UPARSE: highly accurate OTU sequences from microbial amplicon reads

Robert C Edgar

Amplified marker-gene sequences can be used to understand microbial community structure, but they suffer from a high level of sequencing and amplification artifacts. The UPARSE pipeline reports operational taxonomic unit (OTU) sequences with ≤1% incorrect bases in artificial microbial community tests, compared with >3% incorrect bases commonly reported by other methods. The improved accuracy results in far fewer OTUs, consistently closer to the expected number of species in a community.

© 2013 Nature America, Inc. All rights reserved.

A number of recent large-scale studies have taken advantage of next-generation sequencing to characterize microbial community structure and function, including the Human Microbiome Project (HMP)[1] and a survey of the *Arabidopsis thaliana* root microbiome[2]. Many of these projects assess community structure by sequencing amplified markers, such as the 16S ribosomal RNA gene, which are organized into OTUs: groups of sequences that are intended to correspond to taxonomic clades or monophyletic groups. Yet data analysis in this type of study is hampered by ubiquitous artifacts introduced by amplification and sequencing. Current techniques for reducing artifacts include quality filtering of reads[3], denoising of flowgrams[4–6], chimera filtering[6,7] and clustering[8], but many biases and spurious OTUs due to unfiltered artifacts often remain, confounding inferences of community structure and function[9]. A large fraction of OTU representative sequences produced by recommended procedures with commonly used metagenomic sequence analysis pipelines[6,9,10] on artificial ('mock') communities of known composition have <97% identity with true biological sequences, a divergence generally considered sufficient to infer a new species[8], and the number of OTUs often far exceeds the number of expected species.

I have developed a pipeline (UPARSE, http://drive5.com/uparse/ and **Supplementary Software**) for constructing OTUs *de novo* from next-generation reads that achieves high accuracy in biological sequence recovery and improves richness estimates on mock communities. UPARSE works by quality-filtering reads, trimming them to a fixed length, optionally discarding singleton reads and then clustering the remaining reads. Clustering uses UPARSE-OTU, a novel 'greedy' algorithm that performs chimera filtering and OTU clustering simultaneously—unlike previously developed methods, which perform chimera filtering as a separate step, if at all. UPARSE-OTU is essential for the dramatic improvement in accuracy achieved by the pipeline. Unlike QIIME[10], mothur[9] and AmpliconNoise[6], UPARSE does not require technology- or gene-specific parameters (such as an OTU size cutoff), algorithms (such as flowgram denoising) or data (such as a curated multiple alignment), which makes it highly robust with respect to variations in the input data and suggests that UPARSE could be successfully applied to a wide range of marker genes and sequencing technologies.

Definitive assessment of OTU accuracy is not possible because the definition of an OTU for a given experiment is left to the investigator[11]. Here I assessed the accuracy of an OTU by comparing its representative sequence to the closest true biological sequence, considering incorrect or missing bases in the OTU sequence to be errors. I classified OTU sequences as 'Perfect' (identical to the biological sequence), 'Good' (≤1% errors), 'Noisy' (>1% to ≤3% errors), 'Chimeric' (>3% errors and chimeric with high confidence), 'Contaminant' (high-identity match to a species not in the targeted community) or 'Other' (>3% errors or a biological sequence missing from the reference databases). Although these categories are somewhat arbitrary, it is reasonable to regard a method that produces mostly Perfect and Good OTUs as having higher specificity than one that produces many Chimeric and Other OTUs. If the sensitivity of the method is also comparable to or better than that of other methods, then it is reasonable to regard that method as more accurate. I defined sensitivity as the number of detectable species that are assigned to OTUs, and a detectable species as one having at least one read with ≤3% errors.

