Bayesian Modelling of Differential Gene Expression

Alex Lewin,[1,*] **Sylvia Richardson,**[1] **Clare Marshall,**[1] **Anne Glazier**[2] **and Tim Aitman**[2]

[1]Department of Epidemiology and Public Health, Imperial College, Norfolk Place, London W2 1PG, UK

[2]MRC Clinical Sciences Centre, Imperial College, Hammersmith Hospital, London W12 0NN, UK

February 11, 2005

Summary

We present a Bayesian hierarchical model for detecting differentially expressing genes that includes simultaneous estimation of array effects, and show how to use the output for choosing lists of genes for further investigation. We give empirical evidence that expression-level dependent array effects are needed, and explore different non-linear functions as part of our model-based approach to normalization. The model includes gene-specific variances but imposes some necessary shrinkage through a hierarchical structure. Model criticism via posterior predictive checks is discussed. Modelling the array effects (normalization) simultaneously with differential expression gives fewer false positive results. To choose a list of genes, we propose to combine various criteria (for instance, fold change and overall expression) into a single indicator variable for each gene. The posterior distribution of these variables is used to pick the list of genes, thereby taking into account uncertainty in parameter estimates. In an application to

[*] *email:* a.m.lewin@imperial.ac.uk

mouse knockout data, Gene Ontology annotations over and under-represented amongst the genes on the chosen list are consistent with biological expectations.

KEY WORDS: Microarray; Differential Expression; Normalization; Bayesian analysis; Hierarchical model; Model checks; MCMC; WinBUGS

1. Introduction

Analysing gene expression data is a complex process due to the many sources of variability, both obscuring and interesting, that combine to produce the recorded level of transcription. Reviews of the underlying biological process have been written by many authors (e.g. Nguyen et al., 2002, and references therein). In this paper we are concerned with modelling differential gene expression between two experimental conditions.

In early studies, differential gene expression was assessed using the fold change of genes under different conditions (Chen, Dougherty, & Bittner, 1997). This is straightforward, but as genes are known to vary widely under different conditions, raw fold change measures are not directly comparable between genes. Recently, differential expression has been assessed using standardised fold change, which is the fold change divided by an estimate of its variability (e.g. Tusher, Tibshirani, & Gilbert, 2001).

One recognised difficulty is that most microarray experiments have very small numbers of replicate arrays. In order to obtain stable results, further assumptions are needed on the variability of gene expression between replicates under the same condition. The SAM method (Tusher et al., 2001) shrinks the raw variance estimates by adding a constant to the estimated standard deviation. Lönnstedt & Speed (2003) use a hierarchical model with exchangeable gene variances to shrink the variance estimates.

As well as the uncertainty in expression due to biological variability, there are systematic effects in gene expression measurements due to differences between arrays. Adjustment for this is known as normalization. In many cases, it has been found necessary

to go beyond simple (constant) normalization procedures and to make this adjustment dependent on the level of expression (Schadt et al., 2000), and various algorithms have been proposed to deal with this (e.g. Schadt et al., 2000; Kepler, Crosby, & Morgan, 2002; Workman et al., 2002) in a pre-processing step of the analysis.

In this paper, we propose an integrated statistical approach where intrinsic biological variability, systematic array effects and differential expression are modelled simultaneously. Our starting point is an ANOVA formulation, as suggested by Kerr, Martin, & Churchill (2000). This is the first level of a Bayesian hierarchical model, the distribution of the gene variances being modelled at the next level. We estimate our model in a fully Bayesian way, with the WinBUGS software (Spiegelhalter, Thomas, & Best, 1999). This enables us to obtain the posterior distribution of any parameter in the model and any function of parameters. We show how to exploit these posterior distributions to assess differential expression in a new way, using multiple criteria. The uncertainty in parameter estimates is thereby incorporated in a natural manner into a declared list of interesting genes. A Bayesian estimate of the false discovery rate (FDR; Benjamini & Hochberg, 1995; Storey, 2002) is used to find a reasonable cut-off point on the list.

Several people have carried out Empirical Bayes analysis on microarray data (Efron et al., 2001; Lönnstedt & Speed, 2003; Newton et al., 2004). Other fully Bayesian approaches include Ishwaran & Rao (2003) and Bhattacharjee et al. (2004).

Details of the data used to motivate the model in this paper are given in Section 2. Our integrated model for differential expression is presented in Section 3. The model is justified with exploratory work and model checks in Section 4. Section 5 discusses the advantages of the integrated model over a two-step procedure where data is normalized in a pre-processing step. Results obtained from implementing the full model on our data set, with discussion of FDR, decision rules and Gene Ontology annotations, are

given in Section 6. We conclude with a brief discussion.

2. Data

The Cd36 gene has been found to play an important role in the syndrome of insulin-resistance (Aitman et al., 1999). Cd36-deficient mice and rats have defective metabolism of fatty acids and triglycerides. The data used in this paper derives from a set of experiments on three wildtype mice and three mice with Cd36 removed (Febbraio et al., 1999). Samples of peritoneal fat were taken from each animal. RNA from the 6 samples was hybridized on Affymetrix U74A-C chips, making a total of 18 microarrays.

The exploratory work in Section 4 was done on all 18 arrays. Results in Section 6 are shown for the U74A chip data (6 arrays). This set of data can be thought of as 3 repeat measurements of 2 conditions (wildtype and knockout) for each of the 12488 genes represented on the U74A chip. One gene with three identical measurements under one condition is excluded from our analysis, as our model estimates variances on the log scale, thus the dataset used in Section 6 contains 12487 genes.

The expression data we use are the output of the Affymetrix MAS 5.0 software (Hubbell, Liu, & Rui, 2002). Since the distribution of these values is strongly skewed, we model them on the log scale. There has been much work done on alternative data transformations, mostly variants on the log transformation (e.g. Huber et al., 2002). We use the log transformation for simplicity, but other transformations could be considered. The log transformation has the advantage that the parameter δ_g defined in the next section can be interpreted as the log fold difference.

3. Bayesian Hierarchical Model for Differential Expression

3.1 *The Model*

We start with an ANOVA model, as suggested by Kerr et al. (2000), where log gene expression y_{gsr} for gene g, condition $s = 1, 2$ and replicate r is modelled with additive

effects for gene and array:

$$\begin{aligned} y_{g1r} &\sim N(\alpha_g - \frac{1}{2}\delta_g + \beta_{g1r}, \sigma^2_{g1}) \\ y_{g2r} &\sim N(\alpha_g + \frac{1}{2}\delta_g + \beta_{g2r}, \sigma^2_{g2}) \end{aligned} \tag{1}$$

where α_g is the gene effect or overall expression level, β_{gsr} is the array effect (this normalizes the arrays) that depends on g through α_g (see below), and σ^2_{gs} is the gene-specific variance for condition s. The differential effect between conditions is δ_g.

