Trine Aakvik

New broad-host-range Biological Tools for Metagenomics and Screening of Marine DNA Libraries

Thesis for the degree of Philosophiae Doctor

Trondheim, January 2011

Norwegian University of Science and Technology
Faculty of Natural Sciences and Technology
Department of Biotechnology

NTNU – Trondheim
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the degree of Philosophiae Doctor

Faculty of Natural Sciences and Technology
Department of Biotechnology

ISBN 978-82-471-2604-2 (printed ver.)
ISBN 978-82-471-2605-9 (electronic ver.)
ISSN 1503-8181

Doctoral theses at NTNU, 2011:41

Printed by NTNU-trykk

Acknowledgements

The work presented in this thesis has been carried out at the Department of Biotechnology at the Norwegian University of Technology and Science (NTNU), and has been supported by funding from the Research Council of Norway.

I would like to express my deepest gratitude to my supervisor Svein Valla: Your excellent guidance, motivation and positive attitude throughout these years have been very inspiring. Thanks for all your constructive advices and for sharing your scientific knowledge with me, and also for all fruitful discussions.

I am also very grateful to all who have contributed to the work presented in this thesis: Kristin Degnes for keeping the project running during my first maternity leave, Rannveig Dahlsrud for constructing the pRS44 cloning vector, Geir Klinkenberg and Asgeir Winnberg for all help with the robotic screens, and to all the co-authors of the papers in this thesis for their different contributions. A special thank to Anna Lewin for also proof-reading parts of the thesis, and to Mark Liles at the Auburn University for many helpful advices.

To all my current and former co-workers, thanks for all helpful advices during this period and for creating a very good working environment. I have in particular enjoyed our nice coffee-breaks ☺

To my family, thank you for your encouragement and love, and for all help in busy periods. And finally to my three boys at home, thanks for bringing the most important perspectives into my life. Thank you, Lasse, for your encouraging support at times of frustration, and for your everlasting love and care.

Trondheim, January 2011
Trine Aakvik

Abstract

The fact that only a minority of the microorganisms in environmental samples can be cultivated in the laboratory necessitates the use of metagenomics to better explore and exploit the huge source of genetic diversity represented by these organisms. Metagenomic sequencing studies allow for characterization of the microbial community within a selected habitat, whereas functional screening of metagenomic libraries is accomplished with the aim to detect novel biomolecules. The latter approach has the advantage that it does not rely on any sequence-homology to known genes, but it is dependent on successful heterologous expression of environmental-originating genes within the library host.

As the chances for heterologous expression increases when using many different species as hosts for such libraries, the initial objective of the work presented here was to establish a broad-host-range vector system suitable for such metagenomic studies. The resulting RK2-based vector pRS44 is a combined fosmid and BAC vector that can be used for large-insert library construction followed by efficient transfer of intact library clones to numerous hosts through conjugation. The copy number of pRS44 is adjustable meaning that gene expression levels can easily be modified.

It was observed that large plasmids were occasionally modified after conjugal transfer when using the traditional conjugation donor strain *Escherichia coli* S17-1. Analysis of these occurrences revealed that they most probably result from co-transfer of DNA from the chromosomally located *oriT,* followed by homologous recombination events in the recipient host. An improved mobilization system that does not contain a functional *oriT* has therefore been established, and this system also circumvent other problems that have been reported for the S17-1 donor system. It is in addition more flexible as the mobilization helper plasmid is replicating extra-chromosomally through the broad-host-range replicon pBBR1, meaning that it can be transferred to different strains and species for the establishment of new donors.

The pRS44 vector was used to establish a 20000 member metagenomic library using DNA isolated from marine sediments as cloning material. This library was screened mainly for pigment- and antibiotic production using both sequence-based and function-based approaches, the latter on library clones of both *E. coli* and *P. fluorescens*. Characterization of a metagenomic clone producing a pigmented compound, both at gene and product level, revealed that the pigment was a melanin-like polymer and the responsible gene product was very similar to the oxidoreductase rubrerythrin. No connections have to our knowledge earlier been made between this enzyme-class and melanins.

This PhD thesis also includes some work connected to a sequence-based study on metagenomic DNA isolated from a Norwegian oil reservoir. The main findings in this study were that the sequenced DNA was unexpectedly similar to DNA isolated from more accessible locations, and that the dominating microorganisms within this reservoir were sulphur-reducing bacteria and methanogens (archaea). All analyses indicate that the sequence data represents the community at high coverage. As a fraction of the sequences showed similarities to mesophilic species, a gene sequence annotated to encode an enolase from the mesophilic specie *Pelobacter carbinolicus* was synthesized and expressed in *E. coli*. The activity of this enolase proved to be more temperature tolerant than the corresponding *E. coli* enzyme, and it was therefore assumed that the sequence rather originate from a thermophilic *Pelobacter*-like specie.

List of papers

Paper I
Aakvik, T., Degnes, K.F., Dahlsrud, R., Schmidt, F., Dam, R., Yu, L., Völker, U., Ellingsen, T.E. and Valla, S. (2009). A plasmid RK2-based broad-host-range cloning vector useful for transfer of metagenomic libraries to a variety of hosts. *FEMS Microbiol. Lett.* **296:** 149-58

Paper II
Aakvik, T., Degnes, K.F., Lale, R., Lando, M. and Valla, S. New and improved biological tools for conjugal transfer of plasmids. Manuscript.

Paper III
Aakvik, T., Drabløs, F., Andreassen, T. and Valla, S. Production in *Escherichia coli* of a melanin-like polymer from a marine sediment metagenomic clone expressing rubrerythrin. Manuscript.

Paper IV
Kotlar, H.K., Lewin, A., Johansen, J., Throne-Holst, M., Haverkamp, T., Markussen, S., Winnberg, A., Ringrose, P., Aakvik, T., Ryeng, E., Jakobsen, K., Drabløs, F., and Valla, S. High coverage sequencing of DNA from ancient microorganisms living at 250 bars 2.5 kilometers subsurface in a Norwegian sea oil reservoir. Submitted to journal.

Paper V
Aakvik, T., Lale, R., Liles, M. and Valla, S (2011). Metagenomic libraries for functional screening. In Handbook of Microbial ecology I: Metagenomics and complementary approaches. Edited by de Bruijn, F. J., John Wiley & sons. In press.

Contributions to other publications not included in the thesis:
UK patent application, Int. Pub. No.: WO 2007/141540, Int. Appl. No.: PCT/GB2007/02122, Int. Filing Date: 08.06.2007

Table of contents

1 Introduction

The earth is rich in biological material that has not yet been explored. Bioprospecting is a field of science aiming for this kind of explorations and is by the Research Council of Norway defined as the goal-oriented, systematic search for elements, bioactive compounds or genes in organisms, with the intent of developing products of commercial or social value (http://www.forskningsradet.no/en/Newsarticle/New+impetus+for+bioprospecting/1231229970484). Other definitions of this field also include more direct investigation and mapping of the biological diversity existing in nature. Even conservative estimates suggest that the number of unknown species is larger than the number of known species (Global Taxonomy Initiative), and for microorganisms the known portion constitutes presumably a very small fraction (Curtis et al., 2002; Schloss and Handelsman, 2004; Whitman et al., 1998). Thus, an increased knowledge about the biological diversity on the earth is necessary to get a more complete understanding about our environment.

1.1 Natural products

Living organisms have yielded a range of natural products, some of which are being very useful for mankind. In general, these products are derived from all kingdoms of living organisms including microorganisms, plants, invertebrates, insects, fishes, birds, and mammals (Berdy, 2005). According to the Wiley-VCH Antibase, more than 170000 natural products are known today, and every year 700 new structures are added from microorganisms only (Wiley-VCH Antibase 2009). Natural products find their use in a wide range of applications within different fields, such as medicine, agriculture, industry, and academia.

Biological catalysts are examples of natural products with a rapidly increasing range of applications. This is a result of efforts to increase the diversity and applications of their products in the marketplace, improve efficiencies and reduce costs, as well as to reduce the environmental burden of industrial processes (Ferrer et al., 2009). Naturally derived enzymes may often mediate reactions that are difficult or expensive to accomplish using chemical synthesis. In addition they have advantages over non-biological catalysts in that they are often regio- and enantioselective, and yet simultaneously being capable of accepting a wide range of molecules as substrates (Ferrer et al., 2009; Schmid et al., 2001). Large efforts are also spent searching for natural compounds with potential therapeutic applications, including for instance antibiotics, anticancer agents and immonusuppresants (Singh and Pelaez, 2008). The increasing number of bacteria that are developing resistance to the antibiotics in use is one obvious reason why new candidates are needed. Over the last decades, microorganisms have provided most of the antibiotics as well as many other medical agents that have dramatically improved human health (Gillespie et al., 2002), and are therefore an obvious source for further investigation.

1.2 Microorganisms constitute an essential component of the earth's biota

1.2.1 Microbial diversity

It is widely accepted that the genomes of microorganisms represent the major reservoir of genetic diversity on earth (Ferrer et al., 2009; Whitman et al., 1998). Throughout their long existence, microorganisms have diversified greatly, and their metabolism is much more diverse than that of larger eukaryotes. They are responsible for unique transformations in the biogeochemical cycles of the biosphere that are essential for life on earth (Whitman et al., 1998). Microorganisms are ubiquitous in every habitat on Earth, such as water, soil, air, acidic hot springs, glacial ice, highly polluted environments, and deep in the Earth's crust. In addition they are present in organic matter and the living bodies of plants and animals, and an interesting example is that humans have a greater number of bacterial cells (10^{14}) than human cells (10^{13}) in our bodies (Savage, 1977). The total number of bacterial and archaeal cells on earth has been estimated to be 4-6 x 10^{30}, comprising more than 10^{6} different genospecies (distinct taxonomic groups based on gene sequence analysis) within more than 70 large phyla (Curtis et al., 2002; Pace, 2009; Whitman et al., 1998). Further, the total amount of bacterial and arhaeal carbon is estimated to be almost equal to (60-100%) the total carbon of plants. The majority of the microorganisms reside in seawater, soil, and the sediment/soil subsurface, each of these holding approximately 10^4-10^7 cells/ml, 10^6-10^9 cells/gram, and 10^5-10^8 cells/cm^3, respectively. The average turnover times of the microorganisms within the different environments range from less than a day to 2000 years (Whitman et al., 1998).

Thus, there are no doubts that microorganisms constitute an important part of the earth's biota, and increased knowledge about these organisms is essential to fully understand the evolution and sustainability of life on this planet (Singh, 2010). Furthermore, and due to the genetic diversity described for these organisms, they are also obvious targets when screening for novel biomolecules. The traditional way of accessing this diversity is to search for organisms that express a desired trait. However, this method depends on the possibility to cultivate the organisms to be studied, and since current cultivation techniques only cover a small fraction of the diversity in environmental samples (possibly less than 1%), a huge portion of the existing microbial resources is missed (Amann et al., 1995; Rappe and Giovannoni, 2003).

1.2.2 The uncultivable majority

It was not until the 1980s that it became well accepted that most of the microorganisms on the planet resist cultivation. One indication of this was the "great plate count anomaly" -the oft-observed divergence between estimated population sizes when using direct microscopic cell counts and the number of colonies on nutrient agar (Staley and Konopka, 1985). Such estimation methods demonstrated that in natural samples less than one cell in a thousand produced a colony. These observations were also supported by studies involving DNA-DNA reassociation techniques (Torsvik et al., 1990). In 1985, Pace and colleagues introduced a cultivation-independent method involving direct analysis of 5S and 16S rRNA gene sequences in environmental samples to describe the diversity of the present microorganisms (Lane et al., 1985a; Lane et al., 1985b; Stahl et al., 1985). This, together with the development of the PCR were advances that radically changed our understanding of the microbial world. Furthermore, the use of these applications demonstrated that the uncultered majority is highly diverse and contains members that diverge largely from the culturable minority (Handelsman, 2004).

As of today, only about half of the >70 bacterial phyla have any cultured representatives, and most of the remaining phyla are represented by only a few cultured examples (Pace, 2009). There are probably a variety of reasons why an extensive fraction of the bacteria have resisted laboratory cultivation. Some might require special growth conditions (physical and chemical) that are hard or even impossible to imitate in the lab, and for others the interdependence with other organisms in the nature might be crucial. Although the cultivation success rate certainly can be improved (Tyson and Banfield, 2005), the current comprehension is that many organisms will not readily be brought into pure culture. To explore this huge source of genetic diversity, other methods are therefore important.

1.3 Metagenomics

Metagenomics is a culture-independent approach which involves genomic analysis of DNA extracted from its natural environment, and is thus a more inclusive strategy to access microbial genetic reservoirs compared to traditional, culture-dependent approaches (Handelsman et al., 1998). This field is partly analogous to genome library construction and screening, with the difference that the cloned DNA does not originate from a single known microorganism, but rather from the entire population in an environmental sample. It is, however, usually impossible to actually cover the entire population in such samples even when using metagenomic approaches, given the huge number of species within only tiny amounts of sample (see Chapter 1.2.1). Despite of this, metagenomics allows the assessment and exploitation of the taxonomic as well as the metabolic diversity in microbial communities in a highly extended fashion compared to other methods.

Metagenomics involves both sequence-based and function-based approaches (Figure 1.1, also described further below). Sequence-based methods may be used either to study the microbial community in a given habitat, or to search for genes encoding novel products. The latter strategy has the limitation that identification of genes of interest is dependent on a minimum degree of similarity to already known genes.

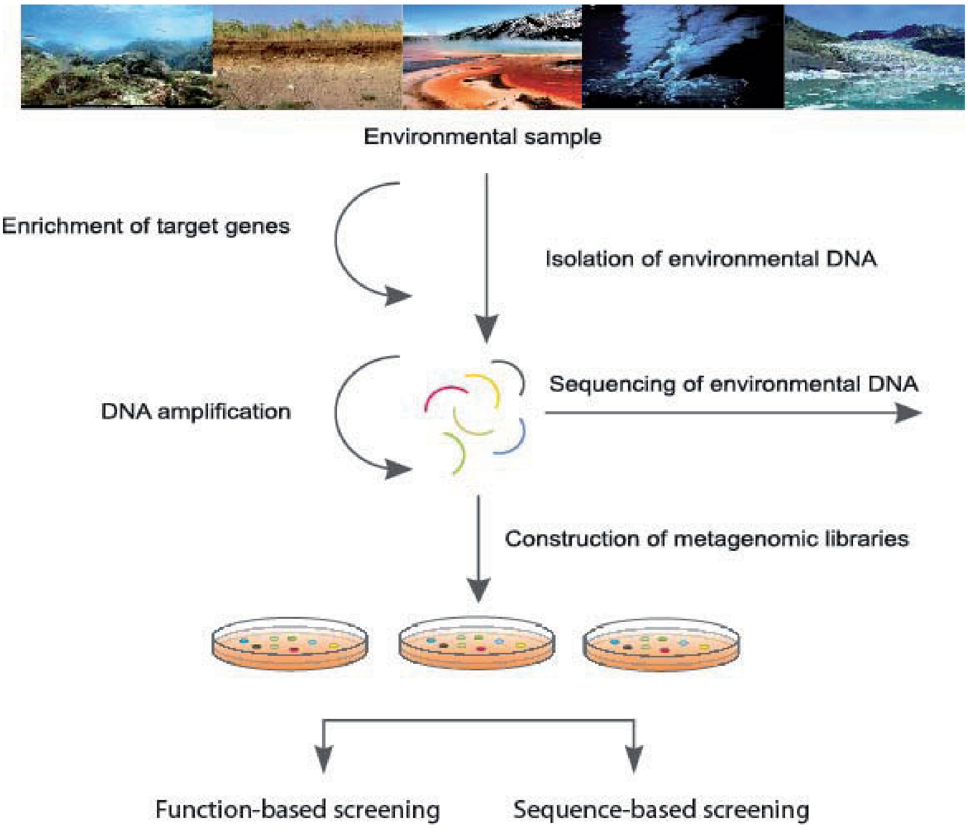

Environmental sample
Enrichment of target genes
Isolation of environmental DNA
DNA amplification
Sequencing of environmental DNA
Construction of metagenomic libraries
Function-based screening
Sequence-based screening

Figure 1.1 Steps involved in metagenomic approaches (modified from Paper V). DNA is first isolated from the habitat of interest, either directly or after enrichment of target genes. In cases of low DNA yield, *in vitro* DNA amplification can be accomplished. The DNA can further be analysed through different strategies, including direct sequencing and/or screening of libraries containing the metagenomic DNA.

A functional metagenomics approach is on the other hand dependent on successful heterologous expression of the gene or genes responsible for their function, as well as detection of gene product function. As this approach does not require a priori sequence information, it represents a more potent strategy for identification of entirely new classes of natural products. This is reflected in that genes discovered through function-based screening are in general more weakly related, or even unrelated, to previously known genes (Tuffin et al., 2009). An illustrative example is the metagenomic study of the Tammar Wallaby gut microbiome (Morrison et al., 2009) where a functional approach led to the discovery of a glycosyl hydrolase that was not one of the more than 800 putative glycosyl hydrolases identified through a sequence-based approach.

A contrast to the novelty often observed in metagenomic studies is the rediscovery rate observed when screening cultivable actinobacteria for novel antibiotics, which can be as high as 99.9% (Zaehner and Fiedler, 1995).

1.3.1 Choice of environment

In principle, any environment can be selected for metagenomic studies given that it is possible to extract nucleic acids from the sample. Three categories of environments are often considered: 1) highly diverse environments (e.g. soil and seawater), 2) environments naturally enriched for the target genes/biocatalyst, or 3) extreme environments. The DNA is either directly extracted from the environmental sample, or the population within the sample is subjected to some kind of pre-enrichment treatment (described in Paper V). With particular reference to the content of this thesis, marine environments including oil reservoirs are further described.

Marine environments

Marine water constitutes the largest contiguous habitat on the globe, occupying more than 70% of the Earth's surface with an average depth of 4 km. Life in this environment is, in contrast to most terrestrial habitats, dominated by microorganisms, both with regard to metabolism and biomass. These microorganisms accomplish many unique steps of the biochemical cycles, and they also represent a huge and dynamic source of genetic variability (Karl, 2007). Thus, this environment represents one of the most significant, but still least understood, microbial environments on Earth. Moreover, marine microbial communities were among the first microbial communities to be studied using metagenomic approaches (DeLong, 2005), and over the last years there have been several studies reporting exploration of microorganisms within marine water and sediments (e.g. Venter et al., 2004, Rusch et al., 2007, Martin-Cuadrado et al., 2007, and Hakvåg et al., 2008).

Offshore oil reservoirs (defined as a part of marine environments by the Research Council of Norway) are extreme environments due to the combination of high temperature, salinity, pressure and extraordinary nutrient conditions. In addition they have physical barriers separating them from surface environments. The first bacteria isolated from these environments were therefore assumed to be of exogenous origin, however, observations during the last couple of decades have shown that the reservoirs are microbial habitats, being occupied by indigenous microbial communities able to cope with these harsh conditions. In spite of this and of the presumable environmental and economical potential of these microorganisms, they have not been extensively investigated. Furthermore, the studies performed have mostly been on the cultivable fraction (e.g. Magot et al., 2000 and Yousset et al., 2009), and one may assume that in such habitats these constitute an even smaller portion of the real microbial community compared to a surface-community due to unknown growth requirements. Culture-independent studies of oil reservoir microbial communities performed up to the time when this PhD project was started have mainly been based on 16S rDNA sequencing (e.g. Orphan et al., 2000 and Voordouw et al., 1996).

1.3.2 Isolation of environmental DNA

Due to the physicochemical diversity in environmental samples, there exists no single universal DNA- (or RNA-) extraction protocol. It is important to choose a method that coincides with the specific genetic sequence and/or functional activity that are sought within the further analysis (see Paper V for detailed criteria). The overall criteria are most often 1) to conserve diversity, 2) to avoid sharing of the DNA, and 3) to avoid contaminating substances. The latter has proven to be a major problem when working with environmental samples, as coextraction of humic substances and polyphenolic compounds have been difficult to avoid. These contaminants will, if not properly removed, disturb the subsequent enzymatic modifications of the isolated DNA (e.g. restriction digestion, ligation, PCR), as well as bacterial transformation. Purification strategies include preprossessing of the sample (e.g. through addition of hexadecyltrimethylammonium bromide, CTAB), agarose gel purification and various chromatographical separations, or a combination of these (Liles et al., 2008; Rajendhran and Gunasekaran, 2008).

One often distinguishes between direct extraction methods where the cells are lysed within the environmental sample matrix, and indirect extraction methods where the cells are separated before lysis. The direct method may give 10-100-fold more DNA, but the DNA is less pure (Gabor et al., 2007). Lysis is achieved using chemical or enzymatic methods, mechanical treatments (thermal shocks, bead-beading, beadmill homogenization, ultrasonification, microwave heating), or a combination of these. The main disadvantage of most mechanical treatments is the resulting DNA sharing (Burgmann et al., 2001; vanElsas et al., 1997). There exist many direct extraction methods which provide DNA of sufficient quality for small-insert metagenomic libraries or PCR/pyrosequencing applications (Ogram et al., 1987; Tsai and Olson, 1991; Zhou et al., 1996), and kits are also available from commercial suppliers. When aiming for high molecular weight (HMW) DNA, it is essential to use DNA extraction methods that are much more gentle to avoid sharing. Immobilization of the cells into agarose plugs prior to lysis is an often used strategy to preserve the HMW DNA (Liles et al., 2008). After cell lysis, deproteinisation and DNA precipitation is carried out most often using organic solvents and isopropanol or ethanol, respectively.

Low biomass samples

Microorganisms living in extreme environments might be of particular interest both for sequence-based studies and in the search for enzymes with increased stability. However, due to the often low biomass in such environments, it may be challenging to extract sufficient amounts of DNA for these purposes (Ferrer et al., 2009). One strategy used to overcome this is “multiple displacement amplification”, which makes it possible to amplify even very tiny amounts of environmental DNA through the use of phi29 DNA polymerase (Abulencia et al., 2006; Kim et al., 2008; Yokouchi et al., 2006). This polymerase has an efficient displacement activity, it is highly processive, and it also has a good proofreading capacity (Dean et al., 2002; Dean et al., 2001). It may synthesize DNA fragments up to 70 kb that are suitable for both library construction and sequencing purposes. Drawbacks of this technique include formation of chimeric artefacts, amplification biases, as well as the potential introduction of point mutations which may preclude subsequent production of functional gene products. The risk for the latter is most serious in cases where clusters of genes representing biochemical pathways are searched for.

1.3.3 Metagenomic libraries

The ligation of restriction-digested or blunt-ended metagenomic DNA into vectors and subsequent transformation into a library host strain are performed using a variety of different strategies, depending on what kind of library is desired (see Figure 1.2). Points to consider include the subsequent screening strategies planned and whether these include sequence-based or function-based approaches.

For shot-gun sequencing studies, small-insert libraries are often sufficient and are usually chosen due to the fact that they are a lot easier to establish. In sequencing projects where longer, continuous sequences are aimed for, construction of large-insert libraries are often necessary.