To assess performance using the 16S gene, I used two mock communities ('Even' and 'Staggered') derived from one set of 21 microbial strains[12], obtaining reads from two technical replicates sequenced on Illumina MiSeq[3] and three on 454 GS FLX Titanium (Roche)[13] for each community (**Supplementary Note 1**). Several species were absent or represented by only one or a few reads, especially in the Staggered community, thus making it reasonable to expect <21 OTUs. I also tested UPARSE on the internal transcribed spacer (ITS) region in two fungal mock communities[14] (**Supplementary Note 1**).

I compared results from UPARSE on the 16S mock communities with those obtained using QIIME, mothur and AmpliconNoise (**Fig. 1** and **Supplementary Note 1**). In all 16S mock data sets, a large majority of UPARSE OTUs were classified as Perfect, Good or Contaminant, which suggests that they are accurate reconstructions of biological sequences. By contrast, from 41 to 71% of mothur OTUs and 23 to 67% of QIIME OTUs were Chimeric. QIIME reported many more OTUs than did UPARSE or mothur on pyrosequenced data sets, with OTUs ranging from 1,900 to 3,647 (see also **Supplementary Note 1**). With AmpliconNoise,

Independent Investigator, Tiburon, California, USA. Correspondence should be addressed to R.C.E. (robert@drive5.com).

RECEIVED 7 DECEMBER 2012; ACCEPTED 15 JULY 2013; PUBLISHED ONLINE 18 AUGUST 2013; DOI:10.1038/NMETH.2604

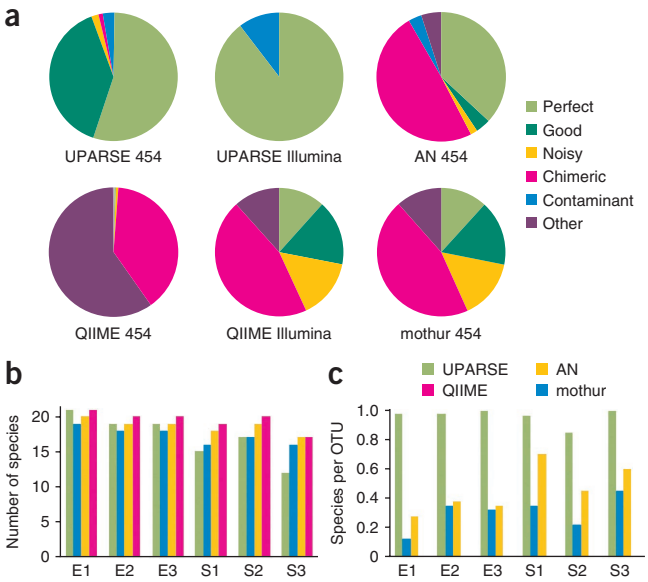
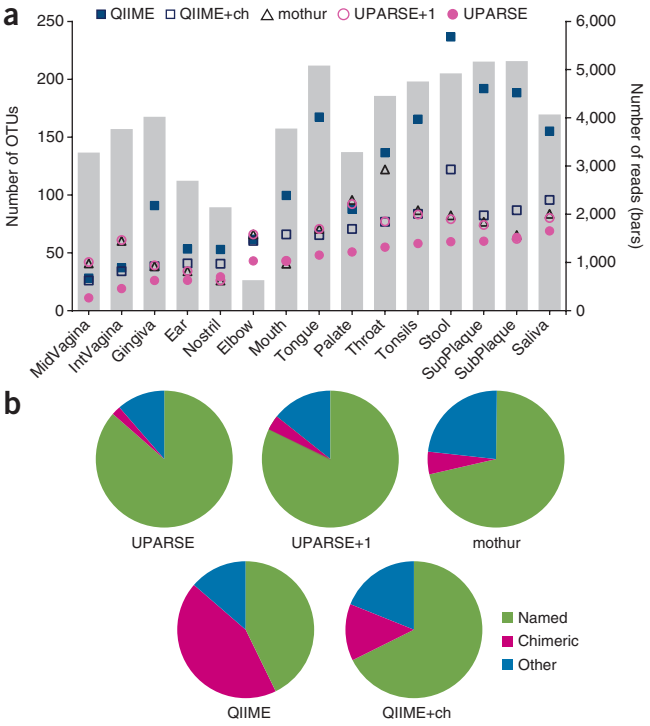


