Bayesian analysis of gene expression data

Sylvia Richardson

with Marta Blangiardo[1], Natalia Bochkina, Anne Mette Hein[2], Alex Lewin[1], Ernest Turro[1] and Peter Green[3]

[1] Centre for Biostatistics, Imperial College,
[2] University of Aarhus, [3]University of Bristol

Introduction

Microarray experiments and gene expression data have a number of characteristics that make them **attractive but challenging** for Bayesian analysis

- Many sources of variability and low signal/noise ratio
- Variability at different levels (array specific, gene specific, ...)

  $\Rightarrow$ accounting for uncertainty is important
- Need to borrow information, e.g. across genes, as typical experiments involve few samples

  $\Rightarrow$ Hierarchical modelling
- Masses of data

  $\Rightarrow$ Need for a variety of data synthesis methods adapted to the level of information processed (probe level signal, gene, posterior probabilities, ...)

Introduction

Bayesian models have been developed to address some of these issues

Models have been formulated at different levels:

- Probe level models aiming to quantify the signal
- Differential expression models aimed at giving 'useful gene lists', accompanied by a measure of false detection rate
- Clustering models of expression profiles aiming to uncover subgroups of genes that co-vary across a range of conditions/treatments

  $\Rightarrow$ **c.f. Peter Green's talk on Monday**
- Methods for synthesing genes lists between different experiments

Cutting across all these developments:

- MCMC issues
- Model checking and criticism

Outline

1. BGX Bayesian Gene eXpression

   MCMC issues

2. Mixture models for differential expression

   Model checking via mixed predictive p-values

3. Synthesising gene lists

Modelling the signal from oligonucleotide arrays (Affymetrix GeneChips)

Data:

Gene g characterised by probe set of J short oligos on array c,

– with Perfect match probes, PM_{jgc},

– and Mis-matched probes, MM_{jgc}, designed to capture non-specific cross hybridisation (but still containing some signal)

Aim :

1. Extract the 'true signal' S_{jgc} and non specific hybridisation H_{jgc} from the pair $\{PM_{jgc}, MM_{jgc}\}$
2. Summarise the expression of each gene by a posterior distribution that combines information from the signal of all the probes $\{S_{jgc}, j = 1, \ldots, J\}$
3. Account for array specific background

1 – BGX Bayesian Gene Expression

BGX model *Hein et al, Biostatistics, 2005, BMC Bioinformatics, 2006*

Can be formulated for replicate arrays or single array per condition (below):

First level

$$\begin{aligned} PM_{jgc} &\sim N(\ S_{jgc} + H_{jgc}, \tau_c^2), \\ MM_{jgc} &\sim N(\phi S_{jgc} + H_{jgc}, \tau_c^2). \end{aligned}$$

Second level

$$\begin{aligned} \log(S_{jgc} + 1) &\sim TN(\mu_{gc}, \sigma_{gc}^2), \\ \log(H_{jgc} + 1) &\sim TN(\lambda_c, \eta_c^2). \end{aligned}$$

where $TN(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2 truncated at 0

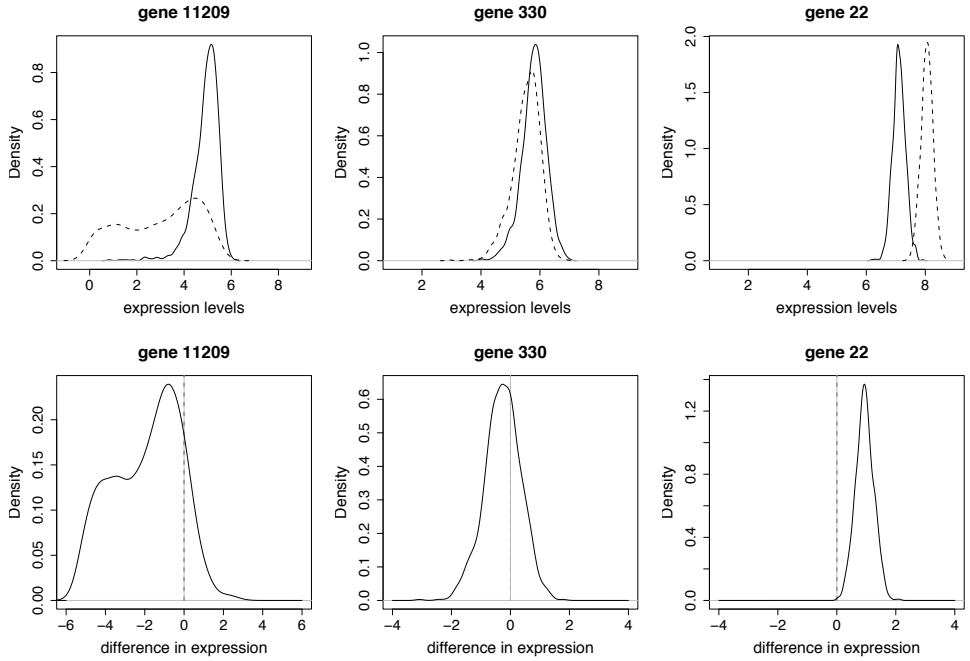
Prior structure

– Exchangeable hierarchical (lognormal) prior structure on the gene specific variability: σ_{gc}^2 (with EB for the hyperparameters)

– weakly informative priors for the other parameters

Inference from BGX

We obtain **posterior distributions** of expression levels μ_{gc} and of between condition differences $d_g = \mu_{g1} - \mu_{g2}$ from which inference on differentially expression genes can be drawn **even with only one replicate per condition**



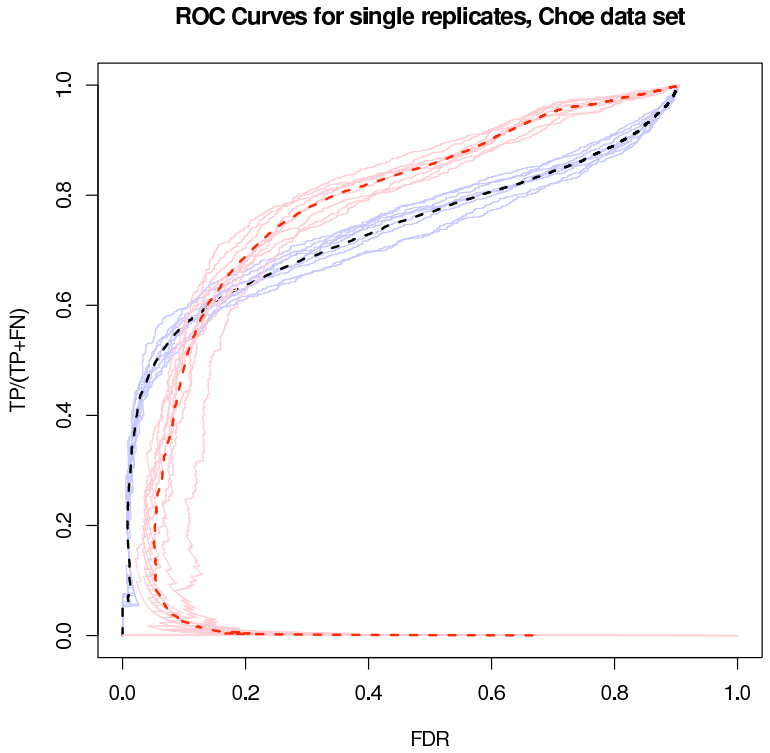
To rank genes:

- Use posterior probability $\text{Prob}(d_g < 0)$ (close to 0 or 1 for DE expressed genes)
- Use $E(d_g)/SD(d_g)$

Performance comparison

Comparison of ranking produced by BGX and GCRMA (Wu & Irizarry 2006) on data from Choe et al. (Genome Biology, 2005). ROC curves: sensitivity (TP/TP + FN) versus FDR (FP /TP + FP) for all 9 single array comparisons.

Blue: BGX, Red: GCRMA



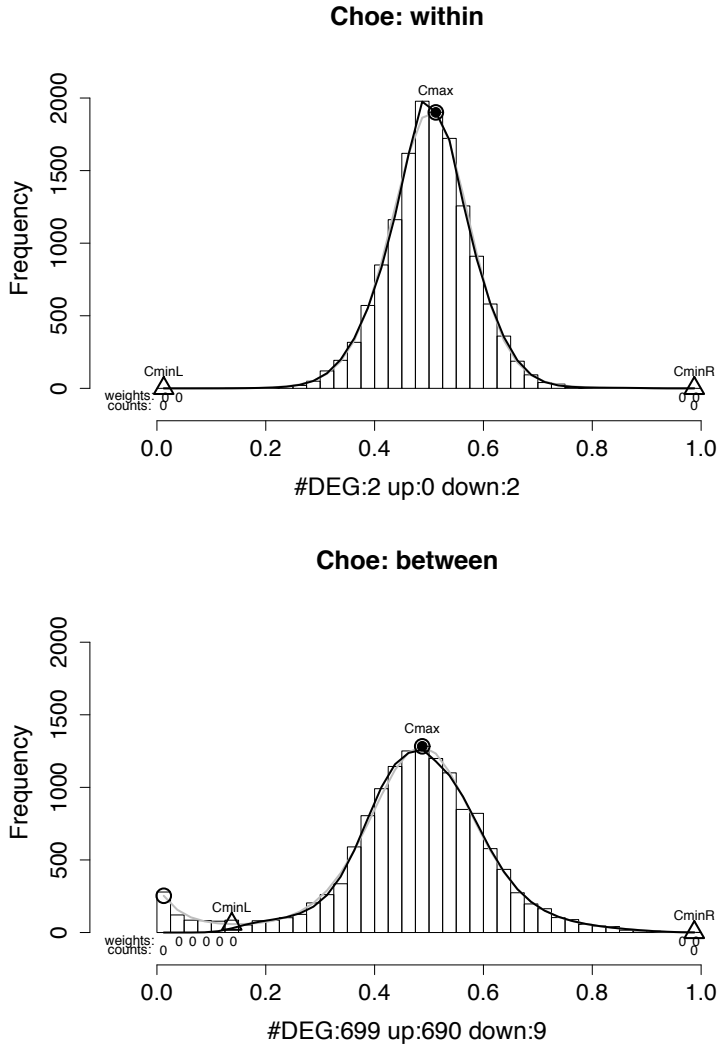
6 Drosophila GeneChip arrays, two conditions, 3 reps each.