The choice of strategy for library construction becomes somewhat complicated when aiming for function-based screening approaches. To be able to detect (novel) activities in metagenomic libraries, the vector borne heterologous gene(s) of interest needs to be successfully expressed, and this requires several criteria to be fulfilled. Firstly, the chosen insert size must be large enough to cover the entire gene or cluster of genes needed to obtain the function of interest. Secondly, a promoter and an appropriately located ribosome binding site (rbs) that are compatible with the expression machinery of the host are necessary. These *cis*-acting factors can either be provided by the cloning vector used, or be internal signals within the cloned DNA fragment. In addition, several *trans* factors need to be provided by the host cell such as proper transcription factors, inducers, precursors, chaperones, cofactors, post-translationally acting factors, and secretion mechanisms. Other possible critical factors include codon usage and the potential toxicity of the heterologous product to the host cell. One potential way to overcome (some of) these complex obstacles is to use vectors that can be transferred to and maintained in a variety of different hosts. This gives the possibility to screen the metagnomic libraries in hosts that are considered likely to express the types of genes that are searched for. Favourable qualities of such vectors include high transfer efficiencies and a broad range of hosts in which they can replicate.

Various vectors are used in metagenomic studies, including small-insert vectors, λ phage based vectors, cosmid-, fosmid-, and BAC vectors. Vectors particularly suitable for metagenomic approaches are also being developed. In Paper V different vectors used in metagenomic studies are presented, with a special focus on those that are broad-host-ranged. Fosmid and BAC vectors are further descriped in Chapter 1.4.1.

1.3.4 Screening of metagenomic libraries

Strategies for function-based and sequence-based screening of metagenomic libraries are reviewed in Paper V and therefore only a short description of these topics are given here (Figure 1.2).

Sequence-based screening

Sequence-based screening approaches are used to identify genes within a metagenomic sample or library on the basis of sequence homology. These approaches include the use of PCR-based or hybridization-based techniques for identification of target genes from primers or probes (respectively) designed from conserved regions of known genes.

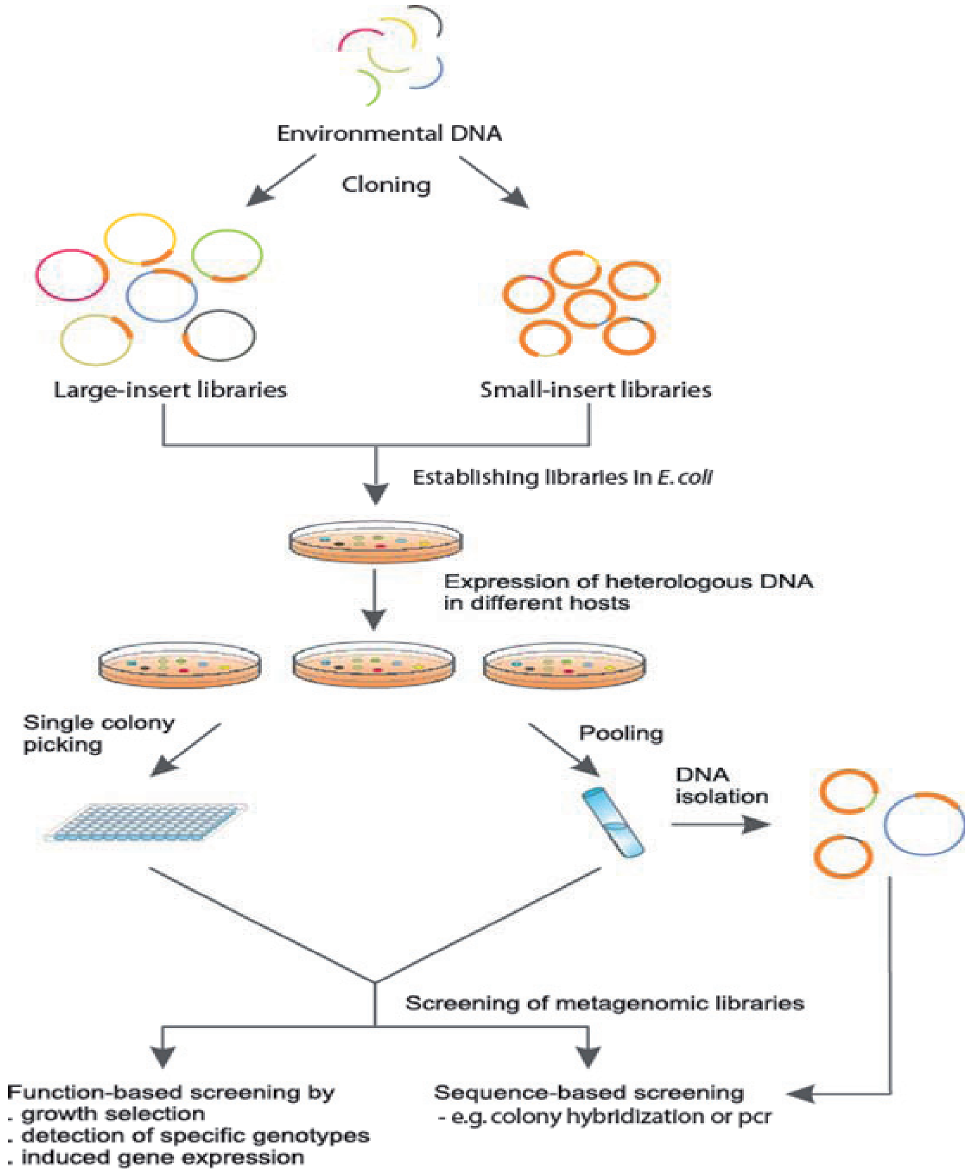

Environmental DNA
Cloning
Large-insert libraries
Small-insert libraries
Establishing libraries in *E. coli*
Expression of heterologous DNA in different hosts
Single colony picking
Pooling
DNA isolation
Screening of metagenomic libraries
Function-based screening by
. growth selection
. detection of specific genotypes
. induced gene expression
Sequence-based screening
- e.g. colony hybridization or pcr

Figure 1.2 Analysis of metagenomic DNA via cloning (modified from Paper V). The DNA can be used as cloning material in the construction of large-insert or small-insert metagenomic libraries. The use of braod-host-range cloning vectors gives the possibility of expressing the metagenomic genes in different hosts. The libraries are either transferred to microtiter plates or pooled. Finally, the libraries can be subjected to different function-based or sequence-based screens.

Function-based screening

A range of different methods have been applied for functional screening of metagenomic libraries. As the frequency of metagenomic clones that express a given trait is low, the method should preferably be either highly sensitive or carried out in a high throughput manner. Some screening approaches can be performed using agar plates supplemented with appropriate substrates, whereas others require the use of multi-well plates. Robotic systems are often needed to accomplish such experiments. The different methods used are listed below:

Growth selection

The presence of a given activity provides a growth advantage, e.g. antibiotic or heavy metal resistance. This strategy might be accomplished using mutant strains that require heterologous complementation for growth under selective conditions.

Phenotypical detection

Expression of metagenomic genes results in altered phenotype of the host, e.g. colour production or formation of halos. This may require the use of specific substrates. Antimicrobial activity may be detected through growth inhibition of suitable indicator organisms.

Screening based on induced gene expression

High-throughput screening methods useful when activities do not result in detectable phenotypes, e.g. SIGEX/ METREX and product sensing reporter systems (Uchiyama et al., 2005; Williamson et al., 2005). These methods involve screening in hosts designed such that expression of specific genes leads to expression of a reporter gene, e.g. *gfp*.

1.3.5 Sequencing metagenomic DNA

The microbial diversity within a metagenomic sample may be explored through PCR amplification and sequencing of 16S rRNA genes or other phylogenetic “anchors” (e.g. 18S rRNA and *recA* genes) present in the sample. Such single gene surveys may provide information concerning which microorganisms are present, but they give little information about the metabolic capabilities of the corresponding microorganisms and what functions they serve in their habitat.

More comprehensive knowledge of microbial communities can be achieved through large scale direct sequencing of metagenomic DNA (Wooley et al., 2010). Such massive sequencing of large metagenomes is now technically feasible due to the relatively recent advances of automated, high throughput sequencing facilities (e.g. 454 technology) and powerful algorithms for sequence assembly (sequencing and assembly technologies are further described below). For simple communities with low diversities, sequencing of a manageable amount of DNA may allow reconstruction of nearly complete genomes (e.g. Tyson et al., 2004). For more diverse ecosystems, however, assembly of the obtained sequences is most often problematic. This was clearly demonstrated by Tringle and co-workers in a study where less than 1% of the 100 MB of sequence obtained from a soil sample overlapped, and no contigs (i.e. longer sequences obtained from overlapping reads) were formed (Tringe et al., 2005). In general, such sequencing projects yield a large number of relatively short reads (between 80 and 1000 bp), assembly of which may lead to formation of artificial contigs containing sequences originating from different hosts (DeLong, 2005; Kunin et al., 2008).

The group of Craig Venter was the first to apply extensive large-scale sequencing for microbial community characterization and identification of new species and genes, using samples from the Sargasso Sea (Venter et al., 2004). The shotgun sequencing data set included 1.0 billion bp and led to the discovery of over 1.2 million new genes. The largest dataset obtained to date is the one generated from the Global Ocean Sampling (GOS) expedition which includes 6.3 billion bp (Rusch et al., 2007; Yooseph et al., 2007). Interestingly, analysis of this dataset indicates that new protein families are discovered at a rate that is approximately linear with the addition of new sequences. This indicates that there are probably many new protein families to be discovered in nature.

Various other environments have been taxonomically described through analysis of sequence data obtained by pyrosequencing and/or Sanger sequence analysis of the present metagenomic DNA, including Antarctic polar Front samples collected at 500 m depth (Moreira et al., 2004; Moreira et al., 2006), an acid mine biofilm (Tyson et al., 2004), deep sea sediments (Hallam et al., 2004), Polar Greenland Sea (Zaballos et al., 2006), glacial ice (Simon et al., 2009), honey bee colonies (Cox-Foster et al., 2007), and the Peru Margin subsea floor (Biddle et al., 2008).

The enormous amount of DNA sequences obtained from such (or similar) studies may also be used as the basis for *in-silico* screening approaches. This may lead to identification of interesting gene sequences, which can further be used for *in vitro* testing (e.g. Höhne et al., 2010)

Next generation sequencing technologies

Traditionally, DNA sequencing has almost exclusively been carried out using the Sanger method (Sanger et al., 1977). Over the past few years, alternative sequencing platforms have become widely available which allow faster sequencing with a reduction in costs by over two orders of magnitude (Shendure and Ji, 2008). The "second generation sequencing technologies" (e.g. the 454 sequencing) circumvent the need for traditional library construction or designing of primers homologous to the template DNA. These new platforms are diverse at many levels, but their work flows are conceptually similar. Summarized their concept consists in sequencing of a dense array of amplified DNA fragments through iterative cycles of enzymatic manipulation and imaging-based data collection (Shendure and Ji, 2008). Relative to other post-Sanger sequencing platforms, the main advantage of the 454 platform is the read-length (Table 1.1).

Even more recent technologies are the so-called "third generation sequencing technologies", which involve sequencing of individual molecules (Xu et al., 2009). Two such third generation sequencing technologies based on fluorescence detection have already been launched (Table 1.1), and others are being developed by different companies (Metzker, 2010; Xu et al., 2009).

Table 1.1 Second and third generation sequencing technologies (adapted from Xu et al. (2009) and Metzker (2010), as well as the references indicated in the table)

Platform	Read-length [bp]	Giga bases pr run	Chemistry
Roche/454's GS FLX Titanium[a]	400-500	0,45 (10h)	PS
Illumina/Solexa's			
Genome Analyzer *IIx*[b]	100	18 (9,5 days)	RTs
Life/APG's SOLiD 4 *hq*[c]	75	Up to 300 (3-14 days[f])	SBL
Helicos BioSciences[d]	25-55	21-35 Gigabases (8 days)	RTs
Pacific Biosciences[e]	Average: 1000	N/A	Real-time

[a]-[c]: Second generation sequencing technologies: [a] www.454.com, [b] www.illumina.com, [c] www.appliedbiosystems.com. [d]-[e]: Third generation sequencing technology: [d] www.helicosbio.com, [e] www.pacificbiosciences.com. [f] depending on individual read lengths. N/A: Not available, PS: pyrosequencing, RT: reverse terminator, SBL: sequencing by ligation.

Due to the still shorter read lengths and/or the higher error rates of these technologies compared to the Sanger procedure, a higher coverage is needed to obtain data of high quality. Nevertheless, this transition to the next-generation sequencing technologies opens the possibility for Gb-scale metagenomic projects, including sequencing of complex communities nearly to saturation (Chistoserdova, 2010). However, successful implementations of such projects are also dependent on improved sampling, DNA extraction, bioinformatical analysis tools, and data storage infrastructures.

Analysis of sequence data

The large amounts of information generated by the above mentioned sequencing platforms must be processed in order to be meaningful for the user, and one of the challenges has been the development of appropriate data-analysis tools. The two most fundamental computational analyses in this context are alignment and assembly: If a reference genome for the sequenced sample exists (or another genome within appropriate evolutionary distance), alignment analyses are mainly used. For DNA samples without a sequenced reference genome such as metagenome-samples, assembly (i.e. combination of sequence reads into contigs) is essential for further analysis (e.g. Newbler, Roche's 454 assembler) (Flicek and Birney, 2009).

In addition, bioinformatic software tools are needed when aiming for phylogenetic analysis of enormous metagenomic derived datasets (reads- or contigs-), and several such tools have been developed for this purpose. Examples include the metagenome analyzer MEGAN which is a recently invented software tool for taxonomic binning of large metagenomic datasets of short DNA fragments typically obtained through 454 sequencing (Huson et al., 2007), and the MG-RAST server that provides a comparative functional and sequence-based analysis for uploaded samples (http://metagenomics.nmpdr.org/).

1.4 Specialized plasmids and strains relevant to this study

1.4.1 Fosmid and BAC vectors

Fosmid and BAC vectors are large-insert cloning vectors with the capabilities of holding up to 40 kb and 300 kb inserts, respectively (Kim et al., 1992; Shizuya et al., 1992). They are both low copy-number vectors based on the *Escherichia coli* F factor replicon, yielding only 1-2 copies per cell. Fosmid vectors can be packaged into lambda phage heads due to the presence of dual *cos*-sites, similarly as its multi-copy precursor cosmid vector. BAC vectors have their strength in that they may hold large biosynthetic gene clusters, such as polyketide or other antibiotic synthesis gene clusters which may be over 40 kb in size (e.g. the 55 kb erythromycin- (Brikun et al., 2004) and the 124 kb nystatin- (Brautaset et al., 2000) clusters). Libraries established in BAC vectors can, however, also contain clones with rather small inserts since there is no selection against small fragments as in the fosmid system. Both BAC and fosmid vectors are frequently used for construction of (meta-) genomic libraries, and due to the F factor system, such large-insert libraries can be established and stably maintained. However, the F factor allows only for replication within *E. coli*, and thus functional screening of libraries can only be done within this host. This is a problem that is addressed in this project.

1.4.2 Cloning vectors with the possibility of conditional copy-number amplification

A single copy/ high copy vector system was developed by Wild and co-workers (Wild et al., 2002) in order to combine the advantages of the stabilizing features of the single copy BAC vectors with a possibility of *in vivo* amplification to obtain higher yields of DNA upon plasmid isolation. This was achieved through a system where the vector carries two origins of replication; *ori2* from the F plasmid and *oriV* from RK2, while the host carries a chromosomally integrated mutant version (copy-up) of the *trfA*-gene (see 1.4.3) that can be activated by expression from the inducible (L-arabinose) *pBAD* promoter. Thus, when conditionally induced with L-arabinose, the copy-number switches from ~1 to ~50-100 copies per cell.

1.4.3 The RK2 replicon

RK2 is a 60 kb large self-transmissible plasmid belonging to the *IncP* incompatibility group (Figure 1.3).

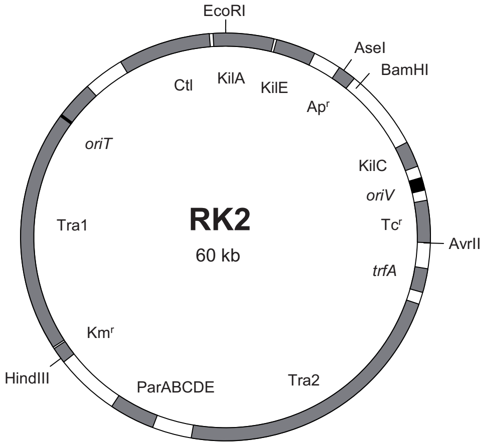


Figure 1.3 Map of the broad-host-range plasmid RK2. Regions responsible for replication (*oriV*, *trfA*), conjugative transfer (*oriT*, Tra1 and Tra2), stabilization (ParABCDE) and antibiotic resistance are depicted, and other genetic regions are indicated according to the abbreviation conventions established in Pansegrau et al. (1994).

It has the capability to replicate in a wide range of bacterial species, including numerous Gram-negative species (Thomas and Helinski, 1989), and it has also been transferred to Gram-positive bacteria, yeasts and mammalian cells (Bates et al., 1998; Poyart and Trieu-Cuot, 1997; Waters, 2001). RK2 contains three different genes for antibiotic resistance (Km^r, Ap^r, Tc^r).

Replication and copy number control

The elements needed for replication of RK2 include the origin of replication, *oriV*, and the *trfA* gene product responsible for replication initiation and copy number control (Perri et al., 1991). The *trfA* gene contains two in-frame translational starts, leading to two gene products of different size. It has been shown that both these TrfA proteins independently can initiate replication in many bacterial hosts (e.g. Durland and Helinski, 1987 and Shingler and Thomas, 1989). The *trfA* products form homo- and heterodimers which bind to short repeated sequences (iterons) in the *oriV* region. The copy number of RK2 is regulated through these interactions according to the "handcuffing model" (Toukdarian and Helinski, 1998), and in *E. coli* the copy number is 4-7 per chromosome (Thomas and Helinski, 1989). Series of *trfA* mutants with divergent properties have been isolated, including for instance *cop* mutants with increased plasmid copy number, and temperature sensitive (ts) mutants for which replication has altered temperature requirements (Durland et al., 1990; Haugan et al., 1992; Haugan et al., 1995; Valla et al., 1991). Of particular interest with respect to this work is that copy number manipulation through the use of *trfA cop* mutants can be done across species barriers (Haugan et al., 1995). Mapped mutations within the *trfA* gene and their effects are illustrated in Figure 1.4.

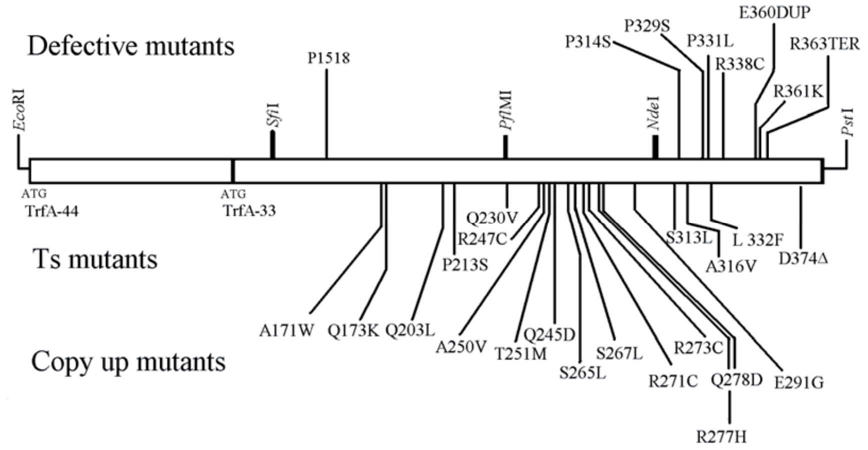


Figure 1.4 Overview over characterized mutations in the *trfA* gene.

Conjugation

Conjugative transfer of RK2 is dependent on the origin of transfer, *oriT*, as well as two separate transfer regions designated Tra1 and Tra2. These two regions encode 27 gene products primarily involved in donor-recipient contact, mating pair formation, DNA transfer and replication (Pansegrau et al., 1994). Transfer of plasmid DNA involves a nick within the *oriT* region, from which the transfer occurs. Conjugal transfer of RK2 is very efficient, particularly compared to transformation of naked DNA.

The stabilization *par* region

The *par* region of RK2 is involved in the maintenance of RK2 plasmids (as well as other replicons) in diverse species. The genes within this region include *parA, B, C, D* and *E*, transcribed as two units (*parABC* and *parDE*). This region is not necessary for expression and function of other regions of the plasmid, but if inactivated the plasmid has been shown to become segregationally unstable in all hosts tested (Pansegrau et al., 1994). The *parDE* genes may ensure stability in a system where plasmid-free cells are killed, as *parE* encodes a lethal polypeptide and *parD* an antagonist (Roberts et al., 1994).

Small cloning vectors based on the RK2 replicon

A series of derivatives of RK2 with considerably reduced size have been made, typically containing the mini RK2 replicon (*trfA* and *oriV*), the origin of transfer (*oriT*), and selected resistance markers, possibly combined with other desirable features (e.g. Blatny et al., 1997a and Blatny et al., 1997b). Such vectors are useful as cloning or expression tools, and are often used in experiments including transfer and expression of genes in non- *E. coli* hosts. As these small-sized vectors do not contain the Tra-regions required for conjugative transfer, they are dependent on the supply of these gene products from another source. This is most often achieved through the use of the *E. coli* donor strain S17-1 or its analogue SM10, as these have the RP4 plasmid (indistinguishable from plasmid RP4 (Burkardt et al., 1979)) integrated into their chromosome (Simon et al., 1983).

1.4.4 The R6K replicon

The R6K origin, *oriR6K*, is a narrow-host-range replication origin that requires the *pir*-encoding Π protein for replication initiation. Suicide plasmids constructed from this replicon carries *oriR6K* but not the gen encoding Π. Thus for extra-chromosomal replication they are dependent on presence of Π, provided *in trans*. The *E. coli* strain S17-1(λ*pir*) is a R6K λ*pir* phage lysogen of *E. coli* S17-1 which enable replication of such plasmids (de Lorenzo et al., 1993). The R6K replicon and strain S17-1(λ*pir*) have been used in this work for the construction of two different suicide plasmids used in experiments involving transposon insertion and homologous recombination, respectively.

2 Aims of the study

The overall aim of this study was to develop biological tools useful for function-based metagenomic studies, to establish metagenomic libraries from marine sources, and to screen these libraries for potentially novel traits, mainly pigment and antibiotic synthesis.

As the use of many different hosts increases the possibility for successful heterologous expression of genes obtained from environmental samples, vectors with a broad host-range should be aspired in function-based metagenomic studies. Construction of a vector system based on elements within the RK2 plasmid would potentially enable this, as these are known to allow for efficient transfer of plasmids to numerous hosts in which they can replicate extra-chromosomally. The vector should also be capable of holding large DNA fragments across species barriers in order to enable expression of entire pathways necessary for synthesis of secondary metabolites such as pigments and antibiotics.

