BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 11 2014, pages 1555–1561
doi:10.1093/bioinformatics/btu088

Gene expression

Advance Access publication February 11, 2014

# Analysis of gene expression data using a linear mixed model/finite mixture model approach: application to regional differences in the human brain

Daniah Trabzuni[1,2], the United Kingdom Brain Expression Consortium (UKBEC)[1] and Peter C. Thomson[3,*]

[1]Department of Molecular Neuroscience, UCL Institute of Neurology, Queen Square, London WC1N 3BG, UK, [2]Department of Genetics, King Faisal Specialist Hospital and Research Centre, Riyadh 11211, Saudi Arabia and [3]ReproGen – Animal Bioscience Group, Faculty of Veterinary Science, The University of Sydney, 425 Werombi Road, Camden, NSW 2570, Australia

Associate Editor: Ziv Bar-Joseph

## ABSTRACT

**Motivation:** Gene expression data exhibit common information over the genome. This article shows how data can be analysed from an efficient whole-genome perspective. Further, the methods have been developed so that users with limited expertise in bioinformatics and statistical computing techniques could use and modify this procedure to their own needs. The method outlined first uses a large-scale linear mixed model for the expression data genome-wide, and then uses finite mixture models to separate differentially expressed (DE) from non-DE transcripts. These methods are illustrated through application to an exceptional UK Brain Expression Consortium involving 12 human frozen post-mortem brain regions.

**Results:** Fitting linear mixed models has allowed variation in gene expression between different biological states (e.g. brain regions, gender, age) to be investigated. The model can be extended to allow for differing levels of variation between different biological states. Predicted values of the random effects show the effects of each transcript in a particular biological state. Using the UK Brain Expression Consortium data, this approach yielded striking patterns of co-regional gene expression. Fitting the finite mixture model to the effects within each state provides a convenient method to filter transcripts that are DE: these DE transcripts can then be extracted for advanced functional analysis.

**Availability:** The data for all regions except HYPO and SPCO are available at the Gene Expression Omnibus (GEO) site, accession number GSE46706. *R* code for the analysis is available in the Supplementary file.

**Contact:** peter.thomson@sydney.edu.au

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

Received on October 20, 2013; revised on January 14, 2014; accepted on February 6, 2014

## 1 INTRODUCTION

Research efforts in molecular and cellular biology are vital to develop our understanding of, for example, the human central nervous system (CNS) function, and to dissect the functional complexity and the progress of diseases. These rapid advancements were mainly determined by the genomic, transcriptomic and diagnostic technological innovations (Geschwind and Konopka, 2009). These technologies, such as expression arrays and RNA sequencing, are applicable at the genome-wide scale, and their application results in more insightful information being produced, and this can lead towards more informed investigations and in turn through to clinical applications. However, the analyses and the volume of the data that are generated from these complex techniques are computationally challenging. To date, there is no standard protocol available to normalize and process the raw data from these experiments into a manageable format that can then be further interrogated on a personal computer and to align and assemble the data to a reference genome by scientists with no great experience in bioinformatics (Auer *et al.*, 2012).

With expression arrays and RNA sequencing increasing in coverage and reducing in cost, they are now widely used with increasingly large sample sizes. Consequently, there is an ongoing need to develop and assess computationally efficient means of analysing 'big data'. Although there is no 'universal' approach to gene expression analysis, the conventional approach is to consider a model for one gene at a time, and consider if this particular gene is differentially expressed (DE) between any of the biological states considered, for example, as implemented in the limma package of *R* (Smyth, 2004). Naturally, this results in an extreme multiple testing situation, which is usually overcome by using an appropriate false discovery rate (FDR) control, such as the q-value procedure. The alternative is to consider a global model for all the expression data and replace the many thousands of null and alternative hypotheses by more global tests to dissect causes of variation in gene expression. Such an approach would have two advantages. First, transcripts are not independent identities, and there is much common information across the genome, so combining this would be expected to increase power of signal detection. Second, although a more philosophical point, we question whether it is best practice to consider many thousands of separate hypotheses: as students, we were taught the importance of writing out the null and alternative hypotheses

