Bayesian modelling of gene expression data

Alex Lewin

Sylvia Richardson (IC Epidemiology)
Tim Aitman (IC Microarray Centre)
Philippe Broët (INSERM, Paris)

In collaboration with
Anne-Mette Hein, Natalia Bochkina (IC Epidemiology)
Helen Causton (IC Microarray Centre)
Peter Green and Graeme Ambler (Bristol)

Contents

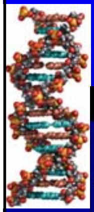
- Introduction to microarrays
- Differential expression
- Bayesian mixture estimation of false discovery rate

Introduction to microarrays

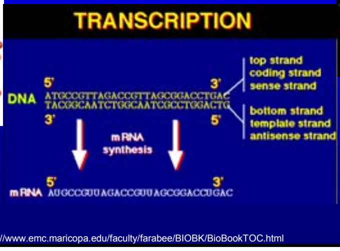
Post-genome Genetics Research

- Challenge: Identify function of all genes in genome
- DNA microarrays allow study of thousands of genes simultaneously

Gene Expression



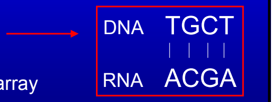
DNA -> mRNA -> protein



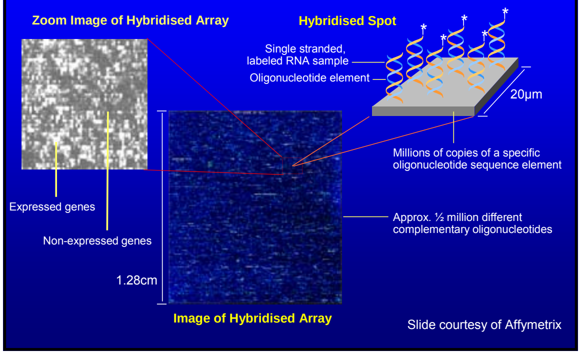
Pictures from http://www.emc.maricopa.edu/faculty/farabee/BIOBK/BioBookTOC.html

Hybridisation

- Known sequences of single-stranded DNA immobilised on microarray
- Tissue sample (with unknown concentration of RNA) fluorescently labelled
- Sample hybridised to array
- Excess sample washed off array
- Array scanned to measure amount of RNA present for each sequence on array



The Principle of Hybridisation


Zoom Image of Hybridised Array
Hybridised Spot
Single stranded, labeled RNA sample
Oligonucleotide element
20µm
Millions of copies of a specific oligonucleotide sequence element
Expressed genes
Non-expressed genes
Approx. ½ million different complementary oligonucleotides
1.28cm
Image of Hybridised Array

Slide courtesy of Affymetrix

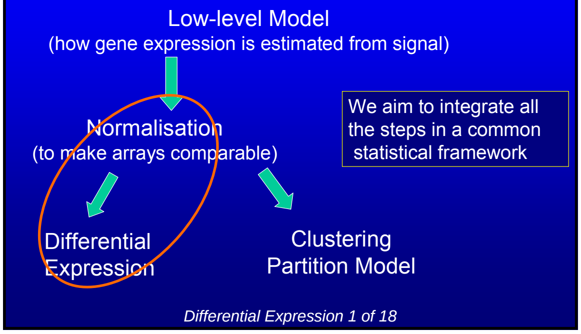
Output of Microarray

- Each gene is represented by several different DNA sequences (probes)
- Obtain intensity for each probe
- Different tissue samples on different arrays so compare gene expression for different experimental conditions

Differential Expression

AL, Sylvia Richardson, Clare Marshall, Anne Glazier, Tim Aitman

Microarray analysis is a multi-step process



Low-level Model
(how gene expression is estimated from signal)

Normalisation
(to make arrays comparable)

We aim to integrate all the steps in a common statistical framework

Differential Expression

Clustering
Partition Model

Bayesian hierarchical model framework

- Model different sources of variability simultaneously, within array, between array …
- Share information in appropriate ways to get better estimates, e.g. estimation of gene specific variability.
- Uncertainty propagated from data to parameter estimates.
- Incorporate prior information into the model.

