J.P. Vanden Heuvel, INDIGO Biosciences Inc., State College PA

Gene Expression: General Concepts

1. Overview

In the present white paper, the basic approaches to understanding how a treatment or condition results in mRNA accumulation will be described. Although the focus and many of the examples will center on gene regulation by *nuclear receptors*, the approaches are applicable to any treatment or condition that alters gene expression. In an accompanying document, the specific means by which nuclear receptors affect gene expression will be described. (For more detail on approaches to study gene expression, the reader is referred to Perdew GH, Vanden Heuvel JP, Peters JM. 2006. *Regulation of gene expression: molecular mechanisms*, pp. ix, 333. Humana Press).

2. Messenger RNA accumulation

a. Concepts

The detection of mRNA levels of a particular gene is one of the cornerstones of molecular biology. Surprisingly, few people give much thought as to whether their methodologies and approaches are appropriate. There are many ways that mRNA can be detected, each with its strengths and weaknesses. However, the quantification and interpretation of results depends on understanding some key points. Three key factors contribute to the complexity and difficulty in examining differentially expressed genes. First, genes are not present at the same abundance and can vary from less than one up to thousands of copies per cell. This has implications for methods that must be used to accurately detect and quantify the expression of an mRNA. Second, the intensity of response varies greatly from gene-to-gene. That is, when comparing two treatments or conditions, an mRNA can be two-fold or several orders of magnitude different between the samples. Certain methods have a robust linear range and can handle both types of responses while others are biased toward the low or high responder. Last, there are many ways to alter gene expression. Some or all of the particular mechanisms may be at play. Approaches must be used that are capable of isolating, or at least accounting for, the competing possibilities such that the hypotheses can be tested. We will briefly discuss these parameters as they pertain to examining altered gene expression, and how these factors impinge on developing an optimal model system.

b. mRNA abundance.

The mammalian genome of 3 x 10^9 base pairs has enough DNA to code for approximately 300,000 genes, assuming a length of 10,000 bp per gene. Obviously, not every gene is expressed per cell (also, not every segment of DNA may be associated with a gene product). In fact, in a mammalian cell, hybridization experiments have shown that approximately 1-

2% of the total sequences of nonrepetitive DNA are represented in mRNA. Thus, if 70% of the total genome is non-repetitive, 10,000 to 15,000 genes are expressed at a given time.

The average number of molecules of each mRNA per cell is called its representation or abundance. Of the 10,000-15,000 genes being expressed, the mass of RNA being produced per gene is highly variable. In fact, usually a few sequences are providing a large proportion of the total mass of mRNA. Hybridization and kinetic experiments between excess mRNA and cDNA in solution identifies several components of mRNA complexity. The majority of the mass of RNA (50%) is being accounted for by a component with few mRNA species. In fact, approximately 65% of the total mRNA may be accounted for in as few as 10 mRNA species. The remaining 35% of the total RNA represents the remaining genes being expressed in that tissue. Of course, the genes present in each category may be present in very different amounts and represent a continuum of expression levels. For means of this discussion, we will divide the three major components into abundant, moderate and scarce representing approximately 100,000 copies, 5000 copies and <10 copies per cell respectively.

There are several reasons for discussing the components of mRNA. First, when doing a differential screen (i.e. microarray) what is actually being compared is two populations of mRNA and you are examining the genes that overlap or form the intersection between groups. When comparing two extremely divergent populations, such as liver and oviduct, as much as 75% of the sequences are the same equating to 10,000 genes that are identical and approximately 3000 genes that are specific to the oviduct. This suggests that there may be a common set of genes, representing required functions, that are expressed in all cell types. These are often referred to as housekeeping or constitutive genes. Second, there are overlaps between all components of mRNA, regardless of the number of copies per cell. That is, differentially expressed genes may be abundant, moderate or scarce. In fact, the scarce mRNA may overlap extensively from cell to cell, on the order of 90% for the liver to oviduct comparison. However, it is worthy to note that a small number of differentially expressed genes are required to denote a specialized function to that cell, and the level of expression does not always correlate with importance of the gene product.

