Trevor C. Charles · Mark R. Liles
Angela Sessitsch *Editors*

Functional Metagenomics: Tools and Applications

Springer

Functional Metagenomics: Tools and Applications

Trevor C. Charles • Mark R. Liles
Angela Sessitsch
Editors

Functional Metagenomics: Tools and Applications

Springer

Editors
Trevor C. Charles
Department of Biology
University of Waterloo
Waterloo, Ontario
Canada

Mark R. Liles
Department of Biological Sciences
Auburn University
Auburn, AL
USA

Angela Sessitsch
Biorescources
AIT Austrian Institute of Technology
Biorescources
Tulln
Austria

ISBN 978-3-319-61508-0 ISBN 978-3-319-61510-3 (eBook)
DOI 10.1007/978-3-319-61510-3

Library of Congress Control Number: 2017955378

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Microbiologists have long understood that most of the microbial world is hidden from view. This has been dramatically emphasized as a result of the application of DNA sequencing to the investigation of microbial communities. The advent of high-throughput DNA sequencing, coupled with accessible public sequence databases, has provided ample fodder for computational analyses of the genomes of uncultivated microbes. It is in this context that we address a major challenge in microbiology—the need for a more complete understanding of gene function that will support models and predictions of cell behavior and community dynamics in particular habitats. Metagenomics, and in particular functional metagenomics, provides a framework within which to address this challenge.

Metagenomics consists of a set of enabling technologies rooted in genomics, microbial genetics, microbial ecology, and bioinformatics, applied to the study of microbial genetic material recovered directly from environmental samples. It provides a deep window into the diversity of life on Earth, which is dominated by microbes. This rapidly emerging field is based on the application of DNA sequencing technology and microbial genetics to the investigation of microbes that to a large extent have not been grown in culture. Microbial communities perform critical services in the environment and are central to processes such as wastewater treatment, bioremediation, food microbiology, and the processes that are critical to the basic understanding of Earth's ecosystems.

Functional metagenomics, which seeks to determine not only what microbes are present but also what they are doing, facilitates the discovery and study of new enzymes or biosynthetic gene clusters without relying on prior cultivation of microbes whose genomes express these enzymes. This is a breakthrough technology whose value cannot be overestimated. It allows access to gene products without having to work with the original microbe that produced the enzyme or metabolite and that may not be able to, or is difficult to, culture. A much better understanding of enzyme function is essential to optimize the processes that occur in microbial communities that are providing essential services. Many enzymes are actually developed as commercial products or used to produce commercial products in

industries such as textile, food, brewing, paper, biofuel, detergent, animal feed, bioremediation, green chemistry, and many more.

It is through this lens that this book was envisioned. A broad set of experimental and computational approaches are being developed to provide traction in the quest toward greater understanding of gene function in microbial communities. Initial chapters provide overview and examples of the application of high-throughput sequencing, bioinformatics tools, and different strategies for production of metagenomic libraries and library screening approaches. This is followed by examples of the application of functional metagenomics to microbial communities from different habitats and ecosystems.

This book is intended as a collection of representative studies and views that will provide the reader with a sense of some of the exciting work currently being done in functional metagenomics. We hope that it contributes to further advances in the field.

Waterloo, ON — Trevor C. Charles
Auburn, AL — Mark R. Liles
Tulln, Austria — Angela Sessitsch

Contents

Chapter 1
Metagenomic Cosmid Libraries Suitable for Functional Screening in Proteobacteria

Jiujun Cheng, Kathy N. Lam, Katja Engel, Michael Hall, Josh D. Neufeld, and Trevor C. Charles

Abstract Functional metagenomics, based on screening/selection of clones from metagenomic libraries, has the potential to make major contributions to our understanding of gene function and the development of biotechnology solutions. However, there are challenges and limitations that must be overcome if that potential is to be realized. These include cloning bias in library construction, host-dependence of gene expression, and library vector host range restrictions. In this chapter, we discuss some of our efforts to improve the quality and availability of metagenomic libraries through the production of a series of metagenomic cosmid libraries from diverse Canadian soils. Although these libraries are suitable for screening in a range of bacteria, they are currently limited to the Proteobacteria. To better capture genes from throughout the diversity of microbial life, it will be desirable to construct and make available metagenomic libraries that are able to support phenotypic screening in correspondingly suitable taxonomic backgrounds. Ongoing work is directed at achieving this important goal.

1.1 Introduction

The depth and breadth of known microbial diversity have been expanded greatly by the application of ever more powerful sequencing technology (Schloss et al. 2016; Hug et al. 2016). Nonetheless, the enormous benefit of accessing data from uncultivated microorganisms is tempered by the acknowledgement that the functions encoded by much of these newly determined DNA sequences cannot be reliably assessed and evaluated. Although the absence of most microbial diversity from pure culture represents a major limitation for gene discovery, functional metagenomics based on phenotypic screening may be the ideal methodological approach for circumventing this limitation.

J. Cheng • K.N. Lam • K. Engel • M. Hall • J.D. Neufeld • T.C. Charles (✉)
Department of Biology, University of Waterloo, Waterloo, ON, Canada, N2L 3G1
e-mail: trevor.charles@uwaterloo.ca

T.C. Charles et al. (eds.), *Functional Metagenomics: Tools and Applications*,
DOI 10.1007/978-3-319-61510-3_1

In general, the use of the term "functional metagenomics" implies a very specific function-based "wet-lab" methodology. Although the term is occasionally co-opted to mean something different, such as sequence-based metagenomics with a focus on gene function (Dinsdale et al. 2008; Roller et al. 2013), or even completely redefined to mean the study of functional members of the microbiota that influence human health (Li et al. 2008), such uses are rare in the scientific literature; the "wet-lab" distinction continues to hold sway. Functional metagenomics, in the traditional sense, involves using DNA that has been isolated from microbial communities to study the functions of proteins and other gene products encoded by that DNA. The process usually consists of extracting DNA directly from environmental samples, cloning that DNA in libraries, introducing those libraries into surrogate hosts where they can be expressed, and selecting or screening for functions of interest. This approach can facilitate the discovery of novel gene products such as enzymes for which DNA sequence is not predictive of function. The information from these analyses can contribute to the annotation of genome and metagenome sequences. In this way, functional metagenomics complements sequence-based metagenomics, similar to the way that molecular genetics of model organisms has provided knowledge of gene function that has been widely applicable in pure culture genomics and systems biology.

1.2 *Escherichia coli* as the Host for Metagenomic Libraries

Traditionally, functional metagenomics has been performed in *Escherichia coli.* This is directly related to the overwhelming dominance of *E. coli* as a model organism and its foundational role as an integral part of the molecular biology toolbox. Most functional metagenomic libraries are constructed in vectors that replicate in *E. coli*. These range from small-insert plasmid libraries, through to medium-insert cosmid and fosmid libraries, and large-insert BAC libraries (Kakirde et al. 2010). Each of these types of libraries has their advantages and disadvantages, but biases in library construction and library maintenance and host dependence of gene expression may have major impacts on the experimental outcomes of functional metagenomics.

