Supplemental Material for 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

1. WinBUGS Code

This is the WinBUGS (Spiegelhalter et al., 1999) code for the model for biological replicates, given by equations (1), (2) and (3) but restricted to one condition s, with the posterior predictive checks on the gene variances. The constraint $\bar{\beta}_{gs.} = 0$ is imposed in WinBUGS using dummy "data" consisting of zeros (see main paper for details). The number of genes is n, where $i = 1, ..., n$ labels the gene, $j = 1, ..., 3$ the array and $k = 1, ..., nk$ the polynomial in the array effect. The code shown here is for the model with piece-wise quadratic array effects; it can easily be modified to include a different functional form. We used version 1.4 of WinBUGS to run this model.

```
model;
{
   ###### 1st level: likelihood and posterior predictive p-values
   for( i in 1 : n ) {
```

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

```
    for( j in 1 : 3) {
       y[i, j] ~ dnorm(x[i, j], tau[i])
       ynew[i, j] ~ dnorm(x[i, j], taunew[i])
       x[i, j] <- alpha[i] + beta[i, j]
    }
    s2[i] <- pow(sd(y[i, ]), 2)
    s2new[i] <- pow(sd(ynew[i, ]), 2)
    pval[i] <- step(s2new[i] - s2[i])
}
###### 1st level: array effects as functions of gene effect
for( i in 1 : n ){
    for( j in 1 : 3 ){
         for( k in 1 : nk ){
             betadum[i,j,k] <- b2[j,k]*pow(alpha[i]-a[j,k],2)
                               *step(alpha[i]-a[j,k])
         }
         beta[i,j] <- b00[j] + b01[j]*(alpha[i]-a0) +
                      b02[j]*pow(alpha[i]-a0,2) + sum(betadum[i,j,])
    }
}
###### 2nd level: exchangeable gene variances
for( i in 1 : n ) {
    tau[i] <- 1.0/sig2[i]
    taunew[i] <- 1.0/sig2new[i]
    sig2[i] <- exp(lsig2[i])
    sig2new[i] <- exp(lsig2new[i])
    lsig2[i] ~ dnorm(mu,etaminus2)
    lsig2new[i] ~ dnorm(mu,etaminus2)
}
###### 3rd level: priors
for( i in 1 : n ){
    alpha[i] ~ dunif( a0, akplus1)
}
mu ~ dnorm( 0.0,1.0E-3)
etaminus2 ~ dgamma(0.01,0.01)
for( j in 1 : 3 ) {
   b00[j] ~ dnorm(0.0,1.0E-1)
   b01[j] ~ dnorm(0.0,1.0E-1)
   b02[j] ~ dnorm(0.0,1.0E-1)
   for( k in 1 : nk ){
     b2[j,k] ~ dnorm(0.0,1.0E-1)
     a[j,k] ~ dunif( a0, akplus1)
```

```
      }
   }
   ###### impose constraints using dummy data zrow
   for( i in 1 : n ) {
      scr[i] <- sum(beta[i, ])
      zrow[i] ~ dnorm( scr[i], 1.0E+6 )
   }
}
```

The gene expression data are called y[i,j] here, and the dummy data all equal to zero are zrow[i] (where i labels genes, j labels replicates).

2. Supplementary Figures

2.1 *Density of data before and after normalisation*

Workman et al. (2002) and Bolstad et al. (2003) show that their quantile normalisation methods bring the empirical distributions of data on the different arrays closer together. Figure 1 shows density plots of our data before and after normalisation (y_{gsr} and $y_{gsr} - \hat{\beta}_{gsr}$ respectively). Our method also brings the distributions into closer agreement.

2.2 *Sharing of information to estimate gene variances*

In our model $\sigma^2_{gs} \sim \text{logNorm}(\mu_s, \eta_s^{-2})$. The parameters μ_s and η_s^2 are estimated as part of the model, so information on variability is shared between genes. For a model with independent variances (a model with μ_s and η_s^2 fixed, not estimated in the model) each variance parameter is estimated using only 3 measurements. With exchangeable variances information from all genes (12487×3 measurements) is shared between all the genes.

Figure 2 shows the variances found by a model with independent variances and the one with exchangeable variances. In the exchangeable model the variances are shrunk towards their mean (on the log scale). The log-Normal prior smooths both low and high

variances. The amount of shrinkage is more than SAM (Tusher et al., 2001) performs on this data set.

2.3 *Directed acyclic graph for predictive model checks*

Figure 3 shows Directed Acyclic Graphs (Lauritzen, 1996) for both the exchangeable and equal variance models, to illustrate the difference in the calculation of the mixed and posterior predictive p-values. The index s for condition is suppressed here. Model parameters are shown as nodes joined by arrows indicating the conditional dependence properties. In Figure 3(a), for example, if α_g, β_{gr} and σ_g are known the distribution of y_{gr} is known. Conditional on $\{\alpha_g, \beta_{gr}, \sigma_g\}$, y_{gr} is independent of $\{\mu, \tau\}$. Rectangles represent data, circles represent stochastic parameters. Double arrows indicate that one quantity is a deterministic function of another (for example S_g^2 is a function of y_{gr}).

