ACTA ABP BIOCHIMICA POLONICA

Vol. 62, No 1/2015
161–166
http://dx.doi.org/10.18388/abp.2014_917

Regular paper

The most widespread problems in the function-based microbial metagenomics

Agnieszka Felczykowska[†✉], Anna Krajewska[†], Sylwia Zielińska[†], Joanna M. Łoś, Sylwia K. Bloch and Bożena Nejman-Faleńczyk

Department of Molecular Biology, University of Gdańsk, Gdańsk, Poland

Metagenomics is a powerful tool to better understand the microbial niches, especially these from extreme habitats like oceans and seas, hot springs or deserts. However, one who is going to face the metagenomic studies should realize the challenges which might occur in the course of experiments. This manuscript indicates common problems in function-driven metagenomics, especially factors that influence gene expression are taken into account. Codon usage bias, internal cell accumulation, correct protein folding or presence of proper initiation factors are discussed and possible ways to overcome these problems are proposed. Finally, the annotation process is described, including possible limitations that one should take under consideration. What is more, the most popular databases for metagenomic data are mentioned and discussed.

Key words: function-based metagenomics; metagenomic library screening; gene expression; microbials; databases; codon usage bias; cell accumulation

Received: 10 October, 2014; revised: 27 October, 2014; accepted: 12 November, 2014; available on-line: 13 February, 2015

INTRODUCTION

Metagenomics is considered to be a powerful tool in research of the microbial world which is based on culture-independent investigation of various habitats through the isolation of DNA directly from the environmental sample and analysis of the target genes (Tringe *et al.*, 2005). Metagenomic studies opened the door to become acquainted with the 99% of bacterial communities that are (yet) believed to be uncultivable under laboratory conditions (Schloss & Handelsman, 2005; Benndorf *et al.*, 2007; Felczykowska *et al.*, 2012). This approach, instead of 16S rDNA-based gene research, permits not only for phylogenetic surveys, but also reveals the complete gene combination of organisms, phylogeny and even evolutionary profiles of a community structure (Thomas *et al.*, 2012). Combined with metaproteomics and metatranscriptomics, metagenomic study could be a remarkable implement in interpretation of gene regulation related to target activities (Moran, 2009; Gilbert & Hughes, 2011).

Metagenomic study involves two different general approaches: the sequence-driven and the functional-driven metagenomics (Kennedy *et al.*, 2011). The first one is based on sequencing of the DNA representing the entire environmental sample and through employment of Next Generation Sequencing methods (NGS) it allows to obtain complex information about the organisms included in the sample. However, researchers involved in the sequenced-based metagenomics are faced with analysis of hundreds, or even thousands of gigabytes of data obtained from sequencing. This could be challenging when specific activities are requested, especially when specific enzymes are desired, where the investigation is usually limited to conserved regions of already well characterized protein families (Kennedy *et al.*, 2008). By contrast, the function-based approach is directed at a particular activity and metagenomic library is screened to find this activity in the clones obtained. Depending on the screening method chosen, the individual clone or a group of clones are investigated at the same time, usually for an enzymatic activity and the clones selected are sequenced to define the target gene. The advantage of the function-driven screening is the certainty that the target compound is biosynthesized correctly and can be produced by the host cell. This aspect is particularly important in the light of reports indicating that the expression of foreign genes in the most popular host cell in the metagenomics study — *Escherichia coli* — is limited to about 40% (McMahon *et al.*, 2012).

However, despite being, admittedly, a great tool in investigation of undiscovered bacterial habitats, activity-based metagenomics remained a challenging field of study for several reasons: (i) isolation of environmental DNA (eDNA), (ii) selection of screening assay, (iii) data analysis and annotation, (iv) limitation of expression properties of the foreign host. This article summarizes the most frequently encountered challenges during metagenomic library construction and analysis. The bias with sampling and isolation of eDNA from soil, water and sludge habitats is discussed. There are many aspects of available assays in searching for the bioactive compounds and their limitations, especially chromogenic and fluorogenic substrates. Then, factors which influence an efficient expression are indicated: recognition of promoters given in the gene library, regulatory agents of the transcription system of the host, toxicity

✉e-mail: agnieszka.felczykowska@biol.ug.edu.pl

[†]These authors contributed equally to this work

Abbreviations: BAC, bacterial artificial chromosome; BATS, Bermuda Atlantic Time-series Study site; CAMERA, The Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis; CDSs, protein coding sequences; CUB, codon usage bias; DOC, dissolved organic carbon; eDNA, environmental DNA; EMBLEBI, European Molecular Biology Laboratory's European Bioinformatics Institute; ENA, European Nucleotide Archive; GOLD, The Genomes On Line Database; IMG, Integrated Microbial Genomes; MGA, MetaGenome Annotator; MG, MetaGene; MG-RAST, metagenomics Rapid Annotation using Subsystems Technology; RBS; ribosomal binding site

