5
High-Density Arrays and Insights into Genome Function

LARS M. STEINMETZ[1*] AND RONALD W. DAVIS[1,2]

[1]*Department of Genetics and* [2]*Biochemistry, Stanford University School of Medicine, Stanford, CA 94305, U.S.A.*

Introduction

Genome projects are producing sequence data at a very fast pace (http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html). The discovery of the complete human genome sequence is only a few years away and a working draft with 90% coverage is promised to appear by the time of this publication (Strategy meeting on human genome sequencing, Cold Spring Harbor, 1999). In addition to the detailed sequence, biologists will receive a list of all 50–100,000 genes in the human genome and the challenge then turns towards organizing the genes and understanding how genes operate and interact to produce a living system. Traditional gene-by-gene analyses are inefficient for obtaining information about the function, regulation, and sequence variation of the thousands of genes in a genome. Highly parallel analyses are needed to be able to survey biology from a global perspective.

One type of tool for studying biology from a global perspective is the high-density array, also known as a microarray, which consists of a miniaturized, high-density array of probes bound to a solid surface. Current applications have been based on DNA probes, although in theory other molecules such as proteins or small molecular weight compounds can also be arrayed at high density. Exploiting the specificity of hybridization, DNA probes on high-density arrays can detect the presence of individual target sequences in complex mixtures. This ability allows for massively parallel hybridization assays for large numbers of genes and sequences, and has been primarily applied to survey genomes for variations in mRNA expression levels or between DNA sequences.

Using high-density DNA arrays for mRNA expression studies, rapid, accurate, and

[*] To whom correspondence may be addressed (larsms@stanford.edu)

Abbreviations: DHPLC, denaturing high performance liquid chromatography; DNA, deoxyribonucleic acid; ddNTP, dideoxynucleoside triphosphate; dNTP, deoxynucleoside triphosphate; mRNA, messenger ribonucleic acid; PCR, polymerase chain reaction; RNA, ribonucleic acid; SNP, single nucleotide polymorphism; STS, sequence tagged site.

Biotechnology and Genetic Engineering Reviews – Vol. 17, August 2000
0264–8725/00/17/109–146 $20.00 + $0.00 © Intercept Ltd, P.O. Box 716, Andover, Hampshire SP10 1YG, U.K.

reproducible information about the transcript level can be obtained in parallel for thousands of genes. This information can be used to organize genes into functional categories, to identify molecular signature patterns, and to search for shared regulatory sequences that are important for understanding the control of biological processes. In a sense, the organization of genes into expression groups for different biological processes can be viewed as a first step towards a global understanding of the molecular composition and operation of cells and organisms. However, a more complete understanding of genome function also requires an understanding of the contribution of DNA sequence variation to phenotype. With the ability to use high-density arrays for DNA variation detection, large amounts of sequence can be surveyed for DNA variation. This information can be used to study the role of DNA variation in human populations. With the further ability to use high-density arrays for genotyping multiple individuals at sites of DNA variation, there is promise that genome-wide association studies will map the genetic variants responsible for complex phenotypes, diseases, and behaviours.

Ultimately, with its current applications centred on DNA hybridization, high-density array data will contribute to an understanding of how RNA transcript levels and DNA sequence variation contribute to differences in phenotypes within and between species. In the future, however, genome-wide experimentation may be extended to measuring protein levels and their interactions with the help of high-density antibody and protein arrays and to measuring levels of other molecules.

This review will focus on the ways high-density DNA array data is generated and analysed, and will be divided into four sections:

(I) The first section will review existing high-density array platforms. The principle of DNA capture by complementary probes will be described with reference to high-density solid surface arrays.

(II) The second section will review the application of arrays for gene expression analysis. This section will begin with the biological applications of global transcription patterns and then proceed to a discussion of more technical matters, such as the assembly of public data repositories and the computational analysis of expression data.

(III) The third section will review the use of arrays for studying DNA sequence variation. This section will open with a description of array-based methods for identifying DNA sequence variants. The discussion will then turn to the application of arrays for genotyping DNA variation.

(IV) The fourth and final section will make brief mention of other high-density array applications.

I. High-density solid surface array platforms

Hybridization to solid surface arrays is a modern successor of filter-based hybridization assays. The demonstration that single-stranded DNA, bound to nitrocellulose, can hybridize to complementary RNA (Gillespie and Spiegelman, 1965) led to the development of colony and plaque hybridization, as well as filter-based gel-blot and dot-blot hybridization assays (Grunstein and Hogness, 1975; Benton and Davis, 1977; Southern, 1975; Kafatos *et al.*, 1979). In colony/plaque hybridization, colonies/plaques growing on a plate are replica-platted onto a nitrocellulose filter, lysed, and

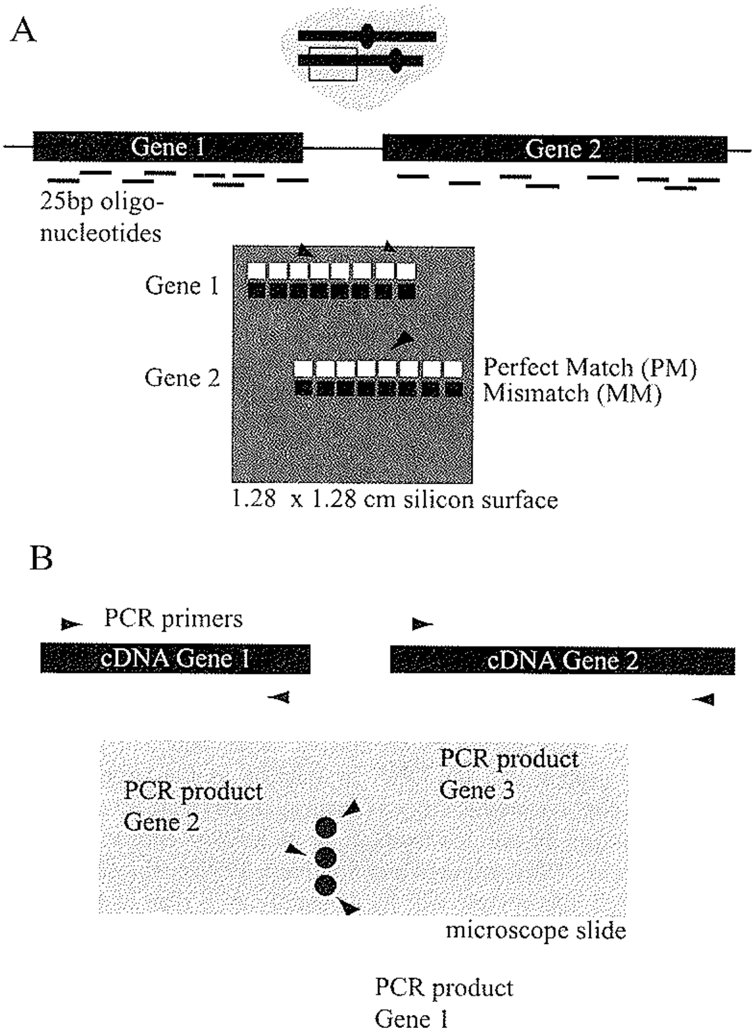

Figure 5.1. Diagram of the two main microarray platforms. (A) Oligonucleotide array. Multiple perfect match (PM) probes complementary to a sequenced transcript are synthesized directly on a silicon surface. Currently, for each complementary probe, a second probe with a single-base mismatch (MM) at its centre position is synthesized directly underneath. (B) cDNA microarray. PCR-amplified products are spotted as probes onto glass microscope slides. Currently, a single probe is used per transcript.

their DNA is fixed to the filter. A colony/plaque with a DNA sequence complementary to a hybridized sample is identified by a hybridization signal at a position on the filter that corresponds to a colony/plaque position on the replica plate. These studies have reached densities of 100 spots/cm^2 and, interestingly, the analyses have often involved comparisons of two replica filters, imaged and superimposed, using two different colours: this is analogous to what is done with current expression array images. In filter-based gel-blot experiments, a random collection of DNA samples, spatially resolved according to size by gel electrophoresis, are blotted to filters and probed for sequence content. The DNA molecule that is complementary to the hybridized sample is identified by its migration rate through the gel. The dot-blot is the first organized and ordered arrangement of DNA samples on a filter. Individual DNA samples are spotted at low density directly onto filters. Because the position of each spot is determined by the experimenter, the spot-density can be controlled, and the identity of the DNA sample at each spot can be inferred directly from its array coordinates. For massively parallel hybridization analyses, large filters (400 and 330 cm^2) have been used to array cDNA libraries at densities of 23 and 91 spots/cm^2 (Lennon and Lehrach, 1991; Drmanac and Drmanac, 1994).

A significant advance in array-based hybridization assays, however, came with an increase in array density, achieved by depositing or synthesizing probes on a solid surface such as glass (Fodor *et al.*, 1991; Khrapko *et al.*, 1991; Maskos and Southern, 1992a; Shalon *et al.*, 1996). High-density arraying is responsible for miniaturizing large-scale hybridization assays, and the use of a glass surface permitted fluorescence detection as a replacement of radioactive, chemiluminescent, or colourimetric detection methods. Although the first high-density array contained 1,024 peptides arrayed at a density of 640 different compounds/cm^2 using spatially directed synthesis (Fodor *et al.*, 1991), current high-density arrays mainly contain surface-bound nucleic acids. Overall, the miniaturization of array technology has allowed for highly parallel hybridization analyses at an unprecedented scale.

The level of throughput that is required for the analysis of entire genomes is currently best met by two types of high-density array platforms (*Figure 5.1*) that can be distinguished by method of probe placement onto the array surface. Following the conventions of array hybridization, nucleic acids bound to the solid surface are referred to as probes, and the DNA in solution as the target. The first type, dominated by the oligonucleotide array, is an array made by direct synthesis in which probes are synthesized by various methods in an arrayed format directly on glass surfaces (Fodor *et al.*, 1991; Maskos and Southern, 1992b; Blanchard *et al.*, 1996). Among these arrays, this review will focus mainly on a type of array upon which probes are synthesized by photolithography and photosensitive oligonucleotide synthesis chemistry (Fodor *et al.*, 1991; Pease *et al.*, 1994; McGall *et al.*, 1996; reviewed by Lipshutz *et al.*, 1999). This technology currently achieves the highest density of surface-bound oligonucleotide probe-spots, with about 180,000 spots/cm^2. The second type of array platform, represented mainly by the cDNA array, is an array in which probes are synthesized and then mechanically spotted or printed in an arrayed format onto glass microscope slides (Schena *et al.*, 1995; Shalon *et al.*, 1996; Guo *et al.*, 1994; Khrapko *et al.*, 1991; Yershov *et al.*, 1996; Sosnowski *et al.*, 1997; Schena *et al.*, 1998). Commercially available robotic spotters and ink-jet printers allow for about 10–25,000 spots/cm^2 (Bowtell, 1999) but, due to the difficulty of achieving consistent, small spot sizes, most spotted arrays are printed at a density of 10,000 spots/cm^2.

In addition to differences in array density, these two processes of array manufacture have different demands on quality control. The process of direct probe synthesis is clearly distinct from probe spotting as it embodies a direct link between a probe sequence and its array position. Information from a sequence database directly controls probe synthesis for each spot, and the identity and location of each probe on the array is therefore known. Spotting probes onto glass involves synthesizing probes separately, tracking samples during processing steps, and then arraying them in an organized format. Minimizing tracking mistakes and intermediate processing errors is critical, and quality control is essential to verify that each probe has its expected location on the array.

The method of array synthesis is also important because it determines the type of sequences that can be placed onto arrays. Direct synthesis of nucleic acid on glass surfaces is currently limited to short sequence lengths (usually 20–25 bp), because the yield of full-length oligonucleotides decreases as the sequence length grows. High-density DNA arrays made by direct synthesis are therefore primarily oligonucleotide arrays that contain probes of short nucleic acids (usually 20–25 bp). In contrast, spotting technology can be applied to any sequence of interest, irrespective of length. As a result, oligonucleotide arrays have also been generated by placing pre-synthesized oligonucleotides onto solid surfaces (Guo *et al.*, 1994; Khrapko *et al.*, 1991; Yershov *et al.*, 1996; Sosnowski *et al.*, 1997), but these arrays have been less frequent, in part because they contain more variation in probe spot consistency and achieve a lower density. Instead, spotting technology is mainly used to synthesize cDNA arrays, consisting of PCR products (of approximately 0.6–2.4 kb) mechanically spotted or printed onto glass microscope slides (Schena *et al.*, 1995; Shalon *et al.*, 1996; reviewed by Cheung *et al.*, 1999). It should be noted that cDNA arrays do not necessarily contain spots of PCR-amplified cDNA but could contain gene sequences amplified directly from genomic DNA or, indeed, other sources. Here, the two main array platforms will be referred to as oligonucleotide arrays and cDNA arrays, respectively.

PRINCIPLES OF DNA HYBRIDIZATION TO HIGH-DENSITY ARRAYS

For both high-density array platforms, an experimental sample, usually DNA, is tagged with fluorescent labels and incubated in solution on the probe array. During the hybridization process, DNA probes and matching targets associate to form complementary base pairs. A scanning confocal fluorescence detector then measures the amount of fluorescence at each probe-spot (Glazer *et al.*, 1990). The amount of target that has hybridized to a probe is measured by the amount of fluorescence signal that is above noise. To detect very low signal, which may result from hybridization to very small probe-spots or low target concentration, it is necessary to increase the ratio of signal to noise. This must be achieved by either amplifying the specific signal or decreasing the noise, and is independent of the probe-spot surface area.

The efficiency of target capture is determined by the hybridization kinetics. In solution, the rate of DNA renaturation is kinetically a second order reaction (Wetmur and Davidson, 1968). The initial formation of a few complementary base pairs is thought to produce an unstable intermediate, which can either dissociate or become stable through the formation of additional base pairs. The rate-limiting step is characterized by a zipper-like process in which additional complementary bases anneal until a stable DNA duplex is formed. Once formed, the annealed duplex is relatively stable and strand separation is very slow, even for relatively short sequences (Maskos and Southern, 1992b).

On the array surface, however, the hybridization kinetics can be more complex: In solution-phase renaturation, both complementary DNA strands diffuse freely in solution. During hybridization on solid supports, probes are bound to the solid surface. Targets in solution must therefore come close enough to probes on the solid surface to interact and initiate base pairing (reviewed by Southern *et al.*, 1999). The rate of this process depends on properties of the target DNA in solution, the surface-bound probes, and the solid surface. In solution, the complexity of target DNA and the diffusion rate of DNA molecules affect the rate at which target molecules approach the array surface. At the surface, probe length and density, as well as the extent to which probes are exposed to the hybridization solution, affect the capture rate of approaching target molecules. Depending on these parameters, DNA hybridization on high-density arrays can ultimately result from a direct collision between targets and their complementary probes. Alternatively, targets can adsorb nonspecifically to the glass surface, and diffuse laterally to a neighbouring probe (Adam and Delbruck, 1968). The relationship between these parameters and the rate of hybridization has been expressed in equation form[1].

