26

Quantification of Gene Expression Based on Microarray Experiment

Samane F. Farsani and Mahmood A. Mahdavi
Department of Chemical Engineering, Ferdowsi University of Mashhad, Azadi Square, Pardis Campus, Mashhad, Iran

1. Introduction

Gene expression is a common process in all forms of living cells including eukaryotes, prokaryotes and viruses to generate the macromolecular requirements for life. The study of gene expression provides a systemic comprehension of the cell function for addressing specific biological questions. This process comprises replication, transcription, RNA splicing, translation and post translational modification of a single protein. At first, DNA serves as a template to replicate itself and the production of RNA (transcription), a copy from the DNA, is mediated by RNA polymerase. In prokaryotes, transcription creates messenger RNA (mRNA) which doesn't need any additional processing for translation but this stage in eukaryotes produces a primary transcript of RNA, which needs further processing prior to becoming a mature mRNA. This step is referred to as RNA splicing that in the proper context, involves the removal of certain sequences called intervening sequences, or introns. Hence, the final mRNA contains the remaining sequences, called exons, which are spliced together (Knapp et al., 1978). In the next stage, so called translation, mRNA separates from DNA strand and serves as a template for protein production that such a process is assisted by ribosomes. Proteins are modified after translation in variety of processes i.e. they are altered at structural level to achieve the final 3D conformation. These modifications are essential for all aspects of biology and can be performed spontaneously or driven by enzyme mediation. Common post-translational modifications include phosphorylation, glycosilation, dimerization or tetramerizaion, etc. (Doyle & Mamula, 2001). Therefore, the transfer of genetic information, from DNA to RNA and to proteins, ending up with the expression of genes in all cells makes up the central dogma of molecular biology (Figure 1) (Crick, 1970).

Genomics information is delivered to the cells in three biochemical datasets including the complete set of mRNA species that result in generating proteins (transcriptomics), the complete collection of proteins (proteomics), and the complete series of metabolites produced in the cell (metabolomics) (Figure 2)(Karakach et al., 2010; van der Werf et al., 2005).

Transcriptomics provides a complete profile of RNAs that appear within the cells, tissues and biological fluids at a specific time. The mRNA levels do vary over time, among diverse cell types and within cells under different conditions while DNA is more or less unchanged over the life cycles. Thus, gene expression based on mRNA mediates cellular function and specifies genes that are turned on or off in different status of cells. As transcriptome

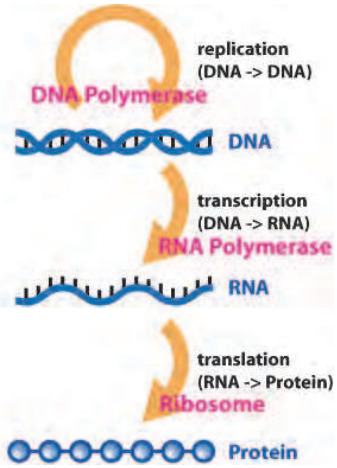


Fig. 1. Central dogma of molecular biology.

represents small percentage of the genome and much more complexity, information carried in the transcriptome has no substantial direct relation to information from the genome (Frith et al., 2005; Tsiridis & Giannoudis, 2006). Proteomics assists to comprehensively characterize quantity, structure and activity of the entire complement of expressed proteins (proteome) in large scale within a cell or tissue at a particular time. In addition, this approach provides the studies of protein-protein interactions and detailed understanding of the complex responses of a living system to stimuli (Beranova-Giorgianni, 2003; Hirsch et al., 2004). However, genome is relatively static while the dynamic proteome changes constantly in response to environmental signals. This is due to many reasons, including different amino acid sequences, alternative splicing of mRNAs and post-translational protein modifications that often give rise to more than one protein per a single gene. Proteomics, therefore, produce large high dimensional datasets that require powerful tools to handle and analyze the data effectively (Hegde et al., 2003; Tsiridis & Giannoudis, 2006).

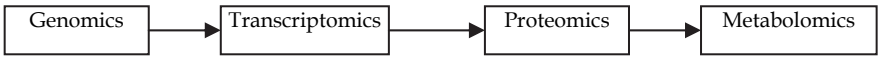


Fig. 2. Biochemical levels of information in gene expression study.

Metabolomics is the study of the entire set of metabolites, low-molecular-weight organic compounds, in the cell (metabolome) assisting the inference of biological functioning (Schaub et al., 2009; van der Werf et al., 2005). It involves the large-scale analysis of changes in metabolites in response to environmental or cellular changes. Metabolomics aims at quantifying every single metabolite and is one step further than metabolic profiling that only elucidates an inventory of the metabolites present in the cell (van der Werf et al., 2005). The transcriptome, proteome and metabolome can change considerably depending on various environmental conditions and directly represent the status of cellular physiology. Hence, these sources are so beneficial in understanding biological performance. Although

omics technologies have been advancing over the years, they still contain some drawbacks. Proteomics and metabolomics offer the holistic and complementary insights into cells because transcriptomics cannot always reflect corresponding protein or metabolite profiling. They are, however, limited in lack of standardized methodologies and poor reproducibility (Pinet, 2009). This is partly due to the heterogeneous characteristics of the compounds identified. In proteomic analysis, the wide range of proteins makes it difficult to design standard protocols for identification of compounds. Likewise, metabolomics suffers from the diverse collection of chemical properties of different metabolites (Karakach et al., 2010). Scientists are not well trained to cope with the large data and limited availability of commercial metabolites. Despite these limitations, going from one biochemical level to the next, information is acquired or lost by regulatory events such as post-transcriptional and post-translational modifications that occur between these levels. Metabolomics, however, is valuable as it is the closest to the function of a cell i.e. the phenotype (Tsiridis & Giannoudis, 2006; van der Werf et al., 2005; Zhang et al., 2010).

Compared to proteomics and metabolomics, transcriptomics is a more robust, large-scale, moderate cost technology of simultaneously measuring thousands of mRNA levels, but most transcriptomic analysis platforms are not routinely set up to systematically detect changes in spliced species as nearly 50% of human genes may undergo alternative splicing (Hegde et al., 2003). Also in some cases mRNA levels are a reasonable proxy for protein abundance, allowing one to make a rational inference regarding the level of protein expression based on the levels of mRNA expression. But, sometimes some caution seems necessary where protein expression is controlled post-transcriptionally by other factors. Since mRNA molecules are relatively more homogeneous than metabolites and proteins, and capture methods based on complementary DNA have been developing, the field of transcriptomics has been more associated with gene expression studies using microarray technology (Karakach et al., 2010).

In conclusion, the study of omics sciences plays an important role in understanding different perspectives of cells to gain knowledge about cellular pathways, mechanisms and functions that eventually make up an expression cycle. The transcriptome is more crucial in expression measurements while the proteome and the metabolome together assist in determining the functionality of expressed genes (van der Werf et al., 2005). Thus, effective integration of omics datasets provides a broader view of systematic changes in expression levels. However, this integration still remains one of the challenges of systems biology and functional genomics.

2. Methods to quantifying mRNA level

Composition and differences of various transcriptomes is specified through mRNA level measurements. There are a number of methods to quantitatively determine this factor including northern blotting, reverse transcriptase polymerase chain reaction (RT-PCR) and DNA microarray. These techniques are briefly discussed in the following.

Northern blotting is a standard method for studying the expression profile of specific genes in mRNA level. It can detect alternatively spliced transcripts and transcript size. In Northern blot analysis, mRNAs are extracted from sample then separated based on size in gel electrophoresis (targets). Probes are a complementary sequence to all or a part of interested mRNAs. Afterwards, targets are transferred to a solid support from an agarose gel to hybridize with radio-labeled probes. If the probe has complemented sequence to an mRNA,

then it will bind to the location of that mRNA on the gel (Trayhuru, 1996). Degree of radiation gives an indication of expression level in gene of interest. This method is a semi-quantitative detection because the amount of radioactivity depends to some extent on the amount of the probe which in turn depends on the amount of mRNA in the sample (Perdew et al., 2006; Trayhuru, 1996). Northern blotting is an appropriate assay especially for laboratories which are limited with the lack of specialized equipments and expertise in molecular biology (Trayhuru, 1996). One of the pitfalls in northern blotting is often sample degradation through the action of RNases, which can be overcome by proper sterilization of glassware and reagents and the employment of RNase inhibitors. Also the used chemicals can be a risk to the researcher.

Polymerase chain reaction (PCR) is an enzymatic assay which produces large amount of a specific DNA sequence from even a small and complex mixture. Also reverse transcriptase (RT)-PCR is a rapid and flexible approach for mRNA examination and quantification. In this method, first the mRNA must be converted to a double-stranded molecule by using the enzyme reverse transcriptase (Perdew et al., 2006). Since small variations of amplification efficiencies between samples can result in significant differences in product yield, quantification of mRNA by RT-PCR is difficult, therefore modified methods have been developed such as quantitative competitive (QC)-PCR, relative RT-PCR and real time RT-PCR. The QC-PCR measures the absolute level of a particular mRNA sequence in a biological sample. It relies on using dilutions of a synthetic RNA called competitors. These competitors compete with the target cDNA for co-amplification. Since competitor molecule differs in size from the target one, the two PCR products can be separated by gel electrophoresis. Although this method provides an accurate result, the design and construction of competitor for each gene is technically complicated. Validation of the results of the technique is also labor intensive (Breljak et al., 2005). Relative or semi-quantitative RT-PCR measures mRNA level using a co-amplified internal control with the gene of interest. Results are reported as ratios of the gene-specific signal to the internal control signal. Although this method requires only common laboratory equipment, it suffers from poor dynamic range of the quantification and being time consuming as well as labor intensive (Lipshutz et al., 1999). A novel approach of PCR, real-time PCR, is the combination of the best features of both relative and competitive PCR. It is much faster, higher throughput and less labor-intensive assay than current quantitative PCR. Furthermore, it combines amplification and detection in one step. Unlike other quantitative PCR methods, real-time PCR does not need preventing carryover contamination of PCR products and PCR processing such as electrophoresis. This approach is carried out through dual labeled fluorogenic probes. The amount of fluorescence emitted is directly proportional to the amount of product produced in each PCR cycle (Breljak et al., 2005; Heid et al., 1996). In spite of outstanding advances performed in the area of real-time RT-PCR, competitive and semi-quantitative RT-PCR may still utilize for relative mRNA quantification especially for small number of samples (Breljak et al., 2005). RT-PCR is much more sensitive, rapid with a large dynamic range of quantification. It requires specialized expensive equipment and ingredient which may be restrictive to some researchers (Perdew et al., 2006; Trayhuru, 1996). Since undesirable primer–primer interactions may happen, RT-PCR is limited in the number of genes to be analyzed each time. Some sources of variation such as template concentration and amplification efficiency make difficult quantification based on RT-PCR (Trayhuru, 1996).

