World J Microbiol Biotechnol (2009) 25:945–954
DOI 10.1007/s11274-009-9971-z

ORIGINAL PAPER

Metagenomics as a new technological tool to gain scientific knowledge

María-Eugenia Guazzaroni · Ana Beloqui · Peter N. Golyshin · Manuel Ferrer

Received: 24 November 2008 / Accepted: 13 January 2009 / Published online: 30 January 2009

Abstract Metagenomics (also Environmental Genomics, Ecogenomics or Community Genomics) is an emerging approach to studying microbial communities in the environment. This relatively new technique enables studies of organisms that are not easily cultured in a laboratory, thus differing from traditional microbiology that relies almost entirely on cultured organisms. Metagenomics technology thus holds the premise of new depths of understanding of microbes and, importantly, is a new tool for addressing biotechnological problems, without tedious cultivation efforts. DNA sequencing technology has already made a significant breakthrough, and generation of gigabase-pairs of microbial DNA sequences is not posing a challenge any longer. However, conceptual advances in microbial science will not only rely on the availability of innovative sequencing platforms, but also on sequence-independent tools for getting an insight into the functioning of microbial communities. This is an important issue, as we know that even the best annotations of genomes and metagenomes only create hypotheses of the functionality and substrate spectra of encoded proteins which require experimental testing by classical disciplines such as physiology and biochemistry. In this review, we address the following question, how to take advantage of, and how can we improve the, metagenomic technology for accommodating the needs of microbial biologists and enzymologists?

M.-E. Guazzaroni · A. Beloqui · M. Ferrer (✉)
CSIC, Institute of Catalysis, Marie Curie 2, 28049 Madrid, Spain
e-mail: mferrer@icp.csic.es

P. N. Golyshin
School of Biological Sciences, Bangor University,
Gwynedd LL57 2UW, UK

P. N. Golyshin
Centre for Integrated Research in the Rural Environment,
Aberystwyth University-Bangor University Partnership
(CIRRE); Environmental Microbiology Laboratory,
HZI-Helmholtz Centre for Infection Research,
38124 Braunschweig, Germany

Keywords Metagenomics · Microbial diversity · Systems microbiology

Introduction

Microbes, the most abundant organisms on Earth, play an essential role in biogeochemical processes and element cycling, maintaining the functioning of the global ecosystem. From this point of view it is crucial to generate a thorough understanding of these key microorganisms and processes they facilitate. However, at present we simply do not know the extent of the functional diversity that microbes encompass: a classical theoretical analysis estimates a population of prokaryotes on Earth of about 10^{30} bacteria, few order of magnitude higher than the number of stars in the known Universe (estimated 10^{22}–10^{24}) (McHardy and Rigoutsos 2007; Lozupone and Knight 2008), with most microbes being members of complex communities. Invertebrate guts are certainly one of the most dense and diverse niches (10^{9}–10^{11} cells per ml of gut fluid; Warnecke et al. 2007), followed by soil (10^{7}–10^{9} cells per g; Schloss and Handelsman 2006), and oligotrophic superficial sea- and freshwater (10^{5}–10^{6} bacteria per ml; DeLong 2005). A non-exhaustive list of questions that should be addressed at microbial level in any of these environmental samples includes: what is the extent of prokaryotic diversity, whether "everything is everywhere", do microorganisms exhibit bio-geographical patterns of distribution, is the relative abundance of a certain group of microorganisms necessarily linked to their importance in the community functioning, which organisms are of pivotal importance in the community, how diverse are metabolic pathways and networks within the given ecosystem, how do microbes and protein-coding genes interact with each other to lead to the overall system function, what is the turnover of nutrients and energy in the community, how many specific microbes are responsible for metabolism of different substrates, how do environmental stimuli impact ecosystem functioning and long-term system stability, and how can we obtain and integrate this information?

Of key relevance here is the ability to analyse large numbers of microbes from the environment, together with phylogenetic, genomic and biochemical analyses. However, any individual survey is limited because of the relatively poor capacity of growth of most microorganisms that is offered even by rather sophisticated resources available for culturing (Ingham et al. 2007). To circumvent this problem, a wide range of approaches collectively described as metagenomics, have been developed to study communities through the analysis of their genetic material without culturing individual organisms (Handelsman 2004). Metagenomics, also referred as environmental genomics, ecogenomics or community genomics, is analogous to genomics with the difference that the genome under study is not from a single microbe, but rather from the entire microbial community present in an environmental sample: it is the community genome. Metagenomics represents a strategic concept that includes investigations at three major interconnected levels, sample processing, DNA sequencing and functional analysis, with an ultimate goal of getting a global view of the functioning of the microbial world. Fig. 1 shows a diagram composed of three overlapping circles representing each of the activities: (1) sample processing (green circle), (2) DNA sequencing and analysis (yellow circle) and (3) functional analysis (blue circle). This Figure is meant to emphasize that a new metagenomic discipline has emerged, and that, indeed, this discipline is contributing to the understanding of microbial communities due to the interdisciplinary nature of different sets of activities.

