Accurate determination of microbial diversity from 454 pyrosequencing data

Christopher Quince[1], Anders Lanzén[2], Thomas P Curtis[3], Russell J Davenport[3], Neil Hall[4], Ian M Head[3], L Fiona Read[3] & William T Sloan[1]

We present an algorithm, PyroNoise, that clusters the flowgrams of 454 pyrosequencing reads using a distance measure that models sequencing noise. This infers the true sequences in a collection of amplicons. We pyrosequenced a known mixture of microbial 16S rDNA sequences extracted from a lake and found that without noise reduction the number of operational taxonomic units is overestimated but using PyroNoise it can be accurately calculated.

Pyrosequencing as implemented by Roche's 454 is a technology that generates a large number of intermediate length DNA reads through a massively parallel sequencing-by-synthesis approach[1]. The GS FLX implementation generates ~400,000 reads of ~250 base pairs in a single run. In many environmental genomics applications of pyrosequencing, DNA is extracted from an entire microbial community, and a particular target region flanked by conserved primers is amplified by PCR before sequencing. This generates an amplicon dataset, in which every read stems from a homologous region, and the sequence variation between the reads reflects the phylogenetic diversity in the community. Because there is no cloning step, resequencing to increase accuracy is not possible and it is therefore vital to disentangle noise from true sequence diversity in this type of data.

During pyrosequencing, each base, in turn, is washed across a plate with hundreds of thousands of wells in which beads attached to multiple copies of a single DNA molecule are localized. If the first unpaired base in a well is complimentary to the incoming base, then synthesis occurs and through a series of chemical reactions light is emitted. Subsequent synthesis and increased light emission will occur if a homopolymer is present. The pattern of light intensities, or flowgram, emitted by each well can then be used to determine the DNA sequence. The major source of noise is that the light intensities do not faithfully reflect the homopolymer lengths. Instead, a distribution of light intensities is associated with each length, and the variance of this distribution increases with length[1]. The standard base-calling procedure is to round the continuous intensities to integers. Consequently, long homopolymers result in frequent miscalls: either insertions or deletions[2].

We calculated the probability distributions of observing a given signal intensity for each homopolymer length by pyrosequencing the V5 region of 23 clones of known sequence (Online Methods). The sequences of these clones differed by at least 7%, and we could therefore unambiguously associate each flowgram with the sequence that generated it. We aligned flowgrams to their parent sequences using an exact Needleman-Wunsch algorithm and then used all signal intensities from each homopolymer length to generate histograms (**Supplementary Fig. 1**). For lengths greater than 5 homopolymers, for which insufficient data was available to construct histograms (**Supplementary Table 1**), we used normal distributions with extrapolated parameters (Online Methods).

The starting point for our algorithm was the realization that we should work with the light intensities associated with each read, or flowgrams, rather than their translations into sequences. Intuitively two sequences can differ substantially, whereas their flowgrams can be similar. To use an example from a real dataset: the true sequence is ACTGGGG, which without noise and with nucleotides flowed in the order TACG, would give the flowgram 0, 1, 1, 0|1, 0, 0, 4, where the | indicates a new series of the four nucleotides. Instead we observed 0.18, 1.03, 1.02, 0.70|1.12, 0.07, 0.14, 4.65, a flowgram that is not that dissimilar but that translates into a sequence, ACGTGGGGG, with two insertions. Using the flowgrams and the distributions of observed intensities, we defined a distance reflecting the probability that a flowgram was generated by a given sequence[3] (Online Methods). These distances were then used in a mixture model to define a likelihood

