Analysis of Replicated Experiments

Statistical Methods in Microarray Analysis Tutorial
Institute for Mathematical Sciences
National University of Singapore
January 2, 2004

Gordon Smyth
Walter and Eliza Hall Institute

1

Some Web Sites

- Technical reports etc
  - http://www.statsci.org/micrarra
  - http://bioinf.wehi.edu.au/marray
  - http://www.stat.berkeley.edu/users/terry/zarray/Html
- Statistical software: R
  - http://www.R-project.org
- Packages within R environment
  - Bioconductor http://www.bioconductor.org
  - LIMMA http://bioinf.wehi.edu.au/limma
  - SPOT (cDNA image analysis) http://experimental.act.cmis.csiro.au/Spot/index.php

2

Five Statistical Issues

- Designing gene expression experiments
- Acquiring the raw data: image analysis
- Summarizing and removing artefacts from the data
- Discovering which genes are differentially expressed
- Discovering which genes exhibit interesting expression patterns

For a review see Smyth, Yang and Speed, "Statistical issues in microarray data analysis", In: *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, Humana Press, March 2003

Lots of other bioinformatics issues …

3

Image Analysis

Yang, Buckley, Dudoit and Speed, "Comparison of methods for image analysis on cDNA microarray data", *Journal of Computational and Graphical Statistics*, January 2002.

4

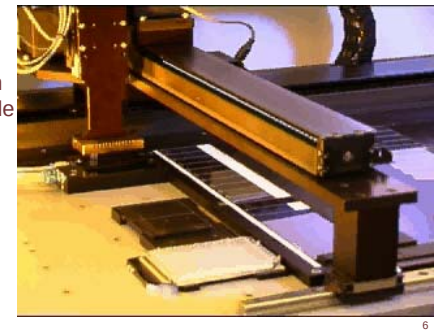
Image Analysis Software

- If you're using Affymetrix arrays, the image analysis will be done as part of the Affy system.
- If you're using spotted arrays, you'll scan your arrays to produce TIFF images. The images will be processed by a program such as SPOT, GenePix, Imagene or Quantarray to acquire intensity measurements

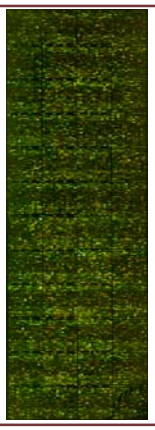
5

Array Printing

Printing custom library on glass slide



6



Microarray Image

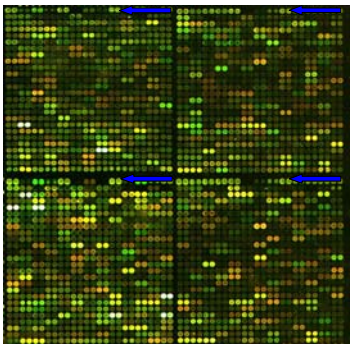
False-coloured image with two channels overlaid:

- more highly expressed in mutant
- equally expressed
- more highly expressed in normal

AGRF NIA 15k mouse array:

12 x 4 pin groups

7



26 x 26 spots in each pin group

Genes printed in duplicate pairs

Arrows indicate print direction

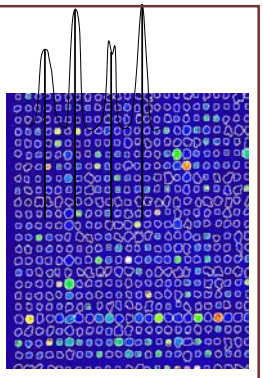
8

Steps in Image Processing

1. **Addressing**: locate centers

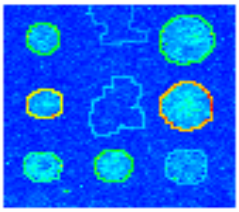
2. **Segmentation**: divide pixels into foreground (part of spot) or background

3. **Information extraction**: calculate signal intensity pairs, background and quality measures for each spot

