MICROARRAY TECHNOLOGY AND APPLICATIONS IN ENVIRONMENTAL MICROBIOLOGY

Jizhong Zhou and Dorothea K. Thompson

Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA

Advances in Agronomy, Volume 82

0065-2113/03 $35.00

I. INTRODUCTION

Microarrays are miniaturized arrays of hundreds to thousands of discrete DNA fragments or synthetic oligonucleotides that have been attached to a solid substrate (e.g., glass) using automated printing equipment such that each spot (element) in a fixed position on the array corresponds to a unique DNA (Schena *et al*., 1998; Schena, 2003). Microarrays are variously referred to as microchips, biochips, DNA chips, or gene chips and have emerged as a widely accepted functional genomics technology for large-scale genomic analysis. In particular, DNA or oligonucleotide arrays have been used to monitor messenger RNA (mRNA or transcript) abundance levels of differentially expressed genes under different cell growth conditions or in response to environmental perturbations or genetic mutations (c.f., Lockhart *et al*., 1996; Schena *et al.*, 1996; DeRisi *et al*., 1997; Wodicka *et al.,* 1997; Richmond *et al*., 1999; Ye *et al*., 2000; Thompson *et al*., 2002; Liu *et al.*, 2003) and to detect specific mutations in DNA sequences (Hacia, 1999; Broude *et al*., 2001). Recently, the potential research applications of microarray technology to studies in microbial ecology have been explored (Zhou and Thompson, 2002; Zhou, 2003).

In principle and practice, microarrays are extensions of conventional membrane-based Southern and Northern hybridization blots, which have been used for decades to detect and characterize nucleic acids in diverse biological samples. Microarray hybridization is based on the association of a single-stranded molecule labeled with a fluorescent tag, or fluorescein, with its complementary molecule, which is covalently attached or immobilized to a solid support, usually glass. In such an assay, the specific hybridization pattern or gene expression profile generated by an unknown (experimental) sample is typically compared with a control (reference) pattern. In microarray terminology, the fluorescein-labeled DNA in solution is generally termed the target, and the DNA strand immobilized on the microarray surface is referred to as the probe. Because the sequence of the arrayed molecule is usually known, it is used to 'probe' or investigate the unknown target molecule in solution. This is directly opposite to the convention established with the development of Southern blot hybridization,

in which target molecules fixed to a porous membrane are interrogated by known solution-phase probes.

The concept of microarrays was first proposed in the late 1980s. One of the first descriptions of DNA microarrays in the literature was provided by Augenlicht and his colleagues, who spotted 4000 complementary DNA (cDNA) sequences on nitrocellulose and used radioactive labeling to analyze differences in gene expression patterns among different types of colon tumors in various stages of malignancy (Augenlicht *et al.,* 1987, 1991). At the same time, four separate research groups simultaneously developed the concept of determining a DNA sequence by hybridization to a comprehensive set of oligonucleotides, i.e., sequencing by hybridization or SBH (Bains and Smith, 1988; Drmanac *et al*., 1989; Khrapko *et al*., 1989; Southern *et al*., 1992). Although SBH is an extremely elegant alternative to conventional DNA sequencing, various inherent problems associated with repeated sequences and the imperfect specificity of hybridization limit the practicality of using SBH for routine sequence determination. Such technical challenges, therefore, have led researchers to focus on the more readily addressable applications of microarray technology, such as gene expression profiling. By the mid-1990s, the reverse dot-blot scheme for monitoring genomic expression was reformulated by several different groups. Both DNA fragments and synthetic oligonucleotides were arrayed on various substrates, including nylon membranes, plastic and glass (Schena *et al.*, 1995; Lockhart *et al.*, 1996). All of them depended on sequence-specific hybridization between the arrayed DNA and the labeled nucleic acids from cellular mRNA. Later studies using the simple eukaryote, yeast, clearly demonstrated that DNA and oligonucleotide arrays are powerful tools for monitoring global gene expression (DeRisi *et al.*, 1997; Wodicka *et al.*, 1997).

Microarray-based genomic technology has greatly benefited from many parallel advances in other fields. Without such advancements, the development of high-density microarrays and the various applications that we see today would not be possible (Eisen and Brown, 1999; Schena and Davis, 2000). For example, large-scale genome sequencing projects have produced the raw sequence information needed for microarray expression profiling, and the development of robotic printers or arrayers has made it possible to fabricate high-density microarrays in a very small area. In addition, recent advances in methods of fluorescent labeling and detection offer significant advantages in speed, data quality, and user safety for microarray-based assays. Together, these technical advancements have enabled microarray-based genomic technologies to revolutionize genetic analyses of biological systems. The widespread, routine use of such genomic technologies will shed light on a wide range of important research questions associated with the genetic programs controlling cell growth and differentiation, bacterial pathogenesis and the host response to infection, antibiotic resistance, specialized metabolic capabilities of microorganisms of bioremediation potential, as well as agricultural and pharmaceutical applications.

In this chapter, we review the technical underpinnings of microarray hybridization and its applications in environmental microbiology, with emphasis on the most important issues of microarray-based assays as outlined in Fig. 1. We first describe the technologies used for microarray fabrication, followed by a discussion on microarray hybridization, fluorescence detection technologies, image processing, and data analysis. In addition, we provide an overview of the recent applications of microarray technologies to study gene expression patterns in environmentally important microorganisms. Finally, we describe various types of microarrays specifically developed for analyzing microbial community composition and function in natural environments. Because glass-based DNA microarrays are currently preferred by most basic research laboratories, our discussion of microarray technologies will focus on this type of array, while other types will be mentioned only briefly. It should be noted that the goal of this chapter is to provide an in-depth description of the basis and principles of

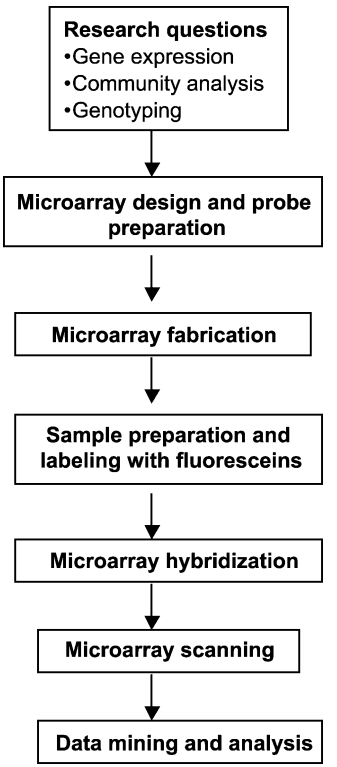


Figure 1 A flow chart of a microarray experiment.

microarray-based technologies rather than an exhaustive review of current microarray technology.

II. MICROARRAY TYPES AND ADVANTAGES

A. Types of Microarrays

Microarrays can be divided into two major formats based on the type of immobilized probes: (1) *DNA microarrays* constructed with DNA fragments typically generated using the polymerase chain reaction (PCR) (Schena *et al.*, 1995; DeRisi *et al.*, 1997; Marshall and Hodgson, 1998); and (2) *oligonucleotide microarrays* constructed with shorter (10- to 40-mer) or longer (up to 120-mer) oligonucleotide sequences that are designed to be complementary to specific coding regions of interest.

DNA microarrays have certain advantages over oligonucleotide microarrays, especially for monitoring gene expression patterns. While oligonucleotide microarrays are limited to array elements of low sequence complexity, the specificity of hybridization for a complex probe is improved with arrays containing DNA fragments that are substantially longer than oligonucleotides (Shalon *et al.*, 1996). In addition, oligonucleotide synthesis requires prior sequence knowledge, whereas DNA arrays do not because DNA fragments of unknown sequences can be amplified from clones using vector-specific primers. For microarrays constructed with PCR-amplified DNA elements, nucleic acids of virtually any length, composition, or origin can be arrayed (Shalon *et al.*, 1996). However, oligonucleotide-based microarrays have the advantage of minimizing the potentially confounding effects of occasional cross-hybridization (Wodicka *et al.*, 1997) and are uniquely suited for detecting genetic mutations and polymorphisms. Since oligonucleotide probes can be commercially synthesized, the handling and tracking of oligonucleotide array elements, unlike PCR products, is generally easier. Amplifying all of the probes with a desired minimum quantity for printing is labor intensive and time-consuming.

Based on probe immobilization and fabrication strategies, there are two general types of oligonucleotide microarrays:

1. Direct parallel synthesis on solid substrates by light-directed or photoactivatable chemistries (Pease *et al.*, 1994; Lipshultz *et al.*, 1999) or standard phosphoramidite chemistries (Southern *et al.*, 1994); or
2. Chemical attachment of pre-made oligonucleotides to solid supports (Khrapko *et al.*, 1989; Beattie *et al.*, 1992, 1995; Eggers *et al.*, 1994; Lamture *et al.*, 1994; Fotin *et al.*, 1998; Guschin *et al.*, 1997a, b; Rehman *et al.*, 1999; Rogers *et al.*, 1999).

Each strategy for oligonucleotide immobilization has its own set of specific advantages and disadvantages (Schena *et al.,* 1998; Hoheisel, 1997). The direct or *in situ* synthesis approach has two major advantages. First, the photoprotected versions of the four DNA bases allow microarrays to be manufactured directly from DNA sequence databases, thereby removing the uncertain and burdensome aspects of sample handling and tracking. Second, the use of synthetic reagents minimizes variations between arrays by ensuring a high degree of precision in each coupling cycle. Costliness is a major disadvantage of the photolithographic approach; photomasks, which direct light to specific areas on the array for localized chemical synthesis, are very expensive and time-consuming to design and build. Also, the yield and length of the synthesized oligonucleotides are subject to wide variation and uncertainty, which could lead to unpredictable effects on hybridization across the microarray. A major advantage of the attachment of pre-synthesized probes is that the concentration and length of each oligonucleotide on the array can be controlled prior to immobilization. Standard synthesis chemistry is also well established for many nucleotide derivatives for which no light-inducible monomer equivalents are available. In addition, the post-synthesis approach is less complicated and can be customized according to the specific needs of the laboratory. The critical drawback of the post-synthesis approach, however, is still the need for the external synthesis and storage of different oligonucleotides prior to array fabrication.

B. Advantages of Microarrays

Microarrays offer the following advantages over conventional nucleic acid-based approaches.

1. High-throughput and Parallel Analysis

Microarray technology allows thousands to hundreds of thousands of array elements or probes to be uniformly deposited in a very small area on the surface of a non-porous substrate. Consequently, the high-density capacity of microarrays permits parallel analysis, in which the expression of the entire gene content of a genome of interest can be monitored, or many constituents of a microbial community can be simultaneously assessed in a single assay using the same microarray. Genomic expression data allows researchers to begin to build a comprehensive, integrated view of a complex biological system.

2. High Sensitivity

High sensitivity can be achieved in probe-target hybridization, because microarray hybridization uses a very small volume of probe and the target nucleic acid is restricted to a small area (Shalon *et al.*, 1996; Guschin *et al.*, 1997a). This feature enables high sample concentrations and rapid hybridization kinetics.

3. Differential Display

Multi-fluorescence detection schemes allow differential display of different biological samples. Different target samples, for example, can be labeled with different fluorescent tags and then hybridized in parallel to the same microarray, allowing the simultaneous analysis of two or more biological samples in a single assay. Multi-color hybridization detection minimizes variations resulting from inconsistent experimental conditions and allows direct and quantitative comparison of target sequence abundance among different biological samples (Shalon *et al.*, 1996; Ramsay, 1998).

4. Low Background Signal Noise

Non-porous surfaces substantially reduce the amount of non-specific hybridization; as a result, organic and fluorescent compounds that attach to microarrays during fabrication and hybridization procedures can be rapidly removed by post-hybridization washing, resulting in considerably less background signal noise than is typically encountered with porous membranes (Shalon *et al.*, 1996).

5. Real-time Data Analysis

Once the microarrays are constructed, hybridization and detection are relatively simple and rapid, allowing real-time data analysis in field-scale heterogeneous environments.

6. Automation

Microarray technology is amenable to automation and therefore, has the potential of being cost-effective compared to traditional hybridization methods (Shalon *et al.*, 1996).

III. MICROARRAY FABRICATION

A. Microarray Substrates

The substrate used for printing microarrays has a large impact ultimately on the quality of the data obtained from microarray hybridizations. Poor surface treatment may result in poor attachment of the DNA probes to the slide, and a non-uniform surface will cause variations in the amount of the attached DNA. Furthermore, residual substances deposited on the slide surface during a microarray experiment can lead to high background fluorescence or noise. Thus, the selection of appropriate substrates for microarray experiments is of critical importance. The substrates used for fabricating microarrays fall into two categories: porous and non-porous.

1. Non-porous Substrate

At present, a non-porous solid surface is the most common type of substrate used for printing arrays. Several non-porous materials, such as glass and polypropylene, are suitable for microarray fabrication. Glass slides are the most widely used substrates, because they are inexpensive, possess physical characteristics *advantageous to* hybridization, and are easily modified for nucleic acid attachment and synthesis (Southern, 2001). In general, non-porous substrates offer a number of advantages over porous substrates (Schena and Davis, 2000). First, small amounts of molecules may be deposited at precise, predefined positions on the substrate surface with little diffusion, thus enabling the high-density capacity of microarrays. Second, hybridization between target and probe molecules occurs at a much faster rate on non-porous solid surfaces than on porous substrates. This is because molecules do not have to diffuse in and out of the pores and as a result, steric inhibition of hybridization is not a problem (Southern, 2001). Third, because small sample volumes can be applied to a non-porous surface under a coverslip, high probe concentrations, rapid hybridization kinetics, and high sensitivity can be achieved. Fourth, non-porous substrates prevent the absorption of reagents and samples into pores, thus allowing unbound labeled materials to be easily removed. This expedites the procedure, improves reproducibility and reduces background. Fifth, a non-porous substrate has low intrinsic fluorescence and thus allows the use of fluorescence detection. Finally, a solid substrate offers a homogeneous attachment surface, and its inherent uniform flatness permits true parallel analysis. However, a major drawback of non-porous substrates, such as glass, is the susceptibility to dust and other airborne particle contamination, which can cause a scanned slide to appear “dirty”. Poor modification of microarray slides is another common cause of poor hybridization

results (Southern, 2001). In addition, because of the planar surface, the capacity for immobilization is limited, and consequently, the sensitivity of the assay is relatively low compared to that of the porous substrates (Afanassiev *et al.*, 2000).

2. Porous Substrates

Porous substrates such as nitrocellulose and nylon membranes have also been used for microarray fabrication (Englert, 2000). The principal advantage of membranes is that larger volumes and concentrations of samples can be immobilized on a small area, because the pores of the substrates provide a larger total surface area for binding. As a result, a relatively higher sensitivity and better dynamic range for quantitative comparison can be achieved. Homogeneous spots are also more readily obtained, because deposited samples are able to distribute immediately into the membrane through capillary flow. In addition, membrane-based microarrays can be reused several times (Beier and Hoheisel, 1999).

However, there are some important disadvantages in using porous membranes for microarray fabrication. The boundaries and shapes of the spots are poorly defined, and membranes swell in solvent, and shrink and distort when dried. Such fragility and flexibility make it difficult to precisely locate probe positions during spotting and image analysis. Also, many membranes have high intrinsic fluorescence and thus higher background noise compared to non-porous substrates. In addition, because the spot sizes on a membrane cannot be reduced to a level comparable with glass slides or other non-porous substrates, much more DNA is required for producing a membrane-based microarray (Beier and Hoheisel, 1999).

Overall, non-porous substrates are preferred for microarray experiments, even though porous substrates have some advantages. This is because the unique physical and chemical characteristics of glass slides (e.g., little diffusion, low intrinsic fluorescence) allow miniaturization and use of fluorescent labeling and detection, which are the most critical requirements for large-scale genomic analysis. The miniaturized microarray format coupled with fluorescent detection represents a fundamental revolution in biological analysis (Schena and Davis, 2000).

B. Surface Modification for the Attachment of Nucleic Acids

1. Attachment Strategies

Appropriate attachment and retainment of nucleic acid probes to an array surface is very important for microarray analysis. For reliable microarray

hybridization, the attachment chemistry must meet the following criteria: (i) nucleic acids must be tightly (covalently) bound to the array surface; (ii) the surface-bound molecules must be accessible for hybridization; and (iii) the attachment chemistry must be reproducible. Both ionic interaction and covalent bonding, which are described in more detail below, are used for attaching nucleic acids to solid surfaces, depending on the size of the nucleic acid molecules.

Electrostatic interaction. Long DNA fragments (in the order of several hundred bases in length) can be immobilized on a glass surface through ionic interaction between the negatively charged phosphodiester backbone and the positively charged slide surface (Fig. 2A). Recent studies showed that synthetic oligonucleotides more than 70 bp in length can also be bound to glass surfaces through ionic interaction (Hughes *et al.*, 2001). Generally, an amine or lysine coating is used to adsorb DNA to glass slides. Because amines have a positive charge at neutral pH, they allow attachment of native DNA molecules through the formation of ionic bonds with the negatively charged phosphate backbone. Electrostatic attachment can be enhanced by exposing the fabricated arrays to ultraviolet light or heat, which induces free radical-based coupling between thymidine residues in the DNA and carbons on the alkyl amine. The combination of electrostatic bonding and non-specific covalent attachment, links native DNA to the substrate surface in a stable manner. Although ionic interaction-mediated attachment is less expensive and more versatile than covalent bonding (Worley *et al.*, 2000), the immobilized DNA molecules are susceptible to removal under high salt and/or high temperature conditions. Therefore, covalent binding methods are preferred.

Covalent bonding. DNA can also be covalently attached to glass surfaces using different attachment chemistries (Fig. 2B). Although long DNA molecules can be attached covalently to the microarray surface by different methods, immobilization of aminated DNA to an aldehyde-coated slide is the usual method of choice (Zammatteo *et al.*, 2000).

Because oligonucleotides are typically short, covalent bonding is generally required for attachment of such molecules to a glass surface. Usually, oligonucleotides are fixed covalently onto solid surfaces at one end of the molecule using a variety of methods. The attachment of biomolecules to a solid phase presents some problems that are unique to homogeneous solutions. Because the bound probe is not free to diffuse, a lower reaction rate is expected. In addition, target molecules in solution may not be able to effectively interact with the bound probes due to steric hindrance from the solid support and the close proximity of other bound probes (Shchepinov *et al.*, 1997). Additional molecules, termed linkers or spacers, are used to tether the probe to the substrate surface, thereby providing a sufficient amount of distance between the

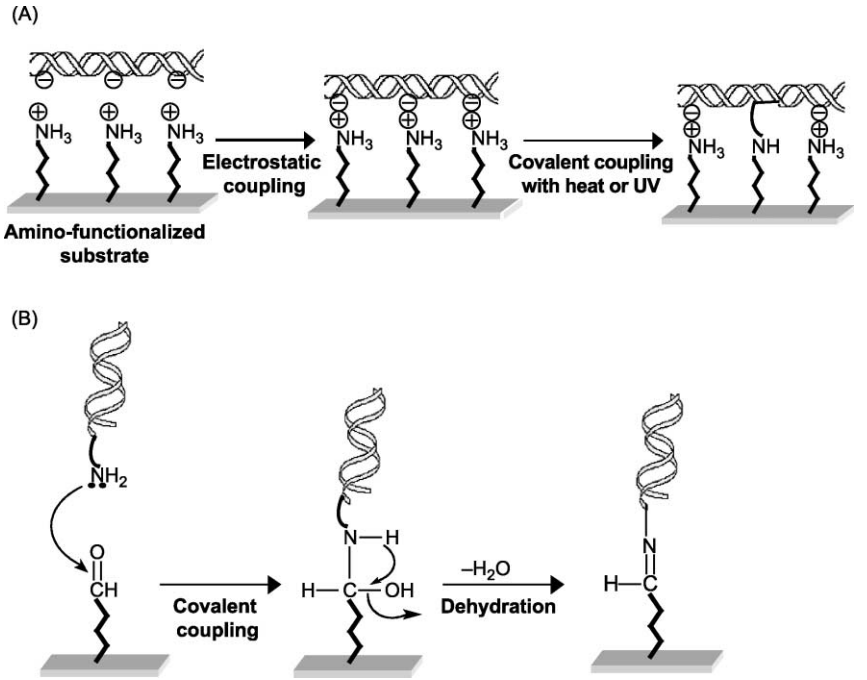


Figure 2 Attachment strategies. (A) Attachment of nucleic acids to solid surfaces through electrostatic interactions. The microarray substrates contain primary amine groups (NH_3^+) attached covalently to the glass surface. The amines carry a positive charge at neutral pH, which permits attachment of native DNA through the formation of ionic bonds with the negatively charged phosphate backbone. Covalent attachment of DNA to the surface can be further achieved by treatment with ultraviolet light or heat. (B) Attachment of nucleic acids to solid surfaces through covalent bonding. The microarray substrates contain primary aldehyde groups attached covalently to the glass surface. Primary amino linkers (NH_2) on the DNA attack the aldehyde groups to form covalent bonds. Such attachment is stabilized with a dehydration reaction by drying in low humidity, which leads to Schiff base covalent bond formation.

oligonucleotide probe and the support to minimize steric interference. To serve as an effective linker, the molecule must meet several criteria (Guo *et al.*, 1994). First, the linkage must be chemically stable under the hybridization conditions used and must be sufficiently long to minimize steric interference. Second, the linker should be sufficiently hydrophilic to be freely soluble in aqueous solution. Third, there should be no non-specific binding of the linker to the support. Shchepinov *et al.* (1997) showed that the optimal linker for immobilizing oligonucleotides at either the 5′ or 3′ terminus should have low negative charge density and a length of 30–60 atoms.

Generally, a linker is coupled to the probe during oligonucleotide synthesis. Various types of linkers have been used, including poly-dT (Guo *et al.*, 1994; Rehman *et al.*, 1999; Afanassiev *et al.*, 2000), poly carbon atoms (Afanassiev *et al.*, 2000), and oligodeoxyribonucleotides with hairpin stem-loop structures (Zhao *et al.*, 2001). Studies have shown that the length of the spacer has a significant impact on the success of hybridization. While virtually no hybridization signal was observed for poly-dT spacers (15 bp), about a 20-fold enhancement of hybridization was obtained for a poly-dT spacer of 15 nucleotides (Guo *et al.*, 1994). The effect of linkers composed of multiple carbon atoms (e.g., C_{36}, C_{18}, C_{12}, and C_6) on microarray hybridization was also examined (Afanassiev *et al.*, 2000). Overall, the signal intensity was improved with a longer C linker. A spacer can also be added by chemically modifying the slide surface (Guo *et al.*, 1994; Beier and Hoheisel, 1999).

Oligonucleotides can be immobilized onto solid supports through homo-bifunctional or hetero-biofunctional cross-linkers. For example, amino-modified oligonucleotides can be covalently attached to glass surfaces containing amine functional groups through homo-biofunctional cross-linkers (Guo *et al.*, 1994) and to glass surfaces containing aldehyde and epoxide through hetero-biofunctional cross-linkers (Lamture *et al.*, 1994; Schena *et al.*, 1996). Thiol-modified or disulfide-modified oligonucleotides can be attached to the glass surface containing amine functional groups *via* hetero-biofunctional cross-linkers (Chrisey *et al.*, 1996). A hetero-biofunctional cross-linker is generally preferred over a homo-biofunctional cross-linker to prevent surface-to-surface linkages and probe-to-probe linkages as opposed to the desired surface-to-probe linkages (Steel *et al.*, 2000). When using a hetero-biofunctional cross-linker, the probe should have a different modification chemistry from the array surface.

While microscope slides made from low-fluorescence glass are suitable substrates for microarray construction, the glass surface must be modified with a chemical coating and cleaned (e.g., free of dust particles) before use. The glass surface must have suitable functional groups for the attachment of target DNA molecules, because DNA does not inherently bind to untreated glass. A hydrophobic surface is essential for achieving high-density spots, because spotted hydrophilic samples will spread less on a hydrophobic surface than on an untreated hydrophilic glass surface (Rose, 2000).

The quality of slide coating ultimately impacts the quality of the microarray data. A poor surface coating can result in poor probe retention. For spot-to-spot consistency, the coating must be uniform and homogeneous without untreated patches. In addition, the coating must be non-fluorescent and capable of resisting harsh physical conditions, such as boiling, baking and soaking. Silanization, dendrimeric linker coating, gel coating, and nitrocellulose coating are types of glass surface modifications that are discussed in greater detail in the following sections.

2. Silanization

Silanes are most commonly used for slide surface modification to provide organic functional groups for the covalent attachment of biomolecules (Shriver-Lake, 1998), and silanized glass slides are commercially available from a number of companies. Glass slides can be modified to contain surface hydroxyls that react with methoxy or ethoxy residues of a silane molecule. Many different commercially available silanes contain various functional reactive groups such as amino, epoxide, carboxylic acid and aldehyde, which are suitable for covalent bonding with appropriately modified biomolecules (Schena, 2003). Most microarray analyses are performed with slide surfaces that contain reactive amine and aldehyde groups (Schena, 2003), which allow attachment of biomolecules *via* electrostatic interactions or covalent bonding. Silanization can be accomplished simply by immersing the slides into a silane-containing solution or by vapor deposition (Steel *et al.*, 2000; Worley *et al.*, 2000). Vapor-phase coating is most effective at uniformly depositing a monolayer of silane on the slide surface (Chrisey *et al.*, 1996; Worley *et al.*, 2000).

3. Dendrimeric Linker Coating

To increase the binding capacity of arrayed probes, a more elaborate chemistry was developed for synthesizing dendrimeric linkers on silanized glass slides to allow covalent attachment of aminated DNA molecules (Beier and Hoheisel, 1999). Such a linker system multiplies the coupling sites by introducing additional reactive groups through branched linker molecules. There are several advantages of this linker system. First, it allows covalent immobilization of both pre-synthesized and *in situ* synthesized oligonucleotides on glass slides. Second, the dendrimeric linker system increases the loading capacity by a factor of 10. Third, it eliminates non-specific attachment of hybridization probes and provides a low fluorescent background. Fourth, covalent bonding is stable, and the microarrays can be reused many times. Finally, bonding through a terminal group of the attached molecules produces no apparent negative effects on hybridization efficiency.

4. Gel Coating

Recently developed attachment strategies that use polyacrylamide or agarose for surface modification combine the advantages of porous and non-porous substrates (Afanassiev *et al.*, 2000; Zlatanova and Mirzabekov, 2001). In this approach, polyacrylamide gel elements or pads, ranging in size from $10 \times 10 \times 5$

to $100 \times 100 \times 20$ μm^3 with volumes varying from picoliters to nanoliters, are affixed to the glass surface. Because each gel-pad is surrounded by a hydrophobic surface that prevents solution diffusion among the elements, they can function independently. The probe molecules are immobilized in the gel-pads by robotic application. Compared to the direct attachment of probes to solid supports, the use of polyacrylamide gel-pad as an immobilization support offers some important advantages (Drobyshev *et al.*, 1999). Three-dimensional immobilization of probes in gel-pads provides a greater density capacity and a more homogeneous environment than heterophase immobilization on glass, leading to higher sensitivity and a faster hybridization rate (Vasiliskov *et al.*, 1999). However, like nylon membrane-based supports, gel-pads can yield higher background levels (Beier and Hoheisel, 1999). In addition, there are restrictions on the size of the molecules that can diffuse into the gel, such that fragmentation of the probe and target DNA may be required to generate molecules of the appropriate size (Englert, 2000).

