The ISME Journal (2011) 5, 777–779

www.nature.com/ismej

COMMENTARY

The Future of microbial metagenomics (or is ignorance bliss?)

Jack A Gilbert, Folker Meyer and Mark J Bailey

The ISME Journal (2011) **5,** 777–779; doi:10.1038/ismej.2010.178; published online 25 November 2010

A recent explosion in the number of studies taking advantage of the power of next-generation sequencing to explore metagenomic or 16S rRNA taxonomic diversity of microbial environments means that, we need to stop and think about how we best interpret these data.

Currently, 16S *rRNA* gene studies provide us with the most effective way of fingerprinting the species richness of a community, with weak links to culture-derived functional relationships. However, the vast increase in the number of bacterial taxa known only from 16S rRNA sequence has started to sever this link, which can only be corrected through more experimental functional characterization requiring improved culturing techniques. A diversity study that uses core-genome processes and key metabolic functions would be an improvement, especially in the absence of the sequence of every genome for every cell in a system. But even then, we know only the potential of the system, not how it is regulated or what is expressed under any given circumstance, for which metatranscriptomics is required (Gilbert *et al.*, 2008). When appropriately applied, core-genome fingerprints could provide a genuine understanding of the population structure with defined niches, insight to functional variation and how these vary between ecosystems. Currently, our understanding is still very limited, but we do have some ideas about how to proceed.

Better bioinformatics

All interpretation of sequence data currently relies on the analysis of sequence similarity, assuming that similar (or near identical) DNA sequences imply similar (or identical) protein function. As numerous studies have shown that the general paradigm is valid, however, our knowledge of the protein universe is less than perfect. Not only are the current annotations of protein-coding genes not comprehensive, but also a large proportion of genes in newly sequenced microbes and viruses cannot be annotated (or even identified; Roberts, 2004), which severely limits our ability to use metagenomics for microbial diversity studies. Although concerted efforts are under way to improve the coverage of known genome-derived proteins from wet-lab derived biochemical annotations, the technology to link the small body of experimentally generated evidence to large 'families' of similar proteins is still in flux. One major factor that hinders the use of the existing annotations is the bias in the genome and protein knowledge bases, for example, the vast majority of sequenced organisms originate from the medical community. The Genomic Encyclopedia of Bacteria and Achaea (GEBA) project has provided a substantial increase to our understanding of microbial genomics from the rest of the phylogenetic matrix, highlighting the importance of whole genomes in exploring evolution and the protein universe. Importantly, this one study significantly increased the number and diversity of novel proteins, expanding our ability to annotate environmental metagenomic data by as much as 4% (Ivanova *et al.*, 2010).

One major concern is the use of different annotation pipelines by each sequencing center, which potentially produce different results. Simple processes like comparative genomics routinely require re-analysis of all data involved in the comparison (Dinsdale *et al.*, 2008). Attempts to create simple exchange vocabularies have not proven useful for microbial genome analysis. This highlights two issues:

(1) With future data volumes (for example, >300 billion base pairs per run on a HiSeq2000 Illumina platform), re-analysis will not be feasible because the data analysis cost will dominate the sequencing cost (Wilkening *et al*, 2009).
(2) Databases used for metagenomic analysis need to be well curated and expanded, the community requires sustained investment into annotation infrastructure.

International coordination of effort and access to sequencers/super computers

The genomic project registry (http://www.genomesonline.org) created by Nikos Kyrpides and colleagues, allows tracking of (meta)genome sequencing projects, avoiding costly repetition of identical experiments. A similar registry will be required for ecologically driven sequencing projects, helping to

avoid duplication of ecosystems, assisting with project design and allowing for the acquisition of comparable data sets. To provide such a project registry, researchers will need a language to express their projects in a computer searchable way. Through the work of the Genomics Standards Consortium (GSC; http://www.gensc.org), the community is now developing controlled vocabularies that allow accurate (and machine readable) descriptions of ecological sequencing projects, enabling questions like: 'Show me all studies of Mediterranean marine sediments in less than 100 meters of water.' This reduces months of paper-searches to seconds of data acquisition.

