<table>
<thead>
<tr>
<th>Item Type</th>
<th>Article</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors</td>
<td>Kodzius, Rimantas; Gojobori, Takashi</td>
</tr>
<tr>
<td>Citation</td>
<td>Marine metagenomics as a source for bioprospecting 2015 Marine Genomics</td>
</tr>
<tr>
<td>Eprint version</td>
<td>Publisher’s Version/PDF</td>
</tr>
<tr>
<td>DOI</td>
<td>10.1016/j.margen.2015.07.001</td>
</tr>
<tr>
<td>Publisher</td>
<td>Elsevier BV</td>
</tr>
<tr>
<td>Journal</td>
<td>Marine Genomics</td>
</tr>
<tr>
<td>Rights</td>
<td>© 2015. This manuscript version is made available under the CC-BY-NC-ND 4.0 license <a href="http://creativecommons.org/licenses/by-nc-nd/4.0/">http://creativecommons.org/licenses/by-nc-nd/4.0/</a></td>
</tr>
<tr>
<td>Download date</td>
<td>2023-11-01 11:34:16</td>
</tr>
<tr>
<td>Link to Item</td>
<td><a href="http://hdl.handle.net/10754/567056">http://hdl.handle.net/10754/567056</a></td>
</tr>
</tbody>
</table>
Marine metagenomics as a source for bioprospecting

Rimantas Kodzius\textsuperscript{a,c,d,+}, Takashi Gojobori\textsuperscript{a,b,d}\textsuperscript{⁎}

\textsuperscript{a} Computational Bioscience Research Center (CBRC), Saudi Arabia
\textsuperscript{b} Biological and Environmental Sciences and Engineering Division (BESE), Saudi Arabia
\textsuperscript{c} Computer, Electrical and Mathematical Sciences and Engineering Division (CEMSE), Saudi Arabia
\textsuperscript{d} King Abdullah University of Science and Technology (KAUST), Saudi Arabia

Abstract
This review summarizes usage of genome-editing technologies for metagenomic studies; these studies are used to retrieve and modify valuable microorganisms for production, particularly in marine metagenomics. Organisms may be cultivable or uncultivable. Metagenomics is providing especially valuable information for uncultivable samples. The novel genes, pathways and genomes can be deducted. Therefore, metagenomics, particularly genome engineering and system biology, allows for the enhancement of biological and chemical producers and the creation of novel bioresources. With natural resources rapidly depleting, genomics may be an effective way to efficiently produce quantities of known and novel foods, livestock feed, fuels, pharmaceuticals and fine or bulk chemicals.

© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Contents
1. Introduction ............................................................... 0
2. Metagenomics as a tool to study the environment .............................................. 0
2.1. Diversity of microorganisms. ............................................... 0
3. Bioprospecting—the process of discovery and commercialization of new products based on biological resources ..................................... 0
3.1. Sequence-based screening .................................................. 0
3.2. Assessment of metabolic potential by function-based screening, ......................... 0
4. Metabolic engineering as a tool for cell factories ................................................ 0
4.1. Established production strains. ............................................ 0
4.1.1. E. coli as a producer .................................................... 0
4.1.2. Algae as a producer .................................................... 0
4.1.3. S. cerevisiae as a producer. ............................................ 0
4.2. Tools to improve the function of strains ........................................... 0
4.2.1. Genetic engineering using molecular biology tools ......................... 0
4.2.2. New organisms created by synthetic biology ............................ 0
5. Conclusion/outlook ............................................................ 0
References .................................................................. 0

1. Introduction
Microbes are ubiquitous and an essential part of all life on Earth. They may be considered a major bioresource, acting as powerful chemical factories that transform environmental chemicals (carbon, hydrogen, nitrogen, oxygen, phosphorus and sulfur) into more biologically accessible molecules, which are then used by higher organisms (Adkins et al., 2012). All higher life forms, including plants and animals,
host over orders of magnitude from millions to trillions of microbes. Some make necessary nutrients and vitamins or help to digest and convert food. Microbes also benefit the environment by removing contaminants and pollutants from soil, groundwater, sediment and surface water. However, our understanding of these complex communities remains limited.

Marine environments comprise over 70% of the earth's surface, ranging from habitats in the freezing Arctic and Antarctic to the warm waters of the tropics. Microorganisms thrive throughout oceans, reaching depths of 11,000 m (mean depth 3200 m), with pressures exceeding 100 MPa and temperatures higher than 100 °C in deep-sea hydrothermal vents (Kennedy et al., 2010). Typically, they associate with other organisms: countless communities comprise bacteria, archaea, protists, fungi and viruses. Furthermore, single-celled and multicellular microorganisms are responsible for 98% of primary productivity in marine ecosystems, making them integral to marine food chains and specifically to carbon and energy cycles (Sogin et al., 2006). Their evolution over the past 3.5 billion years has produced a high diversity in genetic and phenotypic variation. Marine metagenomics is, therefore, an excellent tool for reading the abundance of novel genetic information and unlocking the immensity of metabolic diversity available from microorganisms.

2. Metagenomics as a tool to study the environment

Antoni van Leeuwenhoek observed single-celled organisms discovering microorganisms using microscopic lenses already in the 17th century. Now, microbes are used by industries to cultivate various livestock feeds and nutrients; they are also used as energy sources and in the production of pharmaceuticals. In the late 20th century, prokaryotic communities were typically deciphered using 16S rRNA gene sequencing from total genomic DNA extracted from the whole community. Moreover, 18S rRNA gene sequencing was introduced to translate the eukaryotic community (Rivas et al., 2004).

