The ISME Journal (2012) 6, 1625–1628

www.nature.com/ismej

COMMENTARY

Describing microbial communities and performing global comparisons in the 'omic era

Tom O Delmont, Pascal Simonet and Timothy M Vogel

The ISME Journal (2012) **6,** 1625–1628; doi:10.1038/ismej.2012.55; published online 21 June 2012

Environmental metagenomic approaches emerged with the cloning of DNA fragments extracted directly from the environment into cultivable microorganisms in order to describe the microbial diversity (Pace *et al.*, 1986; Olsen *et al.*, 1986) and to stimulate the discovery of new enzymes and antibiotics (Handelsman *et al.*, 1998). This new field of research has evolved and is now largely focused on massive sequencing of DNA directly (or indirectly) extracted from various environments. The generated metagenomic data sets represent part of the metagenome, that is, part of the entire genetic diversity representing all members of the targeted system. Today, most metagenomic sequencing of complex environments cannot separate different microbial strains from each other, although differences at higher taxonomic levels are often proposed. In addition, except for a few number of habitats (for example, acid mine drainage biofilms, Tyson *et al.*, 2004), the genetic diversity in extracted DNA is several orders of magnitudes too complex to easily assemble all the genomes present with sequences from current sequencing technologies. On the other hand, the sequences already produced are used to define the functional potential of the microbial community represented in the data sets, as well as assemble some genomes. Due in part to the interest in evaluating the community's functional capacity for microbial ecology, the number of available metagenomic data sets is growing and represent sequences from a wide range of microbial communities (for example, data sets on MG-RAST (Meyer *et al.*, 2008) and IMG/M (Markowitz *et al.*, 2008) public annotation platforms). While a majority of studies are done by analyzing data sets from a single ecosystem, considerable information can also be extracted by performing inter-environmental metagenomic data set comparisons (for example, Tringe *et al.*, 2005; Dinsdale *et al.*, 2008; Delmont *et al.*, 2011).

A critical question is the proportion of the metagenome that is sequenced and the effect of incomplete metagenomic data sets on qualitative and quantitative analyses. Confidence concerning the proportion of the metagenome that has been sequenced probably varies inversely to the biodiversity of the environment or ecosystem targeted. Other factors are, of course, sample selection, preparation and DNA extraction, as well as the amount of sequencing done. In general, the more sequencing of a specific ecosystem, environment, biome or sample performed, the higher the proportion of the actual metagenome sequenced. Quantitative and qualitative comparisons assume that the sequenced (and in many cases annotated) fractions of the total metagenome provide accurate information of both phylogenetic and metabolic characteristics. The importance of the missing data is difficult to estimate. Testing the missing data by deeper sequencing and by drastically increasing the number of sequences from different ecosystems might help if there is no systematic bias (during sampling, DNA extraction and sequencing) that would only be repeated.

Relative abundances require a quantitative approach although qualitative analyses are also performed. For the qualitative approach, the recognition of the presence of a taxonomic unit or a metabolic function is used to describe the microbial diversity present in the ecosystem. The quantitative approach is to provide the relative proportion of the taxonomic unit or metabolic function relative to others in the same sample or to other samples or ecosystems. Today, determining both the absence and the true relative proportion seems overly optimistic. They cannot be validated as the entire DNA diversity is not extracted and sequenced in most cases and because each method tested produces metagenomic data sets with different taxonomical and functional distributions. Considering thousands of variations in metagenomic approaches (from DNA extraction to sequence annotation) that could be applied to the same sample, the probability that any one method would provide the true picture of a given microbial diversity seems miniscule. On the other hand, the observation of the real presence of a taxonomical unit or functional subsystem or gene seems more likely (ignoring for the moment the bioinformatics difficulty of phylogenetic or metabolic assignment), but limits the application of 'omic approaches to microbial ecology.

The principal challenge is to connect structure, functions and environmental characteristics at both a local and a global level without losing the biological information during the transformation of

in situ microbial ecology to *in silico* analyses. Current global metagenomic analyses suggest that extraction, sequencing and annotation biases have little effect on comparisons between significantly different ecosystems, such as oceans, soils and human digestive tracts (Delmont *et al.*, 2011, 2012), but this might be due to the large differences in the structure and function of the microbial communities. These differences usually decrease when comparing two similar sample types or ecosystems and then methodological biases might become the overriding factors. One way to measure the acceptable degree of ecosystem or sample similarity at which methodological biases do not hinder global comparisons is to apply different methods to each of the biomes under study and evaluate the variability observed and its effect on the scientific conclusions of the study. In the future, combined conceptual and mathematical models might help evaluate the effect of different methods on different ecosystem metagenome recoveries and biases.