As to be discussed subsequently, the key to developing an effective model for the study of differential gene expression may be to keep the differences in the abundant genes to a minimum. This is due to the fact that a small difference in expression of a housekeeping gene, say 2 fold, will result in a huge difference in the number of copies of that message from cell type-to-cell type (i.e. an increase of 10,000 copies per cell). Also, it is important to have a screening method that can detect differences in the scarce component. If the two populations to be compared have little difference in the abundant genes and you have optimized your screening technique to detect differences in the scarce population, the odds of finding genes that are truly required for a specialized cellular function have increased dramatically.

c. Intensity of response

A basic pharmacologic principle is that drugs and chemicals have different affinities for a receptor and the drug-receptor complex will have different efficacies for producing a biological response, i.e. altering gene

expression. A corollary of this principle states that not every gene being affected by the same drug-receptor complex will have identical dose-response curves. That is, when comparing two responsive genes, the affinity of the drug-receptor complex for the DNA response elements found in the two genes and the efficacy of the drug-receptor-DNA complex at effecting transcription could be quite different. In fact, similar DNA response elements may cause a repression or an induction of gene expression, depending on the context of the surrounding gene. Therefore, when comparing two populations of mRNAs (i.e. control versus treated), there may be orders of magnitude differences in the levels of induction and repression regardless of the fact that all the genes are affected by the same drug-receptor complex.

Needless-to-say, the extent of change is important in the detection of these differences, but not the importance of that deviation. For technical reasons, it is often difficult to detect small changes in gene expression (<2 fold). However, a two-fold change in a gene product may have dramatic effects on the affected cell, especially if it encodes a protein with a very specialized or non-redundant function. Also, the detection of a difference between two cell populations is easier if the majority of the differences are in scarce mRNAs. Once again, this is due to technical aspects of analyzing gene expression whereby the change from 500 to 1000 copies per cell is a dramatic effect compared to a change from $1x10^5$ to 2×10^5, an effect that may be virtually unnoticed.

d. Specificity of response

The last factor we will discuss regarding the complexity of mRNA species, is that regulation of gene expression is multi-faceted. The analysis of differential gene expression is most often performed by comparing steady-state levels of mRNA. That is, the amount of mRNA which accumulates in the cell is a function of the rate of formation (transcription) and removal (processing, stability, degradation). If differences in protein products are being compared, add translation efficiency, processing and degradation to the scenario. With all the possible causes for altered gene expression, the specificity of response must be questioned. Is the difference in mRNA or protein observed an important effect on expression or is it secondary to a parameter in your model system you have not controlled or accounted?

In the best-case situation, the key mechanism of gene regulation that results in the end-point of interest should be known. At least one should have criteria in mind for the type of response that is truly important. With most receptor systems, early transcriptional regulation may predominate as this key event. By assuming that the key event is mRNA accumulation, the true initiating response, such as protein phosphorylation or processing, may be overlooked. Also, the extent and diversity of secondary events, i.e. those that require the initial changes in gene expression, may far exceed the primary events. The amplification of an initial signal (i.e. initial response “gene A” causes regulation of secondary response “gene B”) can confuse the interpretation of altered mRNA accumulation. Once again, one must have a clear understanding of whether a primary or secondary event is the key response and design your model accordingly.

2. Methods and approaches for examining mRNA accumulation for a specific transcript.

The examination of transcript concentrations is one of the fundamental experiments in modern

biology. There are currently several methods for the quantification of mRNA accumulation, each with identifiable strengths and weaknesses. Perhaps the two most common methods are Northern blots and RT-PCR. Northern blots are insensitive and relatively non-quantitative. However, they allow for visualization of transcript size and utilize well-established, straightforward techniques and are inexpensive to perform. In contrast, RT-PCR (quantitative, real-time) are exquisitely sensitive and robust with the capability for high-throughput gene expression analysis. RT-PCR requires expensive equipment and must be performed with care to minimize contamination and tube-to-tube variability. We will focus our discussion on PCR techniques since they have been widely accepted as the most sensitive and quantitative means for addressing mRNA accumulation.