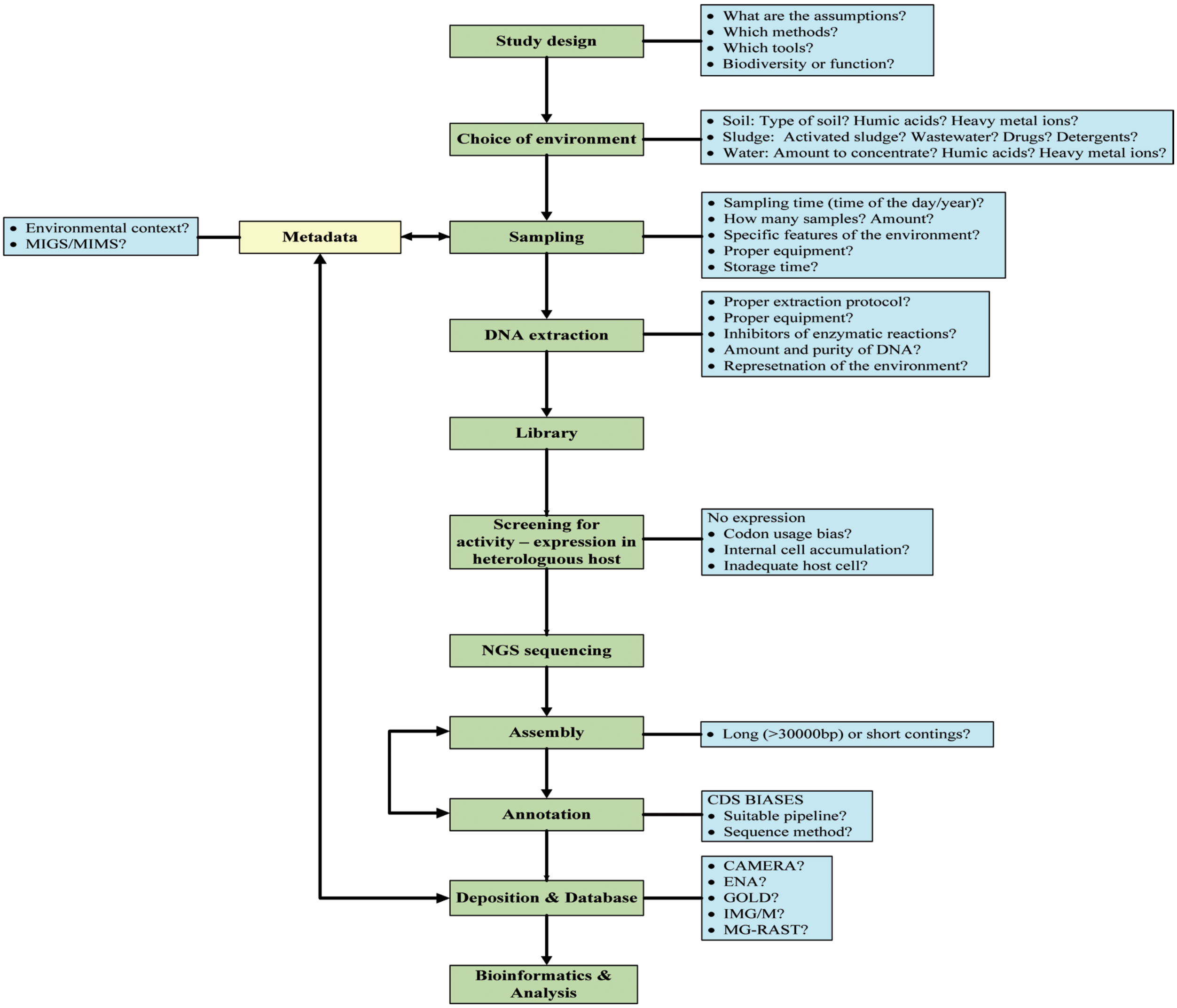
Study design
• What are the assumptions?
• Which methods?
• Which tools?
• Biodiversity or function?
Choice of environment
• Soil: Type of soil? Humic acids? Heavy metal ions?
• Sludge: Activated sludge? Wastewater? Drugs? Detergents?
• Water: Amount to concentrate? Humic acids? Heavy metal ions?
• Environmental context?
• MIGS/MIMS?
Metadata
Sampling
• Sampling time (time of the day/year)?
• How many samples? Amount?
• Specific features of the environment?
• Proper equipment?
• Storage time?
DNA extraction
• Proper extraction protocol?
• Proper equipment?
• Inhibitors of enzymatic reactions?
• Amount and purity of DNA?
• Represetnation of the environment?
Library
Screening for activity – expression in heterologuous host
No expression
• Codon usage bias?
• Internal cell accumulation?
• Inadequate host cell?
NGS sequencing
Assembly
• Long (>30000bp) or short contings?
Annotation
CDS BIASES
• Suitable pipeline?
• Sequence method?
Deposition & Database
• CAMERA?
• ENA?
• GOLD?
• IMG/M?
• MG-RAST?
Bioinformatics & Analysis

