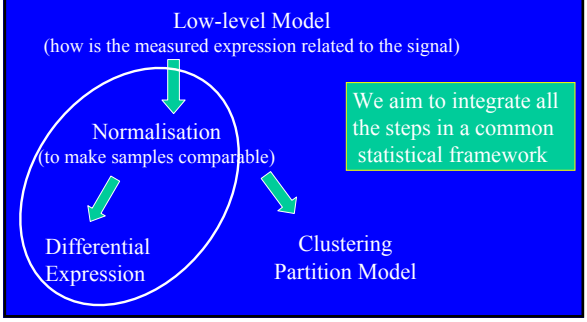
Bayesian modelling of differential gene expression data

Sylvia Richardson, with Alex Lewin
Department of Epidemiology and Public Health, Imperial College

In collaboration with
Anne Mette Hein and Clare Marshall (Imperial)
Philippe Broët (INSERM) and Peter Green (Bristol)
Helen Causton and Tim Aitman (Hammersmith)

Gene expression analysis is a multi-step process



Bayesian hierarchical model framework

- Ability to model various sources of variability:
  e.g. detailed modelling of experimental variability: within array, between array, estimation of gene specific variability …
- Building of all these features into a common model
  uncertainty is propagated
- All unknown quantities are given prior distributions
- Inference is based on their posterior distribution given the data
- Ability to borrow / share information in appropriate ways to get better estimates

Data Set and Biological question

Previous Work (Tim Aitman, Anne Marie Glazier)

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

Good animal model to tease out genes implicated in this syndrome

Microarray Study

- 3 SHR compared with 3 transgenic rats
- 3 wildtype mice compared with 3 knockout mice
- Two tissues: fat and heart
- Affymetrix chips U34A-C and U74A-C
  (≅ 12000 genes)

I -- Building the basic model

Consider replicated arrays (single channel) under one experimental condition

Levels of Variability ?

• Variability of the overall expression level between slides (Possibly non linear)

⇒ Flexible normalisation

• Variability of the expression of each gene between replicates (biological, technical)

⇒ Hierarchical modelling of the gene specific variances

Additive (log scale) model for expression

Notation

- y_{gr} = gene expression measurements ***(log scale)*** for gene g, g=1, …, N, replicate r, r = 1,…, R

Additive model:

$$y_{gr} = \alpha_g + \beta_{r(g)} + \varepsilon_{gr}$$

Here:

-- α_g is the expression level of the g^{th} gene,

-- $\beta_{r(g)}$ is the array effect (normalisation term) possibly dependent on g through the expression level α_g ,
constraints needed to ensure identifiability: $\Sigma_r \beta_{r(g)} = 0$

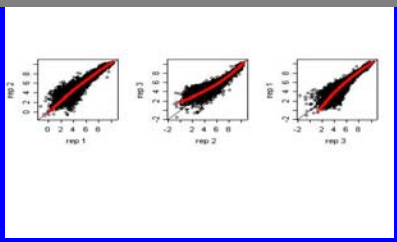
-- ε_{gr} an error term, mean 0, $Var(\varepsilon_{gr}) = \sigma_g^2$

I.A - Exploratory analysis of array effect

What shape for the array effect ?
Constant normalisation term β_r OR non linear dependence on the expression level $\beta_r = \beta_{r(g)}$?

Wildtype mice data

Correlation between the replicates



Exploratory analysis of array effect

What shape for the array effect $\beta_{r(g)}$?

Exploratory analysis:

- Split the data into 6 groups corresponding to ordered values of the mean of (y_{gr}) r = 1,…, R
- Estimate the array effect β_{ri} for each group:

For g in group i: $y_{gr} = \alpha_g + \beta_{ri} + \varepsilon_{gr}$

with $\varepsilon_{gr} \sim N(0, \sigma_i^2)$

- Investigate if β_{ri} changes in the different groups i, i=1, …,6.

