Phylogenetic Microarrays

Oleg Paliy, Vijay Shankar and Marketa Sagova-Mareckova

9

Abstract

Environmental microbial communities are known to be highly diverse, often comprising hundreds and thousands of different species. Such great complexity of these populations, as well as the fastidious nature of many of the microorganisms, makes culture-based techniques both inefficient and challenging to study these communities. The analyses of such communities are best accomplished by the use of high-throughput molecular methods such as phylogenetic microarrays and next generation sequencing. Phylogenetic microarrays have recently become a popular tool for the compositional analysis of complex microbial communities, owing to their ability to provide simultaneous quantitative measurements of many community members. This chapter describes the currently available phylogenetic microarrays used in the interrogation of complex microbial communities, the technology used to construct the arrays, as well as several key features that distinguish them from other approaches. We also discuss optimization strategies for the development and usage of phylogenetic microarrays as well as data analysis techniques and available options.

Introduction

Microbes inhabit diverse environments. Some of these environments include the human intestinal tract and skin, soil, roots, leaf and bark surfaces of plants, ocean waters, deep see vents, and air. The ecosystems of such environments are populated by communities of microorganisms, rather than by individual species, and often contain hundreds and even thousands of different microbial members. Many of these communities play pivotal roles in ecosystem processes such as energy flow, elemental cycling, and biomass production. Energy and nutrients in these systems are processed by intricate networks of metabolic pathways through multiple community members (Duncan *et al.*, 2004; Belenguer *et al.*, 2006; Flint *et al.*, 2008; De Vuyst and Leroy, 2011). The sheer complexity of such networks and the difficulty involved in culturing the individual members of these communities have challenged researchers who have tried to gain a clearer understanding of these interactions. Recent advances in molecular technologies have significantly simplified the analysis of these communities because they remove the need to culture and grow community members individually. Some of the currently available molecular techniques include high-throughput sequencing (discussed in chapter 8 of this book), terminal restriction fragment length polymorphism (discussed in Chapter 6), chequerboard DNA–DNA hybridization, quantitative real-time PCR, fluorescence *in situ* hybridization, and phylogenetic microarrays. Phylogenetic interrogation of small subunit ribosomal RNA (SSU rRNA) molecules using these techniques has led to considerable progress in our understanding of community structure and dynamics of various microbial ecosystems (Suau, 2003; Sekirov *et al.*, 2010). Phylogenetic microarrays, one of the more popular choices among these techniques, have been successfully used to quantitatively profile a variety of microbial communities, including the gastrointestinal tract, sewage sludge, soil, and air (Brodie *et al.*, 2007; Nemir *et al.*, 2010; Val-Moraes *et al.*, 2011; Rigsbee *et al.*, 2012).

Although gene expression analysis was the original motivation behind the development of microarrays, their versatility has allowed researchers to adapt this technology for other uses, including phylogenetic analysis. Several types of microarrays have been developed to characterize the composition and function of microbial communities, including community genome arrays, functional gene arrays, and phylogenetic microarrays. Community genome arrays are constructed using whole-genomic DNA isolated from species in pure culture. They allow detection of individual species and strains in simple and complex communities. Functional gene arrays include probes to genes encoding important enzymes involved in various metabolic processes and are useful for monitoring physiological changes in microbial communities (Waldron *et al.*, 2009; Xie *et al.*, 2010). A good example of a functional gene array is the GeoChip, which contains tens of thousands of oligonucleotide probes for genes involved in biogeochemical cycling of carbon, nitrogen, phosphorus, and sulfur, for genes involved in metal and antibiotic resistance, and for genes coding proteins involved in bioremediation of organic compounds (Zhou *et al.*, 2011). Phylogenetic oligonucleotide microarrays (phyloarrays) contain probes complementary to well conserved and ubiquitous gene sequences (usually the SSU rRNA gene) and are primarily used for the analysis of microbial community composition and variability (Paliy and Agans, 2012). Among different array types, phyloarrays are currently the most popular owing to the availability of a large set of near-full length SSU rRNA sequences deposited in NCBI, EMBL, RDP, and Greengenes databases (see also Chapter 7, 'Repositories of 16S rRNA gene sequences and taxonomies').

The first recognized phylogenetic microarray, developed by Guschin *et al.* (1997), was capable of detecting select genera of nitrifying bacteria. Since then, significant advances have been made with phylogenetic microarrays to improve the breadth of detection (total number of different groups detected), thereby increasing their versatility. Progress has also been made to increase the sensitivity and specificity of phylogenetic microarrays (Hazen *et al.*, 2010; Paliy and Agans, 2012). In this chapter, we will discuss the current developments in the technology, optimization of usage, applications, and potential future trends in the use of phylogenetic microarrays.

Current phylogenetic microarrays

The high-throughput and quantitative nature of phylogenetic microarrays makes them an excellent solution for researchers who seek to determine the composition of their microbial community of interest. Some key features that distinguish different phylogenetic microarrays are the choices of phylogenetic markers utilized for probe design and the experimental platform used to host these probes (Paliy and Agans, 2012). A gene or a group of genes that are ubiquitously present among all or at least the majority of species of interest often make the best target for phylogenetic analysis. A few already utilized examples that fit the above criteria include the SSU rRNA gene (16S in prokaryotes and 18S in eukaryotes), the large ribosomal subunit RNA gene (23S and 28S, respectively), genes coding for the heat shock proteins GroEL and GroES and for ribosomal proteins such as protein S1 (Martens *et al.*, 2007), and in the case of methanogens, the *mcrA* gene which encodes for methyl coenzyme-M reductase (Luton *et al.*, 2002). The SSU rRNA gene is currently the most popular choice in part because it can be fully and selectively amplified from total genomic DNA with a set of primers complementary to the conserved regions at the beginning and the end of the gene. Note, however, that the 16S rRNA gene has substantial limitations as a taxonomic marker when attempting to discriminate between closely related taxa, i.e. below the genus level. This is due to a high level of conservation of this gene sequence across bacterial taxa (Naum *et al.*, 2008). As an alternative to rRNA gene, apart from using the genes mentioned above, one can also utilize more specific metabolic genes for a particular community of interest. For example, to study methanotrophs, methane monooxygenase (*pmoA*) gene can be used (Bodrossy *et al.*, 2003; Stralis-Pavese *et al.*, 2011), while *nifH* gene coding for a component of nitrogenase protein complex can be utilized to profile nitrogen-fixing diazotrophic populations (Zhang *et al.*, 2007).

A typical design process for a microarray specific to a particular ecosystem or community usually involves the acquisition of 16S rRNA genes from members of that community (through clone library sequencing, for example) and subsequent selection of regions within the genes for probe design. Region selection can either be done manually, based on the availability of unique fragments in the hypervariable regions of 16S rRNA sequence, or by using mathematical algorithms. Several software solutions such as ARB, GoArray and PhylArray exist to facilitate this process and provide an optimized automated design of microarray probes (Ludwig *et al.*, 2004; Rimour *et al.*, 2005; Militon *et al.*, 2007). Several technologies are available for the construction of phylogenetic microarrays. A currently popular choice, developed by Affymetrix, Inc. (USA), is to build arrays by probe chemical synthesis through photolithography. In this technique, oligonucleotide probes are directly synthesized on the array glass surface, one nucleotide at a time, using light activation and masking plates. In each round, a light mask is applied to the surface of the array which allows only specific growing oligo sequences to incorporate a particular new nucleotide. After many rounds of masking and nucleotide addition through light activation (typical oligonucleotide length is 20–25 bp), the desired probes are constructed to generate a high-density microarray (Pease *et al.*, 1994). Although expensive to produce compared to other available techniques, the Affymetrix arrays are consistent between batches, have high probe density on the array surface, and display low technical variability (Zakharkin *et al.*, 2005).

In contrast, some laboratories prefer to create 'in-house' glass slide microarrays, where fully constructed oligonucleotide or DNA probes are deposited onto array surfaces using fine-point needles and robotics. The oligo or DNA probes are made and stored in solution, and each individual probe is deposited onto a specific glass surface location (spot) as a small drop. The drops are dried, the probes are subsequently attached to the glass surface, and the microarray is ready for use (Goldmann and Gonzalez, 2000). In addition to the usual glass slide surface, membrane surfaces are sometimes used instead. This microarray construction approach allows for a high level of customization and adaptation. Because no metal masks are required, the array design can be updated frequently, and only a limited number of the arrays can be created at any given time. One of the commercial microarray manufacturers, Agilent, Inc. (USA), uses the process of ink-jet printing to print as many as 185,000 features onto a 1 × 3 inch slide. Recently, microelectrodes have also been used to construct high-density arrays, where probes are aligned and concentrated on the array surface using electrical charges applied to specific sections of the array. This technique reduces the amount of time and labour required for the construction of microarrays. More importantly, due to the fact that oligonucleotide molecules can be concentrated on a small region of the array, this technique permits the use of lower amounts of probe target fragments that are added to the microarray during hybridization (Heller *et al.*, 2000). With the exception of photolithographic synthesis, where each oligonucleotide probe is anchored onto the surface prior to synthesis, the other microarray construction techniques require binding of probes to the array surface. Often, the oligonucleotide probes are chemically bonded onto the microarray surface covered with a coat of silane containing an active functional group (Chiu *et al.*, 2003).

In recent years, significant improvements have been achieved in the design of phylogenetic microarrays, including improvements in the breadth of detection, sensitivity, and specificity. Table 9.1 lists some of the currently available phylogenetic microarrays together with their design parameters and targeted communities. The original phylogenetic microarray designed by Guschin *et al.* was capable of detecting a few genera of nitrifying bacteria (Guschin *et al.*, 1997). The breadth of detection was expanded on the microarray developed by Wang and colleagues to include 40 predominant members of human gut microbiota (Wang *et al.*, 2004). The current leader in the total number of potentially detectable groups, the third generation (G3) PhyloChip array, has been designed to detect as many bacterial phylotypes as possible (Brodie *et al.*, 2006; Hazen *et al.*, 2010). This microarray is based on the Affymetrix GeneChip technology and contains 1.1 million

Table 9.1 A selection of current phylogenetic microarrays

Array name	Target community	Resolution	Technology	Detectable groups	Reference
PhyloChip	All prokaryotes	Varied Species	Photolithography	9000 phylotypes (G2) 50,000 phylotypes (G3)	Brodie *et al*. (2006), Hazen *et al*. (2010)
Microbiota Array	Human intestinal biota	Species	Photolithography	775 phylotypes	Paliy *et al*. (2009)
HOMIM	Human oral biota	Species	Aldehyde slide	272 phylotypes	Preza *et al*. (2009)
V-Chip	Human vaginal biota	Varied	Activated polymer slide	350 groups	Dols *et al*. (2011)
TCE Chip	Soil biota	Varied	Aldehyde slide	742 groups	Nemir *et al*. (2010)
EcoChip	Sewage sludge biota	Species	Amine slide	1,560 phylotypes	Val-Moraes *et al*. (2011)
RHC-PhyloChip	Activated sludge biota	Varied	Aldehyde slide	79 groups	Hesselsoe *et al*. (2009)
Genome Proxy	Marine biota	Species	Poly-L-lysine slide	14 phylotypes	Rich *et al*. (2008)

25-mer probes arranged in a grid of 1,008 rows by columns, with an approximate probe density of 10,000 molecules per μm^2. The array is capable of detecting approximately 50,000 phylotypes (the previous version of the array, G2, contained 500,000 probes and was able to detect approximately 9000 phylotypes) (Brodie *et al.*, 2006). This increase in the breadth of detection allows for wide range applications, evidenced by the recent use of PhyloChip in profiling coastal salt marsh, coral, and several human-associated microbial communities (Cox *et al.*, 2010; Lemon *et al.*, 2010; Wu *et al.*, 2010b; Deangelis *et al.*, 2011; Mendes *et al.*, 2011).