*To whom correspondence should be addressed.

for every statistical test, and this pragmatic approach would seem to fail here. So a global or genome-wide approach can be achieved by fitting linear mixed models (LMMs) to all the data, for example, as has been done by Kerr *et al.* (2000), with the variation between transcripts as another source of variation. Ji and Liu (2010) suggested a similar process using hierarchical models. One potential limitation of the approach of using an LMM fitted to all the expression data is that it may not allow for differences in variability in expression across the biological states of the system under study. For example, in the human brain, some regions of the brain show greater levels of variation in gene expression than in other regions, However, improvements in LMM software, for example, as described in Butler *et al.* (2007) and Pinheiro and Bates (2000), allowing flexible variance structure modelling, make it worthwhile again considering the use of large-scale LMMs in expression data analysis.

Fitting LMMs by themselves does not lead to the determination of which individual transcripts are DE, nor do the LMMs identify clusters of transcripts with similar expression profiles. However, fitting of finite mixture models has been used successfully in this endeavour, to separate out potential DE from non-DE transcripts (McLachlan *et al.*, 2002).

The following article demonstrates how these two approaches can be adopted in the analysis of a large gene expression dataset. The methods are described to allow easy implementation in other situations. While some of this approach has been described in previous papers (Singh *et al.*, 2013), the present article provides additional detail and extends these methods. To illustrate the methods, one of the most comprehensive gene expression databases related to human brain tissue, namely, the UK Brain Expression Consortium (UKBEC), is used, highlighting different expression patterns across the brain regions (Trabzuni *et al.*, 2011).

## 2 METHODS

### 2.1 Data

Assume that we have expression data recorded from a number of arrays ($m$), and each array has expression measures for $n$ transcripts. Further, each array measures expression levels in one of $s$ different biological states (e.g. regions of the brain). Possibly, the number of replicates for each state is equal ($k$), in which case $m = ks$, but this is not essential. So the entire expression dataset of $m \times n$ observations is described as $y_{ij(k)}$, $i = 1, \ldots, m$; $j = 1, \ldots, n$; with $k = 1, \ldots, s$ being specified for a particular array $i$. It is assumed that the expression arrays have been normalized [e.g. using the RMA procedure (Irizarry *et al.*, 2003)], and that appropriate quality control checks have been undertaken. It is also assumed here that $y_{ij(k)}$ is expressed on a logarithmic (base 2) scale.

### 2.2 Linear mixed model

A large-scale LMM is fitted to all the gene expression data simultaneously. This is used to assess sources of variation in gene expression at an overall level, i.e. to estimate and interpret the variance components. The form of the model, in its simplest form, is

$$y_{ij(k)} = \mu + \text{Array}_i + \text{Transcript}_j + \text{Transcript.State}_{jk} + \varepsilon_{ij}$$

where $y_{ij(k)}$ = normalized expression value on a logarithmic scale; $\mu$ = overall mean expression value; $\text{Array}_i$ = random effect due to array, assumed $N(0, \sigma_A^2)$; $\text{Transcript}_j$ = random effect due to transcript, assumed $N(0, \sigma_T^2)$; $\text{Transcript.State}_{jk}$ = random effect due to transcript in a state, assumed $N(0, \sigma_{TS}^2)$; and $\varepsilon_{ijk}$ = random error, assumed $N(0, \sigma_\varepsilon^2)$. Note that although normalization should remove any between-array effects, the random array effect ($\text{Array}_i$) is included to account for any variation not adjusted for by that process. So it is expected that an estimate of $\sigma_A^2$ would be close to 0. The random transcript effects ($\text{Transcript}_j$) account for overall differences in gene expression between transcripts across all states (i.e. some transcripts have a consistently high or low expression value): this variation must be adjusted for, but is not the term of interest in exploring between-state variation in gene expression: it is the transcript × state effects ($\text{Transcript.State}_{jk}$) that are of interest here.

