Generalised empirical Bayesian methods for discovery of differential data in high-throughput biology: Supplemental Materials

Thomas J. Hardcastle *†

March 24, 2015

S1 Qualitative differences between subsets of biomolecular events

In high-throughput sequencing experiments, it is not uncommon to find a subset of biomolecular events that are qualitatively different from the remainder. In mRNA-Seq data, we expect a set of non-expressed genes to which only a small number of reads are assigned, for reasons such as sequencing error, misalignment or very low background levels of expression. Figure S9 shows the distribution of the log of the parameter associated with mean expression in RNA-seq data, assumed to be distributed negative binomially and equivalently across all samples. The tail of data to the left of the modal peak may be considered to represent non-expressed genes.

To distinguish between such qualitatively different events, we can construct additional models in `baySeq` v2. In the example above, we construct one model (M_{NDE}) for expressed but non-differentially expressed genes, and one model (M_{NE}) for non-expressed genes. These two models are identical in terms of their equivalence classes, but will differ in the assumed hyperdistribution.

Two principal options exist for varying the assumed hyperdistributions between models that share the same equivalence classes. Firstly, since the purpose of the two models is to separate two qualitatively different sets of biomolecular events, we may find some function of the values in Θ_q that splits the data. The data shown in Figure S9 can be split by minimising the intra-class variance [3]. Sampled values mapping to the left of the threshold indicated by the vertical red line represent the distribution of data for M_{NE} while those to the right represent the distribution of data for M_{NDE}.

*to whom correspondence should be addressed ¡tjh48@cam.ac.uk¿

†Department of Plant Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EA, United Kingdom

In some cases, the distinction between two quantitatively different models for gene expression introduces a natural variation between the hyperdistributions. For example, in paired data, a substantial proportion of the data may be equivalently expressed within all pairs, and this may be regarded as a qualitatively different scenario to equivalent expression across replicates but divergent expression within each pair. We have previously shown [2] that these cases can be analysed by constructing a model for equivalent expression across replicates. Assuming a beta-binomial distribution with parameters p, the proportion of counts observed in the first member of each pair, and ϕ, the dispersion, a set Θ_q can be constructed by maximum likelihood methods, as discussed in Section 2. We can then construct a second model describing equivalent expression within pairs in which the calculated values for ϕ are used for the dispersions but in which the values for p are set to 0.5, the value which corresponds to a hypothesis of balanced expression between pairs.

We simulate a set of data following [1, 4] in which data from ten thousand genes in ten samples are simulated from a negative binomial distribution, with means sampled from a SAGE dataset. Dispersions for each gene are sampled from a gamma distribution with shape = 0.85 and scale = 0.5. Library sizes for each sample are sampled from a uniform distribution between 30000 and 90000. One thousand of the genes are simulated to have an eight-fold differential expression in either direction between the first and second sets of five samples each. A further one thousand genes have their mean expression reduced by a factor of twenty; these represent a set of unexpressed genes within the data.

The parameters of an assumed negative-binomial distribution are μ_q^h, ϕ^h, where μ_q^h represents the estimated mean (scaled for library size) for some equivalence class q estimated by sampling some genomic event h, and ϕ^h the similarly estimated dispersion. An initial weighting on these parameters for a model M_{NE} of no expression is acquired by considering the log of the μ_q^h estimated in a model of no differential expression. Figure S10 shows the distribution of this random variable and the threshold ψ which, if used to split this variable, minimises the intra-class variance. For a model M_{NE} of no expression, the initial weights used are $w^h_{M_{NE}} = 1$ if $\log \mu_q^h < \psi$ and 0 otherwise, while for all other models, the weights used are $w^h_M = 1 - w^h_{M_{NE}}$. We then iteratively use Eqn 8 to update these weightings and improve the posterior likelihoods acquired for each model. At each iteration, $w^h_{M_{NE}} = p^h_{M_{NE}}$ (the current estimate of the posterior likelihood that gene h is not expressed) and $w^h_M = 1 - w^h_{M_{NE}}$ for all other models. Figure S1 shows the performance of these methods in simultaneously identifying non-expressed and differentially expressed genes in these simulations.

S2 Variable Model Priors

Figure S11 shows a reanalysis of the simulated data used in Soneson *et al* [5]. This figure is derived from simulated data equivalent to 12450 genes from 10 samples, of which approximately 1250 are differentially expressed between the first five and second five samples. In one set of simulations, the differentially expressed genes are equally likely to be up-regulated as down-regulated between the two groups in the data, while in the other, all differential expression is up-regulation of the second group relative to the first. Allowing `baySeq v2` to choose different model priors depending on which group has higher average expression gives a substantial increase in performance in the unbalanced case, while not affecting performance for the balanced data.