The growing interest in the human-associated microbiota has led to the development of several microarrays designed to detect and profile specific human microbial communities. The Microbiota Array, also based on the Affymetrix photolithography technology, was designed to profile microbiota of the human gastrointestinal tract. The array contains 16,223 probes, with multiple probe sets allowing detection and quantification of 775 different human intestinal microbial phylotypes. Each probe set detects a single phylotype (also called operational taxonomic unit or phylogenetic species) and contains between 5 and 11 different probes to that phylotype's 16S rRNA sequences. The Microbiota Array also takes advantage of the Affymetrix microarray design to contain both perfect match probes (provide target quantification) as well as mismatch probes (estimate cross-hybridization amount removed during normalization of probe signals) for each interrogated phylotype. This phyloarray can detect phylotypes that are present at an overall community abundance of less than 0.001% (Paliy *et al.*, 2009). To date the Microbiota Array has been used successfully to accurately profile the microbial communities of the distal gut in healthy adults, adolescents, and adolescents with irritable bowel syndrome (Agans *et al.*, 2011; Rigsbee *et al.*, 2012).

The HOMIM (Human Oral Microbial Identification Microarray), an aldehyde-coated glass-slide microarray, was designed to detect 272 microbial phylotypes from human oral cavity through the interrogation of the 16S rRNA gene. The reverse capture probes in this array consist of 18–20 nucleotides complementary to the target sequence with a spacer sequence of eight thymidines and a 5′-(C6)-amine-modified base for attachment to the slide. The oligonucleotide probes are printed onto a 25 mm × 76 mm aldehyde slide. Each array is separated into five sections to facilitate the parallel processing of five samples, making the overall process more cost effective (Preza *et al.*, 2009). This array has been an effective tool in detecting and profiling the oral microbiota in multiple studies, spanning several

disease states as well as examining oral microbiota in healthy hosts (Preza *et al.*, 2009; Docktor *et al.*, 2012; Luo *et al.*, 2012).

The V-Chip, also called the vaginal microbiota-representing microarray, is another spotted microarray that utilizes polymer-coated slides to house oligonucleotide probes. The array is constructed by employing a high precision robotic dispenser with fine-point quill pins to deliver oligonucleotide probes onto a slide surface. The probes contain a 5′-NH_2-C6 terminal region that is used in the probe attachment. The array surface is coated with a proprietary activated polymer that is responsible for the binding of the probes to the array. The V-Chip array contains a total of 459 probes allowing for the detection of 350 vaginal microbial groups that are spread across multiple taxonomic levels (from species to order level) (Dols *et al.*, 2011). This phylogenetic microarray was designed to profile human vaginal microbiota, and has demonstrated its effectiveness as a diagnostic tool for profiling changes in microbial communities in diseased states such as bacterial vaginosis (Dols *et al.*, 2011).

Several microarrays targeting different soil microbial communities have also been recently developed. A prototype microarray composed of 122 oligonucleotide probes 20 to 25 nt in length was designed to target known microbes from plant rhizospheres, which mostly included representative taxa of Alphaproteobacteria at various taxonomic levels from phyla to species. This microarray was utilized to compare maize rhizospheres and bulk soil samples (Sanguin *et al.*, 2006). This array was further expanded to include 1033 probes targeting specific rhizosphere bacteria known for plant growth promoting or disease suppressive characteristics. It was capable of discriminating between disease suppressive and disease conducive soils for tobacco black root rot (Kyselkova *et al.*, 2009) and wheat take-all disease (Schreiner *et al.*, 2010). A subset of probes from this microarray (113 oligonucleotide probes targeting Actinobacteria, particularly genera known for production of secondary metabolites) was employed in a spatial–temporal study of Actinobacteria in a waterlogged forest (Kopecky *et al.*, 2011). Finally, the same microarray additionally enriched with Gammaproteobacteria and *Pseudomonas* probes has recently been used to assess microbial community structure perturbation as a result of exposure to 1 ppm of trichloroethylene. Microbial groups specifically sensitive to the trichloroethylene addition were determined (Nemir *et al.*, 2010).

The EcoChip, an alternative soil microbiota phyloarray, was developed based on the 16S rRNA clone libraries obtained from different soil types. The clones were chosen from a bank of metagenomic DNA from soil microorganisms. The PCR amplicons (300 to 1000 bp long) were used in replicates for the microarray construction. PCR products were printed on glass slides treated with aminosilane. In total, the EcoChip contains 1,560 distinct partial 16S rRNA gene fragments from soil microorganisms; 43 partial sequences of 18S rRNA genes from fungi were printed to serve as a negative control. This microarray was able to distinguish bacterial communities between various soil sites and could determine the effect of sewage sludge addition on the respective soil bacterial community (Val-Moraes *et al.*, 2011).

Uncultivated microbial phylotypes and their close relatives from marine environments can also be studied with phylogenetic microarrays. To construct a prototype Genome Proxy microarray, probe sets to 14 of the sequenced genome fragments and to genomic regions of the cultivated cyanobacterium *Prochlorococcus* MED4 were designed. Genome fragments consisted of sequenced clones from large-insert genomic libraries from microbial communities in Monterey Bay, the Hawaii Ocean Time station ALOHA, and Antarctic coastal waters. Each probe set contained multiple 70-mers, each targeting an individual open reading frame, and distributed along 40–160 kbp contiguous genomic region. This prototype array correctly identified the presence or absence of the target organisms and their relatives in laboratory mixes, with negligible cross-hybridization to organisms with ≤75% genomic identity (Rich *et al.*, 2008). Furthermore, this microarray can be used for tracking microbial community and population changes in marine environments over time to provide a higher-resolution understanding of the dynamics of marine microbial communities (Rich *et al.*, 2008).

An 'isotope' microarray approach has been developed to allow the measurement of incorporation of labelled substrate into the rRNAs of community members. For this purpose, a 16S rRNA-targeting microarray, RHC-PhyloChip, consisting of 79 nested oligonucleotide probes to most cultured and uncultured Rhodocyclales, was used. The diversity and ecophysiology of Rhodocyclales in activated sludge from a full-scale wastewater treatment plant were analysed. RHC-PhyloChip analysis was performed with fluorescently labelled and fragmented RNA from each activated sludge subsample that was incubated with $^{14}CO_2$ and allylthiourea under different conditions. An activity and substrate-utilization profile of the different Rhodocyclales groups in the activated sludge was created to distinguish between the active and dormant communities (Hesselsoe *et al.*, 2009).

There are several features to take into account when comparing different phylogenetic microarrays. As seen in Table 9.1, microarrays differ in the technology used. The Microbiota Array and the PhyloChip were developed using photolithographic synthesis, which has several advantages including the high degree of efficiency, uniformity, and probe density (Brodie *et al.*, 2006; Paliy *et al.*, 2009). The Affymetrix platform takes advantage of the high probe density to allow these arrays to contain multiple probes per target (phylotype) as well as to enable allocation of mismatch probes that provide means to adjust for target cross-hybridization (Rigsbee *et al.*, 2011). On the other hand, ink-jet and fine-point needle printing allow for cost-effective production and modification of microarrays since expensive tools such as photolithographic masks are not required. Printing on glass slides is still considered the most cost-efficient method currently available. However, the drawback of this type of array manufacturing is the loss of uniformity; therefore, these arrays require more extensive validation tests before they are ready for application.

Phylogenetic microarrays are also distinguished based on their resolution. In order to achieve the degree of resolution seen with Sanger sequencing, a species- or OTU- (operational taxonomic unit) level specificity is required. Profiling communities at this depth allows researchers to understand species-level interactions such as metabolic interdependencies and co-pathogenicity. In many cases the ability of microarrays to measure phylotype abundance is dependent on the complexity of the target community, and several of the currently available microarrays are capable of profiling microbial communities at the phylotype level (Table 9.1). Breadth of detection is yet another variable that differentiates phylogenetic microarrays (Paliy and Agans, 2012). The PhyloChip is an excellent example of a phyloarray specifically designed to detect as many microbial phylotypes as possible across the bacterial and archaeal domains. Its detection breadth makes this phyloarray very versatile, enabling its usage in many environmental and clinical studies. The downside to this type of design strategy is a potential for the high number of false positives due to off-target hybridizations induced by the high number of probes (Midgley *et al.*, 2012). The issue of false positives and cross-hybridization can be ameliorated by optimizing the probe selection process and by assigning strict criteria for signal presence, though a complete resolution of the problem is very difficult. Contrary to such design, phylogenetic microarrays designed for specific communities, such as the Microbiota Array and EcoChip, benefit from the reduced cross-hybridization potential to provide robust estimates of community structure, while maintaining the ability to discriminate different communities with similar efficiency (Kyselkova *et al.*, 2009). The most powerful microarrays might be those that target a very particular microbial community or microbial taxonomic group (Genome Proxy array or RHC-PhyloChip) and thus can be employed to directly test a specific hypothesis.

Phylogenetic microarrays based on non-traditional techniques have also been described in several reports. For example, a fragment ligation reaction based DNA microarray has been developed by Candela *et al.* (Candela *et al.*, 2010). The microarray design involves the use of pairs of oligonucleotides complementary to the adjacent regions of each target sequence. One of the oligonucleotides contains a 5′-fluorescent label and the other has a unique 'zip-code' sequence. The oligonucleotide pair is ligated together only in the presence of the complementary target sequence

binding to both oligos. Since the ligation is carried out by highly selective ligase enzyme, a high level of probe specificity can be achieved with the use of this approach. The quantification of the fluorescently labelled ligated products is accomplished by the use of specially designed 'universal' detection array that houses probes complementary to the tag ('zip-code') sequences present within the ligated products. These universal arrays allow for uniform hybridization conditions and for the use of different ligation probe sets unique to each interrogated community, which enables flexible experimental design. A prototype ligation array developed by Candela and co-workers is capable of quantifying 30 groups of human intestinal microbiota, and the array was used to profile the faecal microbiota of several young adults (Candela *et al.*, 2010). Another non-traditional microarray, referred to as the restriction site tagged microarray, was developed by Zabarovsky *et al.* (2003). The array design was accomplished by developing tag sequences that are complementary to the regions flanking the recognition site of a rare-cutting restriction enzyme. A set of these tags represents a 'passport' for a particular phylotype. In the experimental protocol, genomic DNA is first digested by the restriction enzyme and is allowed to hybridize to tag sequences on the array. Quantification of the hybridization is accomplished through detection of the labelled products. Phylotype differentiation is achieved by constructing a custom microarray containing 'passport' sequences complementary to the enzyme site flanking regions from each phylotype genome. This type of array design allows for the differentiation of even closely related phylotypes. Finally, large subunit ribosomal RNA gene based phylogenetic microarrays have also been developed successfully (Mitterer *et al.*, 2004; Yoo *et al.*, 2009). For example, Mitterer *et al.* (2004) developed a custom glass-slide array that contained genus- and species-specific solid phase primers targeting a single variable region of the 23S rRNA gene (Mitterer *et al.*, 2004). Using universal primers, genomic DNA from environmental samples was subjected to PCR amplification on the glass-slide. The generated PCR products were allowed to bind to the group-specific primers for subsequent elongation accompanied by the incorporation of biotin labelled nucleotides (Mitterer *et al.*, 2004). Quantification was based on fluorescence scanning of the hybridized probe–target pairs. This array was successfully used to identify bacterial communities in cervical swab samples at a high resolution (Mitterer *et al.*, 2004).