High-density oligonucleotide arrays

Due to differences in probe length, the two main high-density array platforms differ in the details of hybridization. Shorter probes on high-density oligonucleotide arrays entail a lower T_m (melting temperature; a measure of probe-target duplex stability) and more variation in T_m between probes. Accordingly, a lower hybridization temperature is used. In addition, to minimize secondary structure formation and to speed diffusion, fragmented targets are used on oligonucleotide arrays.

With current conditions, oligonucleotide arrays can detect target sequences that are present in 1 in 300,000 copies (about 0.05 copies per yeast cell), and detection is quantitative over 3 orders of magnitude (Lockhart *et al.*, 1996; Wodicka *et al.*, 1997). Due to a relatively large incubation chamber (200 µl on current high-density oligonucleotide arrays), large amounts of polyA RNA are needed if labelled directly without amplification (10 µg for yeast hybridizations) (Wodicka *et al.*, 1997). With amplification protocols, sample requirements have been significantly reduced (0.2–5 µg of polyA RNA) (Gene Chip Manual, Affymetrix).

cDNA arrays

Longer probes on cDNA arrays permit higher hybridization temperatures. These conditions minimize formation of secondary structure in target DNAs and increase the diffusion rate. On cDNA arrays, incubations are carried out on microscope slides

[1] The rate J (moles/cm^2/sec) of DNA molecules that are captured by a reverse complementary probe present in excess on a solid surface, can be approximated by $J = (\pi\gamma D_3/4H)C_0$ (Chan *et al.*, 1995; Chan *et al.*, 1997; see also review by McKenzie *et al.*, 1998). C_0 is a familiar measure in solution-phase renaturation studies of complementary DNA fragments. It is the initial concentration of target base pairs in solution. The diffusion coefficient, D_3, determines the frequency at which targets get close enough to their probes to start hybridization. The additional factors in the equation account for differences between solid and solution-phase hybridization: H measures the width of the boundary layer at the plate surface over which the conditions in bulk solution change to conditions at the probes. γ is a measure of how close the capture rate comes to the maximum capture rate obtained when the surface is uniformly covered with probes. The parameter incorporates the contributions by two-dimensional diffusion after nonspecific adsorption, probe density, and probe length. Because the equation assumes that the number of probes available to targets stays constant, it is accurate only as long as a small fraction of all probes on the surface have hybridized (Chan *et al.*, 1995).

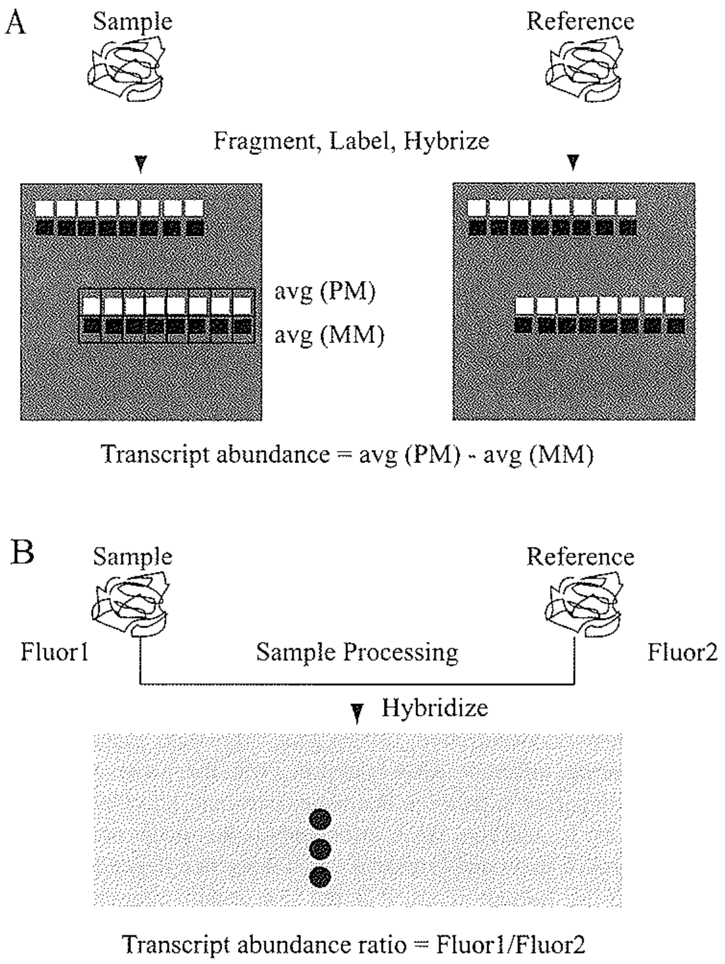

Figure 5.2. Differences in hybridization logistics between array platforms. (A) Calculating transcript abundance with an oligonucleotide array. A single RNA sample is processed, fluorescently labelled, and hybridized. The transcript abundance of a gene is calculated as the average difference between the perfect match (PM) and the mismatch (MM) intensities for all probes that tile across that gene. Changes in expression are obtained by comparing the transcript abundance calculated from independent hybridizations and can be calibrated by spiking in controls at known concentrations. (B) Calculating transcript abundance with a cDNA microarray. To account for differences in the spot-size and probe-concentration of the same probe spot between arrays, a reference sample labelled with a different fluorescent dye is co-hybridized directly to each array. The ratio of the fluorescence intensities represents the ratio of transcript concentration between the sample and its reference.

in a small volume which minimizes the amount of required sample (1–5 μg polyA RNA) (DeRisi *et al.*, 1997; Duggan *et al.*, 1999). The detection limit is at 1 in 100,000 copies (about 0.15 copies per yeast cell) and signal is quantitative over 3 orders of magnitude (Duggan *et al.*, 1999).

COMPARISON OF OLIGONUCLEOTIDE ARRAYS AND CDNA ARRAYS FOR GATHERING GENE EXPRESSION DATA

Both technologies are extremely sensitive for monitoring global mRNA expression levels, but differ in measure of transcript abundance (*Figure 5.2*). On the oligonucleotide array, a single sample is hybridized to each array and the intensities of multiple probes are averaged. The result is a measure of absolute transcript abundance. On the cDNA array, differentially labelled test and reference samples are co-hybridized to each array, and the fluorescence ratio at each probe represents a measurement of the transcript ratio between the sample and its control. The result is a measure of relative transcript abundance. Important advantages and disadvantages are associated with each method with respect to synthesis technology, hybridization method, and probe length.

Synthesizing probes directly on the surface of oligonucleotide arrays increases the reproducibility of probe-spot consistency and, as a result, single fluorescence intensity readings are reproducible and hence reliable. In addition, going directly from a sequence database to probe synthesis at defined sites on the array minimizes the chances for error. The probes synthesized on a surface are short and, as a result, some probes can be designed specifically to distinguish splice variants and members of gene families. In addition, the short duplexes that result from oligonucleotide hybridization are very sensitive to single nucleotide mismatches, making it possible to detect DNA variants by their decreased hybridization signal (discussed in detail later). However, this sensitivity also has disadvantages: a decreased signal can result either because there is a sequence difference in the target DNA or because the target DNA is present at lower concentration. During gene expression analysis, this is not an issue when comparing different RNA samples with the same sequence, but is significant when comparing expression intensities between individuals with frequent sequence differences. In addition, since prior knowledge of the probe sequences is required, having probes that complement a target sequence depends on sequencing quality. Particularly for human genes, it is the low quality sequencing rather than natural, population-based sequence variation that is currently the major source of probe-target sequence discrepancy. Most of these difficulties, however, will go away as the available sequence quality improves and probes can be specifically designed to coding regions that have a low polymorphism rate.

Oligonucleotide arrays made by direct synthesis clearly achieve the highest density, which outweighs most disadvantages. Nevertheless, a fraction of the probes at each spot on the array are truncated and incomplete because the step yield for photochemical synthesis on solid surfaces is 92–94% (McGall *et al.*, 1996). Each probe-spot thus contains a significant amount of incomplete probes, some of which (~ 10% of all probes at a spot) may cross-hybridize with other target sequences under the same conditions used to hybridize the full-length probes. In addition, due to the nature of short DNA duplexes, different capture rates are obtained for probes with different sequence composition (*Figure 5.3*). To overcome this variation, multiple probes are synthesized to the same transcript and are averaged to obtain a measure of transcript abundance. The large number of probes has advantages: by increasing the number of data points per gene, the accuracy and reproducibility of quantitating transcript levels by hybridization rate is increased (*Figure 5.4*), and absolute transcript abundance can be approximated. The redundancy of having multiple probes also increases the

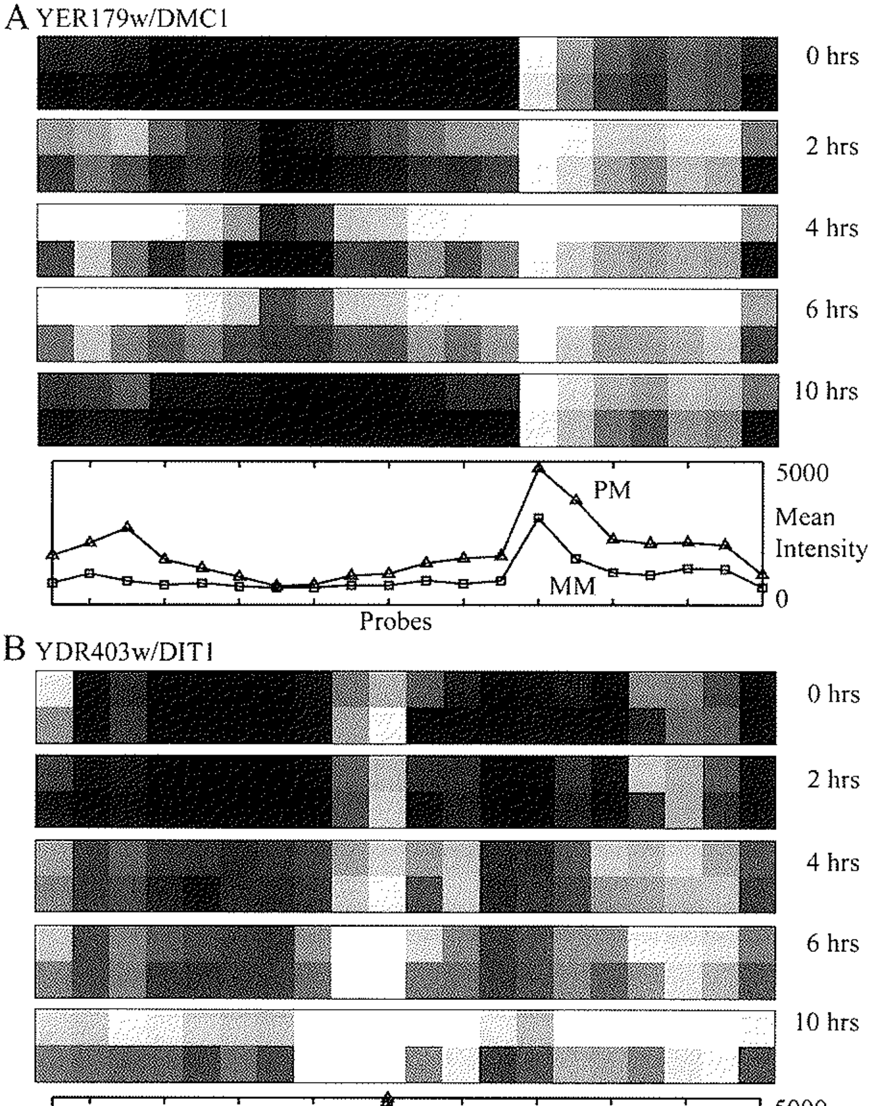

Figure 5.3. Differences in hybridization signal between different oligonucleotide probes. All probe intensities for DMC1 and DIT1, early and late genes in meiosis, were ranked and displayed in grayscale in a reconstructed array image (highest intensity coloured in white). The perfect match row is displayed above the mismatch row for each time point of a time course. (A) DMC1 probe intensities at 5 time points. Variation in probe-target duplex stability is reflected by differences in signal intensity among probes in the same row. The average intensity of each perfect match and mismatch probe over all time points is plotted below each array image. (B) Some analysis for DIT1 probe intensities at 5 time points (data from R. Williams *et al.*, manuscript in preparation).