S3 Simulation of zero-inflated negative binomial data

We base the simulation of these data on previous simulations developed to generate high-throughput sequencing data [1, 4] in which data from ten thousand genes in ten samples are simulated from a negative binomial distribution, with means sampled from a SAGE dataset. Increased sequencing depth can be explored by scaling these sampled means. Dispersions for each gene are sampled from a gamma distribution with shape = 0.85 and scale = 0.5. Library sizes for each sample are sampled from a uniform distribution between 30000 and 90000. One thousand of the genes are simulated to have an eight-fold differential expression in either direction between the first and second sets of five samples. For each gene, we then sample a proportion p_c of zero-inflation from a uniform distribution between 0 and 0.5, and for each sample in that gene, replace the observed value with a zero with probability p_c.

Supplementary Figures

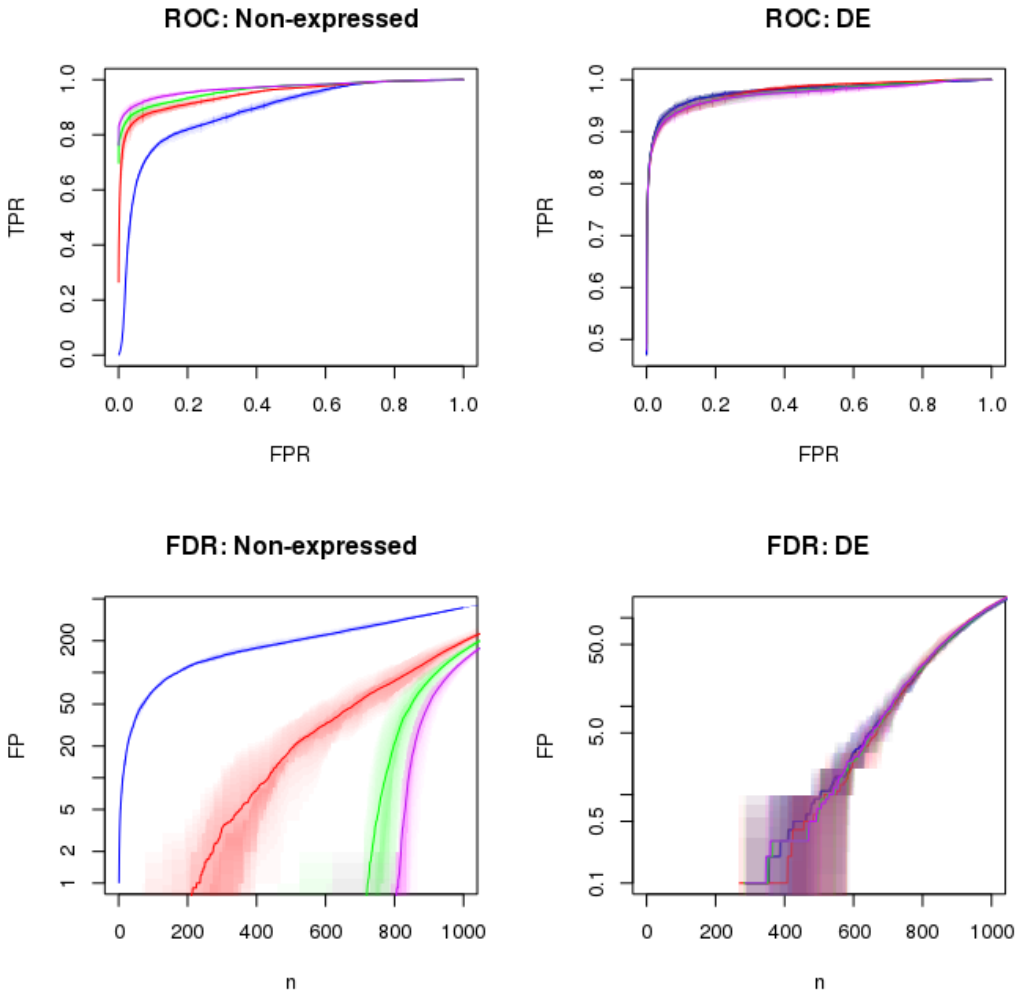

Figure S1: Mean ROC and FDR curves for discovery of non-expressed and differentially expressed data in simulation studies, with no bootstrapping (blue), and two (red), five (green) and ten (purple) cycles of bootstrapping. Percentiles of true discovery rates (for ROC curves) and false discovery rates (for FDR curves) across simulations are shown as transparent areas around curves.

