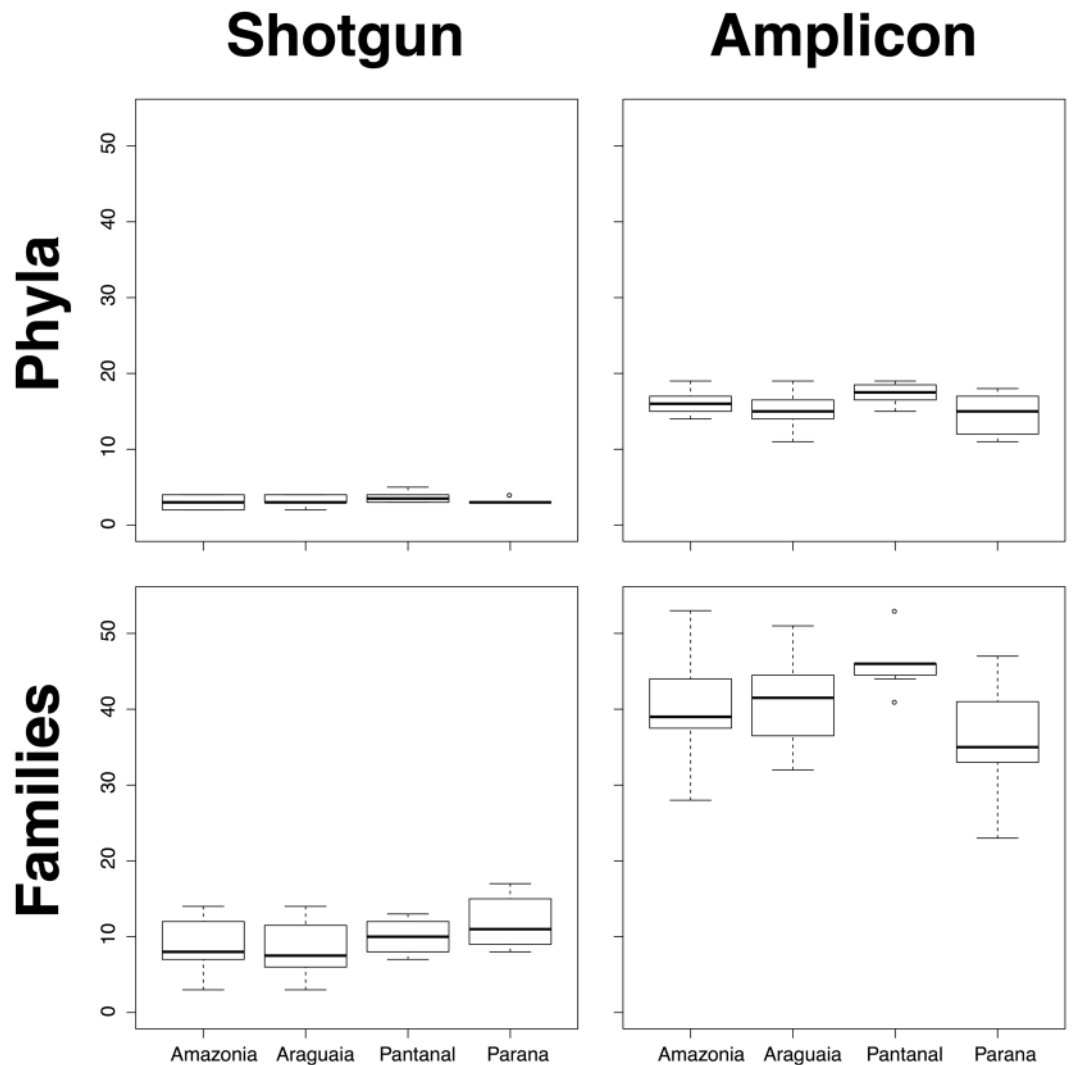


**Figure 2.** Bar plots of the summed proportion of reads (on Y axis) at the phylum level for the two sequencing strategies we have used to review Brazilian sites (labeled as 454 amplicon and Illumina shotgun) and the global comparisons of the meta-analysis<sup>17</sup> (Newton) and our prior global amplicon comparisons<sup>16</sup> (Global). The color code for phyla in the plots is given at the bottom of the figure. Taxonomic nomenclature follows that in the RDP. Note that several of these phyla have since been formally named: TM7 = Saccharibacteria; SR1 = Absconditabacteria; OP10 = Armatimonadetes; OD1 = Parcubacteria.

