Review

Comparing microarrays and next-generation sequencing technologies for microbial ecology research

Seong Woon Roh[1], Guy C.J. Abell[2], Kyoung-Ho Kim[1], Young-Do Nam[1] and Jin-Woo Bae[1]

[1] Department of Life and Nanopharmaceutical Sciences and Department of Biology, Kyung Hee University, HoeGi-Dong 1, DongDaeMun-Gu, Seoul 130-701, Republic of Korea
[2] CSIRO, Marine and Atmospheric Research and Wealth from Oceans, National Research Flagship, Hobart, Tasmania, Australia

Recent advances in molecular biology have resulted in the application of DNA microarrays and next-generation sequencing (NGS) technologies to the field of microbial ecology. This review aims to examine the strengths and weaknesses of each of the methodologies, including depth and ease of analysis, throughput and cost-effectiveness. It also intends to highlight the optimal application of each of the individual technologies toward the study of a particular environment and identify potential synergies between the two main technologies, whereby both sample number and coverage can be maximized. We suggest that the efficient use of microarray and NGS technologies will allow researchers to advance the field of microbial ecology, and importantly, improve our understanding of the role of microorganisms in their various environments.

Introduction

Microbial ecology is a broad study of the relationship between microorganisms and their biotic and abiotic environments, predominantly comprising the analysis of abundance, composition and activity of microbial communities. This discipline is currently undergoing a paradigm shift, driven by the development and application of 'omics' technologies, including genomic and metagenomic tools [1]. With the application of these technologies, in particular next-generation sequencing (NGS) and DNA microarrays, it is apparent that the diversity and population density of microbial communities that inhabit the biosphere are much higher than previous estimates, based on traditional culture-based methods and small-subunit rRNA sequence-based surveys. We are now aware that the low-abundance microorganisms, or so-called 'rare biosphere' that is overshadowed by dominant populations, are highly diverse and largely unexplored [2] and could represent the key to ecosystem resilience.

DNA microarrays comprising hundreds or thousands of DNA fragments arrayed on small glass slides were originally developed for gene expression profiling in 1995 [3]. These were subsequently applied to the study of different aspects of microbial ecology, including methane cycling, total microbial diversity and a range of biogeochemical functions [4–7]. Alternatively, NGS approaches, including pyrosequencing (introduced by 454 Life Sciences, Inc.) as well as other platforms such as Solexa (Illumina, Inc.) and SOLiD (ABI, Inc.), provide cost-effective, rapid and highly parallel sequencing of large numbers of DNA fragments from complex samples or transcriptomes. Pyrosequencing is particularly suited to microbial ecology studies because of its relatively long read length, as compared to other NGS technologies platforms, and therefore has been widely adopted by researchers in microbial ecology [8–11], although other platforms have also been recently applied to the field [12]. Although they are still considered an effective and economical tool for the analysis of complex microbial communities, DNA microarrays are being rapidly superseded by NGS approaches [13]. Although some scientists have utilized pyrosequencing and microarray tools for pathogen genotyping [14] and for human intestinal microbiota exploration [15], Ledford noted that NGS platforms have largely infiltrated the space previously occupied by DNA microarray technologies [13]. Although these two approaches can be considered competing technologies, each tool possesses distinct complementary features suited to different applications in microbial ecology.

In this review, we compare the relative merits, weaknesses and suitability of high-throughput DNA microarrays and NGS as analytical tools in different areas of microbial ecology research. To achieve this comparison, we consider the following: (i) which technique provides the more accurate description of an unknown microbial community? (ii) Which technique is more suitable in terms of sample throughput? (iii) Which technique is more suitable in terms of cost-effectiveness, ease of sample preparation and data analysis? Furthermore, we discuss the suitability of the different approaches to particular research objectives, as well as future directions and potential synergy of these two high-throughput technologies.

Microbial ecological studies using NGS technologies and DNA microarrays

The gene that encodes 16S rRNA is currently regarded as the most versatile phylogenetic marker; it contains hypervariable regions interspersed with highly conserved

Corresponding author: Bae, J.-W. (baejw@khu.ac.kr).

0167-7799/$ – see front matter doi:10.1016/j.tibtech.2010.03.001 Trends in Biotechnology 28 (2010) 291–299

Box 1. 454 pyrosequencing in microbial ecology

The Roche (454) Genome Sequencer generates massive parallel DNA sequence reads from amplified PCR products through a sequencing-by-synthesis approach. Because this technology depends on the detection of pyrophosphate release upon nucleotide incorporation, it is also termed 454 pyrosequencing. Although the 454 pyrosequencing method is more expensive per base and requires more starting material than other next-generation DNA sequencing platforms [e.g. Illumina (Solexa) and Genome Analyzer and Applied Biosystems SOLiD], the 454 platform operates as a high-throughput sequencing tool as a result of the longer read length that is appropriate for *de novo* sequencing – that is, sequencing in which a reference sequence to be aligned with generated sequences is not required.

High-throughput pyrosequencing technology has been used in the various fields of microbial ecology, including microbial diversity and functional genes diversity. The limitation of the capacity for physically partitioning the pyrosequencing sample plate (PicoTiterPlate) for each of the samples to be analyzed has been somewhat circumvented by recent use of sample-specific key sequences called barcodes (or tags). Samples from multiple libraries can be PCR-amplified, tagged with unique barcodes, and combined and sequenced together in a single region of the PicoTiterPlate device. Each of the barcodes ligated to the beginning of a read contains a unique sequence tag that is recognized and sorted bioinformatically. Pyrosequencing technology with sample-specific key sequences was introduced in 2007 through the use of barcode sequences that consisted of 2–10 nucleotides [90,91]. Hamady *et al.* recently have applied error-correcting DNA barcodes of eight-nucleotide sequences using error-correcting codes to overcome the limit in the number of unique barcodes and enhance the ability to recognize sequencing errors that alter sample assignments [92]. In this study, a total of 1544 error-correcting barcodes, which can be tagged only to the 3′ end of the read, were selected after filtering in order to optimize PCR and sequencing performance with a bacteria-specific primer set, 27F and 338R. With this modified approach, the multiplex barcoded pyrosequencing technology allows users to perform studies of microbial ecology that involve in-depth analysis of large numbers of samples simultaneously, at a reduced cost per sample. This is particularly useful when the analysis of large quantity of samples is required as a result of the high spatial or temporal variability of the target community.