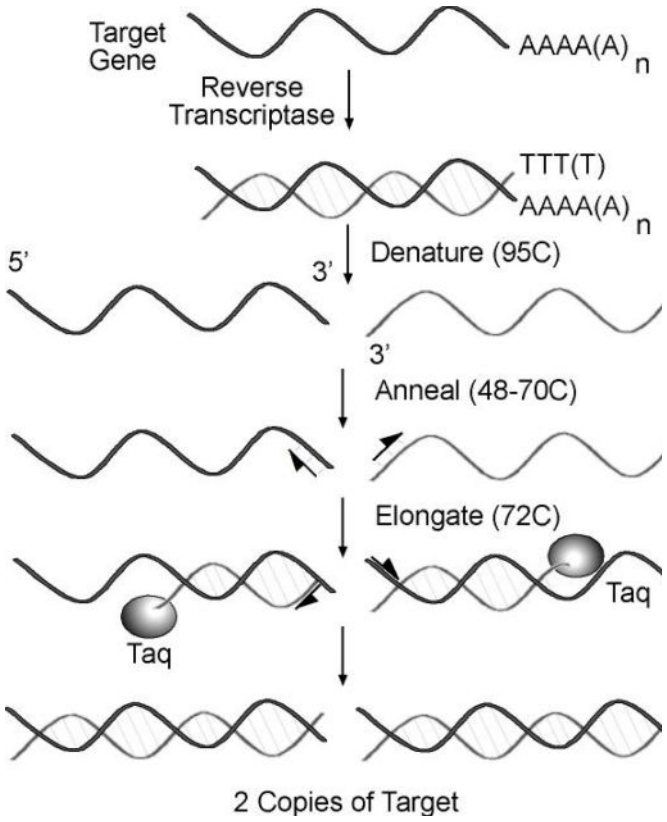

Figure 1. Reverse transcription and one cycle of PCR.

a. General concepts of RT-PCR

The polymerase chain reaction (PCR) is an enzymatic assay which is capable of producing large amounts of a specific DNA fragment starting from a small amount of a complex mixture (Figure 1). In the case of RT-PCR, the messenger RNA must first be converted to a double stranded molecule with the use of the enzyme reverse transcriptase (RT). The thermostable DNA polymerase (i.e. Taq) and the use of specific "primers" are the key features of any PCR reaction. All known DNA polymerases require deoxyribonucleotide triphosphates (dNTPs), a divalent cation (Mg^{2+} or Mn^{2+}), a DNA or cDNA template, and a region of that template that is double-stranded adjacent to a single stranded nick or gap. The double-stranded region is provided by the primer annealing to its complementary region of the DNA template. If the starting mixture includes not only a single stranded polynucleotide template, but also: (a) its complementary strand, and (b) two oligonucleotide primers that hybridize to both strands, copies of *both* of these strands will be produced each cycle and these copies can be used as templates for subsequent cycles. Short DNA fragments whose ends are defined by the position of the two oligonucleotide primers will accumulate in an exponential fashion, i.e., like a chain reaction. If 30 cycles of PCR are performed, theoretically one will achieve a 2^{30} amplification of the target gene's cDNA. The product which is formed is specific for a particular biomarker as dictated by the design of the oligonucleotide primers.

Therefore, RT-PCR is a tool to examine the messenger RNA expression of a target gene with the amount of product formed a function of the amount of starting template. Of course, the examination of mRNA accumulation can be determined in many cases by hybridization

procedures such as Northern blots, dot- or slot-blots, and RNase protection assays. Nonetheless, in terms of amount of sample required, detection of small differences in expression and ability to examine many genes in a large number of samples, RT-PCR stands above the more conventional procedures.

b. Real-time RT-PCR

"Real-time" PCR gets its name from the fact that reaction products are quantitated for each sample in every cycle. The result is a large (10^7-fold) dynamic range, with high sensitivity and speed. Pre-optimized kits for thousands of genes are available for purchase which greatly reduces the time required to generate data. However, the procedure requires expensive equipment and reagents, which may be prohibitive to some investigators.