## Discussion

Our study compares the efficacies of the two NGS sequencing strategies used for eDNA studies (amplicon vs. shotgun) over one of the largest datasets of environmental samples to date. We found the amplicon approach was far more discerning in almost all respects, contrasting general dogma in the field and all but one of eleven





**Figure 3.** Box-and-whisker plots of shotgun vs. amplicon sequencing strategies showing taxon richness at the phylum and family levels. Boxes are middle quartiles divided by the medians, whiskers are 1.5x the interquartile range, and dots are outliers.

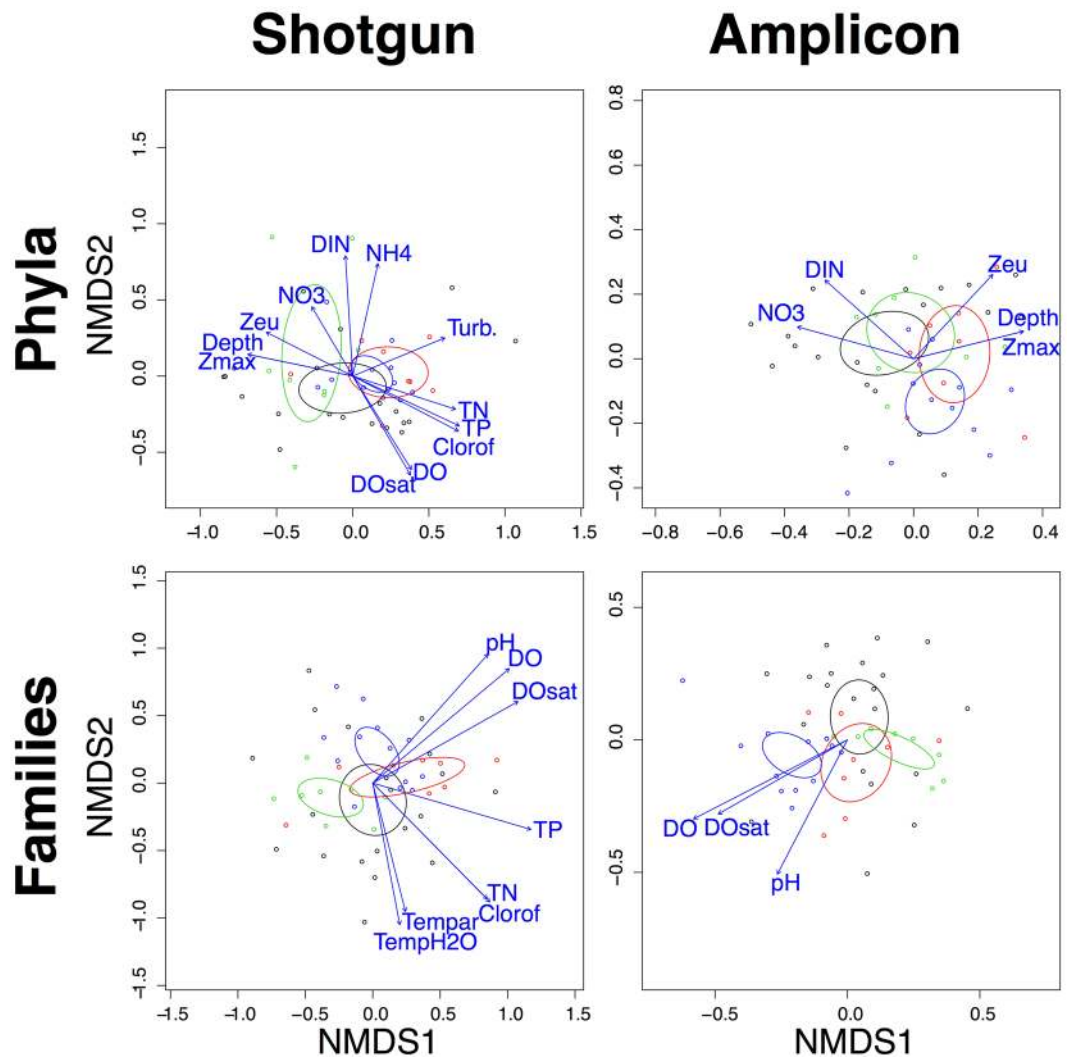
empirical studies in Table 1 that compared these strategies. Unlike our study, that contrasting study differed primarily because of issues with fungal rDNA recovery and deficient databases, rather than due to the systematic biases of the method<sup>12</sup>.