14010 genes: 10150 non-expressed, 3860 spike-in. Of these: 2551 at same conc, 1309 at different conc (FC 1.2-4, same direction)

Note: many diff expressed genes, more realistic data than other spike-in studies

Inference from BGX: Estimating the proportion of DE genes

- Plot histogram of Prob$(d_g < 0)$
- Use central part of the histogram of Prob$(d_g < 0)$ to obtain empirical estimate of the null distribution and proportion of null genes (inspired by Efron, JASA 2004)
- For within-condition, estimate of DE genes near zero
- For between-condition, estimate of DE genes $\simeq$ 700: i.e. picks up $\simeq$ 50% of TP with an FDR $\simeq 6\%$



MCMC issues

Full conditional of many parameters is non standard

Use Random Walk Metropolis, for example to update S_{gjc} and μ_{gc}

In a typical array, more than 200 000 parameters are being updated, with a wide range of variability

How to choose the width γ of the RW proposal for each probe ?

$\Rightarrow$ Experiment with **adaptive MCMC** (Rosenthal and Roberts, 2006):

Convergence of the chain is preserved if use a sequence γ_n of values for γ such that:

- Each kernel P_γ has the right stationary distribution
- Diminishing adaptation: total variation distance between the successive kernels $P_{\gamma_n} \rightarrow 0$ (in probability)
- Bounded convergence condition

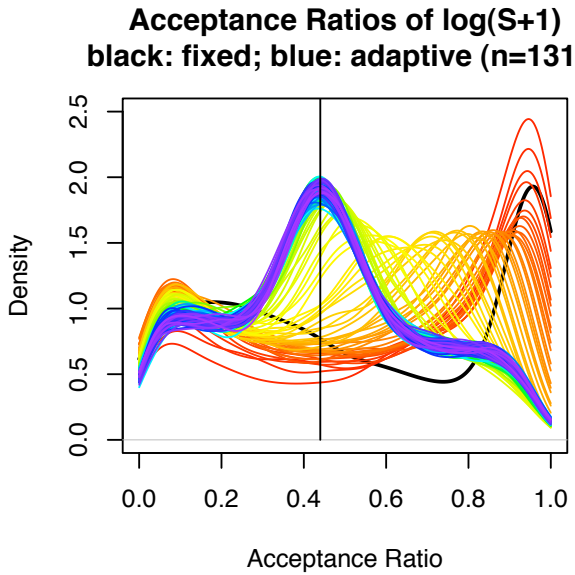
Simple adaptive scheme

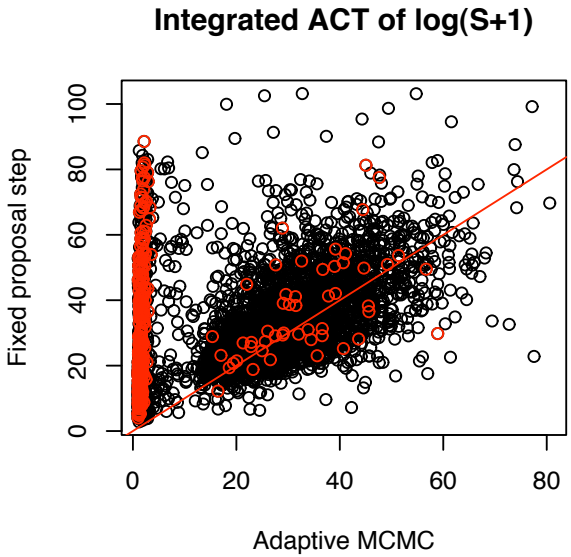
For each variable, Rosenthal and Roberts suggests:

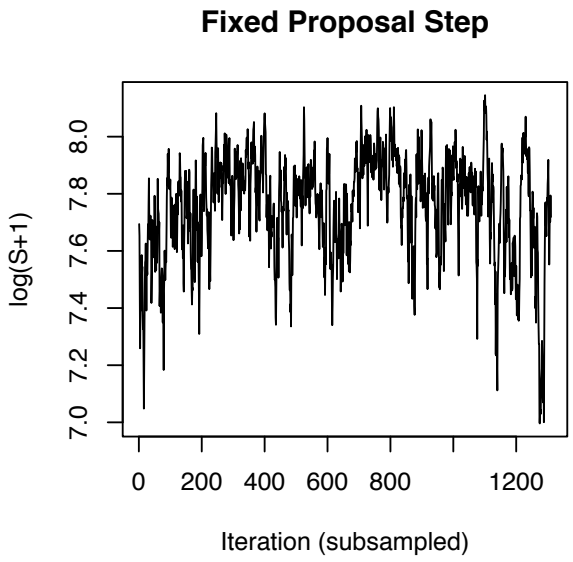
- Start with a RW with normal increment $N(0, \exp(2s))$,
- Choose a sequence $\delta(n) \to 0$

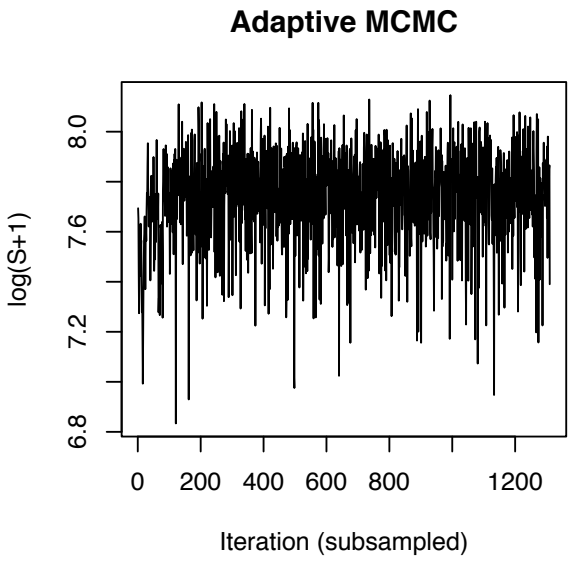
  [for ex. R& R experiment with $\delta(n) = min(0.01, n^{-1/2})$]
- After the nth batch of 50 sweeps, compute the acceptance rate (AR) over last batch of 50 sweeps
- Adapt the RW proposal by adding or subtracting $\delta(n)$ to s, following whether the AR is greater or less than 0.44

MCMC issues

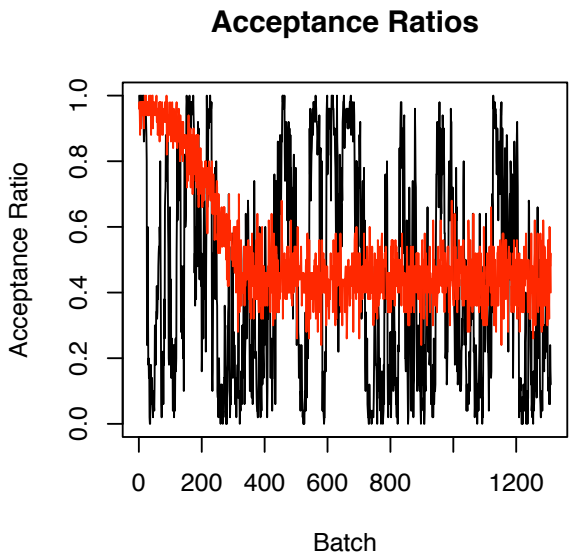

Acceptance Ratios of log(S+1)
black: fixed; blue: adaptive (n=1310)
Density
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0
Acceptance Ratio

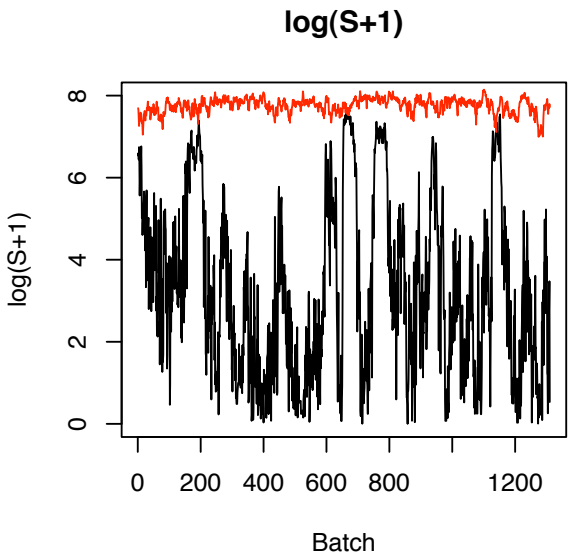

Integrated ACT of log(S+1)
Fixed proposal step
0
20
40
60
80
100
0
20
40
60
80
Adaptive MCMC