The array effect is a function of the expression level, $\beta_{gsr} = f_{sr}(\alpha_g)$. For flexibility, we choose f_{sr} to be a quadratic spline:

$$\beta_{gsr} = b^{(0)}_{sr0} + b^{(1)}_{sr0}(\alpha_g - a_0) + b^{(2)}_{sr0}(\alpha_g - a_0)^2 + \sum_{k=1}^{K} b^{(2)}_{srk}(\alpha_g - a_{srk})^2 I[\alpha_g \geq a_{srk}] \tag{2}$$

where the polynomial coefficients $b^{(p)}_{srk}$ and knots a_{srk} are unknown parameters which are estimated as part of the model. The number of knots K is fixed (but sensitivity to different choices of K can be investigated as part of model checking).

The equations (1) and (2) define the first level of the hierarchical model. At the second level, information is shared between genes to stabilise the variances. The variances are modelled as exchangeable within each condition, *i.e.* the variances are assumed to come from a common distribution, chosen here to be log Normal:

$$\sigma^2_{gs} \sim \text{logNorm}(\mu_s, \eta^2_s). \tag{3}$$

The third level of the model specifies prior distributions for all the unknown parameters, which are intended to be non-informative. The gene-effects α_g and knots a_{srk} are Uniformly distributed on (a_0, a_{K+1}) where a_0 and a_{K+1} are fixed lower and upper limits (chosen to be wide enough not to affect the results). Polynomial coefficients $b^{(p)}_{srk}$ have independent $N(0, 10^2)$ priors and the hyperparameters μ_s and η_s^{-2} have $N(0, 10^3)$

and Gamma$(10^{-2}, 10^{-2})$ priors respectively. In this work, the differential effects δ_g are given independent $N(0, 10^4)$ priors.

The model in equation (1) is unidentifiable as it stands, since constants can be transferred between the gene, array and differential effects for each gene, so constraints must be imposed on the parameters. Our choice is to normalize in a non-linear way *within each condition* by setting $\bar{\beta}_{gs.} = 0 \; \forall \, g, s$, where the dot indicates that we are taking an average over the index r. This fully identifies the model. Normalizing using all genes within condition seems reasonable, as we do not expect systematic differences between genes on replicate arrays.

3.2 *Confounding between Array Effects and Differential Effects*

In our model we normalize within each condition. Alternatively, normalization across replicates and conditions is achieved by setting $\bar{\beta}_{g..} = 0 \; \forall g$ (average taken over s and r). This set of constraints is a subset of the within condition constraints $\bar{\beta}_{gs.} = 0 \; \forall \, g, s$; it is insufficient to identify the model fully. Normalization across replicates and condition is usually completed by setting $\delta_g = 0$ for some control genes (e.g. Kepler et al., 2002). It is important to realise that normalizing across replicates and conditions in a pre-processing step for all genes and not solely for the controls implicitly assumes *all* $\delta_g = 0$ which means the δ_g estimated in a subsequent model are mis-specified.

3.3 *Implementation*

We estimate the model in a fully Bayesian way, using the WinBUGS software (Spiegelhalter et al., 1999) to perform Monte Carlo Markov Chain (MCMC) simulations of the posterior distribution. The WinBUGS code is given in the supplementary material. The constraints $\bar{\beta}_{gs.} = 0$ are imposed by defining quantities $z_{gs} \sim N(\bar{\beta}_{gs.}, 10^{-6})$ and giving the program "dummy data" z_{gs} which are equal to zero for all g and s. Since the variance of the z_{gs} is so small (10^{-6}) this construct forces the $\bar{\beta}_{gs.}$ to be very close

to zero (e.g. for the U74A wildtype data the maximum $\bar{\beta}_{gs.} = 4 \times 10^{-4}$).

We allow 10,000 iterations for the sampler to converge and another 10,000 for sampling the joint posterior. Convergence is checked visually and by using several starting points. The 10,000 posterior samples are thinned to 1000 for estimating the posterior distribution of quantities of interest. For each of the parameters $\alpha_g, \delta_g, \beta_{gsr}, \sigma^2_{gs}$ the Monte Carlo errors (as estimated using the batch method in WinBUGS) are around 3% of posterior standard deviation for most genes, less than 5% for 99% of the genes, and less than 7% for all genes. In Sections 4.2, 5 and 6 we use predictive p-values and posterior probabilities, which are estimated by means of indicator functions. For all genes, the MC errors for these quantities are less than 5%, except for the genes with posterior probability very close to zero, which have larger errors. For our purposes we are not interested in genes with very low probability, however.

The WinBUGS software is particularly user friendly, but due to its general purpose nature can be slow for treating the very large data sets common in genomics applications. For the full model processing 74,922 data points, 1000 iterations take 2 hours 50 minutes on a dual processor 2.4 GHz machine running version 1.4 of WinBUGS under Windows. In the future, we will develop faster, purpose built code.

3.4 *Rules for Selecting Genes*

As we obtain the joint posterior distribution of all parameters, a rich variety of inference criteria can be used. The main biological quantities of interest are the differential effects δ_g and the overall level of expression α_g. In the spirit of a volcano plot, we propose to use the joint posterior on both these parameters to pick genes:

$$p_g \equiv \mathbb{P}(|\delta_g| > \delta_{cut} \text{ and } \alpha_g > \alpha_{cut}| \text{ data }) \tag{4}$$

Genes are selected if $p_g \geq p_{cut}$. This rule is used in Section 6 to analyse our data set. In Section 5, on simulated data, we use a simpler decision rule based only on the

differential effects: if we define $p_g^{'} \equiv \mathbb{P}(\delta_g > \delta_{cut})$, we declare genes positive if $p_g^{'} \geq p_{cut}$. These posterior probabilities are easily estimated by counting the proportion of MCMC samples for which the chosen criteria are true.

The choice of δ_{cut} and α_{cut} correspond to statements of biological interest. The choice of p_{cut} is determined by the evaluation of the False Discovery Rate (and/or False Non-discovery Rate), see Section 6.

