

## 2 BACKGROUND ON MICROARRAY TECHNOLOGY

Recent years have seen the complete sequencing of the human genome. Many other organisms' genomes have been sequenced as well; examples include fruitflies, yeast, and many bacteria. From sequence information, we can learn much of the structure of genes and DNA. However, sequence analysis alone cannot tell us what genes are or how they are used. To reveal more of the functional properties of genes, DNA microarrays measure genome-wide expression. Here, I provide an abbreviated overview of the genetic underpinnings of this technology, and a quick introduction to the technology itself.

### 2.1 Genetic Background

Humans have tens of thousands of genes; taken together, these constitute the human genome. Each *gene* is a unique subsection of the genome and consists of a sequence of a few thousand *nucleotides*. A nucleotide is a special type of molecule that contains four nitrogen bases (some combination of adenine, guanine, cytosine and thymine). From an informatic perspective, a nucleotide can be seen as a member of the four-letter alphabet {A,G,C,T}; a gene can be regarded as a special sequence of these nucleotides (e.g. ...CCTATAGCAACG...).

Genes are important because they code for amino acids, which in turn form proteins – the basic elements of every cell. Genes code for amino acids via a two-step process of *transcription* and *translation*. In transcription, the cell produces a

piece of mRNA that is a complementary copy of the gene. Each section of DNA uniquely corresponds to one section of mRNA, and vice versa. In translation, amino acids are produced directly from the mRNA.[52]

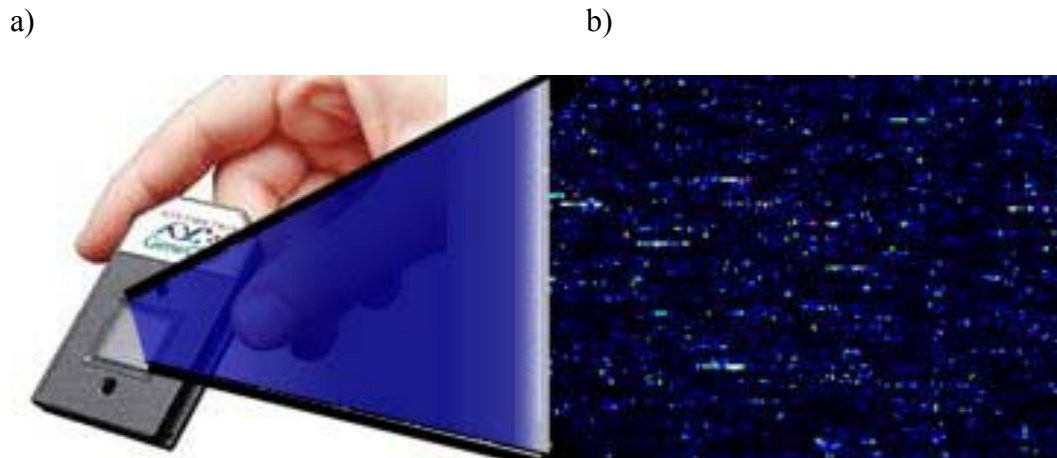
Of course, this is a gross oversimplification of the process, as there are many other factors and molecules involved in transcription and translation. Nonetheless, the important point is that mRNA is a crucial medium that enables the production of amino acids (and therefore proteins) from DNA. As Lander summarized, “The mRNA levels sensitively reflect the state of the cell, perhaps uniquely defining cell types, stages, and responses.”[64]

## **2.2 Microarrays and Expression Levels**

Microarrays measure the presence of mRNA. The mRNA can be extracted from cells, tissues, etc. By analyzing extracted mRNA, one obtains a quantitative assessment of the genetic activity of the location from which the mRNA was extracted. Microarrays derive an *expression level* for each gene – a scalar value corresponding to the amount of mRNA which in turn corresponds to the gene in question. High expression levels indicate a high amount of genetic activity, whereas low expression levels indicate inactivity.

There are three predominant types of microarray technology: high-density oligonucleotide arrays, cDNA microarrays, and SAGE (serial analysis of gene expression). Each technology measures levels of gene expression. Here, I focus on high-density oligonucleotide arrays, though in principle the analysis applies to all three.

Oligonucleotide arrays (Figure 2.1) consist of a high-density grid of a few hundred thousand *oligonucleotides*. Each oligonucleotide, a manufactured sequence of twenty-five bases (also {A,G,C,T}), uniquely corresponds to a specific gene via the same complementarity that relates mRNA to DNA. Using light-directed, solid-phase combinatorial chemical synthesis, these oligonucleotides are spotted onto a glass slide. When immersed in mRNA, the mRNA hybridizes to its oligonucleotide match on the chip. The amount of hybridization is measured by fluorescently staining the array, and subsequently scanning the array to measure the intensity of fluorescence. Thus, on a single high-density array, it is possible to simultaneously obtain expression levels for thousands of individual genes.[20, 67].



**Figure 2.1** *Oligonucleotide arrays.* a) An Affymetrix® GeneChip. b) Scanned image of probe array.

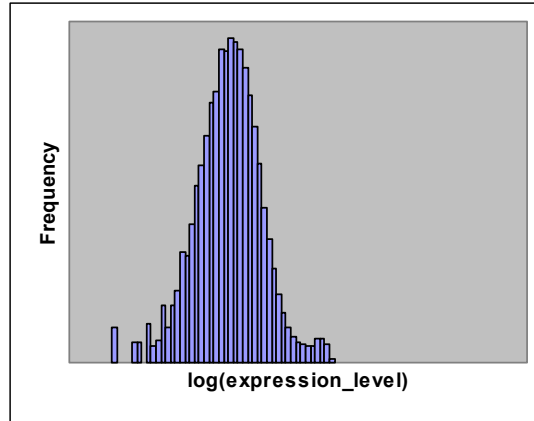
### **2.3 Quantitative Aspects of Microarray Data**

Once the microarray has been processed, the user is left with a scalar expression level for each gene. Each expression level spans three to four orders of

magnitude. In the Affymetrix® arrays referred to in this paper, there are roughly 10,000 such gene expression levels. The distribution of expression levels on each microarray is approximately lognormal (Figure 2.2a) [48, 59], normalized to a mean level specified by the user. Taken together, the 10,000 expression levels comprise a complete expression profile for the sample mRNA.

However, the measurements are not perfect[62]. For instance, though chip technology is improving, background noise and chip defects still contaminate microarray data [34, 50]. Then too, the same mRNA, when washed on two identical chips, does not produce identical expression profiles; a complete model for this error was recently derived by Jensen et al [59]. Aside from chip errors, a common source of error is in the extracted mRNA itself. For example, Ben-Dor *et al.*[8] recently noted that in the colon cancer data used by Alon *et al.*[6], the normal colon biopsy also included smooth muscle tissue from the colon walls. This caused the muscle-related genes to be disproportionately expressed in the normal cells, when compared to the cancerous cells.

Such unwanted variance often leads researchers to use only those genes with high expression levels, standard deviations, and variances – such considerations are discussed in Appendix A, section 3.



**Figure 2.2** *Quantitative Aspects of Microarray Data.* Distribution of logarithm of the expression levels for an Affymetrix® human microarray.