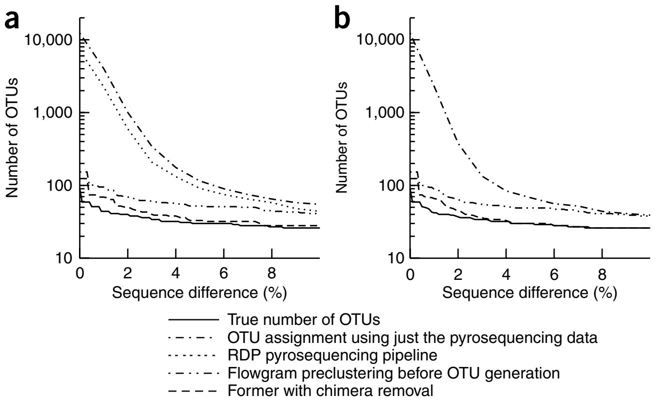

Figure 1 | OTU number as a function of percentage sequence difference for 90 pyrosequenced 16S rRNA gene clones of known sequence. (**a**,**b**) Results are repeated for complete linkage (**a**) and average linkage algorithms (**b**).

[1]Department of Civil Engineering, Rankine Building, University of Glasgow, Glasgow, UK. [2]Department of Biology, University of Bergen, Bergen, Norway. [3]School of Civil Engineering and Geosciences, University of Newcastle upon Tyne, Newcastle upon Tyne, UK. [4]School of Biological Sciences, University of Liverpool, Liverpool, UK. Correspondence should be addressed to C.Q. (quince@civil.gla.ac.uk).

RECEIVED 19 FEBRUARY; ACCEPTED 10 JULY; PUBLISHED ONLINE 9 AUGUST 2009; DOI:10.1038/NMETH.1361

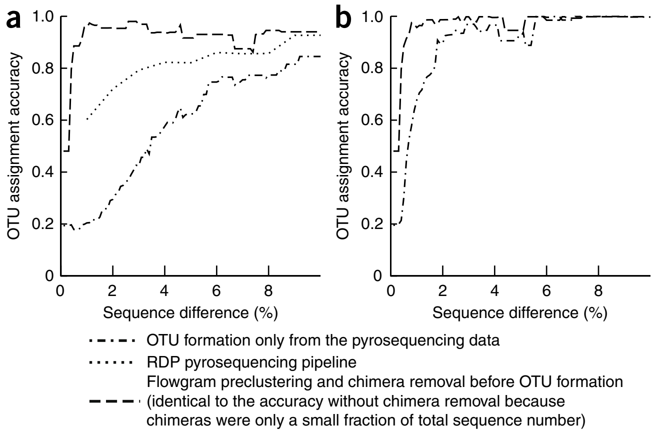

Figure 2 | Proportion of sequences assigned to the correct OTU as a function of percentage sequence difference for pyrosequenced 16S rRNA gene clones of known sequence. (**a**,**b**) Results are repeated for complete linkage (**a**) and average linkage algorithms (**b**).

of observing all the flowgrams assuming that they were generated from a set of true underlying sequences[4]. We used an iterative expectation-maximization algorithm to maximize this likelihood and obtain the de-noised sequences (Online Methods). The algorithm first calculates the most likely set of sequences given the probabilities that each flowgram was generated by each sequence and then recalculates those probabilities given the new sequences. The procedure is then repeated until the algorithm converges. By considering the whole set of flowgrams, this noise removal method, which we refer to as flowgram preclustering, takes the context of a read into account when deducing whether it is noise or a genuinely novel sequence.

To test the algorithm, we applied it to the problem of determining the microbial diversity in a sample. In this case,we amplified a small portion of the 16S rRNA gene by PCR and then pyrosequenced the amplicons. The majority of microbial species have not been taxonomically classified, and their 16S rRNA genes are not known. Diversity must consequently be measured as the number of operational taxonomic units (OTUs) in the sample, defined as the clusters formed at a given level of sequence difference. Typically a complete linkage clustering algorithm is used in which distances between clusters are defined as the maximum distance between their constituent sequences[5]. However, alternative methods of updating distances are possible, in particular average linkage, for which distances are defined as the average of all constituent sequences. Average linkage is referred to as the 'unweighted pair group method with arithmetic mean' in phylogenetic studies. It produces less homogenous clusters than maximum linkage and is therefore not the preferred choice for OTUs[5].