Real-time PCR systems rely upon the detection and quantitation of a fluorescent reporter. Since the products are analyzed in real-time, there is no need for post-PCR manipulation. This decreases the chance for experiment-to-experiment contamination and eliminates the variability inherent in agarose gel electrophoresis and quantitation. Currently there are two major methods that are used in real-time PCR: DNA fluorescent dyes (SYBR green) and Fluorescent resonance energy transfer (FRET, i.e. TaqMan, Molecular Beacons). The amount of fluorescence in each case increases as the amount of PCR product increases. See Figure 2 for a summary of the two methods.

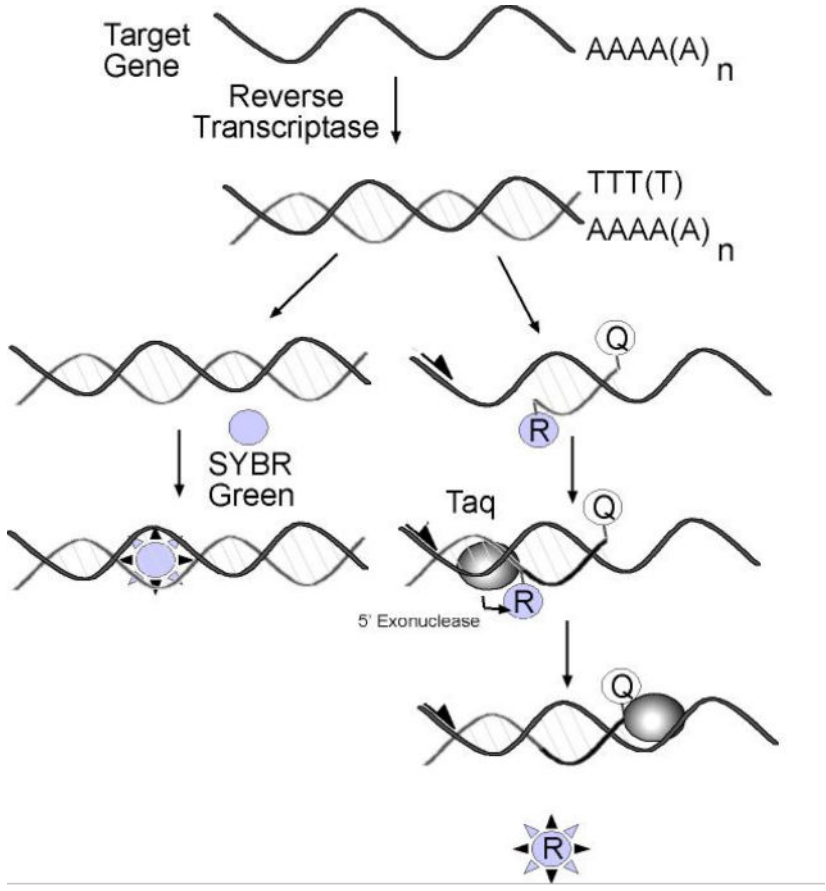

Figure 2. Real time PCR with SYBR green (left side of diagram) or "TaqMan" based (right-side) detection.

Of the DNA fluorescent dyes, SYBR green is the most common. SYBR green is an intercollating agent which binds double-stranded DNA, and in this bound form upon excitation emits light. Thus, as a PCR product accumulates, fluorescence increases. The advantages of SYBR Green are that it's inexpensive, easy to use, and sensitive. The disadvantage is that SYBR Green will bind to any double-stranded DNA in the reaction, including primer-dimers and other non-specific reaction products, which results in an overestimation of the target concentration.