To fit the LMM, a residual maximum likelihood (REML) method is preferable, particularly when the design is unbalanced. Note that REML as a model-fitting procedure is available in some commercial packages commonly used for brain expression analysis, such as Partek® Genomics Suite™; however, the approach implemented there is still 'one gene at a time'. For example, the REML method is available in the nlme (Pinheiro and Bates, 2000), lme4 (http://CRAN.R-project.org/package=lme4) and ASReml-R (Butler *et al.*, 2007) packages in *R*, the latter being a commercial package. REML estimates of $\sigma_A^2, \sigma_T^2$ and $\sigma_{TS}^2$ (as well as $\sigma_\varepsilon^2$) can be obtained and their significance assessed, either in relation to their standard errors (as available in ASReml-R) or the intervals and confint functions in nlme and lme4, respectively. However, a preferable method is using a REML likelihood ratio test (LRT), comparing full and reduced models (by omitting the random effect in being tested).

Estimated effects, or more precisely the best linear unbiased predictions (BLUPs) of the random effects, can then be obtained, although, as mentioned before, it is the BLUPs of the $\text{Transcript.State}_{ij}$ effects that will be most useful. Typically, these will need to be re-formatted into an $n \times s$ array of effects (transcripts × states). These show how expression changes across states, and relatively simple data exploration including boxplots (by state) and pairwise scatterplots (by state) can reveal much information, particularly in patterns of co-expression. This will be returned to later, in Section 3.

### 2.3 Extending the LMM

One important way in which the above LMM can be extended is to include additional factors that may be associated with gene expression, $X_l$, $l = 1, \ldots, p$. For example, expression may differ according to gender or age group. Further, these factors may also differ between states. For example, there may be different age-related expression effects, but these may differ across regions of the brain. This can be accommodated by extending the model as

$$y_{ij(k)} = \mu + \text{Array}_i + \text{Transcript}_j + \text{Transcript.State}_{jk} + \sum_{l=1}^{p} \text{Transcript.X}_{jl} + \sum_{l=1}^{p} \text{Transcript.State.X}_{jkl} + \varepsilon_{ij}$$

where the additional terms are also random effects with $\text{Transcript.X}_{jl} \sim N(0, \sigma_{TX_l}^2)$, $l = 1, \ldots, p$ and $\text{Transcript.State.X}_{jkl} \sim N(0, \sigma_{TSX_l}^2)$. Again, inclusion of additional fixed effects (and random interaction effects of these with transcripts) is not uncommon in expression analysis; however, the approach here is that this information is included in a genome-wide model.

A second important way the LMM can be extended is by including variance heterogeneity. For example, there may be different levels of genetic variation across regions of the brain, so rather than a single variance component $\sigma_{TS}^2$, there may need to be separate variances for each region (state), say $\sigma_{TR,k}^2$, $k = 1, \ldots, s$; further, there may also need to be separate residual variances for each state, i.e. $\sigma_{\varepsilon,k}^2$, $k = 1, \ldots, s$. The need for these can be diagnosed by plots of the BLUPs of the $\text{Transcript.State}_{jk}$ effects, or the residuals.

### 2.4 Computational aspects of fitting LMMs

Fitting a large-scale mixed model can be computationally demanding, especially for a large expression dataset. One solution to this is to fit the model separately to $q$ subsets of the data, each subset having a randomly selected $1/q$ of the transcripts. Note that all records on the one transcript should be retained in the same data subset file. Hence, any differences could be attributed to sampling fluctuations; nevertheless, this is checked by comparing parameter estimates (fixed effect $\mu$, and variance component estimates) from each model fit. The estimates from all subsets can then be combined in a weighted average,

$$\hat{\sigma}_c^2 = \sum_{i=1}^{q} w_i \hat{\sigma}_{i,c}^2 \Big/ \sum_{i=1}^{q} w_i$$