Our study showed weak correlation between the two methods, indicating that while taxonomic overlap exists at both the phylum and family levels the methods are substantially different. Under half the phyla identified from amplicons were found with shotgun; almost all of the phyla recognized by the shotgun approach were also recognized by the amplicon approach. About 30% fewer families were identified from shotgun. This superior performance from amplicons comes despite having <1% of the total reads produced from shotgun. The amplicon results were also far more consistent with prior research on the biodiversity of freshwater systems (Fig. 2). In addition, the Procrustes tests indicated that there is only weak correlation for community composition between the two sequencing strategies using NMDS.

The key difference between the amplicon and shotgun derived data in our study was taxonomic breadth and abundance, whether looking at the overall results or site-by-site. The lower taxon counts for shotgun sequencing appear to be due to issues inherent to the shotgun technique, as well as to the database size. As genome databases are continuously improving and expanding in size, this problem should become less significant. New approaches in multi-enzyme and mechanical shotgun extraction and sequencing techniques may also help<sup>18</sup>. Additionally, shotgun sequencing is complicated by having many reads map to unknown species, which reduces the number of taxonomically-applicable reads (often the majority of reads), and this issue may be more problematic in complex environments such as river basins.

The fundamental issue with the shotgun technique was that taxon richness reached an asymptote on a per site basis at low and unpredictable levels, as compared to amplicon results (Fig. 7A). While this high degree of variability can potentially be overcome by using a large number of sites (note the high variance in total predicted





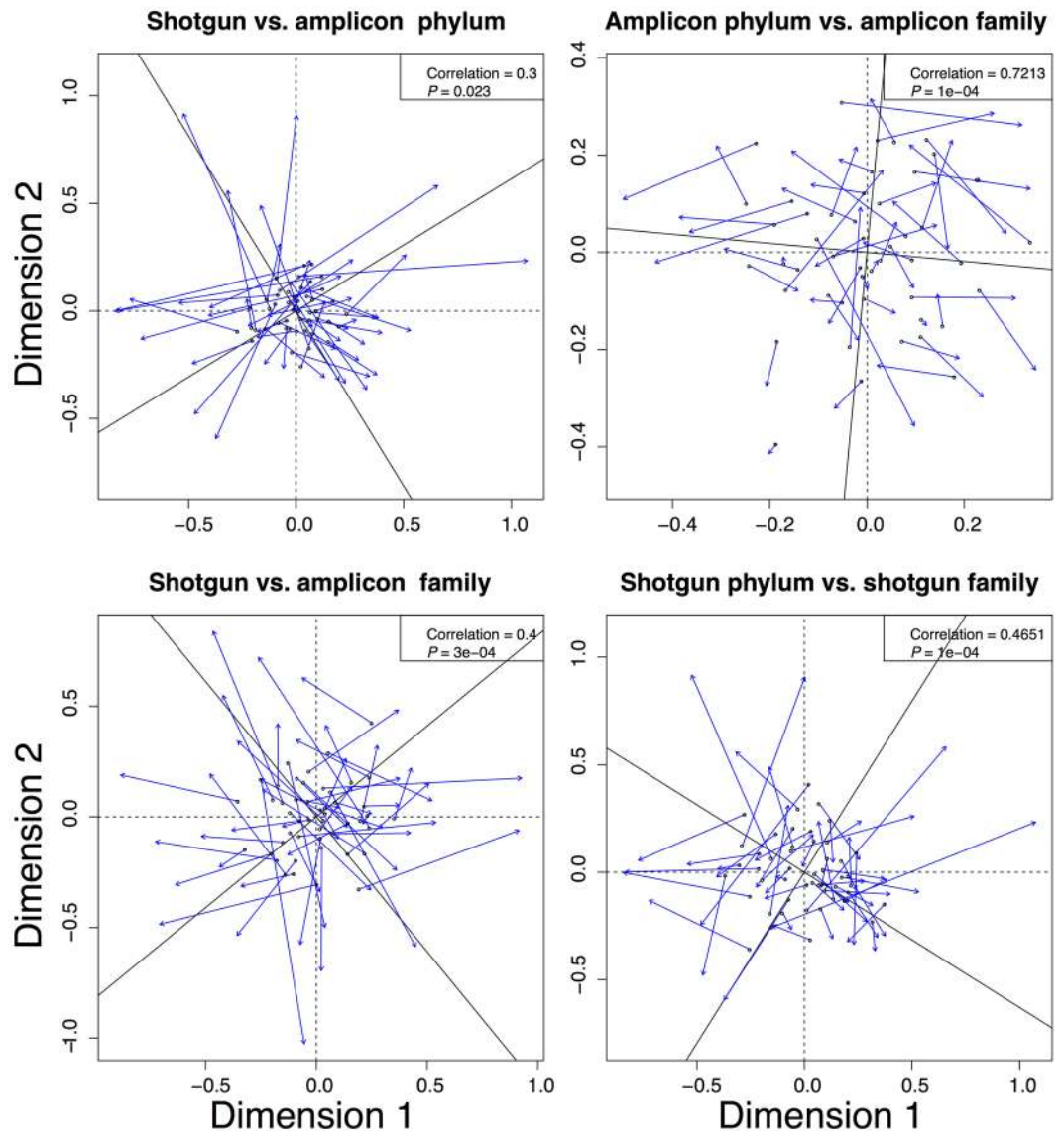
**Figure 5.** NMDS plots for datasets from the shotgun and amplicon techniques for the family and phylum level. Color codes for sites and confidence ellipses are as follows: black = Amazon, red = Araguaia, green = Pantanal, and blue = Paraná. Blue arrows indicate environmental variables that correlate to ordinations. See Supplementary Table 1 for a list of expanded environmental variable names.

