Metagenomics at Grass Roots

Sudeshna Mazumdar-Leighton and Vivek K Choudhary

Metagenomics is a robust, interdisciplinary approach for studying microbial community composition, function, and dynamics. It typically involves a core of molecular biology, microbiology, ecology, statistics, and computational biology. Exciting outcomes anticipated from these studies include unraveling of complex interactions that characterize the ecological milieu of microbial communities. Diverse habitats from which metagenomes have been reported include human guts, caterpillar guts, thermal vents in oceans, ore deposits, polar caps, and even soils that adhere to plant roots. Knowledge generated from metagenomic projects has tremendous potential to benefit human health, agriculture, and ecosystem functions. This article provides a brief history of technical advances in metagenomics, including DNA sequencing methods, and some case studies. A specific example is provided of microbial metagenomes found at the roots of native grass species (family Poaceae) that can grow on degraded lands undergoing revegetation.

Sudeshna Mazumdar-Leighton is a molecular biologist at Delhi University. Her research interests are plant-biotic environment interactions, relevant to marginalized communities.

Vivek K Choudhary has just completed a PhD dissertation on restoration ecology of iron-ore mine sites in Odisha. He is interested in plant community dynamics, revegetation, and soil metagenomics.

Keywords
Microbial ecology, next-generation sequencing, databases, rhizospheric communities.

The term 'metagenomics' was coined in 1998 by Jo Handelsman and her colleagues from University of Wisconsin, Madison and Cornell University, New York, USA [1]. It referred to the simultaneous characterization of genomes of soil microbial communities, using techniques like cloning and DNA sequencing for the general purpose of identifying new sources of medicines and chemicals. Metagenomics is today one of the most prominent areas of research in the life sciences. Etymology of the word 'metagenomics' can be split into 'meta' and 'genomics', where 'meta' refers to the combination of a group or large amount of data, while 'genomics' refers to information gleamed from the genome (total nucleic acid content) of an organism. Metagenomics is used to study not just one organism, but rather multiple microbes (bacte-

ria, fungi, archaea, viruses, and/or micro-eukaryotes) that inhabit any given habitat. Members of microbial communities usually occupy different trophic levels in food webs. They may be primary producers, consumers or decomposers. Hence, metagenomics of microbial communities involves (i) identification of the panoply of abundant and rare taxa, and (ii) predicting the intertwined roles of each component in utilizing its environment.

The first metagenomic study [2] was conducted in 2003 and revealed a tremendous diversity in marine viruses, out of which 80% were previously unreported. Similar findings were obtained recently in 2014, from studies with improved computational methods for analyses of sequence data. These studies addressed important questions like the role of climate change, season, depth of oceans, and distance from the coast on population fluxes of viruses, especially bacteriophages. This is significant as phages regulate the numbers and types of their bacterial hosts. Bacteria, in turn, can sequester carbon and thus, are very important components of oceanic ecosystems. Other classic studies have provided novel insights into the roles of bacteria in biogeochemical cycles in extreme environments like acid mine drainages and the Sargasso sea [3]. An important (and obvious) application of metagenomics has been the ongoing study of microbial communities (or microbiomes) in humans. The Human Microbiome Project (HMP) (www.hmpdacc.org) is an attempt at providing data from various projects on human populations. HMP began in 2008 and is still running.

The first metagenomic study was conducted in 2003 and revealed a tremendous diversity in marine viruses, out of which 80% were previously unreported. Similar findings were obtained recently in 2014, from studies with improved computational methods for analyses of sequence data.

'Shotgun' Metagenomics for Deciphering Microbial Communities

It was realized in the mid 1990s that a vast majority of microbes could not be cultured on artificial growth media traditionally used in labs. An estimate of the number of uncultivable bacteria [3] recalcitrant to laboratory culture in one gram of soil was 109. Hence, a straightforward approach to study microbial communities was to isolate total DNA from environmental samples, and

randomly clone DNA fragments into vectors (like bacterial artificial chromosomes, or BACs, that can accommodate 30kbp inserts). This process was theatrically called 'shotgun cloning'. Sequencing of insert DNA from such metagenomic/clonal libraries was used to generate overlapping sequence reads that identified contiguous regions of a genome (also called 'contigs'). In addition, molecular techniques like Polymerase Chain Reaction (PCR) amplification of 'markers' like ribosomal RNA (rRNA) genes and genes that encoded families of functionally important proteins (e.g., nif, recA, EF-Tu, etc.) provided 'amplicons' that could be sequenced. Sequence analyses provided a rapid estimate of the different types/identities of microbes in a given sample.