and the standard error of the estimate obtained as $\mathrm{se}(\hat{\sigma}_c^2) = (\sum_{i=1}^{q} w_i)^{-1/2}$ where $w_i = [\mathrm{se}(\hat{\sigma}_{i,c}^2)]^{-2}$, with $\hat{\sigma}_{i,c}^2$ and $\mathrm{se}(\hat{\sigma}_{i,c}^2)$ being the estimate of variance component $c$ and its standard error, based on dataset $i = 1, \ldots, q$. Note that while the $\mathrm{se}(\hat{\sigma}_{i,c}^2)$ are available in ASReml-R, they are not available using nlme or lme4. In that instance, they may be approximated as one quarter of the confidence interval width, as returned using the intervals function in the nlme package, or the confint function in the lme4 package. However, an unweighted average would also be appropriate, given that the only differences would be due to sampling variation associated with the random splitting of the full dataset.

There are also benefits of fitting the model to random subsets of the expression data while developing the model, as this will allow a variety of models to be evaluated in a reasonable time. Estimates of variance components are remarkably similar when using only a sample of $m = 1000$ transcripts compared with a full dataset of $m = \sim 20\,000$ transcripts, so inferences made on a smaller sample may be appropriate.

### 2.5 Finite mixture model to assess DE transcripts

Having obtained the model-based predicted values of the transcript × state effects, the next issue is to assess which transcripts are differentially expressed across the states. The definition here of DE is a transcript showing an unusually high or low expression level in that particular state, compared with all the other gene transcript expression levels in that state, i.e. it is not defined in terms of DE between states. Typically, a histogram (or normal Q–Q plot) of the BLUPs of these Transcript.State$_{jk}$ effects (say $g_{jk}$) within any one state $k$ shows extreme tails. This can be modelled as a finite mixture distribution consisting of two components, the one with the larger variance $\sigma_1^2$ consisting of DE transcripts and the other with a smaller variance $\sigma_0^2$ consisting of non-DE transcripts (non-DE). The proportions of transcripts that are DE versus non-DE, i.e. the prior probabilities, are $\pi_1 = P(\mathrm{DE})$ and $\pi_0 = 1 - \pi_1 = P(\text{non-DE})$. Further, it is assumed that the expression values are normally distributed, i.e. $g|\mathrm{DE} \sim N(0, \sigma_1^2)$ and $g|\text{non-DE} \sim N(0, \sigma_0^2)$. This is indicated in Supplementary Figure S2. Fitting the mixture model separately for the transcript effects in state $k$ can be undertaken using maximum likelihood. The log-likelihood is calculated as

$$\log_e L(\sigma_{1k}^2, \sigma_{0k}^2; \mathbf{g}_k) = \sum_{j=1}^{n} \log_e[\pi_{1k} f(g_{jk}; \sigma_{1k}^2) + (1 - \pi_{1k}) f(g_{jk}; \sigma_{0k}^2)]$$

where $f(g; \sigma^2) = (2\pi)^{-1/2} \exp[-g^2/(2\sigma^2)]$ is the probability density function for an $N(0,\sigma^2)$ random variable. This mixture distribution is easily fitted using the expectation–maximization (E–M) algorithm (McLachlan and Krishnan, 1997). The details of implementing the E–M algorithm as well as assessing the fit of the finite mixture model can be found in Supplementary File Section 1.6.

As a by-product of the E–M algorithm, the (posterior) probability ($\tau_{jk}$) of each specific transcript being DE is returned. Any transcript with $\tau_{jk} > 0.5$ is more likely DE than non-DE. However, to reduce false-positive results, a higher threshold is recommended. Typically, a threshold of $\tau_{jk} > 0.8$ is a good choice, and there is usually a distinct subset of transcripts with high DE probabilities (as indicated from a histogram of these values). This choice can be evaluated by calculation of the corresponding FDR, $\mathrm{FDR} = 1 - \mathrm{mean}\{\hat{\tau}_{jk} | \hat{\tau}_{jk} > 0.8\}$, i.e. the average of the posterior probability of a transcript being non-DE, given it was declared as being DE.