The convenience of using the gel-pad attachment strategy is limited by the fact that the method requires activation of gels and probes with labile reactive chemicals (Rehman *et al.*, 1999). A more flexible attachment method using co-polymerization of 5′-terminal modified oligonucleotides with acrylamide monomers was developed (Rehman *et al.*, 1999). The advantages of this method are that probes can be prepared easily using standard DNA synthesis chemistries and probes can be specifically and efficiently immobilized in the absence of highly reactive and unstable chemical crosslinking agents.

Agarose was also examined as a coating material for probe attachment (Afanassiev *et al.*, 2000). Agarose film is activated to produce reactive sites that permit covalent immobilization of molecules with amino groups. Agarose has a higher binding capacity compared to a glass-based planar surface and does not interfere with fluorescent detection. In contrast to acrylamide gels (Zlatanova and Mirzabekov, 2001) and dendrimeric branched systems (Beier and Hoheisel, 1999), this method does not require complex preparation technology.

5. Nitrocellulose Coating

Glass slides coated with a proprietary nitrocellulose-based polymer have also been examined as an immobilization support (Stillman and Tonkinson, 2000). The nitrocellulose-based polymer can bind biomolecules (both DNA and proteins) noncovalently but irreversibly, providing better spot-to-spot consistency, higher binding capacity, and greater dynamic range compared to other glass slide modifications. Nitrocellulose-coated slides are also suitable for fluorescent detection due to their relatively low-light scattering capacity. Although they

have a higher fluorescent background, the background-subtracted signal is significantly higher on this support than the glass support coated with polylysine. However, the potential for miniaturizing the array dimensions of such slides remains to be determined.

C. Arraying Technology

Microarray fabrication involves the printing and stable attachment of DNA probes on the array (Fig. 1). The microarray format is compatible with many advanced printing technologies, of which the most widely used are photolithography, mechanical microspotting, and ink-jet ejection. Each fabrication technology possesses advantages and disadvantages (Schena *et al*., 1998; Schena and Davis, 2000). All three technologies allow the manufacture of microarrays with sufficient density for genetic mutation detection and gene expression profiling applications. The key considerations in selecting a fabrication technology include microarray density and design, biochemical composition and versatility, reproducibility, high-throughput capacity, and cost. Because of its versatility, affordability, and wide applications, microspotting is likely to become the printing technology of choice for the basic research laboratory. Thus, our discussion in this section will focus primarily on microspotting technology.

1. Light-directed Synthesis

In photolithography, oligonucleotides are synthesized *in situ* on a solid surface in a predefined spatial pattern by using a combination of chemistry and photolithographic methods borrowed from the semiconductor industry (Fodor *et al.*, 1991; McGall and Fidanza, 2001) (Fig. 3). Briefly, a glass or fused silica substrate is covalently modified with a silane reagent to obtain a surface containing reactive amine groups, which are then modified with a specific photoprotecting group, namely methylnitropoperonyloxycarbonyl (MeNPOC). Then the specific regions of the surface are activated through exposure to light, and a single base is added to the hydroxyl groups of these exposed surface regions using a standard phosphoramidite DNA synthesis method. The process of photodeprotection and nucleotide addition is iterated until the desired sequences are generated. Typically, the probes synthesized *in situ* on the arrays are 20–25 bp in length. Since the average stepwise efficiency of oligonucleotide synthesis ranges from 90–95%, the proportion of the full-length sequences for 20-mer probes is approximately 10%. However, this should have a relatively minor effect on the performance of microarray hybridization because

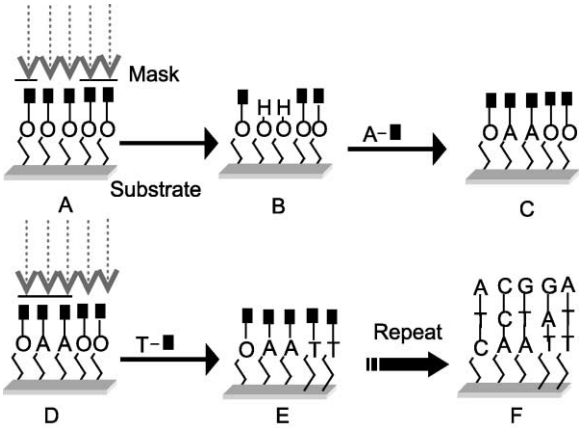


Figure 3 *In situ* light-directed oligonucleotide probe array synthesis. The solid surface contains linkers with a photolabile protecting group X (black box) (e.g., MeNPOC). MeNPOC is resistant to many chemical reagents but it can be removed selectively by using ultraviolet light for a short time. When MeNPOC is removed, the deprotected region on the surface can form chemical bonds with DNA bases containing a MeNPOC photoprotecting group at its 5′ hydroxyl position. In this illustration, light is directed through a photolithographic mask to specific areas of the array surface, which are activated for chemical coupling. The first chemical building block A containing a photolabile protecting group X is then attached. Next, light is directed to a different region of the array surface through a new mask. The second chemical building block T containing a photolabile protecting group X is also added. This process is repeated until the desired product is obtained.

of the high absolute amount of full-length probes on the support (McGall and Fidanza, 2001).

Another emerging light-directed synthesis approach for constructing high-throughput oligonucleotide arrays is to use a digital light processor (DLP), i.e., micromirror (Nuwaysir *et al.*, 2002). This maskless array synthesizer (MAS) technology uses DLP to create "virtual" masks that direct an ultraviolet light beam to discrete locations on a glass substrate for DNA synthesis. Similar to the Affymetrix photolithographic approach, MAS is capable of constructing high-density microarrays containing any desired nucleotide sequence. In contrast, MAS does not require photomasks, which are very expensive and time-consuming to manufacture. The MAS technology makes photolithography much more flexible and user-friendly, although it is still in the early development stages (Nuwyasir *et al.*, 2002).

Photolithographic parallel synthesis offers a very efficient approach to high-density array fabrication, in which the maximum achievable density is ultimately dependent on the spatial resolution of the photolithographic process. Due to steric and/or electrostatic repulsive effects, there is an optimum probe density for maximum hybridization signal. Affymetrix chips currently contain ~250,000

oligonucleotides in an area of 1 cm^2. One of the main advantages of this approach is that microarrays of extremely high-density can be constructed (Ramsay, 1998), but only oligonucleotides can be used in photolithography.

2. Contact Printing

The most commonly used microarray fabrication technology is mechanical microspotting, which uses direct contact of computer-controlled multiple pins, tweezers, or capillaries to deliver picoliter volumes of pre-made biochemical reagents (e.g., oligonucleotides, cDNA, genomic DNA, antibodies, or small molecules) to a solid surface. Currently, more than 1,000 individual cDNA molecules can be deposited in an area of 1 cm^2 using this technology (Rose, 2000). The advantages of microspotting include ease of implementation, low cost, and versatility, while a major disadvantage is that each sample to be arrayed must be prepared, purified, and stored prior to microarray fabrication. In addition, microspotting rarely produces the densities that can be achieved with photolithography. The various pin technologies for microspotting are described below.

Solid pins. Solid pins have either flat or concave tips. Because such tip can accommodate only a relatively small volume of the sample, only one microarray can be printed generally with a single sampling load. Consequently, the overall printing process is slow, making this technology suitable only for constructing low-density arrays. Additionally, loss of sample due to evaporation is a significant problem due to the large surface-to-sample volume ratio of solid pins. Under standard laboratory conditions, about half of a 250 pl volume is lost in 1 s (Mace *et al.*, 2000). To minimize evaporation loss, a highly humid environment is absolutely necessary. However, high humidity may prevent the sample from drying sufficiently on the slide, resulting in sample migration or spreading.

Split pins. Split pins have a fine slot at the end of the pin for sample holding. When the slit pin is dipped into the sample solution, the sample is loaded into the slot, which generally holds 0.2–1.0 μl of sample solution. A small volume of sample (0.5–2.5 nl) is deposited on the microarray by tapping the pins onto the slide surface with sufficient force (Rose, 2000) or touching the pins lightly on the surface like an ink stamp (Martinsky and Haje, 2000). The company, TeleChem International, manufactures a split pin by using digital control, so that there are virtually no variations in mechanical quality from pin to pin (Martinsky and Haje, 2000). Thus, TeleChem pins provide very high printing consistency under conditions of good sample preparation, proper motion control and homogeneous printing substrate. Because the split pins hold a larger sample volume than

the solid pins, more than one microarray can be printed from a single sampling event. Although split pin technology has been used successfully to print microarrays, one of the drawbacks is that dust, particulates, evaporated buffer crystals and/or other contaminants can clog the pin slot. Tapping the pins on the substrate surface, however, is not recommended for various reasons (Martinsky and Haje, 2000). Physical tapping leads to bulk transfer of the sample from the pins and hence causes non-uniformly large spots and spot merging. Also, tapping on the slide surface may lead to deformation of the pin tip, causing larger spots, irregular spot shapes, a larger amount of sample deposition, and poor printing quality (Mace *et al.*, 2000). In addition, tapping may fracture the surface coating and cause irregular spot shapes such as doughnut shapes, in which the center of the spot lacks probe material (Martinsky and Haje, 2000).

Pin and Ring. This is a variation of the pin-based printing process. The sample is taken by dipping the ring into the sample well and then a small volume of sample solution is deposited onto the slide surface by pushing the sample captured in the ring using solid pins (Mace *et al.*, 2000). Different sized rings can be selected to hold 0.5–3.0 μl of sample. Many different spot sizes can be obtained by simply using pins of different diameter. In addition, loss of sample due to evaporation is alleviated by minimizing fluid exposure through specific ring configuration.

The pin-and-ring arraying technology offers a number of advantages. Since the pin is used only for spotting and the sample fluid captured by the ring is relatively large, the deposition of samples on different slides is accomplished in an identical manner for each printing cycle, yielding a microarray fabrication quality that is consistent and reproducible (Mace *et al.*, 2000). In addition, the ring geometry is capable of handling a wide variety of volumes and fluid viscosities. Unlike the split pin, the pin and ring configuration is not susceptible to clogging by the accumulation of dust, particulate matter, high-viscosity fluids, debris, buffer or salts, and other materials. Finally, the pin and ring can deposit samples on soft substrates such as agar, gels and membranes. However, one drawback of this technology is sample loss. The majority of samples captured by the ring cannot be used for spotting and is lost through washing for the next sampling cycle.

3. Non-contact Ink-jet Printing

Ink-jet ejection technologies provide another means of fabricating microarrays. In this approach, the sample is taken from the source plate, and a droplet of sample is ejected from the print head onto the surface of the substrate. Similar

to microspotting, ink-jet ejection allows the spotting of virtually any biological molecule of interest, including cDNA, genomic DNA, antibodies, and small molecules. In contrast to microspotting, ink-jets have the advantage of avoiding direct surface contact but cannot be used to manufacture microarrays as dense as those prepared by photolithography or microspotting approaches.

Currently two types of non-contact ink-jet print technologies, piezoelectric pumps and syringe-solenoid, are used for printing microarrays.

Piezoelectric Pumps. This printing technology utilizes a piezoelectric crystal, which contacts a glass capillary containing the sample liquid (Englert, 2000; Mace *et al.*, 2000; Rose, 2000) and is still in the early stages of development. When the crystal is biased with a voltage and subsequently deformed, the capillary is squeezed and a small volume (0.05–10 nl) of fluid is ejected through the tip from the reservoir. Piezoelectric printing has the advantages of an extremely fast dispensing rate (on the order of several thousand drops per second), very small print volumes, and consistency of droplet size. The main problem with this technology is clogging by air bubbles and particulates, which makes the system less reliable compared to other printing methods. In addition, the void volume of sample solution contained in the capillary is very large (100–500 μl) and not recoverable. It is also difficult to change samples using piezoelectric printing.

Syringe-solenoid Printing Technology. This technology uses a syringe pump and a microsolenoid valve for dispensing samples (Rose, 2000). The sample is taken by a syringe, and sample droplets, ranging from 4 to 100 nl in volume, are ejected by pressure onto the surface through the microsolenoid valve. The main advantages of this technology are reliability and low cost. However, it is not suitable for fabricating high-density microarrays because of the large printing volume and spot size.

D. Critical Issues for Microarray Fabrication

This section highlights some practical issues that are important for microarray fabrication: microarray density, reproducibility, storage time, contamination and printing quality.

1. Microarray Density

Microarray density is one of the most important parameters for microarray fabrication. The number of DNA elements that can be deposited on a slide will depend on spot size and pin configuration.

Spot Size. Microarray density is determined by the size of the DNA spot and depends directly on the volume of sample deposited on the substrate surface. The volume of sample deposited per position on the array generally ranges from 50 to 500 pl, with high and low extremes of 10 pl and 10 nl possible, respectively (Mace *et al*., 2000). Several factors affect the volume of sample that can be applied to an array surface, including the surface properties (e.g., surface energy) of the slides and pins, and the sample solution characteristics (e.g., viscosity). For printing high-density microarrays, a hydrophobic glass surface (e.g., aldehyde-modified slide) is preferred, because spotted hydrophilic samples will spread less on a hydrophobic surface than on a hydrophilic one. Because the pin contact surface area determines the initial contact between the sample and slide, the spot size increases as the pin contact surface area increases. In addition, pin velocity has a great effect on the spot size. The loading sample volume for a split pin (e.g., ChipMaker and Stealth pins from TeleChem) typically ranges from 0.2 to 0.6 μl. Thus, if the pins tap the surface at high speed (>20 mm/s), a large sample volume may be forced out of the pin and large spots will be produced (Rose, 2000).

Pin Configuration. Printing pins are mounted in a print head, which can hold up to 64 pins. The distance between pins on the print head is 4.5 mm and precisely matches the well spacing of a 384-well microtitre plate. DNA samples are first taken from 96-well or 384-well source plates by dipping the pins into the sample wells with either a single pin or multiple pins, and then depositing the sample on the slide surface by gently touching the pins to the surface. Fabrication of arrays using a single pin is the most straightforward approach, but it is also time-consuming. The main advantage of single pin printing is that the DNA samples in the source plates can be positioned on the array in the same order as they occur in the source plate, thus making sample tracking and post-hybridization analysis easier. Another advantage is that pin-to-pin variations are not a problem when using a single pin for microarray printing. Using a single pin and 250-μm spot-to-spot spacing, for example, more than 20,000 spots can be deposited on 22×72 mm^2 printing area.

Multiple pins are generally used for printing high-density microarrays because of the increased printing speed, even though different pins can cause variations in array quality. To print with multiple pins, the pins are dipped into sample wells of a 384-well plate and then touched to the slide surface simultaneously to create separate spots at a 4.5-mm spacing in the first round. Later rounds of printing are achieved by spotting with a predefined spot-to-spot offset distance from the previous location. Each pin deposits samples in a sub-grid. Since some areas within each sub-grid might not be completely filled with spots due to the restriction of the layout, the density of microarrays will generally decrease as the number of pins used increases. Printing microarrays with multiple pins requires

more time in designing the array layout, as well as sophisticated sample tracking and deconvolution in the data analysis phase.

2. Reproducibility

Fabricating microarrays of reproducible quality is of the utmost importance in microarray-based experimentation. For reliable and reproducible data, the uniformity of individual spots across the entire array is very important for simplifying image analysis and enhancing the accuracy of signal detection. Several factors can affect the uniformity of spots, including array substrate, pins, printing buffer, and environmental controls. As mentioned previously, non-homogeneous surface treatment will cause variations in the amount of attached DNA.

Variations in array quality can be caused by differences in pin geometry, pin age and sample solutions. Movement of the pin across the surface in the XY direction may cause the tip to bend (Rose, 2000). Tapping the pins on the surface may result in deformation of the pin tips. In addition, dragging the pin tip across the surface may cause clogging of the pin sample channel. Therefore, great care is needed in handling pins, even though they are robust. Pins should be cleaned with an ultrasonic bath after each printing (Rose, 2000).

Environmental conditions have significant effects on spot uniformity and size (Hegde *et al.*, 2000). Humidity control is absolutely necessary for preventing sample evaporation from source plates and the pin channel during the printing process. Sample evaporation can cause changes in DNA concentration and viscosity. Reducing the extent of evaporation can help the small spotted volume of DNA have more time to bind at equal rates across the entire spot. As a result, DNA spots of high homogeneity will be obtained (Diehl *et al.*, 2001). Generally, the relative humidity is controlled between 65 and 75% (Rose, 2000). Condensation could occur if the relative humidity is greater than 75%.

Producing homogeneous spots on arrays also depends on the printing or deposition buffer. Saline sodium citrate is commonly used as a printing buffer in microarray construction; however, spot homogeneity and binding efficiency with this buffer can be poor. The addition of 1.5 M betaine to the printing buffer can significantly improve spot homogeneity and binding efficiency (Diehl *et al.*, 2001), because betaine increases the viscosity of a solution and reduces the rate of evaporation. More uniform spots can also be achieved with a deposition buffer that contains 50% dimethyl sulfoxide (DMSO) (Hegde *et al.*, 2000; Wu *et al.*, 2001).

3. Storage Time

Another important practical issue concerns the shelf life of unused microarrays. The maximum time that microarrays can be stored prior to use is

currently unknown. The shelf time could depend on the coating chemistry of the slide and the storage conditions. Unprocessed microarrays can be stored in a dessicator for many months without deterioration of performance (Worley *et al.*, 2000).

4. Contamination

To produce high-quality microarrays, the collection of airborne dust and impurities on the slide surface must be eliminated or at least minimized during array fabrication. Dust and particulate matter can settle on the slide surface and cause printing inaccuracies as well as poor quality scanned image displays. Enclosing the array device in a humidity chamber can minimize dust contamination.

Because pins are generally reused for depositing different biological samples, sample carryover during the printing process is a practical concern and can complicate interpretation of hybridization results. Efficient cleaning of the pins is therefore required for the printing process. Generally, the pins are cleaned by dipping them into distilled water or detergent and then using a vacuum to remove the wash solution from the pin channel. Repeating this process three times is generally sufficient to eliminate sample carryover problems. Cross-contamination during sample preparation and handling is another important concern in the microarray printing process. For making microarrays, plasmids containing the desired cDNA clones are generally extracted from bacterial cultures and the desired genes are amplified from the plasmid DNA. Recent studies showed that up to 30% of clones contained the wrong cDNA (Knight, 2001). This is most likely due to bacterial contamination and handling errors during sample preparation. Therefore, great care must be taken to eliminate or minimize such errors. Errors in public sequence databases are also possible and can lead to failures in microarray-based detection. For instance, some mouse sequences in the public databases correspond to the wrong strand of the DNA double helix. As a result, the designed oligonucleotide probes were not able to detect their target mRNAs (Knight, 2001).

5. Evaluation of Printing Quality

After printing, it is important to assess the quality of the arrayed slides prior to hybridization in terms of surface quality, integrity and homogeneity of each DNA spot, the amount of deposited DNA, and consistency among replicated spots. Staining prior to hybridization will identify any problems introduced during the fabrication process. Microarrays can be stained with various fluorescent dyes,

such as PicoGreen, SYBR Green II and Topo Green (Battaglia *et al.*, 2000), followed by fluorescent scanning. Such information is critical in assessing the quality of the printed arrays and the data output from those arrays.

IV. MICROARRAY HYBRIDIZATION AND DETECTION

A. Probe Design and Synthesis

Probe design and synthesis are critical steps in generating high-quality microarrays for gene expression analysis. Three types of probes are used for microarray fabrication: PCR products, cDNA clones, and oligonucleotides. For the construction of cDNA microarrays, individual open reading frames (ORFs) can be amplified using gene-specific primers. Because cross-hybridization among homologous genes is a potential problem, full-length genes cannot be used for microarray construction. Several computer programs are available that can identify DNA fragments specific ($<75\%$ sequence identity) to each ORF by comparing the target gene with all other genes in a genome (Xu *et al.*, 2002). Once the specific fragments are identified, more than one set of primers can be obtained based on the identified unique fragments using the PCR primer design program Primer 3 (Whitehead Institute). The designed primers are generally synthesized commercially.

Optimal forward and reverse primers are generally selected based on the following considerations. First, for genes shorter than 1000 bp, the PCR-amplified unique fragments should be as long as possible. For genes longer than 1000 bp, the optimal amplified fragments should be within 500–1200 bp. Second, each oligonucleotide primer should be 20–28 bp in length and the set of primer pairs (typically stored in 96-well plates) should have an annealing temperature of approximately 65 °C to simplify PCR amplification. If the desired target annealing temperature cannot be obtained, a lower annealing temperature can be used. In the case where specific fragments cannot be identified for some homologous genes, fragments with higher than 75% sequence identity will be selected and appropriate primers can be designed based on these fragments. However, hybridization signals for these genes should be carefully interpreted during microarray data analysis. One of the great practical problems with probe amplification is that PCR product yields vary considerably among different genes and some primers may fail to yield PCR products. This may cause significant variation in the DNA concentration present on the slide surface.

The cDNA clone-based probes are generally derived from whole genes or fragments of genes that are amplified from clone libraries using vector-specific primers. The size of clone probes generally ranges from a few hundred to a few

thousand base-pairs. Generally, since vector-specific primers are used for amplifying the cloned inserts, clone-based probes cannot be specifically designed for regions of low homology to other genes. As mentioned above, a substantial portion of clones may contain the wrong cDNA due to bacterial contamination and mishandling (Knight, 2001).

Oligonucleotide probes are different from other probes in that they can be deposited by printing or synthesized *in situ* on a solid surface. Specific oligonucleotide probes can be designed based on gene sequences. Generally, the sizes of the oligonucleotide probes are shorter than 25 bp, and several different oligonucleotide probes are used per gene for high-density oligonucleotide microarrays. To discriminate mispriming, a probe is designed deliberately to have a single mismatch at the central position (Lockhart *et al.*, 1996; Warrington *et al.*, 2000). Recently, the utility and performance of oligonucleotide microarrays containing 50- to 70-mer oligonucleotide probes were evaluated (Kane *et al.*, 2000). The results indicate that such oligonucleotide microarrays can be used as a specific and sensitive tool for monitoring gene expression.

B. Target Labeling and Quality

Target labeling is another critical step in successful microarray-based experimentation. The methods available for labeling nucleic acids for microarray hybridization can be classified into two categories: direct and indirect labeling.

1. Direct Labeling

In direct labeling, fluorescent tags are directly incorporated into the nucleic acid target mixture before hybridization by enzymatic synthesis in the presence of either labeled nucleotides (e.g., Cy3- or Cy5-dCTP) or PCR primers (Fig. 4A). The most commonly used method is to label the target mRNA or total cellular RNA using reverse transcriptase. In a first-strand reverse transcription reaction, fluorescently labeled cDNA copies of RNA are synthesized by incorporation of a fluorescein-labeled nucleotide analog. Random hexamers, oligo(dT) or gene-specific primers can serve as primers for the initiation of reverse transcription. Since prokaryotic mRNA has no poly-(A) tail, random hexamers are generally used for reverse transcription. In this case, total cellular RNA is used as the template for cDNA synthesis, and hence a greater degree of background fluorescence intensity can occur. Although gene-specific primers can reduce such background levels by copying gene-specific

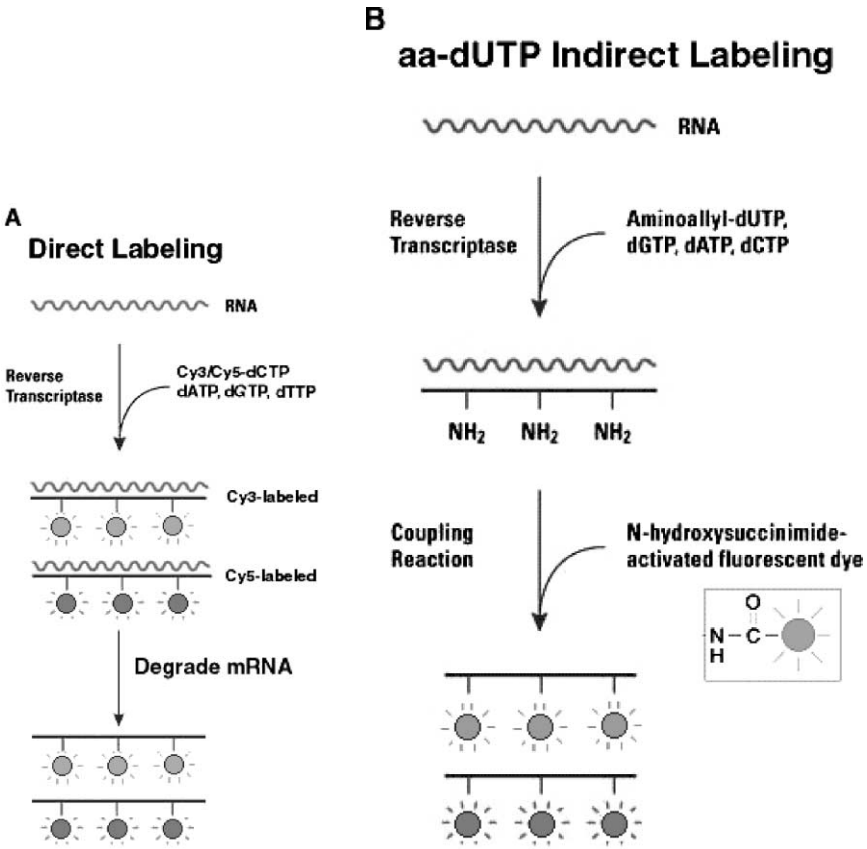


Figure 4 Labeling strategies. (A) Direct incorporation of fluorescent dyes into target sample through reverse transcription. (B) Incorporation of fluorescent dyes into target samples through reverse transcription in the presence of amino-allyl-dUTP, followed by chemical coupling with fluorescent dyes. (C) Dendrimer-based indirect labeling. (Courtesy of Molecular Probes).

fragments, it requires reverse transcription with hundreds or thousands of primers.

A variation of the direct labeling approach is that mRNA is amplified up to 1,000–10,000-fold by T7 polymerase to obtain antisense mRNA (aRNA), and then the aRNA is reverse transcribed to obtain labeled cDNA (Salunga *et al.*, 1999). One of the advantages of the T7 polymerase-based amplification method over other amplification methods is that all mRNAs are almost equally amplified, because amplification with T7 polymerase is a linear process. Another advantage is that mRNA can be labeled easily with reverse transcriptase, which incorporates fluorescent tags much more readily than DNA polymerase.

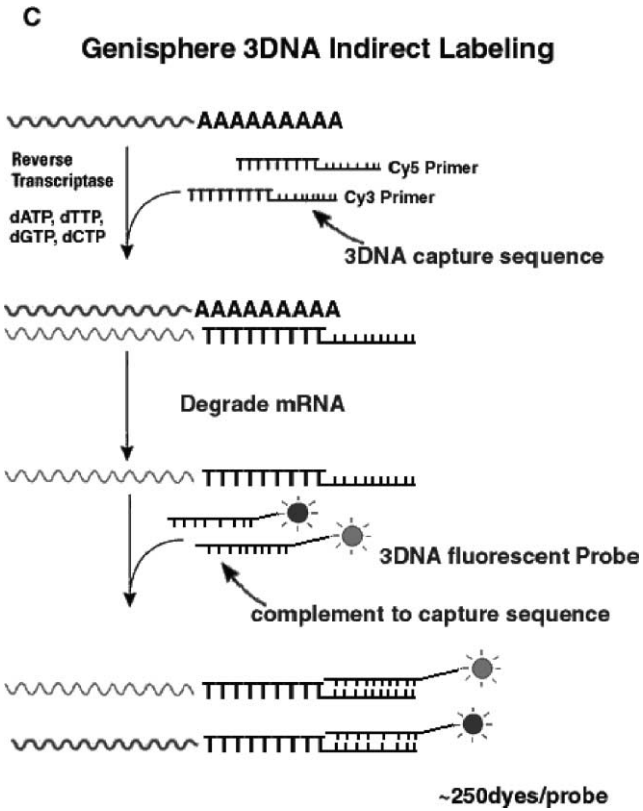


Figure 4 (*continued*)

One of the problems with the reverse transcriptase-based labeling approach is that nucleotides tagged with structurally different fluorescent dyes are differentially and non-uniformly incorporated into cDNA. To resolve this problem, a two-step approach was proposed. The FairPlay™ system developed by Stratagene Corporation uses a two-step chemical coupling method to fluorescently label cDNA. First, an amino allyl-dNTP is uniformly and efficiently incorporated into cDNA by reverse transcriptase (Fig. 4B), because the amino allyl-dNTP does not exhibit steric hindrance. Then, an amine-reactive cyanine is chemically coupled to the amino-modified cDNA. The main advantage of this approach is that this system efficiently produces uniformly labeled cDNA without any dye bias. As a result, this system is highly sensitive (5-fold increase in sensitivity), requires less RNA, and allows detection of low abundance genes. Any labeling bias resulting from fluorescent dye incorporation also appears to be negligible and thus the dual labeling experimental approach is not needed.

