From bacterial genomics to metagenomics: concept, tools and recent advances

Pooja Sharma · Hansi Kumari · Mukesh Kumar · Mansi Verma · Kirti Kumari · Shweta Malhotra · Jitendra Khurana · Rup Lal

Received: 22 January 2008 / Accepted: 23 February 2008

Abstract In the last 20 years, the applications of genomics tools have completely transformed the field of microbial research. This has primarily happened due to revolution in sequencing technologies that have become available today. This review therefore, first describes the discoveries, upgradation and automation of sequencing techniques in a chronological order, followed by a brief discussion on microbial genomics. Some of the recently sequenced bacterial genomes are described to explain how complete genome data is now being used to derive interesting findings. Apart from the genomics of individual microbes, the study of unculturable microbiota from different environments is increasingly gaining importance. The second section is thus dedicated to the concept of metagenomics describing environmental DNA isolation, metagenomic library

construction and screening methods to look for novel and potentially important genes, enzymes and biomolecules. It also deals with the pioneering studies in the area of metagenomics that are offering new insights into the previously unappreciated microbial world.

Keywords Genomics · Metagenomics

Genomics

Genomics as a scientific discipline emerged almost three decades ago. The term genome refers to the complete set of genes of an organism while genomics deals with mapping, sequencing and analysis of the genomes. As scientists became excited by the possibility of sequencing complete genome, there was a boom in the genome sequencing projects that were undertaken across the globe. The first genome to be completely sequenced was that of bacteriophage Φ -X174 in 1977 [1] which was just 5,368 bp in size. Advances in sequencing technology (automation and cost reduction) made the sequencing of bigger genomes possible by the 1980s. In 1995, the first complete microbial genome of *Haemophilus influenzae* (1.8 Mb) was sequenced [2], followed by the sequencing of the complete genome of the bacterium, *Mycoplasma genitalium* in the same year [3]. One of the most exciting developments in genomics has been the completion of the rough draft of the human genome by the Human Genome Project in 2001.

Since microorganisms are the most abundant organisms present in nature, it is not a surprise that the majority of the completed genomes are that of the microbes. While writing this article, 542 bacterial genomes were completely

P. Sharma^{1§} · H. Kumari^{1§} · M. Kumar¹ · M. Verma¹ · K. Kumari¹ · S. Malhotra¹ · J. Khurana² · Rup Lal¹ (✉)

¹Department of Zoology,
University of Delhi,
Delhi - 110 007, India

²Department of Plant and Molecular Biology,
South Campus,
Benito Juarez Marg,
University of Delhi,
Delhi - 110 021, India

[§]The authors have contributed equally to the work

Tel: +91 / 11 / 27666254, Fax: +91 / 11 / 27666254,
E-mail: duzdel@vsnl.com

sequenced and 1343 microbial genomes were in various stages of genome sequencing. Ever since the sequencing of the first bacterial genome in 1995, a great deal of sequence data from microbial genomes and from other organisms has become available. Information regarding complete and ongoing genome sequencing projects is available at <http://www.genomesonline.org>.

It is a well known fact that a vast majority of microorganisms are unculturable by standard laboratory techniques. Recent sequencing technologies have not only facilitated the complete genome sequencing of organisms, but have also enabled the sequencing and analysis of large fragments of DNA from uncultivable microorganisms from a wide variety of niches. The study of genomes from environmental samples has proved valuable for the analysis of microorganisms that are difficult to culture under laboratory conditions. This relatively new field of science called 'metagenomics' offers huge scope and potential for both biotechnological development and basic microbiology.

In this article we will address the aspects of genomics and metagenomics that now represent a major breakthrough in the area of molecular biology. Although our focus during discussion remains on microbial genomics and metagenomics, the underlying tools that have made entire discipline so simple and exciting are also discussed.

DNA sequencing technology: Revolution that continues

DNA sequencing methods are the building blocks in the field of genomics and as evident from the discussion that follows, these techniques have undergone a lot of advances over the past few decades and are still being improved. Before mid 1970s, there was no method by which DNA could be directly sequenced. In 1977, two methods were independently developed for the sequencing of DNA and this remarkable breakthrough earned Nobel Prize for the two group leaders, Professors Walter Gilbert (Harvard

Fig. 1. Maxam Gilbert manual sequencing scheme.

University, USA) and Frederick Sanger (Cambridge University, Great Britain) in 1980.

Allan Maxam and Walter Gilbert [4, 5] developed the chemical cleavage method for the sequencing of DNA. In this method, DNA (either double-stranded or single-stranded) is radioactively labeled at one end of each strand. If double-stranded DNA is used, then the radioactive label must be removed from one end prior to sequencing. The labeled DNA is then separated into four different reactions and treated with chemicals to cleave the DNA at one or two of the four nucleotide bases in each of four reactions (G - dimethyl sulphate or piperidine, A+G - dimethyl sulphate and piperidine in formic acid, C - hydrazine and piperidine in 1.5N NaCl, C+T - hydrazine and piperidine). As a result, a series of radioactively labeled DNA fragments are generated, which are then separated by polyacrylamide gel electrophoresis (on the basis of size), with the four reactions arranged alongside each other. The gel is exposed to X-ray film and the autoradiograph of the gel reveals a pattern of bands from which the 5' to 3' DNA sequence can be read from bottom to top (Fig 1).

At about the same time as Maxam and Gilbert DNA sequencing was being developed, Frederick Sanger and his colleagues [6] devised a different method for the sequencing of DNA called the chain termination method or dideoxy sequencing. This method utilizes dideoxynucleotides (ddNTP's) which are different from deoxynucleotides as they contain a hydrogen atom on the 3' carbon instead of a hydroxyl group (Fig 2). When dideoxynucleotides are incorporated into a sequence, they are unable to form a phosphodiester bond with the next incoming deoxynucleotide which results in the termination of the DNA chain. To perform sequencing by this method all that is required is a DNA template (single-stranded), a radiolabeled DNA primer, a DNA polymerase, deoxynucleotides (dATP, dGTP, dCTP, or dTTP) and dideoxynucleotides (ddATP, ddGTP, ddCTP, ddTTP). The reactions are carried out in four sepa-

rate tubes and only one of the four dideoxynucleotides is added to each reaction. During DNA synthesis, when a dideoxynucleotide is incorporated into the chain instead of a deoxynucleotide, the DNA chain elongation events are terminated. After all the reactions are completed, DNA fragments of varying lengths are produced all ending in the same dideoxynucleotide in one reaction which are then size-separated by polyacrylamide gel electrophoresis, with each of the four reactions arranged next to each other. The gel is then autoradiographed to read the sequence of the nucleotides. Unlike Maxam Gilbert fragments, each lane is base specific making the base calling easier (Fig 3).

Although it was possible to sequence at least 100 nucleotides per reaction using these techniques, Sanger's method was preferred over the chemical sequencing method of Maxam and Gilbert due to its efficiency and use of fewer hazardous chemicals. At that time, it was possible to determine the sequence of ~5,000 bases in a week.

The next significant progression in sequencing technology was the development of instruments which could, with the help of computers, automate the DNA sequencing process and facilitate the sequencing of more DNA in a short span of time. Automated sequencing is based on the same principle of Sanger's DNA sequencing method but makes use of fluorescent labels, laser-induced fluorescence detection and computerized base calling system instead of radioactive labels, autoradiography and manual base calling system that is employed in manual sequencing which is a tedious and an error prone method.

In 1986, Leroy Hood and his colleagues attached different fluorescent dyes to the sequencing primers [7]. A different fluorescent dye was attached to the primers for each of the reactions specific for the bases A, T, C and G. This method is known as 'dye-primer sequencing'. The use of four different fluorescent dyes signifies that, unlike traditional DNA sequencing, all four sequencing reactions can be loaded on a single gel lane. Alternatively, fluorescent

Fig. 2. Difference between deoxyribose and dideoxyribose.

5' pCpApTpApGpCpTpCpGpApTpGpGpCpCpGpCpApTpGpCpCpGpTpApG 3'

Fig. 3. Sanger sequencing scheme.

dyes can also be attached to the dideoxynucleotides instead of the sequencing primers. This meant that the sequencing reaction could be carried out in a single tube with all four terminators present and loaded on just one lane [8]. This approach is known as 'dye-terminator sequencing'. The first "dye set" was based on succinylfluorescein compound. Slight shifts in the emission wavelengths of the dyes were achieved by changing the side groups. The dyes SF505, SF512, SF519, and SF526 were attached to dideoxy terminators ddG, ddA, ddC, and ddT respectively. All four dye labeled terminators could be excited by an argon ion laser at 488nm and each would produce a peak emission that could be distinguished by the detector. A four-color laser induced fluorescence detector examines the DNA as it travels to the bottom of the gel and the color of each dye-labeled fragment is then figured out by the computer as a specific base.

An automated DNA sequencer, ABI Model 370 was built by Leroy Hood and Mike Hunkapiller from Applied

Biosystems, Inc. (ABI) and commercialized in 1987. Using the automated sequencer, it was possible to read over 350 bases per lane. Although subsequent improvements in automated sequencing apparatus like introduction of more sensitive fluorescent dyes, increased gel capacity and refinements in computational hardware and software facilitated considerable enhancements in DNA sequencing throughput, they were still slab gel-based systems and one had to pour the slab gels, mount them on the instrument and load the sequencing reactions prior to the run which was cumbersome, time consuming and error prone.

In 1990, Swerdlow and Gesteland (1990) [9] described the use of polyacrylamide gel-filled capillaries to obtain DNA sequences avoiding the tedious slab-gel based method. The first single capillary sequencer (The ABI PRISM® 310 Genetic Analyzer) was introduced in 1995. Over the years, capillary sequencers containing 4, 16, 48, 96 and 384 capillaries have been developed. With the introduction of capillary DNA sequencers, it became possible to obtain

high quality DNA sequence with long read length (500–1000 bases) in a matter of a few hours. Over 2.8 million bases of sequence data can be generated by using the 384-capillary sequencer (MegaBACE™ 4000 DNA Analysis System) in 24 hours. Currently, a 1024-capillary sequencer called the Monster CAE is being developed at Stanford, USA in collaboration with UC Berkeley, US.

