Review TRENDS in Biotechnology Vol.23 No.6 June 2005

Metagenomic gene discovery: past, present and future

Don Cowan, Quinton Meyer, William Stafford, Samson Muyanga, Rory Cameron and Pia Wittwer

Advanced Research Centre for Applied Microbiology, Department of Biotechnology, University of the Western Cape, Bellville 7535, Cape Town, South Africa

It is now widely accepted that the application of standard microbiological methods for the recovery of microorganisms from the environment has had limited success in providing access to the true extent of microbial biodiversity. It follows that much of the extant microbial genetic diversity (collectively termed the metagenome) remains unexploited, an issue of considerable relevance to a wider understanding of microbial communities and of considerable importance to the biotechnology industry. The recent development of technologies designed to access this wealth of genetic information through environmental nucleic acid extraction has provided a means of avoiding the limitations of culture-dependent genetic exploitation.

Introduction

The total number of prokaryotic cells on earth has been estimated at $4–6\times10^{30}$ [1], thought to comprise between 10^6 and 10^8 separate genospecies (distinct taxonomic groups based on gene sequence analysis) [2]. This diversity presents an enormous (and largely untapped) genetic and biological pool that can be exploited for the recovery of novel genes, entire metabolic pathways and their products [3]. Observations showing that culturing yields a fraction of the microbial diversity evident from microscopic analysis [2] have been consistently supported by the results of phylotypic analyses on community DNA preparations, leading to the concept of 'unculturables' [4]. The apparent underestimation of true microbial diversity derives largely from a reliance on culture-based enumeration methods. There is a growing belief that the term 'unculturable' is inappropriate [5,6] and that in reality we rather have yet to discover the correct culture conditions [7]. The development of metagenomic technologies over the past five years has provided access to much of the prokaryotic genetic information available in environmental samples, independent of culturability.

Sample enrichment

In a metagenomic screening process (e.g. expression screening of metagenomic libraries), the target gene(s) represent a small proportion of the total nucleic acid fraction. Pre-enrichment of the sample thus provides an attractive means of enhancing the screening hit rate. The discovery of target genes can be significantly improved by applying one of several enrichment options (Box 1), ranging from whole-cell enrichment, to the selection and enrichment of target genes and genomes (Figure 1). For example, in the Sargasso Sea genome sequencing project size-selective filtration effectively removed the eukaryotic cell population [8]. Alternatively, differential centrifugation has been used to enrich for *Buchnera aphidicola* and *Cenarchaeum symbiosum* symbionts by removing them from their hosts in preparation for whole genome sequencing [9].

Culture enrichment on a selective medium favours the growth of target microorganisms. The inherent selection pressure can be based on nutritional, physical or chemical criteria, although substrate utilization is most commonly employed. For example, a four-fold enrichment of cellulase genes in a small insert expression library was obtained by culture enrichment on carboxymethylcellulose [10]. Although culture enrichment will inevitably result in the loss of a large proportion of the microbial diversity by selecting fast-growing culturable species, this can be partially minimized by reducing the selection pressure to a mild level after a short period of stringent treatment.

Nucleic acid extraction and enrichment technologies

Numerous community nucleic acid extraction methods have been developed [11,12]. More details on community DNA extraction technologies, can be found in the following papers [11,13,14]. The two principal strategies for the recovery of metagenomic DNA are cell recovery and direct lysis [15]. Extraction of total metagenomic DNA is necessarily a compromise between the vigorous extraction required for the representation of all microbial genomes, and the minimisation of DNA shearing and the co-extraction of inhibiting contaminants. Mechanical bead beating has been shown to recover more diversity compared with chemical treatment [16]. However, chemical lysis is a more gentle method, recovering higher molecular weight DNA. Chemical lysis can also select for certain taxa by exploiting their unique biochemical characteristics.

The technologies for recovering RNA from environmental samples are largely similar to those used for DNA isolation, modified to optimise the yield of intact mRNA by minimising single-stranded polynucleotide degradation [17–20]. Protocols are designed to limit physical

Corresponding author: Cowan, D. (dcowan@uwc.ac.za).
Available online 19 April 2005

 0167-7799/$ - see front matter doi:10.1016/j.tibtech.2005.04.001

Box 1. How feasible are metagenomic enrichment technologies?

Stable isotope probing (SIP) and 5-Bromo-2-deoxyuridine labelling

Many ^{13}C-, ^{18}O- and ^{15}N- labelled fine chemicals are available (e.g. phenol, methanol, ammonia, methane, carbonate etc.) but the wide application of SIP [90] is limited by the commercial availability of complex labelled compounds that require expensive custom synthesis. BrdUTP labelling offers an alternative in cases where SIP labelled compounds are not available. Growth in the presence of BrdUTP and the unlabeled compound accesses metabolically active organisms. These methods are limited by the difficulties in acquiring high labelling efficiency and the recycling of the label in the community resulting in a breakdown in selective enrichment.

Suppressive subtractive hybridisation (SSH)

SSH identifies genetic differences between microorganisms, but the complexity of metagenomes makes this detection difficult. SSH [91] has successfully been used on complex metagenomes [33] and the sensitivity of the process can be increased by using multiple rounds of subtractive hybridisation.

Differential expression analysis (DEA)

DEA targets transcriptional differences in gene expression. Several variations in the basic concept exist [34]. These include selective amplification via biotin and restriction-mediated enrichment (SABRE), integrated procedure for gene identification (IPGI), serial analysis of gene expression (SAGE), tandem arrayed ligation of expressed sequence tags (TALEST) and total gene expression analysis (TOGA). These techniques have been effectively applied for eukaryotic gene discovery but, to the authors knowledge, none have been applied in a metagenomic context. Their high sensitivity and selectivity should enable small differences in expression of single copy genes to be detected.

