Biotechnology and Genetic Engineering Reviews - Vol. 26, 335-352 (2009)

Functional Metagenomics: Recent Advances and Future Challenges

LUDMILA CHISTOSERDOVA*

Department of Chemical Engineering, University of Washington, Seattle WA 98195, USA

Abstract

Metagenomics is a relatively new but fast growing field within environmental biology directed at obtaining knowledge on genomes of environmental microbes as well as of entire microbial communities. With the sequencing technologies improving steadily, generating large amounts of sequence is becoming routine. However, it remains difficult to connect specific microbial phyla to specific functions in the environment. A number of 'functional metagenomics' approaches have been implemented in the recent years that allow high-resolution genomic analysis of uncultivated microbes, connecting them to specific functions in the environment. These include analysis of niche-specialized low complexity communities, reactor enrichments, and the use labeling technologies. Metatranscriptomics and metaproteomics are the newest sub-disciplines within the metagenomics field that provide further levels of resolution for functional analysis of uncultivated microbes and communities. The recent emergence of new (next generation) sequencing technologies, resulting in higher sequence output and dramatic drop in the price of sequencing, will be defining a new era in metagenomics. At this time the sequencing effort will be taken to a new level to allow addressing new, previously unattainable biological questions as well as accelerating genome-based discovery for medical and biotechnological applications.

*To whom correspondence may be addressed (milachis@u.washington.edu)

Abbreviations: WGS, Whole-Genome Shotgun; bp, base pairs; Mb, 1,000,000 base pairs; Gb, 1,000,000,000 base pairs.

Introduction

Metagenomics is a fast growing and diverse field within environmental biology directed at obtaining knowledge on genomes of environmental microbes, without prior cultivation, as well as of entire microbial communities. Other terms are also used to describe this methodology: environmental genomics, ecogenimics, community genomics, megagenomis. For the purposes of this review, all these terms are interchangeable. The term 'functional metagenomics', in a broad sense, is meant to reflect a connection between the identity of a microbe, or a community, uncovered via metagenomics and their respective function(s) in the environment. The power of metagenomics is in allowing one to tap into the vast metabolic potential of uncultivated microbes that represent the majority of microbes on Earth, including entirely novel microbes and novel metabolic pathways. The two main and principally distinct outcomes of the metagenomic approach are the emerging new outlook at the complexity of microbial communities, in terms of both species diversity and community dynamics, and identification of genetic determinants for production of biologically active molecules and processes that carry a potential for medical and biotechnologcial applications. Recent advances in next-generation (ultra-high throughput) sequencing technologies, resulting in a dramatic drop in the price of DNA sequencing, bring about the promise of a new era in metagenomics, when the sequencing effort will be taken to a new level, to allow addressing new, previously unattainable biological questions, as well as accelerating genome-based discovery for biotechnological applications.

Despite its short history, metagenomics has recently become a mainstream approach in every field under the umbrella of biological sciences, and the term itself became a household name. As a reflection of the community interest, the already profound impact, and the future potential of metagenomics, the field is a subject of intense discussion, resulting in a steady stream of reviews covering every topic in the field, from strategies and methodologies of metagenomics (Handelsman 2004; Schloss and Handelsman, 2005; Snyder et al., 2009), to industrial and medical applications (Warnecke and Hess, 2009; Li et al., 2009; Preidis and Versalovic, 2009), to the bioinformatics issues (Markovitz et al., 2008; Kunin et al., 2008; Lapidus 2009). By no means does this review intend to comprehensively cover all the work conducted in the past that involved metagenomic or functional metagenomic approaches. Instead, it will pursue two major goals: highlighting the recent advances specifically addressing a connection between the function and phylogeny in the environment, via high resolution metagenomics, metatranscriptomics and metaproteomics, and offering an outlook at the future of metagenomics that will rely on the newest sequencing technologies as well as on considerably evolved data management infrastructures.

Brief history of metagenomics

As genomics, starting in mid-nineties, has revolutionized the entire range of biological sciences, so did metagenomics a decade or so later. However, the history of metagenomics should probably be traced back to the work of Staley and Konopka (1985), first reporting on 'great plate count anomaly', and the works of the Woese group, identifying the 16S rRNA gene as a marker molecule for assessing microbial diversity (Woese

1987). The practical use of 16S rRNA analysis as a tool for phylogenetic profiling of microbial communities has been pioneered by the Pace group in early nineties (Schmidt et al., 1991), constituting the onset of metagenomics as a sub-field of microbial ecology, albeit without an official name. The term 'metagenomics' was coined almost a decade later by the Handelsman group (Handelsman et al., 1998) and was instantly embraced by the scientific community. In this latter case, the term referred to the functional analysis of mixed environmental DNA captured as large size inserts after a screen for a specific activity, thus also providing one meaning to 'functional metagenomics'. The two seminal works that defined the most widely accepted meaning of metagenomics, the random whole-genome shotgun (WGS) sequencing-based analysis of microbial populations, were published in quick succession in 2004. One describes analysis of an artificially simple community, of a biofilm growing on the surface of an acid mine drainage (Tyson et al., 2004), and the other describes a much more complex community of the Sargasso Sea (Venter at al., 2004). The significance of these two early studies is two-fold. On the one hand, they ultimately defined the path for future metagenomic projects. On the other, they provided important insights into the scale of the sequencing effort that would be required for analyzing communal DNA using this method, spanning a range of scenarios from very simple to very complex. Accordingly, the outcomes of these projects regarding the knowledge on specific members of the communities interrogated were dramatically different. The former (with only 76 Mb sequencing effort) resulted in assembly and analysis of almost complete genomes of the dominant species, including accurate metabolic reconstruction and detection of strain-specific genomic variants. The latter, with a much larger sequencing effort (almost 2 Gb) resulted in very fragmented assemblies even for the most abundant species, with most of the dataset being represented by singleton sequencing reads. The stage has been set for a flood of WGS-sequencing-based projects to follow. At the moment of writing, 167 projects are listed in the GOLD database (Liolios et al., 2008), and results from 57 of these have been already published.

