The ISME Journal (2009) 3, 1314–1317

www.nature.com/ismej

SHORT COMMUNICATION

Systematic artifacts in metagenomes from complex microbial communities

Vicente Gomez-Alvarez[1,3], Tracy K Teal[1,3] and Thomas M Schmidt[1,2]
[1]*Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI, USA and* [2]*Kellogg Biological Station, Michigan State University, East Lansing, MI, USA*

Metagenomics is providing an unprecedented view of the taxonomic diversity, metabolic potential and ecological role of microbial communities in biomes as diverse as the mammalian gastrointestinal tract, the marine water column and soils. However, we have found a systematic error in metagenomes generated by 454-based pyrosequencing that leads to an overestimation of gene and taxon abundance; between 11% and 35% of sequences in a typical metagenome are artificial replicates. Here we document the error in several published and original datasets and offer a web-based solution (http://microbiomes.msu.edu/replicates) for identifying and removing these artifacts.
The ISME Journal (2009) **3,** 1314–1317; doi:10.1038/ismej.2009.72; published online 9 July 2009
Subject Category: integrated genomics and post-genomics approaches in microbial ecology
Keywords: metagenomics; microbial communities; molecular census; pyrosequencing

Metagenomics offers the potential of a relatively unbiased view into the genetic composition of complex microbial communities, a perspective that has been elusive, but is crucial because of the fundamental impact that these communities have on human and animal health and global biogeochemical cycles (Turnbaugh *et al.*, 2006; Dinsdale *et al.*, 2008; Frias-Lopez *et al.*, 2008). Although metagenomes are providing some provocative insights into the metabolic capacity of microbial communities, as with any new technology, it is imperative that we understand both its strengths and its constraints. We have identified an artifact intrinsic to the 454 pyrosequencing technique that routinely leads to the artificial amplification of more than 15% of the original DNA sequencing templates.

In every metagenomic dataset we examined, there were multiple clusters of reads that began at exactly the same position. Clusters consisted of artificial replicates, both identical reads (duplicates) and reads that began at the same position but varied in length or contained a sequencing discrepency (Figure 1a). Metagenomes contained clusters that comprise as many as 4000 artificially replicated reads, but more frequently 100 reads or fewer (Figure 1b). Given the extensive diversity of these communities, it is unreasonable to expect that these replicated sequences were derived independently from the community DNA (see statistical section below). In addition, technical replicates, where DNA was divided before emulsion PCR, yielded similar distributions of clusters (Figure 1b), but the identity of the replicated sequences differed. This shows that the sequences forming the clusters arise randomly; there was no indication that any particular class of genes was preferentially amplified. The best BLAST match and COG affiliation for four of the most abundant clusters in replicate soil metagenomes are shown in Figure 1c. In addition, clustering sequences in the combined set of all the Kellogg Biological Station Long-Term Ecological Research (KBS LTER) metagenomes identified only 31 of 597 338 clusters (0.005%) that contained sequences from more than one site. The distribution and identity of clusters reveals that the artificial amplification of sequences is unbiased and spread across microbial genomes and metabolic pathways.

It has been suggested that multiple reads from a single template occur when amplified DNA attaches to empty beads during emulsion PCR, or when the optical signal during sequencing 'bleeds' into the space of an adjacent empty well (Briggs *et al.*, 2007). For the soil metagenomic datasets, sequences in a given cluster were not in adjacent wells on the sequencing plate, and so we expect that the replicates are generated during emulsion PCR. Regardless of the mechanism, the artifact of artificially replicated sequences is not limited to a specific sequencing center or the GS20 or GS-FLX systems (Table 1).