2. Indirect Labeling

In the indirect labeling approach, fluorescence is introduced into the detection procedure following hybridization. Briefly, epitopes are incorporated into the target samples during cDNA synthesis. After hybridization with the epitope-tagged target samples, the microarray is incubated with a fluorescently tagged protein that binds to the epitopes. The most common indirect labeling method uses a biotin epitope and a fluorescent streptavidin–phycoerythrin conjugate (Warrington *et al.*, 2000). The biotinylated nucleotides are incorporated into cDNA by reverse transcription and hybridized with the microarrays. After hybridization, the array is stained with a streptavidin–phycoerythrin conjugate, which binds to biotin tags and emits fluorescent light when excited with a laser.

Another indirect labeling approach is known as Tyramide Signal Amplification (TSA) (Adler *et al.*, 2000). This approach uses biotin and dinitrophenol (DNP) epitopes as well as streptavidin and antibody conjugates linked to horseradish peroxidase (HRP). In the presence of hydrogen peroxide, HRP catalyzes the deposition of Cy3- or Cy5-tyramide compounds on the microarray surface. By this method, a DNP- or biotin-dCTP analog is first incorporated into cDNA, and then the epitope-tagged cDNA is hybridized with the microarray. Following hybridization, the microarray is incubated with anti-DNP-HRP, and Cy3-tyramide is deposited on the microarray surface, followed by incubation with streptavidin–HRP and deposition of Cy5-tyramide (Adler *et al.*, 2000). The main advantage of this approach is that it can provide10- to 100-fold signal amplification over the direct labeling approach. Thus, this approach can be used effectively to monitor the expression or abundance level of rare transcripts or to analyze samples prepared from small numbers of cells. The main disadvantage of this method is that it is generally less precise for comparative analysis due to variations arising from differences in labeling efficiencies and protein-binding affinities (Schena and Davis, 2000). In addition, the signal intensity is only semi-quantitative because of the involvement of enzymatic signal amplification (Alder *et al.*, 2000).

The third indirect approach is to use DNA dendrimer technology (Stears *et al.*, 2000) (Fig. 4C). Dendrimers are stable, spherical complexes of partially double-stranded oligonucleotides with a determined number of free ends, which are tagged with fluorescent dyes, Cy3 or Cy5. In this technology, the cDNA is first synthesized by reverse transcriptase with primers containing specific capture sequences that can bind the Cy3- or Cy5-tagged dendrimers through sequence complementarity. The synthesized cDNAs are then hybridized to microarrays, and the bound cDNAs on the microarrays are detected by incubating the arrays with the fluorescent dye-tagged dendrimers. The dendrimer detection approach is highly sensitive, requiring up to 16-fold less RNA for probe synthesis. Since the fluorescent dye is attached to the free end of the dendrimers, signal intensity is

independent of probe size and composition. In addition, this detection system has a high signal-to-background ratio and can be used for multiple channel detection on a single microarray.

C. Hybridization

After microarray fabrication, the most important issue in microarray-based analysis is probe–target hybridization. Conceptually, microarray hybridization and detection are quite similar to the traditional membrane-based hybridization (Eisen and Brown, 1999). Before hybridization, the free functional groups (e.g., amine) on the slide should be blocked or inactivated to eliminate non-specific binding, which causes high background and depletion of probes. Any unbound DNA on the slides can be washed away during the pre-hybridization process. Removal of unbound DNA in pre-hybridization is important, because any DNA that washes from the surface during hybridization competes with DNA bound to the slide. Since the rate of hybridization in solution is much faster than that on surfaces, the presence of unbound probe DNA can lead to a dramatic decrease in the measured signals obtained from microarrays.

After pre-hybridization, the microarray is hybridized with fluorescently labeled target DNA or RNA for a certain period of time. Post-hybridization washing then removes unbound labeled material. Regardless of the hybridization format, the hybridization solution should be mixed well so that the labeled targets are evenly distributed across the array surface to obtain the maximum number of optimal target–probe interactions. In addition, the wash solutions should be uniformly distributed to eliminate unbound probes, remove non-specific hybridization, and minimize background signal.

D. Detection

Both the confocal scanning microscope and coupled charge device (CCD) camera have been successfully used for the detection of microarray hybridization signals, and many such devices are commercially available (Hegde *et al.*, 2000). Although the confocal scanning microscope and CCD camera systems both have advantages and disadvantages (as described below), the former is more commonly used.

Generally, a confocal scanner uses laser excitation of a small region of the glass slide (~100 μm^2), and the entire array image is acquired by moving the glass slide, the confocal lens, or both across the slide in two directions (Schermer, 1999). The fluorescence emitted from the hybridized target molecule is gathered with an objective lens and converted to an electrical signal with a photomultiplier (PMT) or an equivalent detector. The main drawbacks of using

a confocal scanning microscope for signal detection is that each excitation wavelength must have its own laser, which can be expensive, and the device is very sensitive to any non-uniformity of the glass slide surface.

The CCD camera exploits many of the same principles as a confocal scanner, but the CCD camera utilizes substantially different excitation and detection technologies (Schermer, 1999). CCD systems typically use broad-band xenon bulb technology and spectral filtration (Basarsky *et al.*, 2000). The key advantage of the CCD camera-based imaging systems is that they allow simultaneous acquisition of relatively large images of a slide (1 cm^2) and hence do not require moving stages and optics, which reduces cost and simplifies instrument design. However, since the CCD camera does not move the optics or stages, several images need to be captured from different fields of the microarray and then stitched together to represent the entire information on the slide. Because most commonly used fluoresceins have a small difference between excitation and emission maxima, it is difficult to effectively separate excitation and emission light in the spectral filtration process.

E. Critical Issues in Hybridization and Detection

This section highlights some important practical issues related to microarray hybridization and detection, namely, probe retention and quantitative hybridization, target labeling and availability, spatial resolution and cross-talk, and photobleaching and scanning parameters.

1. Probe DNA Retention and Quantitative Hybridization

In solution-based hybridization, signal intensity depends on both target and probe DNA concentrations. In gene expression profiling studies, it is assumed that the concentrations of all probe DNAs deposited on the microarrays are much higher than the mRNA concentrations in the fluorescently labeled target samples, so that signal intensity is dependent exclusively on the mRNA concentration in the target samples. Therefore, many factors causing probe deposition variations will have negligible effects on hybridization signal intensity.

For the accurate quantitation of gene expression, it is essential to ensure that the arrayed DNA probes are in excess relative to the labeled target cDNAs. Generally, a DNA concentration of 100–200 ng/μl is used for spotting, which corresponds to 100–200 pg/spot for a 1-nl deposition. The retention is about 20–30% on silanized glass surfaces (Worley *et al.*, 2000). Thus, after boiling and hybridization, this corresponds to approximately 20–60 pg of double-stranded DNA present in each spot for binding. Studies indicate that the arrayed

DNA appears to be in excess for all the protein-coding genes in *Escherichia coli* (Worley *et al.*, 2000). However, probe DNA retention depends on slide surface, coating chemistry, post-processing, hybridization, and washing conditions. Therefore, to ensure accurate quantitative results for highly expressed genes, it is important to understand how much spotted DNA can actually be retained after hybridization.

Whether the probe DNA concentration represented on the array substrate is in excess obviously depends on the amount of target sample used. Typically, 10–20 μg of total cellular RNA is used for monitoring gene expression in prokaryotes. However, for monitoring rare transcripts, higher RNA concentrations (e.g., 50 μg) are generally used. In this case, probes corresponding to abundant transcripts may not be in excess relative to the target samples, resulting in hybridization that is not quantitative. Hence, it is important to select the appropriate amount of RNA to ensure that the microarray signal is within the range of linear response for the system being used.

2. Target Labeling and Availability

The integrity and purity of the RNA are crucial for obtaining high-quality microarray hybridization results. Impurities in RNA preparations could have an adverse effect on both labeling efficiency and the stability of the fluorescent dyes. Thus, the RNA must be free of contaminants such as polysaccharides, proteins and DNA. Many commercial RNA purification kits are available for producing RNA of sufficient purity for microarray studies. In addition, unincorporated nucleotides present in the labeling reaction must be removed to reduce background noise. Finally, both Cy3 and Cy5 are sensitive to light, and thus great caution must be taken to minimize exposure to light during labeling, hybridization, washing and scanning.

The most frequently encountered experimental problem is the variation in hybridization signal between labeling reactions. In many cases, poor hybridization signals result from poor dye incorporation. Very low dye incorporation (<1 dye molecule/100 nucleotides) gives unacceptably low hybridization signal intensities. However, studies showed that very high dye incorporation (e.g., >1 dye molecules/20 nucleotides) is also not desirable, because high dye incorporation significantly destabilizes the hybridization duplex (Worley *et al.*, 2000). Thus, it is important to measure the dye incorporation efficiency prior to hybridization. The specific activity of dye incorporation can be determined by measuring the absorbance at wavelengths of 260 and 550 nm for Cy3 or 650 nm for Cy5. A suitable labeling reaction should have 8–15 A_{260}/A_{550} ratio for Cy3 and 10–20 A_{260}/A_{650} for Cy5.

Another problem encountered routinely is the quality of fluorescent dyes. The labeling efficiency and hybridization vary significantly sometimes from batch to batch, especially for Cy5. Fresh reagents are very important in achieving a high degree of detection sensitivity (Wu *et al.*, 2001).

Microarray hybridization is generally performed in the absence of mixing. Since the diffusion coefficient is very small for large labeled target DNA molecules, the probe at each arrayed spot is in effect hybridizing with its labeled counterpart from its immediate or nearly immediate local environment (Worley *et al.*, 2000). Thus, the target solution should be mixed well and uniformly distributed over the microarray surface area. Otherwise, the availability of the labeled target molecules to the arrayed spots could be significantly different across the microarray surface. As a result, significant differences in signal intensity can be observed.

3. Spatial Resolution and Cross-talk

The spatial resolution of microarray detection systems is usually expressed as a pixel size, the physical "bin" in which a single datum is acquired. The spatial resolution for many commercial systems usually ranges from 5 to 20 μm. The selection of spatial resolution depends on spot size, and in general, the pixel dimension should be less than 1/10 of the diameter of the smallest spot on the array. For example, microarrays containing 100-μm spots require fluorescent detectors with a spatial resolution of 10 μm pixel size.

Cross-talk refers to the phenomenon in which an emission signal from one channel is detected in another channel, resulting in an elevated, erroneous fluorescence reading. Cross-talk is most likely from the shorter wavelength channel into the longer wavelength channel. For example, the fluorescence intensity from the Cy3 channel can contaminate the Cy5 channel but not vice versa. Cross-talk is the most common potential problem in the simultaneous scanning approach, which acquires both images from two channels at the same time (Basarsky *et al.*, 2000). For gene expression experiments, cross-talk should be kept to less than 0.1%. The most common and cost-effective way to minimize cross-talk is to use emission filters that reject light outside the desired wavelengths. Cross-talk can also be greatly minimized by selecting fluorescent dyes and lasers with sufficient differences in wavelength (Schermer, 1999).

4. Photobleaching and Scanning Parameters

Light is emitted from a fluorescent dye when it is illuminated by a light source. Generally, the emitted fluorescence is directly proportional to the power of

the excitation light. Therefore, to increase detection sensitivity, higher power of excitation light is preferred. If the excitation light is excessive, however, the incoming photons can damage the dyes and reduce the fluorescent signals during successive scans, leading to photobleaching of the signal intensity. Photobleaching depends on the duration of sample illumination. More powerful light sources and/or a longer laser exposure time can result in significant photobleaching. When acquiring an array image, it is best to keep photobleaching to less than 1% per scan.

Different dyes have considerable differences in their photostabilities. For example, Cy5 is more sensitive to photobleaching than Cy3. The differences in photostability among different dyes could be a significant problem when multiple dyes are used in an experiment, because the ratios measured can lead to significant quantitative errors. To minimize photobleaching, the Cy5 channel is always scanned first, followed by the Cy3 channel.

Although Cy3 (0.15, no unit) has a lower quantum yield than Cy5 (0.28), Cy3 is more efficiently incorporated into cDNA during reverse transcription. Such dye characteristics can cause variations in the signal intensity obtained in reverse labeling experiments. The signal should be balanced during scanning by using a higher PMT setting for the dye with the weaker signal to allow detection of more spots of low signal intensity.

V. MICROARRAY IMAGE PROCESSING

A. Data Acquisition

The fundamental aim of image processing is to measure the signal intensity of arrayed spots and then quantify gene expression levels based on the signal intensities acquired for each spot. Therefore, spots on the array image must be correctly identified.

The spots on microarrays are arranged in grids. An ideal microarray image for easy spot detection should have the following properties: (i) the location of spots should be centered on the intersections between the row and column lines; (ii) the spot size and shape should be circular and homogeneous; (iii) the location of the grids on the images should be fixed; (iv) the slides should have no dust or other contaminants; and (v) the background intensity should be very low and uniform across the entire image. In practice however, it is difficult to obtain such ideal images. First, the spot position variation occurs because of mechanical limitations in the arraying process, including inaccuracies in robotic systems, the printing apparatus and the platform for holding slides. Second, the shape and size of the spots may fluctuate considerably across the array because of variations in the size of the droplets of DNA solution, DNA and salt concentration in

the printing solution, and slide surface properties. In addition, contamination from airborne dust and impurities on the slide surface is a major problem for processing array images. To obtain accurate measurements of hybridization signals, all of these potential problems should be taken into consideration.

Many methods are available for resolving spot location errors, spot size and shape irregularities and contamination problems (Zhou *et al.*, 2000) in order to accurately estimate spot intensities. Commercial and free software, including ImaGene™ from BioDiscovery (Los Angeles, CA), QuantArray™ from GSI Lumonics, and the software on Axon GenePix™ systems (Bassett Jr. *et al.*, 1999), can be used for microarray image processing. Typically, a user-defined gridding pattern is overlaid on the image, and the areas defined by patterns of circles are used for spot intensity quantification.

The data are extracted and generally expressed as the total (the sum of the intensity values of all pixels in the signal region), mean (the average intensity of pixels), and median (the signal intensity of the median pixel). Microarray output corresponding to the total intensity is not the best measurement of hybridization signal, because it is particularly sensitive to variations in the amount of DNA deposited on the surface and the presence of contamination (Zhou *et al.*, 2000). The mean is probably the best measurement when using an advanced image processing system that permits accurate segmentation of contaminated pixels, because the mean measurement reduces variations caused by differences in the amount of DNA deposited within a spot. However, the mean measurement is vulnerable to outliers (Petrov *et al.*, 2002). The median is a better choice than the mean if the image processing software is not good enough for correctly identifying signal, background and contaminated pixels. The median intensity value is very stable and is close to the mean if the distribution profile of pixels is *uni-modal*. The median is equal to the mean when the distribution is symmetric. An alternative to the median measurement is to use a trimmed mean (the mean of the pixel intensity after a certain percentage of the pixels are removed from both tails of the distribution).

Some comparative studies indicate that the choice of measurements depends on the segmentation techniques used. The mean is the best measurement if the combined and trimmed segmentation techniques are used, whereas the median is the best without trimming (Petrov *et al.*, 2002).

B. Assessment of Spot Quality and Background Subtraction

For some spots, signal intensity data may not be reliable because of the inherently high variation associated with array fabrication, hybridization, and image processing. Thus, the first step in data processing is to assess the quality of spots, with the removal or filtering of unreliable poor spots or outlying spots

(outliers) prior to data analysis (Heyer *et al.*, 1999; Tseng *et al.*, 2001). It is critical to identify problematical slides, because without assessing the quality of the spot signals, conclusions drawn from the analysis of such data could be misleading.

1. Identification of Poor Slides

Due to the multiple steps involved in microarray experiments, it is important to evaluate array slides as a whole and to eliminate unreliable hybridization signals prior to rigorous data analysis. Two measures can be used to assess the overall slide quality if replicate spots are present on the arrays (Worley *et al.*, 2000): one can calculate (i) the average coefficient of variation (CV) of replicates in the spot pairs and (ii) the r^2 value of the regression line from a scatter plot of duplicate spots. Although there is no general consensus on the appropriate threshold value for rejecting slides, slides are generally accepted if the average CV is less than 50%. If there are no replicate spots on the microarrays, slide quality can be assessed by determining the number of spots that are of poor quality. Generally, microarray experiments should be repeated if more than 30% of the spots on the microarray are flagged as poor spots.

2. Identification of Poor Quality Spots

There are no rigorously defined rules for identifying poor spots from a biological or statistical perspective. The spot quality and integrity are generally assessed based on the following criteria:

Spot size and shape. Spots with excessively large or small diameters compared to the majority of spots should be discarded. Discarding such low-quality spots significantly improves the reliability of the data (Zhou *et al.*, 2000).

Spot homogeneity. The distribution of pixels within the spots can be used to assess spot homogeneity. Generally, spots with less than a certain percentage (e.g., 55–60%) of all pixels having intensities greater than the average background intensities (Khodursky *et al.*, 2000) or one standard deviation (SD) above local background are flagged as poor quality spots (Murray *et al.*, 2001).

Spot intensity. Spots with signals not significantly above background should be identified using various standards. For example, spots with median or mean signals less than one to three SDs above background in both channels (Chen *et al.*, 1997; Basarsky *et al.*, 2000; Hegde *et al.*, 2000) are flagged as poor quality spots.

In addition, spots whose signal is not at least 2.5 times higher than the background signal in both channels are excluded (Evertsz *et al.*, 2000).

Another way to define poor spots is based on signal-to-noise ratio (SNR), which is often defined as the ratio of the difference between signal and background divided by SD of background intensity (Verdnik *et al.*, 2002). This ratio indicates how well one can resolve a true signal from the instrumental noises. A commonly used criterion for the minimum signal that can be accurately determined is an SNR value equal to 3. Below that value, the signal cannot be accurately quantified, and such spots are treated as poor spots.

The commercially available software, ImaGene from Biodiscovery, is able to automatically flag poor spots. Spots identified as poor quality are not included in the data analysis. Although the criteria for defining poor spots are based on subjective thresholds rather than statistically robust tests, they take into account the major factors affecting the quality of data and are likely to be very effective in reducing the amount of noise.

3. Removal of Outlying Spots

Outliers represent extreme values in a distribution of replicates. Outlying spots can be caused by uncorrected image artifacts such as dust or by factors undetectable by image analysis such as cross-hybridization. Outliers significantly affect the estimation of expression values and its associated random errors. Thus, removal of outlying spots is an important step in data filtering. However, distinguishing outliers from differentially expressed genes is very challenging, because there is no general definition describing outliers. In this section, we briefly describe several commonly used methods for identifying outliers.

Simple threshold cutoff. A gene whose CV is greater than a certain threshold (Murray *et al.*, 2001), e.g., 30–50%, is excluded from the data analyses.

Intensity-dependent threshold cutoff by windowing procedure. (Tseng *et al.*, 2001). The CV values for individual genes are plotted against the average signal intensity of the two channels [(Cy3 + Cy5)/2]. For each gene, a windowing subset is constructed by selecting a certain number of genes (e.g., 50) whose mean intensities are closest to this gene. If the CV of this gene is within a top certain percentage (e.g., 10%) among genes in its windowing subset, then data on this gene are regarded as unreliable, and hence all replicate data for this gene are discarded. However, in many cases, not all replicate spot data for this gene are unreliable. To salvage some information for this gene, the most outlying spots can be eliminated, and the CV of the intensity ratios of the remaining spots corresponding to this gene can be recalculated. If the CV is significantly reduced

below the threshold level, the data for the remaining spots can be used in subsequent analyses. The CV can also be used for assessing the quality of different slides and different experiments (Tseng *et al*., 2001).

Removal of outliers with jackknife procedure. The jackknife correlation can be used to remove outliers for expression data obtained from time-series microarray experiments (Heyer *et al.,* 1999). In this statistical approach, the correlation coefficient is calculated for each pair of genes using all of the time-series data points. Then the data at one time point are deleted, and the correlation coefficient is recalculated respectively for each pair of genes with all of the time-series data points but one. The jackknife correlation is the minimum correlation coefficients obtained above and can then be used for further cluster analysis. Jackknife correlation is robust and insensitive to single outliers. Applying jackknife correlation reduces false positives, while capturing the shape of an expression pattern. Hyer *et al.* (1999) showed that the genes showing similar expression patterns generally had a jackknife correlation of 0.7 or higher.

Identification of outliers based on pooled error methods. Several methods are used for statistical detection of outliers, but they are generally less adequate for typical microarray studies due to the small number of replicates (Nadon *et al.*, 2001). The random error estimation for each gene based on a small number of replicates is imprecise, which makes statistical tests insensitive. As a result, many replicate spots may be falsely identified as outliers or many true outliers may not be identified (Nadon and Shoemaker, 2002). The potential violation of the normality assumption makes inferences of outliers and gene differential expression less reliable (Nadon *et al.*, 2001).

The pooled error method assumes that all probes or probes of similar intensities within a specific study have the same true random error. Variance estimates therefore can be pooled together across many genes and the precision of error estimation can be greatly improved. Furthermore, it is assumed that the standardized residuals have a normal distribution if the pooled error model is correct. Under these assumptions, the existence of outliers will cause the distribution of the entire data set to deviate from normal. Removal of spots with large residuals will improve the normality of the entire data set. Generally, outliers are identified in an iterative fashion: spots with large absolute residuals are removed from the data set; data are examined for normality and the residuals are calculated again. The process is iterated until the index asymptotes approach a stable value, which indicates that further removal of data values would not improve the normality of the distribution of the remaining data set. Software is available (ArrayStat™) to facilitate array-based statistical analysis (Nadon *et al*., 2001). In this software package, outliers are automatically detected. The pooled error method is a better, more sensitive method for outlier detection and can be used appropriately for microarray experiments having as few as two replicates.

4. Background Subtraction

Subtraction of background fluorescence from hybridization signals is the second step in microarray data processing. Background subtraction is necessary to distinguish actual signals based on hybridization from noise and allows the comparison of specific spots. There are two approaches to background subtraction. The first approach is to take signal intensity levels from blank areas on the array and use this for subtraction. The problem with this approach is that the background varies across the array and thus the background noise among spots can be significantly different. The second approach is to use a local background for the area around each spot for background determination. Local sampling of the background is generally used to specify a threshold that the true signal must exceed. By doing this, it is possible to detect weak signals and extract an average density above the background for each array element (Chee *et al.*, 1996). After removing poor slides, poor quality spots, outliers and background, the microarray data are ready for further normalization and data analysis.

VI. MICROARRAY DATA ANALYSIS

A. Data Normalization

Microarray hybridization possesses intrinsic variation, which can potentially occur at every step in the microarray process. One key question prior to applying statistical analyses is whether such variations represent true random variations in expression values or are due to systematic variations arising from differences in the experimental conditions. Before pursuing further statistical analyses of microarray data, the systematic variations must be removed by normalization to allow statistical comparisons among different slides and different experiments.

1. Sources of Systematic Variations

Systematic variation stems from a number of sources during microarray experiments. The major anticipated sources of variations include the following (Tseng *et al*., 2001).

Variations within a Slide or Spatial Effect. Many studies show that substantial signal variation occurs for the same gene within a slide (Dudoit *et al.*, 2001). Differences in pin geometry, slide homogeneity, hybridization and target fixation could all contribute to variations observed among repeated spots within a slide.

Some systematic differences may occur between different pins due to differences in the length or in the opening of the tips, pin deformation following multiple rounds of printing, and slight differences in surface properties. All of these can lead to differences in target DNA transfer and hence may cause systematic variations in microarray signal intensity. The amount of deposited target DNA also fluctuates for the same pins, while studies showed that the variation among different pins was significantly higher (Dudoit *et al*., 2001). In addition, the fraction of target DNA that is chemically fixed onto the slides is unknown and could vary considerably within slides. For various reasons, the labeled targets may also be distributed unevenly over the slide and/or the hybridization reaction may occur unequally in different parts of the slides. Finally, some areas of a slide may be contaminated and have a high background. The influence of these factors on signal intensity measurement within a slide is generally referred to as *spatial effect*.

Variation among slides or slide effect. Differences in surface properties, microarray fabrication, hybridization and imaging could lead to systematic variations in hybridization signals among different slides. The amount of probe DNA immobilized on the slide during array printing and probe fixation can be substantially different among different slides due to various factors such as differences in slide surface properties and sample evaporation during printing. Also, the amount of cDNA added to the slides, especially when different RNA preparations are used, and the local environment and hybridization conditions, such as temperature, buffer pH, target concentration, incubation and washing time in each hybridization chamber, could be considerably different. Background noise and the local curvature of the surface among different slides may have a large impact on scanning, especially for confocal scanners which are sensitive to focus. The influence of these factors on measuring signal intensity is defined as *slide effect*. Tseng *et al*. (2001) showed that such effects are significant, and normalization is slide-dependent.

Variation from probe labeling or label effect. The most commonly used fluorescent dyes, Cy3 and Cy5, are not equally incorporated into DNA molecules by reverse transcriptase and DNA polymerase. Cy3 is incorporated more efficiently than Cy5 with the same preparation and amount of RNA. While both Cy dyes are relatively unstable, Cy3 and Cy5 have different quantum efficiencies and are detected by the array scanner with different efficiencies. While the detection limit of Cy5 with the scanner is lower than that of Cy3, Cy5 is more sensitive to photobleaching. The influence of these factors on intensity measurements is referred to as *label effect*.

The use of two fluorescent dye labels may also introduce gene-label interactions. For instance, fluorescent labeling may fluctuate systematically, depending on the nucleotide composition of the target sequences, and one or the

other dye may be preferentially incorporated into specific gene sequences. Also, the length of Cy3- and Cy5-labeled cDNA by reverse transcription with random priming could be significantly different from sequence to sequence. This longer labeled cDNA could potentially lead to higher intensity levels for certain arrayed probes. If such interaction occurs, certain sequences will always show higher intensities in one channel than the other channel even under non-differential conditions and after normalization.

Variation in growth conditions and mRNA preparation or sample effect. In a comparative microarray experiment, two RNA samples extracted from cells grown under different conditions are labeled with different fluorescent dyes. Because of the differences in genetic identity (e.g., wild type *versus* mutant strains) and environmental growth conditions, cell biomass and mRNA abundance could fluctuate significantly among different cultures. The RNA purity also could be very different from sample to sample and this could lead to different labeling and hybridization efficiencies. Furthermore, sensitivity to mRNA degradation could be considerably different between preparations. All of these factors affecting signal intensity are referred to as *sample effects*. Due to experimental variations, hybridization signals from microarrays should be normalized prior to comparing data from a single array or multiple arrays.

