Noise and the Accurate Determinination 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 Head[3], L Fiona Read[3] & William T Sloan[1]

[1]*Department of Civil Engineering, Rankine Building, University of Glasgow, Glasgow G12 8LT, UK*

[2]*Computational Biology Unit, Bergen Center for Computational Science, University of Bergen, Bergen, Norway*

[3]*School of Civil Engineering and Geosciences, University of Newcastle upon Tyne, Newcastle upon Tyne NE1 7RU, UK*

[4]*School of Biological Sciences, Biosciences Building, Crown Street, University of Liverpool, Liverpool L69 7ZB, UK*

A microbial community can be characterised by amplifying and sequencing a random sample of SSU - 16S rRNA genes. Diversity is measured as the number of clusters or operational taxonomic units (OTUs) at a given level of sequence difference. Clone based techniques because they are time consuming and expensive are limited to small sample sizes relative to bacterial population numbers. In constrast, pyrosequencing can provide large numbers of reads quickly and cheaply. However, hitherto in calculating OTU numbers from pyrosequencing data sequencing noise has been ignored. We show using samples of known diversity that both pyrosequencing noise and noise from PCR need to be taken into account when calculating the

number of OTUs. We introduce a new algorithm capable of distilling the true sequences from pyrosequenced 16S rRNA gene sequence tag data and adapt an algorithm for PCR chimera detection to large sample sizes. Using these techniques it is possible to accurately quantify microbial diversity. Our results have important consequences for previous pyrosequencing studies of microbial diversity.

The use of small subunit ribosomal RNA sequences - 16S for prokaryotes - to identify microbial taxa has a long history[1]. It was originally developed to determine the phylogenetic relationships between isolated culturable organisms but with the advent of techiques to recover rRNA gene fragments directly from environmental samples it can be used, through sequencing of a sample of 16S gene fragments from the environment, to determine the taxonomic composition of a whole microbial community[2]. Unfortunately cloning of PCR amplified 16S rRNA gene fragments and Sanger sequencing of individual clones is time consuming and expensive. For this reason sequencing of clone libraries of 16S rRNA gene fragments is usually limited to no more than 1000 clones. Microbial communities have very large population sizes, a gram of soil may contain a billion bacteria, and are in many environments very diverse[3,4], consequently clone libraries are only capable of sampling a tiny fraction of individuals and a small proportion of the different sequences and hence taxa present.

Pyrosequencing as implemented by 454 is a new technology that can generate a large number of intermediate length DNA reads through a massively parallel sequencing by synthesis approach[5]. The GSFLX incarnation of the Roche 454 technology generates around 400,000 reads of ca. 250

base pairs (bp) in a single run, and the older GS20 generates ca. 250,000 reads of around 100 bp. This technology has the potential to provide the sample sizes necessary to definitively characterise microbial communities. A small portion of the 16S rRNA gene is amplified by PCR, and the amplicons or tags are then pyrosequenced. This technique has been recently applied to environments such as deep sea vents and soils[6,7]. Of particular interest in these studies are estimates of the total diversity in these communities i.e. the total number of taxa present. Because the majority of microbial species have not been taxonomically classified, and arguably even if they were, a strictly sequence based approach may be more appropriate for microbes, diversity is measured on the basis of the number of operational taxonomic units or OTUs in the sample. OTUs are defined as the clusters formed at a given level of 16S rRNA sequence difference following a hierarchical clustering of the sequences in the sample. Typically a complete linkage clustering algorithm is used where distances between clusters are defined as the maximum distance between their constituent sequences. This method was developed and tested for full length Sanger sequenced clones and very roughly clusters at 3% sequence difference, correspond to species and genera are often delineated at 5% sequence difference[8].