Coordination of data storage and access

Traditionally, DNA sequence data are archived at NCBI's Genbank (Benson *et al.*, 2009). More recently, environmental (metagenomic) sequences have been deposited in the short read archive (SRA) (http://www.ncbi.nlm.nih.gov/sra). However, SRA deposition and querying is not simple. In addition, it is unclear whether NCBI will continue to function as an archive for all DNA reads generated by a democratized sequencing community. Looking at this from the perspective of a microbial ecology data generator and/or data consumer, it seems clear that the community needs a comprehensive sequence archive for all 16S *rRNA* gene and metagenomic sequence reads. This will provide an important resource for the microbial ecology community, no matter how inexpensive sequencing becomes. The effort involved in sample extraction and description alone will make long-term storage and provisioning of the sequencing data worthwhile even as technologies change. The exact specifications needed are already described by current de-facto repositories (for example, VAMPS (http://vamps.mbl.edu/), MG-RAST (Meyer *et al.*, 2008) and CAMERA (http://camera.calit2.net/)). While in the past the community lacked the technology to describe metadata (experimental setup, sampling strategy, and so on.), through the work of the GSC we can now define the required metadata, enabling data creators to mark-up data, and software systems to ingest and provide ways to query and visualize the data.

In this brave new world, one can imagine many portals integrating data relevant for their specific missions, thus, creating de-facto archives by downloading from the data producers directly. However, if long-term storage (beyond funding cycles) is required, resources will need to be dedicated to preserve data sets over long periods of time, which must be through the existing network of the INSDC (http://www.insdc.org).

Designing the next generation of experiments

Of course it is the fundamental question of microbial ecology that will focus future research, and the interplay of different technologies will be paramount in answering these questions. For example, high-throughput 16S *rRNA* gene studies alone can significantly increase our concept of the diversity of life. Now that Rob Knight has shown that short regions of the 16S *rRNA* gene can provide us with as good a picture of microbial diversity as full length reads (Liu *et al.*, 2007), the massive throughput of Illumina can be leveraged to run thousands of parallel 16S *rRNA* gene projects in a single instrument run. Understanding how we apply these techniques to each ecosystem is as important as how we cope with the computational analysis—for example, how do we effectively determine the relevant sample size to accurately determine how ecosystem community structure changes over time or space. For future studies, as sequencing and bioinformatics become less of a bottleneck, it will become important that we examine sampling infrastructure, requiring that communities come together to produce standards associated with sampling volume, technology and application. Understanding the role of spatial scale and sampling volume in capturing microbial interaction and community structure is vital to these studies.

Concluding remarks

The ultimate future goal of our community is to provide a far more detailed understanding of microbial ecology to enable parameterization of ecosystem models, which are predictive and descriptive for diversity and metabolism. To do this, we must improve knowledge transfer and the intelligent interpretation of data at a global scale. Improved exchange of ideas and data will inevitably improve and advance the theory, perhaps, even help to define the basic rules for biological systems beyond the constant of nucleic acid. But to achieve this, ecological practices need to be improved and shared so that metadata and genomic information are based on sound experimentation that is built on statistically relevant design. To do this, we need to provide the support and infrastructure to ensure that samples and information are properly curated and readily accessible.

JA Gilbert is at Argonne National Laboratory, Argonne, IL, USA
JA Gilbert is also at Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA; F Meyer is at Argonne National Laboratory, Argonne, IL, USA
F Meyer is also at Computation Institute, University of Chicago, Chicago, IL, USA and MJ BaileyMJ Bailey is at NERC Centre for Ecology & Hydrology, Crowmarsh Gifford, Wallingford Oxford, UK
E-mail: gilbertjack@anl.gov

References

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. (2009). GenBank. *Nucleic Acids Res* **37**(Database issue): D26–D31.

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.

Gilbert JA, Field D, Huang Y, Edwards R, Li W, Gilna P *et al.* (2008). Detection of Large Numbers of Novel Sequences in the Metatranscriptomes of Complex Marine Microbial Communities. *PLoS ONE* **3**: e3042. journal.pone.0003042.

Ivanova N, Tringe SG, Liolios K, Liu W-T, Morrison N, Hugenholtz P *et al.* (2010). A call for standardized classification of metagenome projects. *Environmental Microbiology* **12**: 1803–1805.

Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R. (2007). Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res* **35**: e120.

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.

Roberts R. (2004). Identifying Protein Function—A Call for Community Action. *PLoS Biol* **2**: e42.

Wilkening J, Desai N, Meyer F, Wilke A. (2009). Using clouds for metagenomics — case study. *IEEE Cluster 2009*; New Orleans.