Correspondence: TM Schmidt, Department of Microbiology and Molecular Genetics, Michigan State University, 6180 Biomedical and Physical Sciences Building, East Lansing, MI 48824-1101, USA.
E-mail: tschmidt@msu.edu
[3]These authors contributed equally to this work.
Received 13 February 2009; revised 28 May 2009; accepted 2 June 2009; published online 9 July 2009

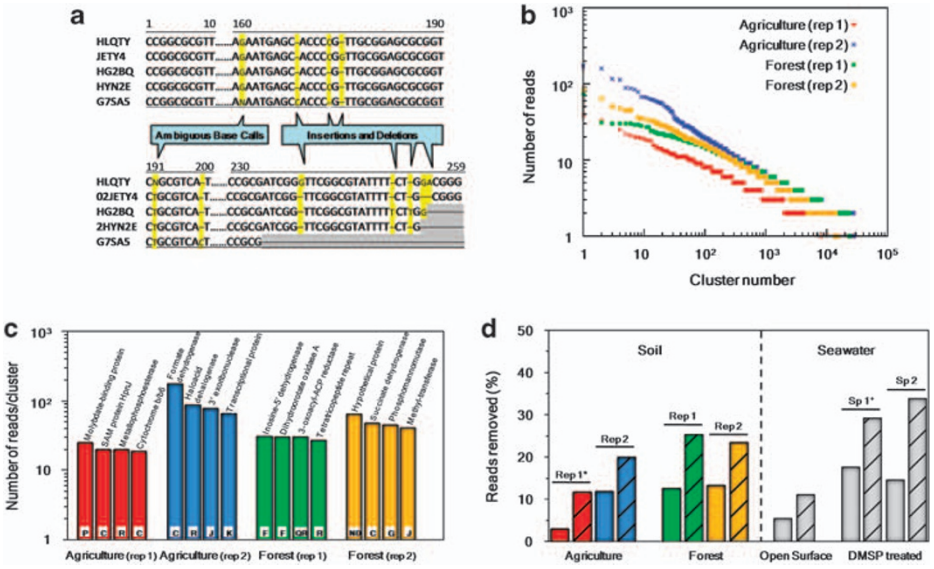

Figure 1 (**a**) Alignment of five sequences in a cluster demonstrates the types of sequencing errors and length variation (highlighted in gray) included in a cluster. (**b**) Number of reads in a cluster versus the cluster number, ordered from the largest to smallest sized cluster; both axes are plotted on a $\log_{10}$ scale. (**c**) The best BLAST match and COG affiliation for four of the most abundant clusters in replicate soil metagenomes. (**d**) Distribution of exact duplicate and all replicate reads in a metagenomic dataset from soil (this study) and seawater metagenomes (Frias-Lopez *et al.*, 2008; Mou *et al.*, 2008). *Rep, technical replicates; $^{+}$Sp, biological replicates. The number of reads in each category is presented in Table 1.

Table 1 Total numbers of reads, exact duplicates and all replicate sequences, including duplicates, from representative metagenomic data sets.

Habitat (metagenome)	*Number of reads*			*Reference (accession no.)*	*Sequencing*
	Total	*Exact duplicates*	*All replicates*		
Agriculture (rep 1)[a]	107 054	3068	12 345	This study	GS-FLX
Agriculture (rep 2)[a]	252 059	29 546	50 085	This study	GS-FLX
Forest (rep 1)[a]	170 708	21 168	43 020	This study	GS-FLX
Forest (rep 2)[a]	118 952	15 583	27 706	This study	GS-FLX
Microbialites	257 573[b]	NA	63 025	Dinsdale *et al.*, 2008 (SEED: 4440061.3)	GS20
Seawater[c] (DMSP treated 1)	66 534[b]	11 686	19 391	Mou *et al.*, 2008 (SEED: 4440364.3)	GS20
Seawater[c] (DMSP treated 2)	58 876[b]	8563	19 913	Mou *et al.*, 2008 (SEED: 4440360.3)	GS20
Seawater (Vanillate treated)	16 444[b]	3998	4201	Mou *et al.*, 2008 (SEED: 4440365.3)	GS20
Seawater (open surface)	414 323	22 258	45 635	Frias-Lopez *et al.*, 2008 (GenBank: SRA000262)	GS20
Mouse	35 053[b]	NA	2637	Turnbaugh *et al.*, 2006 (SEED: 4440325.3)	GS20

Abbreviations: DMSP, dimethylsulphoniopropionate; NA, not available.
[a]Technical replicates.
[b]Exact duplicates were removed prior to data release.
[c]Biological replicates.