Box 2. DNA microarrays in microbial ecology

The term 'environmental microarrays' was introduced by Staffan Kjelleberg [93] in 'Microarrays for environmental studies.' This term was limited to 'microbial diagnostic microarray' in microbial ecology [4], followed by 'microbial ecological microarrays' [6]. DNA microarrays for use in microbial ecology (environmental microarrays, microbial diagnostic microarray or microbial ecological microarrays), have been developed utilizing different types of probes: oligonucleotides, cDNA and microbial genomes. The first oligonucleotide microarray applied to microbial ecology utilized nine probes that comprised 15–20mer oligonucleotides for the discrimination of nitrifying bacteria [94]. This has subsequently led to the development of the high-density 16S rRNA gene-targeting microarray that comprised ~30 000 20mer oligonucleotide probes [7] and ~300 000 25mer probes [37] for the investigation of bacterial and archaeal diversity.

cDNA microarrays, initially fabricated using randomly generated ~1-kb gene fragment probes to reveal taxonomic relationships among *Pseudomonas* species with high sensitivy [95], allow higher sensitivity and better resolution than oligonucleotide microarrays. Alternatively, community genome arrays (CGAs) or genome probing microarrays (GPMs) that employ microbial genome probes have been developed using reverse sample genome probing technology. These arrays can circumvent PCR artifacts and bias and also enhance their specificity and sensitivity. CGA and GPM have been used to characterize the complex microbial composition of soil, river and marine sediments [77], as well as fermented vegetable food monitored over the course of the fermentation process [76].

Another microarray platform called metagenome microarray (MGA) was fabricated for rapid characterization of metagenomic libraries with whole microbial and community genomes. This MGA is distinct in terms of the concept of microarray probe and target. Microarray probes are generally spotted onto a glass slide, whereas the MGA format contains microarray targets arrayed on a glass slide and uses a labeled, specific gene as a probe. The reverse approach used with MGA has made it possible to rapidly screen metagenomic libraries that comprise cosmid clones derived from a groundwater microcosm [96] and fosmid clones from marine sediment [97].

Lastly, the application of the isotope array technique for identifying groups of organisms based on their ability to use radiolabeled substrates [98] represents a significant advancement in the use of microarrays in microbial ecology, and has been used in the physiological analysis of different microbial communities in activated sludge [99].

regions, and is present in all prokaryotes, thus allowing its application to the study of microbial community structure [16]. NGS targeting these hypervariable regions of the gene that encodes 16S rRNA has been widely applied to the exploration of microbial community composition, diversity and distribution, and to the description of the roles of microbes in various ecosystems [2,12,17–20] (Box 1). NGS approaches based not only on single-gene amplification, but also on metagenome or metatranscriptome shotgun sequencing, have been used to discover novel genes and their functions in environmental samples from The Soudan Mine, Minnesota, USA [8], oceanic samples [10] and microbiota of the termite hindgut [9], as well as for metatranscriptomic analysis of marine microbial communities [21], and cultivated bacterial [22,23] and archaeal transcriptomes [24].

High-throughput microarray technology has also been applied to studies of complex microbial communities in various environments, as well as the diversity of functional genes and gene expression [4,5,25,26] using different types of microarray probes, such as oligonucleotides, cDNAs and microbial genomes (Box 2). Recently, He *et al.* developed a comprehensive microarray called GeoChip - a functional gene array that comprises >24 000 oligonucleotide probes that target thousands of functional genes, which are predominantly involved in biogeochemical cycling processes, metal resistance and contaminant degradation [27]. The GeoChip has been successively applied to the analysis of functional microbial communities in deep-sea hydrothermal vents [28] and uranium-contaminated aquifer sites [29]. Microarrays can also be applied to metatranscriptomic analysis, as shown by the application of oligonucleotide arrays to the study of transcriptomics of a phosphorus removal sludge [30].

Quantitative assessment: accuracy and depth of sample coverage

Which high-throughput platform provides a more detailed analysis of an unknown microbial community in a biological sample? The crucial point to consider when addressing this question is the nature of the technology, in that NGS and microarrays represent 'open and closed architecture systems', respectively [31] (Table 1). NGS is suitable for cataloguing gene diversity (including discovery of novel gene diversity), without *a priori* sequence information, through sequencing of 16S rRNA and functional gene

Table 1. Notable features of 454 pyrosequencing and DNA microarray technologies

	454 pyrosequencing	DNA microarray
System type	Open architecture system	Closed architecture system
Depth of sample coverage	Higher	Lower
Appropriate throughput	Lower	Higher
Ease of sample preparation	Simple to prepare PCR products or extracted DNAs	Complex preparation of target and microarray slides
Cost efficiency with multiple samples	More expensive per sample	Less expensive per sample (for short oligo arrays)
Ease of data handling/analysis	Complex annotation and sorting of massive sequence reads	Simple analysis of signal intensities
Applicability to study of species genomic relatedness	Potentially the best method with complete re-sequencing of microbial genomes	Currently the best for the DNA-DNA hybridization method
Recommended application	In-depth studies of unknown microbial community diversity	Routine studies of functional gene diversity across many samples (spatio-temporal mapping of functional gene diversity and distribution)

amplicons (given the limitations of the PCR primers used), or whole-genome shotgun sequence characterization of microbial communities. As such, NGS technologies can be considered 'open' systems. Conversely, microarrays can only account for sequences that are targeted by probes on the array, and therefore, represent 'closed' systems.

Analysis of sequence data from NGS amplicon sequencing projects has demonstrated varying sample coverage, depending on the environment examined. NGS studies of soil revealed between 1000 and 5500 distinct taxa (at 97% similarity) [19,32,33], whereas studies of human fecal samples and oral cavities have demonstrated 1400–1800 and up to 8000 taxa, respectively [12,15]. The number of taxa revealed through NGS is dependent on the targeted region of the gene that encodes 16S rRNA, the number of sequences generated, as well as the stringency of quality control imposed [12,15]. The accuracy of taxa assignment of NGS sequences using pipelines, such as the Ribosomal Database Project (RDP) classifier [34], depends on the variable region used as well as the confidence value threshold applied. The RDP classifier assigns taxonomy down to the genus level with a bootstrap confidence value [35]. Conversely, analysis of 16S rRNA and other gene compositions of environmental samples using microarray platforms such as the PhyloChip and HIT chip platforms [36,37] can only account for taxa for which there are target probes (~10 000 and 1140 distinct taxa, respectively). For microarrays that utilize multiple probes per taxa (multiple probe concept) [38], stringency and confidence of taxa assignment can be controlled by defining the number of positive probes within a taxa set required for detection of those taxa. Microarrays (i.e. the PhyloChip) can be used to identify taxa that vary in abundance by 5+ orders of magnitude [37].