2. Genes Used for Normalization

Two critical issues in the analysis of microarray data, are how to eliminate systematic variations and which genes should be selected as references for normalization. Experimental design largely determines the strategy used for normalization, three of which are described in detail below.

Using all genes on the array or global normalization. Under a certain condition, only a small portion of the genes is expected to be differentially expressed. Thus, the remaining genes should exhibit constant expression levels between two channels and can be used for normalization to calibrate spatial effects (Dudoit *et al*., 2001), slide effects and label effects (Tseng *et al*., 2001). The prerequisite for using almost all genes on the array for normalization is that only a small fraction of the genes are expressed, and the numbers of down- or up-regulated genes are approximately equal.

Using constantly expressed housekeeping genes. The housekeeping genes that are constantly expressed across a variety of conditions can be used for normalization (Duggan *et al*., 1999). Although it can be difficult to identify a set of housekeeping genes that do not change significantly under any condition, it

may be possible to identify small sets of temporary housekeeping genes for particular experimental conditions. A limitation in using housekeeping genes for normalization is that housekeeping genes are generally highly expressed and thus may not be representative of other genes of interest.

Using controls. A third normalization approach is to use spiked controls or a titration series of control sequences. In the spiked control method, DNA sequences from organisms different from the ones being studied are printed on the array and then the mRNAs of the control sequences are mixed with the two different mRNA samples in equal amounts. These spotted controls should have equal Cy5- and Cy3-derived intensities, and thus can be used for normalization. One limitation is that the composition of the control sequences could be considerably different from the target sequences, and as a result, they may not be representative of the genes of interest. Another limitation is that it may be difficult to determine how much mRNA to spike, because there are always varying amounts of rRNA and tRNAs present, and the degree of RNA degradation varies from sample to sample.

In the titration series method, a series of concentrations of the control sequences are printed on the arrays. These control spots are expected to have equal Cy5- and Cy3-derived intensities across a range of concentrations. Genomic DNA could be used in the titration method, because it should have a consistent expression level across various conditions.

3. Experimental Design and Normalization Strategies

Since microarray experiments have inherently high variation, careful experimental design and execution are critical for accurately identifying differentially expressed genes under different conditions. Appropriate normalization is therefore necessary to eliminate different types of systematic variations.

Minimizing spatial effects. To minimize spatial effects, multiple spots for a gene or control DNA should be fabricated on the microarrays. For control sequences, various concentrations of sequences should be spotted on arrays. Multiple spots of genes or control DNAs within the same slide are very useful for identifying contaminated spots, spots having high background noise, and problematical slides in each experiment (Tseng *et al*., 2001). To minimize spatial effects, normalization can be performed for each sector of the microarray-based on all the genes in that sector. Since DNAs in different sectors are deposited by different pins, normalization is an effective way to eliminate pin-to-pin variations (Dudoit *et al.*, 2001). By comparing the normalization results for different genes and

control sequences among different sectors, one should be able to assess and minimize the systematic variations associated with slide surface properties.

Minimizing labeling effects and label–gene interaction. To eliminate systematic variations in probe labeling and gene-label interaction, a reverse labeling experimental design is recommended (Kerr and Churchill, 2001a; Tseng *et al*., 2001). For this, two aliquots of the two RNA samples (A and B) are labeled with Cy3 or Cy5 separately, and then hybridized with two microarrays. The hybridization solution for the first microarray consists of Cy3-labeled sample A and Cy5-labeled sample B, whereas the labeling for the second microarray hybridization is reversed for the two target samples. Then the signal intensity for each microarray is normalized based on all genes on the microarray or on a set of housekeeping genes using different normalization approaches (see below). After normalization, the signal intensities from both channels for each sample are averaged, and the intensity ratios of the two samples are calculated based on the averaged signals. The reverse labeling experimental design is effective in eliminating labeling effects and gene-label interaction (Tseng *et al*., 2001).

Minimizing slide and sample effects. In a typical comparative study, multiple replicated treatments (e.g., 3) under each condition are used and mRNAs from two different conditions are fluorescently labeled and co-hybridized to the same single or multiple slides. Under such an experimental design, the signal intensity is impacted by both slide and sample effects, and it will be difficult to eliminate the resulting systematic variation based on using all arrayed genes for normalization. In this situation, one should identify a sufficient number of non-differentially expressed genes on each slide and use them to construct a normalization curve, because the expression level of the non-differentially expressed genes are expected to remain constant under the experimental conditions tested. One challenge is to select the non-differentially expressed genes. Although predetermined housekeeping genes are good candidates, they may not provide a good fit for normalization due to the high level of expression and natural variability of their expression level.

A rank invariant selection approach (Schadt *et al*., 2000; Tseng *et al*., 2001) can be used for selecting non-differentially expressed genes. This method presumes that for an up-regulated gene, the signal intensity rank for a channel will be significantly higher than the rank in the other channel, and vice versa. Briefly, the signal intensities of individual genes from both channels are ranked. If the ranks of Cy3 and Cy5 intensities for a gene differ by less than a certain threshold value, and the rank of the averaged intensity is not within the known levels of the lowest and highest ranks, then this gene is classified as a non-differentially expressed gene (Tseng *et al.,* 2001). This method works well if the majority of the genes are not differentially expressed. However,

this method may fail if majority of the genes are up- or down-regulated (Tseng *et al*., 2001).

4. Normalization Approaches

Methods for the normalization of microarray hybridization data can generally be categorized as linear or nonlinear. The major difference between these two types is that linear methods multiply all values in one channel by a correction factor, whereas nonlinear methods, which are preferred by most researchers working with microarrays, take the channel intensity into account and therefore are thought be more accurate. Here, we briefly describe the most commonly used normalization methods.

Correction factor based on total intensity. This method calculates a correction factor based on the total measured fluorescence intensity. The primary underlying assumption is that the total amount of RNA labeled with Cy3 and Cy5 is equal because the same amount of RNA from the same sample is used in separate labeling reactions. Although the spot for any one gene in one channel may be higher than that in the other, such variations should be averaged out over thousands of spots on the array. Therefore, the total integrated intensity of all spots should be equal in both channels, and a constant signal correction factor can be derived to rescale the signal intensity of the other channel.

Linear regression method. For differential experiments, it is expected that many genes will be expressed at a nearly constant level under two different growth conditions or treatments. Thus, the slope of the intensity in a scatter plot of both channels should be 1. Based on this assumption, the slope can be calculated by linear regression to obtain a correction factor, and then all values in one channel are multiplied by the correction factor to adjust the slope to 1.

Trimmed geometric mean (TGM). This nonlinear method was initially described by Morrison *et al.* (1999) and is generally recommended for most normalization needs. The method assumes that under a certain condition, only a small proportion of the genes will be differentially expressed. Thus, the remaining genes should display a constant level of expression and can be used for normalization (Beliaev *et al.*, 2002; Thompson *et al*., 2002). The signals from each channel are log transformed and sorted based on the intensity, then 5% of the extreme values (minimum and maximum) are discarded. The log-TGM and the SD of the log-trimmed means are calculated. The normalized value for a gene is obtained by dividing the difference between log intensity and log-trimmed means by the SD of the log-trimmed means. The normalized values are then

converted back from log to normal values, which are then used to calculate expression ratios.

Intensity-dependent nonlinear normalization method. In many cases, the dye bias is dependent on spot density (Dudoit *et al.*, 2001; Tseng *et al.*, 2001). Thus, an intensity-dependent normalization method may be preferable. Yang *et al.* (2001) proposed an intensity-dependent nonlinear normalization method that utilizes most of the genes on an array. Since this method is complicated, the reader is referred to the original paper for details (Dudoit *et al.*, 2001). Briefly, the log intensity ratio and the mean log intensity of both channels are calculated. The normalized intensity ratio is the difference between the actual log intensity ratio and the intensity ratio estimated based on Lowess function. Theoretically, this normalization method should be the most robust.

B. Data Transformation

Prior to statistically analyzing the microarray data, it is important to establish whether the data meet the underlying assumptions of the particular statistical model that will be used. The most common requirements for statistical techniques are that the data have a normal distribution and homogeneous variance. If the data do not meet these assumptions, they may be transformed and reevaluated to determine if they meet the underlying assumptions. If the data do not meet the assumptions, the statistical analyses will not be valid.

Although there are many different approaches to data transformation, the most commonly used approach in microarray studies is taking the logarithm of the quantified expression values. The rationale for this is three-fold. First, the variation in logs of intensities and logs of ratios of intensities are less dependent on absolute magnitude. Log transformation can equalize variability in microarray data with high variability. Second, log transformation evens out highly skewed distributions and thus brings the data closer to a normal distribution. Third, normalization is additive for logs of intensities. Studies show that log transformation is very effective in bringing the microarray data approximately to a normal distribution and is the best approach for the analysis of microarray-based gene expression data (Kalosai and Shams, 2001).

C. Methods for Identifying Differentially Expressed genes

Generally, normalized intensity ratios under two different experimental conditions are used to assess differentially expressed genes. Standard statistical techniques cannot be easily used to determine which level of difference in gene

expression reflects an actual biological difference. This is because of the inherently high variation associated with microarray experiments and low-level replications.

Three basic approaches are currently used for identifying differentially expressed genes. The first approach, which is commonly reported in the literature, is based on arbitrarily assigned fold differences (Schena *et al.*, 1996; Heller *et al.*, 1997). If the average expression level varies by more than a constant factor (e.g., 2) between the treatment and control conditions, then this gene is considered to have changed significantly in its expression. However, such a fixed fold rule is unlikely to identify real biologically differences, because a factor of two has a different significance, depending on the levels of gene expression and variation. The fold rule method is applicable only when the variance among the replicates within a treatment is identical for every gene so that the sample variance can be ignored. However, in practice, the variance differs among genes, and it is critical to incorporate such information into a statistical test.

The second approach is to use standard statistical *t*-test (Baldi and Long, 2001; Beliaev *et al.*, 2001; Thompson *et al.*, 2002) or paired *t*-tests (Rogge *et al.,* 2000) using the intensity ratio or log of the intensity ratio to test whether the fold change is significantly different from 1 or 0. When the t value exceeds a certain threshold, depending on the confidence level selected (typically the 95% confidence level or $P < 0.05$), the gene expression level is considered to be significantly different between two conditions. The *t*-test incorporates variance information and could potentially overcome the drawbacks of the fold rule method. Application of the *t*-test requires that all microarray experiments be highly replicated to obtain accurate estimates of the variance within experimental treatments. However, the level of replication within experimental treatments is often too low to permit *t*-tests, because the microarray experiments are costly and time-consuming to repeat or the amount of biological samples is very limited. A small number of replicates could lead to inaccurate estimation of variance and a correspondingly poor performance of the *t*-test itself (Baldi and Long, 2001).

The third approach is to apply Bayesian probabilistic model-based regularized *t*-test to improve the confidence in interpreting DNA microarray data with a low number of replicates (Baldi and Long, 2001). This method assumes that genes of similar expression levels have similar measurement errors, and that data from all of the genes with similar expression can serve as pseudo-replication of the experiment. Thus, variance of any single gene can be estimated by the weighted average of the variances from a number of genes with similar expression levels. This method has been applied to identify global expression profiles in *E. coli* K12 (Long *et al.*, 2001). The results showed that the Bayesian approach identified a stronger set of genes that were significantly up- or down-regulated and required less replication to achieve the same level of reliability as the *t*-test method. Since this method is computationally demanding, a program for accommodating this

approach, Cyber-T, is available at the Web interface, www.genomics.uci.edu/software.html. Various statistical methods are also available in ArrayStat™ for identifying differentially expressed genes.

D. Microarray Data Analysis

A massive amount of data is generated by microarray hybridization, and the great challenge is how to extract meaningful biological information. One of the key goals for microarray data visualization and analysis is to identify statistically significant up- and down-regulated genes and co-regulated genes exhibiting similar expression patterns. Although many different statistical methods have been used for analyzing microarray data, they are still in the early stages of development. In this section, several current methods will be briefly reviewed.

1. Scatter Plot

Scatter plots are the simplest way to visualize microarray expression data. In a comparative experiment, microarray hybridization is generally performed with two samples from two different conditions. One can use a scatter plot to visualize up- and down-regulated genes by assigning x- and y-axis values to represent signal intensity under the two different conditions. In the scatter plot, genes with equal expression values for two conditions fall along the diagonal identity line, whereas genes that are differentially expressed fall-off the diagonal line; the greater the deviation from the diagonal line, the greater the difference in the expression of a given gene between two samples.

2. Similarity Measurement

In a typical microarray experimental design, multiple experimental conditions at multiple time points are generally compared. In large experiments analyzing thousands of genes, the increased data volume makes it very difficult to identify gene expression patterns using scatter plots. More sophisticated multivariate analysis techniques should therefore be used in such cases. To use different multivariate analysis methods, the relationships among different genes should first be quantified based on signal intensity using appropriate metrics.

Two approaches are generally used for quantifying the relationships among different genes. One approach is to use Euclidean distance, which is defined as the square root of the summation of the squares of the differences between all pair-wise comparisons. This metric measures the absolute distance between two

points in space defined by the two gene expression profiles. In general, such distance measures are suitable when the objective is to cluster genes with similar expression patterns.

The other approach is to use Pearson correlation coefficient. For understanding regulatory networks, it is biologically more interesting to search for genes expressed at different levels but with similar overall profiles. Pearson correlation coefficient is ideal for identifying profiles of similar shape. The values of this correlation coefficient range from −1 (negative correlation) to 1 (positive correlation), and the method can detect both negatively and positively correlated genes. Several variations of the correlation metric have been used such as the correlation coefficient with an offset of zero for specifically taking into account the reference state (Eisen *et a*l., 1998) and jackknife correlation to counter against outlier effects (Heyer *et al.,* 1999).

3. Principal Component Analysis

Principal component analysis (PCA) is an exploratory multivariate statistical method for simplifying data sets that reduces the dimensionality of the variables by finding new variables, which are independent of each other. A few of the new variables, typically 2–3, are selected to explain the majority of variance in the original data. Since each principal component is a linear combination of the original variables, it is often possible to assign meaning to what the principal components represent. For microarray data analysis, genes or experiments can be considered as variables. PCA has been used in a variety of biochemical studies, including the analysis of microarray data in identifying outlier genes and/or experiments (Hilsenbeck *et al.*, 1999). The main advantage of PCA is that it identifies outliers in the data or genes that behave differently than most of the genes across a set of experiments. It can also be used to visualize clusters of genes that behave similarly across different experiments. However, the number of clusters in the data sets is arbitrary and dependent on the user's intuition or experience.

4. Cluster Analysis

One of the most commonly used methods is cluster analysis. Cluster analysis is used to identify groups of genes, or clusters that have similar expression profiles. Clusters and the genes within them can be subsequently examined for commonalities in functions as well as sequences in order to gain a better understanding of how and why they behave similarly. Cluster analysis can help

establish functionally related groups of genes and can predict the biochemical and physiological roles of functionally unknown ORFs.

Despite the emergence of many methods for microarray data analysis, the optimal way of classifying such data is still open to debate. Depending on the way in which the data are clustered, cluster analysis can be divided into hierarchical clustering and non-hierarchical clustering.

(i) *Hierarchical clustering*. This method attempts to group genes and/or experiments in small clusters and then group these clusters into higher-level clusters and so on. As a result of this grouping process, a tree structure called a dendrogram is generated for visualization of the relationships between genes and/or experiments. There are three common options for hierarchical analysis based on the definition of the distance between two clusters: single linkage, average linkage, and complete linkage (Heyer *et al.*, 1999). Although there are numerous versions of the basic algorithm, the most common is known as average linkage. Applications of hierarchical clustering to gene expression data have been described in recent studies (Eisen *et al.*, 1998).

Hierarchical clustering methods are very popular due to their simplicity and analysis speed. However, there are several problems associated with these methods (Heyer *et al.*, 1999). First, decisions to group two elements are based only on the distance between them and once elements are joined, it is impossible for them to be separated. In addition, it is a local decision-making method and does not consider the data as a whole. It suffers from a lack of robustness and solutions may not be unique and dependent on the data order, leading to incorrect clustering overall. Finally, the tree is extremely complex for large data sets, with the performance decreasing with the square of the number of genes requiring classification.

(ii) *Non-hierarchical clustering*. One of the typical non-hierarchical clustering methods is *k*-means clustering, which identifies predetermined *k* points as cluster centers. Each data point is assigned to one of these centers in a way that minimizes the total of distance between all points and their centers. The subsequent centers are chosen by identifying the data points farthest from the centers already chosen, and this process is iterated until the cluster memberships do not change appreciably (Tavazoie and Church, 1998).

The advantage of *k*-means clustering is that it provides sufficient clustering without having to create the full distance and similarity matrix or scan the whole dataset excessively (Zhou *et al.*, 2000). This is particularly useful for microarray data with large numbers of genes and many different experimental conditions. The algorithm converges quickly for good initial choices of the cluster centers. The main disadvantage of this method is that the number of clusters, *k*, must be specified prior to running the algorithm, and the final clustering relies heavily on

the choice of *k*. Generally, the number of clusters is not known in advance. In addition, the quality of the clusters identified by *k*-means is not guaranteed (Heyer *et al*., 1999). Recently, a new version of *k*-means (progressive *k*-means) was proposed to analyze gene expression data. This new procedure identifies the number of different clusters from the data itself and is independent of *a priori* specified number of clusters (Ben-Dor *et al*., 1999; Herwig *et al*., 1999). Despite such limitations, the *k*-means methods appear to perform quite well with a large number of genes (Dopazo *et al*., 2001).

To avoid limitations of hierarchical clustering and *k*-means clustering, another non-hierarchical clustering procedure, quality cluster, was developed (Heyer *et al*., 1999) that focuses on identifying large clusters with a quality guarantee. The quality cluster allows each ORF to initiate a candidate cluster, which is formed by starting each ORF and grouping the ORF with the greatest jackknife correlation coefficient. Other ORFs are iteratively added in a way to minimize the increase in cluster diameter without removing the ORFs, which previous clusters included (Heyer *et al*., 1999). One characteristic of this procedure is that the number of candidate clusters is equal to the ORF numbers and many candidate clusters overlap, with the largest candidate cluster being retained. The ORFs it contains are eliminated and the entire procedure is iterated on the smaller set of ORFs until the largest remaining cluster has fewer than some pre-specified number of elements.

There are several advantages of the quality cluster over both hierarchical and *k*-means clustering. First, the total number of clusters is not required prior to running the algorithm, and the quality of all clusters are guaranteed. Second, although the quality cluster algorithm is similar to the complete linkage hierarchical procedure, the clusters identified at a specified threshold are much larger on average. Third, since each ORF is considered a potential cluster center, local decisions do not have a great impact on the final clustering results. Thus, it is expected that this method is less sensitive than hierarchical approaches to small changes in the data such as removal of ORFs through filtering. Finally, this method is not sensitive to the order in which the similarity or distance data appear. Since this is a new clustering method, its value as an analysis tool remains to be determined.

5. Neural Network Analysis

Since clustering methods have some serious drawbacks in dealing with data with a significant amount of noise, a fundamentally different neural network-based approach has been proposed for microarray data analysis (Tamayo *et al*., 1999; Toronen *et al*., 1999; Herrero *et al*., 2001). Unsupervised neural networks, and in particular self-organizing maps (SOMs), are a more robust and accurate

method for grouping large data sets. The algorithm for neural network analysis works in the following way. First, a two-dimensional grid typically of hexagonal or rectangular geometry is defined. Then, similar to *k*-means clustering, the number of clusters (*k*) is specified to correspond to the representative points in the specified geometrical configuration. Data points are mapped onto the grid and the positions of the representative points are iteratively relocated in a way that each center has one representative point. Clusters close to each other in the grid will be more similar to each other than those further apart.

The main advantage of SOMs is that they are robust to noise. In other words, they are able to handle large data sets containing noisy, poorly defined items with irrelevant variables and outliers. This is particularly useful for analyzing microarray data. SOMs are also reasonably fast and can be easily scaled up to large data sets. One disadvantage of SOMs is that they require pre-determined choices about geometry, like the *k*-means method. The number of clusters is arbitrarily fixed from the beginning and consequently, it is difficult to recover the natural cluster structure of the data set. SOMs also yield non-proportional classification. If irrelevant data or some particular type of profile is abundant, the most interesting profile will be mapped in few clusters and hence their resolution could be low. In addition, it is very difficult to detect higher-order relationships between clusters of profiles due to the lack of a tree structure (Herrero *et al*., 2001).

To overcome some of the limitations of SOMs, an unsupervised neural network with a binary tree topology, termed the self-organizing tree algorithm (SOTA), was proposed (Dopazo and Carazo, 1997). This new algorithm combines the advantages of hierarchical clustering (tree topology) and neural network (accuracy and robustness) and was used to analyze gene expression data (Herrero *et al*., 2001).

VII. USING MICROARRAYS TO MONITOR GENOMIC EXPRESSION

Microarrays have been used widely to quantify and compare global gene expression in a high-throughput fashion. This section briefly reviews the fundamental basis, general approaches to experimental design, and hybridization performance of microarrays in monitoring gene expression levels.

A. General Approaches to Revealing Differences in Gene Expression

Temporal and spatial information concerning gene expression, as well as changes in mRNA abundance levels in response to different environmental

conditions, are important for understanding gene function and regulation. Three comparative approaches have been used for the display of differential gene expression.

1. Differential Display of mRNA Under Different Physiological Conditions

Cells of interest are cultured under different physiological conditions, and the differences in mRNA abundance between the test and reference samples are compared using high-density microarrays. This is the most straightforward and widely used approach for identifying gene expression patterns associated with various physiological states (DeRisi *et al*., 1997; Tao *et al*., 1999; Ye *et al.*, 2000; Beliaev *et al*., 2002).

2. Differential Display of Temporal Gene Expression

Cells of interest are grown under a specific physiological condition and then harvested at different time points during growth. Changes in mRNA levels are revealed using microarrays. Information on the temporal dynamics of gene expression is very useful in understanding when genes are turned on or off and how genes interact with each other (DeRisi *et al.*, 1997; Liu *et al.*, 2003).

3. Comparison of Gene Expression Patterns Between Wild-type and Mutant Cells

Differences in gene expression in response to changing environmental conditions can be very complicated, and oftentimes the expression profiles of many genes are altered as a result. Changes in the expression profiles for many genes present a great challenge to understand the underlying molecular mechanisms controlling these genes. The most effective approach to define the contributions of individual regulatory genes in a complex metabolic process is to use DNA microarrays to identify genes whose expression is affected by mutations in putative regulatory genes (DeRisi *et al*., 1997; Beliaev *et al*., 2002; Thompson *et al*., 2002).

The basic approach to microarray-based gene expression studies is outlined in Fig. 5. In a typical microarray experiment for monitoring gene expression, gene-specific PCR primers are designed based on whole-genome sequence information and synthesized. Gene-specific fragments are then amplified with specific primers, purified, and arrayed on solid substrates. Once the microarrays are ready, total cellular RNA isolated from bacterial cells grown under two different

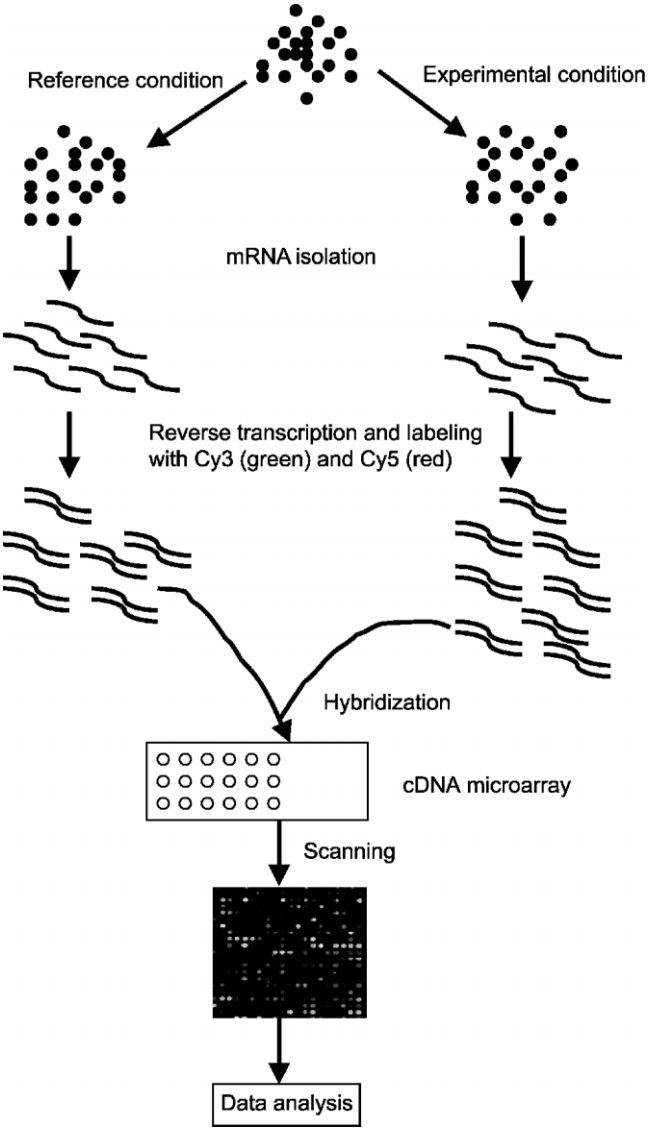


Figure 5 General approach for using microarrays for monitoring gene expression.

conditions (a control and experimental condition) is fluorescently labeled with different dyes (Cy3 or Cy5) *via* the enzyme, reverse transcriptase. The microarray is then simultaneously hybridized with fluorescently tagged cDNA from the test and reference samples. The signal intensity of each fluorescent dye on the array is

then measured with a confocal laser scanning microscope or CCD camera. The quantitative ratio of red (Cy5) to green (Cy3) signal for each spot reflects the relative abundance of that particular gene in the two experimental samples. With appropriate controls, the intensity can be converted into biologically relevant outputs (e.g., the number of transcripts per cell). A series of samples can be compared with each other through separate co-hybridizations with a common reference sample, and the data can be analyzed with various statistical methods. Eisen and Brown (1999) provide a detailed discussion of the technical aspects of microarray experiments for monitoring gene expression.

B. Experimental Design for Microarray-based Monitoring of Gene Expression

Microarray experiments generate massive data sets, which must be analyzed and interpreted in a rapid and meaningful way. To improve the efficiency and reliability of experimental data, careful experimental design is needed. Without this, the collected data may fail to answer the research question of interest or lead to a biased, inadequate interpretation of the experimental results (Yang and Speed, 2002).

The main objective of experimental design is to make the data analysis and interpretation as simple and powerful as possible. For a competitive microarray hybridization experiment in which two fluorescent dyes are used, the most important experimental design issue is how the mRNAs are labeled and which mRNAs are hybridized together on the same slide (Yang and Speed, 2002). In most experiments, several designs can be devised. The selection of the most appropriate design will depend on the particular research questions being asked, the number of comparisons, the number of slides available for hybridization, the amount of mRNAs available, and cost.

Various design schemes have been described in great detail by Yang and Speed (2002) and several designs could be devised for a particular microarray experiment. The microarray experiment design scheme can be classified into the three categories (Fig. 6): reference design, all-pairs design, and loop design. In reference design, all treatment samples are labeled with one dye and are hybridized, respectively, with a common reference sample labeled with another dye (Fig. 6A). This indirect design is used widely in gene expression studies. This design is especially suitable when the amount of mRNA from treatment samples is limited and when many treatment samples are compared. Another advantage of this design is that data analyses and interpretation are easy and do not require sophisticated statistical tools. However, the average variance for this indirect reference design is considerably higher than that for the other designs. Since it is straightforward, the reference design is used much more often than the other designs.