These methods of OTU construction have been developed and tested for full-length dideoxy-sequenced clones for which clusters at 3% sequence difference approximate to species and 5% to genera[5]. Pyrosequencing studies have observed much greater OTU diversity in soils and deep sea vents than previously anticipated[6–8], and estimated diversities including unseen taxa are much larger[9]. However, assignment of OTUs based on sequence divergence has not been tested for pyrosequencing data. Pyrosequencing noise could inflate OTU number by introducing artificial sequence differences. In addition, the effect of PCR errors from the amplification step needs to be considered. Single-base PCR errors will increase the effective per-base sequencing error rate. PCR chimeras[10], which are composed of two or more 'true sequences', need to be treated differently. We adapted the Mallard algorithm[11] to screen for chimeras in the large datasets generated by pyrosequencing (Online Methods).

To quantify the number of OTUs in a sample due to noise, we generated an 'artificial community' of 16S microbial rRNA gene fragments from 90 clones. These clones were isolated from a eutrophic lake (Priest Pot) and dideoxy-sequenced (Online Methods). The number of OTUs observed will depend on the sequence difference used to define the clusters. As sequence difference is increased, clusters merge and OTU number decreases. Accurate OTU construction will only be possible for sequence differences larger than the level of noise. We calculated the number of OTUs that should be observed for this sample by complete linkage clustering of the V5 region of the known clone sequences. We then amplified and pyrosequenced the V5 region of this mixture (**Supplementary Table 2** and Online Methods) and used different algorithms to calculate the number of complete linkage OTUs that were actually observed (**Fig. 1a**). The standard OTU generation method aligning raw sequences to calculate distances (Online Methods) overestimated OTU number; the Ribosomal Database Project (RDP) pipeline that uses quality scores and performs a structural alignment[12] does better but is only accurate at high sequence differences. However, using flowgram preclustering before OTU generation removed the majority of the spurious OTUs, and almost all of the remaining ones were accounted for by chimera removal. We obtained these same results for an experiment in which we considered average linkage (**Fig. 1b**), except for the RDP pipeline, for which this was not an option. We also investigated the effect of noise removal on the accuracy of OTU assignment for complete (**Fig. 2a**) and average (**Fig. 2b**) linkage clustering. Removing noise allowed accurate OTU assignment even at low sequence differences.

These results demonstrate that our algorithm for pyrosequencing noise removal, followed by screening for PCR chimeras, generated sequence data that can be used for the accurate determination of microbial diversity. By comparing the results for complete (**Figs. 1a** and **2a**) and average linkage (**Figs. 1b** and **2b**) clustering it is apparent that the latter is more robust to noise. Consequently if our noise-removal methods, which improve the quality of OTU construction for both complete and average linkage, are not used, average linkage should be preferred. The methods used in previous studies likely overestimate microbial diversity dramatically; applied to the artificial community, they overestimated diversity at least sixfold at the 3% OTU level (**Fig. 1a**). For a sample of pyrosequenced

Table 1 | Numbers of complete linkage OTUs

Method	Sample size	Chimeric number	3% OTUs	5% OTUs
Flowgram preclustering and chimera removal	16,222	479	855	699
Standard method	16,222	–	1,327	877
RDP pipeline	16,222	–	1,208	862

Calculated from pyrosequenced environmental 16S rRNA V5 sequences from Priest Pot lake. –, chimera removal not applied.

environmental 16S rRNA sequences from Priest Pot we found that 3% OTU numbers were reduced by 40% after application of our noise-removal algorithms (**Table 1**). Diversity levels cited in previous reports should thus be treated with caution[6–8].

The rationale for PyroNoise, the flowgram preclustering software, was the problem of accurate construction of OTUs from 16S rRNA sequence data, but it will also aid in the assignment of pyrosequenced reads to known taxa[13]. Applying our algorithm before classification has the advantage that fewer sequences will need to be classified, the relative abundance of different 16S rRNA genes in a sample will be correctly established, and the possibility of noise resulting in an erroneous classification will be reduced. Removing noisy reads will also make phylogenetic tree construction possible for larger datasets and result in more accurate phylogeny based diversity measures[14]. Our algorithm is not restricted to 16S rRNA sequence data: it can be applied whenever a homologous portion of a diverse gene is amplified and pyrosequenced, and could be used to determine eukaryotic microbial diversities or viral diversities in hosts or in population genetics[2].

