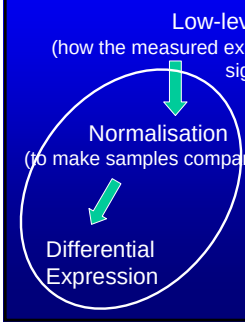
Bayesian modelling of differential gene expression data

Alex Lewin, with Sylvia Richardson, Clare Marshall, Anne Glazier and Tim Aitman (Imperial College)

In collaboration with
Helen Causton (Imperial Microarray Centre)
Anne-Mette Hein (Imperial)
Peter Green and Graeme Ambler (Bristol)

Gene expression analysis is a multi-step process



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

Bayesian hierarchical model framework

- Model different sources of variability simultaneously, within array, between array, estimation of gene specific variability …
- Uncertainty is propagated from data to parameter estimates
- 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)

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)

Bayesian hierarchical model for genes under one condition (I)

Data: y_{gr} = log gene expression for gene g, replicate r (can be any estimate of signal: Affymetrix, Li and Wong etc.)

α_g = gene effect

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

σ_g^2 = gene variance

- 1st level

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

$\beta_{r(g)}$ = function of α_g , parameters {**a**} and {**b**}

Bayesian hierarchical model for genes under one condition (II)

- 2nd level

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

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

We will discuss:

- Array effects (normalisation)
- Bayesian model checks on gene variances
- Confounding of differential and array effects
- Rank statistics

Details of array effects

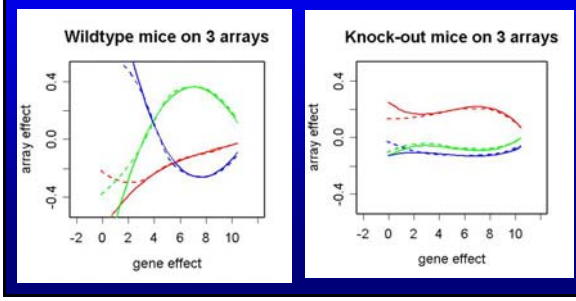
Exploratory work shows need for expression-level dependent normalisation

Piecewise polynomial with unknown break points:
$\beta_{r(g)}$ = quadratic in α_g for $a_{rk-1} \leq \alpha_g \leq a_{rk}$
with coeff ($b_{rk}^{(1)}$, $b_{rk}^{(2)}$), k =1, … #breakpoints

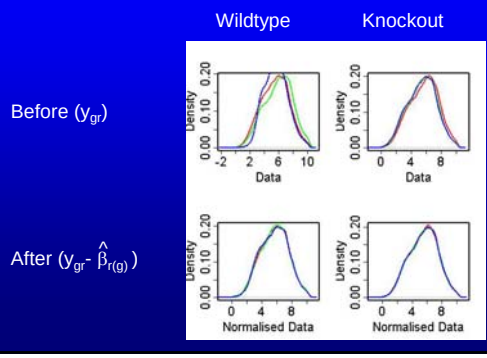
- Locations of break points not fixed
- Must do sensitivity checks on # break points
- Cubic fits well for this data

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

cubic
loess



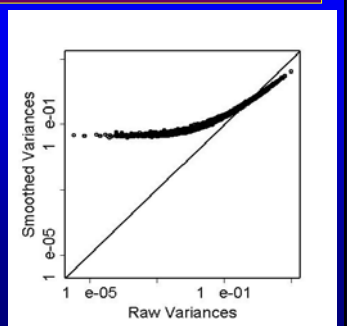
Effect of normalisation on density



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 assumptions on gene variances
- Predict sample variance $S_g^{2\ new}$ from the model for each gene
- Compare predicted $S_g^{2\ new}$ with observed $S_g^{2\ obs}$

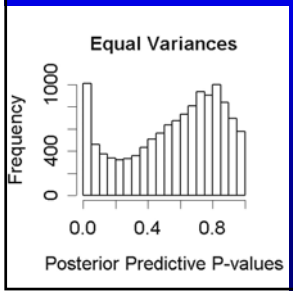
Bayesian p-value Prob($S_g^{2\ new} > S_g^{2\ obs}$)

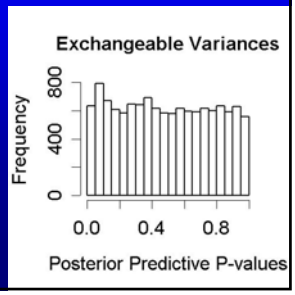
- Distribution of p-values Uniform if model is adequate
- Easily implemented in MCMC algorithm

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 :

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

$y_{g2r} \sim N(\alpha_g + \frac{1}{2} d_g + \beta_{r(g)2}, \sigma_{g2}^2)$, (condition 2)

So $E(\bar{y}_{g2\cdot} - \bar{y}_{g1\cdot}) = d_g$

Prior can be put on d_g directly

Possible Statistics for Differential Expression

$d_g \approx$ log fold change

$d_g^* = d_g / (\sigma^2_{g1}/3 + \sigma^2_{g2}/3)^{1/2}$ (standardised difference)

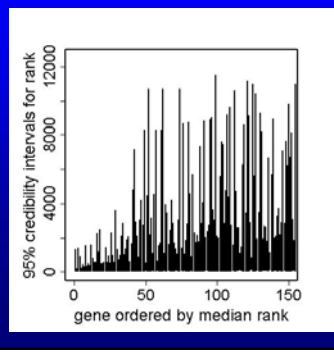
- We obtain the joint distribution of all $\{d_g\}$ and/or $\{d_g^*\}$
- Distributions of ranks

Credibility intervals for ranks

Ranks of modelled log fold change

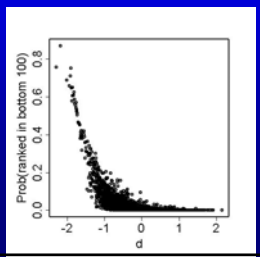
150 genes with lowest rank

Even genes with median rank less than 100 can have large 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

- Model different sources of variability in a single model
- Borrow information from all genes to stabilise estimates of gene specific variances
- Use joint distribution of ranks for inference
- Future work: mixture prior on log fold changes, with uncertainty propagated to mixture parameters