Of course, the sample handling, storage and choice of DNA extraction method have impact on the final multiple community genomic DNA representation. The environmental strains can vary in yield of DNA following even the same extraction protocol. The disruption of the cell depends on the cell wall structure. For example, it is known that Gram-positive bacteria have a much thicker layer (20–80 nm) of peptidoglycan that encircles the cell, compared to Gram-negative bacteria (10 nm). Additionally teichoic acid in Gram-positive bacteria stabilizes the cell wall and makes it stronger. While it is relatively simple to disrupt the Gram-negative cells (for example by bead-beating), additional freezing and thawing are used for more efficient disruption of Gram-positive bacterial envelope. Chemical and enzymatic lysis methods are used in combination with mechanical and temperature control to weaken and disrupt the cell wall (El Bali et al., 2014; Wesolowska-Andersen et al., 2014). The spores, yeast and mycobacteria are even more difficult to break, as the cell wall is more complicated than Gram-negative and Gram-positive bacteria (Vingataramin and Frost, 2015). The extraction methods are adapted to deal with the presence of possible inhibitors in the soil (such as humic acids), as well as metabolites of the cells (Grec et al., 2014). There are additional challenges with possibility to gain additional information by separately isolating intracellular and extracellular DNAs (Alawi et al., 2014). To increase the sample processing throughput, automated DNA extraction platforms are offered by several companies. An article by Oldham et al. describes a comparison of such platforms for the extraction of DNA from seawater and oil samples (Oldham et al., 2012).

While amplicon sequencing is still the dominating method, we foresee that with the development of next generation sequencing methods, the standard will be to conduct shotgun sequencing of microbial community DNA and cDNA (to access expressed genes) (Tringe et al., 2005). Shotgun metagenomic sequencing allows not only the sequencing of microbial, eukaryotic and archaeal genomes, but also of viral genomes. In 2006, Culley et al. analyzed coastal viral communities. However, assembly of sequenced reads was problematic and led to a mosaic of multiple populations (Sharon and Banfield, 2013; Culley et al., 2006). By simplifying the complex environmental population “composite genomes,” can be obtained and sequence variation can be studied (Allen and Banfield, 2005) or genome contents can be inferred (Tyson et al., 2004; Martin et al., 2006). Perhaps not surprisingly, less complex communities were obtained from samples of extreme environments such as acid mines (high acidity) and thermal geysers (high temperature) (Zarraonaindia et al., 2013). Communities can be further fractionated by cell size (filtration through defined pore size filters), DNA composition (GC% gradients), flow cytometry (fluorescence activated cell sorting), affinity purification or by particular metabolic processes (Suenaga, 2012). Recent advancements permit the isolation of single cells that can be lysed to extract genomic DNA (or RNA) followed by amplification and sequencing. Single-cell genomics is an emerging method, getting more and more popularity due to the constantly lowering sequencing costs (Kodzius and Gojobori, accepted for publication). Quake sequenced five single cells of an ammonia-oxidizing archaea Nitrosoarchaeum limnia isolated from low-salinity sediments in the San Francisco Bay (Blaine et al., 2011). By combining single-cell genomic and metagenomic approaches a high-quality draft genome revealing features unique to N. limnia’s low-salinity environment was obtained. Because in single-cell genomics, genes expressed in the cell can be decoded and complete metabolic pathways can be established in a single genome (rather than in multiple distinct genomes). Quake found that N. limnia uses a modified version of the 3-hydroxypropionate/4-hydroxybutyrate pathway for carbon fixation (Blaine et al., 2011). While computational tools can help to solve certain scientific questions, the combination of metagenomics and single-cell genomics on a sample undoubtedly provides the means to more easily assemble contigs and a clearer view of the data. For example, the combined single cell/metagenomic approach enabled Quake to test the single cell assemblies for completeness and the metagenome for contamination (Blaine et al., 2011). At present, there is a shift from a gene-centric towards a more organism-centric approach to encompass community genomics (i.e., ecogenomics or environmental genomics) (DeLong et al., 2006; Worden et al., 2006; McMahon, 2015). For example, DeLong et al. used sequence variation to analyze surface near sea-floor depths of planktonic microbial communities in the North Pacific Subtropical Gyre (DeLong et al., 2006). From the genetic material obtained, they were able to infer both phylogeny and function of the material collected, including uncultivable microbes. Knowledge gained from comparative genomic analysis of microbial communities provides insight into higher-order community organization and dynamics. McMahon recognizes the importance of generating the draft genomes from metagenomes, then pairing these reference genomes with the metatranscriptome datasets. Summarizing the current advancements, McMahon calls all these “Metagenomics 2.0” (McMahon, 2015). Interest is also growing in “high-resolution genomics”, where multiple metagenomic samples (usually from the same source) are sequenced, and sequencing reads are assembled into genes. The groups of genes that covary in abundance are thought to belong to the same genome (Mick and Sorek, 2014). For example, the covariation of gene abundance can be used to resolve a structure of a complex microbial sample. The reconstruction of whole genomes that is possible using data from metagenomics enables not only sequencing of complete genes and metabolic pathways, but also the construction of evolutionary trees. The combination of deep sequencing and bioinformatic approaches allows for metagenome-based genome recovery, even from very complex systems. For example, both bacterial phylum TM7 and phylum bacteriodetes were assembled separately into a single continuous sequence (one contig), basically as single circular chromosome that is complete genome. The relative metagenome abundance of the completely assembled TM7 and bacteriodetes is –(0.47–1.58%) (Albertsen et al., 2013). In addition, complete genomes have been obtained from organisms that constitute 1% of the community in