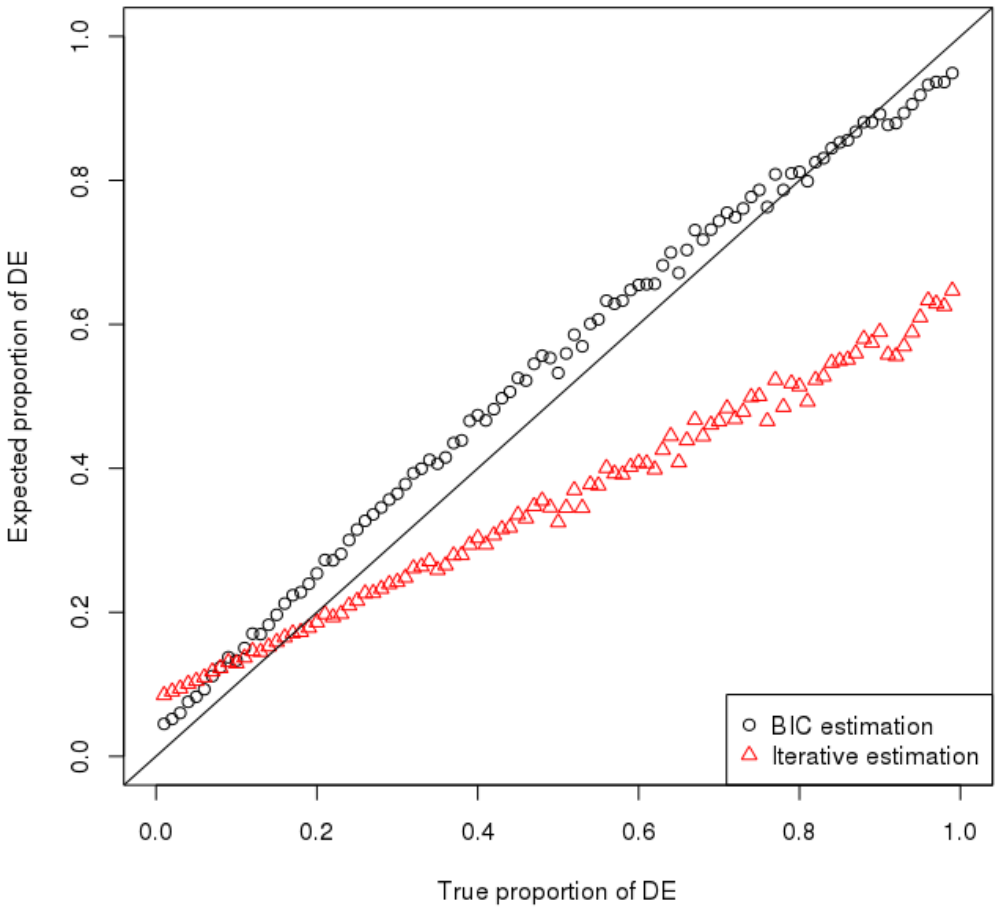

Figure S2: The expected proportion of differentially expressed genes against the true proportion in simulation studies following [1, 4] with 10 libraries. The expected proportion of differentially expressed genes was calculated by summing the posterior likelihoods of differential expression, and dividing by the total number of genes. Model priors were calculated using the BIC method described in 2 or the iterative method described in Hardcastle (2010) [1].