Data Set and Biological question

<u>Previous Work (Tim Aitman, Anne Marie Glazier)</u>

The spontaneously hypertensive rat (SHR): A model of human insulin resistance syndromes.

Deficiency in gene Cd36 found to be associated with insulin resistance in SHR (spontaneously hypertensive rat)

Data Set and Biological question

Microarray Data

3 SHR compared with 3 transgenic rats

3 wildtype (normal) mice compared with 3 mice with Cd36 knocked out

≅ 12000 genes on each array

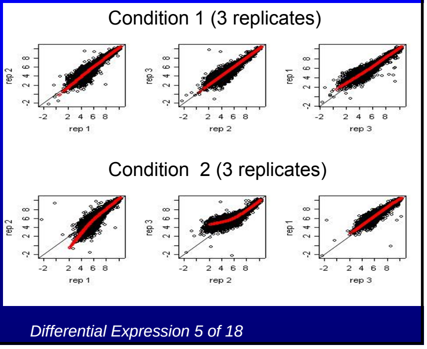
Biological Question

Find genes which are expressed differently in wildtype and knockout mice.

Data

Needs 'normalisation'

Spline curves shown



Model for Differential Expression

- Expression-level-dependent normalisation
- Only 3 replicates per gene, so share information between genes to estimate gene variances
- To select interesting genes, use posterior distribution of ranks

Bayesian hierarchical model for genes under one condition

Data: y_{gr} = log gene expression for gene g, replicate r

α_g = gene effect

$\beta_{r(g)}$ = array effect (expression-level dependent)

σ_g^2 = gene variance

- 1st level

$$y_{gr} \sim N(\alpha_g + \beta_{r(g)}, \sigma_g^2), \quad \Sigma_r \beta_{r(g)} = 0$$

$$\beta_{r(g)} = \text{function of } \alpha_g, \text{ parameters } \{\mathbf{a}\} \text{ and } \{\mathbf{b}\}$$

Bayesian hierarchical model for genes under one condition

- 2nd level

Priors for α_g, coefficients $\{\mathbf{a}\}$ and $\{\mathbf{b}\}$

$$\sigma_g^2 \sim \text{lognormal}(\mu, \tau)$$

Hyper-parameters μ and τ can be influential.
In a full Bayesian analysis, these are not fixed

- 3rd level

$$\mu \sim N(c, d) \qquad \tau \sim \text{lognormal}(e, f)$$

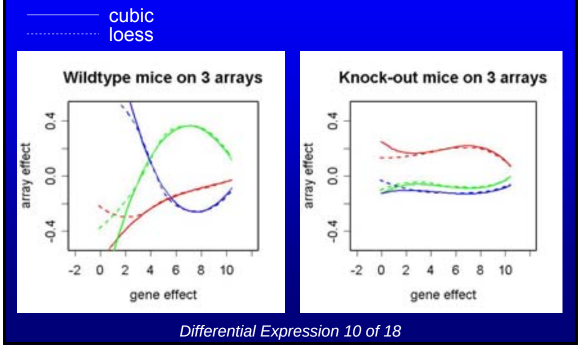
Details of array effects

- Piecewise polynomial with unknown break points:

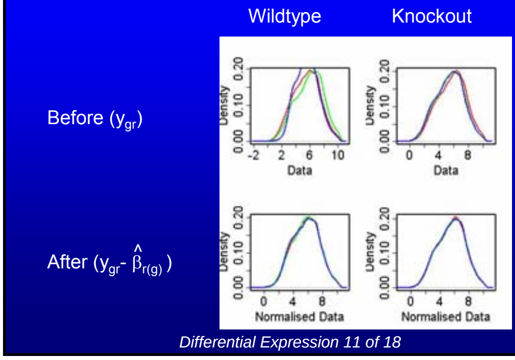
$$\beta_{r(g)} = \text{quadratic in } \alpha_g \quad \text{for } a_{rk-1} \le \alpha_g \le a_{rk}$$
$$\text{with coeff } (b_{rk}^{(1)}, b_{rk}^{(2)}), \quad k = 1, \ldots \#\text{breakpoints}$$

