Transcriptomics

1. Background

Transcriptomics is a subfield of functional genomics that focuses on gene expression, with specific focus on mRNA (transcripts).

The level of gene expression (the amount of mRNA from each gene that is present in a particular cell, tissue, or organism) is a phenotype. It is sometimes considered as an "intermediate phenotype" because it is very close to the genotype in the pathway that determines the final phenotype of an organism.

DNA -> mRNA -> protein -> interactions with internal/external environment -> phenotype

Because mRNAs correspond to particular genes in the genome, it is often possible to establish a link between a genotype and an expression phenotype.

Large-scale, high-throughput methods have been developed for transcriptomic analysis, and it is now possible to examine the expression level of all genes in the genome in a particular sample, or to detect all of the genes that are expressed differently between two samples.

Some questions that can be addressed by transcriptomic methods:
a) How much transcript is there from each gene (expression level)?
b) How does expression level change over development (expression profile)?
c) How does expression differ among different tissues or between sexes?
d) How does environment/treatment affect gene expression?
e) How much variation is there in gene expression levels within natural populations?
f) How does natural (or artificial) selection affect gene expression?

The subject is covered well in Gibson and Muse's *A Primer of Genome Science*.

2. EST sequencing

mRNA is reverse-transcribed into cDNA, then a large number of cDNAs are sequenced. Typically, the full-length of the mRNA is not reverse-transcribed and the full-length of the cDNA is not sequenced. Thus, the resulting sequence fragments are referred to as ESTs (Expressed Sequence Tags). Large-scale EST sequencing projects (ten of thousands or hundreds of thousands) have been performed for model organisms, such as Drosophila, human, mouse, and Arabidopsis.

pro: gives an estimate of absolute mRNA abundance (if cDNA library is random); very useful for gene discovery (annotating expressed regions of genomes and intron/exon boundaries).

con: expensive, time-consuming, requires large-scale sequencing; much of the sequencing is redundant (ESTs from highly-expressed genes are sequenced many times); must sequence 100,000's of ESTs to get good representation genes expressed at low levels; the ESTs only reveal gene expression levels in the particular tissue or sample that was used for mRNA preparation; genes expressed in specific tissues, cells, developmental stages, *etc*. may be missed.

For example, the original EST survey of *D. melanogaster* carried out by the Berkeley Drosophila Genome Project sequenced over 80,000 ESTs. These corresponded to 6,000 non-redundant cDNAs (genes). In total, the Drosophila genome is predicted to have 15,000 genes. At present, over 240,000 ESTs have been sequenced and about 10,000 non-redundant cDNAs identified.

EST databases are often used to estimate expression levels when there is no other experimental evidence. For this, one assumes that the number of "hits" in the EST database is proportional to expression level. That is, the more times an EST corresponding to a particular gene was sequenced, the higher the expression level of that gene.

3. Microarrays

Microarrays are constructed by attaching specific DNA sequences (probes) to a solid surface (often a glass microscope slide). The probes are arranged at very high density. Typically, thousands of probes, each matching a different gene, will fit in the area covered by a microscope cover slip. The probes are often referred to as "spots" and the microarrays are often called "chips".

Different arrays use different types of DNA as probes:
a) cDNA or EST sequences
b) PCR-amplified genomic DNA
c) Oligonucleotides (36-80 bases)

The last two have the advantage that they can be made to all predicted genes in the genome, while the first one requires that a cDNA or EST has been cloned.

The last two also have the advantage that they can be specifically designed to reduce cross-hybridization (by avoiding sequence regions that are similar in two or more genes). However, the first one has the advantage of longer probe sequences, which may give better hybridization signals – especially for cross-species comparisons.

To measure gene expression, hybridizations ("hybs") are performed:
a) RNA is purified from the samples to be compared
b) mRNA is reverse-transcribed to cDNA and labeled with a fluorescent dye (one sample red, the other green)
c) the labeled cDNA solutions are placed together on the array (under a coverslip) in equal amounts and hybridized overnight
d) excess and unbound cDNA is removed by washing, the array is dried
e) array is scanned with laser scanner to create a graphical image
f) image is analyzed to determine relative expression differences (red/green ratio for each spot)

Which genes are expressed differently between two samples? There are two main approaches that are used to determine which genes are differentially expressed:

a) Fold-change – an arbitrary fold-difference is chosen to define genes that are differentially expressed. For example, using a fold change of 2 means that a gene must have an expression level that is at least 2 times higher in one sample than in the other to be considered differentially expressed. Often the ratio of expression between two samples is given on a $\log_2$

scale. In this case, genes with a $\log_2$ ratio greater than 1 or less than –1 would be differentially expressed.

b) Statistical cutoff – a statistical method, such as a t-test, ANOVA, or a Bayesian method is used to calculate a p-value for each gene. The null hypothesis is that the gene is expressed equally in the two samples. If the p-value is below a certain cutoff (critical value), then the null hypothesis can be rejected and the gene considered differentially expressed. Because probes for many genes are on the array, there is a multiple testing problem and traditional p-value cutoffs (such as $p < 0.05$) cannot be used. Often a p-value cutoff is chosen so that the rate of false positives (the fraction of significant genes that are expected due to chance) meets a certain value, such as 5%, 10%, or 20%. This is known as the false discovery rate (FDR).