4. Model Checking

In this section, we present analyses carried out on biological replicate data. These consist of 3 repeat measurements for each gene, obtained from the same strain of mouse under the same experimental condition. The estimated model is thus as given in Section 3 but restricted to one condition ($s = 1$), with $\delta_g = 0$. We carried out the exploratory work in this simpler setting to focus on array effects and not on differential effects. Results are displayed here for one particular set of 3 arrays, the wildtype data on chip U74A, but we reached similar conclusions for the other replicate data sets in this study.

4.1 *Exploratory Analysis of Array Effects*

We explore the form of the array effects by splitting the data into equal sized groups based on average expression level across the 3 arrays. In order to have a sufficient number of genes per group, we split the genes into 6 groups. For each group j, we fit a model with constant array effect β_{j1r} and variances $\sigma^2_{j1}, j \in \{1, ..., 6\}$,

$$y_{g1r} \sim N(\alpha_g + \beta_{j1r}, \sigma^2_{j1}). \tag{5}$$

The array effects are subject to the constraints $\bar{\beta}_{j1.} = 0$ to ensure identifiability.

The posterior mean of the array effects for the 6 groups, $\beta_{j(g)1r}$, are shown as horizontal straight lines in the left hand panel of Figure 1, plotted against the expression levels α_g. The array effects show a clear non-linear trend with expression level, not

compatible with random fluctuations around a constant effect. We have found similar patterns in all sets of three replicated arrays studied. Similarly, Schadt et al. (2000) looked at a large number of Affymetrix data sets and found many array pairs needed a non-linear normalization effect. FIGURE 1 ABOUT HERE

The array effects shown in Figure 1 cross over each other. This means that, on the same array, if genes with relatively low expression are normalized upwards, genes with relatively high expression are normalized downwards, and *vice versa*. A constant array effect would normalize all genes either upwards or downwards together, and so would give qualitatively different results.

We include expression level dependent normalization in our model using equations (1) and (2). For the U74A wildtype data set, we have fitted (2) using three different functions for $f_{sr}(\alpha_g)$, a single quadratic, a quadratic spline with one knot and a single cubic:

$$\beta_{gsr} = b^{(0)}_{sr0} + b^{(1)}_{sr0}(\alpha_g - a_0) + b^{(2)}_{sr0}(\alpha_g - a_0)^2 + b^{(3)}_{sr0}(\alpha_g - a_0)^3. \tag{6}$$

All give similar looking results and are superimposed on the cubic curves which are shown in Figure 1 (middle panel). In order to inform our choice of $f(.)$ in (2), we used the deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002), designed to assess fit versus complexity of different models. For this data, the single cubic has the lowest DIC value of the three. In all our subsequent results for this data set, we therefore use the model with cubic array effects, though we recommend starting with the spline form in equation (2) as a more general non-linear function.

4.2 *Predictive Checks on Prior for Gene Variances*

The Bayesian setting has the desirable feature of allowing us to criticise various aspects of the model from a predictive point of view. Here we choose to investigate the prior distribution of the gene variances as it is an important aspect of the model.

We consider 4 possibilities: an "equal variance model" where $\sigma_{gs}^2 \equiv \sigma_s^2$ and $\sigma_s^2 \sim \text{logNorm}(0, 10^4)$, the exchangeable log Normal variance model presented in Section 3, an exchangeable variance model with a 1-parameter Gamma distribution on the variances: $\sigma_{gs}^{-2} \sim \text{Gam}(2, \beta_s^{prior}); \beta_s^{prior} \sim \text{Gam}(10^{-2}, 10^{-2})$ and an exchangeable model with a 2-parameter Gamma: $\sigma_{gs}^{-2} \sim \text{Gam}(\alpha_s^{prior}, \beta_s^{prior}); \alpha_s^{prior} \sim \text{Gam}(10^{-2}, 10^{-2}); \beta_s^{prior} \sim \text{Gam}(10^{-2}, 10^{-2})$. A plot of the smoothed variances (using the exchangeable log Normal prior) showing the amount of shrinkage is given in the supplementary material.

In order to assess the fit of the variance part of each model, we compare, by way of a Bayesian "p-value", the observed sample variance for each gene (calculated on the 3 replicates) to that evaluated using predictive values for each gene under the model. The distribution of p-values for all 12487 genes is then used to assess the overall fit of the variance part of the model.

For the equal variance model, we use posterior predictive p-values (Gelman, Meng, & Stern, 1996), taking the sample variance as the discrepancy measure. For each gene, new data for the r replicates is predicted from the model, $y_{gsr}^{(\text{pred})} \sim N(\alpha_g + f_{sr}(\alpha_g), \sigma_s^2)$, and the predicted sample variance $S_{gs}^{2(\text{pred})}$ calculated. For the exchangeable models, we use *mixed predictive p-values* (Gelman et al., 1996; Marshall & Spiegelhalter, 2003). These differ from posterior predictive p-values in that a *new variance parameter* $\sigma_{gs}^{2(\text{pred})}$ is first predicted for each gene, and secondly new data is predicted: $y_{gsr}^{(\text{pred})} \sim N(\alpha_g + f_{sr}(\alpha_g), \sigma_{gs}^{2(\text{pred})})$. In each case the distribution of the predicted sample variance is compared to the observed sample variance, $S_{gs}^{2(\text{obs})}$, by computing a p-value $\mathbb{P}(S_{gs}^{2(\text{pred})} > S_{gs}^{2(\text{obs})})$. The supplementary material includes a directed acyclic graph for both exchangeable and equal variance models that highlights the difference between mixed and posterior predictive p-values.

Under the null hypothesis of the model being "true", the distribution of the p-

values is almost uniform (Bayarri & Berger, 2000). Since the data from a particular gene influences its prediction, the p-values are shrunk towards 0.5, *i.e.* they are conservative. Mixed predictive checks have been shown to be much less conservative than posterior predictive checks (Marshall & Spiegelhalter, 2003). FIGURE 2 ABOUT HERE

Figure 2 shows histograms of the p-values for the 4 models for the wildtype data ($s = 1$) on chip U74A. The equal variance model has a large number of small and large p-values, indicating that this model does not allow enough variability for the data. If these p-values were corrected for their conservativeness, they would show even stronger evidence against the equal variance model. The exchangeable variance models fit the data much better. Having only 1 parameter in the prior for the variances still does not allow enough variability, but with 2 parameters both Gamma and log Normal priors seem adequate. Furthermore, there are no extremely small p-values (which would indicate individual outlying genes). We are therefore confident that the exchangeable variance model we propose in Section 3 is appropriate for this data set.

