Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing

Nicholas A Bokulich1–3, Sathish Subramanian4, Jeremiah J Faith4, Dirk Gevers5, Jeffrey I Gordon4, Rob Knight6,7, David A Mills1–3 & J Gregory Caporaso8,9

High-throughput sequencing has revolutionized microbial ecology, but read quality remains a considerable barrier to accurate taxonomy assignment and α-diversity assessment for microbial communities. We demonstrate that high-quality read length and abundance are the primary factors differentiating microbial diversity. ‘Denoising’3,4, an approach developed for amplicon data generated by these rapidly evolving instruments aids this progress. Given unknown sequencing error rates for Illumina GAIIX, HiSeq and MiSeq instruments. We present guidelines for user-defined quality-filtering strategies, enabling efficient extraction of high-quality data and facilitating interpretation of Illumina sequencing results.

Recent advances in high-throughput, short-amplicon sequencing are revolutionizing efforts to describe microbial diversity within and across complex biomes, including the human body1 and Earth’s biosphere2. The greater sequence coverage and lower per-base cost of the Illumina GAIIX, HiSeq and MiSeq instruments relative to the more expensive, lower-coverage platforms aids this progress. Given unknown sequencing error rates for amplicon data generated by these rapidly evolving instruments and changing chemistries, and the potential for PCR error introduced during short-amplicon sample preparation, quality filtering is integral to the analysis of high-throughput sequencing data because it removes erroneous reads that otherwise overestimate microbial diversity. ‘Denoising’3,4, an approach developed for quality-filtering amplicons sequenced on the Roche 454 Life Sciences pyrosequencer, does not scale to Illumina instruments, which generate tens (MiSeq) to hundreds (GAIIX) to thousands (HiSeq) to millions more data per run.

Illumina systems provide Phred quality scores for every nucleotide, which represent the probability that a given base call is erroneous. How best to incorporate these scores in marker-gene–based microbial ecology studies has not been thoroughly investigated, and stringent filtration that discards many reads has been recommended to avoid exaggerated diversity estimates5. Our strategy works on a per-nucleotide basis, truncating reads at the position where their quality begins to drop. It therefore differs from Illumina’s quality-filtering software CASAVA, which filters on a per-read basis. Previous investigation into quality filtering of Illumina data6 focused on whole-genome sequencing applications, where error profiles are expected to differ from those in amplicon-sequencing runs.

We tested the effects of different quality-filtering parameters on taxonomic classification, α-diversity (within-sample diversity) and β-diversity (between-sample diversity comparison) estimates, using the ‘quantitative insights into microbial ecology’ (QIIME)7 pipeline (Fig. 1 and Supplementary Table 1). We tested four different ‘mock’ communities sequenced on the GAIIX (1 sequencing run), HiSeq (2 sequencing runs) and MiSeq (3 sequencing runs) (Supplementary Table 2). These comprised deliberately combined collections of 12–67 bacterial or fungal species whose genomes had been previously sequenced (Supplementary Tables 3–6). We also compared free-living and host-associated communities5,8, which were samples with high β diversity, and wine9 and spontaneous beer fermentation–associated communities10, which were samples with lower β diversity, to evaluate the effects of filtering settings on β-diversity comparisons of different community types. Raw read counts and sample counts for all data sets are presented in Supplementary Table 7.

We evaluated how primary quality-filtering parameters (p, minimal high-quality read length, q Phred score, r maximum of consecutive low-quality calls and n maximum of ambiguous calls allowed) and secondary (c, operational taxonomic unit (OTU) threshold) quality-filtering parameters affect analyses using the following five separate evaluations. Evaluation 1 was an analysis of α diversity and qualitative taxonomic composition, using mock communities, to test which settings best measure true community composition, minimizing spurious additional OTUs (Fig. 2 and Supplementary Figs. 1–7). Evaluation 2 was an analysis of quantitative taxonomic composition, using defined mock communities, to test whether different settings introduce biases in specific taxa (Supplementary Figs. 8–10). Evaluation 3 was an analysis of β diversity, using mock communities, to determine whether different settings cause significant (P < 0.05) differences in phylogenetic composition between identical communities (Supplementary Table 8). Evaluation 4 was an analysis of β diversity, using real communities, to test whether different
settings affect our ability to differentiate sample types in principal coordinates analysis (PCoA) plots (Fig. 2c and Supplementary Figs. 11–16). Evaluation 5 was an analysis of β diversity, using real communities, to test whether differences detected between communities on different sequencing platforms are consistent.