Phage display

Phage-display expression libraries provide a means of isolating a given DNA sequence by affinity selection of the surface-displayed protein to an immobilised ligand. Biopanning involves repeated cycles of binding that will successively enrich the pool. After several rounds of enrichment, individual clones are characterised by DNA sequencing [75]. This method is efficient and amenable to high-throughput screening, offering the potential to enrich even rare DNA sequences in the metagenome, but current phage technology limits expression of proteins $<$50kDa.

Affinity capture

Oligonucleotides covalently immobilised to a solid support can be used to affinity purify target genes. The slow kinetics of hybridisation limit this process, but might be improved by using metagenomic mRNA or single-stranded DNA. This approach is still in development [92].

Microarray

Microarrays allow high-throughput robotic screening for targeting multiple gene products [93]. The cost and availability of microarray technology is rapidly decreasing, making this an increasingly attractive option.

degradation and RNase activity, which are the major causes of yield loss. Samples should be processed or frozen at $-80°C$ immediately after harvesting and additional methods used to minimise RNA degradation, such as the co-precipitation of cellular RNA with proteins (e.g. sulphate salt solution RNAlater, Ambion Inc.; www.ambion.com) and the synthetic capping of the isolated RNA [20–22]. mRNA recovery has been applied extensively to eukaryotes, but has only recently been used in the study of prokaryote metagenomes [23,24]. These techniques provide a feasible route for the construction of metagenomic cDNA libraries for the further identification of functional eukaryotic genes.

Total DNA extracted directly from environmental samples does not typically contain an even representation of the population's genomes within the sample. Rare organisms will contribute a relatively low proportion of the total DNA and the genome population might be overshadowed by a limited number of dominant organisms [25]. This could lead to a selective bias in downstream manipulations such as PCR. This problem can be partially resolved by means of experimental normalisation [26]. Separation of genotypes is achieved by caesium chloride gradient centrifugation in the presence of an intercalating agent, such as bis-benzimide, for the buoyant density separation of genomes based on their %G and C content. Equal amounts of each band on the gradient are combined to represent a normalised metagenome. Normalisation can also be achieved by denaturing fragmented genomic DNA, and re-annealing under stringent conditions (e.g. 68°C for 12–36 h). Abundant ssDNA will anneal more rapidly to generate double-stranded nucleic acids than rare DNA species. Single-stranded sequences are then separated from the double-stranded nucleic acids, resulting in an enrichment of rarer sequences within the environmental sample [26].

Genome and gene enrichment

Genome enrichment strategies can be used to target the active components of microbial populations [27,28]. Stable-isotope probing (SIP) techniques involve the use of a stable isotope-labelled substrate and density gradient centrifugal separation of the 'heavier' DNA or RNA. $^{13}CH_3OH$-labelling of forest soil metagenomic DNA resulted in the identification of both known α-proteobacterial methylotrophs and novel methanol dehydrogenase (*mxaF)* gene variants belonging to Acidobacterial taxa [29]. Similarly, analysis of ^{13}C-phenol enriched anaerobic bioreactor populations by RNA-SIP demonstrated that phenol degradation was dominated by a member of the genus *Thauera*, a group previously unknown as phenol degraders [30]. Actively growing microorganisms can also be labelled with 5-bromo-2-deoxyuridine (BrdU) and the labelled DNA or RNA separated by immunocapture or density gradient centrifugation [27]. Addition of substrates with BrdU selects among the members of the microbial community for enhanced growth on the specific substrate [31]. Limitations of these methods include cross-feeding and recycling of the label within the community, resulting in loss of specific enrichment [9].

Suppressive subtraction hybridisation (SSH) identifies genetic differences between microorganisms and is therefore a powerful technique for specific gene enrichment. Adaptors are ligated to the DNA populations and subtractive hybridization is carried out to select for DNA fragments unique to each DNA sample. This has typically been applied to analyse genetic differences between two

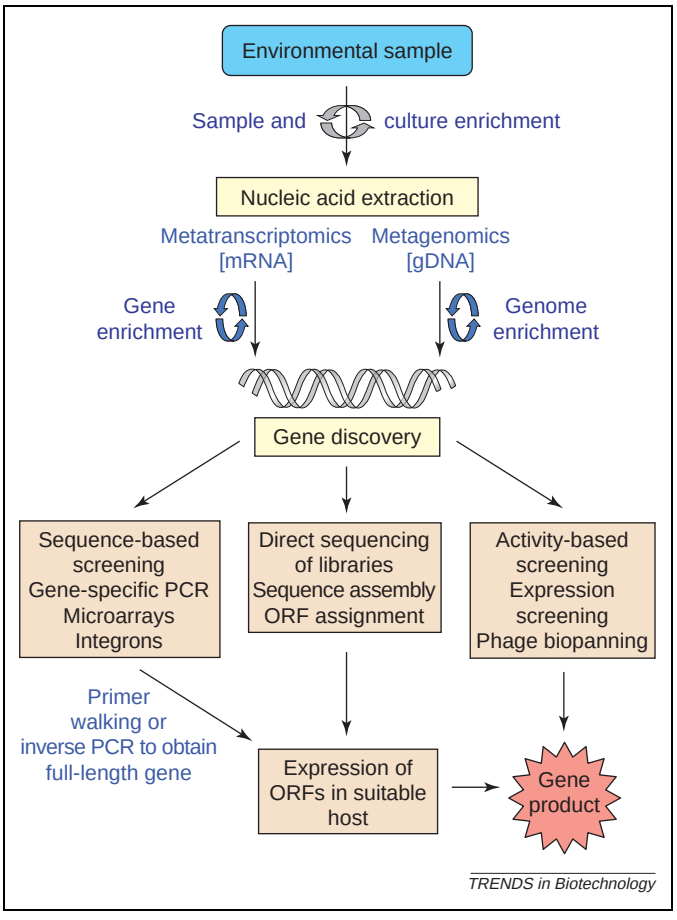

Figure 1. Metagenomic gene discovery. After biotope selection and sample or culture enrichment (if desired), nucleic acid is extracted from the environmental sample. The approach might involve metagenomics (environmental genomic DNA) or metatranscriptomics (environmental mRNA reversed transcribed to complementary DNA, cDNA) and an enrichment or selection can be applied. Gene enrichment selects for differentially expressed genes using techniques such as differential expression analysis (DEA) and gene targeting. Genome enrichment uses techniques such as stable isotope probing (SIP), 5′Bromo-2-deoxyuridine (BrdU)-labelling and suppressive subtractive hybridization (SSH) to enrich or select for genomes of interest. Downstream screening approaches can be activity-based through the screening of expression libraries, sequence-dependent by using gene targeting or can be sequence-independent through the direct sequencing of the metagenome. The final expression requires a full-length open reading frame (ORF) expressed in a suitable host to generate a functional gene product.