Please cite this article as: Kodzius, R., Gojobori, T., Marine metagenomics as a source for bioprospecting. Mar. Genomics (2015), http://dx.doi.org/10.1016/j.margen.2015.07.001
oceans, sediments and even from the adult human gut (Sharon and Banfield, 2013; Albertsen et al., 2013). Although genomes can be obtained from uncultivable cells, the average genome completeness is lower than that for cultivable cells. Rinke et al. recovered 201 partial genomes with 40–55% genome completeness, ranging from a few percent to greater than 90% (Rinke et al., 2013). The variation may be attributed to the bias of the single cell DNA amplification reaction — freshwater and marine samples yielded the highest percentages of successfully amplified genomes (up to 40%), whereas the success rates for soil samples tended to be low (<10%) (Rinke et al., 2014). Single-cell genome sequencing is a complimentary approach to metagenomics (Kodzius and Gojobori, accepted for publication; Rinke et al., 2013; Kashtan et al., 2014). We expect that single-cell genomics will improve by parallel sequencing a few of the same cells, better covering fragmented genomes. It has been suggested that genomes with possible variations be stored as reference genomes rather than as individual sequences. This would allow easier storage, handling of the data and genome comparison for the development of evolutionary trees (Kahn, 2011; Nelson et al., 2010). Because most microbes cannot be predictively cultured, rRNA phylotyping and metagenomics are favorable approaches to explain the microbial community thriving in a particular environment.

2.1. Diversity of microorganisms

Using metagenomics, the genes and pathways of both cultivable and uncultivable organisms can be discovered. Estimates from deposited total 16S rRNA sequences indicate that cultivable organisms constitute much less than 1% (12,000) (Yarza et al., 2013; Cole et al., 2014) of an estimated total microbial populations (4 million) (Quast et al., 2013). Prokaryotes may comprise 10^6 to 10^8 separate genospecies (Sleator et al., 2008). Konstantinidis recommended that prokaryotic organisms be classified (especially the uncultivable) and that their vast but finite diversity be described using metagenomics (Konstantinidis and Rosselló-Móra, 2015). There have been some studies to assess the environmental sample diversity. A study by Zachary Charlop-Powers et al. have investigated soil samples from around the world and found that there is little overlap between samples collected from different global locations (less than 3%); the strongest sequence homology was detected in samples collected in close proximity (Zhang and Moore, 2015; Charlop-Powers et al., 2015). Kashtan et al. applied metagenomics and single-cell genomics by sequencing 90 individual cell genomes on the globally abundant marine cyanobacterium Prochlorococcus in Bermuda (Kashtan et al., 2014). A cell-by-cell genomic diversity assessment evidenced the coexistence of Prochlorococcus subpopulations, where they maintained a relatively stable population size in highly mixed habitats (Kashtan et al., 2014). Using metagenomics approach, there is no need to isolate or cultivate the microorganisms. Directly isolated nucleic acids provide information on the metabolic and functional capacity of a specific microbial community (Simon and Daniel, 2011). Metatranscriptomics helps to explain which metabolic pathways and genes are expressed in a given place at a given time. In parallel, it is possible to prepare and sequence both genomic DNA and total RNA libraries, and there are several reports on metatranscriptome performed on samples taken from marine waters (Mason et al., 2012; Poretsky et al., 2009; Shi et al., 2009). Going forward, we believe that standard sample collection and processing protocols will include not only the analysis of DNA, but also of RNA, proteins and metabolites. In this way not only the taxonomic, but also functional diversity of the environmental communities, as well as the expressed genes and metabolic activity in the chosen environment can be assessed. Metagenomics goes hand in hand with next generation sequencing and high-performance supercomputing, all of which enable broad access to microorganism diversity and function (Knight et al., 2012). Metagenomics is a useful tool to explore and discover the unknown. It is believed that there are other undiscovered branches on the tree of life (Woyke and Rubin, 2014). While the RNA world may still exist in a niche of favorable conditions, recently discovered extremely large DNA viruses such as the pandoravirus and the mimivirus, which are considered to exist in a parasitic fourth domain (Claverie, 2013; Philippe et al., 2013). By mining the Global Ocean Sampling environmental sequence database, the closest relatives to Mimivirus were found in the sea (Claverie, 2013).

3. Bioprospecting—the process of discovery and commercialization of new products based on biological resources

Metagenomics is not a new tool in the science. It is already established in the scientific field, revealing secrets of the earth’s microbial communities. In addition to studies of microbial taxonomic diversity, ecology and evolution, metagenomics is an important tool for environmental resources (earth and life sciences, bioenergy, bioremediation, and biotechnology), agriculture (food production and safety), biomedicine (biomedical sciences, biotechnology, biodefense and microbial forensics), sustainability and ecology (N.R. Council, 2007).