phyla (Supplementary Fig. 1), leaving us with only a minor taxonomic overlap between databases. This discrepancy at the phylum level clearly entails a massive lack of resolution at finer taxonomic levels (e.g., for families reviewed here). Missing a single phylum is disconcerting, let alone 20% of phyla.

Given the 16S vs. genome database discrepancy, many shotgun sequences are surely assigned to inappropriate taxa. These incorrect IDs are most likely close relatives of taxa that have sequenced genomes. Thus, the IDs may still have some merit based on the fact that closely related taxa generally have phylogenetically constrained traits that make them more similar (ecologically, physiologically, etc.) to one another than to more distant relatives<sup>24</sup>. However, ecological analyses using higher taxa as surrogates for species achieve variable results depending on the types of input data<sup>25</sup>. In microbial communities, functional diversity cannot be directly predicted from phylogenetic diversity. For example, while in the macroscopic world it is an accepted paradigm that an ecosystem with a low level of taxonomic richness will also have a reduced functional diversity, this does not seem to apply to microbial communities<sup>20</sup>.

Because of the putative cases of mistaken identity with shotgun sequencing, we chose not to use UniFrac or any of its derivative distances (e.g., weighted and generalized; see ref. 26) for community level analyses. For microbial eDNA community ecology, multivariate analysts now generally favor these phylogenetically adjusted measures rather than simply considering taxa as independent entities. However, without highly accurate identifications, accounting for a specific phylogeny makes little sense: recall that only half the amplicon-recovered phyla were found with shotgun, indicating that many shotgun sequences were identified to incorrect phyla - a phylogenetically gigantic distance.

The biases of close, but not exact, identifications are almost surely less extreme when considered as fully independent entities (i.e., not using UniFrac, but more traditional non-phylogenetic distance matrices). Considering



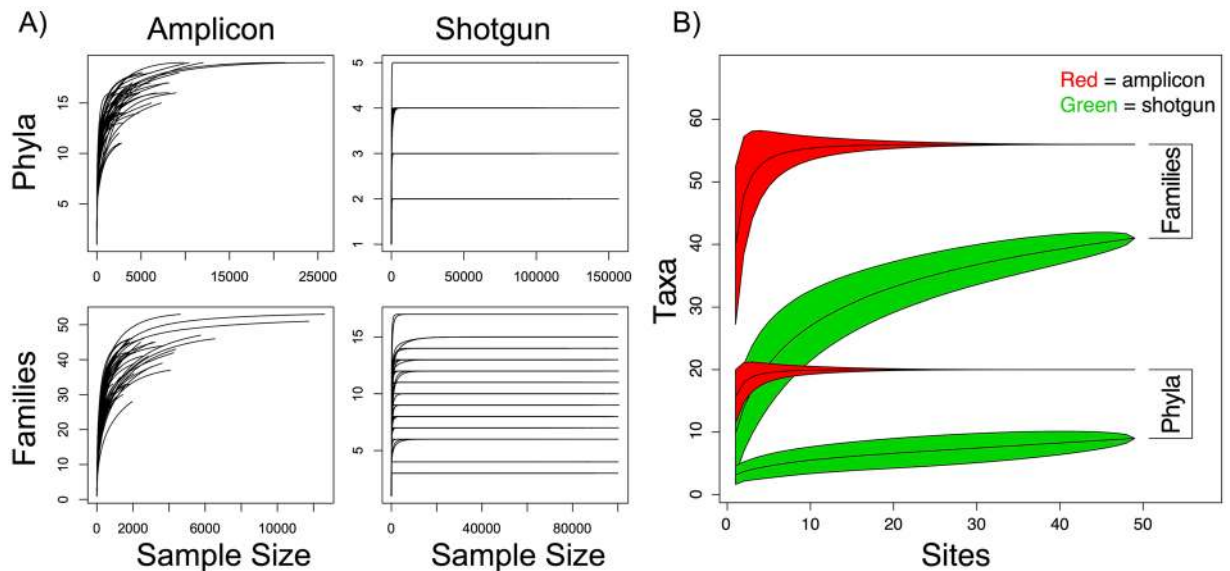
**Figure 6.** Procrustes visualizations of NMDS plots compared at the phylum and family levels for the amplicon vs. shotgun approaches, as well as comparisons of amplicon or shotgun at both taxonomic levels. Test statistics for Procrustes tests are presented for each comparison.