Another approach was to clone random fragments of DNA isolated from microbial communities into expression vectors that could be shuttled into various hosts (like *Escherichia coli*) for the production of recombinant proteins. Subsequently, clones would be identified from the expression of a desired phenotype [3]. However, this technique was rather cumbersome and relied on fortuitous cloning of complete genes and/or operons that could produce a product. Nevertheless, functional characterization of proteins expressed from metagenomic libraries yielded novel antibiotics, enzymes, and hydrolases [3]. Early rewards for metagenomic approaches came with the discovery of novel archaea like Crenarchaeota that participate in the biogeochemical cycles by oxidizing ammonia, phototrophic marine bacteria that use novel rhodopsins as light sensors, as well as the recognition of unculturable Acidobacteria as major components of agricultural soils.

Early rewards for metagenomic approaches came with the discovery of novel archaea like Crenarchaeota that participate in the biogeochemical cycles by oxidizing ammonia, phototrophic marine bacteria that use novel rhodopsins as light sensors, as well as the recognition of unculturable Acidobacteria as major components of agricultural soils.

Metagenomics and Next-Generation Sequencing

The ability to rapidly and cheaply obtain vast amounts of genome sequence information is at the crux of metagenomic studies. The main explosion of metagenomic data came from the identification of microbes in environmental samples using rRNA gene amplicons (16S rRNA genes for prokaryotes and 18S rRNA genes for eukaryotes). Early studies used pyrosequencing and microar-

The ability to rapidly and cheaply obtain vast amounts of genome sequence information is at the crux of metagenomic studies.

rays to identify microbes. The length of DNA reads obtained from pyrosequencing methods are 400–500bp as opposed to 600–800bp reads typically obtained through Sanger sequencing. Even though the reaction run time is similar in both cases, pyrosequencing has no cloning bias and gives 100-fold higher yields of up to 5Mbp [4]. It is also cheaper than Sanger sequencing. However, the inability to read through homo-polymeric stretches of DNA is a major disadvantage resulting in sequencing errors. In the case of microarrays, single-stranded DNA probes are synthesized that are specific to different microbial taxa and protein-encoding genes. These molecules are immobilized on a solid support and hybridized with DNA or RNA isolated from an environmental sample. However, the results obtained are limited to the array, and do not reflect new microbial taxa or novel genes that may be present in the sample.

Subsequent NGS methods generate high volumes of data as short reads ranging from 25–300bp. ABI SOLiD, Illumina GAII, HiSeq-2000, and MiSeq are popular sequencing platforms [4] that can provide adequate coverage of metagenomic libraries. This is because the likelihood of a target DNA being sequenced multiple times per run is very high. Metagenomic DNA fragments that are sequenced can vary in size ranging from about 800bp to 5kbp (depending upon the choice of paired-end or mate-paired libraries). Hence, assembly of results from these platforms into coherent, longer sequence reads (representing any one microbial genome) requires substantial computational effort. Among these platforms, SOLiD is the most error-free and isused in applications requiring re-sequencing a template. An acronym for 'Sequencing by Oligonucleotide Ligation and Detection', this platform is based on the chain ligation method [4]. The Illumina platforms use reversible dye terminators and a 'sequencing by synthesis approach [4]. Other NGS platforms available include the Ion Torrent and PacBio. The Ion Torrent [4] is based on the principle that protons generated during polymerization of DNA molecules can be used to detect nucleotides being incorporated on a template DNA. It produces read lengths of at least 100bp. PacBio

can generate the longest read lengths among all the NGS platforms currently available [4]. It uses DNA isolated from one cell and a single molecule, real-time (SMRT) detection system. However, it is error-prone and under further development. Another method based on single cells and DNA nanopore sequencing is also available. It is particularly suited to resolving repetitive DNA sequences. All these new sequencing methods are free from cloning bias, and some do not involve PCR amplifications (hence can accurately represent abundance of constituent microbes in a given environmental sample). NGS methods have made it possible to sequence hundreds of genomes in a day and tens of microbes in one run. The promise of commercial applications in human medicine like screening for pathogenic microbes, detection of disease-associated sequence motifs (e.g., single nucleotide polymorphisms and haplotypes) has led various companies to develop benchtop versions of next-generation sequencers [4]. It is now possible to use multiple platforms (also called hybrid NGS) to generate very high quality, complete metagenomic data. Most significantly, competition among these companies have driven prices down and enhanced the deliverables available to research laboratories.