5. Comparison of Integrated and Non-integrated Analyses

The problem of estimating functions of the gene effects simultaneously with the gene effects themselves is similar to a measurement error problem, where covariates in a regression are not measured accurately but have some unknown variability. It is well known that in such problems ignoring this variability leads to a bias in estimates of regression coefficients (e.g. Carroll et al., 1995). Therefore, we would expect to obtain biased estimates of the array effects if they are estimated in a pre-processing step, which in turn will lead to worse estimates of the differential effects δ_g. To illustrate this, we compare the results from the full model with those found by pre-normalizing the data using array effects from local regression smoothing (loess).

We simulate a microarray data set with 1000 genes and 3 repeat arrays under 2

conditions. The gene effects α_g range uniformly between 0 and 10, and the array effects are functions of the gene effects:

$$\beta_{gsr} = 5 \times 10^{-3} \times (x_{gsr} - \bar{x}_{gs.}) \tag{7}$$

where $x_{g11}, x_{g12}, x_{g13}, x_{g21}, x_{g22}, x_{g23}$ are $(\alpha_g - 5)^3, -5(0.3\alpha_g - 4)^3, 2(\alpha_g - 5.5)^3, -(\alpha_g - 4)^3, 5(0.05\alpha_g - 7)^3 + 10^3, 2(\alpha_g - 5.5)^3$ respectively. The gene variances are simulated from the model we fit (equation 3), with $\mu_1 = -1.8, \mu_2 = -2.2, \eta_s^2 = 1$ for $s = 1, 2$, giving a similar range of variances to those we observed in the real data. The differential effects δ_g are zero for 900 genes, $N(\log(3), 0.1^2)$ for 50 genes and $N(-\log(3), 0.1^2)$ for the other 50 with the differentially expressed genes uniformly spread over the range of α_g. The right hand panel of Figure 1 shows the array effects and data points for one set of three simulated arrays. The middle panel shows the same plot for the mouse data set used in Section 4.1, using posterior means for the array effects.

To calculate loess estimates of array effects, we use the R function "loess" with $y_{gsr} - \bar{y}_{gs.}$ as a function of $\bar{y}_{g..}$. The array effects $\hat{\beta}_{gsr}^{loess}$ are the values of the loess curve for sample s and array r predicted at $\bar{y}_{g..}$.

We have performed 5 simulations of the above set-up. For each of the 6 arrays we obtain the mean square error (MSE) $\sum_g (\hat{\beta}_{gsr}^{est} - \beta_{gsr}^{true})^2$ for both the loess array effects and for those found in the full model. In 28 out of the 30 array-simulations, the loess MSE is bigger than the full-model MSE. The ratios of loess MSE to full-model MSE for the 6 arrays are 1.5, 1.3, 1.2, 1.2, 1.4, 1.3 (averages for the 5 simulations). As expected, the integrated model obtains estimates of the array effects closer to the true values than those found by loess smoothing.

The quantities of real interest are the differential effects δ_g. In order to assess the effect on these of using a pre-processing step, we fit a model where we pre-normalize the data by subtracting point estimates of the array effects: $y'_{gsr} \equiv y_{gsr} - \hat{\beta}_{gsr}$, and

run our model as in Section 3 without array effects. We have done this for 2 different cases: the first with $\hat{\beta}_{gsr} = \hat{\beta}^{loess}_{gsr}$ and the second with $\hat{\beta}_{gsr} = E(\beta_{gsr}|$ data $)$ from the full model.

The δ_g found by using the two pre-normalized models are very close. The differences between the loess procedure and the full model do not come from the way the array effects are computed but from using a two step procedure as opposed to the full integrated model. Here we show the results from the pre-normalized model using $\hat{\beta}_{gsr} = \hat{\beta}^{loess}_{gsr}$ since this is what is usually done in practice. FIGURE 3 ABOUT HERE

We use the simpler rule proposed in Section 3.4 to pick a list of genes: if $\mathbb{P}(\delta_g > \delta_{cut}|$ data $) \geq p_{cut}$ then gene g is selected. Based on this, we can calculate the number of false positives and negatives. Figure 3 shows the observed false discovery rate (FDR: the number of false positives divided by the number of genes declared positive), and the analogous false non-discovery rate (FNR), for both the full and pre-normalized models, as a function of the cut-off probability p_{cut}, for a choice of $\delta_{cut} = \log(3)$. Graphs shown are curves averaged over the 5 simulations. The full and pre-normalized models have very similar FNR (only a small fraction of genes are differentially expressed), while the FDR is consistently lower for the integrated model than for the pre-normalized model.

The difference shown here between the full and pre-normalized models is fairly small, as the simulation is inspired by the data we use in this paper, which has small array effects. In general, the larger the magnitude of array effects, the larger the difference between the pre-normalized and integrated models.

6. Application to Mouse Data set

6.1 *Discussion of Decision Rules*

So far we have concentrated on producing a flexible model, well-grounded in the data. Here we discuss the output of the model applied to the data set discussed in

Section 2 and illustrate our chosen decision rule. The results shown use a subset of the arrays from the full study (wildtype and knock-out mice on U74A chips) in order to illustrate the generic methodology we propose.

Figure 4 shows point estimates of log fold change $E(\delta_g|$ data $)$ versus overall expression $E(\alpha_g|$ data $)$. It can be seen that genes with low overall expression have a greater range of fold change than those with higher expression. For this reason, it is appropriate to use the rule proposed in Section 3.4 (Equation 4) where $p_g \equiv \mathbb{P}(|\delta_g| > \delta_{cut}$ and $\alpha_g > \alpha_{cut}|$ data $)$ and genes are selected if $p_g \geq p_{cut}$. We use $\alpha_{cut} = 4$ as this corresponds to a background level of 54 on the original scale of the data, and $\delta_{cut} = \ln(2)$ as this is a (conservative) value often used in microarray analysis. FIGURE 4 ABOUT HERE

Genes with probability p_g greater than 0.5 and 0.8 are highlighted in Figure 4. There are 280 genes with $p_g \geq 0.5$, 140 over-expressed and 140 under-expressed, and 46 with $p_g \geq 0.8$, 30 over-expressed and 16 under-expressed. The gene Cd36, which is expected to be under-expressed in this data, is also marked in Figure 4. It stands out clearly from the bulk of the genes. The value of p_g for Cd36 is 0.49.