Figure 1. The function-driven metagenomics scheme.
General steps are shown in green. The blue boxes represent common problems which one could face at each step. Challenges connected to screening for activity, NGS sequencing, assembly, annotation and deposition & databases are discussed in this article. The isolation of eDNA, sampling and metadata problems are described in the article by Felczykowska *et al.* (2015) Sampling, metadata and DNA extraction — important steps in metagenomic studies.

of gene products, codon usage differences (CUB), correct protein folding, presence of proper initiation factors, or the capacity of the host to secret the gene expression product (Craig *et al.*, 2010; Lorenz *et al.*, 2012). This article also highlights some of them and proposes possible ways to overcome them (Fig. 1). Finally, data analysis of clones of interest is mentioned, particularly the annotation of obtained sequences and data storage is considered.

SCREENING FOR ACTIVITY

Selection of adequate clone screening method is of crucial importance in the function-based metagenomics. Generally, there are three different approaches considered to obtain novel bioactive compounds: (i) phenotypical detection of the desired activity (Belogui *et al.*, 2010; Gloux *et al.*, 2010; Liaw *et al.*, 2010), (ii) heterologous complementation of host strains or mutants (Riesenfield *et al.*, 2004; Wang *et al.*, 2006; Chen *et al.*, 2010), and (iii) induced gene expression (Uchiyama *et al.*, 2005; Williamson *et al.*, 2005; Uchiyama & Miyazaki, 2010).

The first method has been commonly employed in searching for an enzymatic activity and applies chromophore-containing derivatives of the enzyme's substrates, which are used to observe the enzymatic reaction. Since the average metagenomic library contains hundred thousands of clones with the potential target activity, the enzyme assay should be designed to investigate the largest number of clones at the same time. Chow *et al.* (2012) discovered and characterized two novel thermostable lipases by the functional-driven screening. The lipase activity was searched for by using cultivation of the metagenomic clones on LB agar plates supplemented with 1% tributyrin as an indicator substrate (Lawrence *et al.*, 1967). The target clone bearing the lipolytic activity occurred to be surrounded by a transparent halo. The plate screening was employed by Lee *et al.* (2007), where LB agar with the addition of 1% of skim milk has been used for investigation of approximately 30000 clones for proteolytic activity. This method allowed for identification of novel zinc-dependent metalloproteinase from the mud in the west coast of Korea.

The phenotypical detection approach has been also adapted to discovery of biomolecules of antimicrobial activity. The most accessible strategy is based on screening of metagenomic library clones for one that generates a zone of growth inhibition in top agar layer on a plate assay (Banik & Brady, 2010). This method led to the discovery of secondary metabolites such as indigo, patellamide, and pederin (Piel *et al*., 2004).

All examples presented prove the concept that simple high throughput assays can be successfully employed to examine the metagenomic library for various activities. However, there is a number of challenges on the way of a researcher whose task is to explore the metagenomic clones.

Internal cell accumulation

One of the problems commonly occurring while working with function-driven metagenomics is the accumulation of the biosynthesized compound inside the host cell which can cause a limitation in detection of the target activity (Kennedy *et al*., 2011). Nonionic detergents, such as Tween or Triton X may be adapted to induce a gentle lysis of the host cell and, as a result, release the target biomolecule. Of particular importance is the fact that nonionic detergents allow to maintain the native (and active) conformations of a compound that is crucial in the case of enzymatic assays (Gloux *et al*., 2007).

Another solution, which can be used when thermostable enzymes are to be detected, is to employ an elevated temperature to elicit the cell lysis. Chow *et al.* (2012) incubated the cosmid clones at plates supplemented with tributyrin (1% v/v) for 1–3 days at 56°C to screen for a novel thermostable lipase. The incubation step was introduced to slowly lyse the *E. coli* cells and to release those enzymes that are active to degrade tributyrin at elevated temperatures and produce a clear halo.

To overcome the cell-internal accumulation of biocompounds, the problem is resolved with the screen of individual clones on 96- or 386-well plates employing colony-picking robots, liquid handlers and microtitre readers (Kennedy *et al*., 2011). Adaptation of professional equipment in the combination with high-throughput screening including fluorophores and chromophores such as nitrophenoles, umbelliferones, fluoresceins, rhodamines and BODIPY dyes, could be the solution to improve exploration of the metagenomic library clones while maintaining the selectivity and sensitivity of the assay method (Reymond *et al*., 2009). Felczykowska *et al.* (2014) tested individual clones belonging to a fosmid metagenomic library made from eDNA isolated from cyanobacteria bloom obtained in the Baltic Sea. Approximately 400 *E. coli* clones bearing environmental DNA were screened for anticancer and antibacterial activity. Every clone was precultivated and the host cells were sonicated to release bioactive compounds. The anticancer activity was determined by the MTT method on a 96-well plate (MTT method determined by Mossman, 1983) and the antibacterial activities against *Micrococcus luteus*, *Staphylococcus aureus*, *Pseudomonas aeruginosa* and *Serratia marcescens* were tested using the microdilute assay method.