Microarray experiment is an emerging technique as such, based on determining expression levels of thousands of genes simultaneously. This approach can be considered as a massive

parallel Northern blotting. DNA microarray gives a holistic picture of gene expression within the cell or the sample in different environmental conditions at a specific time (Tarca et al., 2006). Practically, such high throughput method utilizes an inert surface containing a certain number of spots. Each spot contains a single species of a nucleic acid representing the genes of interest (probe). Hybridization between labeled biological sample (target) and probes creates a signal that represents the level of expression of a gene in a biological sample. The microarrays have become important because they are easier to use and do not require large-scale DNA sequencing. However these studies are still limited by lack of universally accepted standards for data collection, analysis and validation (Bilban et al., 2002; Russo et al., 2003). Microarrays are quite user friendly and usually consistent with results produced from northern blotting and PCR; although, these approaches can measure small levels in gene expression that microarrays cannot. The main advantage of microarrays is visualizing thousands of genes at a time, while other methods are usually quantifying one or a small number of genes (Bilban et al., 2002; Trayhuru, 1996).
Some features of the above mentioned methods have been summarized in Table 1. Regarding the advantages and limitations of each technique, it is concluded that even though the all methods can measure mRNA levels, they differ on their special attributes.

Method	**Pros**	**Cons**
Northern blotting	-Detecting alternatively spliced transcripts -Detecting transcript size -Straightforward -Inexpensive	-Insensitive -RNase contamination -Low throughput -Use of hazardous reagents -Low quality quantification -Needs large quantity of RNA -High background on solid supports
RT-PCR	-High sensitive -Rapid -Wide dynamic range (Real time RT-PCR) -Sensitive and robust -Nearly high throughput -needs small sample	-Expensive equipment -needs expertise in molecular biology -The ease with which minor contamination may yield false-positive results -Post-PCR manipulation except real time PCR
Microarray	-The parallel quantification of thousands of genes from multiple samples -Rapid -Robust -Convenient for directed and focussed studies -Cost effective -Easy to use -do not require large-scale DNA sequencing	-needs verification -Difficult to correlate with absolute transcript number -Sensitive to alternative splicing -many factors can affect microarray result: • chip type • sample preparation • data analysis -Requires bioinformatics for data analysis. -Lack of standard preprocessing methods -Low sensitivity of microarray detection technology

Table 1. Features of conventional techniques to quantifying mRNA level.

Therefore, the selection of methods is performed based on required characteristics in experimental design. It should be noted that although traditional techniques of gene

expression analysis provide valuable biological insights into the living cells, they are probably limited in some ways such as scale, economy, and sensitivity. As a result, compared to the other commonly used techniques, quantification based on microarray is remarkable because of high throughput and cost effective features.

3. Microarray technology

Microarray technology has become one of the most commonly used high-throughput techniques to query a large variety of biological issues. It enables the simultaneous analysis of thousands of parameters within one single experiment. Such miniaturized binding technology is typically divided into DNA, protein, tissue, cellular and chemical compound microarrays (Templin et al., 2002). Some of the arrays such as protein array and tissue array will be described in detail with a special emphasis on DNA arrays.

Protein microarrays assist in characterizing of thousands of proteins in a parallel format. Proteome chips afford researchers a way to address true level of gene function by studying the pair-wise interactions such as protein-protein, protein-DNA, protein-lipid, protein-drug, protein-receptor and antigen-antibody (Hall et al., 2007). In this technique, probes such as aptamers, engineered antibody fragments, affibodies, full-length proteins or protein domains can be spotted on a microscope slide. The array is then probed with a target solution and binding detected using the analytical approaches. Antibody microarray is the most powerful type of protein microarray. Figure 3 shows the detailed view of the steps taken to carry out antibody microarray experiment (Angenendt, 2005). Tissue microarray (TMA) technology was developed in order to evaluate the difference of molecular targets (in the DNA, RNA or protein level) in several thousands of tissue samples at the same time (Kononen et al., 1998; Singh & Sau, 2010). TMA is constructed from paraffin embedded material, frozen tissue, paraffin embedded cell lines or cell blocks (Parsons & Grabsch, 2009). Totally, TMA is made of tissue core samples taken with a precision punching instrument from donor paraffin blocks. These cores of tissue are arrayed into an empty recipient block, TMA block (Figure 4). Afterwards, the TMA block is sectioned by using a device called microtome. The sections are placed on a microscope slide and then analyzed by any standard histological procedure. From a TMA block, approximately 200–300 5-µm sections can be cut and used at independent tests (Parsons & Grabsch, 2009).

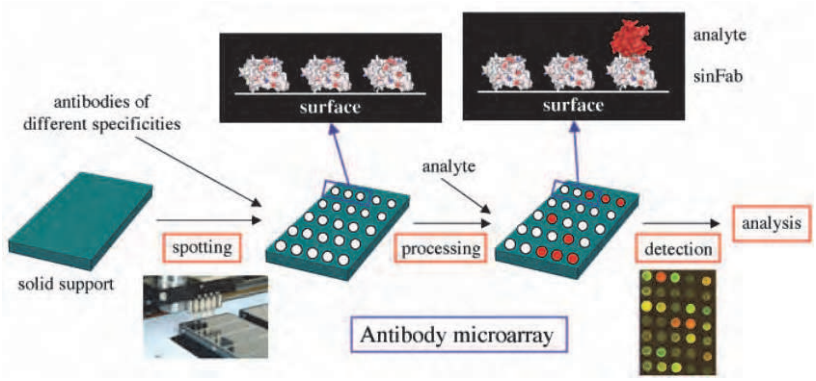


Fig. 3. Schematic diagram of an antibody microarray technology.

DNA microarray is the most popular type of microarray technology that uses nucleic acid–nucleic acid interactions. It allows measuring the amount of mRNA transcripts for thousands of genes in different combinations of sample derived from normal and diseased or treated and non-treated tissues, time courses of treated cells and stages of cell differentiation or development (Karakach et al., 2010). It has been proved that DNA microarrays are extremely valuable in studying of expression profiling, sequence identification and location of transcription factor binding sites (Hall et al., 2007).

4. The DNA microarray experiment

DNA microarrays are currently manufactured using two main techniques: in-situ synthesis and deposition of pre-synthesized probes (spotted arrays). There are various platforms or types of DNA microarrays that are commercially available. Figure 5 summarizes some of these platforms based on different fabrication methods. The two most commonly micarrays are the affymetrix oligonucleotide chips (Lockhart et al., 1996) and spotted cDNA arrays (Schena et al., 1995). Experimental steps and construction process of these arrays are discussed in this section.

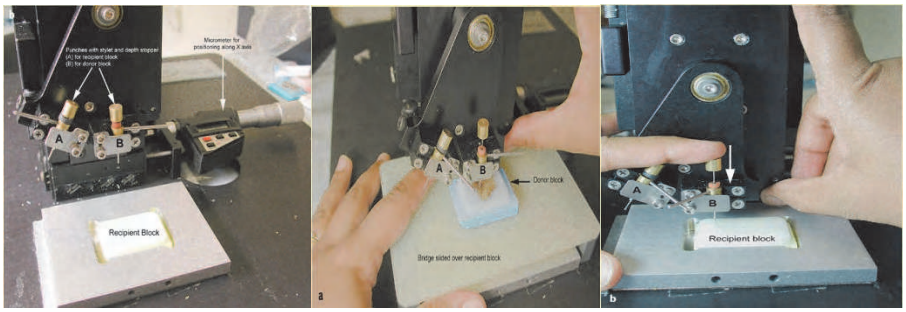


Fig. 4. **a.** Tissue arrayer instrument, **b.** Extraction of the donor core, **c.** Insertion into recipient block (Gulmann & O´Grady, 2003).

4.1 Affymetrix Gene Chips

4.1.1 Fabrication of Affymetrix Gene Chip

The in situ synthesis of oligonucleotides (Affymetrix Gene Chip) can be achieved using a photolithographic method (Fodor et al., 1991). This approach involves adding of adenine (A), cytosine (C), guanine (G) and thymine (T) nucleotides step by step through a set of designed masks. In fabrication process, solid substrate, usually quartz wafer, is washed to provide uniform hydroxylation of the surface and is then placed in a silane bath. Silane molecules are capable to directly react with the hydroxyl groups of the quartz. Therefore starting points are formed to synthesize new oligonucleotide strands. In the following steps,

synthetic linkers are attached to silanes and coated with a light-sensitive protecting group (Figure 6). The first mask is placed over the surface which then exposed to the light source. Masks selectively direct light toward specific areas on the substrate. Afterwards linker molecules are activated at the unprotected position. Next, the first of a series of nucleotides, linked to the light-sensitive agent, is incubated on the surface. Thus, the nucleotides are chemically coupled to the activated sites. Photo labile agents block further nucleotide binding to linkers until light subsequently activates them through a new mask. This chemical cycle is repeated until several hundred thousands of oligonucleotides (probes) with desired lengths and sequences are synthesized at each of sites on the surface of the chip (Lipshutz et al., 1999).

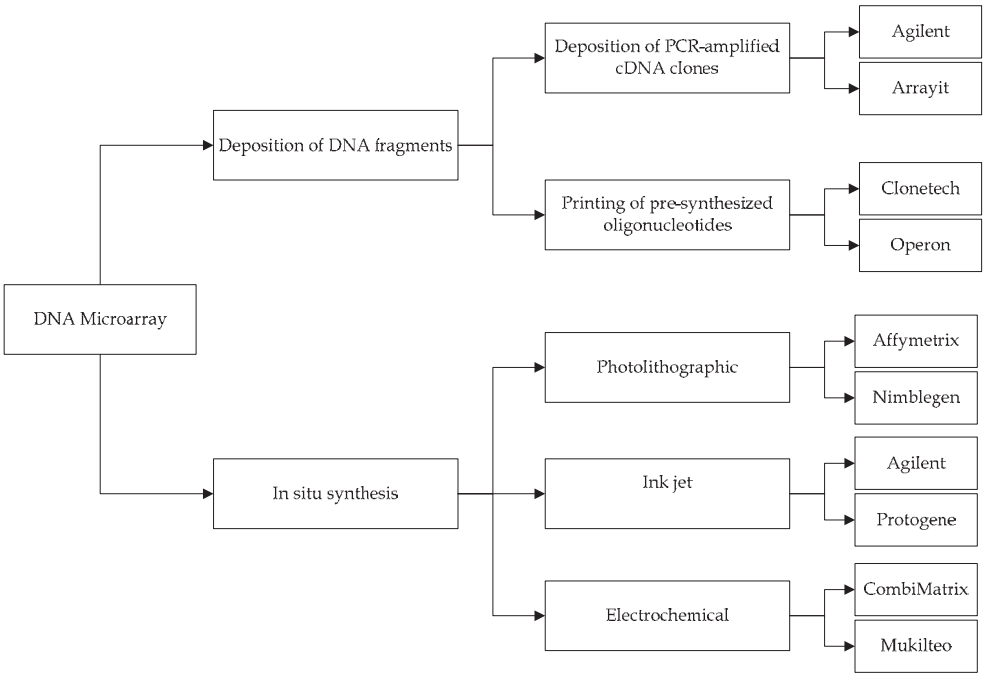


Fig. 5. Different microarray platforms and their fabrication methods.

