1
Introduction to DNA Microarrays

Chris Seidel

1.1
Introduction

All living organisms contain DNA, a molecule that encodes all the information required for the development and functioning of an organism. Finding and deciphering the information encoded in DNA, and understanding how such a simple molecule can give rise to the amazing biological diversity of life, is a goal shared in some way by all life scientists. Microarrays provide an unprecedented view into the biology of DNA, and thus a rich way to examine living systems. DNA is a physical molecule that is able to encode information in a linear structure. Cells express information from different parts of this structure in a context-dependent fashion. DNA encodes for genes, and regulatory elements control whether genes are on or off. For instance, all the cells of the human body contain the same DNA, yet there are hundreds of different types of cells, each expressing a unique configuration of genes from the DNA. In this regard, DNA could be described as existing in some number of states. Microarrays are a tool used to read the states of DNA.

Microarrays have had a transforming effect on the biological sciences. In the past, biologists had to work very hard to generate small amounts of data that could be used to explore a hypothesis with one observation at a time. With the advent of microarrays, individual experiments generate thousands of data points or observations. This turns the experiment from a hypothesis-driven endeavor to a hypothesis-generating endeavor because every experiment sheds light across an entire terrain of gene expression, letting relevant genes reveal themselves, often in surprising ways. Concomitantly, what used to be primarily wet science becomes in large part information science. The highly parallel nature of microarrays that are used to make biological observations signifies that most experiments generate more information than the experimenter could possibly interpret. Indeed, from a statistical point of view, every gene measured on a microarray is an independent variable in a highly parallel experiment. The number of hypotheses to which the data may or may not lend support cannot be known in advance. To take advantage of the excess

Analysis of Microarray Data: A Network-Based Approach. Edited by F. Emmert-Streib and M. Dehmer

ISBN: 978-3-527-31822-3

information in microarray data, repositories have been set up in which people can deposit their experiments, thus making them available to a wide community of researchers with questions to explore (Gene Expression Ominbus GEO, ArrayExpress).

1.1.1 The Genome is an Information Scaffold

Microarrays measure events in the genome. An event may be the transcription of a gene, the binding of a protein to a segment of the DNA, the presence or absence of a mutation, a change in the copy number of a locus, a change in the methylation state of the DNA, or any of a number of states or activities that are associated with DNA or RNA molecules. As a genomic readout, microarrays identify where these events occur.

The idea that one can accurately describe the genome, let alone measure its activity in a comprehensive way, is a relatively novel concept. Several factors have led to the recent enhancement and blending of molecular biology into a field called genomics. The first is genome-sequencing projects. Today, sequencing a genome is considered a routine activity. However, in the late 1980s when sequencing the human genome was first suggested as a serious endeavor, the community was divided. Given the sequencing technology available at the time, the project looked as if it would consume colossal resources over a long time frame that many thought could be put to better use on more practical projects. However, visionaries were banking on two precepts: once given the mandate, the technology would transform itself and new sequencing methods would be invented that would increase the rate of sequence accumulation. The second aspect is that the finished project, full genome sequences, would be a public gold mine of a resource that would pay off for all biologists. Both of these assumptions have come to fruition. Genome sequences accumulate at rates few imagined possible. Biologists can expect the sequence of their model organism to exist in GenBank or to be in someone's sequencing pipeline. More important, having a map of the full genomic sequence of an organism has transformed the way biology is studied.

DNA gives rise to the organism and so is a scaffold for information. The genomic map is like a landscape of code, openly visible to all and for anyone to figure out. Through experimentation, often involving microarrays, DNA is annotated with functional information. In addition, the large-scale sequencing effort served as a kind of space program for biology, whereby the genome was a new frontier [1]. It made possible previously unforeseen possibilities and conceptually paved the way for a host of parallel analysis methods. The unveiling of a unified map begged the creation of microarrays, as well as other large genome-sized projects, such as the systematic deletion of every yeast gene [2], the systematic fusion of every yeast promoter to a reporter gene [3], and many other similar projects [4]. As the invention of the telescope changed how we view the universe, microarrays have changed the way we view the genome.

1.1.2 Gene Expression is Detected by Hybridization

The purpose of a microarray is to examine expression of multiple genes simultaneously in response to some biological perturbation. More generally, a microarray serves to interrogate the concentrations of molecules in a complex mixture and thus can serve as a powerful analytical tool for many kinds of experiments. To understand how this occurs, it may be useful to review the structure of DNA and examine how the unique structure of this molecule plays a role in identifying itself.

Although DNA is remarkably informationally complex, the general structure of the molecule is really quite simple. DNA is made up of four chemical building blocks called bases: adenine, cytosine, guanosine, and thymidine (abbreviated respectively as A, C, G, or T). As individual subunits these building blocks are also referred to as nucleotides. A strand of DNA consists of a sugar phosphate backbone to which these bases are covalently linked such that they form a series. Because these four bases can form sequences, it is possible to use them to encode information based on their patterns of occurrence. Indeed, from an information point of view, DNA has a potential data density of 145 million bits per inch and has been considered as a substrate for computation whereby the sequences are referred to as software [5,6].

Like strings of text in a book, the sequences that make up a strand of DNA have directionality such that information can be encoded in a given direction. The amount of DNA, and thus the amount of sequence, varies from organism to organism. For instance, the microorganism *Escherichia coli* has 4.5 million bases of sequence, whereas human cells have about 3 billion bases. Exactly how much biological information is encoded in these sequences is unknown, representing one of the deepest mysteries of biology, but microarrays provide a way to gain clues.

Cellular DNA most often consists not just of one strand but of two strands antiparallel to each other. The two strands are hydrogen bonded together by interactions between the bases, forming a structure in the cell. The structure is helical, similar to a spiral staircase in which the bases are attached to each side and interact in a plane to form the steps of the staircase. Besides the hydrogen bonds between the bases of opposite strands, the overlapping and proximity of the bases to each other lead to a second kind of noncovalent force called a stacking interaction that contributes to the stability of the double-stranded structure.

The bases of one strand interact with the bases of the other strand according to a set of pairing rules, such that A pairs with T and C pairs with G. Thus, if one knows the sequence of one strand, by definition, one then knows the sequence of the opposite strand. This property has profound consequences in the study of biology. It is also what the cell uses to replicate itself. As the interaction between the bases is noncovalent, consisting only of hydrogen bonds, the strands can essentially be melted apart and separated, thus opening the way for a copying mechanism to read each single strand and re-create the second complementary strand for each half of the pair, resulting in a new double-stranded molecule for each cell. This is also the mechanism by which cells express genes. The strands are opened by the gene expression machinery so that some number of RNA copies of a gene can

be synthesized. The RNA transcript has the same sequence as the gene with the exception that uracil (U) replaces T, though the hybridization pairing rules remain the same (U and T can both pair with A).

This property of complementarity is also what is used for measuring gene expression on microarrays. Just as energy can melt strands apart and separate them into single molecules, the process is reversible such that single strands that are complementary to each other can come together and reanneal to form a double-stranded complex [109]. This process is called hybridization and is the basis for many assays or experiments in molecular biology. In the cell, hybridization is at the center of several biological processes, whereas in the lab complementarity is identity and thus hybridization is at the center of many *in vitro* reactions and analytical techniques. The molecules can come from completely different sources, but if they match, they will hybridize.

1.1.2.1 Hybridization is Used to Measure Gene Expression

Hybridization-based methods for measuring nucleic acid sequences have been used as a standard technique for decades [7]. The most widely accepted and understood standard for measuring gene expression is a hybridization-based assay called the Northern blot [8]. Microarrays are derived from blotting assays. Thus, it is useful to understand the basics of how they work and explain the origin of some of the terminology. Northern blots are also a common method for verifying the results from a microarray study.

The purpose of a Northern blot is to measure the size and abundance of RNA transcribed from a gene. To compare the expression of a gene under various conditions, extracts of cellular RNA are prepared from different batches of cells or tissues and fractionated by size on agarose gels prior to being transferred to a nylon membrane. The membrane binds nucleic acid quantitatively and preserves the pattern of RNA from the gel. The result is a porous membrane with RNA from different samples in discrete regions of the membrane. To quantify the expression level of a gene of interest in each sample, a radioactive DNA probe complementary to the gene of interest is prepared. Upon applying the probe to the membrane, hybridization will occur between the probe and the messenger ribonucleic acid (mRNA) from the gene of interest. In this way, both the location of the RNA within the fractionated sample and the relative abundance will be revealed. The amount of radioactivity captured by each sample on the membrane is a readout of the relative expression level of the gene of interest between the two samples.