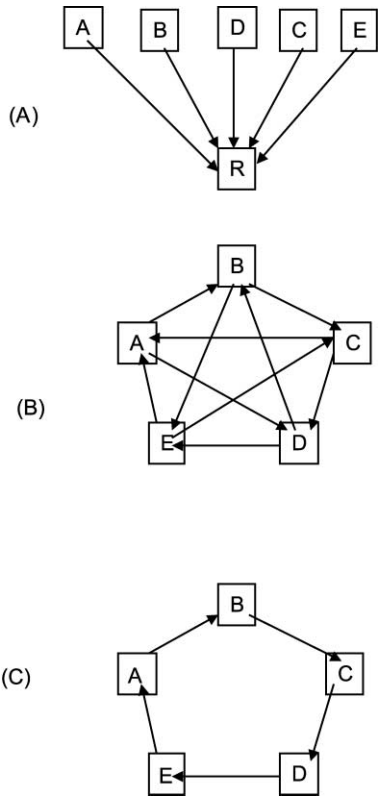


Figure 6 Illustrations of basic types of microarray experimental design schemes with five treatment samples. By convention, the green-labeled sample (Cy3) is placed at the tail while the red-labeled sample (Cy5) is placed at the head of the arrow. (A) Reference design. The five treatment samples (A–E) are labeled with one dye and hybridized, respectively, with the common reference sample R, which is labeled with the other dye. Altogether five hybridizations are needed. (B) All-pair design. Each sample is labeled twice with red and twice with green. Ten pair-wise hybridizations are needed. (C) Loop design. Each sample is labeled once with red and once with green. Five successive pair hybridizations are needed.

In the all-pairs design scheme, all of the treatment samples are labeled with different fluorescent dyes and directly hybridized together in pair-wise fashion (Fig. 6B). The main advantage of this design is that more precise comparisons among different treatment samples can be obtained. However, this design is unlikely to be feasible and desirable when a large number of comparisons are made, due to the constraints on mRNA quantity and cost. Finally, in the loop design, all of the treatments are successively connected as a loop (Fig. 6C) (Kerr and Churchill, 2001a, b). Using the same number of microarrays as the reference design, the loop design obtains twice as much data on the treatments of interest.

The loop design requires far fewer slides than the all-pairs design. However, long paths between some pairs of treatment samples are needed in larger loops, and thus some comparisons are much less precise than others (Yang and Speed, 2002). Another practical problem is that each sample should be labeled with both Cy dyes, which doubles the number of labeling reactions. In addition, the failure of microarray hybridization in one sample will affect the analysis of other samples in the loop.

C. Microarray-based Functional Analysis of Environmental Microorganisms

Genome sequence information for a number of bacteria and archaea of potential environmental or biotechnological relevance is accumulating rapidly and includes representatives from such genera as dissimilatory metal-reducing bacteria (*Shewanella oneidensis* [Heidelberg *et al*., 2002]), extreme radiation-resistant bacteria (*Deinococcus radiodurans* [White *et al*., 1999]), photosynthetic cyanobacteria (*Anabaena* sp. strain PCC 7120 [Kaneko *et al*., 2001], *Synechocystis* sp. strain PCC6803 [Kaneko *et al*., 1996]), thermophilic and hyperthermophilic archaea (*Pyrococcus horikoshii* [Kawarabayasi *et al*., 1998], *Aeropyrum pernix* [Kawarabayasi *et al*., 1999], *Thermotoga maritima* [Nelson *et al*., 1999], *Thermoplasma volcanium* [Kawashima *et al*., 2000], *Pyrococcus furiosus* [Robb *et al*., 2001], *Pyrobaculum aerophilum* [Fitz-Gibbon *et al*., 2002]), thermoacidophilic archaea (*Sulfolobus tokodaii* [Kawarabayasi *et al*., 2001], *Sulfolobus solfataricus* [She *et al*., 2001]), methanogens (*Methanococcus jannaschii* [Bult *et al*., 1996], *Methanobacterium thermoautotrophicum* [Smith *et al*., 1997], *Methanopyrus kandleri* [Slesarev *et al*., 2002], *Methanosarcina acetivorans* [Galagan *et al*., 2002]), sulfate-reducing archaea (*Archaeoglobus fulgidus* [Klenk *et al*., 1997]), and halophilic archaea (*Halobacterium* species NRC-1 [Ng *et al*., 2000]). However, to-date, very few studies have explored the transcriptomes of these organisms using microarray technology. The large majority of microarray-based genomic expression analyses have focused on bacterial pathogens and such model organisms as *E. coli*, *B. subtilis*, and yeast. In this section, we will briefly discuss microarray profiling of gene expression in three organisms of environmental significance, namely, *S. oneidensis*, *D. radiodurans*, and *P. furiosus*, as examples of the application of microarrays to environmental microbiology.

1. *Shewanella oneidensis*, a Dissimilatory Metal-reducing Bacterium

S. oneidensis MR-1 (formerly *Shewanella putrefaciens* strain MR-1 [Venkateswaran *et al*., 1999]) is a facultatively anaerobic γ-proteobacterium

that is noted for its remarkably diverse respiratory capacities. In addition to utilizing oxygen as a terminal electron acceptor during aerobic respiration, *S. oneidensis* can anaerobically respire various organic and inorganic substrates, including oxidized metals (e.g., Mn(III) and (IV), Fe(III), Cr(VI), U(VI)), fumarate, nitrate, nitrite, thiosulfate, elemental sulfur, trimethylamine *N*-oxide, DMSO, and anthraquinone-2,6-disulphonate (Lovley, 1991; Nealson and Saffarini, 1994; Moser and Nealson, 1996). This unusual versatility in the use of alternative electron acceptors for anaerobic respiration is conferred in part by complex electron transport networks, the components of which remain to be elucidated (Richardson, 2000). The metal ion-reducing capabilities of this bacterium, in particular, have important implications with regard to the potential for *in situ* bioremediation of metal contaminants in the environment. However, the effective prediction and assessment of bioremediation performance or activity is complicated due to insufficient knowledge concerning the gene networks and regulatory mechanisms enabling microbial metal reduction.

To expedite understanding of metal reduction by *S. oneidensis* MR-1, its ~5-Mb genome was determined recently by The Institute for Genomic Research (TIGR) under the support of the U.S. Department of Energy (DOE) (Heidelberg *et al*., 2002), making it feasible to apply microarray technology to the study of energy metabolism in this bacterium. The transcriptional response of *S. oneidensis* to different respiratory growth conditions (Beliaev *et al*., 2002b) and to the disruption (inactivation) of genes encoding putative transcriptional regulators (Beliaev *et al*., 2002a; Thompson *et al*., 2002) were examined using DNA microarrays containing 691 arrayed genes. These partial genome microarrays consisted of PCR-amplified MR-1 ORFs putatively involved in energy metabolism, transcriptional regulation, adaptive responses to environmental stress, iron acquisition, and transport systems according to the sequence annotation. These arrays were constructed prior to the closure and publication of the *S. oneidensis* genome sequence.

To identify genes specifically involved in anaerobic respiration, differential mRNA expression profiles of *S. oneidensis* were monitored under aerobic and fumarate-, Fe(III)-, or nitrate-reducing conditions using partial genome microarrays (Beliaev *et al*., 2002b). Gene expression profiling indicated that 121 of the 691 arrayed ORFs showed at least a 2-fold difference in mRNA abundance in response to changes in growth conditions (Beliaev *et al*., 2002b), with a number of genes required for aerobic growth being repressed in the transition from aerobic to anaerobic respiration. Genes induced in a general response to anaerobic respiration, irrespective of the terminal electron acceptor, belonged to several different categories of cellular function: cofactor biosynthesis and assembly, substrate transport, and anaerobic energy metabolism. Of particular importance was the observation that certain genes preferentially displayed increased transcript levels in response to specific electron acceptors. For example, the expression of genes encoding a periplasmic nitrate reductase

(*napBHGA* operon), cytochrome c_{552}, and prismane was elevated 8- to 56-fold specifically in response to the presence of nitrate, while genes encoding a tetraheme cytochrome *c* (*cymA*), a flavocytochrome *c* (*ifcA*), and a fumarate reductase (*frdA*) were preferentially induced 3- to 8-fold under conditions of fumarate reduction. In addition, the mRNA abundance levels for two oxidoreductase-related genes of unknown function and several cell envelope genes involved in multidrug resistance increased specifically under Fe(III)-reducing conditions. This work represented the first attempt to characterize a complex system in *S. oneidensis* on a genome scale. Other microarray-based transcriptomic studies have focused on defining the functions of putative *S. oneidensis* regulatory genes encoding a ferric uptake regulator (*fur*; Thompson *et al.*, 2002) and an electron transport regulator (*etrA*; Beliaev *et al.*, 2002a).

2. *Deinococcus radiodurans*, an Extreme Radiation-resistant Bacterium

D. radiodurans strain R1 is the most characterized member of the DNA-damage resistant bacterial family *Deinococcaceae*, which is comprised of at least seven different species that form a distinct eubacterial phylogenetic lineage (Makarova *et al.*, 2001). *D. radiodurans* is a Gram-positive, non-sporulating bacterium that was originally isolated in 1956 from canned meat that had spoiled following exposure to X-ray sterilization (Anderson *et al.*, 1956). Species in the genus *Deinococcus*, particularly *D. radiodurans*, are extremely resistant to a number of physicochemical agents and environmental conditions that damage DNA, including ionizing and ultraviolet radiation, desiccation, heavy metals, and oxidative stress (reviewed in Minton, 1996; Battista, 1997; Battista *et al.*, 1999). Studies have demonstrated that *D. radiodurans* can survive acute exposures to gamma radiation that exceed 15,000 Gy without lethality or induced mutation (Daly *et al.*, 1994; Daly and Minton, 1995) and flourish in the presence of high-level chronic irradiation (60 Gy/h) (Lange *et al.*, 1998; Venkateswaran *et al.*, 2000). *D. radiodurans* also expresses an intrinsic ability to reduce metals and radionuclides (Fredrickson *et al.*, 2000) and thus has potential applications for the bioremediation of metal- and radionuclide-contaminated sites where the presence of radioactivity prohibitively restricts the activity of more sensitive dissimilatory metal-reducing bacteria such as *Shewanella*.

To enhance the understanding of the molecular basis of extreme DNA damage resistance, the complete genome of *D. radiodurans* R1 was sequenced by TIGR (White *et al.*, 1999) under DOE support. Sequence analysis of this organism's multigenomic content indicates that essentially the entire repertoire of recombinational DNA repair genes identified in *D. radiodurans* has functional homologs in other prokaryotes (White *et al.*, 1999; Makarova *et al.*, 2001), suggesting that the extreme radioresistance of R1 may be attributable to novel genes, repair pathways, and mechanisms yet to be described.

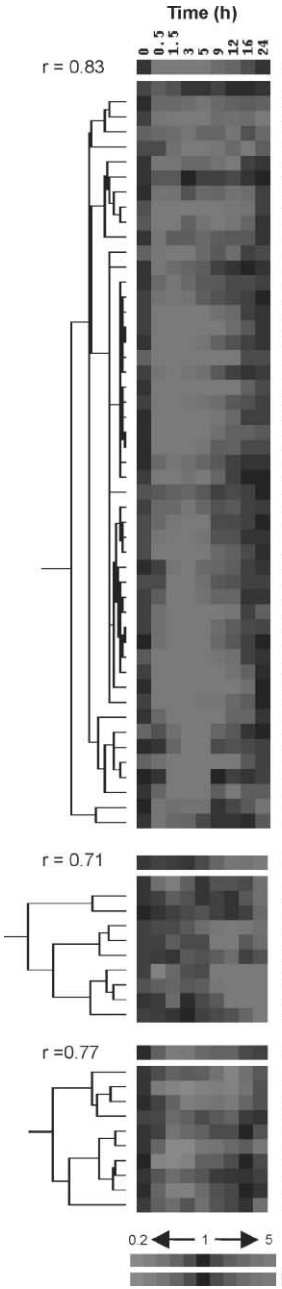
Detailed computational genomic analyses alone, therefore, are unlikely to uncover the fundamental answers underlying the remarkable ability of *D. radiodurans* to withstand DNA-damaging conditions.

The transcriptome dynamics of *D. radiodurans* in response to cellular recovery from acute ionizing radiation was examined using DNA microarrays covering ~94% of the organism's predicted protein-encoding genes (Liu *et al*., 2003). In this time-series study, *D. radiodurans* cells exposed to acute irradiation (15 kGy) were allowed to recover at 37°C for time intervals ranging from 0 to 24 h. *Deinococcus* transcriptome dynamics were monitored in cells representing early (0–3 h), middle (3–9 h), and late (9–24 h) phases of recovery from ionizing radiation and compared to non-irradiated control cells. Microarray analysis of genomic expression patterns revealed a large number of *D. radiodurans* genes responding to acute irradiation: 832 genes (28% of the genome) were induced and 451 genes (15% of the genome) were repressed two-fold or greater at one point during *D. radiodurans* recovery (Liu *et al*., 2003). Genes exhibiting increased transcription in the early phase of cell recovery belonged to a number of broad functional groups, including DNA replication, DNA repair, recombination, cell wall metabolism, cellular transport, and uncharacterized proteins.

Hierarchical clustering of genes showing differential expression revealed similar expression patterns for groups of genes and clusters of presumably co-regulated genes (Fig. 7). Genes responding to recovery from irradiation clustered into three distinct groups: (1) *recA*-like activation pattern (based on the expression profile of *recA*, which is critical for *D. radiodurans* recovery and is substantially upregulated during early-phase recovery and down-regulated before the onset of late phase), (2) growth-related activation pattern, and (3) repressed patterns. Unexpectedly, genes encoding tricarboxylic acid (TCA) cycle components were repressed in the early and middle phases of recovery, whereas genes encoding the glyoxylate shunt pathway were induced during this interval (Liu *et al.,* 2003). In addition, a number of poorly characterized genes showed high induction folds in expression during at least one phase of recovery, thus implicating their encoded proteins in the functional role of cell recovery. The response of metabolic gene systems, however, is not immediately clear and will require further, more focused experimentation. The study by Liu *et al*. (2003) represents the first published description of the application of DNA microarrays to the functional analysis of *D. radiodurans* and suggests that the recovery process for this organism involves the complicated coordination of DNA repair and metabolic functions as well as other cellular functions.

3. *Pyrococcus furiosus*, a Hyperthermophilic Archaeon

P. furiosus is a member of a phylogenetically distinct group of prokaryotes called the Archaea, which constitutes a primary, separate domain in the universal



Gene#	putative function[a]	Ratio (fold)[b]	Time (hr)[c]
A. *recA*-like activation pattern			
DR0911	DNA-directed rna polymerase beta subunit, *rpoC*	1.99 (±1.37)	0.5
DR2220	Tellurium resistance protein TerB	3.13 (±1.49)	5
DR2221	Tellurium resistance protein TerE	5.24 (±2.94)	3
DRB0069	Subtilisin serine protease	3.18 (±1.39)	3
DRB0067	Extracellular nuclease with Fibronectin III domains	4.37 (±1.21)	3
DR0261	8-oxo-dGTPase, *mutT*	3.36 (±1.68)	0.5
DRA0344	LEXA repressor, HTH+protease, *lexA*	1.80 (±1.08)	1.5
DR0099	SsDNA-binding protein, *ssb*	3.01 (±1.20)	0.5
DR2129	Ribosomal component L17, *rplQ*	5.92 (±2.09)	1.5
DR2128	RNA polymerase alpha subunit, *rpoA*	4.03 (±2.80)	1.5
DR0324	Probable glutamate formiminotransferase	3.30 (±1.47)	0.5
DR2337	Uncharacterized protein	7.41 (±5.71)	1.5
DRA0346	PprA protein, involved in DNA damage resistance	3.52 (±1.94)	0.5
DR1825	Protein-export membrane protein	3.21 (±1.48)	1.5
DR1771	UVRA ABC family ATPase, *uvrA-1*	3.52 (±1.15)	1.5
DRA0345	Predicted esterase	10.05 (±4.39)	1.5
DR0422	Trans-aconitate methylase	18.85 (±7.46)	1.5
DR1143	Uncharacterized protein	8.85 (±4.26)	1.5
DR0003	Uncharacterized protein	14.03 (±5.53)	1.5
DR1776	Nudix family pyrophosphatase	4.70 (±2.83)	1.5
DR2340	RecA, *recA*	7.98 (±3.86)	1.5
DR2610	Periplasmic binding protein, *fliY*	4.13 (±1.67)	0.5
DR1645	Teichoic acid biosynthesis protein, *wecG*	5.88 (±2.79)	1.5
DR0696	V-type ATPase synthase, subunit K	7.19 (±2.16)	1.5
DR0421	Uncharacterized protein	4.94 (±2.30)	1.5
DR1775	Superfamily I helicase, *uvrD*	3.30 (±1.69)	1.5
DR1561	UDP-N-acetylglucosamine 2-epimerase, *wecB*	6.00 (±1.40)	1.5
DR2285	MutY, A/G-specific adenine glycosylase, *mutY*	2.36 (±0.40)	3
DR2356	Nudix family hydrolase	3.35 (±0.45)	3
DR2275	Excinuclease ABC subunit B, *uvrB*	4.93 (±1.81)	3
DR0206	Uncharacterized protein	5.45 (±2.65)	3
DR0204	Uncharacterized membrane protein	6.01 (±1.35)	3
DR1354	Excinuclease ABC subunit C, *uvrC*	3.78 (±0.42)	3
DR0203	Uncharacterized membrane protein	3.82 (±0.86)	1.5
DR0205	ABC transporter ATPase	4.10 (±2.45)	3
DR1357	ABC transporter, permease subunit	6.79 (±2.56)	1.5
DR2482	Predicted transcription regulator	5.75 (±2.92)	1.5
DR2483	McrA nuclease	5.43 (±1.22)	1.5
DRA0008	Conserved membrane protein	6.60 (±2.00)	3
DRA0234	Uncharacterized protein,	12.76 (±5.27)	1.5
DR1359	ABC transporter, periplasmic subunit	24.83 (±11.13)	1.5
DR2127	Ribosomal protein S4, *rpsD*	5.40 (±1.50)	3
DR1356	ABC transporter, ATP-binding protein	9.85 (±5.98)	3
DRB0136	Putative DEAH ATP-dependent helicase, *hepA*	5.22 (±0.46)	3
DR1548	Bacillus ykwD ortholog, PRP1 superfamily protein	5.62 (±2.35)	3
DR0207	ComEA related protein, secreted	15.47 (±8.31)	3
DRA0249	Metalloproteinase, leishmanolysin-like	6.47 (±4.43)	3
DR0665	Uncharacterized protein	11.66 (±5.74)	3
DR0596	Resovasome RuvABC, subunit B, *ruvB*	3.22 (±1.31)	0.5
DR0912	DNA-directed rna polymerase beta subunit, *rpoB*	3.19 (±0.80)	0.5
B. Growth-related activation pattern			
DR1172	Lea76/LEa29-like desiccation resistance protein	2.66 (±0.60)	24
DR0461	Bacillus yacB ortholog	2.58 (±0.81)	24
DR1595	6-phosphogluconate dehydrogenase, *gnd*	2.30 (±0.52)	24
DRA0043	TDP-rhamnose synthetase	5.08 (±2.12)	12
DRA0042	Glucose-1-phosphate thymidylyltransferase, *rfbA*	3.70 (±1.19)	12
DRA0031	Glucose-1-phosphate thymidylyltransferase	2.48 (±1.64)	12
DRA0065	Chromosomal protein HU HupA, *hupA*	7.71 (±2.07)	24
DR2263	Bacterioferritin, Iron chelating protein	6.41 (±1.97)	16
DRA0275	Soluble cytochrome C	4.80 (±1.22)	24
DR1279	Superoxide dismutase (Mn)	3.91 (±1.43)	24
C. Repressed pattern			
DR1126	RecJ like DHH superfamily Phosphohydrolase	0.33 (±0.12)	12
DR1337	Transaldolase, *tal*	0.25 (±0.05)	3
DR0728	Fructokinase, *cscK*	0.37 (±0.13)	3
DR0977	Phosphoenolpyruvate carboxykinase, *pckA*	0.48 (±0.22)	1.5
DR1742	Glucose-6-phosphate isomerase, *pgi*	0.42 (±0.12)	1.5
DR1998	Catalase, CATX, *katA*	0.23 (±0.07)	3
DR1146	GSP26 general stress like protein	0.25 (±0.06)	1.5
DR0493	Formamidopyrimidine-DNA glycosidase, *mutM*	0.46 (±0.09)	1.5
DR0674	Argininosuccinate synthase, ASSY, *argG*	0.35 (±0.15)	3
DR2620	Cytochrome oxidase subunit I, COX1, *caaA*	0.45 (±0.25)	5

tree of life (Woese *et al*., 1990; Olsen and Woese, 1997). The Archaea domain is composed of organisms with diverse phenotypes, such as methane-producing methanogens, extreme halophiles, and extremely thermophilic sulfur-metabolizing species (Woese, 1987). Typically, archaeal genes involved in energy production, cell division, cell wall biosynthesis, and metabolism have homologs in bacteria, whereas genes encoding proteins that function in the informational processes of DNA replication, transcription, and translation are more similar to their eucaryal counterparts (Bult *et al*., 1996). Archaea also share certain RNA processing components with Eucarya, such as fibrillarin (a pre-rRNA processing protein) (Bult *et al*., 1996; Belfort and Weiner, 1997) and tRNA splicing endonucleases (Belfort and Weiner, 1997; Kleman-Leyer *et al*., 1997). The mosaic nature of archaea makes this group of organisms extremely interesting from an evolutionary perspective. The sequencing and analysis of archaeal genomes should provide valuable insights into the origin or evolution of eukaryotes, as well as the molecular mechanisms enabling their adaptation to extreme environments.

The hyperthermophilic archaeon *P. furiosus* is able to grow optimally at a temperature of 100°C (Fiala and Stetter, 1986). Studies support a highly regulated fermentative-based metabolism in *P. furiosus* (Adams *et al*., 2001), which can utilize the disaccharide maltose in the presence or absence of elemental sulfur (S^0). In addition, *P. furiosus* can couple the reduction of S^0 to the oxidation of catabolism-generated, reduced ferredoxin, but the molecular mechanism of this metabolic coupling is not presently known (Schut *et al*., 2001). The availability of the complete genome sequence of *P. furiosus* (Robb *et al*., 2001) permits the global analysis of gene function and expression using high-density DNA microarray technology. To investigate the molecular basis of S° metabolism, Schut *et al*. (2001) used DNA microarrays containing 271 ORFs (of the ca. 2200 total ORFs predicted) from the *P. furiosus* genome (1.9 Mb) to measure

Figure 7 Hierarchical clustering analyses of expression profile patterns. Gene expression patterns are displayed graphically. Three distinct patterns are sorted according to the hierarchical clustering analyses, i.e., (A) *recA*-like activation pattern, (B) growth-related activation pattern, and (C) repressed patterns. The top row represents the general pattern of the selected group where a Pearson correlation coefficient (r) is shown on the left side. All displayed graphs are organized in a row/column format. Each row of colored strips represents a single gene whose expression levels are color-recorded sequentially in every column of the same row that represents recovery time intervals. Red color denotes up-regulation, whereas green indicates down-regulation. Black indicates the control level. The variation in transcript abundance is positively correlated with the color darkness. (a) Gene numbers are offered for tracking the primary information of the gene of interest. (b) The maximum (for *recA*-like and growth-related activation pattern) or minimum (for the repressed pattern) expression level for each of the exhibited genes over the 24-h recovery period is presented as the dye intensity ratio of the irradiated sample to the non-irradiated control at (c) the indicated time interval. Values in parentheses show the SD for each mean expression ratio (Courtesy of PNAS).

differential gene expression in cells grown at 95°C on maltose in the presence or absence of S°. The arrayed PCR products represented ORFs with proposed functions in sugar and peptide catabolism, metal utilization, and the biosynthesis of cofactors, amino acids, and nucleotides. This study by Schut *et al*. (2001) represents the first and, to-date, only account describing the application of DNA microarray analysis to a member of the Archaea. Currently published genomic analyses of archaea have been almost exclusively limited to the sequencing, annotation, and *in silico* comparative analysis of archaeal genomes.

DNA microarray analysis revealed a number of ORFs whose expression was dramatically down-regulated (>5-fold decrease) by S°, including 18 genes encoding various subunits associated with three different hydrogenase systems (Schut *et al*., 2001). Other genes displaying decreased transcription, when *P. furiosus* cells were grown with S°, encoded a hypothetical protein and two homologs (ornithine carbamoyltransferase and HypF) involved in hydrogenase biosynthesis. In the presence of S°, the expression of two previously uncharacterized ORFs (encoding products designated SipA and SipB for "sulfur-induced proteins") increased by a striking >25-fold. The encoded proteins of these ORFs were proposed by the authors to be part of a novel S°-reducing, membrane-associated, iron-sulfur cluster-containing complex in *P. furiosus* (Schut *et al*., 2001). The research reported by Schut *et al*. (2001) clearly illustrates the power of DNA microarray analysis in generating new lines of experimentation and in implicating previously uncharacterized ORFs identified by genome sequencing in biological processes. There is little doubt that the continuing determination of archaeal genomes will spawn more microarray-based functional studies of extremophiles.

VIII. APPLICATION OF MICROARRAYS TO ENVIRONMENTAL STUDIES

In addition to monitoring transcription patterns on a genomic scale, microarray-based technology is well suited for detecting microorganisms in natural environments. Many target functional genes involved in biogeochemical cycling in environments are highly diverse, and it is difficult or impossible to identify conserved regions for designing PCR primers or oligonucleotide probes. The microarray-based approach does not require such sequence conservation, because all of the diverse gene sequences from different populations of the same functional group can be fabricated on arrays and used as probes to monitor their corresponding populations.

In contrast to studies using pure cultures, microarray analysis of environmental nucleic acids presents a number of technical challenges that must be overcome. First, target and probe sequences in environmental samples can be

very diverse, and it is not clear whether the performance of microarrays with diverse environmental samples is similar to that with pure culture samples and how sequence divergence is reflected in hybridization signal intensity. Second, environmental samples are generally contaminated with substances such as humic matter, organic contaminants, and metals, which may interfere with the hybridization reaction on microarrays. Third, in contrast to pure cultures, the recoverable biomass in environmental samples is generally low; consequently, it is not clear whether microarray hybridization is sensitive enough to detect microorganisms in all types of environmental samples. Finally, it is uncertain whether microarray-based detection can be quantitative. Environmental and ecological studies require experimental tools that not only detect the presence or absence of particular groups of microorganisms but also provide quantitative data on their *in situ* biological activities.

In the following sections, we discuss three different types of microarray formats that have been developed for use in environmental studies: functional gene arrays (FGAs), phylogenetic oligonucleotide arrays (POAs), and community genome arrays (CGAs).

A. Functional Gene Arrays

Genes encoding functional enzymes involved in various biogeochemical cycling processes (e.g., carbon, nitrogen, sulfate and metals) are very useful as molecular signatures for assessing the physiological status and functional activities of microbial populations and communities in natural environments. Microarrays containing functional gene sequence information are referred to as FGAs, because they are primarily used for the functional analysis of microbial community activities in environments (Wu *et al.*, 2001). Similar to gene expression profiling arrays, both oligonucleotides and PCR-amplified DNA fragments corresponding to functional genes can be used for fabricating FGAs.