Recently, a microbial ecological microarray that uses microbial genomes as probes was developed with uncultivated bacterial genomes retrieved from multiple displacement amplifications based on bacteriophage phi29 polymerase [39]. This study expanded the use of microarrays to analyze unexplored bacterial populations from single cells through serial sample dilutions. However, the application was limited to the genomes of the dominant microorganisms present in a given biological sample. In contrast, NGS permits ultra-deep sequencing with massive DNA sequence reads from amplified PCR products, allowing insight into the diversity contained within the highly divergent 'rare biosphere,' and as a result, has been applied to various environmental niches [2,19,40,41].

Sample throughput

How many samples can be analyzed appropriately with each of the two platforms? Conventional, widely applied molecular tools, such as denaturing gradient-gel electrophoresis (DGGE), fluorescence *in situ* hybridization (FISH) and terminal restriction-fragment length polymorphism (tRFLP), are unsuitable for in-depth analysis of large numbers of samples as a result of their labor-intensive and time-consuming protocols. NGS and microarray platforms are comparable in the number of samples that can be analyzed. Microarray analysis is limited only by the preparation of sufficient target DNA samples and probe-spotted microarray slides; in contrast, NGS analysis is limited by the number of samples processed in a single run by the physical partitioning or sample-specific barcoding approach utilized. The major limitations that govern the throughput of these methods are the labor, time and cost involved in sample analyses. Although it is possible to analyze many samples using multiplexed barcoded NGS, the sample size (i.e. number of reads for the gene that encodes 16S rRNA) is strongly correlated with measures of community evenness (the numerical measure of how even the biodiversity of a community is) and richness (the number of species) [42], whereas for some low-diversity genes, large sample size might not be required to account for the majority of taxa.

The sheer volume of data associated with whole metagenomic or metatranscriptomic analysis of multiple samples - as opposed to a single gene-based community analysis - can be a burden to researchers. Comprehensive functional microarrays such as the GeoChip are alternatives to NGS-based metagenomic analysis. Such arrays can be used to assess functional gene diversity and expression in large numbers of samples in a highly efficient manner. Individual probes on the array can represent a single taxonomic group or a functional gene, therefore, the analysis coverage can be estimated in advance according to the design and arrangement of probes of the microarray. Although the fabrication of microarrays designed to provide comprehensive coverage of all target genes is both time-consuming and costly, microarrays are suitable for

routine studies of the metagenome and metatranscriptome using large numbers (i.e. hundreds) of samples.

Cost-effectiveness, ease of sample preparation and data analysis

For a majority of pyrosequencing studies, researchers only need to prepare PCR products of target genes or extracted DNA from biological samples. For new applications of microarrays analysis, however, it is necessary to prepare and spot the microarray slides, as well as process sample targets for hybridization. Each of these factors can make microarray-based projects time-consuming and expensive (Figure 1). Nevertheless, if an existing, well-evaluated, probe-spotted microarray is available, the platform is suitable for the analysis of hundreds of samples, particularly for the analysis of relatively low-diversity functional genes (e.g. *pmoA*).

Using an array that comprises a few hundred different probes that target a functional gene (e.g. *pmoA* microarray analysis for methanotrophic bacteria [43]), a skillful researcher can analyze as many as 40 environmental samples per day at a low cost-per-sample ($45 USD). This results in small, easy-to-analyze data files that account for the majority of gene diversity. In contrast, the same analysis using NGS could cost more (>$15 000 USD), and involve a longer waiting time for receiving sequence data. Furthermore, the downstream analysis of the resulting pyrosequencing dataset requires sophisticated computer systems, bioinformatics tools and a significant time contribution.

Limitations of pyrosequencing and microarrays in microbial ecology

Pyrosequencing

Although NGS is widely viewed as a method for accessing the rare and unknown biosphere in microbial ecology, it has been demonstrated that the large datasets produced by pyrosequencing contain not only true sequences, but also artifactual sequences that are referred to as pyrosequencing noise [44], systematic artifacts [45] or 'wrinkles' [46]. Pyrosequencing data acquisition is based on the detection of light intensities from pyrophosphates released during nucleotide incorporation. Base-calling of the number of nucleotides within a homopolymer relies on the light intensity, which corresponds to the actual number of incorporated nucleotides [47]. The base-calling method is problematic for long homopolymers because the accuracy of the base-calling decreases as the length of the homopolymer increases. Consequently, pyrosequencing of long homopolymers frequently results in sequence errors in the form of insertions or deletions [48].

Artifactual sequences also stem from multiple templates on a single bead [47], and miscoding lesions [49]. A single base pair difference in the hypervariable region of the 16S rRNA gene sequence can result in the mis-assignment of short sequence reads. Quince *et al*. analyzed the true population diversity in a known artificial mixture of 90 pyrosequenced gene clones for 16S rRNA. As a result of at least sixfold overestimation at the 3% operational taxonomic unit level with the standard RDP pyrosequencing pipeline, it is recommended to use a noise reduction algorithm [44]. Gomez-Alvarez *et al*. reported that artificially replicated sequences from the systematic error subsisted at a high rate (11–35%) in several published and original pyrosequencing datasets [45]. By using a single strain of *Escherichia coli* as the template for 16S rRNA gene sequencing, Kunin and colleagues suggested that a combination of 97% clustering threshold and 0.2% quality trimming is necessary for reliable estimation of community diversity, and also have claimed that the use of unique reads as a diversity estimation tool can overestimate the community diversity by as many as two orders of magnitude [46].

To analyze accurately the diversity of microbial communities in environmental samples, it is necessary to eliminate artificial sequences that can lead to incorrect analyses from true sequence data generated by pyrosequencing [50]. A number of approaches to this have been described, including: Pyrobayes - an advanced base-calling program for pyrosequencing reads [48]; objective criteria to exclude low-quality reads to improve data quality [51]; the PyroNoise algorithm [44]; and a web-based 454 replicate filter [45] for artificial sequence removal. For consistent and accurate phylogenetic assignment of short reads to taxonomic information derived from full-length 16S rRNA gene sequences, Liu *et al*. recommend the automated bioinformatics pipelines [34] - Greengenes and RDP classifier with >250-bp fragments flanked by primer sets that encompass the hypervariable V2 and V3 regions of the gene that encodes 16S rRNA - for rapid, consistent and accurate taxonomic assignment with excellent coverage and recovery.