taxa as fully independent entities is standard for community ecology of large eukaryotic organisms. Yet, despite the acceptability of both methods, it is still a notable difference that shotgun data should not – in our opinion – rely on phylogenetically accountable methods until the databases become larger and the tools more sensitive.

Throughout our study we focused on commonly used bioinformatic pipelines. While the RDP appears to work well for amplicons, our findings of MetaPhlAn having lower quality results for shotgun could be called into question. However, MetaPhlAn is one of the most popular taxonomic categorizers; for instance, it was used in the Human Microbiome Project<sup>27</sup>. More importantly, it relies on clade-specific marker genes, which is crucially important for accurate identifications in bacterial biodiversity studies and is a common algorithmic approach. We believe that current practices for analyzing shotgun data that do not use clade-specific markers may be inappropriate for bacterial taxonomic identifications. Future studies should compare less conservative approaches, such as PhyloSift<sup>28</sup>.

Due to conjugation, horizontal gene transfer is rampant in bacteria. It is equally well established that there is a core set of genes across bacteria that are highly conserved and rarely transferred; this is generally referred to as the core genome<sup>29</sup>. While amplicon-derived analyses take advantage of a single gene in the core genome, shotgun relies on genes across the entire genome. Accordingly, the analytics of shotgun will inevitably lead to avoidable misidentifications if based around genes not found in the core genome. This is a major problem for biodiversity and ecology studies, as confident identifications are paramount. Future shotgun analytics can therefore benefit from limiting taxonomic identifications to sequences from the core genome or clade-specific marker genes (as done by MetaPhlAn<sup>30,31</sup>).





**Figure 7.** Comparisons of the sampling efforts for amplicon and shotgun sequence data at the phylum and family levels using (A) rarefaction curves for individual sites and (B) species accumulation curves for the 49 total sampling sites in Brazil.

Furthermore, while our results could be confounded by the fact that we sequenced amplicons via 454 and shotgun via Illumina, we found the majority of studies in Table 1 comparing the amplicon procedure for 454 vs. Illumina agree that these sequencing platforms give highly similar results. Additionally, while Illumina is the dominant NGS platform, amplicon and shotgun studies generally use different Illumina platforms to meet their goals (e.g., HiSeq and MiSeq, respectively; see Table 1). Thus, we believe that our results and comparisons are valid. It is also worth noting that if there were to be an issue with one of these sequencers, it would be assumed that it would be the 454, as it had fewer than 1% of the reads sequenced for Illumina (as expected) - making our results akin to a fisherman with a single fishing rod catching more fish than a commercial trawler.

The only result that is agnostic towards (or at least difficult to interpret for) shotgun or amplicons was in regards to the environmental correlates of the NMDS ordinations (Supplementary Table 1, Fig. 5), which found shotgun to have more significant variables associated with certain metadata. While this could be in favor of shotgun, it is unlikely as the input matrix was so depauperate in terms of taxon richness and evenness across taxa. More likely, this result could be due to a more simplified ordination space that is largely driven by clear divisions by site for a few taxa, as exemplified by the heat maps (Fig. 4). More correlates were found for the phylum level in both sequencing methods, further supporting the idea that the ordinations driven by fewer taxa could be increasing the number of correlates we found. It is also worth noting that for more thoroughly researched microbial floras that have many genomes sequenced, the shotgun system may outperform the amplicon-based approaches as it will provide useful data for a larger array of questions. This already might be the case for urban environments or the human microbiome<sup>32</sup>.