Both the TaqMan and Molecular Beacons approaches utilize fluorescent probes and FRET for quantitation. This assay utilizes the 5' nuclease activity of Taq polymerase. TaqMan Probes are oligonucleotides that contain a fluorescent dye on the 5' end (i.e. FAM), and a quenching dye on the 3' base (i.e. TAMRA). These probes are designed to hybridize between the standard PCR primers. Prior to enzymatic activity, the excited fluorescent dye (FAM) transfers energy to the nearby quenching dye molecule (TAMRA), which has a much weaker emission, resulting in a nonfluorescent probe. During PCR, when the polymerase replicates a template on which a TaqMan probe is bound, the 5' exonuclease activity of the polymerase cleaves the probe. This separates the fluorescent and quenching dyes and FRET no longer occurs. Fluorescence increases in each cycle, proportional to the rate of probe cleavage.

Probes used as molecular beacons are used in a similar manner to TaqMan, but exonuclease activity is not required. Molecular beacons also contain fluorescent and quenching dyes and utilize FRET, but the two dyes are incorporated within the probe and are not accessible to Taq's exonuclease function. Molecular beacons are designed to adopt a hairpin structure while free in solution, bringing the fluorescent dye and quencher in close proximity. Upon hybridization to the target gene, the fluorescent dye and quencher are separated, interrupting the FRET and enhancing the signal from the probe. Unlike TaqMan probes, molecular beacons are designed to remain intact during the amplification reaction, and must rebind to target in every cycle for signal measurement.

c. Quantitation of Results

Two strategies are commonly employed to quantify the results obtained by real-time RT-PCR; the standard curve method and the comparative threshold method. These are discussed briefly below.

i. Standard Curve Method

In this method, a standard curve is first constructed from an RNA of known concentration, or a sequential dilution of a pooled cDNA sample. This curve is then used as a reference standard for extrapolating quantitative information for mRNA targets of unknown concentrations. Though RNA standards can be used, their stability can be a source of variability in the final analyses. In addition, using RNA standards would involve the construction of cDNA plasmids that have to be in vitro transcribed into the RNA standards and accurately quantitated, a time-consuming process. However, the use of absolutely quantitated RNA standards will help generate absolute copy number data.

In addition to RNA, other nucleic acid samples can be used to construct the standard curve, including purified plasmid dsDNA, in vitro generated ssDNA or any cDNA sample expressing the target gene. Spectrophotometric measurements at 260 nm can be used to assess the concentration of these DNAs, which can then be converted to a copy number value based on the molecular weight of the sample used. cDNA plasmids are the preferred standards for standard curve quantitation. However, since cDNA plasmids will not control for variations in the efficiency of the reverse transcription step, this method will only yield information on relative changes in mRNA expression. This, and variation introduced due to variable RNA inputs, can be corrected by normalization to a housekeeping gene.

ii. Comparative Ct Method

Another quantitation approach is termed the comparative Ct method or $2^{-\Delta\Delta Ct}$ method. This involves comparing the Ct values of the samples of interest with a control or calibrator sample such as

a non-treated sample or RNA from normal tissue. The Ct values of both the calibrator and the samples of interest are normalized to an appropriate endogenous housekeeping gene.

The comparative Ct method is also known as the $2^{-\Delta\Delta Ct}$, where

$\Delta\Delta Ct = \Delta Ct,sample - \Delta Ct,reference$

Here, "ΔCT,sample" is the Ct value for any sample normalized to the endogenous housekeeping gene and "ΔCt, reference" is the Ct value for the calibrator also normalized to the endogenous housekeeping gene. Figure 3 shows a practical example of how to use this procedure to calculate fold change in gene expression.

For the ΔΔCt calculation to be valid, the amplification efficiencies of the target and the endogenous reference must be approximately equal. This can be established by looking at how ΔCt varies with template dilution. If the plot of cDNA dilution versus delta Ct is close to zero, it implies that the efficiencies of the target and housekeeping genes are very similar. If a housekeeping gene cannot be found whose amplification efficiency is similar to the target, then the standard curve method is preferred.