The shorter gene sequence lengths for the 16S rRNA gene (as compared to full-length gene sequences) achieved with pyrosequencing [versus those of traditional Sanger sequencing (<800 bp)], can lead to inaccurate phylogenetic assignment for 16S rRNA. This is further compounded by the PCR primer sets that target various hypervariable regions in the gene that encodes 16S rRNA, which are utilized in the pyrosequencing approach [52]. Although the 59-bp V6 hypervariable region of the 16S rRNA gene is a sufficient length for sequencing reads to classify organisms down to the genus level [53], short NGS reads, such as those produced by the original GS20 instrument (454 Life Sciences, Inc.), cannot guarantee accurate taxonomic assignment at the species level in accordance with sequenced regions of the gene [52]. Recently, the length of sequences achievable with NGS technology has increased with the new GS FLX Titanium series (Roche, USA); it is now possible to sequence more than 1 million reads (>400 bases) during a single run. This approach is still limited by the primer sets available for the gene of interest. Fragment size is a crucial factor in pyrosequencing (amplicons >500 bp do not amplify well in emulsion PCR; Roche Genome Sequencer System Application Note No. 5, 2007); therefore, primer sets that result in large fragments are not suited to pyrosequencing. This can be particularly problematic for the analysis of functional genes (e.g. nitrite reductase *nirS*/*nirK*), for which universal primer pairs that result in fragments of the correct length are often not available or do not provide comprehensive coverage.

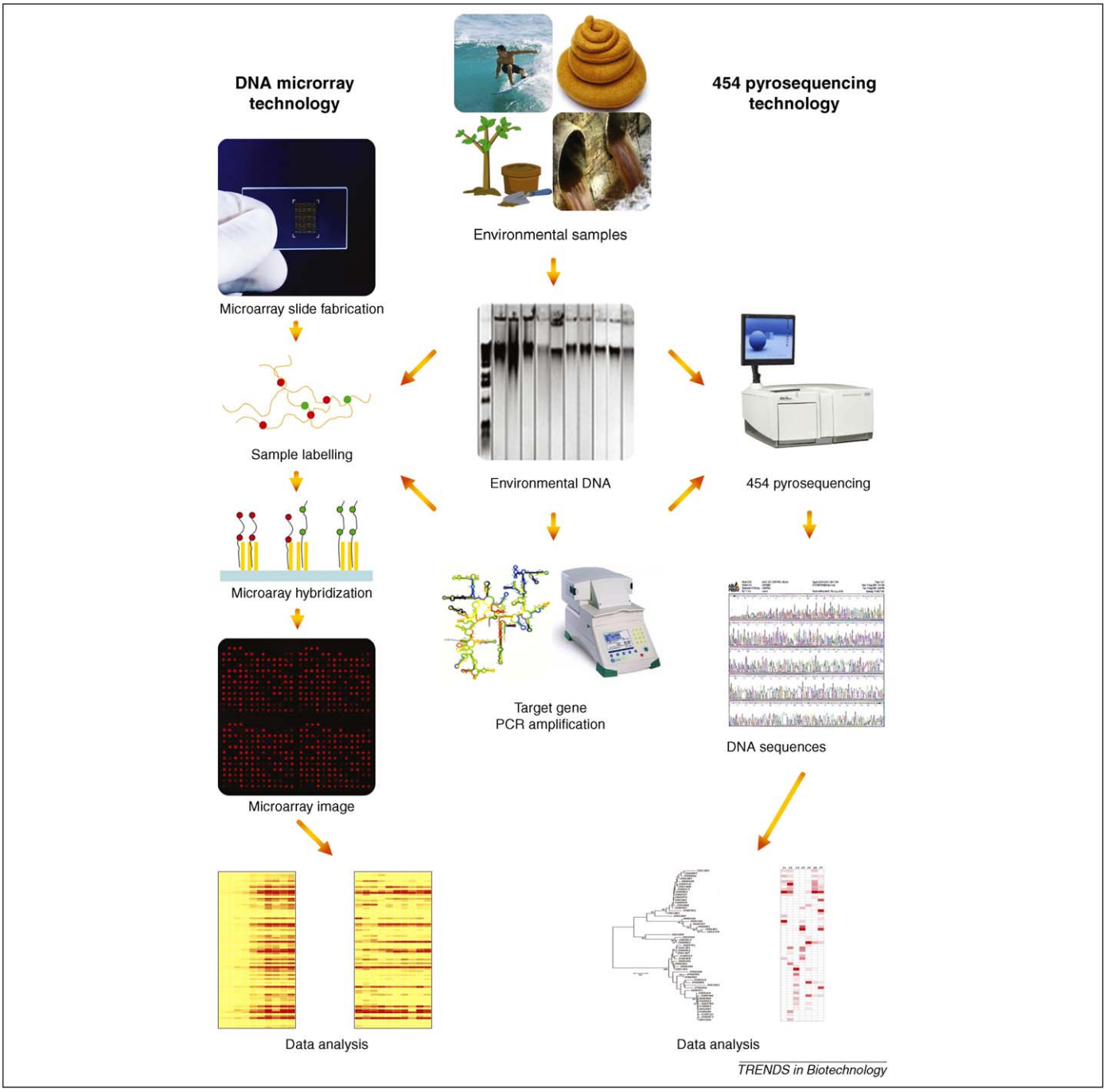

Figure 1. Schematic diagram demonstrating the experimental processes for the application of DNA microarrays and 454 pyrosequencing technologies to microbial ecology. Samples are taken from an environment that encompasses treatment, spatial or temporal scales. From these samples, nucleic acids (RNA or DNA) are extracted, and either analyzed directly with microarray or NGS technologies (metagenomic or metatranscriptomic analyses), or used as templates for PCR amplification of target genes. For microarray analysis, target material is labeled and hybridized onto a pre-fabricated microarray, followed by imaging and data analysis of the resulting images. For NGS analysis, the target template is processed (barcode incorporation, fragmentation, etc.) and applied to the appropriate NGS platform. DNA sequences are processed (quality control, bar-code binning, etc.) and analyzed using the appropriate bioinformatics approaches.

