

# DNA Microarray Experiments: Biological and Technological Aspects

Danh V. Nguyen<sup>1\*</sup>, A. Bulak Arpat<sup>2</sup>, Naisyin Wang<sup>1</sup>, and Raymond J. Carroll<sup>1</sup>

<sup>1</sup>Department of Statistics, Texas A&M University,  
College Station, TX 77843-3143, U.S.A.

<sup>2</sup>Graduate Group in Genetics, University of California,  
Davis, CA 95616, U.S.A.

\**email*: dnguyen@stat.tamu.edu

## SUMMARY

DNA microarray technologies, such as cDNA and oligonucleotide microarray, promise to revolutionize biological research and further our understanding of biological processes. Due to the complex nature and sheer amount of data produced from microarray experiments, biologists have sought the collaboration of experts in the analytical sciences, including statisticians among others. However, the biological and technical intricacies of microarray experiments are not easily accessible to analytical experts. One aim of this review is to provide a bridge to some of the relevant biological and technical aspects involved in microarray experiments. While there is already a large literature on the broad applications of the technology, basic research on the technology itself and studies to understand process variation remain in their infancy. We emphasize the importance of basic research in DNA array technologies to improve the reliability of future experiments.

KEY WORDS: Affymetrix; cDNA; Design of experiments; Gene expression; Image processing; Microarray; Molecular biology; Normalization; Nucleotide labeling; Oligonucleotide; Reverse transcription; Transcription; Variability.

# 1 Introduction

DNA Microarray technologies, such as cDNA array and oligonucleotide array, provide a means of measuring the expression of thousands of genes simultaneously. These technologies have attracted much excitement in the biological and statistical community, and promise to revolutionize biological research and further our understanding of biological processes. (We use the term microarray and array interchangeably.)

The purpose of this article is to provide a review of the literature on DNA microarrays and a tutorial to DNA microarray technology. In this process we hope to (1) provide a bridge to the biological and technical aspects involved in microarray experiments and (2) emphasize the need for basic research in DNA array technologies in order to increase their reliability, and hence the reproducibility of research results.

We begin in Section 2 with the biological principles behind microarray experiments to show *what* is being measured and *why*, and also to review some concepts from molecular biology relevant for a practical understanding of the technologies. Next, we proceed to describe the microarray experimental procedure in Section 3 for cDNA microarrays (Schena et al., 1995), one type of microarray technology. Emphasis is placed on the biological sample (cDNA) labeling process because it is directly related to *what* is being measured. After the actual microarray experiment comes the data collection phase. Data collection for microarray experiments is not a trivial task and requires imaging technology and image processing tools. The data collection procedure for cDNA microarray is described in Section 4. High density oligonucleotide array (e.g., Affymetrix arrays) is another type of microarray technology in wide use: we describe this technology in Section 5. We hope that the material presented here will also make other emerging array technologies accessible to statisticians.

There is a growing literature on the statistical analysis of microarray data. It is not our intention to describe this literature, since our focus is instead on basic background for microarray experiments. Nonetheless, there are some fundamental analytical issues which have a major impact on the final

data amenable for statistical analysis and which we believe deserve more attention, namely variation, normalization and design of experiments. These issues are discussed briefly in Section 6.

## 2 The Biological Principles Behind Microarray Experiment

We begin by providing an elementary answer to the question, “What is a DNA microarray measuring?” The answer is *gene expression*, a term we will expand upon. The reader may wish to refer to Table 1, which contains a small glossary of the major terms used in what follows. Also, supplemental figures illustrating the basic molecular biology concepts described in this Section can be found at <http://stat.tamu.edu/~carroll/techreports.html>.

It is useful to view the primary biological processes as *information transfer* processes. The information necessary for the functioning of cells is encoded in molecular units called *genes*. Messages are formed from genes and the messages contain instructions for the creation (synthesis) of functional structures called *proteins*, necessary for cell life processes.

The information transfer processes are crucial processes mediating the characteristics features or phenotypes of the cells (e.g. cancer and normal cells) . The kind and amount of protein present in the cell depends on the genotype of the cell. Therefore, together with environmental factors, genes “determine” the phenotypes of cells and hence the organism. This simplified model with intermediate products, starting from genes to phenotype, is illustrated below. The model shows how genes (DNAs) are *linked* to organism phenotype and illustrate the reason for measuring mRNA (transcript abundance), the direct product of DNA transcription.

$$\text{DNA} \Rightarrow \text{mRNA} \Rightarrow \text{amino acid} \rightarrow \text{protein} \rightarrow \text{cell phenotype} \rightarrow \text{organism phenotype}$$

Note that there are different levels of gene expression, one at the *transcription level* where RNA is made from DNA and one at the *protein level* where protein is made from mRNA. The relationship between protein and mRNA is not one-to-one, hence the simplified model discussed above is only

an idealization. Gene expression as measured by microarray is at the transcription level, although protein arrays have also been developed (Haab, Dunham, and Brown, 2001).

There are methods for detecting mRNA expression of a single gene or a few genes (e.g. the Southern blot). The novelty of a microarray is that it quantifies transcript levels (mRNA expression levels) on a global scale by quantifying transcript abundance of thousands of genes simultaneously. This novelty has allowed biologists to take a “global perspective on life processes—to study the role of all genes or all proteins at once” (Lander and Weinberg, 2000).

There are three primary information transfer processes in functioning organisms: (1) replication, (2) transcription and (3) translation. We give a brief review of *transcription* in the next section, because it is directly relevant to DNA microarray technologies. For details on the mechanism of DNA transcription the reader is referred to Alberts et al. (1994) and Griffiths et al. (2000).

## 2.1 DNA, Genes, and DNA Transcription

We now know that with few exceptions genes are composed of deoxyribonucleic acid (DNA). DNA consists of four primary types of nucleotide molecules. The common structure of a nucleotide contains a phosphate, a (deoxyribose) sugar, and a nitrogen base. The four types of nucleotides are distinguished from one another by their distinct nitrogen base: adenine (A), guanine (G), cytosine (C) and thymine (T).

DNA exists as a double helix (Watson and Crick, 1953) where each helix is a chain of nucleotides. The two chains or helices are held together primarily by hydrogen bonds. In DNA, base A pairs with T and base G pairs with C exclusively. The specific pairing of DNA bases (A-T, G-C) is called base-sequence complementarity. DNA exists in its native state as a double helical structure. However, with sufficient heating the hydrogen bonds between complementary base pairs break and the DNA double strands separate (denature) into two single strands.

DNA transcription is the information transfer process directly relevant to DNA microarray

experiments because quantification of the type and amount of this copied information is the goal of the microarray experiment. The process of transcription begins with DNA in the nucleus where the DNA template strand is copied. The copied strand is called messenger ribonucleic acid (mRNA) since it carries the set of instructions contained in DNA. More precisely, the RNA goes through some important preprocessing (which we described below) before it becomes mRNA. We refer to the RNA strand prior to preprocessing as pre-mRNA, and after processing as mRNA.

Both DNA and RNA are nucleic acids, however RNA is single stranded, the sugar in its nucleotide is ribose rather than deoxyribose found in DNA, and the pyrimidine base U (uracil) is found in place of T (thymine). Also, U forms hydrogen bonds with A (adenine). In transcription a section of one strand of DNA corresponding to the gene is copied using the base complementarity, namely A-U and G-C. The DNA strand to be copied is called the template strand. The other complementary non-transcribed DNA strand is appropriately called the non-template strand.

Transcription can be subdivided into three stages involving RNA chain (1) initiation, (2) elongation and (3) termination. Regions of DNA that signal for the initiation of transcription are called promoter regions. Promoter regions contain specific DNA sequences which enable/enhance the recruitment of an enzyme (protein) called RNA polymerase II to the transcription initiation site. The RNA polymerase then moves along the DNA and extends the RNA chain by adding free ribonucleotides; adding nucleotides with base A, G, C, or U where a T, C, G, or A is found in the DNA template strand respectively. The enzyme (RNA polymerase) recognizes signals in the DNA sequence for chain termination, resulting in the release of the newly synthesized RNA and enzyme from the DNA template.

As mentioned earlier, the transcription of DNA is carried out in the nucleus. Before the message is transported to the cytoplasm for translation, some important preprocessing (post-transcriptional processing) of the message occurs. For example, an enzyme called polyadenylase adds a sequence of A's to the RNA strand at the 3' end. This sequence of A's is called the poly(A) tail. The poly(A)

tail play an key role in the preparation of mRNA for hybridization in microarray experiments in a process called reverse transcription (Section 3.2). Another important processing of pre-mRNA is called RNA splicing. The DNA segments that encode for a protein, called exons, are interspersed with non-coding segments called introns. RNA splicing is a series of splicing reactions which remove the introns regions and fuses the remaining exon regions together. Thus, the resulting mRNA contains only the coding sequences (exons) and can be identified by the poly(A) tail.

### 3 The cDNA Microarray Experimental Procedure

For exposition we divide the microarray experimental procedure into four primary stages.

1. *Array fabrication* involves preparing the glass slide, obtaining the DNA sequences and depositing (“printing”) the cDNA sequences onto a glass slide.
2. *Sample preparation* consists of processing and preparing the biological samples of interest. This involves isolating total RNA (mRNA and other RNAs) from biological samples such as tissue samples. Due to space limitations, we will not further discuss this particular stage, but note here that much variability may come from sample preparation.
3. *cDNA synthesis and labeling* involves making and labeling *complementary* DNAs (cDNAs) from the experimental and the reference biological samples with a different fluorescent dye.
4. *Hybridization* involves applying the experimental and reference cDNA mixture solution to the array. See Table 1 for a summary of commonly used terminology.

Overviews of cDNA microarray can be found in Duggan et al. (1999) and Kahn et al. (1999).

#### 3.1 Array Fabrication

We give a brief overview of the process of making the arrays to illustrate that many sources of variation can come from this process. Making the arrays first requires selecting the cDNA sequences,

called *probes*, to print on the arrays. A set of potentially relevant genes or of potential relevance to the biological question under investigation are obtained. How does one know which DNA sequences (representing genes) to use and where does one get these sequences? This requires the use of cDNA clones and a cDNA library which we now briefly describe. A DNA clone is a section of DNA that has been inserted, for example, into a plasmid (a vector molecule) and then replicated to form many copies. A *cDNA library*, is a collection of cDNA clones obtained from mRNA (see Section 3.2 for details). Next, the sequence information of the cDNA clones of interest must be determined. There are publicly available databases, such as GenBank, which can be used to annotate the sequence information of custom cDNA libraries or to identify cDNA clones from previously prepared libraries. UniGene is another database, a subset of GenBank, containing *unique* clones (clusters).