While both amplicon and shotgun sequencing methods have their own advantages for microbial studies, amplicon sequencing was clearly superior for the goals of microbial eDNA community ecology in the reviewed lakes of floodplain systems from Brazil. Further studies should strive for comparisons of even larger datasets across a greater number of habitats, as there can be major differences in conclusions drawn based on the type of sequencing conducted<sup>33</sup>. At this point, any large scale studies should at minimum conduct pilot comparisons between these techniques to choose the more appropriate option.

## Methods

**Sample Collection and DNA Isolation.** The samples compared in this study were analyzed with the 454 amplicon approach in a previous publication<sup>16</sup> and detailed information on the collection of the samples can be obtained from that publication. We used the DNA isolated from the water samples in our prior work<sup>16</sup> for comparative sequencing with Illumina-generated shotgun data. Specifically, we matched 49 of the amplicon sequenced samples from our prior study (58 total) with the shotgun data generated here. The list of samples is provided in Supplementary Table 3. Environmental data were also recorded for each site, as detailed in our prior work<sup>16</sup>.

**Amplicon Library Preparation.** 454 library construction, primer design targeting a specific segment of the 16S rRNA gene (per the Earth Microbiome Project), and work up of amplicons (i.e., amplification and sequencing) are as detailed in our previous work<sup>16</sup>.

**Shotgun Library Preparation.** DNA fragments were prepared into sequencing libraries according to modified manufacturer's standard protocols, using the TruSeq Nano DNA library preparation protocols

(FC-121-4001) and the QIAGEN Gene Reader DNA Library Prep I Kit (cat. no. 180984). 50–100 ng of sample DNA went through Covaris fragmentation to ~500 nucleotides. AMPure XP beads were used for size selection (removal of small fragments <200 bp) and removal of excess reagents. DNA was end-repaired to create blunt ends on both 3' and 5' ends. Then A-tailing, or the addition of dATP to the 3' end, was carried out, which increases the stability of the DNA fragments, prevents concatamer formation, and enables ligation to occur with a complementary T nucleotide found on indexes. Next, the DNA fragments were tagged with index ligation tags. Eighteen cycles of Polymerase Chain Reaction (PCR) were then used to amplify sample DNA fragments. An AMPure XP bead wash was then used to purify DNA libraries. Fragments were visualized on a BioAnalyzer 2100 to check quality and average nucleotide length and concentration was measured by Qubit quantification (ng/uL).

**Sequencing.** Using HiSeq (v4) SBS chemistry, we multiplexed 24 samples per lane on a HiSeq 2500 and processed the raw data using the Illumina RTA software and CASAVA 1.8.2. All samples were then checked for standard CASAVA QC parameters (all reads pass filter). Specifically, all samples had high (>Q20) quality values at the median base, low percent alignment to PhiX (<1%), and similar insert size ( $550 \pm \text{SD of } 70 \text{ bp}$ ).

**Sequence Trimming and Quality Control.** The amplicon analysis pipeline is described in our prior work<sup>16</sup>. Concisely, we used a multi-tiered approach to assure the quality of downstream sequence data. We demultiplexed the sequences and implemented five standard 454 quality filters on the GS Junior (Dot, Mixed, Signal Intensity, Primer and TrimBack Valley). Thereafter, `sff_extract` ([http://bioinf.comav.upv.es/sff\\_extract/index.html](http://bioinf.comav.upv.es/sff_extract/index.html)) was used to create .fasta, .fasta.qual, .fastq, and .xml files. Low quality reads were removed and key/adaptor sequences were clipped using `sff_extract`. The results of this filtering were visualized using FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). Two binaries, FASTQ/A Trimmer and FASTQ Quality Trimmer (part of the FASTX toolkit; [http://hannonlab.cshl.edu/fastx\\_toolkit/](http://hannonlab.cshl.edu/fastx_toolkit/)), were used to further trim low-quality regions. The final data set consisted of sequences with bases having a Phred quality score  $\geq 25$ .

**Taxonomic Classification of Sequences.** Diversity at the family and phylum levels for the 454 data set was assessed as in our prior work<sup>16</sup>. Succinctly, we used the RDP categorizer to obtain classifications at broad (phylum) and narrow (family) levels of taxonomic diversity; please see the Discussion section and our prior work<sup>16</sup> for an explanation of why finer (i.e., genus and species level) taxonomic resolution may be inappropriate. Over 50 phyla and 350 families are assessed by the RDP categorizer<sup>34</sup>. MetaPhlAn (v2.0)<sup>30,31</sup> was used to analyze the shotgun data. Samples were run with the `-ignore viruses` parameter to filter out reads matching to phiX that is spiked during some library preparation procedures and becomes a contaminant in the microbiome analysis.