Another approach is to apply the same method everywhere (for example, Venter *et al.*, 2004) and consider that its associated biases apply to all samples in the same way. This approach would lead to comparing thousands of data sets corresponding to various environments or samples using one standardized protocol. Nevertheless, concerns that sampling and DNA extraction biases appear to vary as a function of environment, season, depth, temperature, organic content, and so on would not be addressed if a single standardized protocol is applied. Today, the debate about which methods are the best for qualitative and quantitative analyses suggests that metagenomic approaches are still in their infancy. As each method provides a different image (data set) of the metagenomic structure, the criteria for choosing one specific method have not been established. Even the method that extracts the largest quantity of DNA cannot be defended when its relative biases are unknown and given that the DNA extraction and sequence annotation steps have the potential to influence conclusions concerning relative abundances.

To help alleviate this limitation, environments could be sampled over time and space, and by varying DNA extraction and sequencing and annotation methodologies. Thus, the non-biological factors that influence our perception of the structure of a metagenome could be varied, too. The lack of replicates can also reduce the validity of interpretations of metagenomic studies (Prosser, 2010). Replicate reproducibility is only part of the problem as it provides information about method precision, but not about the accuracy of the results. Given that considerable biases can be hidden behind highly reproducible replicates (an example is presented in Figure 1), replicates might be necessary but insufficient for both quantitative and qualitative analyses; the application of different approaches to access the genetic diversity present in the environment might improve the relevance of metagenomic surveys. We have applied this strategy on multiple samples of the same soil, and although our replicates were relatively similar, our different methods did not provide identical relative abundances of functions (Delmont *et al.*, 2012).

As it is currently impossible to know which approach provides the truest image of the microbial community, multiple methods are warranted even if they are all not applied to all samples. Multiple DNA extraction approaches could be applied using a wide range of strategies depending on environmental characteristics. The interest of applying different methods to the same site is two-fold: (1) it can enable the detection of different portions of the metagenome, and (2) it could increase the number of function and species detected (Delmont *et al.*, 2012). To test the relevance of this approach, artificial mixes of different microorganisms from several different environments and added to sterile samples

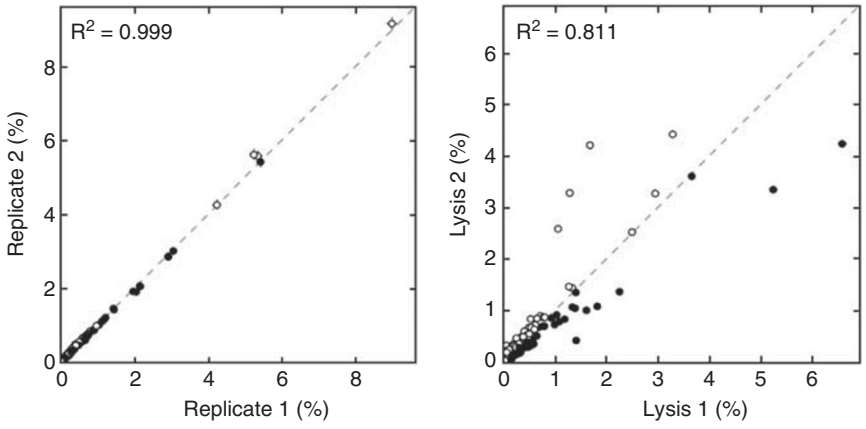

Figure 1 The two graphs represent the relative distribution in percentage of sequences related to genera (using MG-RAST (Meyer *et al.*, 2008), SEED annotation (*E*-value $<10^{-5}$) and STAMP software (Parks and Beiko, 2010)) between two data sets. Replicates 1 and 2 correspond to DNA pools extracted (MP BIO 101) from two distinct control microcosms after 4 months of incubation. Lysis 1 and 2 correspond to the same soil sample and two distinct DNA extraction protocols (see Delmont *et al.*, 2012).