Because metagenomic libraries are almost always constructed and maintained in *E. coli* host strains, and this is not likely to change, screening often also occurs in an *E. coli* background. For example, to isolate clones conferring antibiotic resistance, recombinant host cells can be applied directly to selective media-containing antibiotics. This example, although simple, has been useful for exploring the antibiotic resistance gene reservoir harboured in the human gut microbiota (Sommer et al. 2009). However, screening solely in *E. coli* strains may limit success due to possible incompatibilities that prevent expression in a given background. Depending on the target activity, functional screens can exhibit a low hit rate (Uchiyama and Miyazaki 2009), the reasons for which might include barriers at the levels of both transcription and translation. For example, promoters, codon usage, and regulator elements are all host-dependent factors that influence gene expression. Strategies to improve screening

efficiency in *E. coli* have included the introduction of heterologous sigma factors to direct transcription initiation (Gaida et al. 2015), using T7 RNA polymerase (Terrón-González et al. 2013) and employing hybrid ribosomes (Kitahara et al. 2012). Despite these efforts, it will be necessary to continue the development of different screening hosts, especially for complementation of functions that are not available in *E. coli*. Fortunately, this is an area of investigation that has not been neglected.

1.3 Alternate Hosts for Screening Metagenomic Libraries

We have established that screening in surrogate hosts other than *E. coli* may provide additional success from functional screening due to the variation in regulatory and structural elements required for gene expression between the original organisms and *E. coli*. Though it is arguably difficult to quantify, one estimate of how much of the metagenome is accessible by screening in *E. coli* is ~40%, based on analysis of 32 genomes from different bacteria and archaea, counting ORFs with ribosome-binding sites and promoters that are predicted to be recognized in *E. coli* (Gabor et al. 2004). The fraction of "inaccessible" genes depends, of course, on the particular environmental sample DNA and its underlying microbial community composition. Regardless, to address this limitation, metagenomic libraries can be transferred from *E. coli* to other surrogate hosts that may be more suitable for screening. This may be done efficiently using conjugation or, if the recipient species is amenable, transformation or electroporation. The transferred clones must be able to replicate in the recipient host, either autonomously or after integration into the genome. Also, the issue of possible barriers to transcription and translation in *E. coli* is a particularly important methodological limitation.

Alternative expression hosts that have been used in functional metagenomics include *Agrobacterium tumefaciens*, *Caulobacter vibrioides*, *Rhizobium leguminosarum*, *Ralstonia metallidurans*, *Pseudomonas fluorescens*, *Pseudomonas putida*, *Xanthomonas campestris*, *Burkholderia graminis*, *Sinorhizobium meliloti*, *Bacillus subtilis*, *Thermus thermophilus*, and *Streptomyces albus* (Li et al. 2005; Aakvik et al. 2009; Uchiyama and Miyazaki 2009; Craig et al. 2010; Taupp et al. 2011; Schallmey et al. 2011; Liebl et al. 2014; Leis et al. 2015; Cheng and Charles 2016; Cheng et al. 2017; Iqbal et al. 2016). The vectors used for these metagenomic libraries contain single broad host range *oriV*, multiple *oriV* to support replication in *E. coli* as well as the screening hosts, or recombinase-based systems that facilitate integration into the chromosome of the screening host. Not only do these vectors allow screening in different host backgrounds, but they also make it possible to take advantage of sophisticated genetic analysis, in many cases using specific mutants and strain constructs. Despite what may at first glance appear to be a large number of possible screening hosts for the existing vectors and libraries, they actually only represent a very small proportion of phylogenetic diversity. Hence, there is a need for further expansion of the hosts for metagenomic screening to better represent the entirety of microbial diversity.

1.4 Considerations for Metagenomic Library Construction

There are important considerations for undertaking functional metagenomic approaches. First, consideration must be given to choosing an appropriate environment for the desired target genes. For instance, a rumen sample from a grass-fed cow may be ideal for generating a metagenomic library that is enriched with genes encoding enzymes for cellulose degradation (Gong et al. 2012), whereas a sample from a hot spring site would be more suited to isolation of genes encoding thermostable enzymes (Leis et al. 2015). Second, an appropriate vector must be selected for the library backbone. The choice depends on various factors, such as whether a small-insert or large-insert library is desired and, in the former case, whether expression vectors would be advantageous to help drive gene expression in *E. coli* (Kaddurah-Daouk et al. 2011). Third, surrogate host(s) other than *E. coli* may be considered, as indicated above, for either an attempt to increase the hit rate (Tebbe and Vahjen 1993; Ufarté et al. 2015) or for complementation of specific phenotypes (Wang et al. 2013). Finally, other logistics in the screening strategy have to be considered, such as whether to pool clones for screening or to instead keep clones arrayed and carry out individual clone screening. In the latter case, the achievable throughput must be very carefully considered because, depending on the particular screen, clone-by-clone screening may not be a feasible strategy, although the design of automated microfluidic screening strategies is an exciting area of development (Colin et al. 2015).

Although there are limitations and biases with existing approaches for constructing metagenomic libraries (Ekkers et al. 2012), as there are with all methods, functional metagenomics remains a powerful experimental strategy that has the potential to help improve our understanding of the mechanisms that underlie biological phenomena as well as aid in the functional annotation of the ever increasing number of metagenomes.

1.5 Enrichment of Desired Sequences

Not only can functional selections discover novel gene products, they can also greatly reduce the sheer quantity of genetic material to be sequenced. For example, a high-throughput functional metagenomic approach was used to find enzymes in the human gut involved in dietary fibre catabolism, reducing the amount of metagenomic DNA to be sequenced from 5.4×10^9 bp to 8.4×10^5 bp, a reduction of almost four orders of magnitude, simply by selecting for the growth of library clones on different polysaccharides (Tasse et al. 2010). Using this approach, the authors identified 73 carbohydrate-active enzymes, corresponding to a five-fold enrichment in the target-gene identification over shotgun metagenomic sequencing. If enrichment can be performed prior to sequencing, a great deal of time and resources can be saved, not to mention the value of having experimental data regarding function.

Enrichment can also be effective prior to library construction. This can involve subjecting environmental samples to traditional enrichment culture before DNA is extracted. More recently, it has become possible to enrich for DNA from members

of the community that are performing certain metabolic tasks using techniques based on stable-isotope probing (Neufeld et al. 2007). For example, an enriched metagenomic library constructed from multiple displacement amplification products of DNA pooled from ^{13}C-cellulose incubations of soil microbial communities was used to isolate clones expressing cellulose-degrading ability at a higher rate than previously reported for non-enriched libraries (Verastegui et al. 2014).