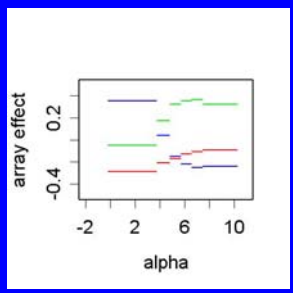
Exploratory analysis of array effect

Wildtype mouse fat data on 3 arrays

Each of the 3 arrays corresponds to a different colour

6 Equal size groups of genes defined by mean expression level

β_{ri}, r =1:3, i=1:6



Flexible model of the array effect

- The 6 groups exploratory analysis suggests to model the array effect as a (smooth) function of α_g
- Choose a piecewise polynomial form with unknown break points, e.g.:
- $\beta_{r(g)}$ = quadratic with coeff ($b_{rk}^{(1)}$, $b_{rk}^{(2)}$), k =0, … #knots (+ cst term) for $a_{rk-1} \leq \alpha_g \leq a_{rk}$

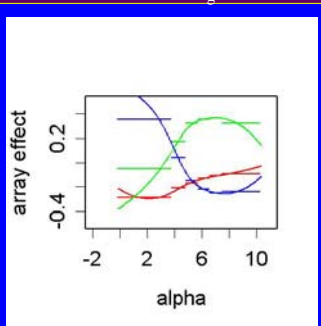
a_{r0} fixed lower limit , (depending on the range)

The breakpoints a_{rk} are random variables, uniform on the range

All coefficients $b_{ri}^{(j)}$ are given centred normal priors,

$\Sigma_r \beta_{r(g)} = 0$ constraint

Non linear fit of array effect as a function of level α_g



I.B - Modelling gene variability

Two extreme cases:

(1) Constant variance $\varepsilon_{gr} \sim N(0, \sigma^2)$

Too stringent Poor fit

(2) Independent variances $\varepsilon_{gr} \sim N(0, \sigma_g^2)$

! Variance estimates based on few replications are highly variable

Need to share information between genes to better estimate their variance, while allowing some variability

⇒ Hierarchical model

Hierarchical structure for gene variability in each condition

- **2nd level of the model** : Exchangeable hierarchical prior:

$$\sigma_g^2 \sim \text{lognormal}(\mu, \tau), \quad g = 1, \ldots N$$

The hyper parameters μ and τ can be influential

In a full Bayesian analysis, these are not fixed

- **3rd level of the model**

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

Where the constants c, d, e and f are chosen so that the third level priors are vague

Summary of the hierarchical model

- 1st level

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

$\beta_{r(g)}$ = quadratic with coeff $(b_{rk}^{(1)}, b_{rk}^{(2)})$, k =0, …#knots

for $a_{rk-1} \leq \alpha_g \leq a_{rk}$

- 2nd level

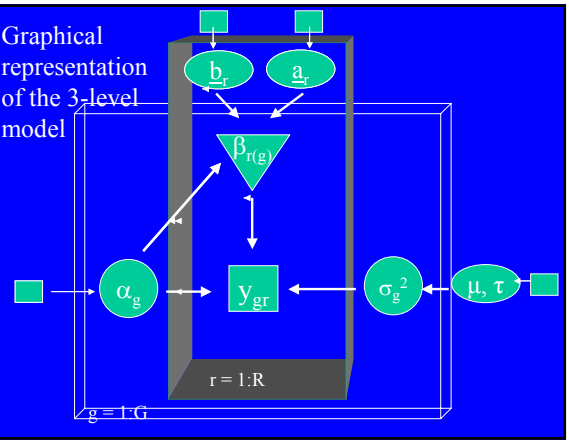
Priors for α_g, coefficients {**b**} and {**a**}

$\sigma_g^2 \sim$ lognormal (μ, τ)

- 3rd level

$\mu \sim N(c, d)$ $\quad \tau \sim$ lognormal (e, f)

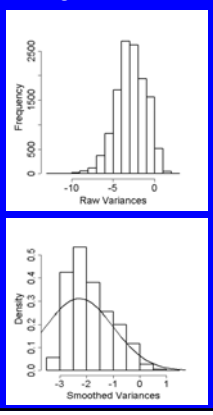
Graphical representation of the 3-level model