3. Methods and approaches for examining mRNA accumulation for multiple transcripts.

We will discuss one of the more common procedures to identify the differences between two cell populations, the gene expression microarray. This approach is often referred to as "transcript profiling" or "transcriptonomics" and is a powerful tool to understand the broad range of affects that initiated by a particular treatment.

	A	B	C	D	E	F	G	H	I
1	**Treatment**	**House-keeping (HK) gene Ct**	**Average CT**	**Target gene Ct**	**Average CT**	**ΔΔCt**	**$2^{-\Delta\Delta Ct}$**	**Average**	**SEM**
2	Control	13.73	13.68	30.84	30.54	0.26	0.83701961	1.03	0.12
3	Control	13.55		30.77		0.37	0.77557238		
4	Control	13.83		30.1		-0.58	1.49830708		
5	Control	13.68		30.47		-0.06	1.04487715		
6	Control	13.59		30.79		0.35	0.78639897		
7	Control	13.71		30.24		-0.32	1.25121814		
8	Treated	14.18		30.52	30.0283333	-0.51	1.42734425	1.93	0.23
9	Treated	13.55		29.15		-1.25	2.38391589		
10	Treated	14.4		29.86		-1.39	2.62684911		
11	Treated	14.2		30.05		-1.00	2.00462632		
12	Treated	13.59		29.45		-0.99	1.99077936		
13	Treated	14.49		31.14		-0.20	1.15135548		
				$\Delta\Delta Ct=(Ct, target-Ct, HK)_{sample}$ - $(Ct, target-Ct, HK)_{avg.control}$					
				ΔΔCt=(D2-B2)-(E2-C2)					

Figure 3. Quantitation of gene expression by comparative Ct

a. Microarray approach

The development of the high density microarray heralded the wide-use of gene expression profiling and the ability to routinely examining gene expression "on a genomic scale". The basic approach to microarrays is similar to that of the "reverse dot-blot", but incorporates several

technological innovations. The production of the microarray generally requires sophisticated equipment. In the spotted array, the probes (prepared from PCR amplification of cloned genes or synthetic oligonucleotides) are arrayed onto coated slides using a robotic device that could print thousands of spots in an 18 mm² area (Figure 4, panel A). The second approach synthesizes probes *in situ* on nylon or glass surfaces using photochemical synthesis or an ink-jet printing process. With photolithographic masks, oligonucleotides of different sequence may be synthesized in 25 µm² areas, yielding a chip with thousands of probes (Figure 4, panel B). As designed by Affymetrix these GeneChip® arrays make use of several (16-20) 20-mer probe pairs for each RNA being monitored, a pair consisting of an oligonucleotide complementary to the RNA, and another with one mismatched base in the sequence. The mismatch probe for each pair serves as an internal control for hybridization specificity. A particular strength of the ink-jet approach is the ability to create different array configurations rather simply, allowing rapid optimization of probe oligonucleotide sequence as well as creation of many custom arrays.

The generation of labeled probe material from RNA utilizes similar approaches regardless of the type of high density microarray (Figure 4, panel C). Reverse transcriptase reaction is performed with fluorescently-labeled nucleotides (or may be post-labeled); RNA from one condition is labeled with one dye (Cy3) while the RNA from the second sample is labeled with a second dye (Cy5). The two probe samples are then competitively hybridized to the cDNA microarray. Finally, the fluorescence at each spot on the slide is determined with a confocal laser scanning microscope; the ratio of the intensities for the two dyes at each spot gives the ratio of abundance for that transcript in the two conditions. The only significant variation with the photolithographic systems (Affymetrix) is that RNA samples are converted to a cDNA using an oligo-dT primer that also incorporates a bacteriophage T7 RNA polymerase promoter sequence. This allows subsequent production of "complementary RNA" in the presence of fluorescent-labeled nucleotides. Also, unlike the approach with the spotted microarrays, only one labeled sample is hybridized at a time. The analysis of spot intensities over all