Metagenomics and Computational Biology

Analyses of metagenomic data typically involve (i) assembly of good quality sequence reads, (ii) binning the sequences into related groups, and (iii) gene annotation of open reading frames. Metagenomic datasets can subsequently be examined by multivariate statistical methods commonly used in ecology to decipher patterns and/or obtain estimates of biodiversity. Assembly of sequence reads into 'scaffolds' (containing contigs) of linear information from one constituent microbial genome can be achieved *de novo*, or with the help of a previously characterized, related 'reference' genome. While the former approach is computationally expensive and can take days, both assemblies require well-curated public domain sequence repositories. After assembly, the sequences are binned into related groups or clusters based upon

NGS methods have made it possible to sequence hundreds of genomes in a day and tens of microbes in one run.

extent of similarity. Assembled and binned contigs are used to search various sequence databases for microbial identity. In certain cases, short reads that span variable domains within highly conserved genes (e.g., rRNA genes) are useful for direct identification of operational taxonomic units (OTUs) in metagenomic samples. Tentative answers become available to questions like – "Who are the constituent microbes in a given environmental samples?" Functional annotations can predict metabolic pathways responsible for physiological processes or adaptations that enable microbes to inhabit a particular environment. Answers become available for the question – "What do microbial communities do in the environment from which they were sampled?"

Launched in 2007–8, MG-RAST, IMG/M, and CAMERA [4] are popular packages that accomplish assembly, binning, and annotation of metagenomic datasets. The MG-RAST (Metagenomics Rapid Annotation Subsystem Technology) pipeline can process pyrosequencing data or short reads of just 75bp and is user-friendly. The IMG/M (Integrated Microbial Genomes and Metagenomes) pipeline is an excellent option for functional annotations, as it can search various databases using BLAST (Basic Local Alignment Search Tool) and profile HMM (Hidden Markov Models) [4]. Both MG-RAST and IMG/M also store metagenomic datasets. CAMERA (Community Cyber-infrastructure for Advanced Microbial Ecology Research and Analysis) is another pipeline that allows metadata (e.g., sampling details, environmental information, etc.) to be included in downstream sequence analyses. In addition to the nonredundant sequence database available at NCBI (http://www.ncbi.nlm.nih.gov/blast-nr), some wellannotated microbial databases include RDP (Ribosomal Database Project, www.rdp.cme.msu.edu), GreenGenes (www.greengenes.lbl.gov), SILVA (http://www.silva-arb.de), AFTOL (Assembling the Fungal Tree of Life, www.aftol.org), and VIROME (www.virome.dbl.udel.edu). These databases are generally devoid of sequencing artifacts like recombinant molecules and chimeras produced as a result of PCR. Some dedicated websites used routinely for functional annotations [4] in the past decade are KEGG (Kyoto Ency-

clopedia of Genes and Genomes (www.genome.jp/kegg), eggNOG (www.egg- nogdbase.embl.de), COGs (Clusters of Orthologous Groups in prokaryotes and eukaryotes, www.ncbi.nlm.nih.gov/COG), and InterPro (www.ebi.ac.uk/interpro). The presence or absence of a particular OTU (or groups of OTUs), associated physiological/biochemical processes, and metabolic pathways in a particular metagenome dataset is indicated from BLAST search results of the NCBI-NR database using threading algorithms or HMM searches of PFAM (www.pfam.xpam.org) databases [4].

Comparison of metagenome datasets can address questions like – "How similar are microbial communities in any given number of environmental samples, and what components of the environment influence microbial communities?" Results from such comparisons can be easily visualized as ordination graph plots of principal component analysis, correspondence analysis, or non-parametric multidimensional scaling. Various open source computational packages are available for assessment of microbial diversity from OTU data representing multiple environmental samples, like QIIME (Quantitative Insights into Microbial Ecology, www.qiime.org), MOTHUR (www.mothur.org), and MEGAN (MEtaGenome ANalyser, www.ab.informatik.uni-tuebingen.de/software/megan6). Similar analyses can also be performed (to a limited extent) using MG-RAST, IMG/M, and CAMERA mentioned earlier. Multivariate statistical packages commonly used to measure populations and communities of plants and animals in terrestrial biodiversity studies like EstimateS (http://viceroy.eeb.uconn.edu/estimates/), CANOCO (www.canoco.com), and PRIMER-E (www.primer-e.com) have also been used to describe microbial communities from environmental samples.