Microarrays

Nucleic acid hybridization-based methods, such as Southern and northern blotting, FISH and microarrays, are sensitive to cross-hybridization of probes to non-target sequences that result from the presence of similar sequences within a sample [54,55]. This can be alleviated to a large degree through the use of multiple probes for each target taxa [56]. Probe specificity is strongly affected by probe length in microarrays: as probe length increases, specificity decreases [57]. As a result, cDNA microarrays that utilize DNA fragments that span hundreds to a few thousand base pairs in length are highly variable in their probe lengths and specificity thresholds. Long oligonucleotides and cDNA arrays both have reported cut-off values that spann the range of 75–87% [58–60]; however, it should be noted that a 100-bp, perfectly matched fragment of a full cDNA fragment might be sufficient to yield a strong, positive signal.

For oligonucleotide microarrays that utilize short oligonucleotides (19mers), single, central, probe-target mismatches have been easily discriminated, producing a signal intensity 15–25% of the perfect match signal, whereas multiple sequence mismatches (3–5 nucleotides) are better discriminated, and show a signal intensity <5% of the perfect match signal [61]. A single-bp mismatch has also been discriminated by using non-equilibrium dissociation rates [62]. Although short oligo-probes (20–25 bp) allow significantly better specificity for identifying microorganisms present in environmental samples, they also have the disadvantage of much lower sensitivity than long oligonucleotide arrays, under the same experimental conditions [63], and therefore, require PCR amplification of the target gene. When considering the choice of oligo-probe length in terms of both specificity and sensitivity, a suitable compromise is to utilize oligo-probes with well-established, non-equilibrium, thermal dissociation for real-time hybridization analysis, thereby allowing the discrimination of perfect-match and mismatch duplexes in microarray platforms [64,65].

Common limitations: PCR-induced artifacts and bias

PCR is a preliminary step in DNA sample analysis in many microbial ecology studies that utilize traditional molecular microbiology methods as well as oligonucleotide microarrays and pyrosequencing. PCR-induced artifacts and biases have been well documented, including chimeras, mutations and heteroduplex molecules [66], and skewed template-to-product ratios [67]. As a result of the fact that artifacts and bias can lead to over- or underestimation of microbial community diversity, high-throughput technology that depends on PCR is sensitive to errors in the estimation of microbial diversity. The disadvantage of PCR-based approaches can be somewhat reduced through methods such as 'reconditioning PCR' [68], modified amplification protocols [69], the use of new specific and multiple PCR primer sets [70], or 'miniprimer PCR' that utilizes novel engineered polymerases (S-Tbr) and short 10mer primers [71].

Sequence data screening for potential chimeric sequences can be performed using tools such as the chimera check program at RDPII, Bellerophon [72] and Pintail [73]. An approach utilized by a number of groups to avoid PCR artifacts and bias is the application of pyrosequencing-based whole metagenome shotgun sequencing to the characterization of microbial communities [74,75] or the use of whole-genome hybridization that employs microarrays without PCR amplification of metagenomic target samples [76,77]. Investigation of microbial diversity without PCR artifacts and bias can be achieved through shotgun sequencing of environmental metagenomic DNA sheared into short fragments and sequenced directly by pyrosequencing. PCR artifacts can also be avoided in a microarray system by direct labeling of environmental rRNA from soil [78] and marine bacterioplankton [79]. All components of metagenomic sequences can be used in the classification through methods such as MEGAN, which utilizes BLAST results and subsequent phylogenetic assignment of sequences [80], or MLTreeMap [81], which uses a set of universally occurring marker genes to assign phylogeny. Although these approaches can help to circumvent PCR artifacts and bias, they require complex analytical methods and only a limited depth of coverage is likely to be achieved. Moreover, some organisms that provide key ecosystem functions (i.e. nitrifying bacteria and archaea) will likely be absent from the sequence data because of their rarity.

Prokaryotic species determination

Species are the basic units of the composition and diversity of microorganisms in ecological studies. The prokaryotic species concept based on genomic relatedness is crucial to the taxonomic assignment of microorganisms. Genomic relatedness among prokaryotes has been analyzed using genomic DNA–DNA hybridization (DDH) methods [82]. The 70% criterion of DDH value is considered the gold standard for the discrimination of prokaryotic species. Conventional DDH methods in which strands of DNA are hybridized onto a micro-well plate or a nylon membrane filter have certain limitations, including limited reproducibility. Chang and colleagues reported that the genome probing microarray, which comprises microbial genomes printed as probes, shows statistical superiority with reduced background signals and a higher degree of reproducibility as compared to the DDH methods [83]. As whole-genome sequencing becomes more rapid and economical, genome-comparison (re-sequencing) using NGS will substitute for conventional and microarray-based DDH methods to identify genomic relatedness among microorganisms, thus providing more comprehensive data based on entire microbial genomes.

Future directions for DNA microarrays

Recent advances in DNA sequencing technology that have resulted in novel NGS platforms can be expected to contribute to the availability of higher sequence density and longer read lengths. DNA microarray technology is also undergoing significant changes with advances in microarray instrumentation, including developments in slides, printing devices, fluidic systems for hybridization, microarray scanners and analytical software technology. In order to improve detection methods, significant effort has been directed at increasing signal detection sensitivity and reliability [84], as well as progress towards real-time monitoring of microarray hybridization events [65,85]. New methods for signal amplification using novel fluorescent agents (i.e. quantum dots and metal colloids), in addition to the development of new and improved optical, electronic and electrochemical detection systems are together expected to improve microarray sensitivity and accuracy. One such development is an oligonucleotide microarray based on the total internal reflection fluorescence detection system that utilizes the excitation of fluorescent dyes in an evanescent field; this detection method has been used to rapidly identify septicemia-causing bacteria in less than 6 h with high sensitivity, using approximately 2 ng of target DNA [86]. Optical fiber sensor arrays promise low detection limits (10 fM target), and enable monitoring of microarray hybridization in real-time [87]. Label-free methods, including imaging surface plasmon resonance and ellipsometry, do not require labeling