The end result of these two processes, i.e. fitting the LMM to obtain genetic effects for each state, and then assessing which transcripts are DE in each state, provides a smaller dataset ($n_0$ transcripts × $k$ states), where likely, $n_0 \ll n$, which can then be used for further targeted analysis. This may involve assessing patterns of co-expression across different states, or leading towards evaluation of functional gene networks. Some of these further explorations are considered in the following section that describes the implementation of these methods outlined above.

## 3 APPLICATION

### 3.1 Description of study and data

To illustrate the methods presented here, use is made of a study involving 101 frozen human post-mortem brain and CNS tissue samples of Caucasian controls with no neurological conditions, and these were obtained from the Medical Research Council (MRC) Sudden Death Brain and Tissue Bank, Edinburgh, UK (Millar *et al.*, 2007). From each brain, tissue samples from up to 12 brain regions were collected, namely, cerebellum (CRBL, from $n = 95$ brains), frontal cortex (FCTX, $n = 96$), hippocampus (HIPP, $n = 93$), hypothalamus (HYPO, $n = 13$), medulla (specifically inferior olivary nucleus, MEDU, $n = 92$), occipital cortex (specifically primary visual cortex, OCTX, $n = 95$), putamen (PUTM, $n = 94$), substantia nigra (SNIG, $n = 68$), spinal cord (SPCO, $n = 13$), temporal cortex (TCTX, $n = 84$), thalamus (THAL, $n = 83$) and intralobular white matter (WHMT, $n = 91$). RNA samples were extracted and expression levels analysed using Affymetrix Exon 917 arrays with 19 215 gene transcripts recorded. This represents a major part of the UKBEC dataset, the most comprehensive CNS expression dataset to date. The expression dataset was normalized using the RMA procedure, and routine quality control checks were applied, including detection above background filtering. Further details of the study are found in Trabzuni *et al.* (2011).

### 3.2 Analysis of the UKBEC data

For the analysis of this dataset, the initial model fitted to the expression data was based on $n = 1000$ randomly selected transcripts. The form of the model was

$$y_{ij(kl)} = \mu + \mathrm{Array}_i + \mathrm{Transcript}_j + \mathrm{Transcript.Region}_{jk} + \mathrm{Transcript.Gender}_{jl} + b_j \mathrm{Age}_i + \varepsilon_{ij}$$

where $y_{ij(kl)}$ is the RMA-normalized data, on a log-2 scale, and all terms except $\mu$ are random effects. Note that gender was included to test if there were gender-specific expression patterns. Also, the age of the brain donor was included as a random covariate effect, i.e. different linear effect of age associated with each gene transcript, $b_j \sim N(0, \sigma_{TAge}^2)$. However, while statistically significant based on crude Wald tests and on LRTs, the magnitude of these two variance components was substantially smaller than all other variance components (including array); consequently, these two effects were subsequently dropped from the model. [Analyses of this dataset where the age and

gender effects have been reported in detail can be found in Kumar *et al.* (2013) (age) and Trabzuni *et al.* (2013) (gender).] However, this initial modelling also indicated that the variation in these gene transcript effects differed across regions, so a common transcript × region variance ($\sigma^2_{TR}$) was replaced by a separate variance for each region ($\sigma^2_{TR,k}$; $k = 1, \ldots 12$). Also as a check, separate residual variances were fitted for each region ($\sigma^2_{\varepsilon,k}$; $k = 1, \ldots 12$), but the estimated variance components were similar across the 12 regions, and this did not improve the fit of the model. This is also supported by the examinations of residuals by brain region (Supplementary Fig. S3).

In the definitive analysis, the data were split randomly into two ($q = 2$) subsets (due to computational limits on the server), each containing half of the 19 215 transcripts. These were labelled Part 1 and Part 2. Consequently, each model involved predicting $½ \times 19\,215 \times 13 \approx 125\,000$ random effects (Transcript + Transcript.Region), as well as 917 array random effects. These BLUPs of the Transcript.Region effects from each fitted model were combined into the one dataset. This dataset then becomes the main dataset for subsequent gene expression analysis (i.e. 19 215 transcripts × 12 regions), including fitting the mixture model within each region. Note that some coding details in *R* are provided in the Supplementary File Section S1, including using the asreml, lmer and lme functions, as well coding of the E–M algorithm for fitting the finite mixture models.