The demand for faster, easier, affordable, high throughput, highly accurate DNA sequencing has increased considerably ever since the importance of sequence information has been recognized by scientists. Sanger DNA sequencing method has been employed for a large proportion of DNA sequencing during the past three decades. Even though enormous effort has gone into developing and refining the DNA sequencing technology to reduce cost and to speed up the sequencing process, this technique faces limitations in terms of cost, time and throughput for future applications. A number of alternative DNA sequencing methods including sequencing by hybridization [10,11,12], pyrosequencing [13, 14], parallel signature sequencing based on ligation

and cleavage [15] and multiplex polony sequencing [16] have been developed. While most of them are at preliminary stages of development, pyrosequencing has already taken off very well thus partly replacing or assisting other traditional tools and technologies. This method is briefly described below as it has already revolutionized the field genomics.

Pyrosequencing is a recently developed DNA sequencing method that is based on the “sequencing by synthesis” principle, where sequencing of a single strand of DNA occurs during the synthesis of its complementary strand. In this method, genomic DNA is sheared into smaller fragments (300–500 base pairs) and short adaptors are ligated, which serve as complementary sequences for primers. The DNA fragments are attached onto emulsion beads under condition where only one fragment is entrapped over one bead. The emulsion PCR is carried out for amplification of single fragment per bead and then the amplified fragments are denatured. These beads containing multiple copies of the same DNA fragment are filled into optic fiber well plate such that

Fig. 4. Sample preparation for pyrosequencing.

only one bead fills one well at a time due to their size compatibility. The wells are also filled with smaller beads which are immobilized with enzymes (DNA polymerase, ATP sulfurylase, luciferase and apyrase) required for chemiluminescent reaction. These steps of genomic DNA preparation for pyrosequencing are depicted in Fig 4. The fluidic assembly (Fig 5) of the instrument provides dNTPs one by one through convection flow. The wells where complementary base is present in template strand incorporate these dNTPs leading to generation of PPi which produce light through a set of enzymatic reactions. Four enzymatic reactions are repeated to perform the sequencing by this method. In the first step, the incorporation of one of the four nucleotides onto the DNA template results in the release of an inorganic pyrophosphate (PPi) molecule; the released inorganic pyrophosphate (PPi) is then quantitatively converted to ATP by ATP sulfurylase in the presence of adenosine 5' phosphosulfate (APS); the ATP generated is used in the luciferase-mediated conversion of luciferin to oxyluciferin that generates visible light in amounts that are proportional to the amount of ATP. Finally, the unincorporated nucleotides and ATP are degraded by the apyrase enzyme before the next nucleotide

in the cycle is added to the reaction mixture. The reactions occurring inside the well are diagrammatically shown in a flowchart (Fig 6).

The major advantage of the first generation instrument Genome Sequencer 20 based on pyrosequencing approach is that 0.2 million sequencing reactions can be carried out parallelly with the average read length of 100bp. Accordingly, 20 million bases of sequence data can be generated in a single run of four hours making it the fastest sequencing method known till date. Besides, this technique does not require any cloning strategy and directly involves emulsion PCR for amplification of DNA fragments. Pyrosequencing is best utilized for sequencing of regions exhibiting strong secondary structures and also lacks any cloning bias as observed in shot gun sequencing.

The sequence reads obtained from pyrosequencing are assembled using 454 life science corporation Newbler assembler. This software assembles reads into unordered and unoriented contigs that cover almost > 99% of the non-repeat region of the genome. The major limitation of the sequence data used for assembly is the lack of paired end reads. Recently a new technique of generation of paired end reads library using type II restriction enzyme has been discovered. In this technique, genomic DNA is fragmented to average size of 2.5Kb, methylated to prevent EcoRI sites with EcoRI methylase and end repaired. Finally an adaptor DNA oligo is ligated to both ends of the DNA fragment and restriction digested with EcoRI and then ligated to form circular fragments. The circular fragments are digested with MmeI which cleaves 20 nucleotides away from two sites present in adaptor DNA and generates fragments with adaptor DNA in the middle flanked by 20 bp sequences of genomic DNA which were originally 2.5Kb apart in genome sequence. Finally these fragments are purified using streptavidin beads and processed through normal steps of pyrosequencing. The whole approach of preparation of paired end library has been well described in Fig 7. Recently, this approach has been applied by Smith *et al.*, (2007) [17] to determine the DNA sequence of *Acinetobacter baumannii*.

As with every technique, pyrosequencing is also not free from limitations. The read length is just 100bp when compared with average read length of 500bp in case of Sanger sequencing and no paired end reads are obtained which makes the process of genome assembly more complicated. During pyrosequencing, the homopolymer region (region where a stretch of same base is present in template DNA molecule) is generally prone to errors. Since the read length is only 100bp, the technique can not be well utilized on repeat rich genomes as the assembly of such genomes is not possible using short read length data. However, pyrosequencing has a wider scope and this can be judged

Fig. 5. Pyrosequencing instrument. The sequencing instrument consists of three major subsystems: (a) a fluidic assembly, (b) a flow chamber that includes the well-containing fibre-optic slide, (c) a CCD camera-based imaging assembly and a computer that provides the necessary user interface and instrument control.

Fig. 6. Illustration of the sequencing-by-synthesis process.

from the number of bacterial genomes that were sequenced by using this technique in the very first year of its invention. The prototype of the technique has been confirmed on genome sequencing of *Mycoplasma genitalium* [18] and the complex genome of barley [19]. Pyrosequencing is also being used in several other areas such as identification of small non-coding RNAs [20] and in EST sequencing and transcriptome sequencing analysis [21]. The technique has immense potential and has opened a new gateway for metagenomics. Using pyrosequencing, genomic potential of deep mine microbial ecology has been explored [22] and relationship between obesity and diversity of gut microflora has been understood [23]. Pyrosequencing has also moved from metagenomics to paleogenomics as large scale sequencing of mammoth DNA [24] and one million base pair genomic DNA of Neanderthal man [25] has been achieved.

In order to further improve in terms of read length and technology throughput, second generation pyrosequencing instrument Genome Sequencer FLX has been developed. Genome Sequencer FLX provides average read length of 200–300 bp and gives almost double throughput of 0.4 million reads as compared to 0.2 million reads as provided from first generation Genome Sequencer 20 instrument. The limitation imposed from short read length has been greatly overcome by improving it from 100bp to 200–300

bp but efforts are still being made to further increase read length to 500bp.

Approaches for Complete Genome Sequencing

There are two main approaches of complete genome sequencing namely hierarchical and shotgun sequencing. The former is a slow and systematic method where a crude physical map of the genome is first constructed before the actual sequencing to later assist in the assembly of the whole genome data. It involves constructing a Bacterial Artificial Chromosome (BAC) library of the complete genome first and assigning identification tags to the BAC clones to determine the order of the fragments. Each BAC clone is further sub cloned into smaller plasmid vectors and the whole library is sequenced from one end generating millions of sequences. The sequences are then assembled to get the sequence of each BAC clone which is finally assembled to represent the whole chromosome/ genome. The second approach of shotgun sequencing is quite fast and goes straight to the sequencing step bypassing the cloning and physical mapping of the genome prior to the sequencing. The whole genome here is randomly shredded into small fragments and libraries of 2–10kb fragments are made in suitable plasmid vectors. The libraries are then sequenced from both

the ends. The sequences are then assembled on the basis of overlaps. Though this method is very fast compared to the first strategy, it really becomes difficult to assemble the complete genome sequence data owing to the repetitive sequences present in the genome.

Before, both the approaches of genome sequencing were based on Sanger's DNA sequencing method. The pyrosequencing technique as discussed above falls into the shotgun sequencing approach without utilizing the Sanger's DNA sequencing method. Although both the traditional Sanger's method and the recently developed pyrosequencing technique have their own advantages and limitations, a consolidated strategy using both Sanger and pyrosequencing could be very successful in sequencing the whole genomes more accurately and faster than by using either of the two techniques alone. Some recent studies using Sanger and pyrosequencing hybrid approach for doing genome sequencing of six marine microbial genomes has highlighted the importance of hybrid approach to reduce the cost and time effort for genome sequencing [26]. A decision tree for using hybrid approach (Sanger's sequencing and pyrosequencing together) based complete sequencing of microbial genomes is shown in Fig 8.

Genome assembly and annotation

The outcome of different sequencing techniques is the generation of sequence reads which need to be assigned to their exact position in the genome. One of the most difficult tasks in genome sequence analysis is genome assembly. Genome assembly is a process wherein a large number of sequences (generated during genome sequencing) are assembled together to generate a representation of the original chromosomes from which the DNA originated. In order to assemble DNA sequence data, a large number of genome assembly programs have been developed like PCAP (Parallel Contig Assembly Program) [27], GAP (Genome Assembly Program) [28], Celera Assembler [29], ARACHNE [30], RePS (repeat-masked Phrap with scaffolding) [31], JAZZ [32], Phusion [33], TIGR (The Institute for Genomic Research) Assembler [34], AMOS (A Modular, Open-Source assembler) (<http://amos.sourceforge.net>) and Phred, Phrap, and Consed-Autofinish. (sequence assembly programs that can be used together or separately, although the combination is highly recommended) (<http://www.phrap.org>).

After assembly, annotation is the process of adding biological information to the raw DNA sequence. It includes

Fig. 7. Illustration of the steps required to generate a paired end library for sequencing.