Optimization of phylogenetic microarrays

Phylogenetic microarrays provide several advantages over some of the other currently available techniques used for profiling microbial communities. These include a highly quantitative nature of the acquired data, an ability to analyse one sample at a time, a short processing time, and an opportunity for multi-probe interrogation of each community member (Paliy and Agans, 2012). Phylogenetic microarrays can be used to identify taxa that vary in abundance by over five orders of magnitude (Roh *et al.*, 2010). Above that, due to a frequent hierarchical organization of microarray probes, the precision of identification is relatively high, and different taxonomic levels of probe targets enable a more comprehensive view of the community structure. Although these attractive features make phylogenetic microarrays a viable option for phylogenetic analysis, there are also some limitations to the technology that must be addressed. Firstly, phylogenetic microarrays typically do not allow for the detection of novel phylotypes. They are only capable of detecting and quantifying phylotypes to which they contain probes. Secondly, microarrays are technically demanding to design, use, and analyse, and thus require rigorous testing, validation, and optimization (Hashsham *et al.*, 2004). To help with the second limitation, a number of approaches that improve the robustness of microarray data have been developed and are discussed below.

Optimization of probe design and hybridization

The design of phylogenetic microarrays requires extensive knowledge and experience in probe selection. A lack of a rigorous probe selection process can lead to issues such as high level of fragment cross-hybridization, which can result in inaccurate or biased community profiles. There are several variables that control the probe–target

hybridization process and the subsequent estimation of signal. One such variable, the size of the probe oligonucleotide or DNA fragment, has a large influence on the hybridization behaviour. In general, the length of the probe is positively correlated with hybridization chance (sensitivity) and is negatively correlated with hybridization specificity (Suzuki *et al.*, 2007). Selecting probes that are small can lead to high specificity but at the cost of low hybridization sensitivity. On the other hand, picking long probes can increase the sensitivity of detection, but risks hybridization of smaller unrelated fragments to each probe. An ideal probe length provides a balance between a high sensitivity and high specificity. Oligonucleotides of lengths between 20 and 30 nucleotides are generally selected in many phylogenetic microarray designs (Brodie *et al.*, 2006; Paliy *et al.*, 2009).

The melting temperature of each probe-target duplex (T_m) is another important variable that should be taken into consideration when designing probes. Since the hybridization efficiency at any given temperature depends on the sequence T_m, it is important to constrict the melting temperatures of all of the probes to a relatively narrow range (He *et al.*, 2005). The resulting consistency will reduce probe hybridization bias due to T_m variability, thereby increasing the validity of the acquired signals. While designing probes for phylogenetic microarrays, it is also important to consider the optimal choice of probe targets. Most phylogenetic microarrays use the SSU rRNA gene for identification and taxonomic analysis of community members. While much of the 16S rRNA gene sequence is highly conserved, the gene contains nine sections commonly referred to as the 'hypervariable' (V) regions that display considerable sequence variability among different microbes (Chakravorty *et al.*, 2007) (Fig. 9.1; see also Chapter 7, 'Marker gene experiments'). Phylogenetic studies tend to exploit the variability within these regions for the detection and identification of microbial members within the analysed community. Many hypervariable regions are flanked by conserved sequences, allowing the use of 'universal' primers for the amplification of these regions from most microbial species. The degree of sequence variability varies among different V regions as shown in Fig. 9.1. As a result, the regions differ in their ability to distinguish among microbial phylotypes and some regions (V3, V6) are slightly better suited to resolve closely related microbial species (Chakravorty *et al.*, 2007). This characteristic emphasizes the need for careful consideration of probe target selection within the 16S rRNA gene. For example, community analysis using a microarray with probes to only a single hypervariable region has a potential to introduce a bias in the microbial community profile. It is generally considered a good practice to design probes to multiple hypervariable regions since such design strategy can adjust for region specific level of variability and any potential hybridization biases.

General strategies for optimizing the design of probes have been previously considered by Letowski and colleagues (Letowski *et al.*, 2004). In that study, the authors explored the effects of sequence mismatch on the destabilization of the probe–target hybridization at different fragment GC% and at different temperatures. One of the objectives of the study was to determine an optimal method for designing probes to closely related target sequences. To obtain quantitative results, the authors designed probes that differed in the number and distribution of mismatches. The probe specificities were determined and compared at various hybridization temperatures. The main conclusion of the study was that the greatest destabilization effect was achieved when

Figure 9.1 Sequence conservation and variability of 16S ribosomal RNA gene in prokaryotes. Sequence entropy is displayed using a gradient scale as shown in the legend. Positions of the variable regions (V1–V9, nucleotide positions are displayed for *Escherichia coli* 16S rRNA sequence) and sequence entropy values are based on the information from Ashelford *et al.* (2005).

mismatches were distributed across the entire sequence of the probe. From that observation the authors inferred that in order to achieve optimal specificity when designing probes to closely related sequences, it is important to choose probes such that the variability is spread along the probe length(s) (Letowski *et al.*, 2004). Conversely, variability concentrated towards the terminal regions of the probes showed greatly reduced specificity and therefore should be avoided. This study also confirmed previous reports of the dependence of the hybridization temperature on the GC% of the probes. In general, optimal specificity was achieved when the hybridization temperature correlated positively with the probe GC% (Letowski *et al.*, 2004).

Hybridization specificity is also dependent on other parameters such as orientation of the immobilized probe, steric hindrance against binding, and secondary structure formation in target molecules. The influence of these parameters on the hybridization specificity as well as methods to curtail their negative impacts have been introduced and discussed by Peplies *et al.* (2003). Probe orientation was tested using variants of select probes immobilized by either their 5′ or 3′ ends. The hybridization of these probes to their target revealed a higher annealing efficiency for the 3′ immobilized probes. The reduction in the hybridization efficiency of the 5′ immobilized probes was likely due to the occurrence of steric hindrance as the target has to bind the probe with its 3′ end facing the array surface. Note that a potential presence of secondary and tertiary structures in the target molecules can complicate the interpretation of these results. The effects of such steric hindrance can be mitigated by the use of spacer sequence in probes positioned between the array surface and the target-specific sequence of the probe. Indeed, Peplies and co-workers determined that there was a linear positive relationship between hybridization signal intensity and the length of the spacer sequence, indicating that larger space sequences significantly reduce steric hindrance (Peplies *et al.*, 2003). Lastly, the use of helper oligonucleotides can resolve secondary and tertiary structures of the target molecules. Helper oligonucleotides are unlabelled sequences designed to bind adjacent to the probe's binding site on the target molecule. By binding to the target molecule, the helper oligonucleotides prevent the target molecule from binding to itself, thereby increasing the efficiency of probe–target hybridization. Other optimization strategies such as selective calibration for particular probes to recover false-negatives and improving specificity through signal-limiting parameters can also be applied (Peplies *et al.*, 2003).

Optimization of sample preparation

Methods to improve the experimental procedures for the use of phylogenetic microarrays have been described. A study by Salonen *et al.* illustrated and compared several methods for the extraction of genomic DNA from faecal samples (Salonen *et al.*, 2010). Interestingly, the study found that the method used for the extraction of the genomic DNA from environmental samples had an effect on the compositional analysis of the community, and thus it is important to choose an extraction method that accurately reflects the actual community composition as well as provide efficient PCR amplification. This study proposed to use DNA quality, amount extracted, and community composition analysis as criteria for selecting and statistically authenticating an optimal method of genomic DNA extraction. The main conclusion from the comparison of methods was that the repeated bead beating approach to cell breakdown performed significantly better than the other methods, likely because it is generally more universal than alternative enzyme and chemical-based techniques (Salonen *et al.*, 2010). The bead beating method was capable of uncovering certain groups of microbes such as the methanogenic Archaea and some Gram-positive bacteria that remained undetected when other commonly utilized extraction protocols were employed. As an alternative to bead beating protocol, a recently developed pressure cycling technology can be utilized. In this approach, microbial or tissue samples are sealed in high-density tubes and are subjected to repeated rounds of high-low pressure fluctuations (Tao *et al.*, 2006). This process not only leads to the breakdown of cells, but can also separate proteins, lipids, and DNA based on their hydrophobicity and ionic properties. Pressure cycling technology was shown to also reduce the

effect of PCR inhibitors (see below), presumably because of the separation of the inhibitors and nucleic acids into different phases (Tao *et al.*, 2006).

A study of microbial community composition typically involves subjection of DNA collected from the community to rounds of target gene (such as 16S rRNA) specific PCR amplification. The goal of this approach is to selectively enrich the DNA pool with the fragments of interest, since 16S rRNA genes, for example, constitute less than 0.5% of total genomic DNA in most microorganisms. In the case of the 16S rRNA gene, primers that bind to universally conserved regions at the start and at the end of the gene or flanking one or several variable regions are used. Methods such as the phylogenetic microarrays and next-generation sequencing are then employed to determine the composition of the amplified library. It is important to keep in mind that environmental communities are composed of a large number of individual phylotypes with sequence differences in the interrogated target gene. Thus, any PCR amplification of such mixture of sequences is multi-template, and it has potential to introduce a skew in the composition of the amplified PCR library compared to that of the original DNA mixture (Polz and Cavanaugh, 1998). Several causes have been proposed to explain this often observed deviation, which include the difference in the template GC% leading to unequal denaturation of template–product pairs during the melting step of the PCR reaction, the higher binding efficiency of the GC-rich variants of the degenerate primer mixtures used to amplify fragments, and the re-annealing of high abundance templates during the annealing step that results in the selection against major templates (Polz and Cavanaugh, 1998). In addition, carrying out successful PCR reaction is always difficult for the genomic DNA obtained from environmental samples due to the presence of PCR inhibitors extracted during DNA isolation process. Faecal material, for example, contains bile salts and complex polysaccharides that are known to inhibit DNA polymerase activity (Lantz *et al.*, 1997; Monteiro *et al.*, 1997). Isolation of high quality DNA from soil presents even greater challenges: not only an efficient lysis of microbial cells is challenging, but the presence of humic acid inhibits most enzymatic reactions (Rock *et al.*, 2010). The problems with PCR inhibitors often necessitate the use of lower amounts of the starting DNA material in the amplification reactions in order to dilute the inhibitor concentration below critical level.

Possible approaches to mitigate such PCR bias have been recently considered by Paliy and Foy (2011). In this study, mathematical modelling of the multi-template PCR amplification of 16S ribosomal RNA genes as well as detection of the PCR products by phylogenetic microarray was used in conjunction with experimentally determined parameters to define optimal amplification conditions that lead to accurate estimations of phylotype levels. One of the most important conclusions from that study was that both the detection and the accuracy of species abundance estimations depended heavily on the number of PCR amplification cycles used. The model predicted that the improvements in the detection and accuracy reached optima between 15 and 20 cycles of PCR amplification. Because of the unequal amplification rate for different templates in the mixture, the accuracy of community composition estimates was negatively affected when DNA was subjected to more than 20 cycles of amplification – at that point gradually increasing PCR bias outpaced any further improvements in phylotype detection (Paliy and Foy, 2011). Modelling the presence of PCR inhibitors in the samples showed that the use of more than 50 ng of starting DNA was detrimental to the overall reaction yield and to the accuracy of phylotype detection and abundance estimates. With higher starting amounts, the higher levels of inhibitors caused a significant reduction in the amplification efficiency, and thus more amplification cycles were needed to reach an appropriate reaction yield, which in turn led to a higher PCR bias. Furthermore, the detection and accuracy of phylotype abundance estimates correlated positively with sample-wide PCR amplification rate but related negatively to the sample template-to-template PCR bias and community complexity (Paliy and Foy, 2011). Although this model was developed based on the simulated interrogation of human intestinal microbiota community and subsequent detection by the Microbiota Array, it can be easily modified to simulate the analysis

of other communities, other available or novel microarray designs as well as other PCR amplification protocols.