closely related bacteria (e.g. in the identification of genetic elements contributing to pathogenesis) [32], but has recently been used to identify differences between complex DNA samples from the rumens of two different animals [33]. Clearly, these techniques could be adapted to target specific genes in related metagenomes. For example, the identification of genes involved in the bioremediation of an environmental pollutant could be identified by comparison of a reference metagenome with a 'perturbed' metagenome (i.e. impacted by a specific pollutant). The relatively crude nature of this approach would only result in the identification of the total genetic difference between the two bacterial populations and would not be specific to genes of interest or genes whose expression was up-regulated on addition of the environmental pollutant.

To selectively enrich for a specific target gene within a metagenome a more practical approach would be to use one of several differential expression technologies that rely on the isolation of mRNA to target transcriptional differences in gene expression. Several innovative methods have been developed (reviewed in [34]). These techniques have so far almost exclusively been used to study patterns in eukaryotic gene expression. Differential expression analysis (DEA) is a particularly effective enrichment tool. The expression profile of a culture grown from a metagenomic sample can be compared pre- and post- exposure to a specific substrate or xenobiotic. In this way the expression of genes up-regulated for the specific activity can be identified. This type of approach was successfully applied to identify bacterial genes up-regulated in the absence of iron [35].

Gene targeting

Gene-specific PCR has been used extensively to probe communities for microorganisms with specific metabolic or biodegradative capabilities. For example, the targeting of genes such as methane monooxygenase, methanol dehydrogenase and ammonia monooxygenase was used to identify methanotrophic [36] and chemolithotrophic ammonium-oxidizing bacteria [37,38]. The biodegradative potential of indigenous microbial populations has been assessed by screening metagenomic extracts for the presence of catechol 2,3-dioxygenase, chlorocatechol dioxygenase and phenol hydroxylase genes [39–41]. Other reported examples include the identification of denitrifying bacteria [42,43] and polyhydroxyalkanoate-producing bacteria [44]. However, as a tool for biocatalyst discovery, gene-specific PCR has two major drawbacks. First, the design of primers is dependent on existing sequence information and skews the search in favour of known sequence types. Functionally similar genes resulting from convergent evolution are not likely to be detected by a single gene-family-specific set of PCR primers. Second, only a fragment of a structural gene will typically be amplified by gene-specific PCR, requiring additional steps to access the full-length genes. Amplicons can be labelled as probes to identify the putative full-length gene(s) in conventional metagenomic libraries. Alternatively, PCR-based strategies for the recovery of either the up- or down-stream flanking regions including universal fast walking [45,46], panhandle PCR [47], random primed PCR [48], inverse PCR and adaptor ligation PCR [49] can be used to access the full-length gene. These methods are technically more difficult to apply at a metagenomic level owing to the increased complexity of a metagenomic DNA sample, but have been used successfully for the recovery of novel gene variants of 2,5 diketo-D-gluconic acid reductase from environmental DNA [50]. Because these approaches can be laborious and time-consuming, innovative alternatives have been developed. Cassette PCR has been used to isolate the central fragment of catechol 2,3-dioxygenase genes from genomic DNA obtained from a phenol and crude oil-degrading bacterial consortium [51]. The internal fragment of a previously cloned full-length copy of a catechol 2,3-dioxygenase gene was then replaced with the PCR-derived internal fragment, thus constructing a novel hybrid catechol 2,3- dioxygenase gene. This approach can be combined with PCR mutagenesis and/or chimeragenesis to generate highly diverse protein

variants incorporating random and directed molecular evolution [52].

The use of RNA might be more effective than DNA for profiling functional microbial communities because RNA is a more sensitive biomarker owing to its high turnover [30]. Reverse transcriptase PCR (RT–PCR) has been used to recover genes from environmental samples, for example in the isolation of naphthalene-degrading enzymes from microorganisms present in a coal tar waste [20]. Although this approach suffers from the technical difficulties associated with mRNA recovery from environmental samples (see previous section), it benefits from wider genomic access (includes structural genes from lower eukaryotes as well as from prokaryotes) and the facility to select for functional genes in response to alterations in environmental conditions.

Integrons are naturally occurring gene capture, dissemination and expression systems that have until recently primarily been associated with antibiotic resistant and pathogenic bacteria [52]. They are widely dispersed in nature and could play a significant role in bacterial genome evolution [53–55]. The key structural features of an integron include a gene cassette integration site (*att1*), an *intI* gene that encodes an integrase and two promoters that drive the expression of the integrase gene and the incorporated gene cassettes [56]. The mobile element in the system is the gene cassette, which consists of one or more open reading frame(s) (ORFs) and associated chromosomal attachment sites (*attC*, also referred to as the 59 base elements; 59-be). The integrase catalyses the insertion of the gene cassette into the integration site controlled by the strong promoter via site-specific recombination using *att1* and *attC* as its substrates [57]. Integrons therefore act as a repository of ORFs coding for many gene products and potentially provide a source of novel genes. Primers designed to target the conserved regions within the 59 base element have successfully been used to recover novel genes homologous to DNA glycosylase, phosphotransferase, methyl transferase and thiotransferase [58]. The specificity of this system for gene targets could be improved by using a primer specific for the gene of interest and one targeting a flanking 59 base element.