Over the same time, breakthroughs in developing alternative sequencing technologies occurred, promising a significantly higher throughput and significantly reduced cost of sequencing, and these new (known as next-generation) sequencing technologies have been immediately tested in metagenomic applications (Edwards et al., 2006). These new sequencing technologies have also enabled a new subfield of metagenomics, termed metatranscriptomics, (i.e. shotgun characterization of environmental transcripts) that is developing incredibly quickly. Another subfield, metaproteomics (i.e. analysis of community protein pools) is emerging, signaling the arrival of a whole new era of metagenomics.

Current state of metagenomics (and how it can be improved)

It has been just over five years since the onset of WGS-sequencing-based metageomics. How has the field progressed towards its maturity and how has it changed since the seminal works published in 2004? About a decade ago, while discussing the fate of microbial genomics, Dr. Woese has lamented the fact that microbiologists have never developed an appropriate, overarching concept for the field, the need for which he thought critical (Woese 1998). The same can be said today about metagenomics.

For the past five years, the field has mostly operated in a 'Wild West' fashion, with no concerted effort, little broad-scale coordination, and in the absence of established 'gold standards'. Individual scientists have been spearheading individual metagenomic projects, mostly on a small scale, as dictated by the economics of sequencing, often times without prior knowledge on the structure or the complexity of the community in question. However, the complexity of natural communities as a challenge to metagenomics has been recognized from the very start. Venter and colleagues (2004) have modeled a sequence coverage level that would be required to identify most of the genomes in the Sargasso Sea sample, concluding that at least an order of magnitude larger sequencing effort was necessary. Another early metagenomic project that interrogated an even more complex community, the one inhabiting soil, resulted in a similar conclusion, admitting significant under-sampling (Tringe et al., 2005). However, the effort toward significantly deeper sampling was deemed not feasible and thus not necessary. As a result, most of the metagenomic projects carried out so far have been following the path of under-sampling, the disregard for community complexity validated by the broad use of a method called gene-centric analysis (Tringe et al., 2005). This method (a poor man's approach to metagenomics) treats a community (mostly represented by singleton reads, frequently of poor quality) as an aggregate, ignoring the context of individual species. Each read is automatically assigned to a functional category, and this way functional profiles of communities can be created. Communities then can be compared to each other in terms of functional profiles (Tringe et al., 2005). This approach performs rather well with singleton sequencing reads generated by the Sanger technology, with approximately 90% of genes being found to encode at least one and sometimes two putative polypeptides. However, the resolution of this approach drops further when it comes to functional gene annotation. This task relies on the content of the current gene and protein databases, which are heavily biased toward model organisms that do not fully represent the diversity of the organisms in the environment (Woese 1998). The problem of annotation of environmental genes persists beyond the lack of close homologs for the genes represented in metagenomic databases. In databases most frequently used to aid in annotation of metagenomic sequences (such as the non-redundant NCBI database, the SEED database), many of the specialized biochemical pathways are poorly annotated. Thus, even if close homologs are present, their most likely functions may be called incorrectly. Even if the functions of genes can be predicted with precision based on a homolog match, placing them into the context of specific metabolic pathways is not always possible with the gene-centric approach and out of the context of a metabiolic make up of an individual organism. For example, the citric acid cycle (whose main role is in energy generation) and the methylcitric acid cycle (whose main role is in propionate utilization) share a number of genes and enzymes in common. The precise differentiation between the two pathways can only be achieved having the knowledge on an entire or almost entire gene complement, and in the context of an individual genome.

The problems described above are even more severe when it comes to the analysis and annotation of the shorter reads currently produced by the 454 sequencing technology (Wommack et al., 2008). In the works published so far, mostly based on the earlier versions of the 454 technology that produced 100 to 200 bp reads, up to 70 per cent of reads could not be classified (Edwards et al., 2006; Dinsdale et al., 2008; Brulc et al., 2009; Thurber et al., 2009). Ironically, despite the hype about the significantly higher throughput and lower cost of 454 sequencing (compared to the Sanger-based sequencing), the metagenomic projects employing this technology tend to produce datasets of relatively small size, deeming the functional profiling and comparative genomics tasks almost useless exercises when it comes to communities of high complexity.

How can this situation be improved? One obvious way is to commeasure the sequencing effort with the complexity of the community. Predictions can be easily made for how much raw sequence is required to obtain good coverage for dominant species (Kunin et al., 2008). If the goal of the study is to obtain insights into the genomes of the minor members of the community, creative approaches such as specific enrichment strategies can be applied. Such approaches, guided by a specific goal, should enable functional insights into the community as a whole or into specific members of the community, resulting in a more complete and meaningful interpretation of the sequence data. This in turn will enable a high-resolution biological knowledge that constitutes functional genomics in a broader sense.

Approaches to functional metagenomics

One of the simple questions typically asked via metagenomics is "Who is there?". Phylogenetic profiling via metagenomics is straightforward as the 16S (or 18S) rRNA genes are easily recognizable and the growing databases of these genes allow for rather precise phylogenetic assignments (Tringe and Hugenholtz, 2008). The more difficult question asked via metagenomics is "What are they doing?". This question can be approached via the gene-centric analysis as a number of functional genes that are hallmarks for major biogeochemical processes are well recognizable (Tringe et al., 2005; Kunin et al., 2008). The most difficult question asked via metagenomics is "Who is doing what?" as this requires establishing a connection between the function and an organism. The resolution of such knowledge can range from as low as a phylum level to as high as strain level. Below, a number of exemplary projects that succeeded in this goal are highlighted. Examples of single species genomes assembled from metagenomic data are presented in Table 1.

