Analysis of Complex Experiments

Statistical Methods in Microarray Analysis Tutorial
Institute for Mathematical Sciences
National University of Singapore
January 3, 2004

Gordon Smyth
Walter and Eliza Hall Institute

1

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

2

Types of microarrays in this talk

- Linear modelling approach in this talk applies to both single channel (Affymetrix) and two-colour arrays
- Need to cover some special features of two-colour arrays
- The examples are two-colour
- Two colour with common reference is virtually equivalent to single channel from an analysis point of view

3

Linear Models

- 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

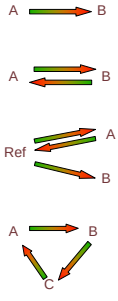
4

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 (Wolfinger, Churchill, Kerr)
- Single channel approach makes new analyses possible but
  - make 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

5

Linear Models for Differential Expression



$$y = \log_2(R) - \log_2(G) \equiv B - A$$

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \beta \qquad \beta \equiv B - A$$

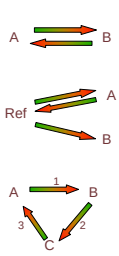
$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \qquad \begin{aligned} \beta_1 &\equiv A - \text{Ref} \\ \beta_2 &\equiv B - A \end{aligned}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \qquad \begin{aligned} \beta_1 &\equiv B - A \\ \beta_2 &\equiv C - A \end{aligned}$$

Allows all comparisons to be estimated simultaneously

6

Matrix Multiplication



$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \beta = \begin{pmatrix} \beta \\ -\beta \end{pmatrix} \qquad \beta \equiv B - A$$

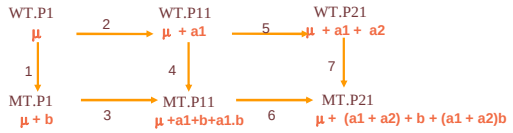
$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ -\beta_1 \\ \beta_1 + \beta_2 \end{pmatrix} \qquad \begin{aligned} \beta_1 &\equiv A - \text{Ref} \\ \beta_2 &\equiv B - A \end{aligned}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ -\beta_1 + \beta_2 \\ -\beta_2 \end{pmatrix} \qquad \begin{aligned} \beta_1 &\equiv B - A \\ \beta_2 &\equiv C - A \end{aligned}$$

Contrast $\beta_2 - \beta_1 \equiv C - B$

7

Slightly larger example:



$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{pmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{pmatrix} a_1 \\ a_2 \\ b \\ a_1 b \\ a_2 b \end{pmatrix}$$

8

Linear Model Estimates

Obtain a linear model for each gene *g*

$$E(\underset{\sim}{y}_g) = X \underset{\sim}{\beta}_g \quad \text{var}(\underset{\sim}{y}_g) = W_g^{-1} \sigma_g^2$$

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

coefficients $\hat{\beta}_{gj}$

standard deviations s_g

standard errors $\text{se}(\hat{\beta}_{gj})^2 = c_{gj} s_g^2$

9

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

10

Hierarchical Model

Normal Model

$$\hat{\beta}_{gj} \sim N(\beta_{gj}, c_{gj}\sigma_g^2)$$

$$s_g^2 \sim \sigma_g^2 \chi_{d_g}^2$$

Prior

$$P(\beta_{gj} \neq 0) = p$$

$$\beta_{gj} \mid \beta_{gj} \neq 0 \sim N(0, c_{0j}\sigma_g^2)$$

$$\sigma_g^2 \sim s_0^2 \left(\chi_{d_0}^2 / d_0\right)^{-1}$$

Generalization of Lönnstedt and Speed 2002

Normality, independence assumptions are wrong but convenient, resulting methods are useful

11

Posterior Statistics

Posterior variance estimators

$$\tilde{s}_g^2 = \frac{s_g^2 d_g + s_0^2 d_0}{d_g + d_0}$$

Moderated t-statistics

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{c_{gj}}} \qquad \sim t_{d_g + d_0} \text{ under null}$$

Eliminates large t-statistics merely from very small *s*

12

Posterior Odds

Posterior probability of differential expression for any gene is

$$\frac{p(\beta \neq 0 \mid \hat{\beta}, s^2)}{p(\beta = 0 \mid \hat{\beta}, s^2)} = \frac{p}{1-p}\left(\frac{c}{c+c_0}\right)^{1/2}\left\{\frac{\tilde{t}^2 + d + d_0}{\tilde{t}^2\frac{c}{c+c_0} + d + d_0}\right\}^{\frac{1+d+d_0}{2}}$$