Figure 1 | Results on 16S mock communities. (**a**) Pie charts constructed by averaging the fraction of OTUs in each category over all data sets from 454 sequencing and for the AllF data set for Illumina (AllF, all forward reads; data sets described in **Supplementary Note 1**). (**b**,**c**) Sensitivity (**b**) and number of species (**c**) per OTU for the 454 data sets. For Illumina, UPARSE OTUs have 0.95 species per OTU averaged over data sets Even1m to Stag3m, whereas QIIME has 0.1 species per OTU on the AllF set (not shown). For 454, the QIIME species per OTU values are too small to be visible, ranging from 0.005 (Stag2P) to 0.014 (Stag1P). AN, AmpliconNoise. Even (E) and Staggered (S) mock communities in **b**,**c** are analyzed by run number (1–3).

from 15 to 64% of OTUs were Chimeric. On the fungal ITS communities, QIIME again produced large numbers of OTUs, none of which was within 3% of a known biological sequence, whereas UPARSE successfully reconstructed many of the expected species and many contaminants, though results on these data sets are harder to interpret (**Supplementary Note 1**).

The much closer agreement between the number of OTUs and the number of mock species (plus detected contaminants) achieved by UPARSE (**Fig. 1c**) suggests that UPARSE OTUs approach a 1:1 correspondence with species sampled *in situ*. This is difficult to establish with certainty, even on mock samples, because additional undetected contaminants cannot be ruled out, and generalizing from mock to real communities—which may have greater diversity—demands caution. UPARSE discards singletons by default, which may reduce sensitivity by eliminating a few rare taxa, especially for samples with low read coverage. However, even if singletons are retained, some of the true diversity will probably still be lost owing to primer mismatches or abundance below the minimum required for detection. Even if all taxa are believed to be present in the sequence data, it is not clear that a reliable estimate of species richness is possible with a pipeline that retains singletons but generates many spurious OTUs. This is especially problematic if the frequency of spurious OTUs depends on the number of reads per sample, community structure, chimera formation rates and sequencing protocol, which are real possibilities that undermine generalizations from mock-community results. Spurious OTUs are greatly reduced by UPARSE, a result suggesting that more robust estimates of richness and diversity are possible.



To assess performance on biological samples collected *in situ*, I randomly selected a set of body-site samples from the HMP that were sequenced using 454 (**Supplementary Note 1**). I classified an OTU as 'Named' if its representative sequence had a global alignment with at least 97% identity to a sequence in the Greengenes Named Isolate database[15], Chimeric if it was classified as a chimera by UCHIME[7] using the Chimera Slayer reference database[12] or as Other (**Fig. 2**). These samples had a relatively small number of reads (from 394 to 5,937), so the artifact frequency as a fraction of unique read sequences was expected to be lower than that in the mock communities (**Supplementary Note 2**). A low artifact frequency implies that pipelines with poor artifact filtering should perform relatively well and that many of the classifications of OTUs as Other were probably due to species being unrepresented in the Greengenes Named Isolate database rather than unfiltered artifacts.

As expected, default QIIME had the largest fraction of chimeric OTUs and the largest overall number of OTUs, as this pipeline

Figure 2 | Results on Human Microbiome Project (HMP) data sets. (**a**) Average number of reads per sample for the site and the average number of OTUs per sample obtained with each method. Samples are sorted in order of increasing number of UPARSE OTUs. Peaks and valleys for other methods therefore indicate cases in which there is a disagreement in the trend in relative diversity measured as the number of OTUs. (**b**) Average fraction of OTUs in three categories: named (within 3% of a sequence in the Greengenes Named Isolate database), chimeric (according to UCHIME) and other. Methods are UPARSE (default parameters), UPARSE+1 (singletons not discarded), QIIME (defaults), QIIME+ch (with chimera filtering) and mothur (recommended HMP pipeline). AmpliconNoise was not assessed on these samples owing to technical difficulties running the software. MidVagina, mid-vagina; IntVagina, vaginal introitus; SupPlaque, supragingival plaque; SubPlaque, subgingival plaque.

