Review

Building on basic metagenomics with complementary technologies

Falk Warnecke and Philip Hugenholtz

Address: Microbial Ecology Program, DOE Joint Genome Institute, Walnut Creek, CA 94598, USA.

Correspondence: Philip Hugenholtz. Email: PHugenholtz@lbl.gov

Published: 28 December 2007

Genome ***Biology*** 2007, **8:**231 (doi:10.1186/gb-2007-8-12-231)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2007/8/12/231

Abstract

Metagenomics, the application of random shotgun sequencing to environmental samples, is a powerful approach for characterizing microbial communities. However, this method only represents the cornerstone of what can be achieved using a range of complementary technologies such as transcriptomics, proteomics, cell sorting and microfluidics. Together, these approaches hold great promise for the study of microbial ecology and evolution.

The majority of microorganisms defy axenic culture in the laboratory and so have eluded study by the classic microbiological approaches [1]. With the advent of cultivation-independent molecular tools, the true extent of microbial diversity has been, and continues to be, revealed [2-4]. Much of that work, however, is based on a single phylogenetic marker gene, small subunit ribosomal RNA (ssu rRNA) [5]. By contrast, metagenomics in principle makes accessible the entire genetic complement of a microbial community - we define metagenomics here as the large-scale application of random shotgun sequencing to DNA extracted directly from environmental samples and resulting in at least 50 megabase pairs (Mbp) of sequence data. It has been barely three years since the publication of the first large-scale metagenomic studies: of an acid mine drainage biofilm [6] and of ocean surface water [7]. Since then, numerous other habitats have been investigated using this 'basic' metagenomic approach (Figure 1, arrow 1), including farmland soil and whale falls (whale carcasses that have fallen to the sea floor) [8], symbionts in a gutless marine worm [9], phosphorus-removing activated sludge [10], the human [11] and termite [12] gut and marine microbial [13,14] and viral [15] samples. In all these cases, metagenomics provided insights into the microbial community under study that probably would have taken much longer to come to light using more directed (nonrandom) approaches. Shotgun sequencing of environmental samples has, however, a number of limitations [16], which can best be addressed by the use of complementary techniques.

Limitations of environmental shotgun sequencing

Three notable limitations of the basic metagenomic approach are low resolution, the inability to classify short metagenomic fragments, and the lack of functional verification. Perhaps surprisingly, the resolution of microbial communities by shotgun sequencing is rather low, with only dominant populations producing sufficient sequence coverage to result in a sequence assembly. For example, assuming no other biases, a population representing 0.1% of a community would account for only 100 kilobase pairs (kbp) of a 100 Mbp metagenome, resulting in very little coverage (0.025X coverage for a 4 Mbp genome). If a recent study on the microbial diversity in the deep sea is an accurate indication of species-abundance distribution [4], rare community members comprising the bulk of the diversity in many environmental samples will be completely missed by current levels of shotgun sequencing.

The second limitation is in identifying the source species of metagenomic fragments. Current methods to classify such fragments do not perform well on sequences of less than 8 kbp [17], that is, the bulk of the sequence data obtained in most metagenomic studies. And third, as with all DNA sequence data, metagenomics can only provide information