Results from the pyrosequencing studies have revealed staggering levels of OTU diversity in soils and deep sea vents[6,7], with true diversities including unseen taxa being potentially much larger[9]. However, these results should be treated with caution as the assignment of OTUs based on particular levels of sequence divergence has hitherto not been tested for pyrosequencing data, and there are three good reasons why it may not be appropriate. Firstly, there is the issue of short read length, the degree to which nucleotides are conserved varies along the 16S rRNA gene, there

are hyper-variable and conserved regions, therefore OTUs derived from 3% sequence difference across the whole gene may not correspond to the OTUs generated from a 250 bp or 100 bp region. Secondly, there is the question of sequencing noise, pyrosequencing generates reads with a high proportion of insertions and deletions associated with long homopolymers. The average per base error rate can be reduced to 0.25% by screening for features associated with noisy reads[10], but this is still sufficiently noisy that assuming independent errors, approximately 3% of reads will differ from the true sequence at a least 4% of nucleotide positions.

These reads would be falsely allocated to a novel OTU using the frequently used 3% divergence cut off for delineating OTUs. For a complete GSFLX run this will generate in excess of over 7,000 spurious OTU predictions and because the DNA fragment sequenced is not isolated as an E. coli clone they can not be resequenced to increase accuracy. Finally, since a small homologous portion of the 16S rRNA gene is amplified then PCR noise needs to be considered too. The rate of PCR point mutation, which will depend on the cycle number, will increase the effective per base error rate. More subtly PCR can introduce chimeras - formed when an incompletely extended sequence hybridizes with a similar but different sequence - comprised of portions of two or rarely more original sequences[11]. If the chimera's 'parents' are sufficiently different this too will create a new spurious OTU.

It is the last two of these issues, the effect of noise on microbial OTU determination from pyrosequencing that forms the focus of this paper. The question of read length and optimal starting point for OTU determination is also important, but it is really about how the diversity of a gene

fragment translates into the diversity of the whole gene. The first step in accurately determining the number of OTUs in a pyrosequenced sample is to accurately measure the diversity of the gene fragment despite the effects of noise. Our approach to quantify the number of OTUs in a sample that can be accounted for by noise, was to generate a mixture of 16S rRNA gene fragments of known sequence and hence OTU number. We then compared these known OTU numbers with the number of OTUs measured from pyrosequencing this mixture.

We extracted microbial DNA from water from Priest Pot a small eutrophic lake in northern England, and constructed a library of 16S rRNA gene clones, which we then Sanger sequenced. The V5/V6 region of a mixture of equal amounts of 23 different clones from this library was sequenced by pyrosequencing and these data were used to characterise noise from the GSFLX sequencer. This information was then processed using a new algorithm we developed, and refer to as flowgram preclustering, to remove noise. Flowgram preclustering should remove pyrosequencing noise but artefacts introduced by the PCR which represent genuine sequences will still remain. For this reason we developed a new high throughput version of the Mallard PCR chimera detection algorithm[12] suitable for applying to these large data sets. We then pyrosequenced a second mixture of 95 clones to test the ability of our algorithms to remove noise and hence allow accurate OTU determination. This second mixture contained potentially similar sequences at concentrations that varied over several orders of magnitude to mimic a natural community. In addition we applied the same V5/V6 pyrosequencing procedure to the original Priest Pot DNA sample and accurately determined OTU diversity in this sample using our noise removal algorithms. The OTU diversity obtained from data subject to the noise removal algorithm was compared with a method applied

previously for OTU assignment[6,7,15] and the RDP pyrosequencing pipeline[13]. Finally we discuss the accuracy of previous studies based on the older GS20 platform.

Results

Flowgram preclustering. An algorithm to remove pyrosequencing noise requires a careful consideration of the origin of that noise. During pyrosequencing each base in turn is washed across the plate. The plate contains many hundreds of thousands of wells where beads attached to multiple copies of a single DNA molecule are localised together with DNA polymerase and a chemiluminescent enzyme complex (luciferase plus ATP sulfurylase). If the first unpaired base in a well is complimentary to the current base then synthesis occurs pyrophosphate is released which reacts with adenosine phosphosulfate to make ATP which drives light emission via luciferase, further synthesis and increased light emission will occur if a homopolymer is present i.e. the base repeats. The pattern of light intensities, or flowgram, emitted by each well can then be used to determine the sequence present. 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. The standard base calling procedure is simply to round the continuous intensities to integer homopolymer lengths, consequently long homopolymers result in frequent miscalls, either insertions or deletions[14].