Fig. 8. Hybrid approach of genome sequencing.

two sequential processes i.e., structural annotation and functional annotation. Structural annotation refers to the identification of open reading frames (ORFs) (or hypothetical genes) in a DNA sequence by means of computational gene discovery software tools. Functional annotation deals with assigning function to the predicted genes by similarity searches against genes of known functions in the database. Some of the annotation tools include AMiGene (Annotation of Microbial Genes) (<http://www.genoscope.cns.fr/agc/tools/amigene/index.html>) [35], GeneMark (<http://opal.biology.gatech.edu/GeneMark/>) [36], GLIMMER (Gene Locator and Interpolated Markov ModelER) (<http://www.tigr.org/softlab>) [37], ORF Finder (Open Reading Frame Finder) (<http://www.ncbi.nlm.nih.gov/gorf/gorf.html>).

The entire process of sequencing, assembly and annotations generates enormous data that act as a storehouse of information which can be used later for various studies and comparisons.

Early years of Genome Sequencing

Till 1990, complete genomes of only a few viruses with relatively small length of genetic codes had been sequenced. Many ambitious sequencing proposals were initiated in

the following years foremost being the introduction of the Human Genome Project (HGP) in 1990 by Department of Energy (DOE) and National Institute of Health (NIH) of US. To meet the challenge of sequencing the ultimate target, the human genome, the genome sequencing projects of smaller model organisms such as *Escherichia coli*, *Mycoplasma capricolum*, and *Caenorhabditis elegans* were initiated in the hope that these would make human genome sequencing easier by increasing the efficiency of sequencing. At that time it was expected that *E. coli* would be the first genome to be sequenced entirely but in 1995, another group led by J. Craig Venter from the Institute for Genomic Research (TIGR) and nobel laureate Hamilton Smith of Johns Hopkins University, sequenced the 1.8 million bp genome of a bacterium, *Haemophilus influenzae* Rd. with new computational methods developed at TIGR's facility in Gaithersburg, Maryland. It became the first genome of a self-replicating, free-living organism to be sequenced [2]. In the same year, the genome of smallest known bacterium *Mycoplasma genitalium* (580,000 bp) [3] associated with reproductive-tract infections was also sequenced by TIGR in a period of eight months (Until the discovery of Nanoarchaeum in 2002, *Mycoplasma genitalium* was considered to be the organism with the smallest genome). In 1996, TIGR subsequently published the first archeal genome of

Methanococcus jannaschii [38]. An analysis of this microbe's genome supported the idea that the archaea represent a third "domain" of life. Not only was this the first member of the archaea to be sequenced, but it was also the first "extremophile" to be sequenced. Subsequently genome sequence of a sulfur-metabolizing organism, *Archaeoglobus fulgidus* [39], a pathogen involved in peptic ulcer disease, *Helicobacter pylori* [40], and a spirochaete of Lyme disease, *Borrelia burgdorferi* [41] were also published by TIGR. The dramatic leadership of TIGR in the field of genome sequencing was finally paralleled with the completion of two large genomes of the bacterium *E. coli* K-12 (4.6 Mb) [42] and the yeast, *S. Cerevisiae* (12.5 Mb) [43] by DOE and Goffeau's European collaboration respectively. The yeast genome was completed with the remarkable international collaboration of more than 600 scientists from over 100 laboratories. *Escherichia coli* K-12 genome though smaller in size was important as it is the most studied organism of molecular biology and is universally adopted for fundamental work in biochemistry, genetics and physiology. In June 1998, complete genome sequence of a medically important organism, *Mycobacterium tuberculosis* (4.4 Mb) was published [44]. The genome sequence of this organism is being studied to develop better therapies, treatment and vaccines against tuberculosis.

The early years of genome sequencing reveal the tough competition and very high ambitions of the scientists that started the microbial genomics era. With the revolution in the sequencing technology, the number of complete microbial genomes sharply rose in the following years. In this article only a few genomes have been discussed and for complete information readers are referred to http://www.ncbi.nlm.nih.gov/genomes/static/eub_g.html. In the following discussion a brief account of some of the recently sequenced genomes is given.

Bacterial Genomics: Scenario in 2007

In the year 2007 alone, 151 bacterial genomes had been completely sequenced by 5th October and scores of new are being added to the list every passing week. The speed at which the database is increasing suggests that at a particular time, thousands of the genomes must be in the various stages of completion across the world. Some institutes like Department of Energy's (DOE) Joint Genome Institute (JGI), TIGR, Sanger's Institute, and J. Craig Venter Institute have taken up the genome sequencing at a mass scale and account for more than 60% of all the genomes sequenced in the year 2007.

The long list of microbes sequenced in 2007 includes the microbes of medical, industrial, agricultural and

environmental concerns. Many microbes implicated in human and animal diseases have been favorite targets for genome sequencing as it has a direct implication on human health and well being. Few of the recently sequenced ones are plague causing *Yersinia pestis* CA88-4125, opportunistic human pathogen *Acinetobacter baumannii*, *Serratia proteamaculans* 568 associated with a serious form of pneumonia in humans, *Shigella dysenteriae* Sd197, a human-specific pathogen that causes endemic dysentery, *Streptococcus pyogenes* str. Manfredo, the strain associated with rheumatic fever and the causative agent of Carrion's disease *Bartonella bacilliformis* KC583.

Some agriculturally important microbes like *Pseudomonas stutzeri* A1501, a nitrogen-fixing organism used in rice cultivation; *Sinorhizobium medicae* WSM419, a soil bacterium capable of forming nitrogen-fixing nodules; *Bacillus amyloliquefaciens* FZB42, a plant-growth-promoting rhizosphere bacterium and many more are the organisms of choice for the complete genome sequencing. The information generated by complete sequencing of these microbes would help improve the agricultural yields to meet the food requirements of the growing population.

Of recently, the growing concern over the rising environmental pollution has led to the complete sequencing of a number of bacteria involved in the degradation of many serious pollutants to get an insight on the molecular pathways of microbial degradation. Few of them sequenced in early 2007 are pentachlorophenol degrading *Pseudomonas mendocina* ymp; hexahydro-1, 3, 5-trinitro-1, 3, 5-triazine (RDX)-degrading bacterium *Shewanella sediminis* HAW-EB3; dibenzo-p-dioxin and dibenzofuran degrading bacterium *Sphingomonas wittichii* RW1.

Another group of organisms that interest scientists are those from extreme environments. For *e.g.* *Sulfurovum* sp. NBC37-1 isolated from a deep-sea hydrothermal vent, *Prochlorococcus marinus* str. MIT 9515 collected from the surface waters of the Equatorial Pacific, *Thermotoga petrophila* RKU-1, an anaerobic thermophile isolated from oil reservoir etc have been sequenced completely. The genome analysis of these strains will give information about the genes and proteins that enable these organisms to survive and flourish in the unfavorable and stressful conditions.

A number of industrially important bacteria such as *Saccharopolyspora erythraea* NRRL 2338 which is a source of the antibiotic erythromycin; *Actinobacillus succinogenes* 130Z, a succinate-producing bacteria used by the pharmaceutical industry and *Clostridium beijerinckii* NCIMB 8052, an industrial solvent-producing bacterium are also completely sequenced.

Though the number of genomes that are sequenced under various contexts is endless, what is really important is the

way this information is being used today. A few examples cited below show how the genome sequence information is being used for new exciting findings and drawing interesting correlations.

Understanding adaptive radiation

The genome sequence of a number of species belonging to a single phylogenetic group which differ in lifestyle can be used to understand the adaptive radiation of the microorganisms. The genetic and metabolic factors responsible for functional differences between the divergent members can throw light on the features enabling a particular lifestyle. The information of the complete genome sequence is being vastly used for such comparative studies. Three different species of the genus *Acidovorax* with as diverse lifestyle as a free living nitroaromatic degrader, opportunistic human pathogen and a symbiotic pathogen namely, *Acidovorax* sp. JS42, *Acinetobacter baumannii* ATCC 17978 and *Acidovorax avenae* subsp. *citrulli* AAC00-1 respectively have been completely sequenced. The underlying idea is to get an important framework for comparisons and to obtain general understanding of the factors responsible for the free living, pathogenic and symbiotic interactions of these closely related organisms. Thus, the genome sequence data of these can be of interest to a broad group of investigators studying not only the specific genus, but those working in the field of host/microbe interactions and also to those working on the bioremediation strategies. While the 5MB sequence *Acidovorax* sp. JS42 is complete but not yet published, the sequence of *Acinetobacter baumannii*, the causative agent of several types of infections including pneumonia, meningitis, septicemia, and urinary tract infections has already been explored [17]. The assembled genome is 3,976,746 base pairs (bp) and has 3830 ORFs as compared to 4858 and 4225 ORFs in closely related but functionally diverse *Acidovorax avenae* subsp. *citrulli* AAC00-1 and *Acidovorax* sp. JS42. A significant fraction of ORFs (17.2%) are located in 28 putative alien islands, indicating that the genome has acquired a large amount of foreign DNA. Supporting its role in pathogenesis, a remarkable number of the islands (16) contain genes implicated in virulence, indicating that the organism devotes a considerable portion of its genes to pathogenesis. The largest island contains elements homologous to the *Legionella*/*Coxiella* Type IV secretion apparatus. Type IV secretion systems have been demonstrated to be important for virulence in other organisms and thus are likely to help mediate pathogenesis of *A. baumannii*.

