# [1] The Affymetrix GeneChip® Platform: An Overview

*By* Dennise D. Dalma-Weiszhausz, Janet Warrington, Eugene Y. Tanimoto, and C. Garrett Miyada

## Abstract

The intent of this chapter is to provide the reader with a review of GeneChip technology and the complete system it represents, including its versatility, components, and the exciting applications that are enabled by this platform. The following aspects of the technology are reviewed: array design and manufacturing, target preparation, instrumentation, data analysis, and both current and future applications. There are key differentiators between Affymetrix' GeneChip technology and other microarray-based methods. The most distinguishing feature of GeneChip microarrays is that their manufacture is directed by photochemical synthesis. Because of this manufacturing technology, more than a million different probes can be synthesized on an array roughly the size of a thumbnail. These numbers allow the inclusion of multiple probes to interrogate the same target sequence, providing statistical rigor to data interpretation. Over the years the GeneChip platform has proven to be a reliable and robust system, enabling many new discoveries and breakthroughs to be made by the scientific community.

## Introduction

Starting in the 1990s, a genomic revolution, propelled by major technological advances, has enabled scientists to complete the sequences of a variety of organisms, including viruses, bacteria, invertebrates, and culminating in the full draft sequence of the human genome (Lander *et al.*, 2001). In the wake of this flood of sequence information, scientists are currently faced with the daunting task of translating genomic sequence information into functional biological mechanisms that will allow a better understanding of life and its disease states and hopefully offer better diagnostics and novel therapeutic interventions. High-density microarrays are uniquely qualified to tackle this daunting task and have therefore become an essential tool in life sciences research. They provide a reliable, fast, and cost-effective method that effectively scales with the ever-increasing amounts of genomic information.

METHODS IN ENZYMOLOGY, VOL. 410 0076-6879/06 $35.00
Copyright 2006, Elsevier Inc. All rights reserved.
DOI: 10.1016/S0076-6879(06)10001-4

Tecan Ex. 2124, Page 1 of 26
QIAGEN Sciences, LLC v. Tecan Genomics - IPR2025-00029

In the last decade, there has been an immense growth in the use of high-throughput microarray technology for three major genetic explorations: the genome-wide analysis of gene expression, SNP genotyping, and resequencing. While many of these studies have focused on human subjects and diseases, microarrays are also being used to study the gene expression and sequence variation of a variety of model organisms, such as yeast, *Drosophila*, mice, and rats. New applications are rapidly emerging, such as the discovery of novel transcripts (from coding and noncoding regions), the identification of novel regulatory sequences, and the characterization of functional domains in the RNA transcript. Integrating all of the information emanating from whole-genome studies will undoubtedly allow a more global understanding of the genome and the regulatory circuits that govern its activity.

The comparison of genome-wide expression patterns provides researchers with an objective and hypothesis-free method to better understand the dynamic relationship between mRNA content and biological function. This method has enabled scientists to discover, for example, the genetic pathways that are changed and disrupted in a wide range of diseases, from cancer (Armstrong *et al*., 2002; Huang *et al*., 2004; Yeoh *et al*., 2002) to multiple sclerosis (Steinman and Zamvil, 2003). Across multiple disciplines, whole-genome expression analysis is helping scientists to stratify disease states, predict patient outcome, and make better therapeutic choices. Some of the recent examples of scientific and medical findings utilizing this technology include the identification of murine longevity genes and the discovery of novel transcripts that question our basic understanding of gene expression (Kapranov *et al*., 2002).

The most recent generation of GeneChip microarrays for DNA sequence analysis allows scientists to genotype single nucleotide polymorphisms on a genome-wide scale (Kennedy *et al*., 2003; Matsuzaki *et al*., 2004a,b). The ability to quickly genotype over 100,000 single nucleotide polymorphisms (SNPs) distributed across the human genome has allowed researchers to conduct linkage analysis and genetic association studies. These new tools for disease mapping studies have already helped scientists pinpoint genes linked to diseases such as sudden infant death syndrome (Puffenberger *et al*., 2004), neonatal diabetes (Sellick *et al*., 2003), and bipolar disorder (Middleton *et al*., 2004). The technology has proven to be scalable, and assays that cover 500,000 SNPs are now available.

Microarrays have revolutionized basic scientific research and are constantly challenging our view of the genome and its complexity. They are finding their way from the research laboratory to the clinic, where they promise the same kind of revolution in patient care. Microarrays used in

clinical research and clinical applications promise to help scientists develop more accurate diagnostics and create novel therapeutics. By standardizing microarray data and integrating it with a patient's existing medical records, physicians can offer more tailored and more successful therapies. The combination of a patient's genetic and clinical data will allow for personalized medicine, which is where GeneChip technology holds the greatest promise to improve health.

## GeneChip Microarrays, a Flexible Platform

GeneChip arrays are the result of the combination of a number of technologies, design criteria, and quality control processes. In addition to the arrays, the technology relies on standardized assays and reagents, instrumentation (fluidics system, hybridization oven and scanner), and data analysis tools that have been developed as a single platform. The key assay steps are outlined in Fig. 1 and are discussed in greater detail in later sections along with array design and manufacturing. The considerable flexibility of the GeneChip system and the manufacturing technology allows the design of the arrays to be dictated by their intended use, such as whole-genome transcriptome mapping, gene expression profiling, or custom genotyping. In addition to GeneChip catalogue microarrays (over 50 arrays and array sets are currently available), a custom program exists, where researchers can design their own arrays for organisms not covered by existing products and for specialized or directed studies. These designs may be based on many of the same design features and manufacturing techniques available in catalogue arrays (probe selection algorithms, manufacturing control tests, etc.) and are expected to provide customers equivalent performance to their commercial counterparts.

## Array Manufacturing

Adapting technologies used in the semiconductor industry, GeneChip array manufacturing begins with a 5-in.$^2$ quartz wafer (Fodor *et al*., 1991; McGall and Christians, 2002). This substrate is first modified covalently with a silane reagent to produce a stable surface layer of hydroxyalkyl groups. Linker molecules with photolabile-protecting groups are then attached covalently to this layer to create a surface that may be spatially activated by light (Fig. 2). A photolithographic mask set that represents the sequence information content on the array is carefully designed. Each mask is manufactured with windows that either block or permit the transmission of ultraviolet light. These windows are distributed over the mask based on the desired sequence of each probe. The mask is carefully aligned

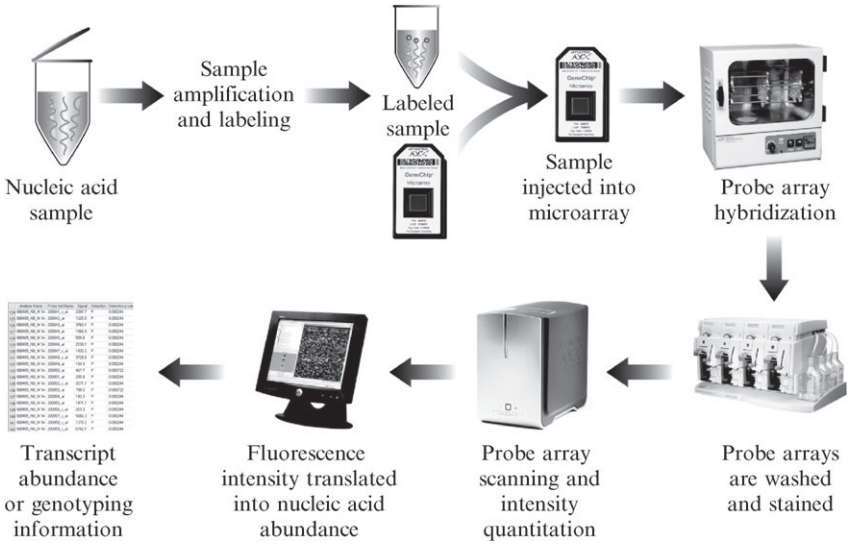


FIG. 1. Flowchart of a GeneChip System microarray experiment. Once the nucleic acid sample has been obtained, target amplification and labeling result in a labeled sample. The labeled sample is then injected into the probe array and allowed to hybridize overnight in the hybridization oven. Probe array washing and staining occur on the fluidics station, which can handle four probe arrays simultaneously. The probe array is then ready to be scanned in the Affymetrix GeneChip scanner, where the fluorescence intensity of each feature is read. Data output includes an intensity measurement for each transcript or the detailed sequence or genotyping (SNP) information.

with the quartz wafer, which ensures that oligonucleotide synthesis is only activated at precise locations on the wafer. When near-ultraviolet light shines through the mask, terminal hydroxyl groups on the linker molecules in exposed areas of the wafer are deprotected, thereby activating them for nucleotide coupling, while linkers in unexposed regions remain protected and inactive. A solution containing a deoxynucleoside phosphoramidite monomer with a light-sensitive protecting group is flushed over the surface of the wafer, and the nucleoside attaches to the activated linkers (coupling step), initiating the synthesis process.