We calculated the probability distributions, denoted $P(f|n)$, of observing a signal of intensity f given a hompolymer of length n for the GSFLX, by pyrosequencing the V5/V6 region of

23 clones of known sequence. These clones differed by at a least 7% and it was therefore possible (after screening for chimeras) to unambigously associate each flowgram with the sequence that generated it. Flowgrams were aligned to their parent sequences using an exact Needleman-Wunsch algorithm (Supplementary Methods online), and then all flows from each homopolymer length used to generate histograms approximating the $P(f|n)$. For long homopolymers where insufficient data was available to construct these distributions we used Gaussians of mean and variance extrapolated by linear regression (Supplementary Methods). The distributions together with equivalent results for the GS20 machine collated from the data of Huse et al.[10] are shown in Figure 1. The increase of noise with hompolymer length is readilly apparent from Figure 1, more surprisingly the newer GSFLX technology actually exhibits more noise than the older GS20, perhaps due to the longer read lengths.

The starting point for our algorithm was the realisation that we should work with the actual light intensities associated with each read - or flowgrams - rather than the translations of those flowgrams into sequences. Intuitively two sequences can differ substantially whilst their flowgrams can be quite similar. Using the flowgrams and the distributions in Figure 1 it was possible to define a distance reflecting the probability that a flowgram was generated by a given sequence and conversely the Bayesian posterior probability that a sequence is consistent with a set of flowgrams (Supplementary Methods). Using these equations we applied an iterative expectation-maximization (EM) algorithm to cluster the flowgrams and produce a set of sequences consistent

with them. 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 it converges. This noise removal method, which we refer to as flowgram preclustering, by considering the whole set of flowgrams, takes the context of a read into account when deducing whether it is noise or a genuinely novel sequence.

Analysis of the artificial community. The number of OTUs observed in a sample will depend on the fractional sequence difference or cut-off used to define the clusters. As cut-off is increased clusters merge and the OTU number decreases. Accurate OTU construction will only be possible for cut-offs larger than the level of sequence noise. Our aim is to be able to accurately determine the number of OTUs and the assignment of individual reads to OTUs at low cut-offs.

In Figure 2 we plot the number of OTUs as a function of cut-off for the artificial community. All the results in Figure 2a are for complete linkage clustering. The black line gives the result of clustering the known V5/V6 regions of the 95 clones. A good algorithm should reproduce these results. The standard OTU generation method aligning raw sequences (see Methods) overestimates OTU number (dot-dashed line), the RDP pipeline does somewhat better (dotted line), but is only accurate at very high cut-offs. Flowgram preclustering prior to OTU formation (dot-dot-dashed line) removes the majority of the spurious OTUs and almost all the rest are accounted for by chimera removal (dashed line). These results are repeated for UPGMA clustering in Figure 2b - except for the RDP pipeline where this is not an option.

The effect of noise removal on the accuracy of OTU assignment (see Methods) is shown

for complete linkage clustering and average linkage in Figure 3. Here we do not show the results for flowgram clustering without chimera removal as most chimeras occur with low frequencies and hence the accuracy is effectively the same as for flowgram clustering with chimera removal. Removing noise allows accurate OTU assignment even at low cut-offs. However, the improvement over the standard method without noise removal is less pronounced when UPGMA is used. This and the slight reduction in the number of false OTUs for the standard method using UPGMA apparent in Figure 2, reflects the fact that UPGMA is more robust to noise than complete linkage. This is because noisy sequences which occur in low frequencies affect the average of all pairs of distances between two clusters less than the maximum distance.