Due to problems with plasmid modifications observed during mobilized conjugal transfer with the extensively used donor strain *E. coli* S17-1, it became also a goal to elucidate the reason for these problems and to establish an alternative, improved mobilization system.

A metagenomic library should be established in the new vector system using DNA isolated from Norwegian marine environments. Metagenomic DNA from different habitats of this environment should be approached for in order to obtain libraries with different content for varying screening purposes. Established libraries would be used in further studies within this work, but would also be stored as resources for further screening programs.

The metagenomic library or libraries should be screened for novel traits including production of pigments and antibiotics, and for selected conserved gene functions through sequence-based strategies. When convenient, it should be aimed at using different hosts in the function-based screens in order to utilize the properties of the new vector system. Possible positive hits should be further investigated, and one cloned environmental DNA fragment should preferably be characterized at the gene and product levels.

As a metagenomic project which aimed at exploring the microbial communities within a Norwegian oil reservoir was initiated during this study, it also became a goal to contribute to this work.

In response to the publication of the new vector system, the group received an invitation to write a review paper for a metagenomic book. It was therefore decided to write this paper as a part of the PhD work.

3 Summary of Results and Discussions

3.1 Development of a broad-host-range metagenome vector system (Paper I)

Since different species represent varying expression capabilities, there is a potential involved in using many different hosts for expression of metagenomic genes (see Chapter 1.3.3). In order to exploit this potential, appropriate biological tools are needed. The initial objective in this project was therefore to develop a vector system for such purposes. The resulting broad-host-range combined fosmid and BAC vectors pRS44/pTA44 are relatively small and functionally well understood vectors (Figure 3.1). As verified through experiments described below, this vector system can be used for large-insert (metagenomic) library construction in *E. coli* followed by efficient transfer of intact library clones to non-*E. coli* hosts. A patent application has been filed for this vector system.

3.1.1 Features of the constructed vectors

The vectors pRS44/pTA44 (hereafter referred to only as "pRS44") were constructed using the commercially available pCC1FOS vector as starting point. pCC1FOS contains both the *ori2* and *oriV* origin of replication, giving the possibility for both single copy and high copy number replication as described in Chapter 1.4.2. In pRS44, *oriV* represents in addition the origin of replication within non-*E. coli* hosts. Three additional genetic elements are included in pRS44: 1) *oriT* from RK2 for efficient transfer to alternative hosts, 2) the stabilization element *parDE*, also from RK2, and 3) the kanamycin resistance gene to provide an additional selection marker.

Replication of pRS44 in non- *E. coli* hosts is (in addition to the presence of *oriV*) dependent on expression of *trfA*. This gene was not included in pRS44, as the single-copy/high-copy feature (i.e. the single copy state determined by *ori2*) then would be lost. Instead, a suicide vector, pRS48 (Figure 3.1), was constructed which allows for insertion of the *trfA* gene into the chromosome of the new hosts (see Paper I for details about pRS48).

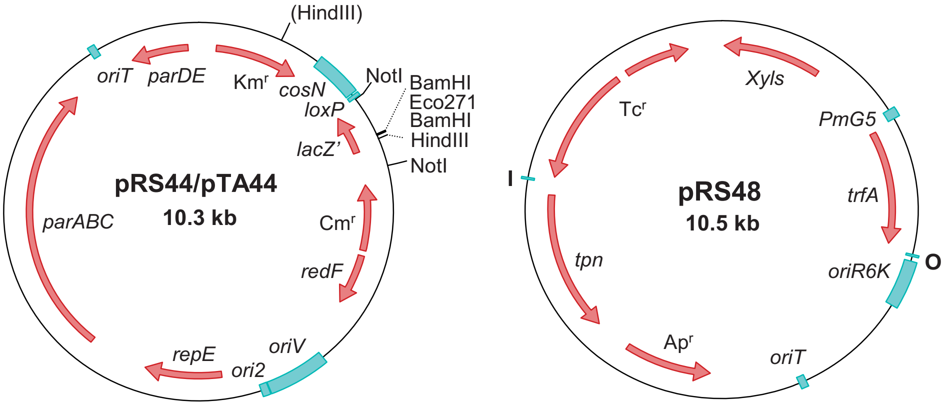


Figure 3.1 Plasmid vectors for use in genomic library constructions (pRS44/pTA44) and for support of vector replication in hosts other than *E. coli* (pRS48). The HindIII site within the Km^R gene of pRS44 is removed in pTA44. (For more details, see legend to Figure 1, Paper I)

In pRS48, the wild type *trfA* is used, giving 4-7 copies pr cell. A similar vector, pTA64, has also been made which instead carries a copy-up mutant trfA-gene (*cop*271C, see Chapter 1.4.3). Correspondingly, any mutant *trfA* can easily be cloned into this suicide vector for further insertion into the chromosome of potential hosts. This means that a given library can be established in different hosts giving different copy-numbers of the library clones. As expression normally correlates with gene dosage, this possibility may represent an important advantage regarding detection of activities from poorly expressed genes, as has also been experimentally observed (Liles et al., unpublished results). Furthermore, as metagenomic libraries are laborious to establish, the possibility to manipulate *one* library through these means is a major advantage in itself.

3.1.2 Verification of the stability and cloning capacity of pRS44

The *parDE* element has been shown to stabilize RK2 vectors across species barriers (Blatny et al., 1997a; Sia et al., 1995), and was included in pRS44 because plasmid stability is a crucial feature for such metagenomic cloning vectors. The positive effect of this stabilization element was clearly confirmed in an experiment where the stability of pRS44 (with and without insert) and pCC1FOS were compared in *E. coli* (Figure 2, Paper I).

The cloning capacity of pRS44, both as a fosmid vector and as a BAC, has been verified through several approaches. The standard 36 kb insert used as a control for the commercially available vectors was ligated, packaged and established in *E. coli* at similar frequencies for pRS44 as for pCC1FOS (data not shown). Its capacities to hold very large fragments was initially examined using total DNA isolated from *Pseudomonas fluorescens* (arbitrarily chosen) as cloning material. Partially digested (Sau3AI) DNA was ligated into the BamHI site of the vector, yielding clones with inserts up to 82 kb (determined with PFGE, data not shown). This capacity was further investigated by a company (BioS&T, Canada) which routinely constructs BAC libraries with large inserts. Their cloning strategy involves HindIII cloning, and therefore pTA44 was constructed, in which the HindIII site within the kanamycin resistance gene is removed (through site specific mutagenesis), thus leaving the remaining HindIII site available as cloning site. Clones with inserts up to 200 kb were obtained in this vector, and ligation and transformation efficiencies were similar to what was observed in parallel experiments with the commercially available BAC cloning vector pIndigoBAC5 (Epicentre). A pulsed field gel electrophoresis analysis of 22 of the obtained clones is given in Figure 4, Paper I.

pRS44 has also proven to function well as a metagenomic cloning vector, as it was used for the construction of a 20000 member fosmid library with insert DNA originating from microorganisms within marine sediments (described further in Chapter 3.3.2). This was an important verification as environmental DNA is known to be both difficult and inefficient to clone. The library was also used to identify the clone described in Paper III.

3.1.3 Transfer of metagenomic clones to non- *E. coli* strains

The results described in the previous chapter confirmed that pRS44 has the features required for metagenomic studies and for holding and maintaining large inserts in *E. coli.* However, to prove the main concept of the vector system it was necessary to demonstrate that insert bearing clones can successfully be transferred to and maintained within non- *E. coli* hosts.

The *E. coli* strain used for construction of metagenomic libraries within pRS44 (EPI300) does not contain the Tra-regions required for *oriT*-mediated conjugative transfer, and therefore library clones have to be transformed into strain S17-1 prior to conjugation (Chapter 1.4.3). The reason why the library was not established in S17-1 in the first place is that transformation frequencies are higher when using EPI300. Transformation of both the fosmid library and selected BAC clones to S17-1 resulted in large numbers of transformants, indicating that this additional step probably does not significantly reduce the representativity of the original libraries. After inserting the *trfA* gene into the chromosome of two selected species, *Pseudomonas fluorescens* and *Xanthomonas campestris*, both fosmid library clones and large-insert BAC clones were efficiently conjugated from S17-1 to these recipients.

To analyse the transconjugants, plasmids were isolated from the non- *E. coli* hosts and – in order to get higher levels of DNA recovery- retransformed into *E. coli* EPI300. Restriction analyses of the plasmids before and after transfer showed that both fosmid- and large-insert BAC clones could be recovered in apparently intact states from the non- *E. coli* hosts (Figure 5, Paper I). It was also observed that a fraction of the transferred plasmids did undergo some kind of modification during the transfer (further described in Chapter 3.2). However, as the modifications did not happen in all transfers for any such clone, this problem can be overcome by simply screening a larger number of transconjugants than would otherwise be required.

The use of the new vector system for heterologous expression of environmental DNA in a non- *E. coli* host has also been demonstrated. This was done by analysing the proteomes of several *P. fluorescens* fosmid clones (Figure 3, Paper I).

Based on the results presented in Chapters 3.1.1-3.1.3, it could be concluded that pRS44 has retained the capacity from the parent plasmid pCC1FOS to hold and maintain very large inserts, and that it is more stable than its parent. Furthermore, and of particular importance, pRS44 has the capacity to be efficiently transferred to and stably maintained in non- *E. coli* hosts, even when holding large inserts.

The use of the RK2 replicon in this vector system provides a host range that is potentially very broad (Chapter 1.4.3), and in metagenomic contexts it is to our knowledge uniquely broad. Particularly interesting is that RK2-derived plasmids have shown to replicate within both thermotolerant (*Methylococcus capsulatus*, Svein Valla, pers.comm.) and psycrophilic species (*Vibrio salmonicida*, Marit Sjo Lorentzen, pers.comm.). This might represent a possibility of successful expression of genes from microorganisms living in hot or cold environments, respectively, as such genes are not necessarily successfully expressed in a mesophilic background. Specifically, proteins from psycrophilic organisms may not be functional at the temperature required for growth of *E. coli*. Another advantage with this vector system is that the efficiency of *oriT*-mediated conjugative transfer is very high, usually much higher than transformation of naked DNA. This means that transfer of metagenomic libraries to new hosts can be done without significant loss of representativity relative to the library present in the *E. coli* donor.

3.2 Construction of an improved plasmid-based mobilization system (Paper II)

When analysing the transconjugants after transfer of large plasmids to non- *E. coli* hosts, it was observed that a fraction of the clones had been modified during the transfer (mentioned in Chapter 3.1.3). Such modifications represent a problem because they may inactivate genes in the metagenomic insert. Even though this can be overcome by increasing the volume of the clones screened, extended screens are very resource-demanding in functional screening. The *E. coli* donor strain that was used, S17-1 (Chapter 1.4.3), is also frequently used as conjugation donor by a range of other research groups, indicated by the nearly 4000 citations (January 2011) to the paper of Simon et al. (1983) describing this strain, as well as the analogous strain SM10. The constant level of citations over the last 20 years is further an indication of a continued interest in these donor strains (Figure 3.2).

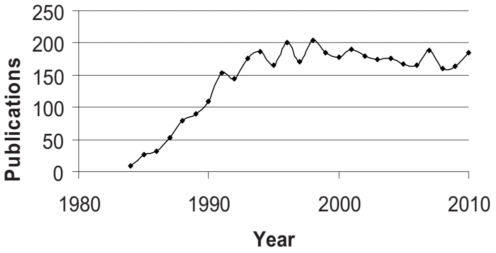


Figure 3.2 Number of citations as a function of year to the paper describing the S17-1/SM10 donor strains (Simon et al., 1983).

As also other complications are reported with the use of S17-1/SM10 as conjugal donors (see Paper II), it was decided to investigate these occurrences further and evaluate whether a new conjugation donor system could eliminate the problems.

3.2.1 Analysis of plasmids which were modified during conjugal transfer

The plasmid modifications were discovered in agarose gel electrophoretic analyses of HindIII digested fosmid clones before and after transfer to two different hosts (*P. fluorescens* and *X. campestris*) (given in Figure 1, Paper II). Plasmids isolated from different transconjugants showed in some cases different restriction patterns even though they originated from the same *E. coli* clone. Furthermore, it was observed that plasmids that had passed through two *different* species could undergo modifications that gave the same restriction patterns (but different from the pattern of the original plasmid). For all cases an inspection of the digestion patterns revealed that the structural modifications involved increases in plasmid sizes.

To investigate the origin of the DNA that seemed to be inserted into the modified plasmids, ten fragment bands appearing on the gel in addition to those of the original plasmid were sliced out, ligated into pLitmus28 and end-sequenced (Sanger procedure). Results from Blast-analyses of the obtained sequences revealed that three of them gave no significant hits against the databases, indicating that these originated from the original metagenomic insert DNA. The remaining sequences were identical to *E. coli* DNA originating from a chromosomal region of ca 35 kb (three sequences), or to a gene within the Tn7 transposon (four sequences). A relevant note here is that Tn7 was used to inactivate the kanamycin resistance gene within RP4 during construction of the conjugation donor strain S17-1 (Simon et al., 1983).

Based on these findings it appeared that the plasmid-modifications involved insertion of chromosomal DNA from the donor into the plasmids. As both the fosmid vector (pRS44) and the S17-1 chromosome contain RK2/RP4 originating DNA, this could possibly have happened through homologous recombination, but as S17-1 is *recA*$^-$ homologous recombination should not happen in this strain. It is, however, well established that also chromosomal DNA may be conjugatively transferred from one strain to another, and as the integrated RP4 within S17-1 has an intact *oriT*, this may obviously happen in this system. Thus, in cases where a fosmid clone and DNA mobilized from the chromosomally located *oriT* is co-transferred to the same, *recA*$^+$ recipient cell, homologous recombination may potentially occur between these molecules in the recipient host. Such a hypothesis may explain how *E. coli* originating DNA ends up within the fosmid clones.

3.2.2 Construction of a plasmid-based mobilization system

Given that the suggested hypothesis is correct, a possible solution to the problem could be to inactivate *oriT* within the donor strain. However, when first developing a new donor-system our intention became also to improve it by other means. Firstly, to avoid the problems experienced with the S17-1/SM10 system (e.g. mobilization of Mu-genomes from donor to recipient) the new system was to be made independent of these specific strains. Further, to make the system more flexible it was decided to keep the mobilizing tra-functions extra-chromosomally, thus giving the possibility for transfer to different strains for establishment of new donors. Such a plasmid-based system required the mobilizing replicon to be changed in order to make it compatible with the many IncP-based vectors used in experiments involving conjugation.

In order to be able to precisely modify the large-sized RK2 plasmid, a system was developed that allowed for homologous recombination between ligated fragments within a suicide plasmid (pTA10) and the target region of RK2, with subsequent selection of RK2-mutants. pTA10 carries the narrow-host-range replication origin *oriR6K* (Chapter 1.4.4), the *sacB* gene for counterselection on sucrose agar, and the chloramphenicol resistance gene. Using this system and the *recA*$^+$ strain *E. coli* ER2566 two minor deletions were made in the RK2 plasmid, resulting in plasmid pTA19 which has both the *oriT* site and the kanamycin resistance gene inactivated (see Paper II for details). To change the replicon of this RK2-derivative a fragment containing *oriV* and the ampicillin- and tetracycline resistance genes was replaced with a fragment containing the pBBR1 replicon and the gentamycin resistance gene. The resulting plasmid, pTA-Mob, is illustrated in Figure 2 in Paper II. pBBR1 also has a broad-host-range, and compatibility studies indicate that it most likely represent a novel incompatibility group, meaning that pTA-Mob can be used within a range of bacterial species in combination with most vectors and plasmids.

pTA-Mob-containing *E. coli* DH10B cells were further transformed with the same fosmid clones that gave varying restriction patterns after transfer from S17-1, in order to repeat the conjugation experiments from this new system. *P. fluorescens* transconjugats were obtained at apparently the same frequencies as from S17-1, indicating that the RK2-derivate pTA-Mob has retained the elements necessary for conjugation. Isolation and restriction analyses of twelve of the obtained transconjugants showed that the above described modifications could not be detected after conjugation from this new system (Figure 3, Paper II).

Based on these results it could be concluded that the developed broad-host-range plasmid pTA-Mob is capable of supporting *in trans* the Tra-genes necessary for conjugation of *oriT* containing plasmids without co-transfer of DNA from a second active *oriT* and the problems that may cause. The complications reported regarding mobilization of the Mu genome within the S17-1/SM10 chromosomes (Babic et al., 2008; Wiater et al., 1994) are also eliminated in the new system. Further, as the system is plasmid-based it provides a flexibility not available with the prior system. pTA-Mob can for instance be established in a strain highly competent as library-construction host prior to library construction, thus avoiding the necessity to transfer library plasmids among *E. coli* strains before transfer. Due to the broad-host-range feature of the helper plasmid, it can also be used to mobilize for instance IncP-based plasmids *from* non-*E. coli* strains. This possibility is intended to be further examined in the near future. Additional applications can also be envisioned, some of which are described in Paper II.

This new mobilization system eliminates the problems reported with the S17-1/SM10 system and opens in addition for new possibilities.

3.3 Construction of metagenomic libraries

It is in general quite challenging to construct metagenomic libraries, particularly when aiming for cloning of large inserts. This is mainly because DNA isolated from environmental samples usually contains contaminating compounds that interfere with the enzymatic reactions involved in library construction. It is also sometimes problematic to obtain sufficient amounts of DNA from the particular habitat of interest, especially DNA of high molecular weight due to the often high degree of degradation during sample preparation. As described in the following chapters, these problems were also experienced in this work, in particular those concerning poor cloning-efficiency and DNA-yield. The molecular weight of the obtained DNA was, on the other hand, in general sufficiently high.

Initially, the plan was to construct metagenomic libraries using DNA isolated from microorganisms within the sea surface microlayer as cloning material, as this particular environment was the focus of an ongoing bioprospecting project involving a large part of our research group (Chapter 3.3.1). However, it turned out to be very laborious to obtain sufficient amounts of DNA from this environment, and the cloning efficiencies were in addition very low. After several attempts it became clear that the resources necessary to continue working with this particular DNA were not available within the frame of this PhD, and it was therefore decided to instead use marine sediments as sampling sources.

During this study another metagenomic project was initiated in our group which involved DNA-isolation from oil reservoirs samples. These samples were unique and very exiting, and it was decided that a part of the activity of this project should be coupled to this PhD. The isolated DNA was unfortunately not of sufficient quality for direct cloning, but a pyro-sequencing study has been successfully completed and has given interesting new information (Chapter 3.5).

3.3.1 Sampling the sea surface microlayer (unpublished results)

The sea surface microlayer has been demonstrated to contain a much higher concentration of organic matter and bacteria than the underlying water, and these bacteria are in addition subjected to much higher levels of UV-radiation (Aller et al., 2005; Bezdek and Carlucci, 1972; Sieburth et al., 1976). Due to these characteristics this habitat has been shown to contain a comparatively high share of bacteria producing pigments and/or antibiotics (Hermansson et al., 1987), which were the main products in focus of this project. An example of an interesting pigment could be astaxanthin, which is heavily used in Norwegian fish farming. Both a metagenomic and a culture-dependent study were planned for microbes within this habitat, and the results were eventually meant to be compared.

Even though the concentration of microorganisms is higher in the surface microlayer than in the underlying water, it is still much lower than in for example sediments. A rather large volume was therefore necessary to obtain sufficient amounts of DNA for cloning (~20 litres, requiring ~250 samplings), and consequently the sampling- and DNA isolation procedures (given in Appendix 1.) became quite laborious. The cloning efficiency of the obtained DNA was in addition very low, particularly when aiming for the large insert libraries that were necessary to include entire pathways for synthesis of pigments and antibiotics. It turned, however, out to be possible to obtain a reasonable amount of small-insert clones when cloning Sau3AI partially digested DNA into the standard high copy-number cloning vector pLitmus28 (BamHI digested). Ninety-one of these clones were used to evaluate the isolated environmental DNA through end-sequencing and BlastN analyses. More than half of the sequences gave no relevant hits to other sequences present in the databases and the best Blast-hits were all against bacteria, several of which are typically found in marine environments (data not shown). The results were therefore consistent with the assumption that the DNA sample mainly originates from marine bacteria, a conclusion that is not necessarily as trivial as it may seem (DeLong, 2005). Further attempts of large-insert cloning in pRS44 resulted in a fosmid library of only about 400 clones. Considering that only a small fraction of the DNA originally present in an environmental sample actually ends up in a metagenomic library, it was assumed that the sampling position was not as important as the possibility to achieve a large library. A decision was therefore made to use marine sediments as sampling source instead, as such samples are more easily obtainable.

3.3.2 Construction of metagenomic libraries using DNA isolated from marine sediments (Paper I and unpublished results)

The sampling of marine sediments from the intertidal zone at the Trondheim fjord and subsequent DNA isolation as described in Paper I resulted in satisfactory amounts of DNA, most of it having a relatively high molecular weight (20-100 kb). While initial attempts of large-insert library construction were not successful, a large-member library with inserts between 5 and 10 kb was quite easily obtained (analogous to the library in pLitmus28 described above, but with larger inserts and more than 20000 clones). Construction of such a small-insert metagenomic library was also within the scope of the bioprospecting project, and this library can be used to search for single genes and their enzymes. A selection of the obtained white colonies (blue-white selection was used) was examined through restriction digest analysis, and also by end-sequencing. The results indicated that the library contained inserts larger than 5 kb of environmental origin (data not shown). White colonies were picked using a Genetix QPix II robotic colony picker and transferred to 384 well plates. Approximately 20000 clones were totally picked, which means that about 150 Mb environmental DNA is stored in this library, assuming an average insert size of 7.5 kb. This library has so far only been screened for resistance against several antibiotics (chloramphenicol, kanamycin, streptomycin and tetracycline), but without detecting promising hits. Further screening approaches were not prioritised for this library, but it is stored and available for possible future applications.

Presence of contaminating compounds in the metagenomic DNA samples was assumed to be the reason for unsuccessful large-insert cloning, and in order to remove more of these compounds the DNA was subjected to an extra agarose-gel purification step. Subsequent cloning of this DNA into pRS44 led to the generation of a fosmid library consisting of about 20000 clones. Restriction digest analysis of a selection of the clones confirmed that the average insert size was about 35 kb, meaning that the library includes approximately 700 Mb of environmental DNA. End-sequencing of the inserts of these clones showed that they contained different inserts, apparently of environmental origin (data not shown). It was therefore concluded that this library could be used in further screening approaches. The library is stored as individual clones in 384 well plates (in 2 parallels, transferred as described for the small-insert library), as well as pooled freeze stocks.