Homologous recombination cloning can be used for single-step gene targeting and screening with only those recombinants containing the gene of interest viable after transformation [59,60]. This method requires the design of an *E. coli* host containing a vector DNA sequence homologous to the 5′- and 3′- sequences flanking the gene of interest. The efficiency of bacterial homologous recombination has been improved and commercial systems are now available (Red/ET system, Gene Bridges; www.genebridges.com). To the best of our knowledge, homologous recombination cloning has not yet been applied to metagenomic gene discovery.

Methods requiring only one gene-specific primer impose less sequence-dependent bias compared with standard twin-primer PCR amplification procedures. An elegant application of this method would be the use of immobilised oligonucleotides [61] designed to target a specific gene fragment or consensus sequences by affinity binding. This approach is, of course, used routinely for recovery of polyA RNA cDNA library construction, but has not been applied to gene targeting. Affinity capture should be equally applicable to either denatured cDNA or genomic DNA fragments and yields could be further enhanced with prior linking of adaptors so that affinity selected DNA fragments could be PCR amplified using linker-specific primers.

Microarrays represent a powerful high-throughput system for analysis of genes. They are typically used to monitor differential gene expression, to quantify the environmental bacterial diversity and catalogue genes involved in key processes [62]. Microarrays of immobilised oligonucleotide gene targets have also been used to select appropriate biotope samples for metagenomic library screening [63]. Such arrays could also be used for the affinity capture of targets as a means of enrichment before construction of metagenomic libraries. Microarray technology could also be used for the pre-selection of genes in metagenomic libraries before shotgun sequencing, thereby reducing the sequencing burden and reducing the proportion of sequences unassigned by database sequence similarity searches [62].

Metagenomic DNA libraries

The basic steps of DNA library construction (generation of suitably sized DNA fragments, cloning of fragments into an appropriate vector and screening for the gene of interest) have been extensively and successfully used for over three decades. As there are no obvious limitations in translating the technologies of genomic library construction and screening to metagenomic libraries, it is perhaps surprising that metagenomics only developed in the mid 1990s with the successful application of library construction to marine metagenomes [64]. Subsequent metagenomic gene mining work by Recombinant Biocatalysis Ltd (now Diversa Corporation) and several other laboratories demonstrated the successful recovery of novel genes from metagenomic gene expression libraries [65–70]. The approach taken by each has been broadly similar, although a variety of vector and host systems have been used (Table 1). Functional expression is commonly used as a method to screen for specific gene classes. However, such libraries are amenable to screening by virtually any method that can be adapted to deal with large clone populations.

DNA fragmentation is a significant problem when constructing metagenomic libraries. The vigorous extraction methods required for high yields of DNA from environmental samples often result in excessive DNA shearing. This precludes the construction of libraries using cohesive ends because highly sheared DNA (e.g. 0.5–5 Kbp fragments) cannot be restricted to generate ligatable sticky ends without significant loss of the total gene complement. An alternative approach uses blunt-end or T–A ligation to clone randomly sheared metagenomic fragments [70].

Cosmid and bacterial artificial chromosome (BAC) libraries have been widely used for the construction of metagenomic libraries [71,72]. The ability to clone large fragments of metagenomic DNA allows entire functional operons to be targeted with the possibility of recovering

Table 1. Characteristics of metagenomic libraries. Examples of libraries constructed for gene targeting are shown[a]

Target gene	Host or vector systems used	Library size (number of independent clones)	Average insert size or range of size (Kbp)	% Prokaryote metagenome represented[e]	Refs
Chitinase	Lambda Zap II/ GigapackIII	750 000[b]	2–10	11[c,d]	[65]
4-hydroxybutyrate dehydrogenase; lipase, esterase; cation/H^+ antiporters	*E. coli* DH5α/ pBluescript	930000	5–8	14[d]	[66–68]
Lipase, amylase, nuclease	*E. coli* DH10B/ pBeloBAC11	3648 24576	27 44.5	0.2[d] 3[d]	[69]
Heme biosynthesis; phosphodiesterase	*E. coli* TOP10/ pCR-XL-TOPO	37000	1–10	0.5[d]	[70]
Polyketide biosynthesis	*E. coli, S. lividans* shuttle cosmid	5000	50	0.7	[74]
Alcohol oxido-reductase	*E. coli*/ pSK+	583 000 360 000 324 000	4.4 3.8 3.5	5 3 2	[89]

[a]Caution is advised in attempting to directly compare metagenomic libraries made in different laboratories using different systems.
[b]Number of clones screened.
[c]1800 genomic species were estimated for an oligotrophic open ocean environment [8]. Owing to the coastal location of the sample used in this study [65], we are assuming a 10-fold higher species diversity.
[d]In making these calculations, we have assumed an average of 10^4 prokaryotic species per environmental sample and an average prokaryotic genome size of 4Mbp.
[e]Chemical lysis methods of DNA extraction from soil samples are relatively non-aggressive and we assume that the contribution from eukaryotic (particularly fungal) genomes is minor. We acknowledge that this assumption might be invalid.

entire metabolic pathways. This approach has successfully been applied to the isolation of several multigenic pathways [73,74] such as that responsible for the synthesis of the antibiotic violacein [73]. Fosmid vectors provide an improved method for cloning and stably maintaining cosmid-sized (35–45 Kbp) inserts in *E. coli* [71].

Phage-display expression libraries provide a means for isolating DNA sequences by affinity selection of the surface-displayed expression product. This method is efficient and amenable to high-throughput screening, offering the potential to enrich even rare DNA sequences in the metagenome. However, phage display is limited by the expression capacity of the bacteriophage, a protein size upper limit of around 50kDa [75].

The limitation of *E. coli* as a host for comprehensive mining of metagenomic samples is highlighted by the low number of positive clones obtained during a single round of screening (typically less than 0.01%). A recent *in silico* study indicates that it is virtually impossible to recover translational fusion products owing to the high number of clones ($>10^7$) that would need to be screened [76]. Intuitively, expression from native promoters and read-through transcription from the vector-based promoter offer the best chance for recovery of heterologously expressing genes. Statistically, for a small insert (<10 Kbp) library, between 10^5 and 10^6 clones need to be screened for a single hit [66]. This suggests that without sample enrichment the discovery of specific genes in a complex metagenome is technically challenging.

