Biotechnol Lett (2010) 32:1351–1359
DOI 10.1007/s10529-010-0306-9

REVIEW

Recent progress and new challenges in metagenomics for biotechnology

Ludmila Chistoserdova

Received: 11 March 2010 / Accepted: 8 May 2010 / Published online: 21 May 2010

Abstract A brief historical perspective on metagenomics is given followed by a discussion of the rapid progress in this field largely defined by transition to the next generation sequencing technologies. Problems and challenges connected to this transition are also addressed. The review focuses on recent literature describing metagenomic approaches connecting sequence information to functionality that are especially relevant to biotechnological applications, including metagenomics of specialized or enriched microbial communities, metagenomics combined with specific labeling techniques, metatranscriptomics and metaproteomics.

Keywords Biotechnology · Metagenomics · Next generation sequencing

Introduction

Metagenomics is a fast growing and diverse field within biological sciences directed at obtaining knowledge on genomes of environmental microbes, as well as of entire microbial communities, omitting the cultivation step. Other terms are used to describe this methodology, such as environmental genomics, ecogenomics, community genomics, megagenomics, and these are all interchangeable. The power of metagenomics for biotechnology is in allowing one to tap into the vast metabolic potential of uncultivated microbes that represent the majority of microbes on Earth, in order to discover novel genes and novel metabolic pathways. For biotechnological applications, two major types of methodologies are traditionally recognized: function-based and sequence-based metagenomics. The two approaches are not mutually exclusive. Instead, their co-existence reflects both the history of metagenomics and the interdependence of the two methodologies. Function-based metagenomics relies on cloning environmental DNA into expression vectors and propagating them in appropriate hosts, followed by appropriate activity screens (Craig et al. 2010; Schmeisser et al. 2007; Streit et al. 2004). After an active clone is identified (referred to as a hit), the sequence of the clone is determined, the gene of interest and its respective product are further analyzed, and their biotechnological potential is explored. This methodology and the recent improvements in cloning and screening technologies have been extensively reviewed (Simon and Daniel 2009; Schmeisser et al. 2007; Streit et al. 2004; Uchiyama and Miyazaki 2009). The sequence-based metagenomics approach relies on the prior knowledge on proteins possessing the activity of interest, and the screening is performed toward the genes that are predicted to encode proteins

L. Chistoserdova (✉)
Department of Chemical Engineering, University of Washington, Box 355014, Seattle, WA 98195, USA
e-mail: milachis@u.washington.edu

with specific sequences/folds/domains indicative of their functionality. In the early days of metagenomics, such searches were conducted employing PCR amplification or hybridization techniques (Simon and Daniel 2009; Streit et al. 2004). However, the experimental component in searching for a specific type of a protein or a functional pathway is no longer essential with this approach, due to the availability of vast gene databases, including metagenomic databases, that continue to grow exponentially (Kyrpides 2009). Instead, searches for genes, proteins or entire metabolic pathways can be conducted in silico. Prominent target genes then can be custom-synthesized, after tailoring them for optimal expression in a host of choice. This approach is known as 'synthetic metagenomics' (Bayer et al. 2009). An idealized workflow of metagenome sequence-based gene discovery is depicted in Fig. 1 and the feasibility of different steps in the workflow is discussed below.

During the last 5 to 6 years, the field of metagenomics has been transformed by the application of a whole genome shotgun (WGS) sequencing technology that a decade or so earlier revolutionized the field of single organism genomics (Fleischmann et al. 1995). More recent advances in next-generation (ultra-high throughput) sequencing technologies, resulting in a dramatic drop in the price of DNA sequencing, are causing yet another revolution within the field, allowing for sequencing efforts on the scale significantly surpassing the scale allowable by the traditional technologies. Not only these developments bring about a possibility of addressing new, previously unattainable questions in biology, but they also have a potential to significantly accelerate genome-based discovery for medical and biotechnological applications by providing a comprehensive and high-resolution blueprint of a variety of biochemical transformations Nature has invented. This review focuses on the recent advances and emerging challenges in the WGS-based metagenomics and their implications for biotechnology.