3.3.3 Sampling a Norwegian oil reservoir (Paper IV / unpublished results)

A project ongoing at our group has the aim to study the microbial communities within a Norwegian oil reservoir, mainly through metagenomic approaches. In this research, samples were collected from a 100 m thick oil-containing layer (250 bars pressure, 85°C) about 2.5 kilometres below the bottom of the Norwegian Sea (Figure 1, Paper IV). This sampling was unique in that a specialized technology was used to avoid contamination problems, and also that the sample pressure was slowly released to avoid cell lysis (see Supplementary information to Paper IV). Thirty µg DNA was isolated from the water phase (5 litres) of this sample essentially as described for the sediment sample in Paper I.

One of the aims within the project has been to construct metagenomic libraries in pRS44 using the DNA isolate as cloning material. This would have constituted a second example of an environmental library established in pRS44 and also a very interesting source for further function-based, as well as sequence-based, investigations. However, the DNA turned out to be almost inclonable into any vectors (even as small inserts) without further treatment. After PCR-amplification of the isolated DNA, a small-insert library containing 35000 clones was established (Lucigen® Middleton; USA). This library is used in the study described in Chapter 3.5. Ongoing work has also shown that whole genome amplified DNA (Chapter 1.3.2) from these isolates can be cloned into fosmid-vectors. Thus, the intention is still to establish libraries in pRS44 using the oil reservoir DNA, but this was unfortunately not attained within the timeframe of this PhD. The library that is used in the screening experiments described in the next chapters is therefore the fosmid library containing DNA isolated from marine sediment samples (Chapter 3.3.2).

3.4 Screening of the marine sediment fosmid library

The metagenomic fosmid library was subjected to different screening approaches in order to test its potential, to investigate the broad-host-range properties of the vector, and also to search for products with possible applications.

Initial sequence-based screens indicated that the library contained DNA of diverse origins and with various features. This was seen as promising for the further screening approaches.

The library was in its entirety transferred to *P. fluorescens* to allow parallel screening in a non- *E. coli* host (when feasible). The focus was mainly on screening for pigmented compounds and antimicrobials in accordance with the parallel ongoing bioprospecting activity in the group (Engelhardt et al., 2010; Hakvag et al., 2008; Jorgensen et al., 2009; Stafsnes et al., 2010). Colonies of both the *E. coli*- and the *P. fluorescens* library were inspected in order to try to detect pigment-production, but unfortunately without any positive hits. Antimicrobial activities were screened for using several different strategies (summarized in Chapter 3.4.2), and although very clear candidates were not found, several clones which showed interesting properties were subjected to sequence analyses.

During the cultivation of individual *E. coli* clones for the antimicrobial screen, a few cultures were observed to be slightly pigmented. These observations indicated successful heterologous expression from the metagenomic DNA, and the clones were selected for further investigations. One clone displayed reproducible pigment-production, and a further study of this clone is described in Chapter 3.4.3 (also Paper III).

The many possibilities for functional screening of the marine sediment library could not be fully exploited within the frame of this PhD, but the library represents a big potential if more resources become available in the future. For example, the use of more hosts and the variable copy-number feature is likely to lead to more hits of interest.

3.4.1 Initial, sequence-based analyses of the library inserts

Exploration of the genetic diversity in the metagenomic library (Paper I)

In order to explore the microbial diversity within the metagenomic fosmid library, regions of 16S rRNA genes were amplified and analysed by DGGE, and 24 of the amplified gene fragments were sequenced. More than 10 different bands could clearly be distinguished on the DGGE gel, indicating that the library contains DNA that originates from many different genotypes (data not shown). Analysis of the 24 obtained sequences resulted in the identification of 10 different genotypes, none of which were identical to existing sequences in the databases. Thus it could be concluded that the library contains environmental inserts of diverse origins.

Screening the metagenomic fosmid library for PKS- and NRPS genes using colony hybridization (unpublished results)

Non-ribosomal peptide synthases (NRPS) and polyketide synthases (PKS) are involved in production of many antimicrobial secondary metabolites and are therefore of particular medical relevance. In order to investigate whether the metagenomic fosmid library included genes involved in the production of such metabolites, the library was subjected to colony hybridization using a mixed probe with sequences from conserved domains within both NRPS- and PKS genes (described in Appendix 2.). The screens were performed on L-agar medium containing L-arabinose, as it was shown that high copy-numbers of the fosmids highly simplified detection of target genes. 12500 colonies were screened, and ten of these gave positive hybridization signals that were confirmed in a rescreen. PCR-, sequencing- and BlastX analyses resulted in the attainment of seven possible NRPS- or PKS gene fragments, three of which displayed significant sequences similarities to NRPS genes. All ten clones were screened for antimicrobial activities, both in *E. coli* and *P. fluorescens*, using overlay-assays as described in the next chapter. Unfortunately, no reproducible activities were detected

Additional candidate clones could probably have been found by expanding this screen. However, as it was considered that identification of the corresponding activities was crucial for such a study it was assumed that this would have been too resource-demanding for the PhD project. Nevertheless, the results from the two sequence-based analyses performed confirmed that the library contains various features from diverse species, which was promising for further screening approaches.

3.4.2 Functional screening of the metagenomic fosmid library for antimicrobial activities (unpublished results)

The metagenomic fosmid library was screened for antibacterial and antifungal activity using the indicator organisms *Micrococcus luteus* (ATCC 9341) and *Candida albicans* (ATCC 10231), respectively. These were chosen due to their general high sensitivities as well as extensive and good experiences in similar screening approaches based on cultivated organisms from environmental samples (e.g. Hakvåk et al., 2008).

Overlay-assays were initially used to screen for antimicrobial activities within library-clones of both *E. coli* and *P. fluorescens* (according to Hakvåg et al., 2008), but without detecting reproducible activities. Another approach was then used, which involved testing DMSO extracts of the individual *E. coli* clones for antagonistic activity against the indicator organisms (according to Jørgensen et al., 2009 and Engelhardt et al., 2010). During preparation of the DMSO extracts, the library clones were cultivated both in the absence and presence of L-arabinose, i.e. at single and high copy number states of the fosmids, respectively. The intention was originally to also apply this latter strategy to different library-containing hosts. However, this required re-establishments of the screening-procedure as well as transfer of non- *E. coli* libraries into well-formats, and due to resource constraints such approaches were not possible within the frame of the PhD.

A third screening approach was taking into consideration that heterologous expression of antimicrobial compounds may potentially affect the growth of the library host in which they appear, given that host-defence mechanisms are not co-expressed. As such activities will not be detected in screens as those described above, it was decided to in parallel investigate the growth of the library-clones at different copy-numbers of the insert-bearing plasmids, assuming that a higher dose of such genes results in poorer host-growth.

Twenty-two clones were in total found to show possible antimicrobial activities within the different described screens, even though reproduction of very clear activities was not possible. One of these was clone 16N21 which displayed very different growth curves in the presence and absence of L-arabinose (Figure 3.3).

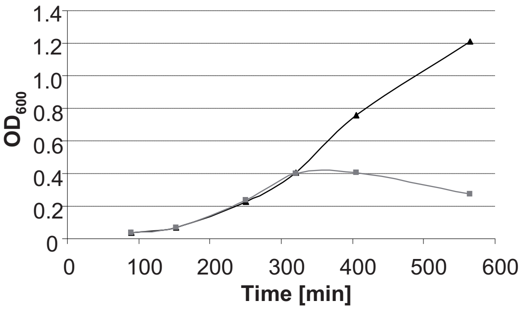


Figure 3.3 Growth curves for clone 16N21 in de absence (black triangles) and presence (grey squares) of L-arabinose. The cultures were inoculated (0.5%) from an over-night culture at time zero.

To evaluate these 22 clones further, a mixture containing all the corresponding plasmids was subjected to sequencing (454 technology, preformed as described in Paper III). Analyses after assembly and annotation (also performed as described in Paper III) revealed that half of the 775 potential ORFs remained un-assigned ("no hits"), and in addition a significant part of the obtained hits were against "putative" or "probable" protein functions (data not shown). This preliminary analysis did not lead to detection of complete operons or pathways linked to production of antimicrobial compounds, though a few interesting single gene functions were assigned. A further and more thorough investigation of the obtained sequence data is required in order to unravel the potential of these clones.

There may be several reasons why no clear antimicrobial activities were confirmed from this library. Complete and intact pathways for such metabolites may not be present in the library, their genes may not have become successfully expressed, or the activities might not be detected through the chosen screening method. Also, reproducibility of successful expression from such metagenomic DNA is particularly challenging, and this complicates the verification of positive candidates (Mark Liles, pers.comm.)

3.4.3 Melanin production from a metagenomic fosmid clone (Paper III)

During cultivation of the library clones for the antimicrobial screen (high copy number state), the cultures within four of the wells were observed to be pigmented. As the host strain does not display such a phenotype itself, this was an indication of successful expression of metagenomic insert-genes. Further investigations revealed that reproduction of the phenotypes -including after retransformation- was only possible for one of these clones (15J7), and only when inducing for high copy numbers of the fosmid. Cultures of this clone displayed different colours from red to bluish-purple and brown, and the cultures became darker upon longer incubation times (Figure 3.4).

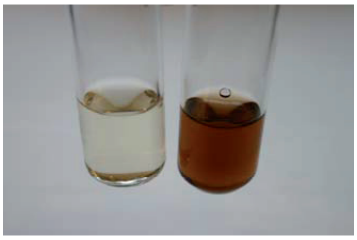


Figure 3.4 Comparison of cell-free culture broths from *E. coli* EPI300 containing clone 15J7 (induced for high copy number of fosmid, right) and empty vector only (left).

Characterization of the coloured compound

Different analytical approaches were used for characterization of the pigment produced by 15J7, and a key initial discovery was that its molecular weight is surprisingly high, estimated to be between 10 and 50 kDa. This, together with the observations that the colour was stably maintained even after rough treatments such as boiling and freezing, indicated that the coloured compound could be a melanin-like polymer. Further characterizations strengthened this hypothesis: A dark precipitate formed upon acidification of the culture broth. This precipitate was insoluble in aqueous acid and organic solvents (such as acetone and EtOH) and soluble in aqueous base, like other melanins. Addition of $FeCl_3$ to the coloured broth resulted in formation of a flocculent precipitate, and addition of $KMnO_4$ led to bleaching. Further, the absorbtion spectrum (200-900 nm), and also the NMR spectrum, were similar to respective melanin spectra reported (see Figures 2 and 3, Paper III).

Melanins are polyphenolic heteropolymers that form a diverse group of pigments. Different melanins are formed from different precursors, and the polymerization process does -in contrast to most other biopolymers- not follow precise patterns, meaning that a given melanin-sample contains molecules with variable structures. Enzymes are needed only to catalyse the first stage(s) of the multistep melanin-formation pathways, whereas the polymerisation itself occurs through spontaneous oxidations (Sanchez-Amat et al., 2010). Melanins are produced by representatives of almost every large taxon, and they are reported to protect the organisms in which they are synthesized against several stress conditions (Plonka and Grabacka, 2006; Sanchez-Amat et al., 2010). (Further information about melanins is given in Paper III.)

To test whether the polymerization of the pigment produced by strain 15J7 also occurs spontaneously, cells were removed from un-pigmented cultures harvested at late exponential phase and the culture broths were incubated at different temperatures (4-80°C). After two hours, the samples at the highest temperatures had become light brown, whereas the rest of the samples where colourless. This trend strengthened upon prolonged incubation at the different temperatures, indicating that the polymerisation of this pigmented material also occurs spontaneously.

Identification of the gene responsible for melanin production

The fosmid clone 15J7 was sequenced using 454 sequencing technology, and assembly of the obtained reads resulted in a continuous sequence of 47312 bp (355x coverage). Analysis of the predicted gene functions after annotation of the insert-sequence did, however, not give any clear candidate gene(s) for the melanin production. In order to locate this gene or genes, the insert was therefore successively reduced in size via sub-cloning procedures. This resulted eventually in a pigment-producing strain containing a plasmid with only one complete ORF (ORF1). The nucleotide sequence of this ORF is 58.2% G+C, and the predicted ORF1-encoded protein contains 318 amino acids (35.4 kDa). BlastX searches revealed high similarities with the oxidoreductase rubrerythrin from several marine and soil bacteria. Most of the rybrerythrins described contain two domains; an N-terminal erythrin (Er) domain with a di-iron site and a C-terminal rubredoxin domain with a FeS_4 site. A further bioinformatic analysis of ORF1 revealed that the corresponding protein contains the N-terminal Er-domain, but not the rubredoxin domain. Instead, there are two somewhat truncated transmembrane VIT1 (vascular iron transport) domains C-terminally (Figure 4 in and Supplementary material to Paper III). Several other sequences (128 in total) with a deduced Er-VIT1 domain architecture can be found in international databases, and their gene products are suggested to be involved in metal transport processes (Andrews, 2010). However, no such enzymes have been studied with respect to activity, and rubrerythrins have generally not been associated with melanin production.

In order to investigate whether the VIT1 domain was necessary for pigment production, the gene sequence for the Er-domain was synthesized (GeneScript USA Inc) and expressed alone in *E. coli* BL21(DE3)pLysS. Interestingly, the pigment was also produced within this culture, indicating that the VIT1 domain is not necessary for this reaction, at least not under the given conditions. The mechanism behind the pigment formation is therefore believed to be that the Er domain has a catalytic activity that leads to the formation of a melanin precursor, either directly or indirectly, and that this precursor further polymerises to melanin through spontaneous oxidations.

To be able to understand in more detail the nature of the production of melanin within clone 15J7, it was desirable to determine the building blocks of the polymer as no such information could be drawn from the deduced responsible gene product. However, such a characterization of melanins is difficult, and very few melanins have actuallty been structurally solved (Jacobson, 2000). Different strategies were tried, including NMR- and LC-MS analyses performed both on the melanin polymer, on $KMnO_4$-oxidized material (according to Ito and Fujita, 1985 and Ito and Wakamatsu, 1994), and on culture broths after different incubation times. Unfortunately, no clear conclusions were obtained from these studies. Contact has therefore been established to a group at the University of Naples, Department of Organic Chemistry and Biochemistry (Italy), which has a long experience with characterization of melanins. Hopefully, this cooperation will lead to a more detailed insight into the polymer structure, and further to the identification of potential precursors. If so, a more specific mechanism for the action of the metagenomic gene product may be proposed.

As stated above, no connections have to our knowledge earlier been made between rubrerythrins and melanin. However, both are reported to be involved in oxidative stress protection, meaning that the functions of the gene product and the final product are actually related. Summarized, the findings within this study represent a successful example of the use of a functional metagenomics approach for the discovery of natural products, and they may further contribute to the assignment of a hitherto unknown function of members of the rubrerythrin enzyme class.

3.5 A metagenomic study of the microbial community within an oil reservoir (Paper IV)

Paper IV describes a bioinformatic analysis of DNA isolated from oil reservoir microorganisms (Chapter 3.3.3). The main conclusions from this study are 1) that the obtained DNA sequences show (to us) surprisingly high levels of similarities to sequences of DNA isolated from more accessible locations, considering that these microorganisms have presumably been separated from the rest of the biosphere for very long periods of time (see Paper IV), and 2) that the dominating microorganisms present in the reservoir are sulphur-reducing bacteria and methane-producing Archaea, known to live either synthrophically or as competitors. We have reasons to believe that this sequence data represents the community at high coverage. In my work I have mainly contributed by confirming this conclusion.

3.5.1 An independent evaluation of the obtained sequence data

The 454 pyro-sequencing results were independently tested by sequencing cloned fragments of the original DNA. The 35000 clone library described in Chapter 3.3.3 was used, from which 17 randomly chosen clones were end sequenced (Sanger procedure). Sixteen of the obtained sequences were identified in the final contig-sequencing dataset from the pyro-sequencing, meaning that only 1/17 of the cloned fragments was not represented in the pyro-sequencing dataset. For nine of these 16 sequences, the entire sequence aligned to a contig. Since most of the sequence-reads from the cloned inserts were longer than the pyro-sequencing reads, this indicated that the assembly process was of high quality. The remaining seven sequences extended beyond the end of their corresponding contigs when aligned to the dataset, and could not be used for this analysis. This observation is consistent with the fact that a large fraction of the contigs is short (i.e. shorter than the Sanger-reads, see Figure S1 in Supplementary information to Paper IV). An interesting note is that the extension of one of these seven sequences aligned to two other (small) contigs end to end in such a way that three contigs were fused without gaps. This observation is an additional indication of the high sequence coverage of the dominating DNA sequences in the sample.

Taken together, these findings suggested that the final datasets represent the DNA within the oil-well sample with high coverage and therefore that the further analysis of the sequences should lead to a good understanding of the corresponding microbial community.

3.5.2 Synthesis, cloning and expression of a gene sequence annotated to encode a *Pelobacter* enolase

When analysing sequences from such habitats it is important to keep in mind that assignments down to species level are problematic due to the limited number of comparable genome sequences available in the databases. This means that sequences annotated as mesophilic species like e.g. *Pelobacter carbinolicus* more likely originates from thermophilic relatives to this species. To test this hypothesis, it was decided to synthesize a gene sequence in the metagenome annotated to encode a known enzyme, express the gene in *E. coli*, and compare temperature stabilities of this enzyme with the corresponding *E. coli* enzyme. A sequence annotated to encode an enolase from *P. carbinolicus* (96% identity at aa level) was chosen for this study. The sequence was codon optimized, synthesized and ligated into the expression vector pGS-21a (GeneScript USA Inc), giving plasmid pNTA1. Expression in *E. coli* BL21(DE3)pLysS was successful and efficient, as confirmed by SDS-PAGE (Figure 3.5).

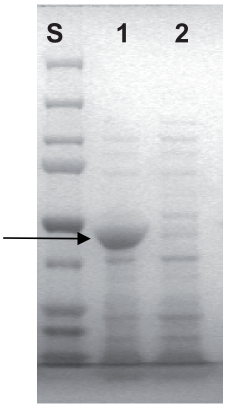


Figure 3.5 Analysis of protein contents within crude extracts of (1) *E. coli* BL21(DE3)pLysS/pNTA1 and (2) the control *E. coli* BL21(DE3)pLysS/pGS-21a. S: Protein standard, Dual Color (Bio-Rad).

Crude extracts of *E. coli* BL21(DE3)pLysS/pNTA1 as well as a control containing an empty pGS-21a vector and thus expressing only the corresponding *E. coli* enolase were used in *in vitro* enolase assays (Boel et al., 2004). The enolase activity was found to be much higher in the cells containing the recombinant gene, confirming the annotation of the sequence, and consistent with the protein gel result indicating a very high level of expression. This crude extract was diluted 70-fold to give approximately the same level of enzyme activity at 37°C as extract from cells with vector only (i.e. as the *E. coli* enolase). The enzyme activity was measured at 37, 55, 65, 70, 75, 80, and 85°C for both samples, and even though the highly diluted extract with the recombinant enzyme was used, the negative control reduced its activity much more rapidly as the temperature was increased (Supplementary information to Paper IV). The endogenous enolase activity was significantly reduced already at 55°C and at 70°C it was completely lost (Figure 3.6), whereas in contrast the recombinant enolase displayed an increased activity at 55°C and in addition showed detectable activity up to 80°C.

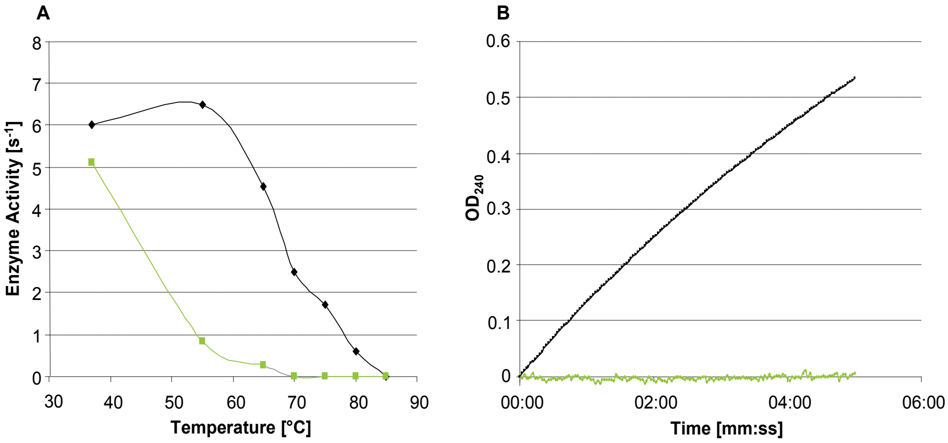


Figure 3.6 Results from enolase assays performed on crude extracts of both BL21(DE3)pLysS/pNTA1 (black) and the control BL21(DE3)pLysS/pGS-21a (grey). A: enzyme activity at different temperatures, B: enolase assays at 70°C.

It is known that the *in vivo* temperature tolerance of enzymes to some extent is positively correlated with elevated pressure (Eisenmenger and Reyes-De-Corcuera, 2009; Hei and Clark, 1994), leading to the assumption that the oil-well originated enolase might be active at even higher temperatures if the assay is performed at high pressure. In conclusion, these results strengthen the original assumtion that sequences assigned to *P. carbinolicus* are in fact rather originating from a thermophilic *Pelobacter*-like specie.

4 Concluding remarks and perspectives

The work presented here involves construction of new biological tools which allow for screening of metagenomic libraries within numerous hosts. These tools include a vector system with a capacity to hold very large insert-fragments (up to 200 kb) that can be efficiently transferred to non- *E. coli* hosts through conjugation. The vector has proven to be remarkable stable, even when holding large inserts. Another advantage with this vector is that its copy number can be manipulated across species barriers, and as expression normally correlates with gene dosage this may represent an important advantage with respect to detection of genes/products expressed at low levels. A new, plasmid-based mobilization system for conjugal transfer of *oriT*-containing plasmids has also been established. This new system has several advantages over the extensively used S17-1/SM10 system and should therefore have the potential to be frequently used.

A 20000 member metagenomic fosmid library has been established in pRS44 using DNA isolated from marine sediments as cloning material. Screening of this library resulted in the identification of a melanin-producing clone, and further investigations revealed that a single protein, rubrerythrin, was sufficient to generate melanin. As no connections have previously been made between rubrerythrins and melanin, these findings may contribute to the assignment of a hitherto unknown function of this enzyme class. The marine fosmid library and also a small-insert library containing DNA from the same habitat are both stored as individual clones and are available for future studies.

What particularly represents a potential for future studies is a further exploration of the broad-host-range features of the constructed vector system. This should be done to evaluate the potential of these tools, for example by screening for selected targets in two or three different hosts, followed by transfer of positive clones to many different species to investigate the frequency by which the activity was observed. The outcomes of such investigations may later be used to design efficient strategies for targeted screening of diverse sets of metagenomic libraries.