The assumption that expression in an *E. coli* host will not impose a further bias is largely untested. Although the *E. coli* transcriptional machinery is known to be relatively promiscuous in recognizing foreign expression signals, a bias in favour of Firmicutes genes has been established [76]. The further development of host screening systems is therefore a fruitful approach for the more effective future exploitation of metagenomes.

Metagenomic cDNA (transcriptomic) libraries

Owing to the presence of intronic sequences, metagenomic expression libraries are generally not suitable for mining eukaryotic genes. The large-scale sequencing of clones from cDNA libraries has long been a rapid means of discovering novel eukaryotic genes [77,78]. Acknowledging the technical difficulties of metagenomic mRNA isolation, there is no inherent reason why these technologies cannot be applied to exploit unculturable eukaryotic enzyme genomes via the construction of metagenomic cDNA libraries. Some caution is nevertheless advised. Metagenomic cDNA libraries cannot be as comprehensive as genomic libraries because they can never represent non-expressed genes. In addition, the process of RT–PCR amplification limits the size of inserts and could impose a large sequence-dependent bias on the library [79].

Metagenome sequencing

The sequencing and analysis of large fragments of genomic DNA from uncultured microorganisms are well established technologies [69,79]. These studies have laid the groundwork for the ultimate in metagenomic gene discovery – the sequencing of complete metagenomes. With the relatively recent advent of automated, high-throughput sequencing facilities and of powerful algorithms for sequence assembly, these projects are now technically feasible, albeit financially ambitious. The scale of the task is not trivial – a gram of soil or litre of seawater contains many thousands of unique viral and prokaryotic genomes, hundreds of lower eukaryote species and DNA derived from higher eukaryotes [80,81]. Using conservative estimates of genome sizes [82,83], soil metagenomes could constitute between 20 and 2000 Gbp of DNA sequences.

The sequencing of 76 Mbp of DNA from an acid mine drainage biofilm was the first reported study of this kind [84]. The low biodiversity of the sample enabled the shotgun sequence assembly of two complete genomes.

More than 4000 putative genes were identified, thereby providing insight to the metabolic pathways of the biofilm community. The sequencing of the Sargasso Sea metagenome [8] was more challenging with the sequencing of >1 Gbp of DNA. Approximately 1.2 million putative genes were identified, clearly illustrating the enormous power of this approach for gene discovery. However, the functional assignment of novel genes (i.e. those with no database homologue) is in a state of infancy, with 'evidence-based' gene finder programs [8,85] having limited success. The high biodiversity of the Sargasso Sea and poor sequencing coverage enabled the assembly of only two near-complete genomes [8]. Whole genome assembly could be improved by normalising abundant sequences using a combination of small, medium and large insert libraries and by increasing the coverage of sequencing (at a cost) [71,86]. Recently, differences in tetranucleotide repeat numbers between genomes have proven useful tools for discrimination, provided that the sample is low in complexity and the genomes are equally represented [87]. However, nucleotide polymorphisms, gene rearrangements, gene duplications and horizontal gene transfer are all factors that will impact on reliable genome assembly [88]. Eukaryotic metagenome sequencing poses even greater challenges owing to the presence of larger genome sizes, introns and 'junk' DNA. The use of metatranscriptomics

Table 2. Commercialisation of metagenomic technologies[a]

Company	Target products	Classes	Products and market	Commercial interest
BASF www.corporate.basf.com	Enzymes	Amylase Hydratase	Acidophilic glucoamylase	Food industry, aiding with the digestion of starch
Bioresearch Italia, SpA (Italy)	Anti-infectives	N.D.	Dalbavancin	Development of human gene targeted therapeutics and novel anti-infective
B.R.A.I.N www.brain-biotech.de	Bioactive peptides and enzymes for pharmaceuticals and agrochemicals	N.D.	Nitrile hydratases Cellulases	Degussa AG Partnership for the industrial processe
Cubist pharmaceuticals http://www.cubist.com/	Anti-infectives	N.D.	N.D.	Various commercial relationships. Variety of products in Stage I, II and III trials
Diversa www.diversa.com	Enzymes	Nitrilase Glycosidase Phytase	Discovery of >100 novel nitrilases Production of Lipitor Pyrolase™ 160 and Pyrolase™ 200; Phyzyme™ XP	Drug, lowering cholesterol levels Broad spectrum β-mannanase and β-glucanase added to animal feed to break down indigestible phytate in grains and oil seeds to release digestible
	Biometabolites	Fluorescent protein	DiscoveryPoint™ Green-FP* and Cyan-FP*	Novel green and cyan fluorescent proteins for potential use in drug discovery, commercial screening and academic research
Diversa and Invitrogen www.invitrogen.com	Enzymes	DNA polymerase	ThermalAce™ and Replicase™ DNA for research and diagnostics	Research and diagnostics
EMetagen www.emetagen.com	Enzymes; antibiotics; small active molecules	Polyketides	eMetagen Gene and Pathway Banks™ Large clone DNA libraries encoding biosynthetic pathways for 5000 to 20 000 secondary metabolites	Food, agriculture, research and other commercial applications Pharmaceuticals: antimicrobial, anticancer and other bioactive properties
Kosan Technology www.kosan.com	Antibiotics	Polyketides	Adriamycin, Erythromycin, Mevacor, Rapamycin, Tacrolimus (FK506), Tetracycline, Rapamycin,	Therapeutic drugs
Genencor www.genencor.com	Enzymes	Lipase Protease	Washing powder and alkaline tolerant protease.	Cleaning industry
Libragen www.libragen.com	Antibiotics and biocatalysis for pharmaceuticals	N.D.	Anti-infective and antibiotic discovery Biocatalysis discovery for pharmaceuticals (partnership with Synkem)	Medicine; synthesis of pharmaceuticals
Prokaria www.prokaria.is/	Enzymes	Rhamnosidase β-1,6 Gluconase; Single stranded DNA ligase	Food and agricultural industry Research and diagnostics	Food industry Anti-phytopathogenic fungal agent
Proteus www.proteus.fr	Enzymes; antibiotics; antigens	Not specified	Products for the agricultural, environmental, food, medical and chemical industries	Development of novel biomolecules
Xanagen www.xanagen.com	Libraries	Gene products	Unspecified	Services in library construction, screening and annotation

[a]Note: Some of the products listed above may have been derived from metagenomic libraries with prior enrichments or from single genomes N.D. - no details available or products still under development.

and cDNA libraries might, to some extent, overcome these limitations.