Although duplicate sequences are occasionally removed from metagenomes (Dinsdale *et al.*, 2008; Mou *et al.*, 2008; Pernthaler *et al.*, 2008), this addresses only part of the problem (3–18%). Between 11% and 35% of 454 metagenomic libraries from complex microbial communities are composed of artificially replicated sequences that need to be removed before further analysis (Figure 1d and Table 1). Failure to remove replicated sequences can lead to incorrect conclusions. For example, when DNA sequences from the soil metagenomes were analyzed using the MG-RAST pipeline (Meyer *et al.*, 2008), the fraction of genes identified as being involved in denitrification did not differ significantly between agricultural and forested sites before and after removing duplicate reads (*t*-test for independent samples, $P>0.05$). However, following removal of all replicate reads in a cluster, a statistically significant difference ($P<0.001$) was detected.

We have developed tools and created a web interface (http://microbiomes.msu.edu/replicates) to identify replicated reads in metagenomic datasets. The analysis incorporates the program CD-HIT (Li and Godzik, 2006), which offers the capacity for rapid clustering of similar sequences, and a filter to determine if reads are starting at the same position to avoid grouping legitimate, partially overlapping

reads. The pipeline returns a file of unique reads along with information about the reads that were clustered. The website also provides the option to customize thresholds and identify and remove only exactly duplicated sequences so that users can determine the extent to which removing the larger problem of replicated sequences affects analysis of a metagenomic dataset. Although CD-HIT has been used to cluster orthologous groups for the purpose of comparing functional groups across communities, the issue of artificial replicates has not been discussed in these studies (Briggs *et al.*, 2007; Li *et al.*, 2008).

Pyrosequencing data being generated to assess gene expression through metatranscriptomics (Urich *et al.*, 2008) or the taxonomic composition of microbial communities through 16S ribosomal RNA gene 'tags' (Sogin *et al.*, 2006) will also be affected by this replication artifact. As reads initiating at the same position are expected in both of these approaches, identifying artificially replicated sequences in these applications will require comparison of technically replicated libraries. For instance, the relative abundance of any sequence (or sequence bin) that changes by more than one standard deviation from the average level of variation identified for the comparison of technical replicates could be flagged as a potential artifact and treated accordingly.

Application of massively parallel sequencing will continue to define the genetic landscape of complex microbial communities without the bias introduced by cultivation, but we need to remain attentive to the inherent biases in data collection and analysis that any new sequencing technology brings.

Materials and methods

Soil metagenomes

Soil samples were collected in December 2006 at the KBS LTER site (http://lter.kbs.msu.edu). Five soil cores (10 cm deep $\times$ 5 cm wide) were collected from each plot and pooled; duplicate plots were sampled from row crop agriculture and deciduous forests sites. Samples were transported to the laboratory on ice, then sieved and frozen at $-80\,^{\circ}C$. DNA was extracted using a direct soil extraction method (Zhou *et al.*, 1996) with a subsequent cesium-chloride gradient purification (Sambrook and Russell, 2001). Pyrosequencing was performed using the 454 Life Sciences GS-FLX system by the Research Technology Support Facility at Michigan State University. Two aliquots from each DNA sample were handled separately during the entire sequencing process: Rep1 and Rep2 (Figures 1b–d). The average read length for all machine runs was 220 bp.