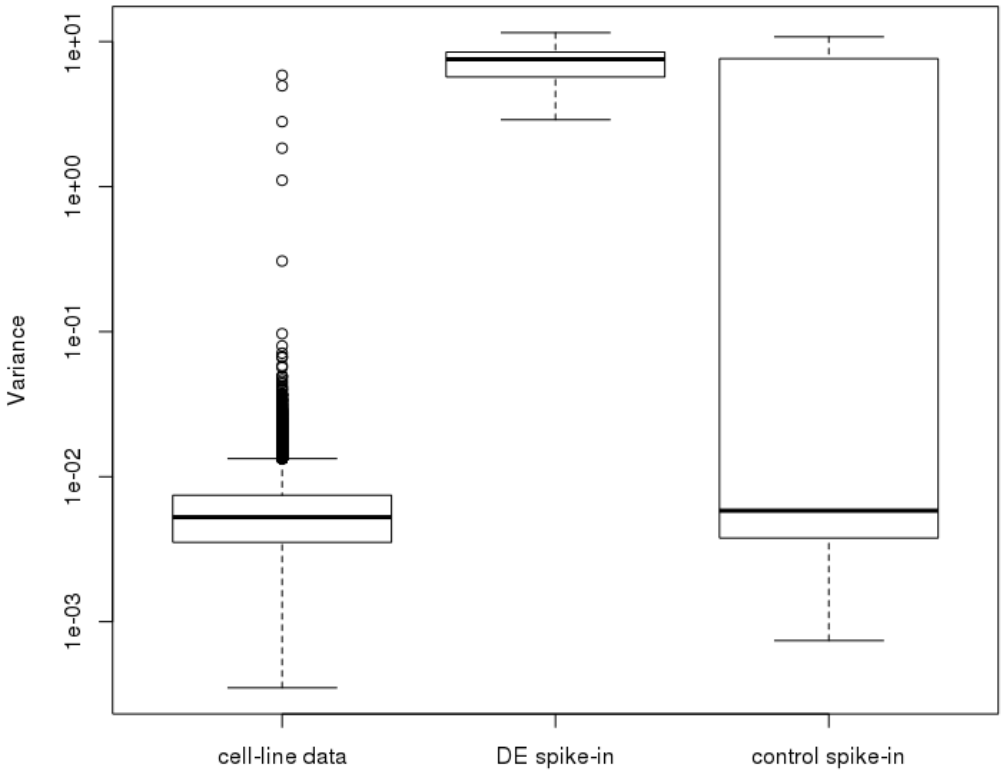

Figure S3: Gene/spike-in variances of expression across all arrays in the Affymetrix HGU133A Latin Square data. Differentially expressed spike-ins have much higher variance than data describing gene expression in cell-lines, as expected. Variance in the non-differentially expressed control spike-ins can be as high as that seen in differentially expressed data, suggesting that these should be removed from differential expression analyses.

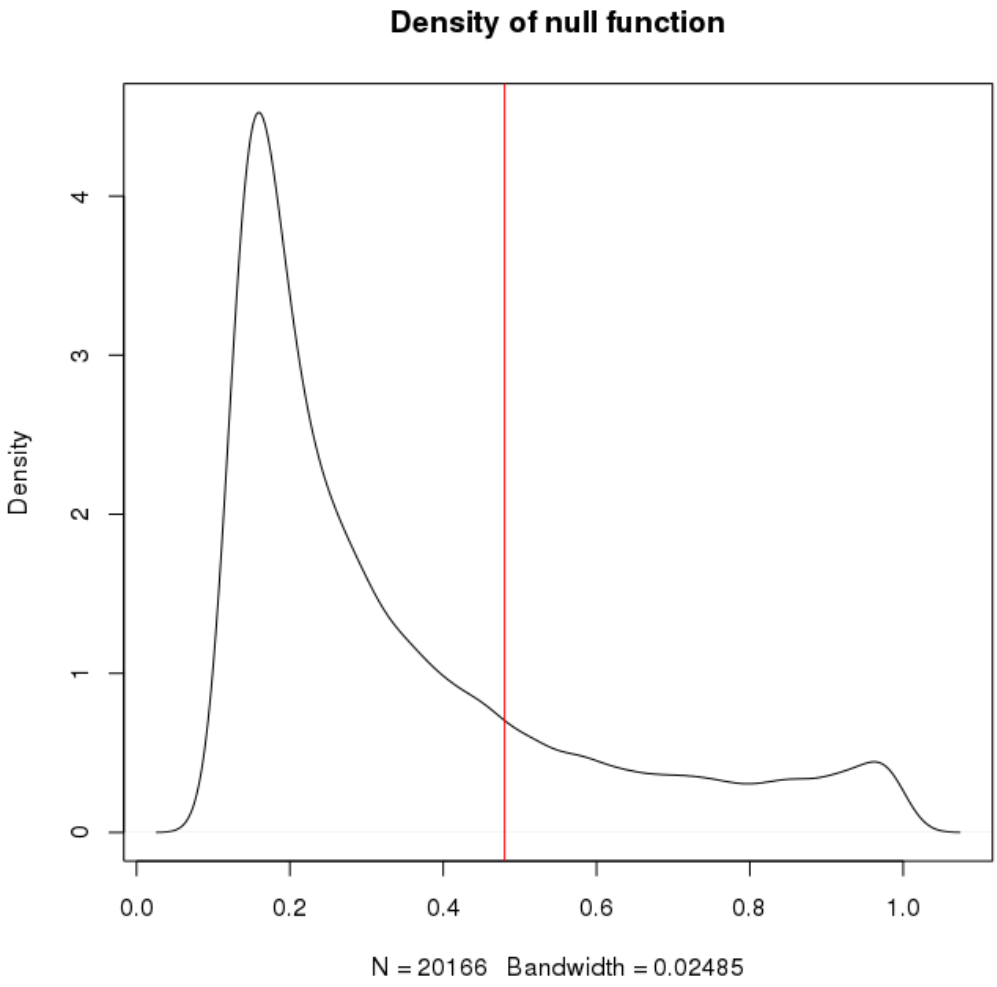

Figure S4: Distribution of p_{q1} in a model of consistent expression between age groups for the observed expression from from ten tissue types (adrenal gland, brain, heart, kidney, liver, lung, muscle, spleen, thymus, and uterus) in female rats, comparing four juvenile (2-week old) to four aged (104-week old) individuals in the Rat BodyMap project [6]. The threshold (red) is chosen to minimise the intra-class variance of the partitioned data.