1.6 Library Construction

The first step of functional genomics is the construction of metagenomic libraries. Many investigators prefer cosmid- or fosmid-based libraries because of their high cloning efficiency and large insert size. DNA fragments are first extracted from the environmental sample of choice, such as soil, faeces, or water. These fragments are typically enriched for those of high molecular weight by size selection using pulsed-field gel electrophoresis (PFGE). Following end repair and ligation to a linearized, blunt-ended cos-based vector, λ phage heads are used to package the resulting ligation mixture through recognition of the vector *cos* site. Transduction of *E. coli* with this packaged ligation mixture results in libraries that typically contain inserts of 25–40 kb, depending on the size of the vector backbone. Among the many advantages of using cos-based vectors and phage transduction to construct clone libraries, two important considerations are the high efficiency of transduction and the reduced likelihood of insert concatemers.

The degree to which the content of metagenomic libraries is an accurate reflection of the content of the source samples is often overlooked. When this is evaluated, biases are often observed. It was first hypothesized that these biases were linked to DNA GC content perhaps due to uneven DNA fragmentation (Temperton et al. 2009; Ghai et al. 2010; Danhorn et al. 2012). However, we recently demonstrated that fragmentation was not a cause of bias for a human gut metagenomic library and that the dominant reason for library cloning bias may be strong spurious transcription from σ^{70}-like sequences, which causes vector instability in *E. coli* (Lam and Charles 2015). Incorporation of transcriptional termination sequences adjacent to the fragment insertion site on the vector may reduce this source of cloning bias.

1.7 An Example of a Collection of Broad Host Range Metagenomic Libraries

In an effort to produce a functional metagenomics resource that could be freely shared with other researchers, we developed the concept of open resource metagenomics (Neufeld et al. 2011). Here, we describe the development of a collection of metagenomic libraries to be made available under this aegis. We collected representative Canadian soils encompassing vast taxonomic diversity and used these

samples to construct metagenomic cosmid libraries. We present these publicly available libraries, in addition to the key methodology used for their construction, as part of the Canadian MetaMicroBiome Library (CM^2BL) project (http://cm2bl.org).

We initiated CM^2BL with the goal of establishing a permanent functional metagenomics resource. Surface soil samples (0–10 cm depth) were collected from 14 locations across Canada spanning multiple biomes and ecozones, including Arctic tundra, oil sands, forest, peatlands, agriculture soils, and municipal compost (Table 1.1). Each sample was collected at three sites that were approximately 5 m apart, combined and sent by courier on ice, or hand delivered, to University of Waterloo. Physical and chemical measurements of the soils were performed at the University of Guelph Agricultural and Food Laboratory (Guelph, Ontario, Canada).

Table 1.1 Canadian MetaMicroBiome soil samples and metagenomic libraries (for details see http://www.cm2bl.org/samples.html)

		Soil characteristics				Metagenomic library		
Sample ID	Habitat	Bulk density	Total carbon (% dry)	Total nitrogen (% dry)	pH	Number of clones	Insert size (kb)	Coverage (×)
1AT	Arctic tundra	0.21	46.9	1.42	3.9	178,100	27.1	1026
2ATN	Arctic tundra	1.05	3.7	0.25	6.7	62,260	31.1	412
4TS	Oil sand 1	1.23	2.1	0.11	7.6	73,000	37.4	581
5BF	Boreal coniferous forest	1.16	1.1	0.08	4.6	56,370	29.7	356
6TD	Temperate deciduous forest	1.10	3.6	0.26	6.4	2,306,580	40.2	19,728
7TR	Temperate rainforest	0.62	10.8	0.35	4.9	68,200	33.7	469
8NP	Northern peatlands	0.38	27.2	1.22	5.5	NA	NA	NA
9WLM	Wetland soil	0.26	43.3	2.21	5.0	64,470	19.7	270
10AS	Agricultural—Soybean	1.10	2.4	0.22	7.6	760,000	37.5	6064
11AW[a]	Agricultural—Wheat	1.10	1.9	0.19	7.4	8,806,400	41.2	77,196
12AC[a]	Agricultural—Corn	1.67	NA	NA	7.8	79,060	33.4	561
13CO	Compost	0.86	11.7	0.92	8.0	42,000	34.2	305
19TS	Oil sand 2	1.12	2.8	0.07	6.0	149,880	33.8	1078
20CG	Community garden	0.87	10.2	0.63	7.6	118,300	36.9	929

The coverage of bacterial genomes was calculated based on the average size of 4.7 Mb microbial genome (Raes et al. 2007)

NA not available

[a]Previously published (Cheng et al. 2014)

The Arctic tundra soil sample (1AT) had the highest content of carbon and nitrogen but the lowest pH of 3.9. The soils from boreal coniferous forest (5BF), temperate rain forest (7TR), northern peatlands (8NP), wetland (9WLM), and oil sands 2 (19TS) were also acidic (pH < 6.0). Compost soil (13CO) had the highest pH of 8.0.

Prior to library construction, metagenomic DNA from the soil samples (http://www.cm2bl.org/samples.html) was isolated using a method described previously (Cheng et al. 2014). Co-extracted humic acids were removed using synchronous coefficient of drag alteration electrophoresis (Engel et al. 2012) or PFGE (Cheng et al. 2014). To explore the diversity of microbes in CM^2BL soil samples, the V3 regions of 16S rRNA genes were sequenced as previously described (Bartram et al. 2011), and the 16S rRNA reads were deposited in the Sequence Read Archive of European Bioinformatics Institute (PRJEB9449).

Taxonomic affiliations of 16S rRNA sequences were assigned through an AXIOME2 (https://github.com/neufeld/AXIOME2) pipeline, including paired-end assembly with PANDAseq v2.5 with a quality threshold of 0.9 (Masella et al. 2012), and sequence clustering at 97% sequence identity with UPARSE using USEARCH v7.0.1090 (Edgar 2013). Taxonomic classifications were predicted by RDP v2.2 (Wang et al. 2007) using the Greengenes 13_8 reference set (McDonald et al. 2012). The OTU table was generated by QIIME 1.8.0 (Caporaso et al. 2010), and taxonomy plots were prepared with ggplot2 (Wickham 2016). The 16S rRNA gene sequence analysis confirmed considerable genetic diversity of the microbial communities, dominated by *Proteobacteria*, *Acidobacteria*, *Actinobacteria*, and *Bacteroides* (Fig. 1.1).