In this scenario, the mRNA sample on the membrane is a complex mixture of molecules of unknown concentrations. Somewhere in the discrete region occupied by the sample RNA is a target transcript, which will be revealed when the probe hybridizes to it. Thus, the probe is known, and the target is what is revealed by hybridization. Although this process is robust, it is limited to questions of one target at a time. To measure a second gene, the membrane must be washed under conditions that disrupt hybridization and strip the membrane of radioactivity, so that it can be reprobed for another target.

1.1.2.2 Microarrays Provide a New Twist to an Old Technique

Microarrays provide a way to measure many genes at once by reversing and parallelizing this process. Instead of labeling what is known and probing a complex mixture of unknowns to highlight a single target, microarray methods label the complex mixture that is in a solution and utilize a two-dimensional surface of known molecules or probes in discrete locations, as a readout. In this context, the cellular RNA mixture is labeled, creating a labeled complex mixture of unknowns. Complementarity between target molecules in the complex mixture and probes arrayed on the solid surface will result in annealing and hybridization, thus capturing the labeled molecules on the surface. Unhybridized molecules are washed away prior to quantification.

Although this technique has been used in the past with membranes to create macroarrays, usually for screening libraries of clones [9], including identification of differentially expressed genes [10], the success of microarrays has to do with miniaturization. The interaction between a probe and its target is a binding assay. As many genes are expressed at low levels, sensitivity is often an issue. In a finding most would consider counterintuitive, a series of studies on binding assays in the mid-1980s found that decreasing the size of a binding target does not reduce the sensitivity or accuracy of an assay and actually leads to an increase in sensitivity [11,12]. Thus, miniaturization of the assay is possible. As the size of a spot decreases, the fractional occupancy of the immobilized probe is dependent only on the concentration of the corresponding target molecule in solution. This is the principle by which microarrays operate.

Other developments have contributed to efficient parallelization through miniaturization. The use of glass substrates instead of porous filters as a medium for capturing nucleic acid in an array format means that small hybridization volumes can be used. The rate of hybridization is concentration dependent. Thus, smaller volumes mean higher rates of hybridization and increased sensitivity. In addition, solid glass surfaces have low inherent fluorescence, allowing the use of fluorescent dyes for labeling samples, and along with methods adapted from confocal microscopy, they allow for efficient quantification of a highly parallel miniaturized assay.

Several groups developed and applied these concepts, combining miniaturization, robots, and genomes to develop new assays for the measurement of gene expression in parallel [13,14].

1.2 Types of Arrays

There are predominantly three kinds of microarray technologies in widespread use among most laboratories: spotted microarrays consisting of presynthesized oligos or PCR products robotically deposited onto a surface, Affymetrix GeneChips composed of relatively short oligonucleotides synthesized on a chip surface, and other *in situ* synthesis platforms such as arrays made by Agilent and NimbleGen. Although each technology effectively serves as a genomic readout, each has unique characteristics

that offer advantages or disadvantages in a given context. Parallel forms of measuring DNA and RNA will continue to change and evolve; however, these three platforms are currently the most ubiquitous.

1.2.1 Spotted Microarrays

Spotted microarrays were the first widely available array platform and continue to enjoy broad use. Originating in the laboratory of Pat Brown, they consist of glass microscope slides onto which libraries of PCR products or long oligonucleotides are printed using a robot equipped with nibs capable of wicking up DNA from microtiter plates and depositing it onto the glass surface with micron precision [13,15]. Since their inception, demand for microarrays has exceeded availability. Because the Brown laboratory expended effort in every aspect of distributing the technology, including plans to build the robot and all protocols required for array manufacture and use, many academic laboratories invest resources into producing these arrays locally. This includes building or purchasing a robot, as well as performing PCR or oligo design and synthesis to create probes for spotting onto glass. The basic principle by which the arrays function is fairly simple, and all the reagents required are available to most researchers with some initial investment. However, apart from praising the benefits of putting technology into the hands of researchers, the reason for highlighting this aspect of spotted arrays is to point out the nonuniform nature of spotted microarrays. Because there is not one manufacturer, one source of materials, or a uniform method of production, variability exists among batches of microarrays and must be considered when planning experiments or when comparing experiments from different array sources.

Spotted microarrays are primarily a comparative technology. They are used to examine relative concentrations of targets between two samples. Complex samples to be compared are labeled with uniquely colored fluorescent tags before being mixed together and allowed to compete for hybridization to the microarray spots (see [16,17] for review). In this way, differences between the samples are observed on a per spot basis because the fractional occupancy of the spot hybridized by each sample reflects the relative concentration of that gene or target in the original complex mixture. Thus, for any probe on the microarray, one gets a readout of the relative concentrations of the target between the two input samples. For this reason, spotted microarrays are often called two-color or two-sample arrays.

1.2.2 Affymetrix GeneChips

Affymetrix GeneChips are the most ubiquitous and long-standing commercial array platform in use. The arrays consist of 25-mer oligonucleotides synthesized *in situ* on the surface of a glass chip. A photolithography mask, similar to that used to construct semiconductor chips, is used to control light-directed DNA synthesis chemistry such that oligo sequences are built up one nucleotide at a time at defined locations

on a solid substrate or glass chip [18,19]. Current chips contain 6.5 million unique probes in an area of 1.28 cm^2. The highly precise nature of the lithographic method allows the construction of compact matrices of square patches of probes.

Instead of using a single sequence to probe expression of each gene, as would be common for a spotted array, Affymetrix employs a set of probes to measure expression of a gene. Probe sets contain two types of probes to measure the gene of interest, perfect match (PM) and mismatch (MM) probes. Perfect match probes are chosen to match the gene exactly and are designed against an exemplar sequence representing the gene. Although each probe is unique, probes may occasionally overlap. Mismatch probes are identical to the perfect match probes except that they contain a single base mismatch in the center of the probe. A single mismatch in a short sequence such as a 25-mer is very disruptive to hybridization. The purpose of the mismatch probe is to serve as a negative control for background hybridization. A typical probe set contains 11 perfect match probes and 11 mismatch probes. The positioning of probes for a single gene on the array is chosen by a random process to protect against local hybridization artifacts that could otherwise affect all the probes for a gene if they were clustered together. As most spotted arrays use only one probe per gene, local hybridization artifacts can be a problem.

To read the expression level of a gene, the perfect match and mismatch probe readings must be summarized. Although Affymetrix has a standard method for summarizing 22 readouts to obtain a single number for gene expression [20], many approaches are available [21].

Affymetrix GeneChips are single sample microarrays (also known as one color or one channel). These arrays measure the relative abundance of every gene in a single sample. In this way, one can examine whether one gene is expressed at a higher or lower level than some other gene in the same sample. If samples are to be compared, a separate chip must be performed for each sample, and the data adjusted by scaling or normalization before comparison.

1.2.2.1 Other *In Situ* Synthesis Platforms

Apart from Affymetrix, two alternative *in situ* synthesis methods exist by which oligonucleotides are built up one nucleotide at a time in successive steps to create probes of length 25–60 nucleotides long [108]. These methods are almost exclusively commercial and different companies take different approaches. Although Affymetrix uses a mask-based photolithographic process to control light-directed DNA synthesis, an alternative method employed by NimbleGen makes use of small rotating mirrors to control light and accomplish a similar task [22]. This approach is called Maskless Photolithography, and uses technology developed by Texas Instruments for projection televisions in which arrays of digitally controlled micromirrors can be used to direct light. In combination with light activated chemistry, light of the appropriate intensity and wavelength can be actuated in patterns required to build up any series of nucleotides into an oligonucleotide on a solid surface [23,24].

The NimbleGen approach has two great advantages over the method worked out by Affymetrix. The first is that it does not require a mask. To build an array of

different oligonucleotides of length N requires a series of $4N$ synthesis steps. Thus, to build a library of unique 25-mers on a surface requires 100 chemical synthesis steps. For Affymetrix, a unique photolithography mask is required to control the chemistry at each step. These masks are expensive to construct; thus, the arrays are very costly. In addition, once a set of masks is constructed, it describes only a single array design. Changing the design requires a whole new set of masks. However, changing a pattern of micromirrors under electronic control is very easy; thus, each array produced by NimbleGen can have a different design.

The second alternative *in situ* synthesis approach to array construction uses traditional oligo synthesis chemistry, but the method of controlling base addition is novel. Ink-jet technology, developed by Hewlett Packard for consumer printers, has been adapted to control the liquid precursors of DNA synthesis. Agilent, a spin-off of Hewlett Packard, uses this technology to synthesize 60-mer oligos on glass slides [22,25]. Like ink-jet printing itself, this technology is very flexible. Every array can be customized and thus possess unique content.