**Comparisons of Amplicon and Shotgun Sequences.** Results from each method were summarized in several formats. Percentages of reads by taxon were visualized for both phyla and family levels. Since our samples are from lakes of floodplain systems, we compared their taxonomic distributions to a major survey of lake microbiota<sup>17</sup> as well as our prior survey of freshwater microbiota<sup>16</sup>. Heatmaps with site and taxon cluster diagrams were produced for each method using the “heatmap” function in R<sup>35</sup>. Species richness was calculated and visualized with box-and-whisker plots in R. To compare the sequence quality in further detail, we produced species accumulation curves (using “specaccum”), rarefaction curves (using “rarefy”), and estimates of true taxon richness (using “specpool”) in R with the `vegan` package<sup>36</sup>.

To compare community level differences between those taxa identified with each sequencing method, NMDS ordinations were constructed using function “`metaMDS`” from the `vegan` package in R<sup>36</sup>; default settings were used except `trymax = 10,000`. For simplicity between comparisons, two-dimensional ordinations were selected. Environmental vectors were fit to the ordination results using “`envfit`” (`vegan`). Separation of floodplains was tested with PERMANOVA analyses conducted with the “`adonis`” function (`vegan`). Non-randomness was tested between the two ordination results with “`protest`” (`vegan`); this was visualized with the “`procrustes`” function (`vegan`). The last three analyses mentioned use permutations; to increase their accuracy total permutations were increased to 9,999.

**Data Availability Statement.** We deposited all 454 sequence data from<sup>16</sup> in NCBI’s Short Read Archive under BioProject ID PRJNA310230 and all Illumina data were deposited under BioProject ID PRJNA389803.

## References

- Majaneva, M., Hyytiäinen, K., Varvio, S. L., Nagai, S. & Blomster, J. Bioinformatic amplicon read processing strategies strongly affect eukaryotic diversity and the taxonomic composition of communities. *PLoS One* **10**, e0130035 (2015).
- Guo, J., Cole, J. R., Zhang, Q., Brown, C. T. & Tiedje, J. M. Microbial Community Analysis with Ribosomal Gene Fragments from Shotgun Metagenomes. *Appl. Environ. Microbiol.* **82**, 157–166 (2015).
- Mitra, S. *et al.* Analysis of the intestinal microbiota using SOLiD 16S rRNA gene sequencing and SOLiD shotgun sequencing. *BMC Genomics* **14**, S16 (2013).
- Schulze-Schweifing, K., Banerjee, A. & Wade, W. G. Comparison of bacterial culture and 16S rRNA community profiling by clonal analysis and pyrosequencing for the characterization of the dentine caries-associated microbiome. *Front. Cell. Infect. Microbiol.* **4**, 164 (2014).
- Mahé, F. *et al.* Comparing high-throughput platforms for sequencing the V4 region of SSU-rDNA in environmental microbial eukaryotic diversity surveys. *J. Eukaryot. Microbiol.* **62**, 338–345 (2015).
- Luo, C., Tsementzi, D., Kyrpides, N., Read, T. & Konstantinidis, K. T. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS One* **7**, e30087 (2012).
- Nelson, M. C., Morrison, H. G., Benjamino, J., Grim, S. L. & Graf, J. Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys. *PLoS One* **9**, e94249 (2014).
- Salipante, S. J. *et al.* Performance comparison of Illumina and ion torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. *Appl. Environ. Microbiol.* **80**, 7583–7591 (2014).
- Pylro, V. S. *et al.* Data analysis for 16S microbial profiling from different benchtop sequencing platforms. *J. Microbiol. Methods* **107**, 30–37 (2014).