Codon usage bias

Roller *et al.* (2013) examined the data obtained from eleven metagenomic samples from eight distinct environments whether the microbes sharing the habitat have the same codon usage bias: the Sargasso Sea, three whale fall carcass samples, Waseca farm soil, human gut microbiome, lean and obese mouse gut microbiomes, an acid mine drainage and two geographically distant enhanced phosphorous removal sludges. They demonstrated that metagenomic-centric bias in codon usage is phylogeny-independent by two parameters (i) the variability of single species' CU across metagenomes; and (ii) the variability of CU in a metagenome upon removal of the dominant phyla. What is more, they provided evidence that microbial communities exhibit codon usage bias similar to that already described for a single microbial species. It suggests that microbial communities sharing an environment are likely to have synchronized regulation mechanisms of translational optimization for expression of environment-specific genes.

There are 61 trinucletide codons encoding 20 different amino acids which means that several codons can carry the information for the insertion of the same amino acid into a protein (Berg & Kurland, 1997). The article of Roller *et al.* (2013) demonstrated that different microbial communities present different preference of codon usage. There is a high correlation between frequency of codon usage in an organism and the pool of cognate tRNAs (Fakruddin *et al*., 2013). Highly expressed genes carry codons for which there is a large pool of cognate tRNAs while regulatory genes commonly use codons for which there is only a very small jackpot of cognate tRNAs (Stoletzki & Eyre-Walker, 2007).

Overexpression of genes with the significant content of rare codons may result in incorrect biosynthesis of the bioactive compound. Since the amount of rare codons in eDNA depends on the sample's habitat, this aspect should be taken into account by one who is about to work with function-based metagenomics. There are reports that expression efficiency of genes containing rare codons can be dramatically improved when the level of cognate tRNAs is increased within the host (Seidel *et al*., 1992; Rosenberg *et al*., 1993). This can be achieved by inserting the wild type tRNA gene in the expression vector itself or when it is placed in a compatible multiple copy plasmid. One of the possibilities is the use of a pRIG plasmid, which contains the *argU*, *ileX*, and *glyT* tRNA genes under their native promoters on a pACYC backbone, which carries the p15a origin of replication (Baca & Hol, 2000). This strategy can be adapted for optimization of expression of genes obtained from organisms with AT or GC rich genomes that is connected to codon usage bias. Baca *et. al* (2000) significantly enhanced the expression of a number of genes derived from AT-rich *Plasmodium* genome using the pRIG plasmid.

The latter possible solution which can be employed when the codon usage bias occurs is the employment of an alternative host. In most studies based on metagenomic techniques, *E. coli* is the standard host. However, as mentioned previously, the use of *E. coli* as a host cell allows expression of only 40% of genes contained in the eDNA of an average sample (Craig *et al*., 2010; McMahon *et al*., 2012). Due to this restriction, alternative host strains from *Bacillus*, *Pseudomonas* or *Streptomyces* genera can be applied (Lorenz & Eck, 2005; Aakvik *et al*., 2009). There are also several archaeal genera (*Methanococcus*, *Pyrococcus*, *Sulfolobus*, *Thermococcus*), which have been successfully employed in designing a stable host-vector expression system (Angelov & Liebl, 2010).

For those who are not able to verify if the eDNA sample contains the rare codons, the safest solution is to apply the broad-host range vectors. There are several reports that the use of different hosts can increase the yield of novel bioactive compounds uncovered by function-driven metagenomics methods (Craig *et al.*, 2010; de Castro *et al.*, 2011). For instance, the application of bacterial genera such as *Pseudomonas*, *Rhizobium* or *Streptomyces*, which have over fifteen RNA polymerase σ factors (*Escherichia coli* has only seven), may be crucial in the expression of genes that require specialized σ factors (Gabor *et al.*, 2004).

DATA ANALYSIS AND STORAGE

When the screening of metagenomic library is successful and sequencing has been performed, the data analysis should begin.

Annotation

Functional annotation includes identification of the features of interest (genes), and then prediction of the function of putative features and determination of taxonomic neighbors (Thomas *et al.*, 2012). There are two possible strategies that could be applied in the functional annotation. The first one is provided for the contings of the length of 30 000 bp or more and commonly applied when reconstructed genomes are the targets. In this case, the existing pipelines for genome annotation, such as IMG/M (Markowitz *et al.*, 2009), MG-RAST (Aziz *et al.*, 2008) or WebMGA (Wu *et al.*, 2011) may be adapted. All three tools tender the complex annotation system combined with the useful features like biodiversity analysis, taxonomic classification and genome comparison (Teeling & Glockner, 2012). The latter approach provides annotation for all the organisms included in the community and is designated for short contings or unassembled reads. In this case the use of genome annotation pipelines is dramatically limited and the specialized tools for metagenomic annotation should be applied (Thomas *et al.*, 2012).