The use of a functional screening approach for the identification of new genes of interest has the advantage that completely new traits can be discovered, even if the sequences of the corresponding genes display no similarity to already known genes. A disadvantage of this approach is the relatively low frequency of positive hits, often in combination with laborious protocols, which means that such screening may become very resource-demanding. Advances in robotics and high through-put screening technologies have reduced this problem somewhat, and the development of improved biological tools (such as those described in this project) is also important. Due to the rapid developments of sequencing technologies lately, DNA can now be sequenced efficiently without prior cloning and at relatively low costs. Because of this, sequence-based approaches are becoming increasingly adopted within metagenomics. The main limitation with this method is that functions of interest that cannot be detected by bioinformatic analyses will not be identified. This is exemplified in a study by Liles and co-workers where many of the antibiotic resistance genes that were identified through a function-based approach lacked significant homology to known antibiotic determinants and would thus not have been found by sequence-based screening (Parsley et al., 2010). The current status is therefore that functional approaches most likely will be important also in the future in order to map more of the functional diversity of microbial life on Earth.

Refererences

Abulencia, C.B., D.L. Wyborski, J.A. Garcia, M. Podar, W. Chen, S.H. Chang, H.W. Chang, D. Watson, E.L. Brodie, T.C. Hazen, and M. Keller. 2006. Environmental whole-genome amplification to access microbial populations in contaminated sediments. *Appl Environ Microbiol*. 72:3291-3301.

Aller, J.Y., M.R. Kuznetsova, C.J. Jahns, and P.F. Kemp. 2005. The sea surface microlayer as a source of viral and bacterial enrichment in marine aerosols. *J Aerosol Sci*. 36:801-812.

Amann, R.I., W. Ludwig, and K.H. Schleifer. 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev*. 59:143-169.

Andrews, S.C. 2010. The Ferritin-like superfamily: Evolution of the biological iron storeman from a rubrerythrin-like ancestor. *Biochim Biophys Acta*. 1800:691-705.

Babic, A., A.M. Guerout, and D. Mazel. 2008. Construction of an improved RP4 (RK2)-based conjugative system. *Res Microbiol*. 159:545-549.

Bates, S., A.M. Cashmore, and B.M. Wilkins. 1998. IncP plasmids are unusually effective in mediating conjugation of Escherichia coli and Saccharomyces cerevisiae: involvement of the tra2 mating system. *J Bacteriol*. 180:6538-6543.

Berdy, J. 2005. Bioactive microbial metabolites. *J Antibiot (Tokyo)*. 58:1-26.

Bezdek, H.F., and A.F. Carlucci. 1972. Surface Concentration of Marine Bacteria. *Limnol Oceanogr*. 17:566-&.

Biddle, J.F., S. Fitz-Gibbon, S.C. Schuster, J.E. Brenchley, and C.H. House. 2008. Metagenomic signatures of the Peru Margin subseafloor biosphere show a genetically distinct environment. *Proc Natl Acad Sci U S A*. 105:10583-10588.

Blatny, J.M., T. Brautaset, H.C. WintherLarsen, K. Haugan, and S. Valla. 1997a. Construction and use of a versatile set of broad-host-range cloning and expression vectors based on the RK2 replicon. *Appl Environ Microb*. 63:370-379.

Blatny, J.M., T. Brautaset, H.C. WintherLarsen, P. Karunakaran, and S. Valla. 1997b. Improved broad-host-range RK2 vectors useful for high and low regulated gene expression levels in gram-negative bacteria. *Plasmid*. 38:35-51.

Boel, G., V. Pichereau, I. Mijakovic, A. Maze, S. Poncet, S. Gillet, J.C. Giard, A. Hartke, Y. Auffray, and J. Deutscher. 2004. Is 2-phosphoglycerate-dependent automodification of bacterial enolases implicated in their export? *J Mol Biol*. 337:485-496.

Brautaset, T., O.N. Sekurova, H. Sletta, T.E. Ellingsen, A.R. StrLm, S. Valla, and S.B. Zotchev. 2000. Biosynthesis of the polyene antifungal antibiotic nystatin in Streptomyces noursei ATCC 11455: analysis of the gene cluster and deduction of the biosynthetic pathway. *Chem Biol*. 7:395-403.

Brikun, I.A., A.R. Reeves, W.H. Cernota, M.B. Luu, and J.M. Weber. 2004. The erythromycin biosynthetic gene cluster of Aeromicrobium erythreum. *J Ind Microbiol Biotechnol*. 31:335-344.

Burgmann, H., M. Pesaro, F. Widmer, and J. Zeyer. 2001. A strategy for optimizing quality and quantity of DNA extracted from soil. *J Microbiol Methods*. 45:7-20.

Burkardt, H.J., G. Riess, and A. Puhler. 1979. Relationship of group P1 plasmids revealed by heteroduplex experiments: RP1, RP4, R68 and RK2 are identical. *J Gen Microbiol*. 114:341-348.

Busti, E., P. Monciardini, L. Cavaletti, R. Bamonte, A. Lazzarini, M. Sosio, and S. Donadio. 2006. Antibiotic-producing ability by representatives of a newly discovered lineage of actinomycetes. *Microbiology*. 152:675-683.

Chistoserdova, L. 2010. Recent progress and new challenges in metagenomics for biotechnology. *Biotechnology Letters*. 32:1351-1359.

Cox-Foster, D.L., S. Conlan, E.C. Holmes, G. Palacios, J.D. Evans, N.A. Moran, P.L. Quan, T. Briese, M. Hornig, D.M. Geiser, V. Martinson, D. vanEngelsdorp, A.L. Kalkstein, A. Drysdale, J. Hui, J. Zhai, L. Cui, S.K. Hutchison, J.F. Simons, M. Egholm, J.S. Pettis, and W.I. Lipkin. 2007. A metagenomic survey of microbes in honey bee colony collapse disorder. *Science*. 318:283-287.

Curtis, T.P., W.T. Sloan, and J.W. Scannell. 2002. Estimating prokaryotic diversity and its limits. *Proc Natl Acad Sci U S A*. 99:10494-10499.

de Lorenzo, V., I. Cases, M. Herrero, and K.N. Timmis. 1993. Early and late responses of TOL promoters to pathway inducers: identification of postexponential promoters in Pseudomonas putida with lacZ-tet bicistronic reporters. *J Bacteriol*. 175:6902-6907.

Dean, F.B., S. Hosono, L. Fang, X. Wu, A.F. Faruqi, P. Bray-Ward, Z. Sun, Q. Zong, Y. Du, J. Du, M. Driscoll, W. Song, S.F. Kingsmore, M. Egholm, and R.S. Lasken. 2002. Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A*. 99:5261-5266.

Dean, F.B., J.R. Nelson, T.L. Giesler, and R.S. Lasken. 2001. Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res*. 11:1095-1099.

DeLong, E.F. 2005. Microbial community genomics in the ocean. *Nat Rev Microbiol*. 3:459-469.

Durland, R.H., and D.R. Helinski. 1987. The sequence encoding the 43-kilodalton trfA protein is required for efficient replication or maintenance of minimal RK2 replicons in Pseudomonas aeruginosa. *Plasmid*. 18:164-169.

Durland, R.H., A. Toukdarian, F. Fang, and D.R. Helinski. 1990. Mutations in the trfA replication gene of the broad-host-range plasmid RK2 result in elevated plasmid copy numbers. *J Bacteriol*. 172:3859-3867.

Eisenmenger, M.J., and J.I. Reyes-De-Corcuera. 2009. High pressure enhancement of enzymes: A review. *Enzyme Microb Tech*. 45:331-347.

Engelhardt, K., K.F. Degnes, M. Kemmler, H. Bredholt, E. Fjaervik, G. Klinkenberg, H. Sletta, T.E. Ellingsen, and S.B. Zotchev. 2010. Production of a new thiopeptide antibiotic, TP-1161, by a marine Nocardiopsis species. *Appl Environ Microbiol*. 76:4969-4976.

Ferrer, M., A. Beloqui, K.N. Timmis, and P.N. Golyshin. 2009. Metagenomics for mining new genetic resources of microbial communities. *J Mol Microbiol Biotechnol*. 16:109-123.

Flicek, P., and E. Birney. 2009. Sense from sequence reads: methods for alignment and assembly. *Nat Methods*. 6:S6-S12.

Gabor, E., K. Liebeton, F. Niehaus, J. Eck, and P. Lorenz. 2007. Updating the metagenomics toolbox. *Biotechnol J*. 2:201-206.

Gillespie, D.E., S.F. Brady, A.D. Bettermann, N.P. Cianciotto, M.R. Liles, M.R. Rondon, J. Clardy, R.M. Goodman, and J. Handelsman. 2002. Isolation of antibiotics turbomycin a and B from a metagenomic library of soil microbial DNA. *Appl Environ Microbiol*. 68:4301-4306.

Hakvag, S., E. Fjaervik, K.D. Josefsen, E. Ian, T.E. Ellingsen, and S.B. Zotchev. 2008. Characterization of Streptomyces spp. isolated from the sea surface microlayer in the Trondheim Fjord, Norway. *Mar Drugs*. 6:620-635.

Hallam, S.J., N. Putnam, C.M. Preston, J.C. Detter, D. Rokhsar, P.M. Richardson, and E.F. DeLong. 2004. Reverse methanogenesis: testing the hypothesis with environmental genomics. *Science*. 305:1457-1462.

Handelsman, J. 2004. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev*. 68:669-685.

Handelsman, J., M.R. Rondon, S.F. Brady, J. Clardy, and R.M. Goodman. 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol*. 5:R245-249.

Haugan, K., P. Karunakaran, J.M. Blatny, and S. Valla. 1992. The phenotypes of temperature-sensitive mini-RK2 replicons carrying mutations in the replication control gene trfA are suppressed nonspecifically by intragenic cop mutations. *J Bacteriol*. 174:7026-7032.

Haugan, K., P. Karunakaran, A. Tondervik, and S. Valla. 1995. The host range of RK2 minimal replicon copy-up mutants is limited by species-specific differences in the maximum tolerable copy number. *Plasmid*. 33:27-39.

Hei, D.J., and D.S. Clark. 1994. Pressure stabilization of proteins from extreme thermophiles. *Appl Environ Microbiol*. 60:932-939.

Hermansson, M., G.W. Jones, and S. Kjelleberg. 1987. Frequency of antibiotic and heavy metal resistance, pigmentation, and plasmids in bacteria of the marine air-water interface. *Appl Environ Microbiol*. 53:2338-2342.

Hohne, M., S. Schatzle, H. Jochens, K. Robins, and U.T. Bornscheuer. 2010. Rational assignment of key motifs for function guides in silico enzyme identification. *Nat Chem Biol*. 6:807-813.

Huson, D.H., A.F. Auch, J. Qi, and S.C. Schuster. 2007. MEGAN analysis of metagenomic data. *Genome Res*. 17:377-386.

Ito, S., and K. Fujita. 1985. Microanalysis of eumelanin and pheomelanin in hair and melanomas by chemical degradation and liquid chromatography. *Anal Biochem*. 144:527-536.

Ito, S., and K. Wakamatsu. 1994. An improved modification of permanganate oxidation of eumelanin that gives a constant yield of pyrrole-2,3,5-tricarboxylic acid. *Pigment Cell Res*. 7:141-144.

Izumikawa, M., M. Murata, K. Tachibana, Y. Ebizuka, and I. Fujii. 2003. Cloning of modular type I polyketide synthase genes from salinomycin producing strain of Streptomyces albus. *Bioorg Med Chem*. 11:3401-3405.

Jacobson, E.S. 2000. Pathogenic roles for fungal melanins. *Clin Microbiol Rev*. 13:708-717.

Jorgensen, H., E. Fjaervik, S. Hakvag, P. Bruheim, H. Bredholt, G. Klinkenberg, T.E. Ellingsen, and S.B. Zotchev. 2009. Candicidin biosynthesis gene cluster is widely distributed among Streptomyces spp. isolated from the sediments and the neuston layer of the Trondheim fjord, Norway. *Appl Environ Microbiol*. 75:3296-3303.

Karl, D.M. 2007. Microbial oceanography: paradigms, processes and promise. *Nat Rev Microbiol*. 5:759-769.

Kim, K.H., H.W. Chang, Y.D. Nam, S.W. Roh, M.S. Kim, Y. Sung, C.O. Jeon, H.M. Oh, and J.W. Bae. 2008. Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Appl Environ Microbiol*. 74:5975-5985.

Kim, U.J., H. Shizuya, P.J. de Jong, B. Birren, and M.I. Simon. 1992. Stable propagation of cosmid sized human DNA inserts in an F factor based vector. *Nucleic Acids Res*. 20:1083-1085.

Kjelleberg, S., T.A. Stenstrom, and G. Odham. 1979. Comparative-Study of Different Hydrophobic Devices for Sampling Lipid Surface-Films and Adherent Microorganisms. *Mar Biol*. 53:21-25.

Kunin, V., A. Copeland, A. Lapidus, K. Mavromatis, and P. Hugenholtz. 2008. A bioinformatician's guide to metagenomics. *Microbiol Mol Biol Rev*. 72:557-578, Table of Contents.

Lane, D.J., B. Pace, G.J. Olsen, D.A. Stahl, M.L. Sogin, and N.R. Pace. 1985a. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci U S A*. 82:6955-6959.

Lane, D.J., D.A. Stahl, G.J. Olsen, D.J. Heller, and N.R. Pace. 1985b. Phylogenetic analysis of the genera Thiobacillus and Thiomicrospira by 5S rRNA sequences. *J Bacteriol*. 163:75-81.

Liles, M.R., L.L. Williamson, J. Rodbumrer, V. Torsvik, R.M. Goodman, and J. Handelsman. 2008. Recovery, purification, and cloning of high-molecular-weight DNA from soil microorganisms. *Appl Environ Microbiol*. 74:3302-3305.

Magot, M., B. Ollivier, and B.K. Patel. 2000. Microbiology of petroleum reservoirs. *Antonie Van Leeuwenhoek*. 77:103-116.

Martin-Cuadrado, A.B., P. Lopez-Garcia, J.C. Alba, D. Moreira, L. Monticelli, A. Strittmatter, G. Gottschalk, and F. Rodriguez-Valera. 2007. Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. *PLoS One*. 2:e914.

Metzker, M.L. 2010. Sequencing technologies - the next generation. *Nat Rev Genet*. 11:31-46.

Moreira, D., F. Rodriguez-Valera, and P. Lopez-Garcia. 2004. Analysis of a genome fragment of a deep-sea uncultivated Group II euryarchaeote containing 16S rDNA, a spectinomycin-like operon and several energy metabolism genes. *Environ Microbiol*. 6:959-969.

Moreira, D., F. Rodriguez-Valera, and P. Lopez-Garcia. 2006. Metagenomic analysis of mesopelagic Antarctic plankton reveals a novel deltaproteobacterial group. *Microbiology*. 152:505-517.

Morrison, M., P.B. Pope, S.E. Denman, and C.S. McSweeney. 2009. Plant biomass degradation by gut microbiomes: more of the same or something new? *Curr Opin Biotechnol*. 20:358-363.

Ogram, A., G.S. Sayler, and T. Barkay. 1987. The Extraction and Purification of Microbial DNA from Sediments. *J Microbiol Meth*. 7:57-66.

Orphan, V.J., L.T. Taylor, D. Hafenbradl, and E.F. Delong. 2000. Culture-dependent and culture-independent characterization of microbial assemblages associated with high-temperature petroleum reservoirs. *Appl Environ Microbiol*. 66:700-711.

Osoegawa, K., P.J. de Jong, E. Frengen, and P.A. Ioannou. 2001. Construction of bacterial artificial chromosome (BAC/PAC) libraries. *Curr Protoc Mol Biol*. Chapter 5:Unit 5 9.

Pace, N.R. 2009. Mapping the tree of life: progress and prospects. *Microbiol Mol Biol Rev*. 73:565-576.

Pansegrau, W., E. Lanka, P.T. Barth, D.H. Figurski, D.G. Guiney, D. Haas, D.R. Helinski, H. Schwab, V.A. Stanisich, and C.M. Thomas. 1994. Complete nucleotide sequence of Birmingham IncP alpha plasmids. Compilation and comparative analysis. *J Mol Biol*. 239:623-663.

Parsley, L.C., E.J. Consuegra, K.S. Kakirde, A.M. Land, W.F. Harper, Jr., and M.R. Liles. 2010. Identification of diverse antimicrobial resistance determinants carried on bacterial, plasmid, or viral metagenomes from an activated sludge microbial assemblage. *Appl Environ Microbiol*. 76:3753-3757.

Perri, S., D.R. Helinski, and A. Toukdarian. 1991. Interactions of plasmid-encoded replication initiation proteins with the origin of DNA replication in the broad host range plasmid RK2. *J Biol Chem*. 266:12536-12543.

Plonka, P.M., and M. Grabacka. 2006. Melanin synthesis in microorganisms--biotechnological and medical aspects. *Acta Biochim Pol*. 53:429-443.

Poyart, C., and P. Trieu-Cuot. 1997. A broad-host-range mobilizable shuttle vector for the construction of transcriptional fusions to beta-galactosidase in gram-positive bacteria. *FEMS Microbiol Lett*. 156:193-198.

Rajendhran, J., and P. Gunasekaran. 2008. Strategies for accessing soil metagenome for desired applications. *Biotechnol Adv*. 26:576-590.

Rappe, M.S., and S.J. Giovannoni. 2003. The uncultured microbial majority. *Annu Rev Microbiol*. 57:369-394.

Roberts, R.C., A.R. Strom, and D.R. Helinski. 1994. The parDE operon of the broad-host-range plasmid RK2 specifies growth inhibition associated with plasmid loss. *J Mol Biol*. 237:35-51.

Rusch, D.B., A.L. Halpern, G. Sutton, K.B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J.A. Eisen, J.M. Hoffman, K. Remington, K. Beeson, B. Tran, H. Smith, H. Baden-Tillson, C. Stewart, J. Thorpe, J. Freeman, C. Andrews-Pfannkoch, J.E. Venter, K. Li, S. Kravitz, J.F. Heidelberg, T. Utterback, Y.H. Rogers, L.I. Falcon, V. Souza, G. Bonilla-Rosso, L.E. Eguiarte, D.M. Karl, S. Sathyendranath, T. Platt, E. Bermingham, V. Gallardo, G. Tamayo-Castillo, M.R. Ferrari, R.L. Strausberg, K. Nealson, R. Friedman, M. Frazier, and J.C. Venter. 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol*. 5:e77.

Sanchez-Amat, A., F. Solano, and P. Lucas-Elio. 2010. Finding new enzymes from bacterial physiology: a successful approach illustrated by the detection of novel oxidases in Marinomonas mediterranea. *Mar Drugs*. 8:519-541.

Sanger, F., S. Nicklen, and A.R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*. 74:5463-5467.

Savage, D.C. 1977. Microbial ecology of the gastrointestinal tract. *Annu Rev Microbiol*. 31:107-133.

Schloss, P.D., and J. Handelsman. 2004. Status of the microbial census. *Microbiol Mol Biol Rev*. 68:686-691.

Schmid, A., J.S. Dordick, B. Hauer, A. Kiener, M. Wubbolts, and B. Witholt. 2001. Industrial biocatalysis today and tomorrow. *Nature*. 409:258-268.

Shendure, J., and H. Ji. 2008. Next-generation DNA sequencing. *Nat Biotechnol*. 26:1135-1145.

Shingler, V., and C.M. Thomas. 1989. Analysis of nonpolar insertion mutations in the trfA gene of IncP plasmid RK2 which affect its broad-host-range property. *Biochim Biophys Acta*. 1007:301-308.

Shizuya, H., B. Birren, U.J. Kim, V. Mancino, T. Slepak, Y. Tachiiri, and M. Simon. 1992. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proc Natl Acad Sci U S A*. 89:8794-8797.

Sia, E.A., R.C. Roberts, C. Easter, D.R. Helinski, and D.H. Figurski. 1995. Different relative importances of the par operons and the effect of conjugal transfer on the maintenance of intact promiscuous plasmid RK2. *J Bacteriol*. 177:2789-2797.

Sieburth, J.M.N., P.J. Willis, K.M. Johnson, C.M. Burney, D.M. Lavoie, K.R. Hinga, D.A. Caron, F.W. French, P.W. Johnson, and P.G. Davis. 1976. Dissolved Organic-Matter and Heterotrophic Microneuston in Surface Microlayers of North-Atlantic. *Science*. 194:1415-1418.

Simon, C., A. Wiezer, A.W. Strittmatter, and R. Daniel. 2009. Phylogenetic diversity and metabolic potential revealed in a glacier ice metagenome. *Appl Environ Microbiol*. 75:7519-7526.

Simon, R., U. Priefer, and A. Puhler. 1983. A Broad Host Range Mobilization System for Invivo Genetic-Engineering - Transposon Mutagenesis in Gram-Negative Bacteria. *Bio-Technol*. 1:784-791.

Singh, B.K. 2010. Exploring microbial diversity for biotechnology: the way forward. *Trends Biotechnol*. 28:111-116.

Singh, S.B., and F. Pelaez. 2008. Biodiversity, chemical diversity and drug discovery. *Prog Drug Res*. 65:141, 143-174.

Stafsnes, M.H., K.D. Josefsen, G. Kildahl-Andersen, S. Valla, T.E. Ellingsen, and P. Bruheim. 2010. Isolation and characterization of marine pigmented bacteria from Norwegian coastal waters and screening for carotenoids with UVA-blue light absorbing properties. *J Microbiol*. 48:16-23.

Stahl, D.A., D.J. Lane, G.J. Olsen, and N.R. Pace. 1985. Characterization of a Yellowstone hot spring microbial community by 5S rRNA sequences. *Appl Environ Microbiol*. 49:1379-1384.

Staley, J.T., and A. Konopka. 1985. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu Rev Microbiol*. 39:321-346.

Stein, J.L., T.L. Marsh, K.Y. Wu, H. Shizuya, and E.F. DeLong. 1996. Characterization of uncultivated prokaryotes: Isolation and analysis of a 40-kilobase-pair genome fragment front a planktonic marine archaeon. *Journal of Bacteriology*. 178:591-599.

Thomas, C.M., and D.R. Helinski. 1989. Vegetative replication and stable inheritance of IncP plasmids. *In* Promiscuous plasmids of gram-negative bacteria. C.M. Thomas, editor. Academic Press, San Diego. 1-25.

Torsvik, V., J. Goksoyr, and F.L. Daae. 1990. High diversity in DNA of soil bacteria. *Appl Environ Microbiol*. 56:782-787.

Toukdarian, A.E., and D.R. Helinski. 1998. TrfA dimers play a role in copy-number control of RK2 replication. *Gene*. 223:205-211.

Tringe, S.G., C. von Mering, A. Kobayashi, A.A. Salamov, K. Chen, H.W. Chang, M. Podar, J.M. Short, E.J. Mathur, J.C. Detter, P. Bork, P. Hugenholtz, and E.M. Rubin. 2005. Comparative metagenomics of microbial communities. *Science*. 308:554-557.

Tsai, Y.L., and B.H. Olson. 1991. Rapid Method for Direct Extraction of DNA from Soil and Sediments. *Appl Environ Microb*. 57:1070-1074.

Tuffin, M., D. Anderson, C. Heath, and D.A. Cowan. 2009. Metagenomic gene discovery: how far have we moved into novel sequence space? *Biotechnol J*. 4:1671-1683.