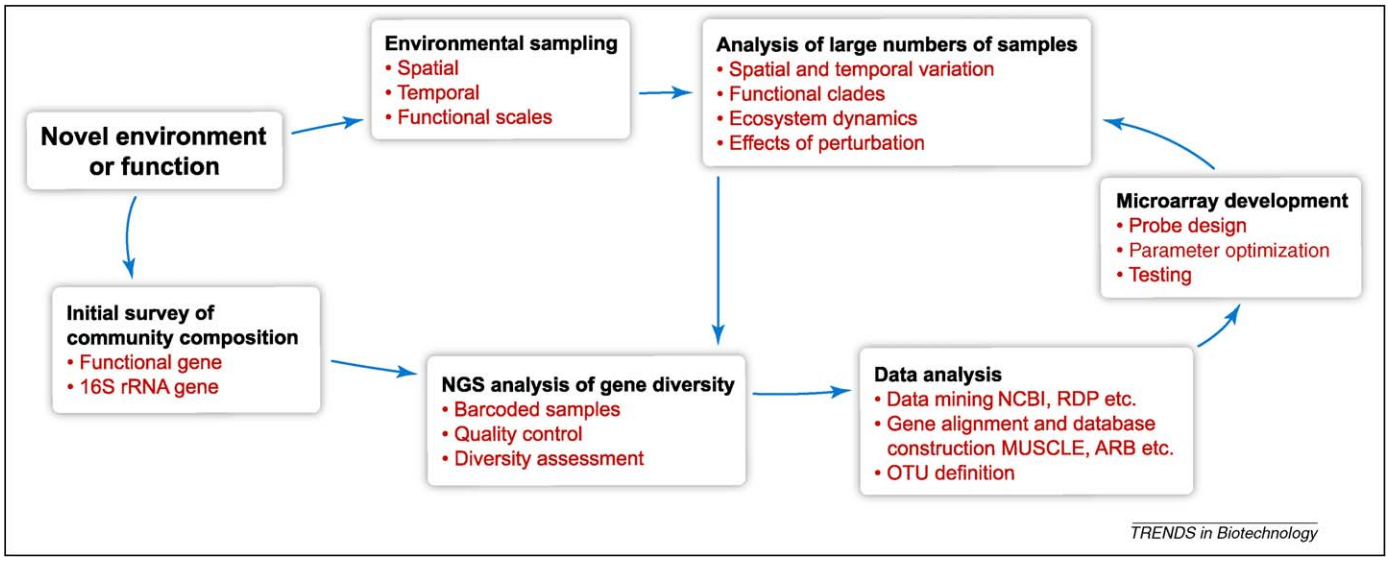

Figure 2. Proposed workflow for the study of a novel environment or functional gene. Pyrosequencing and microarray development are combined to achieve the greatest depth of gene coverage as well as the study of the environment over spatial, temporal and functional scales, thereby leading to an understanding of the ecology of a certain environment or functional group.

strategies with expensive labeling materials and are also not subject to photobleaching or quenching effects [84].

Several electronic detection systems have been developed, including cyclic and square-wave voltammetry, scanning electrochemical microscopy, detection of conduction changes of carbon nanotubes upon protein adsorption, and the scanning Kelvin nanoprobe; such electronic detection technologies pave the way for highly miniaturized and portable microarrays. Along with the innovation of these techniques, it is predicted that single-molecule sensitivity using microarrays will be a possibility in the near future [84]. Blohm and Guiseppi-Elie have predicted a single, ready-to-use piece of equipment in which these new techniques are combined to create a single instrument [88]. Recently, an integrated microarray platform called 'eSensor' was developed, which allows the operator to perform hybridization, detection and analysis on a single instrument. Compared to pyrosequencing, this electrochemical detection system consisting of an economical printed circuit board substrate and microfluidics components represents an emerging cost-effective, 'sample-to-answer' genetic analysis technology in the field of molecular diagnostics [89].

Conclusions and synergy of two high-throughput technologies

The current generation of pyrosequencing technology has the potential to replace microarrays in microbiological applications for which a few environmental samples are to be investigated in detail. However, microarray technology is a powerful method for routine studies of selected target sequences. NGS data enable a detailed understanding of microbial gene diversity, which can be used in the design of microarray probe sets. As such, the two techniques can be combined, which results in the improved target coverage of microarrays and allows confirmation of the specificity of microarray-based results (Figure 2). Recently, NimbleGen has developed 'Sequence Capture Arrays' in which target fragments hybridized to probes can be eluted and used for downstream analysis, such as sequencing. Such a combined approach should permit a detailed examination of environmental microbial communities, and could contribute significantly to improved understanding of the relationship between microorganisms and their environment through the in-depth analysis of the large number of samples that is required to cover the temporal, spatial and functional scales within a given environment.

Acknowledgements

This work was supported by the Environmental Biotechnology National Core Research Center (KOSEF: R15-2003-012-02002-0), TDPAF (Technology Development Program for Agriculture and Forestry), NMC0300837, CAER (Center for Aquatic Ecosystem Restoration) of the Eco-STAR project, 21C Frontier Microbial Genomics and Application Center Program, Eco-Technopia 21 project, the grant (09172KFDA996) from Korea Food & Drug Administration in 2009 and the Office of the Chief Executive, CSIRO. We would also like to thank Stan Robert for assistance with proofreading.

References

1 Xu, J. (2006) Microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances. *Mol. Ecol.* 15, 1713–1731
2 Sogin, M.L. *et al.* (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl. Acad. Sci. U. S. A.* 103, 12115–12120
3 Schena, M. *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467–470
4 Bodrossy, L. and Sessitsch, A. (2004) Oligonucleotide microarrays in microbial diagnostics. *Curr. Opin. Microbiol.* 7, 245–254
5 Zhou, J. (2003) Microarrays for bacterial detection and microbial community analysis. *Curr. Opin. Microbiol.* 6, 288–294
6 Bae, J.W. and Park, Y.H. (2006) Homogeneous versus heterogeneous probes for microbial ecological microarrays. *Trends Biotechnol.* 24, 318–323
7 Wilson, K.H. *et al.* (2002) High-density microarray of small-subunit ribosomal DNA probes. *Appl. Environ. Microbiol.* 68, 2535–2541
8 Edwards, R.A. *et al.* (2006) Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* 7, 57
9 Warnecke, F. *et al.* (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* 450, 560–565
10 Angly, F.E. *et al.* (2006) The marine viromes of four oceanic regions. *PLoS Biol.* 4, e368

