Lecture 3: Differential Expression

IBC2004, 11 July 2004

Gordon Smyth

What’s Your Question?

- What are the targets genes for my knock-out gene?
- Gene discovery, differential expression
- Is a specified group of genes all up-regulated in a specified condition?
- Gene set differential expression
- Can I use the expression profile of cancer patients to predict chemotherapy outcome?
- Class prediction, classification
- Are there tumour sub-types not previously identified? Do my genes group into previously undiscovered pathways?
- Class discovery, clustering

This talk covers first question - differential expression

Log-Ratios or Single Channel Intensities?

- Tradition analysis, as here, treats log-ratios \( M = \log(R/G) \) as the primary data, i.e., gene expression measurements are relative
- Alternative approach treats individual channel intensities \( R \) and \( G \) as primary data, i.e., gene expression measures are absolute (Wolfgang, Churchill, Kerr)
- Single channel approach makes new analyses possible but
  - makes stronger assumptions
  - requires more complex models (mixed models in place of ordinary linear models) to accommodate correlation between \( R \) and \( G \) on same spot
  - requires absolute normalization methods

Replicate Arrays

Need replication to do statistical analysis of differential expression. Most basic possible is series of arrays all comparing the same two RNA sources:

\[
\begin{align*}
&\text{A} \quad \text{B} \\
&\text{A} \quad \text{B} \\
&\vdots \quad \vdots \\
&\text{A} \quad \text{B}
\end{align*}
\]

\( n \) arrays

Gene-wise Summaries

- Each gene gives a series of log-ratios
- Summarize log-ratios by average and standard deviation for each gene, or robust versions of these:

\[
M_1, \ldots, M_n
\]

\[
\bar{M} = \text{ave } M \quad s = \text{std.dev } M
\]

Ranking Criteria

- Average log fold change. Problem: non DE genes with large variances have too much chance of being selected.

\[
t_g = \frac{\bar{M}}{s / \sqrt{n}}
\]

- t-statistics. Problem: genes with very small sample variances are suspect

- Moderated t-statistics. We use a happy compromise between the two

\[
t_g = \frac{\bar{M}}{s / \sqrt{n}}
\]
Moderated t-Statistics

Shrunken standard deviations

\[ \hat{s}_g^2 = \frac{s_g^2 d_g + s_0^2 d_0}{d_g + d_0} \]

Moderated t-statistics

\[ \hat{t}_g = \frac{\hat{M}_g}{\hat{s}_g \sqrt{C}_g} \]

Eliminates large t-statistics merely from very small \( \sigma \)

Shrinkage of Standard Deviations

The data decides whether \( \hat{t}_g \) should be closer to \( \hat{t}_g \) pooled or to \( \hat{t}_g \)

Posterior Odds

Posterior probability of differential expression for any gene is

\[ \frac{p(\beta = 0 \mid \mathbf{R}, \mathbf{s}^2)}{p(\beta = 0 \mid \mathbf{R}, \mathbf{s}^2)} = \frac{p}{1 - p} \left( \frac{\hat{c}}{c + c_0} \right)^{1/2} \left( \frac{\hat{t}_g^2 + d + d_0}{\hat{t}_g^2 + \hat{c} + c_0 + d + d_0} \right)^{1/2} \]

Monotonic function of \( \hat{t}_g^2 \) for constant \( d \)

Estimating Hyper-Parameters

Closed form estimators with good properties are available:
- for \( s_g \) and \( d_g \) in terms of the first two moments of log \( s^2 \)
- for \( c_0 \) in terms of quantiles of the \( |\hat{t}_g| \)

Simulations

- Analyse all arrays together combining information in optimal way
- Combined estimation of precision
- Extensible to arbitrarily complicated experiments
- Design matrix: specifies RNA targets used on arrays
- Contrast matrix: specifies which comparisons are of interest

Linear Models
Designs → Linear Models

\[
\begin{align*}
\beta_1 &= B - A \\
\beta_2 &= C - A
\end{align*}
\]

Allows all comparisons to be estimated simultaneously.