METAGENOMICS OF LOW COMPLEXITY COMMUNITIES

The acid main drainage (AMD) community remains the poster child of metagenomics (Tyson et al., 2004). A modest sequencing effort of 76 Mb has been sufficient for high coverage of the genomes of the dominant species of this low complexity community. Two main lessons learned from assembling separate genomes from a communal sequence pool have been the use of relaxed stringency criteria for sequence alignment, in anticipation of polymorphisms, and the necessity of binning, i.e. assignment of scaffolds to organism types (in this case by the G+C content as well as read depth). As a result, nearly complete genomes (at 10X coverage) were assembled of a *Leptospirillum* group II bacterium and a *Ferroplasma* group II archaeon. Two other genomes (*Leptospirillum* group III and *Ferroplasma* group I) have been covered at 3X. Reconstruction of the metabolism of these species has provided important insights into their ecological roles and the function of the community. Both the *Leptospirillum*

Table 1. Examples of single species genomes extracted from metagenomic sequences

Organism/ environment	Sequencing effort (Mb)	Number of contigs/ scaffolds	Largest contig/ scaffold (Mb)	Sequence coverage (x)	Genome size (Mb)	Reference
C. Cloacamonas acidiminovorans/ reactor	1,120	1	2,246	10.9	2.2	Pelletier et al., 2008
Cenarchaeum symbiosum/ sponge	50	1	2,045	>8	2.0	Hallam et al., 2006
Kuenenia stuttgartensis/ reactor	150	5	2,200	22	4.2	Strous et al., 2006
Leptospirillum sp. group II/ AMD	76	70	137	10	2.2	Tyson et al., 2004
Ferroplasma sp. group II/ AMD	76	59	138	10	1.8	Tyson et al., 2004
Gammaproteobacterium 3/ worm	204	22	1,908	5.2	NA	Woyke et al., 2006
Gammaproteobacterium 1/ worm	204	91	333	3.0	NA	Woyke et al., 2006
Deltaproteobacterium 4/ worm	204	172	252	3.3	NA	Woyke et al., 2006
Deltaproteobacterium 1/ worm	204	226	407	8.4	NA	Woyke et al., 2006
C. Accumulibacter phosphatis/ US sludge	98	33	3,027	8	5.6	García Martín et al., 2006
C. Accumulibacter phosphatis/ OZ sludge	78	254	69	5	5.6	García Martín et al., 2006
Methylotenera mobilis/ lake	60	4078	16	4.4	2.5	Kalyuzhnaya et al., 2008

NA, information not available

and the *Ferroplasma* species appear to possess multiple pathways for carbon fixation, while the *Ferroplasma* species are also equipped for a heterotrophic life style. However, none of the major species are predicted to be able to fix nitrogen, suggesting that a less abundant species, *Leptospirillum* group III is responsible for this activity. A detailed reconstruction of putative electron transfer chains in the *Leptospirillum* and the *Ferroplasma* species has revealed markedly different strategies for harnessing energy from iron oxidation.

Another example of a low complexity, highly specialized community metagenomics is the analysis of symbionts of a marine oligochaete *Olavius algarvensis* (Woyke et al., 2006). In the course of evolution, this worm has lost a mouth, a gut and nephridia, their functions completely taken over by the four symbiotic bacteria, two gammaproteobacterial types and two deltaproteobacterial types. Nearly complete genomes of these four symbionts have been recovered in scaffolds as large as 1.6 Mb, via a sequencing effort of slightly over 200 Mb, allowing for metabolic reconstruction of each species as well as predicting the specific roles for each bacterium in the syntrophic elemental cycling providing metabolic energy and biomass that feed the host. Two species (the gammaproteobacterial strains) were determined to specialize in oxidation of reduced sulphur compounds, while two others (the deltaproteobacterial srains) were predicted to be sulphate reducers. All four species are equipped for autotrophic CO_2 fixation, two former possessing genes for the Calvin-Benson-Bassham cycle and the two latter possessing genes for the reductive acetyl-CoA and the reductive tricarboxylic acid cycles. A suite of metabolic pathways that would enable recycling of the host waste has been identified, including ammonium and urea uptake and metabolism systems. In addition, systems were identified for degradation of osmolytes such as taurine, glycine betaine and trimethylamine N-oxide. The nearly complete genome recovery for the symbionts has also allowed to question whether signatures of symbiont dependence on the host were present, such as genome size reduction, loss of essential metabolic pathways etc. None of these have been found, suggesting that the symbionts of *O. algarvensis* should be able to survive as free-living bacteria. Finally, a scenario has been envisioned of how the symbionts switch between different pathways for energy generation while the host migrates between oxic and unoxic zones in its natural habitat.

In another study, the WGS sequencing approach has been combined with fosmid walking to obtain a complete genome sequence for an uncultivated marine crenarchaeote, a sponge symbiont (Hallam et al., 2006). Based on the initial 2, 779 fosmid end reads, fosmids have been selected containing DNA of the dominant ribotype of *Cenarchaeum symbiosum*, and eventually 155 of the fosmid inserts have been shotgun sequenced, to produce a circular genome representing a composite of closely related but potentially non-identical strains. Metabolic reconstruction revealed important insights into the metabolic pathways of the symbiont that make it useful for the host, such as nitrogen metabolism and CO_2 assimilation, and also revealed metabolic dependence of the symbiont on the supply of essential amino acids and vitamins. While the strategy chosen in this project, of fosmid walking, required a smaller sequencing effort compared to the projects relying on the assembly of shotgun reads, it was a laborious and time consuming effort, taking over four years.

A termite gut is a habitat for a highly specialized microbial community devoted to lignocellulose degradation. Warnecke et al. (2007) targeted a community from a hindgut of a wood-feeding 'higher' termite in order to obtain insights into the diversity of genes enabling cellulose and xylan hydrolysis by the bacterial symbionts. While only a modest sequencing effort was committed (71 Mb) considering the relatively species-rich community, a gene-centric approach resulted in identification of more than 700 glycoside hydrolase catalytic domains, representing 45 different carbohydrate-active enzymes. Compositional binning of the assembled sequences allowed assignment of glycoside hydrolases and other carbohydrate-binding module enzymes to the specific phylogenetic groups, most prominently to *Treponema* and to *Fibrobacter* species. The knowledge gained from metagenomic sequencing was augmented by the proteomic analysis that detected some of the most highly expressed hydrolytic enzymes, as well as by *in vitro* activity tests.

