

CHAPTER 1

DNA Microarray Technology

All living organisms are composed of cells. As a functional unit, each cell can make copies of itself, and this process depends on a proper replication of the genetic material known as deoxyribonucleic acid (DNA). DNA contains genes, and each structural gene functions by transcribing it into the corresponding messenger RNA (mRNA) using DNA as a template and ultimately translating into the corresponding protein using mRNA as a template (Fig. 1.1). The abundance and stability of proteins determine the functions of a cell. Thus, the function or activity of a gene is reflected by synthesis of mRNA (transcription) or protein (translation). DNA microarray technology measures the activity of genes at a transcriptional level.

DNA microarrays (sometimes called DNA chips) are in general characterized by a structured immobilization of DNA targets in the free nucleic acid samples on planar solid supports, on which different types of nucleic acids with known sequences (known as “probes”) are fixed. A probe may be derived from complementary DNA (cDNA), polymerase chain reaction (PCR) products, or synthetic oligomers. In general, applications of DNA microarray technology broadly include (1) gene expression analysis (transcription analysis), which analyzes the transcriptional activity of genes through hybridization between DNA targets and probes; (2) genotyping with oligonucleotide arrays, which is based on the notion of combining the complete sequence of a DNA sample by presenting all possible sequences as a complement on the chip (Drmanac *et al.*, 2002); (3) measurement of enzyme activities on immobilized DNA, which is based on the finding that DNA-modifying enzymes are capable of acting on immobilized DNA templates or oligonucleotides (Bier *et al.*, 1996a; Bier *et al.*, 1996b; Buckle *et al.*, 1996); (4) PCR on the chip, which was

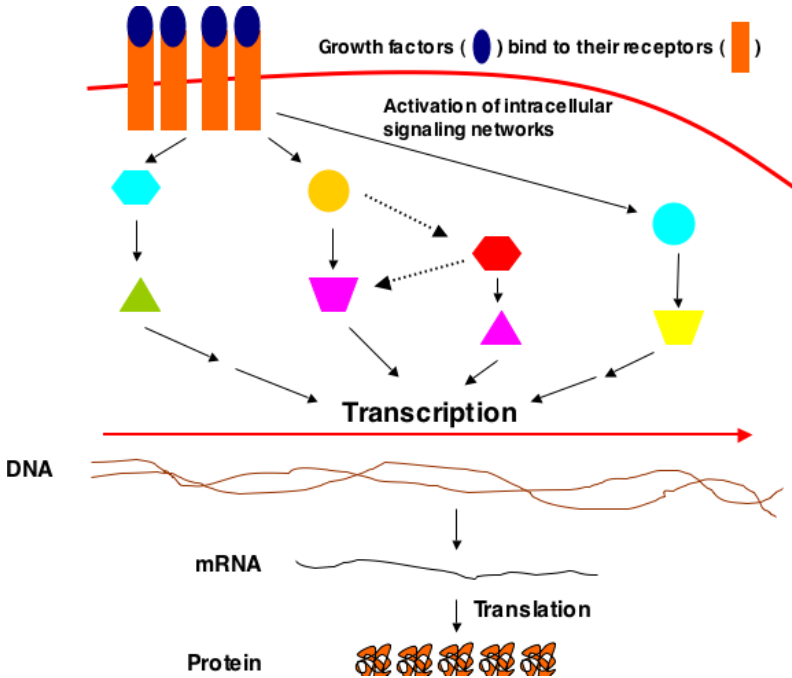


Fig. 1.1. Activation of signaling pathways at different levels in cells that are stimulated by growth factors. Binding of growth factors to the cell surface receptor leads to activation of intracellular downstream known (solid lines) and unknown (broken lines) signaling molecules (colored blocks). The end result of this activation is to turn on the transcription of genes and ultimately the translation of mRNA into proteins that promote or inhibit cell growth. Measurement of synthesis of either mRNA or protein will reflect the function of a gene.

first described in 2000 (Adessi *et al.*, 2000); and (5) transcription on chip, which shows the transcription of a complete gene into mRNA on the chip (Steffen *et al.*, 2005). In this book, we focus on gene expression analysis using DNA microarray technology.