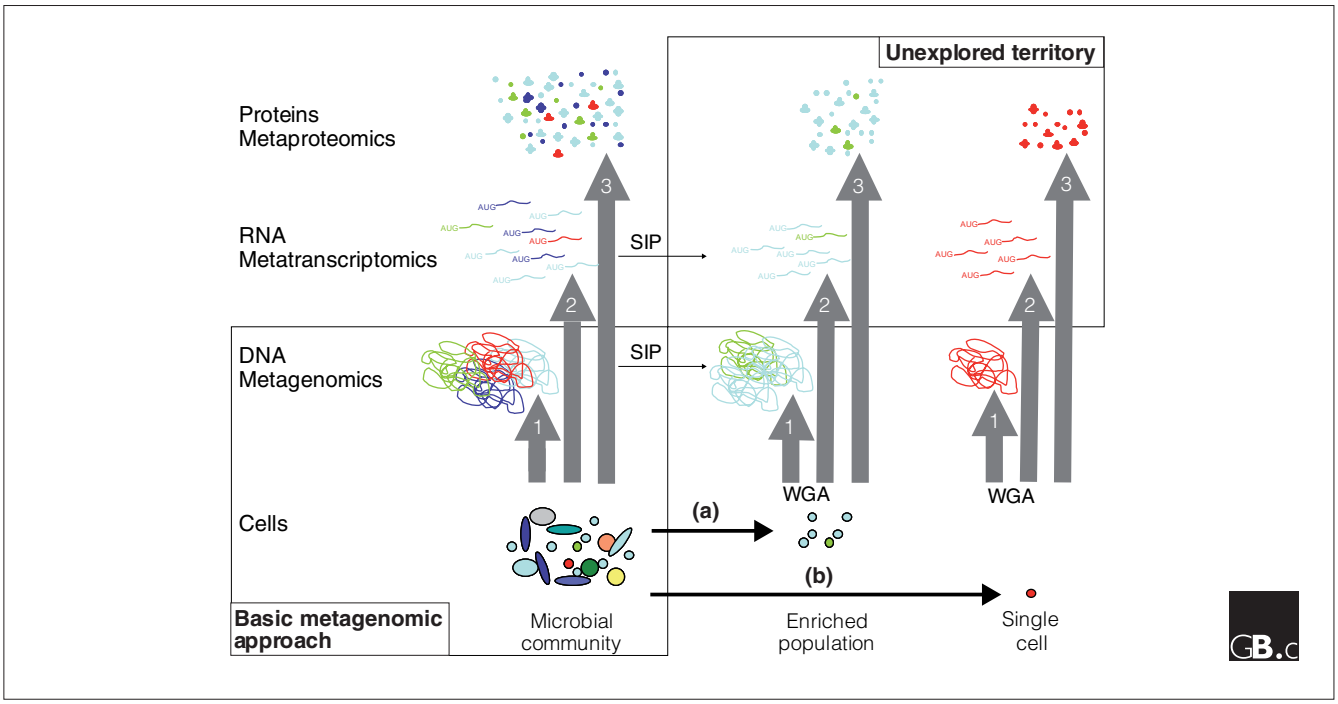

Figure 1
Enhancing the basic metagenomic approach through complementary technologies. The metagenomic analysis of microbial communities by random shotgun sequencing (arrow 1) is being enriched in one dimension by parallel detection and analysis of transcripts ('metatranscriptomics', arrow 2) and of expressed proteins ('metaproteomics', arrow 3). In addition, because of the complexity of most natural microbial communities a separation of the community into populations enriched in a particular group of microorganisms and even into individual cells would be advantageous. Whole-genome amplification (WGA) is beginning to be validated as an approach to metagenomic and metatranscriptomic analysis in such samples, but there are still some methodological constraints to be overcome (see text). The horizontal arrows indicate examples of techniques that can be used to move to the next level of analysis, for example, **(a)** flow sorting and filtration and **(b)** microfluidics and flow sorting. SIP, stable isotope probing.

on metabolic potential, and only for genes with recognizable homology with biochemically characterized proteins.

Divide and conquer

The first two limitations can be addressed by dividing microbial communities into simpler subsets, which facilitates contig identification and greater genomic coverage of populations. Ironically, cultivation of pure strains is an excellent example of this divide-and-conquer approach, as single cells or microcolonies are separated from an environmental inoculum and grown clonally on artificial media. However, directed cultivation of organisms of environmental relevance is typically difficult to achieve [1,18,19], although metagenomic studies can provide valuable guidance for such efforts [20].

Cultivation-independent methods to subdivide microbial communities into enriched populations (see Figure 1, arrow a) often rely on the physical properties of the target cells. For example, populations comprising cells of atypical size can be effectively enriched via filtration. This approach was successfully applied to enrich phylogenetically novel populations of ultra-small archaea using filters with a 0.45 μm pore size [21,22]. Both enriched populations have been the subject of subsequent genome sequencing projects ([23] and B.J. Baker, E.E. Allen and J.F. Banfield, unpublished work; see [24]). In a metagenomic project studying bacterial endosymbionts of a gutless marine oligochete worm, a Nycodenz density-gradient centrifugation was used to separate the bacterial and eukaryotic host-cell populations, improving the recovery of the bacterial genome sequences in subsequent shotgun sequencing [9].