1. Selection of Gene Probes

FGAs are designed for studying functional gene diversity in natural environments. To construct FGAs, the gene probes should be carefully defined and selected based on the specific research questions to be addressed. As an example, microarrays can consist of gene probes that are involved in such biogeochemical processes as nitrification (ammonia monooxygenase, *amoA*), denitrification (nitrite reductases, *nirS* and *nirK*), nitrogen fixation (nitrogenases, *nifH*), sulfite reduction (sulfite reductase, *dsv*A/B), methanogenesis

(methyl coenzyme M reductase genes, *mcrA*), methane oxidation (methane mono-oxygenases, *mmo*), and plant and fungal polymer degradation (cellulases, xylanases, ligin peroxidases, chitinases).

Probes for the construction of FGAs can be generated in three ways. The desired gene fragment can be amplified from genomic DNA extracted from pure bacterial cultures using specific primers or from cloned plasmids containing the desired gene insert using vector-specific primers. However, the availability of pure cultures and plasmid clones can be limited. In the second approach, desired gene fragments are recovered from natural environments using PCR-based cloning methods (Zhou *et al*., 1997). Generally, sequences that show $>85\%$ identity can be used as specific probes for FGAs. These two approaches were used to construct FGAs containing nitrite reductase genes and ammonia monooxygenase genes for monitoring bacteria involved in nitrification and denitrification, respectively (Wu *et al*., 2001). Finally, in the third strategy, oligonucleotides, usually 50–70-mers, are designed based on the functional sequences available in public databases and synthesized for microarray fabrication (Tiquia *et al*., unpublished).

2. Specificity

Hybridization specificity is an important parameter that impacts any detection method. It is influenced by many factors, such as G + C content, degree of sequence divergence, sequence length, secondary structure of the probe, temperature, and salt concentrations. To determine the specificity of DNA microarray hybridization, we have constructed and used FGAs consisting of heme- and copper-containing nitrite reductase genes, ammonia monooxygenase (*amo*A), and methane monooxygenase genes [*pmo*A] (Wu *et al*., 2001). Small subunit (SSU) rRNA genes and yeast genes were used as positive and negative controls, respectively, on the FGAs. Cross-hybridization among different gene groups was not observed at either low (45°C) or high (65°C) stringency. Furthermore, no hybridization was observed with any of the five yeast genes, which served as negative controls for hybridization on the array (Fig. 8A). Based on the sequence similarities, it was estimated that microarray hybridization can differentiate between sequences exhibiting a dissimilarity of approximately 15% at 65°C and 10% at 75°C (Wu *et al*., 2001). In addition, at low stringency, most *nirS*, *nirK* or *amoA* genes hybridized well with their respective homologous target DNA, suggesting that a broad range of detection can be achieved by adjusting the conditions for microarray hybridization. These results indicate that specific hybridization can be achieved using the glass slide-based microarray format with bulk community DNA extracted from environmental samples.

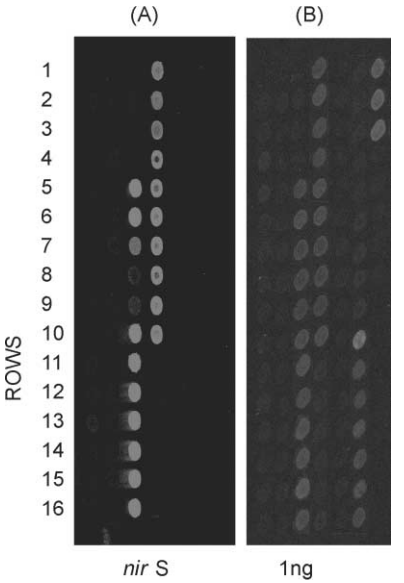


Figure 8 Specificity and sensitivity of DNA fragments-based FGAs. (A) Fluorescence images showing the specificity of *nir*S in DNA microarray hybridization. Target DNA was labeled with either Cy3 from a pure culture using the method of PCR amplification and hybridized separately at high stringency (65°C) to FGAs containing *nir*S, *nir*K, and *amo*A gene probes from both pure bacterial cultures and environmental clones. The 16S rRNA and yeast genes served as positive and negative controls, respectively. (B) Array hybridization images showing the detection sensitivity with labeled pure genomic DNA Genomic DNA from a pure culture of *nir*S-containing *P. stutzeri* E4-2 was labeled with Cy5 using the random primer labeling method. The target DNA was hybridized to the microarrays at total concentrations of 0.5, 1, and 5 ng.

To determine the potential performance of oligonucleotide microarrays for environmental studies, an FGA consisting of 50-mer oligonucleotide probes was constructed and evaluated using 1033 genes involved in nitrogen cycling (*nir*S, *nir*K, *nif*H, *amo*A, and *pmo*A) and sulfite reduction (*dsr*A/B) from public databases and our own sequence collections (Tiquia *et al*, unpublished). Under hybridization conditions of 50°C and 50% formamide, genes having $<86-90\%$ sequence identity were clearly distinguished. As expected, the oligonucleotide-based FGAs showed a higher degree of hybridization specificity than the DNA-based FGAs. Comparison of probe sequences from pure cultures of bacteria involved in nitrification, denitrification, nitrogen fixation, methane oxidation and sulfate reduction indicated that the average similarity of these functional genes at the species level ranged from 74 to 84%. These results suggest that the 50-mer FGAs could provide species-level resolution for analyzing microorganisms involved in these biogeochemical processes.

Compared to DNA-based FGAs, the 50-mer oligonucleotide arrays offer the following main advantages. First, higher hybridization specificity can be achieved because the probe sizes in oligo arrays are much smaller than those used in the DNA-based FGAs. Thus, this type of environmental array could provide a higher level of resolution in differentiating microbial populations. Since the probes can be directly designed and synthesized based on sequence information from public databases, construction of oligonucleotide arrays is much easier than that of the DNA-based FGAs. To construct microarrays containing large DNA fragments, the probes used for array fabrication are generally amplified by PCR from environmental clones or from pure genomic DNA. However, obtaining all the diverse environmental clones and bacterial strains from various sources as templates for amplification can be a considerable challenge. As a result, the construction of comprehensive microarrays representing all functional genes of interest is not practically feasible. With oligo arrays, a great number of genes can easily be arrayed for comprehensive survey of the populations and activities of diverse microbial communities in the environment. In addition, since no PCR amplification is involved in oligonucleotide microarray fabrication, potential cross-contamination due to PCR amplification is minimized.

3. Sensitivity

Sensitivity is another critical parameter that impacts the effectiveness of microarray-based detection of microorganisms. The detection sensitivity of hybridization with a prototype DNA-based FGA was determined using genomic DNA from both pure cultures and soil community samples. At high stringency, strong hybridization signals were observed with 5 ng of DNA for both *nir*S and SSU rRNA genes, whereas hybridization signals were weaker but detectable with 1 ng of DNA (Fig. 8B). The hybridization signals at low DNA concentrations were stronger for SSU rRNA genes than for *nir*S genes. Hybridization signals derived from 0.5 ng of genomic DNA were measurable, but the fluorescence intensity was poor. As a result, the detection limit was estimated to be approximately 1 ng with randomly labeled pure genomic DNA under the tested hybridization conditions.

The detection sensitivity of FGA hybridization was also evaluated using community genomic DNA isolated from surface soil that contained a high level of chromium and organic matter. All of the arrayed genes, with the exception of the five yeast genes, showed hybridization with 50 and 25 ng of labeled community DNA. Only the SSU rRNA genes could be detected when as little as 10 ng of the soil community DNA was used in the hybridization reaction. Thus, the detection sensitivity of *nir*S and SSU rRNA genes in this soil sample was considered to be approximately 25 and 10 ng of the total environmental DNA, respectively. These approximate levels of detection sensitivity should be

sufficient for many studies in microbial ecology and suggest that microarray hybridization can be used as a sensitive tool for analyzing microbial community composition in environmental samples.

The detection limit with 50-mer FGAs was approximately 8 ng of pure genomic DNA. As expected, the sensitivity of the 50-mer FGAs is 10 times lower than the DNA-based FGAs and 100 times lower than CGAs (Tiquia *et al.*, unpublished; Zhou, 2003), which are discussed below. One of the main reasons for the lower sensitivity of the 50-mer FGAs is that the oligonucleotide probes are much shorter than the probes used in DNA-based FGAs and CGAs, which have more binding sites available for capturing the labeled target DNAs. In addition, good hybridizations were obtained with the 50-mer FGAs using 2 μg of bulk community DNA from marine sediments. These results suggest that the amount of DNA sample should not be a major limiting factor in using this type of microarray for environmental studies, because the average DNA yields generally range from 10 to 400 μg of DNA per g (dry weight) for many surface soil and sediment samples.

Although sensitive detection can be obtained with microarray hybridization, the detection sensitivity is dependent on reagents, especially the fluorescent dyes. We found that the sensitivity varies greatly with different batches of fluorescent dyes. In addition, the sensitivity with direct microarray hybridization may still be 100 to 10,000-fold less than with PCR amplification. Microarray hybridization is still not sensitive enough for some environmental studies where the amount of recoverable biomass is very low, thus requiring the development of more sensitive methods.

4. Quantitation

Many environmental and ecological studies require quantitative data on the *in situ* abundance and biological activities of microbial communities. The accuracy of microarray-based quantitative assessment is still uncertain because of the inherently high variation associated with array fabrication, probe labeling, hybridization, and image processing. Comparison of microarray hybridization results with previously known results suggested that microarray hybridization appears to be quantitative enough for detecting differences in gene expression patterns under various conditions (DeRisi *et al.*, 1997; Lockhart *et al.*, 1996; Taniguchi *et al.*, 2001). DNA microarrays have also been used to measure differences in DNA copy number in breast tumors (Pinkel *et al.*, 1998; Pollack *et al.*, 1999) and to detect single-copy deletions or additions (Pollack *et al.*, 1999), suggesting that microarray-based detection is potentially quantitative. A recent study in which lambda (λ) DNA was co-spotted with DNA from reference

bacterial strains also indicated that microarrays could accurately quantify genes in DNA samples (Cho and Tiedje, 2001).

To evaluate whether microarray hybridization can be used as a quantitative tool to analyze environmental samples, the relationship between target DNA concentration and hybridization signal was examined with DNA-based FGAs (Wu *et al*., 2001). A strong linear relationship ($r^2 = 0.96$) was observed between signal intensity and target DNA concentration with DNA from a pure bacterial culture within 1 to 100 ng. Similar to the DNA-based FGAs, a strong linear relationship was observed using the 50-mer oligonucleotide FGAs between signal intensity and target DNA concentrations from 8 to 1000 ng for all six different functional gene groups ($r^2 = 0.96–0.98$) (Tiquia *et al*., unpublished). These results suggest that microarray hybridization is quantitative for pure bacterial cultures within a limited range of DNA concentration. With our optimized protocol, experimental variation between array slides can be reduced to below 15% with environmental samples (Wu *et al*., 2001). This is consistent with the findings of microarray studies on gene expression (Bartosiewicz *et al*., 2000).

Since environmental samples contain a mixture of target and non-target templates, the presence of other non-target templates could affect microarray-based quantification. To determine whether microarray hybridization is quantitative for targeted templates within the context of environmental samples, 11 different genes, exhibiting less than 80% sequence identity, were labeled and hybridized with the DNA-based FGAs. For this mixed DNA population, a linear relationship ($r^2 = 0.94$) was observed between signal intensity and target DNA concentration (Fig. 9), further suggesting that microarray hybridization holds promise as a quantitative tool for studies in environmental microbiology.

The target genes within functional groups present in environmental samples may have different degrees of sequence divergence. Such sequence differences will affect microarray hybridization signal intensities and hence its quantitative power. Although it was shown that microarray hybridization could be used to quantify mixed DNA templates, the difficult challenge in quantifying the abundance of microbial populations in natural environments, based on hybridization signal intensity, is how to distinguish differences in hybridization intensity due to population abundance from those due to sequence divergence. One possible solution is to carry out microarray hybridization under conditions of varying stringency. Based on the relationships among signal intensity, sequence divergence, hybridization temperature, and washing conditions, it should be possible to distinguish, to some extent, the contributions of population abundance and sequence divergence to hybridization intensity (Wu *et al*., 2001). For instance, Wu *et al*. (2001) showed that at about 55–60°C, sequence divergence had little or no effect on signal intensity for *amo*A genes with greater than 80% identity to the labeled target DNA. This suggests that under such hybridization conditions the effect of sequence divergence on signal intensity is negligible for genes with $>80\%$ sequence identity; therefore, any significant differences in

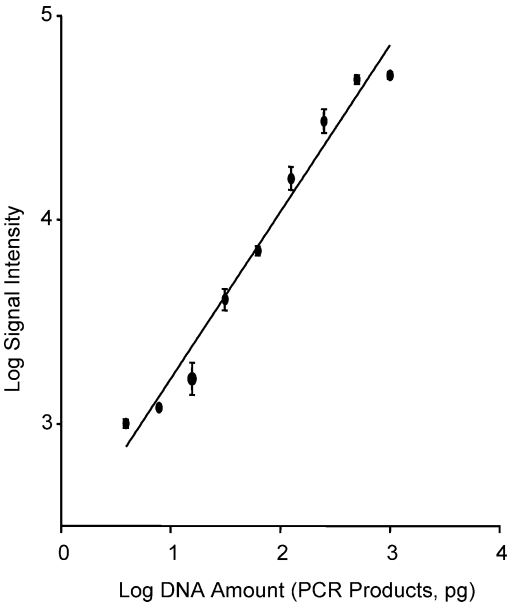


Figure 9 Relationship of hybridization signal intensity to DNA target concentration using a mixture of target DNAs. The PCR products from the following nine strains were mixed together in different quantities (pg): E4-2 (*nirS*), 1000; G179 (*nirK*), 500; wc301-37 (*amoA*), 250; ps-47 (*amoA*), 125; pB49 (*nirS*), 62.5; Y32K (*nirK*), 31.3; wA15 (*nirS*), 15.6; ps-80 (*amoA*), 7.8; wB54 (*nirK*), 3.9. All of these genes are less than 80% identical. The mixed templates were labeled with Cy5. The plot shows the log-transformed average hybridization intensity *versus* the log-transformed target DNA concentration for each strain. The target DNA was prepared by labeling MR-1 genomic DNA with Cy5 using Klenow fragment with random hexamer primers. The data points are mean values derived from three independent microarray slides, with three replicates on each slide (nine data points). Error bars showing the SD are presented.

signal intensity are most likely due to differences in population abundance. Another possible solution to this problem is to use microarrays containing probes that are extremely specific to the target population of interest, such as those used in oligonucleotide microarrays.

5. Applications

FGAs for microbial detection are still in the developmental stages, and thus their applications are still being explored. To demonstrate the applicability of DNA microarrays for microbial community analysis, Wu *et al*. (2001) used FGAs to analyze the distribution of denitrifying and nitrifying microbial populations in marine sediment and soil samples. The prototype FGA revealed differences in

the apparent distribution of *nir*S, *nir*K and *amo*A/*pmo*A gene families in sediment and soil samples. Recently, a 70-mer oligonucleotide microarray containing 64 *nir*S genes (14 from cultured microorganisms and 50 from environmental clones) was evaluated for studying functional gene diversity in the Choptank River-Chesapeake Bay system (Taroncher-Oldenburg *et al*., 2003). Significant differences in the hybridization patterns were observed between the sediment samples from two stations in the Choptank River. The changes in the *nir*S-containing denitrifier population could have been caused by differences in salinity, inorganic nitrogen and dissolved organic carbon between these two stations.

So far, very limited studies have been carried out to evaluate specificity, sensitivity, sequence divergence and quantitation of DNA microarrays for environmental applications. While this tool is potentially valuable for environmental studies, more development is needed, especially for improved sensitivity, quantitation, and the biological meaning of a detectable specificity before it can be used broadly and interpreted meaningfully within the context of microbial ecology.

B. Phylogenetic Oligonucleotide Arrays

Ribosomal RNA genes are powerful molecules for studying phylogenetic relationships among different organisms and for analyzing microbial community structure in natural environments, because these genes exist in all organisms and contain both highly conserved and highly variable regions, which are useful for differentiating microorganisms at different taxonomic levels (e.g., kingdom, phyla, family, genus, species, and strain). A very large database of ribosomal RNA genes exists (http://www.cme.msu.edu), making them ideal molecules for developing microarray-based detection tools. In addition, cells generally have multiple copies of rRNA genes, and the majority ($>95\%$) of total RNA isolated from samples is rRNA. Consequently, the detection sensitivity will be higher for rRNA genes than for functional genes. Therefore, rRNA genes are very useful targets for developing microarray-based detection approaches.

Oligonucleotide microarrays containing information from rRNA genes are referred to as phylogenetic oligonucleotide microarrays (POAs), because such microarrays are used primarily for phylogenetic analysis of microbial communities. The POAs can be constructed for different phylogenetic taxa and used in microbial community analysis studies. The oligonucleotide probes can be designed in a phylogenetic framework to survey different levels of sequence conservation, from highly conserved sequences giving broad taxonomic groupings to hypervariable sequences giving genus- and potentially species- level groupings. Because highly conserved universal primers for amplifying rRNA

genes are available, POA-based hybridization can be easily coupled with PCR amplification, thus enabling the implementation of highly sensitive assays.

1. Challenges of Plylogenetic Oligonucleotide Arrays

Non-rRNA gene-based oligonucleotide microarrays have been used successfully for monitoring genome-wide gene expression (*e.g.*, Lockhart *et al.*, 1996; de Saizieu *et al.*, 1998) and detecting genetic polymorphisms (*e.g.*, Wang *et al.*, 1998). In contrast, rRNA gene-based oligonucleotide arrays present some unique technical challenges concerning hybridization specificity and sensitivity (Zhou and Thompson, 2002; Zhou, 2003).

Specificity. Since rRNA genes are highly conserved and present in all microorganisms, specific detection with rRNA-targeted oligonucleotide microarrays can be difficult. First, the probe length and G + C content can significantly impact microarray hybridization (Guschin *et al.*, 1997a). Second, probe selection is limited by the sequence differences among target genes, and cross-hybridization can be a problem. Oligonucleotide microarrays typically contain many probes. Ideally, all of the oligonucleotides should have similar or identical melting kinetics, so that all of the probes on an array element can be subjected to the same hybridization conditions at once. This can be difficult to achieve, because the melting temperature depends on the length and composition of the oligonucleotide probe as well as the target 16S rRNA molecules in the samples.

Secondary structure. The hybridization of oligonucleotide probes to target nucleic acids possessing stable secondary structures can be particularly challenging, since low stringency conditions (i.e., hybridization temperatures between 0–30°C) are required for stable association of a long target nucleic acid with a short immobilized oligonucleotide probe (Drobyshev *et al.*, 1997; Guschin *et al.*, 1997a, b; Southern *et al.*, 1999). Any stable secondary structure of the target DNA or RNA must be overcome in order to make complementary sequence regions available for duplex formation. The stable secondary structure of SSU rRNA will have serious effects on hybridization specificity and detection sensitivity.

2. Specificity and Sensitivity

In a study by Guschin *et al.* (1997a), gel-pad oligonucleotide microarrays were constructed using oligonucleotides complementary to SSU rRNA sequences from key genera of nitrifying bacteria. The results showed that specific detection could be achieved with this type of microarray. However, the probe specificity depends on various factors, such as probe length. Guschin *et al.* (1997a) showed that, as

the length of the oligonucleotide probe increases, mismatch discrimination is lost; conversely, as the length of the probe decreases, hybridization signal intensity (i.e., sensitivity) is sacrificed. A recent study showed that gel-pad-based oligonucleotide microarrays could also be used to distinguish between *B. thuringiensis* and *B. subtilis* (Bavykin *et al*., 2001). Using glass-based two-dimensional microarrays, Small *et al*. (2001) detected such metal-reducing bacteria as *Geobacter chapellei* and *Desulfovibrio desulfuricans*.

The potential advantage of oligonucleotide probes is that target sequences containing single-base mismatches can be differentiated by microarray hybridization. However, this has not been fully demonstrated with SSU rRNA gene-based probes. To systematically determine whether single mismatch discrimination can be achieved for SSU rRNA genes using microarray hybridization, we constructed a model oligonucleotide microarray consisting of probes derived from three different regions of the SSU rRNA molecule corresponding to different bacterial taxa (X. Zhou and J. Zhou, unpublished data). The probes had 1–5 mismatches in different combinations along the length of the oligonucleotide probe with at least one mismatch at the central position. Hybridization signal intensity with a single-base mismatch was decreased by 10 to 30%, depending on the type of mismatched nucleotide base. The signal intensity of probes with two base mismatches was 5 to 25% of that of the perfect match probes. Probes with three or four base-pair mismatches yielded signal intensities that were 5% of that of the perfect match probes. Maximum discrimination and signal intensity was achieved with 19-base probes. These results indicated that single base discrimination for SSU rRNA genes can be achieved with glass slide-based array hybridization, but complete discrimination appears to be problematic with SSU rRNA genes (Bavykin *et al*., 2001; Small *et al*., 2001; Urakawa *et al*., 2002). Urakawa *et al*. (2002) demonstrated that the single-base-pair near-terminal and terminal mismatches have a significant effect on hybridization signal intensity. With SSU rRNA gene-based oligonucleotide microarrays, the level of detection sensitivity obtained using the *G. chapellei* 16S rRNA gene is about 0.5 μg of total RNA extracted from soils (Small *et al*., 2001).

3. Applications

As with all the arrays developed for environmental applications, SSU rRNA gene-based oligonucleotide arrays are still in the early stages of development, and therefore, only a few studies have applied POAs to the analysis of microbial structure within the context of environmental samples. Using photolithography-based Affymetrix technology, Wilson *et al*. (2002) designed a gene chip (microarray) containing 31,179 and 20-mer oligonucleotide probes specific for SSU rRNA genes. All of the probes are derived from a small SSU rRNA gene

region (i.e., *E. coli* positions 1409 to 1491), which is bound on both ends by universally conserved segments. The gene chip also contained control sequences, which were paired with the probe sequences. A control sequence was identical to the paired probe sequence except that there was a mismatch nucleotide at the 11th position. Thus, the gene chip contained a total of 62,358 features. The number of probes for individual sequences contained in the Ribosomal Database Project (RDP version 5.0, with about 3200 sequences) ranges from 0 to 70. A total of 17 pure bacterial cultures were used to assess the performance of this gene chip, and 15 bacterial species were identified correctly. However, it failed to resolve the individual sequences comprising complex mixed samples (Wilson *et al*., 2002).

Rudi *et al*. (2000) constructed a small microarray containing 10 SSU rRNA probes derived from cyanobacteria, and used it to analyze the presence and abundance of these organisms in lakes with both low and high biomass. The probes were specific to the cultures analyzed, and reproducible abundance profiles were obtained with these lake samples. Relatively good qualitative correlations were observed between the community diversity and standard hydrochemical data, but the levels of correlation were lower for the quantitative data.

Loy *et al*. (2002) developed a microarray containing 132 SSU rRNA-targeted oligonucleotide probes, which represented all recognized groups of sulfate-reducing prokaryotes. Microarray hybridizations with 41 reference strains showed that, under the hybridization conditions used, clear discrimination between perfectly matched and mismatched probes were obtained for most, but not all of the 132 probes. This microarray was used to determine the diversity of sulfate-reducing prokaryotes in periodontal tooth pockets and a hypersaline cyanobacterial mat. The microarray hybridization results were consistent with those obtained using well-established conventional molecular methods. These results suggest that microarray hybridization is a powerful tool in analyzing community structure but great caution is needed in data interpretation because of the potential for cross-hybridization.

C. Community Genome Arrays

Decades of scientific investigations have led to the isolation of many microorganisms from a variety of natural habitats. However, little or nothing is known about the genomic sequences for the majority of these microorganisms. Such a large collection of pure cultures should be very useful for monitoring microbial community structure and composition in natural environments. To exploit such a resource, a novel prototype microarray containing whole genomic DNA, termed community genome array (CGA), was developed and evaluated in my (Zhou's) laboratory.

The CGA is conceptually analogous to membrane-based reverse sample genome probing (RSGP) (Voordouw, 1998), but CGA hybridization is distinctly different from RSGP in terms of the arraying substrate and signal detection strategies. In contrast to RSGP, the CGA uses a non-porous (i.e., glass) surface for fabrication and fluorescence-based detection. The capability of accurate and precise miniaturization with robots on non-porous substrates is one of the two key advances of microarray-based genomic technologies. The miniaturized microarray format coupled with fluorescent detection represents a fundamental revolution in biological analysis. Like RSGP, the main disadvantage of the CGA is that only the cultured components of a community can be monitored, because the construction requires the availability of individual pure isolates, even though CGA-based hybridization itself does not require culturing (Voordouw, 1998). With the recent advances in environmental genomics, high-molecular-weight DNA from uncultivated microorganisms could be accessed through bacterial artificial chromosomes (BACs). BAC clones could also be used to fabricate CGAs, thus allowing the investigation of uncultivated components of a complex microbial community. In the following sections, we will briefly describe the performance of CGA-based hybridization in terms of specificity, sensitivity and quantitation.

1. Specificity

To examine hybridization specificity under varying experimental conditions and to determine the threshold levels of genomic differentiation, a prototype microarray was fabricated that contained genomic DNA isolated from 67 different representative environmental microorganisms classified as α-, β-, and γ-proteobacteria and Gram-positive bacteria. Many of the selected species are closely related to each other based on SSU rRNA and *gyr* B gene phylogenies and belong primarily to three major bacterial genera (*Pseudomonas*, *Shewanella*, and *Azoarcus*). The G + C content of the genomes varies from 37 to 69.3%. By adjusting hybridization temperature and the concentration of additives such as formamide (which increases hybridization stringency), different threshold levels of phylogenetic differentiation could be achieved using the CGAs. For instance, under hybridization conditions of 55°C and 50% formamide, strong signals were obtained for genomic DNAs of corresponding species to the labeled target. Little or no cross-hybridization (~0–4%) was observed for non-target species as well as for negative controls (yeast genes), thus indicating that species-specific differentiation can be achieved with CGAs under the hybridization conditions used. However, different strains of *Pseudomonas stutzeri*, *Azoarcus tolulyticus*, *Bacillus methanolicus*, and *Shewanella algae* could not be clearly distinguished under these conditions (Wu *et al*., unpublished). By further increasing

hybridization temperature (65 and 75°C), strain-level differentiation was obtained for closely related *Azoarcus* strains (Wu *et al*., unpublished).

Due to the complicated nature of microarray hybridization, it is unlikely that such assays will completely eliminate some degree of hybridization to non-target strains. The central question is how to distinguish true hybridization signals from non-specific background noise. One common approach is to determine SNRs and discard values below certain threshold value. Our studies showed that the average SNR for hybridizations with different species within a genus is about 3.35 ± 0.32, which is substantially lower than hybridizations with different strains from the same species. This value is very close to the commonly used threshold value (SNR = 3.0).

CGAs could be used to determine the genetic distance between different bacteria at the taxonomic levels of species and strain. Significant linear relationships were observed between CGA hybridization ratios and sequence similarity values derived from SSU rRNA and *gyr*B genes, DNA-DNA reassociation, or REP- and BOX-PCR fingerprinting profiles ($r^2 = 0.80 - 0.95$) (Wu *et al*., unpublished), suggesting that CGAs could provide meaningful insights into relationships between closely related strains. Because of its high capacity, one can construct CGAs containing bacterial type strains plus appropriately related strains. By hybridizing genomic DNA from unknown strains with this type of microarray, one should be able to quickly and reliably identify unknown strains provided a suitably related probe is on the array. When using CGAs for strain identification, less stringent hybridization conditions (e.g., 45°C and 50% formamide) should be used first to ensure that good hybridization signals can be obtained for distantly related target species. If multiple probes have significant hybridization with the unknown target strains, highly stringent hybridization conditions should then be used.