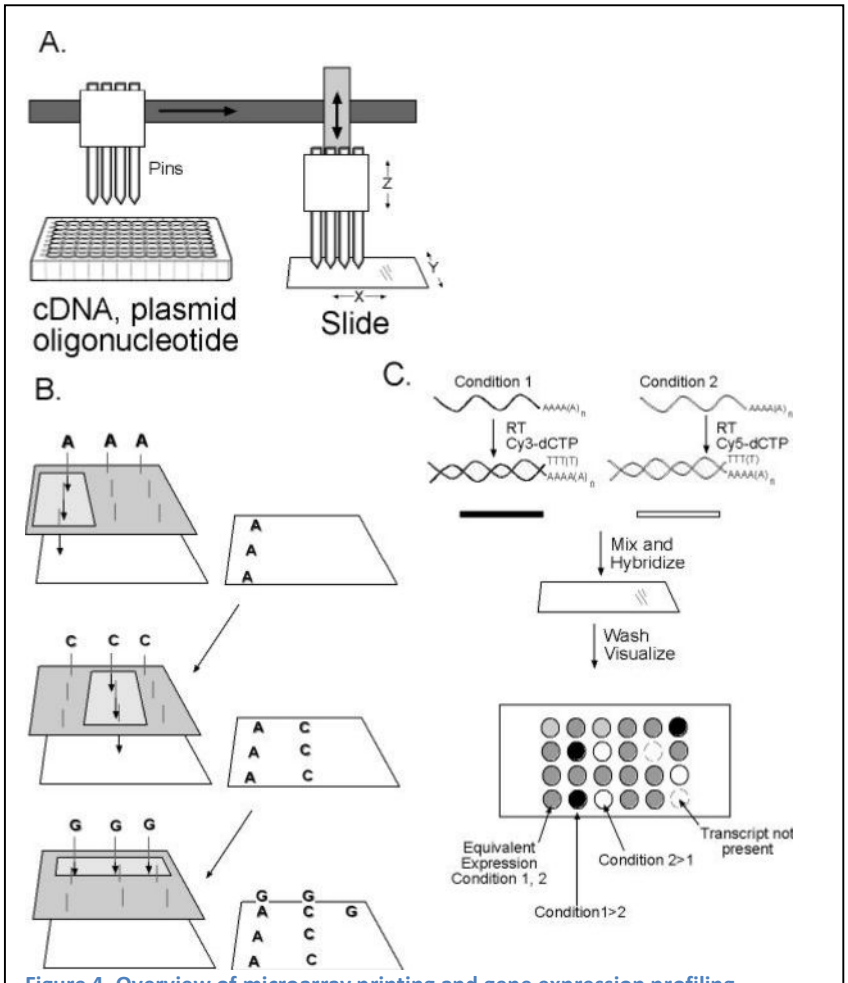

Figure 4. Overview of microarray printing and gene expression profiling

the probe pairs for an RNA allows quantification of the abundance of the transcript in the total sample.

There are many advantages of high density microarrays *versus* their "open-system" counterparts. The most obvious is the speed at which differentially regulated genes can be determined. If the slides are readily available, a fairly comprehensive examination of gene expression can be performed in a matter of days. Currently configurations allow for examination in excess of 30,000 genes in a single experiment. The predominant down-side of this technology is the cost. The equipment from printing to scanning to analysis is specialized and expensive. If high-density microarrays are purchased from commercial sources the cost can be hundreds of dollars for one slide. (However, this approach does save on the expense of probes and the printing equipment). This latter point is in direct conflict with the second major drawback of high-density microarrays, the need for many replicates. The variability in the experimental portions of the procedures, as well as the inherent statistical issues in dealing with thousands of data-points, requires that each condition be examined as many times as possible. Great pains must be taken to assure the proper controls have been included in the array design (internal and external controls) and that proper normalization of data has been performed.