To determine an appropriate probability cut-off p_{cut}, we can calculate an estimate of the false discovery rate (FDR) based on the decision rule above. Here δ_{cut} and α_{cut} determine H_0 (so must be chosen beforehand). This type of null distribution is similar to those used in bioequivalence studies, where a cut-off is chosen to indicate a biologically relevant difference (e.g. Bickel, 2004). It is a composite hypothesis rather than the more usual point hypothesis. We estimate the FDR as in Newton et al. (2004), by $\frac{1}{|S(p_{cut})|}\sum_{g \in S(p_{cut})}(1 - p_g)$, where $S(p_{cut})$ is the group of genes with $p_g \geq p_{cut}$ and $|S(p_{cut})|$ denotes its cardinality. The analogous false non-discovery rate (FNR) is estimated in a similar way. An FDR of 10% corresponds to a probability cut-off of 0.83. The number of genes obtained using this cut-off is 33 and the estimated FNR in this

case is 7%. These values depend on the chosen δ_{cut} and α_{cut}, of course, which determine the null hypothesis under consideration.

The quantities p_g are a convenient way to combine biological significance with statistical significance. Statistical significance is ensured in the usual way through use of the posterior probability. The p_g increase with the standardised log fold difference $t_g \equiv E(\frac{\delta_g}{\sqrt{(\sigma_{g1}^2+\sigma_{g2}^2)/3}} \mid \text{data})$, which is analogous to a t-statistic (though using shrunken variance estimates), so requiring high p_g is partly similar to performing a t-test (see Figure 5). The biological significance is determined by the cut-off on the fold difference δ_{cut}. If genes were selected using only t_g, many genes with low fold difference would be declared as statistically significant. For example, there are 118 genes above background ($E(\alpha_g|$ data $) > 4$) with $|t_g| \geq 2.78$ (corresponding to a 95% confidence interval for a t-test with 4 degrees of freedom). These genes are shown as green triangles in Figure 5). Of these, a large proportion (32/118) have low fold difference $|E(\delta_g|$ data $)| < \log(2)$. This can be seen in the left hand panel of Figure 5: the vertical lines mark $|E(\delta_g|$ data $)| = \log(2)$. FIGURE 5 ABOUT HERE

6.2 *Analysis of Gene Ontology terms*

In order to illustrate the biological significance of the genes found by the above rule (Equation 4), we compare the Gene Ontology (GO) annotations of these genes (the "query group") with the annotations of a reference group made up of the least differentially expressed genes. The query group is defined by $p_g \geq 0.5$, giving 280 genes. The reference group is all genes with $p_g \leq 0.2$ (11171 genes). When these lists are reduced to genes with both Gene Symbol and GO annotation, we retain 95 query versus 4931 reference genes.

The Gene Ontology can be represented as a directed acyclic graph relating biological terms of different degrees of specificity, with directed links from less specific to

more specific terms. Each term can have several parents (broader related terms) and children (more specific related terms). Annotation of a gene to any given term A implies automatic annotation to all ancestors of A (the set of broader terms related to A by directed paths). For each GO term we perform a Fisher's exact test using the FatiGO website (Al-Shahrour et al., 2004), to compare the proportions of genes in the query and reference groups annotated to that term. Our definition of a GO term being "significant" is a p-value less than 0.05 and (for terms over-represented in the query group) annotations for 3 or more genes on the query list. For this interpretative step we do not use multiple testing corrections as there is a large amount of dependency between tests, due both to genes being annotated in several categories and to the graph structure of the Gene Ontology. Since we use a conservative criterion in our definition of p_g this seems reasonable. TABLE 1 ABOUT HERE

The left-hand column of Table 1 shows the most specific GO terms that are found significant according to this criterion. The right-hand column shows the ancestors (broader terms) in the ontology of these terms that are found significant. Also shown are the observed and expected numbers of genes in the query group annotated to each term. Expected numbers are calculated by multiplying the percentage of annotations in the reference group with the number of genes in the query group.

It can be seen that many of the ancestors of "inflammatory response" are significant (in fact only one ancestor of this term is not significant). This result is of interest in view of the emerging observations of the importance of inflammation-related genes in adipocyte biology (Clément et al., 2004). A graph showing the relations between "inflammatory response" and all its ancestors is included in the supplementary material. In addition, the significant over-representation of lipid catabolism and under-representation of protein transport are consistent with the known biological functions

of Cd36 in metabolism (Aitman et al., 1999).

7. Discussion

We have presented a unified Bayesian hierarchical model for differential gene expression incorporating both non-linear array effects and exchangeable gene variances, and have justified our functional choices by exploratory analysis. We have drawn attention to the issue of confounding between array and differential effects, and shown how analysing the data in a two step process can lead to a greater number of false positives. Bayesian joint estimation of differential effects and array effects has also been carried out in a recent paper by Bhattacharjee et al. (2004).

The Bayesian formulation enables us to obtain a much richer output from our model than most current analyses. Our method for choosing lists of genes provides a straightforward way to rank genes using multiple criteria, and a suitable cut-off on the list can be chosen based on the estimated false discovery rate.

Expert opinion is used to set suitable cut-offs on the log fold change and (optionally) overall expression level. The "null hypothesis" formulated here is composite. If a point null hypothesis is required, a mixture model can be used to classify genes and estimate the FDR (Lönnstedt & Speed, 2003; Newton et al., 2004), and we intend to explore this in further work. Another interesting fully Bayesian model has been developed by Ishwaran & Rao (2003). Their approach is complementary to ours: we use a flat prior on the differential effects but shrink the gene variances while they shrink the differential effects, using a "spike and slab" variable selection approach, but use raw variances that in their main model are not gene-specific. Gene lists are then formed using t-statistics.

One interesting alternative would be to look at distributions of gene ranks, for instance based on fold change. Simply ranking genes based on point estimates of fold change is not enough, as some highly ranked genes will have high variability and some

will have low variability. By looking at the distributions of ranks (easily obtained in the Bayesian approach) one can find the highly ranked genes with the lowest variability.

The probability statements made so far have been on a gene by gene basis, but in this framework, we can also make joint statements about the genes that circumvent in some way the problem of multiple testing. For example, for a given list of genes derived from independent biological information, we can make joint statements such as: "the probability (conditional on the data) that 80% of genes $\{g_1, \ldots, g_n\}$ have $\delta_g > \log(2)$ is at least 0.80", which can be readily interpreted by biologists. Statements about the joint distribution of the ranks could also be made.