Similarly, the genus *Burkholderia* comprises of more than 40 different species, which have been found to be

present in a wide range of ecological niches. They are found in soil, water, plant rhizosphere and endophytically in roots and shoots, in and on fungal mycelia and even in the lungs of cystic fibrosis patients [45]. The large and unusual genomes of species under this genus have captured the interest of scientist to study the various phenotypic traits encoded by the multiple replicons that encode diverse range of phenotypic characters within the same genus. Around 26 species of this genus are in various stages of genome sequencing and six have been completed in 2007. The average size of the genome under this genus is 7.5MB and they possess 2-3 large chromosomal replicons. While the first and the second chromosome are characterized by the presence of ribosomal RNA operons along with the majority of essential house-keeping genes defining them as chromosomal, the presence of essential genes on the third replicon is variable within the genus. This together with the acquired foreign DNA makes the genomes of this genus highly flexible and diverse. The wealth of information generated by sequencing the species of this genus can greatly enhance the understanding of the interplay of genes and diversity thus generated at the molecular level. One of the *Burkholderia* species sequenced in 2007 by JGI, *Burkholderia vietnamiensis* G4 species is known to be trichloroethene (TCE) co-oxidizing strain [46]. Along with the its ability to degrade one of the most harmful and widely spread contaminants in groundwaters, this strain has been found in rhizosphere, cystic fibrosis patients and is also known to fix nitrogen gas. The 8.5 MB genome of this species contains three chromosomes of 1241007, 2411759 and 3652814 base pairs and four plasmids pB-VIE03 (226,679 bp), pBVIE04 (107,231 bp), pBVIE02 (265,616 bp) and pBVIE05 (88,096 bp). The whole genome data of this strain can be used for the comparative studies with other *Burkholderia* members and give some clues regarding differentiating traits required for diverse habitats that it inhabits.

Revealing extreme living

The whole genome sequence data can provide the hidden keys to many unanswered questions. The bacterial species surviving highly unfavorable conditions have always been enigmatic to the scientific world and can provide a lot of interesting information regarding genes and factors which facilitate them to do so. *Bacillus pumilus* SAFR-032 is one such organism that produces spores highly resistant to unfavorable conditions like UV radiation, gamma-radiation, oxidative stress, desiccation, chemical disinfection, or starvation and have been sequenced by Baylor College of Medicine [47]. Both the vegetative cells and the spores of this strain are found to be more resistant to UV and H₂O₂ as compared

to other *Bacillus* species and it even survives the decontamination procedures of the Jet Propulsion Lab spacecraft assembly facility. Its 3.7 Mb genome 3687 ORFs, 12 frame-shifts, 38 pseudogenes, 7 rRNA operons, 69 tRNAs, and 21 ncRNAs. The whole genome data also revealed many genes relevant to DNA repair and the oxidative stress response and those were then compared to the less resistant *Bacillus* species *B. subtilis* and *B. licheniformis* [47]. The *B. pumilus* genome includes genes not found in *B. subtilis* or *B. licheniformis* and conserved genes with sequence variations, but surprisingly lacks several genes that function in UV or H₂O₂ resistance in other *Bacillus* species. A more detailed analysis of the genome would reveal the various factors involved in the resistance to unfavourable conditions and a better understanding of this aspect can even be made useful to the mankind.

Fighting environmental pollution through microbial genomics

A lot of emphasis has been laid on the microbial mediated bioremediation of the hazardous and recalcitrant chemical contaminants all over the world. Naturally, the organism involved in degrading these contaminants have been of great importance as their potential has been realized and exploited. However, the reality is that the promise of bioremediation is yet to be realized. There are many setbacks in the implementation of bioremediation technologies and the major one being the lack of enough information on the mechanisms controlling the growth and activity of the microbes in the polluted environments and their interaction and response to the presence of various physical and biological factors. But with the advent of the methods of genome sequencing, analysis of the gene expression and functions under various conditions and comparative genomics, the goal now seem to be attainable. In that direction, genomes of many degraders are being completely sequenced and studied in detail. The information added with each bacterial genome can be useful to get a holistic view of the pattern of interactions among the degraders with their environment. Among the microbial genomes sequenced in 2007 is the complete genome sequence of a hydrocarbon degrading thermophile, *Geobacillus thermodenitrificans* NG80-2, isolated from a deep oil reservoir in Northern China [48]. It consists of a 3,550,319 bp chromosome and a 57,693 bp plasmid. The genome data reveals that this strain NG80-2 possess genes for utilization of a number of energy sources, transporters for the uptake of various nutrients, detoxification of harmful compounds and genes for a flexible respiration system including an aerobic branch comprising five

terminal oxidases and an anaerobic branch with a complete denitrification pathway for quick response to dissolved oxygen fluctuation. All this makes this strain capable of efficiently adapting itself to a number of contaminated environments including oil reservoirs. Interestingly the identification of a nitrous oxide reductase gene has not been previously described in Gram-positive bacteria. The proteome studies further revealed the presence of a long-chain alkane degradation pathway involving a protein called long-chain alkane monooxygenase LadA [48]. The thermophilic soluble monomeric LadA could be an ideal candidate for treatment of environmental oil pollutions and biosynthesis of complex molecules.

Evidence for distributed genome hypothesis (DGH)

According to distributed genome hypothesis, chronic bacterial pathogens exhibit polyclonal infection and reorganization at genetic level to deal with the adaptive host responses. It suggests that free-living bacterial species possess a supragenome that is much larger than the genome of any single bacterium. In a study done by Hoggs *et al* [49], at Allegheny General Hospital, Pittsburgh, nine nontypeable *Haemophilus influenzae* (NTHi) clinical isolates were sequenced and analyzed. Clustering identified a total of 2,786 genes, of which 1,461 were common to all strains, with each of the remaining 1,328 found in a subset of strains. The number of genes ranged from 1,686 to 1,878 per strain. A single strain pair was found to differ from each other by 96 to 585 genes. Comparisons of each of the NTHi strains with the Rd strain revealed between 107 and 158 insertions and 100 and 213 deletions per genome. The mean insertion and deletion sizes were 1,356 and 1,020 base-pairs, respectively, with mean maximum insertions and deletions of 26,977 and 37,299 base-pairs. The large numbers of rearrangements prevalent in these strains is in keeping with the transformation mechanisms that make these pathogens highly adept in tackling the host response. The 1 MB genome sequence of two more strains of *Haemophilus influenzae* strains PittEE and PittGG sequenced later in the year 2007, with 1741-1667 and 1695-1619 genes:protein coding sequences respectively, is also in accordance with the supragenome model for NTHi genomes which explains the distribution of genes among these strains.

There are many more genomes that are being sequenced, annotated and analysed leading to the pool of information that is being generated everyday. The improving sequencing techniques, fast automation and the long list of the microbes being sequenced points to a time when it would be a regular practice to sequence whole genome of all the bacterial species studied for various proposes. Once

the analysis tools become rapid and more comprehensible, many difficulties pertaining to the use/threat by bacterial world will get their answers and we would be able to use the services of prokaryotic kingdom to the welfare of mankind in a more precise and advantageous manner.

Not only the complete microbial genomes, even the DNA sequence from uncultivable microorganisms from a wide variety of niches has opened up a treasure of information from microbial world that has emerged as a field of metagenomics.

Metagenomics

By 1990's, when it became evident that standard laboratory culturing techniques are not able to pick up more than 99% of the bacterial diversity in most environmental samples [50, 51], scientists became thrilled by the thought of identifying these 'uncultured microbes' and unleashing their unlimited potential for biotechnological purposes. Since then, there has been no looking back, only that the goals have been refined. Now, apart from identifying the microbial diversity [52, 53] and exploring it for biotechnological prospects [54–57], understanding the community composition and assigning ecological role to the unculturable microbiota of a particular niche has also gained importance [54, 58, 59]. Apart from the scientists, a number of industries are also gearing up to study the rich biodiversity of the uncultured microbes so as to capture the immense potential of novel enzymes and biomolecules. This study of the collective microbial genomes contained in an environmental sample was first termed as 'metagenomics' by Jo Handelsman [54] and subsequently explained as the statistical concept of *meta*-analysis (the process of statistically combining separate analyses) and genomics (comprehensive analysis of an organism's genetic material) [55, 58, 60]. More specifically, the study may be defined as the culture-independent analysis of the genomes of an assemblage of microorganisms from environmental sample which may include soil [61, 62], aquatic habitats [63], gut [64], oral cavity [65], excreta [66] etc. Several other terms such as 'community genome' [67], 'environment DNA (eDNA) libraries' [63], 'recombinant environmental libraries' [68] etc. were earlier used but the term 'metagenomics' gained precedence over the other terms and probably world wide acceptance after an international conference with this name was held in Darmstadt, Germany in 2003 [59].

Metagenomic DNA

Recovery of high molecular weight DNA from environmental samples is the first and foremost requirement of metagenomic analysis. The metagenomic DNA extraction is often

troublesome and this is especially true for soil ecosystems, which contain inhibitory substances like humic acids and phenolic compounds. The similar physico-chemical properties of humic acids to those of nucleic acids [69] make it difficult to remove them during DNA extraction. These contaminants can inhibit certain restriction endonucleases [70–73] and also Taq polymerase. It becomes almost impossible to obtain PCR amplification from the extracted DNA containing humic acid content as low as 0.08 µg/ml since most of the Taq polymerases are inhibited at this concentration [74]. However, restriction endonucleases can tolerate upto 0.5–17 µg/ml humic acids content in the DNA [74].

The presence of humic acid contamination in the DNA can be detected by the brownish color of the DNA extract and their level can be determined spectrophotometrically by the absorbance ratios. A high 260/230 ratio (>2) is indicative of pure DNA, while a low ratio is indicative of humic acid contamination since humic acids exhibit absorbance at both 230nm and 260nm [75]. As evident, it becomes essential to remove these contaminants, which might require additional clean-up or purification steps before the extracted DNA can be used for other molecular processes.

The size of DNA obtained from environmental samples is also a matter of concern if one has to proceed for the construction of metagenomic library in high capacity vectors. Since most of the DNA isolation protocols especially from soil, involve either bead-beating or vigorous mixing, the DNA obtained is often sheared and the size obtained might not be sufficient to be cloned in large insert vectors such as cosmid, fosmid, Bacterial Artificial Chromosome (BAC) etc.

Though there are various methods available for extraction of DNA from environmental samples [76–79] the choice of method is critical with respect to the type of environmental sample used, type of library to be made (number of clones and coverage) and its application (i.e. kind of natural products to be identified) [80]. If required, the environmental sample can be enriched to increase the representation of a particular group of organisms selectively. This could either be through a compound (e.g. starch, cellulose, aromatic contaminants etc.) alone [55, 81] or along with a label like BrdU or [¹³C] C isotope [55]. Thus, only the selected microbial population flourishes and incorporates the label in its DNA which is then recovered by immunocapture methods (for BrdU) or by using density gradient centrifugation (in case of [¹³C]), the method is called Stable Isotope Probing (SIP) [55].