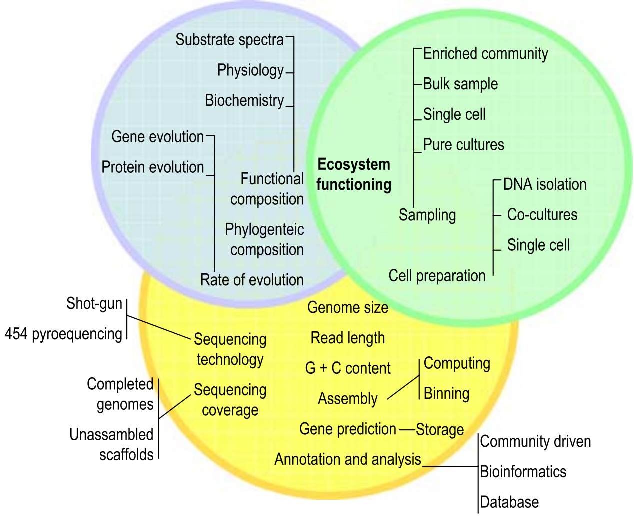

Fig. 1 A generalized diagram with three sets of activities, (1) sampling processing and cell separation, (2) DNA sequencing, annotation and analysis and (3) functional (in terms of activity screens) and phylogenetic analysis, and their intersections. From this representation, the different activity sets are easily observed. Furthermore, if individual elements (i.e. sampling and cell preparation, functional and phylogenetic composition, and rate of evolution, to cite some) are contained in more than one set, the *intersection* indicates a direct view on how those activities are important in different, but related, activities. The interconnection among activities is crucial to get a proper analysis of an environmental microbial community. For example, appropriate sampling and cell-processing strategies are necessary to provide good genome coverage during sequencing and thus, adequate functional activity screens

Whilst many of the technical limitations to processing of samples have been overcome in the last decade (multi-well DNA extractions, single-cell isolation, sequence analysis by technologies such as 454 or Solexa platforms) we believe that major hurdles are still: (1) adequate metagenome coverage, since genes of different organisms are present in very different concentrations in the DNA used to construct the libraries, (2) the integrating and filtering of gene sequences and experimental evidences to facilitate functional assignments of unknown genes, organisms and communities and to recreate functional networks, and (3) the computational aspects of data archiving, analysis and visualization of vast numbers of DNA sequences which are released to databases. In this respect, lessons from 15 years of metagenomics and six of high-throughput DNA sequencing (first analyses of microbial communities through massive sequencing were published in 2004 by Tyson et al. 2004 and Venter et al. 2004) tell that gigabase amounts of environmental sequences can easily be generated to a large extent, but only a fraction of them can properly be annotated in terms of gene functions (~50% of the potential protein-encoding genes lacked any functional assignment). More importantly, DNA sequences per se are not that helpful in linking genes to specific functions, as we know that more than 60% of genes are ubiquitous and have similar housekeeping functions in different organisms (Lombardot et al. 2006). This review looks broadly at current issues in environmental genomics to illustrate the potential of gaining novel knowledge on and to identify the areas of technology to have the greatest promise for critical developments in the near future to move forward culture-independent approaches in the midst of a vast abundance of alternative molecular biology tools.

Diversity of uncultured resources for metagenomics

The principal measure of phylogenetic relatedness, and thus of biodiversity, is the sequence of the 16S ribosomal RNA gene in prokaryotes and its equivalent 18S rRNA gene in eukaryotes. Determination of very large numbers of such sequences has revealed that natural environments contain vast numbers of diverse microorganisms, but only a fraction of them can properly be analysed: many microbes that are present in natural assemblages −99% is often cited- are inactive or unculturable, and thus are not be considered as a part of the microbial community (Dinsdale et al. 2008). Although many reasons may explain this situation (i.e. some species might be found at a particular site merely because they were accidentally transported there but are not capable of functioning), it is clear that culturability estimations are based on the detection methods being used. Therefore, future advances in this topic may help understanding the cultivable percentages within microbial communities. Whatever the case, this “great plate count anomaly” (Staley and Konopka 1985), in fact, observed from early 1930s, stimulated the development of new efficient tools to circumvent problems linked to the cultivation of microbes in artificial media, the so-called metagenomics (Handelsman 2008). These are often described as culture-independent approaches and, in terms of the organisms being accessed and mined, this is the case. However, the need for large amounts of cell biomass for gene and genomic analysis always requires cultivation of a producer microbe, except for DNA sequencing which requires direct separation of cells and bulk DNA. The difference here is that cultivation refers to that of a surrogate organism, the host exploited as a reservoir for archiving the harvested genetic resources. Considering these requirements, metagenomics is often based on a general strategy of producing a large amount of environmental DNA to achieve two goals: (1) discovery of new gene sequences coding for enzymes and drugs (Schmeisser et al. 2007), and (2) randomly sampling and archiving the genomes from a small subset of organisms present in an environment for subsequent in silico analysis (Tringe et al. 2005). The main task of the sequencing projects, both shotgun, Sanger and 454 pyro sequencing, is to obtain maximal sequence information from which one can predict microbial metabolism and predict functional roles (Shendure et al. 2004; Church 2005). The characteristics of all sequencing methods are as follows. Whole genome shotgun sequencing is based on DNA fragmentation followed by cloning in a vector and further sequencing using universal primers. Although still used, it has a major limitation, namely, the size of cloned fragments and difficulties for assembling and price. Sanger-sequencing is based on the electrophoretic separation of deoxyribonucleotide triphosphate (dNTP) fragments with single-base resolution. Using 384-capillary automated sequencing machines, the costs for heavily optimized sequencing centres are currently approaching US $1 per 1,000-bp raw sequencing read and a throughput of ~24 bases per instrument second. Typically, 99.99% accuracy can be achieved with as few as three raw reads covering a given nucleotide. Pyrosequencing, which was introduced in 1996, detects extension through the luciferase-based real-time monitoring of pyrophosphate release. Briefly, the method allows sequencing of up to 100–200 bp single strand of DNA by synthesizing the complementary strand along it, one base pair at a time, and detecting which base was actually added at each step (light is produced only when the nucleotide solution complements the first unpaired base of the template). Ongoing technology allows sequencing of 400×10^6 nucleotides in 10 h, thus completing genome sequencing for about US $7,000. Actually, more than 130 whole metagenomic projects using the 454 technology plattform are running (Liolios et al. 2008). Apart from sequencing, the second research window is based on the possibility of gaining access, through virtual sequence homology screening or by intensive activity screen programmes, to an immense repertoire of millions of known and unknown proteins predicted by the environmental sequence information (Beloqui et al. 2008; Hallin et al. 2008).