Metagenomics of Rhizospheric Soils

Soil is one of the most complex environments on Earth. Metagenomics of soil biota remains very challenging due to the sheer numbers and diversity of constituent archaea, bacteria, fungi, viruses, arthropods, protozoa, nematodes, rotifers, and/or worms. Fur-

Identification of rhizospheric microbial communities that tolerate contaminants and promote transformations that render the soil less toxic, are invaluable for phytoremediation and restoration of degraded lands.

thermore, the composition and distribution of soil biota changes rapidly over spatial and temporal scales [5]. The 'rhizosphere' is a unique zone in the soil immediately proximal to a plant root, extending from the rhizoplane onto a few centimeters of the surrounding soil. It does not include the interior living cells of roots or the 'endophytic zone'. Plant roots produce exudates containing organic carbons that attract a cohort of microbes [5] from the nearby bulk soil that can potentially provide nutrition, minerals, and suppress diseases. Endophytic microbes within any plant are usually a subset of the rhizospheric microbial communities. Rhizospheres of grasses are extremely rich in microbes and soil fauna. It is estimated that while a human gut metagenome may need a million reads on a pyrosequencing machine to identify all the microbes within a sample, unearthing the microbes within a grass root metagenome can need tens of billions of reads [6]. Less is known about rhizospheric metagenomes of grasses that phytostabilize metal-contaminated soils, and sequester toxic metals. Identification of rhizospheric microbial communities that tolerate contaminants and promote transformations that render the soil less toxic, are invaluable for phytoremediation and restoration of degraded lands.

An Example of Bacterial Communities in Grass Rhizospheres

Indian grasses like *Thysanolaena latifolia* (Broom grass) and *Eleusine indica* (Goose grass) can restore degraded sites like overburdened waste dumps from iron ore mines (*Figure* 1). Rhizospheric soil biota of these grasses are implicated in amelioration of metal contaminants like iron, manganese, chromium, and zinc, increase in soil pH, formation of organic matter and ultimately, establishment of aboveground plant communities [8]. Their rhizospheric soil metagenomes contain diverse communities of archaea, bacteria, fungi, and micro-eukaryotes (of which only bacteria are shown in *Figure* 1E). *Figure* 1E shows that proteobacteria were prominent in the grass rhizospheres, albeit at different abundances. Functional annotations of these bacteria suggested presence of metabolic pathways involved in uptake of

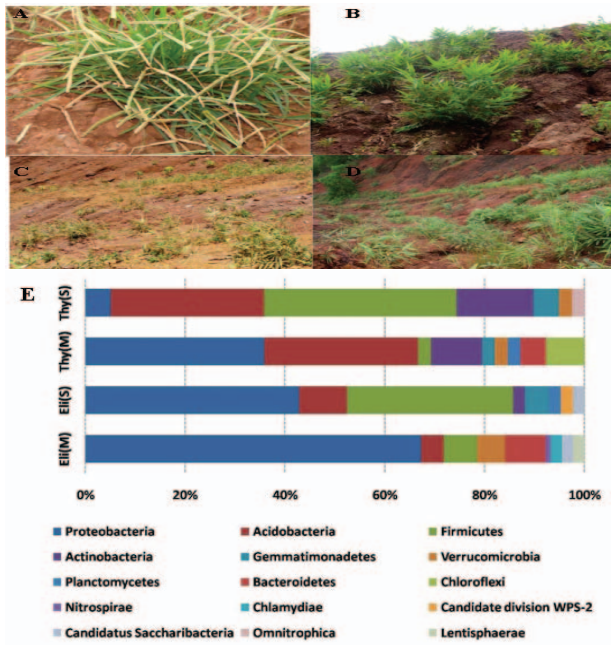

Figure 1. Clumps of grasses of (**A**) *Eleusine indica* and (**B**) *Thysanolaena latifolia* growing on slopes of iron-ore mine dump sites undergoing restoration in (**C**) summer, 2011 and (**D**) monsoon, 2011. (**E**) Relative abundance of prominent phyla of rhizospheric bacterial communities of *T. latifolia* (Thy) and *E. indica* (Eli) sampled in the summer (S) and monsoon (M) from degraded mined-out sites from C and D. Rare bacteria (singletons, <2% of total bacteria) are not shown.