More sophisticated techniques for separating cells from communities are also being applied, including fluorescence-activated cell sorting (FACS [25]) and microfluidics [26] (see Figure 1). FACS can be used to rapidly sort large numbers of cells belonging to specific populations on the basis of cell properties such as size, DNA content, photosynthetic pigments or fluorescently labeled probes targeting the cells [27-29]. Such sorting can provide enough biomass to allow direct extraction of DNA or RNA for the polymerase chain reaction (PCR) and shotgun sequencing. FACS and microfluidics can also be used to separate individual cells, with the caveat that single cells require whole-genome amplification, for example by multiple

strand displacement amplification (MDA [30]), to provide enough genomic DNA for shotgun sequencing.

Co-localization of PCR-amplified marker genes (such as ssu rRNA) and functional genes in single cells has recently been demonstrated in two independent studies. Ottesen and colleagues [31] used highly parallelized microfluidic chambers to separate individual cells and, via PCR, were able to link a key metabolic gene in homoacetogenesis to the ssu rRNA of treponeme spirochetes present in the termite hindgut. Bacterial homoacetogenesis delivers the major carbon and energy source (acetate) for the host termite, and hence represents an important link in this mutualistic symbiosis. Stepanauskas and Sieracki [32] flow sorted single marine planktonic cells into microtiter plates and identified a range of bacteria containing proteorhodopsin and other genes after MDA and PCR. In fact, their results hint at flavobacteria as major carriers of the proteorhodopsin gene. Compared with large-scale shotgun sequencing, this approach represents a rather low-cost alternative for studying the metabolic potential of uncultivated microbes. In summary, both the studies mentioned above mark an important milestone in microbial ecology - the systematic linkage of identity with function in uncultivated microorganisms. PCR-based co-localization of genes is, however, limited by existing sequence data and cannot access novel gene families discovered by random shotgun sequencing.

The holy grail of *de novo* sequencing of sorted cells, and individually sorted cells in particular, is to obtain a finished genome and thus a complete inventory of an organism's genetic potential. The feasibility of genome sequencing from just one or a few cells has been validated by using MDA and partial sequencing of species with known genome sequence (*Escherichia coli* [33] and *Prochlorococcus* [34]). This approach has been applied to members of the candidate bacterial phylum TM7 from the human mouth [35] and from soil [36], yielding some insights into the metabolic potential of novel uncultivated organisms. For example, the presence of genes for type IV pilus biosynthesis in the isolates from both studies [35,36] study may hint at a gliding motility known from some Gram-positive bacteria. However, the majority of genes of the TM7 genomes studied bear little similarity to genes of characterized proteins.

Full genome sequencing from a single microbial cell (Figure 1, arrow 1) remains problematic, however, due to contamination, uneven genome coverage and chimeric sequence formation during MDA [34,37]. A number of solutions have been proposed to somewhat mitigate these limitations. Reducing the reaction volume increases the specific template concentration, leading to fewer chimeric sequences [37]. Microfluidic devices allow MDA reactions at the nanoliter scale, which increases the specific template concentration by three orders of magnitude [35]. Uneven genome coverage, on the other hand, seems random [33] and hence pooling of separate MDA reactions from individual but genomically identical cells [36] should improve coverage.

Going beyond metabolic potential

A major criticism of metagenomics is that it is, to some extent, crystal-ball gazing as one attempts to infer the metabolism of organisms from their DNA sequence alone (the third limitation raised earlier: lack of functional verification). Indeed, purely metagenomic studies often raise more questions than they can answer. Transcriptomic and proteomic analyses have been applied for several years to microbial isolates in order to observe their expressed metabolic potential [38,39]. These approaches have recently been applied in a high-throughput fashion to microbial communities - coining the terms 'metatranscriptomics' and 'metaproteomics'.

A technical difficulty associated with transcriptomics in bacteria and archaea is separating mRNAs from the dominant rRNAs. The poly(A) tail of eukaryotic mRNAs (which facilitates their separation from rRNAs before cDNA synthesis) is not present on bacterial and archaeal transcripts [40]. Leininger and colleagues [41] circumvented this problem to some extent by simply using the brute force of the new massively parallel short-read sequencing technologies to absorb the loss of transcript sequence output due to the predominance of rRNA. Through this approach they provided unexpected evidence for members of the Crenarchaeota being the most active ammonia-oxidizing microorganisms in soil ecosystems [41].