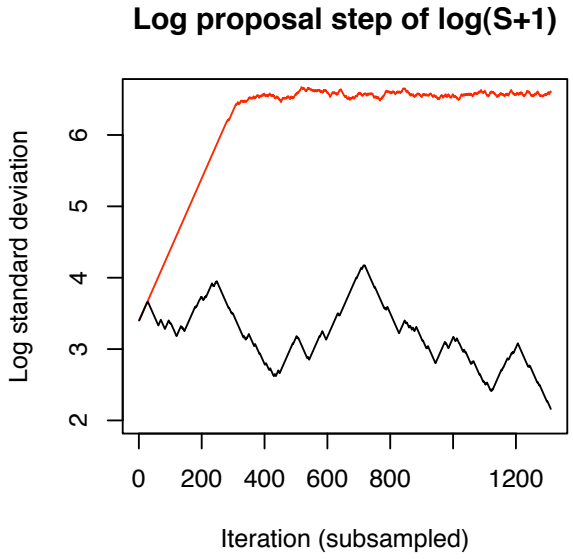

Fixed Proposal Step
log(S+1)
7.0
7.2
7.4
7.6
7.8
8.0
0
200
400
600
800
1200
Iteration (subsampled)

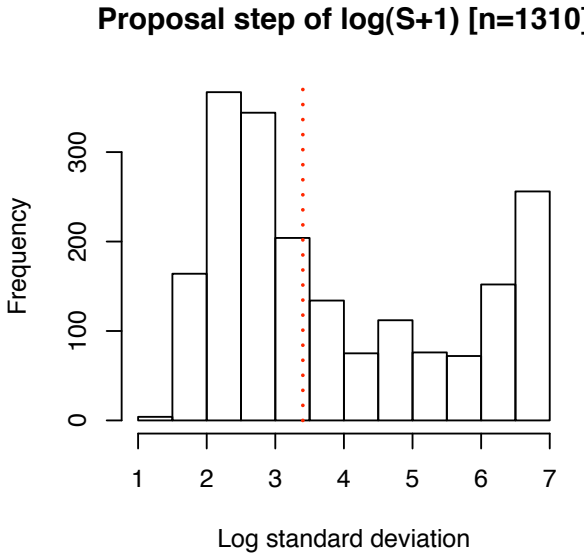

Adaptive MCMC
log(S+1)
6.8
7.2
7.6
8.0
0
200
400
600
800
1200
Iteration (subsampled)

MCMC issues

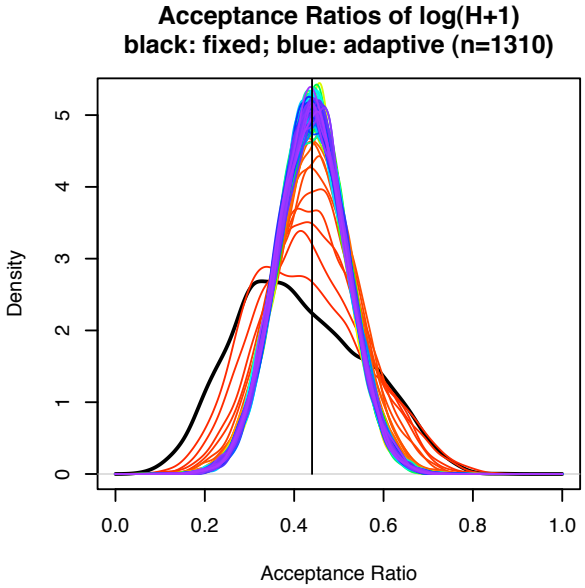

Acceptance Ratios
Acceptance Ratio
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1200
Batch

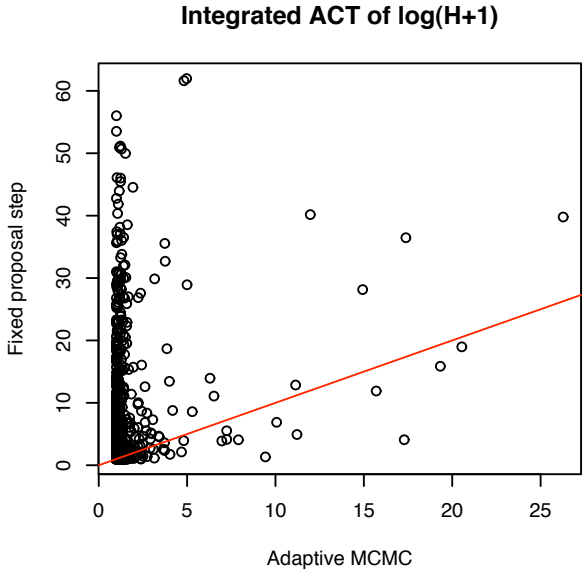

log(S+1)
log(S+1)
0
2
4
6
8
0
200
400
600
800
1200
Batch

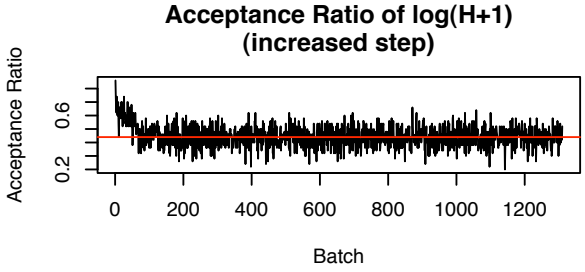

Log proposal step of log(S+1)
Log standard deviation
2
3
4
5
6
0
200
400
600
800
1200
Iteration (subsampled)

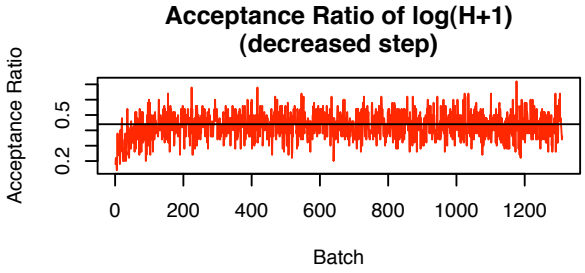

Proposal step of log(S+1) [n=1310]
Frequency
0
100
200
300
1
2
3
4
5
6
7
Log standard deviation

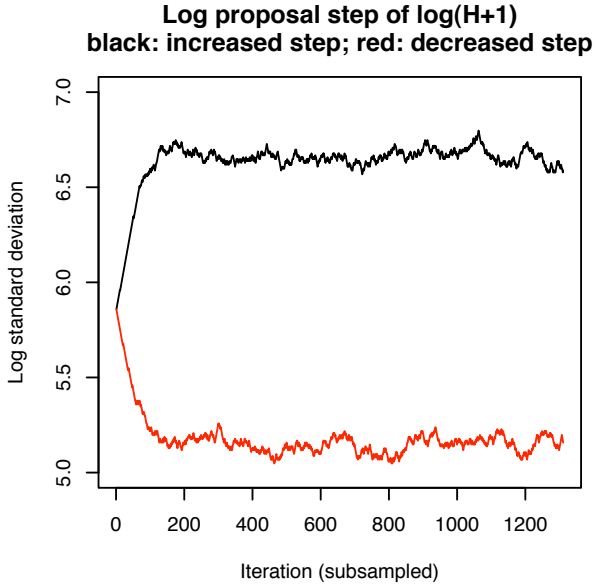
MCMC issues


Acceptance Ratios of log(H+1)
black: fixed; blue: adaptive (n=1310)
Density
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
Acceptance Ratio


Integrated ACT of log(H+1)
Fixed proposal step
0
10
20
30
40
50
60
0
5
10
15
20
25
Adaptive MCMC


Acceptance Ratio of log(H+1)
(increased step)
Acceptance Ratio
0.2
0.6
0
200
400
600
800
1000
1200
Batch


Acceptance Ratio of log(H+1)
(decreased step)
Acceptance Ratio
0.2
0.5
0
200
400
600
800
1000
1200
Batch


Log proposal step of log(H+1)
black: increased step; red: decreased step
Log standard deviation
5.0
5.5
6.0
6.5
7.0
0
200
400
600
800
1000
1200
Iteration (subsampled)

Outline

1. BGX Bayesian Gene eXpression

    MCMC issues

2. **Mixture models for differential expression**

    Model checking via mixed predictive p-values

3. Synthesising gene lists

Mixture models for differential gene expression

A widely used approach to address differential expression problems is to model the expression levels as a mixture of distributions (Lonnstedt & Speed, 2003, ..., Gottardo, 2006)

- latent quantities δ_g characterising the underlying difference between the conditions are introduced
- a mixture prior model for δ_g is defined with point mass at 0 and a parametric distribution for the alternative

Fully Bayesian implementation where the proportion in the null is estimated along with other parameters has only been recently developed

The choice of the alternative distribution in the mixture can be influential

2 – Mixture models for Differential expression

Model for differential gene expression

Data is log gene expression x_{gci}

$g = 1, ..., m$ (gene)

$c = 1, 2$ (experimental condition)

$i = 1, ..., n$ (replicate measure)

First level

$$x_{g1i}|\mu_g, \delta_g, \tau_g \sim N(\mu_g - \frac{1}{2}\delta_g, \tau_g^{-1})$$

$$x_{g2i}|\mu_g, \delta_g, \tau_g \sim N(\mu_g + \frac{1}{2}\delta_g, \tau_g^{-1})$$

rewrite this as

$$\bar{x}_{gc}|\mu_g, \delta_g, \tau_{gc} \sim N(\mu_g \mp \frac{1}{2}\delta_g, (n\tau_g)^{-1})$$

$$S_{gc}|\tau_{gc} \sim Gam(\frac{1}{2}(n-1), \frac{1}{2}(n-1)\tau_g)$$

where $\bar{x}_{gc}$ is mean, S_{gc} is sum of squares $\frac{1}{(n-1)}\sum_i (x_{gci} - \bar{x}_{gc.})^2$.