The flexibility to change the design of an array easily is both a blessing and a curse. The positive aspect is that one can easily change the array design to explore the genome or expression space as required by the experiment. The negative aspect is that data analysis becomes more cumbersome, in general, because one can easily change the probes used to represent a gene from array to array, as well as the content of the array from experiment to experiment. When everyone is using the same chip, as in chips mass produced by Affymetrix, comparisons between data sets are fairly easy. If every chip is unique, comparison between data sets becomes difficult.

Another positive aspect of arrays created using *in situ* synthesis methods is that they do not depend on libraries of clones or molecules created elsewhere. Instead, the content is freshly created with each array. This is good because with spotted arrays one never knows the history of the library, or how many times it has been used to create arrays. However, quality control for *in situ* synthesized arrays remains obscure.

1.2.2.2 Uses of Microarrays

An example of the power of microarrays can be seen in Figure 1.1. Four sets of cells purified in triplicate by the abundance of two cell surface markers exhibit different differentiation fates yet are otherwise indistinguishable. However, microarray analysis and simple clustering of differentially expressed genes reveal previously unknown differences.

As a genomic readout, microarrays can serve many purposes, and novel applications continue to emerge. A common application of microarrays has been the measurement of gene expression, from characterizing cells and processes [26–28] to clinical applications such as tumor classification [29,30]. Another very common use of microarrays is in genotyping and the measurement of genetic variation [31,32].

As microarrays are capable of quantifying complex mixtures of DNA or RNA fragments derived from a variety of experiments, they have been used for a number of purposes. For instance, RNA interacts with many proteins in the cell and is a part

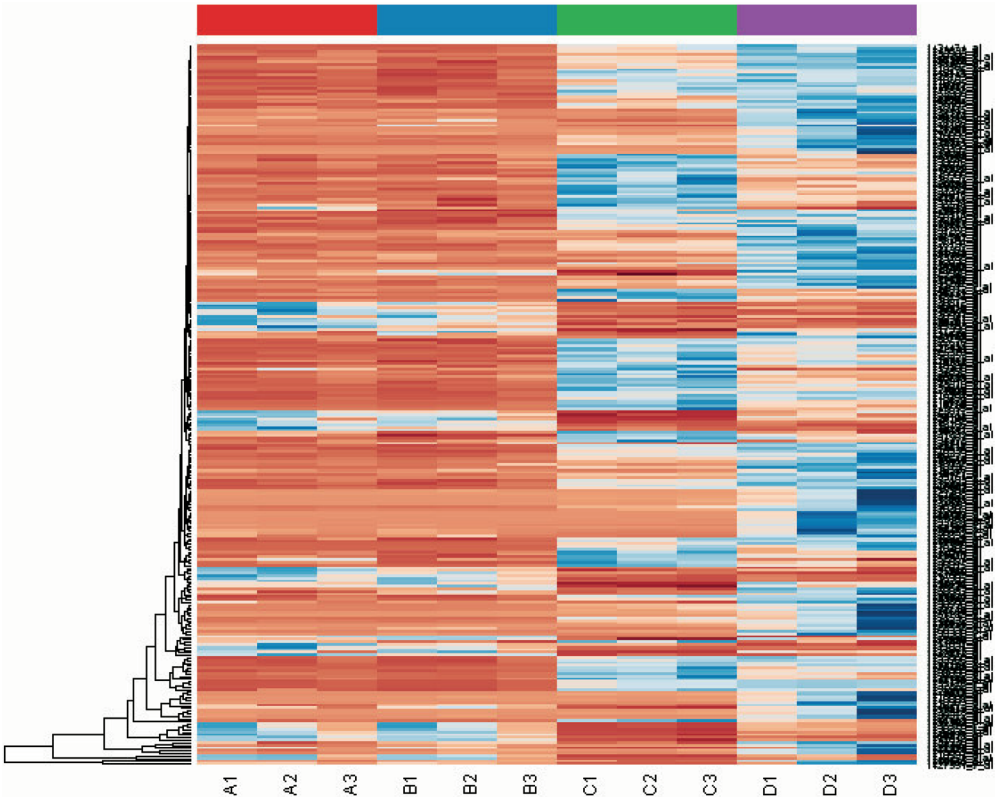

Figure 1.1 Simple clustering to find cell types. Four sets of cells cytometrically sorted in triplicate based on two surface markers are assayed by expression microarray. Clustering of differentially expressed genes defines states of gene expression for each class.

of many molecular complexes, which have traditionally been difficult to characterize. Microarrays have been used to identify the RNA components of various complexes, shedding light on biological mechanisms of RNA translation and transport [33,34]. Recently identified complexes of protein and RNA, called P-bodies, are thought to be involved in gene expression by regulating mRNA in the cytoplasm [35]. Microarrays could be used to monitor and characterize the trafficking of cellular RNA through this complex. Changes in DNA copy number at various loci have been implicated in tumorigenesis and cancer. Using comparative genomic hybridization, microarrays have been used to examine aneuploidy and changes in loci copy number in a variety of cell types [36,37]. Microarrays have been used to examine the progress of replication forks as they copy the genome [38], as well as for genome-wide screens of RNA modifying enzymes [39]. The full range of applications is too numerous to mention, improvements and adaptations are continually being made, and one must continually review the literature (see [40] for review).

However, a rapidly burgeoning area of microarray research over the past few years involves the use of microarrays to identify functional elements in the genome. Expression of a gene in the form of an RNA transcript is but one small slice of the biology of a gene. A fundamental aspect of gene expression currently being explored by microarrays is the revelation of control elements in the genome that are responsible for turning genes on and off. Every gene is under the control of a regulatory code. However, this code is largely unknown. The code is composed of relatively short sequences termed *cis*-acting regulatory sites. These sites are bound by transcription factors, proteins that are responsible for activating or repressing the transcription of a gene. A majority of these sites, and combinations in which they occur, are not known or understood. However, it is possible to map the locations where proteins interact with the genome using a technique called chromatin immunoprecipitation (see [41–43] for review). This technique has been used to map the locations where specific transcription factors bind to the genome under a specific set of conditions [44], as well as to map proteins that interact with the genome on a larger scale, such as cohesins [45] and histones [46].

The general idea is to map protein/nucleic acid interactions by utilizing the specificity of antibodies for a protein of interest. Armed with an antibody against a specific protein, one can target and purify the protein as well as the associated nucleic acid. After purification the nucleic acid is removed, labeled, and hybridized to the microarray to identify which regions of the genome were associated with the protein in the cell. The identification of regulatory elements in this way has had profound implications on our understanding of gene regulatory circuitry [47,48], and provides data for a system's biological approach to understanding the cell.

The ability of chromatin IP experiments to detect functional elements such as protein binding sites in the genome depends on the resolution of probes on the microarray. Early ChIP-chip experiments in yeast used PCR products representing each gene and each intergenic region as a single DNA segment. Thus, the resolution of the technique was limited by the size of the DNA fragment.

However, advances in probe technology, such as the use of long oligos (60–90 bases) for spotted arrays and the ability to create higher and higher density *in situ* synthesized arrays (25–60mers), have allowed for the creation of tiling arrays. In a tiling array, probes are laid down across the genome, either overlapping or at some degree of spacing. Experiments with these arrays have yielded a number of amazing findings, including the ability to view single nucleosome resolution in the genome [49].

Tiling arrays have also had another benefit. High-density arrays consisting of probes that cover both strands of DNA have been used to show that transcription in the genome is ubiquitous [50,51]. The majority of nucleotides that make up the genome are part of a primary transcript, and the amount of antisense transcription is much higher than previously appreciated. In this regard, microarrays serve to directly identify functional components of the genome.

Taken together, the data being collected by tiling arrays for both protein/DNA interactions and identification of novel transcripts in humans are being systematically and jointly analyzed as part of a large consortium termed as the ENCODE

Project with the aim of compiling a comprehensive encyclopedia of DNA elements [52,53]. Microarrays are a fundamental aspect of this effort.

Although tiling arrays are yielding a wealth of productive information, they are still relatively nonstandard both in terms of design and analysis. Depending on the question being asked, a number of analysis and normalization strategies have been developed.