Brief history of metagenomics

The history of metagenomics may be traced back to the work of Staley and Konopka (1985) that first reported on the 'great plate count anomaly', followed by the work of the Woese group that identified the 16S rRNA gene as a marker molecule for assessing microbial diversity (Woese 1987) and by the work of the Pace group that first employed this gene for phylogenetic profiling of microbial communities (Schmidt et al. 1991). The term 'metagenomics' was coined about a decade later by the Handelsman group, referring to the function-based analysis of mixed environmental DNA species (Handelsman et al. 1998). However, a new, and now most widely accepted meaning to the term has emerged as a result of the two seminal works published in 2004, both describing the application of random WGS sequencing (Tyson et al. 2004; Venter et al. 2004) to microbial populations. The significance of these two early studies has been two-fold. On 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 (extreme environment inhabited by few specialists, Tyson et al. 2004) to very complex (environment inhabited by a variety of species, Venter et al. 2004). 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 the 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, over 200 projects are listed in the GOLD database, representing over 450 separate environmental samples (Liolios et al. 2010), and results from 58 of these have been already published.

Over the same time, breakthroughs in developing alternative sequencing technologies occurred, promising a significantly higher throughput and considerably reduced cost of sequencing, thus providing the necessary platform for yet faster acquisition of metagenomic data. These new (known as next-generation; reviewed in Lapidus 2009; Ansorge 2009) sequencing technologies, some of which are

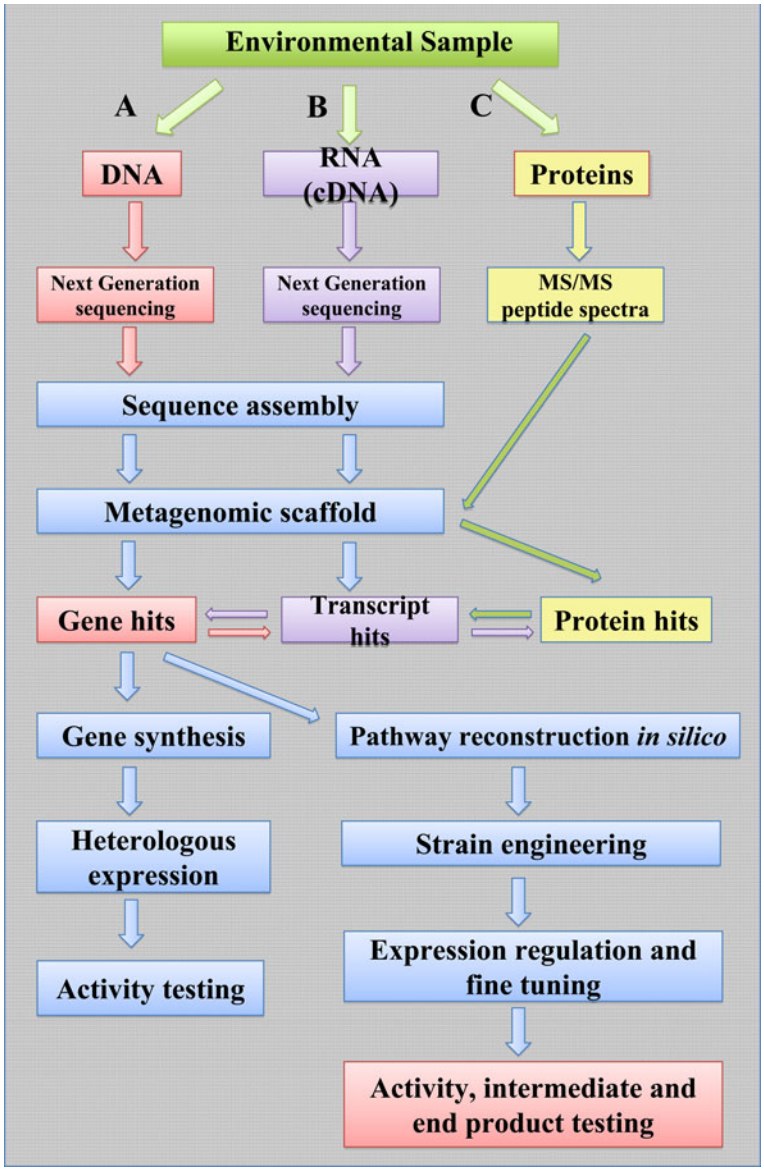

Fig. 1 Workflow of metagenome sequence-based gene discovery. Metagenomics (*A*), metatranscriptomics (*B*), metaproteomics (*C*) or a combination of approaches can be used to identify target genes

already widely utilized (the Roche 454, the Illumina Genome Analyzer and the Applied Biosystems SOLiD platforms) will be defining the future of metagenomics.