Our results revealed general patterns. First, parameters p, q, and c had a marked effect on α diversity and estimates of taxonomic composition, whereas n and r did not (Fig. 2a,b and Supplementary Figs. 1–7). The effects of p and q were variable across runs in an apparently platform-independent fashion (Supplementary Figs. 4–5). All settings except high q values required secondary filtration with c to reach expected taxon counts, but the required level varied between 0.01% to 0.0001% of total sequences, depending on q and p settings. Increasing p also decreased abundance of unassigned sequences and sequences given shallow taxonomic assignment. In all mock data sets studied, extreme settings of q and p, but not r and n, had a marked impact on taxonomic distribution (Supplementary Figs. 8–10 and Supplementary Note). Second, weighted UniFrac distances between mock communities (Supplementary Note) were more robust to changes in parameter settings than unweighted UniFrac distances at low c; however, these differences disappeared at high c. Thus, as expected, differences in low-abundance OTUs have a larger impact on the unweighted metric. We note that any filtering strategies that remove low-abundance reads make it impossible to apply richness estimation metrics such as abundance-based coverage estimator (ACE) and Chao1, which incorporate low-abundance read counts in their calculations. These metrics are unlikely to be accurate, however, if many of these reads actually represent sequencing errors.

Because observations in microbial ecology are often based on PCoA of samples, we applied Procrustes analysis to compare PCoA plots from different parameter settings on both biological and mock communities. We found that conclusions derived from PCoA plots were also robust to differences in parameter settings: the only notable differences occurred at stringent q, p, and c, which resulted in extreme levels of read filtration that blurred the known major distinction between host-associated and free-living communities (Fig. 2c–e and Supplementary Figs. 11, 12), and closely related wine and beer fermentation–associated communities (Supplementary Figs. 13–16 and Supplementary Note).

Finally, these observations generalized from the GAIIX to the HiSeq2000 and MiSeq platforms. We observed the same β-diversity trends (such as separation in host-associated and free-living communities) on all three platforms, and heavily decreased p (p = 0.25) and increased q (q ≥ 20) were the only factors that caused these sample types to erroneously cluster together in the HiSeq data (Supplementary Note).

To calibrate optimal filtering settings, we highly recommend including a standardized mock community in each individual sequencing run. We believe this will be necessary for confident comparison of samples from multiple sequencing runs to normalize run-to-run PCR and sequencing error, but additional work is...
needed to evaluate which factors (such as community composition and complexity) define optimal mock communities for filter calibration under different experimental conditions. For data sets where a mock community is not included for calibration, we recommend a conservative OTU threshold of \( c = 0.005\% \). Additional work is also required to address the impact of filtering strategies on the analysis of paired-end reads.

Users can process sequencing data under specific filtering conditions to support different downstream analyses. For example, users with a majority of high-quality, full-length sequences may wish to increase Phred score \( (q) \) and minimum of high-quality calls \( (p) \) in lieu of limiting OTU abundance \( (c) \), thereby retrieving only full-length sequences with low error rates, potentially increasing the discovery rate of rare OTUs (as sequence selection will be based on length and quality, not count). Users with shorter reads or reads truncated by early low-quality base calls may wish to increase the maximum number of consecutive low-quality calls \( (r) \), lower minimum of high-quality calls \( (p) \) and use a higher OTU threshold. In this way, lower-quality but taxonomically useful reads will be retained, and reliable sequences will be selected based on abundance rather than error probability. Other users may be more interested in maximizing read count for implementation of machine-learning tools, identifying OTUs with different abundances across metadata categories or treatment regimes, or jackknifing or permutational tests for \( \beta \) diversity, all of which benefit from increased sample sizes. In this scenario, reads should be filtered using primary filters of Phred score and minimum high-quality calls instead of OTU abundance, which greatly reduces read count.