We have implemented our model using the WinBUGS software for ease and ubiquity of use and the code is available (see supplementary material). For a much larger number of arrays, it will become unfeasible to use WinBUGS and we are in the process of developing code that would cut the computing time by a substantial fraction.

The ideas in this paper apply to data from cDNA chips as well as Affymetrix, with extra terms included to account for within slide normalization. In future work, we aim to include in the hierarchical specification another layer consisting of the Bayesian signal extraction model for Affymetrix PM's and MM's as proposed in Hein et al. (2005).

Acknowledgements

This work is part of an ongoing BBSRC collaboration. We would like to thank our colleagues Graeme Ambler, Helen Causton, Peter Green and Anne-Mette Hein for their help in shaping this paper. We thank Dr M Febbraio for the gift of Cd36 knockout mouse tissues and Dave Lunn for help with WinBUGS. We also thank the associate editor and anonymous referees for helpful comments which led to many improvements in the paper. This work was supported by BBSRC "Exploiting Genomics" grant 28EGM16093.

References

Aitman, T. J., Glazier, A., Wallace, C., Cooper, L. A., Norsworthy, P. J., Wahid, F. N., Al-Majali, K. M., Trembling, P. M., Mann, C. J., Shoulders, C. C., Graf, D., St. Lezin, E., Kurtz, T. W., Kren, V., Pravenec, M., Ibrahimi, A., Abumrad, N. A., Stanton, L. W., and Scott, J. (1999). Identification of Cd36 (Fat) as an insulin-resistance gene causing defective fatty acid and glucose metabolism in hypertensive rats. *Nature Genetics* **21**, 76–83.

Al-Shahrour, F., Daz-Uriarte, R., and Dopazo, J. (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* **20**, 578–580.

Bayarri, M. J. and Berger, J. (2000). P-values for composite null models. *Journal of the American Statistical Association* **95**, 1127–1142.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Series B* **57**, 289–300.

Bhattacharjee, M., Pritchard, C. C., Nelson, P. S., and Arjas, E. (2004). Bayesian integrated functional analysis of microarray data. *Bioinformatics* **20**, 2943–2953.

Bickel, D. R. (2004). Degrees of differential gene expression: detecting biologically significant expression differences and estimating their magnitudes. *Bioinformatics* **20**, 682–688.

Carroll, R. J., Ruppert, D., and Stefanski, L. A. 1995, Measurement Error in Nonlinear Models (Chapman and Hall/CRC).

Chen, Y., Dougherty, E. R., and Bittner, M. L. (1997). Ratio-Based Decisions and the Quantitative Analysis of cDNA Microarray Images. *Journal of Biomedical Optics* **2**, 364–374.

Clément, K., Vigueire, N., Poitou, C., Carette, C., Pelloux, V., Cyrile, A., Sicard, A., Rome, S., Benis, A., Zucker, J.-D., Vidal, H., Laville, M., Barsh, G. S., Basdevant, A.,

Stich, V., Cancello, R., and Langin, D. (2004). Weight loss regulates inflammation-related genes in white adipose tissue of obese subjects. *FASEB J.* **18**, 1657–1669.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association* **96**, 1151–1160.

Febbraio, M., Abumrad, N. A., Hajjar, D. P., Sharma, K., Cheng, W., Pearce, S. F. A., and Silverstein, R. L. (1999). A null mutation in murine Cd36 reveals an important role in fatty acid and lipoprotein metabolism. *Journal of Biological Chemistry* **274**, 19055–19062.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior Predictive Assessment of Model Fitness via Realized Discrepancies. *Statistica Sinica* **6**, 733–807.

Hein, A.-M. K., Richardson, S., Causton, H. C., Ambler, G. K., and Green, P. J. (2005). BGX: a fully Bayesian gene expression index for Affymetrix GeneChip data. *Biostatistics* (in press).

Hubbell, E., Liu, W., and Rui, M. (2002). Robust estimators for expression analysis. *Bioinformatics* **18**, 1585–1592.

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18**, S96–S104.

Ishwaran, H. and Rao, J. S. (2003). Detecting differentially expressed genes in microarrays using Bayesian model selection. *Journal of the American Statistical Association* **98**, 438–455.

Kepler, T. B., Crosby, L., and Morgan, K. T. (2002). Normalization and analysis of DNA microarray data by self-consistency and local regression. *Genome Biology* **3(7)**, 0037.1–0037.12.

Kerr, M. K., Martin, M., and Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology* **8**, 819–837.

Lönnstedt, I. and Speed, T. (2003). Replicated microarray data. *Statistica Sinica* **12**, 31–46.

Marshall, E. C. and Spiegelhalter, D. J. (2003). Approximate cross-validatory predictive checks in disease mapping models. *Statistics in Medicine* **22**, 1649–1660.

Newton, M., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture model. *Biostatistics* **5**, 155–176.

Nguyen, D. V., Arpat, A. B., Wang, N., and Carroll, R. J. (2002). DNA Microarray Experiments: Biological and Technological Aspects. *Biometrics* **58**, 701–717.

Schadt, E., Li, C., Su, C., and Wong, W. (2000). Analyzing High-Density Oligonucleotide Gene Expression Array Data. *Journal of Cellular Biochemistry* **80**, 192–202.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B* **64**, 1–34.

Spiegelhalter, D. J., Thomas, A., and Best, N. (1999). WinBUGS Version 1.2 User Manual. MRC Biostatistics Unit, software available at http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* **64**, 479–498.

Tusher, V., Tibshirani, R., and Gilbert, C. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences, USA* **98**, 5116–5121.

Workman, C., Jensen, L., Jarmer, H., Berka, R., L., G., Nielsen, H., Saxild, H., Nielsen, C., Brunak, S., and Knudsen, S. (2002). A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biology* **3(9)**, 0048.1–0048.16.

