environmental microbiology

Environmental Microbiology (2010) doi:10.1111/j.1462-2920.2010.02332.x

Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU rRNA gene regions

Anke Behnke,[1†] Matthias Engel,[1†] Richard Christen,[2] Markus Nebel,[3] Rolf R. Klein[4] and Thorsten Stoeck[1*]

Departments of [1]Ecology and [3]Computer Science, University of Kaiserslautern, D-67653 Kaiserslautern, Germany.

[2]Université de Nice et CNRS UMR 654, Laboratoire de Biologie Virtuelle, Nice, F 06108, France.

[4]Seq-IT, Pfaffplatz 10, D-67655 Kaiserslautern, Germany.

Summary

Initial environmental pyrosequencing studies suggested highly complex protistan communities with phylotype richness decisively higher than previously estimated. However, recent studies on individual bacteria or artificial bacterial communities evidenced that pyrosequencing errors may skew our view of the true complexity of microbial communities. We pyrosequenced two diversity markers (hypervariable regions V4 and V9 of the small-subunit rDNA) of an intertidal protistan model community, using the Roche GS-FLX and the most recent GS-FLX Titanium sequencing systems. After pyrosequencing 24 reference sequences we obtained up to 2039 unique tags (from 3879 V4 GS-FLX Titanium reads), 77% of which were singletons. Even binning sequences that share 97% similarity still emulated a pseudodiversity exceeding the true complexity of the model community up to three times (V9 GS-FLX). Pyrosequencing error rates were higher for V4 fragments compared with the V9 domain and for the GS-FLX Titanium compared with the GS-FLX system. Furthermore, this experiment revealed that error rates are taxon-specific. As an outcome of this study we suggest a fast and efficient strategy to discriminate pyrosequencing signals from noise in order to more realistically depict the structure of protistan communities using simple tools that are implemented in standard tag data-processing pipelines.

Received 21 May, 2010; accepted 23 July, 2010. *For correspondence. E-mail stoeck@rhrk.uni-kl.de; Tel. (+49) 631 2052502; Fax (+49) 631 2052496. †Authors contributed equally to this study.

Introduction

Because microbes, distributed among three primary relatedness groups or domains (*Archaea*, *Bacteria* and *Eukarya*), are the most diverse group of organisms on Earth, a profound knowledge of their diversity is fundamental for our understanding of evolution, ecosystem functions and planetary processes (Pace, 1997). Advances in pyrosequencing, such as new generation high-throughput sequencing strategies like the interrogation of hypervariable regions of the small-subunit ribosomal RNA (SSU rRNA) gene, have opened a new window in microbial biodiversity research (Sogin *et al*., 2006; Stoeck *et al*., 2009). Due to the swift nature and relatively low expense of producing millions of environmental small-subunit ribosomal DNA (SSU rDNA) sequences using this method, scientists have, for the first time, an opportunity to gain deep insights into the diversity and complexity of microbial communities. Significant findings obtained by environmental pyrosequencing include for example that natural microbial communities are orders of magnitude more diverse than previously estimated using cultivation or culture-independent dideoxy sequencing methodologies (Sogin *et al*., 2006; Roesch *et al*., 2007; Stoeck *et al*., 2009). The vast majority of this diversity is derived from individual rare pyrosequencing tags, each of which is observed only once (singletons) in an environmental amplicon library of tens to hundreds of thousands of tags. This suggests microbial consortia that are dominated by a relatively small number of different populations, whereas most of the observed phylogenetic diversity refers to taxa that are present at extremely low abundances in most natural microbial assemblages, giving rise to the concept of the 'rare biosphere' (Pedrós-Alió, 2007).

Recently, the 'rare biosphere' concept has evolved into a central hypothesis driving research in our understanding of the mechanisms that create and maintain microbial diversity, of the true extent of microbial phylotype richness, and of the ecological roles microbes play in natural environments (Sogin *et al*., 2006; Pedrós-Alió, 2007;

Caron and Countway, 2009; Dawson and Hagen, 2009; Galand *et al*., 2009; Stoeck and Epstein, 2009).

However, few recent studies accumulated evidence of experimental errors occurring during pyrosequencing, which may give rise to a distorted picture of the true complexity of microbial communities. In an experimental study, Kunin and colleagues (2010) pyrosequenced *Escherichia coli* SSU rDNA reference templates. They found that standard quality filtering of sequence data as performed in environmental studies resulted in spurious 'phylotypes' confounding data interpretation. Likewise, analyses of pyrosequenced SSU rDNA fragments amplified from an artificial bacterial community suggested that sequencing errors result in richness estimates that are at least one order of magnitude too high (Quince *et al*., 2009). Such studies that assessed the effect of pyrosequencing errors were exclusively based on hypervariable regions of SSU rDNA from (individual) bacterial model organisms. These taxon- and region-specific data are not necessarily applicable to taxa from a different domain and other hypervariable gene regions. Therefore, we here for the first time explore the potential source of pyrosequencing errors and their effect on a model protistan community. While in previous studies only the Roche GS-FLX sequencing platform came under scrutiny, we here compared the GS-FLX system with the more recent GS-FLX Titanium flagship sequencing platform, which applies a modified chemistry resulting in longer reads (*c*. 200 bp for GS-FLX and up to 500 bp for GS-FLX Titanium). We sequenced a PCR amplicon library that was generated from 24 reference templates of known sequences that represent a subset of a microbial eukaryote community from an intertidal sediment sample. Furthermore, in this process, we compared two hypervariable regions of the SSU rRNA gene – V4 and V9 – that are used in environmental pyrosequencing (Amaral-Zettler *et al*., 2009; Stoeck *et al*., 2009; 2010) in each of the two sequencing systems. As a result, we present a fast data-processing strategy as part of a standard computational pyrosequence data-processing pipeline that accounts for experimental errors and results in a less biased picture of protistan community structures.

