Transcription

Estimates from ENCODE

How Much of the Genome is Transcribed?

Why Do We Care About Gene Expression?

• Gene expression is the process by which information from a gene is used in the synthesis of a functional product

• Can be modulated
  • Timing
  • Location
  • Amount

• Properties of gene expression contribute to an organism’s phenotype
Why Do We Measure Gene Expression?

• Gene expression profiles provide a snapshot of cellular activity at the molecular scale
• Gene expression profiles provide a snapshot of cumulative interactions of many hard to detect phenomena
• Gene expression is a ‘proxy’ measure for transcription/ translation/ functional events

What Do We Assume About Gene Expression?

• Assume that gene expression levels correspond to functional product levels
• Assume that a normal cell has a standard expression profile/ signature
• Assume that changes in expression profile indicate that something important is happening

Does mRNA Level == Protein Level?

• Comparing protein levels (MS, gels) and RNA levels (SAGE) for 156 genes in yeast (Gygi et al., 1999 Mol Cell Biol)
  • Found that correlation between mRNA and protein levels were insufficient to predict protein levels
• Comparing protein levels (MS, ICAT) and RNA levels (microarray) for 245 genes in yeast on galactose/ ethanol medium (Griffen et al., 2002 Mol Cell Proteomics)
  • Found a significant number of genes show large discrepancies between abundance ratios when measured at the levels of mRNA and protein expression
mRNA vs. Protein Levels in Yeast


Relationships between mRNA transcript abundance and protein abundance in "omics" profiling experiments

Diversity of Regulation
... So Why Do We Measure Gene Expression ... ?

• It’s easier to measure than proteins, metabolites, phenotypes ...

" .. [transcript abundance] doesn’t tell us everything, but it tells us a lot more than we knew before ... "

— Dr Patrick Brown, Stanford Microarray Pioneer

Lecture 1:
• Measuring Gene Expression
  • Targeted
    • Northern
    • RT-PCR
    • In situ Hybridization
  • Reporter Assay
  • High-throughput
    • Microarray
• General Workflow

Lecture 2:
• Transcriptomics
  • RNAseq

OG Way of Measuring Gene Expression: Northern blot
Measuring Gene Expression: Northern blot

**Advantages**
- Inexpensive
- Very sensitive due to use of radioactive probes
- Nearly infinite dynamic range

**Disadvantages**
- Radioactive labeling
- Quality control (non-specific hybridization)
- "Old-fashioned"

---

Measuring Gene Expression: real time qPCR

**Advantages**
- Highly sensitive, quantitative and reproducible
- 'Gold-standard'
- Excellent dynamic range

**Disadvantages**
- Expensive
- Not high-throughput
- Non-specific amplification can lead to false positives
Measuring Gene Expression: *In situ* Hybridization

**Advantages**
- Single-cell sensitivity
- Spatial and temporal analysis
- Can be applicable to archival tissues

**Disadvantages**
- Expensive
- Not high-throughput
- Time consuming; hybridization and permeabilization issues

Measuring Gene Expression: Reporter Assay
Measuring Gene Expression: Reporter Assay

**Advantages**
- *In vivo* applications
- Highly sensitive
- New technology enables longitudinal studies

**Disadvantages**
- Stability issues
- Not high-throughput
- Quantification *in vivo* is affected by many variables

Measuring Gene Expression: Microarray

- Analyze gene expression (mRNA) through large-scale hybridization analysis
- Basically, a high-throughput, non-radioactive Northern technique that uses fluorescence for detection

Microarray: Experimental Steps

- Image processing
  - Laser 1
  - Laser 2
  - Green channel
  - Red channel
  - Overlay images and normalize
  - Image process and analyze
  - Microarray scan on chip = 10 pixels in diameter
Measuring Gene Expression: Microarray

**Advantages**
- High-throughput
- Avoids radioactivity
- Kit systems (easy/accessable)

**Disadvantages**
- Quality and quality control highly variable
- Cross-hybridization
- Analysis and interpretation can be difficult

---

**Experimental Design and General Workflow**

1. How much risk can I afford to take?
2. How much can I trust various components in an experiment?
3. How thoroughly should I plan my experiment?
4. Carrying out the experiment

“Seventy percent of whether your experiment will work is determined before you touch the first test tube” - Sun, Nature Reviews Mol Cell Biol (2004)

---

**General Workflow: Replicates**

What type of replicates?

[Diagram illustrating experimental design]
General Workflow: Power

- Power Analysis
  - A Priori
    - How big a sample size? Cost? Usefulness of experiment?
  - Post-Hoc
    - Can the results be trusted?

The higher the power, the higher the probability of correctly rejecting the null hypothesis.

Errors in Hypothesis Testing

<table>
<thead>
<tr>
<th>Null Hypothesis (H₀) is True</th>
<th>Null Hypothesis (H₀) is False</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject Null Hypothesis</td>
<td>Type-I Error (α) Common values: 0.05, 0.01</td>
</tr>
<tr>
<td></td>
<td>Power (1-β) Common values: 0.8, 0.9</td>
</tr>
<tr>
<td>Fail to Reject Null Hypothesis</td>
<td>Confidence interval (1-α) Common values: 0.95, 0.99</td>
</tr>
<tr>
<td></td>
<td>Type-II Error (β)</td>
</tr>
</tbody>
</table>

Type-I error: probability of finding an effect that is not there
Type-II error: probability of not finding an effect that is there

Parameters for Power Analysis

- Experimental Design
- Level of Significance
- Sample Size
- Effect Size
- Methods
- ... others
Choosing a Sample Size

\[ N = \frac{4\sigma^2(Z_{\alpha/2} + (1 - \beta))^2}{D^2} \]

- \( N \) = total sample size (of both comparison groups)
- \( \sigma \) = the SD of each group (assumed to be equal for both groups)
- \( Z_{\alpha/2} \) = critical value for desired \( \alpha \) (from a table, e.g., 0.05 = 1.96)
- \( (1 - \beta) \) = power
- \( D \) = minimum expected difference between the two means

In a study comparing two groups, the power (sensitivity) of a statistical test must be sufficient to enable detection of a significant difference between the groups if a difference is truly present.

