Metagenomics – the key to the uncultured microbes
Wolfgang R Streit*1,2 and Ruth A Schmitz1

It is widely accepted that up to 99.8% of the microbes present in many environments are not readily culturable. ‘Metagenome technology’ tries to overcome this bottleneck by developing and using culture-independent approaches. From the outset, metagenome-based approaches have led to the accumulation of an increasing number of DNA sequences, but until this time the sequences retrieved have been those of uncultured microbes. These genomic sequences are currently exploited for novel biotechnological and pharmaceutical applications and to increase our knowledge on microbial ecology and physiology of these microbes. Using the metagenome sequences to fully understand how complex microbial communities function and how microbes interact within these niches represents a major challenge for microbiologists today.

Addresses
1 Institut für Mikrobiologie und Genetik, Universität Göttingen, Grisebachstr. 8, 37077, Göttingen, Germany
2 Institut für Grenzflächenbiotechnologie, Molekulare Enzymtechnologie, Universität Duisburg-Essen, Geibelstr. 41, 47057 Duisburg, Germany
‘e-mail: wstreit@gwdg.de

Introduction
Current estimates indicate that more than 99% of the microorganisms present in many natural environments are not readily culturable and therefore not accessible for biotechnology or basic research [1]. In fact, most of the species in many environments have never been described, and this situation will not change until new culture technologies are developed. Additionally, many approaches currently used to explore the diversity and potential of microbial communities are biased because of the limitations of cultivation methods. To overcome the difficulties and limitations associated with cultivation techniques, several DNA-based molecular methods have been developed. In general, methods based on 16S rRNA gene analysis provide extensive information about the taxa and species present in an environment. However, these data usually provide only little if any information about the functional role of the different microbes within the community and the genetic information they contain. By contrast, metagenomics is a new and rapidly developing field, which tries to analyze the complex genomes of microbial niches.

In this review, we discuss the major strategies and working steps used to identify novel genes from uncultured microbes and complex microbial assemblages. Further, we highlight the usefulness of these novel genes for biotechnological or pharmaceutical applications and recent genome observations in metagenomic research. We close with a review on key discoveries from metagenome findings, starting with the initial outline of the concept [2], and their impact on our understanding of fundamental biogeochemical cycles in microbial niches.

Metagenome technologies, DNA extraction, library construction, screening
Metagenome analyses are usually initiated by the isolation of environmental DNAs (Figure 1). A major difficulty associated with the metagenome approach is related to the contamination of purified DNA with polyphenolic compounds, which are copurified with the DNA. These compounds are difficult to remove, and it is well known that polyphenols also interfere with enzymatic modifications of the isolated DNA [3]. In addition, for marine samples often large amounts of water have to be filtered to obtain sufficient DNA for cloning [4]. As a result, the construction of environmentally derived DNA libraries with large inserts is often hindered because of the poor quality of the isolated DNA. These known difficulties associated with the construction of libraries directly derived from environmental DNA samples have forced many laboratories to isolate DNA from the metagenome of a microbial community after pre-cultivation in the laboratory. Although it is expected that laboratory enrichment cultures bear only a limited biodiversity this technique has proven to be highly efficient for the rapid isolation of large DNA fragments and for cloning of operons and genes with high biotechnological value [5*–6–8].

DNA isolation and purification is followed by the construction of DNA libraries in suitable cloning vectors and host strains. The classical approach includes the construction of small insert libraries (< 10 kb) in a standard sequencing vector and in Escherichia coli as a host strain [9]. However, small insert libraries do not allow detection of large gene clusters or operons. To circumvent this limitation researchers have been employing large insert libraries, such as cosmid DNA libraries (mostly in the pWE15 vector of Stratagene) with insert sizes ranging...
from 25–35 kb [7] and/or bacterial artificial chromosome (BAC) libraries with insert sizes up to almost 200 kb [4,10]. Additionally, the construction of fosmid libraries with inserts of 40 kb of foreign DNA has been reported [11]. *E. coli* is still the preferred host for the cloning and expression of any metagenome-derived genes. Only very recently have other hosts such as *Streptomyces lividans* been employed to identify genes involved in the biosynthesis of novel antibiotics [12]. Also, to our knowledge metagenomic libraries are currently developed in other Gram-negative hosts by several laboratories working in the field, which will become available soon. The development of these new host strains will increase screening efficiency tremendously.