In order to have a sufficient amount of each cDNA clone to print on the array, each clone is amplified to get many copies using a technique called polymerase chain reaction (PCR). After amplification, the PCR-product (i.e., liquid containing the amplified cDNA probes) is then deposited on the array using a set of microspotting pins. Each pin contains an uptake channel which can be filled with liquid sample (PCR-product). The set of pins dip into the PCR-product and liquid fills the uptake channel of each pin. A small amount of the PCR-product in each pin is then deposited onto the array. Ideally, the amount deposited should be uniform, but this is not the case in practice. The drops of solution containing cDNA probes form the spots on the array.

Prior to printing, the glass array is treated to enhance binding of the cDNA to the glass surface. The liquid spots are air dried and the cDNAs remain mostly in the spot. However, as we will describe later, the array goes through a series of washes so it is important that the immobilized cDNAs at each spot do not get washed away. Thus, a procedure, called *UV-crosslinking*, is often applied to the spotted array to increase fixation of the cDNA probes to the glass. cDNA probes in solution deposited on the array are double strands. In order for a complementary DNA strand, obtained from a biological sample, to bind to a cDNA probe on the array the double-stranded probe

must be separated. Thus, the array is heated to denature the double-stranded cDNA probe and to enhance binding to the array surface.

Fabrication of cDNA arrays requires a microarray facility for printing. For diagrams of the microarray facility see Kahn et al. (1999). Some guides for setting up a microarray facility are available (DeRisi, Iyer, and Brown, 1999).

### **3.2 Experimental and Reference Target cDNAs Labeling**

An “ideal” microarray to monitor global gene expression is one where probes for *all* genes of a genome of interest are spotted or synthesized onto the array. However, currently all genes of the human genomes are not known. Thus, probes used correspond to known genes, partially sequenced cDNAs (or expressed sequence tags; ESTs), and/or randomly chosen cDNAs from libraries of interest. Efforts are made to minimize the redundancy of probes. Although not completely accurate, we refer to the probes as “genes,” a terminology that is common in the literature.

Given a biological sample of cells there are two possibilities: a set of genes are either expressed in the cells or they are not. If these genes are expressed in the biological sample it is then of interest to quantify accurately the amount of expression under a given cell type or condition. cDNA arrays are designed to measure the expression of cells in an experimental sample relative to a reference (control) sample. In the current practice, cDNAs (representing the mRNA transcripts) from the experimental and reference samples are labeled with different fluorescent dyes, mixed, and hybridized to probes on the array. The measured fluorescent intensity for each should be proportional to transcript abundance, conditional on factors such as spot characteristics, hybridization efficiency, and level of dye incorporation. Thus, the method used to identify or to label each individual transcript is important. Insuring equal labeling per molecule can increase sensitivity and quality of data for downstream analyses. We describe three fluorescence labeling methods currently in use, along with their advantages and disadvantages.

The transcript labeling procedure involves a process called reverse transcription (RT) where cDNA is made from mRNA. The RT process is depicted in Figure 1. We first describe RT in general and then proceed to describe how it is used in microarray experiments.

There are two primary steps involved in RT: (1) recognition of the mRNAs in total RNA; and (2) actual synthesis of the cDNAs from the mRNAs. mRNAs are recognized by their poly(A) tails. Oligo(dT) primers (TTTT...), which recognize mRNA transcripts by binding to their poly(A) tails, are added to solution in RT. An enzyme called reverse transcriptase is added to catalyze the first cDNA strand, complementary to the mRNA. The cDNA strand is synthesized one nucleotide at a time, starting at the primer. The necessary molecular building blocks for synthesizing the cDNAs are denoted by dATP, dGTP, dCTP, dTTP, dUTP or just dNTP (deoxyribo nucleoside triphosphate) if not referring to a specific type. Once a dNTP unit is incorporated into the cDNA strand, it becomes a nucleotide. For example, if a dATP is incorporated then the resulting cDNA strand contains a nucleotide with base A. Sufficient dNTPs are added to solution so that they are available for making cDNA during the RT reaction. Often, in microarray experiments, only the first strand cDNA is made and labeled with a fluorescent dye (Figure 1), although the second strand cDNA can be made by adding the enzyme DNA polymerase. However, labeling targets for Affymetrix arrays (Section 5) requires the second strand cDNA synthesis (Section 5.3).

In the cDNA microarray system, expressions of genes from the experimental cells of interest are measured *relative* to the expressions of the same genes in a fixed reference or control cell type. Thus, for cDNA microarrays, two pools of cDNAs are synthesized, one from mRNAs of the experimental cells and another pool of cDNAs are made from mRNAs of the reference or a control cells. cDNAs obtained from experimental and reference mRNAs are often labeled with a red or green fluorescent dye called Cy5 or Cy3 respectively. We first describe the direct labeling method where the Cy5- and Cy3-labeled nucleotides are incorporated directly into the cDNAs during RT. The other two methods, amino-modified nucleotide and primer tagging, are variations of this method.

### 3.2.1 The Direct Incorporation Labeling Method

The main steps involved in direct labeling of cDNAs (Schena et al., 1995) are depicted in Figure 2. Direct labeling of the nucleotides during RT was the first and still most widely used method.

First, total RNA is isolated from experimental and reference cells, in separate solutions. Next, complementary DNAs (cDNAs) are synthesized from the total RNA using RT, as described in the previous section. We note that during RT, both dUTPs and dTTPs can be used to make the synthetic cDNAs because A can bind with either T or U. That is, if an A base is found in the mRNA template then the corresponding synthesized cDNA base can be T or U, depending on whether a dTTP or dUTP was incorporated in the reaction respectively.

As described in the previous section, the cDNA microarray system involves both experimental and reference cells. Thus, two separate runs of RT are carried out, one for the experimental RNAs and one for the reference RNAs. To distinguish between the two pools of synthesized cDNAs, often Cy5- and Cy3-labeled dUTPs are used for the experimental and reference cDNAs respectively. The experimental and reference cDNAs, labeled with different dyes, are then mixed and hybridized onto the array, a process which we will describe in Section 3.3.

There are some problems associated with direct incorporation. First, the number of labeled U nucleotides present in any cDNA depends on the base composition, the number of A's in the mRNA strand. Second, the number of labeled nucleotides also depends on the length of the transcript. A very long transcript or/and one with many A's resulting in cDNAs with many labeled nucleotides will have a stronger detected fluorescence intensity signal. Thus, in this case, the stronger detected signal does not mean that the gene is expressed more (i.e. many transcripts) because the stronger signal results from a few long transcripts with many labeled nucleotides.

In addition to the fact that the strength of the fluorescence signal will depend the numbers of Cy5 or Cy3-labeled dUTPs in direct labeling, it has been noted that the Cy-dye labeled nucleotides were not efficiently incorporated during the RT reaction, resulting in varying measured fluorescence

intensities depending on the types of dye used.

### 3.2.2 Amino-Modified (Amino-allyl) Nucleotide Method

To address the problem of inefficient or non-uniform incorporation of Cy-dye nucleotides during RT reaction, the amino-modified (amino-allyl) nucleotide method has been proposed. This method is depicted in Figure 3. The first step involves adding *unlabeled* dCTPs, dGTPs, dTTPs and modified dUTPs in *both* experimental and reference total RNA solution for the synthesis of cDNAs. Once the experimental and reference cDNAs have been synthesized (in separate RTs), the second step is to add the Cy5 or Cy3 dyes to the experimental or reference cDNA pools. The dye, added in the second step, couples with the amino-modified dUTP nucleotides, resulting in Cy5- or Cy3-labeled cDNAs. The key point here is that the experimental and reference cDNAs are synthesized using exactly the *same unlabeled* dNTPs, in *both* RT runs. The FairPlay<sup>TM</sup> Microarray Labeling Kit of Stratagene is an example of this method. See Wong et al. (2001) for more details and May et al. (2001) for an application. Note that the problem of the fluorescent intensity depending on the base composition and cDNA length, encountered with direct labeling, still remains.

### 3.2.3 The Primer Tagging Method—DNA Dendrimer Labeling

A third cDNA labeling method called DNA dendrimer labeling is based on “primer tagging,” see Figure 4. The problem of fluorescent intensity depending on the base composition and cDNA length is resolved by not labeling the nucleotides at all.

In this method *unlabeled* and *unmodified* dNTPs are added to experimental or reference cDNA separately and oligo(dT) primers are added to the RNA solution as before. However, the oligo(dT) primers now contain a capture sequence, say TTTTTT — — — — for experimental RNA and TTTTTT + + + + for reference RNA. The two different capture sequences, attached to the oligo(dT) primers above, are denoted by — — — — and + + + +. Next, the synthesized experimental and reference cDNAs are mixed together and hybridized to an array. The array, after hybridization and washing,

is then incubated with Cy5 and Cy3 labeled molecules called dendrimers. Each dendrimer consists of approximately 250 fluorescent Cy5 or Cy3 molecules and contains sequences complementary to the corresponding capture sequences. Thus, there is approximately one intensity signal per cDNA molecule. Note that nucleotides of the synthesized cDNAs are not labeled but rather the primers are tagged and then labeled at the incubation stage.

Primer labeling using DNA dendrimer resolves some of the problems associated with the direct labeling and amino modified nucleotide method. That is, the dye is not incorporated during cDNA synthesis and preparation. Hybridization inefficiencies of cDNAs onto the array are also avoided. We note here that the strong fluorescence signal from dendrimers is one claimed advantage of the method, therefore, the washing step to eliminate unbound cDNA with the capture sequence is important. Also, the rate of “incorrect captures” must be low; that is, the Cy5 dendrimers do not bind to the Cy3 capture sequence and vice versa. For details and experimental protocols, the reader is referred to Stears, Getts, and Gullans (2000) and Genisphere<sup>®</sup> protocol at [www.genishere.com](http://www.genishere.com).

Although the modified nucleotide and the primer tagging approach to cDNA labeling conceptually resolves some of the problems found with direct labeling, both methods could introduce other biases. Some very small experiments have been done, Stears et al. (2000), showing the “advantages” of dendrimer technique, but results are far from a full evaluation of the method.

### **3.3 Hybridization of Experimental and Reference cDNA Targets onto Array containing cDNA Probes**

Hybridization refers to the binding of two complementary DNA strands by base pairing. Suppose that the *mixed* solution of experimental and reference target cDNAs have been applied to the array, which contains the probe cDNAs in each spot. Consider a specific spot on the array. This spot contains cDNA probes for a gene of interest, say gene A. If there are target cDNAs, in the mixed solution, complementary to the probe cDNAs of gene A, they should bind together by base pairing. Suppose that gene A is expressed in both the experimental and the reference cells. In principle,

there will be bound experimental and reference target cDNAs of gene A at the spot and the amount of gene A's expression will show up as Cy5- and Cy3 intensities respectively.