b. Data Analysis

As mentioned above, the statistical examination of microarray data is complicated by the shear numbers of genes being examined as well as the variability from the experimental conditions. Perhaps the foremost challenge in transcript profiling data analysis is the handling of large data sets. Ideally, these data sets have not only the thousands of differential expression values, but also include functional and/or pathway information for the genes examined to allow later functional grouping. Currently, this step is complicated by the limited (but growing) amount of functional annotation for database entries and the need for standard gene and gene function nomenclature (i.e. gene ontology). The next challenge is that of determining patterns of gene expression relevant to the experimental hypotheses. Even if the question is only what genes are truly regulated under a single experimental condition, data analysis is in order. Thus, microarray experiments, despite their expense, must be replicated, with statistical methods applied to determine random variability. In terms of identifying trends in groups of transcripts, a variety of clustering methods has been described, including the original hierarchical clustering, interactive clustering, k-means clustering, and self-organizing maps. Many of these methods have been incorporated into commercially-available software packages, and all remain valuable approaches to data analysis. Several reviews of statistical methods and clustering of microarray data have been published (for example [28, 29]).

The following is a recommendation of how to proceed with data analysis, with the aim of finding biologically pertinent changes in gene expression and gene batteries under coordinate control. Of course, one must start with the simplest, best-designed model system possible. With the relative simplicity and speed with which data can be generated by microarray, researchers have become flippant with model development. This was not the case in approaches such as differential hybridization that could take months-to-years to generate transcript profiles and hence one was very cognizant of the dictum "garbage-in, garbage-out". Thus assuming that the model is appropriate and the experiment was replicated numerous times (n=3-6, for example), here is one set of analysis that can be performed with relative

ease and generate some meaningful, biologically-relevant information.

1. Determine genes that are statistically different between your two conditions. This requires the appropriate normalization of data and inclusion of the proper controls to minimize slide-to-slide variability. In the past, many have used the "two-fold rule". If the gene is either 2-fold higher or lower than the mean for the slide, in the majority of the slides, then it is considered to be different between the samples. However, this approach does not take into consideration the variability of the gene's expression. A more thorough statistical analysis would add confidence to the generation of lists of differentially regulated genes. Several software packages are available that will determine gene's whose expression is statistically different. For example GeneSpring (SiliconGenetics, Redwood City, CA) and DNAStar (Madison, WI) performs a T-test to determine if a gene's expression is different than mean, taking into account variability of the slide as well as across the slide. It is beyond the scope of this article to go into the specifics of the statistical analysis, and often the approach used depends on the investigator's comfort levels with different programs. At some point one must go from normalized, replicated data to lists of genes you think may be differentially regulated.

2. Classify the regulated genes based on function. This is the beginning of functional genomics, where the lists generated above are scrutinized to find an insight into the biological hypothesis being examined. For example, is there a particular signal transduction pathway that is key to the condition under question? Are the responsive genes coordinately affecting cell cycle or apoptosis? Some people also equate this with "hypothesis generation" *versus* "hypothesis testing". Most of the genes present on commercial arrays, or those produced from purchased oligonucleotide or clone sets, are associated with a wide range of biological information. The GenBank accession number can be used to identify a gene's molecular function, cellular localization or biological process. Information on homologues, chromosome localization and polymorphisms can also be easily retrieved. Commonly used programs for performing this functional classification simply retrieve this information and create a database that can be updated as new information is obtained and can be used to sort.

3. Verify a subset of potential target genes. It is not likely that all the genes added to the list of potential target genes can be systematically verified. Thus a subset of genes can be chosen for further scrutiny. This is where the functional classification can have an impact. A subset(s) of genes that share a common functional classification can be chosen for verification. Other factors may influence how the subset is chosen such as strength of response, importance of the product and whether or not it has been shown to be affected by this condition in other experiments. Although any method can be utilized, more and more microarray labs are utilizing real-time PCR due to its quantifiability and throughput.

4. References

Perdew GH, Vanden Heuvel JP, Peters JM. 2006. Regulation of gene expression: molecular mechanisms, pp. ix, 333. Humana Press