Results

The number of quality-checked reads obtained from each experiment was 8917 for the V4-GS-FLX data set, 3879 for V4-GS-FLX Titanium, 40 441 for V9-GS-FLX and 6365 for V9-GS-FLX Titanium. Table S1 summarizes the number of obtained reads from each individual reference template. In all experiments, the number of unique reads (i.e. clusters comprising sequence tags with 100% sequence similarity) obtained outnumbers the species diversity of the microbial model community ($n = 24$) (Fig. 1). We observed 791 unique sequence tags for the V4-GS-FLX data set (9% of total reads), 2039 for V4-GS-FLX Titanium (52%), 1116 for V9-GS-FLX (3%) and 321 for V9-GS-FLX Titanium (5%) (Table S1). Singletons represented the vast majority of these tags, accounting for 74% (V4-GS-FLX), 77% (V4-GS-FLX Titanium), 68% (V9-GS-FLX) and 65% (V9-GS-FLX Titanium) of unique sequences. This indicates that both sequencing platforms are prone to produce more singletons from the V4 region. Removal of singletons from the data sets still resulted in up to 19-fold overestimation of the number of different ribotypes: for example, instead of 24 clones, we observed 462 unique V4-GS-FLX Titanium tags. After clustering tags at decreasing sequence similarities (99%, 98%, 97%, 96%, 95%, 90%, 80%, 70%, 60%) the number of observed clusters (operational taxonomic units, OTUs) and the number of reference OTUs converged (see also Fig. S1 for OTU calling for individual clones). However, the number of observed clusters never corresponded to the number of true OTU richness. There were two exceptions to this rule: first, V9-GS-FLX Titanium OTUs called at 95% sequence similarity (n OTUs in both cases = 18, Fig. 1B) and second, OTUs called at similarities low enough to unite all tags and references in one single OTU (between 60% and 70%, Fig. 1A and B). When calling V4-OTUs at 97% and after removal of singletons, the number of observed OTUs was most similar to the number of reference OTUs while perpetuating most of the original population diversity ($n = 24$). For both V9 data sets, OTUs called at 95% sequence similarity offered the closest comparison with true OTU richness.

A surprisingly low proportion of V4 tags generated by both sequencing systems were identical to the reference templates (65% of all GS-FLX-generated tags and only 6% of all GS-FLX Titanium-generated tags; Fig. 2). Accordingly, the GS-FLX system generated more accurate tags (92%) for V9 compared with the GS-FLX Titanium system (82%); however, the proportion of reference-matching tags was distinctly higher. When allowing for 2–5% errors, all quality-checked tags from both hypervariable regions could be assigned correctly. Thus, the average maximum error rate was between 2% (V4-GS-FLX-generated tags) and 5% (both GS-FLX Titanium-generated data sets). Removal of singletons from the data sets did not have a prominent effect on the proportion of tags that matched the reference sequences (see also Fig. S1 for proportion of reference matching tags for the 24 individual clones).

Table 1 summarizes the error rates for all individual data sets with and without singletons and compares these errors between different taxon groups. The overall error rates varied between 0.09 (V9-GS-FLX) and

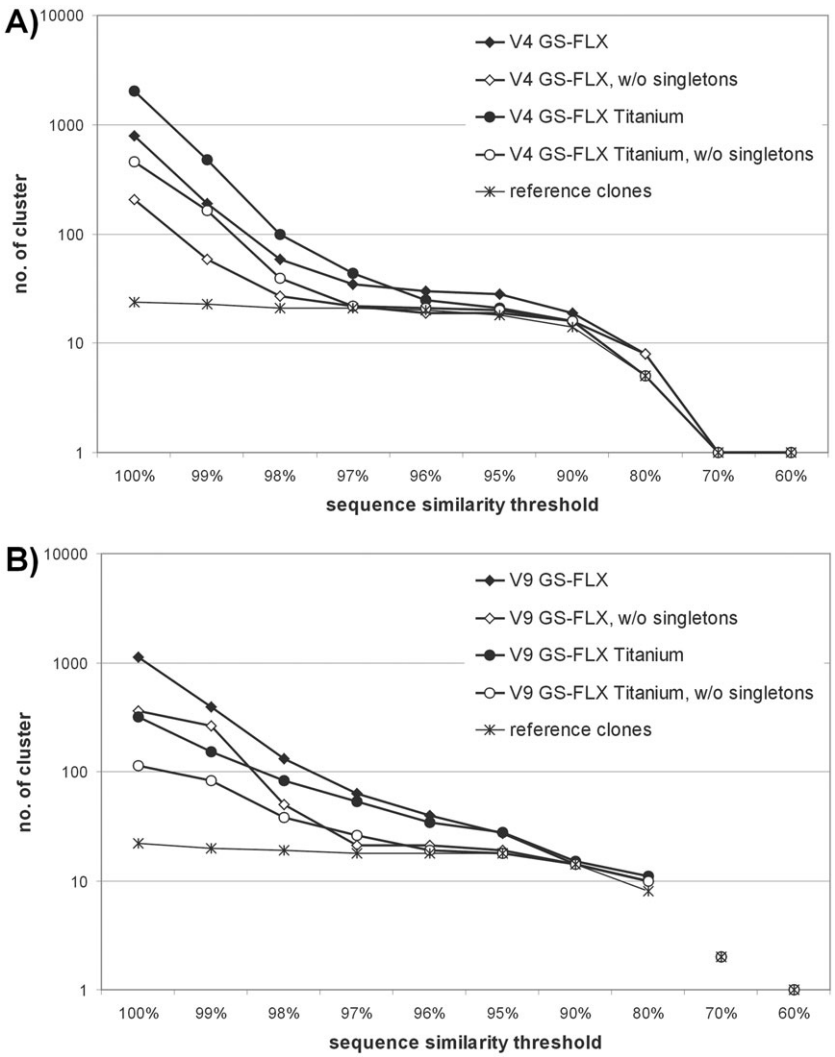

Fig. 1. Clustering profiles for V4 and V9 pyrosequencing data sets. Graphs displaying the effect of singleton filtering and OTU calling threshold on the number of OTUs detected in two PCR amplicon libraries (24 reference templates) using pyrotags from the sequencing of the SSU rDNA V4 (A) and V9 (B) regions. Symbols refer to two different sequencing systems (diamonds: Roche GS-FLX; circles: Roche GS-FLX Titanium). Solid symbols refer to data retrieved after standard filtering of tags; empty symbols refer to results after additional removal of singletons. Crosses refer to number of OTUs called for the respective region of the 24 template clones.