Diversity of Priest Pot. Applying our noise removal algorithms prior to OTU construction allows the accurate determination of OTU numbers and relative abundances. In Table 1 we apply them to the 16S rRNA sequences covering the V5/V6 region recovered from Priest Pot and calculate OTUs at 3% and 5% sequence difference following complete linkage clustering. For comparison we also show OTU numbers after simply performing a MUSCLE alignment of the raw sequences and using the RDP pyrosequencing pipeline. As with the artificial community, noise removal dramatically reduces the number of observed OTUs, with the true OTU number at 3% sequence difference being roughly half that predicted by the standard method without noise removal. In Figure 4 we also show the number of taxa with a given abundance for the standard method and after noise removal. From this graph it is apparent that despite the screening of noisy sequences a large number of OTUs with low abundance remain. This is supported by the Chao estimates of total diversity in Table 1 for the data following nosie removal, which are significantly larger than

the observed diversity. We discuss the implications of this below.

Discussion

We have demonstrated that our algorithm for pyrosequencing noise removal, followed by screening for PCR chimeras, is capable of providing sequence data which can be used for the accurate determination of microbial diversity. We expect that these methods will become standard in this field. It is important to emphasise that if noise removal is not performed then the more robust average linkage or UPGMA algorithm gives more accurate results for OTU assignment than complete linkage. The methods used in studies prior to this are certainly inaccurate, when applied to our data from Priest Pot they overestimated diversity two-fold. Consequently figures cited in previous reports should be treated with a great deal of caution[6,7,15]. These were based on the older GS20 platform with different noise characteristics from the GSFLX so it is difficult to know exactly to what extent the diversities were over-estimated. Given the additional difficulties of aligning the large number of highly divergent short reads considered in these studies the situation could be substantially worse than the situation with data generated using for the GSFLX platform considered here.

To get some idea of the scale of the problem we generated OTUs using complete linkage without performing noise removal for a data set of 99189 GS20 pyrosequencing reads from the

study of Huse et al.[10]. In that study the V6 region of 16S rRNA sequences from a mixture of 43 clones were sequenced to determine pyrosequencing error rates. Consequently the true OTU number should not exceed 43 whereas we found 1340 OTUs at 3% sequence difference, a huge excess of erroneous diversity. However, it is likely that the ranking of diversities cited in previous studies will prove correct, and more importantly the fundamental observation that there are many rare taxa, the so called 'rare biosphere'[15], remains intact. Indeed from Figure 4 we see some evidence that after noise removal the rate at which new taxa are being uncovered actually increases. The cause, extent, and function of the rare biosphere remain therefore vital unanswered questions in microbial ecology. Questions that the algorithms presented here will, by allowing the accurate construction of taxa-abundance distributions, help answer.

We have motivated our noise removal algorithms in terms of the accurate construction of OTUs from 16S rRNA sequence data but they are also useful for assignment of pyrosequence reads to known[16]. Classification is less sensitive to noise than clustering as noisy reads would have a low probability of originating from any known taxa. Applying our noise filtering algorithms also has the advantages that less sequences will need to be classified, the abundances of known taxa are correctly established, and the possibility of noise resulting in an erroneous classification will be reduced. Further, the usefulness of our algorithms are not restricted to microbial 16S rRNA sequence data, they can be applied whenever a homologous portion of a diverse gene is amplified and pyrosequenced, as such they could have application in determining eukaryotic microbial diversities, viral diversities in hosts, and may even be useful for population genetics[14, 17, 18].

Methods

Generation of sequence data. A summary of the sequence data used in this study is given in Table 2. These data sets are available for download from the online supplementary materials.

GS20 data: 111321 sequences generated by pyrosequencing of 43 reference sequences. The reference sequences were 16S rRNA V6 regions chosen to be widely separated in sequence space. The closest pair of sequences are slightly more than 3% divergent. See Huse et al.[10] for details.

GSFLX - Priest Pot: a sample of environmental DNA was extracted from Priest Pot. This sample was then pyrosequenced using primers 787f with a 454 A adaptor at the 5' end and 1492r with the B adaptor. The primers were modified slightly to increase redundancy at some positions and improve coverage across bacterial taxa (787f -ATTAGATACCCNGGTAG: 1492r -GNTACCTTGTTACGACTT) as in Roesch et al.[7]. Pyrosequencing of these amplicons was then performed from the A adaptor and a total of 28361 reads obtained.