Modern proteomic methods based on mass spectrometry allow a fine-scale analysis of the expressed proteins of microbial communities [42]. By combining such techniques with genomic data, Lo *et al*. [43] were able to distinguish strain-specific protein variants differing in only a single amino-acid residue from a different site in the same mine. Interestingly, 48% of the proteins predicted in the genome sequence of the most abundant member in this system, *Leptospirillum* group II, were detected by proteomics. This value is higher than those reported for many proteomic analyses of isolates and may point to a heterogeneity of metabolic states in naturally occurring populations [42].

Unexplored territory

By describing techniques that extend the basic metagenomic approach in two dimensions - gene expression and translation (Figure 1, arrows 2,3) and community fractionation (Figure 1, arrows a,b) - additional combinations become apparent that remain to be explored (see Figure 1 'Unexplored teritory'). Applying transcriptomics and proteomics to separated populations will allow functional

characterization of species that have been inaccessible via cultivation so far. The many phyla in the tree of life without genome-sequenced representatives will provide attractive targets for this type of analysis [2].

The application of transcriptomics and proteomics to enriched populations or even individual microbial cells taken directly from the environment remains technically challenging (see Figure 1, arrows 2,3). However, the technical hurdles may not be insurmountable. For instance, electrospray ionization/mass spectrometry can provide greater sensitivity than the currently standard liquid chromatography mass spectrometry used in proteomics, leading to smaller sample size requirements [44]. Commercial kits are already available for amplifying RNAs from as few as 50 cells (for example, QuantiTect™ from Qiagen) paving the way for single-cell transcriptomics. Such methods would allow functional characterization of single cells, providing insights into the heterogeneity of expression postulated to exist in microbial cell populations [45]. Moreover, if these approaches prove viable, such population expression heterogeneity would be assessable in the context of the community from which the population was derived.

Although there is still great scope for application of the basic metagenomic approach to microbial communities - in making spatial series [14] and in population genomics [46,47] for example - researchers are making concerted efforts to extend and enhance metagenomics using techniques such as flow sorting, microfluidics, transcriptomics and proteomics. There are many other recently developed methods that can similarly be applied to build on or complement the basic metagenomic approach, including stable isotope probing [48], stable isotope mass spectroscopy [49] and subcellular high-resolution imaging [50], guaranteeing a rich and interesting future for those who study microbial ecology and evolution.