Choosing an Effect Size

\[ d = \frac{x_1 - x_2}{s} \]

Cohen’s \( d \) = the difference between two means divided by the pooled standard deviation

<table>
<thead>
<tr>
<th>Effect Size</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very Small</td>
<td>0.01</td>
</tr>
<tr>
<td>Small</td>
<td>0.2</td>
</tr>
<tr>
<td>Medium</td>
<td>0.5</td>
</tr>
<tr>
<td>Large</td>
<td>0.8</td>
</tr>
<tr>
<td>Very Large</td>
<td>1.2</td>
</tr>
<tr>
<td>Huge</td>
<td>2.0</td>
</tr>
</tbody>
</table>

Balancing Sample and Effect Size is Critical

- **Impact of sample size on power**
- **Impact of effect size on power**
General Workflow: RNA Extraction

- RNA extraction
  - RNA is unstable
  - Extracted as total RNA
    - Only 1-2% is mRNA
    - Remainder is rRNA, tRNA, etc
  - RNA extracted from tissue is very heterogeneous
    - Many cells and cell types

Intact vs. Degraded RNA

General Workflow: Normalization

- Try to get a slope ~1 and a correlation of ~1
- Reduces systematic (multiplicative) differences between two channels of a single hybridization or differences between hybridizations
Log Transformation

Channel 2 is too strong and biases the results when the two channels are combined.

Before Normalization

Microarray: Toy Normalization Example

Channel 2 is too strong and biases the results when the two channels are combined.

Before Normalization

Microarray: Toy Normalization Example

Channel 2 is too strong and biases the results when the two channels are combined.

Before Normalization

Microarray: Toy Normalization Example

Channel 2 is too strong and biases the results when the two channels are combined.
After Normalization

Add 0.84 to every value in ch1 to normalize ch1 log2 signal intensity

\[ y = f(x) = x + 0.84 \]

Microarray: Toy Normalization Example

Channel 2 and Channel 1 are Balanced After Normalization

Correlation and Outliers

Experimental error or something important?
General Workflow: Differential Expression

Fisher’s Linear Discriminant: What genes vary the most between your experimental groups?

$$\text{FLD}(X) = \frac{(\overline{x}_1 - \overline{x}_2)^2}{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}$$

where:
- $\overline{x}_1$ = average normalized counts of gene X in group 1
- $\overline{x}_2$ = average normalized counts of gene X in group 2
- $\sigma_1$ = standard deviation of gene X in group 1
- $\sigma_2$ = standard deviation of gene X in group 2

Good class separation within between

General Workflow: Multiple Testing Problem

- Measure the expression of thousands of genes at a time
- Many thousands of statistical tests are performed
  - For each gene we are testing the null hypothesis that there is no differential expression
- Suppose g null hypotheses are being tested
  - The family-wise error rate (FWER) is the probability of rejecting at least one null hypothesis given they are all true (1 false positive)
- Bonferroni Correction
  - $\alpha \rightarrow \frac{\alpha}{g}$
- False Discovery Rate
  - $\pi_{FDR} \rightarrow \alpha$
General Workflow: Can you Identify Patterns?

- Clustering
  - Measure similarity or dissimilarity between objects
  - Assign a threshold value to decide whether an object belongs to a cluster
  - Measures ‘distance’ between two clusters
  - Requires a seed
    - An object to begin the clustering process

General Workflow: Clustering Algorithms (there are many!)

- K-means or Partitioning Methods - divides a set of n objects into k clusters -
  - with or without overlap
  - Assign every object to nearest cluster center
  - Move each cluster center to the mean of its assigned genes
  - Repeat until convergence

- Hierarchical Methods - produces a set of nested clusters in which each pair of objects is progressively nested into a larger cluster until only one cluster remains
  - Find the two closest objects and merge them into a cluster
  - Find and merge the next two closest objects
  - Repeat until one cluster remains

K-means Clustering Example
Heirarchical Clustering Example

Initial cluster                  pairwise               select                   select
compare                     closest             next closest

Hierarchical Clustering

Find 2 most similar gene
express levels or curves

Find the next closest pair
of levels or curves

Iterate

A
B
C
D
E
F

Heat map

General Workflow: Summing It All Up

- Study design and RNA extraction
  - What technology is most appropriate?
  - Power?
- Data QC
  - Are there technical issues? How correlated are your replicates?
- Normalization
  - Choose the method that is appropriate for your experiment.
- Explore your data
  - Calculate distances among your samples. Do the patterns fit your expectations?
- Calculate differential expression and look for patterns
  - Choose the method that is most appropriate for your experiment.
- Interpret the results in terms of biological or physiological knowledge
  - YOU must believe in your data…