Slightly larger example:

<table>
<thead>
<tr>
<th>WT.P1</th>
<th>MT.P1</th>
<th>WT.P11</th>
<th>MT.P11</th>
<th>WT.P12</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu )</td>
<td>( \mu ) + ( a_1 )</td>
<td>( \mu ) + ( a_1 )</td>
<td>( \mu ) + ( a_1 )</td>
<td>( \mu ) + ( a_1 )</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td>6</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\gamma_i &= \begin{pmatrix} y_i \\
\end{pmatrix} \\
\beta &= \begin{pmatrix} \beta_1 \\
\beta_2 \\
\end{pmatrix} \\
\end{align*}
\]

\[
\begin{align*}
\begin{pmatrix} y_1 \\
y_2 \\
y_3 \\
y_4 \\
y_5 \\
y_6 \\
y_7 \\
\end{pmatrix} &= \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 1 & 1 & 1 & 0 \\
\end{pmatrix} \begin{pmatrix} \beta_1 \\
\beta_2 \\
\end{pmatrix} \\
\end{align*}
\]

Matrix Multiplication

\[
\begin{align*}
\gamma_i &= \begin{pmatrix} 1 \\
-1 \\
1 \\
0 \\
0 \\
-1 \\
0 \\
1 \\
\end{pmatrix} \\
\beta &= \begin{pmatrix} \beta_1 \\
\beta_2 \\
\end{pmatrix} \\
\end{align*}
\]

\[
\begin{align*}
\begin{pmatrix} \gamma_1 \\
\gamma_2 \\
\gamma_3 \\
\gamma_4 \\
\gamma_5 \\
\gamma_6 \\
\gamma_7 \\
\end{pmatrix} &= \begin{pmatrix} \beta_1 \\
\beta_2 \\
\beta_3 \\
-\beta_1 \\
-\beta_2 \\
-\beta_3 \\
\end{pmatrix} \\
\end{align*}
\]

Contrast \( \beta_2 - \beta_1 \equiv C - B \)

Linear Model Estimates

Obtain a linear model for each gene \( g \)

\[
E(\gamma_g) = X\beta_g, \quad \text{var}(\gamma_g) = W_g^{-1}\sigma_g^2
\]

Estimate model by robust regression, least squares or generalized least squares to get

- coefficients: \( \hat{\beta}_g \)
- standard deviations: \( s_g \)
- standard errors: \( se(\hat{\beta}_g)^2 = c_g s_g^2 \)

Parallel Inference for Genes

- 10,000-40,000 linear models
- Curse of dimensionality: Need to adjust for multiple testing, e.g., control family-wise error rate (FWE) or false discovery rate (FDR)
- Boon of parallelism: Can borrow information from one gene to another

Hierarchical Model

Prior

\[
\begin{align*}
\beta_g &\sim N(\mu_g, \sigma_g^2) \\
\sigma_g^2 &\sim \sigma_{\tau g}^2 \chi^2_d \\
\end{align*}
\]

Reparametrization of Lönstedt and Speed 2002

Normality, independence assumptions are wrong but convenient, resulting methods are useful
### Posterior Statistics

Posterior variance estimators

\[ \hat{s}_g^2 = \frac{s^2_d + s^2_0}{d_g + d_0} \]

Moderated t-statistics

\[ \tilde{t}_{gi} = \frac{\hat{\beta}_{gi}}{s_g \sqrt{c_{gi}}} \sim t_{d_g + d_0} \]

Eliminates large t-statistics merely from very small \( s \)

### Estimating Hyper-Parameters

Closed form estimators with good properties are available:

- for \( s_g \) and \( d_g \) in terms of the first two moments of \( \log s^2 \)
- for \( c_{gij} \) in terms of quantiles of the \( |\tilde{t}_g| \)

### Within-Array Replicate spots

- Replicate spots of each gene on same array, assume duplicates at regular spacing
- Assume spatial component of correlation between duplicates is same for each gene
- Estimate spatial correlation from consensus estimator across genes
- Greatly improves estimation of precision

### How many genes are differentially expressed?

Assigning absolute significance levels on the basis of probability models is problematic:

- Log-ratios don’t appear to be normally distributed, hard to check
- Log-ratios for different genes are correlated in unknown way
- High level of multiple testing means that very small p-values are required – distributional assumptions must hold in extreme tail
- Little opportunity for usual CLT results to apply

### Ranking Easier Than Testing

- If there was only one gene, a t-test would give a reliable P-value for judging whether the true log-ratio was zero
- With many genes, computed P-values cannot be trusted (unless have > 16 arrays)
- It is more realistic to rank the genes in order of evidence for differential expression

### In Search of Truth

- We treat moderated t and posterior odds as ranking criteria rather than as providers of absolute significance or posterior odds
- To rigorously estimate type I or type II error rates or to compare competing analysis methods, need to construct microarray data with known truth
Spike-In Controls

- Print artificial genes on microarray, then add spike-ins of corresponding RNA in known quantities to sample RNA before labelling and hybridization
- ScoreCard and SpotReport are commercial systems in use at the AGRF

How well do moderated t-statistics distinguish ratio (DE) controls from calibration (non-DE) controls?