ENRICHMENT-BASED METAGENOMICS

Some organisms not available in pure cultures can be enriched in laboratory conditions, or in industrial bioreactors, to reach a relative proportion in mixed population large enough to warrant high coverage for the respective genome(s) when using the traditional WGS sequencing approach. One study applied this strategy to sequence, assemble and annotate the genome of *Kuenenia stuttgartiensis*, a novel bacterium representing a functional guild involved in anaerobic ammonium oxidation (anammox), from a complex bioreactor community. The annamox bacteria were discovered just over a decade ago and were demonstrated to possess specialized biochemical pathways enabling anaerobic oxidation of ammonium. While these organisms are known for very slow growth and none are available in pure culture, their role has been established as key participants of the global nitrogen cycle, contributing up to 50% to the removal of fixed nitrogen from oceans. Strous and colleagues (2006) enriched *K. stuttgartiensis* in a laboratory reactor in which its population comprised approximately 73% of total cell counts. A rather massive sequencing effort was applied (192,713 reads, approximately 154 Mb) allowing not only for assembly, but also for closing most of the gaps in the *K. stuttgartiensis* genome, resulting in only five contigs, with more than 98% of the genome captured. Metabolic reconstruction revealed unexpectedly high metabolic versatility of the organism and high degree of functional redundancy. Over 200 genes were identified involved in catabolism and respiration. Metabolic pathways for ammonium oxidation, electron transfer, energy conservation and carbon metabolism were reconstructed with great precision, and some were confirmed experimentally. Most significantly, candidate gene clusters were identified as being responsible for biological hydrazine metabolism and for ladderane biosynthesis, functions uniquely connected to annamox. To identify these functions in a novel organism, the availability of a complete or nearly complete genome was a prerequisite, as these genes could only be predicted in the context of a specific genome.

A similar approach was used by García Martín and colleagues (2006) to reconstruct nearly complete genomes of two strains of *Candidatus Accumulibacter phosphatis*, a polyphosphate-accumulating bacterium harnessed for inorganic phosphorus removal in wastewater treatment plants. Sludge samples from two laboratory scale reactors containing 60 to 80% of A. phosphatis as a proportion of total cell counts were subjected to WGS sequencing, producing 5x and 8x sequence coverage, respectively, for the two A. phosphatis genomes. Sequence reads were assembled using two alternative assembler tools (to compare their performance, not reviewed here), resulting in contigs as large as 170 kb and scaffolds as large as 3 Mb, covering at least 97% of one of the genomes. The high quality of this genome allowed for obtaining important insights into the biochemical details of polyphosphate biosynthesis as well the as details of other key metabolic pathways. One of the important problems in biologically enhanced phosphate removal has been the lack of understanding of the sources of reducing power for polyhydroxyalkanoate synthesis in the anaerobic phase. Based on the genomic analysis, a novel cytochrome has been proposed to function as a quinol-NAD(P)H reductase, allowing for anaerobic functioning of the tricarboxylic acid cycle. Despite previous evidence for denitrifying capability of A. phosphatis, no traditional respiratory nitrate reductase was encoded, suggesting that a different enzyme may be carrying out the function. Reconstruction of the complete nitrogen fixation pathway was also a surprise. Again, only from complete or nearly complete genomic sequence could the metabolic blueprint be reconstructed with such precision, representing a turning point in the understanding of the genetics of the process of enhanced biological phosphate removal. This knowledge opens ways for predicting the efficiency of bioreactors for phosphorus removal and suggests optimal conditions for their operation.

Another study has demonstrated that, with an adequate sequencing effort, genomes of minor or at least non-dominant members of the communities could also be sequenced to completion (Pelletier et al., 2008). The genome of *Candidatus* Cloacamonas acidaminovorans that is part of a complex community of industrial anaerobic digesters has been targeted in order to obtain insights into the physiology and metabolism of this representative of a candidate division WWE1, with no cultivated members. The strategy implemented was a massive sequencing effort (1.7 million reads, 1.12 Gb) to end-sequence 1 million fosmids, followed by an iterative assembly approach (similar to the one used by Hallam et al., 2006) to reconstruct the entire genome. Indeed the genomic insights uncovered were worth the effort. Analysis of the genome revealed low gene density (81%) unusual for bacteria, 40% of genes being unique. The best matches for the genes in *Candidatus* C. acidiaminovorans were distributed among distantly related taxa, including Proteobacteria, Firmicutes and Planctomycetes. Some of the proteins encoded in the genome were more related to their eukaryotic than their prokaryotic counterparts. The organism was also predicted to use pyrophosphate-dependent enzymes, a relatively rare feature. A deficiency in biosynthesis of 12 amino acids and several vitamins and cofactors were predicted from the genome, offering one explanation for this organism not existing in pure culture. No typical respiratory chains could be predicted from the genome, suggesting that the fermentation metabolism must be responsible for energy production. Overall, the genome of this novel syntrophic bacterium is an important contribution to the gene pool that determines the quality of annotation for the newly emerging genomes and metagenomes.

More recently, a community of a production-scale biogas reactor was analyzed using the 454 pyrosequencing (FLX System) technology, with en average read length of 230 bases, at a 142 Mb sequencing effort (Schlüter et al., 2008). Attempts of assembly resulted in a large number of contigs over 500 bp, some contigs acceding 10 kb in size and the largest contig being 31.5 Kb in size. The contig sequences were matched to the related genomes available in the non-redundant database. This study demonstrated that de novo assembly from shorter 454 sequences is in principle possible for metagenomic sequences, at least in cases of relatively low community complexity.

BROMODEOXYURIDINE LABELING AS A MEANS FOR FUNCTIONAL METAGENOMICS

The principle of this approach is in targeting species actively replicating their DNA in response to the addition of a specific compound, by labeling the newly synthesized DNA with bromodeoxyuridine (an analogue of thymidine). So far this method was employed on a large scale only in one project, as an attempt to identify the species active in utilization of dissolved organic carbon (DOC) in the coastal ocean (Mou at al, 2008). Dimethylsulphoniopropionate (DMSP) and vanillate were used as model DOC compounds in microcosm incubations supplemented with bromodeoxyuridine, followed by pyrosequencing of the DNA captured by immunoprecipitation. However, in this case, no enrichment was observed for known species involved in degradation of the target compounds and no key genes involved in this process were found to be over-represented. From this attempt only, the potential of this method for characterizing natural populations remains uncertain, and further exploration of this approach is required. Some reasons for this failure are obvious, such as the insufficient sampling (only 4 to 10 Mb per sample), while others are less clear. It is possible that the efficiency of the label incorporation is not uniform among different taxa. It is also possible that label incorporation takes place independently of substrate stimulation, thus the results obtained represent random sampling rather than selection for functional types.