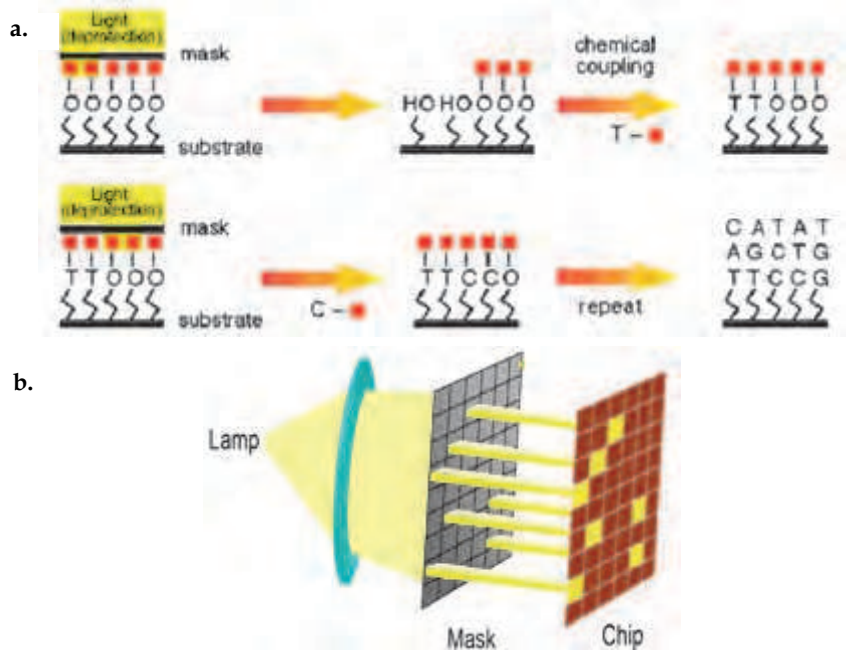


Fig. 6. **a.** Schematic overview on photolithographic fabrication of Gene Chip. **b.** Drawing of the lamp, mask and chip (Lipshutz et al., 1999).

4.1.2 Experiment of Affymetrix Gene Chip

Microarrays use various approaches based on uniqueness and composition design rules to select the 25-nucleotide-length (25mer) probes (Lipshutz et al., 1999). They utilize the Perfect Match/Mismatch probe strategy (Figure 7). Each gene sequence (or expressed sequence tag (EST)) is represented by typically 12-20 different probe pairs. The collection of probes for each gene is referred to as a probeset. Each pair includes a perfect match (PM) oligonucleotide and a mismatch (MM) oligonucleotide. A PM probe is perfectly complementary to the gene sequence of interest (Barrett & Kawasaki, 2003; Lipshutz et al., 1999) while the MM probe has a one-base mismatch in the central base position (the 13th base). The MM probe is used as an internal control to estimate the signal of any non-specific hybridization or contaminating fluorescence within measurement (Lipshutz et al., 1999; Tarca et al., 2006). These probesets are made on array through in situ synthesis and the microarray will be ready to carry out the experiment.

The basic steps in this single-dye experiment are as follows. Total RNA (or mRNA) is extracted from the biological sample, called target. The total RNA is then reversed transcribed to generate double-stranded cDNA. Then, biotin-labeled cRNA is produced from cDNA using *in vitro* transcription. Next, biotin-labeled cRNA is fragmented into smaller segments and hybridized on the array. After a series of washing for removing non-hybridized material, the array is incubated with appropriate fluorescent dyes linked to the biotins on the cRNAs. The array is placed in a scanner and emission of fluorescent staining agent is quantified (Schadt et al., 2001). Measurement of the florescent agent intensity provides an estimate of the level of mRNA within each gene of interest on the chip.

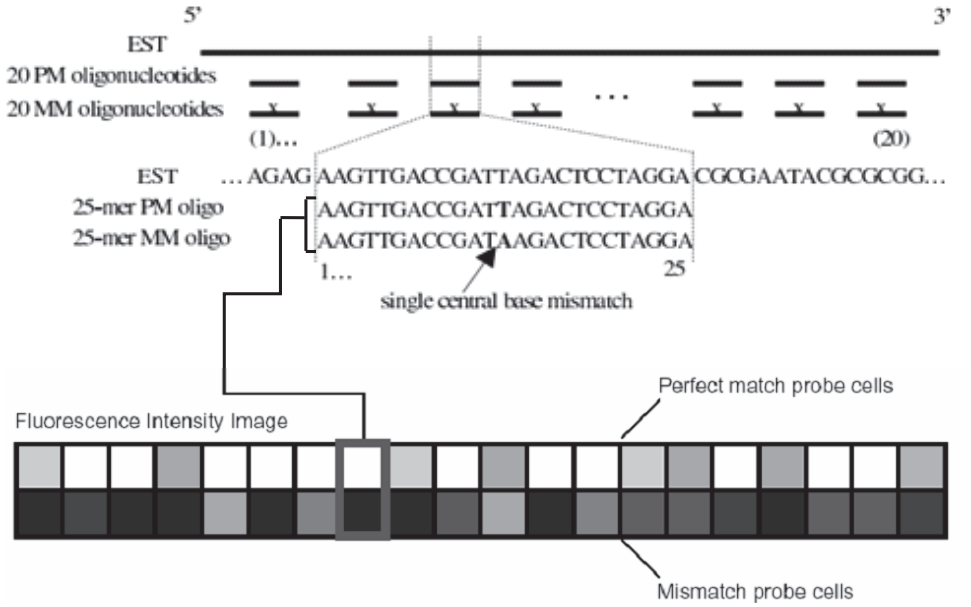


Fig. 7. Design of Affymetrix Gene Chip technology.

4.2 Spotted cDNA array

In spotted technology, probe sequences are synthesized separated from the array. In this technique, the probes correspond to specific genes, expressed sequence tag i.e. a stable cDNA fragment, or cDNAs from libraries of interest (Bilban et al., 2002). If the quantity of available probes is limiting, PCR amplification is performed to make sufficient probes. The PCR products are then analyzed by gel electrophoresis, quantified and eventually spotted using a robotic printing on the microarray surface. Probes are immobilized or attached at fixed locations onto the slides electrostatically, through cross-linking by heat or ultraviolet irradiation and via amines or other active groups on modified slides (Barrett & Kawasaki, 2003). Therefore, the location of each spot on the array can also assist researchers to identify a desired gene sequence.

Since the cDNA probes are double stranded the array is then heated (or alkali treated) until the DNA is separated and hybridized to its complementary strand. In this two color approach there are two samples, a test sample and a reference sample. In order to prepare the targets, cDNAs are synthesized using reverse transcript of mRNAs in the samples. Targets are labeled through variety of labeling methods. The most common approach is labeling with a red and green fluorescent dye, called Cy5 and Cy3, respectively. The labeled targets are combined and deposited on the array. If a gene is present in one or both samples, it will bind to its complementary probe according to the complementary base pairing property of nucleic acids. After washing the array to remove the non-hybridized targets, a

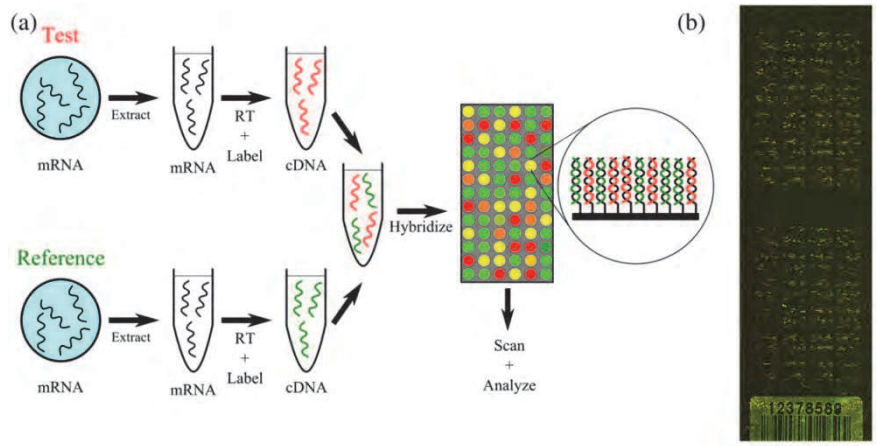


Fig. 8. **a.** Spotted cDNA microarray experiment consists of the following: preparation of target genes, labeling of the targets, hybridizing, scanning, **b.** Scanned image of a cDNA microarray (Karakach et al., 2010).

laser scanner assists to quantify the emission from Cy3 and Cy5 dyes (Figure 8). A green spot indicates that the corresponding gene is more strongly expressed in the reference sample compared to the test sample, while a red spot shows the opposite. A yellow spot reveals a gene in both samples is expressed in the same levels while a black spot shows that the gene is not express in either sample. The fluorescent spot intensity directly gives an estimate of the amount of mRNA concentration at specific condition and cell type. Details of each experimental step have been reviewed elsewhere (Bilban et al., 2002; Karakach et al., 2010).
Characteristics of the two discussed microarrays are summarized in Table 2. This table provides a comparative view of cDNA and affymetrix oligonucleotide microarrays. Selection of desired platform is based on biological question which determines aims of the experiment. In the remainder of the chapter, we will focus mainly on the spotted cDNA microarray because of limited space even though some of the discussions can be generalized to other platforms such as affymetrix oligonucleotide array.

Platforms	Pros	Cons
cDNA microarray	-Low cost -More flexible -Easier to customize and analyse -Wide availability -No sequence information required	-More variability in system -Cross-hybridization -Intensive labour requirement -Frequent failure of array or individual spots
Affymetrix oligonucleotide chip	-More reliable -Easier to use - Speed and specific -Reproducible -Low failure rate -Ability to differentiate between splice variants -Detection of mutant sequences	-More difficult to analyze -Expensive array and reagents -Exact sequence information necessary -Lack of flexibility

Table 2. A comparison between cDNA and oligonucleotides arrays.

5. Image processing

In the microarray experiment, as mentioned earlier, hybridized slides are inserted into a scanner to prepare fluorescent images arranged into a matrix of spots. The next step is processing these images to quantify level of gene expression based on the intensity of each spot and obtain background estimates and quality measures (Istepanian, 2003, Yang et al., 2002a). Accuracy of analysis in this phase has remarkable effect on downstream analyses such as clustering, classification or the identification of differentially expressed genes (Yang et al., 2001). Generally, laser scanning confocal microscopy acquires fluorescent signals emitted by fluorescently labelled targets on the array. Scanners detect and record the signals using photomultiplier tubes (PMT) or charge coupled device (CCD) cameras (Figure 9). These signals are stored in two 16-bit tiff (tagged image file format) images for further analysis (Karakach et al., 2010). Images contain information about each fluorescent dye, typically Cy3 and Cy5. Most of the softwares create a composite image by overlaying the two images corresponding to the individual channels for visualizing different status of genes.

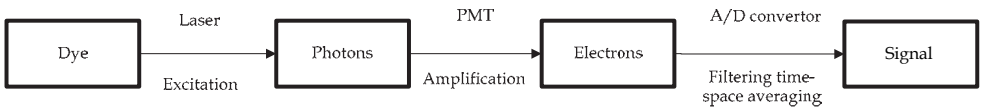


Fig. 9. Photons emit from a fluorescent samples through excitation, enter into a PMT, resulting in the release of electrons. Analogue signals from PMT are converted into digital signals by an analog-to-digital (A/D) converter. (More details in Schena, 2003).