Empirical Cut Off

\[ |\tilde{t}_g| > 4 \]

appears a conservative rule for differential expression with very small false discovery rate for this data

Implications for Design

- Given linear modelling approach, can compute efficiency of various experimental designs
- Need to specify which RNA sources are to compared and which contrasts are of interest
For $k = 3$, efficiency ratio (Design I(a) / Design II) = 3
In general, efficiency ratio = $2k / (k-1)$

In general, efficiency ratio = $k / (k-1)$

For $k = 3$, efficiency ratio (Design I(b) / Design II) = 1.5

Design Choices for $2 \times 2$ Factorial

<table>
<thead>
<tr>
<th>Slides</th>
<th>Indirect</th>
<th>A balance of direct and indirect</th>
</tr>
</thead>
<tbody>
<tr>
<td>N=2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>T2</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>T3</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>T4</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>N=2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>T2</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>T3</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>T4</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>N=2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>T2</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>T3</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>T4</td>
<td>D</td>
<td>D</td>
</tr>
</tbody>
</table>

Table entry: variance

B Cell Lineage Commitment

- Pax5 is a critical gene for B cell development
- Enables development along the B cell lineage and simultaneously inhibits other pathways
How Does Pax5 Work?

- Design a microarray experiment to identify genes downstream from Pax5 in the molecular pathways.

RNA Sources

- Compare RNA from 4 sources:
  - Pax5-/- (knock-out cell line)
  - Rag1-/- (knock-out cell line)
  - Wt ("wild type", i.e., normal)
  - Wt cells with IL-7 removed after initial development commenced
- Rag1-/- and IL-7 removal identify genes turned on or off by halted development rather than by Pax5.

Regression Analysis

- Choose 3 comparisons between the 4 RNA sources to be the coefficients of the linear model, e.g.,
  - PW: Pax5-/- vs Wt
  - RW: Rag1-/- vs Wt
  - IW: IL-7 withdrawn vs Wt
- For each gene, fit a linear model with a coefficient for each contrast
- Any other comparisons of interest can be extracted from the linear model as contrasts

Halted Development

- B cell development can be halted at the pro B stage by:
  - Absence of the Pax5 gene
  - Absence of the Rag1 gene (which activates recombination)
  - Withdrawal of the regulatory cytokine IL-7 (essential growth factor)

Saturated Design

![Diagram showing a saturated design with comparisons between Wt, Pax5-/-, Rag1-/-, and IL-7 removed]
What about Duplicate Spots?

- p between duplicate M values on the same slide
- Gene X: \[ M_{ij} \] \( i \neq j \) and \( X \neq X \)
- p = 0.85
- Use gls procedure in R to fit linear model allowing for correlated spots

RT-PCR Confirmation of DE Genes

<table>
<thead>
<tr>
<th>cDNA</th>
<th>CD19</th>
<th>CD24 (HSA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cDNA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hoxa10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>embigin</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3135e11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IGF2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tulp4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hprt</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

10/15 array positives confirmed by RT-PCR

LIMMA Package for R

- Linear models for microarray data. A software package for the R programming environment. Focus is differential expression including
  - moderated t-statistics
  - methods for duplicate spots
  - classifying F-tests
  - stemmed heat diagrams
- Available from www.bioconductor.org

Acknowledgements

WEHI Biostatistics
- Tony Speed
- Matt Ritchie
- Natalie Thorne
- James Wettenhall
- everyone else

WEHI Scot Lab
- Joelle Michaud
- Catherine Carmichael
- Robert Escher
- Hamish Scott

ASRF
- Steve Wilcox
- Cathy Jensen
- Melanie O’Keefe

WEHI Immunology
- Steve Nutt
- Lynn Girard
- Melissa Holmes
- Mireille Lahoud

Berkeley
- Jean Yang
- Speed Lab
- Ngai Lab