Optimization of data normalization

In order to draw accurate conclusions regarding microbial profiles, raw signal values measured by each microarray have to be normalized and adjusted, so that a valid comparison of signals among multiple samples and arrays can be performed (Fujita *et al.*, 2006). One goal of such signal normalization is to account for technical variability during sample preparation and microarray hybridization that can lead to systemic variations in measured signals. The objective of normalization is therefore to reduce the technical systemic variability among arrays so that it is easier to discern patterns or changes in microbial profiles across arrays. Many different methods of microarray data normalization have been developed over the years, and these approaches are generally applicable to the analysis of phylogenetic microarray data. The best choice of method often depends on the microarray technology used, the type of study, and the error or systemic variation present in the raw data. An interested reader is encouraged to refer to the study by Choe and colleagues who compared the efficiency of different methods of microarray data normalization (Choe *et al.*, 2005).

In general, data normalization procedure encompasses background correction (subtraction of background noise and non-specific general probe binding), subtraction of mismatch and control probe signals where applicable (for example, mismatch probes are used in Affymetrix microarray designs), adjustment of signal distribution within each array to match those of other arrays in the set (across-array normalization), and summation or averaging of signals from multiple probes targeting the same sequence in order to obtain a single estimate of sequence abundance. Examples of software that run these normalizations semi-automatically include Dchip (Corradi *et al.*, 2008), Affymetrix-developed Expression Console (part of Affymetrix analysis suite), and commercially available GeneSpring software suite (Agilent, Inc.). For users who desire control of each step of the process, freely available R-based Bioconductor package allows separate definitions of each normalization step. The authors have used an online-implemented version of this package accessible through the CARMAweb service (Rainer *et al.*, 2006) to successfully normalize Affymetrix and glass slide microarrays.

One type of error that is often present in the phylogenetic microarray data is the occurrence of signal due to off-target fragment hybridization, i.e. cross-hybridization. This issue is especially problematic for 16S rRNA gene based phylogenetic analysis because most probes on such microarrays interrogate a single highly conserved molecule, and thus many fragments in the mixture are likely to possess significant sequence similarity, which leads to increased off-target hybridization and cross-hybridization signal. Without an appropriate method to adjust for cross-hybridization, acquiring accurate estimates of community members' abundances becomes challenging. Microarrays based on Affymetrix design (Microbiota Array, PhyloChip) include a mismatch probe for each interrogating probe. These mismatch probes provide an estimate of potential cross-hybridization that can be removed from the probe set signal estimate during data processing. The situation is more difficult for the designs where such mismatch probes are not incorporated. Several methods have been explored recently to correct for such fragment cross-hybridization. One such approach, described by Rigsbee *et al.* (2011), involved the use of an algorithm for the correction of cross-hybridization of 16S rRNA gene targets among different phylotypes. In this method, the model was first built to estimate the measured total signal for each probeset as a combination of true signal from target–probe hybridization and false signal from cross-hybridizing fragments (Rigsbee *et al.*, 2011). To provide model parameters, the levels of cross-hybridization for different phylotypes were acquired from validation experiments for the Microbiota Array. These cross-hybridization estimates were subsequently incorporated into an adjustment algorithm to calculate true signal from total signal. The resulting true signal was then used instead of the total signal for phylotype abundance calculations. This algorithm was successfully applied to phylogenetic data acquired with Microbiota Array, and the adjusted values were shown to be more consistent with other

estimates of microbial community compositions acquired with alternative molecular techniques (Rigsbee *et al.*, 2011).

Rigsbee and co-authors also introduced a second algorithm to adjust the normalized signal values for the estimated number of 16S rRNA gene copies per phylotype genome (Rigsbee *et al.*, 2011). Since different bacterial species are known to contain a broad range of ribosomal RNA-encoding gene copies per genome (between 1 and 15), the measured true signal of a phylotype represents both its abundance as well as the total number of 16S rRNA gene copies it contains (for most species, 16S rRNA genes within the same organism have nucleotide sequence identity of ≥ 98% and thus would be expected to bind to the same probeset on the microarray) (Rigsbee *et al.*, 2011; Kembel *et al.*, 2012). The known numbers of 16S rRNA gene copies for the various microbial species can be acquired from publicly accessible databases such as rrnDB and NCBI. Adjusting the phylotype signal value by the estimated number of 16S rRNA gene copies allowed for a more accurate inference of each phylotype abundance (Rigsbee *et al.*, 2011).

Improvements in data analysis

Similar to data normalization approaches, standard microarray data analysis tools can be utilized successfully to analyse phylogenetic microarray data. The approaches include various ways to visualize data with heat maps (see Fig. 9.2), box plots, and scatter plots, as well as clustering of different taxonomical groups based on their abundance among samples (Rajilic-Stojanovic *et al.*, 2009; Agans *et al.*, 2011; Rigsbee *et al.*, 2011). Because in many cases abundances of individual taxons are defined relative to the overall community population, such relative abundance data are often presented in stacked columns, stacked

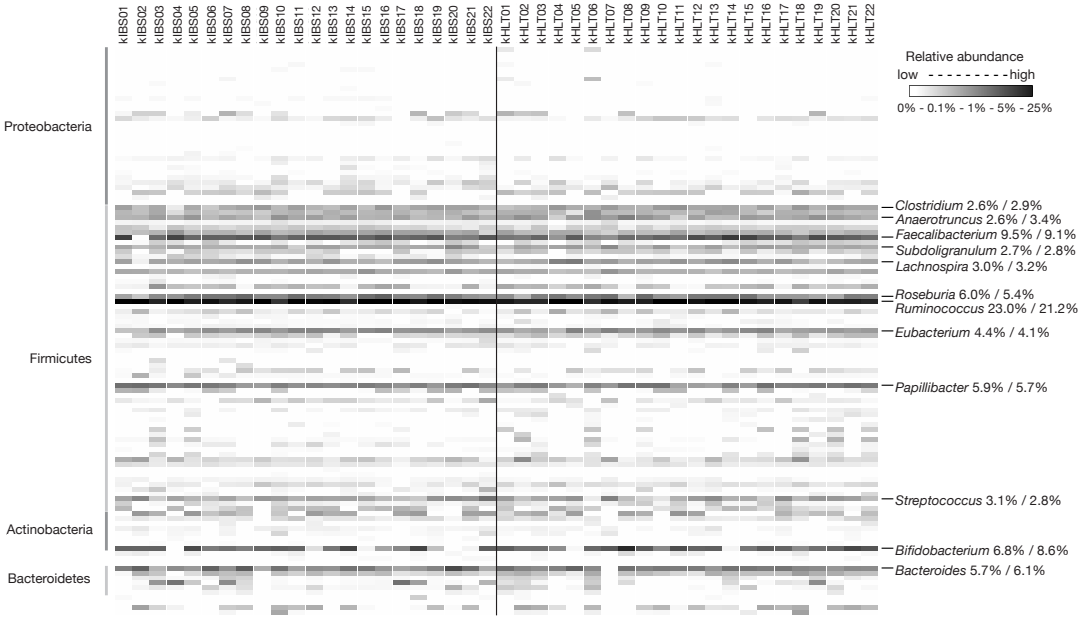

Figure 9.2 Distribution of bacterial relative abundances among samples obtained from healthy children (kHLT) and children diagnosed with IBS (kIBS). Different samples are plotted as columns; microbial genera are plotted as rows. Relative abundances of each genus are displayed using a gradient scale as shown in the legend. Sample designation is shown at the top of each column. Vertical line separates kIBS and kHLT samples. The 12 most abundant genera are displayed on the right side; genus assignments to the four most abundant phyla are shown on the left side of the image. Numbers represent relative average abundance of each genus in kIBS and kHLT samples, respectively. The figure was first published in American Journal of Gastroenterology, issue 107, 2012 (Rigsbee *et al.*, 2012), produced by Nature Publishing Group, a division of Macmillan Publishers Limited.

bars, or pie charts (Wu *et al.*, 2010a; Rigsbee *et al.*, 2012). To assess if different types of samples can be separated based on their community composition, data dimensionality reduction methods such as principal components analysis can be used (Nemir *et al.*, 2010; Agans *et al.*, 2011; Kopecky *et al.*, 2011).

Recently, several studies have explored methods to improve analysis procedures associated with phylogenetic microarrays. A unique feature of the phylogenetic microarray data is the ability to link the presence and abundance of each sequence to the placement of the corresponding species on the phylogenetic tree. This information allows researchers to estimate community ecological parameters such as diversity, richness, and evenness, and to assess the sample separation that takes into account phylogenetic identity of community members (Hazen *et al.*, 2010). For example, Hamady *et al.* described improvements in ecological beta diversity analysis of microarray data using phylogenetic information (Hamady *et al.*, 2010). The approach incorporated evolutionary relationships between taxa to calculate phylogenetic beta diversity, a metric that is used to compare diversity among communities. This type of analysis can uncover underlying patterns of change in diversity that only become evident when phylogenetic relationships are taken into account. The authors developed an online tool, Fast UniFrac, which uses phylogenetic information in conjunction with multivariate statistics to assess if the examined communities are significantly different and to characterize phylogenies of the taxa that are responsible for the differences among communities (Hamady *et al.*, 2010).

Another study, by Schatz and colleagues, introduced a stand-alone software package for the analysis of signal values from the PhyloChip microarray (Schatz *et al.*, 2010). This software, called PhyloTrac, is capable of identifying and quantifying microbial community members from the environmental samples that were interrogated using the PhyloChip microarray. One of the several advantages of this software is the all-inclusive nature of the application. It contains all the necessary dependencies, such as phylogenetic information for assignment of taxonomy, normalization procedures, microarray design information, etc., within the package. This feature provides researchers with an efficient way to go from raw microarray data to comprehensive compositional analysis in a single step. Furthermore, PhyloTrac offers a user-friendly interface for the display of the community composition and taxonomy, which allows for synchronized selection of OTUs across multiple modes of data visualization as well as for filtering of OTUs using any of the standard distance metrics.

Phylogenetic microarray applications

Phylogenetic microarrays have been utilized to successfully carry out many different studies that interrogated a diverse set of microbial environments. These included human associated niches such as the gastrointestinal, oral, and vaginal tracts, as well as communities from ocean waters, soil, and sewage. Examples of such high-throughput analyses using phylogenetic microarrays are discussed in this section.

The Microbiota Array

The faecal microbiome of healthy adolescents and adolescents with diarrhoea-predominant irritable bowel syndrome (IBS) was profiled recently in a study by Rigsbee *et al.* (2012). The objective of the study was to assess the differences in the faecal microbiota profile between the two groups and to potentially identify putative associations among different microbial members. This study took advantage of the quantitative nature of the Microbiota Array to compare relative abundances among the interrogated samples at several taxonomic levels. Microarray data was confirmed with high-throughput 454-based pyrosequencing and fluorescence *in situ* hybridization (FISH). The study showed that the overall structure of the faecal microbiomes was generally similar between healthy and IBS adolescents. In both groups, the phylum Firmicutes was the most abundant, followed by Actinobacteria and Bacteroidetes, with members of these three phyla cumulatively constituting 91% of the overall community composition on average (Fig. 9.2). At the genus level, the relative fractions of the abundant genera in the microbial communities were also similar between

the two groups; the polysaccharide-degrading members of the genus *Ruminococcus* were the most abundant (Rigsbee *et al.*, 2012).