Genetic and biochemical information generated by metagenomics in the last 4 years has been used for the identification of more than 40×10^6 genes (by meaning of database entries), the number exceeding by far the 9×10^5 entries generated by genome or traditionally cultivation and cloning efforts (Fig. 2). Concomitantly, metagenomic data may provide chances, not only for analysing individual genes, but to reconstruct the whole metabolism of the organisms comprising the community, and to predict their functional roles in the ecosystem, subjects that will be discussed later in detail. In this respect, a near-complete genome reconstruction has been achieved for the dominant

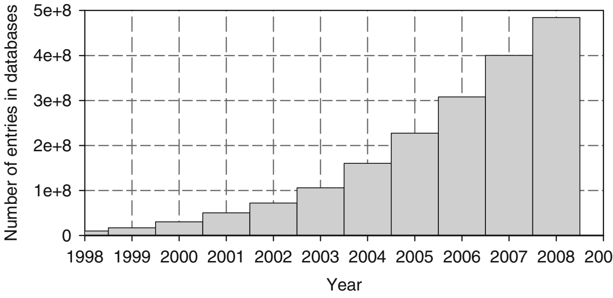

Fig. 2 Exponential increase of the number of entries released to the DNA sequence databases in the last 10 years (when the metagenomic concept was introduced by Handelsman et al. 1998)

Table 1 Analysis of microbial communities published through DNA sequencing technologies

Tool	Sample	Library size*
Shotgun metagenomic sequencing projects	Sargasso Sea	Venter et al. (2004)
	Human feces	Zhang et al. (2006)
	Human distal gut	Gill et al. (2006)
	Soil	Fierer et al. (2007)
	Acid mine drainage biofilm	Tyson et al. (2004)
	Chesapeake Bay virioplankton	Bench et al. (2007)
	Global Ocean	Rusch et al. (2007)
	Soil	Riesenfeld et al. (2004)
	Worm lacking a mouth, gut and nephridia	Woyke et al. (2006)
454 Pyrosequencing technology projects	Coral reef	Krause et al. (2008); Dinsdale et al. (2008)
	Solar saltern	Krause et al. (2008)
	Stromatolite	Krause et al. (2008); Desnues et al. (2008)
	Soudan mine	Edwards et al. (2006)
	Mouse and human distal gut (obese and lean)	Turnbaugh et al. (2006)
	Normal and CCD (Colony collapse disorder) hives	Cox-Foster et al. (2007)
	Coral (*Porites astreoides*)	Wegley et al. (2007)
	North Atlantic deep water and Axial seamount	Sogin et al. (2006)
	Ocean surface waters	Frias-Lopez et al. (2008)
	Global soil	Leininger et al. (2006)
	Human mouth	Marcy et al. (2007)
	Marine virome of four oceanic regions	Angly et al. (2006)
	Northwest Atlantic and Eastern Tropical Pacific seawater	Rusch et al. (2007); Yooseph et al. (2007)
	Surface and hypersaline marine, freshwater samples	Williamson et al. (2008)
	Termite hindgut	Warnecke et al. (2007)

members of communities with a small number of organism types, e.g. acid mine drainage (AMD) communities (Tyson et al. 2004) and for a few highly abundant organisms from diverse communities, e.g. wastewater (García-Martín et al. 2006). Other recent examples of metagenome reconstruction are summarized in Table 1.

Wet-lab sampling procedures for getting a representative sample

It is well recognized that environmental samples are represented by complex microbial communities whose composition and function may differs among them: some organisms are present in high numbers whereas others can be found at concentrations below 10 cells per g of soil (Vieites et al. 2009). In this context, total DNA extracted from environmental samples, in many cases, may not contain even representation of the population's genome, meaning that rare organisms would contribute less to the overall DNA diversity, with the library being dominated by the most abundant organisms (Ferrer et al. 2008, 2009). It is thus of great priority to adapt DNA extraction methods and cloning strategies for normalization of the sample, because the relative abundance of representatives of a certain group of microbes is not necessarily linked to the importance of that group in the community functioning: common organisms may not necessarily play a critical role in a community despite their numbers, and organisms that only muster 0.1% fraction (e.g. nitrogen fixers), can be of pivotal importance (Vieites et al. 2009). What this means in terms of microbial biology, is that the structural and functional information based on individualization studies, that is, classical microbiology and genomics based on single organisms, may not provide appropriate understanding of complex communities. This is an essential part of metagenomic analysis where the global understanding depends essentially on the possibility of (1) isolating the entry bulk DNA (see Fig. 1, green circle) and (2) identifying the genomes (see Fig. 1, yellow circle), genes and proteins (see Fig. 1, blue circle) more relevant to each of the environmental sample under investigation.

To date, much of the research has been focused on bulk DNA. The analysis of such samples, namely at sequence level, has somehow lower resolution, but can access much greater genomic information of untapped microbial biodiversity in a number of environments (see examples in Table 1). By contrast, the second approach shows better options to link specific microbes to specific ecological functions. In one of the first examples, the Sargasso Sea genome equencing project, the authors performed a size-selective filtration for enrichment of the microbes of a certain size (Venter et al. 2004). Newer developments involve the direct separation of cells or preferably enrichment using ^{13}C-labeled compounds directly related to primary ecological functions (Biddle et al. 2008). A particularly elegant strategy combines the extraction of complete uncultured genomes in complex communities (with up to 5,000 species) by high resolution stable isotope probing (SIP) to reconstruct their metabolisms and to link specific microbes, whose DNA is separated by ultracentrifugation, to specific ecological functions (Kalyuzhnaya et al. 2008). Here, authors provided high genome sequence coverage of dominant organisms under dynamic utilization of different nutrients and consequently they were able to link indigenous organisms and processes that are catalysed by these microbes. Although of great potential, the main drawback of cultural enrichment methodologies is the danger of a transient enrichment for fast-growing microbes which reduces the natural diversity in the sample and leads to the transient abundance of microbes not necessarily relevant to the native ecosystem.