11 Brown, M.V. *et al.* (2009) Microbial community structure in the North Pacific ocean. *ISME J.* 3, 1374–1386
12 Lazarevic, V. *et al.* (2009) Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. *J. Microbiol. Methods* 79, 266–271
13 Ledford, H. (2008) The death of microarrays? *Nature* 455, 847
14 Akhras, M.S. *et al.* (2007) PathogenMip assay: a multiplex pathogen detection assay. *PLoS ONE* 2, e223
15 Claesson, M.J. *et al.* (2009) Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. *PLoS ONE* 4, e6669
16 Tringe, S.G. and Hugenholtz, P. (2008) A renaissance for the pioneering 16S rRNA gene. *Curr. Opin. Microbiol.* 11, 442–446
17 Dowd, S.E. *et al.* (2008) Polymicrobial nature of chronic diabetic foot ulcer biofilm infections determined using bacterial tag encoded FLX amplicon pyrosequencing (bTEFAP). *PLoS ONE* 3, e3326
18 Huber, J.A. *et al.* (2007) Microbial population structures in the deep marine biosphere. *Science* 318, 97–100
19 Roesch, L.F. *et al.* (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J.* 1, 283–290
20 Roh, S.W. *et al.* (2010) Investigation of archaeal and bacterial diversity in fermented seafood using barcoded pyrosequencing. *ISME J.* 4, 1–16
21 Gilbert, J.A. *et al.* (2008) Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS ONE* 3, e3042
22 Hewson, I. *et al.* (2009) Microbial community gene expression within colonies of the diazotroph, *Trichodesmium*, from the Southwest Pacific Ocean. *ISME J.* 3, 1286–1300
23 Passalacqua, K.D. *et al.* (2009) Structure and Complexity of a Bacterial Transcriptome. *J. Bacteriol.* 191, 3203–3211
24 Wurtzel, O. *et al.* (2010) A single-base resolution map of an archaeal transcriptome. *Genome Res.* 20, 133–141
25 Lemarchand, K. *et al.* (2004) Molecular biology and DNA microarray technology for microbial quality monitoring of water. *Crit. Rev. Microbiol.* 30, 145–172
26 Loy, A. *et al.* (2002) Oligonucleotide microarray for 16S rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment. *Appl. Environ. Microbiol.* 68, 5064–5081
27 He, Z. *et al.* (2007) GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. *ISME J.* 1, 67–77
28 Wang, F. *et al.* (2009) GeoChip-based analysis of metabolic diversity of microbial communities at the Juan de Fuca Ridge hydrothermal vent. *Proc. Natl. Acad. Sci. U. S. A.* 106, 4840–4845
29 Van Nostrand, J.D. *et al.* (2009) GeoChip-based analysis of functional microbial communities during the reoxidation of a bioreduced uranium-contaminated aquifer. *Environ. Microbiol.* 11, 2611–2626
30 He, S. *et al.* (2010) Metatranscriptomic array analysis of '*Candidatus* Accumulibacter phosphatis'-enriched enhanced biological phosphorus removal sludge. *Environ. Microbiol.*, Epub ahead of print
31 Vieites, J.M. *et al.* (2009) Metagenomics approaches in systems microbiology. *FEMS Microbiol. Rev.* 33, 236–255
32 DeAngelis, K.M. *et al.* (2009) Selective progressive response of soil microbial community to wild oat roots. *ISME J.* 3, 168–178
33 Lauber, C.L. *et al.* (2009) Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl. Environ. Microbiol.* 75, 5111–5120
34 Liu, Z. *et al.* (2008) Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res.* 36, e120
35 Wang, Q. *et al.* (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267
36 Stojanovi, M.R. *et al.* (2009) Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. *Environ. Microbiol.* 11, 1736–1751
37 Brodie, E.L. *et al.* (2006) Application of a high-density oligonucleotide microarray approach to study bacterial population dynamics during uranium reduction and reoxidation. *Appl. Environ. Microbiol.* 72, 6288–6298
38 Loy, A. *et al.* (2005) 16S rRNA gene-based oligonucleotide microarray for environmental monitoring of the betaproteobacterial order "*Rhodocyclales*". *Appl. Environ. Microbiol.* 71, 1373–1386
39 Chang, H.W. *et al.* (2008) Development of microbial genome-probing microarrays using digital multiple displacement amplification of uncultivated microbial single cells. *Environ. Sci. Technol.* 42, 6058–6064
40 Turnbaugh, P.J. *et al.* (2009) A core gut microbiome in obese and lean twins. *Nature* 457, 480–484
41 Fierer, N. *et al.* (2008) The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proc. Natl. Acad. Sci. U. S. A.* 105, 17994–17999
42 Brose, U. *et al.* (2003) Estimating species richness: Sensitivity to sample coverage and insensitivity to spatial patterns. *Ecology* 84, 2364–2377
43 Bodrossy, L. *et al.* (2003) Development and validation of a diagnostic microbial microarray for methanotrophs. *Environ. Microbiol.* 5, 566–582
44 Quince, C. *et al.* (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat. Methods* 6, 639–641
45 Gomez-Alvarez, V. *et al.* (2009) Systematic artifacts in metagenomes from complex microbial communities. *ISME J.* 3, 1314–1317
46 Kunin, V. *et al.* (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ. Microbiol.* 12, 118–123
47 Margulies, M. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380
48 Quinlan, A.R. *et al.* (2008) Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat. Methods* 5, 179–181
49 Briggs, A.W. *et al.* (2007) Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl. Acad. Sci. U. S. A.* 104, 14616–14621
50 Reeder, J. and Knight, R. (2009) The 'rare biosphere': a reality check. *Nat. Methods* 6, 636–637
51 Huse, S.M. *et al.* (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* 8, R143
52 Liu, Z. *et al.* (2007) Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res.* 35, e120
53 Andersson, A.F. *et al.* (2008) Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLoS ONE* 3, e2836
54 Muller, S. *et al.* (1997) A novel source of highly specific chromosome painting probes for human karyotype analysis derived from primate homologues. *Human Genetics* 101, 149–153
55 Wren, J.D. *et al.* (2002) Cross-hybridization on PCR-spotted microarrays. *IEEE Eng. Med. Biol. Mag.* 21, 71–75
56 Loy, A. and Bodrossy, L. (2006) Highly parallel microbial diagnostics using oligonucleotide microarrays. *Clin. Chim. Acta* 363, 106–119
57 Jayaraman, A. *et al.* (2006) Computer simulation study of molecular recognition in model DNA microarrays. *Biophys. J.* 91, 2227–2236
58 Taroncher-Oldenburg, G. *et al.* (2003) Oligonucleotide microarray for the study of functional gene diversity in the nitrogen cycle in the environment. *Appl. Environ. Microbiol.* 69, 1159–1171
59 Kane, M.D. *et al.* (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res.* 28, 4552–4557
60 Wu, L. *et al.* (2001) Development and evaluation of functional gene arrays for detection of selected genes in the environment. *Appl. Environ. Microbiol.* 67, 5780–5790
61 Zhou, X. *et al.* (2004) Fabrication of DNA microarrays on nanoengineered polymeric ultrathin film prepared by self-assembly of polyelectrolyte multilayers. *Langmuir* 20, 8877–8885
62 Urakawa, H. *et al.* (2003) Optimization of single-base-pair mismatch discrimination in oligonucleotide microarrays. *Appl. Environ. Microbiol.* 69, 2848–2856
63 Relogio, A. *et al.* (2002) Optimization of oligonucleotide-based DNA microarrays. *Nucleic Acids Res.* 30, e51
64 Liu, W.T. *et al.* (2001) Optimization of an oligonucleotide microchip for microbial identification studies: a non-equilibrium dissociation approach. *Environ. Microbiol.* 3, 619–629
65 Pozhitkov, A.E. *et al.* (2007) Revision of the nonequilibrium thermal dissociation and stringent washing approaches for identification of mixed nucleic acid targets by microarrays. *Nucleic Acids Res.* 35, e70