Some distinct differences in the microbial profiles were observed at lower taxonomic levels (genus and species). More specifically, the array detected lower levels of the genus *Bifidobacterium* but higher levels of genera *Lactobacillus*, *Veillonella*, and *Prevotella* in adolescents with IBS, which is an observation that is consistent with several other reports (Rigsbee *et al.*, 2012). The array also allowed for the characterization of a set of phylotypes that was present in all or most samples. Such set of phylotypes can be referred to as the core microbiome of that niche, which is often thought to play important roles in the community functional capacity including inter-species and host–microbial interactions. In the combined set of adolescent faecal samples, the array identified a core microbiome of 55 phylotypes. This core microbiome was dominated by genus *Ruminococcus*; members of genera *Bacteroides*, *Clostridium*, *Faecalibacterium*, *Roseburia*, and *Streptococcus* were also present (Rigsbee *et al.*, 2012).

In order to identify putative associations among microbial members, a non-parametric correlation matrix was constructed using the abundance levels of the various genera across all samples. Such relationships can represent potential metabolic interdependencies, where the end-products of metabolism of some community members become energy and carbon sources for other members. The study identified a large number of statistically significant relationships among the genera, which is consistent with our current understanding of the intricate nature of metabolic networks among the community members in the intestinal ecosystem. As an example, abundance of members of genus *Veillonella* correlated with the largest number of other genera, probably because the members of this genus participate in the metabolic cross-feeding pathways (Chalmers *et al.*, 2008). Specifically, *V. parvula* cannot degrade complex or even simple sugars available in the colon and rely on the use of intermediary end-products of carbohydrate fermentation (such as lactate, pyruvate, and fumarate) released by other gut microbes (Gronow *et al.*, 2009). A physical association between *Veillonella* and *Streptococcus* was also observed in dental plaque (Chalmers *et al.*, 2008).

PhyloChip

The G2 version of the PhyloChip was utilized to analyse watershed microbial communities in an attempt to characterize the sensitivity of these communities to perturbations in the environment (Wu *et al.*, 2010a). Three different watershed communities (creek, lagoon, and ocean) were sampled from a coastal area that was known to be prone to faecal contamination. Aside from these environmental samples, faecal samples were also profiled in this study to obtain a direct comparison of community membership. Multi-response permutation procedure using Bray–Curtis diversity distances among the communities revealed significant differences among the four communities. Furthermore, non-parametric multidimensional scaling ordination was successful in separating samples based on their collection site for the majority of the analysed samples (Wu *et al.*, 2010a). Environmental factors were also measured at the sampled sites in order to correlate them with the microbial profiles. Interestingly, among all the measured environmental variables, salinity had the greatest effect on the community composition, evidenced by the fact that in non-parametric multidimensional scaling ordination, lagoon samples that clustered with creek group had salinity levels that resembled those of the creek samples. Specific effects of the environmental factors on the microbial communities were observed at the class level among the four habitats. Of the classes that showed the greatest variability among habitats, Bacilli, Bacteroidetes, and Clostridia were found to have higher relative abundances in faecal samples compared to the creek, lagoon, or marine samples. Conversely, Alphaproteobacteria were found at a lower relative abundance in faecal samples than in the environmental samples. A set of 503 phylotypes, found to be ubiquitous in faecal samples but not in the environmental samples, was used as means to determine which collection sites were prone to heavy faecal contamination (Wu *et al.*, 2010a).

The G3 version of the PhyloChip was used to profile marine microbial communities affected by oil plumes released during the Deep Horizon

oil spill (Hazen *et al.*, 2010). The objective of the study was to characterize the unique features of the communities sampled from deep-sea oil plumes. The 16S rRNA microarray analysis showed that the communities underwent compositional and structural changes upon contact with the oil. Multidimensional scaling ordination using Bray–Curtis beta diversity distance metric was able to differentiate bacterial and archaeal communities from plume and non-plume samples. Since all other factors were not significantly different between the sampled communities, this suggested that changes in microbial community profiles were due to the direct response of the microbes to the existence of oil in the environment. The PhyloChip uncovered a total of 951 individual bacterial taxa spread across 62 phyla from the analysed oil-plume samples. When compared to the non-plume samples, 16 bacterial taxa were found to be significantly enriched in the oil plume samples. All 16 of these taxa belonged to Gammaproteobacteria and most had representative members capable of degrading various hydrocarbons. The bacterial taxa enriched through the presence of oil included a significant number of psychrophilic and psychro-tolerant phylotypes similar to those that have been identified in cold deep-sea ecosystems (Hazen *et al.*, 2010).

HOMIM

Oral microbiota-specific HOMIM array was employed to assess the microbiota profile in the saliva of healthy children and children with dental caries (Luo *et al.*, 2012). The objective of this research project was to determine microbial biomarkers for the onset of dental caries in mixed dentition and to characterize the community profile of the microbial disease. In total, the study identified 86 phylotypes as well as eight clusters of closely related phylotypes. In agreement with several sequencing studies, the microbial community of the saliva was found to be dominated by the phyla Firmicutes and Proteobacteria. The overall relative contribution of different phyla to the total microbial abundance was similar in both sample groups with the exception of the TM7 phylum, which was only detected in the caries-active group. A higher microbial diversity, with 89 detected species, was observed in communities from the caries-active group, compared to the caries-free healthy group that contained on average 59 species. This suggested a shift in microbial community structure in response to the change from a healthy to a diseased oral environment. Examining the relative abundances at the genus level revealed that genus *Streptococcus* was the most abundant, followed by *Prevotella* and *Selenomonas* (Luo *et al.*, 2012).

Surprisingly, at the phylotype level and in contrast with several previous reports, cariogenic species such as *Streptococcus mutans* and members of the cariogenic genus *Lactobacillus* were not highly prevalent in the caries-active group (Luo *et al.*, 2012). Interestingly, these cariogenic groups were substituted by the high prevalence of other streptococci. Examples of phylotypes that were differentially abundant between the two groups included species of *Leptotrichia*, which were found only in caries-active patients, and *Granulicatella* sp. and *Rothia dentocariosa*, which were found at much higher abundance in healthy children. There was a much greater number of phylotypes unique to the caries-active group compared to those unique to the healthy group, likely due to the higher community diversity seen in the caries-active group. A member of the genus *Fusobacterium*, *Fusobacterium nucleatum*, was found to be prevalent in all oral samples, which the authors attributed to the key role this species plays in the establishment of microbial communities in naturally forming dental plaques (Luo *et al.*, 2012).

V-Chip

The vaginal microbiota of African women with or without bacterial vaginosis (BV) was examined by Dols *et al.* (2011) through the use of the vaginal microbiota-representing microarray (V-Chip). The goal of the study was to first test the ability of the microarray to successfully detect microbes found at high prevalence in BV, and to characterize the profiles of the vaginal microbial communities in women in the study group. The microarray results showed that women who were negative for BV had a high prevalence of various species of *Lactobacillus*, a genus that includes many members considered beneficial to human health. The number of detected microbial groups was significantly higher in the BV women than in those with

normal vaginal microbiota. BV-positive women harboured a much larger set of known microbial pathogens as well as more complex microbiota than women from BV negative or intermediate groups. The microarray data also indicated that high prevalence of HIV in many cases correlated with high prevalence of BV. At a species level, the study revealed that *Gardnerella vaginalis* and *Atopobium vaginae* co-occurred in nearly 70% of the women, suggesting potential microbial interaction(s) between these species towards pathogenesis. The presence of *Gardnerella* was also associated with the presence of *Leptotrichia* and *Prevotella* species. Noteworthy, while previous reports found *Gardnerella vaginalis* to be generally associated with BV diagnosis, this species was also present in 24% of BV-negative women profiled in this study. Thus, the microarray data did not support the previous use of the presence of this organism as a diagnostic tool for BV. Instead, the authors proposed to employ the co-occurrence of *Gardnerella vaginalis* and other pathogens such as *Atopobium vaginae* as a criterion for the diagnosis of BV (Dols *et al.*, 2011).

EcoChip

The EcoChip was used to determine an impact of sewage sludge on soil bacterial communities (Val-Moraes *et al.*, 2011). In general, a relatively high variation in community structure was observed from the beginning to the end of the experiment that likely reflected seasonal changes. Consistent with previous reports, microarray data revealed that soil communities were dominated by members from the phylum Acidobacteria, followed by those of Firmicutes, Proteobacteria, and Actinobacteria. Significant alterations in bacterial phyla were observed when bacterial communities were compared before and after sludge application. Sludge amendment containing 25 kg N/ha favoured an increase in the number of members of Acidobacteria, Alphaproteobacteria, Bacteroidetes, Deltaproteobacteria, Firmicutes, Gemmatimonadetes, and Nitrospirae, while Actinobacteria, Planctomycetes, and some Proteobacteria were the most diminished in sludge amendments of 200 kg N/ha. Members of the Epsilonproteobacteria and Spirochaetes were found only in the samples treated with high doses of sludge. The levels of Epsilonproteobacteria correlated well with the levels of sulfate present in the analysed soil – an observation that is consistent with previous reports that claim the presence of Epsilonproteobacteria in sulfate-rich environments such as deep-sea vents (Val-Moraes *et al.*, 2011).

RHC-PhyloChip

A composite microarray-based fingerprint of the Rhodocyclales community present in activated sludge was created with the help of the RHC-PhyloChip (Hesselsoe *et al.*, 2009). Separate microarray hybridization patterns obtained with the fragments after either Rhodocyclales selective or general 16S rRNA gene based PCR amplifications were merged to provide an overall community view. This merged microarray hybridization results indicated the presence of bacteria belonging or related to the *Sterolibacterium* lineage, the 'Candidatus Accumulibacter' cluster, and the genera *Quadricoccus*, *Thauera* and *Zoogloea*. A parallel cloning-sequencing approach provided a validation of the microarray capability to detect uncultured members of Rhodocyclales. A separate RHC-PhyloChip was hybridized with fluorescently labelled and fragmented RNA from each activated sludge subsample. Radioactive signals on the microarray indicated that bacteria represented by several cloned sequences were active under all conditions tested, while other Rhodocyclales groups, for which specific probes were present on the RHC-PhyloChip, displayed more specialized substrate incorporation behaviours. For example, the genus *Zoogloea* was detectable after oxic incubation with butyrate and propionate, but not with toluene (Hesselsoe *et al.*, 2009).

ActinoChip

Actinobacterial community of a waterlogged forest soil was analysed by an Actinobacteria-specific microarray (Kopecky *et al.*, 2011). The goal of the study was to follow bacterial communities at a previously studied site with respect to differences between soil horizons and seasons. The PCA analysis of the microarray data was able to distinguish between communities of the lower and upper horizons along the first ordination axis

(PC1, 49% of dataset variance explained), and the summer and winter communities (especially for the upper horizon) along the second ordination axis (PC2, 10% variance explained), indicating a higher effect of the horizon than season on actinobacterial community composition. The differences between horizons were mostly caused by much higher signals from the *Mycobacterium* probes in the upper horizon, while the differences between the seasons were due to the signals of probes targeting the genera *Asanoa* and *Brevibacterium* (higher in winter), and *Mobiluncus* and *Saccharomonospora* (higher in summer). The upper horizon soil appeared to be mostly influenced by organic matter content in winter and soil moisture in summer, based on the PCA-IV (instrumental variables) analysis (Kopecky *et al.*, 2011).