In order to perform phenotypic screening in a broad range of surrogate hosts, we constructed metagenomic libraries based on the low-copy number IncP Gateway entry cosmid pJC8 (Cheng et al. 2014). Depending on the size distribution of isolated soil DNA, fragment ranges of 25–50 kb or 30–75 kb were excised from PFGE and recovered by electroelution. These fragments were not generated by restriction enzyme digestion to circumvent bias that might arise from sequence-dependent digestion. Following end repair, purified DNA was ligated to cosmid pJC8 that was cut with Eco72I and dephosphorylated. The ligated product was packaged in vitro with Gigapack III XL packaging extracts (Agilent Technologies) and then transduced into *E. coli* HB101. Recombinant cosmid clones were selected on LB agar plates with tetracycline (15 μg/ml), pooled, then saved in 1-ml aliquots at −70 °C in a final concentration of 7% DMSO. *E. coli* library clones were selected randomly for analysis of cosmid DNA. The average sizes of cloned metagenomic DNA and coverage of bacterial genomes were calculated based on average insert sizes of HindIII-EcoRI-BamHI or EcoRI-HindIII fragments and the total number of recombinant library clones.

A total of thirteen metagenomic cosmid libraries were constructed and maintained in *E. coli* HB101 (Table 1.1), including two libraries reported previously (Cheng et al. 2014). Clones recovered for each library ranged from 4.2×10^3 (compost, 13CO) to 2.9×10^6 (agricultural wheat soil, 11AW). Analysis of randomly selected cosmids with restriction enzyme digestion indicated that the average sizes of cloned metagenomic DNA were 20 kb (wetland soil, 9WLM) to 40 kb (temperate deciduous forest, 6TD) (Table 1.1). The generated soil DNA libraries contained

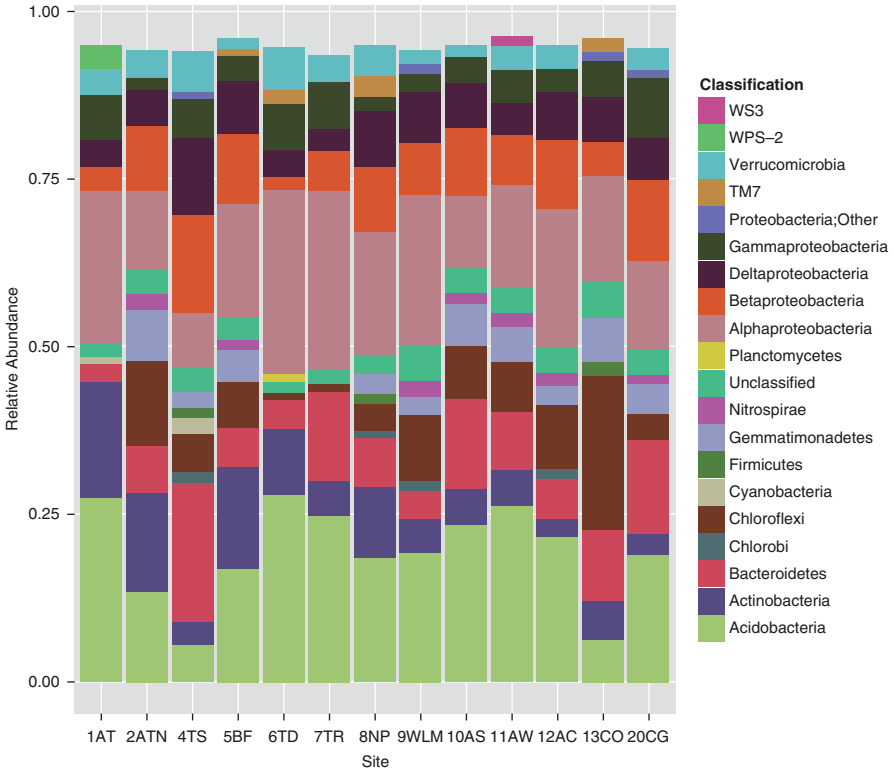

Fig. 1.1 Bacterial community composition of CM^{2}BL samples. Taxonomic profiles of soil samples were analysed based on the sequences of V3 regions of bacterial 16S rRNA genes that were PCR amplified from metagenomic DNA. Any phyla (or class, for *Proteobacteria*) that did not have at least 1% relative abundance in that sample is not shown

1.3–92.7 Gb of metagenomic DNA, which represents approximately 305–19,728 bacterial genomes, assuming an average genome size of 4.7 Mb in a soil community bacteria (Raes et al. 2007). Detailed information describing each metagenomic library and its availability can be accessed through the CM^{2}BL website (http://cm2bl.org/samples.html). This resource thus represents a collection of high-quality metagenomic libraries that are freely available for functional metagenomics in a wide range of Proteobacteria.

1.8 Concluding Statements

The future of functional metagenomics will likely see the development of a greater variety of alternative hosts for functional screening, which will not only lead to an increase in the aggregate hit rates of functional screens but also make available a

broader range of phenotypes for functional complementation. We encourage efforts geared to advance the development of surrogate hosts that better represent the whole of microbial diversity and continue to expand the construction of metagenomic libraries that are suitable for screening in these hosts. This will be necessary if functional metagenomics is to continue its contributions to knowledge of microbial gene function.

Acknowledgements We wish to thank Drs. Kenneth J. Reimer, Paul Grogan, Richard A. Frank, Sylvie A. Quideau, Richard S. Winder, Roland I. Hall, Tim R. Moore, Kari E. Dunfield, and Clark Reichert for collecting the soil samples. This work was partially supported by a Strategic Projects grant to TCC and Discovery Grants to JDN and TCC, both from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Aakvik T, Degnes KF, Dahlsrud R et al (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. doi:10.1111/j.1574-6968.2009.01639.x

Bartram AK, Lynch MDJ, Stearns JC et al (2011) Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads. Appl Environ Microbiol 77:3846–3852. doi:10.1128/AEM.02772-10

Caporaso JG, Kuczynski J, Stombaugh J et al (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7:335–336. doi:10.1038/nmeth.f.303

Cheng J, Romantsov T, Engel K, Doxey AC, Rose DR, Neufeld JD, Charles TC (2017) Functional metagenomics reveals novel ß-galactosidases not predictable from gene sequences. PLOS ONE 12(3):e0172545

Cheng J, Charles TC (2016) Novel polyhydroxyalkanoate copolymers produced in *Pseudomonas putida* by metagenomic polyhydroxyalkanoate synthases. Appl Microbiol Biotechnol 100(17):7611–7627. doi:10.1007/s00253-016-7666-6

Cheng J, Pinnell L, Engel K et al (2014) Versatile broad-host-range cosmids for construction of high quality metagenomic libraries. J Microbiol Methods 99:27–34. doi:10.1016/j.mimet.2014.01.015

Colin P-Y, Kintses B, Gielen F et al (2015) Ultrahigh-throughput discovery of promiscuous enzymes by picodroplet functional metagenomics. Nat Commun 6:10008. doi:10.1038/ncomms10008

Craig JW, Chang F-Y, Kim JH et al (2010) Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse proteobacteria. Appl Environ Microbiol 76:1633–1641. doi:10.1128/AEM.02169-09

Danhorn T, Young CR, DeLong EF (2012) Comparison of large-insert, small-insert and pyrosequencing libraries for metagenomic analysis. ISME J 6:2056–2066. doi:10.1038/ismej.2012.35