After sufficient time is allowed for hybridization the array then goes through a series of washes to eliminate all unbound target cDNAs and solution. The post-hybridization washing procedure must be designed to be stringent enough to wash off "all" extraneous materials and at the same time not too stringent to wash off bound cDNAs, the signals of interest. Washing breaks unstable binding of target and probe cDNAs, which may be a result of cross-hybridizations (binding of non-complementary strands that have sequence similarity to the probe). Designed experiments to determine specificity, i.e. estimate the cross-hybridization rates, for cDNA microarray would be required to understand this issue in more detail.

## **4 Data Collection: Microarray Quantification**

Next, the expression level in the experimental and reference cells of each gene needs to be quantified. The expression levels of a gene in the experimental or reference cells are measured by the spot intensities of the Cy5 dye or Cy3 dye respectively. Dye or fluorescent intensities are obtained by scanning the array using a confocal laser microscope, see Figure 5. The array is scanned at two wavelengths or "channels," one for the Cy5 fluorescent tagged sample and another for the Cy3 tagged sample. For details on confocal scanning and image processing in microarrays the reader is referred to Schermer (1999), Chen, Dougherty, and Bittner (1997), Yang et al. (2002), and the Amersham Pharmacia Biotech technical manual (<http://www.mdyn.com>).

### **4.1 The Observed Data: Images from Confocal Scanning Microscopy**

The product resulting from the array scanning process are two 16-bit tagged image file format (TIFF) images. The scanned array area is divided into equally sized pixels and the resulting image contains fluorescence intensities for corresponding pixels. The primary processes involved in

fluorescence imaging are fluorescence light excitation and detection. These processes and equipment used all affect the resulting image.

#### **4.1.1 Excitation and Emission Wavelengths on Image Quality**

Fluorescence is the emission of light when certain fluorescent dye molecules called fluorophores, such as Cy3 and Cy5, absorb excitation light. For any given dye, the range of excitation wavelengths must be determined for efficient excitation of the dye. Other types of dyes, such as Alexa Fluors, are beginning to be used in microarray experiments. If the excitation wavelength used is too close to the emission peak then the fluorescence signal will be polluted, resulting in lower emitted fluorescence intensities. Excessive excitation light can damage the sample through photobleaching, also resulting in low emission fluorescence intensity.

#### **4.1.2 Confocal Scanning Arrangement and Spot Image**

The term “confocal” in confocal scanning refers to the two focal points of the arrangement, one at the objective lens and the other at the confocal pinhole (Figure 5). The arrangement is designed to limit the detection field in three dimension. This local precision is an advantage, but it is also a disadvantage since a slight deviation from the plane of focus will result in a distorted signal.

#### **4.1.3 Signal Discrimination, Detection and Resulting Image**

In microarray, light excitation is generated by lasers and the excitation light is focused on a small spot on the array, which illuminates a small area of the array. The dye in the focused area absorbs the laser excitation light and emit fluorescence light. Noise fluorescence can come from many sources, including the glass slide itself, chemical treatment of the slide, chemicals from slide washing, dust, streaks, among others. Specifically, light below the focus under the glass slide, which may be caused by a piece of dust, is gathered by the objective lens as well. These fluorescence sources must be separated and ideally only the fluorescence emitted from the dye should be detected.

It should be kept in mind that the fluorescence levels from the dye that must be detected in microarray are very low (picowatt range). At this very low level, materials such as the glass slide itself, chemicals used to treat the slide, residual hybridization solution, and washing chemicals. will fluoresce. In fact, the detected intensities for a blank spot and a spot with liquid but no DNA product in it can be different (data not shown).

## **4.2 The Transformed Data: Image Processing to Extract Information**

For cDNA arrays the raw data consist of two (gray scale) images, one image obtained from Cy5-channel and the other from the Cy3 channel. Next, image processing analysis is required to extract the numerical data. This process involves estimating the location of the spot on the array and then measuring the spot intensity as well as the background intensity based on the area outside the spot.

### **4.2.1 Segmentation: Determining Spot and Background Area**

Ideally, the spots of DNA deposited on the arrays should be of uniform (circular) shape and size with the same amount of DNA deposited on each spot. However, the spots on each array are not uniform, because of variation from processes such as printing and fixation. Thus, there are natural variation in the alignment of the spots, and hence the location of each spot must be estimated. This is often done by superimposing a grid on the image and each resulting square formed by the grid lines defines the area containing a spot, which we denote as the *target area*.

Next, a segmentation technique is used to categorize each pixel in the target area as foreground corresponding to the spot area or as background. The result of segmentation is a spot mask which defines the spot within the target area. Figure 6 illustrates the basic components involved in segmenting the target area of a microarray image. Once the spot mask has been defined inside the target area, a background region (of pixels) is defined for the calculation of local background intensity. For example, as displayed in Figure 6, various regions outside the area defined by the spot mask can be used for background calculation.

## 4.2.2 Measuring Spot and Background Intensities

For a given target area, pixels in the spot area and pixels in the background region are used to compute the spot and background intensity respectively. The mean or median of intensities corresponding to all pixels in the spot area are often used as a measure of spot intensity. The intensity for a given spot is the (1) intensity from fluorescent labeled mRNAs and (2) intensity from all other sources (“background noises”). A background intensity corresponding to a spot is the summary (e.g. mean or median) of the intensities corresponding to pixels in the background region. For example, background calculation can be computed from all pixels in the square excluding spot area, between concentric circle, or in the four diamonds as depicted in Figure 6. These methods represent only some of the methods that have been employed and there are many open questions in this area.

The data that makes it to the analysis is often in the following processed form. Let  $\mathbf{I}_R = (m_{ij}^R)$  and  $\mathbf{B}^R = (b_{ij}^R)$  be  $n \times p$  matrices containing the spot and background intensities of genes  $j = 1, \dots, p$  in samples (arrays)  $i = 1, \dots, n$  from the Cy5-channel (red) image. Similarly,  $\mathbf{I}_G = (m_{ij}^G)$  and  $\mathbf{B}^G = (b_{ij}^G)$  are the corresponding matrices obtained from the Cy3-channel (green) image. Many analyses are based on the background corrected intensities:  $\mathbf{R} = (r_{ij}) = (m_{ij}^R - b_{ij}^R)$  and  $\mathbf{G} = (g_{ij}) = (m_{ij}^G - b_{ij}^G)$  directly or based on the intensity ratios  $\mathbf{X} = (x_{ij})$  where  $x_{ij} = r_{ij}/g_{ij}$ . For example, if the experimental mRNAs are from tumor cells and the reference mRNAs are from non-tumor cells then the ratio  $x_{ij}$  represents the abundance of mRNA transcripts (expression) of gene  $j$  in the tumor cells relative to non-tumor cells observed from array  $i$ . Note that information is lost when raw intensity data values are reduced to ratios. Some analyses, based on raw intensity data, indicate that there is an intensity-dependent trend to ratios (Newton et al., 2001). Such trends would be lost if the starting point of the analysis is the matrix of ratios. Research in utilizing the distribution of spot pixel intensities may yield interesting and useful information.

We have only illustrated a few ways to measure and account for background. Many questions

in this area remain open. There are issues related to image processing and data collection, such as segmentation, signal and background detection, noise patterns on the images, and signal and background computation, all of which statisticians can become familiar with and involved in. Finally, we note that there is another meaning for the term “background,” which refers to the statistical distribution of intensity measurements for genes that are not actually expressed. Thus, the second use of the term background is in the context of measurement error models. Error models for microarray are discussed briefly in Section 6.1.

## 5 Oligonucleotide Array: The Affymetrix Array

Microarray experiments using cDNA array technology are widely used. Also widely used is another competing microarray technology which produces high density oligonucleotide (oligo) arrays using a light-directed chemical synthesis process, illustrated in Figure 7. While there are other oligo arrays, the best known is due to Affymetrix (Lockhart et al., 1996). Unlike cDNA arrays, oligo arrays use a single-dye. We give a general description of the Affymetrix technology here and defer the discussion of single-dye sample labeling to Section 5.3.

Given the sequence information of a gene, the sequence can be synthesized directly onto the array, thus bypassing the need for physical intermediates, such as PCR products, required for making cDNA arrays. Figure 7B-C depicts the light-directed oligonucleotide synthesis process.

Each gene is represented by 14 to 20 features (Lipshutz et al., 1999). Many Affymetrix arrays use 20 features, a convention we now adopt. Each feature is a short sequence of nucleotides, called an oligonucleotide, and it is a perfect match (PM) to a segment of the gene. Paired with the 20 PM oligonucleotides to the gene sequence are 20 other oligonucleotides having the same sequence corresponding to the 20 PMs except for a single mismatch (MM) at the central base of the nucleotide. Figure 7A depicts the sequence of a gene and the corresponding 20 (PM, MM) feature pairs. The 20 PM features serve as unique sequence detectors and the corresponding 20 MM features

as “controls.” Under relatively ideal situations, when the gene is expressed in the cell sample, high intensity is expected for the PM feature and low intensity for the MM feature. It is assumed that differences observed between the PM and MM feature intensities are due to hybridization kinetics of the different feature sequences and nonspecific background RNA hybridizations. We note that the literature for Affymetrix arrays often uses the term “probes” pairs in place of feature pairs. We avoid using this terminology to distinguish it from the usage in cDNA arrays, although this should be clear given the context.

As depicted in Figure 7B, a series of masks define the chip exposure sites. Light is used to deprotect the selected sites defined by the mask, followed by a chemical coupling step resulting in the synthesis of one nucleotide. This masking, light deprotection, and coupling process is then repeated to synthesize the next nucleotide, until the oligonucleotide chain is of the specified length.

Intensity measurements are recorded for each of the 20 feature pairs for a gene on a given array. More precisely, feature intensity values for gene  $j = 1, \dots, p$  on array  $i = 1, \dots, n$  consist of the following 20 (PM, MM) pairs:  $(PM_{ijk}, MM_{ijk})$  for  $k = 1, \dots, 20$ . For example, in a study of transcriptional response to ionizing radiation, human lymphoblastoid cell lines were grown in unirradiated state or irradiated state (Tusher, Tibshirani, and Chu, 2001; Efron et al., 2001). The RNA samples obtained were then labeled and hybridized to Affymetrix arrays to monitor mRNA expression of about  $p = 6,800$  genes in response to radiation. An example of the 20 (PM, MM) pairs for a gene ( $j = 2715$ ), reproduced in Table 2, was given in Efron et al. (2001). Note that for experiments without array replication, the *array* is the *sample*. However, with replicate arrays, a fourth subscript is needed, so that the feature data pairs would be  $(PM_{ijk r}, MM_{ijk r})$  where subscript  $i$  denotes *samples* and  $r$  denotes the *array replicates*.

Given the 20 (PM, MM) feature pairs for a gene how should one quantify or summarize the gene’s expression level? This is a fast evolving area of fundamental research and we review its recent development in the next Section.