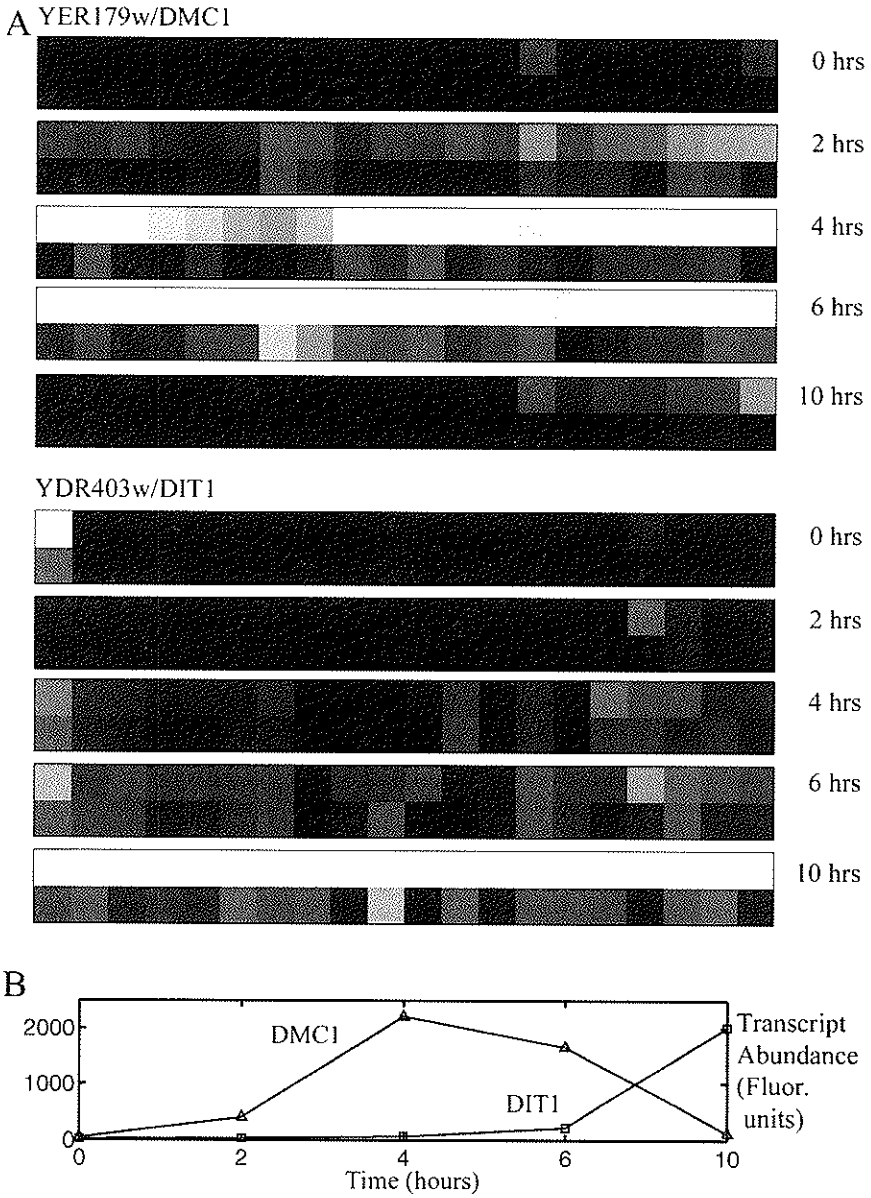

Figure 5.4. Interpreting hybridization signal to measure transcript abundance with 20 short oligonucleotide probes and their corresponding mismatch probes. Although there is considerable difference in hybridization signal for different probes, most probes change signal in accordance with changes in transcript abundance. (A) Array images reconstructed from time course data (same as *Figure 5.3*) for DMC1 and DIT1, early and late genes in meiosis. In contrast to *Figure 5.3* here the highest intensity signal in each column has been coloured in white. This display indicates the timepoint at which each probe is highest in intensity. (B) Plot of the transcript abundance. Comparison of overall transcript abundance with intensity of individual probes from panel A shows that the fluorescence intensity of most of the 20 probes during the time course increases and decreases proportionately with overall transcript abundance as measured by the average difference intensity across all 20 probes (calculated as described in *Figure 5.2*).

possibility that each transcript is represented by probes that are not polymorphic with respect to the target sequence.

With spotting technology, the physical separation of probe preparation and array synthesis allows for errors. Arrays made from cDNA library clones that were identified by sequencing propagate sequence-tracking errors. In addition, a good tracking database is required to minimize errors introduced during array printing. Spotting from microtitre plates raises the possibility that the identity of probes on the array is wrong, either by spotting from the wrong plates or from plates that are out of order or in the wrong orientation. For arrays made from amplified PCR products, accurate tracking of samples, primers, and primer pairs is crucial. In addition, minor PCR contaminants, particularly in products amplified from genomic DNA, that consist of gene sequences that are homologous to an abundant RNA, may confound gene expression measurements. These contaminants may lead to an erroneously strong signal for genes that are expressed at low levels, and good quality control is needed. In addition, with physical placement of samples onto glass surfaces, a single probe-spot on one cDNA array compared to another may differ in size and probe concentration. Single fluorescence intensity readings are therefore meaningless. Nevertheless, spotting technology has the advantage that PCR products that have not been sequenced can be spotted and applied to study the expression level of unsequenced genes. This advantage, however, will be less significant as more genome and gene sequences become available.

On cDNA arrays, two-colour hybridization technology is used to overcome inconsistencies in spot composition. The advantage of a two-colour approach is that it carries a built in control: every measurement is made relative to a co-hybridized reference sample. Ratio measurements, however, become problematic when signals are low; for example, when the reference signal is close to zero, the ratio measurement may not be meaningful. In addition, because an absolute value of transcript abundance cannot be obtained, comparisons of ratio measurements obtained with different reference samples are problematic (discussed in more detail later). Nevertheless, the long probe sequences (0.6–2.4 kb) on cDNA arrays vary little in duplex stability, and a single probe per gene is sufficient. Fewer probes are thus required to measure the same number of genes, but members of closely related gene families cannot be distinguished.

II. Global gene expression monitoring

The ability to measure the expression level of tens of thousands of genes in one experiment has increased the pace of gene expression analyses by several orders of magnitude. This increase in throughput makes genome-wide expression monitoring possible. In yeast and other small genomes in which all genes are known, a measure of the expression level can be obtained for every gene in the genome. In the future, it will be possible to measure the expression level for all human genes. Consequently, high-density DNA arrays are increasingly attractive tools for a variety of biological applications. Gene expression information can be analysed for (1) clues about gene function, (2) clues about genome organization, or (3) to identify signature patterns. These three applications are by no means a comprehensive list of all possible applications of expression data, but provide a framework for understanding the current published literature. A detailed description of each application will follow. Then the more detailed computational issues associated with establishing public gene expression repositories and the statistical analysis of expression data will be discussed.

1. ANALYSING EXPRESSION FOR CLUES ABOUT GENE FUNCTION

Information about the location, level, and timing of gene expression has been important for understanding what a gene does. Many genes that are similarly expressed carry out common functions in the cell. As an example, genes that function specifically in particular stages of the cell-cycle often show a cell-cycle dependent periodicity of expression (Cho *et al.*, 1998; Spellman *et al.*, 1998). There are two arguments for why there might be a tight correlation between expression level and function of a gene. (1) The execution of cellular processes and the activation of molecular pathways are tightly regulated. (2) Evolutionary selection is likely to have limited cellular resources (ie for transcription) to times when protein products are needed. As a result, the transcription profile of an entire genome can be viewed as a detailed description of a cell's active molecular pathways. These arguments make it attractive to test whether expression patterns can be applied to organize genes into parts lists for different cellular processes and predict gene function for many genes that are currently uncharacterized. In drawing conclusions from gene expression profiles, however, biologists are faced with two difficult questions: (1) if a gene is found to be expressed in a particular pathway, is it important for that pathway? And (2) if a gene is not found to be expressed, is it not important for that pathway?

The expression profile of genes during a cellular process, such as the cell cycle, can be used to organize genes into groups (how this is done will be discussed in the context of clustering approaches later). Each resulting group consists of genes with a similar expression profile. To understand the meaning of a set of similarly expressed genes, the relationship between genes within a set can be interpreted with a template list of well-characterized genes. Knowledge of the function of these well-characterized genes is then used to decide whether similarly expressed genes share similar functions. By locating uncharacterized genes amidst sets of well-studied genes, numerous functional clues have been gathered for unknown genes and existing biochemical pathways have been refined (reviewed by Brown and Botstein, 1999).

Although there is debate about how successful the transcript level of genes will be at predicting gene function in specific cases, most biologists concede that additional information is necessary to understand function (Fields, 1997). In some respects, a global understanding of gene regulation is necessary before we can fully understand the significance of changes in expression. Considering that genes compete for cellular resources such as polymerase enzymes and transcription factors, it is likely that an expression change in one gene will affect the transcript rates of all other genes with which it competes. As a result, there is a certain noise level in expression measurements and biologists must be careful not to over-interpret these changes as having biological significance (Brenner, 1999). In addition, not all changes in transcript levels are also changes in protein expression. Non-specific increases in transcript levels may result because a protease produced in response to decreasing one gene's specific activity will degrade other proteins as well, and genes whose function is required at a constant level therefore need to be transcribed at a higher rate. As a result, the most informative genes are the ones that are not differentially expressed under most conditions, except for one. Those genes that change expression only under one particular pathway are most likely pathway-specific and biologically relevant. As more data is collected, these genes will be easier to find.

To address the second question, one must realize that high-density arrays obtain a snapshot of RNA transcript levels. Other determinants of protein function, besides transcription levels, such as rates of translation, protein modification or degradation are not measured. Genes that do not change transcript levels may, therefore, still be regulated in other ways, and hence be important for a particular pathway. So far, expression analysis has focused on large fold changes in transcript levels. Important regulators of cell function, however, may show little variation in expression. The ongoing analysis of gene deletion phenotypes has revealed little correlation between essential genes and their expression pattern (Winzeler *et al.*, 1999). Although genes essential for sporulation in yeast were found to be generally differentially expressed during meiosis, the observed changes were sometimes small and below an arbitrarily-set significance threshold (Chu *et al.*, 1998; R. Williams *et al.*, manuscript in preparation). These data underline the importance of using multiple approaches in genome scale functional studies.

Furthermore, a challenge of the future is to make gene expression measurements on a single cell. Important biological decisions are made on a single cell level, and some gene expression changes that lead to these decisions may not be uncovered by monitoring average expression levels. This is particularly evident for decisions that are triggered by threshold levels of expression. The average expression level of a gene that fluctuates in expression level in a single cell over time has a constant expression level when averaged over multiple cells. By measuring averages, important details may be missed.

Although it seems clear that transcript levels will not be equally important for all genes, it is important to keep in mind that the power of the array-based efforts to understand gene function lies in scale. The value of large-scale gene expression monitoring in assigning function to novel genes will likely not be realized until detailed information for each gene is collected from hundreds and thousands of experiments, combined, and analysed *en masse*. With more detail about how genes are expressed under a variety of conditions, searches will yield genes that are similarly regulated across experiments. With these highly detailed data lists, inferences of shared function become more meaningful.

2. ANALYSING EXPRESSION FOR CLUES ABOUT GENOME ORGANIZATION

Knowledge about individual genes contributes to an understanding of gene function, but expression data can also be applied directly to address more general questions about genomes, such as the correlation between the expression level of genes and sequence similarity, chromosome position, or regulatory structures. In brief, three of these discoveries are outlined:

(1) A study of 100 genes involved in central nervous system development in rats has searched for a correlation between sequence similarity in coding regions and the differential expression of genes. Among members of the same gene family, no significant correlation was found (Michaels *et al.*, 1998; Wen *et al.*, 1998). This observation is in accord with arguments that evolutionary selection, acting on members of duplicated gene sets, has adapted them for different functional purposes.

(2) The largest efforts towards understanding genome structure have focused on understanding the mechanisms of coordinated transcriptional regulation. Studies in bacteria have identified sets of neighbouring genes with a common regulatory mechanism (Jacob and Monod, 1961). Genome-scale expression data from yeast provides other evidence for shared regulatory sequences in genomes with short intergenic regions. Interestingly, about 25% (5-fold higher than expected by random chance) of genes that are periodically expressed during the cell cycle localize directly adjacent to another gene that is expressed during the same phase of the cell cycle (Cho *et al.*, 1998).

(3) Coordinated transcriptional regulation has also been found for genes that are not proximally positioned. Searches for identical regulatory sequence motifs in the non-coding regions of genes has revealed a good correspondence between sets of similarly expressed genes and common regulatory sites. Many common sequence motifs that resemble regulatory sites have been successfully identified among yeast genes that display a similar differential expression profile (DeRisi *et al.*, 1997; Cho *et al.*, 1998; Spellman *et al.*, 1998; Holstege *et al.*, 1998; Chu *et al.*, 1998). A description of the entire control regulatory network in genomes requires a comprehensive search for all functionally active transcription factor-binding sites. The identification of genes that change expression in the presence of transcription factor mutations mark efforts on this front (Holstege *et al.*, 1998).

3. ANALYSING EXPRESSION TO IDENTIFY SIGNATURE PATTERNS

A third application of expression data concerns itself with the overall pattern of transcription, not the transcription of particular genes. In this case, gene function and regulation are not inferred from the expression level of a gene. Instead, the genome-wide transcription pattern is applied to describe the genetic and biochemical identity of a cell, tissue, or organism. The observation that distinct expression patterns are found in disease tissue demonstrated that expression patterns of many genes can be used to recognize disease states from human tissue samples (Heller *et al.*, 1997; Zhang *et al.*, 1997). Additional studies have begun to apply genome-scale expression monitoring to characterize mutations and classify cell types.

It is known that receptor tyrosine kinase signalling proceeds through a set of distinct biochemical pathways activated in the cytoplasm, but it is not clear whether each pathway activates a distinct set of transcriptional targets. By monitoring gene expression in the presence of specific receptor tyrosine kinase mutations, the transcriptional targets of each biochemical pathway have been characterized in isolation. This study revealed that the transcriptional profiles of different signalling pathways are very similar, involving the induction of highly overlapping sets of genes (Fambrough *et al.*, 1999). This finding supports other evidence that suggest that the specificity of each pathway may be achieved by a quantitative rather than a qualitative change in expression (Pawson and Saxton, 1999).

In another context, gene expression profiles have been used to study target genes of drug compounds. The change in expression pattern that results from drug treatment defines a signature pattern in a cell. The absence of a signature pattern in a drug-treated cell with a mutation in a putative target gene demonstrates that the target is required to mediate the differential transcription of genes that define the signature pattern. This approach may be useful for confirming potential drug targets and revealing the presence of alternate pathways through which a drug may exert secondary effects (Marton *et al.*, 1998).

In the clinical setting, expression profiles have been exploited for the ability to make highly accurate phenotype assignments. For years, cancer tumours have been screened for the expression level of specific genes. Using genome-scale expression profiling of breast cancer tumour samples, distinct groups of genes can be identified that correspond to the presence of specific cell types in the solid tumour samples. The comparison of cancerous to normal breast tissue reveals that there exist distinct expression patterns that are found in tumour tissue (Perou *et al.*, 1999). Ultimately, the identified expression patterns may be used as signatures for the diagnosis, classification, and dissection of breast tumour types. Tumour type prediction based on signature patterns has already been demonstrated for the liquid tumours, acute myeloid leukaemia and acute lymphoblastic leukaemia. No single test previously existed for distinguishing between these two tumour types, but signature patterns found by comparing known tumour samples have been successful at defining a class predictor that can be used to accurately assign new samples to one of the two tumour types (Golub *et al.*, 1999).

These studies demonstrate that global expression profiles can be interpreted as molecular readouts of cells. The transcriptional phenotype, that incorporates data for thousands of genes, presents a level of detail and quantitation that is unprecedented in phenotypic analysis. More applications that exploit this power are likely to follow.

DESIGNING LARGE GENE EXPRESSION REPOSITORIES

The many applications of gene expression data demonstrate the utility of global transcription information. However, data lists for gene expression experiments easily run into the hundreds of thousands of values. This volume places new demands on analysing and storing the information (reviewed by Bassett *et al.*, 1999). Only a limited number of analysis tools are currently available to integrate expression data with existing database resources (for example see Ermolaeva *et al.*, 1998). In addition to developing analysis tools, a long-lasting permanent record of gene expression databases is needed. While the data is currently published mainly in private databases, it seems essential that, in the future, databases will be maintained by someone other than the person generating the data. To further public accessibility, the European Bioinformatics Institute has agreed to establish a public database committed to storing array-based expression data (*Editorial*, 1999). This effort will permit cross-validation of data from different experiments, and increase the demand for much needed standards on gene expression analysis and presentation.