TARGETING FUNCTIONAL TYPES VIA STABLE ISOTOPE PROBING

One way to directly link a function in the environment to a specific guild performing this function is to feed the population a substrate of interest, labeled by a heavy isotope, followed by characterization of the heavy fraction of communal DNA that is enriched in DNA of microbes that actively metabolized the labeled substrate. This technique is known as Stable Isotope Probing and it has been effective in identifying microbes involved in specific biogeochemical transformations such as methylotrophy, phenol degradation, glucose metabolism etc. (Friedrich 2006). Typically, small amounts of DNA are isolated from these experiments, and these are used for phylogenetic profiling and detection of key functional genes, after PCR amplification. So far, there is only one example of scaling this method up to obtain amounts of DNA enabling the WGS sequencing approach, applied to communities of a freshwater lake sediment involved in utilization of C1 compounds (methylotrophs; Kalyuzhnaya et al., 2008). The goal of this targeted metagenomic approach has been two-fold: to reduce the complexity of the community that has been estimated at approximately 5000 species and to directly link specific substrate repertoires to functional guilds. Five different labeled substrates have been employed, methane, methanol, methylamine, formaldehyde and formate, resulting in five 'functional' metagenomes (26 to 58 Mb in size). Community complexity in each microcosm was found to be dramatically reduced compared to the complexity of non-enriched community. From the present 16S rRNA genes, the communities shifted toward specific functional guilds that included bona fide methylotroph species as well as organisms distantly related to cultivated species, implicating them in methylotrophy. The methylamine microcosm metagenome (37 Mb) was found least complex and it was dominated by a single species, *Methylotenera mobilis* represented by a number of closely related strains, while in the non-enriched community *Methylotenera* species comprised less than 0.5% of the population. Via compositional binning, a nearly complete genome of this novel organism has been extracted from the metagenome and its metabolism reconstructed, allowing for genome-wide comparisons with a related species. This so far is the most dramatic example of assembling a genome of a species that is a minor member of a community. Thus the method has been dubbed 'high-resolution metagenomics'. In addition, as part of this project, complete genomes of novel bacteriophages have been assembled from the same metagenome and their association with *M. mobilis* has been proposed, suggesting a mechanism for a dynamic control of the *Methylotenera* populations.

METATRANSCRIPTOMICS

Metatranscriptomics, analysis of community transcripts isolated directly from the environment or from microcosms in which the community has been disturbed or manipulated in a certain way, represent the next logical step in the meta- (-omics) approach. This method should enable reaching beyond the community's genomic potential (metagenomic blueprint), and connect more directly the taxonomic make up of the community to its in situ activity (function), via profiling of (most abundant) transcripts and correlating them with specific environmental conditions. For large-scale metatranscriptomics experiments, the next generation sequencing technologies are especially attractive as assembly is not a prerequisite for transcript analysis. The few metatranscriptomic studies published so far (Urich et al., 2008; Frias-Lopez et al., 2008; Gilbert et al., 2008; Poretsky et al., 2009) have employed the 454 sequencing technology, as this technology produces reads of sufficient length to allow for functional predictions based on a single read. These reads were then processed in a gene-centric way. Obviously, all the pitfalls discussed above relating to the analysis and annotation of short metagenomic reads apply to the short metatranscriptomic reads. Thus, with few exceptions (for example when a genome of a cultivated species is well represented in the environment in question and could be used as a scaffold), only general functional predictions can be made, and in most cases no phylogenetic assignments can be made for specific functional genes. In addition, biases and limitations specific to the analysis of RNA molecules apply: often times only very small amounts of the RNA can be isolated, so an amplification step is necessary (Frias-Lopez et al., 2008; Gilbert et al., 2008). The natural abundance of non-messenger RNA can be a blessing (if a careful phylogenetic profiling is desired; Urich et al; 2008) or a curse (if mRNA is the primary target) as efficient separation of mRNA from more abundant ribosomal and transport RNA remains a problem. Of the potential mRNA transcripts, typically only one third can be matched to known genes or functional gene categories while the rest cannot be classified for the lack of any matches in the databases (orphan proteins). Thus, the resolution provided by direct analysis of short reads remains very low.

As a proof of a concept, we tested an approach in which environmental transcripts were matched to a scaffold previously generated for a community from the same study site. A metagenomic scaffold representing the methylotorph community of Lake Washington sediment was employed (Kalyuzhnaya et al., 2008), and transcript sequences from a community sampled from the same study site were generated using the Illumina technology (resulting in ultra-short reads of approximately 40 bp; unpublished data). The metagenomic scaffold we used is mostly represented by contigs of low sequence coverage, reflecting the insufficient sampling that is typical of metagenomic studies (low-resolution metagenome). However, a small part of the metagenome, representing a composite genome of *Methylotenera* species is made up of contigs of much higher sequence coverage (high-resolution metagenome; Kalyuzhnaya et al., 2008). We matched the transcripts separately to the low-resolution metagenome and to the high-resolution metagenome. Only 8% of the almost 25 million Ilumina reads found targets in the metagenome, reflecting that both the metagenome and the metatranscriptome must have been significantly under-sampled. When matching the transcripts against the low-resolution scaffold, we determined that approximately 35% of the metagenome overlapped with the metatranscriptome. However, when matching was done with the much better sampled *Methylotenera* scaffold, we found that over 96% of the composite genome had matches in the metatranscriptome. This result highlights the necessity of well-covered and more complete genomic scaffolds from the environments that are interrogated via metatranscriotpmics, to enable high-resolution analysis.