The posterior predictive checks only require new data points $\{y_{gr}^{(\text{pred})'}\}$ to be predicted. For the mixed predictive checks, new variance parameters $\{\sigma_g^{2\,(\text{pred})}\}$ are first predicted for each gene and then new data points $\{y_{gr}^{(\text{pred})}\}$ are predicted based on the predicted variance parameters.

References

Bolstad, B., Irizarry, R., Astrand, M., and Speed, T. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193.

Lauritzen, S. L. (1996). *Graphical Models.* Oxford University Press, UK.

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.

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.

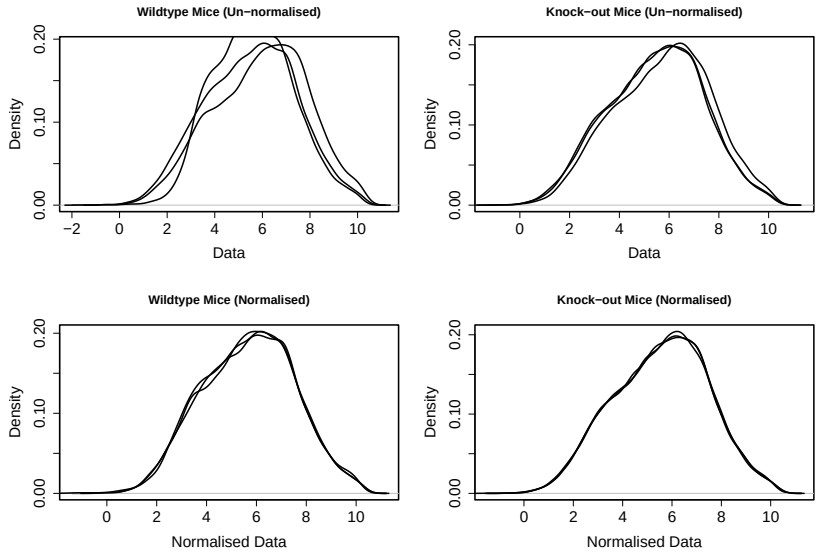


Figure 1. Data from 2 sets of 3 arrays, before and after normalisation. The non-linear normalisation (using array effects which are cubic functions of gene effect) brings the empirical distributions of data on the different arrays together.

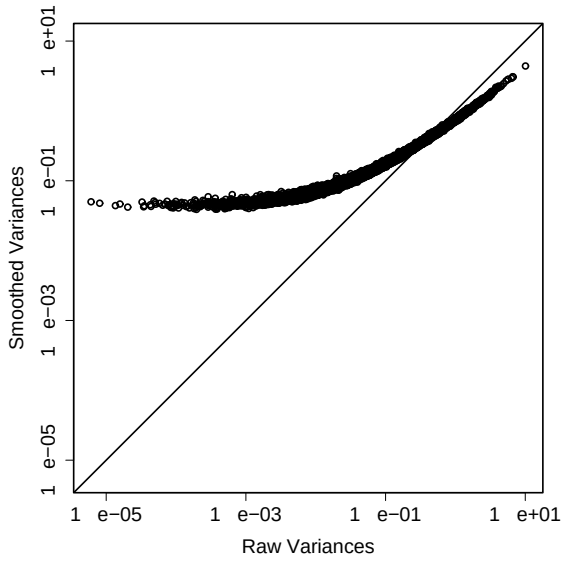


Figure 2. Raw (independent) and smoothed (exchangeable) variances for the wildtype mouse fat data. Both calculations used the same non-linear array effect. The smoothed (exchangeable) variances are shrunk towards the mean.