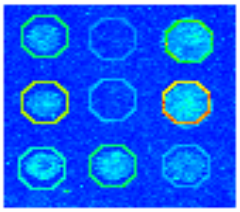


9

Segmentation



Seeded region growing
SPOT



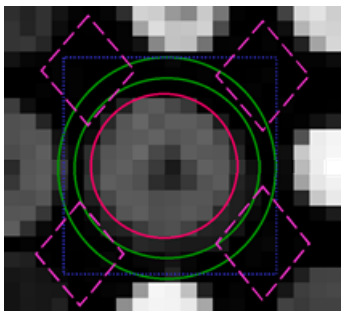
Fixed circle method
Genepix: adaptive circles

Foreground measured from pixels inside the circle
background measured from ambient intensity around spot

10

Local Backgrounds Measures

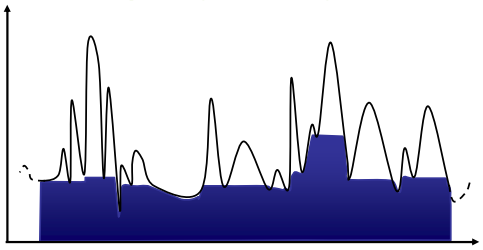
One channel grey scale



Genepix: region including pink diamonds

11

Morphological Background



SPOT estimates background using a nonlinear morphological filter – a lower, less variable measure

12

Quantification of Expression: cDNA

- Foreground red, green $Rf, \quad Gf$
- Background red, green $Rb, \quad Gb$
- Background corrected $R = Rf - Rb, \; G = Gf - Gb$
- log-ratio ("Minus") $M = \log_2 R - \log_2 G$
- Average intensity ("Add") $A = (\log_2 R + \log_2 G)/2$

Lots of issues: which bg is best, to subtract bg or not, quality, filtering

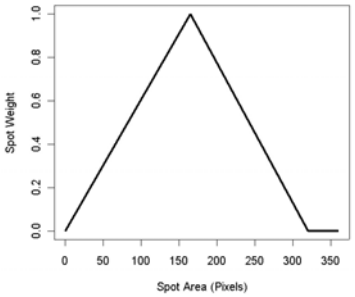
13

Spot Quality Weights

Example of a quality weighting:

downweight spots smaller or larger than nominal size

Many other possibilities



14

Quantification of Expression: Affymetrix

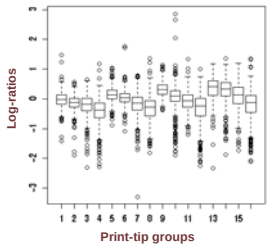
- 12-40,000 probe sets/chip, each consisting of 11-20 Perfect Match (PM) and MisMatch (MM) probe pairs. Consider one set: chips i=1,2,.. probes j=1,2,..
- **Affymetrix summary**:
  log(Signal Intensity) = TukeyBiweight{log(PM_j - MM_j*)}
- **dChip model** (Li, Schadt & Wong):
  $PM_{ij} - MM_{ij} = \theta_i \varphi_j + \varepsilon_{ij}$
- **Robust Multichip Analysis (RMA)**:
  Use median polish or robust regression to fit
  $\log_2(PM\text{-}*BG)_{ij}$ = $chip_i$ effect + $probe_j$ effect + ε_{ij}
  See Bioconductor packages: affy, AffyPLM

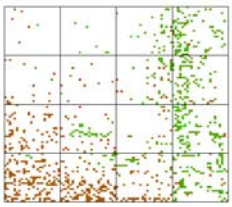
15

Exploratory Data Display

16

Spatial bias in cDNA arrays





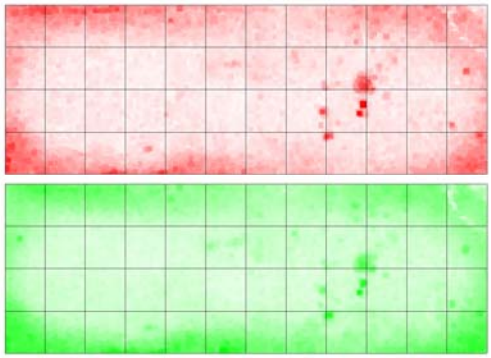
Boxplots of M = $\log_2$R/G by print-tip group (1-16)