Dinsdale EA, Edwards RA, Hall D et al (2008) Functional metagenomic profiling of nine biomes. Nature 452:629–632. doi:10.1038/nature06810

Edgar RC (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods 10:996–998. doi:10.1038/nmeth.2604

Ekkers DM, Cretoiu MS, Kielak AM, Elsas JDV (2012) The great screen anomaly—a new frontier in product discovery through functional metagenomics. Appl Microbiol Biotechnol 93:1005–1020. doi:10.1007/s00253-011-3804-3

Engel K, Pinnell L, Cheng J et al (2012) Nonlinear electrophoresis for purification of soil DNA for metagenomics. J Microbiol Methods 88:35–40. doi:10.1016/j.mimet.2011.10.007

Gabor EM, Alkema WBL, Janssen DB (2004) Quantifying the accessibility of the metagenome by random expression cloning techniques. Environ Microbiol 6:879–886. doi:10.1111/j.1462-2920.2004.00640.x

Gaida SM, Sandoval NR, Nicolaou SA et al (2015) Expression of heterologous sigma factors enables functional screening of metagenomic and heterologous genomic libraries. Nat Commun 6:7045. doi:10.1038/ncomms8045

Ghai R, Martin-Cuadrado A-B, Molto AG et al (2010) Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing. ISME J 4:1154–1166. doi:10.1038/ismej.2010.44

Gong X, Gruninger RJ, Qi M et al (2012) Cloning and identification of novel hydrolase genes from a dairy cow rumen metagenomic library and characterization of a cellulase gene. BMC Res Notes 5:566. doi:10.1186/1756-0500-5-566

Hug LA, Baker BJ, Anantharaman K et al (2016) A new view of the tree of life. Nat Microbiol 1:16048. doi:10.1038/nmicrobiol.2016.48

Iqbal HA, Low-Beinart L, Obiajulu JU, Brady SF (2016) Natural product discovery through improved functional metagenomics in *Streptomyces*. J Am Chem Soc 138(30):9341–9344. doi:10.1021/jacs.6b02921

Kaddurah-Daouk R, Baillie RA, Zhu H et al (2011) Enteric microbiome metabolites correlate with response to simvastatin treatment. PLoS One 6:e25482–e25410. doi:10.1371/journal.pone.0025482

Kakirde KS, Parsley LC, Liles MR (2010) Size does matter: application-driven approaches for soil metagenomics. Soil Biol Biochem 42:1911–1923. doi:10.1016/j.soilbio.2010.07.021

Kitahara K, Yasutake Y, Miyazaki K (2012) Mutational robustness of 16S ribosomal RNA, shown by experimental horizontal gene transfer in *Escherichia coli*. Proc Natl Acad Sci U S A 109:19220–19225. doi:10.1073/pnas.1213609109

Lam KN, Charles TC (2015) Strong spurious transcription likely contributes to DNA insert bias in typical metagenomic clone libraries. Microbiome 3:22. doi:10.1186/s40168-015-0086-5

Leis B, Angelov A, Mientus M et al (2015) Identification of novel esterase-active enzymes from hot environments by use of the host bacterium *Thermus thermophilus*. Front Microbiol 6:275. doi:10.3389/fmicb.2015.00275

Li Y, Wexler M, Richardson DJ et al (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. doi:10.1111/j.1462-2920.2005.00853.x

Li M, Wang B, Zhang M et al (2008) Symbiotic gut microbes modulate human metabolic phenotypes. Proc Natl Acad Sci U S A 105:2117–2122. doi:10.1073/pnas.0712038105

Liebl W, Angelov A, Juergensen J et al (2014) Alternative hosts for functional (meta)genome analysis. Appl Microbiol Biotechnol 98:8099–8109. doi:10.1007/s00253-014-5961-7

Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD (2012) PANDAseq: PAired-enD Assembler for Illumina sequences. BMC Bioinformatics 13:31

McDonald D, Price MN, Goodrich J et al (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J 6:610–618. doi:10.1038/ismej.2011.139

Neufeld JD, Vohra J, Dumont MG et al (2007) DNA stable-isotope probing. Nat Protoc 2:860–866. doi:10.1038/nprot.2007.109

Neufeld JD, Engel K, Cheng J et al (2011) Open resource metagenomics: a model for sharing metagenomic libraries. Stand Genomic Sci 5:203–210. doi:10.4056/sigs.1974654

Raes J, Korbel JO, Lercher MJ et al (2007) Prediction of effective genome size in metagenomic samples. Genome Biol 8:R10. doi:10.1186/gb-2007-8-1-r10

Roller M, Lucić V, Nagy I et al (2013) Environmental shaping of codon usage and functional adaptation across microbial communities. Nucleic Acids Res 41:8842–8852. doi:10.1093/nar/gkt673

Schallmey M, Ly A, Wang C et al (2011) Harvesting of novel polyhydroxyalkanaote (PHA) synthase encoding genes from a soil metagenome library using phenotypic screening. FEMS Microbiol Lett 321:150–156. doi:10.1111/j.1574-6968.2011.02324.x

Schloss PD, Girard RA, Martin T et al (2016) Status of the Archaeal and Bacterial Census: an update. mBio 7:e00201–e00216. doi:10.1128/mBio.00201-16

Sommer M, Dantas G, Church GM (2009) Functional characterization of the antibiotic resistance reservoir in the human microflora. Science 325:1128–1131. doi:10.1126/science.1176950

Tasse L, Bercovici J, Pizzut-Serin S et al (2010) Functional metagenomics to mine the human gut microbiome for dietary fiber catabolic enzymes. Genome Res 20:1605–1612. doi:10.1101/gr.108332.110

Taupp M, Mewis K, Hallam SJ (2011) The art and design of functional metagenomic screens. Curr Opin Biotechnol 22:465–472. doi:10.1016/j.copbio.2011.02.010

Tebbe CC, Vahjen W (1993) Interference of humic acids and DNA extracted directly from soil in detection and transformation of recombinant DNA from bacteria and a yeast. Appl Environ Microbiol 59:2657–2665

Temperton B, Field D, Oliver A et al (2009) Bias in assessments of marine microbial biodiversity in fosmid libraries as evaluated by pyrosequencing. ISME J 3:792–796. doi:10.1038/ismej.2009.32

Terrón-González L, Medina C, Limón-Mortés MC, Santero E (2013) Heterologous viral expression systems in fosmid vectors increase the functional analysis potential of metagenomic libraries. Sci Rep 3:1107. doi:10.1038/srep01107

Uchiyama T, Miyazaki K (2009) Functional metagenomics for enzyme discovery: challenges to efficient screening. Curr Opin Biotechnol 20:616–622. doi:10.1016/j.copbio.2009.09.010

Ufarté L, Potocki-Veronese G, Laville É (2015) Discovery of new protein families and functions: new challenges in functional metagenomics for biotechnologies and microbial ecology. Front Microbiol 6:563. doi:10.3389/fmicb.2015.00563