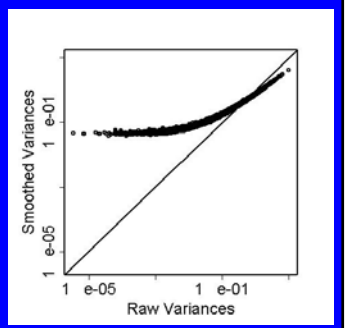
Implementation of the hierarchical model

- Inference on the parameters is based on their posterior distribution | data
- The joint posterior distribution of all parameters is simulated using MCMC algorithms
- The software Winbugs is used for the implementation
- It only requires to specify the distributions involved in the hierarchical model structure

Log scale used



Smoothing of the variances by hierarchical exchangeable prior



Smoothing of the gene specific variances

- Variances are estimated using information from all G x R measurements (typically 8000 X 3) rather than just 3
- Variances are stabilised and shrunk towards average variance
- In particular, some small variance estimates that are incompatible with the overall distribution and are increased

I.C- Bayesian Model Checking

- Different assumptions for the gene variances may give different results

How to choose ?

- Bayesian model-comparison via predictive approaches
  - Based on predicting data points y_{gr}^{new} using the specification of the hierarchical model
  - Use a chosen checking function to compare observed and predicted data points
  - Can be implemented in the MCMC algorithm

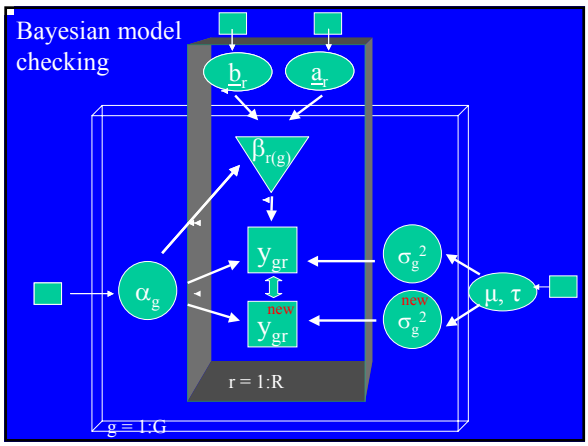
Checking the exchangeable variance model

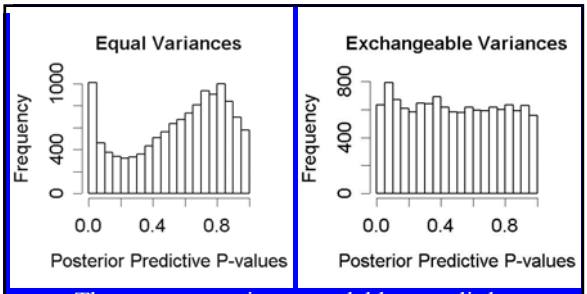
- Generate new variance $\sigma_g^{2\,new}$ from the 2nd level model (lognormal distribution)
- Generate new replicated data point y_{gr}^{new} using α_g , the coefficients **a** and **b**, and $\sigma_g^{2\,new}$
- Checking function:

⇒ Empirical variance calculated over r:

$V(y_{gr})$ and $V(y_{gr}^{new})$

- Calculate Prob $[V(y_{gr}^{new}) > V(y_{gr})]$
- Compare to uniform (the distribution under the null hypothesis that the model is 'true')





- The constant variance model has too little variability for the data
- The gene-specific exchangeable variance model is supported by the data (uniform posterior predictive p values)

II-- Analysing gene expression data

- Gene expression data can be used in several types of analysis:

⇒ Comparison of gene expression under different experimental conditions, or in different tissues

-- Classification of gene expression profiles

-- Exploration of patterns in gene expression matrices

-- Association of gene expression with other factors, e.g. prognosis

Gene expression data matrix

Samples

Genes

Cond 1 Cond 2

y_{g1r} y_{g2r}

Differential expression model

The quantity of interest is the difference between conditions for each gene: d_g , g = 1, …,N

Joint model for the 2 conditions :

$$(1)\ \begin{cases} y_{g1r} = \alpha_g - \tfrac{1}{2} d_g + \beta_{1r(g)} + \varepsilon_{g1r}, & r = 1, \ldots R_1 \\ y_{g2r} = \alpha_g + \tfrac{1}{2} d_g + \beta_{2r(g)} + \varepsilon_{g2r}, & r = 1, \ldots R_2 \end{cases}$$

- α_g is now the overall gene effect over the conditions
- Same assumptions for the distribution of σ^2_{gs} and the modelling of $\beta_{sr(g)}$ as before, s = 1, 2
- All hyper parameters are indexed by the condition

How to select relevant groups of genes?