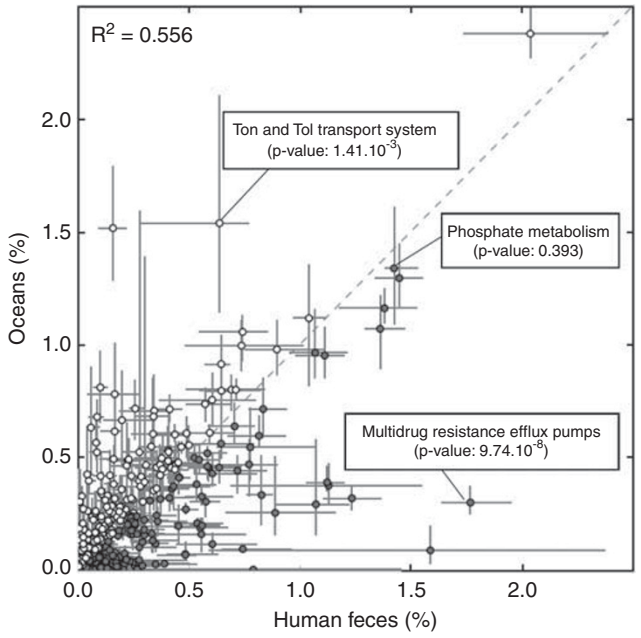

Figure 2 Relative distribution of microbial functions from the ocean surface water (12 data sets) and human feces (11 data sets) (MG-RAST annotation, functional level 3, *E*-value $<10^{-5}$) by integrating spatial, temporal and methodological fluctuations (metagenome information is presented in Delmont *et al.*, 2011) using the STAMP v2.0. s.ds. are presented for each functional subsystem and for the two environments.

(for example, sterile soils, water) could be tested with the different methods as a function of the species and physico-chemical characteristics in order to increase our confidence in the methods chosen. As the data set representing the microbial community varies with physico-chemical conditions (for example, pH), the observed variation could be either the community's response to or the extraction method's dependence on the given condition, although arguments about the artificial nature of these experiments could not be refuted.

Biological replicates can and should be performed to incorporate both method reproducibility and inter-method variability (or result variations) into metagenomic surveys. After sampling, managing and sequencing DNA pools, data sets could then be compiled to represent a global picture of each studied environment. The functional and taxonomical distribution differences so obtained represent both natural and methodological variations. Increasing the variability within the metagenomic data set for a given environment does not provide access to the true composition of the related community and will affect the statistical comparisons of data sets from different environments. This increase in variability might hide true differences, although there is no current method for determining which differences are true. On the other hand, there will be less statistically significant differences between environmental data sets (that is, when no overlaps are observed between the different distributions of the two environments being compared), but these differences should be more robust, hence closer to reality.

The main goals of this combined methodological approach are to help microbial ecologists decrypting communities to understand how microbial structure and function varies from site to site and to perform global metagenomic comparisons with more confidence in their data sets (an example of a possible comparison between human feces and oceans by partially integrating methodological variations is presented in Figure 2). If this strategy is used, then a metagenomic measurement of ecosystem boundaries at the microorganism level could be indirectly defined when inter-environmental distribution differences are globally stronger than intra-environmental variations, whether they are biological, physical or methodological. Technological advances might reduce the methodological difficulties in future metagenomic surveys, but until then, some confidence in the data is needed if we are going to fully explore how 'omic approaches can participate in microbial ecology.

TO Delmont, P Simonet and TM Vogel are at Environmental Microbial Genomics, Ecole Centrale de Lyon, Université de Lyon, Ecully, France
E-mail: tvogel@ec-lyon.fr

References

Delmont TO, Malandain C, Prestat M, Larose C, Monier JM, Simonet P, Vogel TM. (2011). Metagenomic mining for microbiologists. *ISME J* **5**: 1837–1843.

Delmont TO, Prestat E, Keegan KP, Faubladier M, Robe P, Clark IM *et al.* (2012). Structure, fluctuation and magnitude of a natural grassland soil metagenome. *ISME J;* e-pub ahead of print 2 February 2012; doi:10.1038/ismej.2011.197.

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.

Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. (1998). Molecular Biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol* **5**: R245–R249.

Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D *et al.* (2008). IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res* **36**: D534–D538.

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.

Olsen GJ, Lane DJ, Giovannoni SJ, Pace NR, Stahl DA. (1986). Microbial ecology and evolution: a ribosomal RNA approach. *Annu Rev Microbiol* **40**: 337–365.

Pace NR, Stahl DA, Lane DJ, Olsen GJ. (1986). The analysis of natural microbial populations by ribosomal RNA sequences. *Adv Microb Ecol* **9**: 1–55.

Parks DH, Beiko RG. (2010). Identifying biologically relevant differences between metagenomic communities. *Bioinformatics* **26**: 715–721.

Prosser JI. (2010). Replicate or lie. *Environ Microbiol* **12**: 1806–1810.

Tringe SG, Mering CV, Kobayashi A, Salamov AA, Chen K, Chang HW *et al.* (2005). Comparative metagenomics of microbial communities. *Science* **308**: 554–557.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM *et al.* (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37–43.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA *et al.* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. **304**: 66–74.