Oligonucleotide synthesis proceeds by repeating the two basic steps: deprotection and coupling. For each round of synthesis, deprotection generally uses a unique mask from the designed set. The coupling steps alternate through the addition of A-, C-, G-, or T-modified nucleotides. The deprotection and coupling cycle is repeated until all of the full-length probe sequences, usually 25-mers, are completed. Algorithms that optimize

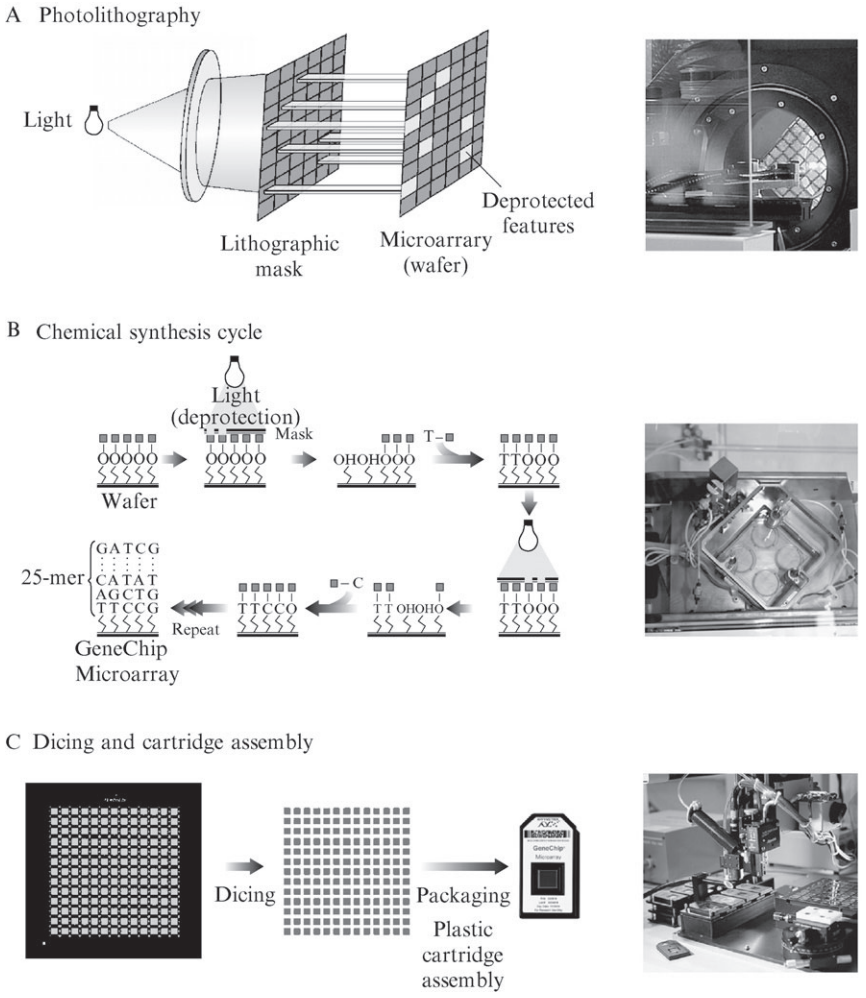


FIG. 2. Manufacture of a GeneChip probe array. (A) Photolithography. (Left) Near-ultraviolet light is passed through a mask containing open windows. The size and the location of each open window delineate the surface on the quartz wafer that will be activated for chemical synthesis. The use of sequential masks in conjunction with the chemical synthesis creates a cycle that directs the precise sequence synthesis of oligonucleotides that compose the array. (Right) The photolithographic process. (B) (Left) Schematic representation of the nucleic acid synthesis cycle. Light removes protecting groups (squares) at defined areas on the array. A single nucleotide is washed over the array and couples to the deprotected areas. Through successive steps, any oligonucleotide sequence can be built on each feature of the array. The number of steps required to build a 25 nucleotide sequence on the array is 100, although the optimization of mask usage has lowered that number to ~75 steps. (Right) The chemical synthesis station, where nucleotide binding occurs. (C) (Left) Complete synthesis on

mask usage allow the creation of the arrays in significantly fewer than the 100 cycles that would normally be required to synthesize all possible 25-mer sequences (Lipshutz *et al.*, 1995). The information density of the array depends on the spatial resolution of the photolithographic process.

Once oligonucleotide synthesis is complete, wafers can be diced in a variety of array sizes and packaged individually into cartridges. Generally, each 5-in. square wafer can yield between 49 and 400 identical GeneChip microarrays, depending on the amount of genetic information required. A typical 1.28-$cm^2$ array (49-format), for example, will contain more than 1.4 million different probe locations, or features, assuming the features are spaced 11 $\mu$m apart. Each of these features contains millions of identical DNA molecules. A reduction of the feature spacing to 5 $\mu$m (as available on the Mapping 500K Array Set released in September of 2005) produces over 6.5 million different features on the same 1.28-$cm^2$ array—an exponential increase in the available data from a single experiment. This demonstrates the power of ''feature shrink'' on the Affymetrix microarray platform. The manufacturing process ends with a comprehensive series of quality control tests to ensure that GeneChip arrays deliver accurate and reproducible data.

## Array Design

Array design is closely coupled to sample preparation and the biological question to be addressed. Specific examples are described in greater detail for expression and genotyping applications. Almost all of the designs utilize two types of probes: (1) probes that have complete complementarity to their target sequence [perfect match probe (PM)] and (2) probes with a single mismatch to the target, centered in the middle of the oligonucleotide [mismatch probe (MM), Fig. 3]. The number of probes used to interrogate a specific SNP or transcript is selected to meet specific performance criteria for each assay.

In addition to the probes specific for a particular assay, arrays contain a number of different control probes. There are probes specific for quality control assays. Another set of probes is arranged in checkerboard patterns on the array. These probes bind to a specific biotinylated oligonucleotide included in the hybridization cocktail. Following scanning, these

the wafer results in many (49–400) identical high-density oligonucleotide microarrays in one wafer. Dicing of the wafer into individual microarrays occurs, and each microarray is inserted into a plastic cartridge. (Right) Machinery used to incorporate the diced microarray into the plastic cartridge.

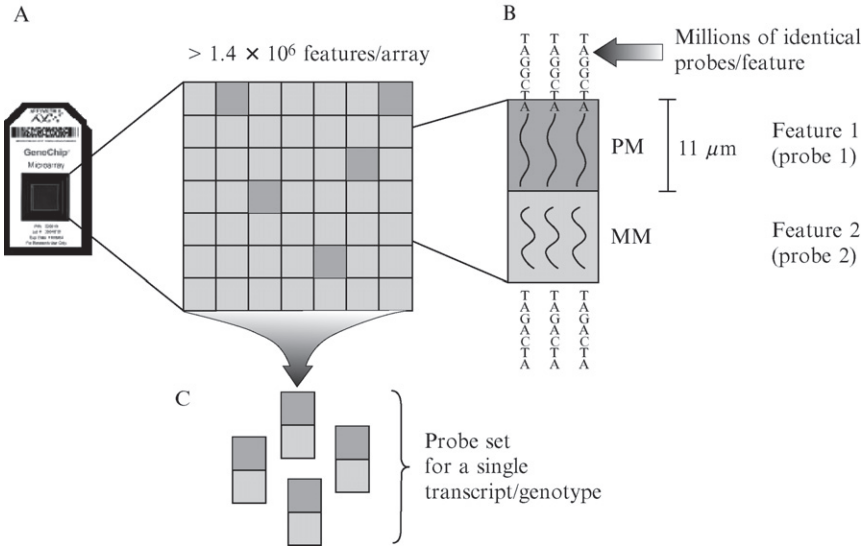


FIG. 3. Dissection of a probe array. (A) Inside the probe array (left) is a piece of quartz, generally containing a synthesis area of 1.28 $cm^2$ and carrying more than a million different features, assuming 11-$\mu$m feature spacing. Each feature, in turn, is composed of millions of oligonucleotide sequences. (B) For every perfect match (PM) feature, a mismatch (MM) feature is included, which is identical to the PM sequence, except for a nucleotide transversion on the 13th nucleotide, the central nucleotide. (C) A probe set refers to all features (PM and MM) that interrogate the same target sequence.

checkerboard patterns provide a means to ensure that signal intensities are properly assigned to the correct feature on the array. Other probes can detect specific controls that are added during sample preparation, providing evidence that the upstream assay was performed properly.