© 2013 Nature America, Inc. All rights reserved.

has no artifact filtering. Default QIIME also had the strongest correlation between number of OTUs and number of reads (Pearson correlation coefficient $r = 0.76$, compared to default UPARSE with lowest $r = 0.42$), supporting the hypothesis that the number of OTUs tends to increase with the number of reads because of artifacts rather than correctly detected species. All methods except default UPARSE agreed closely with the number of OTUs for the 'Elbow' site, which has the smallest number of reads. Here, the sensitivity of UPARSE was probably reduced by the discarding of singleton reads that are accurate biological sequences. These results are consistent with the expectation that the artifact frequency (as a fraction of unique read sequences) increases with the total number of reads and that UPARSE produces estimates of diversity that are more robust to changes in read depth than are estimates of other pipelines. UPARSE+1 results (in which singletons are retained) agreed closely with mothur on OTU number for most samples, probably reflecting similar numbers of unfiltered artifacts due to singletons.

Overall, these results show that UPARSE achieves a substantial improvement in OTU construction over current methods. The same UPARSE pipeline was consistently able to recover biological sequences from mock communities given three different types of reads (454, Illumina unpaired and Illumina paired), a range from 10,000 to >2 million raw reads, and two distinct sequence regions (16S and ITS). UPARSE recovered almost all detectable species with sensitivity sufficient to detect several contaminants. Notably, I did not use parameter tuning, *post hoc* cutoffs or technology-specific algorithms such as flowgram denoising, and I found that computational resource requirements were substantially reduced, especially compared to pipelines that use flowgram denoising (**Supplementary Note 3**).

UPARSE generated OTUs that were perfect reconstructions of sequences representing known species in some samples and that were mostly correct or very close to a known biological sequence in others, though a few chimeras and unclassified sequences remained. In practice, more chimeras and more correct sequences could be detected by using a reference database, whereas this work focused exclusively on *de novo* methods (pre-existing databases were used exclusively for assessment).

METHODS

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

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The author thanks N.A. Bokulich, J.G. Caporaso and D. Gevers for helpful discussions and providing prepublication data; P.D. Schloss for assistance with mothur; C. Quince for assistance with AmpliconNoise; L. Dethlefsen and S.M. Huse for helpful discussions; S. Yourstone for an insightful critique of a draft manuscript; and the Sloan Foundation, Microbiology of the Built Environment Program (M.L. Sogin, Marine Biological Laboratory) for providing compute resources.

AUTHOR CONTRIBUTIONS

R.C.E. conceived of the study, performed the analysis and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The author declares no competing financial interests.

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

1. Human Microbiome Project Consortium. *Nature* **486**, 207–214 (2012).
2. Lundberg, D.S. *et al. Nature* **488**, 86–90 (2012).
3. Bokulich, N.A. *et al. Nat. Methods* **10**, 57–59 (2013).
4. Quince, C. *et al. Nat. Methods* **6**, 639–641 (2009).
5. Reeder, J. & Knight, R. *Nat. Methods* **7**, 668–669 (2010).
6. Quince, C., Lanzen, A., Davenport, R.J. & Turnbaugh, P.J. *BMC Bioinformatics* **12**, 38 (2011).
7. Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C. & Knight, R. *Bioinformatics* **27**, 2194–2200 (2011).
8. Huse, S.M., Welch, D.M., Morrison, H.G. & Sogin, M.L. *Environ. Microbiol.* **12**, 1889–1898 (2010).
9. Schloss, P.D., Gevers, D. & Westcott, S.L. *PLoS ONE* **6**, e27310 (2011).
10. Caporaso, J.G. *et al. Nat. Methods* **7**, 335–336 (2010).
11. Sneath, P.H.A. & Sokal, R.R. *Numerical Taxonomy* (W.H. Freeman, 1973).
12. Haas, B.J. *et al. Genome Res.* **21**, 494–504 (2011).
13. Human Microbiome Project Consortium. *Nature* **486**, 215–221 (2012).
14. Ihrmark, K. *et al. FEMS Microbiol. Ecol.* **82**, 666–677 (2012).
15. DeSantis, T.Z. *et al. Appl. Environ. Microbiol.* **72**, 5069–5072 (2006).