METAPROTEOMICS

Metaproteomics, analysis of protein profiles of microbial communities, presents an even better opportunity to address the function directly, as proteins are the molecules that ultimately perform the function. However, metaproteomics, even more so than metatranscriptomics, rely on quality metagenomic data. The large-scale MS/MS-based metaproteomics approach has been pioneered (Ram et al., 2005) and further perfected by collaborative efforts between the Banfield and the Hettich groups, establishing the current state-of-the-art of the field (VerBerkmoes et al., 2009a). The power of this approach was first demonstrated on a community of low complexity (Ram et al., 2005), an AMD community that was not identical but similar to the community for which the high quality metagenomic sequence has been previously generated (Tyson et al., 2004). Despite differences between the sequences of predicted proteins in the dataset and those in the actual sample, it was possible to match shotgun MS/MS spectra to peptides, resulting in positive identification of over 2000 proteins that belonged to the five most abundant members of the community. For the most abundant organism, *Leptospirillum* group II, expression of 50% of the proteins was detected by proteomics. Important insights into the respective metabolic contributions of the archaeal and the bacterial members have been revealed, including identification of a novel cytochrome that appears to be central to iron oxidation and AMD formation. The group has now produced over 30 datasets from the AMD system and these have been employed to establish the relationship between the species abundance in communities and the efficiency of protein identification, concluding that, while organisms constituting 30-40% of the community can be sampled to saturation by the metaproteomics approach, the most abundant proteins from members constituting as little as 1% of the population can also be detected. These results are important for planning future proteomics projects, including considerations for specific enrichments if members of low-abundant taxa need to be targeted.

Proteomics has been successfully applied to a community of a higher complexity, the enhanced biological phosphorous removal (EBPR) community, dominated by a single species, *Candidatus* A. phosphatis (Wilmes et al., 2008) similar to the ones described by García Martín et al. (2006), so the reference genomes generated in the latter work have been used as scaffolds for peptide matching, resulting in identification of approximately 2300 proteins, of which approximately 700 have been assigned to *Candidatus* A. phosphatis, enabling extensive analysis of the metabolic pathways central to EBPR.

The most complex metaproteome analyzed to date is the metaproteome of human feces (VerBerkmoes et al., 2009b). Even with an unmatched metagenomic database used for protein detection, on the order of 1000 proteins of bacterial origin have been detected per sample (about 30% of the detected proteins were human proteins), and their relative abundances have been estimated.

The future of metagenomics

Metagenomics is coming of age but still gaining momentum. This coincides with rapid improvement of sequencing technologies and with the realization that the next bottleneck in metagenomics will not be the sequence data production but computation and data storage. From the experience of the past five years, the power of metagenomics is obvious, while its potential is still waiting to be fully realized. It is now quite clear that even communities of limited complexity pose major challenges in terms of genomic exploration, highlighting the necessity of much deeper sampling, and the need for special assembly and analysis tools. Such improvements are possible and imminent, given the fast progress in these areas. However, the cost of metagenomic sequencing to high coverage, even when employing the next generation technologies, remains a major challenge. The price of sequencing is typically advertised as a per-base cost. However, sequences generated by different technologies require different depths of coverage. While the Sanger technology has been brought to state of the art by years of perfecting, producing reads of up to 1 kb with very low error rate, the 454 technology that is predicted to produce reads of similar length in the near future inherently has a much higher error rate. Thus a higher coverage is required to obtain data of similar quality. With the yet more per-base cost effective technologies, such as Illumina or SOLiD, the depth of sequencing needs to be yet much higher (probably 40 to 50 fold) to assure sequence quality. Moving to the next level (in sequencing depth i.e. sequence quality) is necessary to truly understand how complex microbial communities operate, how they evolve and how they respond to the changing environment. It is also essential for gene and pathway discovery via metagenomics.

The stage is now set for Gb-scale metagenomic projects, such as sequencing complex communities to (nearly) saturation. Carrying out such projects will not only test the performance of the newly emerging computational tools for sequence analysis, but will ultimately demonstrate whether we can apply the same or similar 'gold standards' to metagenomic sequences as to single genome sequences. It will also ultimately test the predictions of how much sequencing is necessary to enable delineating (via binning and assembly) single-species genomes that are parts of a community gene pool. Without such 'saturation'-level metagenomic sequencing experiments, comparative analyses of communities over time or space will remain of little value, as only (small) parts of under-sampled communities will be compared to one another. While in the future carrying out such experiments may become routine, at this time it will likely require a concerted community effort. A document published by the U.S. National Academies' National Research Council (NRC) in 2007, entitled "The new science of metagenomics: revealing the secrets of our microbial planet" calls for a new Global Initiative to drive advances in the field of Metagenomics, in a way that the Human Genome Project advanced the mapping of our genetic code (see again Woese, 1998). Such an initiative would help move the field of metagenomics to the new level and toward a brighter future.

Acknowledgements

The author acknowledges support from the National Science Foundation (MCB00604269). Many thanks to Dr. N. Ivanova for critical reading of the manuscript.

References

BRULC, J.M., ANTONOPOULOS, D.A., MILLER, M.E., WILSON, M.K., YANNARELL, A.C., DINSDALE, E.A., EDWARDS, R.E., FRANK, E.D., EMERSON, J.B., WACKLIN, P., COUTINHO, P.M., HENRISSAT, B., NELSON, K.E., WHITE BA. (2009) Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proceedings of National Academy of Sciences USA* **106**, 1948-1953.

DINSDALE, E.A., EDWARDS, R.A., HALL, D., ANGLY, F., BREITBART, M., BRULC, J.M., FURLAN, M., DESNUES, C., HAYNES, M., LI, L., MCDANIEL, L., MORAN, M.A., NELSON, K.E., NILSSON, C., OLSON, R., PAUL, J., BRITO, B.R., RUAN, Y., SWAN, B.K., STEVENS, R., VALENTINE, D.L., THURBER, R.V., WEGLEY, L., WHITE, B.A., ROHWER, F. (2008) Functional metagenomic profiling of nine biomes. *Nature* **452**, 629-632.

EDWARD, R.A., RODRIGUEZ-BRITO, B., WEGLEY, L., HAYNES, M., BREITBART, M., PETERSON, D.M., SAAR, M.O., ALEXANDER, S., ALEXANDER, E.C. JR., ROHWER, F. (2006) Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* **7**, 57.