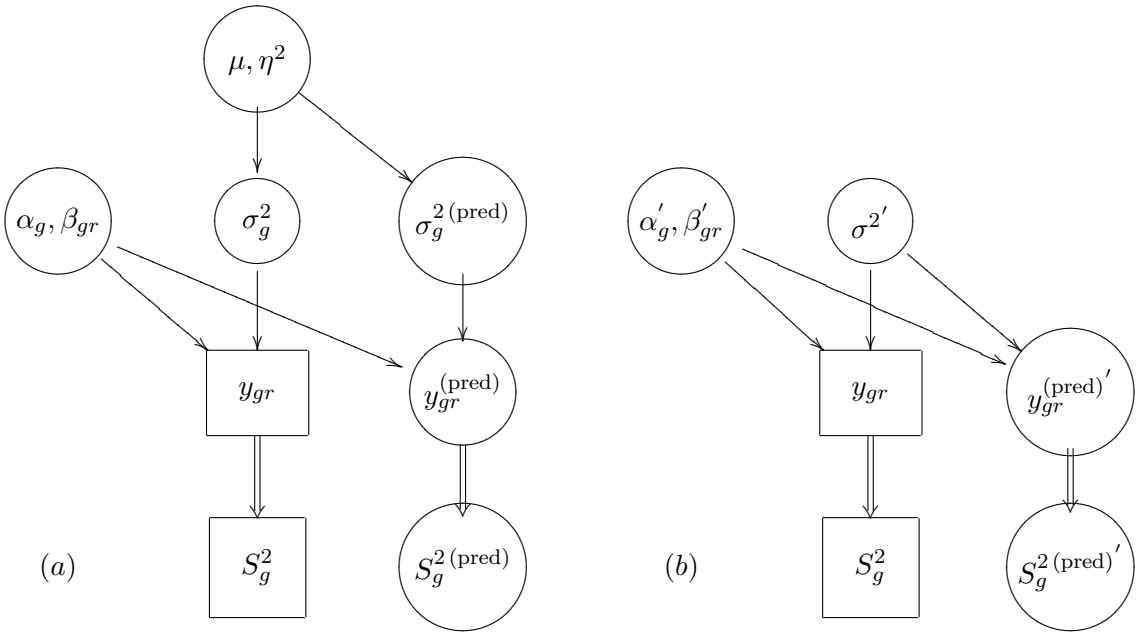
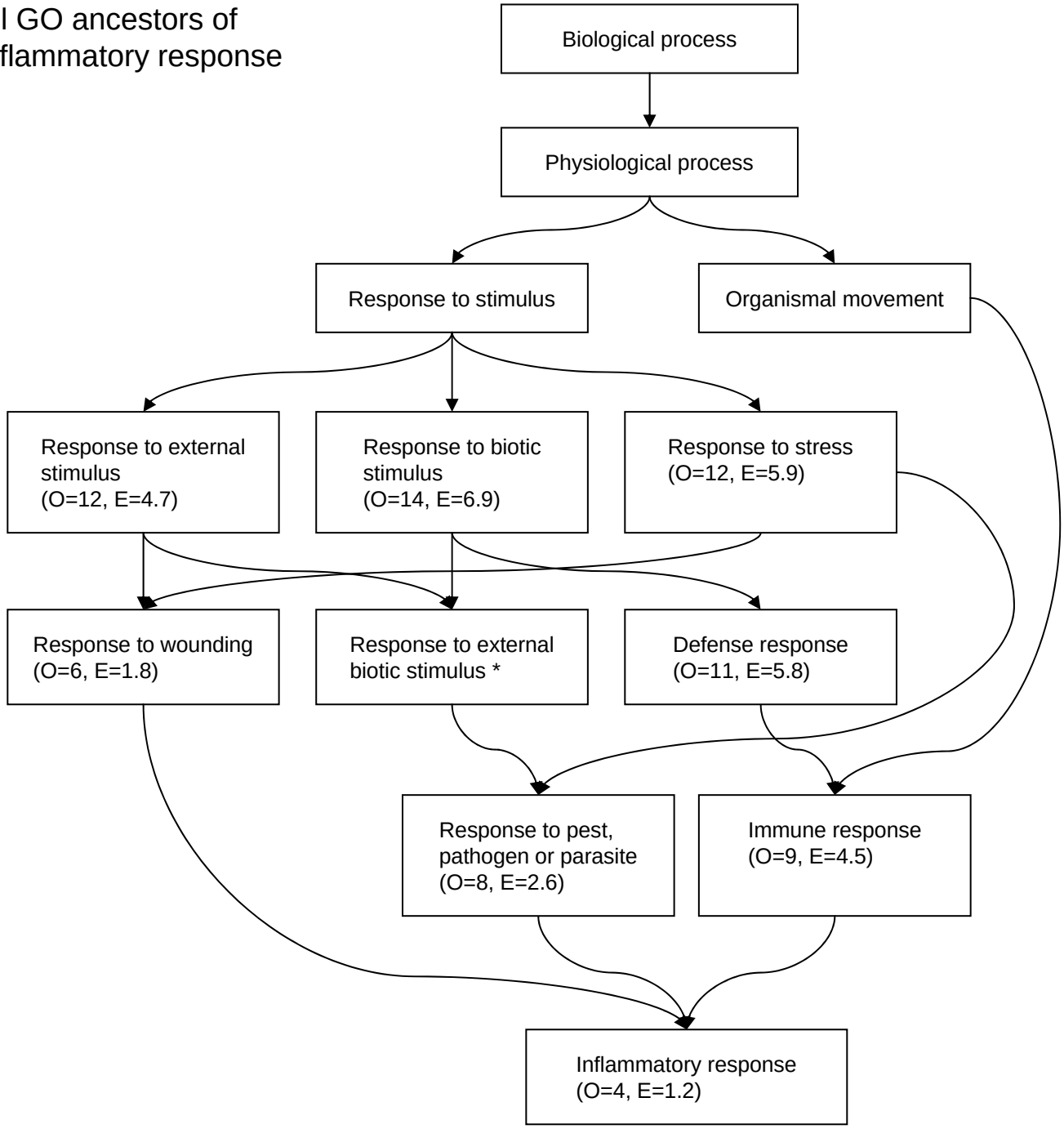


Figure 3. Directed acyclic graphs for the model for biological replicates, with additional parameters used in the calculation of mixed and posterior predictive p-values. (a) Exchangeable variances, (b) Equal variances.

All GO ancestors of Inflammatory response



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.

* This term was not accessed by the FatiGO website.

Relations between GO terms were found using the QuickGO website: http://www.ebi.ac.uk/ego/