- Locations of break points not fixed
- Must do sensitivity checks on # break points
- Cubic fits well for this data

Non linear fit of array effect as a function of gene effect



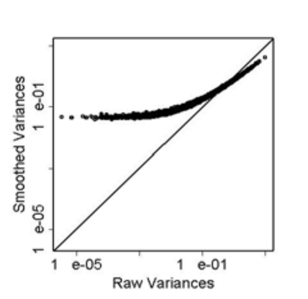
Effect of normalisation on density



Smoothing of the gene specific variances

- Variances are estimated using information from all G x R measurements (~12000 x 3) rather than just 3

- Variances are stabilised and shrunk towards average variance



Bayesian Model Checking

- Check our assumption of different variance for each gene
- Predict sample variance $S_g^{2\ new}$ from the model for each gene
- Compare predicted $S_g^{2\ new}$ with observed $S_g^{2\ obs}$

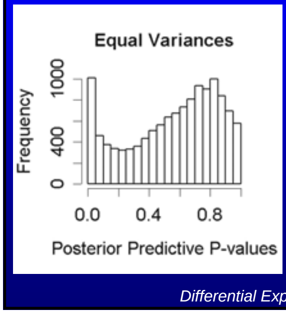
Bayesian p-value Prob($S_g^{2\ new} > S_g^{2\ obs}$)

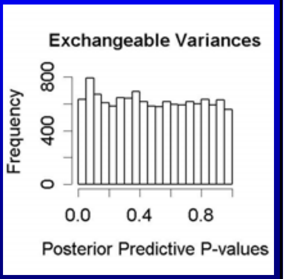
- Distribution of p-values Uniform if model is adequate
- Easily implemented in MCMC algorithm

Bayesian predictive p-values

Control for method: equal variance model has too little variability for the data

Exchangeable variance model is supported by the data





Differential expression model

d_g = differential effect for gene g between 2 conditions

Joint model for the 2 conditions :

$y_{g1r} \sim N(\alpha_g - ½ d_g + \beta_{r(g)1} , \sigma_{g1}^2)$, (condition 1)

$y_{g2r} \sim N(\alpha_g + ½ d_g + \beta_{r(g)2} , \sigma_{g2}^2)$, (condition 2)

Prior can be put on d_g directly

Possible Statistics for Differential Expression

$d_g \approx$ log fold change

$d_g^* = d_g / (\sigma^2_{g1} / 3 + \sigma^2_{g2} / 3)^{½}$ (standardised difference)

•We obtain the joint distribution of all $\{d_g\}$ and/or $\{d_g^*\}$

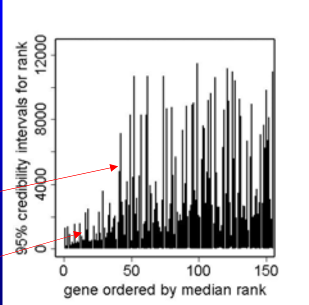
•Distributions of ranks

Credibility intervals for ranks

150 genes with lowest rank (most under-expressed)

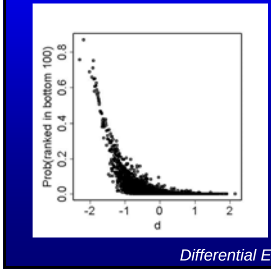
Low rank, high uncertainty

Low rank, low uncertainty



Probability statements about ranks

Under-expression:
probability that gene is ranked in bottom 100 genes



Have to choose rank cutoff (here 100)

Have to choose how confident we want to be in saying the rank is less than the cutoff (eg prob=80%)