0.79 (V4-GS-FLX Titanium). Generally, we found that the GS-FLX Titanium system generated more errors when compared with the GS-FLX system, and that the V4 region is prone to more errors than the V9 region. The relative distribution of error rates for the different gene regions and sequencing systems was the same for tags derived from the taxon groups *Ciliophora*, *Rhizaria* and *Bacillariophyta*. However, absolute error rates differed decisively among different taxon groups: for example for *Rhizaria*, the V9-GS-FLX Titanium error rates were twice as high as for ciliates and four times higher than for bacillariophytes (see also Table S2 for individual error rates of the 24 environmental clones). For V9-GS-FLX and V4-GS-FLX Titanium-derived tags, differences in error rates for the different taxon groups were less pronounced. The removal of singletons had an enormous effect on the quality of all data sets: error rates were reduced up to 56% (*Rhizaria* V9-GS-FLX Titanium). Furthermore, after removal of singletons from the individual data sets, differences regarding error rates were less pronounced among the three taxon groups.

The main sources of pyrosequencing errors were homopolymers, accounting for a cumulative proportion of total errors between 63.8% (V4-GS-FLX) and 80.6% (V4-GS-FLX Titanium) (Table 2). In the V4 reads, homopolymers of five successive bases (5-hps) contributed the highest proportion to these errors. Interestingly, the proportion of 6-hp errors was approximately three times higher in the GS-FLX system (16.3% of all errors) compared with the GS-FLX Titanium system (5.7%). The V9 region did not contain any homopolymers exceeding four nucleotides. While in V9-GS-FLX Titanium, 4-hps were the most error-prone, this was true for 3-hps in the V9-GS-FLX. Interestingly for both gene regions, the GS-FLX system introduced a high number of erroneous insertions (31% of all errors in case of V4 and 25.5% in case of V9).

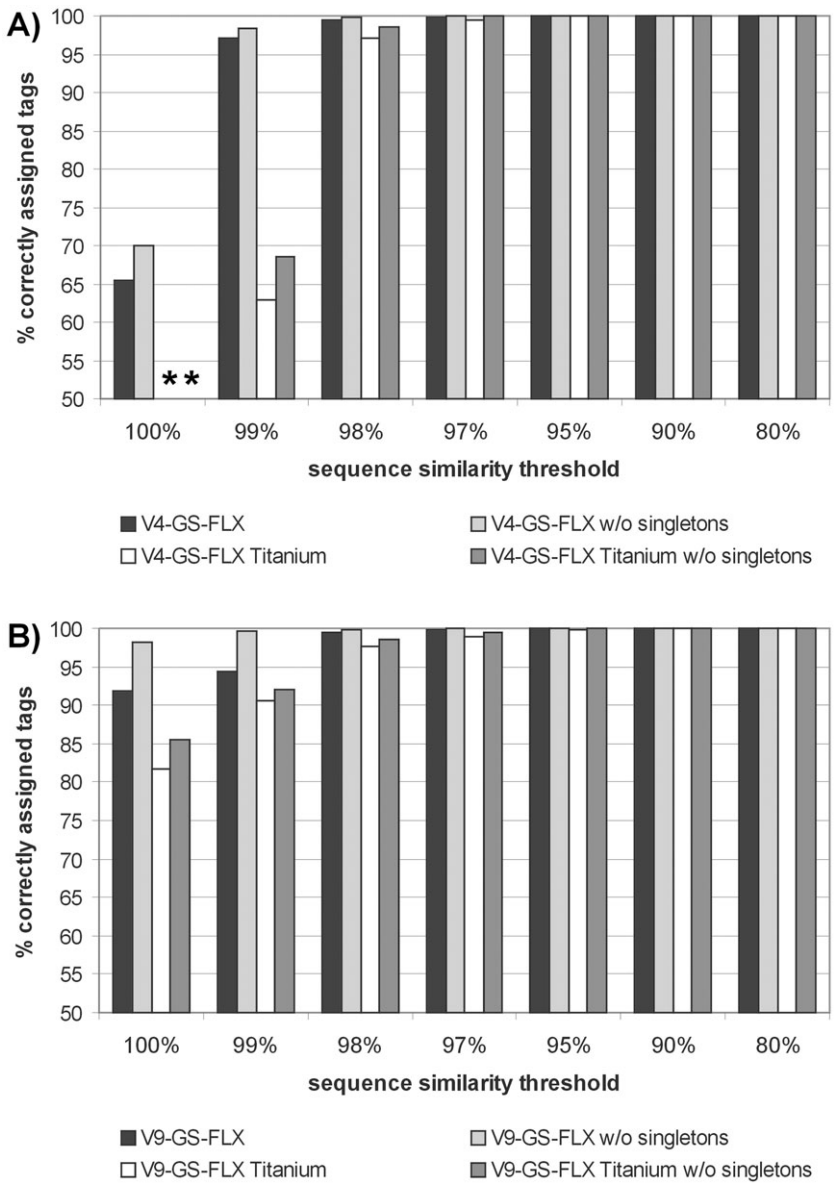

Fig. 2. Proportion of tags with correct assignment to reference clones using different sequence similarity thresholds. Bar graphs displaying the proportion of pyrosequencing tags that were correctly assigned to their respective reference clone (24 reference templates) at different sequence similarity thresholds: (A) V4 region; (B) V9 region. Levels of grey refer to the two different filtering procedures applied and the two different sequencing systems respectively (black: GS-FLX with singletons; light grey: GS-FLX without singletons; white: GS-FLX Titanium with singletons; dark grey: GS-FLX Titanium without singletons). *: V4 GS-FLX Titanium = 6%; V4 GS-FLX Titanium without singletons = 11%.