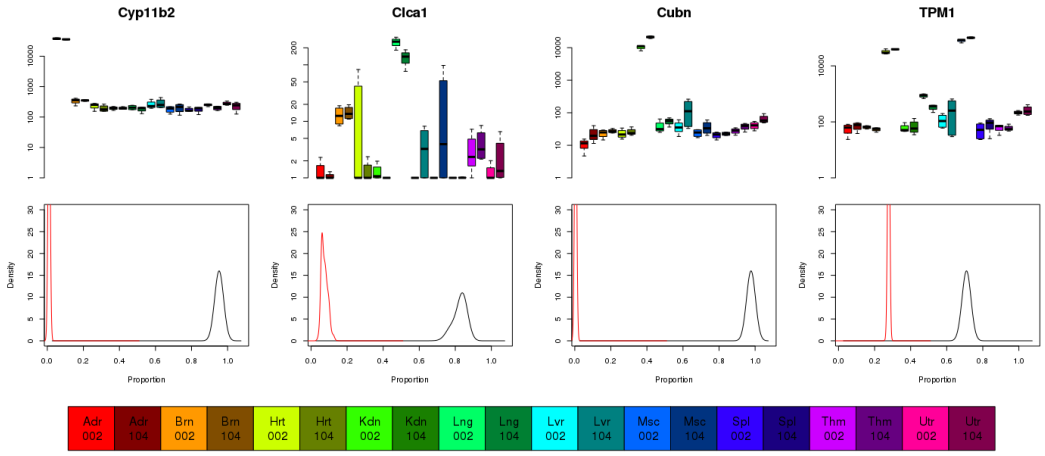

Figure S5: Gene expression plots and posterior distributions of the parameters for the proportion of expression in the tissue of highest (black) and second highest (red) expression. These four genes are the top ranked genes for a change in expression between tissue but not over time.

