

What's the 'meta' with metagenomics?

Grieg F Steward and Michael S Rappé

The ISME Journal (2007) 1, 100–102; doi:10.1038/ismej.2007.25; published online 17 May 2007

Genomics, a field of study concerned with the sequencing and analysis of whole genomes, has traditionally advanced through the accumulation of data from individual sequencing projects, each being devoted to completing the genome of a single strain or an individual. Metagenomics is the application of the methods of genomics to microbial assemblages. For this purpose, microorganisms are harvested from the environment by simple physical means, and their DNA is extracted and processed to create a single community 'shotgun' library. The great appeal of metagenomics is that it circumvents cultivation and provides proportional representation of all of the genomic information in a sample that can then be mined for new insights.

The simplicity of the metagenomic approach also results, unfortunately, in the degradation and loss of information that is difficult, if not impossible, to recover. First, the extraction of DNA *en masse* from a diverse microbial assemblage destroys all the phenotypic information content. Then the remaining genomic information is effectively scrambled in the process of generating smaller, clonable DNA fragments. Although it is clearly inefficient to first scramble information that must be later reassembled, this has been the state of the art because of the limitations of existing cloning and sequencing technology.

Inefficient though it may be, the random shotgun approach has worked well for reconstructing the genomes of the dominant microorganisms harvested from unusual environments having limited microbial diversity (Tyson *et al.*, 2004; Legault *et al.*, 2006). When applied to other natural environments, however, such as soil, sediment and sea water, the yield of assembled genomes has been poor. Nearly eight billion bases of sequence have been generated so far by metagenomic analyses of prokaryote communities, yet only a handful of genomes have been partially reconstructed (Tyson *et al.*, 2004; Venter *et al.*, 2004; Rusch *et al.*, 2007), some of which may be contaminants (DeLong, 2005). Another 0.2 billion bases have been sequenced from metagenomic libraries of marine DNA viruses, but only one small phage genome has been reconstructed (Edwards and Rohwer, 2005; Angly *et al.*, 2006).

Random sampling of metagenomic libraries has resulted in the discovery of many new genes (Venter

et al., 2004) and proteins (Yooseph *et al.*, 2007) and has provided glimpses of how the prevalence of different functional categories of genes may vary among habitats (Tringe *et al.*, 2005; Angly *et al.*, 2006; DeLong *et al.*, 2006; Rusch *et al.*, 2007), but these unassembled sequence data are sorely lacking in genomic and organismal context. A *metagenome* is, ultimately, just a phantom genome whose appearance varies depending upon how, when, and on what spatial scale it is sampled. For a field focused on genomic analysis of natural microbial communities, one of the major goals ought to be the reconstruction of actual genomes.

The current trend in metagenomics is to sequence at increasingly high throughput (Rusch *et al.*, 2007), sometimes at the expense of read length (Angly *et al.*, 2006; Edwards *et al.*, 2006), and to rely on more sophisticated bioinformatics and more powerful computing resources to make sense of the mess (Chen and Pachter, 2005; Seshadri *et al.*, 2007). Investments in improving these downstream processes are needed and are welcomed, but no matter what the sequencing capacity or computing power available, metagenomics will always be an inefficient use of the available resources, because it is built upon a foundation that sacrifices information content for upstream technical convenience.

A more economical way to solve the assembly problems of metagenomics would be to focus on improving upstream processing steps to generate more meaningful and tractable starting material. In particular, isolating individual populations, or even limited consortia of intact microbes, before genomic analysis would allow more efficient reassembly of genomes, especially for rare populations and would facilitate the interpretation of those genomes by allowing parallel analyses of a microorganism's genotype and phenotype. Enrichment or isolation of microbial populations may be achieved through cultivation, but it is also possible to use physical means to separate complex communities into constituent populations or less diverse consortia.