© 2013 Nature America, Inc. All rights reserved.

ONLINE METHODS

16S mock-community data. I analyzed 454 reads using mothur following the recommended procedure[9] at http://www.mothur.org/wiki/Schloss_SOP (downloaded 24 October 2012) and using AmpliconNoise v.1.29 with default parameters at the 0.03 distance cutoff. I did not use mothur for Illumina reads because there was no recommended procedure at the time this work was performed (October 2012). AmpliconNoise is designed specifically for pyrosequencing reads, so it could not be applied to the Illumina reads. I created QIIME OTUs using a recently recommended procedure for Illumina[3] and commands recommended by the QIIME developers for 454 (J.G. Caporaso, Northern Arizona University, personal communication). See **Supplementary Note 3** for software versions and commands.

UPARSE pipeline. The UPARSE workflow includes merging of paired reads, if applicable; read quality filtering; length trimming; merging of identical reads (dereplication); optionally discarding singleton reads; and OTU clustering using the UPARSE-OTU algorithm. These steps are described in more detail below.

Step 1: merging of paired reads. I merged overlapping paired reads using a method similar to PANDAseq[16]. Pairs with mismatches in the overlap were optionally discarded. In the overlapped region, I calculated the most probable base call and combined Phred score (Q) by calculating the posterior probability of each base according to the corresponding position in the forward and reverse reads. I imposed a minimum Q score of 3, truncating one or both reads if necessary. The result of merging was a FASTQ file, which was filtered in step 2.

Step 2: read quality filtering. I performed quality filtering of reads in FASTQ format by imposing a minimum Phred score Q_{min} for all bases in the read, i.e., by truncating at the first base with $Q < Q_{min}$. By default, Q_{min} was 16 (see **Supplementary Note 3**).

Step 3: length trimming. Step 2 produced reads with variable lengths, which can lead to problems due to terminal gaps in alignments. For example, dereplication (merging of identical sequences) is ambiguous when one read 'A' is an exact match to the prefix of a longer read 'B'. If A and B are merged, then the unmatched suffix of B should have lower confidence because its sequence is supported by fewer reads, but this information is lost. If A and B are not merged, then the information that B supports A is lost. Similar issues arise when clustering at a lower identity for generating OTUs. I avoided these problems by truncating reads at a fixed length (L), discarding reads that were shorter. I used the following values: 454 $L = 250$, Illumina forward and merged $L = 250$ and reverse $L = 200$ (**Supplementary Note 3**).

Step 4: dereplication. I identified the set of unique read sequences and recorded the number of occurrences (abundance) for each sequence.

Step 5: discarding singletons. In typical data sets, a large majority of unique read sequences are singletons, most of which are expected to have at least one error. Most such singletons can be discarded without loss of sensitivity, as the correct sequence will also be present. A small fraction typically has >3% errors, and these can induce a large number of spurious OTUs. I therefore generally recommend discarding all singletons in order to improve specificity at the cost of a small possible loss in sensitivity due to discarding a few good singletons that are the only representatives of their taxa. Singletons can be retained if they match a reference database or for later clustering with new reads that may contain additional representatives of their taxa (**Supplementary Note 2**).