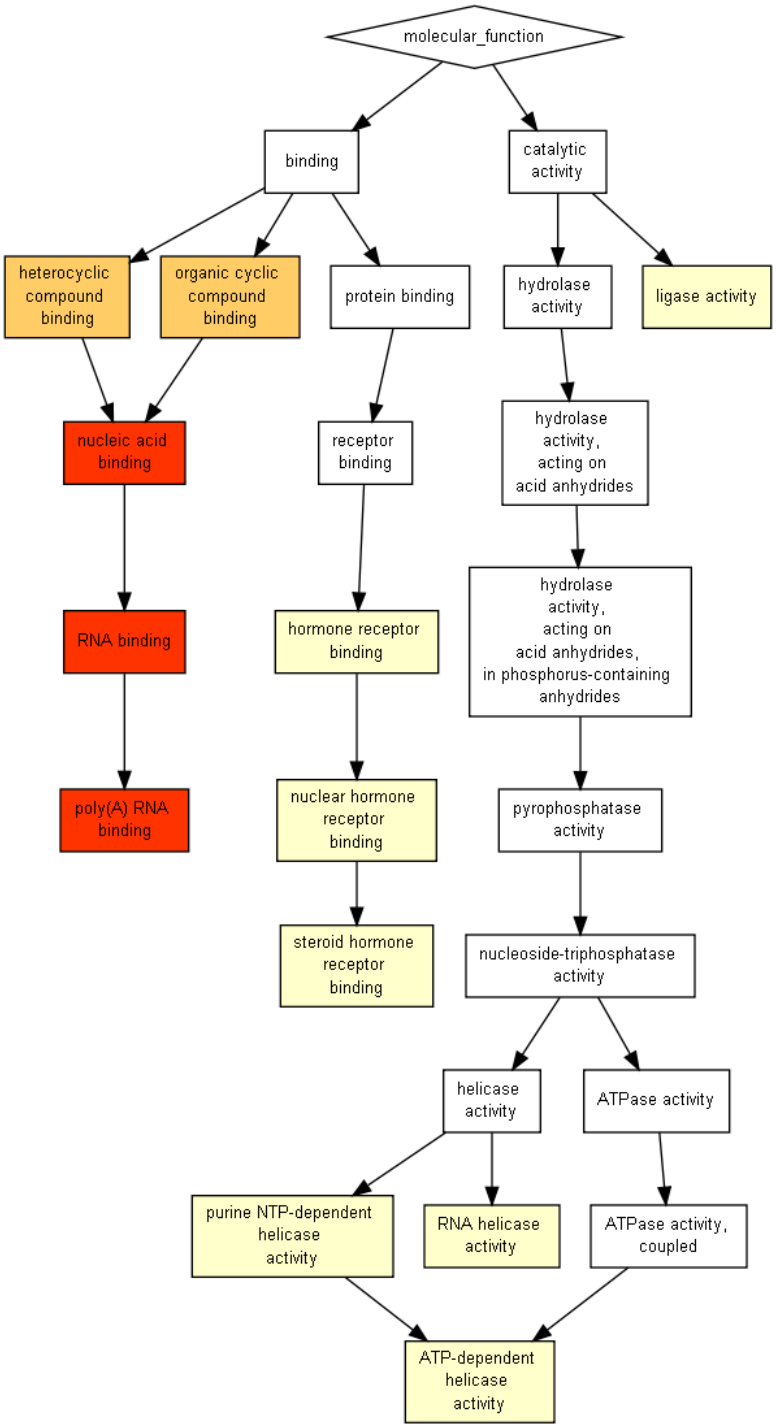

Figure S6: Functional GO terms enriched in the gene set showing a decline in expression in thymus tissue relative to the expression across all other tissues.

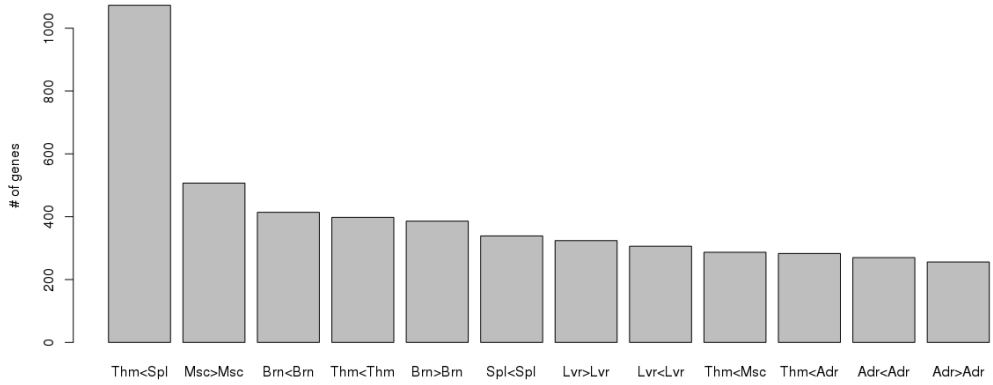

Figure S7: Numbers of genes for which there is a change in ratio of expression between tissues over time, split by tissues of maximum expression in the two time points. Bars are labelled as X¿/¡Y, such that the tissue with the highest proportion of expression in juvenile individuals is X, in aged individuals Y. The higher proportion of expression is indicated by the '¿/¡' symbol.