3.1. Sequence-based screening

In a sequence-driven analysis, the PCR primers are designed based on conserved DNA sequences. The metagenomics library is cloned into vector DNA and transformed into the producer. The clones are screened for the sequence of interest (Schloss and Handelsman, 2003). Of course, such sequence-driven approach has limitation. We can expect more novel information obtained by sequencing environmental samples without prior knowledge on known sequences in databases. For example, metagenome sequencing data from bacterial communities helps to identify genes that encode for natural products. Global sequencing of samples to assess biosynthetic diversity has indicated that the biosynthetic potential of microorganisms remains untapped, particularly because a large number of producers are uncultivable (Zhang and Moore, 2015; Wilson et al., 2014). Often the potential producers are cohabitating in the closed environment. This can be observed in the marine environment, such as microbial symbionts in sponges (Zhang et al., 2015) or corals (Lema et al., 2012). Wilson et al. used single-cell- and metagenomic-based approaches to discover a group of producers known as Entotheonella. The candidate genus Entotheonella is a co-inhabitant of the chemically and microbially rich marine sponge Theonella swinhoei. The Entotheonella producer DNA was detected by PCR primers specific for genes encoding the respective pathways. The authors proved that a single member of the highly diverse microbiome of the sponge host T. swinhoei is the source of almost all the polyketides and peptides that have been isolated from this sponge (Wilson et al., 2014).

Using a metagenomic library (Fig. 1) simplifies the discovery of genes. In fact, it has facilitated the discovery of many novel enzymes and biocatalysts from uncultivable bacteria. There are two main strategies to identify and assign sequence tags to the related species (Sharpton, 2014). In binning, every metagenomic sequence is assigned to a taxonomic group. It is done through comparison to the referential data or clustering into groups based on shared characteristics (such as GC content). In assembly, the sequencing reads are assembled by possibly generating longer sequences, allowing later easier mapping and bioinformatics analysis. Here the strain or species can be identified in metagenomes using genome-specific markers, such as k-mers (Tu et al., 2014). After the sequence pre-processing and clustering, bioinformatics is used for gene prediction and for further functional annotation (Richardson and Watson, 2013). Resolving and quantifying taxonomic diversity allows to understand the biological function of the community, as the presence of specific species will indicate the function of described taxa. We envision improved bioinformatics analysis by combination of short reads (such as Illumina) with long ones (for example from Pacific Biosciences), allowing better species identification and gene prediction.
The development of commercially available compounds favors the use of large sequence databases (Sugawara et al., 2008; Kodama et al., 2012; Kryukov et al., 2012) to look for previously unknown membrane proteins, antibiotics and especially enzymes (Gabor et al., 2004; Gupta et al., 2002). Enzymes are used by industries in foods, agriculture and livestock feed, paper and leather production, textile processing, detergents, personal care products, as well as in the production of fine and bulk chemicals (DeSantis et al., 2002).

While newly discovered enzymes and other useful biomolecules can be found in metagenomic sequence reads and assembled contigs, a sequence-based approach is to design DNA primers derived from conserved regions, amplifying the variety of novel variants of proteins by PCR. Numerous genes have been found to encode for several newly discovered enzymes such as nitrile reductases, chitinases, dioxygenases, hydrogenases, glycerol dehydratases and specific enzyme-degrading compounds (Simon and Daniel, 2011). In the absence of sequencing, quantitative PCR alone can be used to analyze the diversity and abundance of certain genes. For example, Zaprasis found and quantified (per gram of soil) many previously unknown herbicide-degrading dioxygenases by quantitative PCR (Zaprasis et al., 2010). This is done by designing the degenerated primer sets from conserved regions of genes of interest (such as dioxygenase). While the quantitative PCR is a tool to analyze the gene expression of certain genes, the analysis of certain genes might well be used for diversity studies of process-associated bacteria, determining the diversity of genotypes represented in gene libraries (Zaprasis et al., 2010). Furthermore, PCR can be combined with high-throughput sequencing, which allowed Varaljay to obtain 62,000 sequences gathered into >700 clusters of environmental dimethylsulfiniopropionate demethylase (dmdA) sequences (Varaljay et al., 2010). This was all done using primers designed from the DNA of one free-living marine bacterioplankton. Because PCR products are of limited size and sequencing usually reveals only incomplete gene sequences, Iwai et al. introduced “gene-targeted metagenomics” to recover full versions of the target genes. They did this by designing probes from collected sequence information from which the 5’- and 3’-ends of a gene can be sequenced by high-throughput methods to recover many clusters and discover a large variety of genes (Iwai et al., 2010).