Compared to the traditional DNA–DNA reassociation approach, CGAs have several advantages for determining species relatedness. Since many bacterial genomes can be deposited on microarray slides, the tedious and laborious pair-wise hybridizations associated with the traditional DNA–DNA reassociation approach among different species are not needed with CGAs. In contrast to the traditional DNA–DNA reassociation approach, which generally requires about 100 μg DNA, CGA-based hybridization requires only about 2 μg of genomic DNA. This is important for determining the relationships between bacterial species that are recalcitrant to cultivation or grow very slowly.

2. Sensitivity and Quantitative Potential

To determine the detection sensitivity of CGAs, genomic DNA from a pure bacterial culture was fluorescently labeled and hybridized with the CGA at

different concentrations. Under stringent hybridization conditions (i.e., 65°C), the detection limit with randomly labeled pure genomic DNA was estimated to be approximately 0.2 ng, whereas genomic DNA concentrations of 0.1 ng were barely detectable above background levels (Wul *et al.*, unpublished). The level of CGA detection sensitivity should be sufficient for many studies in microbial ecology. The detection sensitivity was approximately 10-fold higher than that of DNA-based FGAs and about 100 times higher than that of the 50-mer FGAs. These results were expected, because the CGA probes represent entire genomes rather than a single gene.

The capacity of CGA hybridization to serve as a quantitative tool was explored by examining the relationship between the concentration of labeled target DNA and hybridization signal intensity. Quantitative potential was determined using labeled genomic DNA from a single pure culture and from 16 targeted bacteria representing different genera and species. In both cases, strong linear relationships between fluorescence intensity and DNA concentration were observed within a certain range of concentrations ($r^2 = 0.92 - 0.95$) (Wu *et al.*, unpublished). The results indicate that CGAs can be used for quantitative analysis of microorganisms in environmental samples. The quantitative feature of CGA is similar to those of the DNA- and oligonucleotides-based FGAs (Wu *et al.*, 2001; Tiquia *et al.*, unpublished).

D. Whole-genome Open Reading Frame Arrays for Revealing Genome Differences and Relatedness

Many microorganisms that are closely related based on SSU rRNA gene sequences show dramatic differences in phenotypic characteristics. One way to understand the genetic basis for such phenotypic differences is to obtain whole-genome sequence information for all closely related species of interest. Patterns of sequence similarity and variability will provide insights on the conservation of gene functions, physiological plasticity and evolutionary processes. However, sequencing the entire genomes of all closely related species is expensive and time-consuming. In addition, it may not be necessary to sequence all closely related genomes once the complete genome sequence for one representative microorganism is available, because substantial portions of the genomic sequence will be common among closely related species. One way to circumvent the need for sequencing multiple genomes of closely related species is to use DNA microarrays containing individual ORFs of a sequenced microorganism to view genome diversity and relatedness of other closely related microorganisms.

The whole-genome ORF array-based hybridization approach has been used to reveal genome diversity and relatedness among closely related organisms in

several studies. Murray *et al*. (2001) used this approach to evaluate the genome diversity and relatedness of several related metal-reducing bacteria within the *Shewanella* genus using partial ORF microarrays for the sequenced metal-reducing bacterium, *S. oneidensis* MR-1. Both conserved and poorly conserved genes were identified among the nine species tested. Under the conditions used in this study, the hybridization results were most informative for the closely related organisms with SSU rRNA sequence similarities greater than 93% and *gyrB* sequence similarities greater than 80%. Above this level of homology, the similarities of microarray hybridization profiles were strongly correlated with *gyrB* sequence divergence. In addition, most genes in operons had high levels of DNA relatedness, suggesting that this approach can be used to identify genes or operons that were horizontally transferred (Murray *et al*., 2001).

Using the ORF arrays for *E. coli* K-12, Dong *et al*. (2001) identified the genes in a common endophyte of maize, *Klebsiella pneumoniae* 342, which is closely related to *E. coli*. About 3000 (70%) of *E. coli* genes were found in strain 342 with greater than 55% identity, whereas about 24% of the *E. coli* genes were absent in strain 342. The genes with high sequence identity were those involved in cell division, DNA replication, transcription, translation, transport, regulatory proteins, energy, amino acid and fatty acid metabolism, and cofactor synthesis, whereas the genes that are less conserved were involved in carbon compound metabolism, membrane proteins, structural proteins, central intermediary metabolism, and proteins involved in adaptation and protection. Genes that were not identified in strain 342 included putative regulatory proteins, putative chaperones, surface structure proteins, mobility proteins, putative enzymes and hypothetical proteins. These results on genomic diversity are consistent with the physiological properties of these two strains, suggesting that the microarray-based whole-genome comparison is a powerful approach to revealing the genomic diversity and relatedness of closely related organisms.

The whole-genome ORF array approach was also successfully used to identify genome differences among 15 *Helicobacter pylori* strains with more and less virulence (Salama *et al.*, 2000) and to detect the deletions existing in other strains of *Mycobacterium tuberculosis* and *M. bovis* (Behr *et al*., 1999). All of these studies suggest that whole-genome ORF arrays will be useful for revealing genome difference and relatedness. Whole-genome ORF arrays are available from many microorganisms and they will be valuable for studying genome diversity and relatedness of closely related microorganisms. For example, the whole-genome arrays for six environmentally important microorganisms, including *S. oneidensis* MR-1, *D. radiodurans* R1, *Rhodopseudomonas palustris*, *Nitrosomonas europaea*, *Desulfovibrio vulgaris,* and *Geobacter metallireducens* are available at Oak Ridge National Laboratory, and we are also currently using these whole-genome ORF arrays to understand the genome diversity and relatedness of some important environmental isolates.

E. Other Types of Microarrays for Microbial Detection and Characterization

DNA microarrays containing random genomic fragments have been used to determine species relatedness in instances where genome sequence information is not available. In this approach, 60–96 genomic fragments of about 1 kb were randomly selected from four fluorescent *Pseudomonas* species as reference genomes for microarray fabrication (Cho and Tiedje 2001). Cluster analysis of hybridization profiles from 12 well-characterized fluorescent *Pseudomonas* species indicated that such types of microarray hybridization could provide species to strain level resolution. This approach could have higher resolution than CGA because extensive component information is obtained rather than an average for the whole-genome. However, this approach is more time-consuming and costly to develop than CGA and such an array would be more limited in scope since many of the array positions would be used for each reference microorganism (L. Wu, personal communication).

Recently, a random nonamer oligonucleotide microarray was developed and evaluated for obtaining fingerprinting profiles among closely related strains instead of using a gel electrophoresis-based method (Kingsley *et al*., 2002). A prototype array containing 47 randomly selected nonamer oligonucleotides was constructed and used to differentiate 14 closely related *Xanthomonas* strains. The REP-PCR was first carried out to obtain the fingerprints from different strains, then the amplified REP-PCR products were hybridized with the nonamer array, and fingerprinting profiles for each strain were obtained based on microarray hybridization. The results showed that the microarray-based fingerprinting methods provide clear resolution among all strains examined, including two strains (*X. oryzae* 43836 and 49072) which could not be resolved using traditional gel electrophoresis of REP-PCR amplification methods. This suggests that the microarray hybridization-based approach could provide higher resolution in strain differentiation than the conventional gel electrophoresis-based fingerprinting approach. This approach is attractive because a universal nonamer array can be developed to generate fingerprints from any microorganisms.

IX. CONCLUDING REMARKS

Microarray is a recently developed functional genomics technology that has powerful applications in a wide array of biological research areas, including the medical sciences, agriculture, biotechnology and environmental studies. Since many universities, research institutions and industries have established microarray-based core facilities and services, microarrays have become a readily accessible, widely used technology for investigating biological systems. As the

technologies for array instrumentation are relatively mature, major trends are emerging in such issues as novel array platforms, attachment strategies and substrates, miniaturization with higher density, novel labeling strategies, scanning technologies and automation (Constans, 2003; Stears *et al*., 2003). Besides the DNA-based array assay, the microarray platform is also being rapidly expanded to include the analysis of other biomolecules such as proteins and carbohydrates (Stears *et al*., 2003). Along with exploration in microarray technology applications, novel strategies and approaches for experimental controls and design are needed to ensure that microarray hybridization data from different samples are comparable, interpretable and biologically significant because of the inherent variability in microarray hybridization signals. Finally, more advanced automatic mathematical and computational tools, such as multivariate analysis, time-series analysis, neural network, artificial intelligence, and differential equation-based modeling approaches, should be extremely useful for rapid pattern recognition, visualization, data mining, cellular modeling, simulation and prediction.

The development and application of microarray-based genomic technology for environmental studies has received a great deal of attention. Because of its high-density and high-throughput capacity, it is expected that microarray-based genomic technologies will revolutionize the analyses of microbial community structure, function and dynamics. Microarray-based assays have great potential as specific, sensitive, quantitative, parallel, and high-throughput tools for microbial detection, identification and characterization in natural environments. However, more rigorous and systematic assessment and development are needed to realize the full potential of microarrays for microbial ecology studies. Several key issues need to be addressed, including novel experimental designs and strategies for minimizing inherent high hybridization variations to improve microarray-based quantitative accuracy, novel approaches for increasing hybridization sensitivity to detect extremely low biomass in natural environments, novel computational tools for microarray data extraction and interpretation, and broad integration and application of microarray technologies with environmental studies to address ecological and environmental questions and hypotheses.

References

Adams, M. W. W., Holden, J. F., Menon, A. L., Schut, G. J., Grunden, A. M., Hou, C., Hutchins, A. M., Jenney, F. E. Jr., Kim, C., Ma, K., Pan, G., Roy, R., Sapra, R., Story, S. V., and Verhagen, M. F. (2001). Key role for sulfur in peptide metabolism and in regulation of three hydrogenases in the hyperthermophilic archaeon *Pyrococcus furiosus*. *J. Bacteriol.* **183**, 716–724.

Adler, K., Broadbent, J., Garlick, R., Joseph, R., Khimani, A., Mikulskis, A., Rapiejko, P., and Killian, J. (2000). MICROMAX™: A highly sensitive system for differential gene expression on microarrays. "*In* Microarray Biochip Technology" (M. Schena, Ed.), pp. 221–230. Eaton Publishing, Natick, MA.

Afanassiev, V., Hanemann, V., and Wolfl, S. (2000). Preparation of DNA and protein micro arrays on glass slides coated with an agarose film. *Nucl. Acids Res.* **28**, E66.

Anderson, A., Nordan, H., Cain, R., Parrish, G., and Duggan, D. (1956). Studies on a *radioresistant micrococcus*, I. Isolation, morphology, cultural characteristics, and resistance to gamma radiation. *Food Technol.* **10**, 575–578.

Augenlicht, L., Taylor, J., Anderson, L., and Lipkin, M. (1991). Patterns of gene expression that characterize the *colonic mucosa* in patients at genetic risk for colonic cancer. *Proc. Nat. Acad. Sci.* **88**, 3286–3289.

Augenlicht, L., Wahrman, M., Halsey, H., Anderson, L., Taylor, J., and Lipkin, M. (1987). Expression of cloned sequences in biopsies of human colonic tissue and in colonic–carcinoma cells induced to differentiate invitro. *Cancer Res.* **47**, 6017–6021.

Bains, W., and Smith, G. (1988). A novel method for nucleic acid sequence determination. *J. Theoret. Biol.* **135**, 303–307.

Baldi, P., and Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics* **17**, 509–519.

Bartosiewicz, M., Trounstine, M., Barker, D., Johnston, R., and Buckpitt, A. (2000). Development of a toxicological gene array and quantitative assessment of this technology. *Arch. Biochem. Biophys.* **376**, 66–73.

Basarsky, T., Verdnik, D., Zhai, J. Y., and Wellis, D. (2000). Overview of a microarray scanner: Design essentials for an integrated acquisition and analysis platform. *In* "Microarray Biochip Technology" (M. Schena, Ed.), pp. 265–284. Eaton Publishing, Natick, MA.

Bassett, D. Jr., Eisen, M., and Boguski, M. (1999). Gene expression informatics — it's all in your mine. *Nat. Genet. Sup.* **21**, 51–55.

Battaglia, C., Salani, G., Consolandi, C., Bernardi, L. R., and De, Bellis, G. (2000). Analysis of DNA microarrays by non-destructive fluorescent staining using SYBR green II. *Biotechniques* **29**, 78–81.

Battista, J. R. (1997). Against all odds: the survival strategies of *Deinococcus radiodurans*. *Annu. Rev. Microbiol.* **51**, 203–224.

Battista, J. R., Earl, A. M., and Park, M.-J. (1999). Why is *Deinococcus radiodurans* so resistant to ionizing radiation? *Trends Microbiol.* **7**, 362–365.

Bavykin, S. G., Akowski, J. P., Zakhariev, V. M., Barsky, V. E., Perov, A. N., and Mirzabekov, A. D. (2001). Portable system for microbial sample preparation and oligonucleotide microarray analysis. *Appl. Environ. Microbiol.* **67**, 922–928.

Beattie, K. L., Eggers, M. D., Shumaker, J. M., Hogan, M. E., Varma, R. S., Lamture, J. B., Hollis, M. A., Ehrlich, D. J., and Rathman, D. (1992). Genosensor technology. *Clin. Chem.* **39**, 719–722.

Beattie, W. G., Meng, L., Turner, S., Varma, R. S., Dao, D. D., and Beattie, K. L. (1995). Hybridization of DNA targets to glass-tethered oligonucleotide probes. *Mol. Biotechnol.* **4**, 213–225.

Behr, M. A., Wilson, M. A., Gill, W. P., Salamon, H., Schoolnik, G. K., Rane, S., and Small, P. M. (1999). Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science* **284**, 1520–1523.

Beier, M., and Hoheisel, J. (1999). Versatile derivatisation of solid support media for covalent bonding on DNA-microchips. *Nucl. Acids Res.* **27**, 1970–1977.

Belfort, M., and Weiner, A. (1997). Another bridge between kingdoms: tRNA splicing in Archaea and Eukaryotes. *Cell* **89**, 1003–1006.

Beliaev, A. S., Thompson, D. K., Fields, M. W., Wu, L., Lies, D. P., Nealson, K. H., and Zhou, J. (2002a). Microarray transcription profiling of a *Shewanella oneidensis etrA* mutant. *J. Bacteriol.* **184**, 4612–4616.

Beliaev, A. S., Thompson, D. K., Khare, T., Lim, H., Brandt, C. C., Li, G., Murray, A. E., Heidelberg, J. F., Giometti, C. S., Yates, III, J., Nealson, K. H., Tiedje, T. M., and Zhou, J. (2002b). Gene and protein expression profiles of *Shewanella oneidensis* during anaerobic growth with different electron acceptors. *OMICS* **6**, 39–60.

Ben-Dor, A., Shamir, R., and Yakhini, Z. (1999). Clustering gene expression patterns. *J. Comput. Biol.* **6**, 281–297.

Broude, N. E., Woodward, K., Cavallo, R., Cantor, C. R., and Englert, D. (2001). DNA microarrays with stem-loop DNA probes: preparation and applications. *Nucl. Acids Res.* **29**, E92.

Bult, C. J., White, O., Olsen, G. J., Zhou, L. X., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J. F., Adams, M. D., Reich, C. I., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., Glodek, A., Scott, J. L., Geoghagen, N. S. M., Weidman, J. F., Fuhrmann, J. L., Nguyen, D., Utterback, T. R., Kelley, J. M., Peterson, J. D., Sadow, P. W., Hanna, M. C., Cotton, M. D., Roberts, K. M., Hurst, M. A., Kaine, B. P., Borodovsky, M., Klenk, H. P., Fraser, C. M., Smith, H. O., Woese, C. R., and Venter, J. C. (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273**, 1058–1073.

Chee, M., Yang, R., Hubbell, E., Berno, A., Huange, X. C., Stern, D., Winkler, J., Lockhart, D. J., Morris, M. S., and Fodor, S. P. A. (1996). Accessing genetic information with high-density DNA arrays. *Science* **274**, 610–614.

Chen, Y., Dougherty, E. R., and Bittner, M. L. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics.* **2**, 364–374.

Cho, J. C., and Tiedje, J. M. (2001). Bacterial species determination from DNA–DNA hybridization by using genome fragments and DNA microarrays. *Appl. Environ. Microbiol.* **67**, 3677–3682.

Chrisey, L. A., Lee, G. U., and O'Ferrall, C. E. (1996). Covalent attachment of synthetic DNA to self-assembled monolayer films. *Nucl. Acids Res.* **24**, 3031–3039.

Clewell, D. R. (1993). Sex pheromones and the plasmid encoded mating response in *Enterococcus faecalis*. *In* "Bacterial conjugation" (D. B. Clewell, Ed.), pp. 349–367. Plenum Press, New York, NY.

Constans, A. (2003). Microarray instrumentation. *The Scientist* **17**, 37–38.

Daly, M. J., Ling, O., and Minton, K. W. (1994). Interplasmidic recombination following irradiation of the radioresistant bacterium *Deinococcus radiodurans*. *J. Bacteriol.* **176**, 7506–7515.

Daly, M. J., and Minton, K. W. (1995). Interchromosomal recombination in the extremely radioresistant bacterium *Deinococcus radiodurans*. *J. Bacteriol.* **177**, 5495–5505.

Dassy, B., and Fournier, J. M. (1996). Respiratory activity is essential for post-exponential-phase production of type 5 capsular polysaccharide by *Staphylococcus aureus*. *Infect. Immun.* **64**, 2408–2414.

de Saizieu, A., Certa, U., Warrington, J., Gray, C., Keck, W., and Mous, J. (1998). Bacterial transcript imaging by hybridization of total RNA to oligonucleotide arrays. *Nat. Biotechnol.* **16**, 45–48.

DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686.

Diehl, F., Grahlmann, S., Beier, M., and Hoheisel, J. D. (2001). Manufacturing DNA microarrays of high spot homogeneity and reduced background signal. *Nucleic Acids Res.* **29**, E38.

Dong, S., Wang, E., Hsie, L., Cao, Y., Chen, X., and Gingeras, T. R. (2001). Flexible use of high-density oligonucleotide arrays for single-nucleotide polymorphism discovery and validation. *Genome Res.* **11**, 1418–1424.

Dopazo, J., and Carazo, J. M. (1997). Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. *J. Molecul. Evol.* **44**, 226–233.

Dopazo, J., Zanders, E., Dragoni, I., Amphlett, G., and Falciani, F. (2001). Methods and approaches in the analysis of gene expression data. *J. Immunol. Meth.* **250**, 93–112.

Drmanac, R., Labat, I., Brukner, I., and Crkvenjakov, R. (1989). Sequencing of megabase-plus DNA by hybridization: Theory of the method. *Genomics* **4**, 114–128.

Drobyshev, A., Mologina, N., Shik, V., Pobedimskaya, D., Yershov, G., and Mirzabekov, A. (1997). Sequence analysis by hybridization with oligonucleotide microchip: identification of beta-thalassemia mutations. *Gene* **188**, 45–52.

Drobyshev, A. L., Zasedatelev, A. S., Yershov, G. M., and Mirzabekov, A. D. (1999). Massive parallel analysis of DNA-Hoechst 33258 binding specificity with a generic oligodeoxyribonucleotide microchip. *Nucl. Acids Res.* **27**, 4100–4105.

Dudoit, S. Y., Yang, H., Gallow, M. J., and Speed, T. P. (2001). Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments. *Statist Sincia* **12**, 111–139.

Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P., and Trent, J. M. (1999). Expression profiling using cDNA microarrays. *Nat. Genet.* **21**, 10–14.

Dunman, P. M., Murphy, E., Haney, S., Palacios, D., Tucker-Kellogg, G., Wu, S., Brown, E. L., Zagursky, R. J., Shlaes, D., and Projan, S. J. (2001). Transcription profiling-based identification of *Staphylococcus aureus* genes regulated by the *agr* and/or *sarA* loci. *J. Bacteriol.* **183**, 7341–7353.

Eckmann, L., Smith, J. R., Housley, M. P., Dwinell, M. B., and Kagnoff, M. F. (2000). Analysis of high density cDNA arrays of altered gene expression in human intestinal epithelial cells in response to infection with the invasive enteric bacteria *Salmonella*. *J. Biol. Chem.* **275**, 14084–14094.

Eggers, M. D., Hogan, M. E, Reich, R. K., Lamture, J., Ehrlich, D., Hollis, M., Kosicki, B., Powdrill, T., Beattie, K., Smith, S., Varma, R., Gangadharan, R., Mallik, A., Burke, B., and Wallace, D. (1994). A microchip for quantitative detection of molecular utilizing luminescent and radioisotope reporter groups. *Biotechniques.* **14**, 516–525.

Eisen, M., and Brown, P. (1999). DNA microarrays for analysis of gene expression. *Meth. Enzymol.* **303**, 179–205.

Englert, D. (2000). Production of microarrays on porous substrates using noncontact piezoelectric dispensing. *In* "Microarray Biochip Technology" (M. Schena, Ed.), pp. 231–246. Eaton Publishing, Natick, MA.

Evertsz, E., Starink, P., Gupta, R., and Watson, D. (2000). Technology and applications of gene expression microarrays. *In* "Microarray Biochip Technology" (M. Schena, Ed.), pp. 149–166. Eaton Publishing, Natick, MA.

Fiala, G., and Stetter, K. O. (1986). *Pyrococcus furiosus* sp. nov. represents a novel genus of marine heterotrophic archaebacteria growing optimally at 100°C. *Arch. Microbiol.* **145**, 56–61.

Fitz-Gibbon, S. T., Ladner, H., Kim, U.-J., Stetter, K. O., Simon, M. I., and Miller, J. H. (2002). Genome sequence of the hyperthermophilic crenarchaeon *Pyrobaculum aerophilum*. *Proc. Natl. Acad. Sci. USA* **99**, 984–989.

Fodor, S. P. A., Read, J., Pirrung, M., Stryer, L., Lu, A., and Solas, D. (1991). Light-directed, spatially addressable parallel chemical synthesis. *Science* **251**, 767–773.

Fotin, A., Drobyshev, A., Proudnikov, D., Perov, A., and Mirzabekov, A. (1998). Parallel thermodynamic analysis of duplexes on oligodeoxyribonucleotide microchips. *Nucl. Acids Res.* **26**, 1515–1521.

Fredrickson, J. K., Kostandarithes, H. M., Li, S. W., Plymale, A. E., and Daly, M. J. (2000). Reduction of Fe(III), Cr(VI), U(VI), and Tc(VII) by *Deinococcus radiodurans* R1. *Appl. Environ. Microbiol.* **66**, 2006–2011.

Galagan, J. E., Nusbaum, C., Roy, A., Endrizzi, M. G., Macdonald, P., FitzHugh, W., Calvo, S., Engels, R., Smirnov, S., Atnoor, D., Brown, A., Allen, N., Naylor, J., Stange-Thomann, N., DeArellano, K., Johnson, R., Linton, L., McEwan, P., McKernan, K., Talamas, J., Tirrell, A., Ye, W. J., Zimmer, A., Barber, R. D, Cann, I., Graham, D. E., Grahame, D. A., Guss, A. M., Hedderich, R., Ingram-Smith, C., Kuettner, H. C., Krzycki, J. A., Leigh, J. A., Li, W. X., Liu, J. F., Mukhopadhyay, B., Reeve, J. N., Smith, K., Springer, T. A., Umayam, L. A., White, O., White, R. H., de, Macario, E. C., Ferry, J. G., Jarrell, K. F., Jing, H., Macario, A. J. L., Paulsen, I., Pritchett, M., Sowers, K. R., Swanson, R. V., Zinder, S. H., Lander, E., Metcalf, W. W., and Birren, B. (2002). The genome of *M. acetivorans* reveals extensive metabolic and physiological diversity. *Genome Res.* **12**, 532–542.

Guo, Z., Guilfoyle, R. A., Thiel, A. J., Wang, R., and Smith, L. M. (1994). Direct fluorescence analysis of genetic polymorphisms by hybridization with oligonucleotide arrays on glass supports. *Nucl. Acids Res.* **22**, 5456–5465.

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

Guschin, D., Yershov, G., Zaslavsky, A., Gemmell, A., Shick, V., Proudnikov, D., Arenkov, P., and Mirzabekov, A. (1997b). Manual manufacturing of oligonucleotide, DNA, and protein microchips. *Anal. Biochem.* **250**, 203–11.

Hacia, J. G. (1999). Resequencing and mutational analysis using oligonucleotide microarrays. *Nat. Genet.* **21**, 42–47.

Hegde, P., Qi, R., Abernathy, K., Gay, C., Dharap, S., Gaspard, R., Hughes, J. E., Snesrud, E., Lee, N., and Quackenbush, J. (2000). A concise guide to cDNA microarray analysis. *Biotechniques* **29**, 548–560.

Heidelberg, J. F., Paulsen, I. T., Nelson, K. E., Gaidos, E. J., Nelson, W. C., Read, T. D., Eisen, J. A., Seshadri, R., Ward, N., Methe, B., Clayton, R. A., Meyer, T., Tsapin, A., Scott, J., Beanan, M., Brinkac, L., Daugherty, S., DeBoy, R. T., Dodson, R. J., Durkin, A. S., Haft, D. H., Kolonay, J. F., Madupu, R., Peterson, J. D., Umayam, L. A., White, O., Wolf, A. M., Vamathevan, J., Weidman, J., Impraim, M., Lee, K., Berry, K., Lee, C., Mueller, J., Khouri, H., Gill, J., Utterback, T. R., McDonald, L. A., Feldblyum, T. V., Smith, H. O., Venter, J. C., Nealson, K. H., and Fraser, C. M. (2002). Genome sequence of the dissimilatory metal ion-reducing bacterium *Shewanella oneidensis*. *Nature Biotechnol.* **20**, 1118–1123.

Heller, H. M., Schena, M., Chai, A., Shalon, D., Bedilion, T., and Gilmore, J. (1997). Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc. Natl. Acad. Sci. USA* **94**, 2150–2155.

Herrero, J., Valencia, A., and Dopazo, J. (2001). A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics* **17**, 126–136.

Herwig, R., Poustka, A. J., Muller, C., Bull, C., Lehrach, H., and O'Brien, J. (1999). Large-scale clustering of cDNA-fingerprinting data. *Genome Res.* **9**, 1093–1105.

Heyer, L. J., Kruglyak, S., and Yooseph, S. (1999). Exploring expression data: identification and analysis of coexpressed genes. *Genome Res.* **9**, 1106–1115.

Hilsenbeck, S. G., Friedrichs, W. E., Schiff, R., O'Connell, P., Hansen, R. K., Osborne, C. K., and Fuqua, S. A. (1999). Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. *J. Natl. Cancer Inst.* **91**, 453–459.

Hoheisel, J. (1997). Oligomer-chip technology. *Trends Biotech.* **15**, 465–469.

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M., and Friend, S. H. (2000). Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126.

Kalocsai, P., and Shams, S. (2001). Use of bioinformatics in arrays. *Meth. Mol. Biol.* **170**, 223–236.

Kane, M. D., Jatkoe, T. A., Stumpf, C. R., Lu, J., Thomas, J. D., and Madore, S. J. (2000). Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucl. Acids Res.* **28**, 4552–4557.

Kaneko, T., Nakamura, Y., Wolk, C. P., Kuritz, T., Sasamoto, S., Watanabe, A., Iriguchi, M., Ishikawa, A., Kawashima, K., Kimura, T., Kishida, Y., Kohara, M., Matsumoto, M., Matsuno, A., Muraki, A., Nakazaki, N., Shimpo, S., Sugimoto, M., Takazawa, M., Yamada, M., Yasuda, M., and Tabata, S. (2001). Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. *DNA Res.* **8**, 205–213.

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M., and Tabata, S. (1996). Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* **3**, 109–136.