Image processing techniques can be divided into the following steps: gridding, segmentation, quantification and spot quality assessment (Istepanian, 2003; Yang et al., 2001). Over the last years, a number of commercial and free softwares have been developed which can perform each step of image processing in a particular approach. These steps are discussed in more details in following sections.

5.1 Image gridding

The basic layout of a microarray image is determined by the robotic printing devices (arrayers) as it is known in advance (Figure 10). The arrayer itself consists of a series of pins arranged as a print tip (also referred to as sub-arrays or grids). The pins pick up reagents and deposit them on the array. Hence, the spots on array are organized in several print tips that each one is composed of spots printed with one pin (Karakach et al., 2010). Gridding (addressing) is the process of finding location of the spots on images. This is carried out using a simple model based on layout of scanned image. In order to enhance the reliability, manual intervention is utilized in association with automatic procedures (semi-automatic) (Yang et al., 2001). However, this can probably make the process very time consuming and introduce user bias and loss of consistency. At first, the user manually specifies the positions of spots on the image. Then, a suitable grid pattern is automatically provided from the indicated positions (Gjerstad et al., 2009; Yang et al., 2002a).

5.2 Image segmentation

Grid spots are partitioned into foreground (within printed spot) and background regions through a process referred to as segmentation. Foreground pixels represent the true signal

while pixels in the background area correspond to signals not due to hybridization of target molecules (noise or artifacts) (Yang et al., 2001; Yang et al., 2002a). The most common segmentation methods are classified based on whether they place restrictions on the spot geometry. Fixed circle and adaptive circle segmentation methods assume circular spot shapes, while the histogram and adaptive shape segmentation approaches apply no restrictions on the shapes of the spots in the estimation of the spot masks. Each segmentation method generates a spot mask which consists of a set of foreground pixels for each spot (Karakach et al., 2010; Yang et al., 2001). The simplest method is fixed circle that assigns a circle with constant diameter to all spots. It characterizes the pixels within the circle as true signal and the pixels out of the circle as background pixels. Adaptive circles segmentation estimates the circles' diameters separately for each spot (Yang et al., 2002a; Yang et al., 2001). Since this approach requires the user to adjust spot sizes, it can be time-consuming for an array with thousands of spots. Furthermore it will be hard to distinguish a transition between the foreground and background if the signal strength is low (Yang et al., 2002a). Although most spot shapes are expected to be circular, in practice non-commercial arrayers rarely print the perfect circular shapes of spots resulting in poor estimates of fluorescent intensities for hybridized targets. Thus, novel approaches known as "adaptive shape segmentation" methods has been developed which try to find the best shape of the spot (Yang et al., 2001; Yang et al., 2002a). These methods are commonly based on the watershed transform (Beucher & Meyer, 1993) and the seeded region growing algorithm (SRG) (Adams & Bischof, 1994) which successfully detect different sizes and shapes of segmented spots (Karakach et al., 2010).

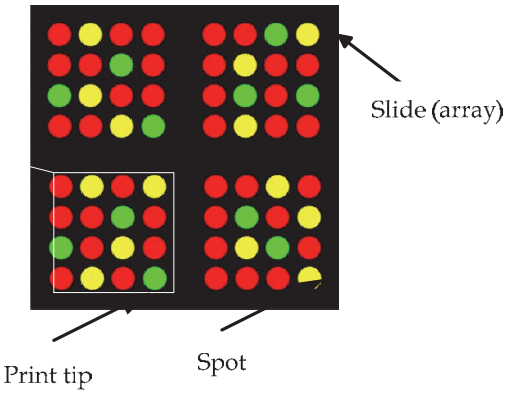


Fig. 10. Common structure of a cDNA microarray slide.

The most widely used method for segmenting spots, without restricting to particular shapes, is the histogram-based technique (Yang et al., 2001). It defines a target spot mask whose size is larger than any other spot. Foreground and background intensities of each spot are estimated from histogram of the pixels within this mask in various ways (Yang et al., 2002a). This technique directly quantifies values and needs no spot quantification stage. The discussed segmentation methods are implemented in most softwares to perform primary level processing of microarray images (Table 3) (Yang et al., 2001).

Segmentation method	Software Implementing Method
Fixed Circle	ScanAnalyze, GenePix, QuantArray
Adaptive Circle	QuantArray, GenePix, Dapple, Agilent Feature Extraction
Histogram	ImaGene, QuantArray and DeArray
Adaptive shape	Spot

Table 3. Different segmentation methods in different image processing softwares.

5.3 Image quantification

After detecting the location of spot and classifying pixels, it is necessary to compute red and green foreground intensities as well as red and green background fluorescent values for each spot on the array (Yang et al., 2002a).

5.3.1 Foreground quantification

In fact, the aim of the spot quantification is estimating a quantitative measure which is a combination of pixel intensity values (Yang et al., 2001). There are different statistics to compute this measure. Simple sum of pixel intensities is not a good statistic because it dependent on the size of the spot. Thus, values obtained from spots with different densities cannot be compared directly. Most microarray imaging softwares estimate the foreground intensity as the mean or median of pixel values within the segmented spot mask (Yang et al., 2002a). The median value is more robust to possible outlier pixels; hence it is preferable over the mean. Also interquartile range (IQR) (i.e., the difference between the 25th and 75th percentiles) of foreground may be computed for each channel as pixel variation estimation.

5.3.2 Background quantification

Background estimation is generally considered necessary for the aim of performing background correction (Yang et al., 2002a). Background estimation methods can be classified into four categories: local, morphological, constant and no adjustment background (Yang et al., 2001). In the first category, background intensities are computed by focusing on small regions around the spot mask. Different softwares utilize variety of shapes for these areas such as square, diamond-shape (referred to as the valley) and circles with different diameters. Usually, the background measure is estimated by the median of pixel values within these specific regions; however, it is possible to calculate mean, standard deviation, and interquartile range of pixels (Yang et al., 2001). Also, there are two types of morphological filters. The first one corresponds to a non-linear filter called morphological opening that is obtained by applying a form of local minimum filter (an erosion process) followed by a local maximum filter (a dilation process) with the same window for each image (for more details, see (Soli, 1999)). The second one corresponds to a combination of a closing followed by an opening that removes small dark regions as a better estimate (Wang, 2007; Yang et al., 2001). Constant background is a global method which estimates the mean or median intensity of the whole image background as a constant background for all spots. The fourth option is possibility of no background correction (Yang et al., 2001).

5.4 Spot quality assessment

The quality assessment step facilitates to diagnose possible quality problems or even mistakes that occurred during microarray fabrication and experiment. If this step does not report any serious irregularities, it will allow performing the following preprocessing steps.

After calculation of foreground and background intensities, quality measures are estimated to assess spot quality and reliability. These include variability of pixel values within each spot, spot area, a circularity measure, relative signal to background intensity (signal to noise ratio) and flag (Yang et al., 2001; Yang et al., 2002a). Each quality measure can be interpreted as follows. In most arrays the spots should be of the same size, thus very large or very small spots may be an indication of problems (Wang, 2007). Eliminating or marking of poor-quality and low-intensity spots is called flagging. This is zero if the spot is good, but will take different values if the spot has problems. Different image processing software uses different flag values for different problems, but the typical flagged spots are:

1. Bad spot: The pixel standard deviation is considerably higher than the pixel mean.
2. Dark spot: The signal of the spot is very weak.
3. Negative spot: The signal of the spot is less than the background value.
4. Manually flagged spot: The user has flagged the spot using the image processing software (Stekel, 2003).

Performing the four steps of image processing, quantitative parameters are generated in an output file of the software as shown in Table 4. These measures exhibit some of the location information, foreground and background quantifications and quality measures in a microarray experiment.

6. Preprocessing of cDNA microarray data

Prior to identification of DEGs, the data collected from image processing step needs to be preprocessed. This important step in microarray data analysis removes non-biological variations, makes data more meaningful, transforms data into an appropriate scale for analysis and enhances the quality of subsequent analysis. There are a number of approaches for preprocessing such as background correction, logarithm transformation and normalization of microarray data. It should be mentioned that spot quality assessment (section 5.4) in image processing could also be considered as a preprocessing step.

6.1 Background correction

Background correction is a necessary step in preprocessing of cDNA microarray data since the quantified fluorescence intensity of a spot contains background noise which does not reflect the true hybridization of the target to the probe. Background noise results from several sources such as non-specific hybridization of labeled target to the array surface, autofluorescence from the array surface or detection instrument, spatial heterogeneity across the arrays (Ritchie et al., 2007; Tarca et al., 2006). For the purpose of background correction, it is conventionally assumed that the background signals are additive to the foreground signals (Ritchie et al., 2007). Also, the standard approach for correction is subtracting an estimate of the local background intensity from the foreground intensity. Despite spread implementation of this approach in different software packages, it may cause problems. It generates negative corrected intensities resulting in missing log ratios, if the background intensity is larger than the foreground intensity. Even when there is no missing, it results highly variable log-ratios for low intensity spots (Kooperberg et al., 2002). Also it may cause some difficulties in the identification of differentially expressed genes (Yang et al., 2001). To overcome aforementioned limitations, alternative approaches have been proposed such as subtractive correction using an estimate of the global instead of the local background and

Index	grid.r	grid.c	spot.r	spot.c	Area	Gmean	Gmedian	GIQR	bgGmean
1	1	1	1	1	95	22028.26	23219	0.564843	372.6964
2	1	1	1	2	85	25613.2	20827	0.672128	928.8974
3	1	1	1	3	77	22652.39	17498	0.939413	1371.86
4	1	1	1	4	21	8929.286	5270	1.975485	250.5417
5	1	1	1	5	21	8746.476	7396	2.518724	262.0417
6	1	1	1	6	112	37010.08	41539	0.943238	499.1722

bgGmed	bgGSD	Valley	morphG	morphG.erode	morphG.close.open	Logratio	Perimeter	Circularity	Badspot
307	0.252131	306	182	153	289	-0.17171	40	0.746128	0
299	0.390198	280	171	153	278	-0.16341	36	0.824183	0
339	0.820078	275	153	136	278	-0.15408	34	0.837033	0
244	0.270411	258	153	132	271	0.80675	16	1.030835	0
235	0.275412	244	153	132	216	-0.10662	16	1.030835	0
304	0.740031	244	139	120	224	-0.44679	44	0.72698	0
381	1.05041	243	138	120	224	-0.21073	34	0.739198	0

Table 4. Partial output file from Spot software for green (Cy3) channel: Location information (spot index, grid row, grid column, spot row and spot column), Area: the number of foreground pixels for each spot, Gmean: the average of foreground pixel values, Gmedian: the median of foreground pixel values, GIQR: the interquartile range of foreground pixel, bgGmean: the average of background pixel values, bgGmed: the median of background pixel values, bgGSD: the standard deviation of background pixel values , valleyG: the background intensity estimate from the local background valley method, morphG: background estimate using morphological opening, morphG.erode: green background estimate using morphological erosion, morphG.close.open: green background estimate using morphological closing-opening, Logratio: the log-ratio for each spot is calculated as $\frac{\log_2(Rmean - bgmedR)}{\log_2(Gmean - bgmedG)}$, Circularity: Shape of spot defined as $\frac{4 \times \pi \times Area}{primeter^2}$, Badspot: If the spot has problem, it equals to 1, otherwise 0.

morphological opening filters which provide less variable log ratios (Ritchie et al., 2007; Yang et al., 2001). Some methods utilize statistical models, other than subtraction, to adjust the background estimate. A simpler background correction method was proposed to avoid negative corrected values. This model adjusts the foreground intensities by subtracting the background when the difference between the foreground and background is larger than a threshold value. However, when the difference is less than the threshold, subtraction is replaced by a smooth monotonic function. Kooperberg et al., 2002 proposed an empirical bayes model to correct background noise. A remarkable feature of this method is only the use of the mean, median and standard deviation statistics for each spot that are provided through the scanning software. In other methods, the models based on variance stabilizing transformations were proposed for incorporating additive components which prevent negative intensities. The Models use an arcsinh function instead of the logarithm transformation of the data. Also background correction and normalization are simultaneously performed on all the arrays together (Kooperberg et al., 2002; Ritchie et al., 2007). It is notable that no background correction has been recommended. Sometimes, local

background methods show greater variability around the low intensity spots rather than no background adjustment (Yang et al., 2001).