PyroNoise source code is available as **Supplementary Software**. Data and source code are also available at http://people.civil.gla.ac.uk/~quince/PyroNoise.html.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Accession codes. NCBI Short Read Archive: study, SRP000570 and samples, SRS002051–SRS002053.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank K. Ashelford (University of Liverpool) for the preliminary bioinformatics analysis of the sequencing data and for making the Mallard code publicly available, S. Huse (Marine Biological Laboratory, Woods Hole) for providing the quickdist algorithm and R. Knight for commenting on the manuscript. C.Q. is supported by a Lord Kelvin Glasgow University fellowship, W.T.S. by an Engineering and Physical Sciences Research Council Advanced Fellowship, R.J.D. by a Royal Society fellowship. The Natural Environment Research Council funded sequencing.

AUTHOR CONTRIBUTIONS

T.P.C., R.J.D., I.M.H., C.Q. and W.T.S. designed the study. C.Q. devised algorithms and wrote software. A.L. and C.Q. performed analysis. R.J.D. and L.F.R. performed experiments. N.H. oversaw sequencing. T.P.C., R.J.D., N.H., I.M.H., A.L., C.Q., L.F.R. and W.T.S. wrote the paper.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Margulies, M. *et al. Nature* **437**, 376–380 (2005).
2. Quinlan, A.R., Stewart, D.A., Stromberg, M.P. & Marth, G.T. *Nat. Methods* **5**, 179–181 (2008).
3. Vacic, V., Jin, H., Zhu, J.-K. & Lonardi, S. *Pac. Symp. Biocomput.* **13**, 75–86 (2008).
4. Fraley, C. & Raftery, A.E. *Comput. J.* **41**, 578–588 (1998).
5. Schloss, P.D. & Handelsman, J. *Appl. Environ. Microbiol.* **71**, 1501–1506 (2005).
6. Huber, J.A. *et al. Science* **318**, 97–100 (2007).
7. Roesch, L.F. *et al. ISME J.* **1**, 283–290 (2007).
8. Sogin, M.L. *et al. Proc. Natl. Acad. Sci. USA* **103**, 12115–12120 (2006).
9. Quince, C., Curtis, T.P. & Sloan, W.T. *ISME J.* **2**, 997–1006 (2008).
10. Wang, G.C.Y. & Wang, Y. *Microbiology* **142**, 1107–1114 (1996).
11. Ashelford, K.E., Chuzhanova, N.A., Fry, J.C., Jones, A.J. & Weightman, A.J. *Appl. Environ. Microbiol.* **72**, 5734–5741 (2006).
12. Cole, J.R. *et al. Nucleic Acids Res.* **37**, D141–D145 (2009).
13. Liu, Z.Z., DeSantis, T.Z., Andersen, G.L. & Knight, R. *Nucleic Acids Res.* **36**, 11 (2008).
14. Lozupone, C. & Knight, R. *Appl. Environ. Microbiol.* **71**, 8228–8235 (2005).
15. Eriksson, N. *et al. PLOS Comput. Biol.* **4**, 13 (2008).

ONLINE METHODS

Distance between a flowgram and a sequence. We will denote the probability that a signal of intensity f is observed when n nucleotides are incorporated as $P(f\,|\,n)$ (**Supplementary Fig. 1**). These distributions were approximated for the GSFLX implementation by pyrosequencing known sequences. Flowgrams were aligned to their parent sequences using an exact Needleman-Wunsch algorithm (see below), and then all signal intensities from each homopolymer length used to generate the histograms. If less than 10,000 signal intensities were observed for a given homopolymer length then rather than using the histogram we substituted a Gaussian of mean n. This occurred for all homopolymer lengths with $n > 5$. The s.d., σ_n, for these Gaussians was extrapolated from a linear regression of the observed values as a function of n for $n = 1,\ldots,5$, to give $\sigma_n = 0.04 + 0.03n$ for $n > 5$.