Same slide

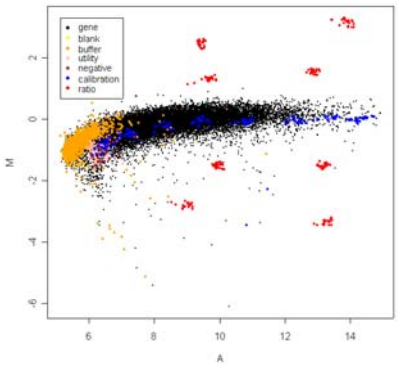
Locations of spots with extreme 5% M: high, low

17

Spatial Plots: Background Intensities

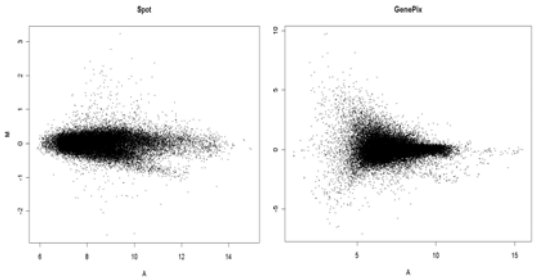


MA Plot



19

Background Affects Variability

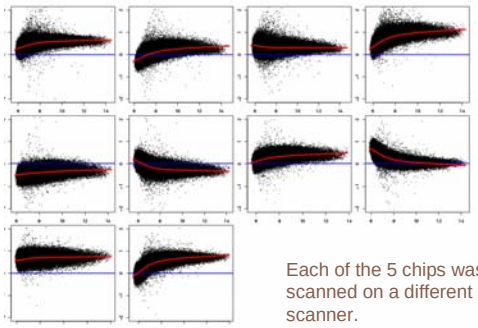


From Spot: note lack of fanning (not same data)

From GenePix: note fanning out at low intensities

20

MA plots of 10 Affymetrix replicate chip pairs



Each of the 5 chips was scanned on a different scanner.

Normalization

Smyth and Speed "Normalization of cDNA microarray data", in: *METHODS: Selecting Candidate Genes from DNA Array Screens*, December 2003

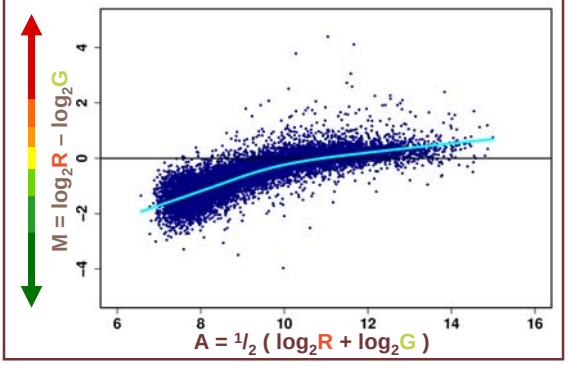
22

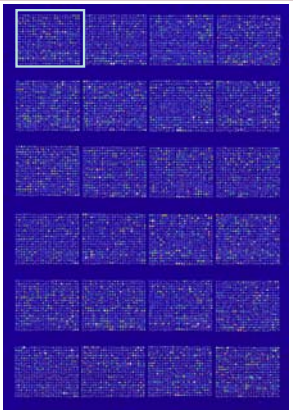
No Absolute Measurement Scale

- The spotted DNA sequences differ in binding efficiency, so absolute intensities are not directly comparable between genes
- The response of the dyes is subject to arbitrary rescaling for each dye on each array – need to normalize the red/green balance for each array (to make M-values unbiased)