1.3 Array Content

1.3.1 ESTs Are the First View

Microarrays are designed to measure active parts of the genome. The most obviously functional parts of the genome are genes. These are also the parts where sequencing efforts were initially focused and used to start building sequence-based maps of the genome. Organisms express the genome by copying parts of the DNA into a form called mRNA, which is easy to isolate and obtain sequence from. Thus, the first sequenced parts of a genome that have traditionally been available are usually the expressed parts. RNA isolated from cells is converted to DNA and sequenced; however, due to certain technical aspects of the enzymes involved, it is often difficult to get a complete sequence for each RNA molecule. Thus, the resulting sequences are termed as expressed sequence tags (ESTs) because they represent short reads of an expressed sequence. It is through the collection and assembly of ESTs that we get our first view of the genome, and from which gene models are built [54]. Even after genomes are completely sequenced through systematic conventional means, ESTs are mapped back onto the genome to define or indicate which parts of the genome are expressed as genes. Microarrays used to measure gene expression have traditionally been constructed with probes designed against assemblies of ESTs.

The human body contains hundreds of cell types. Consider for a moment how even a single tissue such as skin can be different depending on where from the body it is isolated. As all the tissues and cell types arise from the same DNA, each has a unique gene expression program. As mentioned above, RNA is easy to isolate and convert to DNA. The RNA isolated from cells or tissues represents the result of gene expression occurring in that cell or tissue at the time of isolation. Once RNA has been converted to DNA, techniques exist for cloning the DNA fragments into plasmids or other vectors such that each molecule can be treated as a separate isolate, and thus can be amplified or manipulated at will. The process of converting a complex population of molecules into a collection of individually manipulatable molecules is called library construction. Many microarrays are constructed by individually amplifying the DNA segments represented by a library and printing them onto glass. In this way, one can study the activity of genes expressed in a certain tissue by performing array analysis using arrays constructed from a library of genes isolated from that tissue [55]. In some cases, libraries are constructed from organisms that have not been sequenced. For instance,

microarray studies of the human pathogen Histoplasma began before a sequence of the genome was known. Libraries of Histoplasma genes were constructed and used to create microarrays that led to the identity of genes active during infection [56]. In this situation, the identity of the spots on the array is often not known at the outset. Through array analysis, genes that appear to change under certain conditions can then be sequenced to get a foothold into the biology of the organism.

However, another way in which ESTs are used to create microarrays is through the use of a rational design process to create and design microarray probes from assemblies of ESTs. As also mentioned above, ESTs are assembled to create gene models *in silico*. Thus, with a collection of gene models, it is possible to design complementary DNA sequences that can be synthetically produced for microarray construction and that can be used to identify the expression of the target gene uniquely. Affymetrix arrays, since they are produced *in situ*, have always been based on a rational design process.

1.3.1.1 **Probe Design**

The ability of a microarray to report on the identity of a hybridization target depends on the probes attached to the microarray surface. Spotted arrays have traditionally been produced with experimentally derived libraries or with PCR products representing known gene segments [57]. Many genes share some degree of sequence identity; thus, cross-hybridization of labeled transcripts can occur between them, making it difficult to interpret results. However, most genes can be uniquely distinguished using optimized probes, which can be accomplished using synthetic DNA oligonucleotides. The wealth of sequence information available and the realization that long oligonucleotides could be substituted for PCR products with little modification to existing array production protocols [58] have changed the content of most spotted arrays to rationally designed probes.

Although PCR products serve as effective array reagents, because they are easy to produce and they stick well to glass without advanced chemistry, they can also be a source of error and variability. For instance, early attempts to create yeast genome microarrays representing more than 6000 yeast genes involved weeks and months of effort to perform 6000 PCR reactions with a typical success rate of 85%. The resulting library represented a collection of printable DNA fragments at various concentrations and of varying purity. This kind of variation differs from amplification to amplification and from group to group. Systematic errors such as erroneous plate mappings are not uncommon. Some widely used libraries have well-documented problems of misidentified or contaminated clones [59].

The length of a PCR product makes it a sensitive hybridization reagent, as after a certain length DNA molecules converge on a general melting temperature. However, length also serves a disadvantage because long pieces of DNA are likely to have a region of homology or similarity to more than one place in a genome. Experiments with differing lengths of long oligonucleotides found that long oligos were capable of sticking to glass, and exhibited sensitivity similar to or better than PCR products in a spotted array format [60]. A number of groups have evaluated the use of long oligos as array reagents [61,62].

Long oligos are synthetically produced array reagents, and thus are amenable to bioinformatic optimization. General considerations when designing probes for measuring gene expression are the uniqueness in the genome, temperature of hybridization, secondary structure, and proximity to the 3′ end of a gene [63,64]. To find probes targeted to measure the expression of a single gene, design targets, usually consisting of cDNA sequences or unigene clusters, are screened against a genomic background to find unique regions from which probes can be designed. The oligos are chosen to have similar melting temperatures, so that they behave similarly under a single hybridization condition on the array. They are also chosen to be free of secondary structures that would interfere with hybridization. Long oligo probes for gene expression are also chosen with a bias toward the 3′ end of a gene. The labeling procedure for most gene expression experiments in eukaryotes utilizes the polyA tail at the end of the transcript resulting in a 3′-end bias, so the probe is chosen to be near the 3′ end to have a higher chance of detecting the labeled transcript. Although these criteria have made for good quality data collection on gene expression, it is likely that design criteria can be improved further [65].

Advances in oligo synthesis technology have lowered costs significantly such that design and production of a genome-sized reagent is within reach of an organized group of researchers. A variety of long oligo design programs have become available. An additional advantage of printing spotted arrays with long oligonucleotides is the practical matter that the oligos are synthesized by machine; thus, many of the pitfalls associated with PCR products are avoided. For example, every oligo is guaranteed to exist, plate location errors are avoided, and the concentration of oligos is made uniform for printing.

Since Affymetrix arrays are a commercial reagent, oligo design for the arrays is not usually in the hands of researchers. However, some of the principles are worth noting both for issues of data analysis and for design considerations when using other *in situ* synthesis platforms. In early work, Affymetrix performed a careful examination of characteristics required for the design of probes from which hybridization signals could be used to determine relative expression level of a gene from a complex sample. An array consisting of 16 000 probes was used to examine the expression of just 12 genes [14]. By making hundreds of probes per gene and then examining the characteristics of the probes that most accurately reported the expression level of the gene, a set of heuristic probe design criteria was derived. These results were then applied to larger numbers of genes. In addition, a minimum number of probes required was determined, leading to a current set size of 11–20 probes per gene.

Tiling arrays have been designed both with and without regard to bioinformatic optimization. The presence of repeat regions and elements with sequence similarity in multiple places is likely to cause problems and should be taken into account in some way. Fortunately, such issues are being examined and addressed, and tools are beginning to emerge [66].

The importance of good probe design has to do with hybridization. Hybridization signal is used to infer the expression level of a gene. By comparing hybridization signals between different genes, quantitative statements can be made about the

expression levels of those genes. However, the signal derived from hybridization is affected by many factors, and unless those factors are identical from one gene to the next on a chip, our inferences about gene levels may be incorrect. In addition, the hybridization characteristics of a DNA probe are unique and sequence dependent. Each type of array that utilizes hybridization as a readout suffers from the effects of hybridization differences between DNA sequences representing genes, and measures are taken to minimize them. Affymetrix and NimbleGen address this problem by using multiple probes to represent a gene.

For spotted arrays the problem of hybridization differences between probes is less of an issue because comparisons are not made on a gene-by-gene basis within a sample. Instead, spotted arrays are used for comparing gene expression differences between two samples. The samples are uniquely labeled and then mixed together prior to array hybridization. Thus, for any given gene, what is observed on the array is not an absolute hybridization signal for a gene, but rather a ratio of hybridization signals for a gene – one from each sample. Whatever the characteristics of the hybridization probe on the chip surface, uniquely labeled complements from each sample are competing for hybridization to the spot on the array surface, and the ratio of hybridization is the primary quantity of interest.

1.4 Normalization and Scaling

Microarray experiments, whether utilizing one-channel or two-channel technology, are comparative experiments involving populations of measurements, with the end goal being to compare abundance of targets in complex populations. Most of the time, targets are compared between populations. To compare two populations to find meaningful differences, one must make assumptions about the populations. As such, the set of fluorescent measurements representing a population must undergo a process of normalization so that the parts that have changed can be distinguished from the parts that have not changed.

In the generation and collection of microarray data, many factors can give rise to systematic differences in the population. For instance, samples to be compared are not always labeled with the same efficiency. Purification steps for each sample often have differing yields. Samples to be compared on an array are not always mixed in equal proportions prior to hybridization (either by mass or by moles of dye). The dyes themselves have differing rates of decay from atmospheric elements such as ozone [67], as well as photo bleaching during the scanning of the array. Thus, if the labeled transcripts from one sample are uniformly less bright than those of the other sample, the uniform nature of such an effect allows for correction by adjusting the brightness of the dim sample by a correction factor. One goal of normalization is to remove these systematic biases [68].