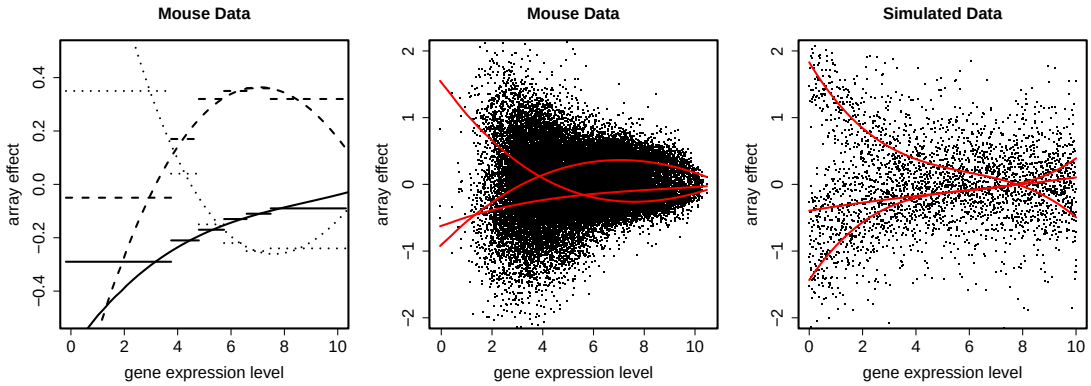

Figure 1. Left panel: Array effects as functions of gene expression level for the wildtype mouse fat hybridized to chip U74A. The straight lines are the posterior means of the array effects found by fitting model 5 on the relevant expression level range for the data divided into 6 groups. Dashed, dotted and dot-dash lines represent one array each. The smooth curves (dashed, dotted and dot-dash) are the posterior means of the cubic array effects from our model (equations 1, 3 and 6). Middle panel: Array effects as functions of gene expression level for the wildtype mouse fat hybridized to chip U74A. Lines are the posterior means of the cubic array effects from our model. Points are the data points with the overall gene expression level subtracted for each gene ($y_{gsr} - E(\alpha_g|$ data $)$). Right panel: Array effects β_{gsr} (lines) used for three of the simulated arrays in section 5 of the main paper. The points on the plot are the simulated data points with the overall gene expression level subtracted for each gene ($y_{gsr} - \alpha_g$). Note the similarity of spread and array effects between simulated and mouse data.

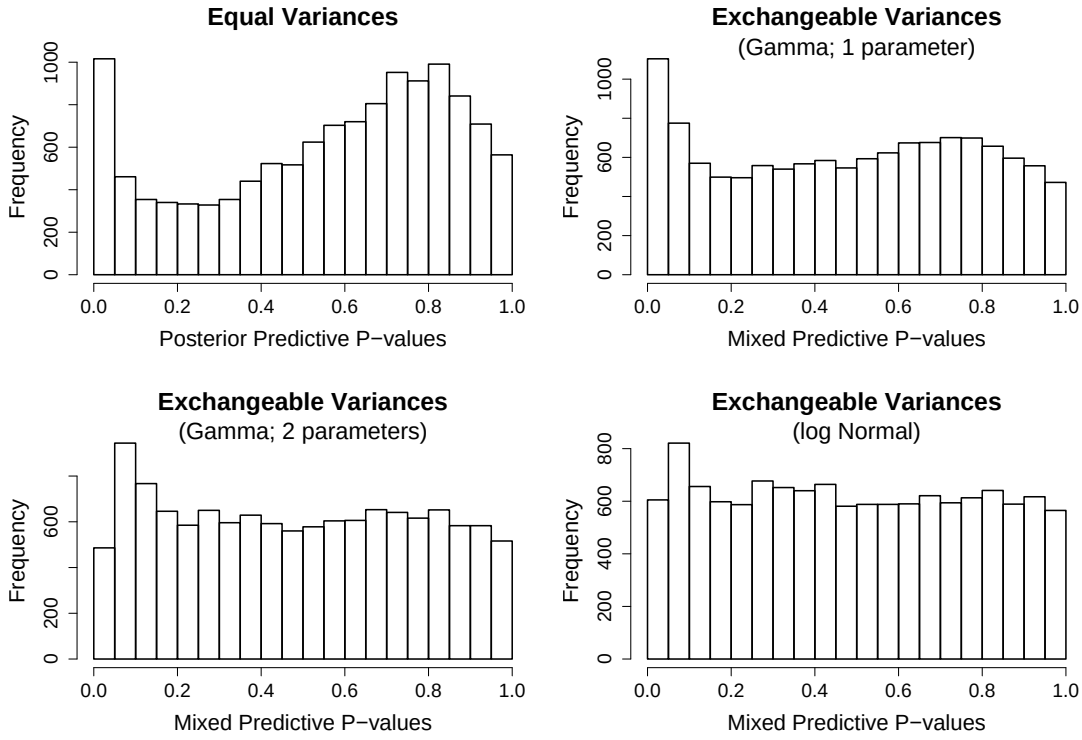

Figure 2. Mixed and posterior predictive p-values for the wildtype mouse U74A data (s=1) obtained from fitting the model with equal variances, exchangeable variances with a 1-parameter Gamma prior, exchangeable variances with a 2-parameter Gamma prior and exchangeable variances with a 2-parameter log Normal prior (as discussed in Section 3).

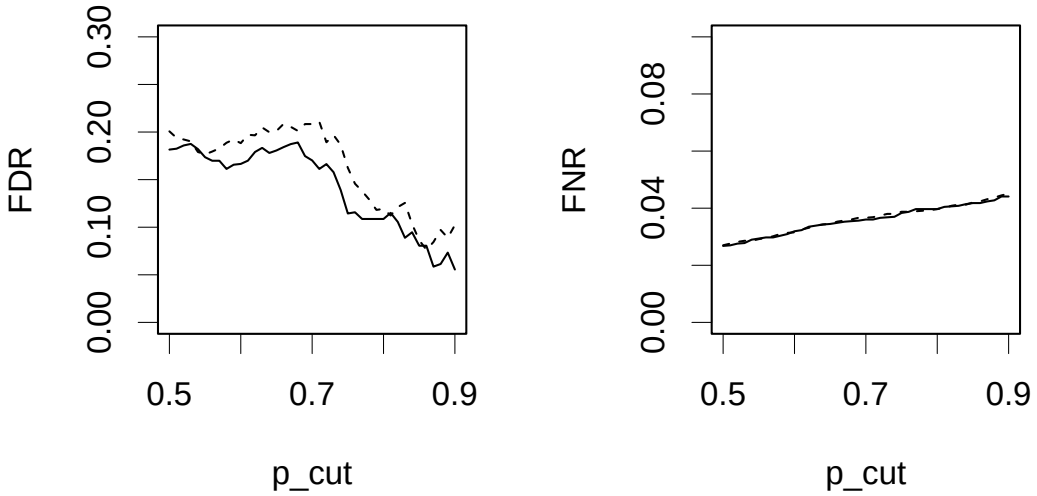

Figure 3. Observed false discovery rate (FDR) and false non-discovery rate (FNR) versus probability cut-off p_{cut} for the simulated data set described in Section 5. The solid line is for the full, integrated model and the dashed line is for the pre-normalized model. All curves are averages over 5 simulations.