- Statistics usually considered are:

The differences d_g (≈ log fold change) or

The standardised differences $d_g^* = d_g / (\sigma^2_{g1}/R_1 + \sigma^2_{g2}/R_2)^{1/2}$

- We obtain the joint distribution of all $\{d_g\}$ or $\{d_g^*\}$

In particular, we can

⇒ Process the output to have the distributions of the ranks

$\{r(d_g), g = 1, \ldots, N\}$

⇒ Model the distribution of d_g flexibly to allow a mixture of (small) subgroups of genes with 'extreme' d_g (H_1) and a (large) group of genes with d_g around 0 (H_0)

How to use the joint distribution of the differential expression measures ?

- Compute the median rank for each gene as well as its 2.5% to 97.5% credibility interval
- Plot these by median rank order
- Make probability statements such as

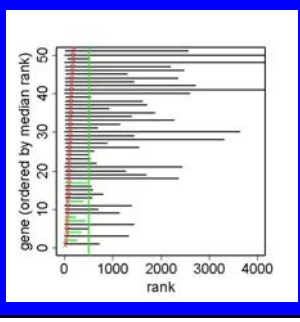
« Probability that genes $\{g_1, \ldots g_n\}$ are among the top 100 is at least 0.95 »

Joint distribution of $\{\text{rank}(d_g)\}$ **(unstandardised measure)**

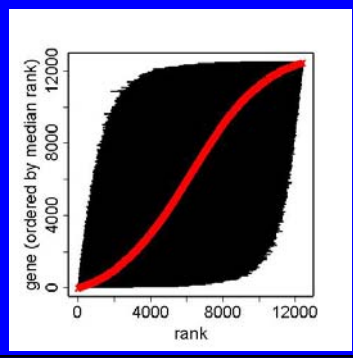
3 wildtype mice compared to 3 knockout mice

For ex, we find 6 genes that are ranked in the most under expressed 500 with 95% confidence

2.5% - 97.5% rank intervals for each gene
Showing the 50 genes with the smallest d_g



Joint distribution of $\{\text{rank}(d_g)\}$ All 12000 genes



Proposed approach for modelling the distribution of d_g

- Mixture model framework but with an 'unknown' number of components
- Fully Bayesian hierarchical framework (the number of components is a random variable)
- Based on Green (1995) and Richardson and Green (1997) papers
- MCMC algorithms
- Applied in an experimental context concerning bladder cancer

(Broët, Richardson and Radvanyi; Journal of Computational Biology, 9, 671-683, 2002)

Bayesian mixture model (1)

Finite normal mixtures with an unknown number of states

A gene can be in different states:

down regulated, …, unaffected, …, up-regulated

↑

the central state corresponding to $d_g \approx 0$

Bayesian estimation of posterior distribution:

- for number of states (besides the unaffected one)
- for the allocation of genes to the states

⇒ classification of states in the 'extreme components' or in the central one, based on their posterior probability

Bayesian mixture model (2)

- Mixture model specification takes into account the particular context of differential gene expression

$$d_g \sim w_0\, N(0, \lambda^2\eta_0^2) + \Sigma_{j=1:k}\, w_j\, N(\mu_j, \eta_j^2)$$

Prior setting

$\lambda^2 > 1$ expresses that the central component is expected to have a larger variance

