RAMA UNIVERSITY

www.ramauniversity.ac.in

FACULTY OF AGRICULTURE SCIENCES AND ALLIED INDUSTRIES

(Principles of Biotechnology)

For

M.Sc. Ag (GPB)

Course Instructor

Dr Shiv Prakash Shrivastav

FASAI(Genetics and Plant Breeding)

Rama University, Kanpur

Transcriptomics

1. Background

Transcriptomics is a subfield of functional genomics that focuses on gene expression, typically with a focus on mRNA (transcripts), although non-coding RNA can also be analyzed.

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?

2. EST sequencing

This was the first gnome-wide method used to investigate gene expression.
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 a good representation of 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 (BDGP) sequenced over 80,000 ESTs. These corresponded to 6,000 non-redundant cDNAs (genes). In total, the *Drosophila* genome is predicted to have around 14,000 genes. Thus, cDNAs were cloned and sequenced for less than half of the genes in the genome. At present, over 240,000 ESTs have been sequenced by BDGP and about 10,000 non-redundant cDNAs have been cloned and sequenced (70% of all genes).

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, many 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 sometimes called "chips".

Microarrays can use different types of DNA as probes, including:
a) cDNA or EST sequences
b) PCR-amplified genomic DNA
c) Synthesized oligonucleotides (typically 36–80 bases long)

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 for each gene that is used.

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. For example, if a microarray designed for *D. melanogaster* is used to measure gene expression in *D. simulans*.

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 same microarray (under a coverslip) in equal amounts and hybridized overnight
d) the excess and unbound cDNA is removed by washing, the array is dried
e) the array is scanned with laser scanner to create a graphical image
f) image is analyzed to determine relative expression differences (red/green signal intensity 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 considered as differentially expressed.

b) Statistical cutoff – a statistical method, such as a t-test, binomial 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 thousands of 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). Very often a FDR of 5% is used.

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 long) 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 10–15 bases are sequenced from each cDNA. Before sequencing, the tags are concatenated so that many of them can be sequenced in a single Sanger sequencing reaction. For this method, 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 primer to synthesize double-stranded cDNA
c) cut cDNA with a restriction enzyme, such as *Nla*III which recognizes the sequence CATG and cuts, on average, every 256 bp
d) purify only the 3' poly dT ends of the cut cDNA in a streptavidin column (binds to biotin attached to the oligo dT primer)
e) ligate an adapter (short synthesized DNA sequence) to the cut end. The adapter contains a restriction site for the restriction enzyme *Bsm*FI (recognizes GGGAC, but cuts 15 bp away from this sequence into the cDNA fragment)
f) ligate two adapter ends to each other tail-to-tail to create "ditags". PCR amplify the ditags with primers complementary to the to adapter sequence
g) cut again with *Nla*III to remove adaptors, leaving a 30 bp ditag
h) ligate many ditags end-to-end (up to 1 Kb total length), then sequence 1000's of these. You can typically sequence 30-40 tags per Sanger sequencing 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 from (or too close to) the polyA tail.

To quantify the expression level of a gene, simply count the number of times that the tag for that gene is sequenced. At least 10,000–50,000 tags should be sequenced to get an accurate estimate of expression (≈300–1000 sequencing reactions).

Pro: gives an estimate of absolute transcript abundance; more efficient than large-scale EST sequencing, because many fewer sequencing reactions are required.

Con: still 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 (usually 50–300 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.