Comparing data from different array platforms

The construction of an expression database raises important concerns about the compatibility of experiments performed with different array technologies. Ideally, comparison between high-density array data sets would be in units of transcript copies per cell. In practice, however, sequence-dependent differences in the hybridization efficiency of probes on oligonucleotide arrays and differences in spot consistency on cDNA arrays make measurements of absolute concentration difficult. On oligonucleotide arrays, absolute transcript concentration can currently be approximated by averaging the signal from multiple probes, in addition to calibrating the signal with spiked-in controls (Wodicka *et al.*, 1997; Holstege *et al.*, 1998). Because the oligonucleotide array permits a measure of transcript abundance directly from the hybridization intensity of a single sample (*Figure 5.4*), fluorescence measurements from different high-density array experiments can be directly compared *ad hoc*. For cDNA arrays only, the ratio of transcript concentration between two or more samples can currently be quantitated with accuracy and efficiency. A direct comparison between hybridizations is therefore meaningful only as long as the same reference sample has been used to obtain a relative measure of transcript abundance. To make more widespread comparisons possible, both within and between laboratories, it has been suggested that a common reference sample be introduced (Bassett *et al.*, 1999). Reaching agreement on a common reference sample, however, is difficult and will take time.

In addition, it needs to be pointed out that measures of absolute transcript levels on oligonucleotide arrays are accurate only as long as the hybridized target sequences are identical in sequence to the probes on the array. Likewise, ratio measures on cDNA arrays are accurate only as long as the co-hybridized targets have identical sequences. Transcript intensity measurements for human samples are complicated because the sequence of the hybridized sample often is polymorphic with respect to the probe sequences. While relative measures of expression level from samples taken from a single individual or a single cell line are accurate, comparisons between multiple polymorphic targets, such as in cross-species comparisons, are more complex. It would therefore be most useful to develop internal controls with which to calibrate the hybridization signal so that experiments on all platforms can be directly compared.

Reproducibility of high-density array data

A database of expression experiments raises important issues about the reproducibility of high-density array data. As several authors point out, there is currently greater variation between the same experiment performed consecutively, than there is between repeated hybridizations of the same sample (Wodicka *et al.*, 1997; Holstege *et al.*, 1998; Lander, 1999). Using high-density oligonucleotide arrays, about 10 of 6,200 genes had intensity differences of more than two-fold between consecutive hybridizations of the same sample (Wodicka *et al.*, 1997); 30 to 70 genes varied more than two-fold when comparing independently prepared samples (Wodicka *et al.*, 1997; Holstege *et al.*, 1998). These differences shift attention to the variation of independent biological measurements. Each measurement may vary as a result of differences in gene expression in different samples, as well as measurement error. For each gene, one needs to determine what sort of natural variation is expected. For instance, a two-fold change may be significant for one gene, but not for another. Multiple samples (triplicate or greater) are likely to be essential for judging the significance of fold changes in the expression level of individual genes. Public databases made from similar experiments carried out in different laboratories will provide a measure of variance for each gene. This variation can then be taken into account during the analysis of individual experiments.

COMPUTATIONAL ANALYSIS OF GENE EXPRESSION DATA

With large amounts of expression values filling public databases, attention has focused on computational methods to make sense of the data. The data needs to be standardized and interpreted. Standardization corrects for differences introduced by experimentation. There are no standards for normalizing gene expression data (for a statistical treatment of cDNA array image data see Chen *et al.*, 1997). The form of data analysis that follows depends on the design of the experiment and the result expected. Although different in the details, all studies aimed at addressing one of the above applications face the same technical hurdles.

When analysing an expression experiment, how do we identify and organize the subset of genes that are biologically relevant? Often, the first step in an analysis is the identification of genes that are significantly induced or repressed. For this purpose, fold-change thresholds have been primarily used (for example Chu *et al.*, 1998; Fambrough *et al.*, 1999). In some cases, the stringency of a fold-change threshold can be evaluated with existing knowledge of gene function: one of the relevant questions may be whether a significance threshold selects genes whose function makes them likely to be differentially expressed. In most cases, however, there are limited prior expectations and a significance threshold is placed arbitrarily at a value thought to be greater than the expected natural sample-to-sample variability. However, since each gene is likely to have different degrees of natural variation, significance thresholds need to be applied on a gene-by-gene basis. To minimize false-positives and false-negatives, different measurement scales are needed for genes that display different degrees of variability (Wittes and Friedman, 1999). The natural variability of each gene can be evaluated by repeating the same experiment multiple times. In cases, such as limited human tissue samples for which repetition is not feasible, it may be possible to develop alternate approaches. Transcript level variances calculated from a database of gene expression from normal human tissue samples might be useful. Further development of this stage of data analysis is needed.

Cluster analysis and the search for patterns

Once lists of significant genes are generated, the second step in data analysis turns toward organizing genes according to patterns in expression trajectories. The list of genes is searched for non-random patterns in expression profile. The most obvious patterns can be detected by eye, as done for cell cycle periodic transcripts (Cho *et al.*, 1998), but this approach is inappropriate as the data grows and becomes more complex. Currently, the most widely used method for detecting underlying patterns in large information sets is a cluster analysis. Applied to gene expression data, clustering divides genes into groups according to similarity in expression profile. The expression profile of a single gene over a set of n samples (such as a time course) is represented as a single point in n-dimensional space. The genome-wide expression profile of yeast occupies about 6,200 points in space. Clustering algorithms locate the points that are closest to each other and group these into clusters. There are several clustering techniques that differ in measure of similarity.

Hierarchical cluster analysis calculates distance for pairs of genes by either a correlation coefficient or by Euclidean distance (two similar measures that are linearly equated). The results are displayed in a nested hierarchical tree, where distance to the nodes reflects distance in expression space, and genes connected by the nearest node are most similar in expression pattern (Eisen *et al.*, 1998). Although hierarchical clustering has been useful for the interpretation of several experiments (for example Iyer *et al.*, 1999; Chu *et al.*, 1998; Michaels *et al.*, 1998; Wen *et al.*, 1998; Spellman *et al.*, 1998), its structure places a hierarchical order on genes, some of which may be correlated in more complicated ways.

Several other clustering methods have been applied to the analysis of gene expression data, including k-means method (Tavazoie *et al.*, 1999) and self-organizing maps (Tamayo *et al.*, 1999; Toronen *et al.*, 1999). Two articles have helped to clarify the distinction between clustering methods, and make three key points (Bittner *et al.*, 1999; Chen *et al.*, 1999):

(1) For data that naturally falls into distinct groups and is well separated, all methods produce the same gene clusters, but if the data is more uniformly distributed, each algorithm places the cluster boundaries differently. In effect, the position of genes with large differences in expression pattern are placed into different groups by any method, but the positions of genes with little difference are variable.

(2) The statistical confidence in clusters depends on the number of data values in an expression trajectory. To make accurate comparisons between genes, many data values per gene are required (for 95% confidence, 20 to 30 sample points for genes with a correlation coefficient of 0.7). The confidence in comparisons depends on the correlation between expression profiles, and more data values are required for comparing genes with little similarity.

(3) Comparisons by the aforementioned cluster analyses fail to detect more complex relationships between genes, because they are based on pairwise comparisons. For example, a gene's expression may be correlated with the sum of the transcript levels of five other genes (see Chen *et al.*, 1999). Other models and statistical approaches are required to recognize these more complex interactions.

The utility of more complex algorithms

More complex algorithms have been described that could be used to complement a hierarchical clustering approach (see Michaels *et al.*, 1998). Are these algorithms justified? Although the utility of these approaches remains to be established, experimenting with more complex approaches could lead to the discovery of novel links between genes that provide clues to gene function and regulation.

The biologist needs to evaluate the utility of these approaches. A well-established standard is to test how well a novel clustering approach organizes genes for which the function and regulatory interactions are known. Clusters of genes are used to understand biochemical processes and regulatory networks in cells; it therefore remains to be determined whether complex algorithms that search for more complex interactions will be efficient at detailing biologically relevant interactions. Many complex gene expression patterns may turn out to be described by a combination of multiple, but individually simple, linear interactions among genes.

The emergence of large-scale expression monitoring has created a need for information science in biology. Already, authors are beginning to speculate that the discovery of a distinct number of gene expression clusters may allow the construction of a model for specific complex biomolecular networks that could be used to make testable predictions (Michaels *et al.*, 1998). However, before such feats can be realized, more needs to be learned about the regulation of genes under a variety of experimental conditions. At this point in time, the main focus lies in testing new and improved ways of utilizing genome-wide expression data to address medically relevant questions.

III. Identifying and genotyping DNA variation

High-density arrays have also been used to study DNA sequences. For many model organisms, there is a defined laboratory strain whose sequence has been designated as a reference, and with respect to which sequence variants can be assigned. For the human genome, there is no single, defined reference, and current sequencing efforts are yielding the equivalent of a single human sample. With a draft of a human genome sequence forthcoming, the discovery of DNA variation among different individuals has drawn increasing interest. A representative reference sequence of the human species needs to include the sequence variants found between individuals, as well as between different human populations. Given the high cost of comprehensive genome sequencing, however, it is unlikely that more than one genome will be sequenced in human, as well as in most other species. Instead, techniques for large-scale and massively parallel sequence comparisons are needed. Most established methods, however, cannot easily meet the demands for rapid, cost-effective, and large-scale sequence analysis (see Schafer and Hawkins, 1998), but the high-density format of the DNA array has proven extremely useful.

Methods for identifying and genotyping DNA variants have a wide range of applications: differences in DNA sequence underlie the genetic basis of hereditary traits and diseases and, in part, explain some of the phenotypic differences within and between human populations and related species. For example, by characterizing changes in DNA sequence between different human populations, it has been possible to trace the history and origins of the human species. In addition, sequence variations have served as landmarks throughout the human genome for identifying mutations underlying Mendelian phenotypes. The next step is to extend positional mapping to traits and diseases with complex inheritance patterns (reviewed by Schafer and Hawkins, 1998; Plomin *et al.*, 1994). For this purpose, at least 500,000 allelic variants may need to be identified and catalogued (Kruglyak, 1999; Risch and Merikangas, 1996). The high-density format of the DNA array could be applied to identify and genotype large numbers of DNA variants needed for these studies. However, the use of arrays for variation detection and genotyping is still in its early stages, and the issues surrounding large-scale sequence analysis are currently primarily of a technical kind.

In gene expression analyses, both cDNA arrays and oligonucleotide arrays have been used. cDNA arrays, however, cannot be applied to detect allelic variants in strictly hybridization-based approaches because a single nucleotide mismatch among hundreds of complementary base pairs has little effect on hybridization rate. Short probes (usually 20–25 bp) are sufficiently sensitive, and high-density oligonucleotide arrays have therefore primarily been applied. In this review, the study of DNA variation by hybridization will be divided into two sections: first, the application of arrays to scan genomes for DNA sequence variation and second, the application of arrays to genotype individuals at sites of DNA variation.

IDENTIFYING DNA SEQUENCE VARIATION

The success of array-based variation detection can be addressed in the context of two different approaches. Each strategy will be discussed in detail. The first is a re-sequencing approach in which the hybridization pattern of a DNA sample is directly interpreted to infer the exact sequence of base pairs. The second is a comparative approach in which data analysis has shifted attention to homing in on the key differences between the hybridization pattern of two samples.

Single sample re-sequencing approaches

Initially, arrayed combinatorial libraries of complete sets of n-mer oligonucleotide probes have been conceptualized for the *de novo* DNA sequence reconstruction of a DNA segment of any composition. According to theory, the hybridization pattern of a DNA sample hybridized to a set of all possible oligonucleotides of length n can be decoded and the sequence identity inferred (Bains and Smith, 1988; Khrapko *et al.*, 1989; Drmanac *et al.*, 1989; Southern *et al.*, 1992). In practice, however, this task is difficult; in most part due to the computational complexity of assembling data that is confounded by cross-hybridization, interactions among targets, and different hybridization rates of different probes (for example, see *Figure 5.3*). The *de novo* sequencing approach, as initially proposed has therefore not been successfully implemented. However, with prior knowledge of a specific reference sequence, hybridization to an n-mer array has been successful in a more specialized application: the re-sequencing of DNA segments. In these applications, n-mer arrays contain all possible olignucleotides of length n arrayed at high-density on a solid surface. From the complete n-mer set, only probes that complement the known reference sequence can be interpreted to assemble the sequence of a hybridized sample. By modifying the detection method (hybridization followed by ligation to a co-hybridized anchor probe), a 9-mer array has been successful at re-sequencing 1.2 kb with 99.9% accuracy. This demonstration suggests the possibility of using a single generic array to re-sequence a DNA target of any composition. However, most likely due to the short probe sequence lengths on the arrays, longer targets are re-sequenced with considerably lower accuracy (2.5 kb at 94.5%) (Gunderson *et al.*, 1998).