We should also point out that community genomics is not limited to prokaryotes: eukaryotic microbial diversity is also enormous, and hence of great interest for exploration of functional diversity. Because of the problem of introns in eukaryotes, considerable effort has been invested in the generation of cDNA from RNA, rather than dealing with genomic DNA. This requires isolation of full-length mRNAs, reverse transcribing them, and analysis of the cDNA (Bodrossy et al. 2006). Here, the RNA extraction technique is critical, since it needs to extract RNA from thick-walled organisms, like fungi and yeasts (and their spores). Further, as a complement to the long-standing trend towards reductionism, metagenomics seeks to treat the community as a whole. However, this is not an easy task, especially for sample processing, as we know that environmental samples also contain pico-eukaryotes (size <2–3 μm) whose composition varies dynamically in response to both seasonal and spatial gradient in ecosystems properties (Bench et al. 2007; Lozupone and Knight 2008). Therefore, a general strategy for sample processing could be recommended for metagenomic studies in the future, in which multiple microbial groups are processesed separately by using single micro-droplets, cell-free translation systems and cell-sorting ("single-cell genomics") and integrate this data with those obtained using mixed microbial communities (Huson et al. 2007; Lasken 2007; Ishoey et al. 2008). Finally we should consider that genome coverage is an ephemeral term, since different community members are present in different numbers in a sample and their genomes are extracted with different efficiencies, so genes of different organisms will be present in very different concentrations in the DNA. For this reason, attempts to obtain (or even calculate the size of) a sample providing good coverage of all genomes present in a sample are rare and limited to samples from extreme environments known to contain microbial communities of very limited complexity and diversity (Ferrer et al. 2007). Further advances in this topic may be required to describe appropriately metagenomic samples.

Data mining in DNA datasets

Inexpensive and ultra-fast sequencing technologies for obtaining nucleotide sequences, usually several tens of bases long, are daily generating enormous amounts of sequencing data (Margulies et al. 2005; Bentley 2006; Eisen 2007). On one hand, this opens up an unprecedented opportunity to dig into the goldmine of "new" sequences; on the other, such large datasets raise several processing problems, and drive current bioinformatic tools to their limits (Fig. 1, yellow circle). In the beginning (see below), DNA sequencing technology was applied to pure cultures, but in the new perspective of sequencing uncultured microbial communities emerge new tools for data analysis. Genome sequencing of single organisms has made a great contribution to our understanding of individual components of microbial communities, whereas community genomics approaches, in which the genomes of a group of organisms are lumped and studied together, open up new horizons to expand our knowledge of the community as a whole. A metagenomic library contains DNA sequences for majority of the genes in the microbial community which with the aid of powerful assembler computer programs, the snippets of DNA sequences are aligned and reassembled into their original order. For such purpose freely available software is used for DNA assembly, i.e. DNA baser and AMOScmp, to mention two.

Three fundamental differences between a cultured microbe and an environmental microbial sample have to be considered when analysing sequence data. First, in the case of cultured organisms, the cells used for DNA isolation represent a clonal population and will have the same genomic sequence. In environmental samples, even for organisms that represent the same "species" there are many independent lineages that result in varying degrees of sequence variation (polymorphisms) within that “species” population (Johnson and Slatkin 2006). That heterogeneity has a significant impact on the assessment of sequence quality, especially on sequence assemblies since to obtain individual whole genome assemblies or the different lineages may not be feasible even with sufficient sequence coverage. The second fundamental difference between cultured organisms and environmental samples is that natural microbial communities are usually highly diverse at multiple taxonomic levels. While several of the recently analysed communities were simple and allowed genomic and metabolic reconstruction of most of their species members, others inferred diversities in the thousands of species, without containing dominant ones (Tringe et al. 2005). What this means in terms of assembly is that the more complex a community is, the less chance there is of getting larger contigs (the raw sequencing reads are trimmed and assembled into contiguous sequences) of any single represented genomes. For example, in the case of a 150 Mbp of soil sequence data, which represented an estimated 3,000 species with no dominant ones, the largest scaffold is <10 kb and over 99% of the sequence reads do not assemble into contigs (Tringe et al. 2005). To circumvent the variability of factors affecting data processing, the overall goals of a metagenomic project have to be balanced with the diversity and structure of the community, allowing an estimation of the necessary amount of sequence to reach a desired assembly depth (Johnson and Slatkin 2006; Huson et al. 2007). Third, metagenomics confers the potential to map the metabolisms of microbes in space and time (for example see Woyke et al. 2006), whereas genome analysis is restricted to single metabolic analysis whose variation may further be studied by alternative tools.

Whatever the case, we should emphasize that more genomes need to be fully sequenced and annotated, and more genes should be characterized to serve as references for metagenomic analysis. This is a critical point, since to interpret sequences and potential genes in the environmental context is a prerequisite for being able to transform the wealth of sequence data into biological understanding. However, with every new genome or meta genome sequenced ~50% of the potential protein-encoding genes lacks any functional assignments. This huge set of functionally untapped protein diversity probably codes for determinants of specific relevant adaptations of the organisms. New catalytic mechanisms and processes can be uncovered, which could be of value for biotechnological applications. Integrating complement datasets to facilitate functional assignments of unknown genes is required. For example, by integrating data sources like environmental parameters, expression data (“metatranscriptomics”), protein level (“metaproteomics”), metabolite level (Cakir et al. 2006) and even structural information (“structural metagenomics”), the patterns of gene occurrence might emerge from an otherwise amorphous cloud of sequence data and provide the first hints of functional role (Bailly et al. 2007; Benndorf et al. 2007; Lo et al. 2007; Farber and Lusis 2008; Ferrara et al. 2008). Obviously, this analysis cannot be experimentally proven for every single gene of the 40 × 10^6 deposited in databases, but rather will require some more focussed approaches, in which bioinformatics can be used as a filter to reduce the millions of candidate genes for a given environmental function to a smaller set which may be primary targets for wet-lab experiments (Noguchi et al. 2006; Raes et al. 2007). Additionally, in order to infer the biological functions of a microbial community from sequences, a process named “binning” to group unassembled DNA sequence fragments and small contigs into biologically meaningful “bins,” may also be used (Chan et al. 2008). Nevertheless, the analysis of thousands of environmental sequences and their comparison will never be trivial and will require future appropriate repository infrastructures which range from appropriate sequence technology and assembly to gene annotation and function prediction as a whole. Following on from this, existing methods for enzymatic activity detection based on changes in spectroscopic properties should give rise to high-throughput strategies. This type of information will be extremely useful for ascribing functions to genetic sequences from environmental samples, thus minimizing annotation mistakes and suggesting biotechnological potential. Finally, in order to compare metagenomic data it should be also important to define relations among meta genomic projects and define standards. An example of this research can be found in CAMERA (Community Cyber-structure for Advanced Marine Microbiology and Ecology Research and Analysis) that started in 2006 with the release of the first set of data from the Global Ocean Survey project: 8 × 10^6 reads from 48 sample sites (http://camera.calit2.net/).