With these guidelines, users can confidently extract more, higher-quality sequences and decrease OTU filtration thresholds, increasing acuity for rare OTU discrimination and \( \beta \)-diversity comparisons. The Earth Microbiome Project (EMP)\(^2\) is adopting these guidelines for routine analysis of all small-subunit rRNA gene sequencing on the Illumina HiSeq and MiSeq systems, facilitating deeper, more efficient insight into how microbial diversity varies over spatial and temporal scales across our planet. The conclusions drawn from this study are conserved for data from HiSeq2000, MiSeq and GAIIx systems, supporting confident cross-platform data handling. In addition, we recommend new default settings for Illumina processing in QIIME \(( r = 3; p = 0.75 \text{ total read length}; q = 3; n = 0; c = 0.005\% ; \text{Supplementary Note})\), incorporated in the recent release of QIIME 1.5.0. Finally, although quality parameters tested here were evaluated using QIIME, these conclusions are relevant to Illumina amplicon quality filtering across all bioinformatics pipelines for improved diversity estimates in all taxa and environments.

**METHODS**

Methods and any associated references are available in the online version of the paper.

**ACKNOWLEDGMENTS**

We thank G. Giannoukos (Broad Institute of MIT and Harvard), I. Rasolonjatovo (Illumina), M. Gebert (University of Colorado, Boulder) and L. Wegener Parfrey (University of Colorado, Boulder) for contributing mock community sequencing data used in this study, and S. Huse and A. Gonzalez for useful feedback and discussions of this manuscript. This work was supported in part by grants from the US National Institutes of Health (NIH DK78660 to J.I.G., NIH R01HD059127 to D.A.M. and NIH U54HG004969 to D.G.), the Juvenile Diabetes Research Fund (D.G.), the Crohn’s and Colitis Foundation of America (J.I.G. and D.G.), and the Howard Hughes Medical Institute. N.A.B. was supported by the 2012–2013 Dannon Probiotics Fellow Program (The Dannon Company) and a Wine Spectator scholarship.

**AUTHOR CONTRIBUTIONS**

N.A.B., J.G.C., D.A.M. and R.K. conceived and designed the experiments; N.A.B. performed the experiments and data analysis. All authors contributed sequencing data sets and wrote the manuscript.

**COMPETING FINANCIAL INTERESTS**

The authors declare no competing financial interests.

Published online at [http://www.nature.com/doifinder/10.1038/nmeth.2276](http://www.nature.com/doifinder/10.1038/nmeth.2276).

Reprints and permissions information is available online at [http://www.nature.com/reprints/index.html](http://www.nature.com/reprints/index.html).

1. Yatsunenko, T. et al. Nature **486**, 222–227 (2012).
2. Gilbert, J.A. & Meyer, F. *ASM Microbe* **7**, 64–69 (2012).
3. Reeder, J. & Knight, R. *Nat. Methods* **7**, 668–669 (2010).
4. Quince, C. et al. *Nat. Methods* **6**, 639–641 (2009).
5. Caporaso, J.G. et al. *Proc. Natl. Acad. Sci. USA* **108**, 4516–4522 (2011).
6. Minoche, A.E. et al. *Genome Biol.* **12**, R112 (2011).
7. Caporaso, J.G. et al. *Nat. Methods* **7**, 335–336 (2010).
8. Caporaso, J.G. et al. *ISME J.* **6**, 1621–1624 (2012).
9. Bokulich, N.A. et al. *PLoS ONE* **7**, e36357 (2012).
10. Bokulich, N.A., Bamforth, C.W. & Mills, D.A. *PLoS ONE* **7**, e35507 (2012).
11. Lozupone, C. & Knight, R. *Appl. Environ. Microbiol.* **71**, 8228–8235 (2005).
ONLINE METHODS

Data availability. Raw data can be found in the QIIME database (http://microbio.me/qiime/) under the following study identifiers, where data set number can be found in Supplementary Table 7: data set 1, 719; data set 2, 1685; data set 3, 1686; data set 4, 1626; data set 5, 1687; data set 6, 1688; data set 7, 1683; data set 8, 1684; data set 9, 1689; and data set 10, 1690.