Step 6: OTU clustering. UPARSE-OTU is a greedy clustering method that uses a single representative sequence to define each cluster (OTU), using the following algorithm. A database of OTU sequences is initially empty. Unique read sequences are considered in order of decreasing abundance, motivated by the expectation that more abundant reads are more likely to be correct amplicon sequences[6–8]. If the read matches an existing OTU within the identify threshold (default 97%), the OTU abundance is updated but the database is otherwise unchanged. Otherwise, a model of the read is constructed by the UPARSE-REF algorithm (below) using the current OTU database as a reference. If the model is chimeric, the read is discarded; otherwise, the read is added to the database and thus becomes the representative sequence for a new OTU.

UPARSE-REF algorithm. Given a reference database D of sequences in the sample that is assumed to be complete and correct, UPARSE-REF infers errors in a sequence using parsimony. Similar algorithms have previously been used to validate AmpliconNoise and to find candidate parents in Chimera Slayer[12]. UPARSE-REF was used (i) for identifying errors in mock-community reads using a database containing biological sequences in the sample (for validation, not for OTU construction) and (ii) as a subroutine in UPARSE-OTU, where a database of OTUs was constructed *de novo* from the reads. The goal of UPARSE-REF is to derive a given sequence S with the fewest possible events starting from D. Here, events are mutations that arise from PCR or sequencing errors. This is done by constructing a model sequence M using one or more sequences from the database ('refseqs'). Typically, M is a single refseq representing a nonchimeric amplicon. Otherwise, M is made from m refseq segments that are concatenated to represent a chimeric amplicon. If M has one segment, i.e., is a single refseq, then the distance between M and S is defined to be the number of mismatches $d(S, M)$. These differences are interpreted as sequencer or PCR errors. For modeling chimera formation, crossover points may be introduced. If there are m segments, there are $(m - 1)$ crossover events, and the total distance between S and M is then

$$\Phi(S,M) = d(S,M) + (m-1) \quad (1)$$

When there are insertions or deletions (indels), $d(S, M)$ counts each gap in the pairwise alignment of S and M as one additional difference. Gaps are assumed to arise from an incorrect number of bases in the read or an indel mutation during PCR. The score Φ is thus the total number of base changes, indels and chimeric crossovers needed to explain how S was obtained from sequences in D via amplification and sequencing according to a given model M. Finding an optimal model, i.e., an M that minimizes Φ, thus gives the most parsimonious explanation of S. UPARSE-REF finds an optimal M using dynamic programming, as follows. First, S is aligned to each refseq using the textbook Needleman-Wunsch algorithm, giving N pairwise alignments. Assuming no gaps, then for position j in S, let d_{jk} be 0 or 1 to indicate whether there is a difference with refseq k in its pairwise alignment at that position (0 = match, 1 = mismatch). Let Φ_{jk} be the score of the best model

© 2013 Nature America, Inc. All rights reserved.

of $S_{1...j}$ that terminates at the column containing position S_j in the alignment of S with refseq k. Then

$$\Phi_{j+1,k} = min_{k'=1...N}\{\Phi_{jk'} + d_{j+1,k} + \mathbf{1}(k' \neq k)\} \quad (2)$$

Here, $\mathbf{1}(x)$ is 1 if x is true or 0 if x is false. One may account for gaps by considering them to be differences in the next ungapped column, so with gaps d_{jk} is $\mathbf{1}$(mismatch) + (number of gaps immediately preceding j). This definition ensures that Φ counts both gaps and mismatches as differences between M and S. The recursion relation (equation (2)) is used to minimize Φ by dynamic programming. The model sequence M, number of segments m and number of simple differences d are recovered by trace-back through the dynamic programming matrix Φ_{jk}. M is interpreted as the most parsimonious explanation of S given D.