Commercial successes

Since the inception of two pioneering commercial metagenomics ventures in the late 1990s (Recombinant Biocatalysis Ltd of La Jolla and TerraGen Discovery Inc. of Vancouver; www.cubist.com) these technologies have been taken up by several of the biotechnology giants, and have been the focal area of several start-up companies (see Table 2). Recombinant Biocatalysis Ltd, now Diversa Corporation; www.diversa.com), is the acknowledged leader in the field with impressive lists of libraries derived from global biotopes and of cloned enzymes in a range of enzyme classes (Table 2). Several other smaller biotechnology companies appear to be competing in the same market sector (Table 2).

The relatively small size of the industrial enzyme market compared with the pharmaceuticals market suggests that a switch in product focus might not be unexpected. Although the authors are unaware of any successfully commercialised therapeutics derived from metagenomic screening programs, the normal timelines for the identification, development, evaluation and approval of products for the pharmaceutics market are longer than the existence of metagenomics as a research field.

Conclusion

It is probably too early to state that metagenomic gene discovery is a technology that has 'come of age'. New approaches and technological innovations are reported on a regular basis and many of the technical difficulties have yet to be fully resolved. However, there can be little doubt that the field of metagenomic gene discovery offers enormous scope and potential for both fundamental microbiology and biotechnological development.

Acknowledgements

The authors gratefully acknowledge the National Research Foundation Institutional Research Development program, the Department of Science and Technology Lead program and the University of the Western Cape for financial support.

References

1 Whitman, W.B. *et al*. (1998) Prokaryotes: the unseen majority. *Proc. Natl. Acad. Sci. U. S. A.* 95, 6578–6583
2 Amman, R.L. *et al*. (1995) Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiol. Rev.* 59, 143–169
3 Cowan, D.A. (2000) Microbial genomes - the untapped resource. *Trends Biotechnol.* 18, 14–16
4 Hugenholtz, P. *et al*. (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J. Bacteriol.* 180, 4765–4774
5 Rappe, M.S. and Giovannoni, S.J. (2003) The uncultured microbial majority. *Annu. Rev. Microbiol.* 57, 369–394
6 Stevenson, B.S. *et al*. (2004) New Strategies for cultivation and detection of previously uncultured microbes. *Appl. Environ. Microbiol.* 70, 4748–4755
7 Hugenholtz, P. and Pace, N.R. (1996) Identifying microbial diversity in the natural environment: a molecular phylogenetic approach. *Trends Biotechnol.* 14, 190–197
8 Venter, J.C. *et al*. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304, 66–74
9 Schloss, P.D. and Handelsman, J. (2003) Biotechnological prospects from metagenomics. *Curr. Opin. Biotechnol.* 14, 303–310
10 Rees, H.C. *et al*. (2004) Diversity of Kenyan soda lake alkaliphiles assessed by molecular methods. *Extremophiles* 8, 63–71
11 Miller, D.N. *et al*. (1999) Evaluation and optimization of DNA extraction and purification procedures for soil and sediment samples. *Appl. Environ. Microbiol.* 65, 4715–4724
12 Zhou, J. *et al*. (1996) DNA recovery from soils of diverse composition. *Appl. Environ. Microbiol.* 62, 316–322
13 Bürgmann, H. *et al*. (2001) A strategy for optimizing quality and quantity of DNA extracted from soil. *J. Microbiol. Methods* 45, 7–20
14 Krsek, M. and Wellington, E.M.H. (1999) Comparison of different methods for the isolation and purification of total community DNA from soil. *J. Microbiol. Methods* 39, 1–16
15 Roose-Amsaleg, C.L. *et al*. (2001) Extraction and purification of microbial DNA from soil and sediment samples. *Appl. Soil Ecol.* 18, 47–60
16 Niemi, R.M. *et al*. (2001) Extraction and purification of DNA in rhizosphere soil samples for PCR-DGGE analysis of bacterial consortia. *J. Microbiol. Methods* 45, 155–165
17 Alm, E. *et al*. (2000) The presence of humic substances and DNA in RNA extracts affects hybridization results. *Appl. Environ. Microbiol.* 66, 4547–4554
18 Frischer, M.E. *et al*. (2000) Whole-cell versus total RNA extraction for analysis of microbial community structure with 16S rRNA-targeted oligonucleotide probes in salt marsh sediments. *Appl. Environ. Microbiol.* 66, 3037–3043
19 Griffiths, R.I. *et al*. (2000) Rapid method for coextraction of DNA and RNA from natural environments for analysis of ribosomal DNA-and rRNA-based microbial community composition. *Appl. Environ. Microbiol.* 66, 5488–5491
20 Wilson, M.S. *et al*. (1999) In situ, real-time catabolic gene expression: extraction and characterization of naphthalene dioxygenase mRNA transcripts from groundwater. *Appl. Environ. Microbiol.* 65, 80–87
21 Carninci, P. *et al*. (1996) High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics* 37, 327–336
22 Edery, I. *et al*. (1995) An efficient strategy to isolate full-length cDNAs based on an mRNA cap retention procedure (CAPture). *Mol. Cell. Biol.* 15, 3363–3371
23 Lamar, R.T. (1995) Quantitation of fungal mRNAs in complex substrates by revese transcription PCR and its application to phanerochaete chrysosporium-colonized soil. *Appl. Environ. Microbiol.* 61, 2122–2126
24 Zhu, Y. *et al*. (1996) Synthesis of high-quality cDNA from nanograms of total or poly A+ RNA with the CapFinderTM PCR cDNA Library Construction Kit. *CLONTECHniques* XI, 30–31
25 Bohannan, B.J.M. and Hughes, J. (2003) New approaches to analyzing microbial biodiversity data. *Curr. Opin. Microbiol.* 6, 282–287
26 Short, J.M. and Mathur, E.J. (1999) Production and use of normalized DNA libraries. *US Patent 6001574*
27 Urbach, E. *et al*. (1999) Immunochemical detection and isolation of DNA from metabolically active bacteria. *Appl. Environ. Microbiol.* 65, 1207–1213
28 Borneman, J. (1999) Culture-independent identification of microorganisms that respond to specified stimuli. *Appl. Environ. Microbiol.* 65, 3398–3400
29 Radajewski, S. *et al*. (2002) Identification of active methylotroph populations in an acidic forest soil by stable-isotope probing. *Microbiology.* 148, 2331–2342
30 Manefield, M. *et al*. (2002) RNA stable isotope probing, a novel means of linking microbial community function to phylogeny. *Appl. Environ. Microbiol.* 68, 5367–5373
31 Yin, B. *et al*. (2000) Bacterial functional redundancy along a soil reclamation gradient. *Appl. Environ. Microbiol.* 66, 4361–4365
32 Bart, A. *et al*. (2000) Representational difference analysis of *Neisseria meningitides* identifies sequences that are specific for the hypervirulent lineage III clone. *FEMS Microbiol. Lett.* 188, 111–114
33 Galbraith, E.A. *et al*. (2004) Suppressive subtractive hybridisation as a tool for identifying genetic diversity in an environmental metagenome: the rumen as a model. *Environ. Microbiol.* 6, 928–937
34 Green, C.D. *et al*. (2001) Open systems: panoramic views of gene expression. *J. Immunol. Methods* 250, 67–79
35 Bowler, L.D. *et al*. (1999) Representational difference analysis of