Define $d_g = \bar{x}_{g2.} - \bar{x}_{g1.}$

2 – Mixture models for Differential expression

Second level

Exchangeable prior for gene precision

$$\tau_g \sim Gam(\alpha, \beta)$$

Mixture prior for log ratios

$$\delta_g | z_g, \lambda, \eta \sim \begin{cases} Gam^{(-)}(\lambda_-, \eta_-) & z_g = -1 \\ \delta(0) & z_g = 0 \\ Gam(\lambda_+, \eta_+) & z_g = 1 \end{cases}$$

where $Gam^{(-)}(x|\lambda_-, \eta_-) = Gam(-x|\lambda_-, \eta_-)$.

Each component characterizes log ratios for groups of genes:

$z_g = 0$ for non-differentially expressed genes (null hypothesis)

$z_g = 1, -1$ for over and under expressed genes

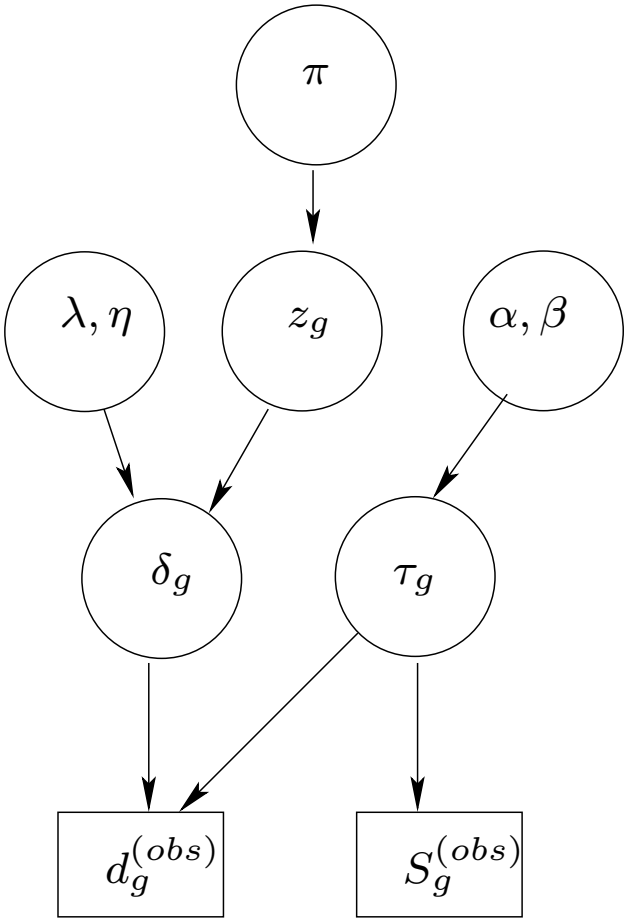
2 – Mixture models for Differential expression

Third/fourth levels

$z_g = -1, 0, 1$ have prior $\mathbb{P}(z_g = j|\pi_j) = \pi_j$.

π_j (mixture weights) have Dirichlet prior.

$\alpha, \beta, \lambda, \eta$ are given Gamma priors.



MCMC Implementation

Joint update for the δ_g and z_g for each feature g.

Performance of mixture model *Lewin, Bochkina & SR, in preparation*

Using realistic simulated data (noise distribution from real data) where differential expression follows a variety of distributions (not those used in the mixture model), we found:

- 2-sided Gamma alternative is flexible and seems to adapt to a range of shapes of simulated alternatives. **How can we check this?**
- Posterior allocation probabilities of the fitted mixture: Prob$(z_g = i|$ data) can be used to build classification rules (not necessary Bayes rule), i.e. to define a set S_{rej} of genes classified in the alternative
- Good estimates of False Discovery Rate can be obtained for any rule:

$$1/\#S_{rej} \sum_{g \in S_{rej}} Prob(z_g = 0|data)$$

2 – Mixture models for Differential expression

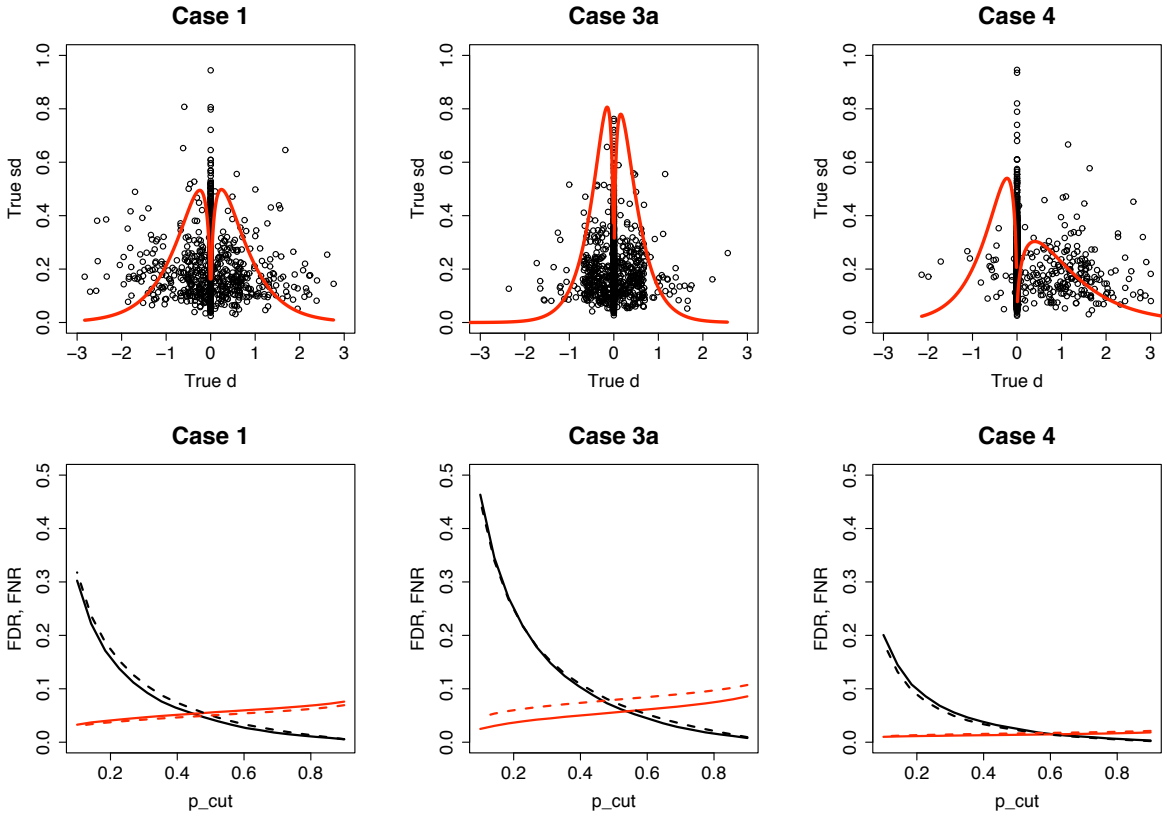
The 3 cases differ in the shape of the alternative (a mixture of uniform and normal) and values of π_0:

Case 1: $\pi_0 = 0.8$, alternative close to normal

Case 2: $\pi_0 = 0.8$, alternative close to uniform

Case 4: $\pi_0 = 0.9$, asymmetric alternative

Bottom plots: Estimated and 'true' FDR and FNR



Model checking via mixed predictive p-values

Aims of model checking

1. Get some idea of how well a given model fits the data, without considering alternative models $\longrightarrow$ **Bayesian predictive p-values**
2. Want measure of model fit for each feature so we can use the ensemble to judge model fit/look for outliers $\longrightarrow$ **p-value for each feature**

Prediction Scheme

Cross-validation: For each feature, take data out, predict data from model fit to rest of data. Not practical here.

Posterior predictive distribution: Predict data for each feature simultaneously (no data removed). Conservative.

Mixed prediction: In hierarchical model, predict new intermediate level parameters before predicting new data. Much less conservative. (Gelman, Meng and Stern, 1996; Marshall and Spiegelhalter, 2003). Use in gene expression models (Lewin et al, 2006).

Choice of parameters to predict

Differences d_g are dependent on 3 sets of gene specific parameters:

δ_g, τ_g and z_g.