3.2. Assessment of metabolic potential by function-based screening

In sequence-based screening, the isolation of newly discovered enzymes is based on homological sequences, which often reveal only partial sequences: it is limited to known sequences of interest that do not allow for the discovery of information about the biochemical functions of the encoded enzymes. Activity-based assays can be performed to isolate genes that encode for novel biomolecules or biological activities. This “function-based screening” uses metagenomic libraries constructed in expression vectors to express in a chosen host; sequence information is not required. On the other hand functional screens enable the identification of novel gene classes that encode for previously unknown or known functions (Ferrer et al., 2009; Handelsman, 2005; Riesenfeld et al., 2004). However, function-driven approaches are much slower because genes must be expressed in a selected vector, where enzymes are ensured to be correctly folded. Although function-based approaches are becoming more popular, their set-up and development are time consuming and tedious; and because hit rates are low, high-throughput screening is required. Miniaturization and the application of microfluidic technologies are key to the success of genomics in the future (Wu et al., 2012, 2014; F.Q. Li et al., 2014). Reliable and sensitive enzymatic assays also require sophisticated substrates and highly sensitive analytical methods (such as high-performance liquid chromatography) to detect biomolecules. On the other hand, function-based screening permits the detection of clones that contain previously unknown, active enzymes, from which conclusions about enzyme physicochemical parameters and activities can be drawn (Rabausch et al., 2013; Parachin and Gorwa-Grauslund, 2011). Subsequently, protein engineering through in vitro evolution can be used to create enzymes with improved properties by mutational exploration and selection of the best candidates.
In addition to enzymes, compound screens have also been shown to be successful. For example, a study that combined single-cell- and metagenomic-based approaches discovered that the bacterial symbiont *Entotheonella* species expresses unique chemical bioactive compounds including bioactive polyketides and peptides (Wilson et al., 2014).

### 4. Metabolic engineering as a tool for cell factories

#### 4.1. Established production strains

Industries are using various strains of bacteria, yeast and algae to deliver valuable products such as cheese, wine and beer as well as other biotechnological and pharmaceutical products on a large scale. For this reason, cost-saving approaches have the potential for a huge impact on industries. Before the 1990s, microorganisms were genetically modified by chemicals, and the mutant strain that overexpressed the desired metabolite was used in production. However, random mutation is problematic because the modified genes or DNA location was not known, and untargeted pathways were affected (Torres and Voit, 2002). Other early technologies used the random insertion of specific genes into a living cell. This blind insertion method is similarly limited by unknown effects on unrelated pathways and subsequent unwanted effects. Any changes in cell metabolic pathways may have drastic effects on the viability of the cell (Vemuri and Aristidou, 2005). With the advancement of genome engineering and genome-editing technologies, it is now possible to remove, insert or replace specific DNA sequences in the cell. Instead of a direct deletion or an overexpression of the genes that encode the related metabolic enzymes, it is now more common to target regulatory networks. In this way, particular pathways and/or the maximum output of a desired metabolite can be enhanced or suppressed in a cell (even nonnative pathways), with the possibility of scaling up entire processes, such that the cell remains viable under production conditions.

Typically, the desired product to be synthesized is identified, and the related reactions and pathways that can produce the desired product are selected. The most suitable organism to synthesize the product is chosen from genetic and chemical information. It is important to consider how easily the organism’s pathways can be modified, how easily the organism is to grow and maintain and the volume, value and cleanliness of the final product. A wide range of microorganisms established as cell factories for industrial purposes are currently used for metabolic engineering: recombinant bacterial strains of *Escherichia coli*, filamentous fungi *Aspergillus niger* and *Aspergillus oryzae*, various microalgae (e.g., *Arthrospira*, *Botryococcus*, *Chlorella*, *Chlorella protothecoides*, *Chlorella vulgaris*, *Clamydomonas*, *Cyanobacteria*, *Dunaliella*, *Nannochloris*, *Nannochloropsis*, *Ostreococcus*, *Porphyridium*, *Porphyridium cruentum*, *Porphyridium reticulatum*, *Porphyridium scumatum* and *Synechocystis*). The availability of references and databases about the genomic and chemical information, reactions and metabolic pathways for the production of product or a wanted result is essential. Each of the producers has its own advantages and suffers from some problems. For example, *B. subtilis* is capable producing riboflavin, butanediol, isobutanol, celluolase and other substances. Production of these substances can be enhanced by gene modification (Hao et al., 2013). The genome-scale metabolic models of *B. subtilis* are very useful for the rapid and accurate prediction of the cellular response to gene knock-out, media conditions and environmental changes. It is important to be able to scale up the production from the laboratory to the industrial scale. The volume of a product produced by engineered cells varies. High-value products (and usually low volume) include pharmaceuticals (recombinant proteins, statins and other natural products) and food ingredients (vitamins, antioxidants and flavors). On the other hand, biorefineries seek larger volume products (even if the value of the product is lower) such as fine chemicals (antibiotics, enzymes and chiral building blocks), fuels and bulk chemicals (ethanol, solvents, polymer building blocks and feed additives like amino acids).

#### 4.1.1. *E. coli* as a producer

*E. coli* is a well-studied organism with a complex system of 4627 genes, 5827 regulatory interactions (including transcription initiation, transcription attenuation, regulation of translation and enzyme modulation) and 2528 unique compounds (Keseler et al., 2013; Feist et al., 2007); its metabolism is well understood. *E. coli* is an ideal candidate for both metabolic engineering and industrial-scale production of desirable bioproducts. *E. coli* encompasses excellent properties such as rapid doubling time and growth rate, facilitation of high-cell-density fermentation and low-cost production (Chen et al., 2013). *E. coli* is known for the production of various biofuels (hydrogen, bioethanol, 1-butanol, isopropanol, 1-propanol, acetate, pyruvate, lactic and succinic acids); amino acids (phenylalanine, threonine, tryptophan, tyrosine and valine); sugars and alcohols (mannitol, xylitol); diols and polymers (such as precursors for polyesters) (Chen et al., 2013). As biosynthetic pathways in microorganisms thriving in marine waters are uncovered, we will likely see *E. coli* used more and more as a chemical producer. Ongley demonstrated a high-tier heterologous production of lyngbyatoxin in *E. coli*, a protein kinase C activator from an uncultivable marine cyanobacterium (Ongley et al., 2013). The following are a few examples of companies using *E. coli* to produce high volumes of various metabolites for industrial purposes: Dupont (1,3-propanediol, butanol and ethanol from lignocellulosics); DSM (the antibiotic cephalaxin); BASF (vitamin B2 and riboflavin); Novozymes and Cargill (3-hydroxypropionic acid); and Gevo (isobutanol) (Hong, 2012; Hong and Nielsen, 2012).