Metagenomics in transition

It has been just over 5 years since the onset of WGS sequencing-based metageomics. This is a short time for a new field, so short in fact that it has not yet allowed for establishing a type of the gold standard in metagenomic sequencing and analysis, as has been previously done for single genome sequencing (Chain et al. 2009). Instead, metagenomic projects have been mostly conducted by individual investigators on a small scale, as dictated by the economics of sequencing, and often times without prior knowledge on the structure or the complexity of the community in question. As a result, most of the early metagenomic projects on have followed 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 treats a community

(mostly represented by singleton reads, sometimes 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 reasonably 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 (Liolios et al. 2010; Wu et al. 2009). 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 (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 metabolic make up of an individual organism.

While these challenges persist, a major shift has already occurred toward employing next-generation sequencing technologies to metagenomic data acquisition, the major technology being the Roche 454 producing longer reads and, with sufficient sampling, allowing for sequence assembly (Schlüter et al. 2008). However, recent developments in de novo assembly from short reads generated by the Illumina or SOLiD technologies indicate their potential in metagenomics (Hiatt et al. 2010; Qin et al. 2010). The challenges in data analysis encountered with the unassembled Sanger reads as described above are more severe when it comes to shorter reads (Gomez-Alvarez et al. 2009; Wommack et al. 2008). In the works published so far, some based on the earlier versions of the 454 technology that produced 100–200 bp reads, up to 70% of the reads could not be classified (Brulc et al. 2009; Dinsdale et al. 2008; Edwards et al. 2006; Vega Thurber et al. 2009). In addition, new generation metagenomic data are prone to systematic artifacts that bias data interpretation (Gomez-Alvarez et al. 2009).

The cost of sequencing still remains a major challenge. While the Sanger technology has been brought to the 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 high quality. With the yet more per-base cost effective technologies, such as Illumina or SOLiD, the depth of sequencing needs to be, however, much higher to assure sequence quality. This has not become a common practice so far. Unsurprisingly, some of the metagenomic datasets produced recently using the new technologies are not only smaller than metagenomic datasets produced with the traditional Sanger technology, but, as one would expect, are of lower quality due to low coverage. However, moving to the next level (in sequencing depth and respectively in sequence quality) is crucial to truly understand the structure of the complex microbial communities, which will translate into the knowledge on how they function and evolve and how they respond to the changing environment. High quality sequence data are also imperative for gene and pathway discovery via metagenomics.

Connecting metagenomics to functionality

While metagenomic databases are growing steadily, at this time they represent only a very small fraction of uncultivated microbes. To overcome this problem, a number of approaches have been implemented in order to increase the resolution of metagenomic data and to connect them to functionality, thus allowing to target specific genes and address specific biological questions or specific biotechnological applications. A few of such examples are discussed below.

Metagenomics of niche-specialized or enriched 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. As a result, nearly complete genomes (at 10× 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) were covered at 3×. Reconstruction of the metabolism of these species provided important insights into their ecological roles and the function of the community. This study remains an idealized model of a desired outcome of a metagenomic project.

Examples of highly specialized communities that are more complex but are of great interest as potential sources of enzymes for biotechnological use include symbiotic communities of plant-digesting organisms. 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, using the Sanger sequencing technology), 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.

Similarly, a metabolic potential of communities of the fiber-adherent bovine rumen has been investigated in terms of the abundance and predicted specificity of glycoside hydrolases. With a smaller sequencing effort, using the Roche 454 sequencing technology, a large number of enzymes able to degrade plant biomass were uncovered. Fundamental differences have been noted between the predicted specificity of glycosyl hydrolases from bovine rumen compared to termite hindgut, attributed to the difference in respective diets (forages versus wood; Brulc et al. 2009).

Another target for discovery of glycoside hydrolases has been undertaken via metagenomics of a switchgrass-adapted compost community. In this case, the community in question has been specifically enriched for species most active in lignocellulose degradation. However, even at a larger sequencing effort (225 Mb of a 454-titanium pyrosequence data), akin to the two former studies, few complete glycosyl hydrolase genes have been recovered, thus providing few targets for functional expression and activity testing (Allgaier et al. 2010). While at a lower proportion of total genes, glycosyl hydrolase candidates have been identified in metagenomes originating from a variety of other environments (Li et al. 2009), suggesting additional resources for gene mining for biofuel production.