Choice of **what quantities to predict** to implement mixed predictive checks:

δ_g: **yes** - these have the most influence on the data

τ_g: **no** - find results similar whether or not these are predicted

z_g: **no** - want to look at separate mixture components

Mixed Prediction for Log ratios

$$\delta_g^{pred}|z_g, \lambda, \eta \sim H_{z_g}(\lambda, \eta)$$ (H = parametric form of the mixture component)

$$d_g^{mixpred} \sim N\left(\delta_g^{pred}, \left(\frac{n\tau_g}{2}\right)^{-1}\right)$$

Compute $\nu_{gj} \equiv \mathbb{P}(d_g^{mixpred} > d_g^{obs}|\mathbf{x}^{obs}, z_g = j)$

Model checking

Graphical representation of predictive schemes:

the mixed predictive $d_g^{mix-pred}$ does not condition on δ_g

the post predictive $d_g^{post-pred}$ conditions on δ_g and is very conservative

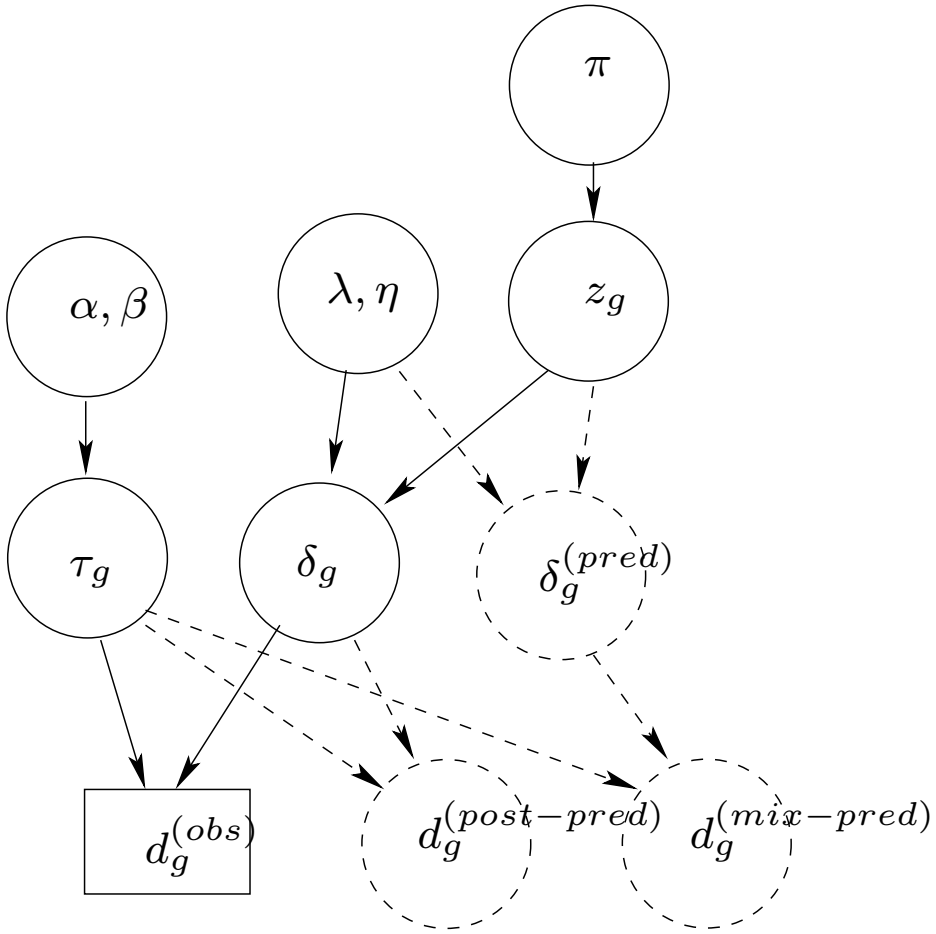


Figure 1: Model (solid), prediction (dashed)

Model checking

Note we do not expect approximately uniform ν_{gj} p-values for misclassified genes.

$\Rightarrow$ Useful to restrict the plots of ν_{gj} to genes with high posterior allocation, e.g.

$pr(z_g = j|\mathbf{x}^{obs}) > 0.5$

On the RHS, typical predictive plots when the simulated and fitted mixture models are the same

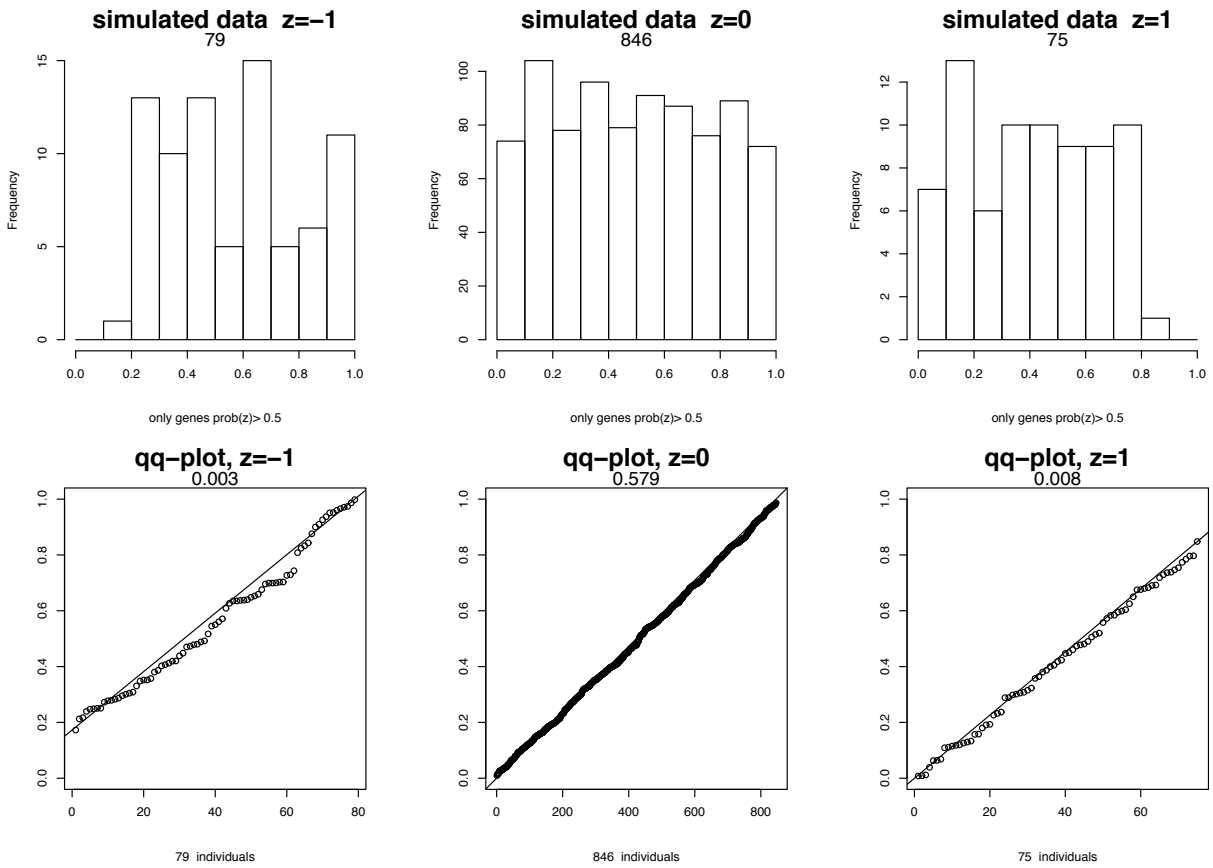


Illustration on gene expression data set

Experiment: 6 knock-out (knock-out gene involved in insulin resistance) and 5 wildtype mice.

Processed using MAS5.0 software

Mixture models:

1. $\delta_g \sim \pi_0 \delta_0 + \pi_{+1}$Unif(0,3) $+ \pi_{-1}$Unif(-3,0)
2. $\delta_g \sim \pi_0 \delta_0 + \pi_{+1}\text{Gam}^{+}(\lambda_+, \eta_+) + \pi_{-1}\text{Gam}^{-}(\lambda_-, \eta_-)$
3. $\delta_g \sim \pi_0 N(0, \tau_\epsilon) + \pi_{+1}\text{Gam}^{+}(\lambda_+, \eta_+) + \pi_{-1}\text{Gam}^{-}(\lambda_-, \eta_-)$

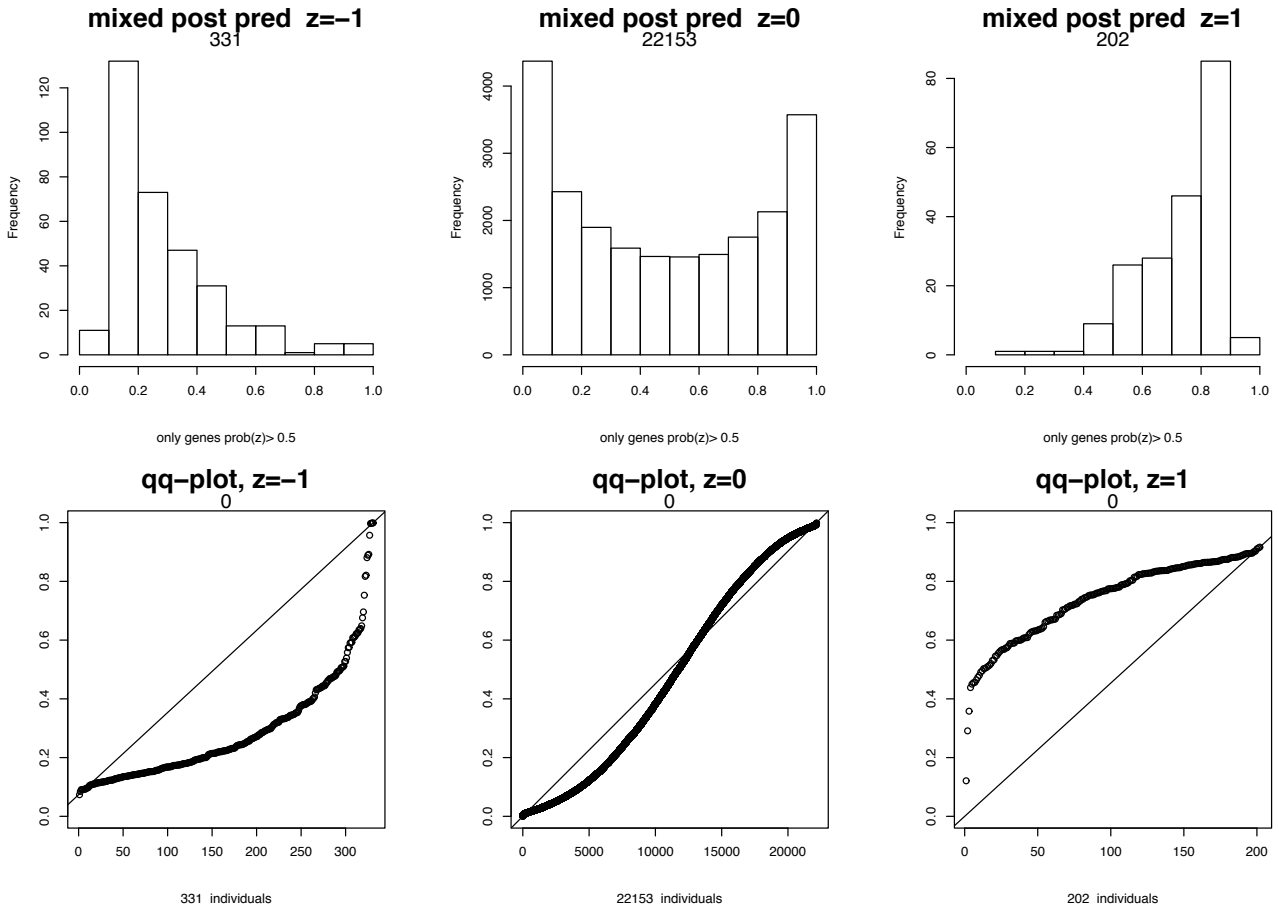
These 3 mixture models lead to different results for classification of differentially expressed genes:

Model 1: π_0 = 0.96 Model 2: π_0 = 0.68 Model 3: π_0 = 0.99

Model checking

Model 1: Mixture of point mass and Uniform:

$\delta_g \sim \pi_0\delta_0 + \pi_{+1}$U(0,3) $+ \pi_{-1}$U(-3,0).

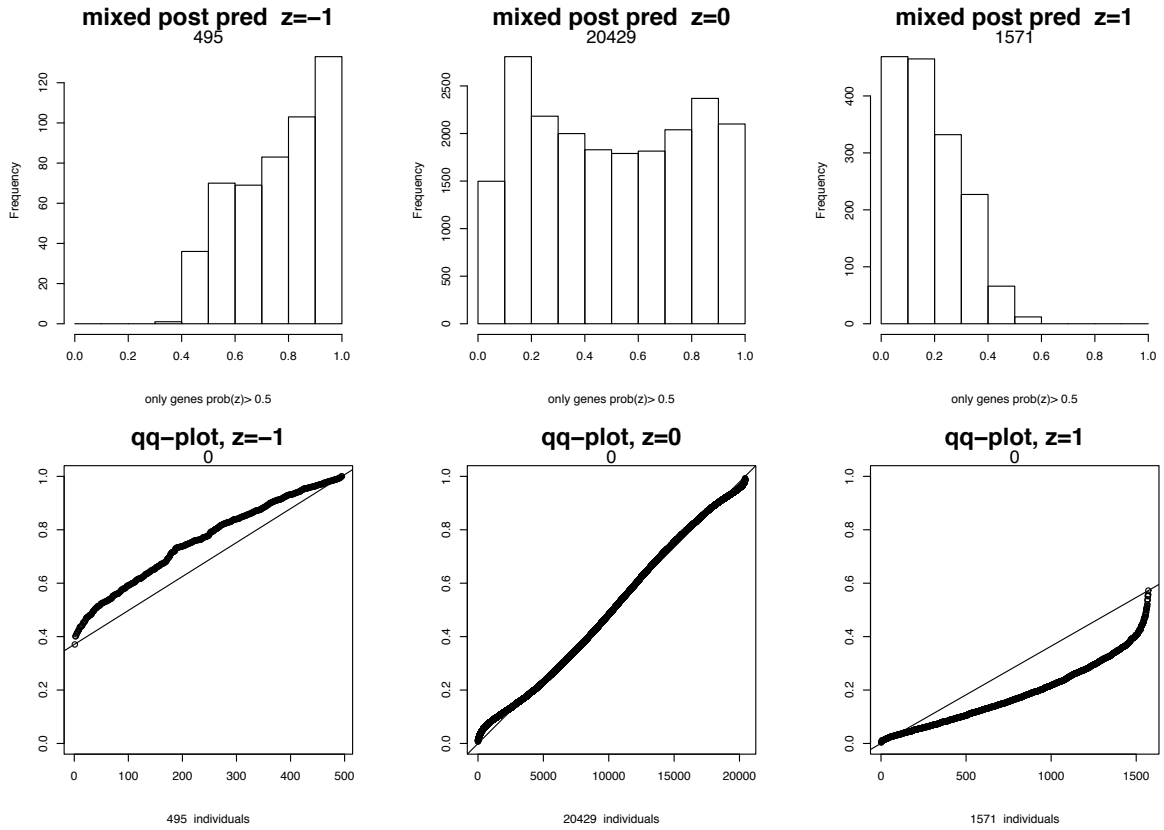


Excess of extreme values in the null, very skewed distribution in the outer components $\Rightarrow$ **Alternative has too much weight on extreme values, bad fit**

Model checking

Model 2: Mixture of point mass and gammas:

$$\delta_g \sim \pi_0\delta_0 + \pi_{+1}\text{Gam}^{+}(\lambda_+, \eta_+) + \pi_{-1}\text{Gam}^{-}(\lambda_-, \eta_-)$$

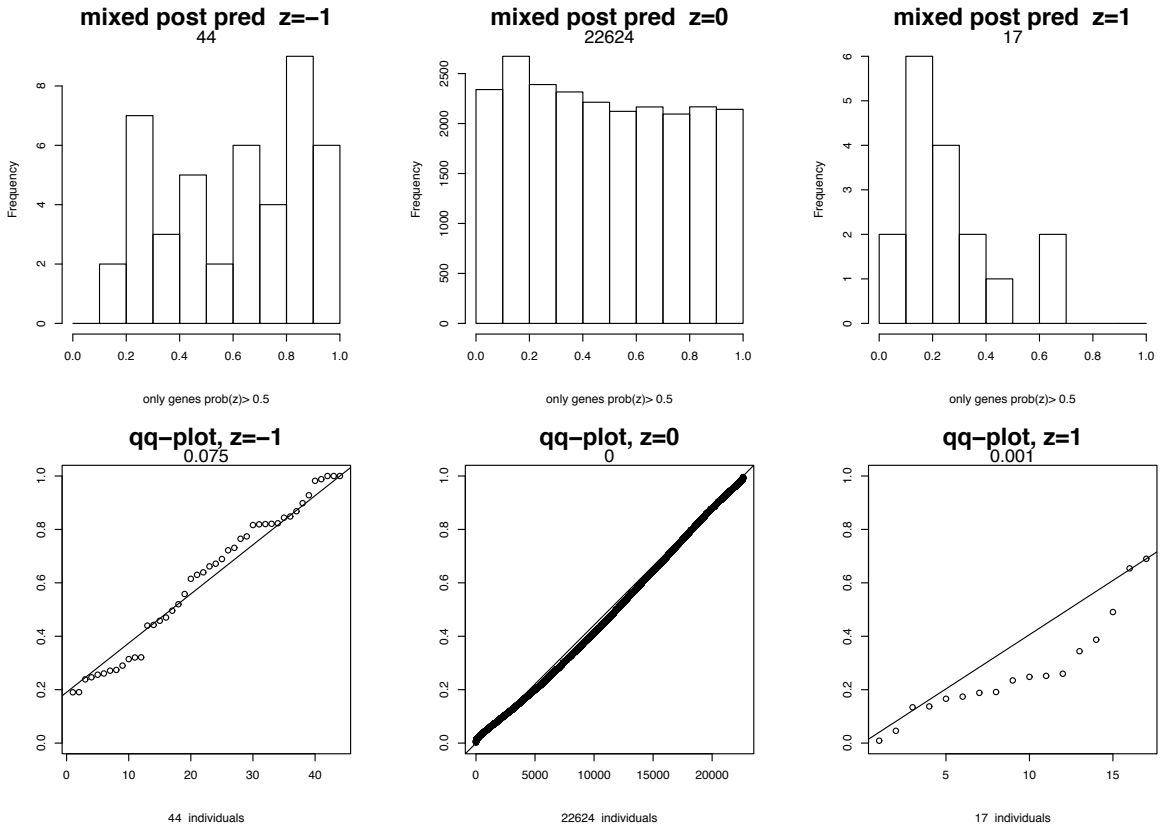


Null component better, still some excess of extreme p-values. Deficit of small $\nu_{g,-1}$ and large $\nu_{g,1}$ due to overlap between mixture components or too narrow null?

Model checking

Model 3: Mixture with 'nugget null' and gammas:

$$\delta_g \sim \pi_0 N(0, \tau_\epsilon^{-1}) + \pi_{+1}\text{Gam}^{+}(\lambda_+, \eta_+) + \pi_{-1}\text{Gam}^{-}(\lambda_-, \eta_-)$$



Better fit! Few genes with high post prob so more difficult to judge the histograms in the outer components.