Arrays with considerably longer probes (usually 20–25 bp) achieve better hybridization quality, presumably because the specificity of probe-target interactions is increased. Although it would, in theory, be possible to make an n-mer array with all possible probes of 20 bp in length, this would require more probe spots (4^20) than can currently be manufactured on a single array. Consequently, most re-sequencing arrays are limited to probes designed for a specific reference sequence. High-density oligonucleotide arrays, consisting of a tiled array of probes to the known sequence, interrogate each position on a strand of a reference sequence with four different probes that only differ by a single base at their centre position (*Figure 5.5*) (Chee *et al.*, 1996). Also known as variant detector arrays, or VDA in short, tiling arrays have been mainly used for the purpose of detecting DNA polymorphisms. To detect all

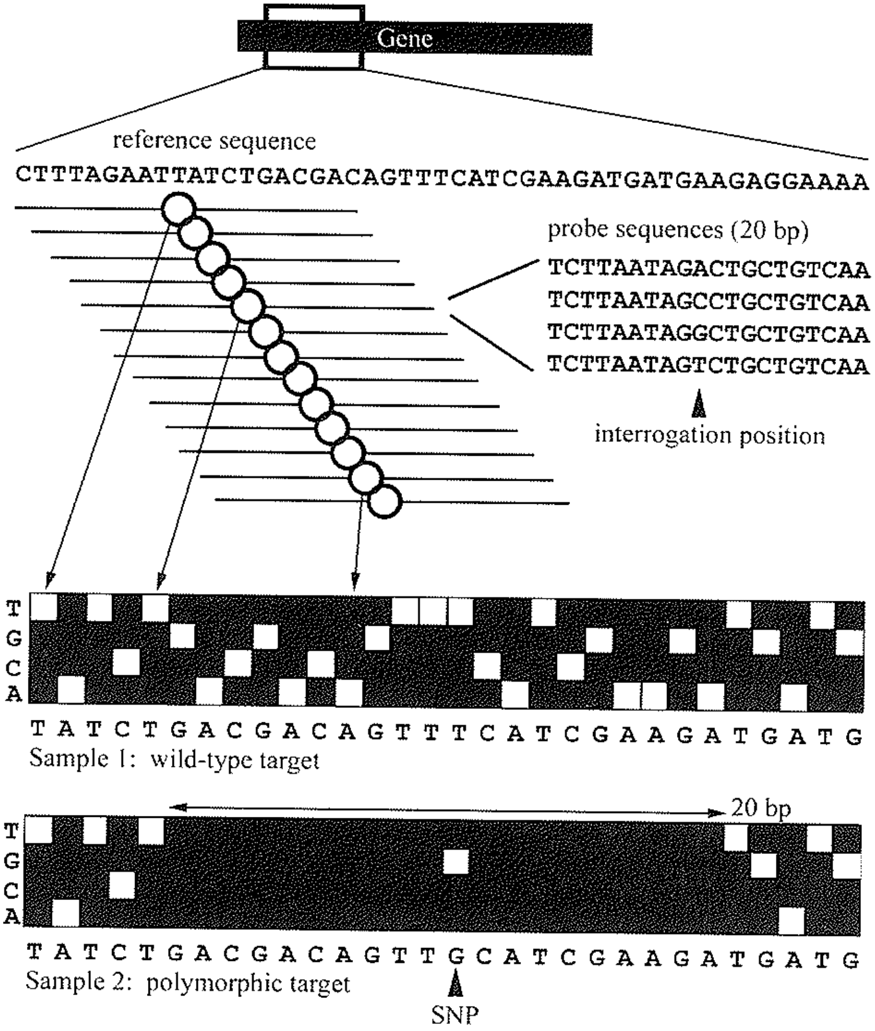
Gene
reference sequence
CTTTAGAATTATCTGACGACAGTTTCATCGAAGATGATGAAGAGGAAAA
probe sequences (20 bp)
TCTTAATAGACTGCTGTCAA
TCTTAATAGCCTGCTGTCAA
TCTTAATAGGCTGCTGTCAA
TCTTAATAGTCTGCTGTCAA
interrogation position
T
G
C
A
T A T C T G A C G A C A G T T T C A T C G A A G A T G A T G
Sample 1: wild-type target
20 bp
T
G
C
A
T A T C T G A C G A C A G T T G C A T C G A A G A T G A T G
Sample 2: polymorphic target
SNP

Figure 5.5. Scanning for polymorphisms on a tiling array. Four 20 bp oligonucleotides that are complementary to the reference sequence but have either an A, C, G, or T at their centre position interrogate each position in the reference sequence in turn. These probes are positioned in a single column on the tiling array. Shown here is a small fragment of a hybridized tiling array (hybridization signal coloured white). Sample1 is identical to the reference sequence and, consequently, the reference probe at each position hybridizes. Sample 2 has an SNP (G instead of T) and among 20 probes that overlap the non-reference base, the only probe that hybridizes is the one with the correct base substitution at the polymorphic position. For each sample the sequence inferred from the hybridization pattern is shown below the tiling block.

possible single nucleotide differences between two sequences of length N, a total of 4 N probes are required. Other types of allelic variation can also be detected by adding, for example, an additional set of N probes that complement all possible deletions of a given length (Chee *et al.*, 1996; Hacia *et al.*, 1996). With current production techniques yielding about 300,000 distinct probes on a 1.28 × 1.28 cm^2 solid surface, a single tiling array with 4 N probes, in theory, is capable of scanning about 75 kb of sequence for all possible single-nucleotide substitutions (reviewed by Lipshutz *et al.*, 1999).

Although arrays designed for a specific reference sequence avoid the difficulty of having to assemble hybridization data *de novo*, the main challenge lies in developing ways to achieve uniform hybridization characteristics for all probes. Variation in hybridization rate exists for different probes on expression arrays (*Figure 5.3*), but is particularly critical for tiling array experiments. The probe sequence has to be complementary to the sequence around the polymorphism on either DNA strand, and hence cannot easily be manipulated to yield optimal hybridization results. These difficulties are reflected in the array-based re-sequencing results. 98.26% of 33,858 bp of total sequence from the human immunodeficiency virus-1 genome (297 bp from 114 samples) were in agreement with the sequence predicted with conventional dideoxy sequencing (Kozal *et al.*, 1996). With respect to polymorphism detection, a 1.73% error rate of a re-sequencing approach is currently too inefficient as it would require a second step to determine which polymorphisms are true-positives. Since the difficulties associated with re-sequencing increase with target complexity, this approach has been difficult to apply on a larger scale. In an alternative strategy, tiling arrays have been applied in a comparative analysis that avoids the difficulty of having to infer sequence directly from the hybridization pattern of a single sample. Instead, a hybridization pattern is interpreted in the context of a hybridized reference sample.

Comparative re-sequencing approaches

A significant advance towards tackling more complex targets came with the demonstration that polymorphisms can be detected by comparing hybridization patterns on oligonucleotide arrays (Chee *et al.*, 1996; Hacia *et al.*, 1996; Winzeler *et al.*, 1998). In a comparative two-colour hybridization, differentially labelled test and perfect match reference DNAs are co-hybridized to the same array, and the fluorescence ratios provide a probe-by-probe measure of DNA sequence variation. This analysis provides a built-in control for potentially confusing signals resulting from cross-hybridization and multiple probe-target interactions. The same analysis approach has also been applied to reference samples hybridized to separate arrays, due to the high reproducibility of arrays synthesized with photolithography. Both approaches have been successful.

Searches for single nucleotide polymorphism (SNP) signatures in two-colour hybridization data successfully detected DNA variation in 16.6 kb of the human mitochondrion, and 3.5 kb from the hereditary breast and ovarian cancer gene BRCA1 (Chee *et al.*, 1996; Hacia *et al.*, 1996). In addition, a tiling array designed for 705 bp of the rpoB gene of *Mycobacterium tuberculosis* used similar algorithms to successfully assign strains of non-tuberculosis mycobacteria to species (Gingeras *et al.*, 1998). With a reference sample hybridized to a separate array, and by considering exclusively probes that complement the reference sequence, an 8-mer array has also proven functional in comparative hybridization to detect polymorphisms. 90% of the DNA variation in a 2.5 kb target were detected with a false positive rate of < 0.03% (Gunderson *et al.*, 1998). In larger SNP surveys, 149 different arrays designed to match 2 Mb of distinct sequence from 16,000 human sequence tagged sites (STS) were used to scan a total of 14 Mb of DNA for polymorphisms (Wang *et al.*, 1998).

This study offers the first large-scale evaluation of array-based variation detection screens and reports a 90% sensitivity (percentage of SNPs that were detected) and specificity (percentage of true positives). This level of data quality is equal to that of single-pass dideoxy sequencing. In another study, two groups have screened 22 Mb (196 kb from 114 chromosomes) and 28 Mb (190 kb from 148 chromosomes) of sequence, and report sensitivities of 85% and 92%, and specificities of 55% and 83% (Cargill *et al.*, 1999; Halushka *et al.*, 1999). A summary of tiling array polymorphism detection is given in *Table 5.1*.

Performance of array-based variation detection and the feasibility of surveying entire genomes

Comparative tiling array approaches can detect polymorphisms with extreme efficiency. The labour and cost required to manufacture high-density arrays is substantial, but the subsequent analysis is higher throughput than conventional dideoxy sequencing. At an estimated six-fold greater efficiency in cost, time, and effort (Chee *et al.*, 1996), one study has analysed a total of 30 kb of sequence in parallel by hybridization to a single array (Wang *et al.*, 1998). Nevertheless, the specificity and sensitivity is currently still lower than for dideoxy sequencing.

The sensitivity of SNP detection by tiling arrays depends on the complexity of the target sequence. Polymorphisms located in sequences with optimal hybridization behaviour are detected preferentially. This requirement often eliminates SNPs located near other sequence polymorphisms and SNPs in the heterozygous state (Hacia *et al.*, 1999). As a result, the sensitivity of polymorphism detection may be low in cases where high specificity is demanded (*Table 5.1*). When identifying sequence variants in the human genome for the purpose of linkage mapping, a low sensitivity is not detrimental because SNPs are highly abundant (0.5–10 per every 1 kb). For other human studies, however, it may be critical to detect and score all polymorphisms, and exhaustive screens may be needed. Unfortunately, little is known currently about why tiling array signal is sometimes difficult to interpret. With an improved understanding, it may be possible to optimize probe properties or hybridization conditions to achieve a uniform hybridization for all probes. Improvements based on empiricism may yield insights (Gentalen and Chee, 1999; Nguyen *et al.*, 1999; Hacia *et al.*, 1998).

Issues of specificity are easier to resolve. With modest sensitivity, tiling arrays achieve a specificity of polymorphism detection that can be compared to dideoxy sequencing. Under stringent SNP selection criteria, 90% to 100% of the SNPs tagged as certain are accurate (*Table 5.1*) (Wang *et al.*, 1998; Cargill *et al.*, 1999; Halushka *et al.*, 1999). In cases where a modest false-positive rate is acceptable (0–10%), such as for SNP detection screens, or in cases where genotypes are not reported to study subjects, tiling arrays are more efficient than traditional methods (Hacia, 1999). In addition, SNPs that are detected by tiling arrays are easily adapted for hybridization-based genotyping (discussed below).

Current tiling arrays have feature sizes of about 20 μm (Lipshutz *et al.*, 1999), but these may be reduced with advances in photolithographic resolution. With the 2 μm feature sizes of experimental array versions (Lipshutz *et al.*, 1999), the entire human genome can be surveyed on a single 22 × 22 cm^2 array (4 probes per base position). Although synthesis should be possible at even higher density (Fodor *et al.*, 1991),

Table 5.1. Performance of tiling arrays in identifying SNPs

Study	Sequence length (kb)	Number of alleles scanned[a]	True variants expected	Sensitivity estimate (%)	Specificity estimate (%)	Reference
HIV-1	0.382	167	156[b]	NA[c]	NA[c]	(Kozal *et al.*, 1996)
Mycobacteria	0.705	44	12	NA	100	(Gingeras *et al.*, 1998)
Mitochondria	16.6	10	505[d]	NA	100	(Chee *et al.*, 1996)
BRCA-1	3.45	70	22[e]	93	100	(Hacia *et al.*, 1996)
Blood pressure genes	190	148	726[f]	92	83[g]	(Halushka *et al.*, 1999)
Various coding	196	114	560[h]	85	55[i]	(Cargill *et al.*, 1999)
STS	2,000	14	2,473[j]	90	90	(Wang *et al.*, 1998)

[a] For all studies except HIV, bacteria, mitochondria, the number of individuals is half the number of alleles.

[b] 41% of nucleotides are variable.

[c] Chip and dideoxy sequencing were 1.7% discordant.

[d] 180 additional substitution reported in 2.5 kb scans on different arrays in 12 samples. 2.5 kb from each sample were sequence confirmed.

[e] 14/15 expected polymorphisms confirmed and 8 others found.

[f] Estimate based on specificity and the 874 SNPs detected.

[g] SNPs that were predicted as certain were 100% accurate. Since specificity estimates are based on single-pass sequencing, false-positive rate can be anywhere between 11–21%.

[h] Detected by hybridization or DHPLC.

[i] SNPs predicted as certain were 90% accurate.

[j] Estimate based on specificity and the 2748 SNPs detected.

issues of signal detection require thought. As feature size becomes smaller, higher resolution array scanners, or other signal-to-noise detection systems, are required. In addition, with fewer probe molecules at a spot, the average number of molecules hybridized at a spot is reduced for the same hybridization conditions. If these numbers get low enough, variances in the actual number of hybridized targets may become significant. With these considerations in mind, however, the savings in cost and efficiency compared to dideoxy sequencing may increase dramatically as hybridization to small feature sizes becomes feasible and accurate. Current applications in yeast demonstrate the utility of whole-genome variation scans.

In a different approach, a high-density oligonucleotide array designed for gene expression monitoring has been applied to scan about 20% of the genome for DNA variation. The DNA hybridization patterns of two divergent yeast strains have been compared to identify probes with high-intensity signal for the reference and low-intensity for the polymorphic strain. Since the location in the genome of each probe on the array is known, a biallelic marker map consisting of 3714 markers spaced on average every 3.5 kb in the yeast genome, has been constructed. This density of markers permits the construction of a high-resolution inheritance map that locates meiotic breakpoints for the entire yeast genome (Winzeler *et al.*, 1998).

Unfortunately, yeast is currently the only organism for which total genomic DNA (12 Mb) has been hybridized directly to high-density arrays to scan an entire genome for sequence variation. Before whole-genome scans can be realized in humans (6 Gb of a diploid genome or 500-fold more base pairs than a haploid yeast cell), difficulties associated with hybridizations of complex DNA samples need to be overcome. As for all other traditional variation detection methods, PCR amplification is needed to increase target concentration and decrease sample complexity. Nevertheless, it has been possible to hybridize total human RNA to oligonucleotide arrays (total of 120 Mb) and total human genomic DNA to cDNA arrays (Pollack *et al.*, 1999). An increased understanding of the hybridization behaviour of complex DNA targets may make it possible in the future to scan total mammalian genomic DNA for DNA variants.

GENOTYPING DNA VARIATION

The second application of high-density arrays in the study of DNA has focused on genotyping allelic variants. With the large number of DNA variants already identified (*Table 5.1*), biologists can begin to empirically test hypothesis about DNA variants in human populations. With a catalogue of human DNA sequence variation, many questions about the history of the human species, its origin, migration, and the contribution of common DNA variation to natural phenotypes and diseases, can be addressed (reviewed by Chakravarti, 1999). To answer some of these fundamental questions, a tool is needed for cost-efficient and rapid population-based genotyping of thousands of markers.