The crucial step in functional annotation is determination of protein coding sequences (CDSs). There is a number of available algorithms, adapted for the annotation of full genome sequences with the estimated accuracy of about 95% in the prediction of CDSs (Lukashin & Borodovsky, 1998). There are several tools adjusted to predict the CDSs, such as MetaGeneMark (McHardy *et al.*, 2007), FragGeneScan (Rho *et al.*, 2010) or MetaGene Annotator (Noguchi *et al.*, 2008). All tools listed are based on internal information, such as codon usage, to categorize the sequence fragment as coding or noncoding (Thomas *et al.*, 2012).

MetaGene (MG) is a package software which is able to predict the gene in metagenomic sequence in two steps. Firstly, the ORFs are extracted from the sequence. In the second stage, all the ORFs are scored according to their base composition and length by a specialized log-odds scoring scheme (Yok & Rosen, 2011). Despite the fact that MetaGene is a useful software for one who works with metagenomic data, there are two significant drawbacks of employing the MetaGene tool: the lack of a ribosomal binding site (RBS) model, and a low sensitivity to genes with different codon usage compared to the "typical" genes (Noguchi *et al.*, 2008). A new version of the program — MetaGenome Annotator — was designed to overcome those limitations and to expand the application of the software. MGA has statistical models of prophage genes which can be useful to reveal lateral gene transfers or phage infections. Moreover, an adapted RBS model based on complementary sequences of the 30 bp tail of 16 S ribosomal RNA enables for prediction of gene start sites even when input genomic sequences are short and anonymous sequences (Yok & Rosen, 2011).

The functional annotation is still the most challenging computational step in a number of metagenomic projects. There are reports that only 20 to 50% of metagenomic data can be annotated (Gilbert *et al.*, 2010). The greatest limitation of the accessible tools applied for data analysis is based on the comparison to the actual database resources. In the light of this circumstance it is of significant importance to provide the free flow of information between metagenome analysis platforms for the best analysis of metagenomic data.

Databases for metagenomic data

Next generation sequencing techniques like 454 pyrosequencing or Illumina system produce a large amount of data in a short period of time with relatively low cost. Data obtained from heterogeneous microbial communities can contain information from even more that 10 000 species but the sequencing data can be partial and of low quality (Wooley *et al.*, 2010), and it can be described as one of the metagenomic challenges. It is even a greater challenge to collect all sequences generated by metagenomic projects for nucleic acid sequence data archives and it can be even more difficult to possess specialized databases that will be able to offer consistent storage and querying of the metagenomic data, which also allow access to widen the context of a specific project (Handelsman *et al.*, 2007).

Over the last few years, we have had to deal with a number of initiatives to build an infrastructure dedicated for metagenomic sequences and associated metadata. One of them is the CAMERA project (The Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis, http://camera.calit2.net/) (Sun et al., 2011) a joint venture of the University of California, San Diego and J. Craig Venter Institute (JCVI). The aim of this project is to provide cyber infrastructure tools, resources and bioinformatics expertise in order to facilitate on-line collaboration to share and forward data. The four main pillars of the CAMERA project are (i) transferring data directly from a sequencing centre, (ii) community or user contributions, (iii) data exchange with public data resources, (iv) integration of reference data sets (Amid *et al.*, 2012). In the database design, MIMS and MIGS, established by the Genomics Standards Consortium (Field *et al.*, 2008) have been adapted to handle samples' metadata. Unfortunately, due to the lack of support, CAMERA no longer accepts user's submissions, but continues to maintain free and open access of data already submitted.

The European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena) is maintained and developed at the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBLEBI). ENA is dedicated to gain and present information that is connected to experimental workflows of nucleotide sequencing; typically, it includes the isolation and preparation of material for sequencing, producing sequencing data and bioinformatic analysis. ENA records implemented information into data model that covers input information, output machine data and interpreted information. ENA also provides data from various international databases, provides submission tools as well as many program and search technologies, in order to provide access to the data sets stored (Amid *et al.,* 2012). Users, by working with provided tools, can create studies, samples and experiments to facilitate creation of metadata and submit them for public use. It intends to provide permanent scientific records as a complementation to published literature and allows sharing of pre-publication data. ENA assures the basis for bioinformatics infrastructure.