## 5.1 Quantifying Gene Expression Levels

To quantify expression level (index) of a gene in a particular sample, Affymetrix originally proposed the average difference  $x = \text{avg}\{d_k = (PM_k - MM_k), k = 1, 2, \dots, 20 = K\} \equiv \text{AvDiff}$ , where the subscripts have been suppressed. Actually, the above average is usually based only on differences,  $d_k$ , within 3 standard deviations from the mean of  $d_{(2)}, \dots, d_{(K-1)}$ , where  $d_{(k)}$  is the  $k$ th smallest difference. For Affymetrix arrays, most reported results are based on analyses that use the average differences of the PM and MM, but with various ways to filter the outliers (see for example Wodika et al., 1997). Efron et al. (2001) investigated  $x = \text{avg}\{d_k = \log(PM_k) - c \log(MM_k), k = 1, 2, \dots, 20\}$ , for various scale factors  $c$ , as a measure of mRNA expression in comparing gene expressions in irradiated and unirradiated cells. Naef et al. (2001) questioned the assumed information content of the MM features as a measure of non-specific binding. Noting that the information content of the MM features is not clear they proposed expression indexes using only the PM features. Recently, Irizarry et al. (2001) investigated the quantities MM, PM, PM/MM, and PM-MM using spike-in studies where different quantities of RNA are added to the hybridization mixture at different concentrations. It was observed that the MM features rise with concentration levels, indicating that the MM features are detecting signal as well as non-specific binding.

As mentioned above, many “high-level” analyses accept the AvDiff (or slight variants of it) as the measure of gene expression. Li and Wong (2001a) brought attention to the fact that AvDiff, as a measure of expression, has not been studied extensively and emphasized the importance of “low-level” (probe level) analyses to obtain more sensitive measure of gene expression. They proposed a model-based estimate of expression index, namely the least squares estimates of  $\theta_i$  ( $i = 1, \dots, n$ ) from fitting the model  $PM_{ik} - MM_{ik} = \theta_i \phi_k + \epsilon_{ik}$ . Here  $\phi_k$  denotes the feature specific parameter and  $\epsilon_{ij} \sim N(0, \sigma^2)$  is the random error term. The estimation procedure proposed by Li and Wong also contains rules for outlier detection.

Recognizing problems with the AvDiff measure, the latest release of Affymetrix’s software (MAS

5.0) reports  $\text{signal} = \text{Tukey Biweight}\{PM_k - CT_k\}$ , where  $CT_k$  is a quantity computed from  $\{MM : MM < PM\}$ . Irizarry et al. (2001) compared various measures of expression, using spike-in and dilution studies. In these comparisons the expression measures are evaluated based on their ability to detect the known levels of gene expression. They also proposed their own measure of expression based on a background plus signal model.

## 5.2 Image Processing

Our discussion of Affymetrix arrays has assumed that the intensities for the 20 feature pairs are given. How are these intensities obtained? In the Affymetrix set up, there is no reference sample. Briefly, cDNA is made from experimental RNA, then cRNA (see Table 1) from cDNA and labeled with a single dye and hybridized onto the array: details are given in the next Section. The Affymetrix software uses a gridding procedure based on alignment features on the array to determine the locations of the features. Intensity value for a feature is computed as the 75th percentile of the pixel intensities for the feature, excluding the boundary pixels (Lockhart et al., 1996). Each feature consist of about 64 pixels. As in cDNA arrays, the signal intensities are corrected for background noise intensities. Due to limited space we refer the reader to Schadt et al. (2000) for details.

## 5.3 Affymetrix Sample Labeling

Affymetrix is a one-color array platform. We briefly describe the typical labeling protocol and discuss some advantages to the one-color detection scheme. Targets in one-color oligo systems, including Affymetrix, are labeled cRNA rather than cDNA. The basic steps are: (1) synthesis of double-stranded cDNAs using reverse transcription (RT), (2) Synthesis of targets, biotinylated cRNAs, using *in vitro* transcription (IVT), and (3) fragmentation of the biotinylated cRNAs. As in cDNA synthesis for two-color system, first strand cDNAs are synthesized via RT from total RNA (step 1). In particular, T7-linked oligo-dT primer is used in first strand cDNA synthesis, followed by second strand cDNA synthesis using DNA polymerase (Figure 1). The synthesized cDNAs in

step 1 serve as templates to make target cRNAs using IVT (step 2). The IVT is performed using biotinylated nucleotides to “label” the cRNAs. After IVT, these modified cRNAs are fragmented into smaller segments and hybridized on the array. The array is then washed and incubated with appropriate fluorescent dyes which couples with the biotins on the cRNAs.

Note that with a one-dye system unequal incorporation between dyes is not an issue. Spectral overlap between dyes is also not an issue. The reference sample is no longer needed, thus reducing the required biological materials for experiments. Also, possible genomic DNA being labeled during RT is avoided since modified nucleotides are incorporated during IVT. Fragmentation of the target cRNAs ensures that the most target lengths are within a reasonable range, thus avoiding target folding.

Finally, we note that there are other alternatives to the oligo array from Affymetrix, including Motorola CodeLink microarray, for example, which synthesizes probe oligos onto spots on the array similar to cDNA array (Ramakrishnan et al., 2002). New microarray systems, combining attributes of both “cDNA” and “oligo” technologies, with different oligos synthesis methods, provide a fertile ground for statisticians interested in image and other quality issues.

## 6 Variation, Normalization, and Design of Experiments

As pointed out in the introduction, the emphasis of this paper is not on the rapidly expanding literature on the statistical analysis of microarray data. Instead, we discuss a few basic issues that affect the quality of the data available for analysis: variation, normalization, and design of experiments. Some experimental designs have been proposed to understand of these issues.

Understanding the basic sources of variation within arrays and between arrays is important, both for improving analyses and for identifying areas of the experimental procedure that can lead to improvements. Normalization is crucial to a sensible comparison across arrays. In the following sections, we summarize some of the existing literature on these topics. Clearly, more research in

necessary.

## 6.1 Gene Expression Variation

Progress in understanding the basic variation of gene expression data is due to microarray experiments performed with replication. We distinguish three types of replication: (1) spot to spot, (2) array to array, and (3) subject to subject. The replication of spots (i.e. genes) is achieved by depositing probes for the same genes multiple times on the array. Array to array replication refers to multiple hybridizations using the same mix of RNA source. The third type of replication is sampling multiple individuals. The first assesses within array variation (spot-to-spot variation), the second between array variation, and the third biological variation. We next give some examples of experiments with replication and the information obtained from such experiments.

Data from microarray experiments where a gene is spotted (replicated) multiple times on the same array has revealed informative patterns of variation, reminiscent of other measurement technologies: one at higher intensity and another at low intensity measurements. See Ideker et al. (2000), who observed different error structure, depending on the observed intensity level in an experiment that contained both replication (1) and (2). They proposed an error model with both multiplicative and additive errors. This is based on the observation that the standard deviation (s.d.) from replicates is proportional to the means for high intensity measurements. However, due to background noise issues, the s.d. does not decrease to zero as the intensity measurement decreases to zero. Similar results and models have been reported by Rocke and Durbin (2001) based on data from Bartosiewicz et al. (2000), whose replication is of type (1).

Another experiment with replication type (1) and (2) is that of Lee et al. (2000). Here 3 hybridizations to cDNA arrays from the same RNA source with each array were used. Due to variability from one array to another, classification error rates based on an individual array is high compared to classification based on combined data from all 3 arrays.

Observed variation from microarray experiments can be due to numerous sources, such as the target labeling and purification procedure, hybridization conditions, array fabrication, and data collection. The exact factors contributing to variation are not well understood, due to lack of experiments devoted to understanding the experimental procedure. Replication experiments have shown some of the sizes and properties of variation, and it has become clear that replication of both types (1) and (2) is essential.

## 6.2 cDNA Array Normalization

Many experimental steps, such as target labeling, can also introduce systematic biases. The current approach to cope with such problems is through array normalization, which appears to be necessary in order to compare across arrays.

To determine normalization factors, the set of genes to use for normalization must be determined. The choice of the gene set is often determined from a “biological hypothesis”, e.g., a group of genes that are believed to be constantly expressed, sometimes called housekeeping genes. These housekeeping genes establish a baseline for normalization. Another approach uses all genes on the array for normalization. This approach is used when the majority of genes are believed to be not differentially expressed in the experiment (Yang et al., 2001). The use of various types of control sequences spotted on the arrays can also be used for normalization. The basic idea here, whether using synthetic DNA control sequences, cross-species DNA control sequences, or a series of titration control sequences, is that intensities for both channels in the control spots should be equal, hence, can be used for normalization. Unfortunately, it is very difficult, if not impossible, to find a set of natural or control genes that are constantly expressed.

After a normalization gene set has been chosen, the normalization procedure is applied to each array separately. For a given array let  $(r_j, g_j)$  be the Cy5 and Cy3-channel intensity values for genes  $j = 1, \dots, p$ . A widely used normalization method is global normalization where the center of the

distribution of the log-ratios is shifted to zero. Here the intensities that are assumed to be related by a simple constant factor,  $k$ ,  $r_j = kg_j$ . Under this assumption,  $\log(r_j/kg_j) = \log(r_j/g_j) - c$ , where  $c = \log(k)$ . Thus, a constant shift normalization is made for all spots,  $\log r_j/g_j \leftarrow \log r_j/g_j - c$ . Various methods of estimating  $c$  have been used (Ideker et al., 2000), while some researchers do not take medians (Schummer et al., 1999). See also Tseng et al. (2001) for normalization of cDNA arrays.

In many experiments, such normalization has now been found to be less than adequate. Yang et al. (2001) observed that the dye bias is dependent on spot intensity,  $r_j = k_j g_j$ , and suggested a nonlinear normalization procedure which depends on the average log intensity,  $k_j = k(A_j)$  where  $A_j = 0.5(\log r_j + \log g_j)$ . The normalization is  $\log r_j/g_j \leftarrow \log r_j/g_j - c(A_j)$ , where  $c(A_j) = \log k(A_j)$  is the loess fit to the scatter plot of  $(M_j, A_j)$  and  $M_j = \log r_j/g_j$ . A similar method was later proposed by Tseng et al. (2001). Yang et al. (2001) extend their methods to allow for normalization due to systematic bias arising from print-tip groups. There remains considerable interest in array normalization, and more research is needed.

### 6.3 Design of Experiments

The importance of design of experiments (DOE) for microarray studies was emphasized by Kerr and Churchill (2001), where the use of ANOVA models for cDNA normalization was also proposed. The amount of data gained, quality of data, assessment of the sources of variation, estimation of error variation, and precision of estimates among others are intricately related to DOE. Carefully designed experiments can contribute at many levels, including assessment of the technology to eventual understanding of biological processes which generate the data. However, despite the tremendous activities ranging from low-level to high-level analysis, the use of DOE in microarray experiments is lacking. Due to limited space we briefly review and discuss where DOE can play a more significant role.