Mixed predictive checks

- Promising tool to explore different aspects of model specification
- For mixture priors, useful to condition on the component allocation. But some arbitrariness in choice of cut-off on $pr(z_g = j|\mathbf{x}^{obs})$
- More work is needed to investigate patterns of departure from uniformity

Outline

1. BGX Bayesian Gene eXpression

    MCMC issues

2. Mixture models for differential expression

    Model checking via mixed predictive p-values

3. **Synthesising gene lists in related experiments**

Synthesising gene lists

Synthesising lists of differentially expressed features in related experiments *Blangiardo & SR, 2006*

- System biology investigations are often interested in the comparison of two or more similar experiments .

  e.g. effect of a knock out gene on several tissues or species

- The aim is to find **common denominators between these experiments**

  i.e. a parsimonious list of features (e.g. genes, biological processes) for which there is strong evidence that the listed features are commonly perturbed in all the experiments

- Joint re-analysis of all experiments not always possible and time consuming,

  $\Rightarrow$ Consider ranked list of features for each experiments

  Define intersections of the lists and assess the strength of association

Synthesising gene lists

Suppose we have two experiments, each reporting a measure of strength of evidence of differential expression on a probability scale (e.g. p-value) :

Small p value $\Longrightarrow$ MOST 'differentially expressed'

p value nearer 1 $\Longrightarrow$ NOT 'differentially expressed'

Experiment A	Experiment B
p_{A1}	p_{B1}
p_{A2}	p_{B2}
$\ldots$	$\ldots$
p_{An}	p_{Bn}

- Consider sequence of cut offs q on the p-values
- Count the number of differentially expressed genes in common
- Compute sequence of Ratios: Observed in common to Expected in common under independence of the lists

Synthesising gene lists

Bayesian model

For each probability threshold q, starting from the 2×2 table, we specify a multinomial distribution for the vector of joint frequencies $Multi(\mathbf{O} \mid \boldsymbol{\theta}, n)$

		Exp B DE	Exp B $\overline{DE}$	
Exp A DE		$O_{11}(q)$	$O_{1+}(q) - O_{11}(q)$	$O_{1+}(q)$
Exp A $\overline{DE}$		$O_{+1}(q) - O_{11}(q)$	$n - O_{+1}(q) - O_{1+}(q) + O_{11}(q)$	$n - O_{1+}(q)$
		$O_{+1}(q)$	$n - O_{+1}(q)$	n

The vector of parameters $\boldsymbol{\theta}$ is given a Dirichlet(0.25,0.25,0.25,0.25) prior, so that the posterior distribution of $\boldsymbol{\theta} \mid \mathbf{O}, n$ is again Dirichlet.

The quantity of interest is the ratio of the probability that a gene is in common, to the probability that a gene is in common by chance:

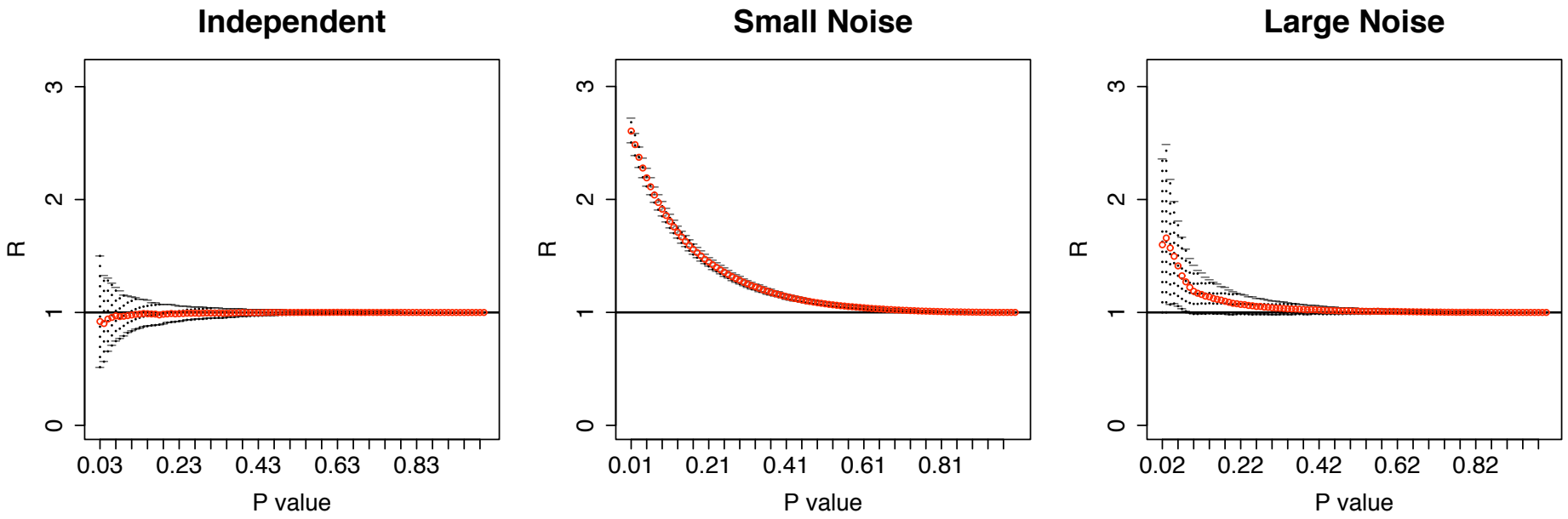
$$R(q) = \frac{\theta_{O_{11}(q)}}{\theta_{O_{1+}} \times \theta_{O_{+1}}}$$

Compute Median and 95% Credibility intervals for $R(q) \mid \mathbf{O}, n$ for each threshold q.

Synthesising gene lists

Ratios $R(q)$ and Credibility intervals

3 simulated cases. Average results of 50 repetitions



Synthesising gene lists

Decision rules

Typically, many Credibility Intervals exclude 1

How to select useful thresholds q and associated lists $O_{11}(q)$?

Different 'utilities':

1. Parsimonious list (small q) giving the least false positive, i.e. genes wrongly declared to be commonly perturbed

   OR

2. Larger list that achieves a balance between false positives and false negatives?

Synthesising gene lists

Parsimonious list: consider

$$q_{max} = \arg\max\{Median(R(q) \mid \mathbf{O}, n)\}$$

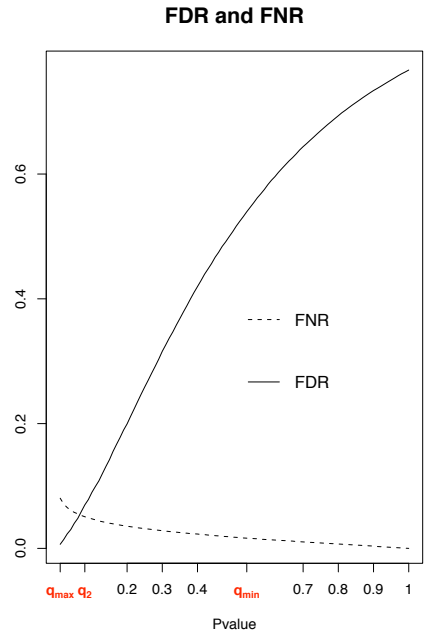
over the set of values of q for which $CI_{95}(q)$ excludes 1.

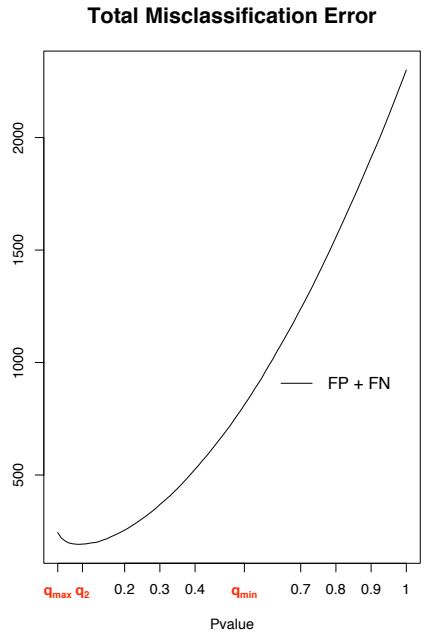
Balanced list: consider

$$q_2 = \max\{q\ ; Median(R(q) \mid \mathbf{O}, n) \geq 2$$

and $CI_{95}(q)$ excludes $1\}$

q_2 is the largest threshold where the number of genes called in common **at least doubles** the number of genes in common under independence





Synthesising gene lists

In extensive simulation studies, we found

– q_{max} to be **highly specific but conservative**

– q_2 to be **close to report lists with minimum global misclassification errors**

For example, for 2 lists of 3000 genes generated so that

– There are 1000 DE genes for Experiment A and 800 DE genes for Experiment B

– 700 genes are truly in common

	Joint Bayesian Model						
Small noise	q	$R(q)$	$95\%CI$	O_{11}	FP	FN	Minimum error
q_{max}	0.01	2.88	2.76-3.02	464	3	241	
q_2	0.08	2.01	1.94-2.09	588	41	153	192
Larger noise							
q_{max}	0.01	4.13	3.83-4.46	187	2	515	
q_2	0.09	2.02	1.91-2.13	348	39	391	428

Synthesising gene lists

Case study: Effect of High Fat Diet vs Normal Fat Diet on gene expression in 2 tissues: fat and muscle in mice

Data from Diabetes Genome Anatomy project.