Systems understanding of microbial communities through metagenomics

The broad aim of systems microbiology, a subset of systems biology, is to gain the knowledge on the relationships between the individual components (genes, proteins, macromolecules, small molecules and cell organelles) that build a cellular organism or a community and the environment. As a complement to the long-standing trend towards reductionism, systems microbiology seeks to treat the community as a whole, integrating fundamental biological knowledge to ultimately create an integrated picture of how a microbial cell or community operates. This is not an easy task, as we know that the majority of microbes will never be cultured and the number of interactions, processes and activities in the living cells is enormous, i.e. 1 g of soil may contain up to 10^{12} interconnected putative reactions (Vieites et al. 2009). Thus, it now accepted that metagenomics can complement the toolbox of "*Systems Microbiology*" to gain a knowledge of the relationships between the individual components that build an uncultured cellular organism and its community.

For systems understanding of the functioning microbial communities through metagenomics three consecutive efforts are required (1) the isolation of a representative DNA material from a number of environments which differ in regard to the species richness and main environmental constraints to construct the libraries (or for direct sequencing), (2) the DNA sequencing, assembling and in silico annotation of these community genomes and (3) the experimental validation of gene function in the communities. The last step poses the major hurdle in metagenomics, since only a few surveys are dealing with the study of the functional composition of certain environmental niches, the majority of them based on sequence-like hypotheses (see examples by Sauer 2006; Dinsdale et al. 2008; Schloss and Handelsman 2008). For example, whole genome shotgun sequencing and metabolic pathway reconstruction revealed that the symbionts from the marine oligochaete *Olavius algarvensis*, a worm lacking a mouth, gut and nephridia, are sulphur-oxidizing and sulphate-reducing bacteria, all of which are capable of carbon fixation, thus providing the host with multiple sources of nutrition (Woyke et al. 2006). In another prominent example, new insights into other important symbiotic functions including H_2 metabolism, CO_2-reductive acetogenesis and N_2 fixation are also provided by this first system-wide gene analysis of a wood-feeding higher termite microbial community specialized towards plant ligno-cellulose degradation (Warnecke et al. 2007). Another strategy combines the extraction of shotgun sequences of microbial communities harboring few dominant organisms under dynamic utilization of different nutrients (Kalyuzhnaya et al. 2008). Accordingly, environment-specific organisms and processes catalysed by the corresponding microbe may be pinpointed. Even though some technical issues could be improved in the future, all in all, these works can be considered as a significant contribution towards the understanding the system as a whole. This power of metagenomics can examine globally how the functional composition of a given community, in terms of microbe numbers, phylogeny and catalytic activities may change in response to certain number of parameters. Nevertheless, to date, no single technique is available to reveal the functions of different taxa of microbes from a specific niche and to evaluate the ratios of the number of microbes and processes in each category that detect predominant members and thereby predominant biochemical transformations. Novel techniques that allow us a numerical description of the specific biological functions unique to specific niches and acting against particular elements are required. Here, experimental platforms for testing, dynamically analysing, surveying and visualizing whatever type of metabolic activity in metagenomic DNA of any origin which are not just based on large-scale sequence analysis, are therefore strongly demanded (Raes and Bork 2008). This should be done in a combination with extensive activity screens of metagenome libraries since the global outcome of microbial adaptation and evolution greatly depends on the ability of their protein machinery to adapt and evolve rapidly to new environmental conditions (Poelarends et al. 2006). In this respect, metabolic processes could be far more diverse than one can imagine simply counting all existing protein families. This, on the other hand, may offer the chance to discover many bioconversions which are not amenable to the existing enzymes. For this reason, studies that critically analyse and compare the mechanism and evolutionary trajectory of metagenomic enzymes are highly desired. For the purpose, primary enzyme discovery in an expression library, followed by identification of the same gene in a large insert library and genome walking on the identified fragment, constitutes a powerful means of maximizing the discovery process and identifying the interesting new organisms that are producing such enzymes.

Conclusions

Recent studies suggest that not only is the functional space of recently accessed genes far greater than that previously known from cultivated microbes, but that we have so far only explored a tiny part of the actual diversity space. This suggests that it will be far more efficient and productive to seek new data from metagenomes than to tweak the existing genomes. We started out by stating that the grand challenge of metagenomics is to access, analyse and exploit the enormous biodiversity of the microbial world. Two major bottlenecks to progress are the paucity of good high quality functional analysis of the rapidly expanding number of genomic sequences and genes being discovered, and the large number of misleading annotations in current databases, many of which are wrong and which perpetuate incorrect annotations that have many knock-on effects (in bioinformatics and systems biology, to cite some), and that cannot be corrected without experimental validation. Databases are only as useful as the quality of the data they contain; bioinformatics is only as good as the information fed into the computer. If we ignore this problem, we increasingly waste significant financial resources and staff effort. For this reason, it is furthermore clear that to access the microbes in their natural milieu there is a strong need for elaboration of a Systems biology concept that builds on the combination of multiple strategies to understand the functioning of microbial communities as a whole, with metagenomic tools playing a pivotal role.