10. Poretzky, R., Rodriguez-R, L. M., Luo, C., Tsementzi, D. & Konstantinidis, K. T. Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. *PLoS One* **9**, e93827 (2014).
11. Ranjan, R., Rani, A., Metwally, A., McGee, H. S. & Perkins, D. L. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem. Biophys. Res. Commun.* **469**, 967–977 (2016).
12. Tedersoo, L. *et al.* Shotgun metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding analyses of fungi. *MycKeys* **10**, 1–43 (2015).
13. Chan, C. S., Chan, K.-G., Tay, Y.-L., Chua, Y.-H. & Goh, K. M. Diversity of thermophiles in a Malaysian hot spring determined using 16S rRNA and shotgun metagenome sequencing. *Front. Microbiol.* **6**, 177 (2015).
14. Clooney, A. G. *et al.* Comparing apples and oranges?: Next generation sequencing and its impact on microbiome analysis. *PLoS One* **11**, e0148028 (2016).
15. Jovel, J. *et al.* Characterization of the gut microbiome using 16S or shotgun metagenomics. *Front. Microbiol.* **7**, 459 (2016).
16. Tessler, M. & Brugler, M. R. *et al.* A global eDNA comparison of freshwater bacterioplankton assemblages focusing on large-river floodplain lakes of Brazil. *Microb. Ecol.*, **73**, 61–74 (2017).
17. Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D. & Bertilsson, S. A guide to the natural history of freshwater lake bacteria. *Microbiol. Mol. Biol. Rev.* **75**, 14–49 (2011).
18. MetaSUB International Consortium. The Metagenomics and Metadesign of the Subways and Urban Biomes (MetaSUB) International Consortium inaugural meeting report. *Microbiome* **4**, 24 (2016).
19. Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6**, 1621–1624 (2012).
20. Fierer, N. *et al.* Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc. Natl. Acad. Sci. USA* **109**, 21390–21395 (2012).
21. Cole, J. R. *et al.* Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* **42**, D633–42 (2014).
22. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–6 (2013).
23. McDonald, D. *et al.* An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* **6**, 610–618 (2012).
24. Logares, R. *et al.* Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ. Microbiol.* **16**, 2659–2671 (2014).
25. Neeson, T. M., Van Rijn, I. & Mandelik, Y. How taxonomic diversity, community structure, and sample size determine the reliability of higher taxon surrogates. *Ecol. Appl.* **23**, 1216–1225 (2013).
26. Chen, J. *et al.* Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* **28**, 2106–2113 (2012).
27. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
28. Darling, A. E. *et al.* PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* **2**, e243 (2014).
29. Vernikos, G., Medini, D., Riley, D. R. & Tettelin, H. Ten years of pan-genome analyses. *Curr. Opin. Microbiol.* **23**, 148–154 (2015).
30. Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9**, 811–814 (2012).
31. Truong, D. T. *et al.* MetaPhlan2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
32. Afshinnikoo, E. *et al.* Geospatial resolution of human and bacterial diversity with city-scale metagenomics. *Cell Syst* **1**, 72–87 (2015).
33. Mason, C. E., Afshinnikoo, E., Tighe, S., Wu, S. & Levy, S. International standards for genomes, transcriptomes, and metagenomes. *J. Biomol. Tech.* **28**, 8–18 (2017).
34. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**, 5261–5267 (2007).
35. R Core Team. *A language and environment for statistical computing* R Foundation for Statistical Computing, Vienna, Austria (2013).
36. Oksanen, J. *et al.* *Vegan: Community Ecology Package. R-package version 2* (2013).
37. Yergeau, E. *et al.* 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 (2012).
38. Shakya, M. *et al.* Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. *Environ. Microbiol.* **15**, 1882–1899 (2013).
39. Quail, M. A. *et al.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**, 341 (2012).

## Acknowledgements

R.D. acknowledges the Korein Foundation and the Sackler Institute for Comparative Genomics at the AMNH for support. M.R.B. acknowledges the Gerstner Family Foundation for providing support. We would like to thank the Genomics Core Facility at Weill Cornell Medicine, as well as funding from the Irma T. Hirschl and Monique Weill-Caulier Charitable Trusts, Bert L. and N. Kuggie Vallee Foundation, the WorldQuant Foundation, and the Bill and Melinda Gates Foundation (OPP1151054). We would also like to thank the Brazilian National Council of Technological and Scientific Development (CNPq)/SISBIOTA-BRASIL for providing financial support to L.F.M.V., B.T.S. and F.A.L.T.

## Author Contributions

L.F.M.V., B.T.S., F.A.L.T., and M.L. collected samples. M.T., R.D., C.E.M., and M.R.B. conceived and designed the experiments. M.T., E.A., M.P., R.H., R.D., and M.R.B. performed the experiments. M.T., J.S.N., E.A., M.P., R.D., and M.R.B. analyzed the data. M.T., J.S.N., R.D., C.E.M., and M.R.B. wrote the paper. All authors read and approved the final manuscript.

## Additional Information

**Supplementary information** accompanies this paper at doi:10.1038/s41598-017-06665-3

**Competing Interests:** The authors declare that they have no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017