Tyson, G.W., and J.F. Banfield. 2005. Cultivating the uncultivated: a community genomics perspective. *Trends Microbiol*. 13:411-415.

Tyson, G.W., J. Chapman, P. Hugenholtz, E.E. Allen, R.J. Ram, P.M. Richardson, V.V. Solovyev, E.M. Rubin, D.S. Rokhsar, and J.F. Banfield. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*. 428:37-43.

Uchiyama, T., T. Abe, T. Ikemura, and K. Watanabe. 2005. Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. *Nat Biotechnol*. 23:88-93.

Valla, S., K. Haugan, R. Durland, and D.R. Helinski. 1991. Isolation and properties of temperature-sensitive mutants of the trfA gene of the broad host range plasmid RK2. *Plasmid*. 25:131-136.

vanElsas, J.D., V. Mantynen, and A.C. Wolters. 1997. Soil DNA extraction and assessment of the fate of Mycobacterium chlorophenolicum strain PCP-1 in different soils by 16S ribosomal RNA gene sequence based most-probable-number PCR and immunofluorescence. *Biol Fert Soils*. 24:188-195.

Venter, J.C., K. Remington, J.F. Heidelberg, A.L. Halpern, D. Rusch, J.A. Eisen, D. Wu, I. Paulsen, K.E. Nelson, W. Nelson, D.E. Fouts, S. Levy, A.H. Knap, M.W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y.H. Rogers, and H.O. Smith. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*. 304:66-74.

Voordouw, G., S.M. Armstrong, M.F. Reimer, B. Fouts, A.J. Telang, Y. Shen, and D. Gevertz. 1996. Characterization of 16S rRNA genes from oil field microbial communities indicates the presence of a variety of sulfate-reducing, fermentative, and sulfide-oxidizing bacteria. *Appl Environ Microbiol*. 62:1623-1629.

Waters, V.L. 2001. Conjugation between bacterial and mammalian cells. *Nat Genet*. 29:375-376.

Wawrik, B., L. Kerkhof, G.J. Zylstra, and J.J. Kukor. 2005. Identification of unique type II polyketide synthase genes in soil. *Appl Environ Microbiol*. 71:2232-2238.

Whitman, W.B., D.C. Coleman, and W.J. Wiebe. 1998. Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A*. 95:6578-6583.

Wiater, L.A., A. Marra, and H.A. Shuman. 1994. Escherichia coli F plasmid transfers to and replicates within Legionella pneumophila: an alternative to using an RP4-based system for gene delivery. *Plasmid*. 32:280-294.

Wild, J., Z. Hradecna, and W. Szybalski. 2002. Conditionally amplifiable BACs: switching from single-copy to high-copy vectors and genomic clones. *Genome Res*. 12:1434-1444.

Williamson, L.L., B.R. Borlee, P.D. Schloss, C. Guan, H.K. Allen, and J. Handelsman. 2005. Intracellular screen to identify metagenomic clones that induce or inhibit a quorum-sensing biosensor. *Appl Environ Microbiol*. 71:6335-6344.

Wooley, J.C., A. Godzik, and I. Friedberg. 2010. A primer on metagenomics. *PLoS Comput Biol*. 6:e1000667.

Xu, M., D. Fujita, and N. Hanagata. 2009. Perspectives and challenges of emerging single-molecule DNA sequencing technologies. *Small*. 5:2638-2649.

Yokouchi, H., Y. Fukuoka, D. Mukoyama, R. Calugay, H. Takeyama, and T. Matsunaga. 2006. Whole-metagenome amplification of a microbial community associated with scleractinian coral by multiple displacement amplification using phi29 polymerase. *Environ Microbiol*. 8:1155-1163.

Yooseph, S., G. Sutton, D.B. Rusch, A.L. Halpern, S.J. Williamson, K. Remington, J.A. Eisen, K.B. Heidelberg, G. Manning, W. Li, L. Jaroszewski, P. Cieplak, C.S. Miller, H. Li, S.T. Mashiyama, M.P. Joachimiak, C. van Belle, J.M. Chandonia, D.A. Soergel, Y. Zhai, K. Natarajan, S. Lee, B.J. Raphael, V. Bafna, R. Friedman, S.E. Brenner, A. Godzik, D. Eisenberg, J.E. Dixon, S.S. Taylor, R.L. Strausberg, M. Frazier, and J.C. Venter. 2007. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol*. 5:e16.

Youssef, N., M.S. Elshahed, and M.J. McInerney. 2009. Microbial processes in oil fields: culprits, problems, and opportunities. *Adv Appl Microbiol*. 66:141-251.

Zaballos, M., A. Lopez-Lopez, L. Ovreas, S.G. Bartual, G. D'Auria, J.C. Alba, B. Legault, R. Pushker, F.L. Daae, and F. Rodriguez-Valera. 2006. Comparison of prokaryotic diversity at offshore oceanic locations reveals a different microbiota in the Mediterranean Sea. *FEMS Microbiol Ecol*. 56:389-405.

Zaehner, H., and H.P. Fiedler. 1995. The need for new antibiotics: possible ways forward. *In* Fifty years of antimicrobials: past perspectives and future trends. P.A. Hunter, G.K. Darby, and N.J. Russel, editors. Cambridge University Press, Cambridge. 67-84.

Zhou, J., M.A. Bruns, and J.M. Tiedje. 1996. DNA recovery from soils of diverse composition. *Appl Environ Microbiol*. 62:316-322.

Appendix 1.

Sample collection and DNA extraction from the sea surface microlayer

Collection of water from the surface microlayer at six different locations along the Trondheim fjord was performed essentially as described by Kjelleberg et al. (1979). Approximately 250 samplings were required to collect a 20-litre sample. The samples were pre-filtered though a 250 µm grid to remove the largest particles, and then filtered through nylon filters with 60 µm pore size (Millipore). The filtrates were concentrated in two steps, first by tangential flow filtration using a Pellicon XL50 ultra filtration system, followed by further concentration in an Amicon stirred cell (Millipore). The samples were then diluted in 2 x STE-buffer (1 M NaCl, 0.1 M Na_2EDTA, 10 mM Tris-HCl, pH 8.0) to give a final DNA concentration of 150-250 µg/ml. The cell suspensions were mixed with an equal volume of 2% InCert agarose (Cambrex) in PBS-buffer (0.8% NaCl, 0.02% KCl, 0.144% Na_2HPO_4, 0.024 % KH_2PO_4, pH 7.4) and agarose plugs were molded in disposable plug molds (BioRad).

High molecular weight DNA was extracted according to the method of Stein *et al.* (1996), but with some modifications. The agarose plugs were lysed in 25 ml lysis buffer (10 mM Tris-HCl, 50 mM NaCl, 0.1 mM Na_2EDTA, 1% N-Lauroyl Sarcosine sodium salt, 0.2% sodium deoxycholate, 1 mg/ml lysozyme, pH 8.0) for 3 hours. The plugs were transferred to 25 ml EPS buffer (1% N-Lauroyl Sarcosine sodium salt and 1 mg/ml proteinase K in 0.5 M Na_2EDTA, pH 8) and incubated at 55°C for 16 hours. Proteinase K was inactivated and the plugs were dialysed and stored as described by Osoegawa *et al.* (Osoegawa et al., 2001).

Appendix 2.

Methods involved in detection of NRPS- and PKS genes in the fosmid library

Probe synthesis

PCR-fragments were synthesized using degenerated primers targeting conserved domains of the PKS I β-ketoacyl synthase (KS) domain (primers KSMA-F and KSMB-R (Izumikawa et al., 2003)), the PKS II KS domain (primers 540F and 1100R (Wawrik et al., 2005)), and the NRPS adenylation domain (primers ADEdom5 and ADEdom3 (Busti et al., 2006)). Plasmid DNA isolated from a mixture of the entire fosmid library was used as template for the PCR reactions. No PCR-product was obtained for the PKS II primer-pair, even after several attempts. The resulting PCR fragments from the other reactions were ligated into the pDrive cloning vector (Qiagen), and selected insert-bearing plasmids were sequenced (Sanger procedure). Plasmids with inserts annotated to originate from PKS- or NRPS- genes, respectively, were selected as templates for the probe synthesis using the PCR DIG probe synthesis kit (Roche)

Colony hybridization

Library clones were plated from the pooled freeze-stocks onto eight L-agar plates containing L-arabinose such that each plate contained approximately 1500 colonies. A positive control, i.e. pRS44 containing a NRPS annotated PCR-product, was simultaneously plated onto L-agar plates with and without L-arabinose to evaluate the relevance of the fosmid copy-number. Autoclaved positively charged nylon membranes (Roche) were overlaid the pre-cooled plates and then lifted off. After drying the colony-containing membranes, they were placed on 3 mm Whatman paper drenched in denaturation solution (0.5 M NaOH, 1.5 M NaCl) for 10 min, then on Whatman paper drenched in neutralization solution (3.0 M NaCl, 0.5 M Tris-HCl, pH 7.5) for 10 min and finally on Whatman paper drenched in 2xSSC (18 g/l NaCl, 7.7 g/l Na-Citrate, pH 7.0). Fixation of the DNA to the membranes was done by incubating the membranes at 120°C for 30 min. Hybridization of probes and detection of positive clones were conducted using the Easy DIG hybridization solution and DIG Nucleic acid Detection kit, respectively, according to the supplier's instructions (Roche).

Paper I

Paper II

New and improved biological tools for conjugal transfer of plasmids

Trine Aakvik[1], Kristin Fløgstad Degnes[2], Rahmi Lale[1], Malin Lando[1] and Svein Valla[1*]

[1]Department of Biotechnology, Norwegian University of Science and Technology, 7491 Trondheim, Norway; [2]Department of Biotechnology, SINTEF Materials and Chemistry, Trondheim, Norway.

*Corresponding author

Phone: +47 73598694

Fax : +47 73591283

E-mail: svein.valla@biotech.ntnu.no

Is not included due to copyright

Paper III

Production *in Escherichia coli* of a melanin-like polymer from a marine sediment metagenomic clone expressing rubrerythrin

Trine Aakvik[1], Finn Drabløs[2], Trygve Andreassen[1] and Svein Valla[1*]

[1]Department of Biotechnology, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway.

[2]Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway.

*Corresponding author
Phone: +47 73598694
Fax : +47 73591283
E-mail: svein.valla@biotech.ntnu.no

Paper IV

High coverage sequencing of DNA from microorganisms living at 250 bars 2.5 kilometers subsurface in a Norwegian Sea oil reservoir.

Hans K. Kotlar[4#], Anna Lewin[1#], Jostein Johansen[2], Mimmi Throne-Holst[3¤], Thomas Haverkamp[5], Sidsel Markussen[3], Asgeir Winnberg[3], Philip Ringrose[4], Trine Aakvik[1], Einar Ryeng[2], Kjetill Jakobsen[5], Finn Drabløs[2], Svein Valla[1*]

[1] Dept. of Biotechnology, NTNU, Trondheim, Norway
[2] Dept of Cancer Research and Molecular Medicine, NTNU, Trondheim, Norway
[3] Sintef Material and Chemistry, Biotechnology, Trondheim, Norway
[4] Statoil ASA, Trondheim, Norway
[5] CEES and MERG, Dept of Biology, Oslo, Norway
[#] HKK and AL should be seen as equal contributors to the work described.
[¤] current address Statoil ASA, Trondheim, Norway.
* corresponding author

Is not included due to copyright

Paper V

Metagenomic libraries for functional screening

Trine Aakvik[1], Rahmi Lale[1], Mark Liles[2] and Svein Valla[1*]

[1]Department of Biotechnology, Norwegian University of Science and Technology, 7491 Trondheim, Norway.

[2]Department of Biological Sciences, Auburn University, Auburn, AL 36894, USA

*Corresponding author

Phone: +47 73598694

Fax : +47 73591283

E-mail: svein.valla@biotech.ntnu.no

Number of words:

Number of tables: 1

Number of figures: 1

Running title: Metagenomics and functional screening

Keywords: metagenomics, metagenomic library, metagenomic vectors, function-based screening

Abstract

The need for metagenomics to fully exploit the global genetic diversity of microorganisms follows from the fact that the majority of them cannot be readily cultivated in the laboratory. Functional screening of metagenomic libraries has proven to be a successful approach for discovery of novel biomolecules resulting from the activity of gene products lacking significant sequence similarities to those previously known. The first critical step is the isolation of DNA from the environment of choice. DNA quality is measured in terms of molecular weight, purity and yield, with all three criteria being critical for construction of large-insert metagenomic libraries, while small insert libraries are much easier to establish. The vectors in which the libraries are constructed are chosen on the basis of insert size, the target of interest, and the host selected for expression. *Escherichia coli* is normally used as host for library constructions, but to enhance the probability of detecting heterologously expressed gene products alternative hosts are now frequently being used for screening. Numerous screening protocols with varying complexity are available, ranging from direct selection of the clones of interest, identification by colony phenotypes on solid media, to sometimes quite costly and laborious procedures involving advanced robotics and measurements of biological activities in each separate clone. In conclusion the main advantage of functional screening is that it requires no prior knowledge of the sequence-function relation, and this approach is therefore likely to be used in the years to come in parallel with the rapidly developing high-capacity sequencing procedures.

Introduction

The genomes of Eubacteria represent the major reservoir of genetic diversity on Earth [Whitman et al., 1998; Ferrer et al., 2009a]. This diversity represents a huge resource that can be exploited for the recovery of novel biomolecules, and its exploration is also important for the understanding of microbial ecology. It is well known that the vast majority of microorganisms (possibly more than 99%) cannot be readily cultured in a laboratory setting [Amann et al., 1995; Rappé and Giovannoni, 2003]. Although the cultivation success rate can certainly be improved [Tyson and Banfield, 2005], the use of a metagenomic approach involving the direct extraction, cloning and analysis of DNA from its natural environment [Handelsman et al., 1998] is a more inclusive strategy to access microbial genetic reservoirs. The strength of the metagenomics approach for the identification of novel metabolites is reflected in that most of the genes encoding pathways for synthesis of new biomolecules discovered from metagenomic libraries are either weakly related or unrelated to previously known genes [Daniel, 2005]. In contrast, when for instance searching for new/novel antibiotics among cultivable actinobacteria, the rediscovery rate may be as high as 99.9% [Zaehner and Fiedler, 1995].

DNA isolated from the environment can be explored either by direct sequencing, by sequencing of gene libraries prepared from the isolated DNA, or by functional screening of such libraries. Sequence-based methods have resulted in identification of many novel natural products, however, identification of genes of interest by this method depends on the existence of some kind of similarity to already known genes or gene products. Functional screening does not require pre-existing sequence information, and is thus the best strategy to identify entirely new classes of natural products. A functional metagenomics approach does, however, depend on successful heterologous expression of the gene or genes responsible for their respective function, and also that the gene products are active in the host used for expression. In this review we describe some of the tools used to construct metagenomic libraries, as well as the array of different screening technologies for detection of different (and potentially novel) functions that may be expressed from metagenomic clones.

Isolation of environmental DNA and enrichment of target-genes

The choice of metagenomic DNA extraction method and cloning strategy adopted should coincide with the specific genetic sequence and/or functional activity that are sought within a metagenomic library. Specifically, the criteria that should be considered early on in this process are 1) the likelihood of detecting a particular gene or gene family using a strict sequence-based approach, 2) the availability of a functional screen (or ideally, a selection) for the respective gene product, 3) the predicted size of the genetic loci (or pathway), 4) the previous knowledge of heterologous expression of the targeted gene within an *Escherichia coli* or other host, and 5) the availability of resources and/or time required for constructing and screening a large-insert metagenomic library compared to less labor- and resource-intensive strategies. For example, expression of an enzyme for which there is a plate-based screen (i.e., halo formation due to substrate conversion) would indicate a preferred strategy of constructing a small-insert library within an expression vector. On the other hand, screening for the presence of antimicrobial active clones would be dependent in most cases on expression of biosynthetic pathways contained on large-insert clones, preferably with the option of heterologous expression in multiple bacterial hosts. Therefore, careful consideration of metagenomic strategy is critical prior to initiating work in this field. The different steps involved in metagenomics for functional screening approaches are illustrated in Figure 1.

For small-insert metagenomic libraries or PCR/pyrosequencing applications, many direct DNA extraction methods are available (e.g., bead-beating) that provide a sufficient degree of cell lysis and subsequent purification [Ogram et al., 1987; Tsai and Olson, 1991; Zhou et al., 1996], and kits are available from many commercial suppliers. The physicochemical characteristics of the environmental sample (i.e., pH, organic content) should be taken into account in order to use the correct sampling scale for DNA extraction. It should also be considered that alternative extraction methods might improve DNA yield and/or purity [Sagova-Mareckova et al., 2008]. It is advisable to use the efficiency of PCR amplification of extracted DNA (e.g., of 16S rRNA genes) as a surrogate for monitoring the suitability of the DNA for cloning. Given that any one method for DNA extraction may have an inherent bias in the representation of specific bacterial taxa as revealed by PCR [Martin-Laurent et al., 2001], there is no single perfectly suitable extraction method for any environmental sample. When using a previously uncharacterized environmental sample, it is therefore wise to use multiple extraction methods and to select a method for metagenomic analysis based on empirical results of DNA yield, ribotype richness, and suitability of the DNA template for PCR amplification and cloning [see also Chapters 10 and 11, Vol. II].

Large-insert cloning requires adoption of DNA extraction methods that are considerably less harsh to prevent DNA shearing in order to procure high molecular weight (HMW) metagenomic DNA. The use of indirect DNA extraction methods, in which microbial cells are first recovered and washed prior to cell lysis, have advantages for the removal of the environmental matrix and/or eukaryotic cells, and in preserving genomic DNA integrity [Steffan et al., 1988; Torsvik et al., 1990; Krsek and Wellington, 1999]. The primary disadvantage of such methods for extraction of high molecular weight DNA is the inability to use organic extraction and other purification schemes that are highly effective in removing

DNA-associated contaminants (e.g., humic acids) due to the resultant DNA fragmentation. Other potential disadvantages in the use of indirect extraction methods include decreased DNA yield and ribotype richness as compared to direct extraction [Steffan et al., 1988; Krsek and Wellington, 1999]. Monitoring of the microbial populations of interest (e.g., Actinobacteria) during the cell recovery and lysis steps is recommended in adopting an indirect extraction method. Purification of HMW metagenomic DNA embedded within agarose plugs may be achieved through use of polyvinylpolypyrrolidone or formamide solutions to remove DNA-associated contaminants and still permit downstream molecular manipulation [Steffan et al., 1988; Liles et al., 2008]. For further discussion of DNA purification strategies for metagenomic applications please refer to Chapters 10 and 11, vol. II.

The frequency of metagenomic clones expressing specific traits of interest is low in libraries constructed on the basis of total DNA extraction. This means that in the screening protocols which involve a significant amount of resource input per clone, the vast majority of the clones are expected to give a negative result. One way to reduce this problem is to enrich the environmental sample for the population that carries the genes of interest prior to library construction. This may be achieved by exposing the microorganisms in the sample (*in situ*) to a selective pressure based on nutritional, physical or chemical criteria. Such enrichment steps result in loss of diversity in the resulting metagenomic library, however, this has proven to be an advantageous strategy when searching for a specific activity [Entcheva et al., 2001; Knietsch et al., 2003a, b; Rees et al., 2003; Gabor et al., 2004b]. Another possibility is to enrich for high G+C% content in the cloned DNA, since many G+C-rich microorganisms are of particular interest for secondary metabolite expression. The DNA from these organisms can be enriched through an ultracentrifugation step after DNA extraction. An even more elegant enrichment strategy is the incorporation of labelled nucleotides into the DNA of metabolically active organisms. This can be accomplished through the use of a stable isotope-labelled substrate (stable isotope probing) or a substrate containing bromodeoxyuridine, followed by separation of the labelled DNA through ultracentrifugation or immunocapture, respectively (various enrichment technologies are reviewed in Cowan et al. [2005]).

For identification of enzymatic activities with specific tolerances to environmental extreme conditions, it may be useful to isolate microorganisms living in a respective extreme environment and search for biomolecules that are stable under such hostile conditions. Due to the often low biomass in such environments, it may be challenging to extract sufficient DNA for a metagenomic approach [Ferrer et al., 2009a]. One way to overcome such obstacles is by the use of whole genome amplification which makes amplification of even very tiny amounts of environmental DNA possible through the use of phi29 DNA polymerase [Abulencia et al., 2006; Yokouchi et al., 2006; Kim et al., 2008; Simon et al., 2009]. Yamada et al. [2008] also demonstrated that this polymerase could successfully be used together with target-specific primers (containing locked nucleic acids) to enrich an environmental sample for target DNA prior to inverse PCR (IPCR) and cloning. Drawbacks of whole genome amplification include amplification biases and formation of chimeric sequences, and also the possible introduction of point mutations that may reduce the chances of cloning a functional gene product.

Vectors for function-based metagenomic studies

The ability to detect novel activities in metagenomic libraries is reliant upon successful expression of the vector borne heterologous gene(s) of interest, and this requires several criteria to be fulfilled. First of all, the chosen insert size must be large enough to include the gene or entire cluster of genes needed to express the function of interest. Secondly, a promoter and ribosome binding site that are compatible with the expression machinery of the host are necessary for successful expression of the cloned genes. In addition, several factors may need to be provided by the host cell such as proper transcription factors, inducers, chaperones, cofactors, post-translationally acting factors, and secretion mechanisms. Codon usage may also become a limiting factor for expression, as for example G+C% content of bacteria can vary significantly. The potential toxicity of the final gene product is yet another possible obstacle to a successful screening result. One potential way to solve these complex problems is to use vectors that can be transferred to and maintained in a variety of different hosts for expression of the metagenomic DNA, such that one can screen a library within a host that is considered likely to express the specific gene product(s) of interest.

Vectors for small-insert metagenomic libraries

Successful heterologous expression of metagenomic DNA is usually simplest to achieve if only one gene needs to be expressed. In such cases a promoter known to function well in the host used for screening can be a part of the vector such that it transcribes the cloned DNA. One may also use a promoter transcribing in the opposite direction from the other end of the insert and thus become independent of the orientation of the gene searched for, as for example described in Lämmle et al. [2007]. Even in such cases one has to consider prior to library construction at what level one would like the desired gene to be expressed, especially if the gene product has a potential to be toxic to the host. The use of gradually inducible promoters may solve this problem. Vectors with varying copy numbers can also be used to modulate expression levels, but very high copy number vectors may in some cases display toxic levels of expression even in the absence of induction. Instead of using plasmids one may use λ phage based vectors (phagemids). Since such expression vector system are accompanied by lysis of the host cell during expression, the problem of gene product toxicity commonly experienced with high copy number vectors is avoided [Ferrer et al., 2005].