cDNA for the detection of differential gene expression in bacteria: development using a model of iron-regulated gene expression in *Neisseria meningitidis*. *Microbiology* 145, 3529–3537
36 Henckel, T. *et al*. (1999) Molecular analyses of the methane-oxidizing microbial community in rice field soil by targeting the genes of the 16S rRNA, particulate methane monooxygenase, and methanol dehydrogenase. *Appl. Environ. Microbiol.* 65, 1980–1990
37 Henckel, T. *et al*. (2000) Molecular analyses of novel methanotrophic communities in forest soil that oxidize atmospheric methane. *Appl. Environ. Microbiol.* 66, 1801–1808
38 McDonald, I.R. *et al*. (1995) Detection of methanotrophic bacteria in environmental samples with the PCR. *Appl. Environ. Microbiol.* 61, 116–121
39 Futamata, H. *et al*. (2001) Group-specific monitoring of phenol hydroxylase genes for a functional assessment of phenol-stimulated trichloroethylene bioremediation. *Appl. Environ. Microbiol.* 67, 4671–4677
40 Mesarch, M.B. *et al*. (2000) Development of catechol 2,3-dioxygenase-specific primers for monitoring bioremediation by competitive quantitative PCR. *Appl. Environ. Microbiol.* 66, 678–683
41 Watanabe, K. *et al*. (1998) Molecular detection, isolation, and physiological characterization of functionally dominant phenol-degrading bacteria in activated sludge. *Appl. Environ. Microbiol.* 64, 4396–4402
42 Braker, G. *et al*. (1998) Development of PCR primer systems for amplification of nitrite reductase genes *(nirk* and *nirs*) to detect denitrifying bacteria in environmental samples. *Appl. Environ. Microbiol.* 64, 3769–3775
43 Hallin, S. and Lindgren, P. (1999) PCR detection of genes encoding nitrite reductase in denitrifying bacteria. *Appl. Environ. Microbiol.* 65, 1652–1657
44 Sheu, D. *et al*. (2000) Rapid detection of polyhydroxyalkanoate-accumulating bacteria isolated from the environment by colony PCR. *Microbiology* 146, 2019–2025
45 Mishra, R.N. *et al*. (2002) Directional genome walking using PCR. *Biotechniques* 33, 830–834
46 Myrick, K.V. and Gelbart, W.M. (2002) Universal fast walking for direct and versatile determination of flanking sequence. *Gene* 284, 125–131
47 Megonigal, M.D. *et al*. (2000) Panhandle PCR for cDNA: a rapid method for isolation of MLL fusion transcripts involving unknown partner genes. *Proc. Natl. Acad. Sci. U. S. A.* 97, 9597–9602
48 Liu, Y. and Whittier, R.F. (1995) Thermal asymmetric interlaced PCR: automatable amplification and sequencing of insert end fragments from pi and yac clones for chromosome walking. *Genomics* 25, 674–681
49 Ochman, H. *et al*. (1993) Use of polymerase chain reaction to amplify segments outside boundaries of known sequences. *Methods Enzymol.* 218, 309–321
50 Eschenfeldt, W.H. *et al*. (2001) *DNA from Uncultured Organisms as a Source of* 2,5-Diketo-D-Gluconic Acid Reductases. *Appl. Environ. Microbiol.* 67, 4206–4214
51 Okuta, A. *et al*. (1998) PCR isolation of catechol 2,3-dioxygenase gene fragments from environmental samples and their assembly into functional genes. *Gene* 212, 221–228
52 Rowe-Magnus, D.A. and Mazel, D. (1999) Resistance gene capture. *Curr. Opin. Microbiol.* 2, 483–488
53 Heidelberg, J.F. *et al*. (2000) DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* 406, 477–483
54 Rowe-Magnus, D.A. and Mazel, D. (2001) Integrons: natural tools for bacterial genome evolution. *Curr. Opin. Microbiol.* 4, 565–569
55 Smalla, K. *et al*. (2000) PCR-based detection of mobile genetic elements in total community DNA. *Microbiology* 146, 1256–1257
56 Bennet, P.M. (1999) Integrons and gene cassettes: a genetic construction kit for bacteria. *J. Antimicrob. Chemother.* 43, 1–4
57 Collis, C.M. and Hall, R.M. (1992) Site specific deletion and rearrangement of integron insert genes catalyzed by the integron DNA integrase. *J. Bacteriol.* 174, 1574–1585
58 Stokes, H.W. *et al*. (2001) Gene cassette PCR: sequence-independent recovery of entire genes from environmental DNA. *Appl. Environ. Microbiol.* 67, 5240–5246
59 Zhang, Y. *et al*. (2000) DNA cloning by homologous recombination in *E.coli*. *Nat. Biotechnol.* 18, 13214–13217
60 Bubeck, P. *et al*. (1993) Rapid cloning by homologous recombination *in vivo*. *Nucleic Acids Res.* 21, 3601–3602
61 Stull, D. and Pisano, J.M. (2001) Purely RNA: New innovations enhance the quality, speed, and efficiency of RNA isolation techniques. *Scientist* 15, 29–31