23

MA – Plot

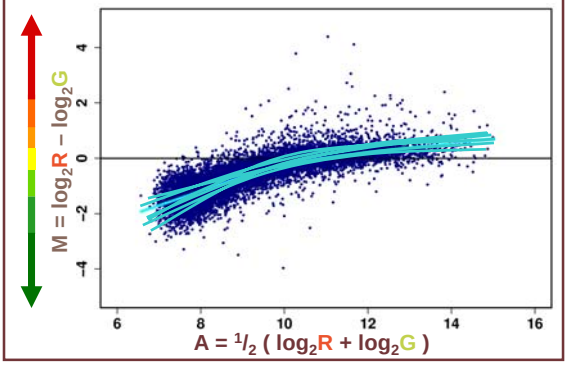




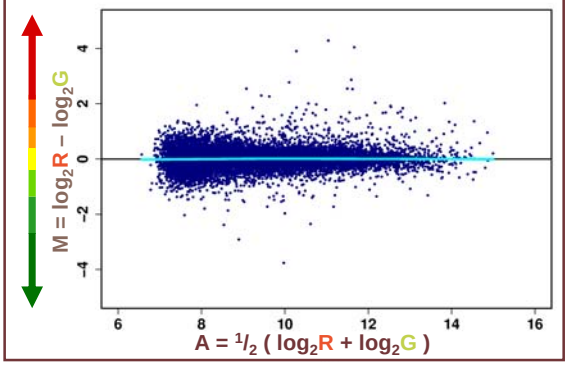
6 x 4
pin-groups

25

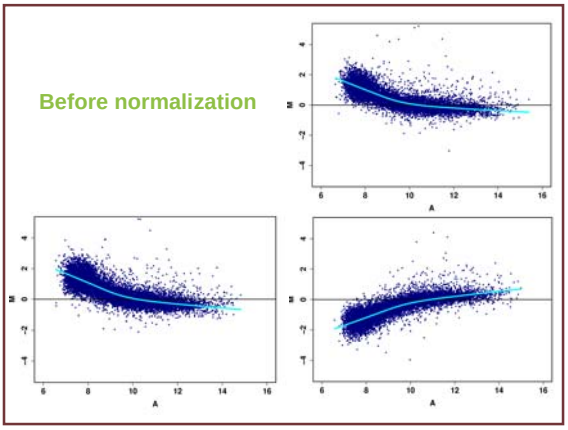
Print-tip-loess Normalization



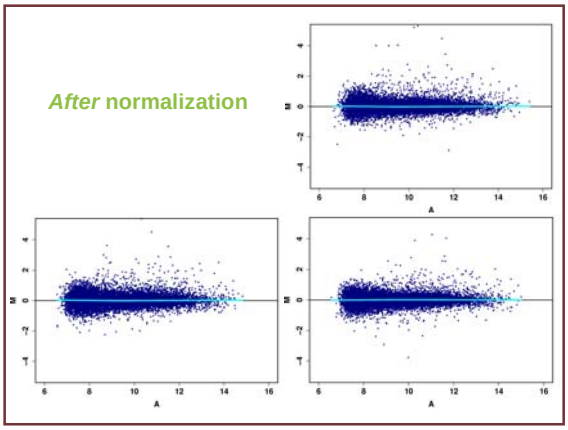
After Normalization



Before normalization



After normalization



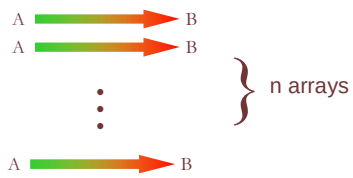
Which Genes are Differentially Expressed?

Smyth, "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments", July 2003

30

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:



31

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:

$$M_1, \ldots, M_n$$

$$\bar{M} = \text{ave}\, M \qquad s = \text{std.dev}\, M$$

32

Ranking Criteria

- Average log **fold change**. $\bar{M}_g$
  Problem: non DE genes with **large variances** have too much chance of being selected.
- **t-statistics** $t_g = \dfrac{\bar{M}_g}{s_g / \sqrt{n}}$
  Problem: genes with very **small sample variances** are suspect
- **Moderated t-statistics**. We use a happy compromise between the two $\tilde{t}_g = \dfrac{\bar{M}_g}{\tilde{s}_g / \sqrt{n}}$

33

Moderated t-Statistics

Shrunk standard deviations

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

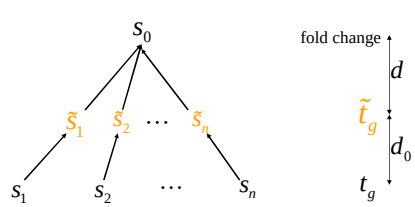
Moderated t-statistics

$$\tilde{t}_g = \frac{\bar{M}_g}{\tilde{s}_g \sqrt{c_g}}$$

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

34

Shrinkage of Standard Deviations



The data decides whether $\tilde{t}_g$ should be closer to $t_{g,\text{pooled}}$ or to t_g