Acknowledgments This research was supported by the Spanish MEC BIO2006-11738, CSD2007-00005, GEN2006-27750-C-4-E, CDTI I + DEA (CENIT-2007-1031) and PET2006_0052 projects. A. Beloqui thanks the Spanish MEC for a FPU fellowship. P. N. Golyshin acknowledges the Grant 0313751 K from the Federal Ministry for Science and Education (BMBF) within the GenoMikPlus initiative.

References

Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, Chan AM, Haynes M, Kelley S, Liu H, Mahaffy JM, Mueller JE, Nulton J, Olson R, Parsons R, Rayhawk S, Suttle CA, Rohwer F (2006) The marine viromes of four oceanic regions. PLoS Biol 11:e368. doi:10.1371/journal.pbio.0040368

Bailly J, Fraissinet-Tachet L, Verner MC, Debaud JC, Lemaire M, Wésolowski-Louvel M, Marmeisse R (2007) Soil eukaryotic functional diversity, a metatranscriptomic approach. ISME J 7:632–642. doi:10.1038/ismej.2007.68

Beloqui A, de María PD, Golyshin PN, Ferrer M (2008) Recent trends in industrial microbiology. Curr Opin Microbiol 11:240–248. doi:10.1016/j.mib.2008.04.005

Bench SR, Hanson TE, Williamson KE, Ghosh D, Radosovich M, Wang K, Wommack KE (2007) Metagenomic characterization of Chesapeake Bay virioplankton. Appl Environ Microbiol 23:7629–7641. doi:10.1128/AEM.00938-07

Benndorf D, Balcke GU, Harms H, von Bergen M (2007) Functional metaproteome analysis of protein extracts from contaminated soil and groundwater. ISME J 3:224–234. doi:10.1038/ismej.2007.39

Bentley DR (2006) Whole-genome re-sequencing. Curr Opin Genet Dev 16:545–552. doi:10.1016/j.gde.2006.10.009

Biddle JF, Fitz-Gibbon S, Schuster SC, Brenchley JE, House CH (2008) Metagenomic signatures of the Peru Margin subseafloor biosphere show a genetically distinct environment. Proc Natl Acad Sci USA 105:10583–10588. doi:10.1073/pnas.0709942105

Bodrossy L, Stralis-Pavese N, Konrad-Köszler M, Weilharter A, Reichenauer TG, Schöfer D, Sessitsch A (2006) mRNA-based parallel detection of active methanotroph populations by use of a diagnostic microarray. Appl Environ Microbiol 72:1672–1676. doi:10.1128/AEM.72.2.1672-1676.2006

Cakir T, Patil KR, Onsan Z, Ulgen KO, Kirdar B, Nielsen J (2006) Integration of metabolome data with metabolic networks reveals reporter reactions. Mol Syst Biol 2:50. doi:10.1038/msb4100085

Chan CK, Hsu AL, Tang SL, Halgamuge SK (2008) Using growing self-organising maps to improve the binning process in environmental whole-genome shotgun sequencing. J Biomed Biotechnol 2008:513701. doi:10.1155/2008/513701

Church JM (2005) The personal genome project. Mol Syst Biol 1:20050030

Cox-Foster DL, Conlan S, Holmes EC, Palacios G, Evans JD, Moran NA, Quan PL, Briese T, Hornig M, Geiser DM, Martinson V, vanEngelsdorp D, Kalkstein AL, Drysdale A, Hui J, Zhai J, Cui J, Hutchison L, Simons JF, Egholm M, Pettis JS, Lipkin WI (2007) A metagenomic survey of microbes in honey bee colony collapse disorder. Science 318:283–287. doi:10.1126/science.1146498

DeLong EF (2005) Microbial community genomics in the ocean. Nat Rev Microbiol 3:459–469. doi:10.1038/nrmicro1158

DeLong EF (2006) Archaeal mysteries of the deep revealed. Proc Natl Acad Sci USA 103:6417–6418. doi:10.1073/pnas.0602079103

Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T, Haynes M, Liu H, Furlan M, Wegley L, Chau B, Ruan Y, Hall D, Angly FE, Edwards RA, Li L, Thurber RV, Reid RP, Siefert J, Souza V, Valentine DL, Swan BK, Breitbart M, Rohwer F (2008) Biodiversity and biogeography of phages in modern stromatolites and thrombolites. Nature 452:340–343. doi:10.1038/nature06735

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, McDaniel L, Moran MA, Nelson KE, Nilsson C, Olson R, Paul J, Brito BR, Ruan Y, Swan BK, Stevens R, Valentine DL, Thurber RV, Wegley L, White BA, Rohwer F (2008) Functional metagenomic profiling of nine biomes. Nature 452:629–632. doi:10.1038/nature06810

Edwards RA, Rodriguez-Brito B, Wegley L, Haynes M, Breitbart M, Peterson DM, Saar MO, Alexander S, Alexander EC Jr, Rohwer F (2006) Using pyrosequencing to shed light on deep mine microbial ecology. BMC Genomics 7:57. doi:10.1186/1471-2164-7-57

Eisen JA (2007) Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. PLoS Biol 5:e82. doi:10.1371/journal.pbio.0050082

Farber CR, Lusis AJ (2008) Integrating global gene expression analysis and genetics. Adv Genet 60:571–601. doi:10.1016/S0065-2660(07)00420-8

Ferrara CT, Wang P, Neto EC, Stevens RD, Bain JR, Wenner BR, Ilkayeva OR, Keller MP, Blasiole DA, Kendziorski C, Yandell BS, Newgard CB, Attie AD (2008) Genetic networks of liver metabolism revealed by integration of metabolic and transcriptional profiling. PLoS Genet 3:e1000034. doi:10.1371/journal.pgen.1000034