Combining metagenomics with substrate-specific labeling

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 the DNA of the 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- or multiple displacement amplification (Chen and Murrell 2010). 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 (26–58 Mb in size). Community complexity in each microcosm was found to be dramatically reduced compared to the complexity of

the 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. Via compositional binning, a nearly complete genome of a novel organism *Methylotenera mobilis* 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 (<0.5%) of a community. Thus the method has been dubbed 'high-resolution metagenomics'. The method can be used with virtually any substrate. A modification of the method employing metagenomic PCR-amplified and cosmid libraries has recently been used to assess diversity of biphenyl dioxygenase genes from a polychlorinated biphenyl-contaminated river sediment (Sul et al. 2009).

An alternative type of labeling for uncovering functionality uses bromodeoxyuridine. The principle of this approach is in targeting species actively replicating their DNA in response to a test compound. These species will incorporate bromodeoxyuridine that is an analogue of thymidine, into their newly synthesized DNA. 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 et al. 2008). Dimethylsulphoniopropionate and vanillate were used as model DOC compounds in microcosm incubations supplemented with bromodeoxyuridine, followed by pyrosequencing of the DNA captured by immunoprecipitation. This method also has a potential to be utilized with a variety of natural or anthropogenic substrates, and the resulting metagenomic data can be mined for either known enzymes or novel enzymes relevant to the process based on their over-representation in the active population.

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 reach 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 advantage of metatranscriptomics for gene discovery, especially in targeted environments (specialized communities as described above or enriched communities) as it requires a much smaller sequence space compared to metagenomics, focusing on the expressed (or over-expressed if transcriptomes from different conditions are compared) subset of genes (Warnecke and Hess 2009). The few metatranscriptomic studies published so far (Frias-Lopez et al. 2008; Gilbert et al. 2008; Urich et al. 2008) employed the Roche 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. Other biases and limitations specific to the analysis of RNA molecules exist: 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. A principally different approach to metatranscriptomics would involve matching the transcripts to a specific metagenomic scaffold (preferably from the same sample or at least from the same environment), as it is becoming a popular practice with single-organism transcriptomics (Sorek and Cossart 2010; van Vliet 2009).

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. Conversely, high quality proteomic data can be used for refining genomic annotations (Armengaud 2009). As a reflection of this interdependence, the term proteogenomics is becoming increasingly popular (Delmotte et al. 2009; Denef et al. 2010; Wilkins et al. 2009). Not surprisingly, the best examples so far of large-scale MS/MS-based metaproteomics approaches include low-complexity models, most prominently AMD communities for which high-quality metagenomic sequences are available (Baker et al. 2010; Denef et al. 2010; Goltsman et al. 2009; Lo et al. 2007; Ram et al. 2005), and these determine the current state of the art in the field (VerBerkmoes et al. 2009a). From these analyses, organisms constituting 30–40% of the community can be sampled to saturation by the metaproteomics approach. However, the most abundant proteins from members constituting as little as 1% of the population can also be detected. These results highlight the importance of careful planning for proteomics-based projects, including considerations for specific enrichments, in case members of low-abundant taxa need to be targeted.

Although at somewhat lower resolution, the proteogenomics approach has been successfully expanded to communities of much higher complexities, such as an enhanced biological phosphorous removal community (Wilmes et al. 2008), a human feces community (VerBerkmoes et al. 2009b), a uranium bioremediation community (Wilkins et al. 2009) and phyllosphere communities (Delmotte et al. 2009). The proteogenomics approach is still gaining momentum and promises to play an increasingly prominent role in identification of proteins with biological activities desirable for biotechnological applications.

Concluding remarks

While the power of metagenomics is obvious as applied to discovery of the multitude of enzymes and biochemical pathways for biotechnological applications, the field is still gaining momentum and its potential is waiting to be fully realized. With the transition to the next generation sequencing technologies, the stage is now set for Gb-scale metagenomic projects, such as sequencing complex communities to (nearly) saturation. However, 'saturation'-level metagenomics projects will require much deeper sampling that is currently the norm, along with special assembly, analysis tools and data storage infrastructures. Such projects are still unattainable by single research groups and require concerted community efforts. Such efforts are already under way for the analysis of human microbiomes (Qin et al. 2010; Turnbaugh et al. 2007) and soil metagenomes (Vogel et al. 2009), but they need to be expanded to other environments. Major initiatives and integrated solutions are also needed for data sharing that is essential for mining the ever-growing databases (Martin and Martin 2010). When this is achieved, synthetic metagenomics will near a state of the art allowing for not only high-resolution gene mining but for assembling entire metabolic pathways in silico, akin to pathway construction from the well studied enzyme systems (Atsumi et al. 2009; Dueber et al. 2009).