35

Posterior Odds

Posterior probability of differential expression for any gene is

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

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

Reparametrization of Lönnstedt and Speed 2002

36

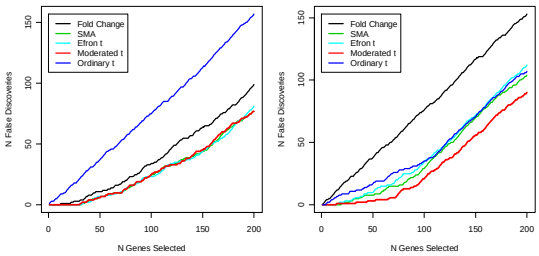
Estimating Hyper-Parameters

Closed form estimators with good properties are available:

- for s_0 and d_0 in terms of the first two moments of log s^2
- for c_0 in terms of quantiles of the $|\tilde{t}_g|$

37

Simulations



σ^2 similar

σ^2 very different

38

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

39

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

40

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*

41

Spike-Ins

- Spike-in artificial RNA cocktail to induce known fold changes in control genes
- Spike-ins can give an objective basis for choosing cut-off for differential expression

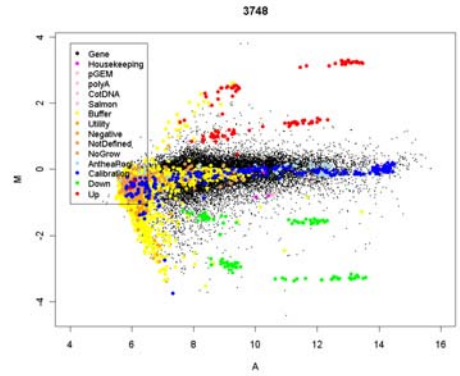
42

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

43

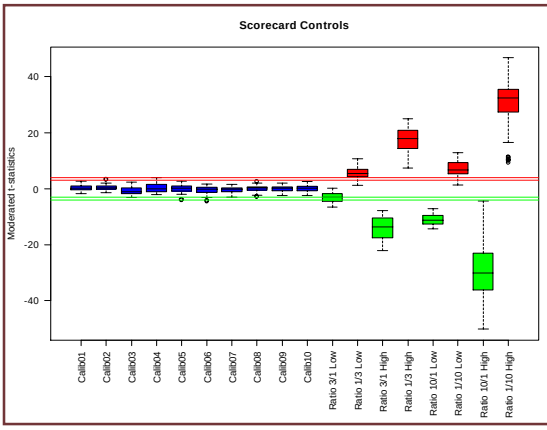
MA-Plot with spike-in controls highlighted



44

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

45



Empirical Cut Off

$$|\tilde{t}_g| > 4$$

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

47

Other Data with Constructed Truth

Assessing Precision:

- Replicate arrays
- Same vs same hybridizations
- Whole library pool titration series – dynamic range but not differentially expressed

Assessing Bias:

- Independent measurement of differential expression – PCR, independent arrays
- Spike-in ratio and calibration controls
- Mixture and dilution experiments

48

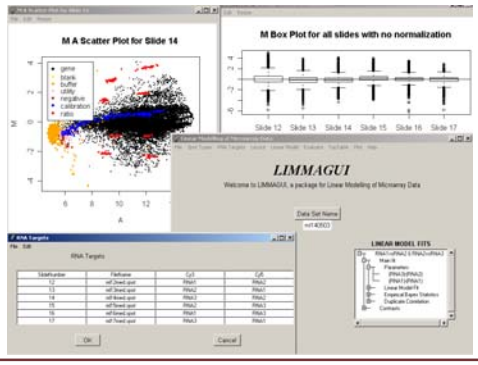
LIMMA Package for R

- 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

49

limma GUI …



50

Acknowledgements

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

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

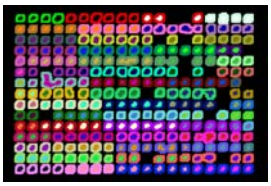
AGRF
- Steve Wilcox
- Cathy Jensen
- Melanie O'Keefe

WEHI Immunology
- Steve Nutt
- Lynn Corcoran
- Mirielle Lahoud
- Melissa Holmes

Berkeley
- Jean Yang
- Speed Lab
- Ngai Lab

51

When the chips are done …



WEHI Bioinformatics

52