Metagenomic libraries

Metagenomic libraries are the systematic representation of the microbial diversity of a particular niche in suitable vectors (Fig. 9). The idea of directly cloning

environmental DNA was first suggested by Norman Pace in 1985 [82] and successively applied by his group in 1991 for cloning 16S rRNA gene from ocean metagenome [83]. Landmark achievement in this direction was the construction of fosmid vector library of entire metagenomic DNA from oceanic sample by DeLong and coworkers in 1996 [63].

The process of construction of a metagenomic library requires extraction of DNA from environmental sample either directly [80] or following enrichment [55]. The quality of DNA in terms of size, amount and presence of contaminants should be considered before the library is made as poor quality DNA can lead to low coverage of environmental genomes in the library [80]. Extracted DNA is purified if required and cloned in suitable vectors, such as plasmid [84, 85], cosmid [86], fosmid [63, 87] or BAC [88, 89]. As small insert libraries do not permit detection of large gene clusters and operons, large insert libraries in cosmid/ fosmid/ BAC vectors are preferred [80]. The cloned DNA can be transformed or transfected to *E. coli* host cells. Alternatively, libraries can be made directly in expression vectors, wherein cloned fragments are linked to a promoter for obtaining expression of foreign genes. However, direct expression of genes from metagenomic samples in heterologous hosts such as *E. coli* is not free from problems. Many genes from environmental samples might not express in a particular bacterial host owing to its toxicity or lack of suitable regulatory factors/transcription/translation machinery/protein folding enzymes [80]. Enhancing expression of genes in metagenomic libraries may lead to discovery of a wide array of natural products. This may be achieved by taking libraries to alternate hosts such as *Streptomyces* [90], *Bacillus*, *Rhizobium*, *Pseudomonas* [54] etc. and has been the basis for discovery of antibiotic terragins [90]. In addition, novel shuttle vectors that would facilitate the simultaneous expression of environmental DNA in alternative hosts are currently being investigated [80].

Screening of metagenomic libraries

The next formidable task is screening of these libraries containing a huge collection of clones for selecting specific clones containing desired gene or activity. This can be done by either sequence based or function based approaches as detailed below.

Sequence-based analysis of metagenomic libraries: As the name suggests, this method employs conserved DNA sequences to detect genes showing homology to the sequences already present in the database [55]. Conserved sequences are used for designing hybridization probes or

PCR primers that are then used to screen metagenomic libraries. For instance, novel polyketide synthase (PKS) I cluster was detected from soil metagenome using this approach [91, 92]. One glitch this approach suffers from is that it omits the isolation of members of unknown families and can only be used to screen genes belonging to previously described families [59, 58, 80].

The approach also involves sequencing of clones containing universal phylogenetic markers such as 16S rRNA gene that helps in linking a particular genomic fragment to a particular taxa [58, 59]. In one such report Beja and co-workers [93] sequenced a clone from seawater that contained a rhodopsin-like gene but was of γ -proteobacterial origin. This was first such report in which bacteriorhodopsin was being reported from a proteobacteria and disproved the generalization that this gene is limited to archaea.

Another microarray-based method called ‘metagenomic profiling’ requires hybridization of metagenomic library with labeled DNA from various reference strains, isolates, or communities and thus is a rapid method for identifying clones that belong to unidentified microbial taxa [94].

Function-based analysis of metagenomic libraries: This approach requires identification of clones expressing a desired trait, followed by their sequencing and biochemical analysis [55, 58]. The drawback of the sequence-driven approach is overcome here, as both novel and previously described genes can be identified through this method. Some genes that have been detected by virtue of functional screening are oxidoreductases [81], amylases [84], lipases [81], and esterases [95, 96]. The identification of clones which express beneficial natural products or proteins find direct application in the industry and emphasizes on the need of high-throughput screening technologies specific for required products.

Furthermore, modified function-based methods such as SIGEX and METREX have been developed for intracellular screening of clones. SIGEX (Substrate-Induced Gene Expression) was put forth by Uchiyama and co-workers [97] to trap genes involved in catabolic pathways by their (pathway) specific effectors. The underlying principle is to clone metagenomic DNA upstream of a promoterless *gfp* and then screen for clones showing GFP production in the presence of specific effectors by using Fluorescence-Activated Cell Sorter (FACS). This approach facilitates detection of promoters of inducible catabolic gene(s) which respond to their specific substrate thereby producing a reporter (GFP) response. The reporter response will however not be noticed if the substrate is unable to enter the host cell or the clone contains some terminator sequences which prevent GFP expression or the host lacks regulatory elements necessary for substrate-promoter interplay [97, 98].

Fig. 9. Metagenomic analysis.

While SIGEX traps promoters responsive to certain effectors, METREX (Metabolite-Regulated Expression) is an offshoot of SIGEX which detects the effectors that activate such promoters [99]. In this technique, cloned metagenomic DNA is introduced to a host cell containing a biosensor for compounds that induce quorum sensing. Now if the clone produces a signaling compound (such as acylated homoserine lactones or other related compounds), it induces quorum sensing and directs expression of reporter gene (*gfp*) producing the reporter (GFP) response [58, 99].

Major Discoveries

Metagenomics has given us a powerful tool which can trap the potential of inaccessible microbial world that was completely unimaginable few years back. Contrary to the earlier studies based on individual microorganisms, metagenomics adopts a holistic approach of community analysis which is in agreement with the basic fact that microbial activities

are a result of complex interactions between them. Today, metagenomics has revolutionized our thinking of the biological world and has helped us to use the full potential of microbial wealth that was untapped till now. The information generated by metagenomics has not only enriched our understanding and knowledge but also find direct application in industry, therapeutics and environment sustainability. Some of these discoveries are discussed below.

Enzyme and natural products

A number of novel enzymes, genes and natural products of commercial interest have been identified through metagenomic studies. Several amylases, lipases, esterases, chitinases, amidases etc. that show very low homology to the proteins in the database have been reported from metagenomic analysis in the last decade. Table 1 summarizes some of these discoveries. For instance, A novel amylolytic enzyme (AmyM) showing less than 50%

Table 1. Table showing discovery of some of the novel genes or gene products by Metagenomic analyses.

Gene/ Natural Products	Source of DNA extraction	Active clones/No. of clones analyzed	Vector used	Host strain	Reference
Turbomycin A and B	Soil	3/ 24,546	BAC (pBeloBAC1)	<i>E.coli</i> DH10B	101
Terragine	Soil	18/ 1020	???	<i>S. lividans</i>	90
Alcohol oxidoreductase activity	Soil from a sugar beet field		Plasmid (pBluescript SK+)	<i>E. coli</i> strains DH5a & ECL707	81
Lipase/Esterase	Forest top soil	15/ 33,700	Fosmid (pEpiFOS-5)	<i>E. coli</i> coli EPI-100 strain	87
	Pond water	11/30,000	Plasmid (pUC19)	<i>E.coli</i> DH10B <i>E.coli</i> EPI300TM-T1 ^R strain & <i>E.coli</i> BL21 (DE3) for functional screening	85
	Hot water spring sediments	4/2,000	Fosmid (p=CC1FOS)	<i>E.coli</i> EPI300 TM -T1 ^R strain	96
	Metagenome pools (mud fats, beaches, & forests)	5/60,000	Fosmid (pCC1FOS)		95
Amylase	Soil	1/30,000	Plasmid (pUC19)	<i>E.coli</i> DH5a	84
	Soil	8/3,648	BAC (pBeloBAC11)	<i>E.coli</i> DH10B	62
Nuclease	Soil	1/3,648	BAC (pBeloBAC11)	<i>E.coli</i> DH10B	62
Antibacterial agent and indirubin and indigo blue compounds	Two Forest soils	2/ 80,500 + 33,200	Fosmid (pEpiFOS-5)	<i>E. coli</i> EPI-100	111

identity with known genes in the Genbank database has been isolated using metagenomic approach which possesses starch hydrolysis and transglycosylation properties [84]. Similarly a number of lipolytic genes or esterases have been isolated. In one such report, out of the 33700 fosmid clones from a metagenomic library prepared from top forest soil, 8 unique lipolytic clones were identified on the basis of clearing zone on LB-tributyryn medium [87]. The subcloning of the lipolytic clones in the small insert library yielded six novel genes that encoded for lipolytic enzymes with 34–48% similarity to known enzymes. A new family of bacterial lipolytic enzymes was identified and characterized after the discovery of novel lipase-encoding gene, *lipG* tidal flat metagenomic library from Korea [100]. In another report, 12 unique lipolytic genes with 25 to 70% amino acid identity with known proteins in the database were discovered from pond water metagenome [85]. Functional screening for ester-hydrolyzing enzymes from a metagenomic library resulted in the isolation of a novel esterase (Est25) which hydrolyzed ketoprofen ethyl ester efficiently [96]. The amino acid sequence of the deduced protein showed only moderate identity (64.8%) to known esterases/lipases in the database. A gene coding for a thermostable esterase (*estE1*) has been isolated from mud and water rich sediment of a hot water spring [95]. The enzyme showed high thermal stability and was active between 30° and 95° C. It showed 64% homology to an enzyme from a hyperthermophilic archaeon, *Pyrobaculum calidifontis*.