Acknowledgement The author acknowledges support from the National Science Foundation (MCB00604269).

References

Allgaier M, Reddy A, Park JI et al (2010) Targeted discovery of glycoside hydrolases from a switchgrass-adapted compost community. PLoS One 5:e8812

Ansorge WJ (2009) Next-generation DNA sequencing techniques. Nat Biotechnol 25:195–203

Armengaud J (2009) A perfect genome annotation is within reach with the proteomics and genomics alliance. Curr Opin Microbiol 12:292–300

Atsumi S, Higashide W, Liao JC (2009) Direct photosynthetic recycling of carbon dioxide to isobutyraldehyde. Nat Biotechnol 27:1177–1180

Baker BJ, Comolli LR, Dick GJ et al (2010) Enigmatic, ultrasmall, uncultivated Archaea. Proc Natl Acad Sci USA 107(19):8806–8811

Bayer TS, Widmaier DM, Temme K et al (2009) Synthesis of methyl halides from biomass using engineered microbes. J Am Chem Soc 131:6508–6515

Brulc JM, Antonopoulos DA, Miller ME et al (2009) Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. Proc Nat Acad Sci USA 106:1948–1953

Chain PS, Grafham DV, Fulton RS et al (2009) Genomics. Genome project standards in a new era of sequencing. Science 326:236–237

Chen Y, Murrell JC (2010) When metagenomics meets stable-isotope probing: progress and perspectives. Trends Microbiol 18:157–163

Craig JW, Chang FY, Kim JH et al (2010) Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse proteobacteria. Appl Environ Microbiol 76:1633–1641

Delmotte N, Knief C, Chaffron S et al (2009) Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. Proc Natl Acad Sci USA 106:16428–16433

Denef VJ, Kalnejais LH, Mueller RS et al (2010) Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. Proc Natl Acad Sci USA 107:2383–2390

Dinsdale EA, Edwards RA, Hall D et al (2008) Functional metagenomic profiling of nine biomes. Nature 452:629–632

Dueber JE, Wu GC, Malmirchegini GR et al (2009) Synthetic protein scaffolds provide modular control over metabolic flux. Nat Biotechnol 27:753–759

Edwards RA, Rodriguez-Brito B, Wegley L et al (2006) Using pyrosequencing to shed light on deep mine microbial ecology. BMC Genomics 7:57

Fleischmann RD, Adams MD, White O et al (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–498

Frias-Lopez J, Shi Y, Tyson GW et al (2008) Microbial community gene expression in ocean surface waters. Proc Nat Acad Sci USA 105:3805–3810

Friedrich MW (2006) Stable-isotope probing of DNA: insights into the function of uncultivated microorganisms from isotopically labeled metagenomes. Curr Opin Biotechnol 17:59–66

Gilbert JA, Field D, Huang Y et al (2008) Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. PLoS One 3:e3042

Goltsman DS, Denef VJ, Singer SW et al (2009) Community genomic and proteomic analyses of chemoautotrophic iron-oxidizing "*Leptospirillum rubarum*" (Group II) and "*Leptospirillum ferrodiazotrophum*" (Group III) bacteria in acid mine drainage biofilms. Appl Environ Microbiol 75:4599–4615

Gomez-Alvarez V, Teal TK, Schmidt TM (2009) Systematic artifacts in metagenomes from complex microbial communities. ISME J 3:1314–1317

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

Hiatt JB, Patwardhan RP, Turner EH et al (2010) Parallel, tag-directed assembly of locally derived short sequence reads. Nat Methods 7:119–122

Kalyuzhnaya MG, Lapidus A, Ivanova N et al (2008) High-resolution metagenomics targets specific functional types in complex microbial communities. Nat Biotechnol 26:1029–1034

Kyrpides NC (2009) Fifteen years of microbial genomics: meeting the challenges and fulfilling the dream. Nat Biotechnol 27:627–632

Lapidus A (2009) Genome sequence databases (overview): sequencing and assembly. In: Schaechter M (ed) The encyclopedia of microbiology. New York, Elsevier, pp 196–210