*Array Designs for Gene Expression*

The probe selection strategy used for gene expression arrays is dictated by the intended use of the array. For example, probes can be selected that identify unique transcripts, common transcript sequence segments, multiple splice sites, or polyadenylation variants. Bioinformatics techniques are used to assemble sequences from various public sources such as GenBank, dbEST, and RefSeq. Genome sequence alignments allow the selection of high-quality sequence data, as well as the consolidation of redundant transcripts and the identification of splice variants. The use of cDNA assemblies over exemplar sequences results in a higher quality

design based on all of the empirical sequence data. cDNA sequence orientation is determined using a probabilistic model applied to genetic annotations, genomic splice-site usage, polyadenlyation sites, and sequence observations. This combination of metrics ensures that probes are selected against the correct strand. Annotations are generated for each target and are then prioritized for inclusion in the final array design.

In the *in vitro* transcription (IVT) assay the probe selection region is typically defined as the first 600 bases proximal to the polyadenylation site (3′ end) (Fig. 3). Probe selection requires applying a multiple linear regression model to identify those probes whose hybridization intensities respond in a linear fashion to the relative abundance of the target (Mei *et al*., 2003). The algorithm is based on a thermodynamic model of nucleic acid duplex formation modified with empirically derived parameters. Probes are also selected to minimize the effects of cross hybridization and to maximize spacing between the probes. Typically 11 probe pairs are selected per 600-bp probe selection region. A probe pair consists of a PM probe and its corresponding MM probe (Fig. 3).

Expression assays based on random priming methods can be applied to both prokaryotes and eukaryotes, and probe selection regions need not be restricted to the 600 bases proximal to the 3′ end of the gene. In the case of prokaryotes, arrays are usually designed using open reading frames as the probe selection region. For eukaryotes the probe selection region is defined by potential exons. This type of design permits expression analysis over the

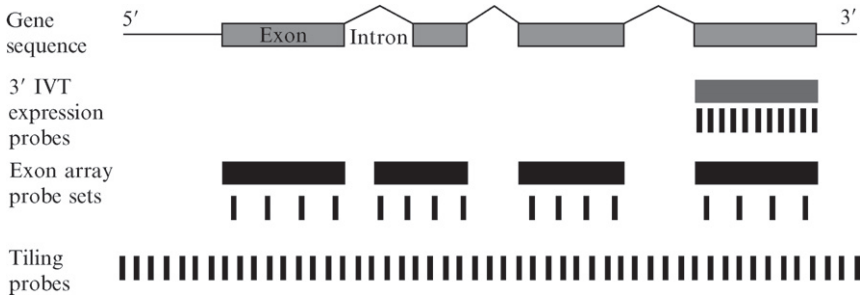


FIG. 4. Gene expression array design strategies. The different expression strategies for probe selection are represented. The gene sequence shown at the top represents an example of a target transcript. Rectangles represent exons, while the connecting lines represent introns. The 3′ IVT expression probes target sequences are at the extreme 3′ end and are adjacent to the poly(A) tail of the mRNA. This strategy is the most commonly used for commercial whole genome transcriptome designs. Exon array probe sets include probes that are within exon sequences. For tiling arrays, probes are placed sequentially throughout the genome at the same approximate distance from each other.

entire transcript and allows the identification of alternatively spliced transcripts. Such a design is illustrated in Fig. 4.

A third type of design is used in expression analysis. Tiling arrays interrogate the genome at regular intervals without regard to gene annotations (Fig. 4). Originally this type of design was applied to a pair of human chromosomes. Currently the entire human genome can be interrogated at 35-bp intervals (measured center to center from adjacent probes) using 14 arrays that contain features spaced 5$\mu$m apart. This type of design has proven useful in the identification of novel transcripts but its utility stretches beyond RNA mapping. For example, the chromosomal location of binding sites for DNA-binding proteins have been identified by applying chromatin-immunoprecipitated material to these arrays. It should be noted that with the exception of tiling, the array designs can be improved with better gene annotations.

### *Array Design for DNA Analysis*

High-density oligonucleotide arrays enable rapid analysis of sequence variation (resequencing) and analysis of single nucleotide polymorphisms (genotyping). A different set of strategies is used to select probes for DNA analysis. The design of the array relies on multiple probes to interrogate individual nucleotides in a sequence. For sequence variation analysis (or resequencing), the identity of a target base can be deduced using four identical probes that vary only in the target position, each containing one of the four possible bases. For SNP genotyping, arrays with many probes for each allele can be created to provide redundant information. The probe tiling strategy for SNP genotyping is provided in greater detail later.

For any given SNP with alleles A and B, probes are synthesized on the array to represent both potential variants (Fig. 5). Each SNP is represented on the array by a probe set that consists of multiple probe pairs. The probe pairs differ in the location of the SNP within the oligonucleotide sequence. In addition to the PM and MM pair that contain the SNP on the central position of the probe (position 0), there are probes for each SNP that are shifted either upstream (+1, +3, +4 nucleotides) or downstream (−1, −2, −4 nucleotides) relative to the probe containing the SNP at the central position. Each of the 7 probes is empirically tested on a pilot microarray, and a total of 5 probe pairs are ultimately selected for inclusion on the final array product. Additionally, for each position, probes are included from the sense and the antisense strand. Therefore, there are a total of 20 probes interrogating each allele for a total of 40 probes per SNP. Following hybridization to the arrays, one can determine the identity of the particular SNP location as homozygous (AA or BB) or heterozygous (AB).

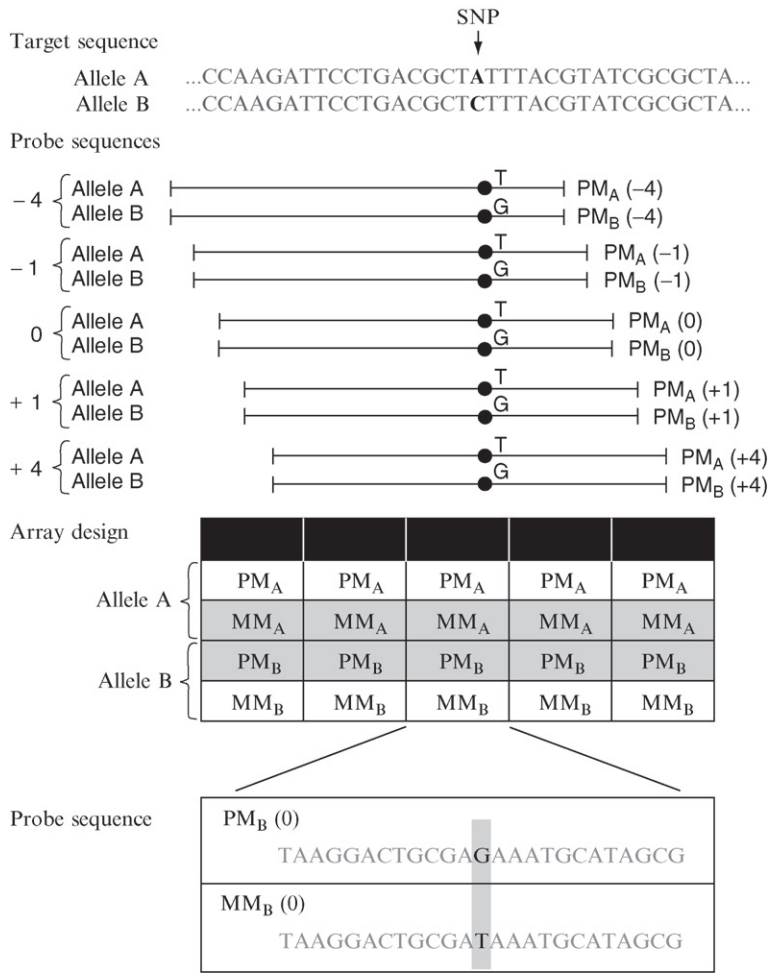


FIG. 5. Array design for DNA analysis. The top of the figure shows two possible alleles of the nucleic acid sequence to be analyzed (target). The probe sequence selection strategy for SNP genotyping includes probes that are centered on the SNP location (0), as well as probes that are shifted to the left (−4, −1) and to the right (+1, +4) of the central SNP location. The array design contains interrogation probes for both alleles and, similar to expression designs, includes a PM and MM probe pair. Depicted at the bottom of this schematic are two features representing the B allele, one harboring the centered PM probe and the one below representing its partner MM probe.