62 Sebat, J.L. *et al*. (2003) Metagenomic profiling: microarray analysis of a metagenomic DNA library. *Appl. Environ. Microbiol.* 69, 4927–4934
63 Kim, C.C. *et al.* (2002) Improved analytical methods for microarray based genome composition analysis. *Genome Biol.* 3, RESEARCH0065
64 Stein, J.L. *et al*. (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
65 Cottrell, M.T. *et al*. (1999) Chitinases from uncultured marine microorganisms. *Appl. Environ. Microbiol.* 65, 2553–2557
66 Henne, A. *et al*. (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
67 Henne, A. *et al*. (2000) Screening of environmental DNA libraries for the presence of genes conferring lipolytic activity on *Escherichia coli*. *Appl. Environ. Microbiol.* 66, 3113–3116
68 Majernik, A. *et al*. (2001) Screening of environmental DNA libraries for the presence of genes conferring Na+(Li+)/H+ antiporter activity on *Escherichia coli*: characterization of the recovered genes and the corresponding gene products. *J. Bacteriol.* 183, 6645–6653
69 Rondon, M.R. *et al*. (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl. Environ. Microbiol.* 66, 2541–2547
70 Wilkinson, D.E. *et al*. (2002) Efficient molecular cloning of environmental DNA from geothermal sediments. *Biotechnol. Lett.* 24, 155–161
71 Beja, O. (2004) To BAC or not to BAC: marine ecogenomics. *Curr. Opin. Biotechnol.* 15, 187–190
72 Daniel, R. (2004) The soil metagenome- a rich resource for the discovery of novel natural products. *Curr. Opin. Biotechnol.* 15, 199–204
73 Brady, S.F. *et al*. (2001) Cloning and heterologous expression of a natural product biosynthetic gene cluster from cDNA. *Org. Lett.* 3, 1981–1984
74 Courtois, S. *et al*. (2003) Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products. *Appl. Environ. Microbiol.* 69, 49–55
75 Crameri, R. and Suter, M. (1993) Display of biologically active proteins on the surface of filamentous phages: a cDNA cloning system for the selection of functional gene products linked to the genetic information responsible for their production. *Gene* 137, 69–75
76 Gabor, E.M. *et al*. (2004) Quantifying the accessibility of the metagenome by random expression cloning techniques. *Environ. Microbiol.* 6, 879–886
77 Rebel, F. *et al*. (1995) PCR-generated cDNA libraries from reduced numbers of mouse oocytes. *Zygote* 3, 241–250
78 Starkey, M.P. *et al*. (1998) Reference cDNA library facilities available from European sources. *Mol. Biotechnol.* 9, 35–57
79 Beja, O. *et al*. (2000) Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ. Microbiol.* 2, 516–529
80 Handelsman, J. *et al*. (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* 5, R245–R249
81 Torsvik, V. and Ovreas, L. (2002) Microbial diversity and function in soil: from genes to ecosystems. *Curr. Opin. Microbiol.* 5, 240–245
82 Ball, K.D. and Trevors, J.T. (2002) Bacterial genomics: the use of DNA microarrays and bacterial artificial chromosomes. *J. Microbiol. Methods* 49, 275–284
83 Farrelly, V. *et al*. (1995) Effect of genome size and *rrn* gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species. *Appl. Environ. Microbiol.* 61, 2798–2801
84 Tyson, J.W. *et al*. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428, 37–43
85 Aggarwal, G. and Ramaswamy, R. (2002) *Ab initio* gene identification: prokaryote genome annotation with GeneScan and GLIMMER. *J. Biosci.* 27, 7–14

86 Schmidt, T.M. *et al*. (1991) Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J. Bacteriol.* 173, 4371–4378
87 Teeling, H. *et al*. (2004) Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ. Microbiol.* 6, 938–947
88 Nelson, K.E. (2003) The future of microbial genomics. *Environ. Microbiol.* 5, 1223–1225
89 Knietsch, A. *et al*. (2003) Metagenomes of complex microbial consortia derived from different soils as sources for novel genes conferring formation of carbonyls from short-chain polyols on *Escherichia coli*. *J. Mol. Microbiol. Biotechnol.* 5, 46–56
90 Radajewski, S. *et al*. (2003) Stable-isotope probing of nucleic acids: a window to the function of uncultured microorganisms. *Curr. Opin. Biotechnol.* 14, 296–302
91 Sagerstrom, C.G. *et al*. (1997) Substrate cloning. Past, present and future. *Annu. Rev. Biochem.* 66, 751–783
92 Demidov, V.V. *et al*. (2000) Duplex DNA capture. *Curr. Issues Mol. Biol.* 2, 31–35
93 Wu, L. *et al*. (2001) Development and evaluation of functional gene arrays for the election of selected genes in the environment. *Appl. Environ. Microbiol.* 67, 5780–5790