Summary: Differential Expression

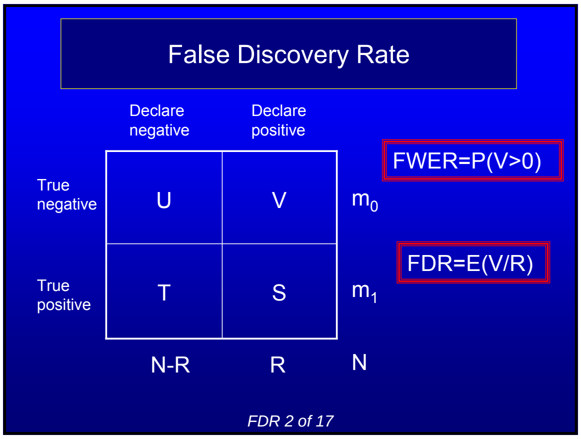
- Expression-level-dependent normalisation
- Only 3 replicates per gene, so share information between genes to estimate gene variances
- To select interesting genes, use posterior distribution of ranks

Bayesian estimation of False Discovery Rate

Philippe Broët, AL, Sylvia Richardson

Multiple Testing

- Testing thousands of hypotheses simultaneously
- Traditional methods (Bonferroni) too conservative
- Challenge: select interesting genes without including too many false positives.

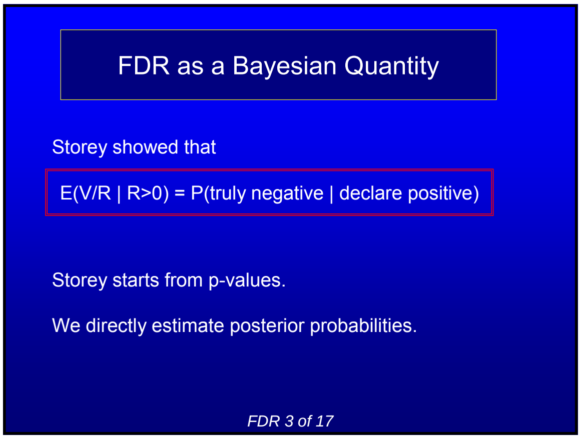


False Discovery Rate

	Declare negative	Declare positive	
True negative	U	V	m_0
True positive	T	S	m_1
	N-R	R	N

FWER=P(V>0)

FDR=E(V/R)



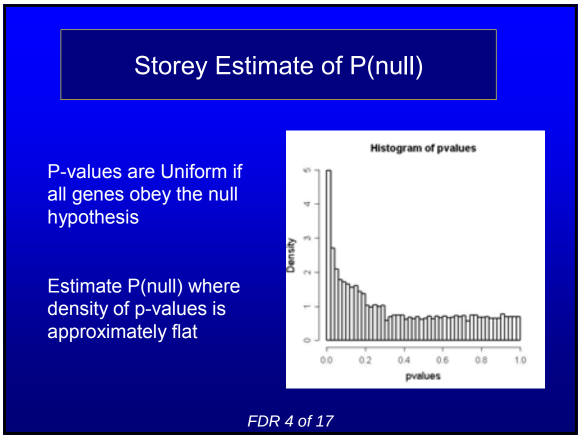
FDR as a Bayesian Quantity

Storey showed that

E(V/R | R>0) = P(truly negative | declare positive)

Storey starts from p-values.

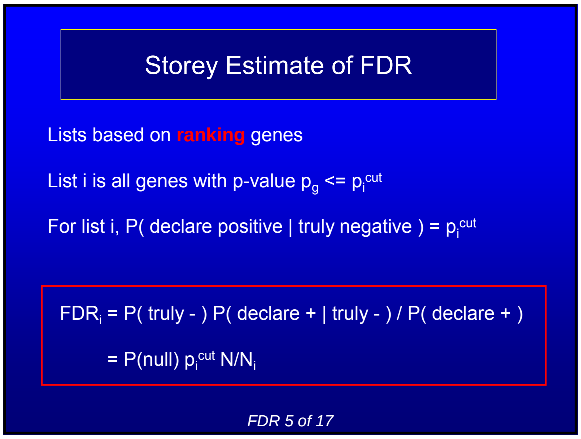
We directly estimate posterior probabilities.