Kawarabayasi, Y., Hino, Y., Horikawa, H., Yamazaki, S., Haikawa, Y., Jin-no, K., Takahashi, M., Sekine, M., Baba, S., Ankai, A., Kosugi, H., Hosoyama, A., Fukui, S., Nagai, Y., Nishijima, K., Nakazawa, H., Takamiya, M., Masuda, S., Funahashi, T., Tanaka, T., Kudoh, Y., Yamazaki, J., Kushida, N., Oguchi, A., Aoki, K., Kubota, K., Nakamura, Y., Nomura, N., Sako, Y., and Kikuchi, H. (1999). Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix K1*. *DNA Res.* **6**, 83–101.

Kawarabayasi, Y., Hino, Y., Horikawa, H., Jin-no, K., Takahashi, M., Sekine, M., Baba, S., Ankai, A., Kosugi, H., Hosoyama, A., Fukui, S., Nagai, Y., Nishijima, K., Otsuka, R., Nakazawa, H., Takamiya, M., Kato, Y., Yoshizawa, T., Tanaka, T., Kudoh, Y., Yamazaki, J., Kushida, N., Oguchi, A., Aoki, K., Masuda, S., Yanagii, M., Nishimura, M., Yamagishi, A., Oshima, T., and Kikuchi, H. (2001). Complete genome sequence of an aerobic thermoacidophilic crenarchaeon, *Sulfolobus tokodaii* strain7. *DNA Res.* **8**, 123–140.

Kawarabayasi, Y., Sawada, M., Horikawa, H., Haikawa, Y., Hino, Y., Yamamoto, S., Sekine, M., Baba, S., Kosugi, H., Hosoyama, A., Nagai, Y., Sakai, M., Ogura, K., Otsuka, R., Nakazawa, H., Takamiya, M., Ohfuku, Y., Funahashi, T., Tanaka, T., Kudoh, Y., Yamazaki, J., Kushida, N., Oguchi, A., Aoki, K., and Kikuchi, H. (1998). Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii*. *DNA Res.* **5**, 55–76.

Kawashima, T., Amano, N., Koike, H., Makino, S., Higuchi, S., Kawashima-Ohya, Y., Watanabe, K., Yamazaki, M., Kanehori, K., Kawamoto, T., Nunoshiba, T., Yamamoto, Y., Aramaki, H., Makino, K., and Suzuki, M. (2000). Archaeal adaptation to higher temperatures revealed by genomic sequence of *Thermoplasma volcanium*. *Proc. Natl. Acad. Sci. USA* **97**, 14257–14262.

Kerr, M. K., and Churchill, G. A. (2001a). Statistical design and the analysis of gene expression microarray data. *Genet. Res.* **77**, 123–128.

Kerr, M. K., and Churchill, G. A. (2001b). Experimental design for gene expression microarrays. *Biostatistics* **2**, 183–201.

Khodursky, A. B., Peter, B. J., Cozzarelli, N. R., Botstein, D., Brown, P. O., and Yanofsky, C. (2000). DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **97**, 12170–12175.

Khrapko, K. R., Lysov, Y. P., Khorlyn, A. A., Ivanov, I. B., Yershov, G. M., Vasilenko, S. K., Vlorentiev, V. L., and Mirzabekov, A. (1989). An oligonucleotide hybridization approach to DNA sequencing. *FEBS Lett.* **256**, 118–122.

Kingsley, M. T., Straub, T. M., Call, D. R., Daly, D. S., Wunschel, S. C., and Chandler, D. P. (2002). Fingerprinting closely related *Xanthomonas* pathovars with random nanamer oligonucleotide microarrays. *Appl. Environ. Microbiol.* **68**, 6361–6370.

Kleman-Leyer, K., Armbruster, D. W., and Daniels, C. J. (1997). Properties of *H. volcanii* tRNA intron endonuclease reveal a relationship between the archaeal and eucaryal tRNA intron processing systems. *Cell* **89**, 839–847.

Klenk, H.-P., Clayton, R. A., Tomb, J.-F., White, O., Nelson, K. E., Ketchum, K. A., Dodson, R. J., Gwinn, M., Hickey, E. K., Peterson, J. D., Richardson, D. L., Kerlavage, A. R., Graham, D. E., Kyrpides, N. C., Fleischmann, R. D., Quackenbush, J., Lee, N. H., Sutton, G. G., Gill, S., Kirkness, E. F., Dougherty, B. A., McKenney, K., Adams, M. D., Loftus, B., Peterson, S., Reich, C. I., McNeil, L. K., Badger, J. H., Glodek, A., Zhou, L. X., Overbeek, R., Gocayne, J. D., Weidman, J. F., McDonald, L., Utterback, T., Cotton, M. D., Spriggs, T., Artiach, P., Kaine, B. P., Sykes, S. M., Sadow, P. W., Dandrea, K. P., Bowman, C., Fujii, C., Garland, S. A., Mason, T. M.,

Olsen, G. J., Fraser, C. M., Smith, H. O., Woese, C. R., and Venter, J. C. (1997). The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* **394**, 364–370.

Knight, J. (2001). When the chips are down. *Nature* **410**, 860–861.

Lange, C. C., Wackett, L. P., Minton, K. W., and Daly, M. J. (1998). Engineering a recombinant *Deinococcus radiodurans* for organopollutant degradation in radioactive mixed waste environments. *Nat. Biotechnol.* **16**, 929–933.

Lamture, J. B., Beattie, K. L., Burke, B. E., Eggers, M. D., Ehrlich, D. J., Fowler, R., Hollis, M. A., Kosicki, B. B., Reich, R. K., and Smith, S. R. (1994). Direct detection of nucleic acid hybridization on the surface of a charge coupled device. *Nucl. Acids Res.* **22**, 2121–2125.

Lipshultz, R. J., Fodor, S. P. A., Gingeras, T. R., and Lockhart, D. J. (1999). High density synthetic oligonucleotide arrays. *Nat. Genet.* **21**, 20–24.

Liu, Y., Zhou, J., Omelchenko, M., Beliaev, A., Venkateswaran, A., Stair, J., Wu, L., Thompson, D. K., Xu, D., Rogozin, I. B., Gaidamakova, E. K., Zhai, M., Makarova, K. S., Koonin, E. V., and Daly, M. J. (2003). Transcriptome dynamics of *Deinococcus radiodurans* recovering from ionizing radiation. *Proc. Natl. Acad. Sci. USA* **100**, 4191–4196.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **14**, 1675–1680.

Long, A. D., Mangalam, H. J., Chan, B. Y., Tolleri, L., Hatfield, G. W., and Baldi, P. (2001). Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in *Escherichia coli* K12. *J. Biol. Chem.* **276**, 19937–19944.

Lovley, D. (1991). Dissimilatory Fe(III) and Mn(IV) reduction. *Microbiol. Rev.* **55**, 259–287.

Loy, A., Lahner, A., Lee, N., Adamczyk, J., Meier, H., Ernst, J., Schleifer, K. H., and Wagner, M. (2002). Oligonucleotide microarray for 16S rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment. *Appl. Environ. Microbiol.* **68**, 5064–5081.

Mace, M. L., Montagu, J., Rose, S. D., and McGuinnes, G. (2000). Novel microarray printing and detection technologies. *In* "Microarray Biochip Technology" (M. Schena, Ed.), pp. 39–64. Eaton Publishing, Natick, MA.

Makarova, K. S., Aravind, L., Wolf, Y. I., Tatusov, R. L., Minton, K. W., Koonin, E. V., and Daly, M. J. (2001). Genome of the extremely radiation-resistant bacterium *Deinococcus radiodurans* viewed from the perspective of comparative genomics. *Microbiol. Mol. Biol. Rev.* **65**, 44–79.

Martinsky, T., and Haje, P. (2000). Microarray tools, kits, reagents, and services. *In* "Microarray Biochip Technology" (M . Schena, Ed.), pp. 201–220. Eaton Publishing, Natick, MA.

McGall, G. H., and Fidanza, J. A. (2001). Photolithographic synthesis of high-density oligonucleotide arrays. *In* "DNA Arrays — Methods and Protocols" (J. B. Rampal, Ed.), pp. 71–101. Humana Press, Totowa, NJ.

Minton, K. W. (1996). Repair of ionizing-radiation damage in the radiation resistant bacterium *Deinococcus radiodurans*. *Mutat. Res.* **363**, 1–7.

Morrison, N., Rattray, M., Oliver, S. G., Hayes, A., Zhang, N., Brutsche, M., Penkett, C., Lockey, J., Rao, S., Hays, I., Jupp, R., and Brass, A. (1999). Robust normalization of microarray data over multiple experiments. Nature Genetics Microarray Meeting: Technology, Application and Analysis.

Moser, D., and Nealson, K. H. (1996). Growth of the facultative anaerobe *Shewanella putrefaciens* by elemental sulfur reduction. *Appl. Environ. Microbiol.* **62**, 2100–2105.

Murray, A. E., Lies, D., Li, G., Nealson, K., Zhou, J., and Tiedje, J. M. (2001). DNA/DNA hybridization to microarrays reveals gene-specific differences between closely related microbial genomes. *Proc. Natl. Acad. Sci. USA* **98**, 9853–9858.

Nadon, R., Shi, P., Skandalis, A., Woody, E., Hubschle, H., Susko, E., Rghei, N., and Ramm, P. (2001). Statistical inference methods for gene expression arrays. *In* "Proceedings of SPIE, Microarrays: Optical Technologies and Informatics" (Bittner, Ed.), V4266, pp. 46–55. International Society for Optical Engineering, Bellingham, WA.

Nadon, R., and Shoemaker, J. (2002). Statistical issues with microarrays: processing and analysis. *Trends Genet.* **18**, 265–271.

Nelson, K. E., Clayton, R. A., Gill, S. R., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, L. D., Nelson, W. C., Ketchum, K. A., McDonald, L., Utterback, T. R., Malek, J. A., Linher, K. D., Garrett, M. M., Stewart, A. M., Cotton, M. D., Pratt, M. S., Phillips, C. A., Richardson, D., Heidelberg, J., Sutton, G. G., Fleischmann, R. D., Eisen, J. A., White, O., Salzberg, S. L., Smith, H. O., Venter, J. C., and Fraser, C. M. (1999). Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399**, 323–329.

Nealson, K. H., and Saffarini, D. A. (1994). Iron and manganese in anaerobic respiration: environmental significance, physiology, and regulation. *Ann. Rev. Microbiol.* **48**, 311–343.

Ng, W. V., Kennedy, S. P., Mahairas, G. G., Berquist, B., Pan, M., Shukla, H. D., Lasky, S. R., Baliga, N. S., Thorsson, V., Sbrogna, J., Swartzell, S., Weir, D., Hall, J., Dahl, T. A., Welti, R., Goo, Y. A., Leithauser, B., Keller, K., Cruz, R., Danson, M. J., Hough, D. W., Maddocks, D. G., Jablonski, P. E., Krebs, M. P., Angevine, C. M., Dale, H., Isenbarger, T. A., Peck, R. F., Pohlschroder, M., Spudich, J. L., Jung, K. H., Alam, M., Freitas, T., Hou, S. B., Daniels, C. J., Dennis, P. P., Omer, A. D., Ebhardt, H., Lowe, T. M., Liang, R., Riley, M., Hood, L., and DasSarma, S. (2000). Genome sequence of *Halobacterium* species NRC-1. *Proc. Natl. Acad. Sci. USA* **97**, 12176–12181.

Nuwaysir, E. F., Huang, W., Albert, T. J., Singh, J., Nuwaysir, K., Pitas, A., Richmond, T., Gorski, T., Berg, J. P., Ballin, J., McCormick, M., Norton, J., Pollock, T., Sumwalt, T., Butcher, L., Porter, D., Molla, M., Hall, C., Blattner, F., Sussman, M. R., Wallace, R. L., Cerrina, F., and Green, R. D. (2002). Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Genome Res.* **12**, 1749–1755.

Olsen, G. J., and Woese, C. R. (1997). Archaeal genomics: an overview. *Cell* **89**, 991–994.

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

Petrov, A., Shah, S., Draghici, S., and Shams, S. (2002). Microarray image processing and quality control. *In* "DNA Array Image Analysis — Nuts & Bolts" (S. Shah and G. Kamberova, Eds.), pp. 99–130. DNA Press, LLC, Eagleville, PA.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W. L., Chen, C., Zhai, Y., Dairkee, S. H., Ljung, B. M., Gray, J. W., and Albertson, D. G. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* **20**, 207–11.

Pollack, R. J., Perou, C., Alizadeh, A., Eisen, M., Pergamenschikov, A., Williams, C., Jeffrey, S., Botstein, D., and Brown, P. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Am.* **23**, 41–46.

Ramsay, G. (1998). DNA chips: state-of-the art. *Nat. Biotechnol*. **1**, 640–644.

Rehman, F. N., Audeh, M., Abrams, E. S., Hammond, P. W., Kenney, M., and Boles, T. C. (1999). Immobilization of acrylamide-modified oligonucleotides by co-polymerization. *Nucl. Acids Res* **27**, 649–655.

Richardson, D. J. (2000). Bacterial respiration: a flexible process for a changing environment. *Microbiol.* **146**, 551–571.

Robb, F. T., Maeder, D. L., Brown, J. R., DiRuggiero, J., Stump, M. D., Yeh, R. K., Weiss, R. B., and Dunn, D. M. (2001). Genomic sequence of hyperthermophile Pyrococcus furiosus: implications for physiology and enzymology. *Meth. Enzymol.* **330**, 134–157.

Rogers, Y., Jiang-Baucome, P., Huange, Z., Bogdanov, V., Anderson, S., and Boyce-Jacino, M. T. (1999). Immobilization of oligonucleotides onto a glass support via disulfide bonds: A method for preparation of DNA microarrays. *Ann. Biochem.* **266**, 23–30.

Rose, D. (2000). Microfluidic technologies and instrumentation for printing DNA microarrays. *In* "Microarray Biochip Technology" (M. Schena, Ed.), pp. 19–38. Eaton Publishing, Natick, MA.

Rudi, K., Skulberg, O. M., Skulberg, R., and Jakobsen, K. S. (2000). Application of sequence-specific labeled 16S rRNA gene oligonucleotide probes for genetic profiling of cyanbacterial abundance and diversity by array hybridization. *Appl. Environ. Microbiol.* **66**, 4004–4011.

Salama, N., Guillemin, K., McDaniel, T. K., Sherlock, G., Tompkins, L., and Falkow, S. (2000). A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. *Proc. Natl. Acad. Sci. USA* **97**, 14668–14673.

Salunga, R. C., Guo, H., Luo, L., Bittner, A., Joy, K. D., Chambers, J. R., Wan, J. S., Jackson, M. R., and Erlander, M. G. (1999). Gene expression analysis via cDNA microarrays of laser capture microdissected cells from fixed tissue. *In* "DNA Microarrays: A Practical Approach" (M. Schena, Ed.), pp. 121–136. Oxford University Press, New York, NY.

Schadt, E. E., Li, C., Su, C., and Wong, W. H. (2000). Analyzing high-density oligonucleotide gene expression array data. *J. Cell Biochem.* **80**, 192–202.

Schena, M. (2003). "Microarray analysis." John Wiley & Sons, New York, NY.

Schena, M., and Davis, R. W. (2000). Technology standards for microarray research. *In* "Microarray Biochip Technology" (M. Schena, Ed.), pp. 1–18. Eaton Publishing, Natick, MA.

Schena, M., Heller, R., Theriault, T., Konrad, K., Lachenmeier, E., and Davis, R. W. (1998). Microarrays: biotechnology's disovery platform for functional genomics. *Trends Biotech.* **16**, 301–306.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science.* **270**, 467–470.

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P., and Davis, R. W. (1996). Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. USA* **93**, 10614–10619.

Schermer, M. J. (1999). Confocal scanning microscopy in microarray detection. *In* "DNA Microarrays: A Practical Approach" (M. Schena, Ed.), pp. 17–42. Oxford University Press, New York, NY.

Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H., and Herzel, H. (2000). Normalization strategies for cDNA microarrays. *Nucl. Acids Res.* **28**, E47.

Schut, G. J., Zhou, J., and Adams, M. W. W. (2001). DNA microarray analysis of the hyperthermophilic archaeon *Pyrococcus furiosus*: Evidence for a new type of sulfur-reducing enzyme complex. *J. Bacteriol.* **183**, 7027–7036.

Shalon, D., Smith, S. J., and Brown, P. O. (1996). A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* **6**, 639–645.

Shchepinov, M. S., Case-Green, S. C., and Southern, E. M. (1997). Steric factors influencing hybridisation of nucleic acids to oligonucleotide arrays. *Nucl. Acids Res.* **25**, 1155–1161.

Shriver-Lake, L. C. (1998). Silane-modified surfaces for biomaterial immobilization. *In* "Immobilized Biomolecules in Analysis" (T. Cass and F. S. Ligler, Eds.), pp. 1–14. Oxford University Press, New York, NY.

Slesarev, A. I., Mezhevaya, K. V., Makarova, K. S., Polushin, N. N., Shcherbinina, O. V., Shakhova, V. V., Belova, G. I., Aravind, L., Natale, D. A., Rogozin, I. B., Tatusov, R. L., Wolf, Y. I., Stetter, K. O., Malykh, A. G., Koonin, E. V., and Kozyavkin, S. A. (2002). The complete genome of hyperthermophile *Methanopyrus kandleri* AV19 and monophyly of archaeal methanogens. *Proc. Natl. Acad. Sci. USA* **99**, 4644–4649.

Small, J., Call, D. R., Brockman, F. J., Straub, T. M., and Chandler, D. P. (2001). Direct detection of 16S rRNA in soil extracts by using oligonucleotide microarrays. *Appl. Environ. Microbiol.* **67**, 4708–4716.

Smith, D. R., Doucette-Stamm, L. A., Deloughery, C., Lee, H., Dubois, J., Aldredge, T., Bashirzadeh, R., Blakely, D., Cook, R., Gilbert, K., Harrison, D., Hoang, L., Keagle, P., Lumm, W., Pothier, B., Qiu, D., Spadafora, R., Vicaire, R., Wang, Y., Wierzbowski, J., Gibson, R., Jiwani, N., Caruso, A., Bush, D., Reeve, J. N., et al., (1997). Complete genome sequence of *Methanobacterium thermoautotrophicum* △H: Functional analysis and comparative genomics. *J. Bacteriol.* **179**, 7135–7155.

Southern, E. M. (2001). DNA microarrays. History and overview. *Meth. Mol. Biol.* **170**, 1–15.

Southern, E. M., Case-Green, S. C., Elder, J. K., Johnsone, M., Mir, K. U., Wang, L., and Williams, J. C. (1994). Arrays of complementary oligonucleotides for analyzing the hybridizaiton behavior of nucleic acids. *Nucl. Acids Res.* **22**, 1368–1373.

Southern, E. M., Maskos, U., and Elder, J. K. (1992). Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models. *Genomics* **13**, 1108–1107.

Southern, E., Mir, K., and Shchepinov, M. (1999). Molecular interactions on microarrays. *Nat. Genet.* **21**(Suppl. S), 5–9.

Stears, R. L., Getts, R. C., and Gullans, S. R. (2000). A novel, sensitive detection system for high-density microarrays using dendrimer technology. *Physiol. Genomics* **3**, 93–99.

Stears, B. L., Martinsky, T., and Schena, M. (2003). Trends in microarray analysis. *Nat. Med.* **9**, 140–145.

Steel, A. B., Levicky, R. L., Herne, T. M., and Tarlo, M. J. (2000). Immobilization of nucleic acids at solid surfaces: effect of oligonucleotide length on layer assembly. *Biophys.* **J79**, 975–981.

Stillman, B. A., and Tonkinson, J. L. (2000). FAST slides: a novel surface for microarrays. *Biotechniques* **29**, 630–635.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* **96**, 2907–2912.

Taniguchi, M., Miura, K., Iwao, H., and Yamanaka, S. (2001). Quantitative assessment of DNA microarrays — comparison with Northern blot analyses. *Genomics* **71**, 34–39.

Tao, H., Bausch, C., Richmond, C., Blattner, F. R., and Conway, T. (1999). Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media. *J. Bacteriol.* **181**, 6425–6440.

Taroncher-Oldenburg, G., Griner, E. M., Francis, C. A., and Ward, B. B. (2003). Oligonucleotide microarray for the study of functional gene diversity in the nitrogen cycle in the environment. *Appl. Environ. Microbiol.* **69**, 1159–1171.

Tavazoie, S., and Church, G. M. (1998). Quantitative whole-genome analysis of DNA-protein interactions by in vivo methylase protection in *E-coli*. *Nat. Biotech.* **16**, 566–571.

Thompson, D. K., Beliaev, A. S., Giometti, C. S., Tollaksen, S. L., Khare, T., Lies, D. P., Nealson, K. H., Lim, H., Yates, III, J., Brandt, C. C., Tiedje, J. M., and Zhou, J. (2002). Transcriptional and proteomic analysis of a ferric uptake regulator (Fur) mutant of *Shewanella oneidensis*: Possible involvement of Fur in energy metabolism, transcriptional regulation, and oxidative stress. *Appl. Environ. Microbiol.* **68**, 881–892.

Toronen, P., Kolehmainen, M., Wong, G., and Castren, E. (1999). Analysis of gene expression data using self-organizing maps. *FEBS Lett.* **451**, 142–146.

Tseng, G. C., Oh, M. K., Rohlin, L., Liao, J. C., and Wong, W. H. (2001). Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucl. Acids Res.* **29**, 2549–2557.

Urakawa, H., Noble, P. A., El Fantroussi, S., Kelly, J. J., and Stahl, D. A. (2002). Single-base-pair discrimination of terminal mismatches by using oligonucleotide microarrays and neural network analyses. *Appl. Environ. Microbiol.* **68**, 235–244.

Vasiliskov, A., Timofeev, E., Surzhikov, S., Drobyshev, A., Shick, V., and Mirzabekov, A. (1999). Fabricaiton of microarray of gel-immobilized compounds on a chip by copolymerization. *Biotechniques* **27**, 592–606.

Venkateswaran, A., McFarlan, S. C., Ghostal, D., Minton, K. W., Vasilenko, A., Makarova, K. S., Wackett, L. P., and Daly, M. J. (2000). Physiologic determinants of radiation resistance in *Deinococcus radiodurans*. *Appl. Environ. Microbiol.* **66**, 2620–2626.

Venkateswaran, K., Moser, D., Dollhopf, M., Lies, D. P., Saffarini, D. A., MacGregor, B. J., Ringelberg, D. B., White, D. C., Nishijima, M., and Sano, H. (1999). Polyphasic taxonomy of the genus *Shewanella* and description of *Shewanella oneidensis* sp. *nov. Int. J. Syst. Bacteriol.* **49**, 705–724.

Verdnik, D., Handran, S., and Pickett, S. (2002). Key considerations for accurate microarray scanning and image analysis. *In* "DNA Array Image Analysis — Nuts & Bolts" (S. Shah and G. Kamberova, Eds.). DNA Press, LLC, Eagleville, PA.

Voordouw, G. (1998). Reverse sample genome probing of microbial community dynamics. *ASM News* **64**, 627–633.

Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, E., Hubbell, E., Robinson, M., Mittmann, M. S., Morris, N., Shen, D., Kilburn, T., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T. J., Lander, E. S., and *et al.* (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082.

Warrington, J. A., Dee, S., and Trulson, M. (2000). Large-scale genomic analysis using Affymetrix GeneChip® probe arrays. *In* "Microarray Biochip Technology" (M. Schena, Ed.), pp. 119–148. Eaton Publishing, Natick, MA.

White, O., Eisen, J. A., Heidelberg, J. F., Hickey, E. K., Peterson, J. D., Dodson, R. J., Haft, D. H., Gwinn, M. L., Nelson, W. C., Richardson, D. L., Moffat, K. S., Qin, H. Y., Jiang, L. X., Pamphile, W., Crosby, M., Shen, M., Vamathevan, J. J., Lam, P., McDonald, L., Utterback, T., Zalewski, C., Makarova, K. S., Aravind, L., Daly, M. J., Minton, K. W., Fleischmann, R. D., Ketchum, K. A., Nelson, K. E., Salzberg, S., Smith, H. O., Venter, J. C., and Fraser, C. M. (1999). Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science* **286**, 1571–1577.

Wilson, K. H., Wilson, W. J., Radosevich, J. L., DeSantis, T. Z., Viswanathan, V. S., Kuczmarski, T. A., and Andersen, G. L. (2002). High-density microarray of small-subunit ribosomal DNA probes. *Appl. Environ. Microbiol.* **68**, 2535–2541.

Wodicka, L., Dong, H., Mittmann, M., Ho, M.-H., and Lockhart, D. J. (1997). Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat. Biotechnol.* **15**, 1359–1367.

Woese, C. R. (1987). Bacterial evolution. *Microbiol. Rev.* **51**, 221–271.

Woese, C. R., Kandler, O., and Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA* **87**, 4576–4579.

Worley, J., Bechtol, K., Penn, S., Roach, D., Hanzel, D., Trounstine, M., and Barker, D. (2000). A systems approach to fabricating and analyzing DNA microarrays. *In* "Microarray Biochip Technology" (M. Schena, Ed.), pp. 65–85. Eaton Publishing, Natick, MA.

Wu, L. Y., Thompson, D., Li, G.-S., Hurt, R., Huang, H., Tiedje, J. M., and Zhou, J.-Z. (2001). Development and evaluation of functional gene arrays for detection of selected genes in the environment. *Appl. Environ. Microbiol.* **67**, 5780–5790.

Xu, D., Li, G., Wu, L., Zhou, J.-Z., and Xu, Y. (2002). PRIMEGENS: A computer program for robust and efficient design of gene-specific targets on microarrays. *Bioinformatics* **18**, 1432–1437.

Yang, Y. H., and Speed, T. (2002). Design issues for cDNA microarray experiments. *Nat. Rev. Genet.* **3**, 579–588.

Ye, R., Tao, W., Bedzyk, L., Young, T., Chen, M., and Li, L. (2000). Global gene expression profiles of Bacillus subtilis grown under anaerobic conditions. *J. Bacteriol.* **182**, 4458–4465.

Zammatteo, N., Jeanmart, L., Hamels, S., Courtois, S., Louette, P., Hevesi, L., and Remacle, J. (2000). Comparison between different strategies of covalent attachment of DNA to glass surfaces to build DNA microarrays. *Anal. Biochem.* **280**, 143–150.

Zhao, X., Nampalli, S., Serino, A. J., and Kumar, S. (2001). Immobilization of oligodeoxyribonucleotides with multiple anchors to microchips. *Nucl. Acids Res.* **29**, 955–959.

Zhou, J.-Z. (2003). Microarrays for bacterial detection and microbial community analysis. *Curr. Opin. Microbiol.* **6,** 288–294.

Zhou, Y.-X., Kalocsai, P., Chen, J.-Y., and Shams, S. (2000). Information processing issues and solutions associated with microarray technology. *In* "Microarray Biochip Technology" (M. Schena, Ed.), pp. 167–200. Eaton Publishing, Natick, MA.

Zhou, J., Palumbo, A. V., and Tiedje, J. M. (1997). Sensitive detection of a novel class of toluene-degrading denitrifiers, *Azoarcus tolulyticus*, with small-subunit rRNA primers and probes. *Appl. Environ. Microbiol.* **63**, 2384–2390.

Zhou, J., and Thompson, D. K. (2002). Challenges in applying microarrays to environmental studies. *Curr. Opin. Biotechnol.* **13**, 204–207.

Zlatanova, J., and Mirzabekov, A. (2001). Gel-immobilized microarrays of nucleic acids and proteins: production and application for macromolecular research. *Meth. Mol. Biol.* **170**, 17–38.