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

$\{\eta_0^2 \cdot \eta_j^2\}$ exchangeable, Gamma distributed

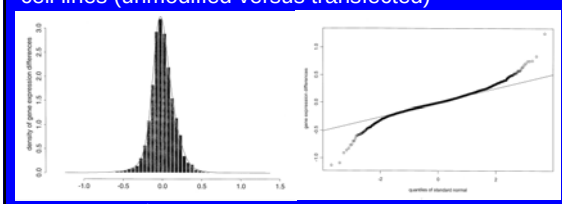
weights $\{w_0, w_j, j=1:k\} \sim$ Dirichlet,

k, unknown number of components

Biological Context

- Data from the Curie Institute/Research Section (F. Radvanyi team, CNRS/Curie)
- Aim: to study transcriptional changes induced by a defined DNA transfection in a cell line
- The cell line: T24 bladder cell lines
- The transfection: FGFR2 cDNA (fibroblast growth factor receptor 2)
- Comparison of 2 cell lines: Unmodified and Modified (transfected by FGFR2 cDNA)
- Material: Nylon microarray, 4608 gene expression from the same batch, ^{33}P labelling, 4 replicates, same experimenter, same day

Comparison of gene expression in two bladder cancer cell lines (unmodified versus transfected)



Distribution of d_g and QQ plot

d_g : Differential expression for gene g after taking into account array and cell line main effects. Negative values of d_g correspond to up-regulated genes.

Results (1)

- ***Posterior distribution of the number of mixture components***

Components left of central one

P(0) = 0.00	P(1) = 0.11	**P(2) = 0.73**
P(3) = 0.14	P(4) = 0.01	P(5) = 0.00

Components right of central one

P(0) = 0.00	P(1) = 0.46	**P(2) = 0.49**
P(3) = 0.04	P(4) = 0.052	P(5) = 0.00

➡ Support for a model with 5 components, 2 on the left and 2 on the right of the central one

Results (2)

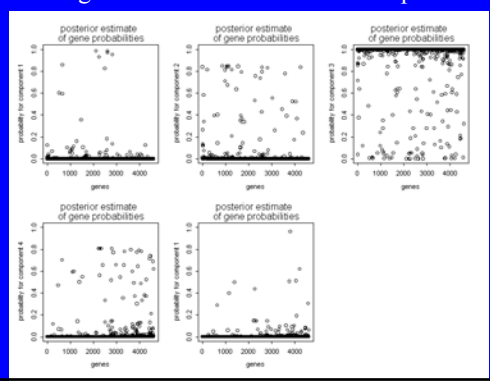
- Estimate the posterior probabilities of each gene *g* of belonging to the *j*th component

- Classification: Maximum *a posteriori* rule

From the posterior probabilities for each gene g, we allocate a gene to the group *k* such as:

$$L_k = \arg\max_j \{ j:\ P(\text{gene}_g \in j \mid d_g)\}$$

Posterior estimate of the probability for each gene to be in each of the 5 components

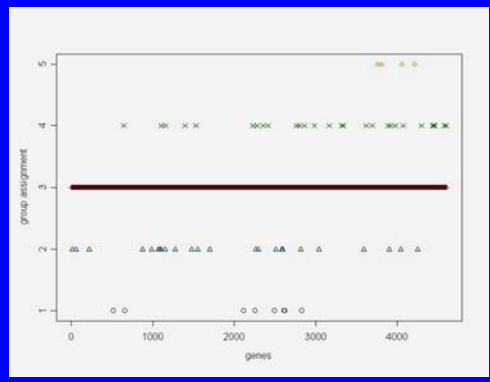


Results (3)

- Classification group:

Group	1	2	3	4	5
Nb	**9**	**26**	**4540**	**29**	**4**
Mean	-0.82	-0.51	0	0.51	0.86
SD	0.16	0.08	0.14	0.10	0.17
Weight	0.2%	0.8%	98%	0.8%	0.2%

➡ So this analysis highlights 9 up-regulated genes (expressed after transfection of the receptor) and 4 down-regulated ones, with an indication of 2 further subgroups showing some evidence of differential expression



Summary

- Model different sources of variability into a single model
- Borrow information from all genes to stabilise estimates of gene specific variances under replication
- Exploit the joint distribution of the differential expression measure through ranks or mixture models: useful for overcoming multiple-testing problems
- Further work is under way
  - on modelling of the low level Affy probe data
  - on more general clustering algorithms

We are avertising for a post-doc position in my department (Imperial) to work on analysing gene expression data Please contact me if you could be interested

Nous cherchons à recruter un postdoc à Imperial sur l'analyse statistique des biopuces Contactez moi si cela vous intéresse

sylvia.richardson@imperial.ac.uk