Functional searches for novel genes in metagenomic libraries have often been performed using highly sophisticated picking and pipetting robots. In many of the recent publications large clone-libraries have been screened. Often several hundred thousand clones have been analyzed to detect less than ten active clones in a single screen [9,13,14]. This is mainly owing to a lack of efficient transcription of the metagenome-derived genes in the host strain. This effect might be worsened by a weak translation in combination with a poor secretion of the foreign protein by the employed host strain. Also, in many cases it can be expected that the desired protein is not folded correctly because required chaperones are absent in the host strain. Furthermore, cofactors might

The identification of novel genes and the analysis of complete metagenomes from microbial communities. Analyses of complete metagenomes involves at least four major working steps: (a) DNA isolation from microbial niches (i) soils, (ii) sediments, or (iii) laboratory enrichments; (b) construction of DNA libraries; (c) mining for clones and DNA sequences of interest and; (d) accumulation of desired clones and DNA sequences. Metagenomics thereby pursues two major goals: the identification of novel biocatalyst genes and increasing our understanding on microbial ecology.

from 25–35 kb [7] and/or bacterial artificial chromosome (BAC) libraries with insert sizes up to almost 200 kb [4,10]. Additionally, the construction of fosmid libraries with inserts of 40 kb of foreign DNA has been reported [11]. *E. coli* is still the preferred host for the cloning and expression of any metagenome-derived genes. Only very recently have other hosts such as *Streptomyces lividans* been employed to identify genes involved in the biosynthesis of novel antibiotics [12]. Also, to our knowledge metagenomic libraries are currently developed in other Gram-negative hosts by several laboratories working in the field, which will become available soon. The development of these new host strains will increase screening efficiency tremendously.

Functional searches for novel genes in metagenomic libraries have often been performed using highly sophisticated picking and pipetting robots. In many of the recent publications large clone-libraries have been screened. Often several hundred thousand clones have been analyzed to detect less than ten active clones in a single screen [9,13,14]. This is mainly owing to a lack of efficient transcription of the metagenome-derived genes in the host strain. This effect might be worsened by a weak translation in combination with a poor secretion of the foreign protein by the employed host strain. Also, in many cases it can be expected that the desired protein is not folded correctly because required chaperones are absent in the host strain. Furthermore, cofactors might

The identification of novel genes and the analysis of complete metagenomes from microbial communities. Analyses of complete metagenomes involves at least four major working steps: (a) DNA isolation from microbial niches (i) soils, (ii) sediments, or (iii) laboratory enrichments; (b) construction of DNA libraries; (c) mining for clones and DNA sequences of interest and; (d) accumulation of desired clones and DNA sequences. Metagenomics thereby pursues two major goals: the identification of novel biocatalyst genes and increasing our understanding on microbial ecology.

from 25–35 kb [7] and/or bacterial artificial chromosome (BAC) libraries with insert sizes up to almost 200 kb [4,10]. Additionally, the construction of fosmid libraries with inserts of 40 kb of foreign DNA has been reported [11]. *E. coli* is still the preferred host for the cloning and expression of any metagenome-derived genes. Only very recently have other hosts such as *Streptomyces lividans* been employed to identify genes involved in the biosynthesis of novel antibiotics [12]. Also, to our knowledge metagenomic libraries are currently developed in other Gram-negative hosts by several laboratories working in the field, which will become available soon. The development of these new host strains will increase screening efficiency tremendously.

Functional searches for novel genes in metagenomic libraries have often been performed using highly sophisticated picking and pipetting robots. In many of the recent publications large clone-libraries have been screened. Often several hundred thousand clones have been analyzed to detect less than ten active clones in a single screen [9,13,14]. This is mainly owing to a lack of efficient transcription of the metagenome-derived genes in the host strain. This effect might be worsened by a weak translation in combination with a poor secretion of the foreign protein by the employed host strain. Also, in many cases it can be expected that the desired protein is not folded correctly because required chaperones are absent in the host strain. Furthermore, cofactors might
not be synthesized by the foreign host strain and/or not
inserted correctly into the recombinant metagenomic
protein. Finally, a different codon usage could be a reason
for poor protein expression and low activities. Several
laboratories are currently working on solutions for some of
these problems and constructing novel vectors and
strains. Other avenues to pursue include the development
of highly sensitive screening methods using fluorogenic
substrates.