66 Qiu, X. *et al.* (2001) Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. *Appl. Environ. Microbiol.* 67, 880–887
67 Polz, M.F. and Cavanaugh, C.M. (1998) Bias in template-to-product ratios in multitemplate PCR. *Appl. Environ. Microbiol.* 64, 3724–3730
68 Thompson, J.R. *et al.* (2002) Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by 'reconditioning PCR'. *Nucleic Acids Res.* 30, 2083–2088
69 Acinas, S.G. *et al.* (2005) PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Appl. Environ. Microbiol.* 71, 8966–8969
70 Hong, S. *et al.* (2009) Polymerase chain reaction primers miss half of rRNA microbial diversity. *ISME J.* 3, 1365–1373
71 Isenbarger, T.A. *et al.* (2008) Miniprimer PCR, a new lens for viewing the microbial world. *Appl. Environ. Microbiol.* 74, 840–849
72 Huber, T. *et al.* (2004) Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics* 20, 2317–2319
73 Ashelford, K.E. *et al.* (2005) At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl. Environ. Microbiol.* 71, 7724–7736
74 Tito, R.Y. *et al.* (2008) Phylotyping and functional analysis of two ancient human microbiomes. *PLoS ONE* 3, e3703
75 Dinsdale, E.A. *et al.* (2008) Microbial ecology of four coral atolls in the Northern Line Islands. *PLoS ONE* 3, e1584
76 Bae, J.W. *et al.* (2005) Development and evaluation of genome-probing microarrays for monitoring lactic acid bacteria. *Appl. Environ. Microbiol.* 71, 8825–8835
77 Wu, L. *et al.* (2004) Development and evaluation of microarray-based whole-genome hybridization for detection of microorganisms within the context of environmental applications. *Environ. Sci. Technol.* 38, 6775–6782
78 Small, J. *et al.* (2001) Direct detection of 16S rRNA in soil extracts by using oligonucleotide microarrays. *Appl. Environ. Microbiol.* 67, 4708–4716
79 Peplies, J. *et al.* (2004) Application and validation of DNA microarrays for the 16S rRNA-based analysis of marine bacterioplankton. *Environ. Microbiol.* 6, 638–645
80 Huson, D.H. *et al.* (2007) MEGAN analysis of metagenomic data. *Genome Res.* 17, 377–386
81 von Mering, C. *et al.* (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* 315, 1126–1130
82 Wayne, L.G. *et al.* (1987) Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *Int. J. Syst. Bacteriol.* 37, 463–464
83 Chang, H.W. *et al.* (2008) Statistical superiority of genome-probing microarrays as genomic DNA-DNA hybridization in revealing the bacterial phylogenetic relationship compared to conventional methods. *J. Microbiol. Methods* 75, 523–530
84 Bally, M. *et al.* (2006) Optical microarray biosensing techniques. *Surf. Interface Anal.* 38, 1442–1458
85 Marcy, Y. *et al.* (2008) Innovative integrated system for real-time measurement of hybridization and melting on standard format microarrays. *Biotechniques* 44, 913–920
86 Francois, P. *et al.* (2006) Rapid bacterial identification using evanescent-waveguide oligonucleotide microarray classification. *J. Microbiol. Methods* 65, 390–403
87 Ferguson, J.A. *et al.* (2000) High-density fiber-optic DNA random microsphere array. *Anal. Chem.* 72, 5618–5624
88 Blohm, D.H. and Guiseppi-Elie, A. (2001) New developments in microarray technology. *Curr. Opin. Biotechnol.* 12, 41–47
89 Liu, R.H., and Longiaru, M. (2009) ESensor: An electrochemical detection-based DNA microarray technology enabling sample-to-answer molecular diagnostics. In *Proceedings of SPIE - The International Society for Optical Engineering*
90 Binladen, J. *et al.* (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE* 2, e197
91 Parameswaran, P. *et al.* (2007) A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Res.* 35, e130
92 Hamady, M. *et al.* (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat. Methods* 5, 235–237
93 Kjelleberg, S. (2002) Environmental biotechnology. *Curr. Opin. Biotechnol.* 13, 199–203
94 Guschin, D.Y. *et al.* (1997) Oligonucleotide microchips as genosensors for determinative and environmental studies in microbiology. *Appl. Environ. Microbiol.* 63, 2397–2402
95 Cho, J.C. and Tiedje, J.M. (2001) Bacterial species determination from DNA-DNA hybridization by using genome fragments and DNA microarrays. *Appl. Environ. Microbiol.* 67, 3677–3682
96 Sebat, J.L. *et al.* (2003) Metagenomic profiling: microarray analysis of an environmental genomic library. *Appl. Environ. Microbiol.* 69, 4927–4934
97 Park, S.J. *et al.* (2008) Metagenome microarray for screening of fosmid clones containing specific genes. *FEMS Microbiol. Lett.* 284, 28–34
98 Adamczyk, J. *et al.* (2003) The isotope array, a new tool that employs substrate-mediated labeling of rRNA for determination of microbial community structure and function. *Appl. Environ. Microbiol.* 69, 6875–6887
99 Hesselsoe, M. *et al.* (2009) Isotope array analysis of *Rhodocyclales* uncovers functional redundancy and versatility in an activated sludge. *ISME J.* 3, 1349–1364