Verastegui Y, Cheng J, Engel K et al (2014) Multisubstrate isotope labeling and metagenomic analysis of active soil bacterial communities. mBio 5:e01157–14. doi:10.1128/mBio.01157-14

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73:5261–5267. doi:10.1128/AEM.00062-07

Wang L, Hatem A, Catalyurek UV et al (2013) Metagenomic insights into the carbohydrate-active enzymes carried by the microorganisms adhering to solid digesta in the rumen of cows. PLoS One 8:e78507. doi:10.1371/journal.pone.0078507

Wickham H (2016) ggplot2: Elegant graphics for data analysis. Springer-Verlag, New York

Chapter 2
Expression Platforms for Functional Metagenomics: Emerging Technology Options Beyond *Escherichia coli*

Anna Lewin, Rahmi Lale, and Alexander Wentzel

Abstract *Escherichia coli* is the prime workhorse for various metagenomic applications due to the multitude of efficient tools available for genetic manipulation and controlled heterologous gene expression. However, metagenome-based bioprospecting efforts continuously target a wider spectrum of ecological niches in order to harvest new enzymes and bioactive compounds for industrial and medical applications from the enormous pool of natural microbial diversity. Consequently, the development of robust and flexible screening platforms that allow functional evaluation of an expanded fraction of the highly diverse metagenomic information is widely addressed in Functional Metagenomics research. The heterologous recognition of transcriptional regulators and promotors, diverse codon usages among environmental microorganisms, and sufficient supply of precursors for secondary metabolite formation are major challenges that are addressed by an increasing spectrum of alternative expression and host systems. This includes optimized broad host-range transfer and expression vectors, screening hosts for improved gene expression and metabolite formation, as well as cell-free expression systems to cover proteins that due to toxicity are inaccessible by in vivo screening methods. In this chapter, we provide a current overview of the state of the art of selected expression systems and host organisms useful for functional metagenome screening for new enzymes and bioactive metabolites, as emerging options beyond what is currently available in and for *E. coli*.

A. Lewin • A. Wentzel (✉)
Department of Biotechnology and Nanomedicine, SINTEF Materials and Chemistry, Trondheim, Norway
e-mail: Alexander.Wentzel@sintef.no

R. Lale
Department of Biotechnology and Food Science, PhotoSynLab, Norwegian University of Science and Technology, Trondheim, Norway

T.C. Charles et al. (eds.), *Functional Metagenomics: Tools and Applications*,
DOI 10.1007/978-3-319-61510-3_2

2.1 Introduction

Metagenomics has since its introduction in the late 1990s (Handelsman et al. 1998) proven to be a powerful tool for describing microbial communities and their metabolic potentials irrespectively of cultivability. Over the years, both sequence- and function-based screening approaches have led to the discovery of numerous new enzymes and metabolites fulfilling various academic and industrial needs (Ferrer et al. 2015; Fernandez-Arrojo et al. 2010; Novakova and Farkasovsky 2013). The pipeline for Functional Metagenomics spans from sampling, isolation of high-quality environmental DNA (eDNA), and its cloning (including vector design) to metagenomic library construction (including host transformation and transfer), heterologous gene expression, and production of functional molecules in amounts sufficient for detection in high throughput screening (Fig. 2.1). The function-based screening route of metagenome-based bioprospecting therewith complements the sequence-based route, in which eDNA is sequenced using next-generation sequencing methods and resulting sequence datasets mined bioinformatically for genes of interest (Lewin et al. 2013).

Irrespective of the chosen screening route, successful bioprospecting of a metagenomic library starts with the isolation of the eDNA. Its quality and quantity are of major importance for the achievable number of clones of the constructed library and consequently the representation of biodiversity in an environmental sample (Zhou et al. 1996). In order to capture as much of the biodiversity as possible, the applied DNA isolation procedures need to be highly effective in sampling from the diverse microorganisms inhabiting the selected environment (Kakirde et al. 2010). In addition, isolated DNA needs to have a high degree of purity and be free of contaminating substances, such as humic acids that are often present in soil and hamper efficient library construction (Tebbe and Vahjen 1993). Several studies document eDNA isolation procedures that resulted in contamination-free high molecular weight (HMW) DNA (Zhou et al. 1996; Brady 2007; Liles et al. 2008; Pel et al. 2009; Cheng et al. 2014). Contaminating compounds co-isolated with the eDNA can also be successfully removed by gel electrophoretic methods, including conventional (Craig et al. 2010), pulse-field (Cheng et al. 2014), or nonlinear electrophoresis (Pel et al. 2009), followed by size selection of the random fragmented DNA, prior to cloning.

New and improved enzyme discovery is currently the largest field of application for Functional Metagenomics tools. Aside from the catalytic function itself, beneficial properties like robustness under harsh conditions or high activity at low temperatures are often required in industrial applications. Consequently, dependent on the aims of a bioprospecting approach, different environments might serve as eDNA sources (Taupp et al. 2011). The microbial habitat to be sampled usually reflects the desired properties, i.e., subjecting a metagenomic library originating from a thermal vent or a hot deep subsurface oil reservoir to thermostable enzyme screening is likely to have a higher success rate compared to subjecting a