6.2 Logarithm transformation

Before normalization, a logarithmic transformation is often performed on microarray data. This transformation is successful at reducing some of the variations, and makes the multiplicative noise of the data additive. Also data is transformed into a symmetrical and normal data distributed around zero through taking log transformation. This means that up- and down-expressed genes are treated in identical way (Quackenbush, 2002). However, the log transformed ratios limit subsequent analyses and the amount of information gained from the data (Zhao et al., 2007). The ratios do not provide information about the absolute expression levels. Also, the use of the ratios remarkably dependents on the choice of the reference sample, which is uncharacterized and not accurately reproduced. This will make it difficult to compare between data sets that use different reference samples (Zhao et al., 2007).

6.3 Normalization

There is variety of variations from the beginning of the experimental process through generation of raw data in microarray experiment. Two sources of variations are biological variations and procedural variations. Biological variations are the consequence of environmental changes or biological differences of the studied genes on the array. These are desired variations and represent the true changes in expression cycle. Procedural variations can be attributed to many sources such as microarray fabrication, mRNA preparation, reverse transcription, labelling, amplification, pin geometry, fluctuations in target volumes, target fixation, hybridization parameters, overshining, and image analysis. Detail description of each variation source is presented elsewhere (Schuchhardt et al., 2000; Yang et al., 2002b). Procedural variations can be removed (or minimized) using statistical approaches, so that biological variations are more accurately detected. The processes and transformations for the purpose of adjusting data are referred to as normalization. Hence, normalization is a crucial step in microarray data preprocessing, since data interpretation and identification of DEGs depends on the choice of normalization method (Yang et al., 2002b).

Different biases arise from variations in the microarray data. The most common is dye bias i.e. imbalance between the two channels due to differences between physical properties of dyes and detection efficiencies between the fluorescent dyes. Other biases such as print tip bias and spatial bias may arise from variation between spatial positions on the array due to differences between the print-tips on the arrayer (Smyth & Speed, 2003). In order to remove biases, numerous normalization approaches have been proposed. These algorithms can be applied either globally to an entire data set or locally to a subset of the data. For cDNA spotted microarray, local normalization is often applied to each print tip (Quackenbush, 2002). Normalization methods can be divided into two main categories: within-array normalizations and between-array normalizations. Within-array normalization has to be performed to adjust procedural variations for each single microarray. Some of more common approaches are as follows.

6.3.1 Global normalization

Global normalization is the simplest and most common within-array normalization method. It assumes the red and green intensities are related by a constant factor k, namely R=kG. The

log-ratios are corrected by subtracting a constant c to get normalized values. ($\log R$, $\log G$) are background corrected red and green intensities and then:

$$\left[\log_2\left(R_i/G_i\right)\right]_{normalized} = \log_2\left(R_i/G_i\right) - c = \log_2\left(R_i/G_i\right) - \log_2(k) \quad (1)$$

The global constant c is usually estimated from the mean or median log ratios over a subset of the genes assumed to be not differentially expressed, although variety of strategies have been proposed for estimating this global constant (Quackenbush, 2002; Smyth & Speed, 2003). Global method is limited in adjusting intensity-dependent dye bias and spatial bias.

6.3.2 Intensity-dependent linear normalization

In most cases, the dye bias appears to be dependent on spot intensity linearly or nonlinearly. Linear normalization assumes the relation between M and A is linear based on model $M = \beta_0 + \beta_1 A$, where (β_0, β_1) can be estimated by least squares estimation. The most common method to visualize behavior of two channels is MA plot which uses log intensity ratios (M) and log intensity averages (A) where M and A are usually defined for each gene as $M_i = \log_2\left(R_i/G_i\right)$ and $A_i = \frac{1}{2} \times \log_2(R_i \times G_i)$

6.3.3 Intensity dependent nonlinear normalization

The most efficient and widely used nonlinear normalization approach was proposed by Yang et al., 2002b. It considers the relation between M and A as a function of A i.e. M = c(A), instead of a linear relation. The estimation of c(A) is made by using a loess (locally weighted scatter plot smoother) function to operate a local scatter plot smoothing to the MA plot. The scatter plot smoother performs local linear fits in overlapping windows on the data and then combines the regressions to produce a smooth curve. This method can be divided into three categories based on the type of the treatment performed on the data. These categories include: global loess, print tip loess, and two-dimensional loess. Global Loess (Gloess) normalization method uses the loess function to perform a local A-dependent analysis:

$$\left[\log_2\left(R_i/G_i\right)\right]_{normalized} = \log_2\left(R_i/G_i\right) - c(A) = \log_2\left(R_i/G_i\right) - \log_2(k(A)) \quad (2)$$

Where c(A) is the loess fit to the MA plot for all printed genes (Smyth & Speed, 2003).
Print tip loess (PTloess) is performed within each of the print tip groups separately as follows:

$$\left[\log_2\left(R_i/G_i\right)\right]_{normalized} = \log_2\left(R_i/G_i\right) - c_p(A) = \log_2\left(R_i/G_i\right) - \log_2(k_p(A)) \quad (3)$$

Where $c_p(A)$ is the loess fit as a function of A for the p^{th} print tip. By fitting separate loess lines for each group and correcting the intensity by its corresponding loess lines, not only the dye bias will be removed, but it can also correct the print tip bias. Two dimensional loess (twoDloess) method fits a smooth two-dimensional surface to the data which is a function of overall row position r and column position c of the spot on the array. The intensity-based

trend is assumed to be global rather than varying across the array as for print-tip loess normalization.

$$\left[\log_2(R_i/G_i)\right]_{normalized} = \log_2(R_i/G_i) - loess(r,c) \quad (4)$$

Where loess(r,c) is a loess fit calculated based on the position of the spots. The three techniques remove different biases arose from the experiment. Gloess removes the dye bias dependent to spot intensity. PTloess removes spatial bias introduced from print tips and twoDloess removes the spatial bias on the overall slide.

The above normalization methods are applied to a single microarray. But in order to be able to facilitate comparison and integration of different microarrays, it is required to remove the variability caused by using multiple microarrays. It can be performed through the following approaches. Differences between arrays may arise from differences in print quality or from differences in ambient conditions when the plates are processed (Smyth & Speed, 2003).

6.3.4 Scale normalization

This method is a simple scaling of the data on multiple arrays so that each array has identical median absolute deviation (MAD). It aims to remove scale differences in the data and assumes that the log ratios on the array follow a normal distribution with mean zero and variance $a_j^2\sigma^2$ where σ^2 is the variance of the true log ratios and a_j is the scale factor for array j with n denoting the total number of arrays (Yang et al., 2002b).

$$a_j = MAD_j \Big/ \sqrt[n]{\prod_{j=1}^{n} MAD_j} \quad (5)$$

Where MAD_j denote the median absolute deviation for array i. Then

$$MAD_j = median_j\left\{\left|M_j - median(M_j)\right|\right\} \quad (6)$$

Finally, all log ratios are scaled through dividing by the same scale factor for each array. It is notable that scale normalization can also be applied to data within a microarray locally at the print tip level.

6.3.5 Quantile normalization

Quantile normalization was initially developed for the Affymetrix single channel chip, and then extended for two color cDNA microarrays. The goal of this method is to produce the same empirical distributions of expression levels on all arrays analyzed. It relies on the assumption that the probe intensities among arrays are always exactly the same, regardless of biology or study design. Clearly the situation where all samples have equal amounts of expressed genes is the exception, not the rule, making it the rare case where quantile normalization will normalize data without introducing errors. Quantile normalization is carried out through the following steps: Suppose that we have the (log base 2 transformed) probe level expression values from p genes and n arrays in a p × n matrix $X = \{X_{1j}\}$ with i = 1,2,…,p and j = 1,2,…,n. First, each column of X separately is ranked to generate a p × n

matrix $Y = \{Y_{1j}\}$. Next, the average of each row of Y is computed and generated X_m . X_m is assigned to each column of Y to get a matrix denoted as X_{sort} . Finally, the normalized genes for each array is provided by rearranging each column of X_{sort} to have the same ordering as the corresponding column of the matrix X so that empirical distributions of the normalized genes are the same across arrays. Because the algorithm consists of only sorting and averaging operations, it runs quickly, even with large data sets (Bolstad et al., 2003). (More details in (Stafford, 2008))

All above normalization methods utilize certain critical biological and statistical assumptions about data distribution which may not be valid in practice. The main assumption is that the most genes on the array are non-differentially expressed between the two samples and the number of up-regulated genes approximately equals the number of down-regulated genes. In such cases, above-mentioned normalization methods may yield unreliable results. Xiong et al., 2008 proposed a novel statistical method based on the Generalized Procrustes Analysis (GPA) algorithm free of assumption (Xiong et al., 2008).

7. Differentially gene expression

Once the data is normalized, further analysis is necessary to obtain biologically meaningful results. In fact, the main purpose of microarray experiment is to identifying genes that are significantly differentially expressed under different biological, and/or clinical conditions. A growing number of approaches have been presented to fulfill this purpose that can be divided in three categories: marginal filters, wrappers, and embedded methods. The wrapper and embedded methods are a type of search algorithms by which subsets of genes that are useful to define a good predictor are generated. Evaluation of a specific subset of genes is provided by running a specific classification model on the subset. The filter approaches are scalable and fast methods and independent of the classification algorithm including t-tests and nonparametric tests and analysis of variance (ANOVA) (Saeys et al., 2007). We will provide a brief overview on some of the popular statistical differential expression methods. It is notable that various methods usually identify different ordered list of significant genes since each approach is based on a specific set of assumptions, and takes certain features of dataset into account.