At the Department of Energy`s Joint Genome Institute, there is an existing microbial genome database project (IMG — Integrated Microbial Genomes), which serves as a community resource for comparative analysis of publicly available genomes in a comprehensive integrated context (Markowitz *et al.*, 2012), expanded to cope with metagenomic data (IMG/M). Primarily, IMG/M was developed as an experimental system, but subsequently it has been extended in terms of metagenome data content and metagenome specific analytical tools (Markowitz *et al.*, 2008). The aim of this platform is to support comparative metagenome analysis in the context of microbial genome and metagenomic data, which are often generated with the use of various sequencing technology platforms, as well as diverse data processing methods. IMG/M system has been extended by regular updates since its first release, and it is available at http://img.jgi.doe.gov/m. A companion IMG/M system provides support for annotation and expert review of unpublished metagenomic data sets (IMG/M ER: http://img.jgi.doe.gov/mer) (Markowitz *et al.*, 2012).

In addition to those already mentioned, there are several other projects around the world that developed various specific data models and interfaces, sometimes build on an already existing database and often constructed for metagenomic data. Examples are The Genomes On Line Database (GOLD, http://www.genomesonline.org) (Liolios *et al.*, 2010) and the metagenomics RAST server (MG-RAST, metagenomics Rapid Annotation using Subsystems Technology; http://metagenomics.nmpdr.org) (Meyer *et al.*, 2008). GOLD contains information from both complete and ongoing projects, also providing metadata associated with them, build in accordance with the MIGS/MIMS. The metagenomics RAST is an open source system based on MIGS specification as well. This platform was designed to upload raw sequence data in a fast format and provides several methods to access the different data types, including phylogenetic and metabolic reconstructions.

Due to emergence of the next generation sequencing techniques, it is necessary to create new databases and tools needed for analysing and data sharing. Each metagenome project has its own assumptions and requires specific analytical tools and software. The variety of available options gives a choice of the appropriate tools required for a particular project with a user-friendly interface.

Acknowledgements

This work was supported by European Commission (grant no. 245137, project acronym: MAREX, large scale integrating project within KBBE-2009-3-2-01) and Ministry of Science and Higher Education, Poland (grant no. 1544/7.PR UE/2010/7). This work was also supported by National Science Center, Poland (grant no. 2011/01/D/NZ2/04817).

REFERENCES

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

Amid C, Birney E, Bower L, Cerdeño-Tárraga A, Cheng Y, Cleland I, Faruque N, Gibson R, Goodgame N, Hunter C, Jang M, Leinonen R, Liu X, Oisel A, Pakseresht N, Plaister S, Radhakrishnan R, Reddy K, Rivière S, Rossello M, Senf A, Smirnov D, Hoopen PT, Vaughan D, Vaughan R, Zalunin V, Cochrane G (2012) Major submissions tool developments at the European nucleotide archive. *Nucleic Acids Res* **40**: D43–D47.

Angelov A, Liebl W (2010) Heterologous gene expression in the hyperthermophilic archaeon *Sulfolobus solfataricus*. *Methods Mol Biol* **668**: 109–116.

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Overbeek RA, McNeal LK, Paarmann D, Paczian T, Parello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**: 75.

Baca AM, Hol WGJ (2000) Overcoming codon bias: a general method for high-level overexpression of *Plasmodium* and AT-rich parasite genes in *Escherichia coli*. *Int J Parasitology* **30**: 113–118.

Banik JJ, Brady SF (2010) Recent application of metagenomic approaches toward the discovery of antimicrobials and other bioactive small molecules. *Curr Opin Microbiol* **13**: 603–609.

Beloqui A, Nechitaylo TY, López-Cortés N, Ghazi A, Guazzaroni MG, Polaina J, Strittmatter AW, Reva O, Waliczek A, Yakimov MM, Golyshina OV, Ferrer M, Golyshin PN (2010) Diversity of glycosyl hydrolases from cellulosedepleting communities enriched from casts of two earthworm species. *Appl Environ Microbiol* **76**: 5934–5946.

Benndorf D, Balcke GU, Harms H, von Bergen M (2007) Functional metaproteome analysis of protein extracts from contaminated soil and groundwater. *ISME J* **1**: 224–234.

Berg OG, Kurland CG (1997) Growth rate-optimised tRNA abundance and codon usage. *J Mol Biol* **270**: 544–550.

Chen IC, Thiruvengadam V, Lin WD, Chang HH, Hsu WH (2010) Lysine racemase: a novel non-antibiotic selectable marker for plant transformation. *Plant Mol Biol* **72**: 153–169.

Chow J, Kovacic F, Antonia YD, Krauss U, Fersini F, Schmeisser C, Lauinger B, Bongen P, Pietruszka J, Schmidt M, Menyes I, Bornscheuer UT, Eckstein M, Thum O, Liese A, Mueller-Dieckmann J, Jaeger KE, Streit WR (2012) The Metageome-Derived Enzymes LipS and LipT Increase the Diversity of Known Lipases. *PLoS ONE* **7**: e47665.