Table 1. Error rates calculated for the four different pyrosequencing data sets retrieved from two hypervariable SSU rDNA regions with two different systems (V4-GS-FLX, V4-GS-FLX Titanium, V9-GS-FLX, V9-GS-FLX Titanium).

		Error rate: average (min.–max.)			
Data set	Quality filtering procedure	Total (all clones)	Ciliates	Diatoms	Rhizaria
V4-GS-FLX	With singletons	0.15 (0.06–0.37)	0.11 (0.06–0.16)	0.16 (0.10–0.16)	0.22 (0.13–0.37)
	Without singletons	0.09 (0.04–0.15)	0.07 (0.04–0.11)	0.10 (0.06–0.13)	0.11 (0.09–0.14)
V4-GS-FLX Titanium	With singletons	0.79 (0.54–1.13)	0.74 (0.56–1.00)	0.87 (0.63–1.13)	0.77 (0.60–0.94)
	Without singletons	0.56 (0.21–0.83)	0.52 (0.29–0.80)	0.57 (0.43–0.83)	0.57 (0.38–0.80)
V9-GS-FLX	With singletons	0.09 (0.04–0.16)	0.07 (0.04–0.19)	0.08 (0.06–0.14)	0.09 (0.06–0.1)
	Without singletons	0.09 (0.04–0.18)	0.07 (0.04–0.18)	0.09 (0.05–0.14)	0.07 (0.06–0.1)
V9-GS-FLX Titanium	With singletons	0.25 (0.05–0.77)	0.22 (0.05–0.77)	0.09 (0.06–0.12)	0.41 (0.17–0.71)
	Without singletons	0.16 (0–0.57)	0.16 (0–0.57)	0.08 (0.01–0.5)	0.18 (0.01–0.42)

Rates (errors/100 nt) are given for the four combined data sets (total) as well as for three different taxon groups (ciliates, diatoms, rhizaria). Furthermore, for each data set, error rates are calculated for both of the different quality filtering procedures (with and without singletons). Numbers refer to average error rate; values in brackets refer to minimum and maximum error rate calculated for the different reference clones.

Table 2. Sources of error for the four different pyrosequencing data sets retrieved from two hypervariable SSU rDNA regions with two different systems (V4-GS-FLX, V4-GS-FLX Titanium, V9-GS-FLX, V9-GS-FLX Titanium).

	V4-GS-FLX		V4-GS-FLX Titanium		V9-GS-FLX		V9-GS-FLX Titanium	
	No.	% of total errors	No.	% of total errors	No.	% of total errors	No.	% of total errors
2-Homopolymers	237	6.3	2 133	9.3	719	18.0	401	18.9
3-Homopolymers	143	3.8	3 151	13.7	1080	27.0	563	26.5
4-Homopolymers	620	16.6	5 471	23.8	988	24.7	730	34.4
5-Homopolymers	777	20.8	6 460	28.1	–	–	–	–
6-Homopolymers	609	16.3	1 312	5.7	–	–	–	–
Insertions	1160	31.0	6	0.0	1019	25.5	301	14.2
Transitions	57	1.5	665	2.9	135	3.4	79	3.7
Transversions	141	3.8	3 803	16.5	59	1.5	48	2.3
Sum	3744		23 001		4000		2122	

For each type of error the total number of occurrences and the respective proportion are given.

Discussion

Our data suggest that the GS-FLX Titanium platform is more error-prone compared with the GS-FLX system. For the GS-FLX system, homopolymers (hps) have been identified as the major source of errors (Margulies *et al*., 2005; Huse *et al*., 2007; Kunin *et al*., 2010; Quince *et al*., 2009). We found that the GS-FLX Titanium system is even more defective when it comes to longer stretches of identical nucleotides. We assume that this is the cost for the increased read lengths and number that comes with the GS-FLX Titanium system. The principle of the sequencing chemistry (Margulies *et al*., 2005) is the same for GS-FLX and GS-FLX Titanium: hundreds of thousands of microbeads, each carrying millions of copies of a unique single-stranded DNA molecule, are sequenced in parallel by synthesis of a complementary strand. If an incoming nucleotide is complementary to the template strand, the polymerase extends the existing DNA strand by adding nucleotide(s). The addition of one (or more) nucleotide(s) results in an enzymatic reaction that generates a light signal that is recorded by a CCD camera in the instrument. The signal strength is proportional to the number of nucleotides incorporated in a single nucleotide flow. As described in more detail elsewhere (Quince *et al*., 2009), these light intensities do not faithfully reflect homopolymer lengths due to intensity saturation and the base-calling procedure applied by default results in serious hp errors in the GS-FLX system. From the GS-FLX system to the GS-FLX Titanium system, the well diameter of the picotitre plate carrying the individual beads (one bead per well) decreased from *c*. 50 μm to 35 μm allowing the accommodation of 3.6 billion wells in one GS-FLX Titanium compared with 1.8 billion wells in one GS-FLX picotitre plate. This increased the number of reads from *c*. 400 000 (GS-FLX) to *c*. 1.2 billion (GS-FLX Titanium) per plate. Therefore, the CCD imaging system has to record a higher number of signals with the same number of pixels, resulting in even earlier overall saturations of emitted light intensities and an increased impairment of the standard base-calling procedure.

Comparing the different diversity markers with each other, we noted higher error rates for V4 tags than for the V9 tags. The relative number of hps in both gene regions is similar (*c*. 0.5 per base) and equally distributed over these regions (Fig. S2). Thus, they can hardly be responsible for this observation. However, most errors in the V4 fragment occurred in hps with ≥ five successive nucleotides. Interestingly, none of the V9 fragments that were randomly chosen as reference templates possessed hps longer than four nucleotides, eliminating one major source of error for V9 fragments. Furthermore, the formation of secondary structures in the single-stranded templates to be sequenced may give rise to false sequence signals (Gharizadeh *et al*., 2004; 2006). The formation of secondary structures strongly correlates with the number of nucleotides (Andronescu *et al*., 2008). Thus, single-stranded V4 amplicons (367–391 nucleotides for the 24 environmental clones) are more likely to form secondary structures resulting in false sequencing signals than the shorter V9 regions (159–170 nucleotides).