1.1. Experimental Procedure

The procedure of a DNA microarray experiment includes multiple steps from sample preparation to data analysis (Fig. 1.2), among which

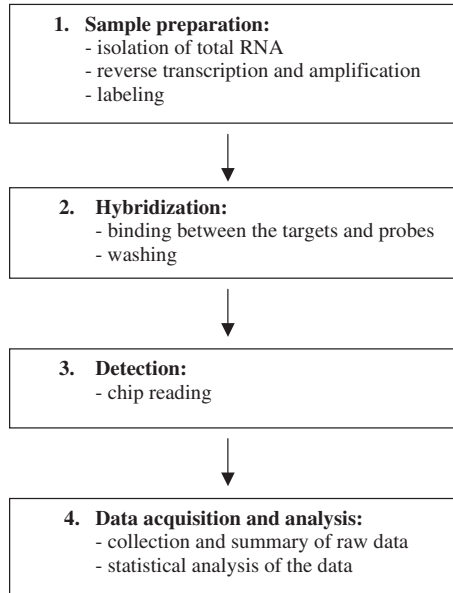


Fig. 1.2. Workflow of a DNA microarray experiment.

hybridization is a central process. The sample that contains the targets to be investigated is added to the DNA chip to allow their binding to the probes fixed on the chip, resulting in a characteristic-binding pattern representing the levels of gene expression of the sample. The sample itself is labeled prior to the hybridization, and the most-often-used labels are fluorescence labels that allow detection of the binding event. After the hybridization, the chip is washed and then fluorescence intensities on the chip are read and recorded by a scanning or imaging device. Raw data reflecting the fluorescence intensities are statistically analyzed and often shown by fold changes as compared to control.

In our laboratory, we performed DNA microarray experiments to study gene expression profiling in mouse leukemia cells. Here is an example of the procedures for carrying out the microarray experiments. Briefly, cells are dissolved in RNAlater (Ambion, Austin, TX, USA) and homogenized in RLT Buffer (RNeasy Micro Kit; Qiagen, Valencia, CA, USA). Total RNA is isolated by following the protocol for the RNeasy Micro Kit, and quality is assessed using a 2100 Bioanalyzer instrument and RNA 6000 Pico

LabChip assay (Agilent Technologies, Palo Alto, CA, USA). Utilizing the GeneChip Whole Transcript Sense Target Labeling Assay kit (Affymetrix, Santa Clara, CA, USA), 100–300 ng of total RNA undergoes reverse transcription with random hexamers tagged with T7 sequence. The double-stranded cDNA that is generated is then amplified by T7 RNA polymerase to produce cRNA. Second-cycle first-strand cDNA synthesis then takes place, incorporating dUTP, which is later used as sites where fragmentation occurs by utilizing a uracil DNA glycosylase and apurinic/apyrimidinic endonuclease 1 enzyme mix. The fragmented cDNA is then labeled by terminal transferase, attaching a biotin molecule using Affymetrix proprietary DNA Labeling Reagent. Approximately 2.0 μg of fragmented and biotin-labeled cDNA is then hybridized onto a Mouse Gene ST 1.0 Array (Affymetrix, Santa Clara, CA, USA) for 16 hours at 45°C. Posthybridization staining and washing are performed according to the manufacturer's protocols using the Fluidics Station 450 instrument (Affymetrix). Finally, the arrays are scanned with a GeneChip Scanner 3000. Images are acquired and CEL files generated, which are then used for data analysis. Figure 1.3 shows an example of the fluorescence intensities on the chip recorded by an imaging device.

1.2. Experimental Design

Based on our experience, the most critical factor in the experimental design of a DNA microarray experiment is to have a correct control for comparison, which means that we need to appropriately choose a cell or tissue source for isolating control RNA. Choosing a correct control can be extremely challenging, and sometimes it may not be possible to find a “perfect” control. For example, DNA microarray technology has been widely used in comparing gene expression profiling between tumor cells or tissues and corresponding normal cells or tissues in humans. Typically, total RNA is isolated from tumor cells/tissues that are heterogeneous in origin, and control RNA isolated from cells/tissues adjacent to the tumor cells/tissues or corresponding normal cells/tissues is also heterogeneous in most cases; cellular compositions of tumor and normal tissues are not the same or sometimes not even closely similar for appropriate comparisons. Although this cellular difference between tumor and normal tissues is a