We define the perfect flowgram $\mathbf{U}$ for a sequence $\mathbf{S}$ as the flowgram that would be generated in the absence of noise, and we denote this mapping $\mathbf{S} \rightarrow \mathbf{Y}$. Therefore all intensities are equal to the homopolymer lengths. This will be a vector of length M with elements U^l. The probability that $\mathbf{S}$ would generate a real flowgram $\mathbf{f}$ is

$$P(\mathbf{f}\,|\,\mathbf{U}) = \prod_{l=1}^{M} P(f^l\,|\,n = U^l)$$

assuming signal intensities are independent. In general these probabilities will be very small. Numerically it is easier to work with the negative logarithm of this quantity. We will refer to this as a 'distance' between a flowgram and a sequence, although it fails to meet the mathematical criteria of a true distance. The distance between a flowgram and a sequence is then the sum of the distances at each flow:

$$d(\mathbf{f}, \mathbf{S} \rightarrow \mathbf{U}) = -\log\left(\prod_{l=1}^{M} P(f^l\,|\,n = U^l)\right) = \sum_{l=1}^{M} -\log(P(f^l\,|\,n = U^l)) = \sum_{l=1}^{M} d(f^l, U^l) \tag{1}$$

Alignment of flowgrams to sequences. Equation 1 does not allow for the possibility of deletions or insertions during the pyrosequencing process. Given the underlying mechanics, this is probably reasonable but we want a distance measure that applies even when a flowgram has not been generated by exactly that sequence to account for single base errors introduced during PCR amplification. For this reason, we adapted the global Needleman-Wunsch alignment algorithm to align flowgrams to sequences[16]. We used this dynamic programming method to exactly find the alignment that minimized the distance between the flowgram and the sequence. We defined the cost of matching a flow of intensity f to a homopolymer of length n as $d(f, n) = -\log(P(f\,|\,n))$ and used a fixed gap cost G of 15.0, chosen to be just larger than $d(1.0\,|\,0)$ so that introducing a gap was preferred to matching a signal to no signal. The only subtlety is that gaps must be introduced in groups of four to reflect the period of nucleotide flows across the plate. Internal gap costs are included in the distance measure. It is useful to normalize the distances in equation 1 by the alignment length M between a flowgram and a sequence so that a short flowgram is not considered more likely to be generated from the same sequence than a longer one $d'(\mathbf{f}, \mathbf{S}) = d(\mathbf{f}, \mathbf{S}) \,/\, M$.

An algorithm to remove pyrosequencing noise by flowgram preclustering. Our method to remove pyrosequencing noise used model-based clustering. We assumed that the likelihood of the observed flowgrams is described by a mixture model[4], and each component of the mixture corresponds to a different sequence. Assume that L sequences, $\mathbf{S}_j \rightarrow \mathbf{U}_j$, each indexed by j are present with relative frequency τ_j. To define the likelihood we need the density of the observed flowgrams about each sequence. It is natural to use our distances to define this density. We assumed that the flowgrams are distributed as exponentials about the sequences with a characteristic cluster size σ, to give

$$F(\mathbf{f}_i\,|\,\mathbf{S}_j) = \frac{\exp\left(-\dfrac{d'(\mathbf{f}_i, \mathbf{S}_j)}{\sigma}\right)}{\sigma}$$

This form for the density function was chosen as it was a reasonable fit to the distribution of distances about sequences observed in our 'divergent sequences' dataset. The likelihood of the dataset $\mathbf{D}$ of N flowgrams indexed i is then:

$$L(\mathbf{D}\,|\,L; \tau_1, \ldots, \tau_L; \mathbf{S}_1, \ldots, \mathbf{S}_L) = \prod_{i=1}^{N}\left(\sum_{j=1}^{L} \tau_j F(\mathbf{f}_i\,|\,\mathbf{S}_j)\right) \tag{2}$$