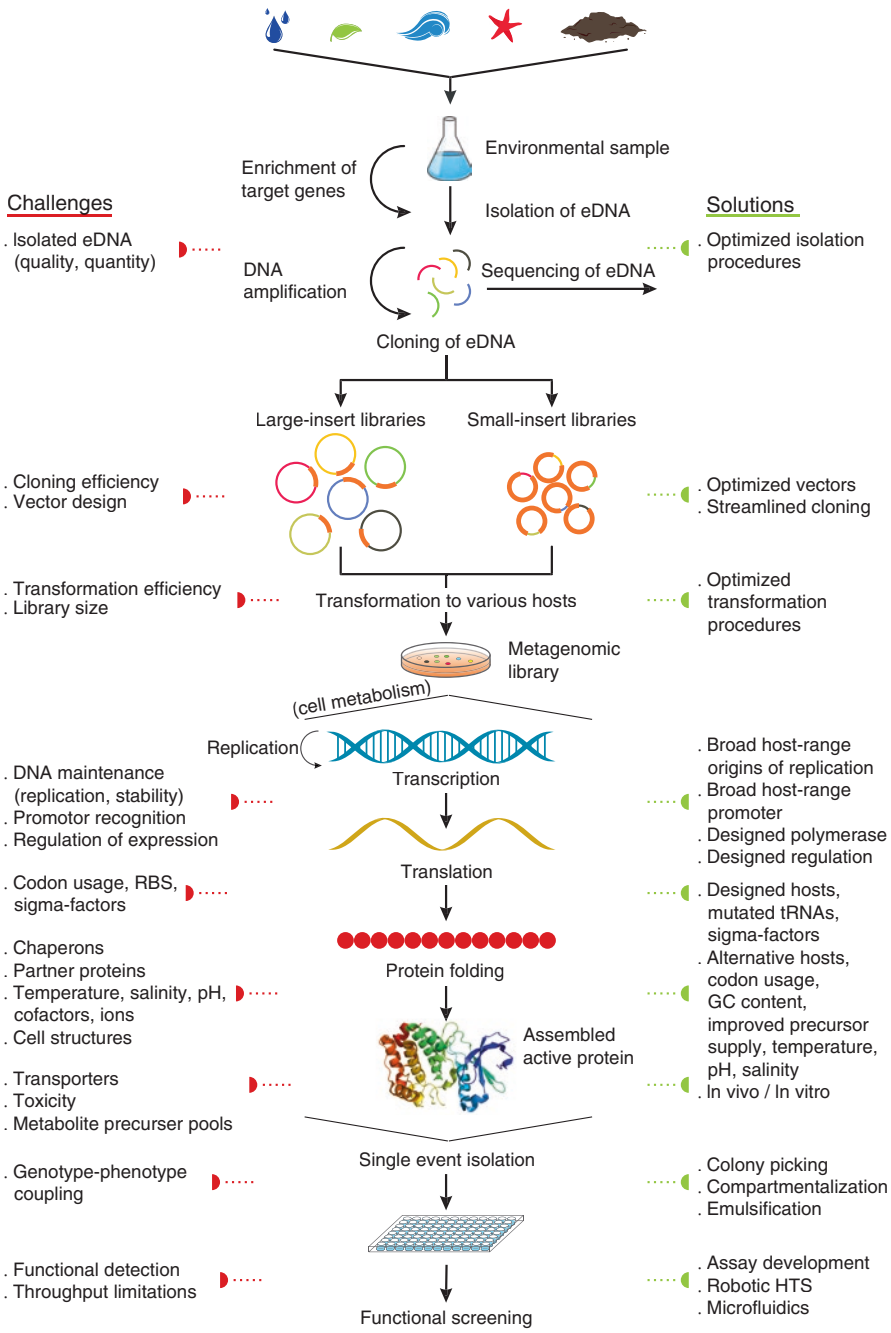

Fig. 2.1 Graphical representation of the Functional Metagenomics biodiscovery pipeline with its key challenges and potential solutions

glacier or permafrost soil-originating library to the same screening. In many examples such directed metagenomics sampling strategies aiming to increase probability of finding the desired properties have proven successful (Vester et al. 2015; Taupp et al. 2011). Selected examples are, among many others, a cold-adapted esterase enzyme from Antarctic desert soil (Hu et al. 2012), hydrolytic enzymes from cow rumen metagenome (Ferrer et al. 2007), and thermostable lipolytic enzymes from water, sediment, and biofilm samples from the Azores, Portugal (Leis et al. 2015a). However, due to often lower microbial density within some, particularly extreme environments, sufficient DNA yields may not be readily obtainable (Kennedy et al. 2008; Vester et al. 2015; Kotlar et al. 2011). In such cases, isolated metagenomic DNA can be subjected to isothermal amplification (like Phi29 whole genome amplification, WGA) in order to increase DNA yields prior to cloning (Rodrigue et al. 2009; Zhang et al. 2006). However, the challenge of this technology with respect to the formation of amplification artifacts, like chimeras, duplications, and inversions, needs to be considered. Therefore it is well suited for small-insert libraries for the purpose of enzyme discovery, but less suitable for large-insert library cloning where intact biosynthetic gene clusters are targeted.

Following sampling and successful isolation, the eDNA is usually either sequenced directly or cloned in suitable vectors for functional screening approaches (Sect. 2.2). The choice of the vector usually depends on the envisioned eDNA insert sizes, as well as the screening targets and methodology. However, for successful expression of genetic information contained in metagenomic DNA libraries, several additional factors need to be taken into account (Fig. 2.1). Suitable vector systems need to carry host-compatible selection markers, replicate stably and autonomously (ideally in combination with the possibility to control the copy number), may contain functional gene regulatory elements like inducible promotors for high level expression, and preferably enable vector transfer to other host organisms. Suitable host organisms in turn need to provide functionality of the vector elements involved in the production of functional products and allow efficient transcription and translation (Sect. 2.3). In addition, proper folding, possible cofactor supply, sufficient precursor availability for metabolite product formation, as well as means for nontoxic product localization, like secretion mechanisms, are needed. In order to meet the different demands for functional expression, such as codon usage, different assay temperatures, precursor requirements, etc. (Lam and Charles 2015; Uchiyama and Miyazaki 2009), different approaches can be applied in order to maximize the probability of successful expression (Fig. 2.1). *E. coli* systems designed and optimized for this purpose have so far been most widely used and extensively covered elsewhere (Guazzaroni et al. 2015). The scope of this chapter is therefore to summarize developments of various hosts and heterologous expression systems for functional metagenome screening beyond the common systems available for *E. coli* only.

2.2 Cloning and Expression Vectors for Environmental DNA

Selection of a suitable vector system for random metagenomic library construction will largely be guided by (1) the expected DNA size encoding the targeted compound of interest, (2) the envisioned subsequent screening approach involving one or more expression hosts, (3) the desired design of the library to be established, as well as in some occasions (4) the quantity of DNA available (Sect. 2.2.1). For approaches to identify new enzymes, small-insert libraries with eDNA sizes of 5–10 kb will in most cases be sufficient to obtain a sufficiently large number of complete gene sequences. Isolation of DNA for small-insert libraries is normally straightforward, since DNA shearing is not a major concern. However, it needs to be considered that a library with an average insert size of 10 kb will require 3–20 times more clones compared to a library with inserts of 30–40 kb to cover the same amount of genetic potential (Sabree et al. 2009). Hence comparably larger amounts of DNA are needed. To identify encoded functions that rely on single genes or small gene loci (e.g., enzyme function or genetic determinants of antibiotic resistance (Riesenfeld et al. 2004)), small-insert libraries are normally sufficient (Kakirde et al. 2010; Sabree et al. 2009). However, in cases where a desired function depends on multiple gene products, libraries harboring larger inserts are needed. These are normally constructed as cosmid, fosmid (30–40 kb), or bacterial artificial chromosome (BAC) libraries (up to ≥100 kb). The construction of comprehensive large-insert libraries can be very laborious, both with respect to the isolation of HMW DNA and successful cloning and transformation of the host. In addition, the lower stability of large inserts in the generated library needs to be considered. Also, the aspect of a higher degree of degradation of low guanine + cytosine (G + C) content DNA and some DNA modifications, which can impair cloning of HMW DNA, can result in a bias within large-insert libraries (Danhorn et al. 2012).