**Table 1.** Parameter estimates for the two randomly selected subsets of transcripts and their combined estimates

| Parameter | Subset 1 Estimate ± SE | Subset 2 Estimate ± SE | Combined Estimate ± SE |
|---|---|---|---|
| μ | 5.305 ± 0.015 | 5.297 ± 0.015 | 5.301 ± 0.011 |
| $\sigma^2_A$Array | 0.007 ± 0.000 | 0.007 ± 0.000 | 0.007 ± 0.000 |
| $\sigma^2_T$Transcript | 2.040 ± 0.030 | 2.025 ± 0.029 | 2.032 ± 0.021 |
| $\sigma^2_{TR,1}$CRBL | 0.795 ± 0.012 | 0.803 ± 0.012 | 0.799 ± 0.008 |
| $\sigma^2_{TR,2}$FCTX | 0.222 ± 0.004 | 0.227 ± 0.004 | 0.225 ± 0.003 |
| $\sigma^2_{TR,3}$HIPP | 0.117 ± 0.002 | 0.131 ± 0.002 | 0.123 ± 0.001 |
| $\sigma^2_{TR,4}$HYPO | 0.042 ± 0.001 | 0.041 ± 0.001 | 0.042 ± 0.001 |
| $\sigma^2_{TR,5}$MEDU | 0.143 ± 0.002 | 0.134 ± 0.002 | 0.138 ± 0.002 |
| $\sigma^2_{TR,6}$OCTX | 0.246 ± 0.004 | 0.258 ± 0.004 | 0.252 ± 0.003 |
| $\sigma^2_{TR,7}$PUTM | 0.174 ± 0.003 | 0.190 ± 0.003 | 0.181 ± 0.002 |
| $\sigma^2_{TR,8}$SNIG | 0.123 ± 0.002 | 0.125 ± 0.002 | 0.124 ± 0.002 |
| $\sigma^2_{TR,9}$SPCO | 0.176 ± 0.003 | 0.167 ± 0.003 | 0.171 ± 0.002 |
| $\sigma^2_{TR,10}$TCTX | 0.244 ± 0.004 | 0.251 ± 0.004 | 0.247 ± 0.003 |
| $\sigma^2_{TR,11}$THAL | 0.141 ± 0.002 | 0.142 ± 0.002 | 0.142 ± 0.002 |
| $\sigma^2_{TR,12}$WHMT | 0.419 ± 0.006 | 0.404 ± 0.006 | 0.411 ± 0.004 |
| $\sigma^2_\varepsilon$Residual | 0.202 ± 0.000 | 0.202 ± 0.000 | 0.202 ± 0.000 |

## 4 RESULTS

The estimates of the overall expression mean ($\mu$) and the variance components from each random subset, as well as their combined estimates, are shown in Table 1. It is clear that the estimates from both subsets are similar and consistent with sampling fluctuations.

There is very little between-array variation, as indicated by its small estimated variance component ($\hat{\sigma}^2_A = 0.007 \pm 0.0002$). There was a relatively large estimated variance component to assess between-transcript variation ($\hat{\sigma}^2_T = 2.032 \pm 0.0209$). (BLUPs of overall transcripts effects can be obtained to highlight differing overall expression levels between transcripts.) For the transcript-specific effects in individual regions, the level of variation differed widely: CRBL showed the greatest level of variation ($\hat{\sigma}^2_{TR,1} = 0.799 \pm 0.0083$), followed by WHMT ($\hat{\sigma}^2_{TR,12} = 0.411 \pm 0.0044$), with HYPO displaying the least variation ($\hat{\sigma}^2_{TR,4} = 0.042 \pm 0.0008$). Note the formal REML LRTs indicate the heterogeneous model was superior to a homogeneous model with a common between-transcript variance: for Part 1, LRT = 21 044 and for Part 2, LRT = 20 327, and compared with a chi-square distribution with 11 df returns $P = 0$ in both. Estimated residual variation ($\hat{\sigma}^2_\varepsilon = 0.202 \pm 0.0001$) accounted for between 6.6 and 8.8% of the variation in expression levels (depending on the region). Supplementary Figure S4 shows a boxplot of the BLUPs of the transcript × region interaction effects. It corroborates the large estimated variance for CRBL and WHMT, and the small variance for HYPO.