In environmental microbial diversity studies, sequences are binned into OTUs, a measure of diversity used for microbial species of which the majority has not been taxonomically classified (Schloss and Handelsman, 2005; Hong *et al*., 2006; Jeon *et al*., 2006; Caron *et al*., 2009). Furthermore, clustering sequences accounts for experimental errors. However, pyrosequencing-induced error rates are unequal among different taxon groups (Table 1, Table S2). For example, when clustering all V4-GS-FLX derived sequence tags at 96%, the number of ciliate OTUs equals the number of ciliates in the model community ($n = 9$). In case of rhizaria, we still recovered seven OTUs (from five taxa in the model community) when clustering rhizaria tags at 91% sequence similarity (Fig. S1). Thus, relying on one

individual similarity value for different taxon groups when binning sequences into OTUs might easily result in overestimation or underestimation of whole community diversity. Of course, the present study provides only first information about distinct error rates of different protistan taxon groups, and a deeper and wider taxon set has to be pyrosequenced to test if these results are confirmed. But if patterns like this are consistently observed, refinements of OTU calling should take into account these different errors rates by first sorting tags by taxon groups and subsequently calling OTUs.

Varying error rates in different taxon groups may be a result of higher numbers of long hp stretches that are specifically prone to errors (Table 2). Indeed, only two out of nine ciliate V4 sequences from our model community exhibited a homopolymer stretch exceeding five nucleotides. In contrast, all rhizarian V4 fragments exhibited at least one such homopolymer. Furthermore, specific taxon groups, which form secondary structures with a higher number of loops and stems, helix lengths and base pairs within helices, may be an additional source of defective sequence reads (Nickrent and Sargent, 1991; Wuyts *et al*., 2000). The V4 domain is the most variable region of the eukaryotic SSU rDNA. The variability of this area is not restricted to high substitution rates, but also involves higher indel rates, which may result in the deletion or insertion of entire helices, influencing the complexity of the domain (Wuyts *et al*., 2000). Thus, the V4 varies in length from *c*. 230 bases (most common in non-protistan taxa) to *c*. 500 bases (in most protists) (Nickrent and Sargent, 1991). Some flagellates, however, have an extremely reduced V4 region or lack this domain entirely (Sogin *et al*., 1989). Unfortunately, there are hardly any data available to compare ciliate V4 secondary structures with the ones from rhizaria in order to explain higher pyrosequencing error rates in the rhizarian V4 region. Nevertheless, the high complexity of the V4 region with the presence of additional hairpins and branched structures in a number of different taxa most certainly gives rise to taxon-specific pyrosequencing error rates that require specific attention in sequence data correction and clustering.

The most striking finding of our study, however, is the large proportion of unique tags that were observed only once in the sequence data sets (singletons), regardless of the targeted domain or sequencing system used. As a rule, these singletons were experimental artefacts. Not only do they inflate diversity richness estimates – because extrapolations of the total species number from observed data rely heavily on the number of species observed only once (Hong *et al*., 2006) – but they also artificially inflate a 'rare biosphere'. Even binning sequence tags into OTUs at 97% sequence similarity did not eliminate artefactual richness, which remained 3.5 times that of the number of actual templates (Fig. 1). Therefore, we conclude that previous surveys of protistan diversity may have significantly overestimated the protistan richness in environmental samples. Reanalysing environmental pyrosequencing data sets targeting eukaryotic V4 and V9 domains in a marine plankton sample revealed that the removal of singletons resulted in protistan communities that were up to 6.5 times (unique OTUS) and up to three times (calling OTUs at 97% sequence similarity) less complex than previously suggested (Stoeck *et al*., 2010) (Fig. S3). We would like to stress that sorting and clustering of tags according to taxon groups very likely would have furthermore improved diversity analyses of the samples; however, as we have only (preliminary) information about error rates from three different taxa (ciliates, diatoms, and rhizarians) we decided to choose the same sequence similarity cut-off for all tags.

The strategy to account for experimental artefacts as implemented in this study (binning sequences into clusters based on primary structure similarities and removal of singletons that remain after clustering) provides an efficient and fast approach to handle pyrosequencing data for diversity analyses. In addition, there are complementary techniques available that may help to furthermore improve data quality as for example PyroNoise (Quince *et al*., 2009) that reduces hp-derived errors. However, our data show that PyroNoise cannot serve as a stand-alone alternative, as errors other than hp-derived errors may comprise up to 36.3% of all pyrosequencing errors (Table 2). Furthermore, the current 'ready-to-run' version of PyroNoise (http://people.civil.gla.ac.uk/~quince/Software/PyroNoise.html) allows only for the processing of up to 10 000 tags [200-bp-sized tags in 24 h on 128 processors (Quince *et al*., 2009); computational complex and time-exhausting workarounds for larger data sets are available], while the strategy described here allows processing of millions of tags on individual desktop computers in a reasonable amount of time (e.g. 340 000 tags of 400 bp in 4 h and 40 min on one Intel core i7 processor with 2.67 GHz and 6 GB ram). PyroNoise is unsuitable for implementations in automated data-processing pipelines, while algorithms for the clustering/singleton removals strategy can be easily implemented in any standard pyrosequencing data-processing pipeline. Finally, PyroNoise is optimized only for the GS-FLX platform, while the clustering/singleton removal process is independent of the sequencing platform used (e.g. GS-FLX, GS-FLX Titanium or any future system). In sum, our strategy provides users with a much needed advance that can handle more types of potential errors, on millions of sequence tags, in a computationally efficient manner, with data derived from different sequencing platforms.

Applying the clustering/singleton removal strategy to environmental data sets is likely to result in a lower-bound estimate of microbial community complexity, as inevitably not all singletons are experimental artefacts. However, we note that this underestimation (loss of 'true' phylotypes by removing singletons) is insignificant as our experiments suggest that only 0.25% (max) of all observed singletons reflect truly existing template sequences.