Having a microorganism in pure culture offers unique advantages that make this an attractive option whenever it is achievable. The ability to relate physiological responses to gene expression under controlled conditions, for example, provides information still not obtainable in any other way. Cultivation methodologies have advanced in leaps and bounds in recent years and efforts to isolate additional representative microbes from diverse environments for whole genome sequencing, as well

as sequencing more of those isolates already in hand, should be high priorities. New techniques may be required to cultivate many key species (Giovannoni and Stingl, 2005), but it is well recognized that additional whole genome sequences of isolates will be invaluable for improving assembly and interpretation of metagenomic data (Rusch *et al.*, 2007).

There are practical limitations to obtaining and maintaining pure cultures, of course, just as there are certain advantages in analyzing microbes extracted directly from their natural habitat (Banfield *et al.*, 2005). For these reasons, cultivation-independent approaches will always be key to understanding the ecological implications of genomic diversity. Metagenomics is one such approach, but its principal failing stems from the fact that in circumventing cultivation, it also dispenses with isolation. Since the two are not synonymous, this need not be the case. It is possible to distinguish and physically separate intact microbes from one another based on size, buoyant density, surface chemistry, optical properties or a host of other characteristics. Diverse fractionation techniques operating on these principles have been successfully used to separate, and even purify, populations of intact eukaryotic cells, prokaryotes and viruses. Despite a wealth of available techniques that could be readily adapted to the purpose, filtration is typically the only fractionation procedure applied in the preparation of samples for metagenomic analysis.

While it may not be possible to cleanly isolate every population from a complex community by either cultivation or physical fractionation, creative use of multiple techniques in series, each separating based on a different characteristic, could dramatically reduce the complexity within individual fractions. The benefits of reducing complexity are best illustrated by the successful assemblies that have been achieved by targeting subsets of an assemblage either based on genomic characteristics (Dale *et al.*, 2005; Culley *et al.*, 2006) or by enriching for specific populations of intact cells via cultivation (Erkel *et al.*, 2006) or physical fractionation (Hallam *et al.*, 2006). Fractionation, by separating abundant and rare species from one another, also has the potential to improve the efficiency of mining novel genes and products from the environment once the sampling of the numerically dominant species begins to reach saturation.

For some types of analyses, large amounts of starting material will be required to achieve sufficient biomass of a physically fractionated target population. However, with amplification techniques now permitting the genomic analysis of a single cell (Zhang *et al.*, 2006), the sample size needed just to sequence and assemble the genome of even a rare population could be quite small. With a decreased DNA requirement, even highly discriminating, semi-preparative fractionation techniques, such as sorting flow cytometry, could be used as part of a

multi-dimensional fractionation procedure. In many cases, physical fractionation could result in sufficient purity such that genome, proteome and morphology of targeted uncultivated microbes can be directly and unambiguously compared. Depending on the specific separation techniques employed, some populations of microbes may even retain viability after fractionation, thereby permitting physiological characterization (Moissl *et al.*, 2003) and facilitating cultivation efforts (Crosbie *et al.*, 2003).

Metagenomics is now a field unto itself, and metaproteomics (Wilmes and Bond, 2004) and metatranscriptomics (Poretzky *et al.*, 2005) are close behind, but the value of performing these types of analyses should be critically examined in the light of the scientific questions to be answered, the nature of the system to be analyzed, and the other analysis options available. With a little innovation and concerted effort, the field of microbial ecology need not remain stuck in its 'meta' phase. It is time to move beyond 'phantomics' and realize the exciting prospects of achieving coupled genotypic and phenotypic analyses of the actual microorganisms that make up even the most complex natural communities.