Novel genes

Metagenomic libraries have also been used for mining a number of novel genes or products of pharmacological importance. Handelsman and co-workers identified a potential antibiotic-Turbomycin A and B exhibiting broad-spectrum antibiotic activity against gram-negative and gram-positive organisms, from soil metagenome [101] in 2002. Brady and colleagues (2001) reported the discovery of a broad-spectrum antibiotic violacein from soil metagenomic library [102]. Sequencing of this active metagenomic clone revealed that though the genetic organization of the biosynthetic operon was similar to that of a known violacein producer, *Chromobacterium violaceum*, the sequence of the genes showed differences suggesting that the metagenomic clone was derived from some other organism [58, 102]. Davies and co-workers in 2000, identified a novel metabolite, Terragine from soil metagenome [90]. More recently, Diaz-Torres *et al.* (2006) screened four metagenomic libraries made from the oral microflora of 20 adult human beings to identify novel tetracycline, amoxicillin and gentamicin resistant genes [65]. Piel and group reported a set of

putative antitumor polyketide, pederin biosynthesis genes from a metagenomic library of *Paederus fuscipes* beetles and its associated bacteria [92]. Although pederin has been known to be synthesized and used by the beetles as a chemical defense [103], the identified gene cluster was found to be located in the genome of an as yet unculturable *Pseudomonas* sp. bacterium.

A number of novel genes involved in degradation of organic compounds have also been identified through metagenomic approach. Hennes *et al.* have identified novel genes with activity similar to 4-hydroxybutyrate dehydrogenases [61]. These genes are significantly different to the already known genes involved in the metabolism of 4-hydroxybutyrate and their gene products might facilitate the utilization of 4-hydroxybutyrate in a way different to the already known pathway [61]. Kube *et al.* 2005 identified a 79 kb contiguous non-redundant sequence from the fosmid library of a microbial mat from the northwest shelf of Black Sea. The sequence was found to contain seventy ORFs (36 with assigned function and 34 of unknown function) and one potential tRNA^{Ser}. A complete pathway of anaerobic benzoate degradation as well as genes involved in anaerobic 4-hydroxybenzoate degradation was located on the 79kb stretch of DNA [104].

Exploring microbiota of human gut

Apart from the discovery of many novel genes and enzymes, metagenomic analysis is leading to some very interesting findings. In a recent study, researchers have explored the microbiota in the gut of mice models using metagenomic approach to find if any correlation could be drawn between obese and lean subjects (mice or humans) [64, 105]. The study depicted that the microbial population inhabiting the gastrointestinal tract of both humans and mice belongs to two groups, the firmicutes and the bacteroidetes and the proportion of firmicutes group is greater than bacteroidetes in obese subjects; while the reverse is true for lean subjects [64, 105]. Such findings emphasize on the utility of comparative metagenomics in understanding the complex nature of interactions between humans and microbial communities [64, 105].

Exploring microbiota of extreme environments

16S rRNA gene analysis data from a number of environmental metagenomic samples [51, 58, 59, 61–63, 67, 80, 82, 83] has unveiled the enormous microbial diversity present in nature. Furthermore metagenomics has also facilitated the study of unexplored microbiota present in extreme environmental conditions such as in hot water springs, sea, ice, extreme pH and hyper salinity. One such example

is the study of natural acid mine drainage (AMD) biofilm in California which was found to be composed of *Leptospirillum*, *Ferroplasma*, *Sulfobacillus* and *Acidimicrobium* [67]. Such studies are important in identifying the types of organisms prevalent in extreme conditions and will possibly help us in understanding of biogeochemical cycles and lifestyle of extremophiles [67, 106]. In addition; shot-gun sequencing of metagenomic clones from Sargasso Sea traced 148 completely new species from nearly 1800 genomic species detected. Such a data is sufficient to represent the magnitude of potential of the unexplored zones of our environment [107].

Metagenome sequencing and Genome reconstitution

The scope of metagenomics has been expanded by the sequencing of complete community genomes and reconstitution of multiple genomes directly from environmental samples. As the community structure of samples from extreme environments is expected to be less diverse, it is easier to sequence and reconstitute genomes from these samples. One such sample chosen by Tyson and coworkers is natural acid mine drainage (AMD) biofilm, at Richmond mine in California [67]. AMD represents a nutrient deficient extremely harsh environment with very low pH (~0.5), high FeS₂ content and restricted energy source. The group subjected metagenome of the biofilm to random shotgun sequencing. The AMD metagenome data determined was somewhat more than 10 Mb of sequences in ~ 2500 contigs and nearly 8,000 putative protein sequences [108]. The sequenced data was used to reconstruct the near-complete genomes of the dominant microbes, *Leptospirillum* group II and *Ferroplasma* type II and partial recovery of three other genomes. The analysis of the gene complement for each organism revealed the metabolic pathways for carbon and nitrogen fixation and energy generation. Genes for biosynthesis of isoprenoid-based lipids and for a variety of proton efflux systems were also identified [67]. This kind of data not only provided important insights into the microbial community structure but also lead to the understanding of survival strategies of the resident microbes in extreme acidic environments.

One of the most extensive microbial metagenomic studies till date has been the sequencing of metagenome of Sargasso Sea in the Atlantic Ocean near Bermuda. This herculean task was taken up by Craig J. Ventor and his group [106]. The approach involved shotgun sequencing of the microbial consortium from the Sea and more than 1 million kb non-redundant sequences were obtained. The data generated almost 2 million sequence reads, yielding over 1.6 million bp of raw DNA sequence and revealed at

least 1800 genomic species of which 148 were unknown. Around 1.2 million potential unique proteins were identified out of which 782 were putative rhodopsin-like photoreceptors. Sargasso Sea metagenome data includes 788 Mb of sequences in 8, 09, 112 contigs and nearly a million putative protein sequences [107]. Two complete genomes of bacterial strains *Shewanella* and *Burkholderia* have already been reconstituted from this extensive data by Ventor and co-workers [106]. De Long, however had reservations on this data and propounded these as contaminants rather than inhabitants of the Sea [108]. This kind of controversy puts forth the need for rigorous methods of validation of results arising from such a 'meta-analysis'. More recently the fidelity of results obtained by employing various methods for processing metagenomic sequences has been put to test by simulation studies conducted by Mavromatis *et al* [109]. In their study, they created data sets by randomly combining reads from more than 100 genomes to model real metagenomes with respect to their complexity and phylogenetic analysis. Statistical methods can also be employed to do comparative analysis to determine the prevalence of certain genes in varied environments [107]. Such studies are important as these serve as yardsticks to determine the correctness of already existing methods and improve their efficacy. The lack of reference genomes makes the annotation of sequences from uncultured organisms a difficult task [110]. The daunting task of metagenome assembly and reconstitution of individual genomes will be easy if more reference genomes are made available in the database.

The scope presented by genomics and metagenomics is likely to change the face of biological studies in future. The challenge right now is to come up with easy to use tools for sequencing and assembly of genomes and metagenomes and to put this information together to understand the function of novel genes, proteins and organisms and extrapolate their relevance with logical comparative analyses.

Acknowledgements

The authors PS, HK, MK, MV, KK and SM are thankful to CSIR and ICMR, Government of India, for providing doctoral fellowships. Financial support from funding agencies DBT and DST is gratefully acknowledged.

References

1. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocumbe PM & Smith M (1977) Nucleotide sequence of bacteriophage phi X174 DNA. Nature 265:687–695

2. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA & Merrick JM (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269(5223): 496–512
3. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, Fritchman RD, Weidman JF, Small KV, Sandusky M, Fuhrmann J, Nguyen D, Utterback TR, Saudek DM, Phillips CA, Merrick JM, Tomb JF, Dougherty BA, Bott KF, Hu PC, Lucier TS, Peterson SN, Smith HO, Hutchison CA 3rd & Venter JC (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science* 270(5235):397–403
4. Maxam AM & Gilbert W (1977) A new method for sequencing DNA. *Proc Natl Acad Sci* 74:560–564
5. Maxam AM & Gilbert W (1980) Sequencing end-labelled DNA with base specific chemical cleavages. *Methods Enzymol* 65:499–560
6. Sanger F, Nicklen S & Coulson AR (1977) DNA sequencing with chain terminating inhibitors. *Proc Natl Acad Sci* 74: 5463–5467
7. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SBH & Hood LE (1986) Fluorescence detection in automated DNA sequence analysis. *Nature* 321:674–679
8. Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, Zagursky RJ, Cocuzza AJ, Jensen MA & Baumeister K (1987) A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* 238(4825): 336–341
9. Swerdlow H & Gesteland R (1990) Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Res* 18(6):1415–1419
10. Bains W & Smith GC (1988) A novel method for nucleic acid sequence determination. *J Theor Biol* 135(3): 303–307
11. Drmanac R, Labat I, Brukner I & Crkvenjakov R (1989) Sequencing of megabase plus DNA by hybridization: theory of the method. *Genomics* 4(2):114–128
12. Khrapko KR, Lysov YuP, Khorlyn AA, Shick VV, Florentiev VL & Mirzabekov AD (1989) An oligonucleotide hybridization approach to DNA sequencing. *FEBS Lett* 256 (1-2):118–122
13. Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P. Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem* 1996; 242(1):84–89
14. Ronaghi M, Uhlén M & Nyrén P (1998) A sequencing method based on real-time pyrophosphate. *Science* 281 (5375):363–365
15. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, Roth R, George D, Eletr S, Albrecht G, Vermaas E, Williams SR, Moon K, Burcham T, Pallas M, DuBridge RB, Kirchner J, Fearon K, Mao J & Corcoran K (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol* 18(6):630–634
16. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD & Church GM (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309:1728–1732
17. Smith MG, Gianoulis TA, Pukatzki S, Mekalonas JJ, Ornston LN, Gersein M, Snyder M (2007) New insights into *Acinetobacter baumannii* pathogenesis revealed by high-density pyrosequencing and transposon mutagenesis. *Genes and development* 21:601–614
18. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhi-jani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF & Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437(7057):376–380
19. Wicker T, Schlagenhauf E, Graner A, Close TJ, Keller B & Stein N (2006) 454 sequencing put to test using the complex genome of barley. *BMC Genomics* 7:275
20. Berezikov E, Thuemmler F, van Laake LW, Kondova I, Bontrop R, Cuppen E & Plasterk RH (2006) Diversity of microRNAs in human and chimpanzee brain. *Nat Genet* 38: 1375–1377
21. Emrich SJ, Barbazuk B, Li L & Schnable PS (2007) Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res* 17:69–73
22. Edwards RA, Rodriguez-Brito B, Wegley L, Haynes M, Breitbart M, Peterson DM, Saar MO, Alexander S, Alexander Jr EC & Rohwer F (2006) Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* 7:57
23. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER & Gordon JI (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444: 1027–1031
24. Poinar HN, Schwarz C, Qi J, Shapiro B, Macphee RD, Buigues B, Tikhonov A, Huson DH, Tomsho LP, Auch A, Rapp M, Miller W & Schuster SC (2006) Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science* 311(5759):392–394
25. Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, Simons JF, Du L, Egholm M, Rothberg JM, Paunovic M & Paabo S (2006) Analysis of one million base pairs of Neanderthal DNA. *Nature* 444:330–336
26. Goldberg SM, Johnson J, Busam D, Feldblyum T, Ferreira S, Friedman R, Halpern A, Khouri H, Kravitz SA, Lauro FM, Li K, Rogers YH, Strausberg R, Sutton G, Tallon L, Thomas T, Venter E, Frazier M & Venter JC (2006) Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proc Natl Acad Sci* 103(30):11240–11245
27. Huang X, Wang J, Aluru S, Yang S P & Hillier L (2003) PCAP: A Whole-Genome Assembly Program. *Genome Res* 13:2164–2170
28. Bonfield JK, Smith K & Staden R (1995) A new DNA sequence assembly program. *Nucleic Acids Res* 23(24): 4992–4999
29. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH,