Monotonic function of $\tilde{t}^2$ for constant *d*

Generalization of Lönnstedt and Speed 2002

13

Within-Array Replicate spots

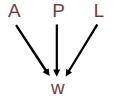
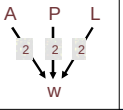
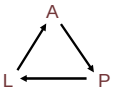
- Replicate spots of each gene on same array, assume duplicates at regular spacing
- Assume spatial component of correlation between duplicates is same for each gene
- Estimate spatial correlation from **consensus** estimator across genes
- Greatly improves estimation of precision

14

Implications for Design

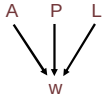
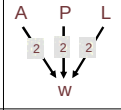
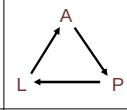
- Given linear modelling approach, can compute efficiency of various experimental designs
- Need to specify which RNA sources are to compared and which contrasts are of interest

15

Comparing 3 RNA Sources	I (a) Common reference	I (b) Common reference	II Direct comparison
			
Number of Slides	N = 3	N=6	N=3
Ave. variance	2		0.67
Units of material			
Ave. variance			

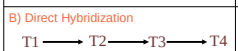
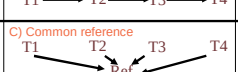
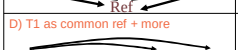
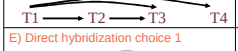
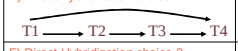
For k = 3, efficiency ratio (Design I(a) / Design II) = 3
In general, efficiency ratio = 2k / (k-1)

16

Comparing 3 RNA Sources	I (a) Common reference	I (b) Common reference	II Direct comparison
			
Number of Slides	N = 3	N=6	N=3
Ave. variance	2		0.67
Units of material	A = B = C = 1	A = B = C = 2	A = B = C = 2
Ave. variance		1	0.67

For k = 3, efficiency ratio (Design I(b) / Design II) = 1.5
In general, efficiency ratio = k / (k-1)

17

Design Choices in Time Series		t vs t+1			t vs t+2			
		T1T2	T2T3	T3T4	T1T3	T2T4	T1T4	Ave
N=3	A) T1 as common reference 	1	2	2	1	2	1	1.5
	B) Direct Hybridization 	1	1	1	2	2	3	1.67
N=4	C) Common reference 	2	2	2	2	2	2	2
	D) T1 as common ref + more 	.67	.67	1.67	.67	1.67	1	1.06
	E) Direct hybridization choice 1 	.75	.75	.75	1	1	.75	.83
	F) Direct Hybridization choice 2 	1	.75	1	.75	.75	.75	.83

18

Design Choices for 2 x 2 Factorial

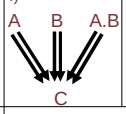
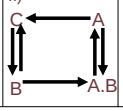
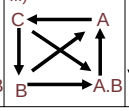
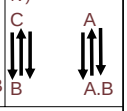
	Indirect	A balance of direct and indirect		
	I) 	II) 	III) 	IV) 
# Slides	N = 6			
Main effect A	0.5	0.67	0.5	NA
Main effect B	0.5	0.43	0.5	0.3
Interaction A.B	1.5	0.67	1	0.67

Table entry: variance

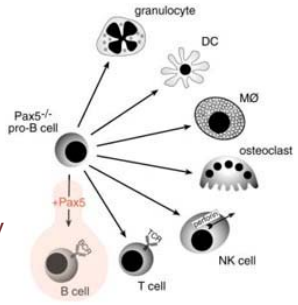
19

Case Study: B Cell Lineage Commitment

20

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



21

How Does Pax5 Work?