Expectation-maximization algorithm. To find a solution to the set of sequences that generated the flowgrams we used an expectation-maximization algorithm (EM). In EM for clustering the complete data are considered to include both the observed data and the mapping of data points to clusters, referred to as the unobserved data. In our case these are which sequences generated which flowgrams. This can be represented by a matrix $\mathbf{Z}$ with rows for each flowgram and columns for each sequence, so that $z_{i,j} = \delta_{i,m(i)}$, where $m(i)$ gives the sequence that generated flowgram i. The complete data likelihood is then:

$$LC(\mathbf{D}, \mathbf{Z}\,|\,L; \tau_1, \ldots, \tau_L; \mathbf{S}_1, \ldots, \mathbf{S}_L) = \prod_{i=1}^{N}\prod_{j=1}^{L} (\tau_j F(\mathbf{f}_i\,|\,\mathbf{S}_j))^{z_{i,j}} \tag{3}$$

assuming that each row of $\mathbf{Z}$, the vector $\mathbf{z}_i$, is independently and identically distributed according to a multinomial over L categories with probabilities $\tau_1,\ldots,\tau_L$. We then define the quantity $z'_{i,j}$ as the conditional expectation of $z_{i,j}$ given the model parameters, that is, the sequences and their abundances, under the complete data likelihood. These are the conditional probabilities that sequence j generated flowgram i. The EM iterates between an E step, where the $z'_{i,j}$ are computed given the model parameters and an M step, where the model parameters are calculated so as to maximize equation 3 with the $z_{i,j}$ replaced by their estimates $z'_{i,j}$. This process will, under quite general conditions, converge to a local maximum of equation 2 (ref. 4). Our EM algorithm is: M step: given the $z'_{i,j}$ generate sequences $j = 1,\ldots, L$ using the aligned flowgrams, such that at each position $l = 1, \ldots, M$:

$$U_j^l = \min_n\left(\sum_{i=1}^{N} z'_{i,j}\, d'(f_i^l, n)\right)$$

doi:10.1038/nmeth.1361

If at the position l a flowgram alignment has a gap, then it is ignored in the calculation of the true homopolymer length at that point. Gaps in the sequence alignment are irrelevant. Define new relative frequencies as:

$$\tau_j = \sum_{i=1}^{N} \frac{z'_{i,j}}{N}$$

This generates sequences and their frequencies which maximize the complete data likelihood, eqution 3, given **D** and **Z**.

Align flowgrams to sequences to obtain new distances $d'(\mathbf{f}_i, \mathbf{U}_j)$.

E step: calculate new $z'_{i,j}$ as:

$$z'_{i,j} = \frac{\tau_j \exp\left(-\frac{d'(\mathbf{f}_i, \mathbf{S}_j)}{\sigma}\right)}{\sum_{k=1}^{L} \tau_k \exp\left(-\frac{d'(\mathbf{f}_i, \mathbf{S}_k)}{\sigma}\right)}$$

Repeat until convergence.

Expectation-maximization algorithms because they only find local optimum are sensitive to initial conditions. These and the number of sequences L were determined using a preliminary hierarchical clustering between flowgrams. The distances between pairs of flowgrams for this hierarchical clustering were defined as the negative logarithm of the probability that both flowgrams were generated by the same sequence divided by the probability that each was generated independently. This is a Bayesian test of the hypothesis H that a pair of flowgrams i and j are from the same sequence:

$$P(H \mid \mathbf{f}_1, \mathbf{f}_2) = \prod_{l=1}^{M} \frac{\sum_{n=0}^{\infty} P(f_1^l \mid n) P(f_2^l \mid n) P(n)}{\left(\sum_{n=0}^{\infty} P(f_1^l \mid n) P(n)\right)\left(\sum_{n=0}^{\infty} P(f_2^l \mid n) P(n)\right)}.$$

The priors for the homopolymer lengths $P(n)$, were defined as their normalized frequencies in the 'divergent sequences' dataset provided they occurred over 10,000 times (**Supplementary Table 1**), otherwise $(1/4)^n$ was used. The distances obtained by taking the negative logarithm of this equation were normalized by the alignment length M and used as the input to a complete linkage clustering algorithm. The clusters formed at a given cut-off distance c were used as the input to the EM.