TCE Chip

Soils contaminated with trichloroethylene (TCE) were examined in response to different doses of fresh TCE amendments at four concentrations (1 ppb, 100 ppb, 1 ppm and 25 ppm) after exposure of 2 h, 2 days, 14 days, 35 days, and 151 days in a study by Nemir and others (Nemir *et al.*, 2010). Changes in bacterial communities were determined with the TCE Chip. TCE presence in the microcosms for only 2 h was sufficient to elicit changes in microbial composition. It was possible to discriminate between bacterial communities containing either 1 ppm or 10 ppm TCE from samples treated with lower TCE concentration. This trend continued over time, with visible separation between contaminated and control samples. After 151 days, however, the community structure regained homogeneity across concentrations. There was no significant difference between wet and dry negative controls tested at 2 h and 151 days time points, showing that the effect of adding water to the samples was negligible when compared to the effect of adding TCE. An apparent threshold at which the microbial community structure was significantly affected was determined to be at TCE concentration of about 1 ppm. Bacterial taxa associated with TCE contamination included, among others, Planctomycetes, Acidobacteria, and various groups of Proteobacteria (Nemir *et al.*, 2010).

Future trends and outlook

High-throughput techniques such as phylogenetic microarrays and next-generation sequencing provide us extensive knowledge regarding the composition of complex microbial communities. This knowledge enables us to understand which members are present in the community as well as to predict their potential role. Examples of the phyloarray applications that have been described in the previous section of this chapter highlight a multitude of questions that can be answered through the use of phylogenetic microarrays. A diverse set of microbial communities that include those found in human-associated niches such as gut, airways, and vaginal canal, as well as environmental ecosystems such as marine, soil, and sewage sludge, have been analysed qualitatively and quantitatively by phylogenetic microarrays.

The intricate nature of the microarray design process and the extensive validation procedures have been limiting factors towards the wider use of phylogenetic microarrays. Nonetheless, there already exists an assortment of phylogenetic microarrays capable of analysing a variety of microbial ecosystems (see Table 9.1). The improvements in cost efficiency and the highly quantitative nature of phyloarrays make them an excellent choice for high-throughput compositional analysis of microbial communities. A particularly attractive application is the use of both phylogenetic microarrays and next-generation sequencing for the analysis of the same microbial community (Ahn *et al.*, 2011; Crielaard *et al.*, 2011; van den Bogert *et al.*, 2011; Rigsbee *et al.*, 2012). The phyloarrays provide quantitative data for the comparison of abundances across groups of samples, while the 16S rRNA amplicon sequencing allows for the identification of novel members of the community.

The future trends in the use of phylogenetic microarrays are likely to be defined by a shift towards integrative approaches to community analysis. Current studies have helped us understand the composition of microbial communities. Using this information in combination with new molecular tools, future studies will likely focus on the interactions among members of the microbial communities as well as between microbiota and the environment. There is also a growing interest

towards understanding the link(s) between the function and the activity of microbiota in various environmental niches or disease states. In integrative approaches, the use of phylogenetic microarrays can be augmented with other high-throughput methods such as metabolomics, meta-genomics, meta-transcriptomics, and meta-proteomics to construct a more comprehensive model of the analysed community ((Klaassens *et al.*, 2007; Booijink *et al.*, 2010; Martin *et al.*, 2010; see also Chapter 7). A combination of these techniques would allow us to determine the profile of the community composition, total gene content, and expression levels of the genes and proteins, and we would be able to relate this data to the metabolite profiles of the environment and community members. Such an approach will enable us to understand the intricate relationships and the roles the members of the microbiota play within different microbial ecosystems.

Thanks to the advancements in technology and our knowledge of microbial communities, several enhancements to the design and use of phylogenetic microarrays can also be conceived. Programs such as the Human Microbiome Project (Peterson *et al.*, 2009) and the MetaHIT initiative (Qin *et al.*, 2010) have made available a substantial number of genome sequences of human-associated microbiota members. The availability of such resources has given rise to the possibility of designing phylogenetic detection arrays based on functionally conserved genes such as *groEL*, *rpoB*, *gyrA* and *tufA* (Loy and Bodrossy, 2006). Specific pathogen detection arrays have a potential to play a vital role in the field of microbial forensics for the rapid detection and identification of pathogens in the environment. Furthermore, phylogenetic microarrays can also be designed to contain probes to functional genes to enable simultaneous analysis of community structure and function (Louis and Flint, 2007). In a clinical setting, phylogenetic microarrays can be used as diagnostic tools, where their ability to detect human-associated microbiota members at a species level in a relatively short period of time can help in the diagnosis of various pathological states and rapid selection of treatment procedures that are most likely to succeed (Loy and Bodrossy, 2006).

References

Agans, R., Rigsbee, L., Kenche, H., Michail, S., Khamis, H.J., and Paliy, O. (2011). Distal gut microbiota of adolescent children is different from that of adults. FEMS Microbiol. Ecol. *77*, 404–412.

Ahn, J., Yang, L., Paster, B.J., Ganly, I., Morris, L., Pei, Z., and Hayes, R.B. (2011). Oral microbiome profiles: 16S rRNA pyrosequencing and microarray assay comparison. PLoS One *6*, e22788.

Ashelford, K.E., Chuzhanova, N.A., Fry, J.C., Jones, A.J., and Weightman, A.J. (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.

Belenguer, A., Duncan, S.H., Calder, A.G., Holtrop, G., Louis, P., Lobley, G.E., and Flint, H.J. (2006). Two routes of metabolic cross-feeding between Bifidobacterium adolescentis and butyrate-producing anaerobes from the human gut. Appl. Environ. Microbiol. *72*, 3593–3599.

Bodrossy, L., Stralis-Pavese, N., Murrell, J.C., Radajewski, S., Weilharter, A., and Sessitsch, A. (2003). Development and validation of a diagnostic microbial microarray for methanotrophs. Environ. Microbiol. *5*, 566–582.

van den Bogert, B., de Vos, W.M., Zoetendal, E.G., and Kleerebezem, M. (2011). Microarray analysis and barcoded pyrosequencing provide consistent microbial profiles depending on the source of human intestinal samples. Appl. Environ. Microbiol. *77*, 2071–2080.

Booijink, C.C., Boekhorst, J., Zoetendal, E.G., Smidt, H., Kleerebezem, M., and de Vos, W.M. (2010). Metatranscriptome analysis of the human fecal microbiota reveals subject-specific expression profiles, with genes encoding proteins involved in carbohydrate metabolism being dominantly expressed. Appl. Environ. Microbiol. *76*, 5533–5540.

Brodie, E.L., Desantis, T.Z., Joyner, D.C., Baek, S.M., Larsen, J.T., Andersen, G.L., Hazen, T.C., Richardson, P.M., Herman, D.J., Tokunaga, T.K., *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.

Brodie, E.L., DeSantis, T.Z., Parker, J.P., Zubietta, I.X., Piceno, Y.M., and Andersen, G.L. (2007). Urban aerosols harbor diverse and dynamic bacterial populations. Proc. Natl. Acad. Sci. U.S.A. *104*, 299–304.

Candela, M., Consolandi, C., Severgnini, M., Biagi, E., Castiglioni, B., Vitali, B., De Bellis, G., and Brigidi, P. (2010). High taxonomic level fingerprint of the human intestinal microbiota by ligase detection reaction--universal array approach. BMC Microbiol. *10*, 116.

Chakravorty, S., Helb, D., Burday, M., Connell, N., and Alland, D. (2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. J. Microbiol. Methods *69*, 330–339.

Chalmers, N.I., Palmer, R.J., Jr., Cisar, J.O., and Kolenbrander, P.E. (2008). Characterization of a Streptococcus sp.-Veillonella sp. community

micromanipulated from dental plaque. J. Bacteriol. *190*, 8145–8154.

Chiu, S.K., Hsu, M., Ku, W.C., Tu, C.Y., Tseng, Y.T., Lau, W.K., Yan, R.Y., Ma, J.T., and Tzeng, C.M. (2003). Synergistic effects of epoxy- and amine-silanes on microarray DNA immobilization and hybridization. Biochem. J. *374*, 625–632.

Choe, S.E., Boutros, M., Michelson, A.M., Church, G.M., and Halfon, M.S. (2005). Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. Genome Biol. *6*, R16.

Corradi, L., Fato, M., Porro, I., Scaglione, S., and Torterolo, L. (2008). A Web-based and Grid-enabled dChip version for the analysis of large sets of gene expression data. BMC Bioinformat. *9*, 480.

Cox, M.J., Allgaier, M., Taylor, B., Baek, M.S., Huang, Y.J., Daly, R.A., Karaoz, U., Andersen, G.L., Brown, R., Fujimura, K.E., *et al.* (2010). Airway microbiota and pathogen abundance in age-stratified cystic fibrosis patients. PLoS One *5*, e11044.

Crielaard, W., Zaura, E., Schuller, A.A., Huse, S.M., Montijn, R.C., and Keijser, B.J. (2011). Exploring the oral microbiota of children at various developmental stages of their dentition in the relation to their oral health. BMC Med. Genomics *4*, 22.

Deangelis, K.M., Allgaier, M., Chavarria, Y., Fortney, J.L., Hugenholtz, P., Simmons, B., Sublette, K., Silver, W.L., and Hazen, T.C. (2011). Characterization of trapped lignin-degrading microbes in tropical forest soil. PLoS One *6*, e19306.

De Vuyst, L., and Leroy, F. (2011). Cross-feeding between bifidobacteria and butyrate-producing colon bacteria explains bifdobacterial competitiveness, butyrate production, and gas production. Int. J. Food Microbiol. *149*, 73–80.

Docktor, M.J., Paster, B.J., Abramowicz, S., Ingram, J., Wang, Y.E., Correll, M., Jiang, H., Cotton, S.L., Kokaras, A.S., and Bousvaros, A. (2012). Alterations in diversity of the oral microbiome in pediatric inflammatory bowel disease. Inflamm. Bowel Dis. *18*, 935–942.

Dols, J.A., Smit, P.W., Kort, R., Reid, G., Schuren, F.H., Tempelman, H., Bontekoe, T.R., Korporaal, H., and Boon, M.E. (2011). Microarray-based identification of clinically relevant vaginal bacteria in relation to bacterial vaginosis. Am. J. Obstet. Gynecol. *204*, 305. e301–307.

Duncan, S.H., Louis, P., and Flint, H.J. (2004). Lactate-utilizing bacteria, isolated from human feces, that produce butyrate as a major fermentation product. Appl. Environ. Microbiol. *70*, 5810–5817.

Flint, H.J., Bayer, E.A., Rincon, M.T., Lamed, R., and White, B.A. (2008). Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis. Nat. Rev. *6*, 121–131.

Fujita, A., Sato, J.R., Rodrigues Lde, O., Ferreira, C.E., and Sogayar, M.C. (2006). Evaluating different methods of microarray data normalization. BMC Bioinformat. *7*, 469.

Goldmann, T., and Gonzalez, J.S. (2000). DNA-printing: utilization of a standard inkjet printer for the transfer of nucleic acids to solid supports. J. Biochem. Biophys. Methods *42*, 105–110.

Gronow, S., Welnitz, S., Lapidus, A., Nolan, M., Ivanova, N., Glavina Del Rio, T., Copeland, A., Chen, F., Tice, H., Pitluck, S., *et al.* (2009). Complete genome sequence of *Veillonella parvula* type strain (Te3). Stand. Genomic Sci. *2*, 57–65.