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

22

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)

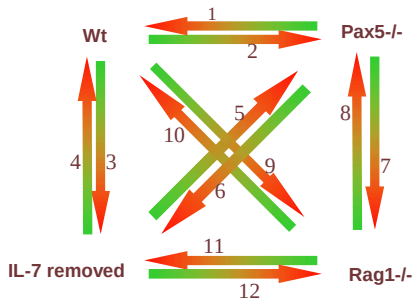
23

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

24

Saturated Design



25

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

26

$$E\begin{pmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \\ m_5 \\ m_6 \\ m_7 \\ m_8 \\ m_9 \\ m_{10} \\ m_{11} \\ m_{12} \end{pmatrix} = \begin{pmatrix} -1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 1 \\ 1 & 0 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 1 & -1 \end{pmatrix} \bullet \begin{pmatrix} PW \\ RW \\ IW \end{pmatrix}$$

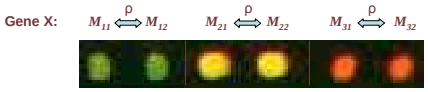
$$\hat{\beta} = (X'\,X)^{-1} X'\,M$$

Full design matrix with duplicate spots is double this

27

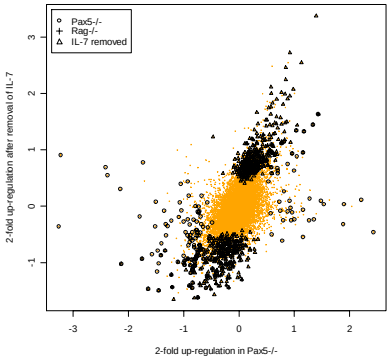
What about Duplicate Spots?

- ρ between duplicate M values on the same slide



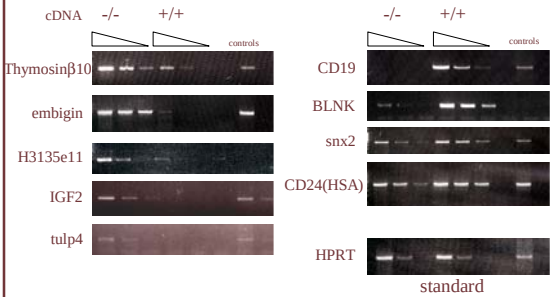
- ρ ≈ 0.85
- Use gls procedure in R to fit linear model allowing for correlated spots

28



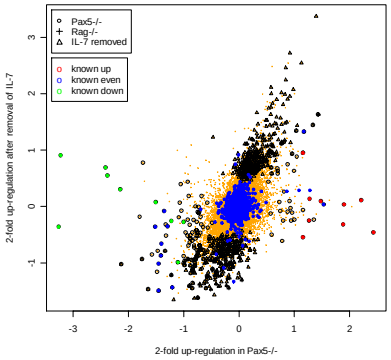
29

RT-PCR Confirmation of DE Genes



10/15 array positives confirmed by RT-PCR

30



31

Agreement with PCR

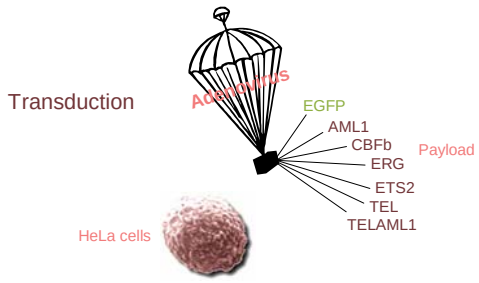
- Observe average rank of known DE genes relative to known non-DE genes
- Moderated t-statistic and ordinary t-statistic do virtually the same on this data
- Both do better than fold change

32

Case Study: Transcription Factor Targets

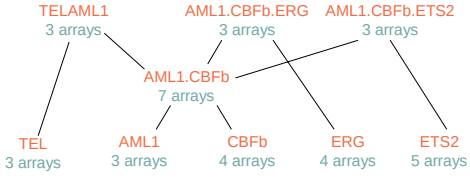
33

Experiment with Transcription Factors



34

Experimental Design I



All arrays use reference **EGFP**

Two GAL files

35

Experimental Design II - Controls

Serum Effects

AML1.CBFb
vs EGFP
w/o serum
3 arrays

AML1
vs EGFP
w/o serum
3 arrays

Virus Effects

No Virus
vs EGFP
4 arrays

AML1
300 viruses
vs 100 viruses
2 arrays

EGFP
300 viruses
vs 100 viruses
2 arrays

36

Comparisons of Interest

- Ordinary comparisons with EGFP:
  N, A, C, R, T, AC, RAC, TAC, TEL, TELAML1
- Comparisons with no virus condition:
  A-N, C-N, R-N, T-N, AC-N, RAC-N, TAC-N, TEL-N, TELAML1-N
- Interaction comparisons:
  AC-A, AC-C, RAC-R, RAC-AC, TAC-T, TAC-AC, TELAML1-TEL, TELAML1-AC
- Control comparisons:
  ACwos, Awos, Awos-A, ACwos-AC, G300-G100, A300-A100