GF Steward is at the Department of Oceanography, University of Hawaii at Manoa, Honolulu, HI, USA and MS Rappé is at the Hawaii Institute of Marine Biology, University of Hawaii at Manoa, Kane'ohe, HI, USA. E-mail: grieg@hawaii.edu

References

- Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C *et al.* (2006). The marine viromes of four oceanic regions. *PLoS Biol* **4**: 2121–2131.
- Banfield JF, VerBerkmoes NC, Hettich RL, Thelen MP. (2005). Proteogenomic approaches for the molecular characterization of natural microbial communities. *OMICS* **9**: 301–333.
- Chen K, Pachter L. (2005). Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Comput Biol* **1**: e24 (doi:10.1371/journal.pcbi.0010024).
- Crosbie ND, Pöckl M, Weisse T. (2003). Rapid establishment of clonal isolates of freshwater autotrophic picoplankton by single-cell and single-colony sorting. *J Microbiol Meth* **55**: 361–370.
- Culley AI, Lang AS, Suttle CA. (2006). Metagenomic analysis of coastal RNA virus communities. *Science* **312**: 1795–1798.
- Dale C, Dunbar H, Moran NA, Ochman H. (2005). Extracting single genomes from heterogenous DNA samples: a test case with *Carsonella ruddii*, the bacterial symbiont of psyllids (Insecta). *J Insect Sci* **5**: 3.
- DeLong EF. (2005). Microbial community genomics in the ocean. *Nat Rev Microbiol* **3**: 459–469.

- DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N-U *et al.* (2006). Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**: 496–503.
- Edwards RA, Rodriguez-Brito B, Wegley L, Haynes M, Breitbart M, Peterson DM *et al.* (2006). Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* **7**: 57 (doi:10.1186/1471-2164-1187-1157).
- Edwards RA, Rohwer F. (2005). Viral metagenomics. *Nat Rev Microbiol* **3**: 504–510.
- Erkel C, Kube M, Reinhardt R, Liesack W. (2006). Genome of rice cluster I Archaea – the key methane producers in the rice rhizosphere. *Science* **313**: 370–372.
- Giovannoni SJ, Stingl U. (2005). Molecular diversity and ecology of microbial plankton. *Nature* **437**: 343–348.
- Hallam SJ, Konstantinidis KT, Putnam N, Schleper C, Watanabe Y, Sugahara J *et al.* (2006). Genomic analysis of the uncultivated marine crenarchaeote *Cenarchaeum symbiosum*. *Proc Natl Acad Sci USA* **103**: 18296–18301.
- Legault BA, Lopez-Lopez A, Alba-Casado JC, Doolittle WF, Bolhuis H, Rodriguez-Valera F *et al.* (2006). Environmental genomics of ‘*Haloquadratum walsbyi*’ in a saltern crystallizer indicates a large pool of accessory genes in an otherwise coherent species. *BMC Genomics* **7**: 171 (doi:10.1186/1471-2164-1187-1171).
- Moissl C, Rudolph C, Rachel R, Koch M, Huber R. (2003). *In situ* growth of the novel SM1 euryarchaeon from a string-of-pearls-like microbial community in its cold biotope, its physical separation and insights into its structure and physiology. *Arch Microbiol* **180**: 211–217.
- Poretsky RS, Bano N, Buchan A, LeClerc G, Kleikemper J, Pickering M *et al.* (2005). Analysis of microbial gene transcripts in environmental samples. *Appl Environ Microbiol* **71**: 4121–4126.
- Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007). The *Sorcerer II* global ocean sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5**: e77 (doi:10.1371/journal.pbio.0050077).
- Seshadri R, Kravitz SA, Smarr L, Gina P, Frazier M. (2007). CAMERA: a community resource for metagenomics. *PLoS Biol* **5**: e75 (doi:10.1371/journal.pbio.0050075).
- Tringe SG, von Mering C, 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 PW *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. *Science* **304**: 66–74.
- Wilmes P, Bond PL. (2004). The application of two-dimensional polyacrylamide gel electrophoresis of a mixed community of prokaryotic microorganisms. *Environ Microbiol* **6**: 911–920.
- Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K *et al.* (2007). The *Sorcerer II* global ocean sampling expedition: expanding the universe of known protein families. *PLoS Biol* **5**: e16 (doi:10.1371/journal.pbio.0050016).
- Zhang K, Martiny AC, Reppas NB, Barry KW, Malek J, Chisholm SW *et al.* (2006). Sequencing genomes from single cells by polymerase cloning. *Nat Biotechnol* **24**: 680–686.