Sample data. Data analyzed in this study were generated in ten separate sequencing runs on the Illumina GAIx (3 sequencing runs), HiSeq2000 (3 sequencing runs) and MiSeq (4 sequencing runs) (Supplementary Table 7). In the first phase of this study, this consisted of six different sequencing runs analyzing four different bacterial and fungal mock communities (Supplementary Table 2); of these, only data set 1 comprised previously published data and contained biological samples in addition to mock community samples. In the second phase of this study, β-diversity comparisons were made between biological samples analyzed in four different, previously published studies using Illumina GAIx system (data sets 1, 9, 10; Supplementary Table 7), and HiSeq2000 system (data set 7) and MiSeq system (data set 8). DNA extraction, PCR and sequencing for all sequencing runs were performed described previously except that singleton PCRs were performed, the PCRs contained only 25 cycles and the extension step was extended by 5 min. Additionally, data set 2 was prepared using the Illumina TruSeq v2 paired-end library preparation kit, and data set 3 was prepared with the TruSeq v1 paired-end library kit. Mock communities analyzed in this study were derived from six total sequencing runs on the GAIx, HiSeq and MiSeq with reads between 90 nt and 250 nt (Supplementary Table 2). The taxonomic compositions of the four mock communities analyzed by these runs are presented in Supplementary Tables 3–6. These sequencing runs were performed on different instruments at different sites with the goal of assessing the impact of filtering parameters across a broad set of sequencing conditions: Illumina Cambridge (data sets 1, 3, 7 and 8), Broad Institute (data sets 2 and 4), Washington University School of Medicine (data sets 5 and 6) and University of California DNA Technologies Core (data sets 9 and 10). Although all sequencing runs in this study were paired-end runs, only the forward reads were analyzed for the purposes of this study, as QIIME’s filtering pipeline currently handles each read independently and does not use a scheme for aligning or concatenating paired-end reads.

Sequence analysis. Raw Illumina fastq files were de-multiplexed, quality filtered and analyzed using QIIME (v. 1.4.0-dev). Reads were filtered using variable manual settings, as modulated by the parameters p, q, r and n (Supplementary Table 1). OTUs were assigned using the QIIME UCLUST wrapper, with a threshold of 97% pair-wise nucleotide sequence identity, and the cluster centroid for each OTU was chosen as the OTU representative sequence. OTU representative sequences were then classified taxonomically using the QIIME-based wrapper of the Ribosomal Database Project (RDP) naïve Bayesian classifier retrained on the Greengenes 16S rRNA gene database prefiltered at 97% identity, using a 0.80 confidence threshold for taxonomic assignment. After taxonomic assignment, variable OTU minimum abundance thresholds (c) were applied to remove any OTU representing fewer sequences than the defined threshold. Representative sequences were aligned using PyNAST against a template alignment of the Greengenes database, filtered at 97% identity, and phylogenetic trees were constructed using FastTree. β-diversity estimates were calculated within QIIME using UniFrac distances between samples, with even subsampling at 2,000 sequences per sample with 1,000 Monte Carlo iterations. Procrustes analysis was performed on UniFrac distance matrices with 1,000 Monte Carlo randomizations to compute goodness of fit (M2) and visualized using PCoA.

12. Edgar, R.C. Bioinformatics 26, 2460–2461 (2010).
13. Wang, Q., Garrity, G.M., Tiedje, J.M. & Cole, J.R. Appl. Environ. Microbiol. 73, 5261–5267 (2007).
14. DeSantis, T.Z. et al. Appl. Environ. Microbiol. 72, 5069–5072 (2006).
15. Caporaso, J.G. et al. Bioinformatics 26, 266–267 (2010).