Fold Change (FC) cutoff is one of the early approaches for DEG identification that is still widely used to rank genes in microarray assays. In this method, when ratio of two color intensities from each gene exceeds a pre-set threshold is said to be differentially expressed. Usually a threshold of twofold up- or down-regulation is considered as cutoff value in most biological studies. This method ranks genes based on the ratio of average gene expression under two different groups or conditions. Simplicity is a main reason for popularity of fold change approach. Also a major drawback is that it does not consider variance of the expression values quantified. Hence, in order to cope with this problem, it will be used in combination with other statistical methods (Tarca et al., 2006). There also exists variety of statistical tests instead of using a fold change cutoff, for a correct selection of deferentially expressed genes. A simple but popular method is the t-test and its variants (Cui & Churchill, 2003). The t-test performs according to the simple estimation of the population variance for a gene through the sample variance of its expression levels. It typically compares the difference between the mean expression levels among the two groups, considering the variability of genes in their ranking (Tarca et al., 2006). T-test depends on the type of

distribution of the gene expression data. Thus, may not properly perform when data exhibit a strong departure from the normal distribution. Also the performance of t-test will be poor when sample sizes are small, because variance estimation is more challenging (Yan et al., 2005).

The ANOVA approach is a generalization of the t-test that can be used when more than two conditions are compared. The idea underlying ANOVA is to make a model that considers the variation sources that affect measurements. Then variance of each individual variable in the model is computed using expression data (Tarca et al., 2006). In order to improve the performance of the ordinary t-test and produce more stable results, modified t-statistics are alternatively proposed. The main difference between an ordinary t-statistic and these novel statistics is that the latter estimate variability regarding to information not only from the gene tested, but also from other genes displaying a similar magnitude of expression level (Smyth, 2004). Two commonly used approaches, i.e. the modified t-statistic methods (empirical Bayes and SAM), will be described in more detail as fallows.

7.1 The empirical Bayes t-test LIMMA

This empirical Bayes t-tset has been implemented in the limma R statistical package. In this approach, gene-wise linear models are separately made to represent the design of a microarray experiment. Next, the coefficients of each linear model are estimated through the expression data. After quantification of coefficients of model and standard errors, moderate t, F and B (log-odds) statistics of differential expression are computed using empirical Bayes approach. It is equivalent to reduction of the gene-wise sample variance towards a pooled estimate producing more stable result when the number of measurements is small in experiments. Finally, genes can be ranked based on one of the chosen statistics. A more detailed derivation can be found in (Smyth, 2004).

7.2 Significance analysis of microarrays (SAM)

SAM is a statistical technique, proposed by Tusher et al., 2001. It utilizes a non-parametric statistics, since the expression data may not be normally distributed. Modified t-statistic used in this method is essentially similar to the moderated t-statistic used in limma but have no associated distributional theory. Also the empirical bayes method provides a more complex model of the gene variance. SAM assigns a score to each gene based on change in gene expression relative to the standard deviation over repeated measurements for that gene (Smyth, 2004). Genes with scores greater than a threshold are considered differentially expressed. The threshold significance is determined by the user based on the FDR. The proportion of such genes identified by chance (false positives) is the false discovery rate (FDR). To estimate the FDR, nonsense genes are specified using random permutations of the repeated measurements (Tusher et al., 2001; Yan et al., 2005).

8. R and bioconductor packages

Microarray experiments produce large and highly complex datasets. Access to an efficient statistical computing environment is a critical aspect of the analysis of these gene expression datasets. There are a lot of free and commercial software. In most cases, the microarray kits come with the software that adequately analyses microarray data. One of the best options for data analysis is the R statistical programming environment (www.rproject.org) where

the open-source Bioconductor R packages (www.bioconductor.org) are resourceful and effective in dealing with these microarray data.

There are plenty of packages such as limma, marray and arrayQuality for two-color spotted arrays or affy, affyPLM, affyPara and gcrma for Affymetrix array and Agi4x44PreProcess and AgiMicroRna for Agilent chips. The complete documentation of Bioconductor packages can be found on the Bioconductor project web site at: http://www.bioconductor.org/help/bioc-views/release/bioc/. Bioconductor packages remove noise from measurements of microarray experiments through preprocessing of data. Also they specialize in various related tasks in handling microarray data. Some packages are dedicated to facilitation and automation of array data input and applied to detection of spatial and dye effects on arrays via a variety of diagnostic plots and graphs. In addition to the primary fluorescence intensity data, these packages also extract textual information on probe sequences and target samples, such as gene annotations, layout array, target sample descriptions and hybridization conditions, etc.

Limma package implements tools for data quality assesment, background correction, normalization and identification of DEGs in microarray experiments. Marray package also provides alternative functions for reading microarray data into R, normalization data and diagnostic plots of different measurements. Limma and marray packages share some features (Smyth & Speed, 2003; Yang et al., 2002b).

In the following, in a case study, we will demonstrate a microarray analysis flow using Bioconductor R packages in experimental design, data preprocessing, and differential expression detection. This analysis is performed using Bioconductor Release 2.7 based on R Version 2.9.

8.1 Step by step microarray analysis

The publicly available dataset from Swirl zebra fish two-color spotted microarray experiment was used as a typical example in this analysis. Swirl is a point mutant in the BMP2 gene that affects the dorsal/ventral body axis. In this experiment, two sets of dye-swap were prepared. On the first array, the wild type and mutated samples were labelled with Cy5 and Cy3 dyes, respectively. On the second array, the Cy5 and Cy3 dyes were swapped for the samples. The next two arrays were replicates of the first two arrays, respectively (Table 5). Thus, four arrays were prepared. Each array consisted of 16 print tips (4 by 4) and each print tip comprised 22 by 24 spots. Therefore, each array accounted for 8448 spots. Once the experimental steps were carried out, each array was scanned and then analyzed by SPOT software (Buckley, 2000). The main purpose of the Swirl experiment is to find genes with altered expression in the swirl mutant compared to wild type zebra fish.

Date	FileName	Slide number	Conditions
2001/9/20	Swirl.1.spot	81	Swirl(Cy3), wild type (Cy5)
2001/9/20	Swirl.2.spot	82	Swirl(Cy5), wild type (Cy3)
2001/11/8	Swirl.3.spot	93	Swirl(Cy3), wild type (Cy5)
2001/11/8	Swirl.4.spot	94	Swirl(Cy5), wild type (Cy3)

Table 5. All experiments for the study of Swirl mutant.

In order to analyze the microarray data, a directory of all the image processing output files should be created (.spot files). This directory includes a file containing experiment description (SwirlSamples.txt file) and a file describing information on probe sequences, such as gene names, spot ID (fish.gal). Then R is started in the desired working directory. The following command will load Limma and marray packages for preprocessing swirl data.

```
>library (limma)
>library (marray)
```

Information about the hybridizations and the raw fluorescent intensities data are provided through the following commands

```
>targets <- readTargets (“SwirlSamples.txt”)
>RG <- read.maimages (targets$FileName, source=”spot”)
```

In order to identify gene names the following command may be used,

```
>Genes <- readGAL(“fish.gal”)
```

and the layout information of slides uses this command,

```
>Layout <- getLayout (Genes)
```

Qualitative assessment of arrays can be performed using different plots and graphs in microarray experiments. Therefore, serious quality problems and sources of artifacts will be identified in the data. In this step, the background signal, different biases such as dye bias and spatial bias are evaluated using visualization techniques. According to the results of the quality assessment, the need for each preprocessing method is clearly revealed. Firstly, the background signal distribution is evaluated to identify whether there is any region with non-uniformity distribution.

```
>imageplot (log2(RG$Rb[,1]), Layout, low="white", high="red")
>imageplot (log2(RG$Gb[,1]), Layout, low="white", high="green")
```

Figure 11 shows that the background signals in both red and green channels are unreliably high in some region of array. It can be concluded as that there is spatial non-uniformity.

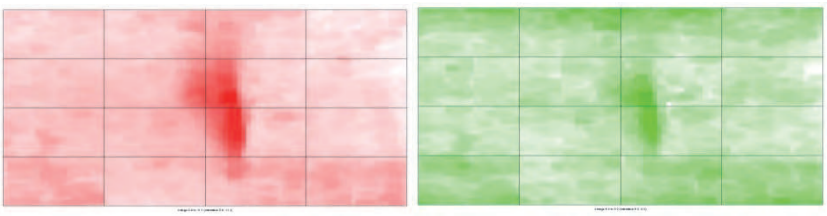


Fig. 11. Image of green and red channel background intensities for slide 1.

Therefore background correction is performed on swirl data. Since data have no negative value, subtractive methods can be carried out. Background corrected M and A-values are generated through subtraction method as follows

```
>MA <- normalizeWithinArrays (RG, method=”none”)
```

Secondly, we visualize the intensity range of M-values for each individual microarray using MA-plots. The signals include both background signals and foreground signals. These plots are generated using the following commands:

```
> plotMA (MA[,1], main="slide 1", ylim=c(-3,3))
> plotMA (MA[,2], main="slide 2", ylim=c(-3,3))
```

Figure 12 shows MA-plots of raw data of two slides of swirl experiment. Swirl experiment satisfies major assumption in microarray experiment i.e. a small percentage of genes are expected to be differentially expressed. Therefore, the majority of the points on the y axis (M-value) would be located at 0, since log (1) is 0. The shape of the curve on slide 1 shows more non–linear dependence on the overall spot intensity than slide 2. Therefore, normalization will attempt to remove the curvature of the spread and centralize the data around zero axis.

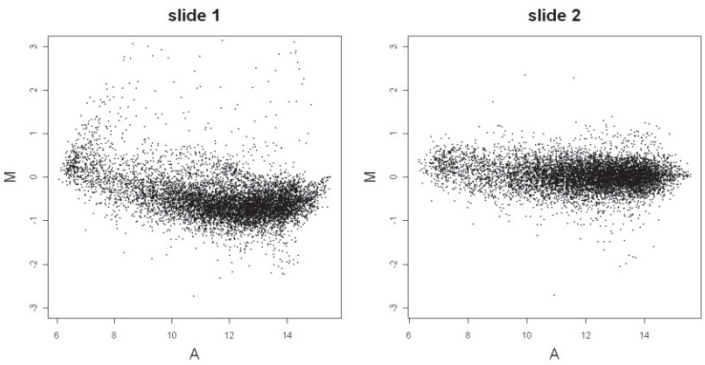


Fig. 12. MA plots for slide 1 and slide 2 in swirl experiment.

Another diagnostic plot is boxplot which can be useful for comparing M-values and homogeneity between print tip group and slides. Boxplot in different print tip is plotted using marray package after generation background corrected data by maNorm function.

```
>swirl.norm <- maNorm (swirl, norm="none")
>boxplot (swirl.norm[,1], xvar="maPrintTip", yvar="maM", main="slide
1")
```

The following command also produces a boxplot of the pre–normalization M-values for all four arrays in the swirl experiment.

```
>boxplot (MA$M~col (MA$M), xlab="slides", ylab="M")
```

A boxplot shows graphically 5-number summary of data, the median, the upper and lower quartiles, the range, and individual extreme values. The central box in the plot represents the interquartile range (IQR), which is specified as the difference between the 75th percentile and 25th percentile. The width of a box represents the variability of the data and solid line in the middle of the box represents the median. Extreme values, greater than 1.5×IQR above the 75th percentile and less than 1.5×IQR below the 25th percentile, are plotted individually (54).