Craig JW, Chang FY, Kim JH, Obiajulu SC, Brady SF (2010) Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse *Proteobacteria*. *Appl Environ Microbiol* **5**: 1633–1641.

de Castro AP, Quirino BF, Allen H, Williamson LL, Handelsman J, Krüger RH (2011) Construction and validation of two metagenomic DNA libraries from Cerrado soil with high clay content. *Biotechnol Lett* **33**: 2169–2175.

Fakruddin M Mazumdar RM, Mannan KSB, Chowdhury A, Hossain MN (2013) Critical factors affecting the success of cloning, expression, and mass production of enzymes by recombinant *E.coli*. *SRN Biotechnol* **2013**: 590587.

Felczykowska A, Bloch SK, Nejman-Faleńczyk B, Barańska S (2012) Metagenomic approach in the investigation of new bioactive compounds in the marine environment. *Acta Biochim Pol* **59**: 501–505.

Felczykowska A, Dydecka A, Bohdanowicz MG, Gąsior T, Soboń M, Kobos J, Bloch S, Nejman-Faleńczyk BE, Węgrzyn G (2014) The use of fosmid metagenomic libraries in preliminary screening for various biological activities. *Microb Cell Fact* **13**: 105.

Felczykowska A, Krajewska A, Zielińska S, Łoś JM (2015) Sampling, metadata and DNA extraction — important steps in metagenomic studies. *Acta Biochim Pol* **62**: 151–160.

Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, Ashburner M, Axelrod N, Baldauf S, Ballard S, Boore J, Cochrane G, Cole J, Dawyndt P, De Vos P, DePamphilis C, Edwards R, Faruque N, Feldman R, Gilbert J, Gilna P, Glöckner FO, Goldstein P, Guralnick R, Haft D, Hancock D, Hermjakob H, Hertz-Fowler C, Hugenholtz P, Joint I, Kagan L, Kane M, Kennedy J, Kowalchuk G, Kottmann R, Kolker E, Kravitz S, Kyrpides N, Leebens-Mack J, Lewis SE, Li K, Lister AL, Lord P, Maltsev N, Markowitz V, Martiny J, Methe B, Mizrachi I, Moxon R, Nelson K, Parkhill J, Proctor L, White O, Sansone SA, Spiers A, Stevens R, Swift P, Taylor C, Tateno Y, Tett A, Turner S, Ussery D, Vaughan B, Ward N, Whetzel T, San Gil I, Wilson G, Wipat A (2008) The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol* **26**: 541–547.

Gabor EM, Alkema WBL, Janssen DB (2004) Quantifyling the accessibility of the metagenome by random expression cloning techniques. *Environ Microbiol* **6**: 879–886.

Gilbert JA, Field D, Swift P, Thomas S, Cummings D, Temperton B, Weynberg K, Huse S, Hughes M, Joint I, Somerfield PJ, Mühling M (2010) The taxonomic and functional diversity of microbes at a temperate coastal site: a multi-omic study of seasonal and diel temporal variation. *PLoS One* **5**: e15545.

Gilbert JA, Hughes M (2011) Gene expression profiling: metatranscriptomics. *Methods Mol Biol* **733**: 195–205.

Gloux K, Leclerc M, Iliozer H, L'Haridon R, Manichanh C, Corthier G, Nalin R, Blottière HM, Doré J (2007) Development of high-throughput phenotyping of metagenomic clones from the human gut microbiome for modulation of eukaryotic cell growth. *Appl Environ Microbiol* **73**: 3734–3737.

Gloux K, Berteau O, El Oumami H, Béguet F, Leclerc M, Doré J (2010) Microbes and Health Sackler Colloquium: a metagenomic_-glucuronidase uncovers a core adaptive function of the human intestinal microbiome. *Proc Natl Acad Sci USA* doi:10.1073/pnas.1000066107.

Handelsman J *et al.* (2007) (Committee on Metagenomics: Challenges and Functional Applications, National Research Council). The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet; The National Academies Press, Washington, DC, USA.

Kennedy J, Marchesi JR, Dobson ADW (2008) Marine Metagenomics: strategies for the discovery of novel enzymes with biotechnological applications from marine ecosystems. *Microb Cell Fact* **7**: 27.

Kennedy J, O'Leary ND, Kiran GS, Morrissey JP, O'Gara F, Selvin J, Dobson ADW (2011) Functional metagenomic strategies for the discovery of novel enzymes and biosurfactants with biotechnological applications from marine ecosystems. *J Appl Microbiol* **111**: 787–799.

Lawrence RC, Fryer TF, Reiter B (1967) Rapid method for the quantitative estimation of microbial lipases. *Nature* **213**: 1264–1265.

Lee DG, Jeon JH, Jang MK, Kim NY, Lee JH, Kim SJ, Kim GD, Lee SH (2007) Screening and characterization of a novel fibrinolytic metalloprotease from a metagenomic library. *Biotechnol Lett* **29**: 465–472.