The challenge of genotyping DNA variants is rather different from that of identifying DNA variation among a pool of individuals. It is extremely important that variants genotyped on individuals are accurate because these may be used in diagnostic tests of disease mutations. As a result, the success of SNP scoring is measured by the accuracy of individual genotype assignments. Oligonucleotides in solution have initially been hybridized to filter-immobilized DNA to genotype specific, known sequence variants (Wallace *et al.*, 1979; Wallace *et al.*, 1981). Then, increasing the number of sequence loci that can be analysed in parallel, oligonucleotides have been immobilized in a low-density array format onto nylon membranes in reverse-dot blots. In a first demonstration, labelled PCR products of HLA-DQA or β-globin were hybridized to immobilized oligonucleotides complementary to six types of HLA-DQA alleles or nine β-thalassemia mutations respectively (Saiki *et al.*, 1989). With probes synthesized directly on glass surfaces, 37 known mutations in the coding region of the cystic fibrosis transmembrane conductance regulator were interrogated (Cronin *et al.*, 1996). In another study, three different mutations in β-thalassemia were detected with oligonucleotides immobilized in gel-coated glass slides (Yershov *et al.*, 1996). Currently, three array technologies have promise for large-scale genotyping. With the capacity to achieve the current highest throughput, tiling arrays tailored for scoring markers have been applied in hybridization-based genotyping. This application will be discussed first. Then an application in which oligonucleotide arrays and generic bar code arrays have been used for minisequencing will be described.

Genotyping with tiling arrays

As for identifying DNA variation, the difficulty of inferring sequence from hybridization patterns is also evident in array-based variation scoring, but there are ways for optimization:

(1) To score heterozygotes that have stoichiometrically only one wild-type allele, tiling arrays designed for genotyping contain a second tiling block with probes complementary to the alternate allele (*Figure 5.6*).
(2) SNPs with similar hybridization requirements can be grouped, and groups placed onto different arrays.
(3) In addition, oligonucleotides on the array surface can be customized in length and direction of flanking sequence, to optimize hybridization results.
(4) Most importantly, however, SNPs can be selected and repeatedly tested on known biological controls to achieve robust and accurate genotype assignments (*Figure 5.6*).

These principles (particularly 1 and 4) have been tested. From an initial number of 558 candidate human SNPs and 412 *Arabidopsis thaliana* SNPs, 68% and 57% could be scored accurately on known controls by hybridization to a tiling array (Wang *et al.*, 1998; Cho *et al.*, 1999). As expected, a comparison of the two studies suggests that the fraction of accurately scored SNPs is higher for SNPs originally identified by hybridization than for those identified by denaturing high-performance liquid chromatography (DHPLC) or sequencing (*Table 5.2*). The difference in numbers reflects the fact that variants identified by tiling array have undergone prior selection for favourable probe-target duplex stability.

Once true markers have been tested on trial samples to confirm their ability to genotype known sequence variants, array-based marker scoring achieves an accuracy of 99.9% on unknown sequences (1611/1613) (Wang *et al.*, 1998). As a result, tiling arrays have been applied to localize an *A. thaliana* mutation to its chromosomal position in a whole genome-mapping study (Cho *et al.*, 1999). In addition, tiling

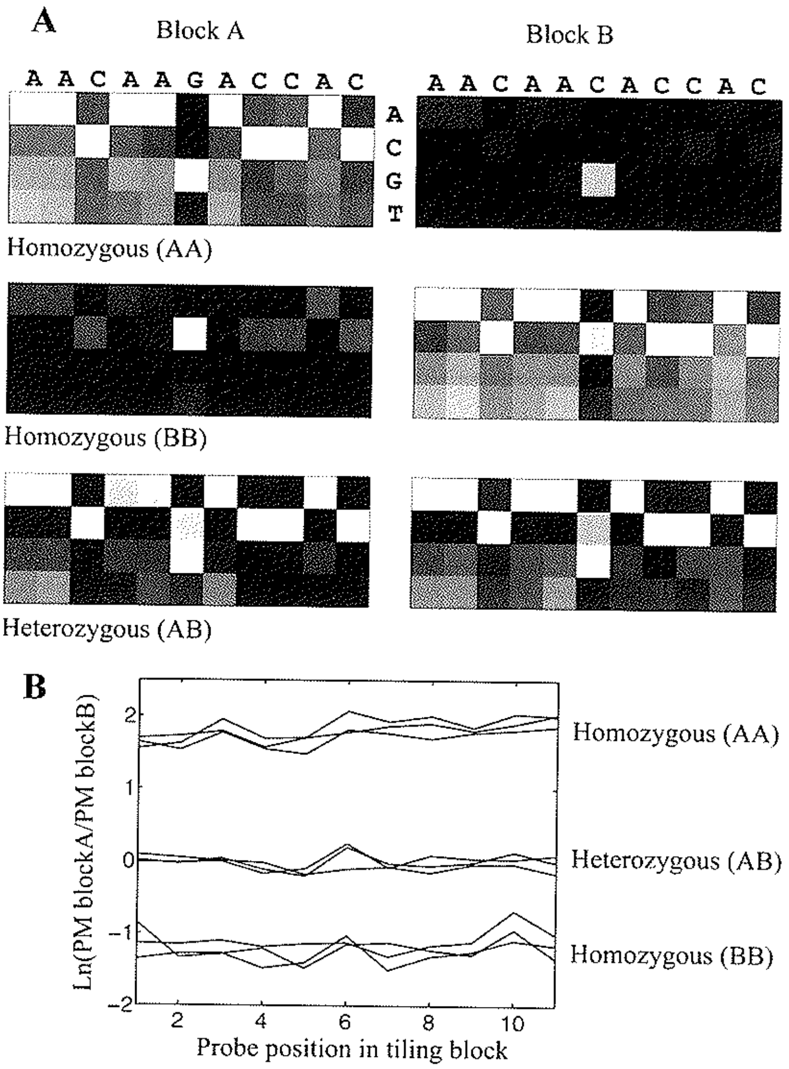

Figure 5.6. Scoring SNPs on a tiling array. (A) Reconstructed array images (highest intensity probe in each column coloured in white). Two probe blocks, one for each SNP allele, show the hybridization patterns expected for homozygous AA, homozygous BB, and heterozygous AB samples.The identity of the base in the reference sequence that is interrogated by the probes in each column is shown above each block. In this case homozygous AA has genotype G/G, homozygous BB, genotype C/C, and heterozygous AB, genotype G/C. (B) Plot of the hybridization intensity ratios versus position on a tiling array for nine samples. For each base position, the intensity ratio has been calculated by dividing the signal intensities for the perfect match reference probes of block A by the perfect match reference probes of block B. Homozygous AA, homozygous BB, and heterozygous AB samples cluster into distinct groups and can be clearly distinguished (data from Cho *et al.*, 1999).

arrays designed for human SNPs have been scored in chimpanzees and gorillas to determine ancestral alleles in the human population. Of 397 human SNPs scored, the most common allele in human populations seems to be ancestral about 76% of the time (Hacia *et al.*, 1999). This value confirms predictions from population genetics (Clark, 1999).

Table 5.2. Efficiency of successive tests to select robust scoring SNPs for tiling array-based marker scoring (Cho *et al.*, 1999; Wang *et al.*, 1998)

	Number of SNPs after successive selections (%)	
Organism	***Arabidopsis thaliana***	***Homo sapiens***
Detection method	DHPLC & dideoxy sequencing	Tiling array
Total number of candidate SNPs	412 (100)	558 (100)
Good signal after singleplex PCR	351 (85)	547 (98)
Good signal after multiplex PCR	332 (81)	500 (90)
Robust scoring SNPs	235 (57)	378 (68)

Combining minisequencing and hybridization to DNA arrays

In addition to hybridization-based variation scoring, powerful minisequencing assays have coupled target hybridization with enzymatic primer extension reactions. In this approach, oligonucleotides synthesized one base short of the polymorphic position are hybridized to PCR amplified target sequences. In a single-base primer extension reaction, a labelled nucleoside triphosphate complementary to the polymorphic base is incorporated. Initially, extensions were executed with genomic DNA template and minisequencing primers, both in solution, and with a single type of radiolabelled deoxynucleoside triphosphate (dNTP) per reaction (Sokolov, 1990). Then, template DNA was PCR amplified and rendered single-stranded by attachment to an avidin matrix (Syvanen *et al.*, 1990). The attachment of templates to microtitre wells permitted the parallel screening of multiple nucleotide variants in multiple individuals (Syvanen *et al.*, 1993). In a modified approach, oligonucleotide primers were fixed to the microtitre wells instead, and primer extensions were carried out with two differentially labelled dideoxynucleoside triphosphates (ddNTP). By fixing the primers to the solid surface, template DNA could be removed, thereby eliminating possible nonspecific signals from 3' extensions of the template. The use of two different colourimetric assays permitted two target nucleotides to be interrogated in a single extension reaction. In addition, extensions with ddNTPs ensured that only one nucleotide is added to each primer (Nikiforov *et al.*, 1994). Then primers were arrayed onto glass surfaces, where arrayed primers could be extended in parallel in a single multiplex extension reaction. Although applied mainly with radioactive dNTPs (Shumaker *et al.*, 1996) and ddNTPs (Pastinen *et al.*, 1997), glass surfaces make possible the use of multiple differentially fluorescent ddNTPs in the same reaction. In principle, a high-density array of primers and four differentially labelled ddNTPs could be used to interrogate many sequence loci simultaneously for all possible target nucleotides.

Minisequencing seems to overcome some of the difficulties, such as low signal, experienced with hybridization. Only a single primer is needed for each SNP. In addition, primer-template annealing can be performed at lower temperatures

(20–37°C) and single-base extension can be carried out at higher temperatures (~ 60°C) than for allele-specific hybridization. The high temperature resolves secondary structures in targets and increases specificity. At the high temperature, the polymerase extension either proceeds so fast that short-lived duplexes do not matter, or stabilizes the probe-target duplex so that comparable signal is obtained for different probes (in a review by Southern *et al.*, 1999). The selectivity of polymerase assures that only perfectly annealed 3' ends of probes are extended. In addition, since the genotype assignment depends on the identity of the extended base and polymerases have a high fidelity, high levels of allele discrimination can, in principle, be achieved. As a result, genotyping signal and discrimination in cases has been better than by direct hybridization (Pastinen *et al.*, 1997), but larger studies are required for meaningful comparisons: including an assessment of the number of markers that fail because the flanking sequence environment precludes the design of appropriate extension primers.

In a more complex sequence analysis, mutations in a 33-base region of p53 have been identified and scored by re-sequencing the entire interval with a set of tiled primers immobilized on glass slides. By spotting a second oligonucleotide complementary to the antisense stand, some redundancy has been added and further confidence in base calls obtained (Head *et al.*, 1997). Nevertheless, a transition to higher density array formats is needed to scale to the genome. Such a transition is currently complicated, as probes on oligonucleotide arrays manufactured by photolithography and photosensitive oligonucleotide synthesis chemistry cannot be extended in a minisequencing reaction because the oligonucleotides are synthesized 3' to 5'; the wrong sense to act as a primer. As alternate synthesis methods should be possible, this need not be a permanent limitation.

Although a direct primer extension is not possible on high-density oligonucleotide arrays, these arrays have, nevertheless, been applied in an alternate strategy. Bar code arrays that contain 20-mer oligonucleotides of similar sequence composition and hybridization efficiency (Shoemaker *et al.*, 1996) provide a means of adapting the minisequencing approach to a generic high-density oligonucleotide array format. Primers are synthesized by attaching a flanking sequence, complementary to a unique bar code on the array, to a minisequencing primer. Several primers are pooled and extended in a minisequencing reaction with PCR amplified targets and fluorescent nucleoside triphosphates. Each primer extension assay is then decoded by hybridization to the bar code array (Sklar and Hirschhorn, unpublished data reported in Lander, 1999).

Scaling to the genome

Minisequencing reactions have promise for being high-throughput, but more applications are needed to prove their suitability for large-scale genotyping. Tiling arrays have been successfully tested, but current studies need to be increased considerably to reach genome capacity. The major limitation is the requirement for PCR amplification. The number of primers that can be combined in a multiplex PCR reaction is limited. 45 primer pairs in a single reaction yield about a 90% success rate (Wang *et al.*, 1998; Cho *et al.*, 1999). Under these conditions, 500,000 SNPs that are needed per individual for genome-wide association studies (Kruglyak, 1999), can be amplified in about 11,000 multiplex PCR reactions or 29 multi-well (384) plates! With current methods for PCR amplification and costs of synthesizing 1,000,000 unique oligonucleotide primers, it will be difficult to analyse multiple individuals. The possibility of coupling high-density solid surface oligonucleotide synthesis with PCR amplification may yield new advances.

Another concern is the number of arrays that is required. With standard synthesis techniques yielding 300,000 probes on a 1.28 × 1.28 cm^2 array and redundant probe arrangements (80 probes per SNP, 20 per strand) (Lipshutz *et al.*, 1999), about 3,700 SNPs can be scored on a single tiling array. 500,000 SNPs would currently require about 130 array designs. Although feasible for proof-of-principle experiments, the need for smaller feature size or fewer probes is apparent for high-throughput, genome-wide linkage disequilibrium studies in which a large number of individuals need to be genotyped. With the 2 μm feature sizes of current experimental arrays (Lipshutz *et al.*, 1999) a single 1.28 × 1.28 cm^2 array is sufficient to hold the highly redundant set of 40,000,000 probes. However, it is already known that 80 probes per marker is excessive. Optimally, 1–2 probes per SNP should suffice, as single probes have been sufficient for scoring polymorphisms in haploid yeast (Winzeler *et al.*, 1998). Under these conditions, all probes for 500,000 SNPs could fit on 2–4 arrays of current size. Alternatively, because minisequencing also requires only a single probe per SNP, major advances in minisequencing could also solve the current array problem.

IV. Other applications of solid surface high-density array platforms

Although gene expression analysis and DNA variation detection have been the primary applications of array technology, other array applications deserve mention and will give the reader an idea of the wide range of array-based experimentation that is underway:

(1) In a strategy that does not involve scoring DNA variation throughout the genome, cDNA arrays have been applied to map inherited loci. Without genotyping DNA variation, hybridization to a cDNA array has been successful at identifying chromosomal segments that are identical by descent between related genomes (Cheung *et al.*, 1998). Genomic mismatch scanning exploits the ability of mismatch enzymes to recognize and cleave mismatched DNA strands and, after several biochemical steps, the sample is enriched for DNA that is identical by descent (Nelson *et al.*, 1993). Although technically demanding and labour intensive, because genomic mismatch scanning does not require a map of DNA variation, it can be used to map inherited loci in organisms for which sequence information is not available.