Bioinformatics

To identify clusters of artificially replicated sequences, a program incorporating the open source program CD-HIT was developed (Li and Godzik, 2006). CD-HIT uses a short-word filter such that pairwise alignments are not required, and it has been used frequently as a clustering algorithm. It provides a significant increase in speed over an approach such as an all-by-all blast, and clustering sequences that have only a short local alignment can be avoided. Sequences clustered by CD-HIT are then analyzed to determine if the initial base pairs of each sequence in the cluster are identical. Only sequences grouped by CD-HIT that have the identical initial bases remain in the cluster; other sequences are grouped according to their start position. The representative sequence for each cluster is the longest sequence in that cluster. These representative sequences comprise the set of unique sequences from the dataset. A FASTA file of all the unique sequences, a FASTA file with the sequences in each cluster and a summary file are returned. To facilitate the dereplication of metagenomic libraries, we created a web interface to accept FASTA files of sequences and return this information to the user (http://microbiomes.msu.edu/replicates/). Through this interface, the cutoff value, length restrictions and initial base pair match requirement can be modified to customize the analysis for each dataset, as communities with different levels of complexity may require different cutoff values. Additionally a user can modify the requirement for the number of initial base pairs required to match. As sequencing errors can occur in the initial base pairs, requiring this match produces a conservative estimate. For the analyses presented here, a cutoff of 0.9, no length difference requirement and an initial base pair match of 3 base pairs was used. When clusters identified with this method were evaluated for KBS LTER soil samples, all artificially replicated sequences were accurately categorized into clusters with sequences that started at the same position.

Statistical analysis

The probability of multiple reads occurring at the same position at random, given independent sampling with replacement, is determined by focusing on that probability for the most abundant member of the community. This probability is $R \times (p \times q)^{n-1}$, where R is the number of reads, q is the percent of the community that is the most abundant member, p is $1/L$, where L is the length of the genome of the most abundant member, and n is the number of reads that start at the same position. So, for a community where the most abundant OTU is 5% of the population and is taken to have a genome of 3 Mb, if there are 400 000 reads, the probability of three reads starting at the same position is $400\,000 \times (0.05 \times 1/3 \times 10^6)^2$ or 1.1×10^{-10}. Given the extremely low probability of replicate reads in our soil samples and in other complex communities, it is very unlikely that replicate reads are occurring biologically at the frequencies documented here.

Acknowledgements

This work was funded by a grant from the National Science Foundation (MCB 0731913) to TMS. Special thanks to the WK Kellogg Biological Station personnel, The Josephine Bay Paul Center (ML Sogin lab) and The Research Technology Support Facility at MSU for their valuable comments.

References

Briggs AD, Stenzel U, Johnson PLF, Green RE, Kelso K, Prüfer K *et al.* (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *Proc Natl Acad Sci USA* **104**: 14616–14621.

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM *et al.* (2008). Functional metagenomic profiling of nine biomes. *Nature* **452**: 629–632.

Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW *et al.* (2008). Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci USA* **105**: 3805–3810.

Li W, Godzik A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658–1659.

Li W, Wooley JC, Godzik A. (2008). Probing metagenomics by rapid cluster analysis of very large datasets. *PLoS ONE* **3**: e3375.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M *et al.* (2008). The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**: 386–394.

Mou X, Sun S, Edwards RA, Hodson RE, Moran MA. (2008). Bacterial carbon processing by generalist species in the coastal ocean. *Nature* **451**: 708–711.

Pernthaler A, Dekas AE, Brown CT, Goffredi SK, Embaye T, Orphan VJ. (2008). Diverse syntrophic partnerships from deep-sea methane vents revealed by direct cell capture and metagenomics. *Proc Natl Acad Sci USA* **105**: 7052–7057.

Sambrook J, Russell DW. (2001). *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, New York, USA.

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR *et al.* (2006). Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc Natl Acad Sci USA* **103**: 12115–12120.

Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**: 1027–1031.

Urich T, Lanzén A, Qi J, Huson DH, Schleper C, Schuster SC. (2008). Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS ONE* **3**: e2527.

Zhou J, Bruns MA, Tiedje JM. (1996). DNA recovery from soils of diverse composition. *Appl Environ Microbiol* **62**: 316–322.