#### 4.1.2. Algae as a producer

In the presence of sunlight, CO₂, water and nutrients (nitrogen, potassium and phosphorous) algae can produce an oil used for biofuels. Algae can be single-celled or multicellular; some algae have short sexual cycles and rapid growth that allows for easy manipulation, and some algae can grow in wastewater, salty water or nonarable land (Hannon et al., 2010). Algal strains vary in their requirements for growth medium, their growth rate and their biomass productivity; they are not all photosynthetic, and they have various lipid profiles and levels of resistance to pathogens (Georgianna and Mayfield, 2012). Algal strains should be optimized for tolerance to high temperatures, various light conditions, salinity, high oxygen concentration and nutrient composition. Ideally, cells should be able to grow and produce lipids simultaneously, and it would be best if the lipids produced were excreted outside cells. From the total 40,000 algal species (with some estimates of up to 72,500 algal species worldwide (Guiry, 2012)), only a few thousand strains (~3000) are kept in collections and only half-a-dozen are cultivated in industrial quantities. To date, only approximately 10 different algal species can be transformed and fewer than 50 whole genomes have been sequenced (Walker et al., 2005; Wijffels and Barbosa, 2010).

#### 4.1.3. *S. cerevisiae* as a producer

*S. cerevisiae*’s genome was the first eukaryotic organism to be sequenced, and, therefore, it has been used widely as a model organism in many fundamental studies. The yeast *S. cerevisiae* naturally metabolizes hexoses (glucose, fructose and galactose) and a few dimers such as sucrose and maltose. Yeasts can tolerate high ethanol and sugar concentrations (resistant to osmotic stress), a high cell density, high temperatures, low pH values and low sulfite concentrations (Basso et al., 2008). For these reasons yeasts are able to grow under environmental conditions where other organisms cannot survive. In addition, yeasts are immune to contamination by phages (compared to *E. coli*). However, there are certain limitations to *S. cerevisiae*, such as an inability to metabolize pentoses present in hemicelluloses, their conversion and growth rates are lower than those for *E. coli* and it is more difficult to engineer yeast than *E. coli* (Hong and Nielsen, 2012). Despite
these limitations, many strains of *S. cerevisiae* have been developed for the production of biofuels (ethanol, biodiesels and butanol); bulk chemicals (pyruvic acid, succinic acid, 1-lactic acid, 1,2-propanediol, d-ribose and ribitol and polyhydroxy-alkanoates); fine chemicals (1-ascorbic acid, antifungal diterpene, methylmalonyl-coenzyme A, nonribosomal peptides and Se-methylselenocysteine); and protein drugs (immunoglobulin G, hepatitis B virus surface antigen, epidermal growth factor, insulin-like growth factor, single-chain antibodies and glucagon) (Hong and Nielsen, 2012). Many genes discovered in marine life can be expressed in *S. cerevisiae*; for example, a gene encoding fatty acid desaturase was cloned from the marine fungus *Thraustochytrium* sp. and expressed at a very high conversion rate (56.40%) in *S. cerevisiae* (Huang et al., 2011).

### 4.2. Tools to improve the function of strains

We seek to improve wild-type producers to better suit our needs. The evolution of agriculture, for example, began around 10,000 BC, when humans transitioned from hunting to farming. Over time, crops beneficial to agricultural output have been selected. Although unaware of genomic DNA, people began to understand that genetic material is passed from parents to their offspring. Therefore, genetic material can be altered by either random mutation events, such as irradiation by X-rays, or by genome-targeted editing technologies, to make improvements to strains in a short period of time (Fig. 2).

Similarly, established microbial strains have been improved in many ways—from random strain mutations to advanced methods, such as metabolic engineering, when specific genes and pathways are changed. Advances in molecular biology and genetic sciences have developed genome engineering, which enables targeted mutations in genomes. In 2009, Church automated a process called multiplex-automated genome engineering, which generated a diverse set of genetic changes (mismatches, deletions and/or insertions) that enabled *E. coli* to be engineered to overproduce (5-fold) the normal amount of isoprenoid lycopene, an antioxidant linked to anticancer properties that is produced industrially. His work was completed in three days using a complex pool of synthetic DNA, creating over 4.3 billion combinatorial genomic variants (Wang et al., 2009). The accelerated growth of *E. coli* was integral to the speed at which Church was able to improve and discover new properties (Wang et al., 2009).

Producers in marine waters have been discovered and improved. For example, a marine actinomycete *Marinactinospora thermotolerans* found in deep-sea marine sediment was genetically modified to increase its yield by ~25-fold compared to the wild-type strain for its use in the production of antimicrobial nucleoside antibiotic A201A. *M. thermotolerans* was also shown to be a source of antibacterial agents (Zhu et al., 2012).