Storey Estimate of P(null)

P-values are Uniform if all genes obey the null hypothesis

Estimate P(null) where density of p-values is approximately flat



Storey Estimate of FDR

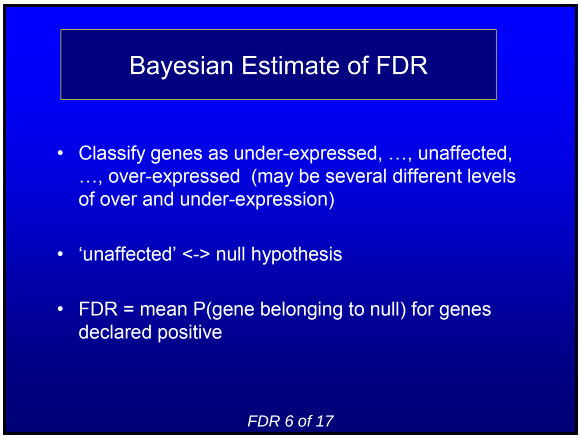
Lists based on ranking genes

List i is all genes with p-value $p_g <= p_i^{cut}$

For list i, P(declare positive | truly negative) = p_i^{cut}

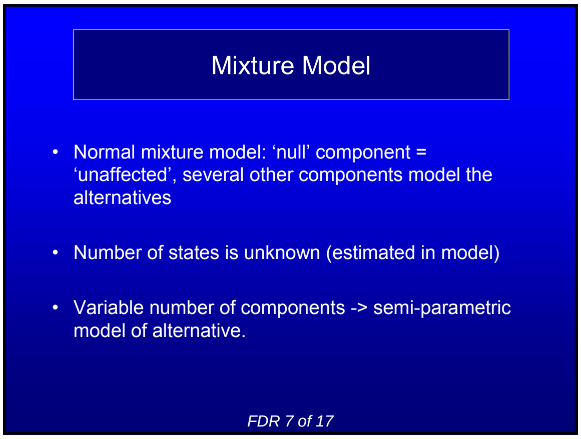
FDR_i = P(truly -) P(declare + | truly -) / P(declare +)

= P(null) p_i^{cut} N/N_i



Bayesian Estimate of FDR

- Classify genes as under-expressed, …, unaffected, …, over-expressed (may be several different levels of over and under-expression)
- ‘unaffected’ <-> null hypothesis
- FDR = mean P(gene belonging to null) for genes declared positive



Mixture Model

- Normal mixture model: ‘null’ component = ‘unaffected’, several other components model the alternatives
- Number of states is unknown (estimated in model)
- Variable number of components -> semi-parametric model of alternative.

Bayesian mixture model

Mixture model specification NULL ALTERNATIVE

$$d_g \sim w_0\, N(0, \sigma_0^2) + \Sigma_{j=1:k}\, w_j\, N(\mu_j, \sigma_j^2)$$

$$\mu_j \begin{cases} \mu_j^+ > 0 \text{ , ordered, uniform on upper range} \\ \mu_j^- < 0 \text{ , ordered, uniform on lower range} \end{cases}$$

k, unknown number of components -> alternative is modelled semi-parametrically

Bayes Estimate of FDR

Latent variable z_g = 0, 1, …., k with prob $w_0, w_1, \ldots, w_k$

P(gene g in null | data) calculated from the z_g

For **any** given list L containing N_L genes,

$FDR_L = 1/N_L \Sigma_{g \text{ on list } L}$ P(gene g belonging to null | data)

Compare Estimates of FDR

Storey $FDR_i = 1/N_i\ \ p_i^{cut}\ \ P(\text{null})\ N$

Bayes $FDR_i = 1/N_i\ \Sigma_{\text{list } i}$ P(gene in null | data)

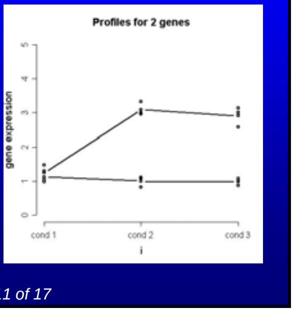
NB Bayes estimate can be calculated for any list of genes, not just those based on ranking genes