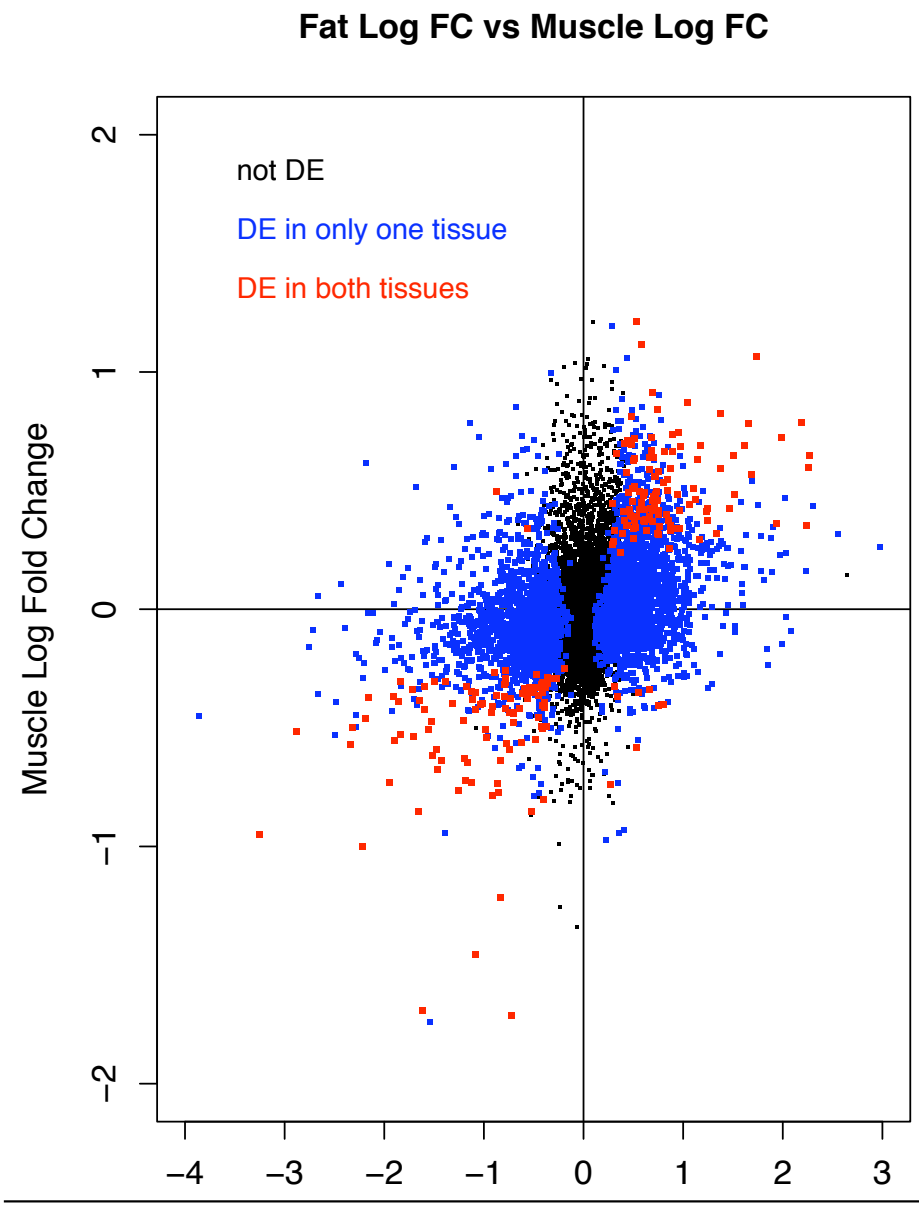
3 replicate array for each tissue and each condition

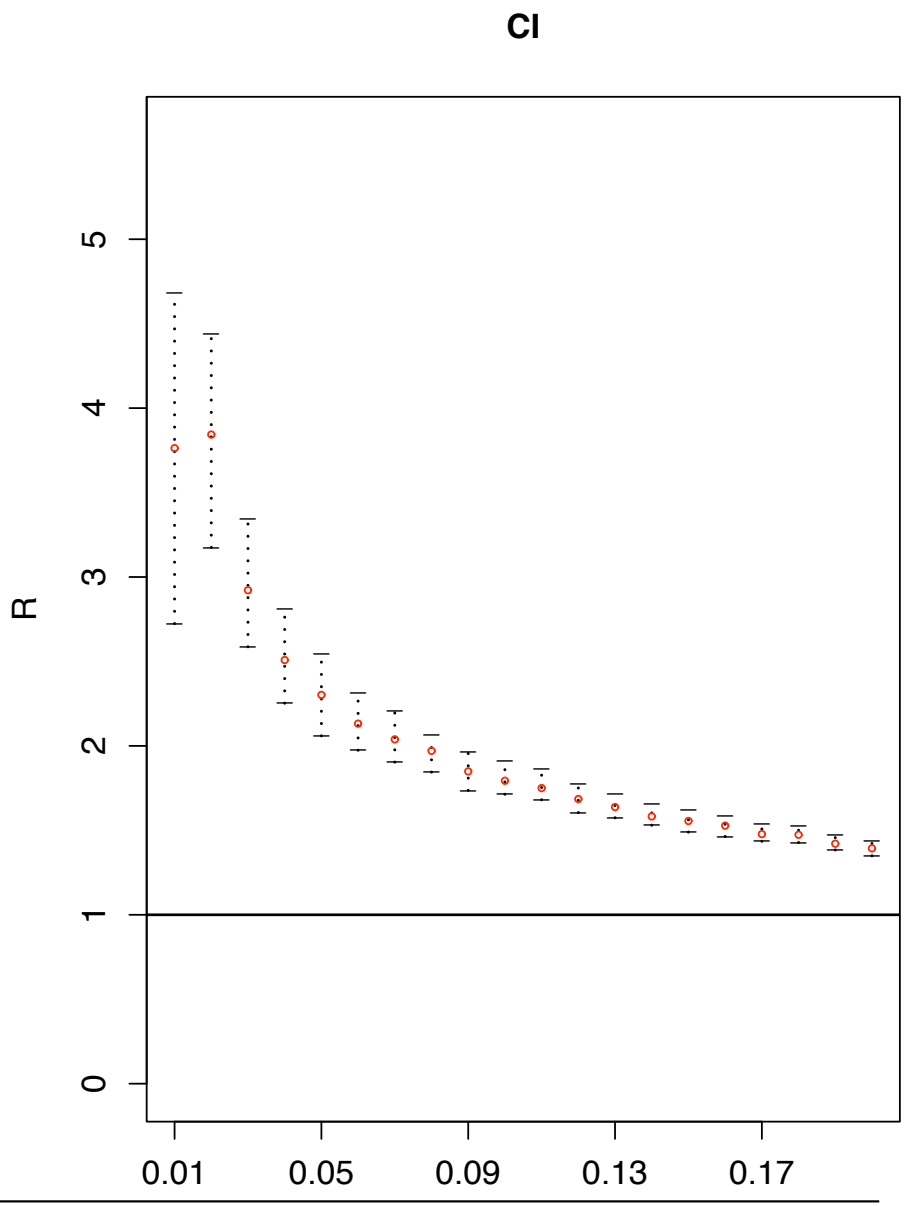
List of p-values obtained separately for each tissue by Cyber-T

Aim: investigate common perturbation of biochemical processes induced by HFD in the 2 tissues

	Joint Bayesian Model					
Rules	q	$R(q)$	O_{11}	O_{1+}	O_{+1}	$95\%CI$
max	0.01	3.76	20	1482	44	2.72-4.68
2	0.07	2.04	226	3059	452	1.90-2.21

Synthesising gene lists





Synthesising gene lists

Results from the synthesis of High Fat diet in two tissues

GO Annotation for 226 genes found by rule q_2:

- 122 involved in **biological processes** categories
- 124 in **molecular function** categories
- 127 in **cellular components** categories

172 in **KEGG pathways**

Cellular Components: "Mitochondria", organelles where oxidative reactions take place (responsible for ATP synthesis).

Molecular Functions: "Oxidative Reaction"

$\longrightarrow$ confirms an involvement of oxidation and energy production related to the switch of diet in more than one tissue

$\longrightarrow$ Both are linked to chronic obesity and diabetes in literature.

Coherent results with KEGG pathways: "Cytokine-Cytokine receptor interaction", regulates extra cellular signals transmitted to the nucleus of the cell, e.g. inflammation as a result of an HFD.

Concluding remarks

Integrated gene expression analysis via Bayesian modelling

- Uses the natural hierarchical structure of the data
- Is able to effectively synthesise information at many levels

  e.g. from low level probe-based models to models for combining summary information from different experiments
- Provide realistic quantification of uncertainty

  $\longrightarrow$ Posterior distributions can be exploited for inference with no or few replicates

  $\longrightarrow$ Interesting questions concerning the choice of decision rules
- Model checking becomes integral part of the modelling process
- Model based classification, e.g. through mixtures, provides interpretable output and a structure to deal with multiplicity

Conclusion

Implementation is challenging

- Convergence issues in very high dimensional space
- Adaptive methods seem promising and warrant further exploration
- Parallelising is important

Beyond the benefits in gene expression analysis, useful framework for investigating general questions of

– interplay between style of modelling and level of information

– decision rules to exploit the rich output

– benefits of fully integrated analysis versus ‘piecemeal’ analysis

Thank You

Collaborators at Imperial: Marta Blangiardo, Natalia Bochkina, Anne Mette Hein (now at Aarhus), Alex Lewin, Ernest Turro

Co-PI of the BGX project: Peter Green (Bristol)

Colleagues in Biology: Tim Aitman, Helen Causton

BBSRC Exploiting Genomics Initiative, Wellcome Trust Functional Genomics Initiative

Papers and tech reports at `www.bgx.org.uk`