Embedded within the goal of normalization is an assumption about gene expression and how large populations of gene expression measurements should be compared. For two-color experiments to be compared on one chip, the assumption is that

in response to some factor – whether it is genotype – or treatment, the majority of genes will remain unchanged, and some small fraction of genes (e.g., 2%) involved in pathways responsive to the effect being studied will be regulated up or down in roughly equal proportion [69]. According to this assumption, the average signal coming from each population as measured on the array should be equal, whereas small numbers of specific genes will change their levels and show ratios of expression significantly different from the overall population. Of course, this assumption is not always true.

Early normalization methods consisted of calculating a median of each population after excluding the upper and lower ends of the distribution of the data, and then adjusting one population through multiplication of a constant to bring the means of the populations to equality. However, in plotting one population against the other, one often notices that the populations too do not differ by a single constant across their range; that is, an intensity dependence is often observed for differences between populations. This is illustrated in Figure 1.2a. A common method of viewing two-color microarray data is to examine the ratio and the intensity together on a microarray plot [70]. In this kind of plot, the ratio of colors (R for red and G for green) is represented on the *y*-axis by M and is calculated as $M = \log_2(R/G)$, and the combined intensity from each fluorescent dye is represented on the *x*-axis by A, which is calculated as $A = (1/2)\ (\log_2(R) + \log_2(G))$. Figure 1.2a illustrates raw data from a pair of wild-type (wt) and mutant samples hybridized on a two-color array. The observed skew in the ratio is not a function of the difference between wt and mutant gene expression, but rather comes from a systematic difference arising in the process from the microarray experiment. Such plots are not uncommon and can

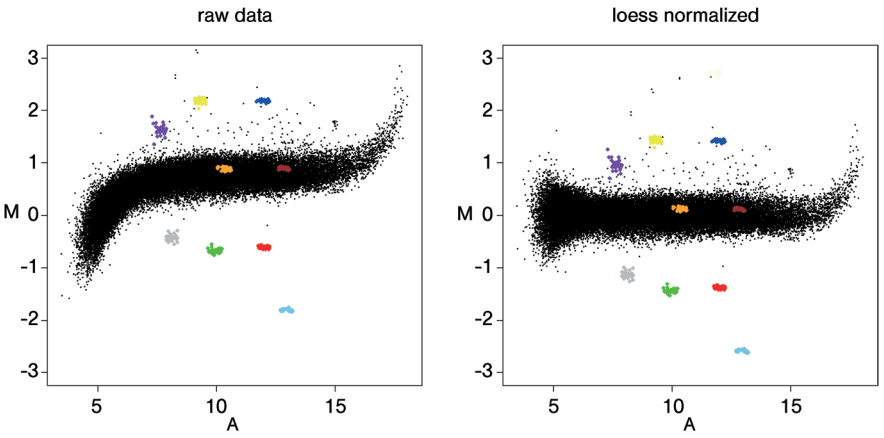

Figure 1.2 Lowess normalization of data. MA plot of a single-spotted array of expression data from a wt versus a mutant cell. Systematic skew in the data is corrected by lowess normalization. Control targets mixed at known ratios are indicated by color. (a) Left panel: unnormalized data. (b) Right panel: normalized data.

be seen even from single samples that are split in two, separately labeled and hybridized.

To alleviate this effect, many methods of normalization have been developed. The most popular is a lowess normalization method utilizing local regression to fit each population in an intensity-dependent fashion [69]. Figure 1.2b illustrates the same data as in Figure 1.2a after lowess normalization. Much of the skew from the original data set has been removed. A variety of methods have been developed for the normalization of two-color array data [71–75].

The methods above assume that the populations to be normalized are roughly equally distributed, and that the number of genes differentially expressed is small and the direction of expression is symmetric. This assumption may not be valid in some cases. For instance, small custom arrays consisting of only a subset of genes can invalidate the assumption that the effect being studied is evenly and symmetrically distributed across the gene set. However, in simulated experiments lowess normalization has been found to be robust even when 20% of the genes show differential expression in just one direction [76]. Chromatin immunoprecipitation experiments represent another area where normalization methods for the populations being compared may require special consideration [42]. In this regard, the relatively new emergence of tiling arrays has prompted the development of new methods for normalization [77]. Current methods do not take sequence composition into account, yet is clear that sequence effects will contribute to hybridization signals and should be accounted for.

Affymetrix GeneChips use hybridization intensities of single samples as a readout of gene expression. Since many factors unrelated to gene expression can affect the hybridization properties of a probe, each gene is represented not by one probe (like most other types of arrays) but by a population of probes. Summarizing a readout of several probes into a single value for gene expression adds a layer of complexity to data analysis because there are several ways probe sets can be polled and opinions differ on which method is best. Indeed, dozens of methods have been developed [21].

In dealing with Affymetrix arrays, there are two basic steps involved in data analysis. The first is summarizing the probe sets representing a gene into a readout of expression for that gene. If gene expression between two samples is to be compared, the results from each chip must be scaled to each other in a second step to account for differences in labeling and scanning of the samples. There are numerous different approaches for carrying out these steps. As of 2006, more than 30 methods had been identified [21].

The most popular methods are MAS5 [20], RMA (Robust Multichip Average) [78], and dCHIP [79]. Some methods such as GCRMA (GC Robust Multichip Average) take probe sequence into account [80]. No method is clearly the best [81], each has trade-offs and makes different assumptions about the data. The Affymetrix algorithm (MAS5) includes a background subtraction that considers both the perfect match and mismatch probes, whereas some algorithms such as RMA and GLA (Generalized Logarithm Average) [82] do not consider data from the MM probes. In a small data set with many values close to background, such differences may lead to

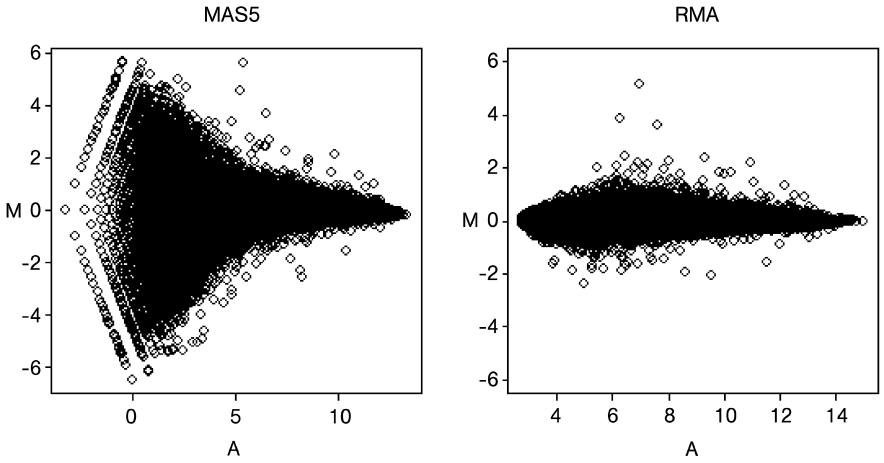

Figure 1.3 MAS5 versus RMA comparison of two data sets. Two Affymetrix expression data sets on amplified mouse cell RNA comparing a wt and mutant mouse using either MAS5- or RMA-treated data. Replicates should be performed as small data sets can exhibit large differences with different data treatments.

very different views of the data. Figure 1.3 shows MA plots of a pair of arrays summarized with MAS5 or RMA. An unsophisticated user taking only ratios into account might be misled. However, such a difference should be seen as a cautionary tail that replicates should be performed, as the two methods would otherwise have more agreement.

Control data sets have been generated and analyzed for the comparison and validation of different analysis methods [83], and then analyzed by others with different conclusions [84]. Validated data sets will be an important component in deciphering which method to apply to the analysis of raw microarray data. It is important to note the existence of different approaches, and realize trade-offs of various methods. However, most agree that the biology of the system is usually the noisiest component in the analysis. Regardless of the analysis method, microarrays are sensitive, and it is important to be a good experimentalist.

Microarrays have a difficult job to do. Expression of RNA in the cell can vary over several orders of magnitude. It is estimated that the majority of mRNA extracted from the cell contains transcripts from a small minority of genes [85]. The majority of transcripts are considered rare and expressed at very low levels [86,87]. Thus, it is not surprising that a typical Affymetrix GeneChip hybridized with a mammalian sample and analyzed with Affymetrix software reports that half of the probes are considered ''absent'' such that the signal was too low to be reliably detected. In addition, many protocols for array hybridization are suboptimal. Historically, quantitative PCR and Northern blots represent a standard method for measuring gene expression, and are used as a way to verify microarray results. It is common when performing Northern blot analysis, or qPCR, to optimize the probe

and hybridization conditions. In a microarray experiment, all the probes are different, yet they all experience identical hybridization conditions. Few experimentalists take time to optimize one of the most important steps: hybridization and wash conditions. Most spotted array protocols call for suboptimal hybridization times [88]. In addition, the specificity of hybridization is established in the washing steps when nonhybridized material is washed away. Unlike Northern blotting, few optimize this step, yet doing so would decrease noise from background hybridization, and increase the reproducibility of expression values [89].