Liaw RB, Cheng MP, Wu MC, Lee CY (2010) Use of metagenomic approaches to isolate lipolytic genes from activated sludge. *Bioresour Technol* **101**: 8323–8329.

Liolios K, Chen IA, Mavromatis K, Tavernarakis N, Hugenholtz P, Markowitz VM, Kyrpides NC (2010) The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* **38**: D346–D354.

Lorenz P, Eck J (2005) Metagenomics and industrial applications. *Nature Rev Microbiol* **3**: 510–516.

Lorenz P, Liebeton K, Niehaus F, Eck J (2012) Screening for novel enzymes for biocatalytic processes: accessing the metagenome as a resource of novel functional sequence space. *Curr Opin Biotechnol* **13**: 572–577.

Lukashin AV, Borodovsky M (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res* **26**: 1107–1115.

Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D, Chen I, Grechkin Y, Dubchak I, Anderson I, Lykidis A, Mavromatis K, Hugenholtz P, Kyrpides NC (2008) IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res* **36**: D534–D538.

Markowitz VM, Mavromatis K, Ivanova NN, Chen IM, Chu K, Kyrpides NC (2009) IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics* **25**: 2271–2278.

Markowitz VM, Chen I, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Mavromatis NN, Kyrpides NC (2012) IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res* **40**: D115–D122.

McHardy ACZ, Martin HGL, Tsirigos A, Hugenholtz P, Rigoutsos IB (2007) Accurate phylogenetic classification of variable- length DNA fragments. *Nat Methods* **4**: 63–72.

McMahon MD, Guan C, Handelsman J, Thomasa MG (2012) Metagenomic analysis of *Streptomyces lividans* reveals host-dependent functional expression. *Appl Environ Microbiol* **6**: 3622–3629.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA (2008) The metagenomics RAST server — a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**: 38.

Moran MA (2009) Metatranscriptomics: Eavesdropping on Complex Microbial Communities. *Microbe* 4: 329–335.

Mossman T (1983) Rapid colorimetric assay for cellular growth and survival: Application to proliferation and cytotoxicity assays. *J Immunol Methods* **65**: 5–63.

Noguchi H, Taniguchi T, Itoh T (2008) MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res* **15**: 387–396.

Piel J, Hui D, Wen G, Butzke D, Platzer M, Fusetani N, Matsunaga S (2004) Antitumor polyketide biosynthesis by an uncultivated bacterial symbiont of the marine sponge *Theonella swinhoei*. *Proc Natl Acad Sci USA* **101**: 16222–16227.

Reymond JL, Fluxà VS, Maillard N (2009) Enzyme assays. *Chem Commun* **7**: 34–46.

Rho M, Tang H, Ye Y (2010) FragGeneScan: predicting genes in short and error- prone reads. *Nucleic Acids Res* **38**: e191.

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

Roller M, Lucić V, Nagy I, Perica T, Vlahovicek K (2013) Environmental shaping of codon usage and functional adaptation across microbial communities. *Nucleic Acids Res* **44**: 8842–8852.

Rosenberg AH, Goldman E, Dunn JJ, Studier FW, Zubay G (1993) Effects of consecutive AGG codons on translation in Escherichia coli, demonstrated with a versatile codon test system. *J Bacteriol* 175: 716–722.

Schloss PD, Handelsman J (2005) Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome Biology* **6**: 229.

Seidel HM, Pompliano DL, Knowes JR (1992) Phosphonate biosynthesis: molecular cloning of the gene for phosphoenolpyruvate mutase from Tetrahymena pyriformis and overexpression of the gene product in *Escherichia coli*. *Biochemistry* 31: 2598–2608.

Stoletzki N, Eyre-Walker A (2007) Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. *Mol Biol Evol* **24**: 374–387.

Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S, Stocks K, Allen EE, Ellisman M, Grethe J, Wooley J (2011) Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Res* **39**: D546–D551.

Teeling H, Glockner FO (2012) Current opportunities and challenges in microbial metagenome analysis-a bioinformatic perspective. *Briefings Bioinform* **13**: 728–742.

Thomas T, Gilbert J, Meyer F (2012) Metagenomics — a guide from sampling to data analysis. *Microb Inform Exp* **2**: 3.

Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW (2005) Comparative metagenomics of microbial communities. *Science* **308**: 554–557.

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.

Uchiyama T, Miyazaki K (2010) Product-induced gene expression (SIGEX): a product-responsive reporter assay for enzyme screening of metagenomic libraries. *Appl Environ Microbiol* **76**: 7029–7035.

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.

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.

Wooley JC, Godzik A, Friedberg I (2010) A primer on Metagenomics. *Computal Biol* **2**: 1–13.

Wu S, Zhu Z, Fu L, Niu B, Li W (2011) WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genomics* **12**: 444.

Yok NG, Rosen GL (2011) Combining gene prediction methods to improve metagenomic gene annotation. *BMC Bioinformatics* **12**: 20.