FRIAS-LOPEZ, J., SHI, Y., TYSON, G.W., COLEMAN, M.L., SCHUSTER, S.C., CHISHOLM, S.W., DELONG, E.F. (2008) Microbial community gene expression in ocean surface waters. *Proceedings of National Academy of Sciences USA* **105**, 3805-3810.

FRIEDRICH, M.W. (2006) Stable-isotope probing of DNA: insights into the function of uncultivated microorganisms from isotopically labeled metagenomes. *Current Opinion in Biotechnology* **17**, 59-66.

GARCÍA MARTÍN, H., IVANOVA, N., KUNIN, V., WARNECKE, F., BARRY, K.W., MCHARDY, A.C., YEATES, C., HE, S., SALAMOV, A.A., SZETO, E., DALIN, E., PUTNAM, N.H., SHAPIRO, H.J., PANGILINAN, J.L., RIGOUTSOS, I., KYRPIDES, N.C., BLACKALL, L.L., MCMAHON, K.D., HUGENHOLTZ, P. (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nature Biotechnology* **24**, 1263-1269.

GILBERT, J.A., FIELD, D., HUANG, Y., EDWARDS, R., LI, W., GILNA, P., JOINT, I. (2008) Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS ONE* **3**, e3042.

HALLAM, S.J., KONSTANTINIDIS, K.T., PUTNAM, N., SCHLEPER, C., WATANABE, Y., SUGAHARA, J., PRESTON, C., DE LA TORRE, J., RICHARDSON, P.M., DELONG, E.F. (2006) Genomic analysis of the uncultivated marine crenarchaeote *Cenarchaeum symbiosum*. *Proceedings of National Academy of Sciences USA* **103**, 18296-18301.

HANDELSMAN, J. (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews* **68**, 669-685.

HANDELSMAN, J., RONDON, M.R., BRADY, S.F., CLARDY, J., GOODMAN, R.M. (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry and Biology* **5**, R245-249.

KALYUZHNAYA, M.G., LAPIDUS, A., IVANOVA, N., COPELAND, A.C., MCHARDY, A.C., SZETO, E., SALAMOV, A., GRIGORIEV, I.V., SUCIU, D., LEVINE, S.R., MARKOWITZ, V.M., RIGOUTSOS, I., TRINGE, S.G., BRUCE, D.C., RICHARDSON, P.M., LIDSTROM, M.E., CHISTOSERDOVA, L. (2008) High-resolution metagenomics targets specific functional types in complex microbial communities. *Nature Biotechnology* **26**, 1029-1034.

KUNIN, V., COPELAND, A., LAPIDUS, A., MAVROMATIS, K., HUGENHOLTZ, P. (2008) A bioinformatician's guide to metagenomics. *Microbiology and Molecular Biology Reviews* **72**, 557-578.

LAPIDUS, A. (2009) Genome sequence databases (overview): sequencing and assembly. In *The Encyclopedia of Microbiology*. M. Schaechter, pp196-210. New York: Elsevier.

LI, L.L., MCCORKLE, S.R., MONCHY, S., TAGHAVI, S., VAN DER LELIE, D. (2009) Bioprospecting metagenomes: glycosyl hydrolases for converting biomass. *Biotechnology for Biofuels* **2**, 10.

LIOLIOS, K., MAVROMATIS, K., TAVERNARAKIS, N., KYRPIDES, N.C. (2008) The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research* **36**, D475-479.

MARKOWITZ, V.M., IVANOVA, N.N., SZETO, E., PALANIAPPAN, K., CHU, K., DALEVI, D., CHEN, I.M., GRECHKIN, Y., DUBCHAK, I., ANDERSON, I., LYKIDIS, A., MAVROMATIS, K., HUGENHOLTZ, P., KYRPIDES, N.C. (2008) IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Research* **36**, D534-538.

MOU, X., SUN, S., EDWARDS, R.A., HODSON, R.E., MORAN, M.A. (2008) Bacterial carbon processing by generalist species in the coastal ocean. *Nature* **451**, 708-711.

PELLETIER, E., KREIMEYER, A., BOCS, S., ROUY, Z., GYAPAY, G., CHOUARI, R., RIVIÈRE, D., GANESAN, A., DAEGELEN, P., SGHIR, A., COHEN, G.N., MÉDIGUE, C., WEISSENBACH, J., LE PASLIER D. (2008) "*Candidatus* Cloacamonas acidaminovorans": genome sequence reconstruction provides a first glimpse of a new bacterial division. *Journal of Bacteriology* **190**, 2572-2579.

PORETSKY, R.S., HEWSON, I., SUN, S., ALLEN, A.E., ZEHR, J.P., MORAN, M.A. (2009) Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. *Environmental Microbiology* **11**, 1358-1375.

PREIDIS, G.A. AND VERSALOVIC, J. (2009) Targeting the human microbiome with antibiotics, probiotics, and prebiotics: gastroenterology enters the metagenomics era. *Gastroenterology* **136**, 2015-2031.

RAM R.J., VERBERKMOES, N.C., THELEN, M.P., TYSON, G.W., BAKER, B.J., BLAKE, R.C. 2ND, SHAH, M., HETTICH, R.L., BANFIELD, J.F. (2005) Community proteomics of a natural microbial biofilm. *Science* **308**, 1915-1920.

SCHLOSS PD AND HANDELSMAN, J. (2005) Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome Biology* **6**, 229.

SCHLÜTER, A., BEKEL, T., DIAZ, N.N., DONDRUP, M., EICHENLAUB, R., GARTEMANN, K.H., KRAHN, I., KRAUSE, L., KRÖMEKE, H., KRUSE, O., MUSSGNUG, J.H., NEUWEGER, H., NIEHAUS, K., PÜHLER, A., RUNTE, K.J., SZCZEPANOWSKI, R., TAUCH, A., TILKER, A., VIEHÖVER, P., GOESMANN, A. (2008) The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. *Journal of Biotechnology* **136**, 77-90.