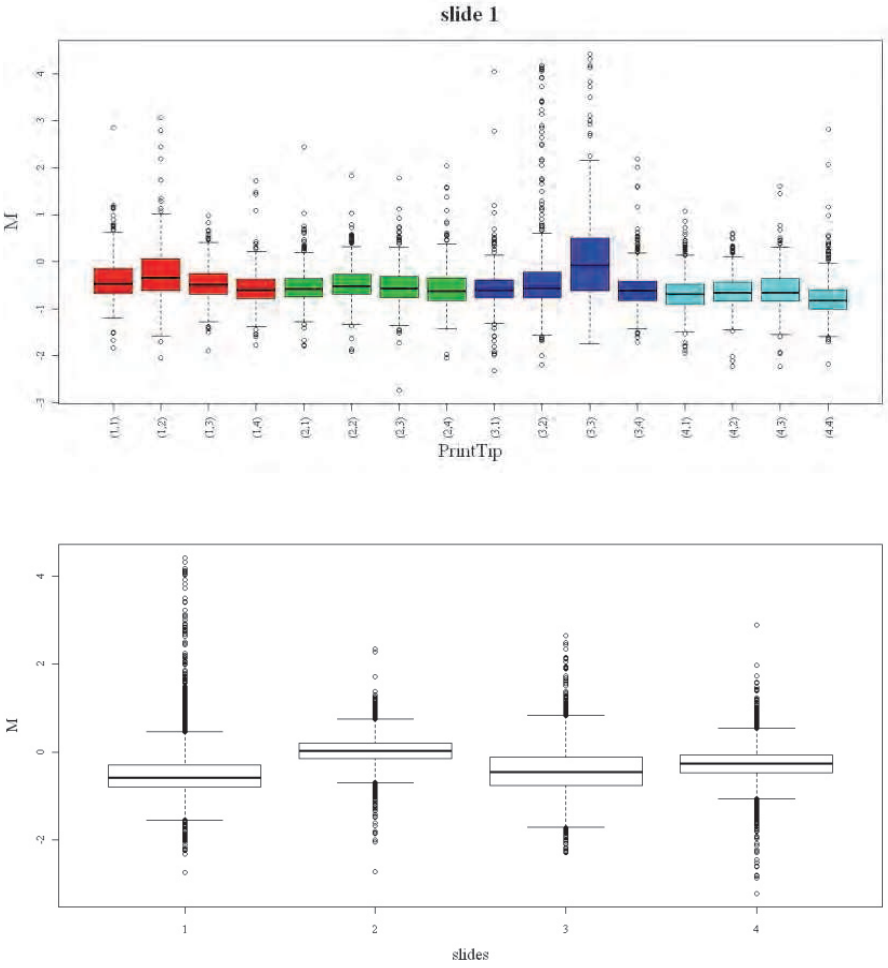


Fig. 13. Boxplots in different slides and print tips of slide 1.

In the next step, we can normalize data through different within and between array normalization using both marray and limma packages. The pre–normalization MA–plot and boxplot for slide 1 in Figures 12 and 13 illustrate the non–linear dependence of the M-value on the overall spot intensity A and the existence of spatial biases. We thus perform PTloess normalization on this data. In the following scale normalization will be performed on the swirl data because four slides have different spread of M-values (Figure 14).

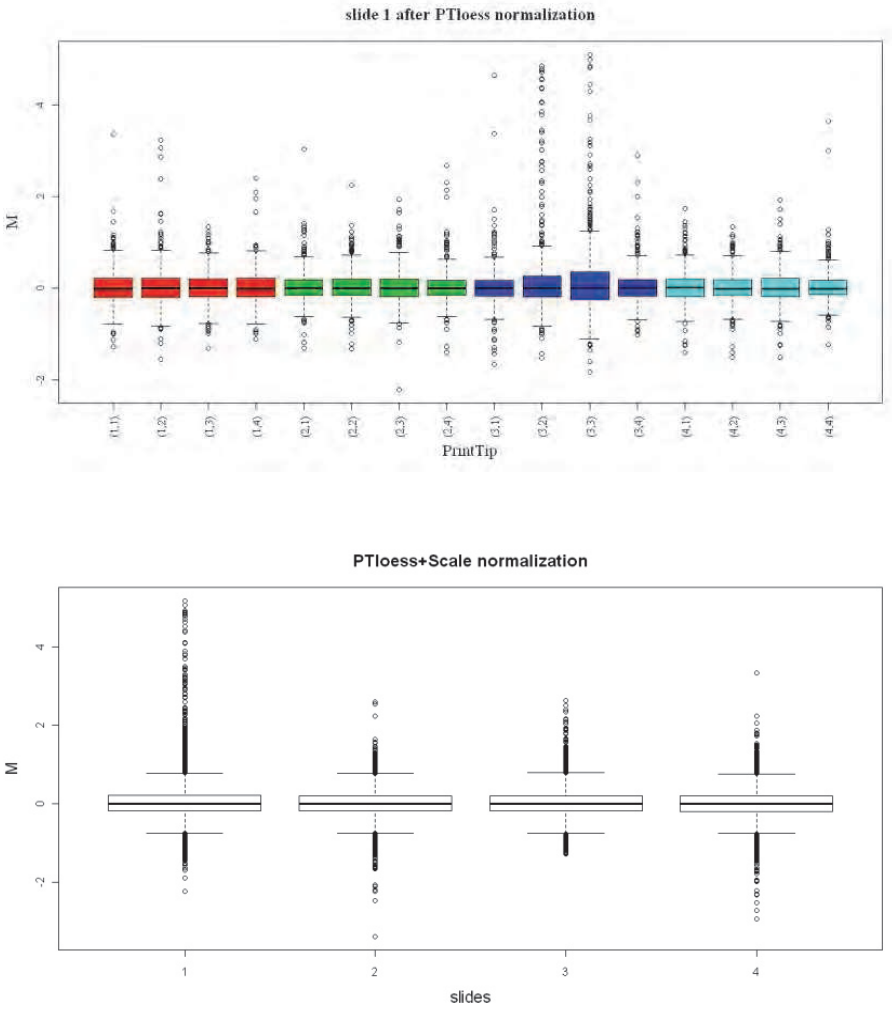


Fig. 14. Swirl data after PTloess normalization and PTloess fallowing Scale normalization.

```
>MA <- normalizeWithinArray(RG)
>MA <- normalizeBetweenArrays (MA, method="scale")
```

The closer the solid line to zero line in Figure 14 the more centrality of the data after normalization. In boxplot plotted between arrays when the width of the rectangles are approximately the same the distribution of the spots on replicate arrays are the most similar that means between-array normalization method has been selected appropriately.
In order to detect genes with differential expression between wild type and mutant samples, linear model and empirical bayes methods in limma package are used (Smyth, 2004). Dye swap samples are specified using the design matrix, which allows calculating of the average M- values across multiple arrays.

```
>design <- c(-1,1,-1,1)
```

M-values are estimated between these two samples using the lmFit function.

```
>fit <- lmFit(MA,design)
```

Moderated t-statistics and log–odds (B-statistics) of expression data are calculated using empirical Bayes methods

```
> fit <- eBayes (fit)
```

A summary table of some statistics for the top genes will be obtained using the following command.

```
>topTable(fit, number=10, adjust="fdr", sort.by="t")
```

ID	Name	M-value	Moderated-t	B	Adj.P.val
Control	BMP2	-2.205288	-21.06952	7.960750	0.0003572816
Control	BMP2	-2.296045	-20.28697	7.778330	0.0003572816
Control	Dlx3	-2.184900	-20.01066	7.710959	0.0003572816
Control	Dlx3	-2.180471	-19.63599	7.710959	0.0003572816
fb94h06	20-L12	1.271119	14.08467	7.617005	0.0020666932
fb40h07	7-D14	1.347207	13.52924	5.535983	0.0020666932
fc22a09	27-E17	1.266129	13.41339	5.483567	0.0020666932
fb85f09	18-G18	1.275686	13.39543	5.475386	0.0020666932
fc10h09	24-H18	1.195126	13.23722	5.402676	0.0020666932
fb85a01	18-E1	-1.287128	-13.07059	5.324819	0.0020666932

Table 6. Top 10 genes from the Swirl data.

The moderated t-statistic with adjusted p-values can identify differentially expressed genes. As seen in Table 6, it can sort both copies of the gene BMP2 knocked out and both copies of Dlx3, which is a known target of BMP

9. Conclusion

Gene expression is a common process in all forms of living cells to generate the macromolecules which are necessary for life. Systemic comprehension of the cell function is provided using study of gene expression. Investigation of molecular dynamics of the cell can be performed in three biochemical levels, transcriptomics, proteomics, metabolomics. Compared to others, transcriptomics is a more robust, large-scale, moderate cost technology of simultaneously measuring thousands of mRNA level. There are various techniques for quantifying gene expression based on mRNA. However gene expression traditional techniques provide valuable biological information, they are limited in some ways such as scale, economy and sensitivity. Therefore, compared to the other commonly used techniques, quantification based on microarray is really remarkable because of being high throughput and cost effective. It enables the simultaneous analysis of thousands of genes within one single experiment. Such miniaturized binding technology is typically divided into DNA, protein, tissue, cellular and chemical compound microarrays. DNA microarrays are the most popular type of this technology which currently manufactured through two main approaches: in situ synthesis and deposition of pre-synthesized probes (spotted arrays). We focused mainly on the spotted cDNA microarray. After microarray experiment, slides are inserted into scanner. The output data are fluorescent images arranged into a matrix of spots. Then, images are processed to quantify level of gene expression based on the intensity of each spot and obtain background estimates and quality measures. It is performed in gridding, segmentation, quantification and spot quality assessment stages. The output data from image processing stage needs to be preprocessed to eliminate non-biological variations, transform data into a suitable scale and improve the quality of downstream analysis. These are performed using background correction, logarithm transformation and normalization of microarray data. Finally, identification of genes that are significantly differentially expressed under different conditions can be carried out using marginal filters, wrappers, and embedded methods. We pointed some of the filter approaches such as t-test and its variants such as moderated t-test and SAM approach and analysis of variance (ANOVA). In summary, in order to analyze microarray data, the R statistical programming environment is chosen where the Bioconductor R packages such as limma and marray are effective in processing these microarray data. These packages address data input, production of diagnostic plots to detection of different biases, the statistical methods of removing experimental noises and errors on the spots within and between arrays. Finally, limma package is also used as a powerful tool to identification of differentially expressed genes.