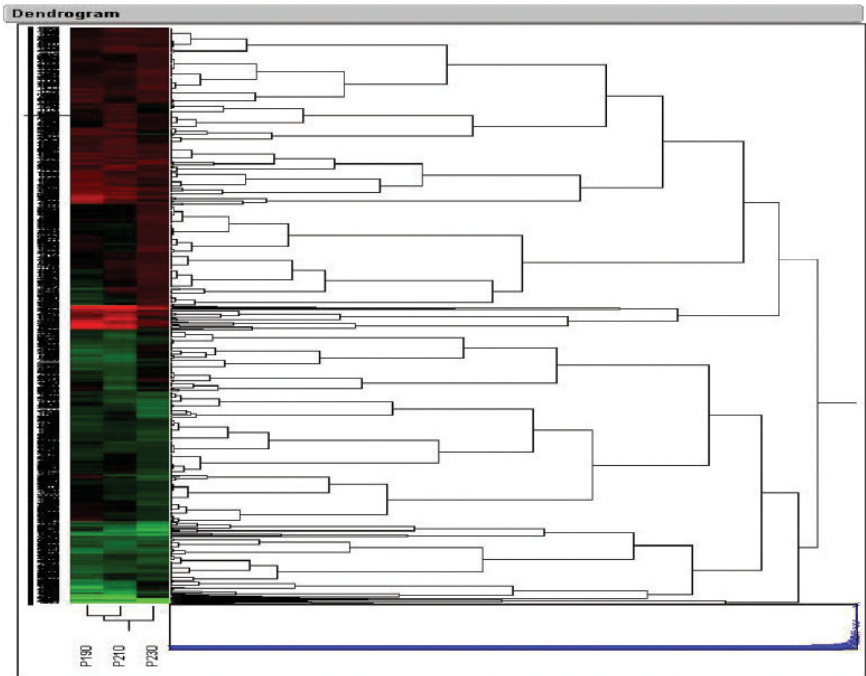


Fig. 1.3. DNA microarray analysis of gene expression profiles in pre-B leukemic cells expressing P190, P210, or P230 BCR-ABL. Total RNA was isolated from the parental ENU pre-B cell line and the same cell line expressing P190, P210, or P230. A total of 24 000 genes were measured using Affymetrix GeneChip, and the data were analyzed by F2 test. Expression of 2086 genes in BCR-ABL-expressing cells was found to be significantly different from that for the reference (parental cell control) ($\alpha < 0.05$). The scientific background of the pre-B leukemic cells expressing P190, P210, or P230 BCR-ABL will be introduced in Chapter 2.

difficult technical problem in microarray studies, the current approach may have done its best to find the most appropriate controls available. Some of the findings obtained from this type of studies are real and useful, although some genes that play critical roles in tumor formation might not be detected due to the inappropriate controls used. A better way to improve this situation is to isolate subpopulations of tumor cells with antibodies recognizing phenotypical cell surface markers. This approach can be taken much more easily when studying human blood cancer cells. However, the

availability and amount of human cells for sorting out a particular cell population can be difficult. In parallel, the different genetic background of each human patient will cause some differences in gene expression when a comparison is made among different patients with the same type of cancer. Regardless of these difficulties, DNA microarray technology has been found to be useful in the clinic, although there is still a long way to go before DNA microarray results can be used to support regulatory decision making or accurate and consistent prediction of patient outcomes.

Here is an excellent example of how to perform a DNA microarray study using human cancer patient samples. In an elegant study of gene expression profiles of human breast cancer cells, a comparison of gene expression was made between breast cancer cells with higher tumorigenic capacity and normal breast epithelium (Liu *et al.*, 2007). Specifically, the tumor cells were low or undetectable levels of CD24 (CD44⁺CD24^{-/low}), whereas the phenotype of cells from normal breast epithelium was unknown. A 186-gene invasion-associated signature obtained from this study suggests that there is a significant association between tumor invasion and both overall and metastasis-free survival in breast cancer patients. A critical question to ask is whether the normal breast epithelium is an appropriate control in this DNA microarray study. No matter what the answer is, it is clear that the normal breast epithelium is one of the best controls available.

1.3. Quality Control

A major issue in DNA microarray technology is its repeatability and reproducibility. Repeatability refers to the ability to provide closely similar results from replicate samples processed in parallel at the same test location using the same gene expression assay. Reproducibility refers to the ability to provide closely similar results from replicate samples processed with different microarray platforms or at different test locations using the same gene expression assay. To achieve high repeatability and reproducibility, quality control has become a key issue in DNA microarray studies.