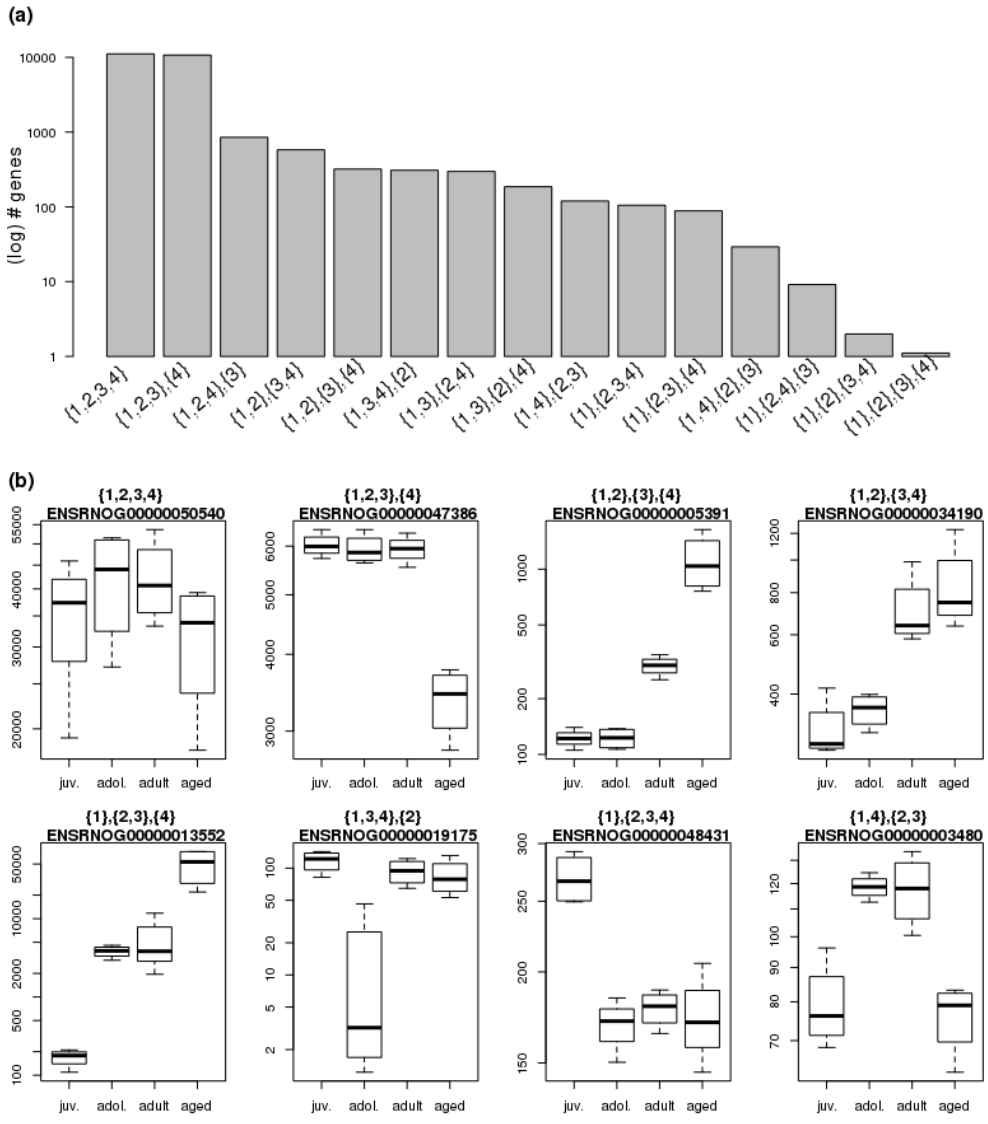

Figure S8: The expected number of genes belonging to each model (a). Normalised expression values of the top ranked gene for each of the eight models with highest expected number of genes, summarised by age group (b). Models are defined such that, e.g. 1,2,34 indicates that the data from age groups 1,2 and 3 are equivalently distributed, and differ from age group 4. Age groups are identified in order, with 1 corresponding to juvenile individuals and 4 to aged individuals.

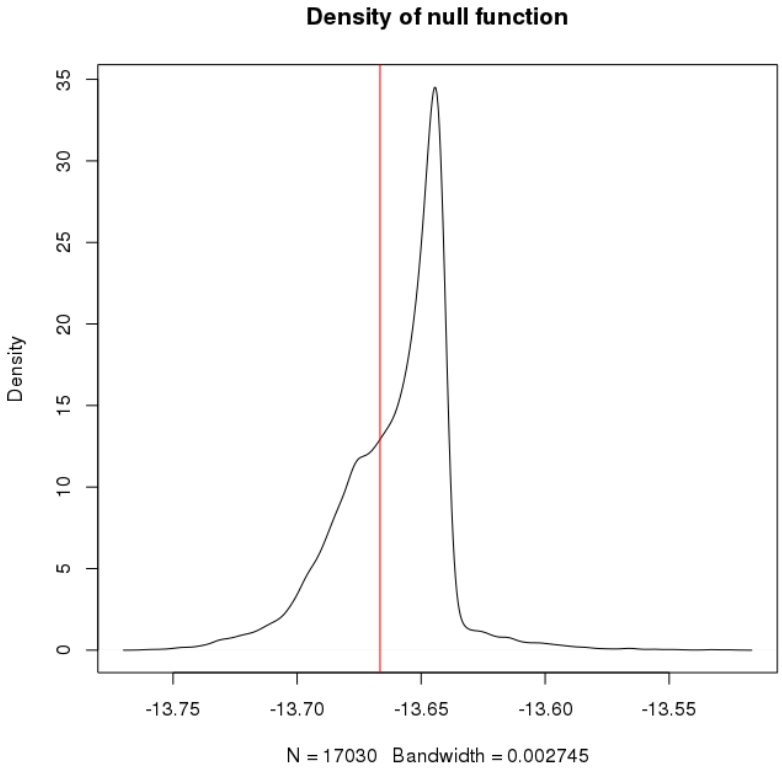

Figure S9: Distribution of the log of the parameter associated with the mean expression (scaled by library scaling factor and gene length) in RNA-seq data derived from rat thymus in juvenile female individuals [6]. The tail of data to the left of the modal peak may be considered to represent non-expressed genes. The red line indicates the threshold level which minimises the intra-class variance.

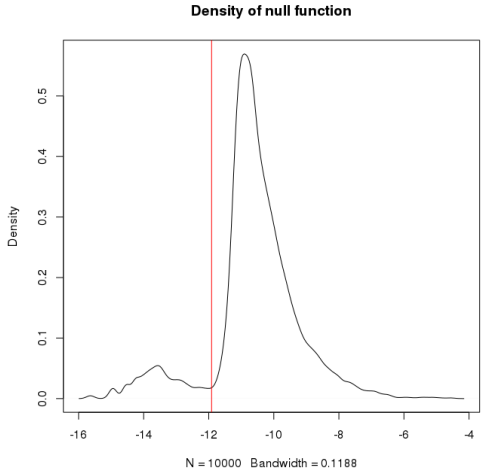

Figure S10: Distribution of the log of the parameter associated with the mean expression (scaled for library scaling factor) in a set of simulated RNA-Seq data. The tail of data to the left of the modal peak may be considered to represent non-expressed genes. The red line indicates the threshold level which minimises the intra-class variance.