#### 4.2.1. Genetic engineering using molecular biology tools

It is known that many of the organisms identified through metagenomics are unculturable. Advances in metagenomics and microbiology allow to identify new media, expanding the list of potential producers. Genomic mining of the metagenomics data allows to identify biochemical diversity of specific gene families. For example, 31 cyanobactin gene clusters (family of cyclic ribosomal peptides) from 126 genomes of cyanobacteria were identified by Leikoski et al. (2013). The gene families and pathways found can be inserted into the known producer strains.

#### 4.2.1.1. Meganucleases as a tool for genome engineering. Meganucleases are enzymes capable of recognizing and cutting large DNA sequences (from 12 to 40 bp); they provide a more sophisticated approach to genome engineering than do random mutation techniques (Stoddard, 2005). However, naturally occurring meganucleases do not serve for the purpose of engineering because it is impossible to match an exact meganuclease to act on a specific DNA sequence. Therefore, many meganucleases are created (Smith et al., 2006) by either by mutagenesis

![Fig. 2. The typical steps followed to improve a producer strain using genetic engineering tools to modify genomic DNA by mutagenesis and pathway perturbation, followed by high-throughput screening. The best producer is used to scale up the production process.](http://dx.doi.org/10.1016/j.margen.2015.07.001)
(Seligman et al., 2002) or by combinatorial assembly, whereby protein subunits from different enzymes are fused (Arnould et al., 2006). Currently, a fully rational design process (“directed nuclease editor”) is being used to create highly specific engineered meganucleases for any location in a genome (Gao et al., 2010). The more stringent the genome sequence recognition is of the meganuclease, the less toxic it is for the cell.

Actinomycetes are Gram-positive mycelial bacteria, known to produce a wide variety of industrially and medically relevant compounds (antibiotics, chemotherapeutics, fungicides, herbicides and immunosuppressants). Using meganuclease I-Scel from S. cerevisiae, the gene cluster for the red-pigmented tripyrrole compound that remains associated with the mycelium was deleted in actinomycetes Streptomyces coelicolor. The changed genotype and phenotype were confirmed (Fernandez-Martinez and Bibb, 2014).

4.2.1.4. CRISPR as a tool for genome engineering. Zinc finger nucleases (ZFNs) and transcription activator-like effectors (TALEs) (Baker, 2012) are two DNA sequence-recognizing peptides that can be combined with a FokI endonuclease catalytic domain, which have different recognition and cleaving sites. As a result, a construct with multiple recognition domains for recognizing multiple composite sequences over 24 bp can be created. Although ZFN (Kim et al., 1996) and TALEN (Li et al., 2011; Christian et al., 2010) are time consuming and expensive technologies, they have been used successfully to genetically engineer many species, validating their role as protein engineers.

TALEN has successfully been applied to genetically modified sea anemone Nematostella vectensis (Ikmi et al., 2014), sea urchin embryos (Hosoi et al., 2014), amphioxus Branchiostoma belcheri (G. Li et al., 2014), marine annelid Platynereis dumerilii (Bannister et al., 2014) and ascidian Ciona intestinalis (Yoshida et al., 2014).

4.2.1.3. RE-TFOs as a tool for genome engineering. However, ZNF and TALEN, are limited by the fact that their constructs must be re-engineered for each new DNA target. On one hand, restriction enzyme triple-helix-forming oligonucleotide (RE-TFO) conjugates offer an easier approach to performing genome engineering (Silanskas et al., 2012; Eisenschmidt et al., 2005), where only DNA triple-helix-forming nucleotides are used for binding specificity (no protein engineering is required). On the other hand, RE-TFO conjugates are restricted by the complicated methodology required to conjugate oligonucleotides to the nuclease module, by the slow formation of the triple helix and by the low-targeting frequency.

4.2.1.4. CRISPR as a tool for genome engineering. The latest development in genome-editing technology is the CRISPR/Cas system. CRISPR is an RNA-based bacterial defense mechanism (in other words bacterial and archaeal immune systems) designed to recognize and eliminate foreign DNA from invading bacteriophages and plasmids (Horvath and Barrangou, 2010). RNA-guided DNA endonuclease use guide RNA for target-site recognition, and Cas endonuclease cleaves the DNA. The main advantage to using CRISPR is the speed and simplicity of this assay — no protein engineering is required and only synthesized RNA is used. Multiplexing is possible, which would allow for more changes in the genome with one experiment (Cong et al., 2013; Gasananas and Siksnyis, 2013). The CRISPR approach has shown high potential and genome-editing flexibility that can be applied to numerous DNA targeting applications including transcriptional control (Barrangou and Marraffini, 2014). To date, CRISPR has been primarily used in the engineering of eukaryotic cell lines; however, we can expect more applications for genetic engineering in microorganisms in the near future. Although there is evidence that CRISPR-mediated cleavage of chromosomes leads to cell death in many bacteria (Marraffini and Sontheimer, 2010; Edgar and Qinmon, 2010; Bikard et al., 2012) and archaea (Gudbergsdottir et al., 2011; Fischer et al., 2012), precise mutations have been successfully introduced in the genomes of Streptococcus pneumoniae and E. coli using CRISPR system (Jiang et al., 2013). Currently, targeted genome editing by CRISPR can be applied to microorganisms harboring inducible plasmid (Jiang et al., 2013).