37

Linear Models

- Design matrix is straightforward here because of use of common reference
- Lots of contrasts of interest
- Raises question of simultaneous inference across the contrasts, as well as across genes

38

Moderated F-tests

Can combine several t-tests together in an F-test to test several hypotheses simultaneously

If

$$\beta_g = 0$$

then

$$\frac{\hat{\beta}_g^T X^T W X \hat{\beta}_g}{\tilde{s}_g^2} \sim F_{k,d+d_0}$$

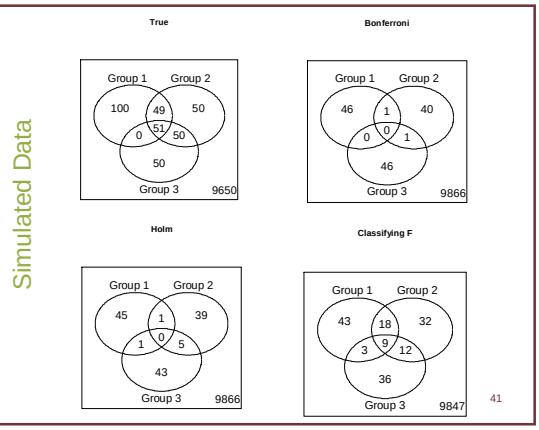
Non-null prior on β doesn't enter

39

Classifying Genes

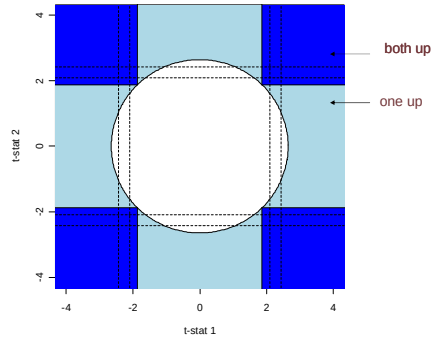
- Any method of classifying genes as up, down or neutral for each transcription factor individually will underestimate the number of genes co-regulated by two or more transcription factors
- **Classifying F-test** method classifies each gene over any number of comparisons arising from a linear model
- More realistic idea of co-regulation

40



41

F-Tests as Classification Problem



42

Composite Classification Method

- F-test classification is not very powerful for detecting genes which respond to one condition (TF) only when there are many comparisons
- Final classification method was a composite of classifying F and individual t

43

Filtering of EGFP Responders

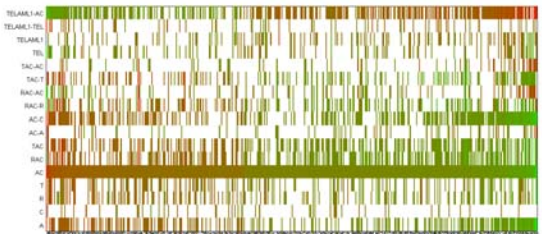
Include TF vs EGFP differences only if they are not reproduced by the no virus vs EGFP comparison

No virus response (N)	AML1 response (A)	AML1 vs no virus contrast (A-N)	Keep?
0	1		Yes
1	1	1	Yes
1	1	0	No

44

Stemmed Heat Diagram

Large scale experiments allow differential expression results to be compared across conditions:



45

Summary

- Analyse data all at once
- Use standard deviances not just fold changes
- Use ensemble information to shrink variances
- Assess differential expression for all comparisons together

46

Analysis Strategies

- Stable background estimation
- Intensity/spatial normalization
- Robustness
- Automatic spot quality weights
- Estimate variability
- Smoothing across genes
- Linear modelling
- Duplicate spots
- Differential expression as classification

Allows high-through put analysis

47

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

48

Acknowledgements

WEHI Bioinformatics
- Terry Speed
- Matt Ritchie
- Natalie Thorne
- James Wettenhall

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

AGRF
- Steve Wilcox
- Cathy Jensen
- Melanie O'Keefe

WEHI Immunology
- Steve Nutt

UC San Francisco
- Jean Yang

49