Ferrer M, Golyshina O, Beloqui A, Golyshin PN (2007) Mining enzymes from extreme environments. Curr Opin Microbiol 10:207–214. doi:10.1016/j.mib.2007.05.004

Ferrer M, Beloqui A, Timmis KN, Golyshin PN (2008) Metagenomics for mining new genetic resources of microbial communities. J Mol Microbiol Biotechnol 16:109–123. doi:10.1159/000142898

Ferrer M, Beloqui A, Vieites JM, Guazzaroni ME, Berger I, Aharoni A (2009) Interplay of metagenomics and in vitro compartmentalization. Microbiol Biotechnol 2:31–39. doi:10.1111/j.1751-7915.2008.00057.x

Fierer N, Breitbart M, Nulton J, Salamon P, Lozupone C, Jones R, Robeson M, Edwards RA, Felts B, Rayhawk S, Knight R, Rohwer F, Jackson RB (2007) Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. Appl Environ Microbiol 73:7059–7066. doi:10.1128/AEM.00358-07

Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW, Delong EF (2008) Microbial community gene expression in ocean surface waters. Proc Natl Acad Sci USA 10:3805–3810. doi:10.1073/pnas.0708897105

García-Martín H, Ivanova N, Kunin V, Warnecke F, Barry KW, McHardy AC, Yeates C, He S, Salamov AA, Szeto E, Dalin E, Putnam NH, Shapiro HJ, Pangilinan JL, Rigoutsos I, Kyrpides NC, Blackall LL, McMahon KD, Hugenholtz P (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. Nat Biotechnol 24:1229–1230. doi:10.1038/nbt1006-1229

Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE (2006) Metagenomic analysis of the human distal gut microbiome. Science 312:1355–1359. doi:10.1126/science.1124234

Hallin PF, Binnewies TT, Ussery DW (2008) The genome BLAST-atlas-a GeneWiz extension for visualization of whole-genome homology. Mol Biosyst 5:363–371. doi:10.1039/b717118h

Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. Mol Biol Rep 68:669–685

Handelsman J (2008) Metagenomics is not enough. DNA Cell Biol 27:219–221. doi:10.1089/dna.2008.1503

Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol 5:245–249. doi:10.1016/S1074-5521(98)90108-9

Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Res 3:377–386. doi:10.1101/gr.5969107

Ingham CJ, Sprenkels A, Bomer J, Molenaar D, van den Berg A, van Hylackama Vlieg JE, de Vos WM (2007) The micro-petri dish, a million-well growth chip for the culture and high-throughput screening of microorganisms. Proc Natl Acad Sci USA 104:18217–18222. doi:10.1073/pnas.0701693104

Ishoey T, Woyke T, Stepanauskas R, Novotny M, Lasken RS (2008) Genomic sequencing of single microbial cells from environmental samples. Curr Opin Microbiol 11:198–204. doi:10.1016/j.mib.2008.05.006

Johnson PLF, Slatkin M (2006) Inference of population genetic parameters in metagenomics: a clean look at messy data. Genome Res 16:1320–1327. doi:10.1101/gr.5431206

Kalyuzhnaya MG, Lapidus A, Ivanova N, Copeland AC, McHardy AC, Szeto E, Salamov A, Grigoriev IV, Suciu D, Levine SR, Markowitz VM, Rigoutsos I, Tringe SG, Bruce DC, Richardson PM, Lidstrom ME, Chistoserdova L (2008) High-resolution metagenomics targets specific functional types in complex microbial communities. Nat Biotechnol 26:1029–1034. doi:10.1038/nbt.1488

Krause L, Diaz NN, Goesmann A, Kelley S, Nattkemper TW, Rohwer F, Edwards RA, Stoye J (2008) Phylogenetic classification of short environmental DNA fragments. Nucleic Acids Res 36:2230–2239. doi:10.1093/nar/gkn038

Lasken RS (2007) Single-cell genomic sequencing using multiple displacement amplification. Curr Opin Microbiol 10:510–516. doi:10.1016/j.mib.2007.08.005

Leininger S, Urich T, Schloter M, Schwark L, Qi J, Nicol GW, Prosser JI, Schuster SC, Schleper C (2006) Archaea predominate among ammonia-oxidizing prokaryotes in soils. Nature 442:806–809. doi:10.1038/nature04983

Liolios K, Mavromatis K, Tavernarakis N, Kyrpides NC (2008) The genomes on line database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 36:D475–D479. doi:10.1093/nar/gkm884

Lo I, Denef VJ, Verberkmoes NC, Shah MB, Goltsman D, DiBartolo G, Tyson GW, Allen EE, Ram RJ, Detter JC, Richardson P, Thelen MP, Hettich RL, Banfield JF (2007) Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. Nature 7135:537–541. doi:10.1038/nature05624

Lombardot T, Kottmann R, Pfeffer H, Richter M, Teeling H, Quast C, Glöckner FO (2006) Megx.net-database resources for marine ecological genomics. Nucleic Acids Res 34:390–393. doi:10.1093/nar/gkj070

Lozupone CA, Knight R (2008) Species divergence and the measurement of microbial diversity. FEMS Microbiol Rev 32:557–578. doi:10.1111/j.1574-6976.2008.00111.x

Marcy Y, Ouverney C, Bik EM, Lösekann T, Ivanova N, Martin HG, Szeto E, Platt D, Hugenholtz P, Relman DA, Quake SR (2007) Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. Proc Natl Acad Sci USA 104:11889–11894. doi:10.1073/pnas.0704662104

Margulies EH, NISC Comparative Sequencing Program, Maduro VV, Thomas PJ, Tomkins JP, Amemiya CT, Luo M, Green ED (2005) Comparative sequencing provides insights about the structure and conservation of marsupial and monotreme genomes. Proc Natl Acad Sci USA 102:3354–3359. doi:10.1073/pnas.0408539102