Guschin, D.Y., Mobarry, B.K., Proudnikov, D., Stahl, D.A., Rittmann, B.E., and Mirzabekov, A.D. (1997). Oligonucleotide microchips as genosensors for determinative and environmental studies in microbiology. Appl. Environ. Microbiol. *63*, 2397–2402.

Hamady, M., Lozupone, C., and Knight, R. (2010). Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. ISME J. *4*, 17–27.

Hashsham, S.A., Wick, L.M., Rouillard, J.M., Gulari, E., and Tiedje, J.M. (2004). Potential of DNA microarrays for developing parallel detection tools (PDTs) for microorganisms relevant to biodefense and related research needs. Biosens. Bioelectron. *20*, 668–683.

Hazen, T.C., Dubinsky, E.A., DeSantis, T.Z., Andersen, G.L., Piceno, Y.M., Singh, N., Jansson, J.K., Probst, A., Borglin, S.E., Fortney, J.L., *et al.* (2010). Deep-sea oil plume enriches indigenous oil-degrading bacteria. Science *330*, 204–208.

He, Z., Wu, L., Fields, M.W., and Zhou, J. (2005). Use of microarrays with different probe sizes for monitoring gene expression. Appl. Environ. Microbiol. *71*, 5154–5162.

Heller, M.J., Forster, A.H., and Tu, E. (2000). Active microeletronic chip devices which utilize controlled electrophoretic fields for multiplex DNA hybridization and other genomic applications. Electrophoresis *21*, 157–164.

Hesselsoe, M., Fureder, S., Schloter, M., Bodrossy, L., Iversen, N., Roslev, P., Nielsen, P.H., Wagner, M., and Loy, A. (2009). Isotope array analysis of Rhodocyclales uncovers functional redundancy and versatility in an activated sludge. ISME J. *3*, 1349–1364.

Kembel, S.W., Wu, M., Eisen, J.A., and Green, J.L. (2012). Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. PLoS Comp. Biol. *8*, e1002743.

Klaassens, E.S., de Vos, W.M., and Vaughan, E.E. (2007). Metaproteomics approach to study the functionality of the microbiota in the human infant gastrointestinal tract. Appl. Environ. Microbiol. *73*, 1388–1392.

Kopecky, J., Kyselkova, M., Omelka, M., Cermak, L., Novotna, J., Grundmann, G.L., Moenne-Loccoz, Y., and Sagova-Mareckova, M. (2011). Actinobacterial community dominated by a distinct clade in acidic soil of a waterlogged deciduous forest. FEMS Microbiol. Ecol. *78*, 386–394.

Kyselkova, M., Kopecky, J., Frapolli, M., Defago, G., Sagova-Mareckova, M., Grundmann, G.L., and Moenne-Loccoz, Y. (2009). Comparison of rhizobacterial community composition in soil suppressive or conducive to tobacco black root rot disease. ISME J. *3*, 1127–1138.

Lantz, P., Matsson, M., Wadstrom, T., and Radstrom, P. (1997). Removal of PCR inhibitors from human faecal samples through the use of an aqueous two-phase system for sample preparation prior to PCR. J. Microbiol. Methods *28*, 159–167.

Lemon, K.P., Klepac-Ceraj, V., Schiffer, H.K., Brodie, E.L., Lynch, S.V., and Kolter, R. (2010). Comparative analyses of the bacterial microbiota of the human nostril and oropharynx. mBio *1*, e00129–00110.

Letowski, J., Brousseau, R., and Masson, L. (2004). Designing better probes: effect of probe size, mismatch position and number on hybridization in DNA oligonucleotide microarrays. J. Microbiol. Methods *57*, 269–278.

Louis, P., and Flint, H.J. (2007). Development of a semiquantitative degenerate real-time pcr-based assay for estimation of numbers of butyryl-coenzyme A (CoA) CoA transferase genes in complex bacterial samples. Appl. Environ. Microbiol. *73*, 2009–2012.

Loy, A., and Bodrossy, L. (2006). Highly parallel microbial diagnostics using oligonucleotide microarrays. Clin. Chim. Acta *363*, 106–119.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, Buchner, A., Lai, T., Steppi, S., Jobb, G., *et al.* (2004). ARB: a software environment for sequence data. Nucleic Acids Res. *32*, 1363–1371.

Luo, A.H., Yang, D.Q., Xin, B.C., Paster, B.J., and Qin, J. (2012). Microbial profiles in saliva from children with and without caries in mixed dentition. Oral Dis. *18*, 595–601.

Luton, P.E., Wayne, J.M., Sharp, R.J., and Riley, P.W. (2002). The mcrA gene as an alternative to 16S rRNA in the phylogenetic analysis of methanogen populations in landfill. Microbiology *148*, 3521–3530.

Martens, M., Weidner, S., Linke, B., de Vos, P., Gillis, M., and Willems, A. (2007). A prototype taxonomic microarray targeting the rpsA housekeeping gene permits species identification within the rhizobial genus Ensifer. Syst. Appl. Microbiol. *30*, 390–400.

Martin, F.P., Sprenger, N., Montoliu, I., Rezzi, S., Kochhar, S., and Nicholson, J.K. (2010). Dietary modulation of gut functional ecology studied by fecal metabonomics. Journal of proteome research *9*, 5284–5295.

Mendes, R., Kruijt, M., de Bruijn, I., Dekkers, E., van der Voort, M., Schneider, J.H., Piceno, Y.M., DeSantis, T.Z., Andersen, G.L., Bakker, P.A., *et al.* (2011). Deciphering the rhizosphere microbiome for disease-suppressive bacteria. Science *332*, 1097–1100.

Midgley, D.J., Greenfield, P., Shaw, J.M., Oytam, Y., Li, D., Kerr, C.A., and Hendry, P. (2012). Reanalysis and simulation suggest a phylogenetic microarray does not accurately profile microbial communities. PLoS One *7*, e33875.

Militon, C., Rimour, S., Missaoui, M., Biderre, C., Barra, V., Hill, D., Mone, A., Gagne, G., Meier, H., Peyretaillade, E., *et al.* (2007). PhylArray: phylogenetic probe design algorithm for microarray. Bioinformatics *23*, 2550–2557.

Mitterer, G., Huber, M., Leidinger, E., Kirisits, C., Lubitz, W., Mueller, M.W., and Schmidt, W.M. (2004). Microarray-based identification of bacteria in clinical samples by solid-phase PCR amplification of 23S ribosomal DNA sequences. J. Clin. Microbiol. *42*, 1048–1057.

Monteiro, L., Bonnemaison, D., Vekris, A., Petry, K.G., Bonnet, J., Vidal, R., Cabrita, J., and Megraud, F. (1997). Complex polysaccharides as PCR inhibitors in feces: Helicobacter pylori model. J. Clin. Microbiol. *35*, 995–998.

Naum, M., Brown, E.W., and Mason-Gamer, R.J. (2008). Is 16S rDNA a reliable phylogenetic marker to characterize relationships below the family level in the enterobacteriaceae? J. Mol. Evol. *66*, 630–642.

Nemir, A., David, M.M., Perrussel, R., Sapkota, A., Simonet, P., Monier, J.M., and Vogel, T.M. (2010). Comparative phylogenetic microarray analysis of microbial communities in TCE-contaminated soils. Chemosphere *80*, 600–607.

Paliy, O., and Agans, R. (2012). Application of phylogenetic microarrays to interrogation of human microbiota. FEMS Microbiol. Ecol. *79*, 2–11.

Paliy, O., and Foy, B. (2011). Mathematical modeling of 16S ribosomal DNA amplification reveals optimal conditions for the interrogation of complex microbial communities with phylogenetic microarrays. Bioinformatics *27*, 2134–2140.

Paliy, O., Kenche, H., Abernathy, F., and Michail, S. (2009). High-throughput quantitative analysis of the human intestinal microbiota with a phylogenetic microarray. Appl. Environ. Microbiol. *75*, 3572–3579.

Pease, A.C., Solas, D., Sullivan, E.J., Cronin, M.T., Holmes, C.P., and Fodor, S.P. (1994). Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proc. Natl. Acad. Sci. U.S.A. *91*, 5022–5026.

Peplies, J., Glockner, F.O., and Amann, R. (2003). Optimization strategies for DNA microarray-based detection of bacteria with 16S rRNA-targeting oligonucleotide probes. Appl. Environ. Microbiol. *69*, 1397–1407.

Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J.A., Bonazzi, V., McEwen, J.E., Wetterstrand, K.A., Deal, C., *et al.* (2009). The NIH Human Microbiome Project. Genome Res. *19*, 2317–2323.

Polz, M.F., and Cavanaugh, C.M. (1998). Bias in template-to-product ratios in multitemplate PCR. Appl. Environ. Microbiol. *64*, 3724–3730.

Preza, D., Olsen, I., Willumsen, T., Boches, S.K., Cotton, S.L., Grinde, B., and Paster, B.J. (2009). Microarray analysis of the microflora of root caries in elderly. Eur. J. Clin. Microbiol. Infect. Dis. *28*, 509–517.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., *et al.* (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature *464*, 59–65.

Rainer, J., Sanchez-Cabo, F., Stocker, G., Sturn, A., and Trajanoski, Z. (2006). CARMAweb: comprehensive R- and bioconductor-based web service for microarray data analysis. Nucleic Acids Res. *34*, W498–503.

Rajilic-Stojanovic, M., Heilig, H.G., Molenaar, D., Kajander, K., Surakka, A., Smidt, H., and de Vos,

W.M. (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.

Rich, V.I., Konstantinidis, K., and DeLong, E.F. (2008). Design and testing of 'genome-proxy' microarrays to profile marine microbial communities. Environ. Microbiol. *10*, 506–521.

Rigsbee, L., Agans, R., Foy, B.D., and Paliy, O. (2011). Optimizing the analysis of human intestinal microbiota with phylogenetic microarray. FEMS Microbiol. Ecol. *75*, 332–342.

Rigsbee, L., Agans, R., Shankar, V., Kenche, H., Khamis, H.J., Michail, S., and Paliy, O. (2012). Quantitative profiling of gut microbiota of children with diarrhea-predominant Irritable Bowel Syndrome. Am. J. Gastroenterol. *107*, 1740–1751.

Rimour, S., Hill, D., Militon, C., and Peyret, P. (2005). GoArrays: highly dynamic and efficient microarray probe design. Bioinformatics *21*, 1094–1103.

Rock, C., Alum, A., and Abbaszadegan, M. (2010). PCR inhibitor levels in concentrates of biosolid samples predicted by a new method based on excitation-emission matrix spectroscopy. Appl. Environ. Microbiol. *76*, 8102–8109.

Roh, S.W., Abell, G.C., Kim, K.H., Nam, Y.D., and Bae, J.W. (2010). Comparing microarrays and next-generation sequencing technologies for microbial ecology research. Trends Biotechnol. *28*, 291–299.

Salonen, A., Nikkila, J., Jalanka-Tuovinen, J., Immonen, O., Rajilic-Stojanovic, M., Kekkonen, R.A., Palva, A., and de Vos, W.M. (2010). Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. J. Microbiol. Methods *81*, 127–134.

Sanguin, H., Remenant, B., Dechesne, A., Thioulouse, J., Vogel, T.M., Nesme, X., Moenne-Loccoz, Y., and Grundmann, G.L. (2006). Potential of a 16S rRNA-based taxonomic microarray for analyzing the rhizosphere effects of maize on Agrobacterium spp. and bacterial communities. Appl. Environ. Microbiol. *72*, 4302–4312.