Experimental procedures

DNA isolation and PCR amplification of eukaryotic SSU rRNAs

DNA was isolated from an environmental sediment sample (Sylt, Germany) with the Power Soil DNA Kit (MO BIO Laboratories, Carlsbad, CA). The integrity of total DNA was checked by agarose gel electrophoresis (1%), and the DNA yield was quantified photometrically with the Nanodrop ND-1000 UV-VIS Spectrophometer of Nanodrop Technologies (Wilmington, DE, USA). Nearly full-length SSU rRNA genes were PCR amplified by using a eukaryotic-specific primer set [EUK360FE: 5′-CGGAGA(AG)GG(AC)GC(AC)TGAGA-3′; EukB: 5′-TGATCCTTCTGCAGGTTCACCTAC-3′ (Medlin *et al*., 1988)] as well as a ciliate-specific primer set [CilF: 5′-TGGTAGTGTATTGGAC(AT)ACCA-3′ (Lara *et al*., 2007), and EukB]. Each PCR mixture contained 10–20 ng of template DNA, 5 U of HotStar *Taq* DNA polymerase (Qiagen, Valencia, CA), 1× CoralLoad PCR Buffer (containing 1.5 mM $MgCl_2$), 200 μM concentrations of each deoxynucleotide triphosphate, and 0.5 μM concentrations of each oligonucleotide primer. The final volume was adjusted to 50 μl with sterile water. The PCR protocol for SSU rRNA gene amplification consisted of an initial hot start incubation (5 min at 95°C) followed by 35 identical amplification cycles (denaturation at 95°C for 45 s, annealing at 56°C for 45 s and extension at 72°C for 2 min) and a final extension at 72°C for 5 min. PCR products were checked by agarose gel electrophoresis (1%).

Cloning, amplified ribosomal DNA restriction fragment analysis (ARDRA) and sequencing

PCR products were used to construct a clone library by using the TA cloning kit (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions. Positive clones were identified by blue-white screening, and a colony PCR of overnight cultures was performed to check for the presence of the target insert (primer set M13F and M13R, for PCR mixture see above). The protocol for the colony PCR consisted of an initial denaturation (2 min at 95°C) followed by 30 identical amplification cycles (denaturation at 95°C for 45 s, annealing at 53°C for 45 s and extension at 72°C for 90 s) and a final extension at 72°C for 7 min. Between 200 and 400 ng of positive amplification products of the target size were digested with 7.5 U of the restriction endonuclease HaeIII (New England Biolabs, Beverly, MA) for 90 min at 37°C, followed by an inactivation step for 20 min at 80°C. The resulting bands were separated by agarose gel electrophoresis (1%). Clones were identified with identical amplified ribosomal DNA restriction fragment analysis (ARDRA). One clone of each unique ARDRA pattern was fully sequenced at Seq-It laboratories (Seq-It GmbH, Kaiserslautern, Germany) using an Applied Biosystems 3730 DNA Stretch sequence with the XL upgrade and an Applied Biosystems Prism BigDye Terminator version 3.1 cycle sequencing Ready Reaction kit. After MegaBLAST analysis (Altschul, 1999) of the environmental sequences against the NCBI nr-database we selected 24 clones from different protistan taxon groups (10 alveolates, nine stramenopiles, five rhizaria) as reference templates for the construction of amplicon libraries.

Amplicon library construction and pyrosequencing

The 24 Sanger-sequenced protistan clones were used as reference templates for pyrosequencing the variable V4 and V9 region of the eukaryotic SSU rDNA using the Roche systems GS-FLX and GS-FLX Titanium. As described previously (Stoeck *et al*., 2010), for the V4 region, the primer pair TAReuk454FWD1 (5′-CCAGCA(GC)C(CT)GCGGTAATTCC-3′) and TAReukREV3 (5′-ACTTTCGTTCTTGAT(CT)(AG)A-3′) were used and the V9 region was amplified with 454V9F (5′-GTACACACCGCCCGTC-3′) and 454V9R (5′-TGATCC TTCTGCAGGTTCACCTAC-3′). For GS-FLX-sequencing, adapters A (5′-GCCTCCCTCGCGCCATCAG-3′) and B (5′-GCCTTGCCAGCCCGCTCAG-3′) were linked to the 5′ end of the forward and reverse primers respectively. For GS-FLX Titanium sequencing, we added SG linkers to the 5′ ends of the forward and the reverse primer (5′-CCATCTCAT CCCTGCGTGTCTCCGACTCAG-3′ and 5′- CCTATCCCC TGTGTGCCTTGGCAGTCTCAG-3′ respectively). For each protistan clone, pyrosequencing system and gene region we performed separate PCR reactions. PCR mixtures contained 2 U of Phusion Hot Start high-fidelity tag polymerase (New England Biolabs GmbH, Frankfurt/Main, Germany), 1× Phusion high-fidelity buffer, 200 μM dNTPs, 0.5 μM of each primer and 3–10 ng of template DNA in a volume of 50 μl. The amplification started with an initial activation step at 98°C for 1 min, followed by 30 cycles consisting of 98°C for 15 s, 60°C for 30 s and 72°C for 15 s; and a final 5 min extension at 72°C. PCR products were checked on a 1% low-melting-point agarose gel and cleaned up by using the MinElute PCR purification kit (Qiagen, Hilden, Germany). All tags (GS-FLX and GS-FLX Titanium, both regions) were sequenced from the forward primer (5′ end).