OTU assessment on mock communities. An OTU representative sequence S for a mock community was assessed by comparison with the reference database (D) containing its known biological sequences. I aligned S to every database sequence and determined the highest pairwise identity (V). If $V = 100\%$, I classified S is as Perfect; if $100\% > V \geq 99\%$, S was Good; or if $99\% > V \geq 97\%$, S was Noisy. If $V < 97\%$ and UCHIME or UPARSE-REF reported S as chimeric by comparison with D, then I classified S as Chimeric, noting that some Good and Noisy sequences may be also be chimeras but that these would be less likely to disrupt downstream analysis. In the case of UPARSE-REF, a high false positive rate is expected with default parameters (**Supplementary Note 3**). When I used UPARSE-REF in assessment, I reduced false positives by setting more stringent parameters: (i) I assigned a breakpoint penalty of 3, giving a higher weight to this event vs. the default of 1 corresponding to unweighted parsimony, and (ii) the model was required to be ≤1% different from S because with higher divergences, parsimony is expected to be less reliable because alternative explanations become more likely. If S was not Chimeric and $V < 95\%$, then I searched S against the NCBI Nucleotide collection (nt) database[17] using MEGABLAST. If I found a hit with identity ≥98% covering ≥98% of S, then I classified S as a Contaminant. I required V to be <95% rather than <97% for a contaminant because I found that with a 97% threshold, OTUs derived from mock species with 3–4% errors often had matches to nt, whereas with $V < 95\%$, these false positive identifications were greatly reduced. If none of the above categories applied, then I classified S as Other. A sequence classified as Other might be a correct novel biological sequence missing from the reference databases but is more likely to be a mock-community sequence with >3% errors. There is thus some uncertainty in the Chimera and Contaminant classifications and in the interpretation of Others (**Supplementary Note 1**). Therefore, the Chimeric, Contaminant and Other categories should be regarded as indicating meaningful trends only if large differences are observed between pipelines, but they should not be considered definitive for individual sequences or used for ranking pipelines if differences are minor.

Sensitivity on mock communities. It is difficult to assess sensitivity because contaminants cannot be fully accounted for. I took the approach of determining how many of the designed species in the mock community were represented in the OTUs. Because some of those species were present in only one or very few of the 454 reads (**Supplementary Note 1**), this number provided a reasonable indication of sensitivity. In the case of Illumina data sets, the minimum species abundance was 11 reads, so these data did not provide a good sensitivity test. I constructed global pairwise alignments for all OTUs with all reference sequences for the designed mock-community species, considering a species having at least one reference sequence aligning to at least one OTU with a pairwise identity ≥97% to be found, and I defined sensitivity as the total number of species found in the OTUs.

Species per OTU on mock communities. Attempting to measure specificity raises the additional complication that the intended or expected taxonomic scope of an OTU is typically not specified, so, for example, a pipeline might generate OTUs approximately corresponding to species, genera or a mix of different taxonomic levels. Any monophyletic group might reasonably be considered a valid OTU. Bearing in mind that no measure can be definitive, I used the ratio of detected species to the number of OTUs as a proxy for specificity, assuming (i) a method that produces approximately one OTU per species is desirable in some applications, and (ii) methods that produce a consistent number of OTUs per species will produce more stable estimates of species richness. As most methods produce many OTUs per species, I used the inverse ratio, i.e., number of species divided by number of OTUs, to produce a measure that usually falls in the range 0–1. I estimated the total number of detected species as the sum of the number of mock species found (as defined for sensitivity above) plus the number of OTUs classified as Contaminant.

16. Masella, A.P., Bartram, A.K., Truszkowski, J.M., Brown, D.G. & Neufeld, J.D. *BMC Bioinformatics* **13**, 31 (2012).
17. Sayers, E.W. *et al. Nucleic Acids Res.* **40**, D13–D25 (2012).

© 2013 Nature America, Inc. All rights reserved.
npg

doi:10.1038/nmeth.2604