Another type of array for DNA analysis is used for resequencing. Some experimental approaches, such as sequencing large genomic regions, analyzing the sequence variants of a candidate gene, analyzing the genetic

variability within a clinical trial population, and even assessing the sequence alterations among the genome of a pathogen are well served by this array design. This design provides a highly efficient analysis of up to 30 kb of double-stranded sequence, for a total of 60 kb. The array design includes tiling four different probes for each base interrogated per strand, for a total of eight probes per nucleotide position, which provides the redundancy for analysis of sequence variation and genotype determination.

### *Other Array Designs*

As the foregoing demonstrates, Affymetrix core technology may be used to interrogate genetic material in numerous different assays to answer a broad range of different biological questions. In addition to the current gene expression and genotyping assays, two additional assays that demonstrate the flexibility of the technology are worth describing. Both of these are generic arrays to which a number of different assay or targets can be applied.

The GenFlex probe array contains over 2000 generic capture probes, which were selected for their lack of homology to existing genomic and cDNA sequences and for their similar hybridization behavior. This idea has been expanded up to 20,000 capture probes in the universal tag arrays, which are designed to work with the molecular inversion probe assay (Hardenbol *et al*., 2003), which is designed to genotype flexible panels of SNPs that can be selected by the researcher.

Another generic array is the all *n*-mer design (Lipshutz *et al*., 1995). For example, all possible 10-mer sequences can be synthesized in 40 steps on a single (1.28-$cm^2$) array with 12-$\mu$m feature spacing. These arrays may be used for differentiating variants of a known sequence.

## Target Preparation

Most target preparation protocols start with a purified nucleic acid sample that is usually amplified and then labeled and fragmented. RNA targets are prepared by *in vitro* transcription, which provides amplification of the target. Biotinylated nucleotides or analogues are incorporated into the target during the IVT process. The labeled RNA is then purified and fragmented by hydrolysis. In the case of DNA targets the purified material is first purified and then fragmented by DN'ase I. At this point the DNA fragments are labeled with terminal deoxynucleotidyl transferase (TdT) and a biotinlyated nucleotide analogue. The material is now used directly in hybridizations.

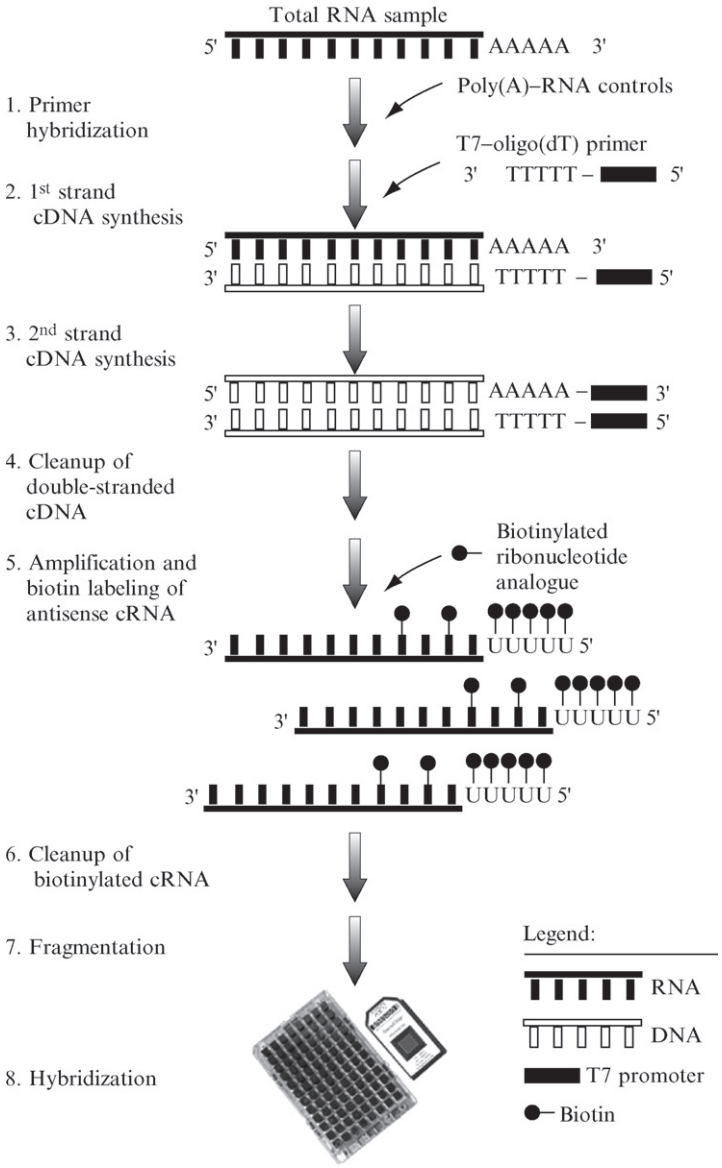


FIG. 6. One-cycle sample preparation for gene expression profiling. The flowchart depicts the steps by which eukaryotic samples are prepared for gene expression profiling. Briefly, total RNA or poly(A)–RNA is isolated. A primer that includes a poly(T) tail and a T7 polymerase-binding site [T7–oligo(dT) primer] is used for reverse transcription, resulting in synthesis of the first strand complementary DNA (cDNA). The second cDNA strand is completed, resulting in a double-stranded cDNA. In the one-cycle method, the double-stranded cDNA is

Gene expression assays have made use of both RNA and DNA targets. The most widely used sample preparation for gene expression utilizes the IVT reaction as originally described by Eberwine and colleagues (Van Gelder *et al*., 1990). In this assay cDNA synthesis is initiated from an oligo(dT) primer that is also coupled to a T7 RNA polymerase primer. In this case cDNA synthesis starts adjacent to the poly(A) tail of the mRNA. After second strand synthesis, a double-stranded cDNA copy of each mRNA is created attached to the T7 RNA polymerase primer. An IVT reaction is then carried out to create a biotinylated RNA target. A schematic of the assay is shown in Fig. 6. A variation of this technique utilizes two rounds of IVT amplification and is used to create a target from very small amounts (100 ng or less) of starting material.

Gene expression assays have also been described that utilize random priming of cDNA synthesis for target preparation. This style of target preparation is used in the case of prokaryotic expression, where mRNAs lack poly(A) tails and in instances where the entire transcript is interrogated. Examples of the latter include targets for either tiling or exon designs. The final target after random priming is either single- or double-stranded cDNA. In either case the target is fragmented by DN'ase I digestion and labeled using TdT and a biotinylated nucleotide analogue.

Chromatin immunoprecipitation (ChIP) represents another sample preparation technique where the final product may be applied to tiling arrays. Proteins are first cross-linked to chromosomal DNA by formaldehyde. The cross-linked chromatin is then fragmented and immunoprecipitated with antibodies specific for the protein of interest. The associated DNA fragments are released from the immunoprecipitated material, purified, and amplified by a polymerase chain reaction (PCR). The PCR products are labeled using techniques described previously and the final target is hybridized to the array.

The whole genome sampling analysis assay for SNP analysis does not require site-specific primers, is highly scalable, and enables the creation of hybridization target starting with as little as 250 ng of chromosomal DNA (Fig. 7). The assay starts with the digestion of the DNA sample with a single restriction enzyme, followed by ligation of a common primer and amplification by PCR. The PCR conditions are optimized for the selective amplification of fragments that are 250–2000 nucleotides in length. The

used as a template for *in vitro* transcription with biotinylated ribonucleotides, resulting in a biotin-labeled RNA sample. After cRNA fragmentation, the sample is ready to be hybridized to the array.