Principles of DOE were emphasized primarily by Kerr and Churchill (2001; 2001b). For example, they examined sources of expression variation due to array, dye, treatment (“variety”), gene, and a few interactions from a “reverse-labeling” design (“dye-swap”) and a “reference” design with two arrays. The term “reference” design refers to the original design where every sample is compared to a common reference, as described in Section 3.2.1. In the reverse-labeling design, the second experiment (array) is replicated with the Cy5 and Cy3 labeling reversed. Imbalance in the dyes is revealed in the reverse-labeling design. As pointed out by the authors, in the reference design, most information is collected on the reference (control) sample which is not the quantity of real interest. However, for the reverse-labeling design two measurements per gene are taken and only one for the reference design. For more than two varieties, Kerr and Churchill (2001) proposed a loop structure design which collects twice as much data on the varieties of interest than the reference design. In these designs varieties are balanced with respect to dyes.

Despite potential gains, carefully designed experiments have not been widely adopted in microarray studies. One reason is that reference design, although inefficient, can be easily extended to more samples by simply adding another array using the same reference. Other reasons are associated cost, physical limitations of the experimental procedure, and innovations in sample preparation, labeling, and detection. The lack of DOE is also due to the lack of statisticians devoted to the area. It is our hope that this paper will generate more interest from statisticians.

Array technologies are still evolving and DOE can play more prominent roles in assessing the technologies. For example, carefully designed experiments are needed to more rigorously evaluate the sample labeling protocol, extent of cross-hybridization, relationship between true mRNA concentrations and measure intensities, and quality of microarrays among others. Microarray can also be thought of as similar to other assays. Oligo and cDNA arrays are in principle similar to radioimmunoassay, ELISA assay, HPLC assay, and other bioassays. In optimizing assays, we have to identify the factors that affect assay quality, before DOE can provide effective improvements.

Concepts in these assays, like working range, quality limits, detection limits, among others, can be adopted. Microarray is still evolving. DOE in this area may not currently be very exciting, but it is useful and can be potentially important.

## 7 Discussion

Since the introduction of the cDNA array (in 1995) and the Affymetrix array (in 1996), the use of the technologies have spread to many fields of research. There is already a large and rapidly expanding literature on applications of the technologies and statistical analyses. There has also been progress in the evaluation of the experimental process, mainly through basic experiments with replications. In addition, efforts have been made on improving “low-level” analysis starting from the raw images.

However, the technologies are evolving slowly, with only a few new technical innovations. A review of the literature shows few publications on the technical issues and on the evaluation of the technologies compared to their applications, until recently. Basic research, starting from selecting the materials, such as dyes and substrate material, to labeling, and hybridization can gain much from better designed experiments. Statistics, particularly DOE, can play an important role in evaluating the technologies and process improvement.

Understanding the fundamental biological and technological aspects of microarray experiments can expand on ways statisticians contribute to experimental science. For example, in addition to the methodological developments, statisticians have significantly contributed to elucidating experimental techniques and biological assumptions. This includes, for instance, nucleotide labeling using RT and the assumption that a single central nucleotide mismatch in the MM features on Affymetrix arrays measure non-specific binding.

## Acknowledgments

We are grateful to three referees and an associate editor for very helpful comments. Our research was supported by National Cancer Institute grants CA90301, CA57030 and CA74552, and by the National Institute of Environmental Health Sciences (P30-ES09106).

## References

- Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. and Watson, J. D. (1994). *Molecular Biology of The Cell*. Yew York: Garland Publishing Inc.
- Amersham Pharmacia Biotech. *Fluorescence Imaging: Principles and Methods*. Sunnyvale, California.  
<http://www.mdyn.com>.
- Bartosiewicz, M., Troustine, M., Barker, D., Johnson, R. and Buckpitt, A. (2000). Development of a toxicological gene array and quantitative assessment of this technology. *Archives of Biochemistry and Biophysics*, **376**, 66–73.
- 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.
- DeRisi, J., Iyer, V., and Brown, P. O. (1999). *The MGuide: A Complete Guide to Building Your Own Microarrayer version 2.0*. Stanford, California: Biochemistry Department, Stanford University.  
<http://cmgm.stanford.edu/pbrown/mguide>.
- Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P. and Trent, J. M. (1999). Expression profiling using cDNA microarrays. *Nature Genetics*, **21**, 10–14.
- Efron, B., Tibshirani, R., Storey, J. D., and Goss, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, **96**, 1151-1160.

Genisphere, *3DNA Submicro Expression Array Detection Kit* (Protocol). Hatfield, Pennsylvania. <http://genisp.com>

Griffiths, A. J. F., Miller, J. H., Suzuki, D. T., Lewontin, R. C. and Gelbart, W. M. (2000). *An Introduction to Genetic Analysis*. New York: W.H. Freeman and Company.

Haab, B. B., Dunham, M. J. and Brown, P. O. (2001). Protein Microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biology*, **2**(2):research 0004.1–0004.13.

Ideker, T., Thorsson, V., Siegel, A. F., and Hood, L. R. (2000). Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *Journal of Computational Biology*, **6**, 805–817.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2001). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Manuscript. Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland.

Kahn, J., Saal, L. H., Bittner, M. L., Chen, Y., Trent, J. M. and Meltzer, P. S. (1999). Expression profiling in cancer using cDNA microarrays. *Electrophoresis*, **20**, 223–229.

Kerr, M. K. and Churchill, G. A. (2001). Experimental design for gene expression microarrays. *Biostatistics*, **2**, 183–202.

Kerr, M. K. and Churchill, G. A., (2001b). Statistical design and analysis of gene expression microarrays. *Genetical Research*, **77**, 123–128.

Lander, E. S., and Weinberg, R. A. (2000). Journal to the center of biology. *Science*, **287**, 1777–1782.

Lee, M. T., Cuo, F. C., Whitmore, G. A., and Sklar, J. (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridization. *Proceeding of the National Academy of Sciences, USA*, **97**, 9834–9839.

- Li, C. and Wong, W. H. (2001a). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceeding of the National Academy of Sciences, USA*, **98**, 31–36.
- Li, C. and Wong, W. H. (2001b). Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology*, **2**, 1–11.
- Lipshutz, R. J., Fodor, S. P. A., 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 of monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, **14**, 1675–1680.
- May, B. J., Zhang, Q., Li, L. L., Whittam, T. S., and Kapur V. (2001). Complete genomic sequence of *Pasteurella multocida*, Pm70. *Proceeding of the National Academy of Sciences, USA*, **98**, 3460–3465.
- Naef, F., Lim, D. A., Patil, N., and Magnasco, M. O. (2001). Manuscript. The Laboratories of Mathematical Physics, Rockefeller University, New York, New York.
- Newton, M. A., Kendzierski, C. M., Richmond, C. S., Blattner, F. R., Tsui, K. W. (2001). On differential variability of the expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, **8**, 37–52.
- Ramakrishnan, R., Dorris, D., Lublinsky, A., Nguyen, A., Domanus, M., Prokhorova, A., Gieser, L., Touma, E., Lockner, R., Tata, M., Zhu, X., Patterson, M., Shippy, R., Sendera, T. J., and Mazumder, A. (2002). An assessment of Motorola CodeLink<sup>TM</sup> microarray performance for gene expression profiling applications. *Nucleic Acids Research*, **30**, e30.
- Rocke, D. M. and Durbin, B. (2001). A model for measurement error for gene expression arrays. *Journal of Computational Biology*, **8**, 557–569.

- Schadt, E. E., Li, C., Su, C., and Wong H. (2000). Analyzing high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry*, **80**, 192–202.
- 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.
- Schermer, M. J. (1999). Confocal scanning microscopy in microarray detection. In *DNA Microarrays: A Practical Approach*, Schena, M. (ed), 17–42. New York: Oxford University Press.
- Stears, R. L., Getts, R. C. and Gullans, S. R. (2000). A Novel, sensitive detection system for high-density microarrays using dendrimer technology. *Physiol Genomics*, **3**, 93–99.
- Tseng, G. C., Oh, M., Rohlin, L., Liao, J. C and Wong, W. H. (2001). Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research*, **29**, 2549–2557.
- Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to ionizing radiation response. *Proceedings of the National Academy of Sciences, USA*, **98**, 5116–5121.
- Watson, J. D. and Crick, F. H. C. (1953). Genetical implications of the structure of deoxyribonucleic acid. *Nature*, **171**, 964–967.
- Wodicka, L., Dong, H., Mittmann, M., Ho. M-H. and Lockhart, D., (1997). Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nature Biotechnology*, **15**, 1359–1467.
- Wong, D. T., Basehore, S., Buchanan, M. A., Sorge, J. A. and Mullinax, M. (2001). Preparation of labeled cDNA without bias using FairPlay™ microarray labeling kit. *Strategies Newsletter*, **14**, 62–63.
- Yang, Y. H., Buckley, M. J., Dudoit, S. and Speed, T. P. (2002). Comparison of methods for Image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics*, **11**, 108–136.

Yang, Y. H., Dudoit, S., Luu, P. and Speed, T. P. (2001). Normalization for cDNA microarray data.  
In *Microarrays: Optical Technologies and Informatics* Bittner, M. L., Chen, Y., Dorsel, A. N., and  
Dougherty, E. R. (eds). San Jose: Proceedings of SPIE, Society for Optical Engineering.

Table 1: Description of common terminologies in the *context* of microarrays. Definitions of some terminologies have been simplified.