Vectors for large-insert metagenomic libraries

If the product of interest is a compound that requires many gene products, the use of large-insert vectors is essential. The chosen insert size must also be large enough to include all the genes needed to express the function of interest. Since bacterial genetic pathways required for secondary metabolite synthesis (e.g. antibiotics) are very often clustered on a genome, a large-insert metagenomic approach is technically feasible. On the other hand it should be remembered that fragments are randomly cloned when constructing metagenomic libraries. This means that the insert sizes must be significantly bigger than the expected size of a cluster, as it is otherwise unlikely that an entire cluster will be present in a single clone. Note also that for expression of complex secondary metabolite biosynthetic pathways, the lack of a single essential gene may be sufficient to abolish the synthesis of the final product. For large genetic operons, vector promoters may not achieve expression of all required genes. The transcription of a cloned genetic operon will therefore often depend on native or cryptic promoters already present in the environmental DNA and that are used naturally in the host of origin to express the genes; therefore, heterologous expression of large-insert clones within different hosts may substantially improve the probability that operons may be expressed and their respective natural product(s) detected.

The first large-insert cloning vector described was a cosmid [Collins and Hohn, 1978], which is a multi-copy cloning vector for inserts up to 40 kb that can be packaged into lambda phage heads due to the presence of dual *cos*-sites. The observed instability of genomic libraries constructed in these multi-copy vectors led to the development of the fosmid vector, which is a low copy-number cosmid vector based on the *E. coli* F factor replicon [Kim et al., 1992]. The copy number of this replicon is strictly controlled in *E. coli*, yielding only 1-2 copies per cell. Simultaneously, a bacterial artificial chromosome (BAC) vector based on the same replicon was constructed, but with the capacity of cloning inserts with sizes up to around 300 kb [Shizuya et al., 1992]. The strength of BAC vectors is that large biosynthetic gene clusters can be cloned, such as polyketide or antibiotic synthesis gene clusters which may be over 40 kb in size (e.g. the 55 kb erythromycin- [Brikun et al., 2004] and the 124 kb nystatin- [Brautaset et al., 2000] clusters). BAC libraries can, however, also contain rather small inserts since there is no selection against small fragments as in the fosmid system. Due to the F factor system, both fosmid- and BAC libraries can be established and stably maintained, and these vectors display a low level of chimerism. Wild et al. [2002] combined the stabilizing features of the F factor during the maintenance phase with a possibility of *in vivo* amplification for achievement of higher levels of DNA recovery and purity upon plasmid isolation (referred to as the copy-control fosmid/-BAC vectors, now distributed by Epicentre and Lucigen). Due to the possibility of inducing higher gene dose of the metagenomic DNA, this copy control system also increases the chances for detection of activities from gene products expressed at low levels.

Host range of vectors useful in metagenomics

Construction of metagenomic libraries is most easily done in *E. coli* due to its well established genetics and great transformation competence. Moreover, using *E. coli* as an expression host for function-based screens has also resulted in detection of many novel activities, with a very diverse phylogenetic origin of the DNA sources [Handelsman et al., 1998; Tirawongsaroj et al., 2008; Heath et al., 2009; Simon et al., 2009]. This may be attributable to the fact that *E. coli* has relatively relaxed requirements regarding promoter recognition and translation initiation compared to other well-established expression hosts such as *Bacillus subtilis* and *Streptomyces lividans* [Gabor et al., 2007]. Another advantage of using *E. coli* as host is that it is commonly used in industrial fermentation, thus easing the possible commercial production of the identified product(s). In spite of the benefits from adopting *E. coli* as a heterologous host, it is also clear that many genes will not be expressed from their native promoters in *E. coli* and it is therefore potentially advantageous to expand the library host expression range. This has been demonstrated in several studies, as for example the one by Martinez et al. [2004] where it was reported that *E. coli*, *Pseudomonas putida* and *Streptomyces lividans* differed in their abilities to express heterologous gene clusters. For more examples see Li et al., [2005], Wexler et al., [2005], Angelov et al., [2009], and Craig et al., [2009]. An *in silico* study by Gabor and co-workers predicted that in average only 40% of the genes from 32 different genomes had expression signals that could be recognized by the expression machinery of *E. coli* [Gabor et al., 2004a]. Furthermore, in an experimental study by Warren et al. [2008] the transcription levels of foreign DNA in *E. coli* were compared by RT-PCR, and it was, not surprisingly, found that the degree of expression of genes from phylogenetically diverse organisms was higher the closer the source organism was related to *E. coli*.

The host specificity of the different vectors developed hitherto to improve the chances for heterologous gene expression includes a wide range of both Gram-negative and Gram-positive hosts (see Table 1). The vectors are different with regard to how they are transferred to- and maintained in the non-*E. coli* host. Some vectors contain an origin for conjugal transfer, usually *oriT* from RK2 (e.g., pRK7813, see Table 1), which allows the transfer of constructed libraries from *E. coli* to other hosts with high efficiencies. In the new hosts the vectors with inserts are either integrated into the chromosome (e.g., pCT3FK and pMBD14, see Table 1) or they replicate extra-chromosomally via a vector-provided replicon suitable for the host (e.g., pJWC1 and pUvBBAC, see Table 1). The autonomously replicating vectors have the potential for a higher copy-number than those integrated into the chromosome, often enhancing the chances for product detection during screening. The pRS44 vector described in Aakvik et al. [2009] is a vector capable of mobilized conjugation and that has both the F factor replicon and the broad host range RK2 replicon, meaning that a (large-insert) library constructed in this vector can be efficiently transferred to and screened in many different hosts. A particularly important finding in this study was that very large-inserts could be established and maintained in species different from *E. coli.*

Expression in bacterial hosts is usually limited to Eubacterial genes, but metagenomic DNA can also contain DNA of archeal and eukaryotic origin [Gabor et al., 2003]. To circumvent the lack of expression of eukaryotic DNA in bacteria partly due to the presence of non-coding sequences it is also possible to construct libraries via cDNA [Grant et al., 2006; Bailly et al., 2007]. These libraries will, however, only represent expressed genes, and the reverse transcription process will limit the possible insert size. An alternative method to the use of non-*E. coli* hosts is to develop *E. coli* such that its expression capabilities becomes expanded. One such example is given by Bernstein [2007] who reported up to a 12-fold higher expression level from G+C-rich DNA in *E. coli* by directed evolution of the *E. coli* ribosomal protein S1. Other strategies to enhance *E. coli* heterologous expression include introduction of tRNA genes for rare amino acid codons [Rosano and Ceccarelli, 2009] and co-expression of specific chaperone proteins [Wall and Plückthun, 1995; Nishihara et al., 1998; Ferrer et al., 2004].

Cloning strategies

The ligation of restriction digested or blunt-ended metagenomic DNA into vectors and subsequent transformation into a library host strain are performed using a variety of different strategies depending on what kind of library is desired. For construction of small-insert libraries the DNA is most often partially restriction digested and cloned according to standard techniques, see for example Henne et al., [1999], Riesenfeld et al., [2004], and Lämmle et al., [2007]. Waschkowitz et al. [2009] compared the use of a T4 DNA ligase-based and a topoisomerase-based cloning method for construction of such small-insert metagenomic libraries, finding that both higher amounts of clones and larger insert sizes were obtained per gram of DNA using the topoisomerase based method. When aiming for large-insert libraries, the cloning strategy depends to a greater extent on the molecular weight of the isolated DNA. It has been demonstrated that the use of optimized extraction methods makes it possible to obtain libraries with average insert sizes up to 100 kb even after restriction treatment of the metagenomic DNA [Liles et al., 2008; Ouyang et al., 2009]. When the isolated DNA is not of high molecular weight, an alternative approach is to employ blunt-end or T-A ligation to avoid further decrease in insert-size, see Wilkinson et al., [2002], and Lee et al., [2004] for examples. In general, however, construction of large libraries with large inserts is often challenging due to problems with the quality of DNA from many environmental sources.

Screening of metagenomic libraries

Several different methods exist for functional screening, and as the frequency of the metagenomic clones that express a given activity is low, the method should be either highly sensitive or carried out in a high throughput manner, or preferably both. Whereas some screening approaches can be carried out on agar plates supplemented with appropriate substrates, other approaches require the use of 96- or 384- (or possibly even higher) well formats, which enable separate cultivation of the individual clones with subsequent assaying. The accomplishment of such high throughput experiments requires the use of robotic systems, which not only improves the reproducibility and reduces data scattering, but also allows for miniaturisation that may reduce costs (due to reduced amounts of both the individual clones and the components of the assay). Handling of clone pools in liquid cultures tends to result in biased libraries, but Hrvatin and Piel [2007] suggested a method for handling primary libraries that reduces this problem, involving semisolid media for generation of growth in three dimensions.

Screening by growth selection

The most convenient screening methods are based on growth selection, where the presence of a given activity provides a growth advantage to the organism in which the gene of interest is expressed. Selection for antibiotic or heavy metal resistance is one example [Diaz-Torres et al., 2003; Riesenfeld et al., 2004; Mirete et al., 2007; Mori et al., 2008; Kazimierczak et al., 2009], whereas another approach is the use of mutant host strains that require heterologous complementation for growth under selective conditions. Examples of the latter screening approach are detection of phosphonate utilization pathways [Martinez et al., 2009], DNA polymerase I [Simon et al., 2009], lysine racemases [Chen et al., 2009], naphthalene dioxygenase [Ono et al., 2007], enzymes involved in poly-3-hydroxybutyrate metabolism [Wang et al., 2006], glycerol dehydratases [Knietsch et al., 2003a], operons for biotin biosynthesis [Entcheva et al., 2001], and genes encoding Na+/H+ antiporters [Majerník et al., 2001].

Screening by detection of specific phenotypes

When the function of interest does not provide the basis for selection, an alternative approach is through detection of specific phenotypes. For this approach the individual clones in the library need to be physically separated, either on agar media or in liquid phase in microtiter plates, and assayed individually for the given trait preferably in a high throughput manner. To be able to identify enzymatic activities, chemical dyes and insoluble, chromogenic or fluorogenic substrates can be incorporated into the growth medium [Daniel, 2005]. Recent examples are identification of metalloproteases [Waschkowitz et al., 2009] and esterases [Elend et al., 2006; Chu et al., 2008; Wu and Sun, 2009] that are identified due to formation of clear halos on agar plates containing skimmed milk or tributyrin, respectively, and also extradiol dioxygenases [Suenaga et al., 2007] identified through formation of a yellow coloured product. Antimicrobial or antifungal activity may be detected through growth inhibition of a suitable indicator organism which is either overlaid colonies of the library in a soft agar medium or grown in microtiter plates in the presence of extracts of the clones [Rondon et al., 2000; Courtois et al., 2003; Brady et al., 2004; Chung et al., 2008; Craig et al., 2009]. Detection assays may also be carried out in mutated strains where heterologous complementation of a certain trait results in a detectable phenotype. In general, accomplishment of detection assays on solid/agar media may be limited in sensitivity as soluble products often diffuse away from the colony, thus leading to detection of only very highly expressing clones.

Screening based on induced gene expression

Not all activities/functions can be easily linked to a detectable phenotype, and for those cases "substrate-induced gene expression screening" (SIGEX) can be an alternative [Handelsman, 2005; Uchiyama et al., 2005]. SIGEX is a high throughput screening approach for identification of catabolic genes through the use of an operon trap *gfp*-expression vector. In the vector, metagenomic inserts are cloned upstream of a promoterless *gfp* gene, thus placing expression of green fluorescent protein (GFP) under the control of promoters in the metagenomic DNA. When the clones are incubated in the presence of a target substrate that is acting as an inducer, positive clones are identified by fluorescence cell sorting (FACS). A pre-screen without the substrate allows for elimination of false positive clones. Limitations of the application are that it misses catabolite genes that are not induced upon the substrate or do not have transcriptional regulators localized close to them, and also those that are either cloned in opposite direction of the reporter gene or have a transcription terminator between the catabolite genes and the *gfp* gene. Uchiyama et al. [2005] successfully applied SIGEX to isolate 35 aromatic hydrocarbon-induced genes from a groundwater metagenomic library. Another, similar screening technique designated metabolite-related expression (METREX) has been developed by Williamson et al. [2005]. Here, a biosensor that detects small molecules inducing quorum-sensing is inside the same cell as the vector carrying the metagenomic insert. If the clone produces a quorum sensing inducer, the cell produces GFP, and the fluorescent clone can be identified by fluorescence microscopy or FACS.

Detection systems that are induced by the product of a catabolic reaction have also been reported [Mohn et al., 2006; van Sint Fiet et al., 2006]. In these systems a transcriptional regulator that is activated by the product in the reaction of interest is cloned downstream of a reporter gene (*lacZ* or *tetA*). van Sint Fiet et al. [2006] reported such a system that detects biocatalysts responsible for the formation of benzoate and 2-hydroxybenzoate from their aldehydes, and in the system described by Mohn et al. [2006] biocatalysts converting γ-hexachlorocyclohexan to 1,2,4-trichlorobenzene are detected. This product-sensing reporter system can be modified to extend the range of biocatalysts that is possible to detect.

Sequence-based screening

Sequence-based screening approaches include the use of PCR-based or hybridization-based techniques for identification of target genes through the use of primers or probes (respectively) designed from conserved regions of known genes or gene products. A high throughput alternative to the traditional colony hybridization is the use of microarrays for screening of metagenomic libraries. In such a “metagenome microarray” (MGA) the metagenomic library plasmids are spotted on a slide and specific, labelled gene probes are used for hybridization [Sebat et al., 2003; Park et al., 2008]. Sequence-based approaches may also include direct sequencing of the insert DNA followed by bioinformatics analysis of the obtained sequences [Kunin et al., 2008; Sleator et al., 2008]. In any case, the identification of novel genes are based on predictions made on the basis of already known gene sequences, thus limiting the total potential. Other limitations are that positive clones may not harbour complete genes or pathways, nor give rise to functional gene products. On the other hand, the advantage is that successful expression is not necessary, thus harbouring the potential of identifying genes that will never be expressed in heterologous hosts. Jogler et al. [2009] used a hybridization-based method to identify two fosmids containing operons with homology to magnetosome islands of magnetotactic bacteria. Other recent examples reported by Banik and Brady [2008] and Kim et al. [2007], include identification of two novel glycopeptide-encoding gene clusters and a cytochrome P450 monooxygenase gene, respectively, using PCR-based screening methods. There are also examples of using target non-specific primers, such as those designed for amplification of gene cassettes within integrons [Stokes et al., 2001; Elsaied et al., 2007; Koenig et al., 2009]. Such cassettes harbour open reading frames often encoding important features (e.g. antibiotic resistance determinants [Rowe-Magnus and Mazel, 1999]) that are flanked by *attC* sequences required for the translocation and recombination of integrons [Hall, 1997]. Rowe-Magnus [2009] also describes a sequence-independent method for recovery of such gene cassettes through the use of a three-plasmid genetic strategy. Clones containing integron gene cassettes within a metagenomic library are here fused to another plasmid due to recombination at the *attC* site, and selection of fusions is accomplished through subsequent conjugation to a second strain and plating on selective media.

Perspectives

Identification of new genes of interest in metagenomic libraries by functional screening has the advantage that completely new traits can be discovered, even if the sequences of the corresponding genes display no similarity to already known genes. The main disadvantages of this approach are that the protocols are often laborious and screening can consequently in some cases be very costly, even though advances in robotics have reduced this problem. Since DNA can now be sequenced efficiently without prior cloning and to a lower and lower price, it is likely that such an approach will become increasingly adopted in the future. The disadvantage of this method is that functions of interest that cannot be detected by bioinformatics analyses will not be identified. Additionally, most biosynthetic pathways will not be assembled from a metagenomic sequence collection into complete contiguous genomic fragments, nor will these pathways be expressed to produce a secondary metabolite. These disadvantages for a purely sequence-based approach will remain an insurmountable obstacle to discovery of much of the genetic diversity present in natural environments.

The metagenomic surveys of microbial and viral communities that are completed so far suggest that the diversity of environmental communities is exceptional and that more than 60% of the sequences are novel sequences with unknown functions [Ferrer et al., 2009b]. This indicates that functional screening may become quite important for many years in order to map more of the gene sequence space that exists in nature. There is little doubt that both direct sequencing and functional screening of libraries will be used in the years to come, and completely new functions and sequences discovered by functional screening will in the long run contribute to improved bioinformatics predictions. We therefore believe that the long-term tendency is likely to move more in the direction of direct sequencing coupled to bioinformatics analyses, but that functional screening will also be important in the foreseeable future. The use of a combined sequence- and function-based metagenomic approach can overcome some of these biases that are inherent in a metagenomic approach, permitting access to the as-yet-uncultured extant functional diversity of microbial life on Earth.

References

Aakvik T, Degnes KF, Dahlsrud R, Schmidt F, Dam R, Yu L, Völker U, Ellingsen TE, Valla S. 2009. A plasmid RK2-based broad-host-range cloning vector useful for transfer of metagenomic libraries to a variety of bacterial species. FEMS Microbiol. Lett. 296:149-158.

Abulencia CB, Wyborski DL, Garcia JA, Podar M, Chen W, Chang SH, Chang HW, Watson D, Brodie EL, Hazen TC and others. 2006. Environmental whole-genome amplification to access microbial populations in contaminated sediments. Appl. Environ. Microbiol. 72:3291-3301.

Amann RI, Ludwig W, Schleifer KH. 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev 59:143-169.

Angelov A, Mientus M, Liebl S, Liebl W. 2009. A two-host fosmid system for functional screening of (meta)genomic libraries from extreme thermophiles. Syst. Appl. Microbiol. 32:177-185.

Bailly J, Fraissinet-Tachet L, Verner M-C, Debaud J-C, Lemaire M, Wésolowski-Louvel M, Marmeisse R. 2007. Soil eukaryotic functional diversity, a metatranscriptomic approach. ISME J. 1:632-642.

Banik JJ, Brady SF. 2008. Cloning and characterization of new glycopeptide gene clusters found in an environmental DNA megalibrary. Proc. Natl. Acad. Sci. U. S. A. 105:17273-17277.

Bernstein JR, Bulter T, Shen CR, Liao JC. 2007. Directed evolution of ribosomal protein S1 for enhanced translational efficiency of high GC Rhodopseudomonas palustris DNA in *Escherichia coli*. J. Biol. Chem. 282:18929-18936.

Brady SF, Chao CJ, Clardy J. 2004. Long-chain N-acyltyrosine synthases from environmental DNA. Appl. Environ. Microbiol. 70:6865-6870.

Brautaset T, Sekurova ON, Sletta H, Ellingsen TE, Strøm AR, Valla S, Zotchev SB. 2000. Biosynthesis of the polyene antifungal antibiotic nystatin in *Streptomyces noursei* ATCC 11455: analysis of the gene cluster and deduction of the biosynthetic pathway. Chem. Biol.7:395-403.

Brikun IA, Reeves AR, Cernota WH, Luu MB, Weber JM. 2004. The erythromycin biosynthetic gene cluster of *Aeromicrobium erythreum*. J. Ind. Microbiol. Biotechnol.31:335-344.

Chen I-C, Lin W-D, Hsu S-K, Thiruvengadam V, Hsu W-H. 2009. Isolation and characterization of a novel lysine racemase from a soil metagenomic library. Appl. Environ. Microbiol. 75:5161-5166.

Chu X, He H, Guo C, Sun B. 2008. Identification of two novel esterases from a marine metagenomic library derived from South China Sea. Appl. Microbiol. Biotechnol. 80:615-625.

Chung EJ, Lim HK, Kim J-C, Choi GJ, Park EJ, et al. 2008. Forest soil metagenome gene cluster involved in antifungal activity expression in *Escherichia coli*. Appl. Environ. Microbiol. 74:723-730.

Collins J, Hohn B. 1978. Cosmids: a type of plasmid gene-cloning vector that is packageable in vitro in bacteriophage lambda heads. Proc. Natl. Acad. Sci. U. S. A. 75:4242-4246.

Courtois S, Cappellano CM, Ball M, Francou F-X, Normand P, et al. 2003. Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products. Appl. Environ. Microbiol. 69:49-55.

Cowan D, Meyer Q, Stafford W, Muyanga S, Cameron R, Wittwer P. 2005. Metagenomic gene discovery: past, present and future. Trends Biotechnol. 23:321-329.

Craig JW, Chang F-Y, Brady SF. 2009. Natural products from environmental DNA hosted in *Ralstonia metallidurans*. ACS Chem. Biol.4:23-28.

Daniel R. 2005. The metagenomics of soil. Nat. Rev. Microbiol. 3:470-478.

Diaz-Torres ML, McNab R, Spratt DA, Villedieu A, Hunt N, Wilson M, Mullany P. 2003. Novel tetracycline resistance determinant from the oral metagenome. Antimicrob. Agents. Chemother. 47:1430-1432.

Elend C, Schmeisser C, Leggewie C, Babiak P, Carballeira JD, Steele HL, Reymond J-L, Jaeger K-E, Streit WR. 2006. Isolation and biochemical characterization of two novel metagenome-derived esterases. Appl. Environ. Microbiol. 72:3637-3645.

Elsaied H, Stokes HW, Nakamura T, Kitamura K, Fuse H, Maruyama A. 2007. Novel and diverse integron integrase genes and integron-like gene cassettes are prevalent in deep-sea hydrothermal vents. Environ. Microbiol.9:2298-2312.

Entcheva P, Liebl W, Johann A, Hartsch T, Streit WR. 2001. Direct cloning from enrichment cultures, a reliable strategy for isolation of complete operons and genes from microbial consortia. Appl. Environ. Microbiol. 67:89-99.

Ferrer M, Beloqui A, Timmis KN, Golyshin PN. 2009a. Metagenomics for mining new genetic resources of microbial communities. J. Mol. Microbiol. Biotechnol. 16:109-123.

Ferrer M, Beloqui A, Vieites JM, Guazzaroni ME, Berger I, Aharoni A. 2009b. Interplay of metagenomics and in vitro compartmentalization. Microbial Biotechnol. 2:31-39.

Ferrer M, Chernikova TN, Timmis KN, Golyshin PN. 2004. Expression of a temperature-sensitive esterase in a novel chaperone-based *Escherichia coli* strain. Appl. Environ. Microbiol. 70:4499-4504.

Ferrer M, Golyshina OV, Chernikova TN, Khachane AN, Reyes-Duarte D, Santos VAPMD, Strompl C, Elborough K, Jarvis G, Neef A and others. 2005. Novel hydrolase diversity retrieved from a metagenome library of bovine rumen microflora. Environ. Microbiol. 7:1996-2010.

Gabor E, Liebeton K, Niehaus F, Eck J, Lorenz P. 2007. Updating the metagenomics toolbox. Biotechnol. J. 2:201-206.

Gabor EM, Alkema WBL, Janssen DB. 2004a. Quantifying the accessibility of the metagenome by random expression cloning techniques. Environ. Microbiol.6:879-886.

Gabor EM, de Vries EJ, Janssen DB. 2004b. Construction, characterization, and use of small-insert gene banks of DNA isolated from soil and enrichment cultures for the recovery of novel amidases. Environ. Microbiol.6:948-958.