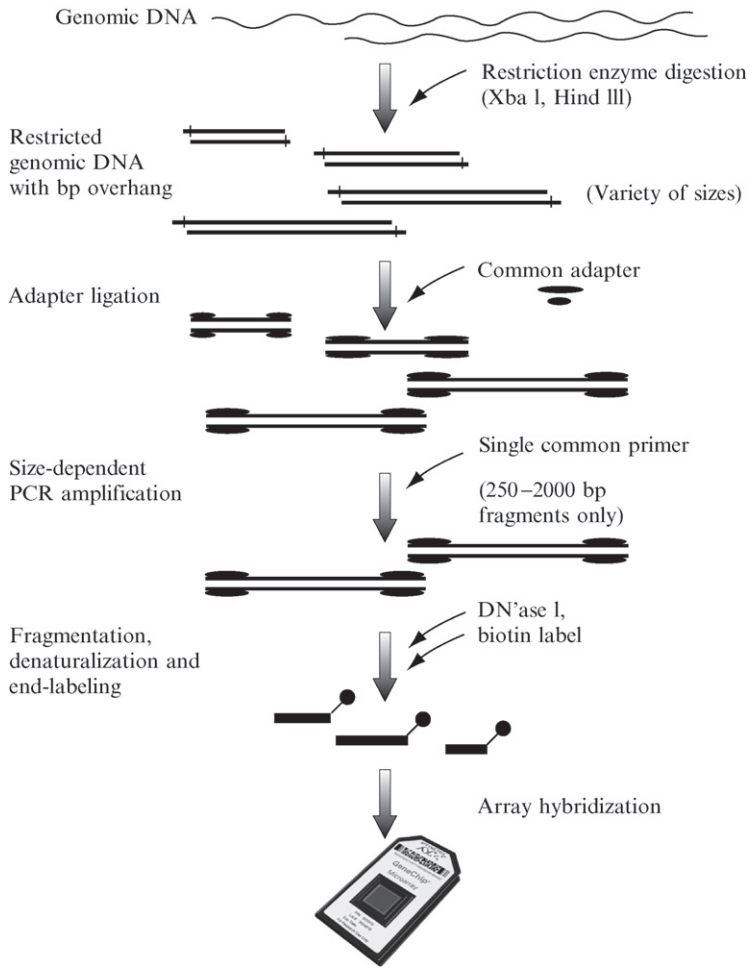


FIG. 7. Whole genome sampling assay. Schematic representation of the experimental procedure used to create a sample amenable to SNP analysis. Genomic DNA is subjected to restriction enzyme digestion, which results in varied size fragments. A common adapter is linked to the restriction overhangs and is used as a primer for PCR. PCR is conducted under controlled conditions were only fragments of 250–2000 bp are amplified, which results in a dramatic reduction of genome complexity. PCR product fragmentation and labeling result in a sample that is ready for microarray hybridization.

combination of restriction digestion and size-selective PCR amplification creates a sample of reduced complexity relative to the entire genome, which results in more accurate genotyping. The amplicons are fragmented and labeled as described previously for other DNA targets.

Resequencing assays start with a PCR amplicon or amplicons specific for the region of interest. When combining amplicons, the relative molar amount of each amplicon can be normalized to ensure a relatively uniform signal over the array. The PCR products are fragmented and labeled as described previously prior to hybridization.

### GeneChip Instrument Components and Associated Assay Steps

The GeneChip instrument components include a hybridization oven, the fluidics station, an optional autoloader, and a scanner. All of these instruments are designed to work together and, with the exception of the hybridization oven, are directed by the GeneChip operating software (GCOS). The hybridization oven can hold up to 64 probe arrays and provides continuous rotation and consistent temperatures over the 16 h that are typically required for hybridization. The temperature is tunable to cover the different array applications and is usually selected between 40 and 50°.

After hybridization the arrays are transferred to the fluidics station. The fluidics station performs washing and staining operations for GeneChip microarrays, a crucial step in the assay that impacts data consistency and reproducibility. It washes and stains up to four probe arrays simultaneously. Unbound nucleic acid is washed away through a combination of low and high stringency washes. The stringency of the wash is determined by the salt concentration of the buffer and the temperature and duration of the wash, with the temperature and duration controlled by the fluidics station. The fluidics station contains inlets for two different buffers and heats buffers up to 50°, permitting temperature-controlled washes.

In the next step, bound target molecules are ''stained'' with a fluorescent streptavidin–phycoerythrin conjugate (SAPE), which binds to the biotins incorporated during target amplification. Most protocols also include an additional signal amplification process where biotinylated anti-streptavidin antibodies are bound to the initial SAPE molecules and then stained with a second SAPE addition. In the latest fluidics station, the 450 Model, wash and stain steps proceed in an automated fashion, ending with an array that is ready for scanning. The fluidics station is controlled by a computer workstation running GCOS. Different array applications require predetermined fluidics scripts, which can also be modified for custom protocols.

The AutoLoader is a front-loading sample carousel that can be added to the latest generation scanners as an option. The AutoLoader increases throughput by permitting unattended scanning for up to 48 arrays. Arrays are maintained at 15° prior to and after scanning. The instrument also

includes a bar code reader that identifies the arrays, permits sample tracking, and aids in high-throughput analysis.

The current scanner is a wide-field, epifluorescent, confocal microscope that uses a solid-state laser to excite fluorophores bound to hybridized nucleic acids. The scanning mechanism incorporates a ''flying objective,'' which employs a large numerical aperture objective that eliminates the need for multiple array scans. The most recent version of the scanner has a pixel resolution of 0.7 $\mu$m and is able to scan features with 5-$\mu$m spacing. The scanner can resolve more than 65,000 different fluorescence intensities. During the scan process a photomultiplier tube collects and converts fluorescence values into an electronic signal, which is then converted into the corresponding numerical values. These numerical values represent the fluorescence intensities, which are stored as pixel values that comprise the image data file (.dat file).

### Image and Data Analysis

The next step in analysis is the assignment of pixels that make up the image (.dat) file to the appropriate feature. Previous methods have used a global gridding method in which the four corners of the array, defined by checkerboard patterns, serve as anchors for the grid. Features are then created by evenly dividing the area defined by the anchored corners into the known number of features for a given array. As the number of pixels per feature continued to decrease, an additional step called Feature Extraction was implemented to assign pixels to features in a more robust manner. In Feature Extraction the original pixels assigned to a feature are shifted as a block, a pixel at a time, and the coefficient of variation (CV) of pixel intensities for the shifted feature is computed. After allowing the feature pixels to shift up to a predetermined distance, the feature is defined where the pixel intensity CV is a minimum. Following Feature Extraction the intensity of each feature is calculated and stored in a .CEL file.

Regardless of application, the feature intensities found in .CEL files are used by analysis software to detect sequence variation or to differentiate gene expression levels of transcripts. During analysis, the use of multiple probes per genotype or gene is combined with standard statistical methods to provide a transparent and robust conversion of probe intensities to biological information.

For gene expression, a variety of algorithms exist to summarize multiple probe intensities (including PM or MM probes in a probe set) into an aggregate signal estimate that is correlated to the relative abundance of the transcript in the experimental sample. Detection calls are made by

Affymetrix software through an arithmetic vote of probe pairs within a set designed to detect a specific transcript (GeneChip MAS5 and GCOS software). More widely used is an estimation of relative transcript abundance by a probe set signal and the trend has shifted away from median probe intensity-based algorithms such as MAS5 to probe modeling algorithms such as dCHIP (Schadt *et al.*, 2004), RMA (Irizarry *et al.*, 2003), and PLIER Estimation (Affymetrix Technical Note, 2005). The probe modeling analysis software considers intrinsic probe behavior to account for systematic nonsystematic biases, error, and allows for true replicate analysis.

It is still common for algorithms to use both PM and MM probes; however, PM-only algorithms are popular. Subtraction of MM probe intensity from the PM intensity or subtraction of modeled background estimates in PM-only analyses serves the same purpose, which is to estimate the true probe intensity by subtraction of background from the raw PM probe intensity. Raw probe intensity (PM or MM) is the sum of a true hybridization signal, specific cross-hybridization signal, nonspecific binding signal, and small amounts of signal generated by system noise. Background consists of everything but the true signal, and most would agree that, for an unbiased or true measurement of probe intensity, background must be subtracted from the raw perfect match probe intensity. For most Affymetrix expression measurements, subtraction of MM probe intensity is an accurate method to remove background. However, background can also be estimated in the absence of MM probes, for example, RMA. Continued discussion around this topic is indicative of the maturing thought in this area. Despite differences in precision, accuracy, and bias, most signal estimate-algorithms (PM, MM, or PM only) result in similar biological interpretations from the same data sets.