Large-scale sequencing projects such as the one initiated
by Craig Venter for the metagenome of the Sargasso Sea
resulted in the identification of numerous novel genes
and is a very famous example of sequence-based meta-
genome analyses [15]. Similar approaches have been
initiated by European laboratories to sequence complete
or partial metagenomes of a phylogenetically highly
diverse biofilm [16]. Complementary to this, Tyson
and coworkers described the nearly complete sequence
analyses of the metagenome of an acidophilic biofilm,
which was characterized by a low biodiversity [17].
A completely novel strategy has been developed by search-
ing through metagenome DNA sequences and by using
microarray technology [18]. The microarray profiling
offers an effective approach for rapidly identifying and
characterizing many clones. It could potentially be used
for the isolation of large numbers of conserved genes.
Similarly, other laboratories have been using degenerated
PCR primers for the isolation of metagenome-derived
genes. Among these are genes involved in the degrada-
tion of α-halocarboxylic acids [19] and genes encoding for
novel hydrolases [20].

Metagenomics — a key technology for
the development of novel biotechnological and
pharmaceutical products

Because of the overwhelming majority of non-cultured
microbes in most microbial niches, metagenome search-
es will always result in the identification of hitherto
unknown genes and proteins. Thus, the probability of
uncovering novel sequences makes this approach more
favorable than searches in already cultivated microbes.
Since its introduction, metagenomics has identified a
significant number of novel genes encoding for biocata-
lysts or molecules with high potential for use in phar-
caceutical products or production processes (Figure 1).
Interestingly, searches in metagenome-derived DNA
libraries have mainly focused on a rather small group
of enzymes. Among these are lipases and esterases
[5,13,16,20]. Lipases and esterases are of importance
for biotechnological applications because many enzymes
of this class remain active in organic solvents; they
usually display exquisite chemo-, regio-, and stereoselec-
tivities and they do not require cofactors [21–23].
Oxidoreductases are another example of useful catalysts
with a high enantioselectivity that have been identified
in metagenome searches [6,9,24,25]. Oxidoreductases
are frequently used for the synthesis of carbonyl com-
pounds, hydroxy acids, amino acids and chiral alcohols,
which are often difficult to prepare [26]. Of particular
interest, nicotinamide-dependent alcohol reductases
are employed for the preparation of deuterium or tritium
labeled compounds, production of dihydroxyacetone
and as tools for enzymatic analysis of serum lipids
[27]. Similarly, polysaccharide-modifying enzymes such as
the starch modifying enzymes are of considerable
interest to industry. Therefore, a significant number of
metagenome searches have identified polysaccharide-
modifying genes [5,8,10,16,28,29]. Further, the iso-
lation of enzymes useful for the production of bulk che-
icals [30], proteases [31] and nitrilases [32] has been
reported. Metagenome searches have also focused on the
isolation of genes involved in vitamin biosynthesis.
Interest in 2,5-diketo-d-gluconic acid synthesis is related
to a new biotechnological process for the production of
vitamin C (ascorbic acid) using glucose as a substrate
[25]; and interest in biotin biosynthesis genes is linked to
the construction of biotin overproducing bacteria for
large-scale fermentation of this vitamin [33]. Finally,
the isolation of genes encoding for novel therapeutic
molecules is a very valuable area of research [34–38].
The genes of interest within these searches are often
type I and type II polyketide synthases (β-ketoacyl
synthetases) [12,39], which are key genes involved in
the synthesis of polyketide antibiotics by Gram-positive
Streptomyces and are often part of large biosynthetic gene
clusters [40].

Recent discoveries on the physiology of
non-cultured microbes and their significance
for biogeochemical cycles

Sequence analysis of large insert libraries with environ-
mental DNA combined with genetic and functional anal-
ysis has the potential to provide significant insight into
the genomic potential and ecological roles of cultured
and uncultured microbes (Figure 1). The importance of this
potential for understanding complex environments can
be estimated by the following six very recent examples.
First, bacteriorhodopsins capable of generating a chemi-
omsotic membrane potential in response to light have
been demonstrated only for halophilic archaea [41]. How-
ever, recent analysis of genome fragments recovered
directly from marine bacterioplankton suggested the pre-
sence of a new bacterial rhodopsin, proteorhodopsin
[4,42]. Biochemical and biophysical analyses of this y-
proteobacterial rhodopsin protein expressed heterolo-
gously in E. coli and analyses of the native protein present
in ocean surface waters demonstrated its ability to func-
tion as a light-driven proton pump. The discovered wide-
spread distribution of proteorhodopsin genes among
divergent marine bacterial taxa [42], the high abundance
and spectra adaptation of proteorhodopsin proteins to
different habitats combined with the genetic and biophys-
ical data indicate that proteorhodopsin-based bacterial
phototrophy is a globally significant oceanic microbial process [4,11,43–46].