amino acids, breakdown of root exudates, and the ability to form biofilms. Various proteobacteria in the two grass rhizospheres resembled iron oxidizing bacteria (FeOB) with genes like *PioA cytochromes* that act as iron oxidases, pumping electrons to the periplasm. Acidobacteria known to metabolize a variety of carbon sources, break down complex plant carbohydrates, and withstand water stress were also present. Nitrospirae or nitrite oxidizing bacteria (NOB) co-occurred with ammonia oxidizing Thaumarchaeota, indicating a capacity to complete the nitrogen cycle in these metal-contaminated soils. Firmicutes like bacilli were predominant in the rhizospheres of both grasses in the summer. These bacteria are known as 'multi-taskers' enabling plant roots to withstand toxic metals and water stress. Bacteroidetes like *Flavobacterium* were observed in the monsoon season. These bacteria can promote plant growth by solubilizing rock phosphates and secreting phytohormones like auxins. Other prominent rhizospheric bacteria belonging to Choloroflexi, Gemmatimonadetes and Planctomycetes were present. These bacteria can utilize different sugars and adapt to variations in soil moisture, organic mat-

ter, and pH. All bacterial sequences from the two grass rhizospheres could be assigned to putative taxonomic phyla, including unassigned groups like *Candidatus saccharibacteria*. However, their unequivocal assignment to specific genera and/or species was often difficult. Several OTUs represented rare bacteria of low abundance. A high degree of intraspecific variation was also evident in certain bacterial lineages. While exact implications need further investigation, these results resemble reports [5] on bacteria from forest soils. *Figure* 1E also shows seasonal shifts in abundance of rhizospheric bacterial communities of the two grasses, adding yet another layer of complexity. Similar trends were evident from fungal and faunal OTUs identified in these grass rhizospheres, where many OTUs had no significant match in the sequence databases, and could not be classified taxonomically. This untamed frontier in microbiology was aptly described in a recent publication [7] by 48 scientists entitled 'Back to the future of soil metagenomics':

A comprehensive catalog of soil microorganisms and functional genes does not yet exist for any soil. We still do not know the extent of what we do not know. There are more than a million times as many soil microorganisms on our planet than stars in the universe and we argue that the time has come for humans to tackle the challenge of soil microbial diversity.

Conclusion

As mentioned in the seminal paper by Handelsman *et al*., 1998, potential applications of metagenomic projects are truly remarkable. Ecological restorations of degraded mined-out areas involve amelioration from previously metal-contaminated and unhealthy soil conditions, towards reference healthy forest soil conditions. Changes in structure of soil bacterial and fungal populations, as well as establishment of trophic linkages with soil fauna/microeukaryotes are clearly implicated [5, 8]. Rhizospheric metagenomes of grasses from sites undergoing revegetation can identify 'turnkey' belowground ecological processes that control ecosystem devel-

opment at degraded mined-out sites. Further work is necessary to decipher and compare rhizospheric metageno- mes of grasses and other plant types, especially legumes that characterize complex changes in aboveground plant communities at sites undergoing restoration.

Acknowledgements

Professor Emeritus C R Babu, University of Delhi, is thanked for the suggestion to explore soil metagenomics at degraded sites undergoing restoration. PhD research scholars, Ms. Parul Bhardwaj and Ms. Aashima Mehra made valuable suggestions to improve the manuscript. The reviewer is thanked for a very helpful critique of the article.

Suggested Reading

[1] J Handelsman, M R Rondon, S F Brady, J Clardy, and R M Goodman, Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products, *Chemistry and Biology*, Vol.5, No.10, pp.R245–249, 1998.

[2] P Hugenholtz and G W Tyson, Metagenomics, *Nature*, Vol.455, pp.481–483, 2008.

[3] P D Schloss and J Handelsman, Biotechnological prospects from metagenomics, Current Opinion in Biotechnology, Vol.14, pp.303–310, 2003.

[4] T Thomas, J Gilbert and F Meyer, Metagenomicsa guide from sampling to data analysis, Microbial Informatics and Experimentation, Vol.2, No.3, 2012. doi:10.1186/2042-5783-2-3.

[5] R D Bardgett and D A Wardle, Aboveground-Below-ground linkages: Biotic Interactions, Ecosystem processes and Global change, Chapter 3, pp.62–110, 2010.

[6] A C Howe, J K Jansson, S A Malfatti, S G Tringe, J M Tiedje and C T Brown, Tackling soil diversity with the assembly of large, complex metagenomes, *Proceeding of the National Academy of Sciences* (USA), Vol.111, pp.4904–4909, 2014.

[7] J Nesme et al., Back to the future of soil metagenomics, Frontiers in Microbiology, Vol.7, No.73, 2016.

[8] V K Choudhary, Microbial diversity at Iron ore mined-out sites in Odisha and its significance in Ecological Restoration, PhD thesis, University of Delhi, 2016.

Address for Correspondence
Sudeshna Mazumdar-Leighton* and Vivek K Choudhary
Plant-Biotic Interaction Lab,
Department of Botany
Chhattra Marg
University of Delhi
Delhi 110 007, India.
Email:
*smazumdar@botany.du.ac.in
vivekchy007@gmail.com