References

1. Kaeberlein T, Lewis K, Epstein SS: **Isolating "uncultivable" microorganisms in pure culture in a simulated natural environment.** *Science* 2002, **296:**1127-1129.
2. Hugenholtz P: **Exploring prokaryotic diversity in the genomic era.** *Genome Biol* 2002, **3:**reviews0003.1-0003.8.
3. Rappe MS, Giovannoni SJ: **The uncultured microbial majority.** *Annu Rev Microbiol* 2003, **57:**369-394.
4. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ: **Microbial diversity in the deep sea and the underexplored "rare biosphere".** *Proc Natl Acad Sci USA* 2006, **103:**12115-12120.
5. Pace NR: **A molecular view of microbial diversity and the biosphere.** *Science* 1997, **276:**734-740.
6. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF: **Community structure and metabolism through reconstruction of microbial genomes from the environment.** *Nature* 2004, **428:**37-43.
7. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu DY, Paulsen I, Nelson KE, Nelson W, *et al.*: **Environmental genome shotgun sequencing of the Sargasso Sea.** *Science* 2004, **304:**66-74.
8. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, *et al.*: **Comparative metagenomics of microbial communities.** *Science* 2005, **308:**554-557.
9. Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, Gloeckner FO, Boffelli D, Anderson IJ, Barry KW, Shapiro HJ, *et al.*: **Symbiosis insights through metagenomic analysis of a microbial consortium.** *Nature* 2006, **443:**950.
10. Garcia Martin H, Ivanova N, Kunin V, Warnecke F, Barry KW, McHardy AC, Yeates C, He S, Salamov AA, Szeto E, *et al.*: **Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities**. *Nat Biotechnol* 2006, **24:**1263.
11. Gill SR, Pop M, DeBoy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE: **Metagenomic analysis of the human distal gut microbiome.** *Science* 2006, **312:**1355-1359.
12. Warnecke F, Luginbühl P, Ivanova N, Ghassemian M, Richardson TH, Stege JT, Djordjevic G, Aboushadi N, Sorek R, Tringe SG, *et al.*: **Functional metagenomics implicates termite hindgut bacteria as major catalysts in wood hydrolysis.** *Nature* 2007, **450:**560-565.
13. DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N-U, Martinez A, Sullivan MB, Edwards R, Brito BR, *et al.*: **Community genomics among stratified microbial assemblages in the ocean's interior.** *Science* 2006, **311:**496-503.
14. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, *et al.*: **The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific**. *PLoS Biol* 2007, **5:**e77.
15. Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, Chan AM, Haynes M, Kelley S, Liu H, *et al.*: **The marine viromes of four oceanic regions.** *PLoS Biol* 2006, **4:**e368.
16. Tyson GW, Hugenholtz P: **Environmental shotgun sequencing**. In *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics.* New York: John Wiley & Sons; 2005.
17. Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E, McHardy AC, Rigoutsos I, Salamov A, Korzeniewski F, Land M, *et al.*: **Use of simulated data sets to evaluate the fidelity of metagenomic processing methods.** *Nat Methods* 2007, **4:**495.
18. Hahn MW: **Isolation of strains belonging to the cosmopolitan *Polynucleobacter necessarius* cluster from freshwater habitats located in three climatic zones.** *Appl Environ Microbiol* 2003, **69:**5248-5254.
19. Rappe MS, Connon SA, Vergin KL, Giovannoni SJ: **Cultivation of the ubiquitous SAR11 marine bacterioplankton clade.** *Nature* 2002, **418:**630-633.
20. Tyson GW, Lo I, Baker BJ, Allen EE, Hugenholtz P, Banfield JF: **Genome-directed isolation of the key nitrogen fixer *Leptospirillum ferrodiazotrophum* sp. nov. from an acidophilic microbial community.** *Appl Environ Microbiol* 2005, **71:**6319-6324.
21. Baker BJ, Tyson GW, Webb RI, Flanagan J, Hugenholtz P, Allen EE, Banfield JF: **Lineages of acidophilic archaea revealed by community genomic analysis.** *Science* 2006, **314:**1933-1935.
22. Huber H, Hohn MJ, Rachel R, Fuchs T, Wimmer VC, Stetter KO: **A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont.** *Nature* 2002, **417:**63-67.
23. Waters E, Hohn MJ, Ahel I, Graham DE, Adams MD, Barnstead M, Beeson KY, Bibbs L, Bolanos R, Keller M, *et al.*: **The genome of *Nanoarchaeum equitans*: insights into early archaeal evolution and derived parasitism.** *Proc Natl Acad Sci USA* 2003, **100:**12984-12988.
24. **Joint Genome Institute: why sequence *Euryarchaeota* in acid mine drainage?** [http://www.jgi.doe.gov/sequencing/why/CSP2006/Euryarchaeota.html]
25. Brehm-Stecher BF, Johnson EA: **Single-cell microbiology: tools, technologies, and applications.** *Microbiol Mol Biol Rev* 2004, **68:**538-559.
26. Weibel DB, DiLuzio WR, Whitesides GM: **Microfabrication meets microbiology.** *Nat Rev Microbiol* 2007, **5:**209-218.
27. Fuchs BM, Zubkov MV, Sahm K, Burkill PH, Amann R: **Changes in community composition during dilution cultures of marine bacterioplankton as assessed by flow cytometric and molecular biological techniques.** *Environ Microbiol* 2000, **2:**191-202.
28. Robertson BR, Button DK, Koch AL: **Determination of the biomasses of small bacteria at low concentrations in a mixture of species with forward light scatter measurements by flow cytometry.** *Appl Environ Microbiol* 1998, **64:**3900-3909.