For genotypic sequence variation detection, a dynamic model-mapping algorithm has been developed by Affymetrix. In recent applications as few as six probe quartets (24 probes) are used to generate a genotype call and confidence score for all genotypes called. The dynamic model-based approach provides a highly accurate genotype calling method, is effective for SNP screening, is robust against changes in experimental conditions, is flexible to experiment designs, and is scalable to more SNPs (Di *et al.*, 2005). For resequencing, a unique base-calling algorithm derived from the work of Cutler *et al.* (2001) is employed.

## Current Applications

To date, there are more than 12,000 peer-reviewed publications based on microarray technology. Given that the use of microarrays became feasible in the late 1990s, it is easy to imagine how researchers from a myriad of fields have quickly leveraged this technology for a variety of

scientific endeavors. This section touches briefly on some examples of the applications of this technology, initially those based on gene expression profiling, and later on those based on whole-genome DNA analysis, or genotyping.

*Expression*

Gene expression profiling studies are performed with the goal of comparing tissues, tissue types, and cellular responses to a variety of stimuli such as altered growth conditions, cancer, and infectious processes to gain biological insight into basic biochemical pathways or molecular mechanisms of disease and its regulatory circuits. To date, whole-genome expression analysis has already helped scientists stratify disease, predict patient outcome, compare strains with varying virulence, study the relationship between host and parasite, and understand the affected molecular pathways of certain diseases.

Cancer research is one of the clinical fields in which microarrays have had an unquestionable impact. Whole genome expression profiles of cancerous cells have already allowed scientists to classify cancer subtypes, predict a patient's prognosis, select between alternate therapies, and even identify new classes of tumors. In a now classic example, Armstrong and colleagues (2002) studied the gene expression profile of cells isolated from patients with acute lymphoblastic leukemia (ALL) and acute myelogenous leukemia (AML). The current diagnostic methodology for these diseases includes a microscopic assessment of the morphology of the cells. Given that the morphology of these two cell types can sometimes be very similar, it is difficult to differentiate ALL from AML. Gene expression profiling of these two cell populations resulted in a unique molecular signature for each one. Even more surprising was this group's finding of a unique molecular signature that was distinct from that of ALL and of AML among the diagnosed patients. This unique signature corresponded to a new leukemia subtype, namely mixed lineage leukemia (MLL). Upon review of the MLL patients' clinical histories it was noted that not only had all of them failed standard ALL therapy, but they also had a poor prognosis compared to ALL patients. The latter was the first whole-genome transcriptome study that showed the effects that a translocation, such as that of the MLL gene, can specify a unique expression signature. This information has allowed this research group to expand on their studies and, for example, study the effect of this translocation on the hematopoietic properties of granulocyte/macrophage progenitors (Wang *et al*., 2005).

Another novel application of high-density gene expression microarrays includes the unbiased study of the transcription that occurs throughout the genome, independent of considerations such as open reading frames and annotations. Most genetic studies have focused on regions that code for proteins, which compose around 2% of the human genome. However, given the 3.1 billion base pairs in our genome, it now seems striking that the rest of the 98% of the genome would be nonfunctional. There are new and collaborative efforts, such as the ENCODE (Encyclopedia of DNA Elements) project that are now attempting to study those neglected regions of DNA. For example, a group led by Thomas Gingeras has used tiling arrays on chromosomes 21 and 22 and discovered that there is widespread transcription, that is, they found far more transcriptional activity than could be accounted for by known genes that express proteins. This work has raised the possibility that genome function and regulation is far more complex than previously thought (Kapranov *et al*., 2002).

Additionally, this group was able to identify transcription-binding sites on all the nonrepetitive sequences of these two chromosomes. They were able to identify a large and unexpected number of binding sites for three common transcription factors, Sp1, cMyc, and p53, distributed across chromosomes 21 and 22, suggesting a far more complex network of transcriptional regulation (Cawley *et al*., 2004). Most of the transcriptional binding sites were not located at random, but rather at the start of novel, noncoding transcripts, embedded within or between known coding genes. These novel transcripts are expressed simultaneously with the coding transcripts and are regulated similarly, suggesting that coding and noncoding counterparts function in concert. The group of transcripts may actually be the genetic functional unit. As additional transcription factor-binding sites are studied across the whole genome, a better understanding of the complex regulatory networks that govern genome function will undoubtedly be discovered.

In addition to gene expression profiling in cancer and in the basic study of the genome function, there is an extensive collection of exciting examples covering fields such as infectious disease (Apidianakis *et al*., 2005; Comer *et al*., 2005; Fan *et al*., 2005), cardiovascular disease (Boerma *et al*., 2005; Kong *et al*., 2005), and psychiatric disorders (Hekmat-Scafe et al., 2005; Iwamoto *et al*., 2005). Additionally, there are numerous examples of gene expression profiling applications based on a variety of different species, such as *Drosophila* (Girardot *et al*., 2004; Hekmat-Scafe *et al*., 2005), *Caenorhabditis elegans* (Dinkova *et al*., 2005; Reinke *et al*., 2004), and *Arabidopsis* (Davletova *et al*., 2005; Gomez-Mena *et al*., 2005). [The reader is invited to visit the Affymetrix Web site, where a database of all applications based on Affymetrix GeneChip technology are listed and classified.]

*Genotyping*

The following selection of applications is based on DNA analysis (genotyping) rather than on expression profiling. There are an estimated $3 \times 10^6$ nucleotide differences between any two humans, which only accounts for 1 out of every 1000 bases in the human genome. The ability to analyze 100,000–500,000 SNPs at once across the whole genome enables scientists to create detailed genetic maps and, among other things, discover the gene(s) responsible for disease. Additionally, pharmacogenomics—the use of genomic information to study a patient's response to drugs—has also been enabled by high-density DNA analysis microarrays. An understanding of the enzymatic mechanisms underlying the pharmacology and pharmacokinetics associated with every drug for each individual patient allows a more personalized approach to the administration of pharmacologic treatments and could potentially avoid the trial-and-error process currently employed for drug selection.

The GeneChip Mapping 100K Set is already bearing its scientific fruit. Klein and colleagues (2005) used this array set to study age-related macular degeneration (AMD), a major cause of blindness in the elderly. Even though family-based and candidate gene studies had been undertaken, causative genes or gene mutations were hard to find. Because performing an association study requires typing hundreds of thousands of SNPs, Klein and colleagues (2005) used the high-density microarrays to study the whole genome SNP variations between AMD patients and healthy subjects. This study led to the identification of an intronic and common variant in the complement factor gene (CFH) that puts patients at higher risk for AMD. This gene is located on chromosome 1, consistent with chromosomal regions previously identified as being linked to AMD. The identification of this risk factor may be used in the future, for example, for diagnostics and for preventive therapies in patients at high risk of AMD.

Many diseases include an alteration in the normal number of chromosomes or chromosome segments, as well as mutations, deletions, or amplifications of more succinct sequence fragments. For example, Down syndrome results from a trisomy (triplication) of chromosome 21 (Korenberg, 1993), while a loss of a fragment of chromosome 17 (17q25.1) is characteristic of ovarian cancers (Presneau *et al*., 2005). Information stemming from the characterization of this altered DNA copy number is crucial to the understanding of the mechanisms underlying the disease. There are two experimental approaches for DNA analysis that have been used to study the chromosomal stability of cancer biopsies: chromosomal copy number and loss of heterozygosity (LOH). Commonly used methods for addressing these issues include fluorescence *in situ* hybridization, Northern blotting, microsatellites, and

comparative genomic hybridization (CGH), among others. However, high-density SNP microarrays enable scientists to interrogate the genome at far higher resolution than these techniques allow.

For example, it is known that chromosomal amplifications and deletions frequently contribute to cancer. Loss of heterozygosity refers to the loss of one allele caused by either a mutation or a deletion resulting in homozygosity. When this occurs at a tumor suppressor gene locus, for example, it may result in a neoplastic transformation.

One of the first studies to use SNP arrays to study genomic alteration such as LOH was conducted by the Meyerson group at the Dana Farber Cancer Institute. Their initial studies validated the large-scale genotyping of SNPs on small cell carcinoma cells on the first-generation high-density SNP array and showed that the loss of LOH data was consistent with previous CGH results (Lindblad-Toh *et al*., 2000).