(2) cDNA arrays have also been applied to survey DNA copy-number variation. In a comparative genomic hybridization, test and reference samples with different fluorescent labels are co-hybridized to an array of DNA clones. Fluorescence ratios provide a measure of DNA copy-number variation for each locus (Solinas-Toldo *et al.*, 1997; Pinkel *et al.*, 1998; Geschwind *et al.*, 1998). In such an analysis, a cDNA array with 30,000 human genes has been applied to measure DNA copy-number variation in breast cancer cell lines and human tumours (Pollack *et al.*, 1999). An advantage of using a cDNA array is that gene

expression levels of the same sample can be monitored in parallel on an array of the same format. This application promises to help identify important genes in tumour samples for diagnosis and the development of treatments.

(3) In another application of high-density oligonucleotide arrays, generic bar code arrays have been applied in a comprehensive and systematic deletion phenotype analysis of the yeast genome (Shoemaker *et al.*, 1996; see also Hensel *et al.*, 1995). Deletion strains, each with one of the 6,200 yeast genes deleted and molecularly tagged with unique 20 bp bar code sequences, can be pooled and grown competitively. After selection, a common sequence flanking each bar code is used to amplify all bar codes in a single PCR with a single primer pair; effectively solving the multiplexing problem. The amplified sequences are then hybridized to a high-density array containing probes complementary to each tag which has been selected for common, optimal hybridization characteristics. The change in hybridization intensity over time can be used to quantitate the growth rates for all bar coded strains in parallel. This strategy shifts traditional deletion phenotype analysis into high gear, and has successfully identified genes essential for growth in complete and minimal media (Winzeler *et al.*, 1999). With this technology, it is feasible to perform rapid phenotypic selections under a variety of environmental conditions. A similar technique has also been employed to identify drug targets. Deletion strains, heterozygous for all potential drug targets, have been screened against a drug compound. The strain most sensitive to the compound has provided strong clues to the drug target (Giaever *et al.*, 1999).

Conclusion

The broad range of array-based experimentation illustrates the utility of a global approach to biology. Many experiments that were unthinkable only a few years ago are now in progress. This advance illustrates how technological breakthroughs allow for new ways of investigating biology. Nevertheless, the technology and its applications for studying the DNA sequence of an entire genome are in their infancy, and many challenges lie ahead.

In the research setting, the main challenge resides in data analysis. Existing tools are still rudimentary, particularly for the interpretations of gene expression patterns. Global patterns of transcription can be applied to achieve different goals – the identification of functionally or regulatory related genes and of genes which define a transcriptional phenotype – and analysis tools need to be tailored accordingly. A variety of statistical tools already exist, and may prove applicable in distinct biological settings. In addition to expanding applications of data analysis, consideration must be given to the accuracy of whole genome approaches. A small number of data points can be readily checked by hand, but parallel assays that generate thousands of data values pose a substantially greater challenge. A measure of confidence associated with each data point is required. Such standards for genome analysis will make it meaningful to integrate the expression level of genes and their allelic variation with existing knowledge of genetic, protein-protein, and small molecule interactions. This integration is essential to put a specific discovery involving a small set of genes into the context of an entire genome.

In particular, in the clinical setting, it is often difficult with current technology to obtain high-quality data from limited amounts of human cells or tissues. This limitation increases in importance considering that high-density array technology has promise for point-of-care disease diagnosis. Reliable sample amplification techniques or hybridization procedures that operate with less starting material are needed. These hurdles, among others, need to be overcome to be able to proceed from a detailed list of allelic variation in genomes to a mechanistic understanding of how genotype and phenotype relate in humans. In addition, genome science yields an opportunity to uncover the effects of the environment, as revealed by the cellular levels of RNA, protein, or small molecular weight compounds. With the future of medicine pointed in the direction of genome-scale genotyping, it may ultimately become possible to tailor disease treatments on the basis of information revealed by a patient's genome composition as well as by the presence of distinct expression signatures, describing a patient's environmental past.

Undoubtedly, the future will see further refinement in array miniaturization that will allow for an even higher level of throughput and a further advance in global experimentation. Other possible detection methods may become attractive as spot size is decreased. Akin to other main technological advances in biology, such as commercial oligonucleotide synthesis, it seems likely that, as arrays become standard tools for biological investigation, they will also be purchased. As their use is increased, so will the demand for inexpensive, customizable, and highly reliable array platforms. These technical improvements in array technology will bear immediate fruits for biology. As the applications discussed in this review demonstrate, high-density, array-based, global approaches to biology could represent a rapid, powerful, and comprehensive first step towards understanding the entire DNA sequence content of the human genome.

Acknowledgements

We thank our colleagues for helpful discussion and Kevin White for critical comments on the manuscript. This work was supported by a grant from NIH (2R01 HG00198) and by a Howard Hughes Medical Institute predoctoral fellowship awarded to L.M. Steinmetz.

References

ADAM, G. AND DELBRUCK, M. (1968). Reduction of dimensionality in biological diffusion processes. In: *Structural chemistry and molecular biology*. Eds. A. Rich and N. Davidson, pp 198–215. San Francisco: W. H. Freeman.

BAINS, W. AND SMITH, G.C. (1988). A novel method for nucleic acid sequence determination. *Journal of Theoretical Biology* **135**, 303–307.

BASSETT, D.E., JR., EISEN, M.B. AND BOGUSKI, M.S. (1999). Gene expression informatics – it's all in your mine. *Nature Genetics* **21**, 51–55.

BENTON, W.D. AND DAVIS, R.W. (1977). Screening λgt recombinant clones by hybridization to single plaques in situ. *Science* **196**, 180–182.

BITTNER, M., MELTZER, P. AND TRENT, J. (1999). Data analysis and integration: of steps and arrows. *Nature Genetics* **22**, 213-215.

BLANCHARD, A.P., KAISER, R.J. AND HOOD, L.E. (1996). Synthetic DNA arrays. *Biosensors and Bioelectronics* **11**, 687–690.

BOWTELL, D.D. (1999). Options available – from start to finish – for obtaining expression data by microarray. *Nature Genetics* **21**, 25–32.

BRENNER, S. (1999). Sillycon valley fever. *Current Biology* **9**, R671.

BROWN, P.O. AND BOTSTEIN, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nature Genetics* **21**, 33–37.

CARGILL, M., ALTSHULER, D., IRELAND, J., SKLAR, P., ARDLIE, K., PATIL, N., LANE, C.R., LIM, E.P., KALAYANARAMAN, N., NEMESH, J., ZIAUGRA, L., FRIEDLAND, L., ROLFE, A., WARRINGTON, J., LIPSHUTZ, R., DALEY, G.Q. AND LANDER, E.S. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics* **22**, 231–238.

CHAKRAVARTI, A. (1999). Population genetics – making sense out of sequence. *Nature Genetics* **21**, 56–60.

CHAN, V., GRAVES, D.J. AND MCKENZIE, S.E. (1995). The biophysics of DNA hybridization with immobilized oligonucleotide probes. *Biophysical Journal* **69**, 2243–2255.

CHAN, V., GRAVES, D.J., FORTINA, P. AND MCKENZIE, S.E. (1997). Adsorption and surface diffusion of DNA oligonucleotides at liquid/solid interfaces. *Langmuir* **13**, 320–329.

CHEE, M., YANG, R., HUBBELL, E., BERNO, A., HUANG, X.C., STERN, D., WINKLER, J., LOCKHART, D.J., MORRIS, M.S. AND FODOR, S.P. (1996). Accessing genetic information with high-density DNA arrays. *Science* **274**, 610–614.

CHEN, Y., DOUGHERTY, E.R. AND BITTNER, M.L. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics* **2**, 364–374.

CHEN, Y., BITTNER, M.L. AND DOUGHERTY, E.R. (1999). Issues associated with microarray data analysis and integration. *Nature Genetics*, supplementary to Bittner *et al.*, 1999.

CHEUNG, V.G., GREGG, J.P., GOGOLIN-EWENS, K.J., BANDONG, J., STANLEY, C.A., BAKER, L., HIGGINS, M.J., NOWAK, N.J., SHOWS, T.B., EWENS, W.J., NELSON, S.F. AND SPIELMAN, R.S. (1998). Linkage-disequilibrium mapping without genotyping. *Nature Genetics* **18**, 225–230.

CHEUNG, V.G., MORLEY, M., AGUILAR, F., MASSIMI, A., KUCHERLAPATI, R. AND CHILDS, G. (1999). Making and reading microarrays. *Nature Genetics* **21**, 15–19.

CHO, R.J., CAMPBELL, M.J., WINZELER, E.A., STEINMETZ, L., CONWAY, A., WODICKA, L., WOLFSBERG, T.G., GABRIELIAN, A.E., LANDSMAN, D., LOCKHART, D.J. AND DAVIS, R.W. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell* **2**, 65–73.

CHO, R.J., MINDRINOS, M., RICHARDS, D.R., SAPOLSKY, R.J., ANDERSON, M., DRENKARD, E., DEWDNEY, J., REUBER, T.L., STAMMERS, M., FEDERSPIEL, N., THEOLOGIS, A., YANG, W., HUBBELL, E., LASHKARI, D., LEMIEUX, B., DEAN, C., LIPSHUTZ, R.J., AUSUBEL, F.M., DAVIS, R.W. AND OEFNER, P.J. (1999). Genome-wide mapping with biallelic markers in Arabidopsis thaliana. *Nature Genetics* **23**, 203–207.

CHU, S., DERISI, J., EISEN, M., MULHOLLAND, J., BOTSTEIN, D., BROWN, P.O. AND HERSKOWITZ, I. (1998). The transcriptional program of sporulation in budding yeast. *Science* **282**, 699–705.

CLARK, A.G. (1999). Chips for chimps. *Nature Genetics* **22**, 119–120.

CRONIN, M.T., FUCINI, R.V., KIM, S.M., MASINO, R.S., WESPI, R.M. AND MIYADA, C.G. (1996). Cystic fibrosis mutation detection by hybridization to light-generated DNA probe arrays. *Human Mutation* **7**, 244–255.

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

DRMANAC, R., LABAT, I., BRUKNER, I. AND CRKVENJAKOV, R. (1989). Sequencing of megabase plus DNA by hybridization: theory of the method. *Genomics* **4**, 114–128.

DRMANAC, S. AND DRMANAC, R. (1994). Processing of cDNA and genomic kilobase-size clones for massive screening, mapping and sequencing by hybridization. *Biotechniques* **17**, 328–336.

DUGGAN, D.J., BITTNER, M., CHEN, Y., MELTZER, P. AND TRENT, J.M. (1999). Expression profiling using cDNA microarrays. *Nature Genetics* **21**, 10–14.

Editorial (1999). Array data go public. *Nature Genetics* **22**, 211–212.

EISEN, M.B., SPELLMAN, P.T., BROWN, P.O. AND BOTSTEIN, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 14863–14868.

ERMOLAEVA, O., RASTOGI, M., PRUITT, K.D., SCHULER, G.D., BITTNER, M.L., CHEN, Y., SIMON, R., MELTZER, P., TRENT, J.M. AND BOGUSKI, M.S. (1998). Data management and analysis for gene expression arrays. *Nature Genetics* **20**, 19–23.

FAMBROUGH, D., MCCLURE, K., KAZLAUSKAS, A. AND LANDER, E.S. (1999). Diverse signaling pathways activated by growth factor receptors induce broadly overlapping, rather than independent, sets of genes. *Cell* **97**, 727–741.

FIELDS, S. (1997). The future is function. *Nature Genetics* **15**, 325–327.

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.

GENTALEN, E. AND CHEE, M. (1999). A novel method for determining linkage between DNA sequences: hybridization to paired probe arrays. *Nucleic Acids Research* **27**, 1485–1491.

GESCHWIND, D.H., GREGG, J., BOONE, K., KARRIM, J., PAWLIKOWSKA-HADDAL, A., RAO, E., ELLISON, J., CICCODICOLA, A., D'URSO, M., WOODS, R., RAPPOLD, G.A., SWERDLOFF, R. AND NELSON, S.F. (1998). Klinefelter's syndrome as a model of anomalous cerebral laterality: testing gene dosage in the X chromosome pseudoautosomal region using a DNA microarray. *Developmental Genetics* **23**, 215–229.

GIAEVER, G., SHOEMAKER, D.D., JONES, T.W., LIANG, H., WINZELER, E.A., ASTROMOFF, A. AND DAVIS, R.W. (1999). Genomic profiling of drug sensitivities via induced haploinsufficiency. *Nature Genetics* **21**, 278–283.

GILLESPIE, D. AND SPIEGELMAN, S. (1965). A quantitative assay for DNA-RNA hybrids with DNA immobilized on a membrane. *Journal of Molecular Biology* **12**, 829–842.

GINGERAS, T.R., GHANDOUR, G., WANG, E., BERNO, A., SMALL, P.M., DROBNIEWSKI, F., ALLAND, D., DESMOND, E., HOLODNIY, M. AND DRENKOW, J. (1998). Simultaneous genotyping and species identification using hybridization pattern recognition analysis of generic Mycobacterium DNA arrays. *Genome Research* **8**, 435–448.

GLAZER, A.N., PECK, K. AND MATHIES, R.A. (1990). A stable double-stranded DNA-ethidium homodimer complex: application to picogram fluorescence detection of DNA in agarose gels. *Proceedings of the National Academy of Sciences of the United States of America* **87**, 3851–3855.

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**, 531–537.

GRUNSTEIN, M. AND HOGNESS, D.S. (1975). Colony hybridization: a method for the isolation of cloned DNAs that contain a specific gene. *Proceedings of the National Academy of Sciences of the United States of America* **72**, 3961–3965.

GUNDERSON, K.L., HUANG, X.C., MORRIS, M.S., LIPSHUTZ, R.J., LOCKHART, D.J. AND CHEE, M.S. (1998). Mutation detection by ligation to complete n-mer DNA arrays. *Genome Research* **8**, 1142–1153.

GUO, Z., GUILFOYLE, R.A., THIEL, A.J., WANG, R. AND SMITH, L.M. (1994). Direct fluorescence analysis of genetic polymorphisms by hybridization with oligonucleotide arrays on glass supports. *Nucleic Acids Research* **22**, 5456–5465.

HACIA, J.G. (1999). Resequencing and mutational analysis using oligonucleotide microarrays. *Nature Genetics* **21**, 42–47.

HACIA, J.G., BRODY, L.C., CHEE, M.S., FODOR, S.P. AND COLLINS, F.S. (1996). Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis. *Nature Genetics* **14**, 441–447.

HACIA, J.G., WOSKI, S.A., FIDANZA, J., EDGEMON, K., HUNT, N., MCGALL, G., FODOR, S.P. AND COLLINS, F.S. (1998). Enhanced high density oligonucleotide array-based sequence analysis using modified nucleoside triphosphates. *Nucleic Acids Research* **26**, 4975–4982.