Gabor EM, Vries EJ, Janssen DB. 2003. Efficient recovery of environmental DNA for expression cloning by indirect extraction methods. FEMS Microbiol. Ecol. 44:153-163.

Grant S, Grant WD, Cowan DA, Jones BE, Ma Y, Ventosa A, Heaphy S. 2006. Identification of eukaryotic open reading frames in metagenomic cDNA libraries made from environmental samples. Appl. Environ. Microbiol. 72:135-143.

Hain T, Otten S, von Both U, Chatterjee SS, Technow U, Billion A, Ghai R, Mohamed W, Domann E, Chakraborty T. 2008. Novel bacterial artificial chromosome vector pUvBBAC for use in studies of the functional genomics of *Listeria* spp. Appl. Environ. Microbiol. 74:1892-1901.

Hall RM. 1997. Mobile gene cassettes and integrons: moving antibiotic resistance genes in gram-negative bacteria. Ciba Found. Symp. 207:192-202; discussion 202-5.

Handelsman J. 2005. Sorting out metagenomes. Nat. Biotechnol. 23:38-39.

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:R245-R249.

Heath C, Hu XP, Cary SC, Cowan D. 2009. Identification of a novel alkaliphilic esterase active at low temperatures by screening a metagenomic library from antarctic desert soil. Appl. Environ. Microbiol. 75:4657-4659.

Henne A, Daniel R, Schmitz RA, Gottschalk G. 1999. Construction of environmental DNA libraries in *Escherichia coli* and screening for the presence of genes conferring utilization of 4-hydroxybutyrate. Appl. Environ. Microbiol. 65:3901-3907.

Hrvatin S, Piel Jr. 2007. Rapid isolation of rare clones from highly complex DNA libraries by PCR analysis of liquid gel pools. J. Microbiol. Methods 68:434-436.

Jogler C, Lin W, Meyerdierks A, Kube M, Katzmann E, Flies C, Pan Y, Amann R, Reinhardt R, Schüler D. 2009. Toward cloning of the magnetotactic metagenome: identification of magnetosome island gene clusters in uncultivated magnetotactic bacteria from different aquatic sediments. Appl. Environ. Microbiol. 75:3972-3979.

Kakirde, K.S., J. Wild, R. Godiska, D.A. Mead, A.G. Wiggins, R.M. Goodman, W. Szybalski, and M.R. Liles. 2010. Gram negative shuttle BAC vector for heterologous expression of metagenomic libraries. Gene. In press.

Kazimierczak KA, Scott KP, Kelly D, Aminov RI. 2009. Tetracycline resistome of the organic pig gut. Appl. Environ. Microbiol. 75:1717-1722.

Kim BS, Kim SY, Park J, Park W, Hwang KY, Yoon et al. 2007. Sequence-based screening for self-sufficient P450 monooxygenase from a metagenome library. J. Appl. Microbiol. 102:1392-1400.

Kim K-H, Chang H-W, Nam Y-D, Roh SW, Kim M-S, et al. 2008. Amplification of uncultured single-stranded DNA viruses from rice paddy soil. Appl. Environ. Microbiol. 74:5975-5985.

Kim UJ, Shizuya H, de Jong PJ, Birren B, Simon MI. 1992. Stable propagation of cosmid sized human DNA inserts in an F factor based vector. Nucleic Acids Res. 20:1083-1085.

Knietsch A, Bowien S, Whited G, Gottschalk G, Daniel R. 2003a. Identification and characterization of coenzyme B12-dependent glycerol dehydratase- and diol dehydratase-encoding genes from metagenomic DNA libraries derived from enrichment cultures. Appl. Environ. Microbiol. 69:3048-3060.

Knietsch A, Waschkowitz T, Bowien S, Henne A, Daniel R. 2003b. Construction and screening of metagenomic libraries derived from enrichment cultures: generation of a gene bank for genes conferring alcohol oxidoreductase activity on *Escherichia coli*. Appl. Environ. Microbiol. 69:1408-1416.

Koenig JE, Sharp C, Dlutek M, Curtis B, Joss M, Boucher Y, Doolittle WF. 2009. Integron gene cassettes and degradation of compounds associated with industrial waste: the case of the Sydney tar ponds. PLoS One 4:e5276.

Krsek M, Wellington EM. 1999. Comparison of different methods for the isolation and purification of total community DNA from soil. J. Microbiol. Methods 39:1-16.

Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P. 2008. A bioinformatician's guide to metagenomics. Microbiol. Mol. Biol. Rev. 72:557-78.

Lee S-W, Won K, Lim HK, Kim J-C, Choi GJ, Cho KY. 2004. Screening for novel lipolytic enzymes from uncultured soil microorganisms. Appl. Microbiol. Biotechnol. 65:720-726.

Li Y, Wexler M, Richardson DJ, Bond PL, Johnston AWB. 2005. Screening a wide host-range, waste-water metagenomic library in tryptophan auxotrophs of *Rhizobium leguminosarum* and of *Escherichia coli* reveals different classes of cloned trp genes. Environ. Microbiol.7:1927-1936.

Liles MR, Williamson LL, Rodbumrer J, Torsvik V, Goodman RM, Handelsman J. 2008. Recovery, purification, and cloning of high-molecular-weight DNA from soil microorganisms. Appl. Environ. Microbiol. 74:3302-3305.

Lämmle K, Zipper H, Breuer M, Hauer B, Buta C, Brunner H, Rupp S. 2007. Identification of novel enzymes with different hydrolytic activities by metagenome expression cloning. J. Biotechnol. 127:575-592.

Majerník A, Gottschalk G, Daniel R. 2001. Screening of environmental DNA libraries for the presence of genes conferring Na(+)(Li(+))/H(+) antiporter activity on *Escherichia coli*: characterization of the recovered genes and the corresponding gene products. J. Bacteriol. 183:6645-6653.

Martin-Laurent F, Philippot L, Hallet S, Chaussod R, Germon JC, Soulas G, Catroux G. 2001. DNA extraction from soils: old bias for new microbial diversity analysis methods. Appl. Environ. Microbiol. 67:2354-2359.

Martinez A, Kolvek SJ, Yip CLT, Hopke J, Brown KA, MacNeil IA, Osburne MS. 2004. Genetically modified bacterial strains and novel bacterial artificial chromosome shuttle vectors for constructing environmental libraries and detecting heterologous natural products in multiple expression hosts. Appl. Environ. Microbiol. 70:2452-2463.

Martinez A, Tyson GW, Delong EF. 2009. Widespread known and novel phosphonate utilization pathways in marine bacteria revealed by functional screening and metagenomic analyses. Environ. Microbiol. In press.

Mirete S, de Figueras CG, González-Pastor JE. 2007. Novel nickel resistance genes from the rhizosphere metagenome of plants adapted to acid mine drainage. Appl. Environ. Microbiol. 73:6001-6011.

Mohn WW, Garmendia J, Galvao TC, de Lorenzo Vc. 2006. Surveying biotransformations with Ã la carte genetic traps: translating dehydrochlorination of lindane (gamma-hexachlorocyclohexane) into lacZ-based phenotypes. Environ. Microbiol.8:546-555.

Mori T, Mizuta S, Suenaga H, Miyazaki K. 2008. Metagenomic screening for bleomycin resistance genes. Appl. Environ. Microbiol. 74:6803-6805.

Nishihara K, Kanemori M, Kitagawa M, Yanagi H, Yura T. 1998. Chaperone coexpression plasmids: differential and synergistic roles of DnaK-DnaJ-GrpE and GroEL-GroES in assisting folding of an allergen of Japanese cedar pollen, Cryj2, in *Escherichia coli*. Appl. Environ. Microbiol. 64:1694-1699.

Ogram A, Sayler GS, Barkay T. 1987. The extraction and purification of microbial DNA from sediments. J. Microbiol. Methods 7:57-66.

Ono A, Miyazaki R, Sota M, Ohtsubo Y, Nagata Y, Tsuda M. 2007. Isolation and characterization of naphthalene-catabolic genes and plasmids from oil-contaminated soil by using two cultivation-independent approaches. Appl. Microbiol. Biotechnol. 74:501-510.

Ouyang Y, Dai S, Xie L, Kumar MR, Sun W, Sun H, Tang D, Li X. 2009. Isolation of High Molecular Weight DNA from Marine Sponge Bacteria for BAC Library Construction. Mar. Biotechnol. 12:318-325.

Park S-J, Kang C-H, Chae J-C, Rhee S-K. 2008. Metagenome microarray for screening of fosmid clones containing specific genes. FEMS Microbiol. Lett. 284:28-34.

Rappé MS, Giovannoni SJ. 2003. The uncultured microbial majority. Annu. Rev. Microbiol. 57:369-394.

Rees HC, Grant S, Jones B, Grant WD, Heaphy S. 2003. Detecting cellulase and esterase enzyme activities encoded by novel genes present in environmental DNA libraries. Extremophiles 7:415-421.

Riesenfeld CS, Goodman RM, Handelsman J. 2004. Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. Environ. Microbiol.6:981-989.

Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, et al. 2000. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl. Environ. Microbiol. 66:2541-2547.

Rosano GnL, Ceccarelli EA. 2009. Rare codon content affects the solubility of recombinant proteins in a codon bias-adjusted *Escherichia coli* strain. Microb. Cell Fact. 8:41.

Rowe-Magnus DA. 2009. Integrase-directed recovery of functional genes from genomic libraries. Nucleic Acids Res. 37:e118.

Rowe-Magnus DA, Mazel D. 1999. Resistance gene capture. Curr. Opin. Microbiol. 2:483-488.

Sagova-Mareckova M, Cermak L, Novotna J, Plhackova K, Forstova J, Kopecky J. 2008. Innovative methods for soil DNA purification tested in soils with widely differing characteristics. Appl. Environ. Microbiol. 74:2902-2907.

Sebat JL, Colwell FS, Crawford RL. 2003. Metagenomic profiling: microarray analysis of an environmental genomic library. Appl. Environ. Microbiol. 69:4927-4934.

Shizuya H, Birren B, Kim UJ, Mancino V, Slepak T, Tachiiri Y, Simon M. 1992. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. Proc. Natl. Acad. Sci. U. S. A. 89:8794-8797.

Simon C, Herath J, Rockstroh S, Daniel R. 2009. Rapid identification of genes encoding DNA polymerases by function-based screening of metagenomic libraries derived from glacial ice. Appl. Environ. Microbiol. 75:2964-2968.

Sleator RD, Shortall C, Hill C. 2008. Metagenomics. Lett. Appl. Microbiol. 47:361-366.

Sosio M, Giusino F, Cappellano C, Bossi E, Puglia AM, Donadio S. 2000. Artificial chromosomes for antibiotic-producing actinomycetes. Nat. Biotechnol. 18:343-345.

Steffan RJ, Goksøyr J, Bej AK, Atlas RM. 1988. Recovery of DNA from soils and sediments. Appl. Environ. Microbiol. 54:2908-2915.

Stokes HW, Holmes AJ, Nield BS, Holley MP, Nevalainen KM, Mabbutt BC, Gillings MR. 2001. Gene cassette PCR: sequence-independent recovery of entire genes from environmental DNA. Appl. Environ. Microbiol. 67:5240-5246.

Suenaga H, Ohnuki T, Miyazaki K. 2007. Functional screening of a metagenomic library for genes involved in microbial degradation of aromatic compounds. Environ. Microbiol.9:2289-2297.

Tirawongsaroj P, Sriprang R, Harnpicharnchai P, Thongaram T, Champreda V, Tanapongpipat S, Pootanakit K, Eurwilaichitr L. 2008. Novel thermophilic and thermostable lipolytic enzymes from a Thailand hot spring metagenomic library. J. Biotechnol. 133:42-49.

Torsvik V, Goksøyr J, Daae FL. 1990. High diversity in DNA of soil bacteria. Appl. Environ. Microbiol. 56:782-787.

Tsai YL, Olson BH. 1991. Rapid method for direct extraction of DNA from soil and sediments. Appl. Environ. Microbiol. 57:1070-1074.

Tyson GW, Banfield JF. 2005. Cultivating the uncultivated: a community genomics perspective. Trends Microbiol. 13:411-415.

Uchiyama T, Abe T, Ikemura T, Watanabe K. 2005. Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. Nat. Biotechnol. 23:88-93.

van Sint Fiet S, van Beilen JB, Witholt B. 2006. Selection of biocatalysts for chemical synthesis. Proc. Natl. Acad. Sci. U. S. A. 103:1693-1698.

Wall JG, Plückthun A. 1995. Effects of overexpressing folding modulators on the in vivo folding of heterologous proteins in *Escherichia coli*. Curr. Opin. Biotechnol. 6:507-516.

Wang C, Meek DJ, Panchal P, Boruvka N, Archibald FS, Driscoll BT, Charles TC. 2006. Isolation of poly-3-hydroxybutyrate metabolism genes from complex microbial communities by phenotypic complementation of bacterial mutants. Appl. Environ. Microbiol. 72:384-391.

Warren RL, Freeman JD, Levesque RC, Smailus DE, Flibotte S, Holt RA. 2008. Transcription of foreign DNA in *Escherichia coli*. Genome Res. 18:1798-1805.

Waschkowitz T, Rockstroh S, Daniel R. 2009. Isolation and characterization of metalloproteases with a novel domain structure by construction and screening of metagenomic libraries. Appl. Environ. Microbiol. 75:2506-2516.

Wexler M, Bond PL, Richardson DJ, Johnston AWB. 2005. A wide host-range metagenomic library from a waste water treatment plant yields a novel alcohol/aldehyde dehydrogenase. Environ. Microbiol.7:1917-1926.

Whitman WB, Coleman DC, Wiebe WJ. 1998. Prokaryotes: the unseen majority. Proc. Natl. Acad. Sci. U. S. A. 95:6578-6583.

Wild J, Hradecna Z, Szybalski W. 2002. Conditionally amplifiable BACs: switching from single-copy to high-copy vectors and genomic clones. Genome Res. 12:1434-1444.

Wilkinson DE, Jeanicke T, Cowan DA. 2002. Efficient molecular cloning of environmental DNA from geothermal sediments. Biotechnol. Let. 24:155-161.

Williamson LL, Borlee BR, Schloss PD, Guan C, Allen HK, Handelsman J. 2005. Intracellular screen to identify metagenomic clones that induce or inhibit a quorum-sensing biosensor. Appl. Environ. Microbiol. 71:6335-6344.

Wu C, Sun B. 2009. Identification of novel esterase from metagenomic library of Yangtze river. J. Microbiol. Biotechnol. 19:187-193.

Yamada K, Terahara T, Kurata S, Yokomaku T, Tsuneda S, Harayama S. 2008. Retrieval of entire genes from environmental DNA by inverse PCR with pre-amplification of target genes using primers containing locked nucleic acids. Environ. Microbiol.10:978-987.

Yokouchi H, Fukuoka Y, Mukoyama D, Calugay R, Takeyama H, Matsunaga T. 2006. Whole-metagenome amplification of a microbial community associated with scleractinian coral by multiple displacement amplification using phi29 polymerase. Environ. Microbiol.8:1155-1163.

Zaehner H, Fiedler FP. 1995. Fifty years of antimicrobials: past perspectives and future trends. In: Hunter PA, Darby GK, Russell NJ, editors. The Need for New Antibiotics: Possible Ways Forward. Fifty-Third Symposium of the Society for General Microbiology. Cambridge, UK: Cambridge University Press. p 67-84.

Zhou J, Bruns MA, Tiedje JM. 1996. DNA recovery from soils of diverse composition. Appl. Environ. Microbiol. 62:316-322.

Table 1. Vectors used (or proposed used) for expression of metagenomic libraries in non-*E. coli* hosts.*

Vector	Insert size	Properties/Host specificities	Reference
Cosmid-/ Fosmid vectors:			
Cosmid pOS700I	Not explicitly stated	Replicates in *E. coli* and can be transformed to and integrated into the chromosome of *S. lividans*	Courtois et al., 2003
Cosmid pLAFR3	Ca 25 kb	Contains a broad-host-range RK2-replicon and can be conjugated to a variety of hosts. A library is in this study transferred to *Rhizobium leguminosarum.*	Wexler et al., 2005 Li et al., 2005
Cosmid pKS13S	Ca 25 kb	Contains a broad-host-range RK2 replicon. Is in this study transferred to *Pseudomonas putida* by electroporation.	Ono et al., 2007
Cosmid pRK7813	Ca 33 kb	Contains a broad-host-range RK2 replicon and can be conjugated to a variety of hosts. Is in this study transferred to *Sinorhizobium meliloti*.	Wang et al., 2006
Fosmid pCT3FK	Ca 40 kb	Replicates in *E. coli* and can be transformed to and integrated into the chromosome of *Thermus thermophilus*.	Angelov et al., 2009
Cosmid pJWC1	Not explicitly stated	Contains a broad-host-range RK2 replicon. Is in this study transferred to *Ralstonia metallidurans* by electroporation.	Craig et al., 2009
Fosmid and BAC pRS44	Fosmid clones: ca 35 kb	Contains both an F-replicon and a broad-host-range RK2 replicon. Can be conjugated to a variety of hosts,	Aakvik et al., 2009

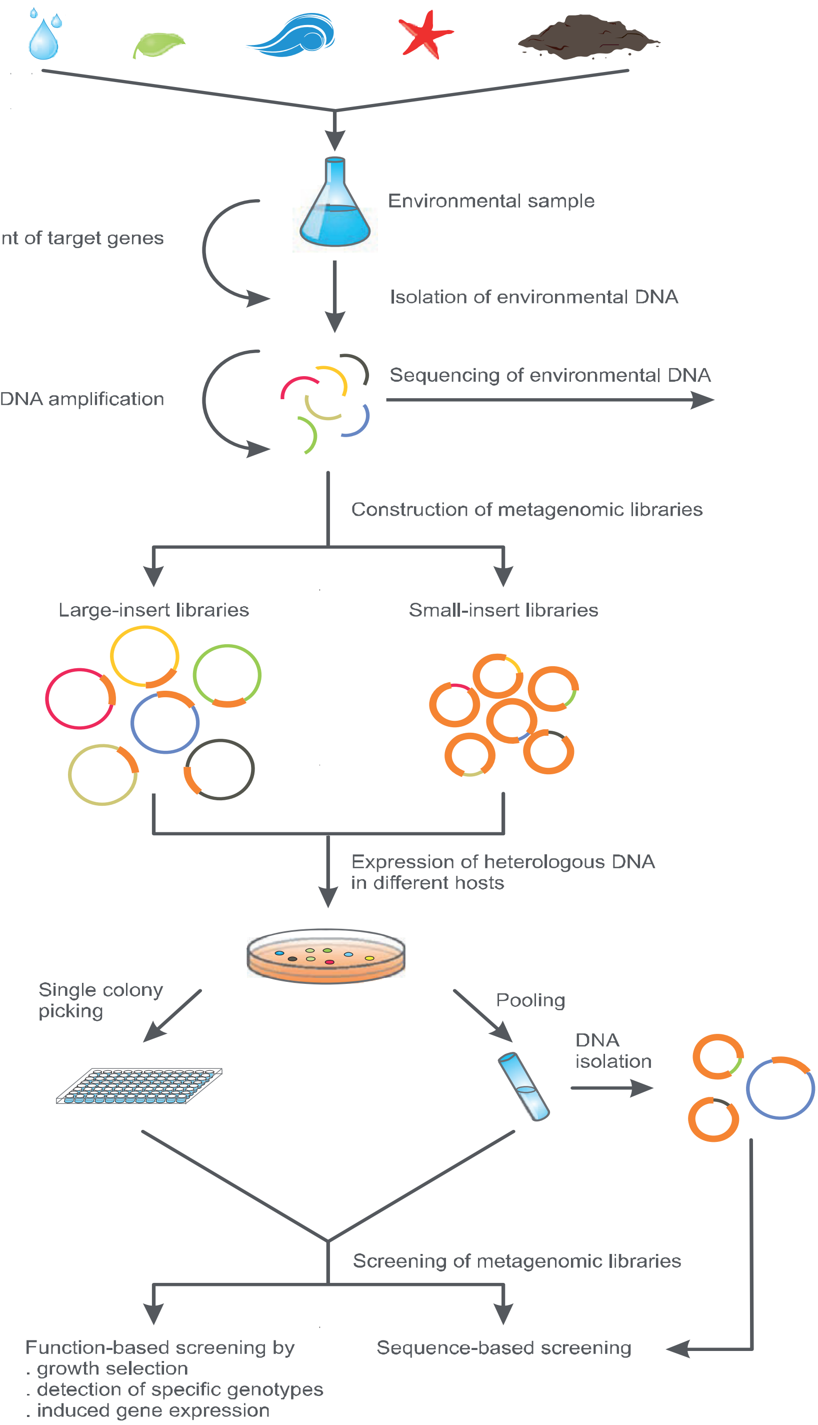
	BAC clones: up to 195 kb	and is in this study transferred to *Pseudomonas fluorescens* and *Xanthomonas campestris.*	
BAC vectors:			
BAC pPAC-S1 ("ESAC")	Up to 140 kb	Replicates in *E. coli* and can be transformed to and integrated into the chromosome of *Streptomyces* hosts. Is in this study transferred to *S. lividans*.	Sosio et al., 2000
BAC pMBD14	Up to 85 kb	Replicates in *E. coli* and can be conjugated to and integrated in the chromosome of *S. lividans* and *Pseudomonas putida*	Martinez et al., 2004
BAC pUvBBAC	Up to 178 kb	Replicates in *E. coli* and can be transformed to *Listeria* spp. and possibly other G+ bacteria. Is in this study transferred to *L. innocua.*	Hain et al., 2008
BAC pBAC-1003	Average: 100 kb	A shuttle vector between *E. coli* and *Streptomyces* sp. (Is in this study not transferred to other hosts.)	Ouyang et al., 2009
BAC pGNS-BAC	> 80 kb	A G- shuttle vector. Contains both an F-replicon and a broad-host-range RK2 replicon.	(Kakirde et al., 2010)

* The classification of the vectors as fosmid/cosmid- or BAC vectors is based on the indicated references.

Figure legend:

Figure 1. Steps involved in metagenomics for functional screening. DNA can be isolated from various sources (water, soil, plants, air, biota, etc.). If required, target genes can be enriched prior to DNA isolation. The isolated DNA can then either be subjected to sequencing or in cases where the quantity of the DNA is low DNA amplification can be performed. Depending on the goals either small- or large-insert libraries can be constructed by the use of different vectors. Once the libraries are established in different hosts, they can either be transferred to microtiter plates or pooled. Finally, these libraries can be subjected either to function- or sequence-based screenings. For further details, see the text.

Figure 1


Environmental sample
Enrichment of target genes
Isolation of environmental DNA
DNA amplification
Sequencing of environmental DNA
Construction of metagenomic libraries
Large-insert libraries
Small-insert libraries
Expression of heterologous DNA
in different hosts
Single colony
picking
Pooling
DNA
isolation
Screening of metagenomic libraries
Function-based screening by
. growth selection
. detection of specific genotypes
. induced gene expression
Sequence-based screening