Despite the caveats of different analysis methods, and the general difficulty of measuring gene expression, comparisons of different kinds of microarrays and across different laboratories show that the technology is robust [90–94], though issues in lack of concordance have been observed [95,96]. Most would agree that the majority of noise is associated with biological variability rather than with microarray measurement platform. Often, the question arises as to which microarray platform to use in an experiment. In terms of sensitivity, specificity, and reproducibility, a large comprehensive study recently indicated that both single and double sample platforms are equivalent [97].

1.4.1
Be Unbiased, Be Complete

Microarrays afford a new luxury that changes the way science is done. Just a short time ago, the measurement of gene expression was performed one or two genes at a time, and almost always with a hypothesis about the gene in mind. These days, however, this would be considered a biased approach. Since microarrays confer the ability to measure virtually all the genes, one can take an unbiased approach to a problem by examining all genes and identifying those genes or elements that exhibit change under some condition. Thus, the idea that one can be unbiased and complete when looking for a gene expression response emerges.

However, even here technological advances show us how biased we still are. Because microarrays will measure only what they are designed to measure, if they contain only sequences corresponding to known genes, they will miss novel biological responses that occur in unexpected places.

For example, a genomic probe set from Operon Technologies has been designed to perform whole genome analysis of gene expression in yeast, but since it consists of 70-mer oligonucleotides for 6300 known and predicted genes, the probes represent only 1.8% of the genome when taking both strands into account. Given the ubiquity of genomic transcription recently observed on tiling arrays, we can see that there are many transcriptional events that will not be detected.

1.4.2
Sequence Counts

There are two ways to identify a nucleic acid sequence: match it to something that is known or sequence it directly. Microarrays identify DNA by

hybridizing it to a known complement. However, new advances in sequencing technologies allow large populations of molecules to be identified by direct sequencing.

As mentioned earlier, advances in sequencing technology seeded a paradigm shift by creating an appetite for parallel gene expression. The switch from serial to parallel has extended into other areas such as protein arrays [98], cellular arrays [99], and phenotype arrays [100]. Recent advances in sequencing technology are poised to take over for microarrays and change the way we measure gene expression, or quantify nucleic acids in complex mixtures, yet again. Bead-based sequencing technologies allow the collection of hundreds of millions of sequence reads on individual samples [101–103]. Thus, rather than label fragments and hybridize them to an array of probes, one can simply gather sequences and count the occurrence of tags as a readout of expression.

This technology fundamentally changes what is measured in a gene expression experiment. It does not depend on gene models or building a probe, rather it counts the nucleotides directly, thus directly reading the results of gene expression. This opens the door to a long sought goal of understanding splice variation in the genome. It also sidesteps many of the technical issues behind hybridization-based methods such as designing good probes or carefully controlling hybridization conditions. Much like SAGE (Serial Analysis of Gene Expression) techniques that provide a more digital view of transcription [104] and have nucleotide level resolution, high volume sequencing techniques will do the same but are much more efficient.

As mentioned previously, the highly parallel nature of microarray data means that more is measured in a typical experiment than may be realized by the experimenter. To keep data available as a resource for mining, standards have been created for the capture and description of microarray data [105], and public repositories such as GEO [106] and ArrayExpress [107] have sprung up to archive data for public mining. The bottleneck is no longer in the generation of data, but in our ability to make sense of it.

References

1 Lander, E.S. (1996) The new genomics: global views of biology. *Science*, **274** (5287), 536–539.

2 Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H., Chu, A.M., Connelly, C., Davis, K., Dietrich, F., Dow, S.W., El Bakkoury, M., Foury, F., Friend, S.H., Gentalen, E., Giaever, G., Hegemann, J.H., Jones, T., Laub, M., Liao, H., Liebundguth, N., Lockhart, D.J., Lucau-Danila, A., Lussier, M., M'Rabet, N., Menard, P., Mittmann, M., Pai, C., Rebischung, C., Revuelta, J.L., Riles, L., Roberts, C.J., Ross-MacDonald, P., Scherens, B., Snyder, M., Sookhai-Mahadeo, S., Storms, R.K., Véronneau, S., Voet, M., Volckaert, G., Ward, T.R., Wysocki, R., Yen, G.S., Yu, K., Zimmermann, K., Philippsen, P., Johnston, M. and Davis, R.W. (1999) Functional characterization of the *S. cerevisiae* genome by gene deletion and

parallel analysis. *Science*, **285** (5429), 901–906.

3 Dimster-Denk, D., Rine, J., Phillips, J., Scherer, S., Cundiff, P., DeBord, K., Gilliland, D., Hickman, S., Jarvis, A., Tong, L. and Ashby, M. (1999) Comprehensive evaluation of isoprenoid biosynthesis regulation in *Saccharomyces cerevisiae* utilizing the Genome Reporter Matrix. *Journal of Liposome Research*, **40** (5), 850–860.

4 Brenner, C. (2000) A cultivated taste for yeast. *Genome Biology*, **1** (1), 1031–1034.

5 Adleman, L.M. (1994) Molecular computation of solutions to combinatorial problems. *Science*, **266** (5187), 1021–1024.

6 Amos, M. (2005) *Theoretical and Experimental DNA Computation*, Springer.

7 Gillespie, D. and Spiegelman, S. (1965) A quantitative assay for DNA–RNA hybrids with DNA immobilized on a membrane. *Journal of Molecular Biology*, **12** (3), 829–842.

8 Allwine, J.C., Kemp, D.J. and Stark, G.R. (1977) Method for detection of specific RNAs in agarose gels by transfer to diazobenzylmethyl-paper and hybridization with DNA probes. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 5350–5354.

9 Gergen, J.P., Stern, R.H. and Wensink, P.C. (1979) Filter replicas and permanent collections of recombinant DNA plasmids. *Nucleic Acids Research*, **7** (8), 2115–2136.

10 Riggleman, R.C., Fristensky, B. and Hadwiger, L.A. (1985) The disease resistance response in pea is associated with increased levels of specific mRNAs. *Journal Plant Molecular Biology*, **4** (2–3), 81–86.

11 Ekins, R.P. and Chu, F.W. (1991) Multianalyte microspot immunoassay – microanalytical ''compact disk'' of the future. *Clinical Chemistry*, **37** (11), 1955–1967.

12 Ekins, R.P. (1998) Ligand assays: from electrophoresis to miniaturized microarrays. *Clinical Chemistry*, **44** (9), 2015–2030.

13 Shalon, T.D. (1995) DNA Microarrays: a new tool for genetic analysis.Stanford University. Ph.D. thesis.

14 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. *Nature Biotechnology*, **14** (13), 1675–1680.

15 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** (5235), 467–470.

16 Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P. and Trent, J.M. (1999) Expression profiling using cDNA microarrays. *Nature Genetics*, **21** (1 Suppl.), 10–14.

17 Brown, P.O. and Botstein, D. (1999) Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, **21** (1 Suppl.), 33–37. Review.

18 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** (4995), 767–773.

19 Pease, A.C., Solas, D., Sullivan, E.J., Cronin, M.T., Holmes, C.P. and Fodor, S.P. (1994) Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **91** (11), 5022–5026.

20 Affymetrix((2002) Statistical Algorithms description document. Technical Report.

21 Irizarry, R.A., Wu, Z. and Jaffee, H.A. (2006) Comparison of Affymetrix

GeneChip expression measures. *Bioinformatics*, **22** (7), 789–794.

22 Blanchard, A.P. and Friend, S.H. (1999) Cheap DNA arrays – it's not all smoke and mirrors. *Nature Biotechnology*, **17** (10), 953.

23 Singh-Gasson, S., Green, R.D., Yue, Y., Nelson, C., Blattner, F., Sussman, M.R. and Cerrina, F. (1999) Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nature Biotechnology*, **17** (10), 974–978.

24 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 Research*, **12** (11), 1749–1755.

25 Hughes, T.R., Mao, M., Jones, A.R., Burchard, J., Marton, M.J., Shannon, K.W., Lefkowitz, S.M., Ziman, M., Schelter, J.M., Meyer, M.R., Kobayashi, S., Davis, C., Dai, H., He, Y.D., Stephaniants, S.B., Cavet, G., Walker, W.L., West, A., Coffey, E., Shoemaker, D.D., Stoughton, R., Blanchard, A.P., Friend, S.H. and Linsley, P.S. (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnology*, **19** (4), 342–347.