Schatz, M.C., Phillippy, A.M., Gajer, P., DeSantis, T.Z., Andersen, G.L., and Ravel, J. (2010). Integrated microbial survey analysis of prokaryotic communities for the PhyloChip microarray. Appl. Environ. Microbiol. *76*, 5636–5638.

Schreiner, K., Hagn, A., Kyselkova, M., Moenne-Loccoz, Y., Welzl, G., Munch, J.C., and Schloter, M. (2010). Comparison of barley succession and take-all disease as environmental factors shaping the rhizobacterial community during take-all decline. Appl. Environ. Microbiol. *76*, 4703–4712.

Sekirov, I., Russell, S.L., Antunes, L.C., and Finlay, B.B. (2010). Gut microbiota in health and disease. Physiol. Rev. *90*, 859–904.

Stralis-Pavese, N., Abell, G.C., Sessitsch, A., and Bodrossy, L. (2011). Analysis of methanotroph community composition using a pmoA-based microbial diagnostic microarray. Nat. Protoc. *6*, 609–624.

Suau, A. (2003). Molecular tools to investigate intestinal bacterial communities. J. Pediatr. Gastroenterol. Nutr. *37*, 222–224.

Suzuki, S., Ono, N., Furusawa, C., Kashiwagi, A., and Yomo, T. (2007). Experimental optimization of probe length to increase the sequence specificity of high-density oligonucleotide microarrays. BMC Genomics *8*, 373.

Tao, F., Li, C., Smejkal, G., Lazarev, A., Lawerence, N., and Schumacher, R. (2006). Pressure Cycling Technology (PCT) Applications in Extraction of Biomolecules from Challenging Biological Samples. High Pressure Biosci. Biotechnol. *1*, 166–173.

Val-Moraes, S., Marcondes, J., Carareto Alves, L., and Lemos, E. (2011). Impact of sewage sludge on the soil bacterial communities by DNA microarray analysis. World J. Microbiol. Biotechnol. *27*, 1997–2003.

Waldron, P.J., Wu, L., Van Nostrand, J.D., Schadt, C.W., He, Z., Watson, D.B., Jardine, P.M., Palumbo, A.V., Hazen, T.C., and Zhou, J. (2009). Functional gene array-based analysis of microbial community structure in groundwaters with a gradient of contaminant levels. Environ. Sci. Technol. *43*, 3529–3534.

Wang, R.F., Beggs, M.L., Erickson, B.D., and Cerniglia, C.E. (2004). DNA microarray analysis of predominant human intestinal bacteria in fecal samples. Mol. Cell Probes *18*, 223–234.

Wu, C.H., Sercu, B., Van de Werfhorst, L.C., Wong, J., DeSantis, T.Z., Brodie, E.L., Hazen, T.C., Holden, P.A., and Andersen, G.L. (2010a). Characterization of coastal urban watershed bacterial communities leads to alternative community-based indicators. PLoS One *5*, e11285.

Wu, X., Ma, C., Han, L., Nawaz, M., Gao, F., Zhang, X., Yu, P., Zhao, C., Li, L., Zhou, A., *et al.* (2010b). Molecular Characterisation of the Faecal Microbiota in Patients with Type II Diabetes. Curr. Microbiol. *61*, 69–78.

Xie, J., He, Z., Liu, X., Liu, X., Van Nostrand, J.D., Deng, Y., Wu, L., Zhou, J., and Qiu, G. (2010). GeoChip-based analysis of the functional gene diversity and metabolic potential of microbial communities in acid mine drainage. Appl. Environ. Microbiol. *77*, 991–999.

Yoo, S.M., Lee, S.Y., Chang, K.H., Yoo, S.Y., Yoo, N.C., Keum, K.C., Yoo, W.M., Kim, J.M., and Choi, J.Y. (2009). High-throughput identification of clinically important bacterial pathogens using DNA microarray. Mol. Cell Probes *23*, 171–177.

Zabarovsky, E.R., Petrenko, L., Protopopov, A., Vorontsova, O., Kutsenko, A.S., Zhao, Y., Kilosanidze, G., Zabarovska, V., Rakhmanaliev, E., Pettersson, B., *et al.* (2003). Restriction site tagged (RST) microarrays: a novel technique to study the species composition of complex microbial systems. Nucleic Acids Res. *31*, e95.

Zakharkin, S.O., Kim, K., Mehta, T., Chen, L., Barnes, S., Scheirer, K.E., Parrish, R.S., Allison, D.B., and Page, G.P. (2005). Sources of variation in Affymetrix microarray experiments. BMC Bioinformat. *6*, 214.

Zhang, L., Hurek, T., and Reinhold-Hurek, B. (2007). A nifH-based oligonucleotide microarray for functional diagnostics of nitrogen-fixing microorganisms. Microb. Ecol. *53*, 456–470.

Zhou, J.Z., He, Z.L., Van Nostrand, J.D., and Deng, Y. (2011). Development and applications of functional gene microarrays in the analysis of the functional diversity, composition, and structure of microbial communities. Front Environ. Sci. En. *5*, 1–20.

Genetic Barcoding of Bacteria and its Microbiology and Biotechnology Applications

10

Oleg N. Reva, Wai Y. Chan, Oliver K.I. Bezuidt, Svitlana V. Lapa, Larisa A. Safronova, Lilija V. Avdeeva and Rainer Borriss

Abstract

A wide variety of genetic data about organisms of interest has become available with the advancement to next generation sequencing (NGS). For many potential new users, to process this huge amount of genetic data released by NGS and to utilize this information to resolve practical questions remains a challenge. Genetic barcoding of microorganisms is the first obvious area where NGS has met the requirements of applied microbiology. In general, barcoding in microbiology is a comparative genome approach to differentiate between species or strains that are hard to distinguish by traditional methods. In this chapter, we introduce the conceptual background of bacterial barcoding and present several basic bioinformatics tools and approaches to provide solutions to NGS data handling. While working with a putative industrial strain or potentially hazardous pathogen, the following questions arise: (i) is this strain unique and if so, what makes it unique genetically or practically speaking?; (ii) how can it be detected in the environment?; (iii) are there any genetic markers for its extraordinary activity? The possibility of barcoding of whole bacterial communities is considered and both the benefits and limitations of the traditional 16S rRNA based barcoding and multi-locus sequence typing are discussed.

The history of barcoding of microorganisms

When Antonie van Leeuwenhoek looked through his miracle microscope for the first time, he was amused by the multiformity of an unseen world that he had discovered (van Zuylen, 1981). In fact the creatures that had impressed Leeuwenhoek's imagination were protists – single-celled organisms, not the usually uniform rod and coccal bacterial cells, which do not look so impressive. After the discovery of the microscope and the introduction of techniques of bacterial cultivation on solid growth media by Robert Koch and Koch's assistant J. R. Petri (Weiss, 2005), the characterization of bacteria by the morphology of the colony became common practice. Very soon it became obvious that morphology of bacterial cells and colonies is not a robust taxonomic property and that the real dimension of bacterial versatility is in their biochemistry. One of the very first biochemical tests used in microbiology was Gram-staining. This technique was developed by H.C. Gram in 1884 for the identification of pathogenic bacteria, particularly for *Typhus bacilli* (Gram, 1884). Subsequently, many diagnostic tests have been developed for the identification of various bacterial species. A comprehensive regularly updated overview of all the identification procedures used in microbiology since 1923 has been published in *Bergey's Manual of Determinative Bacteriology* (visit the Springer Web-site www.springer.com for the latest issues of the manual). A common belief of researchers was that a larger diagnostic test set would provide more reliable species identification. New approaches based on the comparison of multiple independent tests have been developed and termed as numeric taxonomy. Fuelled by the advances in computer technologies in the early 1970s, the concept of numeric taxonomy had reflected a general conceptual shift in science. As descriptive and narrative diagnostic tests used

in bacteriology gave way to digital rows of data designed for arithmetic and computer based processing, the concept of numeric taxonomy was first developed, introduced and further elaborated in works by Sokal and Sneath (1963) and Sneath and Sokal (1973). This concept was used for barcoding in microbiology where the first barcode was based on sets of biochemical tests. The early concept of barcoding was related to the phenetic approach of bacterial classification proposed by Sokal and Sneath (1963). In contrast to the cladistic approach that uses sets of hierarchical diagnostic tests for a bipalmate branching of organisms for the different taxonomic levels, the phenetics is used to search for similarities between organisms by comparison between the patterns of multiple and independent variables, i.e. barcodes. The advent of numeric taxonomy required developments of new multiplex facilities to generate massive datasets of biochemical traits and also new approaches to deal with these enormous arrays of data. Far before the introduction of sequencing techniques, numeric taxonomy had trailed bioinformatics as a new scientific discipline.

The numeric taxonomy approach was challenged (i) by the problem with standardization of experimental conditions, which at later stage was resolved to some extent by the introduction of highly standardized commercial analytical profile index (API) systems; (ii) by the biochemical versatility of bacterial species which hindered the correct species identification even by commercial test systems (Inglis *et al.*, 1998); (iii) to a greater extent also by the extraordinary plasticity of microorganisms, with the rapid evolution of bacteria under the pressure of changed environmental conditions, which may significantly differ from the parent organisms. Typical examples are small colony variants of pathogenic bacteria, which rapidly evolve and often are associated with chronic bacteraemia and a long-term persistence in host cells (Proctor *et al.*, 2006). Advancements in molecular biology and gene amplification have changed the paradigm of bacterial taxonomy from operating with rows of experimental data to comparative studies of molecular residues in biopolymers, i.e. DNA and protein molecules.

To conclude, barcodes may be defined as 400–800 bp DNA fragments serving as unambiguous species identifiers; whereas DNA barcoding is an approach for rapid species identification based on DNA sequences (Kress and Erickson, 2008). DNA barcoding in bacteriology was pioneered using 16S rRNA sequences as the taxonomic markers (Weisburg *et al.*, 1991), followed by the use of several other housekeeping protein coding gene sequences as potential barcodes (Case *et al.*, 1997). For eukaryotes, the internal transcribed spacer (ITS) region of the nuclear ribosomal DNA was proposed as a genetic species marker for fungi (Nilsson *et al.*, 2008); whereas the mitochondrial gene cytochrome *c* oxidase I (COI) was established as a universal DNA barcode for animals (Hebert *et al.*, 2003). In this chapter we discuss the applications of molecular biology and sequencing techniques for bacterial species identification and barcoding of individual organisms.

16S ribosomal RNA sequence – a universal barcode of bacterial species

The introduction of the polymerase chain reaction (PCR) amplification technique by Mullis (1993) has revolutionized molecular biology. PCR provided scientists the ability to obtain multiple copies of precisely selected DNA fragments. Both the quantity and quality of PCR product can be further analysed by electrophoresis, direct sequencing and other alternative methods. PCR amplification has become easy to standardize and also guaranteed reproducible results in different laboratories. A gene encoding the small 16S rRNA ribosomal subunit was found to be a universal target for phylogenetic studies (Pace, 1997). The application of this method aided in classifying bacteria at levels from prokaryotic domains to individual strains (Woese and Fox, 1977; Dalevi *et al.*, 2007). Notably, this method provided researchers with a universal genetic barcode for bacteria. The 16S rRNA gene is extremely conserved in Archaea and eubacteria and allows the construction of universal primers that enclose several informative variable regions (Coenye and Vandamme, 2003). 16S rRNA remains one of the most sequenced DNA fragments for species identification (see Chapter 8). For its unprecedented