SCHMIDT, T.M., DELONG, E.F., PACE, N.R. (1991) Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *Journal of Bacteriology* **173**, 4371-4378.

SNYDER, L.A., LOMAN, N., PALLEN, M.J., PENN, C.W. (2008) Next-generation sequencing-the promise and perils of charting the great microbial unknown. *Microbial Ecology* **57**:1-3.

STALEY, J.T. AND KONOPKA, A. (1985) Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual Review of Microbiology* **39**, 321-346.

STROUS, M., PELLETIER, E., MANGENOT, S., RATTEI, T., LEHNER, A., TAYLOR, M.W., HORN, M., DAIMS, H., BARTOL-MAVEL, D., WINCKER, P., BARBE, V., FONKNECHTEN, N., VALLENET, D., SEGURENS, B., SCHENOWITZ-TRUONG, C., MÉDIGUE, C., COLLINGRO, A., SNEL, B., DUTILH, B.E., OP DEN CAMP, H.J., VAN DER DRIFT, C., CIRPUS, I., VAN DE PAS-SCHOONEN, K.T., HARHANGI, H.R., VAN NIFTRIK, L., SCHMID, M., KELTJENS, J., VAN DE VOSSENBERG, J., KARTAL, B., MEIER, H., FRISHMAN, D., HUYNEN, M.A., MEWES, H.W., WEISSENBACH, J., JETTEN, M.S., WAGNER, M., LE PASLIER, D. (2006) Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature* **440**, 790-794.

THURBER, R.V., WILLNER-HALL, D., RODRIGUEZ-MUELLER, B., DESNUES, C., EDWARDS, R.A., ANGLY, F., DINSDALE, E., KELLY, L., ROHWER, F. (2009) Metagenomic analysis of stressed coral holobionts. *Environmental Microbiology* **11**, 2148-2163.

TRINGE, S.G., VON MERING, C., KOBAYASHI, A., SALAMOV, A.A., CHEN, K., CHANG, H.W., PODAR, M., SHORT, J.M., MATHUR, E.J., DETTER, J.C., BORK, P., HUGENHOLTZ, P., RUBIN, E.M. (2005) Comparative metagenomics of microbial communities. *Science* **308**, 554-557.

TRINGE, S.G., AND HUGENHOLTZ, P. (2008) A renaissance for the pioneering 16S rRNA gene. *Current Opinion in Microbiology* **11**, 442-446.

TYSON, G.W., CHAPMAN, J., HUGENHOLTZ, P., ALLEN, E.E., RAM, R.J., RICHARDSON, P.M., SOLOVYEV, V.V., RUBIN, E.M., ROKHSAR, D.S., BANFIELD, J.F. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37-43.

URICH, T., LANZÉN, A., QI, J., HUSON, D.H., SCHLEPER, C., SCHUSTER, S.C. (2008) Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS One* **3**, e2527.

VENTER, J.C., REMINGTON, K., HEIDELBERG, J.F., HALPERN, A.L., RUSCH, D., EISEN, J.A., WU, D., PAULSEN, I., NELSON, K.E., NELSON, W., FOUTS, D.E., LEVY, S., KNAP, A.H., LOMAS, M.W., NEALSON, K., WHITE, O., PETERSON, J., HOFFMAN, J., PARSONS, R., BADEN-TILLSON, H., PFANNKOCH, C., ROGERS, Y.H., SMITH, H.O. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66-74.

VERBERKMOES, N.C., DENEF, V.J., HETTICH, R.L., BANFIELD, J.F. (2009) Systems biology: Functional analysis of natural microbial consortia using community proteomics. *Nature Reviews Microbiology* **7**:196-205.

VerBerkmoes, N.C., Russell, A.L., Shah, M., Godzik, A., Rosenquist, M., Halfvarson, J., Lefsrud, M.G., Apajalahti, J., Tysk, C., Hettich, R.L., Jansson, J.K. (2009) Shotgun metaproteomics of the human distal gut microbiota. *ISME Journal* **3**, 179-189.

Warnecke, F. and Hess, M. (2009) A perspective: metatranscriptomics as a tool for the discovery of novel biocatalysts. *Journal of Biotechnology* **142**, 91-95.

Warnecke, F., Luginbühl, P., Ivanova, N., Ghassemian, M., Richardson, T.H., Stege, J.T., Cayouette, M., McHardy, A.C., Djordjevic, G., Aboushadi, N., Sorek, R., Tringe, S.G., Podar, M., Martin, H.G., Kunin, V., Dalevi, D., Madejska, J., Kirton, E., Platt, D., Szeto, E., Salamov, A., Barry, K., Mikhailova, N., Kyrpides, N.C., Matson, E.G., Ottesen, E.A., Zhang, X., Hernández, M., Murillo, C., Acosta, L.G., Rigoutsos, I., Tamayo, G., Green, B.D., Chang, C., Rubin, E.M., Mathur, E.J., Robertson, D.E., Hugenholtz, P., Leadbetter, J.R. (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* **450**, 560-565.

Wilmes, P., Andersson, A.F., Lefsrud, M.G., Wexler, M., Shah, M., Zhang, B., Hettich, R.L., Bond, .P.L, VerBerkmoes, N.C., Banfield, J.F. (2008) Community proteogenomics highlights microbial strain-variant protein expression within activated sludge performing enhanced biological phosphorus removal. *ISME Journal* **2**, 853-864.

Woese, C.R. (1987) Bacterial evolution. *Microbiological Reviews* **51**, 221-271.

Woese, C.R. (1998) A manifesto for microbial genomics. *Current Biology* **8**, R781-R783.

Wommack, K.E., Bhavsar, J., Ravel, J. (2008) Metagenomics: read length matters. *Applied and Environmental Microbiology* **74**, 1453-1463.

Woyke, T., Teeling, H., Ivanova, N.N., Huntemann, M., Richter, M., Gloeckner, F.O., Boffelli, D., Anderson, I.J., Barry, K.W., Shapiro, H.J., Szeto, E., Kyrpides, N.C., Mussmann, M., Amann, R., Bergin, C., Ruehland, C., Rubin, E.M., Dubilier, N. (2006) Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* **443**, 950-955.