The two approaches can be displayed graphically by a "volcano plot". This plot shows the fold-change in expression between two samples (on a $\log_2$ scale) on the X-axis, and the p-value (typically on a $-\log_{10}$ scale) on the Y-axis.

Summary of microarrays
pro: can quickly, cost-efficiently do many comparisons and replicates
con: do not measure absolute, but relative abundance; statistical interpretation may be difficult

4. Affymetrix "Affy" GeneChips™
The company Affymetrix produces and sells microarrays for several model species, including human, mouse, Drosophila, *C. elegans*, and Arabidopsis. These are known as GeneChips™ . The arrays are made by a process called photolithography, in which specific oligonucleotide probes (25 bases) are synthesized directly on the array surface. For each gene, 20 different probes corresponding to different regions of the transcript are present on the array. Additionally, 20 "mismatch controls" that are identical to the above probes except for a single mismatched nucleotide at the center of the sequence (base 13) are present. The expression level of each gene is estimated by the intensity difference between the match and the mismatch probes, averaged over all 20 probes per genes. Thus, there is not a competitive hybridization of two different samples. Only one sample is hybridized per array.

pro: can buy pre-made chips; high quality control; high standardization; easy to use
con: can be expensive; requires Affymetrix machines; short probes (25 bases) are not good for divergent species; useful only for model species for which a GeneChip is commercially available.

5. SAGE
SAGE (Serial Analysis of Gene Expression) is a method that is similar to EST sequencing, but is more efficient because only short "tags" of around 15 bases are sequenced from each cDNA. Before sequencing, the tags are concatenated so that many of them can be sequenced in a single sequencing reaction. It is necessary to have a good sequence and annotation of the genome, so that the tags can be accurately mapped back to their corresponding genes.

Summary of the procedure:
a) Purify mRNA (poly-A) from sample
b) Use biotinylated oligo dT to make ds cDNA
c) cut cDNA with 4-cutter restriction enzyme (*Nla*III) - CATG, cuts on avg. every 256 bp
d) purify only 3' poly dT ends in a streptavidin column

e) ligate adapter to cut end. Adapter has restriction site for *Bsm*FI GGGAC (cuts 15 bp from site into cDNA fragment)
f) ligate adapter ends tail-to-tail (ditags). PCR amplify with primers to adapter sequence
g) cut again with *Nla*III to remove adaptors, leaving 30 bp ditag
h) ligate many ditags end-to-end (up to 1-kb), then sequence 1000's of these @ 30-40 tags/reaction.

Each 15-bp tag should give a unique match to a transcript in the genome (the odds of a match at random are $\approx 1/4^{15}$ or ≈ 1 in a billion). Furthermore, it should come just after the 3' most *Nla*III site in a gene. A few genes may be missed if they have no *Nla*III site or if the site is too far or too close to the polyA tail.

To quantify expression of a gene, simply count the number of times that the tag for that gene is sequenced. Requires at least 10,000-50,000 tags for accurate estimate (≈300-1000 sequencing reactions).

pro: gives an estimate of absolute transcript abundance; more efficient than large-scale EST sequencing
con: requires much sequencing, which can be expensive; not accurate for rare transcripts; sometimes difficult to map tags to genes; must be repeated for each sample (tissue, sex, treatment, *etc*.)

6. RNA-seq

High throughput RNA sequencing (RNA-seq) follows the same scheme outlined above for EST sequencing. The difference is that a pool of cDNA is used for next generation sequencing. With this approach hundreds of millions of short EST sequences (50-250 bases in length) can be generated quickly and at low cost. These sequences can be mapped back to their corresponding gene in the genome and used to quantify gene expression. RNA-seq can also be applied to species with unknown or un-annotated genomes to assemble the transcriptome *de novo*.

pro: produces direct "counts" of gene expression with very large sample sizes (number of reads per gene), which is good for statistical analysis and comparison of expression between samples. Can be applied to non-model organisms. Can assemble transcriptomes *de novo*.
con: may be more expensive than microarray technologies (but the costs of RNA-seq are dropping rapidly). Requires more complex bioinformatic analysis than arrays.