- Remington KA, Anson EL, Bolanos RA, Chou HH, Jordan CM, Halpern AL, Lonardi S, Beasley EM, Brandon RC, Chen L, Dunn PJ, Lai Z, Liang Y, Nusskern DR, Zhan M, Zhang Q, Zheng X, Rubin GM, Adams MD & Venter JC (2000) A whole-genome assembly of *Drosophila*. *Science* 287(5461):2196–2204
30. Batzoglu S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauce-li E, Berger B, Mesirov JP & Lander ES (2002) ARACHNE: a whole-genome shotgun assembler. *Genome Res* 12(1): 177–189
 31. Wang J, Wong GK, Ni P, Han Y, Huang X, Zhang J, Ye C, Zhang Y, Hu J, Zhang K, Xu X, Cong L, Lu H, Ren X, Ren X, He J, Tao L, Passey DA, Wang J, Yang H, Yu J & Li S (2002) RePS: a sequence assembler that masks exact repeats identified from the shotgun data. *Genome Res* 12(5): 824–831
 32. Aparicio S, Chapman J, Stupka E, Putnam N, Chia J, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, Gelpke M D S, Roach J, Oh T, Ho I Y, Wong M, Detter C, Verhoef F, Predki P, Tay A, Lucas S, Richardson P, Smith S F, Clark M S, Edwards Y J K, Doggett N, Zharkikh A, Tavtigian S V, Pruss D, Barnstead M, Evans C, Baden H, Powell J, Glusman G, Rowen L, Hood L, Tan Y H, Elgar G, Hawkins T, Venkatesh B, Rokhsar D & Brenner S (2002) Whole-Genome Shotgun Assembly and Analysis of the Genome of *Fugu rubripes*. *Science* 297:1301–1310
 33. Mullikin JC & Ning Z (2003) The phusion assembler. *Genome Res* 13(1):81–90
 34. Sutton G G, White O, adams M D & Kerlavage A R (1995) TIGR Assembler: a new tool for assembling large shotgun sequencing projects. *Genome Science and Technology* 1:9–19
 35. Bocs S, Cruveiller S, Vallenet D, Nuel G & Médigue C (2003) AMIGene: Annotation of Microbial Genes. *Nucleic Acids Research* 31(13):3723–3726
 36. Besemer J & Borodovsky M (2005) GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res* 33(Web Server issue):W451–W454
 37. Delcher AL, Harmon D, Kasif S, White O & Salzberg SL (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* 27(23):4636–4641
 38. Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD, Kerlavage AR, Dougherty BA, Tomb JF, Adams MD, Reich CI, Overbeek R, Kirkness EF, Weinstock KG, Merrick JM, Glodek A, Scott JL, Geoghagen NS & Venter JC (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* 273:1058–1073
 39. Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD, Richardson DL, Kerlavage AR, Graham DE, Kyrpides NC, Fleischmann RD, Quackenbush J, Lee NH, Sutton GG, Gill S, Kirkness EF, Dougherty BA, McKenney K, Adams MD, Loftus B, Peterson S, Reich CI, McNeil LK, Badger JH, Glodek A, Zhou L, Overbeek R, Gocayne JD, Weidman JF, McDonald L, Utterback T, Cotton MD, Spriggs T, Artiach P, Kaine BP, Sykes SM, Sadow PW, D'Andrea KP, Bowman C, Fujii C, Garland SA, Mason TM, Olsen GJ, Fraser CM, Smith HO, Woese CR & Venter JC (1997) The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* 390(6658):364–370
 40. Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, Klenk HP, Gill S, Dougherty BA, Nelson K, Quackenbush J, Zhou L, Kirkness EF, Peterson S, Loftus B, Richardson D, Dodson R, Khalak HG, Glodek A, McKenney K, Fitzgerald LM, Lee N, Adams MD, Hickey EK, Berg DE, Gocayne JD, Utterback TR, Peterson JD, Kelley JM, Cotton MD, Weidman JM, Fujii C, Bowman C, Watthey L, Wallin E, Hayes WS, Borodovsky M, Karp PD, Smith HO, Fraser CM & Venter JC (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* 388(6642):539–547
 41. Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, White O, Ketchum KA, Dodson R, Hickey EK, Gwinn M, Dougherty B, Tomb JF, Fleischmann RD, Richardson D, Peterson J, Kerlavage AR, Quackenbush J, Salzberg S, Hanson M, van Vugt R, Palmer N, Adams MD, Gocayne J, Weidman J, Utterback T, Watthey L, McDonald L, Artiach P, Bowman C, Garland S, Fuji C, Cotton MD, Horst K, Roberts K, Hatch B, Smith HO & Venter JC (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* 390(6660):580–586
 42. Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B & Shao Y (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* 277:1453–1474
 43. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H & Oliver SG (1996) Life with 6000 genes. *Science* 274(5287):546, 563–567
 44. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE 3rd, Tekai F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Krogh A, McLean J, Moule S, Murphy L, Oliver K, Osborne J, Quail MA, Rajandream MA, Rogers J, Rutter S, Seeger K, Skelton J, Squares R, Squares S, Sulston JE, Taylor K, Whitehead S & Barrell BG (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393(6685):537–544
 45. Coenye T & Vandamme P (2003) Diversity and significance of *Burkholderia*. *Environ Microbiol* 5(9):719–729
 46. Fries, MR, Forney LJ & Tiedje JM (1997) Phenol- and toluene-degrading microbial populations from an aquifer in which successful trichloroethene cometabolism occurred. *Appl Environ Microbiol* 63:1523–1530
 47. Gioia J, Yerrapragada S, Qin X, Jiang H, Igboeli OC, Muzny D, Dugan-Rocha S, Ding Y, Hawes A, Liu W, Perez L, Kovar C, Dinh H, Lee S, Nazareth L, Blyth P, Holder M, Buhay C, Tirumalai MR, Liu Y, Dasgupta I, Bokhetache L, Fujita M, Karouia F, Eswara Moorthy P, Siefert J, Uzman A, Buzumbo P, Verma A, Zwiya H, McWilliams BD, Olowu A, Clinkenbeard KD, Newcombe D, Golebiewski L, Petrosino JF, Nicholson WL, Fox GE, Venkateswaran K, Highlander SK & Weinstock GM (2007) Paradoxical DNA Repair and