26 DeRisi, J., Penland, L., Brown, P.O., Bittner, M.L., Meltzer, P.S., Ray, M., Chen, Y., Su, Y.A. and Trent, J.M. (1996) Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics*, **14** (4), 457–460.

27 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** (5338), 680–686.

28 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** (1), 109–126.

29 Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286** (5439), 531–537.

30 Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson, J. Jr, Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O. and Staudt, L.M. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403** (6769), 503–511.

31 Winzeler, E.A., Richards, D.R., Conway, A.R., Goldstein, A.L., Kalman, S., McCullough, M.J., McCusker, J.H., Stevens, D.A., Wodicka, L., Lockhart, D.J. and Davis, R.W. (1998) Direct allelic variation scanning of the yeast genome. *Science*, **281** (5380), 1194–1197.

32 Mägi, R., Pfeufer, A., Nelis, M., Montpetit, A., Metspalu, A. and Remm, M. (2007) Evaluating the performance of commercial whole-genome marker sets for capturing common genetic variation. *BMC Genomics*, **118**, 159.

33 Johannes, G., Carter, M.S., Eisen, M.B., Brown, P.O. and Sarnow, P. (1999)

Identification of eukaryotic mRNAs that are translated at reduced cap binding complex eIF4F concentrations using a cDNA microarray. *Proceedings of the National Academy of Sciences of the United States of America*, **96** (23), 13118–131123.

34 Takizawa, P.A., DeRisi, J.L., Wilhelm, J.E. and Vale, R.D. (2000) Plasma membrane compartmentalization in yeast by messenger RNA transport and a septin diffusion barrier. *Science*, **290** (5490), 341–344.

35 Parker, R. and Sheth, U. (2007) P bodies and the control of mRNA translation and degradation. *Molecules and Cells*, **25** (5), 635–646.

36 Pollack, J.R., Perou, C.M., Alizadeh, A.A., Eisen, M.B., Pergamenschikov, A., Williams, C.F., Jeffrey, S.S., Botstein, D. and Brown, P.O. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics*, **23** (1), 41–46.

37 Shadeo, A. and Lam, W.L. (2006) Comprehensive copy number profiles of breast cancer cell model genomes. *Breast Cancer Research*, **8** (1), R9.

38 Khodursky, A.B., Peter, B.J., Schmid, M.B., DeRisi, J., Botstein, D., Brown, P.O. and Cozzarelli, N.R. (2000) Analysis of topoisomerase function in bacterial replication fork movement: use of DNA microarrays. *Proceedings of the National Academy of Sciences of the United States of America*, **97** (17), 9419–9424.

39 Hiley, S.L., Jackman, J., Babak, T., Trochesset, M., Morris, Q.D., Phizicky, E. and Hughes, T.R. (2005) Detection and discovery of RNA modifications using microarrays. *Nucleic Acids Research*, **733** (1), e2.

40 Hoheisel, J.D. (2006) Microarray technology: beyond transcript profiling and genotype analysis. *Nature Reviews Genetics*, **7** (3), 200–210.

41 Hanlon, S.E. and Lieb, J.D. (2004) Progress and challenges in profiling the dynamics of chromatin and transcription factor binding with DNA microarrays. *Current Opinion in Genetics and Development*, **14** (6), 697–705.

42 Buck, M.J. and Lieb, J.D. (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, **83** (3), 349–360.

43 Wu, J., Smith, L.T., Plass, C. and Huang, T.H. (2006) ChIP-chip comes of age for genome-wide functional analysis. *Cancer Research*, **66** (14), 6899–6902.

44 Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T.L., Wilson, C.J., Bell, S.P. and Young, R.A. (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290** (5500), 2306–2309.

45 Glynn, E.F., Megee, P.C., Yu, H.G., Mistrot, C., Unal, E., Koshland, D.E., DeRisi, J.L. and Gerton, J.L. (2004) Genome-wide mapping of the cohesin complex in the yeast *Saccharomyces cerevisiae*. *PLoS Biology*, **2** (9), E259.

46 Schübeler, D., MacAlpine, D.M., Scalzo, D., Wirbelauer, C., Kooperberg, C., van Leeuwen, F., Gottschling, D.E., O'Neill, L.P., Turner, B.M., Delrow, J., Bell, S.P. and Groudine, M. (2004) The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes and Development*, **18** (11), 1263–1271.

47 Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.B., Volkert, T.L., Fraenkel, E., Gifford, D.K. and Young, R.A. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298** (5594), 799–804.

48 Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., Gifford,

D.K., Melton, D.A., Jaenisch, R. and Young, R.A. (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, **122** (6), 947–956.

49 Liu, C.L., Kaplan, T., Kim, M., Buratowski, S., Schreiber, S.L., Friedman, N. and Rando, O.J. (2005) Single-nucleosome mapping of histone modifications in *S. cerevisiae*. *PLoS Biology*, **3** (10), e328.

50 Johnson, J.M., Edwards, S., Shoemaker, D. and Schadt, E.E. (2005) Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends in Genetics*, **21** (2), 93–102. Review.

51 David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L., Jones, T., Davis, R.W. and Steinmetz, L.M. (2006) A high-resolution map of transcription in the yeast genome. *Proceedings of the National Academy of Sciences of the United States of America*, **103** (14), 5320–5325.

52 ENCODE Project Consortium,((2004) The ENCODE (ENCyclopedia of DNA Elements) Project. *Science*, **306**, 636–640.

53 ENCODE Project Consortium,((2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

54 Davies, K. (1993) The EST express gathers speed. *Nature*, **364**, 554.

55 Altmann, C.R., Bell, E., Sczyrba, A., Pun, J., Bekiranov, S., Gaasterland, T. and Brivanlou, A.H. (2001) Microarray-based analysis of early development in *Xenopus laevis*. *Developmental Biology*, **236** (1), 64–75.

56 Hwang, L., Hocking-Murray, D., Bahrami, A.K., Andersson, M., Rine, J. and Sil, A. (2003) Identifying phase-specific genes in the fungal pathogen *Histoplasma capsulatum* using a genomic shotgun microarray. *Molecular Biology of the Cell*, **14** (6), 2314–2326.

57 Tomiuk, S. and Hofmann, K. (2001) Microarray probe selection strategies. *Briefings in Bioinformatics*, **2** (4), 329–340.

58 Chen, M., ten Bosch, J., Beckman, K., Saljoughi, S., Seidel, C., Tuason, N., Larka, L., Lam, H., Sinibaldi, R. and Saul, R. (1999) Covalent attachment of sequence optimized PCR products and oligos for DNA microarrays. Abstract TIGR Conference.

59 Halgren, R.G., Fielden, M.R., Fong, C.J. and Zacharewski, T.R. (2001) Assessment of clone identity and sequence fidelity for 1189 IMAGE cDNA clones. *Nucleic Acids Research*, **29** (2), 582–588.

60 Sinibaldi, R., O'Connell, C., Seidel, C. and Rodriguez, H. (2001) Gene expression analysis on medium-density oligonucleotide arrays. *Methods in Molecular Biology*, **170**, 211–222.

61 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 (50 mer) microarrays. *Nucleic Acids Research*, **28** (22), 4552–4557.

62 Chou, C.C., Chen, C.H., Lee, T.T. and Peck, K. (2004) Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression. *Nucleic Acids Research*, **12** (32), e99.

63 Bozdech, Z., Zhu, J., Joachimiak, M.P., Cohen, F.E., Pulliam, B. and DeRisi, J.L. (2003) Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray. *Genome Biology*, **4** (2), R9.

64 Kreil, D.P., Russell, R.R. and Russell, S. (2006) Microarray oligonucleotide probes. *Methods in Enzymology*, **410**, 73–98.

65 Pozhitkov, A.E., Tautz, D. and Noble, P.A. (2007) Oligonucleotide microarrays: widely applied poorly understood.

Briefings in Functional Genomics and Proteomics, **6** (2), 141–148.

66 Graf, S., Nielsen, F.G., Kurtz, S., Huynen, M.A., Birney, E., Stunnenberg, H. and Flicek, P. (2007) Optimized design and assessment of whole genome tiling arrays. *Bioinformatics*, **23** (13), i195–i204.

67 Fare, T.L., Coffey, E.M., Dai, H., He, Y.D., Kessler, D.A., Kilian, K.A., Koch, J.E., LeProust, E., Marton, M.J., Meyer, M.R., Stoughton, R.B., Tokiwa, G.Y. and Wang, Y. (2003) Effects of atmospheric ozone on microarray data quality. *Analytical Chemistry*, **75** (17), 4672–4675.