29. Sekar R, Fuchs BM, Amann R, Pernthaler J: **Flow sorting of marine bacterioplankton after fluorescence *in situ* hybridization.** *Appl Environ Microbiol* 2004, **70:**6210-6219.
30. Hosono S, Faruqi AF, Dean FB, Du Y, Sun Z, Wu X, Du J, Kingsmore SF, Egholm M, Lasken RS: **Unbiased whole-genome amplification directly from clinical samples.** *Genome Res* 2003, **13:**954-964.
31. Ottesen EA, Hong JW, Quake SR, Leadbetter JR: **Microfluidic digital PCR enables multigene analysis of individual environmental bacteria**. *Science* 2006, **314:**1464-1467.
32. Stepanauskas R, Sieracki ME: **Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time.** *Proc Natl Acad Sci USA* 2007, **104:**9052-9057.
33. Abulencia CB, Wyborski DL, Garcia JA, Podar M, Chen W, Chang SH, Chang HW, Watson D, Brodie EL, Hazen TC, *et al.*: **Environmental whole-genome amplification to access microbial populations in contaminated sediments.** *Appl Environ Microbiol* 2006, **72:**3291-3301.
34. Zhang K, Martiny AC, Reppas NB, Barry KW, Malek J, Chisholm SW, Church GM: **Sequencing genomes from single cells by polymerase cloning.** *Nat Biotechnol* 2006, **24:**680-686.
35. Marcy Y, Ouverney C, Bik EM, Losekann T, Ivanova N, Martin HG, Szeto E, Platt D, Hugenholtz P, Relman DA, *et al.*: **Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth.** *Proc Natl Acad Sci USA* 2007, **104:**11889-11894.
36. Podar M, Abulencia CB, Walcher M, Hutchison D, Zengler K, Garcia JA, Holland T, Cotton D, Hauser L, Keller M: **Targeted access to the genomes of low-abundance organisms in complex microbial communities.** *Appl Environ Microbiol* 2007, **73:**3205-3214.
37. Hutchison CA, III, Smith HO, Pfannkoch C, Venter JC: **Cell-free cloning using ϕ29 DNA polymerase.** *Proc Natl Acad Sci USA* 2005, **102:**17332-17336.
38. Völker U, Hecker M. **From genomics via proteomics to cellular physiology of the Gram-positive model organism *Bacillus subtilis*.** *Cell Microbiol* 2005, **7:**1077-1085.
39. Thompson A, Rowley G, Alston M, Danino V, Hinton JC. ***Salmonella* transcriptomics: relating regulons, stimulons and regulatory networks to the process of infection.** *Curr Opin Microbiol* 2006, **9:**109-116.
40. Poretsky RS, Bano N, Buchan A, LeCleir G, Kleikemper J, Pickering M, Pate WM, Moran MA, Hollibaugh JT: **Analysis of microbial gene transcripts in environmental samples.** *Appl Environ Microbiol* 2005, **71:**4121-4126.
41. Leininger S, Urich T, Schloter M, Schwark L, Qi J, Nicol GW, Prosser JI, Schuster SC, Schleper C: **Archaea predominate among ammonia-oxidizing prokaryotes in soils.** *Nature* 2006, **442:**806.
42. Ram RJ, VerBerkmoes NC, Thelen MP, Tyson GW, Baker BJ, Blake RC, II, Shah M, Hettich RL, Banfield JF: **Community proteomics of a natural microbial biofilm.** *Science* 2005, **308:**1915-1920.
43. Lo I, Denef VJ, VerBerkmoes NC, Shah MB, Goltsman D, DiBartolo G, Tyson GW, Allen EE, Ram RJ, Detter JC, *et al.*: **Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria.** *Nature* 2007, **446:**537-541.
44. Ibrahim Y, Tang KQ, Tolmachev AV, Shvartsburg AA, Smith RD: **Improving mass spectrometer sensitivity using a high-pressure electrodynamic ion funnel interface.** *J Am Soc Mass Spectrom* 2006, **17:**1299-1305.
45. Newman JR, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, DeRisi JL, Weissman JS: **Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise.** *Nature* 2006, **441:**840-846.
46. Johnson PLF, Slatkin M: **Inference of population genetic parameters in metagenomics: a clean look at messy data.** *Genome Res* 2006, **16:**1320-1327.
47. Whitaker RJ, Banfield JF: **Population genomics in natural microbial communities**. *Trends Ecol Evol* 2006, **21:**508-516.
48. Dumont MG, Murrell JC: **Stable isotope probing - linking microbial identity to function.** *Nat Rev Microbiol* 2005, **3:**499-504.
49. Lechene C, Hillion F, McMahon G, Benson D, Kleinfeld AM, Kampf JP, Distel DL, Luyten Y, Bonventre J, Hentschel D, *et al.*: **High-resolution quantitative imaging of mammalian and bacterial cells using stable isotope mass spectrometry.** *J Biol* 2006, **5:**20.
50. McDonald KL, Auer M: **High-pressure freezing, cellular tomography, and structural cell biology.** *Biotechniques* 2006, **41:** 137,139,141 passim.