Lung cancer is one of the leading causes of cancer deaths in the United States (Jemal *et al*., 2003). Many studies have focused on the LOH patterns of lung cancer; however, this is such a complex disease that a correlation between LOH analyses and clinical outcome has been challenging. Given that SNPs occur at a frequency of once every thousand base pairs, the study of their identity allows a higher LOH mapping resolution. A group of researchers at Harvard studied the LOH patterns in human cancer cell lines. By using the 10K SNP array, in conjunction with the dChipSNP informatics software package, these investigators were able to compare and confirm LOH patterns to those obtained previously with microsatellites. Moreover, this effort also resulted in the identification of previously undetected LOH regions that were smaller and unattainable by other methods (Janne *et al*., 2004).

More recent studies stemming from this research group detected genomic regions with an altered DNA copy number and LOH. By hybridizing breast and lung carcinoma cell DNAs and measuring the fluorescence intensity of the allele-specific hybridization to certain segments, genomic amplifications and deletions were identified, as well as some LOH events. Some of these alterations were consistent with previous data, although some were novel, and could serve as new diagnostic markers (Zhao *et al*., 2004). Studies such as the examples given earlier demonstrate that the combination of SNP analysis and copy number analysis provides insight into the genetic alterations and molecular mechanisms responsible for cancer. The advent of technologies such as the GeneChip Mapping 100K Set enables researchers to study whole genome chromosomal copy number changes, as well as LOH markers simultaneously, at an unprecedented efficiency.

## Advancing the Future of Genomics

Since their inception, high-density microarrays have followed the same trend as computer microprocessors. In 1965, Moore predicted that the power of microprocessors would double every 18 months. This trend has held true for the computer industry, where much faster and smaller central processing units are constantly being produced. High-density oligonucleotide arrays have evolved similarly. More and more genetic information is being included into a smaller and smaller surface area. This allows scientists to analyze vast amounts of genetic information at an unprecedented efficiency. This trend, in combination with the wealth of information generated by the sequencing of the human and other genomes, has generated a unique opportunity in the advancement of clinical and life sciences research. Global views of the genome will undoubtedly accelerate the understanding of complex diseases such as psychiatric and cardiovascular ailments and drug response.

In addition to feature-size reduction, the overall microarray platforms are changing in other ways. High-throughput automation systems are currently being developed that provide the convenience of hybridizing 96 arrays at a time. Microarray systems have also been developed for diagnostic purposes, an example being the Roche system for the detection of polymorphisms in cytochrome P450 genes. This gene family controls how individuals respond to different drugs, and knowledge of an individual's genotype should aid in prescribing proper doses, reducing side effects. Controls and standard practices are also developing with microarrays in mind. The establishment of controls and standard practices will allow greater acceptance of microarray assays into the clinical and diagnostic fields.

In a little over a decade the microarray has evolved from a research publication to a mainstream tool of life science research. At the time of completion for this manuscript, several new applications are being introduced commercially: a 500K SNP genotyping assay and an exon-based expression assay. What remains certain is that microarray assays and technology will continue to evolve in the future and further expand from life sciences research into the clinical and diagnostic communities.

## Acknowledgments

The authors thank Sean Walsh and Glenn McGall for critically reviewing this manuscript and Andy Lau and Dan Bartell for providing the graphics. In addition, the authors thank their many colleagues at Affymetrix for contributing the ideas, methods, and products that made this review possible.

## References

Affymetrix Technical Note (2005). Guide to Probe Logarithmic Intensity Error (PLIER) Estimation, http://www.affymetrix.com.

Apidianakis, Y., Mindrinos, M. N., Xiao, W., Lau, G. W., Baldini, R. L., Davis, R. W., and Rahme, L. G. (2005). Profiling early infection responses: *Pseudomonas aeruginosa* eludes host defenses by suppressing antimicrobial peptide gene expression. *Proc. Natl. Acad. Sci. USA* **102,** 2573–2578.

Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., Sallan, S. E., Lander, E. S., Golub, T. R., and Korsmeyer, S. J. (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat. Genet.* **30,** 41–47.

Boerma, M., van der Wees, C. G., Vrieling, H., Svensson, J. P., Wondergem, J., van der Laarse, A., Mullenders, L. H., and van Zeeland, A. A. (2005). Microarray analysis of gene expression profiles of cardiac myocytes and fibroblasts after mechanical stress, ionising or ultraviolet radiation. *BMC Genom.* **6,** 6.

Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A. J., Wheeler, R., Wong, B., Drenkow, J., Yamanaka, M., Patel, S., Brubaker, S., Tammana, H., Helt, G., Struhl, K., and Gingeras, T. R. (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116,** 499–509.

Comer, J. E., Galindo, C. L., Chopra, A. K., and Peterson, J. W. (2005). GeneChip analyses of global transcriptional responses of murine macrophages to the lethal toxin of *Bacillus anthracis*. *Infect. Immun.* **73,** 1879–1885.

Cutler, D. J., Zwick, M. E., Carrasquillo, M. M., Yohn, C. T., Tobin, K. P., Kashuk, C., Mathews, D. J., Shah, N. A., Eichler, E. E., Warrington, J. A., and Chakravarti, A. (2001). High-throughput variation detection and genotyping using microarrays. *Genome Res.* **11,** 1913–1925.

Davletova, S., Rizhsky, L., Liang, H., Shengqiang, Z., Oliver, D. J., Coutu, J., Shulaev, V., Schlauch, K., and Mittler, R. (2005). Cytosolic ascorbate peroxidase 1 is a central component of the reactive oxygen gene network of Arabidopsis. *Plant Cell* **17,** 268–281.

Di, X., Matsuzaki, H., Webster, T. A., Hubbell, E., Liu, G., Dong, S., Bartell, D., Huang, J., Chiles, R., Yang, G., Shen, M. M., Kulp, D., Kennedy, G. C., Mei, R., Jones, K. W., and Cawley, S. (2005). Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays. *Bioinformatics* **21,** 1958–1963.

Dinkova, T. D., Keiper, B. D., Korneeva, N. L., Aamodt, E. J., and Rhoads, R. E. (2005). Translation of a small subset of *Caenorhabditis elegans* mRNAs is dependent on a specific eukaryotic translation initiation factor 4E isoform. *Mol. Cell. Biol.* **25,** 100–113.

Fan, W., Bubman, D., Chadburn, A., Harrington, W. J., Jr., Cesarman, E., and Knowles, D. M. (2005). Distinct subsets of primary effusion lymphoma can be identified based on their cellular gene expression profile and viral association. *J. Virol.* **79,** 1244–1251.

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

Girardot, F., Monnier, V., and Tricoire, H. (2004). Genome wide analysis of common and specific stress responses in adult *Drosophila melanogaster*. *BMC Genom.* **5,** 74.

Gomez-Mena, C., de Folter, S., Costa, M. M., Angenent, G. C., and Sablowski, R. (2005). Transcriptional program controlled by the floral homeotic gene AGAMOUS during early organogenesis. *Development* **132,** 429–438.

Hardenbol, P., Baner, J., Jain, M., Nilsson, M., Namsaraev, E. A., Karlin-Neumann, G. A., Fakhrai-Rad, H., Ronaghi, M., Willis, T. D., Landegren, U., and Davis, R. W. (2003).

Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat. Biotechnol.* **21,** 673–678.

Hekmat-Scafe, D. S., Dang, K. N., and Tanouye, M. A. (2005). Seizure suppression by gain-of-function escargot mutations. *Genetics* **169,** 1477–1493.

Huang, J., Wei, W., Zhang, J., Liu, G., Bignell, G. R., Stratton, M. R., Futreal, P. A., Wooster, R., Jones, K. W., and Shapero, M. H. (2004). Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum. Genom.* **1,** 287–299.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31,** e15.

Iwamoto, K., Bundo, M., and Kato, T. (2005). Altered expression of mitochondria-related genes in postmortem brains of patients with bipolar disorder or schizophrenia, as revealed by large-scale DNA microarray analysis. *Hum. Mol. Genet.* **14,** 241–253.

Janne, P. A., Li, C., Zhao, X., Girard, L., Chen, T. H., Minna, J., Christiani, D. C., Johnson, B. E., and Meyerson, M. (2004). High-resolution single-nucleotide polymorphism array and clustering analysis of loss of heterozygosity in human lung cancer cell lines. *Oncogene* **23,** 2716–2726.

Jemal, A., Murray, T., Samuels, A., Ghafoor, A., Ward, E., and Thun, M. J. (2003). Cancer statistics, 2003. *CA Cancer J. Clin.* **53,** 5–26.

Kapranov, P., Cawley, S. E., Drenkow, J., Bekiranov, S., Strausberg, R. L., Fodor, S. P., and Gingeras, T. R. (2002). Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296,** 916–919.