HACIA, J.G., FAN, J.B., RYDER, O., JIN, L., EDGEMON, K., GHANDOUR, G., MAYER, R.A., SUN, B., HSIE, L., ROBBINS, C.M., BRODY, L.C., WANG, D., LANDER, E.S., LIPSHUTZ, R., FODOR, S.P. AND COLLINS, F.S. (1999). Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nature Genetics* **22**, 164–167.

HALUSHKA, M.K., FAN, J.B., BENTLEY, K., HSIE, L., SHEN, N., WEDER, A., COOPER, R., LIPSHUTZ, R. AND CHAKRAVARTI, A. (1999). Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genetics* **22**, 239–247.

HEAD, S.R., ROGERS, Y.H., PARIKH, K., LAN, G., ANDERSON, S., GOELET, P. AND BOYCE-JACINO, M.T. (1997). Nested genetic bit analysis (N-GBA) for mutation detection in the p53 tumor suppressor gene. *Nucleic Acids Research* **25**, 5065–5071.

HELLER, R.A., SCHENA, M., CHAI, A., SHALON, D., BEDILION, T., GILMORE, J., WOOLLEY, D.E. AND DAVIS, R.W. (1997). Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 2150–2155.

HENSEL, M., SHEA, J.E., GLEESON, C., JONES, M.D., DALTON, E. AND HOLDEN, D.W. (1995). Simultaneous identification of bacterial virulence genes by negative selection. *Science* **269**, 400–403.

HOLSTEGE, F.C., JENNINGS, E.G., WYRICK, J.J., LEE, T.I., HENGARTNER, C.J., GREEN, M.R., GOLUB, T.R., LANDER, E.S. AND YOUNG, R.A. (1998). Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**, 717–728.

IYER, V.R., EISEN, M.B., ROSS, D.T., SCHULER, G., MOORE, T., LEE, J.C.F., TRENT, J.M., STAUDT, L.M., HUDSON, J., JR., BOGUSKI, M.S., LASHKARI, D., SHALON, D., BOTSTEIN, D. AND BROWN, P.O. (1999). The transcriptional program in the response of human fibroblasts to serum. *Science* **283**, 83–87.

JACOB, F. AND MONOD, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology* **3**, 318–356.

KAFATOS, F.C., JONES, C.W. AND EFSTRATIADIS, A. (1979). Determination of nucleic acid sequence homologies and relative concentrations by a dot hybridization procedure. *Nucleic Acids Research* **7**, 1541–1552.

KHRAPKO, K.R., LYSOV, YU P., KHORLYN, A.A., SHICK, V.V., FLORENTIEV, V.L. AND MIRZABEKOV, A.D. (1989). An oligonucleotide hybridization approach to DNA sequencing. *FEBS Letters* **256**, 118–122.

KHRAPKO, K.R., KHORLIN, A.A., IVANOV, I.B., CHERNOV, B.K., LYSOV, YU P., VASILENKO, S.K., FLORENT'EV, V.L. AND MIRZABEKOV, A.D. (1991). Hybridization of DNA with oligonucleotides immobilized in a gel: a convenient method for recording single base replacements. *Molekuliarnaia Biologiia* **25**, 718–730.

KOZAL, M.J., SHAH, N., SHEN, N., YANG, R., FUCINI, R., MERIGAN, T.C., RICHMAN, D.D., MORRIS, D., HUBBELL, E., CHEE, M. AND GINGERAS, T.R. (1996). Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays. *Nature Medicine* **2**, 753–759.

KRUGLYAK, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics* **22**, 139–144.

LANDER, E.S. (1999). Array of hope. *Nature Genetics* **21**, 3–4.

LENNON, G.G. AND LEHRACH, H. (1991). Hybridization analyses of arrayed cDNA libraries. *Trends in Genetics* **7**, 314–317.

LIPSHUTZ, R.J., FODOR, S.P., GINGERAS, T.R. AND LOCKHART, D.J. (1999). High density synthetic oligonucleotide arrays. *Nature Genetics* **21**, 20–24.

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**, 1675–1680.

MARTON, M.J., DERISI, J.L., BENNETT, H.A., IYER, V.R., MEYER, M.R., ROBERTS, C.J., STOUGHTON, R., BURCHARD, J., SLADE, D., DAI, H., BASSETT, D.E., JR., HARTWELL, L.H., BROWN, P.O. AND FRIEND, S.H. (1998). Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Medicine* **4**, 1293–1301.

MASKOS, U. AND SOUTHERN, E.M. (1992a). Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ. *Nucleic Acids Research* **20**, 1679–1684.

MASKOS, U. AND SOUTHERN, E.M. (1992b). Parallel analysis of oligodeoxyribonucleotide (oligonucleotide) interactions. I. Analysis of factors influencing oligonucleotide duplex formation. *Nucleic Acids Research* **20**, 1675–1678.

MCGALL, G., LABADIE, J., BROCK, P., WALLRAFF, G., NGUYEN, T. AND HINSBERG, W. (1996). Light-directed synthesis of high-density oligonucleotide arrays using semiconductor photoresists. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 13555–13560.

MCKENZIE, S.E., MANSFIELD, E., RAPPAPORT, E., SURREY, S. AND FORTINA, P. (1998). Parallel molecular genetic analysis. *European Journal of Human Genetics* **6**, 417–429.

MICHAELS, G.S., CARR, D.B., ASKENAZI, M., FUHRMAN, S., WEN, X. AND SOMOGYI, R. (1998). Cluster analysis and data visualization of large-scale gene expression data. *Pacific Symposium on Biocomputing*, 42–53.

NELSON, S.F., MCCUSKER, J.H., SANDER, M.A., KEE, Y., MODRICH, P. AND BROWN, P.O. (1993). Genomic mismatch scanning: a new approach to genetic linkage mapping. *Nature Genetics* **4**, 11–18.

NGUYEN, H.K., FOURNIER, O., ASSELINE, U., DUPRET, D. AND THUONG, N.T. (1999). Smoothing of the thermal stability of DNA duplexes by using modified nucleosides and chaotropic agents. *Nucleic Acids Research* **27**, 1492–1498.

NIKIFOROV, T.T., RENDLE, R.B., GOELET, P., ROGERS, Y.H., KOTEWICZ, M.L., ANDERSON, S., TRAINOR, G.L. AND KNAPP, M.R. (1994). Genetic Bit Analysis: a solid phase method for typing single nucleotide polymorphisms. *Nucleic Acids Research* **22**, 4167–4175.

PASTINEN, T., KURG, A., METSPALU, A., PELTONEN, L. AND SYVANEN, A.C. (1997). Minisequencing: a specific tool for DNA analysis and diagnostics on oligonucleotide arrays. *Genome Research* **7**, 606–614.

PAWSON, T. AND SAXTON, T.M. (1999). Signaling networks – do all roads lead to the same genes? *Cell* **97**, 675–678.

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**, 5022–5026.

PEROU, C.M., JEFFREY, S.S., VAN DE RIJN, M., REES, C.A., EISEN, M.B., ROSS, D.T., PERGAMENSCHIKOV, A., WILLIAMS, C.F., ZHU, S.X., LEE, J.C., LASHKARI, D., SHALON, D., BROWN, P.O. AND BOTSTEIN, D. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 9212–9217.

PINKEL, D., SEGRAVES, R., SUDAR, D., CLARK, S., POOLE, I., KOWBEL, D., COLLINS, C., KUO, W.L., CHEN, C., ZHAI, Y., DAIRKEE, S.H., LJUNG, B.M., GRAY, J.W. AND ALBERTSON, D.G. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics* **20**, 207–211.

PLOMIN, R., OWEN, M.J. AND MCGUFFIN, P. (1994). The genetic basis of complex human behaviors. *Science* **264**, 1733–1739.

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**, 41–46.

RISCH, N. AND MERIKANGAS, K. (1996). The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517.

SAIKI, R.K., WALSH, P.S., LEVENSON, C.H. AND ERLICH, H.A. (1989). Genetic analysis of amplified DNA with immobilized sequence-specific oligonucleotide probes. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 6230-6234.

SCHAFER, A.J. AND HAWKINS, J.R. (1998). DNA variation and the future of human genetics. *Nature Biotechnology* **16**, 33–39.

SCHENA, M., SHALON, D., DAVIS, R.W. AND BROWN, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470.

SCHENA, M., HELLER, R.A., THERIAULT, T.P., KONRAD, K., LACHENMEIER, E. AND DAVIS, R.W. (1998). Microarrays: biotechnology's discovery platform for functional genomics. *Trends in Biotechnology* **16**, 301–306.

SHALON, D., SMITH, S.J. AND BROWN, P.O. (1996). A DNA microarray system for analyzing complex DNA samples using two- color fluorescent probe hybridization. *Genome Research* **6**, 639–645.

SHOEMAKER, D.D., LASHKARI, D.A., MORRIS, D., MITTMANN, M. AND DAVIS, R.W. (1996).

Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nature Genetics* **14**, 450–456.

SHUMAKER, J.M., METSPALU, A. AND CASKEY, C.T. (1996). Mutation detection by solid phase primer extension. *Human Mutation* **7**, 346–354.

SOKOLOV, B.P. (1990). Primer extension technique for the detection of single nucleotide in genomic DNA. *Nucleic Acids Research* **18**, 3671.

SOLINAS-TOLDO, S., LAMPEL, S., STILGENBAUER, S., NICKOLENKO, J., BENNER, A., DOHNER, H., CREMER, T. AND LICHTER, P. (1997). Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes, Chromsomes and Cancer* **20**, 399–407.

SOSNOWSKI, R.G., TU, E., BUTLER, W.F., O'CONNELL, J.P. AND HELLER, M.J. (1997). Rapid determination of single base mismatch mutations in DNA hybrids by direct electric field control. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 1119–1123.

SOUTHERN, E., MIR, K. AND SHCHEPINOV, M. (1999). Molecular interactions on microarrays. *Nature Genetics* **21**, 5–9.

SOUTHERN, E.M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology* **98**, 503–517.

SOUTHERN, E.M., MASKOS, U. AND ELDER, J.K. (1992). Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models. *Genomics* **13**, 1008–1017.

SPELLMAN, P.T., SHERLOCK, G., ZHANG, M.Q., IYER, V.R., ANDERS, K., EISEN, M.B., BROWN, P.O., BOTSTEIN, D. AND FUTCHER, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell* **9**, 3273–3297.

SYVANEN, A.C., AALTO-SETALA, K., HARJU, L., KONTULA, K. AND SODERLUND, H. (1990). A primer-guided nucleotide incorporation assay in the genotyping of apolipoprotein E. *Genomics* **8**, 684–692.

SYVANEN, A.C., SAJANTILA, A. AND LUKKA, M. (1993). Identification of individuals by analysis of biallelic DNA markers, using PCR and solid-phase minisequencing. *American Journal of Human Genetics* **52**, 46–59.

TAMAYO, P., SLONIM, D., MESIROV, J., ZHU, Q., KITAREEWAN, S., DMITROVSKY, E., LANDER, E.S. AND GOLUB, T.R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 2907–2912.

TAVAZOIE, S., HUGHES, J.D., CAMPBELL, M.J., CHO, R.J. AND CHURCH, G.M. (1999). Systematic determination of genetic network architecture. *Nature Genetics* **22**, 281–285.

TORONEN, P., KOLEHMAINEN, M., WONG, G. AND CASTREN, E. (1999). Analysis of gene expression data using self-organizing maps. *FEBS Letters* **451**, 142–146.

WALLACE, R.B., SHAFFER, J., MURPHY, R.F., BONNER, J., HIROSE, T. AND ITAKURA, K. (1979). Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch. *Nucleic Acids Research* **6**, 3543–3557.

WALLACE, R.B., JOHNSON, M.J., HIROSE, T., MIYAKE, T., KAWASHIMA, E.H. AND ITAKURA, K. (1981). The use of synthetic oligonucleotides as hybridization probes. II. Hybridization of oligonucleotides of mixed sequence to rabbit beta-globin DNA. *Nucleic Acids Research* **9**, 879–894.

WANG, D.G., FAN, J.B., SIAO, C.J., BERNO, A., YOUNG, P., SAPOLSKY, R., GHANDOUR, G., PERKINS, N., WINCHESTER, E., SPENCER, J., KRUGLYAK, L., STEIN, L., HSIE, L., TOPALOGLOU, T., HUBBELL, E., ROBINSON, E., MITTMANN, M., MORRIS, M.S., SHEN, N., KILBURN, D., RIOUX, J., NUSBAUM, C., ROZEN, S., HUDSON, T.J., LIPSHUTZ, R., CHEE, M. AND LANDER, E.S. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082.

WEN, X., FUHRMAN, S., MICHAELS, G.S., CARR, D.B., SMITH, S., BARKER, J.L. AND SOMOGYI, R. (1998). Large-scale temporal gene expression mapping of central nervous system development. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 334–339.

WETMUR, J.G. AND DAVIDSON, N. (1968). Kinetics of renaturation of DNA. *Journal of Molecular Biology* **31**, 349–370.

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**, 1194–1197.

WINZELER, E.A., SHOEMAKER, D.D., ASTROMOFF, A., LIANG, H., ANDERSON, K., ANDRE, B., BANGHAM, R., BENIDO, R., BOEKE, J.D., BUSSEY, H., CHU, A.M., CONNELLY, C., DAVIS, K., DIETRICH, F., WHELEN DOW, S., ELBAKKOURY, 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., STORMS, R.K., VERONNEAU, 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**, 901–906.

WITTES, J. AND FRIEDMAN, H.P. (1999). Searching for evidence of altered gene expression: a comment on statistical analysis of microarray data. *Journal of the National Cancer Institute* **91**, 400–401.

WODICKA, L., DONG, H., MITTMANN, M., HO, M.H. AND LOCKHART, D.J. (1997). Genome-wide expression monitoring in Saccharomyces cerevisiae. *Nature Biotechnology* **15**, 1359–1367.

YERSHOV, G., BARSKY, V., BELGOVSKIY, A., KIRILLOV, E., KREINDLIN, E., IVANOV, I., PARINOV, S., GUSCHIN, D., DROBISHEV, A., DUBILEY, S. AND MIRZABEKOV, A. (1996). DNA analysis and diagnostics on oligonucleotide microchips. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 4913–4918.

ZHANG, L., ZHOU, W., VELCULESCU, V.E., KERN, S.E., HRUBAN, R.H., HAMILTON, S.R., VOGELSTEIN, B. AND KINZLER, K.W. (1997). Gene expression profiles in normal and cancer cells. *Science* **276**, 1268–1272.