McHardy AC, Rigoutsos I (2007) What's in the mix: phylogenetic classification of metagenome sequence samples. Curr Opin Microbiol 10:449–503

Noguchi H, Park J, Takagi T (2006) MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. Nucleic Acids Res 34:5623–5630. doi:10.1093/nar/gkl723

Poelarends GJ, Almrud JJ, Serrano H, Darty JE, Johnson WH Jr, Hackert ML, Whitman CP (2006) Evolution of enzymatic activity in the tautomerase superfamily: mechanistic and structural consequences of the L8R mutation in 4-oxalocrotonate tautomerase. Biochemistry 45:7700–7708. doi:10.1021/bi0600603

Raes J, Bork P (2008) Molecular eco-systems biology: towards an understanding of community function. Nat Rev Microbiol 6:693–699. doi:10.1038/nrmicro1935

Raes J, Foerstner KU, Bork P (2007) Get the most out of your metagenome: computational analysis of environmental sequence data. Curr Opin Microbiol 10:490–498. doi:10.1016/j.mib.2007.09.001

Riesenfeld CS, Goodman RM, Handelsman J (2004) Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. Environ Microbiol 6:981–989. doi:10.1111/j.1462-2920.2004.00664.x

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S et al (2007) The Sorcerer II global ocean sampling expedition: northwest. Atlantic through eastern tropical Pacific. PLoS Biol 5:398–431. doi:10.1371/journal.pbio.0050077

Sauer U (2006) Metabolic networks in motion: ^{13}C-based flux analysis. Mol Syst Biol 2:62. doi:10.1038/msb4100109

Schloss PD, Handelsman J (2006) Toward a census of bacteria in soil. PLOS Comput Biol 7:e92. doi:10.1371/journal.pcbi.0020092

Schloss PD, Handelsman JA (2008) A statistical toolbox for metagenomics: assessing functional diversity in microbial communities. BMC Bioinformatics 9:34. doi:10.1186/1471-2105-9-34

Schmeisser C, Steele H, Streit WR (2007) Metagenomics, biotechnology with non-culturable microbes. Appl Microbiol Biotechnol 75:955–962. doi:10.1007/s00253-007-0945-5

Shendure J, Mitra RD, Varma C, Church GM (2004) Advanced sequencing technologies: methods and goals. Nat Rev Genet 5:335–344. doi:10.1038/nrg1325

Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, Arrieta JM, Herndl GJ (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci USA 103:12115–12120. doi:10.1073/pnas.0605127103

Staley JT, Konopka A (1985) Measurements of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu Rev Microbiol 39:321–346. doi:10.1146/annurev.mi.39.100185.001541

Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM (2005) Comparative metagenomics of microbial communities. Science 308:554–557. doi:10.1126/science.1107851

Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444:1027–1031. doi:10.1038/nature05414

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37–43. doi:10.1038/nature02340

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66–74. doi:10.1126/science.1093857

Vieites JM, Guazzaroni ME, Beloqui A, Golyshin PN, Ferrer M (2009) Metagenomics approaches in systems microbiology. FEMS Microbiol Rev 33:236–255. doi:10.1111/j.1574-6976.2008.00152.x

Warnecke F, Luginbühl P, Ivanova N, Ghassemian M, Richardson TH, Stege JT, Cayouette M, McHardy AC, Djordjevic G, Aboushadi N, Sorek R, Tringe SG, Podar M, Martin HG, Kunin V, Dalevi D, Madejska J, Kirton E, Platt D, Szeto E, Salamov A, Barry K, Mikhailova N, Kyrpides NC, Matson EG, Ottesen EA, Zhang X, Hernández M, Murillo C, Acosta LG, Rigoutsos I, Tamayo G, Green BD, Chang C, Rubin EM, Mathur EJ, Robertson DE, Hugenholtz P, Leadbetter JR (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. Nature 450:560–565. doi:10.1038/nature06269

Wegley L, Edwards R, Rodriguez-Brito B, Liu H, Rohwer F (2007) Metagenomic analysis of the microbial community associated with the coral *Porites astreoides*. Environ Microbiol 9:2707–2719. doi:10.1111/j.1462-2920.2007.01383.x

Williamson SJ, Rusch DB, Yooseph S, Halpern AL, Heidelberg KB, Glass JI, Andrews-Pfannkoch C, Fadrosh D, Miller CS, Sutton G, Frazier M, Venter JC (2008) The Sorcerer II global ocean sampling expedition: metagenomic characterization of viruses within aquatic microbial samples. PLoS ONE 1:e1456. doi:10.1371/journal.pone.0001456

Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, Gloeckner FO, Boffelli D, Anderson IJ, Barry KW, Shapiro HJ, Szeto E, Kyrpides NC, Mussmann M, Amann R, Bergin C, Ruehland C, Rubin EM, Dubilier N (2006) Symbiosis insights through metagenomic analysis of a microbial consortium. Nature 443:950–955. doi:10.1038/nature05192

Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, Jaroszewski L, Cieplak P, Miller CS, Li H, Mashiyama ST, Joachimiak MP, van Belle C, Chandonia JM, Soergel DA, Zhai Y, Natarajan K, Lee S, Raphael BJ, Bafna V, Friedman R, Brenner SE, Godzik A, Eisenberg D, Dixon JE, Taylor SS, Strausberg RL, Frazier M, Venter JC (2007) The Sorcerer II global ocean sampling expedition: expanding the universe of protein families. PLoS Biol 5:e16. doi:10.1371/journal.pbio.0050016

Zhang T, Breitbart M, Lee WH, Run JQ, Wei CL, Soh SW, Hibberd ML, Liu ET, Rohwer F, Ruan Y (2006) RNA viral community in human feces: prevalence of plant pathogenic viruses. PLoS Biol 4:e3. doi:10.1371/journal.pbio.0040003