Kennedy, G. C., Matsuzaki, H., Dong, S., Liu, W. M., Huang, J., Liu, G., Su, X., Cao, M., Chen, W., Zhang, J., Liu, W., Yang, G., Di, X., Ryder, T., He, Z., Surti, U., Phillips, M. S., Boyce-Jacino, M. T., Fodor, S. P., and Jones, K. W. (2003). Large-scale genotyping of complex DNA. *Nat. Biotechnol.* **21,** 1233–1237.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., and Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science* **308,** 385–389.

Kong, S. W., Bodyak, N., Yue, P., Liu, Z., Brown, J., Izumo, S., and Kang, P. M. (2005). Genetic expression profiles during physiological and pathological cardiac hypertrophy and heart failure in rats. *Physiol. Genom.* **21,** 34–42.

Korenberg, J. R. (1993). Toward a molecular understanding of Down syndrome. *Prog. Clin. Biol. Res.* **384,** 87–115.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921.

Lindblad-Toh, K., Tanenbaum, D. M., Daly, M. J., Winchester, E., Lui, W. O., Villapakkam, A., Stanton, S. E., Larsson, C., Hudson, T. J., Johnson, B. E., Lander, E. S., and Meyerson, M. (2000). Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. *Nat. Biotechnol.* **18,** 1001–1005.

Lipshutz, R. J., Morris, D., Chee, M., Hubbell, E., Kozal, M. J., Shah, N., Shen, N., Yang, R., and Fodor, S. P. (1995). Using oligonucleotide probe arrays to access genetic diversity. *Biotechniques* **19,** 442–447.

Matsuzaki, H., Dong, S., Loi, H., Di, X., Liu, G., Hubbell, E., Law, J., Berntsen, T., Chadha, M., Hui, H., Yang, G., Kennedy, G. C., Webster, T. A., Cawley, S., Walsh, P. S., Jones, K. W., Fodor, S. P., and Mei, R. (2004a). Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat. Methods* **1,** 109–111.

Matsuzaki, H., Loi, H., Dong, S., Tsai, Y. Y., Fang, J., Law, J., Di, X., Liu, W. M., Yang, G., Liu, G., Huang, J., Kennedy, G. C., Ryder, T. B., Marcus, G. A., Walsh, P. S., Shriver, M. D., Puck, J. M., Jones, K. W., and Mei, R. (2004b). Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Res.* **14,** 414–425.

McGall, G. H., and Christians, F. C. (2002). High-density genechip oligonucleotide probe arrays. *Adv. Biochem. Eng. Biotechnol.* **77,** 21–42.

Mei, R., Hubbell, E., Bekiranov, S., Mittmann, M., Christians, F. C., Shen, M.-M., Lu, G., Fang, J., Liu, W.-M., Ryder, T., Kaplan, P., Kulp, D., and Webster, T. A. (2003). Probe selection for high density oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* **100,** 11237–11242.

Middleton, F. A., Pato, M. T., Gentile, K. L., Morley, C. P., Zhao, X., Eisener, A. F., Brown, A., Petryshen, T. L., Kirby, A. N., Medeiros, H., Carvalho, C., Macedo, A., Dourado, A., Coelho, I., Valente, J., Soares, M. J., Ferreira, C. P., Lei, M., Azevedo, M. H., Kennedy, J. L., Daly, M. J., Sklar, P., and Pato, C. N. (2004). Genomewide linkage analysis of bipolar disorder by use of a high-density single-nucleotide-polymorphism (SNP) genotyping assay: A comparison with microsatellite marker assays and finding of significant linkage to chromosome 6q22. *Am. J. Hum. Genet.* **74,** 886–897.

Moore, G. (1965). Cramming more components onto integrated circuits. *Electronics* **38,** 114–117.

Presneau, N., Dewar, K., Forgetta, V., Provencher, D., Mes-Masson, A. M., and Tonin, P. N. (2005). Loss of heterozygosity and transcriptome analyses of a 1.2 Mb candidate ovarian cancer tumor suppressor locus region at 17q25.1–q25.2. *Mol. Carcinog.* **43,** 141–154.

Puffenberger, E. G., Hu-Lince, D., Parod, J. M., Craig, D. W., Dobrin, S. E., Conway, A. R., Donarum, E. A., Strauss, K. A., Dunckley, T., Cardenas, J. F., Melmed, K. R., Wright, C. A., Liang, W., Stafford, P., Flynn, C. R., Morton, D. H., and Stephan, D. A. (2004). Mapping of sudden infant death with dysgenesis of the testes syndrome (SIDDT) by a SNP genome scan and identification of TSPYL loss of function. *Proc. Natl. Acad. Sci. USA* **101,** 11689–11694.

Reinke, V., Gil, I. S., Ward, S., and Kazmer, K. (2004). Genome-wide germline-enriched and sex-biased expression profiles in *Caenorhabditis elegans*. *Development* **131,** 311–323.

Schadt, E. E., Edwards, S. W., GuhaThakurta, D., Holder, D., Ying, L., Svetnik, V., Leonardson, A., Hart, K. W., Russell, A., Li, G., Cavet, G., Castle, J., McDonagh, P., Kan, Z., Chen, R., Kasarskis, A., Margarint, M., Caceres, R. M., Johnson, J. M., Armour, C. D., Garrett-Engele, P. W., Tsinoremas, N. F., and Shoemaker, D. D. (2004). A comprehensive transcript index of the human genome generated using microarrays and computational approaches. *Genome Biol.* **5,** R73.

Sellick, G. S., Garrett, C., and Houlston, R. S. (2003). A novel gene for neonatal diabetes maps to chromosome 10p12.1-p13. *Diabetes* **52,** 2636–2638.

Steinman, L., and Zamvil, S. (2003). Transcriptional analysis of targets in multiple sclerosis. *Nat. Rev. Immunol.* **3,** 483–492.

Van Gelder, R. N., von Zastrow, M. E., Yool, A., Dement, W. C., Barchas, J. D., and Eberwine, J. H. (1990). Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proc. Natl. Acad. Sci. USA* **87,** 1663–1667.

Wang, J., Iwasaki, H., Krivtsov, A., Febbo, P. G., Thorner, A. R., Ernst, P., Anastasiadou, E., Kutok, J. L., Kogan, S. C., Zinkel, S. S., Fisher, J. K., Hess, J. L., Golub, T. R., Armstrong, S. A., Akashi, K., and Korsmeyer, S. J. (2005). Conditional MLL-CBP targets GMP and models therapy-related myeloproliferative disease. *EMBO J.* **24,** 368–381.

Yeoh, E. J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., Mahfouz, R., Behm, F. G., Raimondi, S. C., Relling, M. V., Patel, A., Cheng, C., Campana, D., Wilkins, D., Zhou, X., Li, J., Liu, H., Pui, C. H., Evans, W. E., Naeve, C., Wong, L., and Downing, J. R. (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* **1,** 133–143.

Zhao, X., Li, C., Paez, J. G., Chin, K., Janne, P. A., Chen, T. H., Girard, L., Minna, J., Christiani, D., Leo, C., Gray, J. W., Sellers, W. R., and Meyerson, M. (2004). An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res.* **64,** 3060–3071.

# [2] The Agilent *In Situ*-Synthesized Microarray Platform

*By* Paul K. Wolber, Patrick J. Collins, Anne B. Lucas, Anniek De Witte, and Karen W. Shannon

## Abstract

Microarray technology has become a standard tool in many laboratories. Agilent Technologies manufactures a variety of catalog and custom long-oligonucleotide (60-mer) microarrays that can be used in multiple two-color microarray applications. Optimized methods and techniques have been developed for two such applications: gene expression profiling and comparative genomic hybridization. Methods for a third technique, location analysis, are evolving rapidly. This chapter outlines current best methods for using Agilent microarrays, provides detailed instructions for the most recently developed techniques, and discusses solutions to common problems encountered with two-color microarrays.

## Introduction

During the last decade, microarrays have evolved from a promising technology for exploring a variety of genomic problems (Hughes *et al.*, 2001; Kuhn *et al.*, 2001; Miki *et al.*, 2001; Nacht *et al.*, 1999) into a workhorse technology for investigating important questions in cancer research (Chang

METHODS IN ENZYMOLOGY, VOL. 410 0076-6879/06 $35.00
Copyright 2006, Elsevier Inc. All rights reserved.
DOI: 10.1016/S0076-6879(06)10002-6