Implementation. The EM algorithm was implemented as a message parsing interface (MPI) program to run on Linux clusters. The maximum dataset size that can be processed in a reasonable amount of time (~1 d) on a cluster with 128 processors is 10,000 flowgrams. To denoise larger datasets we divided them into the clusters formed at 35% sequence difference and then denoised the flowgrams in the individual clusters before recombining. There are two adjustable parameters in the flowgram preclustering: σ, the cluster size and c, the cut-off in the hierarchical clustering used to initialize the EM. Both the datasets were run with $\sigma = 1/15$ and $c = 0.05$, values that gave good results. During each run the clusters with high initial weight typically draw in smaller clusters that are close by and the number of sequences with nonzero weight τ_j decreases. The final number of sequences is therefore less than or equal to L and generated automatically. The algorithm was determined to have converged once the maximum change in $z'_{i,j}$ was smaller than 10^{-6} or a thousand iterations had been exceeded. After running the algorithm each flowgram i is assigned to the sequence j for which $z'_{i,j}$ was largest. Output consists of a set of denoised sequences with integer weights giving the number of flowgrams mapping to that sequence.

Chimera detection. The generation of 16S rRNA sequence libraries by pyrosequencing typically contains a PCR step where a portion of the 16S rRNA gene is amplified with the 454 A and B adaptors before sequencing. The possibility of PCR chimeras being generated during this step has hitherto been ignored even though the conditions, many different sequences often closely related, are highly conducive to chimera generation. Flowgram pre-clustering generates a set of denoised sequences which in general is much smaller than the original number of reads. It is therefore possible to search these sequences for PCR chimeras. To do this we adapted the Mallard algorithm[11]. Rather than performing pair-wise comparisons of all sequences as in the original algorithm, we first considered all sequences with weights greater than ten, these were compared to each other and all judged chimeric by the Mallard algorithm with 100% outlier cut-off were removed. This generated a set of reference sequences that we could be confident in. The rest of the sequences were then compared to a random subset of 200 of these reference sequences in blocks of 200: if more than one chimeric interaction was observed with an outlier cut-off of 99.9% then this sequence was removed. This procedure reflects our greater confidence in sequences with higher weights and avoids having to do large multiple alignments and too many pairwise comparisons. In addition, our program differed from the original Mallard implementation in not considering pairwise comparisons with a sequence difference greater than 25%, in our experience comparisons based on sequences more different than this were often falsely judged chimeric. These two changes resulted in a substantial reduction in the false positive rate over the default implementation. The multiple alignments for the Mallard algorithm were performed using the program MAFFT-G-INS-i[17].

Generation of sequence data. Supplementary Table 2 provides a summary of the sequence data used in this study.

A 5-l sample of Priest Pot lake water was collected in winter, transported on ice and filtered through a 0.2-μm membrane filter to concentrate the microorganisms. Microbial biomass was washed from the 0.2-μm filter using 50 ml of the filtrate and the DNA was extracted using a FastDNA kit for soils according to the manufacturer's instructions (MP Biomedicals). This sample was then amplified using modified primers 787f and 1492r, with improved coverage across bacterial taxa[7], with attached 454 adaptors A and B, respectively, at the 5′ end. We denote the combined oligonucleotide sequences A-787f and B-1492r (**Supplementary Table 3**). The PCR conditions were 95 °C for 5 min, 30 cycles of 95 °C for 30 s; 57 °C for 30 s; 72 °C for 1 min; and finally 72 °C for 7 min. The polymerase used was standard Taq polymerase. Several PCR amplifications were carried out in order to obtain a total of 500 ng of DNA for pyrosequencing, the bulked products of which were cleaned-up using a QIAquick PCR purification kit according to the manufacturer's instructions (Qiagen Ltd.). Pyrosequencing

doi:10.1038/nmeth.1361

of these amplicons was then performed from the A adaptor and a total of 28,361 reads obtained.