Second, culture-independent partial or nearly complete recovery of microbial genomes from an environmental sample by an extended random shotgun-sequencing approach offers a highly intriguing approach to study natural microbial communities. A recent example gave significant insights into the community structure and the metabolism of a natural acidophilic biofilm growing on the surface of a flowing acid mine drainage [17**]. This was mainly possible through reconstruction of the microbial genomes present in this niche. For this purpose the near-complete genomes of Leptospirillum group II and Ferroplasma type II were reconstructed. A detailed analysis of these genomes allowed pathway reconstruction of carbon fixation and energy generation and provided insights into survival strategies in the extreme acidic environment; for example, genes for biosynthesis of isoprenoid-based lipids and for a variety of proton efflux systems were identified. However, this information has to be confirmed by biochemical and biophysical approaches.

Third, another field for which significant insights in structure, metabolism and gene expression of mixed microbial communities are of high medical interest are biofilms of indwelling medical devices. Recently, Wenderoth and colleagues [47] investigated the community composition of biliary stent biofilms by a cultivation-independent approach, i.e. PCR amplification of 16S rRNA/rDNA, SSCP (single-stranded conformation polymorphism) fingerprinting of the amplicons and sequencing the major SSCP species. Using this approach, novel uncultured bacteria were identified as the major members of biliary stent biofilms and a sequential colonization process of the stent surface was observed, with Pseudomonas aeruginosa being the pioneer colonizer followed by Klebsiella pneumoniae and an uncultured microbe [47].
Fourth, another example is the analysis of the complex metagenome of model biofilms growing on rubber-coated valves within drinking water networks using the cultivation-independent approach. This analysis revealed significant insights into phylogenetic, catabolic and metabolic abilities of the analyzed microbial community [16*]. Concerning the potential health risk of the studied biofilm, no DNA or protein sequences directly linked to pathogenic traits were identified.

Fifth, the major sequencing approach of the Sargasso Sea has identified a by far greater number of DNA sequences than described in any of the above cited references [15*,48]. This study is most likely to open up the possibility of new investigations of the microbial world and of its evolution. However, it will require new bioinformatics tools and it also highlights how generalist databases such as Genbank or EMBL are not adapted to the overflow of new sequences. In brief, the functions of these many novel sequences and their significance for the ecosystems need to be elucidated over the next decade.

Sixth, most recently mixed microbial communities of interest have been investigated by monitoring gene expression using DNA microarrays [49,50]. This approach however, identifies only those genes transcribed in the analyzed community and is therefore very valuable for the identification of functional genes with relevance to the ecosystem. Though several challenges still remain to be overcome to most effectively apply microarray technology to monitoring gene expression in complex microbial ecosystems (e.g., increasing the sensitivity), the potential and capability of this promising in situ technology appears to be highly attractive.

Finally, a completely different but highly intriguing approach was recently reported by Kruger et al. [51**]. Cultivation-independent biochemical studies of microbial mats of a northwestern Black Sea shelf, which oxidize methane under anaerobic conditions, resulted in the identification of a prominent nickel-containing protein. Biochemical analyses of this nickel protein isolated from the natural system and its abundance in the methane-oxidizing mat (7% of extracted proteins) indicated that it is likely to catalyze the crucial step in anaerobic methane oxidation (methane activation). This approach is not only remarkable because of the biochemical and ecological findings but also because it first uncovered the function of the novel protein and subsequently applied metagenome technologies to determine the corresponding gene.

Conclusions

It can be expected that the number of novel genes identified through metagenome technologies will exceed the number of genes identified through sequencing individual microbes (Figure 2). However, the challenge now not only lies in accumulating these sequences but more-over in understanding the function of these novel genes and proteins within the microbial niches and their role in the global cycles.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

* of special interest
** of outstanding interest


This publication is remarkable because of the enormous number of DNA sequences generated and the number of novel genes detected.


This is a landmark publication as it is the first report of partial metagenome sequencing of a phylogenetically highly diverse microbial community. Together with the random snap shot sequences the authors identified more than 100 kb of continuous environmental DNA sequences.


This is one of the first publications to present a near-complete set of metagenomic sequences from a mixed but low diversity microbial community.


This publication is remarkable as it identifies the functional role of a novel protein involved in the global carbon (methane) cycle and it identifies the DNA sequence only subsequently.