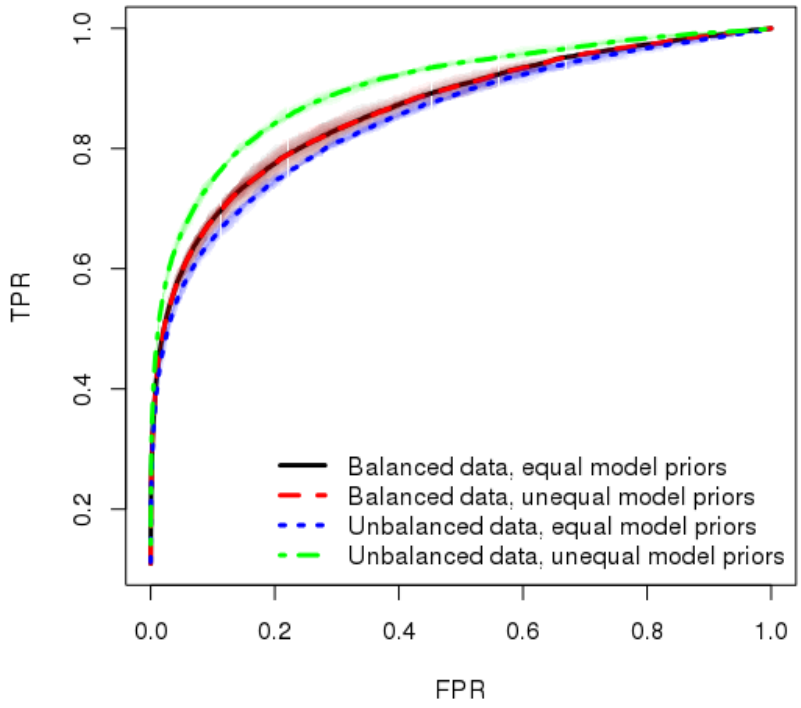

Figure S11: Average ROC curves showing performance of `baySeq v2` on balanced and unbalanced differentially expressed data. Allowing unequal model priors for different sets of the data increases performance for unbalanced data. Percentiles of true positive rates across samplings are shown as transparent areas around curves.

Supplementary Tables

GO term	Description	P-value	FDR q-value
GO:0003723	RNA binding	3.17E-13	4.14E-10
GO:0044822	poly(A) RNA binding	1.66E-10	1.08E-7
GO:0003676	nucleic acid binding	2.69E-10	1.17E-7
GO:1901363	heterocyclic compound binding	5.69E-6	1.86E-3
GO:0097159	organic cyclic compound binding	7.25E-6	1.89E-3
GO:0008026	ATP-dependent helicase activity	2.12E-4	4.61E-2
GO:0070035	purine NTP-dependent helicase activity	2.12E-4	3.95E-2
GO:0051427	hormone receptor binding	2.86E-4	4.66E-2
GO:0035257	nuclear hormone receptor binding	3.05E-4	4.42E-2
GO:0035258	steroid hormone receptor binding	3.91E-4	5.09E-2
GO:0016874	ligase activity	7.28E-4	8.63E-2
GO:0003724	RNA helicase activity	8.34E-4	9.07E-2

Table S1: Functional GO terms enriched in the gene set showing a decline in expression in thymus tissue relative to the expression across all other tissues.

References

[1] T. J. Hardcastle and K. A. Kelly. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11(1):422, Jan. 2010.

[2] T. J. Hardcastle and K. A. Kelly. Empirical Bayesian analysis of paired high-throughput sequencing data with a beta-binomial distribution. *BMC Bioinformatics*, in press, 2013.

[3] N. Otsu. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

[4] M. D. Robinson and G. K. Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–7, Nov. 2007.

[5] C. Soneson and M. Delorenzi. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, 14(1):91, Jan. 2013.

[6] Y. Yu, J. C. Fuscoe, C. Zhao, C. Guo, M. Jia, T. Qing, D. I. Bannon, L. Lancashire, W. Bao, T. Du, H. Luo, Z. Su, W. D. Jones, C. L. Moland, W. S. Branham, F. Qian, B. Ning, Y. Li, H. Hong, L. Guo, N. Mei, T. Shi, K. Y. Wang, R. D. Wolfinger, Y. Nikolsky, S. J. Walker, P. Duerksen-Hughes, C. E. Mason, W. Tong, J. Thierry-Mieg, D. Thierry-Mieg, L. Shi, and C. Wang. A rat RNA-Seq transcriptomic BodyMap across 11 organs and 4 developmental stages. *Nature Communications*, 5:3230, Jan. 2014.