Gene Expression Profiles

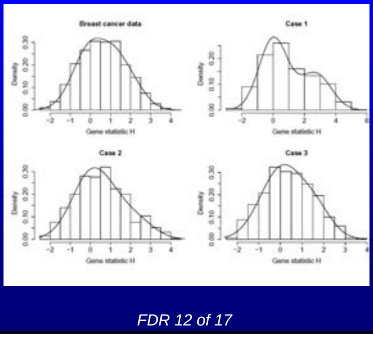
Each gene has repeat measurements under several conditions: gene profile

Summarize profile by F-statistic (one for each gene)

Transform -> approx. Normal if no change across conditions



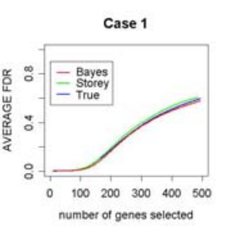
Simulated Data

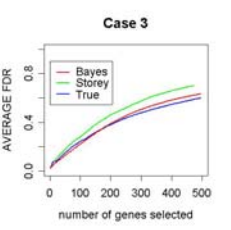


Results for Simulated Data

Usual methods (Storey q-value and SAM) overestimate FDR

Bayes mixture estimate of FDR is closer to true value



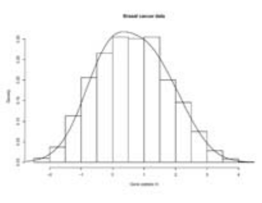


Breast Cancer Data

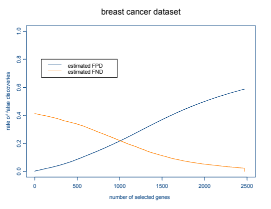
- Study of gene expression changes among 3 types of tumour: BRCA1, BRCA2 and sporadic tumours.
- Gene profiles across tumours summarized by F-statistics, transformed.
- Estimate FDR **and FNR** (false non-discovery rate)

Results for Breast Cancer Data (Ordered Lists of Genes)

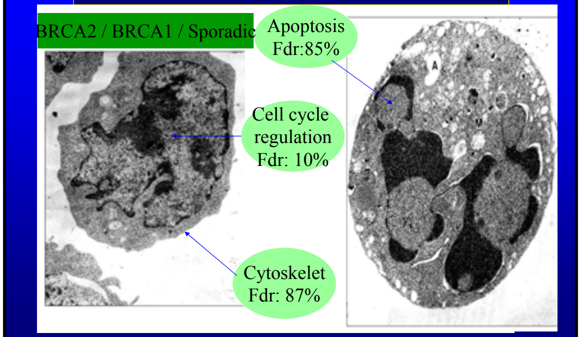
Gene statistics



FDR, FNR



Results for subsets of genes


BRCA2 / BRCA1 / Sporadic

Apoptosis Fdr:85%

Cell cycle regulation Fdr: 10%

Cytoskelet Fdr: 87%

Slide from Philippe Broët

Summary: FDR

- Good estimate of FDR and FNR
- Semi-parametric model for differentially expressed genes.
- Obtain posterior probability for each gene.
- Can calculate FDR, FNR for any list of genes.

Summary

Differential Expression

Expression-level-dependent normalisation

Borrow information across genes for variances

Joint distribution of ranks

False Discovery Rate

Flexible mixture gives good estimate of FDR

Future work

Mixture prior on log fold changes, with uncertainty propagated to mixture parameters

Two papers submitted:

Lewin, A., Richardson, S., Marshall C., Glazier A. and Aitman T. (2003) *Bayesian Modelling of Differential Gene Expression.*

Broët, P., Richardson, S., Lewin, A., Dalmasso, C. and Magdelenat, H. (2004) *A model-based approach for detecting distinctive gene expression profiles in multiclass response microarray experiments.*

Available at
http ://www.bgx.org.uk/