10. References

Adams, R., & Bischof, L. (1994). Seeded region growing. *IEEE transactions on pattern analysis and machine intelligence,* Vol.16, No.6, pp. 641–647

Angenendt, P. (2005). Progress in protein and antibody microarray technology. *DDT*., Vol.10, No.7, pp. 503-511

Barrett, J.C., & Kawasaki, E.S. (2003). Microarrays: the use of oligonucleotides and cDNA for the analysis of gene expression. *DDT*., Vol.8, No.3, pp. 134-141

Beranova-Giorgianni, S. (2003). Proteome analysis by two-dimensional gel electrophoresis and mass spectrometry: strengths and limitations. *Trends in Analytical Chemistry,* Vol.22, No.5, pp. 273-281

Beucher, S., & Meyer, F. (1993). The morphological approach to segmentation: the watershed transformation. *Processing of mathematical morphology in image,* NewYork, pp. 433–481

Bilban, M., Buehler, L.K., Head, S., Desoye, G., & Quaranta, V. (2002). Normalizing DNA Microarray Data. *Curr. Issues Mol. Biol.,* Vol.4, pp. 57-64

Bolstad, B.M., Irizarry, R.A., Astrand, M., & Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics,* Vo.19, No.2, pp. 185-193

Breljak, D., Ambriović-Ristov, A., Kapitanović, S., Čačev, T., & Gabrilovac, J. (2005). Comparison of Three RT-PCR Based Methods for Relative Quantification of mRNA. *Food Technol. Biotechnol,* Vol.43, No.4, pp. 379–388

Buckley, M.J. (2000). Spot User's Guide, CSIRO Mathematical and Information Sciences, Sydney, Australia, Available from: <http://www.cmis.csiro.au/iap/Spot/spotmanual.htm>

Crick, F. (1970). Central dogma of molecular biology. *Nature,* Vol.227, pp. 561–563

Cui, X., & Churchill, G.A. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol,* Vol.4, pp.210–219

Doyle, H.A., & Mamula, M.J. (2001). Post-translational protein modifications in antigen recognition and autoimmunity. *TRENDS in Immunology,* Vol. 22, No.8, pp. 443-449

Fodor, S.P.A., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T., & Solas, D. (1991) Light-directed, spatially addressable parallel chemical synthesis. *Science*, Vol.251, pp. 767–773

Frith, M.C., Pheasant, M., & Mattick, J.S. (2005). The amazing complexity of the human transcriptome. *European Journal of Human Genetics,* Vol.13, pp. 894–897

Gjerstad, Ø., Aakra, Å., Snipen, L., & Indahl, U. (2009). Probabilistically assisted spot segmentation-with application to DNA microarray images. *Chemometrics and Intelligent Laboratory Systems,* Vol.98, pp. 1–9

Gulmann, C., & O'Grady, A. (2003). Tissue microarray: an overview. *Current Diagnostic Pathology,* Vol.9, pp. 149 -154

Hall, D.A., Ptacek, J., & Snyder, M. (2007). Protein Microarray Technology. *Mech Ageing Dev,* Vol.128, No.1, pp. 161–167

Hegde, P.S., White, I.R., & Debouck, C. (2003). Interplay of transcriptomics and proteomics. *Current Opinion in Biotechnology,* Vol.14, No.6, pp. 647–651

Heid, C.A., Stevens, J., Livak, K.J., & Williams, P.M. (1996). Real Time Quantitative PCR. *Genome Research,* Vol.6, pp. 986-994, ISSN 1054-9803/96

Hirsch, J., Hansen, K.C., Burlingame, A.L., & Matthay, M.A. (2004). Proteomics: current techniques and potential applications to lung disease. *Am J Physiol Lung Cell Mol Physiol,* Vol. 287, No.1, pp. L1–L23

Istepanian, R.S.H., (2003). Microarray Image Processing: Current Status and Future Directions. *IEEE Trans. Nanobioscience,* Vol.2, No.4, pp. 173-175

Karakach, T.K., Flight, R.M., & Douglas, S. (2010). An introduction to DNA microarrays for gene expression analysis. *Chemometrics and Intelligent Laboratory Systems,* Vol.104, No.1, pp. 28–52

Knapp, G., Beckwith, J.S., Johnson, P.F., Fuhrman, S.A., & Abelson, J. (1978). Transcription and processing of intervening sequences in yeast tRNA genes. *Cell 14,* pp. 221–236

Kononen, J., Bubendorf, L., Kallioniemi. A., Barlund, M., Schraml, P., Leighton. S., Torhorst. J., Mihatsch, M.J., Sauter, G., & Kallioniemi, O.P. (1998). Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med*, Vol.4, pp. 844–847

Kooperberg, C., Fazzio, T.G., Delrow, J.J., & Tsukiyama, T. (2002). Improved Background Correction for Spotted DNA Microarrays. *Journal of computational biology,* Vol.9, No.1, pp. 55-66

Lipshutz, R.J., Fodor, S.P.A., Gingeras, T.R., & Lockhart, D.J. (1999). High density synthetic oligonucleotide arrays. *Nature genetics supplement,* Vol.21, pp. 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., Norton, H., & Brown, E.L. (1996). DNA expression monitoring by hybridization of high density oligo-nucleotide arrays. *Nature Biotechnology*, Vol.14, pp. 1675–1680

Parsons, M., & Grabsch, H. (2009). How to make tissue microarrays. *Diagnostic histopathology,* Vol.15, No.3, pp. 142-150

Perdew, G.H., Vanden Heuvel, J.P., & Peters, J.M. (2006). Regulation of Gene Expression: Molecular Mechanisms. *Humana Press,* pp. 11-30

Pinet, F. **(**2009). Identifying patients at risk of progressive left ventricular dysfunction. *Heart Metab,* Vol. 42**,** pp. 10–14

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature genetics supplement*, Vol.32, pp. 496–501

Ritchie, M.E., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A., & Smyth, G.K. (2007). A comparison of background correction methods for two-colour microarrays. *Bioinformatics,* Vol.23, No.20, pp. 2700–2707

Russo, G., Zegar, C., & Giordano, A. (2003). Advantages and limitations of microarray technology in human cancer. *Oncogene,* Vol.22, pp. 6497–6507

Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics,* Vol.23, No.19, pp. 2507–2517

Schadt, E.E., Li, C., Su, C., & Wong, W.H. (2001). Analyzing High-Density Oligonucleotide Gene Expression Array Data. *Journal of Cellular Biochemistry,* Vol.80, pp. 192–202

Schaub, M.C., Lucchinetti, E., & Zaugg, M. (2009). Genomics, transcriptomics, and proteomics of the ischemic heart. *Heart Metab,* Vol.42, pp. 4–9

Schena, M., Shalon, D., Davis, R., & Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science,* Vol.270, pp. 467-470

Schena, M. (2003). *Microarray analysis*. John Wiley & Sons, New Jersey

Schuchhardt, J., Beule, D., Wolski, E., Eichhoff, H., Leharch, H., & Herzel, H. (2000). Normalization strategies for cDNA microarrays. *Nucleic Acids Research,* Vol.28, No.10, pp. e47

Singh, A., & Sau, A. K. (2010). Tissue Microarray: A powerful and rapidly evolving tool for high-throughput analysis of clinical specimens. *IJCRI,* Vol.1, No.1, pp. 1-6

Stekel, D. (2003). *Microarray Bioinformatics.* Cambridge University Press, Cambridge

Smyth, G.K., & Speed, T. (2003). Normalization of cDNA microarray data. *Methods,* Vol.31, pp. 265-273

Smyth, G.K. (2004). Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology,* Vol.3, No.1, Article 3

Soli, P. (1999). *Morphological image Analysis: Principles and Applications.* Springer-Verlag Berlin, Heidelberg, New York

Stafford, P. (2008). *Methods in microarray normalization*. Taylor and Francis CRC Press, 978-1-4200-5278-7, Boca Raton, London, NewYork

Tarca, A.L., Romero, R., & Draghici, S. (2006). Analysis of microarray experiments of gene expression profiling. *American Journal of Obstetrics and Gynecology,* Vol.195, pp. 373–388

Templin, M.F., Stoll, D., Schrenk, M., Traub, P.C., Vöhringer, C.F., & Joos, T.O. (2002). Protein microarray technology. *TRENDS in Biotechnology,* Vol.20, No.4, pp. 160-166

Trayhuru, P., (1996). Northern blotting. *Proceedings of the Nutrition Society*, Vol.55, pp. 583-589

Tsiridis, E., & Giannoudis, P.V. (2006). Transcriptomics and proteomics: Advancing the understanding of genetic basis of fracture healing. *Injury, Int. J. Care Injured,* Vol. 37S, pp. S13—S19

Tusher, V.G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences,* Vol.98, No.9, pp. 5116–5121

van der Werf, M.J., Jellema, R.H., & Hankemeier, T. (2005). Microbial metabolomics: replacing trial-and-error by the unbiased selection and ranking of targets. *J Ind Microbiol Biotechnol,* Vol. 32, pp. 234–252

Wang, D. (2007). Spot: cDNA Microarray Image Analysis Users Guide. CSIRO Mathematical and Information Sciences, Australia, Available from: < http://spot.cmis.csiro.au/spot/doc/Spot.pdf>

Xiong, H., Zhang, D., Martyniuk, C.J., Trudeau, V.L., & Xia, X. (2008). Using Generalized Procrustes Analysis (GPA) for normalization of cDNA microarray data. *BMC Bioinformatics,* Vol.9, No.25

Yan, X., Deng, M., Fung, W.K., & Qian, M. (2005). Detecting differentially expressed genes by relative entropy. *Journal of Theoretical Biology,* Vol.234, pp. 395–402

Yang, Y.H., Buckley, M.J., & Speed, T.P. (2001). Analysis of cDNA microarray images. *Briefings in bioinformatics,* Vol.2, No.4, pp. 341-349

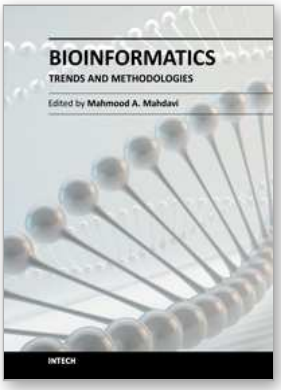
Yang, Y.H., Buckley, M.J., Dudoit, S., & Speed, T.P. (2002a). Comparison of methods for image analysis on cDNA microarray data. *J. Comput. Graph. Statist.,* Vol.11, pp. 108–136

Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J., & Speed, T.P. (2002b). Normalization for cDNA microarray data: a robust composite method addressing

single and multiple slide systematic variation. *Nucleic Acids Research,* Vol.30, No.4, pp. e15

Zhang, L., Zhang, X., Ma, Q., Ma, F., & Zhou, H. (2010). Transcriptomics and Proteomics in the Study of H1N1 2009. *Genomics Proteomics Bioinformatics,* Vol.8, No.3, pp. 139-144

Zhao, H., Engelen, K., De Moor, B., & Marchal, K. (2007). CALIB: a BioConductor package for estimating absolute expression levels from two-color microarray data. *Bioinformatics*, Vol.23, No.13, pp. 1700-1701



Bioinformatics - Trends and Methodologies
Edited by Dr. Mahmood A. Mahdavi

ISBN 978-953-307-282-1
Hard cover, 722 pages
Publisher InTech
Published online 02, November, 2011
Published in print edition November, 2011

Bioinformatics - Trends and Methodologies is a collection of different views on most recent topics and basic concepts in bioinformatics. This book suits young researchers who seek basic fundamentals of bioinformatic skills such as data mining, data integration, sequence analysis and gene expression analysis as well as scientists who are interested in current research in computational biology and bioinformatics including next generation sequencing, transcriptional analysis and drug design. Because of the rapid development of new technologies in molecular biology, new bioinformatic techniques emerge accordingly to keep the pace of in silico development of life science. This book focuses partly on such new techniques and their applications in biomedical science. These techniques maybe useful in identification of some diseases and cellular disorders and narrow down the number of experiments required for medical diagnostic.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821