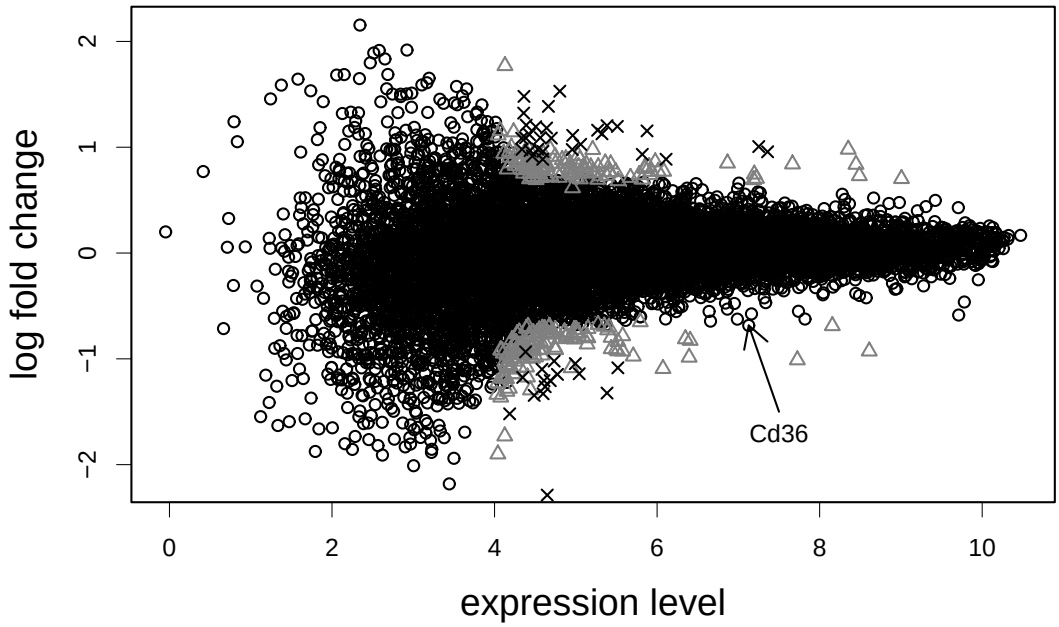

Figure 4. Posterior means of log fold change $E(\delta_g|$ data $)$ and gene expression level $E(\alpha_g|$ data $)$ for the U74A chip. Genes with $p_g \geq 0.5$ are plotted with triangles and those with $p_g \geq 0.8$ with crosses. The arrow indicates the gene Cd36, whose p_g is 0.49.

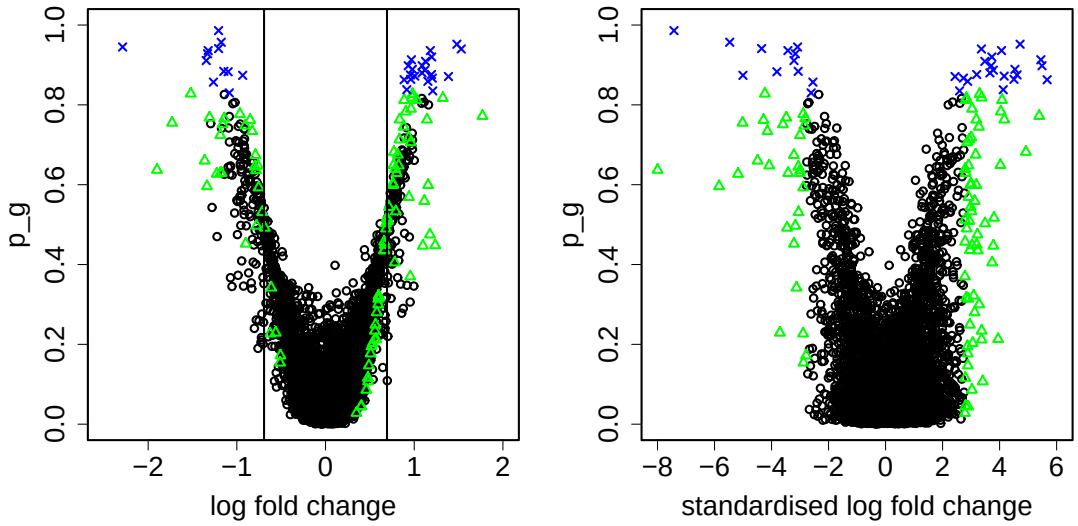

Figure 5. Posterior probabilities p_g versus log fold difference (left hand panel) and versus standardised log fold difference (right hand panel) for genes above background ($E(\alpha_g|$ data $) > 4$). Genes with $p_g \geq 0.83$ (corresponding to an estimated FDR of 10%) are shown as blue crosses, while genes with $|t_g| \geq 2.78$ (but not $p_g \geq 0.83$) are shown as green triangles. All except two of the genes with $p_g \geq 0.83$ have $|t_g| \geq 2.78$. The vertical lines show where $|E(\delta_g|$ data $)| = \log(2)$.

Most specific "significant" terms	Ancestor terms also "significant" (GO level > 3)
inflammatory response (O=4, E=1.2)	immune response (O=9, E=4.5) response to pest, pathogen or parasite (O=8, E=2.6) response to wounding (O=6, E=1.8) defense response (O=11, E=5.8) response to external stimulus (O=12, E=4.7) response to biotic stimulus (O=14, E=6.9) response to stress (O=12, E=5.9)
neuromuscular physiological process (O=5, E=1.3)	neurophysiological process (O=8, E=2.6) organismal movement (O=7, E=1.9)
transmission of nerve impulse (O=5, E=1.3)	neurophysiological process (O=8, E=2.6) cell-cell signaling (O=5, E=1.9)
DNA repair (O=4, E=1.3)	response to stress (O=13, E=5.9) nucleobase, nucleoside, nucleotide and nucleic acid metabolism (O=15, E=23.8)
lipid catabolism (O=3, E=0.6)	-
protein transport (O=0, E=5.0)	-
fertilization (O=3, E=0.2)	-

Table 1

The left column shows the most specific GO terms found significant. The right-hand column shows the ancestors (broader terms) in the ontology of these terms that are also found significant. Numbers in brackets indicate observed (O) and expected (E) numbers of genes in the query list (those most differentially expressed) annotated to the term. Expected numbers are calculated by multiplying the percentage of annotations in the reference group with the number of genes in the query group.