Priest Pot clones: in addition a library of 94 16S rRNA clone products was prepared using primers 787f and 1492r from the same environmental DNA sample. A TOPO TA cloning kit was used for sequencing with a pCR4-TOPO vector according to the manufacturer's instructions (Invitrogen). The clone products were then amplified using T7 and T3 primers and Sanger sequenced using the 454 adaptor A as a primer. Two mixtures of these PCR products were then prepared. (i) Divergent sequences: a sample comprising 23 clone products mixed in equal proportions. These clones differed by at least 7% to allow unambiguous classification of pyrosequencing reads. (ii) Artificial community: a sample comprising 90 products from the clone library mixed in proportions that varied by two orders of magnitude and were determined to approximate 3% OTU abundances in the Priest Pot pyrosequencing data. This sample provides an approximation to a real community with variable abundances and sequences that can be very similar.

These two mixtures were then pyrosequenced following amplification with A-787f adaptor primer and the B-1492r primer as above. Read numbers were 57,902 and 46,249 for the two datasets respectively.

OTU construction: initial noise removal. Noise in pyrosequencing data can be greatly reduced by removal of short reads and reads containing noisy bases, defined as a signal intensity between 0.5 and 0.7 (ref. 18). We therefore curtailed flowgrams when a noisy read was observed and removed all flowgrams where this gave a sequence of less than 200 bases. In addition we removed all reads which did not possess a perfect copy of the primer sequence.

OTU generation: standard methods. The 'standard method' of OTU generation for pyrosequencing data begins with a multiple alignment of the unique sequences in the dataset[6–8]. We used MUSCLE with arguments, - maxiters 2 - diags, to do this[19]. These parameters, which restrict the number of iterations, were necessary because of the large size of the datasets. The multiple alignment was used to define distances between reads as the percentage base-pair difference using the quickdist algorithm[6]. Terminal gaps were ignored and internal gaps counted as one base pair difference regardless of length. This distance measure was used throughout this study. These distances were used to perform a hierarchical clustering of sequences and OTUs were defined at a given level of sequence dissimilarity. This is identical to the OTU construction procedure used in the package DOTUR[5]. We used two hierarchical clustering algorithms, complete linkage, in which distances between clusters are defined as the maximum distance between all their members, and average linkage (unweighted pair group method with arithmetic mean (UPGMA)) for which distances are the average between members[5]. In the latter case the frequencies of the unique sequences were taken into account, in the former they are irrelevant. Complete linkage is typically used in diversity estimation from 16S rRNA sequence data. In addition we processed our data (both sequences and qualities) using the pyrosequencing pipeline on the RDP 10 web server, in which noise removal and alignment of sequences is followed by a complete linkage clustering to determine OTUs[12].

OTU generation after noise removal. After flowgram preclustering, and PCR chimera removal, multiple alignments of the denoised sequenced were generated by either MAFFT-G-INS-i[17] for the artificial community and MUSCLE[19] for Priest Pot samples. Distance matrices were then calculated, and OTUs formed by hierarchical clustering using either the complete or average linkage algorithms.

OTU assignment accuracy. To test the accuracy of the assignment of reads to OTUs we performed a BLAST[20] search of each sequence in the artificial community dataset against the 90 clones and classified sequences according to their closest clone sequence. From the clustering of the known V5/V6 clone sequences we determined what the true assignment to OTUs at any given cut-off should be. We then labeled the reads with these assignments. A good OTU generation algorithm should reconstruct this labeling. Starting with the largest OTU, we associated it with the most frequent true OTU label that was unassigned among its reads. The accuracy of OTU construction was defined as the number of reads whose labels matched that of their OTU.

16. Needleman, S.B. & Wunsch, C.D. *J. Mol. Biol.* **48**, 443–453 (1970).
17. Katoh, K., Kuma, K., Toh, H. & Miyata, T. *Nucleic Acids Res.* **33**, 511–518 (2005).
18. Huse, S.M., Huber, J.A., Morrison, H.G., Sogin, M.L. & Mark Welch, D. *Genome Biol.* **8**, 9 (2007).
19. Edgar, R.C. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
20. Altschul, S.F. *et al.* *Nucleic Acids Res.* **25**, 3389–3402 (1997).

doi:10.1038/nmeth.1361