68 Quackenbush, J. (2002) Microarray data normalization and transformation. *Nature Genetics*, **32** (Suppl.), 496–501. Review.

69 Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J. and Speed, T.P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, **30** (4), e15.

70 Dudoit, S., Yang, Y.H., Callow, M. and Speed, T. (2002) Statistical methods for identifying genes with differential expression in replicated cDNA experiments. *Statistica Sinica*, **12**, 111–139.

71 Dabney, A.R. and Storey, J.D. (2007) Normalization of two-channel microarrays accounting for experimental design and intensity-dependent relationships. *Genome Biology*, **8** (3), R44.

72 Wilson, D.L., Buckley, M.J., Helliwell, C.A. and Wilson, I.W. (2003) New normalization methods for cDNA microarray data. *Bioinformatics*, **19** (11), 1325–1332.

73 Baird, D., Johnstone, P. and Wilson, T. (2004) Normalization of microarray data using a spatial mixed model analysis which includes splines. *Bioinformatics*, **20** (17), 3196–3205.

74 Wit, E. and McClure, J. (2004) Normalization in *Statistics for Microarrays: Design, Analysis and Inference*, John Wiley & Sons, Ltd., Chapter 4.

75 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. *Nucleic Acids Research*, **29** (12), 2549–2557.

76 Oshlack, A., Emslie, D., Corcoran, L.M. and Smyth, G.K. (2007) Normalization of boutique two-color microarrays with a high proportion of differentially expressed probes. *Genome Biology*, **8** (1), R2.

77 Royce, T.E., Rozowsky, J.S. and Gerstein, M.B. (2007) Assessing the need for sequence-based normalization in tiling microarray experiments. *Bioinformatics*, **23** (8), 988–997.

78 Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed, T.P. (2003) Exploration normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4** (2), 249–264.

79 Li, C. and Wong, W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences of the United States of America*, **98** (1), 31–36.

80 Wu, Z., Irizarry, R.A., Gentleman, R., Francisco, M.M. and Spencer, F. (2004) A Model Based Background Adjustment for Oligonucleotide Expression Arrays. Johns Hopkins University Department of Biostatistics Working Papers. Working Paper 1.

81 Qin, L.X., Beyer, R.P., Hudson, F.N., Linford, N.J., Morris, D.E. and Kerr, K.F. (2006) Evaluation of methods for oligonucleotide array data via quantitative real-time PCR. *BMC Bioinformatics*, **177**, 23.

82 Zhou, L. and Rocke, D.M. (2005) An expression index for Affymetrix GeneChips based on the generalized logarithm. *Bioinformatics*, **21** (21), 3983–3989.

83 Choe, S.E., Boutros, M., Michelson, A.M., Church, G.M. and Halfon, M.S. (2005) Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biology*, **6** (2), R16.

84 Dabney, A.R. and Storey, J.D. (2006) A reanalysis of a published Affymetrix GeneChip control dataset. *Genome Biology*, **7** (3), 401.

85 Holland, M.J. (2002) Transcript abundance in yeast varies over six orders of magnitude. *Journal of Biological Chemistry*, **277** (17), 14363–14366.

86 Axel, R., Feigelson, P. and Schutz, G. (1976) Analysis of the complexity and diversity of mRNA from chicken liver and oviduct. *Cell*, **7** (2), 247–254.

87 Bishop, J.O., Morton, J.G., Rosbash, M. and Richardson, M. (1974) Three abundance classes in HeLa cell messenger RNA. *Nature*, **250** (463), 199–204.

88 Sartor, M., Schwanekamp, J., Halbleib, D., Mohamed, I., Karyala, S., Medvedovic, M. and Tomlinson, C.R. (2004) Microarray results improve significantly as hybridization approaches equilibrium. *Biotechniques*, **36** (5), 790–796.

89 Korkola, J.E., Estep, A.L., Pejavar, S., DeVries, S., Jensen, R. and Waldman, F.M. (2003) Optimizing stringency for expression microarrays. *Biotechniques*, **35** (4), 828–835.

90 Wang, Y., Barbacioru, C., Hyland, F., Xiao, W., Hunkapiller, K.L., Blake, J., Chan, F., Gonzalez, C., Zhang, L. and Samaha, R.R. (2006) Large scale real-time PCR validation on gene expression measurements from two commercial long-oligonucleotide microarrays. *BMC Genomics*, **7**, 59.

91 de Reynies, A., Geromin, D., Cayuela, J.M., Petel, F., Dessen, P., Sigaux, F. and Rickman, D.S. (2006) Comparison of the latest commercial short and long oligonucleotide microarray technologies. *BMC Genomics*, **7**, 51.

92 Petersen, D., Chandramouli, G.V., Geoghegan, J., Hilburn, J., Paarlberg, J., Kim, C.H., Munroe, D., Gangi, L., Han, J., Puri, R., Staudt, L., Weinstein, J., Barrett, J.C., Green, J. and Kawasaki, E.S. (2005) Three microarray platforms: an analysis of their concordance in profiling gene expression. *BMC Genomics*, **56** (1), 63.

93 Yauk, C.L., Berndt, M.L., Williams, A. and Douglas, G.R. (2004) Comprehensive comparison of six microarray technologies. *Nucleic Acids Research*, **32** (15), e124.

94 Draghici, S., Khatri, P., Eklund, A.C. and Szallasi, Z. (2005) Reliability and reproducibility issues in DNA microarray measurements. *Trends in Genetics*, **22** (2), 101–109. Review.

95 Kothapalli, R., Yoder, S.J., Mane, S. and Loughran, T.P. (2002) Microarray results: how accurate are they? *BMC Bioinformatics*, **233**, 22.

96 Tan, P.K., Downey, T.J., Spitznagel, E.L., Jr, Xu, P., Fu, D., Dimitrov, D.S., Lempicki, R.A., Raaka, B.M. and Cam, M.C. (2003) Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Research*, **31** (19), 5676–5684.

97 Patterson, T.A., Lobenhofer, E.K., Fulmer-Smentek, S.B., Collins, P.J., Chu, T.M., Bao, W., Fang, H., Kawasaki, E.S., Hager, J., Tikhonova, I.R., Walker, S.J., Zhang, L., Hurban, P., de Longueville, F., Fuscoe, J.C., Tong, W., Shi, L. and Wolfinger, R.D. (2006) Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. *Nature Biotechnology*, **24** (9), 1140–1150.

98 Haab, B.B., Dunham, M.J. Brown, P.O., (2001) Protein microarrays for highly

parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biology*, **2** (2), 1–13.

99 Wheeler, D.B., Carpenter, A.E. and Sabatini, D.M. (2005) Cell microarrays and RNA interference chip away at gene function. *Nature Genetics*, **37** (Suppl.), S25–S30. Review.

100 Homann, O.R., Cai, H., Becker, J.M. and Lindquist, S.L. (2005) Harnessing natural diversity to probe metabolic pathways. *PLoS Genetics*, **1** (6), e80.

101 Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F. and Rothberg, J.M. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437** (7057), 376–380.

102 Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129** (4), 823–837.

103 Johnson, D.S., Mortazavi, A., Myers, R.M. and Wold, B. (2007) Genome-wide mapping of *in vivo* protein–DNA interactions. *Science*, **316** (5830), 1497–1502.

104 Velculescu, V.E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M.A., Bassett, D.E., Jr, Hieter, P., Vogelstein, B. and Kinzler, K.W. (1997) Characterization of the yeast transcriptome. *Cell*, **88** (2), 243–251.

105 MGED Society (2002) Microarray standards at last. *Nature*, **419**, 323.

106 Edgar, R., Domrachev, M. and Lash, A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, **30** (1), 207–210.

107 Parkinson, H., Sarkans, U., Shojatalab, M., Abeygunawardena, N., Contrino, S., Coulson, R., Farne, A. and Lara, G.G., Holloway, E., Kapushesky, M. Lilja, P., Mukherjee, G., Oezcimen, A., Rayner, T., Rocca-Serra, P., Sharma, A., Sansone, S. and, Brazma, A. (2005) ArrayExpress – a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, **33**,D553–D555.

108 Blanchard, A.P., Kaiser, R.J. and Hood, L.E. (1996) High-density oligonucleotide arrays. *Biosensors and Bioelectronics*, **6/7**, 687–690.

109 Marmur, J. and Doty, P. (1961) Thermal renaturation of deoxyribonucleic acids. *Journal of Molecular Biology*, **3**, 585–594.