Priest Pot clones: in addition a library of 95 16S rRNA clones was prepared using primers 787f and 1492r from the same environmental DNA sample. These clones were then amplified and Sanger sequenced. Two mixtures of these PCR products were then prepared:

1. Divergent sequences: a sample comprising 23 clones mixed in equal proportions. These clones differed by at least 7% to allow unambigous classification of pyrosequencing reads.
2. Artificial community: a sample comprising all 95 clones 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 787f - A adaptor primer and the 1492r - B primer as above to give the GSFLX - divergent and GSFLX - community data sets. Read numbers were 57902 and 46249 for the two data sets respectively.

Initial noise removal. Noise in pyrosequencing can be greatly reduced by removal of short reads and reads containing noisy bases, defined as a flow intensity between 0.5 and 0.7[10]. We therefore curtailed flowgrams when a noisy read was observed and removed all flowgrams where this gave a sequence of less than 80 bases (GS20) and 200 bases (GSFLX). 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 alignment of the unique sequences in the data set [6,7,9,15]. We used MUSCLE with arguments, -maxiters 2 -diags, to do this. These parameters, which restrict the number of iterations, were necessary because of the large size of the data sets. 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 dissimiliarity. We used two hierarchical clustering algorithms, complete linkage - where distances between clusters are defined as the maximum distance between all their mem-

bers - and average linkage or UPGMA where distances are the average between members. 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 sequence data.

In addition we processed our data (both sequences and qualities) using the pyrosequencing pipeline on the RDP 10 web server, where noise removal and alignment of sequences is followed by a complete linkage clustering to determine OTUs[13].

OTU generation following noise removal. Following flowgram clustering, and PCR removal (Supplementary Methods), OTUs were generated as above by hierarchical clustering of distance using either the complete linkage or UPGMA algorithms.

Accuracy of OTU generation. To test the accuracy of the assignment of reads to OTUs we performed a BLAST search of each sequence in the artifical community data set against the 95 clones and classified sequences according to their closest clone sequence[19]. From the clustering of the known V5/V6 clone sequences we were then able to determine what the true assigment to OTUs at any given cut-off should be. We then labelled the reads with these assignments. A good OTU generation algorithm should reconstruct this labelling. Starting with the largest OTU, we associated it with the most frequent true OTU label which was unassigned amongst its reads. The accuracy of OTU construction was then defined as the number of reads whose labels matched that of their OTU.

1. Tringe, S. G. & Hugenholtz, P. A renaissance for the pioneering 16S rRNA gene. *Curr. Opin. Microbiol.* **11** (2008).

2. Hugenholtz, P. Exploring prokaryotic diversity in the genomic era. *Gen. Biol.* **3**, 0003.1–0003.8 (2002).

3. Torsvik, V. & Ovreas, L. Microbial diversity and function in soil: from genes to ecosystems. *Curr. Opin. Microb.* **5**, 240–245 (2002).

4. Gans, J., Wolinsky, M. & Dunbar, J. Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science* **309**, 1387–1390 (2005).

5. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).

6. Huber, J. A. *et al.* Microbial population structures in the deep marine biosphere. *Science* **318**, 97–100 (2007).

7. Roesch, L. F. W. *et al.* Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME* **1**, 283–290 (2007).

8. Schloss, P. D. & Handelsman, J. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl. Environ. Microbiol.* **71**, 1501–1506 (2005).

9. Quince, C., Sloan, W. T. & Curtis, T. C. The rational exploration of microbial diversity. *ISME* **in press** (2008).

10. Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L. & Welch, D. M. Accuracy and quality of massively parallel DNA pyrosequencing. *Gen. Biol.* **8**, R143 (2007).

11. Wang, G. C.-Y. & Wang, Y. The frequency of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from different bacterial species. *Microbiol.* **142**, 1107–1114 (1996).

12. Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J. & Weightman, A. J. New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras. *Appl. Environ. Microb.* **72**, 5734–5741 (2006).

13. Cole, J. R. *et al.* The ribosomal database project: improved alignments and new tools for rRNA analysis. *Nucl. Acids Res.* **Advance Access**, doi:10.1093/nar/gkn879 (2008).

14. Quinlan, A. R., Stewart, D. A., Stromberg, M. P. & Marth, G. T. Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nature Meth.* **5**, 179–181 (2008).

15. Sogin, M. L. *et al.* Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl. Acad. Sci. USA* **103**, 12115–12120 (2006).

16. Liu, Z., DeSantis, T. Z., Andersen, G. L. & Knight, R. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucl. Acids Res.* **36**, e120 (2008).

17. Gharizadeh, B. *et al.* Identification of medically important fungi by the pyrosequencing (TM) technology. *Mycoses* **47**, 29–33 (2004).

18. Eriksson, N. *et al.* Viral population estimation using pyrosequencing. *PLOS Comp. Biol.* **4**, e1000074 (2008).

19. Altschul, S. F. *et al.* Gapped blast and psi-blast: a new generation of protein database search. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

20. Chao, A. Estimating the population-size for capture recapture data with unequal catchability. *Biometrics* **43**, 783–791 (1987).

Acknowledgements Put acknowledgements here.

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

Correspondence Correspondence and requests for materials should be addressed to Christopher Quince (email: quince@civil.gla.ac.uk).

	Sample size	Clean no.	3% OTUs	3% Chao	5% OTUs	5% Chao
Noise removed	16222	15378	721	1683	604	1252
Standard method	16222	16222	1327	2254	877	1442
RDP pipeline	16222	16222	1208	2219	862	1432

Table 1: Summary of OTU numbers for pyrosequenced environmental 16S rRNA V5/V6 sequences from Priest Pot lake. Results are shown for OTU generation following flowgram preclustering and chimera removal, for the standard method using just sequence information, and for the RDP pyrosequencing pipeline. All results are for complete linkage clustering. The Chao estimates of total diversity for the samples are also shown [20].

Name	Source	Read no.	Clean no.
GS20 data	43 pyrosequenced clones[10]	111321	99189
Priest Pot clones	Sanger sequenced environmental	n.a.	95
GSFLX - Priest Pot	Pyrosequenced environmental	28361	16222
GSFLX - divergent	23 pyrosequenced clones	57902	38351
GSFLX - community	95 pyrosequenced clones	46249	34308

Table 2: Summary of the Data sets used in this study with read number and clean read number following initial noise removal.

Figure 1 Probability distribution of flow intensities for different lengths of hompolymer (marked on figure). The solid line shows the results for the GSFLX implementation and the dashed line the GS20.

Figure 2 OTU number as a function of cut-off for a pyrosequenced 'artificial community' of 95 16S rRNA gene clones of known sequence. The black line is the true number of OTUs. The dot-dashed line shows the result of OTU assignment from the pyrosequencing data using just the sequences, the dotted line is the output from the RDP pyrosequencing pipeline. Flowgram preclustering prior to OTU generation is given by the dot-dot-dashed line, and the dashed line the same procedure but with chimera removal too (see Methods). Results are repeated for complete linkage (a) and average linkage algorithms (b).

Figure 3 Proportion of sequences correctly clustered as a function of cut-off for a pyrosequenced mixture of 95 16S rRNA gene clones of known sequence. The dot-dashed line shows the result of OTU formation from the pyrosequencing data using only sequence information, the dotted line is the output from the RDP pyrosequencing pipeline. The dashed line shows the result of flowgram preclustering and chimera removal prior to OTU formation. Results are repeated for complete linkage (a) and average linkage algorithms (b).

Figure 4 Taxa abundance distribution for 3% complete linkage OTUs derived from the Priestpot environmental sample using either flowgram preclustering and chimera removal or the standard method of no noise removal prior to OTU generation. Data points were

aggregated so that observed counts were greater than twenty. Both axes have been scaled logarithmically.