The choice of suitable vector systems is usually also related to the available expression host organism for subsequent screening experiments (Sect. 2.3). Moreover, for some targets, screening in multiple hosts can increase the hit rates (Mullany 2014). Hence library transfer and broad host-range capabilities of an expression vector (Sect. 2.2.2) can be desired characteristics (Craig et al. 2010; Aakvik et al. 2009; Kakirde et al. 2010).

2.2.1 Small- and Large-Insert Random Cloning Vectors

Cloning vectors useful for small-insert metagenomic library construction usually contain a defined promoter for transcription of the inserted DNA sequence. In some cases they are even equipped with two promoters (dual promotor vectors), flanking both sides of the cloning site in order to achieve gene expression regardless of insert orientation (Lammle et al. 2007). The promoters can have different, independent

induction mechanisms in order to achieve expression in only one direction at a time to prevent potential mRNA duplex formation that may result in lower protein production (Lale et al. unpublished). For cloning and construction of small-insert metagenomic libraries in *Escherichia coli* as primary host organisms, standard cloning vectors, such as pUC derivatives, pBluescript SK(+), and pTOPO, or their derivatives (Mullany 2014; Sabree et al. 2009) are frequently used.

In order to allow metagenomic library clones to cover entire biosynthetic pathways, like secondary metabolite clusters, large-insert libraries are required. Such libraries can be generated as cosmids or fosmids based on phage packaging of the eDNA ligated to a respective vector fragment or for very large inserts (up to 100 kb or more) as BACs (Kakirde et al. 2010; Danhorn et al. 2012). Fosmid and cosmid cloning vectors carry inserts of 30–40 kb, and both approaches utilize phage-based transfer of the cloned DNA into the host, usually *E. coli*. Consequently, the resulting library clones carry inserts within a narrow size range, determined by the packing capacity of the phage particle, and generally rely on gene expression from promoters included in the cloned insert. Cosmids are hybrid plasmids containing *cos* sequences from the λ phage, whereas fosmids are based on the F-factor replicon from *E. coli*. Compared to cosmids, fosmids are more tightly regulated with respect to copy number and are hence more stable (Kim et al. 1992; Kakirde et al. 2010). Both cosmids and fosmids are designed to carry antibiotic resistance markers and have broad host-range capabilities (Craig et al. 2010; Cheng et al. 2014; Aakvik et al. 2009; Wexler et al. 2005). Due to the frequent use of both cosmid and fosmid systems for metagenomic library construction, several variants (including commercial ones) are available (Lam et al. 2015; Mullany 2014; Kim et al. 1992; Parks and Graham 1997; Li et al. 2011; Terron-Gonzalez et al. 2013).

For random cloning of very large inserts, 40–100 kb and above, BACs are normally used, relying on the F-factor replicon (Danhorn et al. 2012; Shizuya et al. 1992). BAC vectors have been used in several metagenomic studies (Brady 2007) using, e.g., soil samples (Rondon et al. 2000) and murine bowel microbiota (Yoon et al. 2013). Similar to fosmids and cosmids, there are different BAC systems available, with some of them allowing inducible high copy numbers (Mullany 2014; Warburton et al. 2009; Wild et al. 2002) and/or having broad host-range capability (Mullany 2014; Aakvik et al. 2009; Kakirde et al. 2010). The US-based company Lucigen Corp. (Madison, WI; www.lucigen.com) has developed dedicated broad host-range vector systems for use in Functional Metagenomics. The pBAC-SBO and pSMART-BAC-S vectors both attribute efficient library construction in *E. coli* and are transferable to both Gram-positive and Gram-negative hosts. They have features allowing selection in several host organisms and gene expression from both insert-flanking regions, and are inducible in copy number (see Chap. 1). pSMART-BAC-S vector provides integration in the host genome only, whereas the pBAC-SBO vector allows both chromosomal integration, as well as extrachromosomal propagation in the recipient.

For DNA experiencing superhelical stress due to, e.g., regions dense in tandem and/or inverted repeats, cloning into circular plasmids can be challenging. In such cases, linear plasmids, such as the pJAZZ vector series (Lucigen), have been

designed which can carry large DNA inserts and contain features like transcriptional terminators flanking the cloning site to hinder vector-insert transcriptional interference (Godiska et al. 2010).

2.2.2 Broad Host-Range Expression Vectors

Depending on the desired activity, functional screening in different (or several) hosts can be of high value. As mentioned, *E. coli* is the most commonly used host both for library construction and functional screening. However, for certain screening activities, such as thermostable enzymes, or for bioactive secondary metabolite production, hosts like *Thermus thermophilus* (Angelov et al. 2009) and *Streptomyces* (or other *Actinobacteria*), respectively, might be beneficial due to their inherent features (Kakirde et al. 2010; Martinez et al. 2004) (see Sects. 2.3.1 and 2.3.2). Metagenomic libraries can be constructed directly in the host where they will be screened. However, the number of transformants obtained is often much lower in such hosts compared to the number of clones that can be obtained in *E. coli*. Thus, the common method is to utilize shuttle and/or broad host-range vectors for library construction in *E. coli*, which allows library transfer and screening in the host organism of choice. There are various such vectors available, both for small and large inserts. *E. coli–Bacillus subtilis* shuttle systems (plasmid and BAC) have been used for screening soil metagenomes for antimicrobial activities (Biver et al. 2013), and the pMDB14 vector (McMahon et al. 2012) can be shuttled between *E. coli*, *Pseudomonas putida*, and *Streptomyces lividans*, allowing gene expression in different hosts, similar to other systems reported (Sosio et al. 2000; Martinez et al. 2004). For development of psychrophilic expression systems, *E. coli* shuttle vectors such as a pGEM derivative and a pJRD215 derivative have been constructed, allowing the transfer of constructed libraries from *E. coli* to, e.g., *Psychrobacter* sp. and *Shewanella livingstonensis* (Cavicchioli et al. 2011; Miyake et al. 2007; Tutino et al. 2001). Also, *E. coli–T. thermophilus* shuttle systems have been designed (Angelov et al. 2009; Leis et al. 2015b). Apart from these, several other broad host-range systems have been developed. The pUvBBAC system supports replication in both Gram-positive and Gram-negative bacteria and allows functional screening in *Listeria* hosts (Hain et al. 2008). pGNS-BAC-1 presents opportunities for a copy induction in *E. coli*, as well as replication and functional screening in a broad spectrum of Gram-negative species (Kakirde et al. 2010). The pRS44 plasmid system (Aakvik et al. 2009) has been constructed both as fosmid and BAC system, which enables induction based on control on the vector copy number in *E. coli* and conjugative transfer into other hosts. In addition to the transferable BAC systems, several broad host-range cosmid vectors have also been reported (Craig et al. 2010; Cheng et al. 2014; Wexler et al. 2005).

In order to exploit the benefits of metagenomic library screening in several hosts with complementary features (Martinez et al. 2004; Leis et al. 2015a, b), efficient library transfer between host strains is of high importance. Though library vector