- Peroxide Resistance Gene Conservation in *Bacillus pumilus* SAFR-032. *PLoS ONE* 2(9):e928
48. Feng L, Wang W, Cheng J, Ren Y, Zhao G, Gao C, Tang Y, Liu X, Han W, Peng X, Liu R & Wang L (2007) Genome and proteome of long-chain alkane degrading *Geobacillus thermodenitrificans* NG80-2 isolated from a deep-subsurface oil reservoir. *Proc Natl Acad Sci* 104(13):5602–5607
 49. Hogg JS, Hu FZ, Janto B, Boissy R, Hayes J, Keefe R, Post JC & Ehrlich GD (2007) Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biol* 8(6):R103
 50. Torsvik V, Goksoyr J & Daae FL (1990) High diversity in DNA of soil bacteria. *Appl Environ Microbiol* 56:782–787
 51. Amann KI, Ludwig W & Schleifer KH (1995) Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiol Rev* 59:143–169
 52. Pace NR (1997) A molecular view of microbial diversity and its biosphere. *Science* 276:734–740
 53. Hugenholtz P (2002) Exploring prokaryotic diversity in the genomic era. *Genome Biol* 3:1–8
 54. Handelsman J, Rondon MR, Brady SF, Clardy J & Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol* 5:R245–249
 55. Schloss PD & Handelsman J (2003) Biotechnological prospects from metagenomics. *Curr Opin Biotechnol* 14: 303–310
 56. Schmeisser C, Steele H & Streit WR (2007) Metagenomics, biotechnology with non-culturable microbes. *Appl Microbiol Biotechnol* 75:955–962
 57. Lorenz P & Eck J (2005) Metagenomics and Industrial applications. *Nature* 3:510–516
 58. Handelsman J (2004) Genomics of uncultured microorganisms. *Mol Microbiol Rev* 68:669–685
 59. Riesenfeld CS, Schloss PD & Handelsman J (2004) Metagenomics: Genomic Analysis of microbial communities. *Annu Rev Genet* 38:525–552
 60. Kimura N (2006) Metagenomics: Access to Unculturable microbes in the environment *Microbes Environ* 21:201–205
 61. Henne A, Daniel R, Schmitz RA & Gottschalk G (1999) Construction of environmental DNA libraries in *Escherichia coli* and screening for the presence of genes conferring utilization of 4-hydroxybutyrate. *Appl Environ Microbiol* 65:3901–3907
 62. Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, Liles MR, Loiacono KA, Lynch BA, MacNeil IA, Minor C, Tiong CL, Gilman M, Osburne MS, Clardy J, Handelsman J & Goodman RM (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl Environ Microbiol* 66:2541–2547
 63. Stein JL, Marsh TL, Wu KY, Shizuya H & DeLong EF (1996) Characterization of uncultivated prokaryotes: Isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J Bacteriol* 178:591–599
 64. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER & Gordon JI (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444:1027–1031
 65. Diaz-Torres ML, Villedieu N, Hunt R, McNab DA, Spratt, Allan E, Mullany P & Wilson M (2006) Determining the antibiotic resistant potential of the indigenous oral microbiota of humans using a metagenomic approach. *FEMS Microbiol Lett* 258:257–262
 66. Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, Salamon P & Rohwer F (2003) Metagenomic Analyses of an Uncultured Viral Community from Human Feces. *J Bacteriol* 185:6220–6223
 67. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS & Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428:37–43
 68. Courtois S, Cappellano CM, Ball M, Francou FX, Normand P, Helynck G, Martinez A, Kolvek SJ, Hopke J, Osburne MS, August PR, Nalin R, Guérineau M, Jeannin P, Simonet P & Pernodet JL (2003) Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products. *Appl Environ Microbiol* 69:49–55
 69. Harry M, Gambier B, Bourezgui Y & Garnier-Sillam E (1999) Evaluation of purification procedures for DNA extracted from organic rich samples: interference with humic substances. *Analyst* 27:439–442
 70. Porteous LA & Armstrong JL (1991) Recovery of bulk DNA from soil by a rapid, small-scale extraction method. *Curr Microbiol* 22:345–348
 71. Tsai YL & Olson BH (1991) Rapid method for direct extraction of DNA from soil and sediments. *Appl Environ Microbiol* 57(4):1070–1074
 72. Tsai YL & Olson BH (1992) Detection of low numbers of bacterial cells in soils and sediments by polymerase chain reaction. *Appl Environ Microbiol* 58:754–757
 73. Jacobsen, Carsten S & Rasmussen Ole F (1992) Development and Application of a New Method To Extract Bacterial DNA from Soil Based on Separation of Bacteria from Soil with Cation-Exchange Resin. *Appl Environ Microbiol* 58(8): 2458–2462
 74. 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
 75. Yeates MR, Gillings AD, Davison, Altavilla N & Veal DA (1997) PCR amplification of crude microbial DNA extracted from soil. *Lett Appl Microbiol* 25 (4):303–307
 76. Krsek M & Wellington EMH (1999) Comparison of different methods for the isolation and purification of total community DNA from soil. *J Microbiol Methods* 39:1–16
 77. Zhou J, Bruns MA & Tiedje JM (1996) DNA recovery from soils of diverse composition. *Appl Environ Microbiol* 62: 316–332
 78. Berry AE, Chiochini C, Selby T, Sosio M & Wellington EMH (2003) Isolation of high molecular weight DNA from soil for cloning into BAC vectors. *FEMS Microbiol Lett* 223:15–20
 79. Daniel R (2005) The metagenomics of soil. *Nature* 3:470–477
 80. Handelsman J, Liles M, Mann D, Riesenfeld C & Goodman RM (2002) Cloning the metagenome: Culture-independent access to the diversity and functions of the uncultivated microbial world. *Meth in Microbiol* 33(13):241–255

81. Knietzsch A, Waschkwitz T, Bowien S, Henne A & Daniel R (2003) Construction and Screening of Metagenomic Libraries Derived from Enrichment Cultures: Generation of a Gene Bank for Genes Conferring Alcohol Oxidoreductase Activity on *Escherichia coli*. *Appl Environ Microbiol* 69: 1408–1416
82. Pace NR, Stahl DA, Lane DJ & Olsen GJ (1985) Analyzing natural microbial populations by rRNA sequences. *ASM News* 51:4–12
83. Schmidt TM, DeLong EF & Pace NR (1991) Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J Bacteriol* 173:4371–4378
84. Yun J, Kang S, Park S, Yoon H, Kim MJ, Heu S & Ryu S (2004) Characterization of a Novel Amylolytic Enzyme Encoded by a Gene from a Soil-Derived Metagenomic Library. *Appl Environ Microbiol* 70:7229–7235
85. Ranjan R, Grover A, Kapardar RK & Sharma R (2005) Isolation of novel lipolytic genes from uncultured bacteria of pond water. *Biochem Biophys Res Comm* 335:57–65
86. Voget S, Leggewie C, Uesbeck A, Raasch C, Jaeger K-E & Streit WR (2003) Prospecting for novel biocatalysts in a soil metagenome. *Appl Environ Microbiol* 69:6235–6242
87. Lee SW, Won K, Lim HK, Kim JC, Choi GJ & Cho KY (2004) Screening for novel lipolytic enzymes from uncultured soil microorganisms. *Appl Microbiol Biotechnol* 65: 720–726
88. Béja O, Suzuki MT, Koonin EV, Aravind L, Hadd L, Nguyen LP, Villacorta R, Amjadi M & Garrigues C (2000) Construction of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ Microbiol* 2:516–529
89. Béja O (2004) To BAC or not to BAC: marine ecogenomics. *Curr Opin Biotechnol* 15:187–190
90. Wang GY, Graziani E, Waters B, Pan W, Li X, McDermott J, Meurer G, Saxena G, Anderson RJ & Davies J (2000) Novel natural products from soil DNA libraries in a Streptomycete host. *Org Lett* 2:2401–2404
91. Ginolhac A, Jarrin C, Gillet B, Robe P, Pujic P, Tuphile K, Bertrand H, Vogel TM, Perrière G, Simonet P & Nalin R (2004) Phylogenetic analysis of polyketide synthase I domains from soil metagenomic libraries allows selection of promising clones. *Appl Environ Microbiol* 70:5522–5527
92. Piel J, Hui D, Fusetani N & Matsunaga S (2004) Targeting modular polyketide synthases with iteratively acting acyltransferases from metagenomes of uncultured bacterial consortia. *Environ Microbiol* 6:921–927
93. Béja O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP, Jovanovich SB, Gates CM, Feldman RA, Spudich JL, Spudich EN & DeLong EF (2000) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* 289:1902–1906
94. Sebat JL, Colwell FS & Crawford RL (2003) Metagenomic profiling: microarray analysis of an environmental genomic library. *Appl Environ Microbiol* 69:4927–4934
95. Rhee JK, Ahn DG, Kim YG & Oh JW (2005) New thermophilic and thermostable esterase with sequence similarity to the hormone-sensitive lipase family, cloned from a metagenomic library. *Appl Environ Microbiol* 71:817–825
96. Kim YJ, Choi GS, Kim SB, Yoon GS, Kim YS & Ryu YW (2006) Screening and characterization of a novel esterase from a metagenomic library. *Prot Express Purif* 45(2):315–323
97. Uchiyama T, Abe T, Ikemura T & Watanabe K (2005) Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. *Nature Biotechnol* 23:88–93
98. Jia Y & Ryu S (2005) Screening for novel enzymes from metagenome and SIGEX, as a way to improve it. *Microbial cell factories* 4: Article number 8, 5p
99. 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
100. Lee MH, Lee CH, Oh TK, Song JK & Yoon JH (2006) Isolation and Characterization of a novel lipase from a meta-genomic library of tidal flat sediments: Evidence for a new family of bacterial lipases. *Environ Microbiol* 72: 7406–7409
101. Gillespie DE, Brady SF, Bettermann AD, Cianciotto NP, Liles MR, Rondon MR, Clardy J, Goodman RM & Handelsman J (2002) Isolation of Antibiotics Turbomycin A and B from a Metagenomic Library of Soil Microbial DNA. *Appl Environ Microbiol* 68:4301–4306
102. Brady SF, Chao CJ, Handelsman J & Clardy J (2001) Cloning and Heterologous Expression of a Natural Product Biosynthetic Gene Cluster from eDNA. *Org Lett* 3 (13):1981–1983
103. Kellner RLL & Dettner K (1996) Differential efficacy of toxic pederin in deterring potential arthropod predators of *Paederus* (Coleoptera: Staphylinidae) offspring. *Oecologia* 107:293–300
104. Kube M, Beck A, Meyerdieck A, Amann R, Reinhardt R & Rabus R (2005) A catabolic gene cluster for anaerobic benzoate degradation in methanotrophic microbial Black Sea mats. *Syst and Appl Microbiol* 28:287–294
105. Ley RE, Turnbaugh PJ, Klein S & Gordon JI (2006) Human gut microbes associated with obesity. *Nature* 444:1022–1023
106. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson K E, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH & Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304:66–74
107. Rodriguez-Brito B, Rohwer F & Edwards RA (2006) An application of statistics to comparative metagenomics. *BMC Bioinformatics* 7: art. no. 162
108. DeLong EF (2005) Microbial community genomics in the ocean. *Nat Rev* 3:459–469
109. Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltzman E, McHardy AC, Rigoutsos I, Salamov A, Korzeniewski F, Land M, Lapidus A, Grigoriev I, Richardson P, Hugenholtz P & Kyrpides NC (2007) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods* 4:479–480
110. Schloss PD & Handelsman J (2005) Metagenomics for studying unculturable microorganisms: cutting the Gordian Knot. *Genome Biol* 6 (8):229
111. Lim HK, Chung EJ, Kim JC, Choi GJ, Jang KS, Chung YR, Cho KY & Lee SW (2005) Characterization of a forest soil metagenome clone that confers indirubin and indigo production on *Escherichia coli*. *Appl Environ Microbiol* 71: 7768–7777