Sequence data processing

Sequencing resulted in four different data sets (V4-GS-FLX, V4-GS-FLX Titanium, V9-GS-FLX, V9-GS-FLX Titanium) that were subsequently subjected to two quality filters. The first filter, corresponding to the standard filter for environmental pyrosequencing data sets, selected for tags that displayed perfect matches to forward and reverse primers and contained no ambiguous nucleotide. In the case of V4-GS-FLX reads, none of the tags reached the reverse primer. Therefore, these tags were considered high quality when they displayed a perfect match to the forward primer, no ambiguous nucleotide and a minimum length of 200 nucleotides. Furthermore, GS-FLX V4 tags were 3′ end-trimmed to the same sequence

position (nt position 805 of *Saccharomyces cerevisiae*, Accession No. J01353) based on error entropy plots in order to remove the most error-prone part of GS-FLX reads. The second quality filter additionally removed all singletons (unique sequences occurring only once) from the data sets. Subsequently, each data set was subjected to cluster analyses by means of the UPGMA algorithm to determine total numbers of clusters at different sequence similarity thresholds. Sequence similarities were calculated from pairwise alignments (making use of the IUB scoring scheme), as it has been demonstrated that distances derived from pairwise alignments provide more accurate estimates of microbial diversity (Sun *et al.*, 2009). Clusters were constructed using the average-linkage algorithm, as it is more robust to noise produced by pyrosequencing compared with the complete-linkage algorithm (Quince *et al.*, 2009). Both the pairwise alignments and the clustering are part of the functionality of the software package JAguc (Nebel, 2010; http://wwwagak.informatik.uni-kl.de/research/JAguc/).

To investigate if different taxonomic groups display differences regarding the number of OTUs detected, we assigned all tags to their respective reference clone by means of the JAguc software. We used the implemented BLAST against a reference database comprising the 24 reference clones using the following BLAST parameters: -m 7 -r 5 -q -4 -G 8 -E 6. The resulting 24 data sets were checked manually for correct assignment of 454 tags. Subsequently, these data sets were subjected to cluster analyses at different sequence similarity thresholds, according to the analyses of combined data sets, with and without taking singletons into account. Finally, tags were compared with their respective reference sequence to determine error rates and sources of errors, using a Needleman–Wunsch algorithm. Detected errors were categorized and manually sorted into homopolymer errors (2-, 3-, 4-, 5- and 6-hps), insertions, transversions and transitions. This procedure was performed separately for all four original data sets, V4-GS-FLX, V4-GS-FLX Titanium, V9-GS-FLX and V9-GS-FLX Titanium.

The original data sets generated in this study have been deposited at GenBank's Short Read Archive (SAR) under Accession No. SRA022989.1.

Acknowledgements

The authors would like to thank S. Epstein (NEU, Boston) and M. Dunthorn (University of Kaiserslautern) for helpful discussions and comments on the manuscript, H.-W. Breiner (University of Kaiserslautern) and C. Buschbaum (AWI, Sylt) for help with sampling. This study was financed by the Deutsche Forschungsgemeinschaft (DFG) with Grant STO414/3-1 to T.S. and funds from the University of Kaiserslautern. This article is a publication that partly emerged from the Biomarks consortium. We thank the consortium for inspiring this study and for helpful discussions on the subject.

References

Altschul, S. (1999) Hot papers – bioinformatics – gapped BLAST and PSI-BLAST: a new generation of protein database search programs by S.F. Altschul, T.L. Madden, A.A. Schaffer, J.H. Zhang, Z. Zhang, W. Miller, D.J. Lipman – comments. *Scientist* **13:** 15.

Amaral-Zettler, L.A., McCliment, E.A., Ducklow, H.W., and Huse, S.M. (2009) A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS ONE* **4:** e6372.

Andronescu, M., Bereg, V., Hoos, H.H., and Condon, A. (2008) RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC Bioinformatics* **9:** 340.

Caron, D.A., and Countway, P.D. (2009) Hypotheses on the role of the protistan rare biosphere in a changing world. *Aquat Microb Ecol* **57:** 227–238.

Caron, D.A., Countway, P.D., Savai, P., Gast, R.J., Schnetzer, A., Moorthi, S.D., *et al.* (2009) Defining DNA-based operational taxonomic units for microbial-eukaryote ecology. *Appl Environ Microbiol* **75:** 5797–5808.

Dawson, S.C., and Hagen, K.D. (2009) Mapping the protistan 'rare biosphere'. *J Biol* **8:** 105.

Galand, P.E., Casamayor, E.O., Kirchman, D.L., and Lovejoy, C. (2009) Ecology of the rare microbial biosphere of the Arctic Ocean. *Proc Natl Acad Sci USA* **106:** 22427–22432.

Gharizadeh, B., Eriksson, J., Nourizad, N., Nordstrom, T., and Nyren, P. (2004) Improvements in Pyrosequencing technology by employing Sequenase polymerase. *Anal Biochem* **330:** 272–280.

Gharizadeh, B., Akhras, M., Nourizad, N., Ghaderi, M., Yasuda, K., Nyren, P., and Pourmand, N. (2006) Methodological improvements of pyrosequencing technology. *J Biotechnol* **124:** 504–511.

Hong, S.H., Bunge, J., Jeon, S.O., and Epstein, S.S. (2006) Predicting microbial species richness. *Proc Natl Acad Sci USA* **103:** 117–122.

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

Jeon, S.O., Bunge, J., Stoeck, T., Barger, K., Hong, S.-H., and Epstein, S. (2006) Synthetic statistical approach reveals a high degree of richness of microbial eukaryotes in an anoxic water column. *Appl Environ Microbiol* **72:** 6578–6583.

Kunin, V., Engelbrektson, A., Ochman, H., and Hugenholtz, P. (2010) Wrinkles in the rare biosphere: pyrosequencing errors lead to artificial inflation of diversity estimates. *Environ Microbiol* **12:** 118–123.

Lara, E., Berney, C., Harms, H., and Chatzinotas, A. (2007) Cultivation-independent analysis reveals a shift in ciliate 18S rRNA gene diversity in a polycyclic aromatic hydrocarbon-polluted soil. *FEMS Microbiol Ecol* **62:** 365–373.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376–380.

Medlin, L., Elwood, H.J., Stickel, S., and Sogin, M.L. (1988) The characterization of enzymatically amplified eukaryotic 16S-like rRNA-coding regions. *Gene* **71:** 491–499.

Nebel, M. (2010) *JAguc – A Software Package for Processing Pyrosequencing Data*. Kaiserslautern, Germany: AG Algorithms and Complexity, University of Kaiserslautern.