A major criticism voiced about DNA microarray studies has been the lack of accuracy and reproducibility of the microarray data. The quality

of DNA microarray results is associated with technical, instrumental, computational, and interpretative factors (Casciano and Woodcock, 2006). For the same samples tested, results obtained at different locations and between different microarray platforms could be different, making it difficult to use and interpret microarray data. Optimization and standardization of microarray procedures are critical steps. In this regard, the US Food and Drug Administration (FDA) has initiated a MicroArray Quality Control (MAQC) project among researchers in academic, government, and industrial institutions to seek to experimentally address the key issues surrounding the reliability of DNA microarray data. As part of this effort, since 2004 the FDA has started accepting voluntary genomic data submission with accompanying information related to the number and scope of DNA microarray-based expression data (Anonymous, 2006). The MAQC project aims to establish quality control metrics and thresholds for an objective assessment of the performance achievable by different microarray platforms, and for evaluation of the merits and limitations of various data analysis methods (Casciano and Woodcock, 2006). The specific concerns of the MAQC project relate to the impact of microarray data quality on genomic data submission (Frueh, 2006; Ji and Davis, 2006), the framework for the use of genomics data (Dix *et al.*, 2006), comparisons of different commercial microarray platforms (Canales *et al.*, 2006; Patterson *et al.*, 2006; Shippy *et al.*, 2006), the use of external RNA controls (Tong *et al.*, 2006), interplatform and intraplatform reproducibility of gene expression measurements (Shi *et al.*, 2006), and analytical consistency across microarray platforms (Guo *et al.*, 2006).

The main conclusion of the MAQC project is that, with careful experimental design and appropriate data transformation and analysis, microarray data can indeed be reproducible and comparable between different formats and laboratories, and that fold change results from microarray experiments correlate closely with results from well-accepted assays such as quantitative reverse transcription–polymerase chain reaction (qRT-PCR) (Anonymous, 2006). However, there is still a long way to go before we can answer the key remaining question: when can microarray data be used in a regulatory decision-making process?

1.4. Interpretation of DNA Microarray Data

While expensive, DNA microarray experiments have not yielded sufficient information that allows us to draw decisive conclusions. Often, what we have in the end is a long list of more than 30 000 genes with fold changes of their mRNA expression comparing control and experimental groups. An easy explanation for the fold change of mRNA expression of a particular gene, which is actually a way of obtaining information from DNA microarray data by most people, is that mostly changed genes in expression are likely involved in the biological process being studied. A “hot” list is often provided by statisticians to researchers to describe which genes are mostly upregulated or downregulated, and those genes with minor or no changes in mRNA expression are often considered as not involved or not important by the researchers. However, there are numerous examples showing a disassociation between the abundance of mRNA and the level of translated protein for a gene of interest, and between the abundance of mRNA and the effect of the gene of interest on a particular biological process.

From our own experience, a general guess on the biological role of a particular gene based on the fold change is sometimes agreeable to the results of experiments testing the function of this gene, but it is not rare to see misleading or wrong conclusions drawn based solely on the magnitude of the fold changes. Importantly, we learned that upregulation or downregulation of a gene does not necessarily mean a positive or negative role of this gene in the biological process being studied. A novel idea we intend to propose here is that a gene with no change in its expression, as determined by DNA microarray analysis, may still play a significant role in the biological process. In Chapter 2, we describe in much more detail and emphasize these ideas about the interpretation of DNA microarray data, with specific examples from our own studies.

1.5. Advantages and Disadvantages

One of the biggest advantages of DNA microarray technology is that it can evaluate simultaneously the relative expression of thousands of genes by using small amounts of materials, providing gene signatures for particular

disease situations. In addition, the procedures can easily be automated. Furthermore, the capacity of measurement of gene expression by DNA microarray is huge, allowing researchers to take the expression of all genes from an individual into consideration for disease analysis in so-called “personalized medicine”.

One of the major disadvantages of DNA microarray technology is that it only evaluates gene expression at a transcriptional, but not translational, level, as posttranscriptional modifications (such as phosphorylation) often play significant roles in the regulation of protein functions. In addition, DNA microarray technology is still not mature enough for decision making based on the microarray data. In this book, we introduce our new way of interpreting and analyzing microarray data, which will hopefully bring us closer to success in decision making using the information obtained through DNA microarray technology (see Chapter 4).