Today, the CRISPR system is recognized as the most efficient and easiest system for genome engineering. CRISPR should be used to create and modify producer strains and to improve and enhance the production of a chemical of interest. A high-throughput system environment is essential for CRISPR system function, and combined with single-cell technologies, for example, the best performing producers could be quickly and easily selected. In the near future, we expect these methods to lead to more producer strains for various industrially relevant output products. Marine metagenomics is not only a source to explain bacterial and archaeal CRISPR-related gene sequences and their acquired resistance to viruses (Heidelberg et al., 2009; Anderson et al., 2011), but it is also a source for identifying ways to improve parts of the CRISPR system (Anderson et al., 2011; Zhang et al., 2014; Smedile et al., 2013); new producer strains can be obtained from marine sources and modified/improved using CRISPR. CRISPR/Cas systems can contain genes that encode highly divergent proteins (Makarova et al., 2011). All these are a valuable source for the synthetic biology. There are at least two studies which presented CRISPR-based transcriptional cascades for synthetic circuits, such as transcriptional activators and repressors (Vogt, 2014). Voltage-dependent changes led to the change to fluorescence resonance energy transfer (FRET) signals or changes to endogenous protein fluorescence. While this was done in mammalian cells, it is also applicable for transfer to prokaryotes or other cells. The shift from observation to application is beginning to take place.

4.2.2. New organisms created by synthetic biology

Synthetic biology opens the door to a new era of producer improvement that could even lead to the creation of new organisms (Channon et al., 2008). Advances in oligonucleotide synthesis, where longer pieces are being produced, are enabling the recreation of entire genomic DNA from certain cells. In addition to the development of pathways and components, biological parts can be standardized (Baker et al., 2006), and the functional units are being introduced into organisms, with the potential to construct entire organisms de novo.

In 2000, a 9.6 kbp hepatitis C virus genome was synthesized (Bligt et al., 2000). Two years of work later, a synthetic 7.7 kbp poliovirus genome was developed (Covzin, 2002). The next year, the synthesis and construction of a 5.4 kbp bacteriophage Phi X 174 genome took only two weeks (Smith et al., 2003). In 2006, the J. Craig Venter Institute constructed a synthetic genome of a newly discovered minimal bacterium Mycoplasma laboratorium (Giibson et al., 2008). In 2010, Venter demonstrated the synthetic assembly of a 1.08-M-bp Mycoplasma mycoides genome (Giibson et al., 2010). Eukaryotic algal chromosomes of up to 500 kb have been assembled in yeast at the Craig Venter Institute (Karas et al., 2013). It is likely that fully synthetic eukaryotic producers will soon be synthesized and assembled. Other developments include expanding the natural genetic code (which consists of four bases A, C, G and T) to an extra base pair (denoted by X and Y artificial nucleotides) (Malychev et al., 2014). The production of new life forms and new proteins encoded from previously unknown and from developed amino acids is expected in the future (Suzuki et al., 2001; Nakamura et al., 2000; Suzuki and Gojobori, 1999).

The first synthetic cell created by Venter cost 40 million USD (Sleator, 2010). However, as DNA synthesis and sequencing technologies improve and prices fall, we can expect not only further discovery of naturally occurring organisms, but also the development of more engineered microorganisms. Many companies are involved in creating synthetic cells to capture CO2 and produce renewable fuels such as Joule Unlimited in Cambridge, LS9 Inc. in San Francisco, Amyris Biotechnologies in California, Synthetic Genetics Inc. in California (a company founded by Dr. Venter) and the Exxon Mobil Corporation in Texas. Synthetic biology serves to benefit from marine metagenomics because
lots of the components for synthetic biology are derived from marine microorganisms (Wang and Mattheos, 2009).

5. Conclusion/outlook

With the constant advancement of sequencing technologies, delivering longer reads at lower costs, we can expect a flood of information about the genes, pathways and whole genomes of microorganisms in various environmental populations. The limiting step will be reaching extreme locations to collect some of the most specially adapted organisms to discover not only new genomes, but also potentially undiscovered branches on the tree of life (Woyke and Rubin, 2014).

High-throughput sequencing utilizing microfluidic droplet technology will allow us to select the best enzymes and best biomolecular producers. Many opportunities remain to catalog genomes and pathways for the simplified selection of necessary parts for genetic engineering. Researchers at the Massachusetts Institute of Technology recently presented CRISPR-based biological parts to perform logical functions for synthetic biology (Rusk, 2014). Both the CRISPR-based transcription activators (Nissin et al., 2014) and repressors (Kiani et al., 2014) were designed and demonstrated to be successful for synthetic circuits in mammalian cells; very likely this technology will soon be applied to microbial cells. Both the standardization of the sample collection technique and bioinformatics pipeline will speed up data production and analysis (Table 1.). Large quantities of data are expected to become a challenge for bioinformatics; solutions so far include storing reference genomes with their possible variations (Kahn, 2011; Nelson et al., 2010).

We will see more synthetic organisms using environmental pollutants (e.g., CO₂) or industrial waste to produce food, livestock feed, fuels, small molecules and pharmaceuticals. Considerable interest from industry is expected to speed up solutions for scaling up production. Interest from both industry and academia promises to move developments in marine metagenomics along at an impressive pace.

References