Nickrent, D.L., and Sargent, M.L. (1991) An overview of the secondary structure of the V4 region of eukaryotic small-subunit ribosomal RNA. *Nucleic Acids Res* **19:** 227–235.

Pace, N.R. (1997) A molecular view of microbial diversity and the biosphere. *Science* **276:** 734–740.

Pedrós-Alió, C. (2007) Ecology. dipping into the rare biosphere. *Science* **315:** 192–193.

Quince, C., Lanzen, A., Curtis, T.P., Davenport, R.J., Hall, N., Head, I.M., *et al*. (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* **6:** 639–641.

Roesch, L.F., Fulthorpe, R.R., Riva, A., Casella, G., Hadwin, A.K., Kent, A.D., *et al*. (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J* **1:** 283–290.

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

Sogin, M.L., Gunderson, J.H., Elwood, H.J., Alonso, R.A., and Peattie, D.A. (1989) Phylogenetic meaning of the kingdom concept: an unusual ribosomal RNA from *Giardia lamblia*. *Science* **243:** 75–77.

Sogin, M.L., Morrison, H.G., Huber, J.A., Welch, D.M., Huse, S.M., Neal, P.R., *et al*. (2006) Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc Natl Acad Sci USA* **103:** 12115–12120.

Stoeck, T., and Epstein, S. (2009) Protists and the rare biosphere. Crystal Ball. *Environ Microbiol Rep* **1:** 20–22.

Stoeck, T., Behnke, A., Christen, R., Amaral-Zettler, L., Rodriguez-Mora, M.J., Chistoserdov, A., *et al*. (2009) Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. *BMC Biol* **7:** 72.

Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M.D., Breiner, H.W., and Richards, T.A. (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol Ecol* **19:** 21–31.

Sun, Y., Cai, Y., Liu, L., Yu, F., Farrell, M.L., McKendree, W., and Farmerie, W. (2009) ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Res* **37:** e76.

Wuyts, J., De Rijk, P., Van de Peer, Y., Pison, G., Rousseeuw, P., and De Wachter, R. (2000) Comparative analysis of more than 3000 sequences reveals the existence of two pseudoknots in area V4 of eukaryotic small subunit ribosomal RNA. *Nucleic Acids Res* **28:** 4698–4708.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. Clustering profiles of V4 and V9 pyrosequencing data sets with regard to the 24 different reference templates and proportion of tags identical to respective reference clone. Curve charts display the effect of quality filtering and OTU calling threshold on the number of OTUs detected in four different pyrosequencing data sets (A: V4-GS-FLX; B: V4-GS-FLX Titanium; C: V9-GS-FLX; D: V9-GS-FLX Titanium) with regard to the 24 different environmental clones. Colours refer to taxonomic groups (blue: ciliates; purple: dinoflagellates; green: bacillariophytes; mauve: *Pseudopirsonia*; pink: rhizaria), symbols are given consecutively per taxonomic group. Bar graphs display the proportion of pyrosequencing tags that were identical to their respective reference clone (colour coding according to curve charts).

Fig. S2. Distribution of homopolymers in the two hypervariable SSU rDNA regions V4 and V9. Graph displaying the number and distribution of homopolymers with regard to sequence length for the two hpyervariable SSU rDNA regions (A: V4; B: V9). Colours refer to taxonomic groups (blue: ciliates; purple: dinoflagellates; green: bacillariophytes; mauve: *Pseudopirsonia*; pink: rhizaria).

Fig. S3. Clustering profiles of V4 and V9 pyrosequencing data retrieved from an environmental sample (Framvaren Fjord, Norway). Graphs displaying the effect of singleton removal and OTU calling threshold on the number of OTUs detected in two different pyrosequencing data sets (A: V4; B: V9) retrieved from an environmental sample (Framvaren Fjord, Norway). To test the effect of our filtering procedures, we used data retrieved from a marine sample (Stoeck *et al*., 2010), hailing from the oxic/anoxic interface of the Norwegian Framvaren Fjord sampled in May 2008. This data set comprised 271 197 V4-GS-FLX and 330 873 V9-GS-FLX tags. For reanalysis, tags were subjected to the following filtering: V9 tags had to (i) display correct matches to forward and reverse primers and (ii) contain no ambiguous nucleotide. V4 tags were considered high quality when they displayed (i) correct match to the forward primer, (ii) no ambiguous nucleotide and (iii) a minimum length of 200 nucleotides. Furthermore, these tags were end-trimmed to their average length (239 nucleotides). Remaining tags were clustered at different OTU calling thresholds (100%, 99%, 98% and 97%), based on distances derived from pairwise alignments and using the average-linkage algorithm. Subsequently, all clusters represented by a single sequence only (singletons) were removed. Crosses refer to results obtained after original analysis and diamonds refer to results of reanalysis in this study.

Table S1. Number of obtained sequence reads for the two hypervariable SSU rDNA regions (V4 and V9) pyrosequenced with two different systems (GS-FLX and GS-FLX Titanium). Numbers are given for each of the individual reference clones as well as for the combined data sets (total). For the combined four data sets the number and percentage of unique reads and singletons is provided. *: amplification for pyrosequencing was not successful; §: clones could not be distinguished based on the sequence information of the V9 region.

Table S2. Error rates calculated for the four different pyrosequencing data sets retrieved from two hypervariable SSU rDNA regions (V4 and V9) with two different pyrosequencing systems (V4-GS-FLX, V4-GS-FLX Titanium, V9-GS-FLX, V9-GSFLX Titanium). Rates (errors/100 nt) are given separately for each of the 24 pyrosequenced environmental clones (C1, 2, 3 4, 5, 6, 7, 8, 10, 11, 12, 14, 15, 17, E3, 6, 11, 12, 13,

14, 15, 16, 18 and 20), derived from different taxonomic groups (ciliates, diatoms, rhizaria, dinoflagellates and *Pseudopirsonia*). Error rates refer to data sets including singletons.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.