Li LL, McCorkle SR, Monchy S et al (2009) Bioprospecting metagenomes: glycosyl hydrolases for converting biomass. Biotechnol Biofuels 2:10

Liolios K, Chen IM, Mavromatis K et al (2010) The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 38:D346–D354

Lo I, Denef VJ, Verberkmoes NC et al (2007) Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. Nature 446:537–541

Martin NF, Martin F (2010) From Galactic archeology to soil metagenomics—surfing on massive data streams. New Phytol 185:343–347

Mou X, Sun S, Edwards RA et al (2008) Bacterial carbon processing by generalist species in the coastal ocean. Nature 451:708–711

Qin J, Li R, Raes J, Arumugam M et al (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464:59–65

Ram RJ, VerBerkmoes NC, Thelen MP et al (2005) Community proteomics of a natural microbial biofilm. Science 308:1915–1920

Schlüter A, Bekel T, Diaz NN et al (2008) The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. J Biotechnol 136:77–90

Schmeisser C, Steele H, Streit WR (2007) Metagenomics, biotechnology with non-culturable microbes. Appl Microbiol Biotechnol 75:955–962

Schmidt TM, DeLong EF, Pace NR (1991) Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. J Bacteriology 173:4371–4378

Simon C, Daniel R (2009) Achievements and new knowledge unraveled by metagenomic approaches. Appl Microbiol Biotechnol 85:265–276

Sorek R, Cossart P (2010) Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. Nat Rev Genet 11:9–16

Staley JT, Konopka A (1985) Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Ann Rev Microbiol 39:321–346

Streit WR, Daniel R, Jaeger KE (2004) Prospecting for biocatalysts and drugs in the genomes of non-cultured microorganisms. Curr Opin Biotechnol 15:285–290

Sul WJ, Park J, Quensen JF III et al (2009) DNA-stable isotope probing integrated with metagenomics for retrieval of biphenyl dioxygenase genes from polychlorinated biphenyl-contaminated river sediment. Appl Environ Microbiol 75:5501–5506

Tringe SG, von Mering C, Kobayashi A (2005) Comparative metagenomics of microbial communities. Science 308:554–557

Turnbaugh PJ, Ley RE, Hamady M et al (2007) The human microbiome project. Nature 449:804–810

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

Uchiyama T, Miyazaki K (2009) Functional metagenomics for enzyme discovery: challenges to efficient screening. Curr Opin Biotechnol 20:616–622

Urich T, Lanzén A, Qi J et al (2008) Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. PLoS One 3:e2527

van Vliet AH (2009) Next generation sequencing of microbial transcriptomes: challenges and opportunities. FEMS Microbiol Lett 302:1–7

Vega Thurber R, Willner-Hall D, Rodriguez-Mueller B et al (2009) Metagenomic analysis of stressed coral holobionts. Environ Microbiol 11:2148–2163

Venter JC, Remington K, Heidelberg JF (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66–74

VerBerkmoes NC, Denef VJ, Hettich RL, Banfield JF (2009a) Systems biology: functional analysis of natural microbial consortia using community proteomics. Nat Rev Microbiol 7:196–205

VerBerkmoes NC, Russell AL, Shah M et al (2009b) Shotgun metaproteomics of the human distal gut microbiota. ISME J 3:179–189

Vogel TM, Simonet P, Jansson JK et al (2009) TerraGenome: a consortium for the sequencing of a soil metagenome. Nat Rev Microbiol 7:252

Warnecke F, Hess M (2009) A perspective: metatranscriptomics as a tool for the discovery of novel biocatalysts. J Biotechnol 142:91–95

Warnecke F, Luginbühl P, Ivanova N et al (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. Nature 450:560–565

Wilkins MJ, Verberkmoes NC, Williams KH et al (2009) Proteogenomic monitoring of *Geobacter* physiology during stimulated uranium bioremediation. Appl Environ Microbiol 75:6591–6599

Wilmes P, Andersson AF, Lefsrud MG et al (2008) Community proteogenomics highlights microbial strain-variant protein expression within activated sludge performing enhanced biological phosphorus removal. ISME J 2: 853–864

Woese CR (1987) Bacterial evolution. Microbiol Rev 51: 221–271

Wommack KE, Bhavsar J, Ravel J (2008) Metagenomics: read length matters. Appl Environ Microbiol 74:1453–1463

Wu D, Hugenholtz P, Mavromatis K et al (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature 462:1056–1060