<b>DNA</b>	deoxyribonucleic acid contains base A, G, C, or T; is double stranded
<b>denaturation</b>	the separation of the two DNA strands when heated
<b>RNA</b>	ribonucleic acid contains base A, G, C, or U; is single stranded
<b>base complementarity</b>	the pairing of base G w/ C, A w/ T in DNA, but with U replacing T in RNA
<b>nucleotides</b>	molecular units making up DNA and RNA; a nucleotide is s-p-Base, p=phosphate, s=deoxyribose sugar for DNA & ribose sugar for RNA
<b>amino acid</b>	the basic building block of proteins (or polypeptides)
<b>mRNA</b>	messenger RNA is a RNA strand complementary to a DNA template
<b>transcription</b>	the process where the DNA template is copied/transcribed to mRNA
<b>gene expression</b>	a gene is expressed if its DNA has been transcribed to RNA and gene expression is the level of transcription of the DNA of the gene
<b>RT</b>	reverse transcription is an experimental procedure to synthesize a DNA strand complementary to a mRNA template, namely cDNA
<b>cDNA/cRNA</b>	complementary DNA is DNA synthesized from mRNA during RT and similarly, in the context of oligo arrays, complementary RNA is RNA synthesized during <i>in vitro</i> transcription
<b>dNTP</b>	deoxyribo nucleoside triphosphate; denotes any of dUTP, dTTP, dATP or dGTP; molecular building blocks for making DNAs in RT, PCR, or <i>in vitro</i> replication; dNTPs in solution, not incorporated into the nucleic acid strand yet, as molecules w/ 3 phosphates provide the necessary energy for cDNA synthesis
<b>rNTP</b>	used for synthesis of cRNA; see above description for dNTP
<b>primer</b>	a short single strand RNA or DNA which can initiate chain growth from a template
<b>oligo(dT)</b>	primer with sequence TTTT... used to initiate cDNA synthesis in RT
<b>reverse transcriptase</b>	an enzyme which catalyzes the synthesis of cDNA during RT
<b>poly(A) tail</b>	a sequence of A (AAA...) at the 3' end of mRNA; oligo(dT) is used in RT to recognize mRNA by its poly(A) tail
<b>target cDNAs</b>	mixture of cDNAs obtained from the experiment and reference mRNAs
<b>probe cDNAs</b>	immobilized cDNA printed on the array
<b>hybridization</b>	process of bringing into contact the target and probe for binding in microarray, but refers to the binding two DNA strands generally
<b>PCR</b>	polymerase chain reaction is a procedure to amplify a segment of DNA

Table 2: **20 (PM-MM) feature pairs.** Given are intensity of the 20 feature pairs,  $(PM_{ijk}, MM_{ijk})$ ,  $k = 1, \dots, 20$ , for gene  $j = 2715$  in array  $i$ . Adopted from Efron et al. (2001).

Feature ( $k$ )	1	2	3	4	5	6	7	8	9	10
PM	1054	3242	1470	4050	1356	1476	561	606	1307	1057
MM	793	2333	826	1912	561	558	942	526	699	1060
Feature ( $k$ )	11	12	13	14	15	16	17	18	19	20
PM	974	1584	802	1399	1670	2514	2096	6592	5662	2244
MM	829	1771	601	569	840	950	700	8717	1484	668

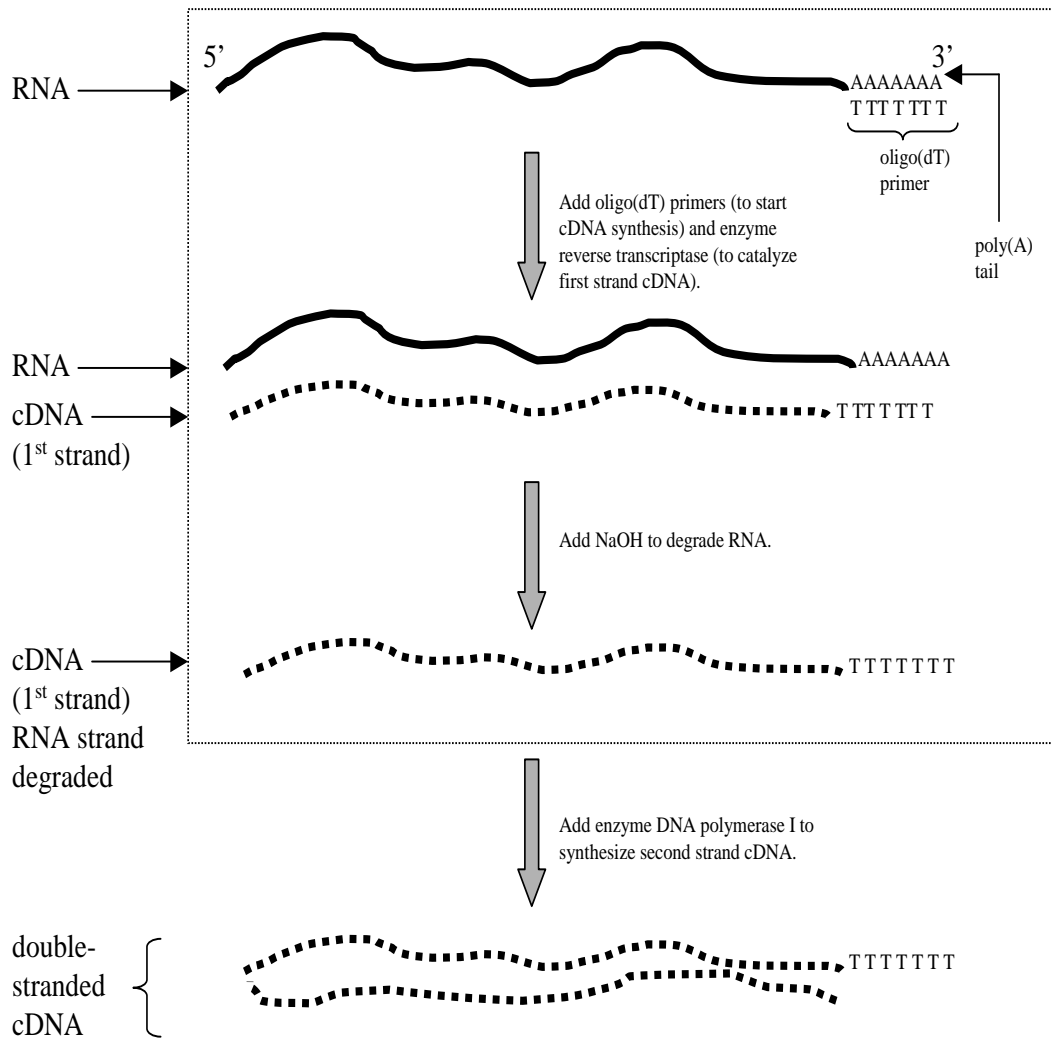


Figure 1:

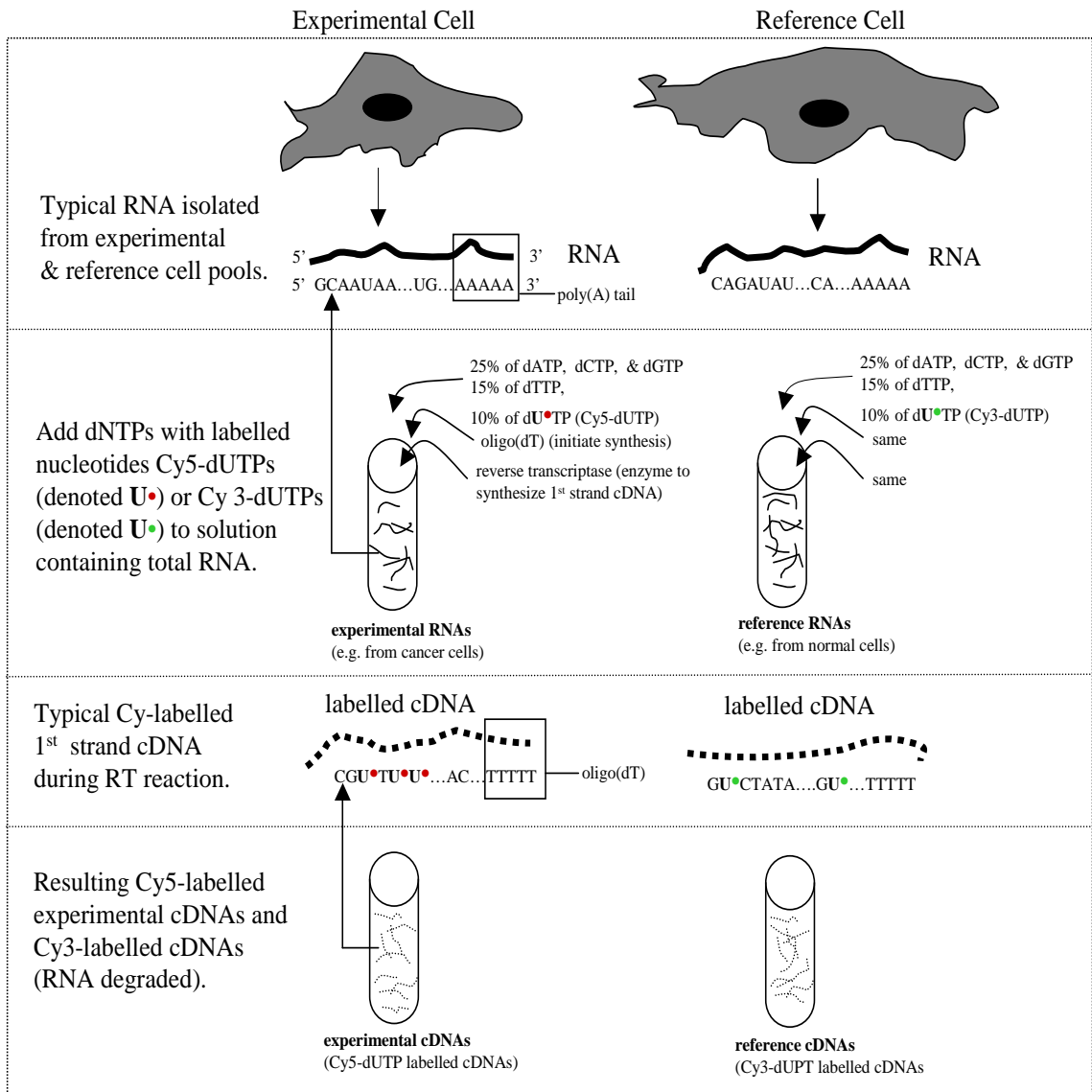


Figure 2:

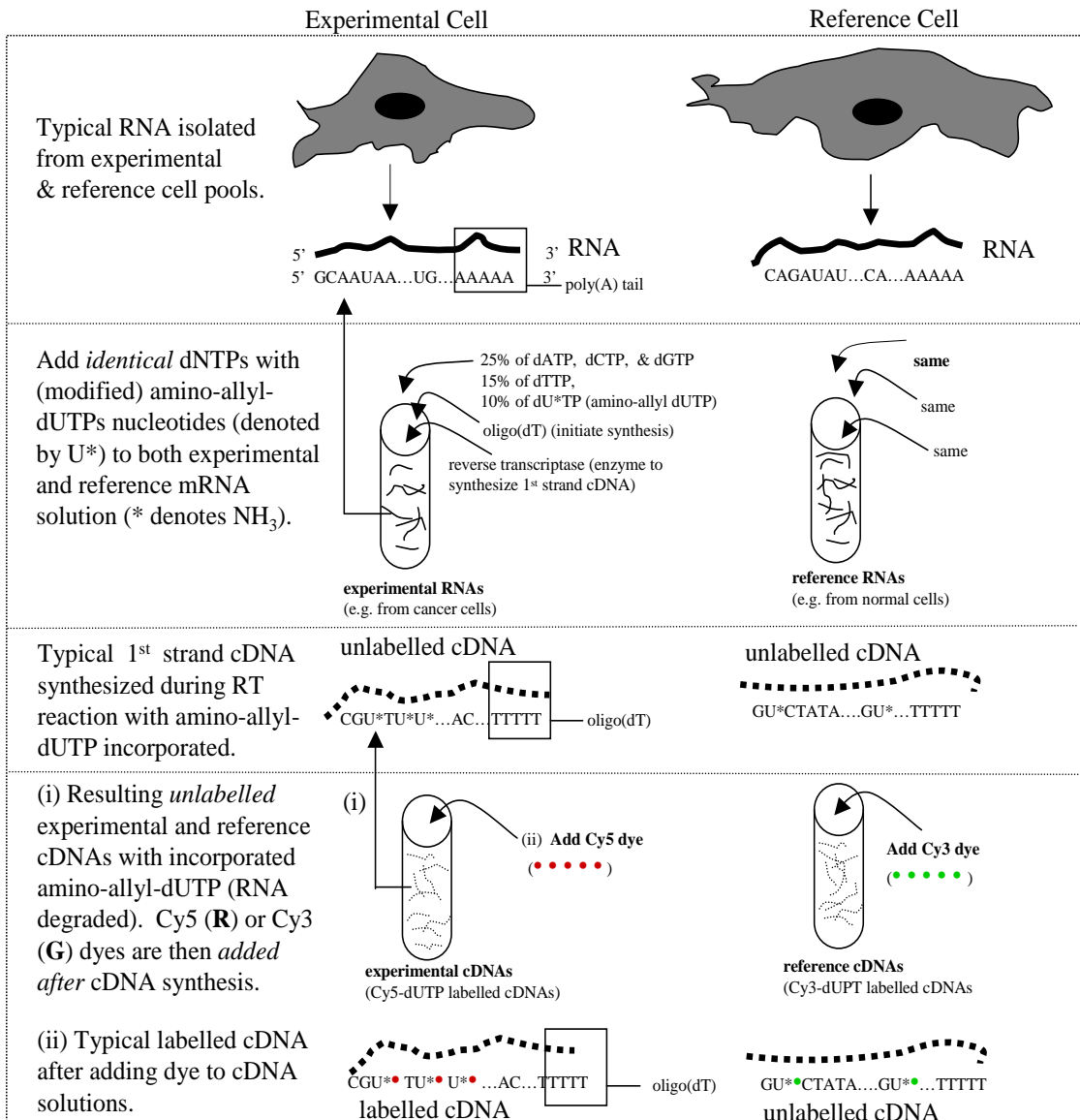


Figure 3:

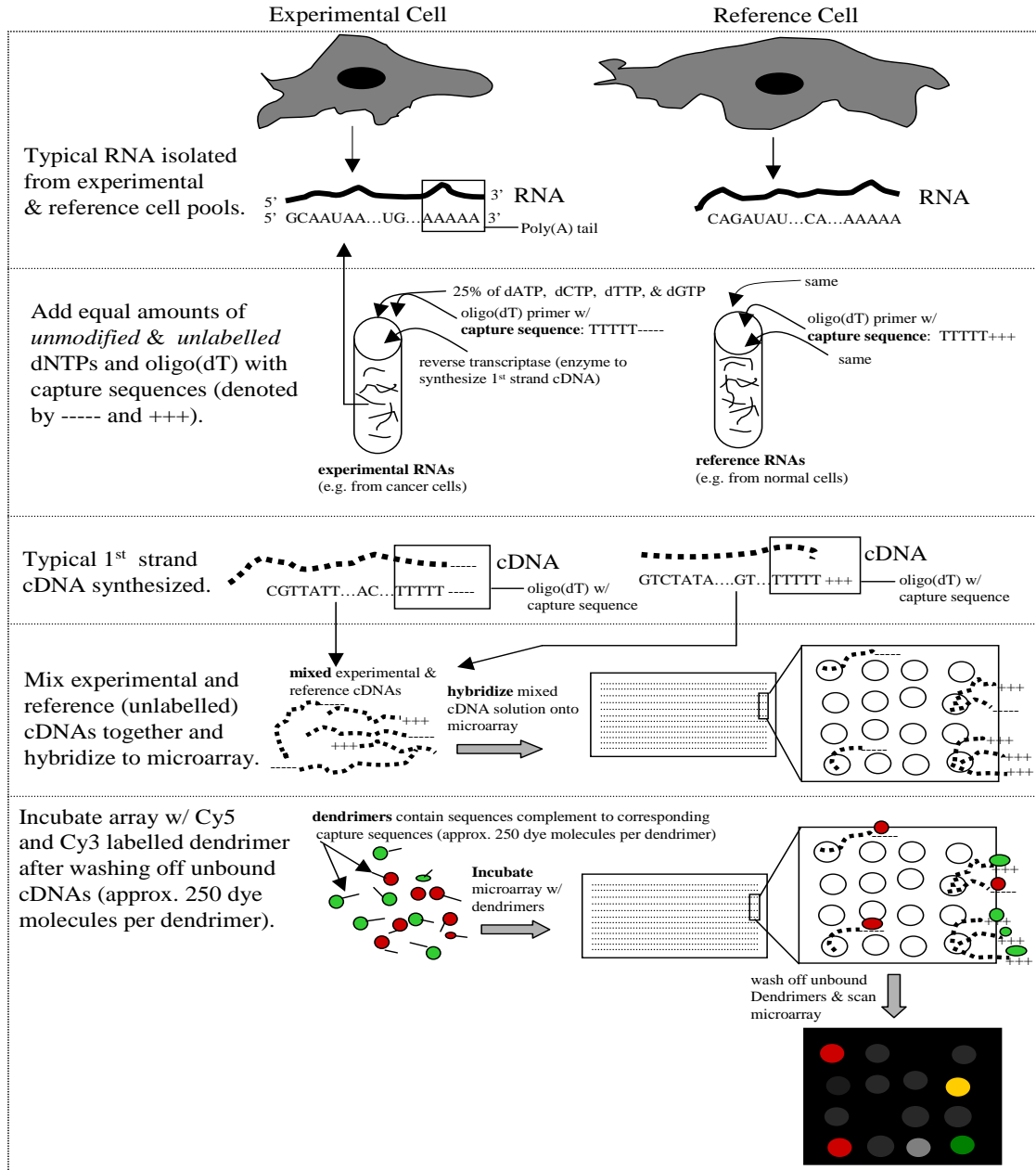


Figure 4:

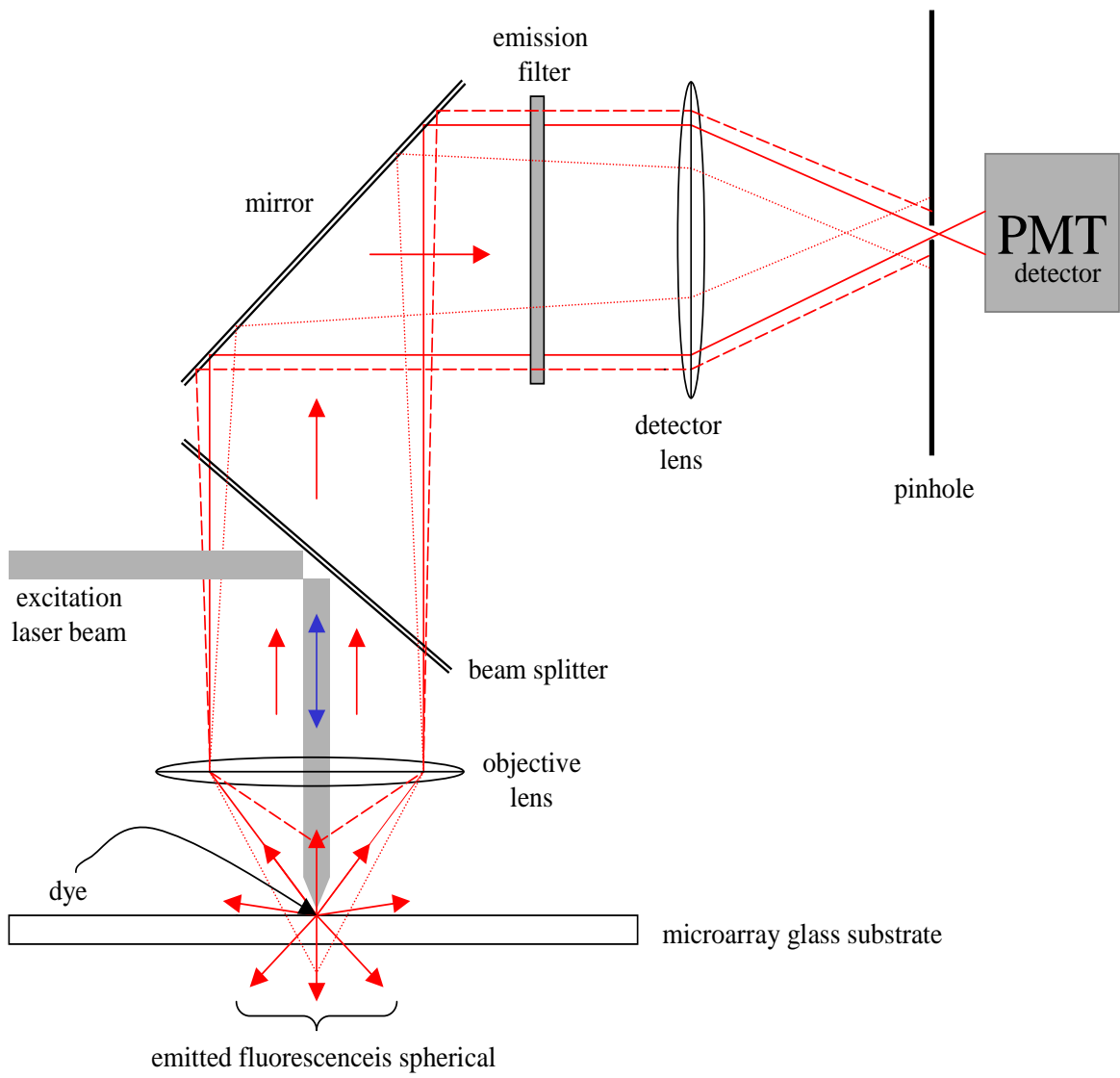


Figure 5:

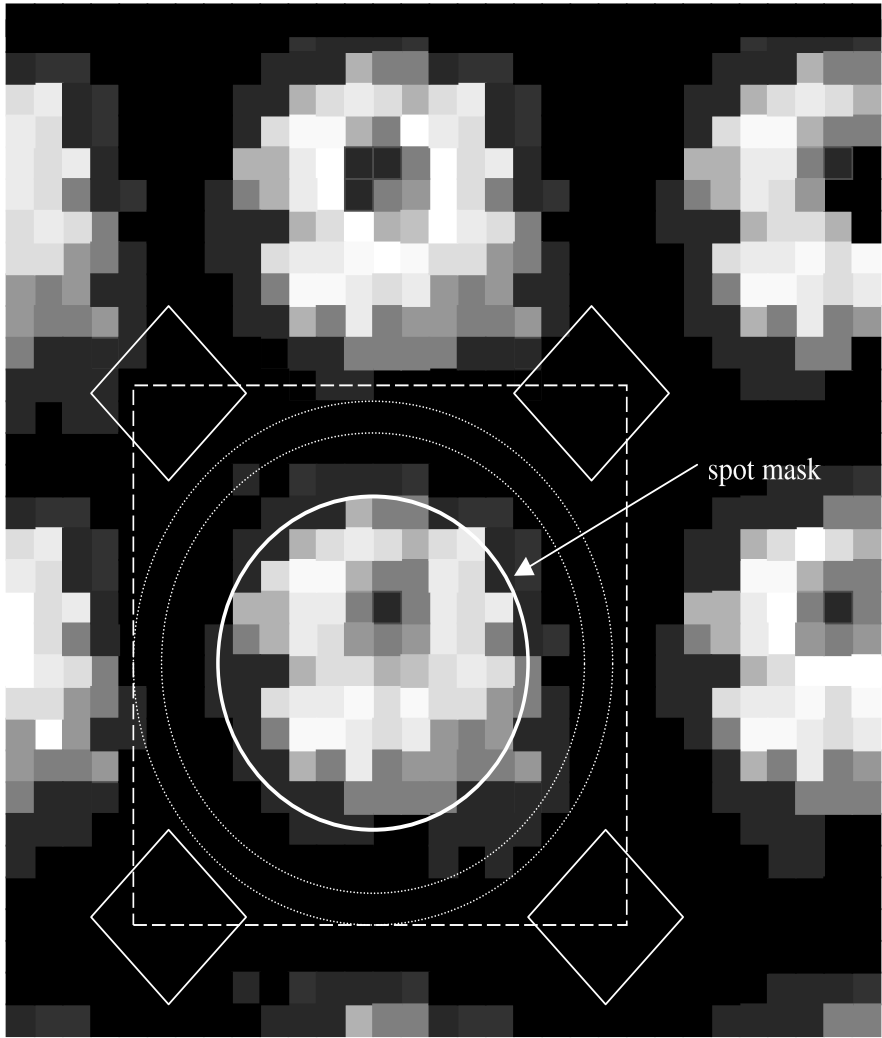


Figure 6:

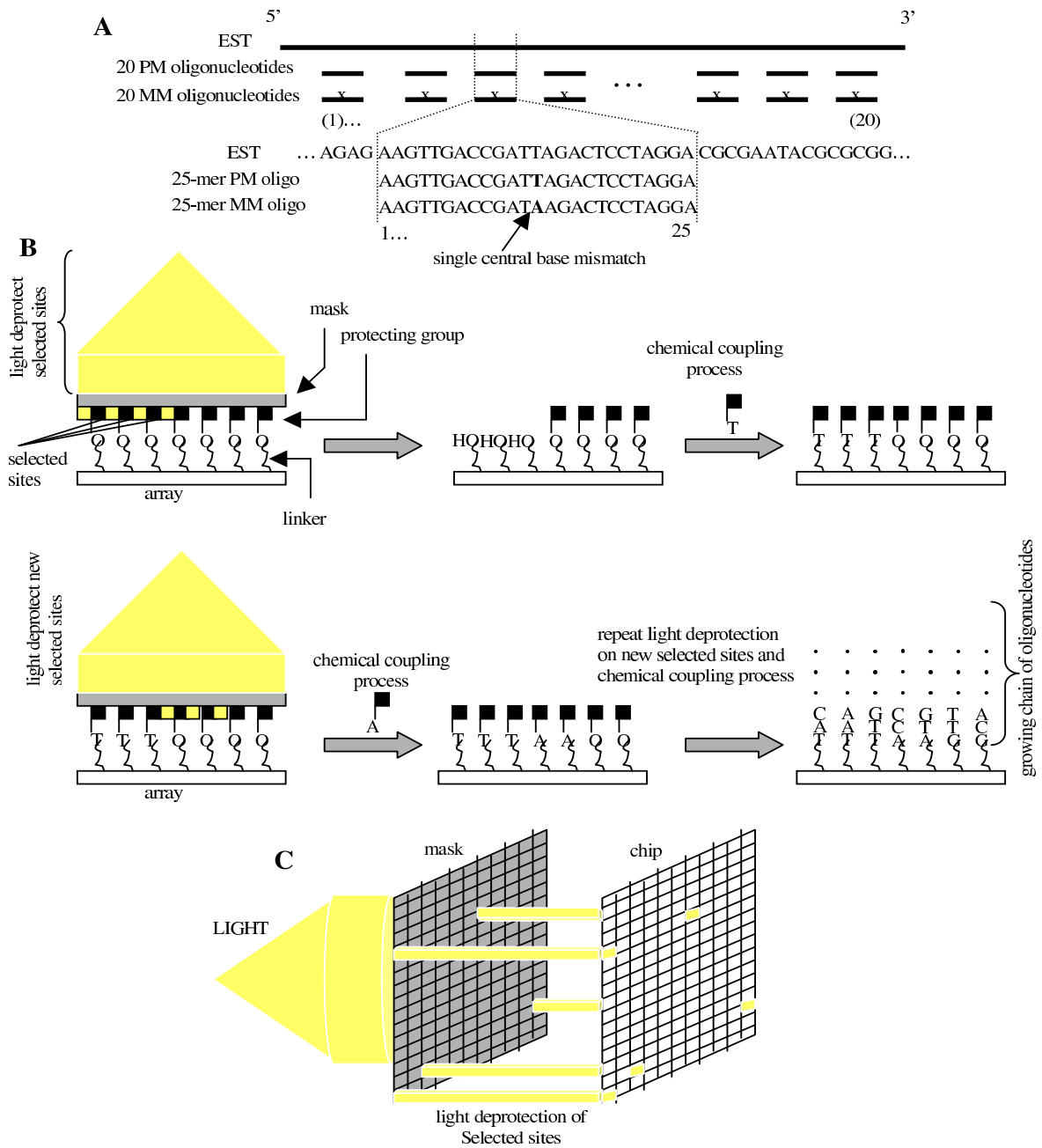


Figure 7:

## FIGURE CAPTIONS

1. RT is an experimental technique where mRNA strands are used to make (synthesize) double-stranded complementary DNAs (cDNAs). There are two main ingredients that are required to initiate and carry out the synthesis of cDNAs: (1) oligo(dT) primers and (2) reverse transcriptase (enzymes). (1) Oligo(dT) primers recognizes the complementary “signature” poly(A) tails of mRNAs in solution. (2) Reverse transcriptase catalyzes first strand cDNA synthesis. In cDNA microarray experiments, only the first strand cDNA is usually made and the main steps involved in first strand cDNA synthesis are boxed above. For oligo arrays, both cDNA strands are synthesized and then cRNAs are made for the cDNAs.
2. Total RNA are obtained from the experimental and reference cell groups separately (**top**). In separate RT reactions (**second** panel) *labeled* dUTPs (Cy5-labeled dUTPs for the experimental RNAs and Cy3-labeled dUTPs for the reference RNAs) are added to the respective solutions. Thus, the resulting experimental and reference cDNA strands are Cy5- and Cy3-labeled respectively (**third & bottom** panel). mRNA and unused nucleotides are eliminated so that mostly what remains in solution are the labeled cDNAs (**bottom** panel). It has been noted that Cy-dyes are not uniformly and efficiently incorporated into the cDNA. Usually a Cy3-labeled cDNA have higher fluorescent intensity than a Cy5-labeled cDNA from the same RNA source. Furthermore, the intensity would depend on the base composition as well as the cDNA length.
3. It is believed that the Cy-labeled dNTPs exhibit some steric hindrance contributing to non-efficient and non-uniform incorporation into the cDNA. Therefore, instead of having different labeled nucleotides, *identical* and *unlabeled* dNTPs are added to *both* experimental and reference samples (**second** panel). Thus, the cDNAs synthesized during RT are unlabeled and both pools of cDNAs contain modified dUTPs (amino-allyl-dUTPs) (**third & bottom** panel (i)). (It is claimed that the amino-allyl is small, have less steric hindrance, and hence, the amino-allyl-dUTPs are uniformly incorporated into the cDNAs with higher frequency.) Next (**bottom** panel), Cy5 dyes are then added to the experimental cDNAs and the Cy3 dyes added to the reference cDNAs (**bottom** panel). The added Cy-dyes couple with the cDNAs containing amino-allyl-dUTPs resulting in Cy5-labeled experimental and Cy3-labeled cDNAs (**bottom** panel (ii)). Note that the problem of the fluorescence intensity depending on the base composition and cDNA length, encountered with direct labelling, still remains.
4. To eliminate measured intensities depending on base composition and length, nucleotide labeling is not used. Instead, the primers added during RT are tagged with a capture sequence, one for the experimental and another for the reference RT run (**second** panel). Thus, the resulting synthesized cDNAs (in both samples) contain *unmodified* and *unlabeled* nucleotides (**third** panel). Two pools of cDNAs are mixed and hybridized to the array (**fourth** panel). Finally, hybridized arrays are incubated in a mixture of Cy5- and Cy3-dendrimer, each containing about 250 dyes molecules (**bottom** panel). The dendrimers contain sequences complementary to the capture sequence attached to the cDNA.

5. Depicted are the basic components of a laser confocal microscope used in microarray detection. Emission source (photons) are converted into electric current using a detector (e.g. a PMT (photomultiplier tube) detector) in a scanner. Given are three basic types of emission sources: (1) focused emission source (—), (2) out-of-focus (over) emission source (...), and (3) out-of-focus (under) emission source (---). In both out-of-focus sources, most of emission light is deflected and does not enter the pinhole, thus, does not reach the PMT detector.
6. A segmentation method is used to determine the spot mask (solid white circle). Pixels inside the spot mask are used to calculate the spot intensity. There are various ways that have been used to calculate the background intensity. Examples include using (1) pixels outside the spot area within the square (---), (2) pixels between the concentric circles (...), or (3) pixels inside the four diamonds (—). Adopted from Yang et al. (2002).
7. **(A)** An gene sequence is represented by 20 subsequences of the gene, each of length 25-bp (oligonucleotides). These are subsequences that are perfect match (PM) to the subsequences of the gene. Another 20 subsequences with the same bases as the PMs, except for one mismatch (MM) at the central base (arrow) is used. **(B)** Depicted is the light-directed process of synthesizing the oligonucleotides on the chip (array). **(C)** The schematics of the light, mask, and array in the oligonucleotide synthesis process. Adopted from Lipshultz et al. (1999), *Nature Genetics*, **21**, 20–24.