Supplementary Table S2 shows the correlation matrix (Pearson correlation, *r*) between these region-specific transcript effects. Correlations (almost) >0.5 in absolute value have been coloured (red: positive; blue: negative), and those >0.75 are shown in bold. It is noteworthy that the expression levels in CRBL were not substantially correlated with those in other regions, the greatest being with OCTX ($r = 0.318$). The highest correlations (all positive) were between the three cortex regions (FCTX, OCTX and TCTX). To illustrate, Figure 1 shows a smoothed scatter plot (using the smoothScatter function in the geneplotter package of *R*) indicating the strong correlations between FCTX and OCTX ($r = 0.86$). Supplementary Figure S5 shows similar pairs of regions for which $|r| > 0.5$, including a negative correlation between TCTX and SPCO ($r = -0.75$).

Figure 2 shows details of the region-specific transcript effects for HYPO. (Similar results for the other 11 regions can be found in the Supplementary Figs S6–S9). Figure 2A shows the histogram of the distributions of transcripts effects, supported by a corresponding Q–Q plot in Figure 2B. They show evidence of some extreme effects, i.e. differential gene expression, both upregulated and downregulated, as indicated by the departures from the line in the Q–Q plot. Similar findings are found in the other regions.

The two-component mixture model was fitted to the transcript effects within each region, and these are summarized in Table 2. On average, just <30% of transcripts show differential expression in any one region, although this varies from 16 (HYPO) to 45% (OCTX). In terms of variation in expression, the DE transcripts show on average 3.3× greater standard deviations than non-DE transcripts, and this ratio is fairly consistent across regions ($r = 0.90$). Note that using results for mixture distributions, the overall REML variance estimates in Table 1 can be approximately reproduced here, as $\sigma^2_{TR,k} = \pi_1\sigma^2_1 + (1 - \pi_1)\sigma^2_0$ (because the means of both mixture components can be taken as 0). To illustrate here for CRBL, from Table 2, $0.338 \times 1.306^2 + 0.662 \times 0.559^2 = 0.784$, compared with $\hat{\sigma}^2_{TR,1} = 0.799$ in Table 1.

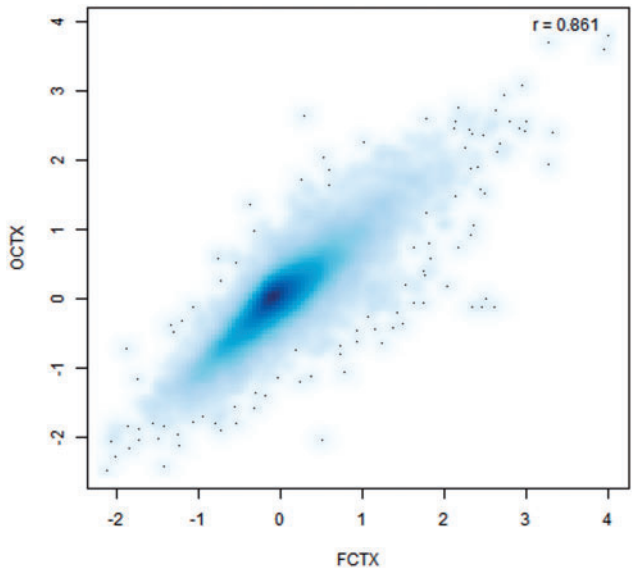


**Fig. 1.** Smoothed scatter plot of the estimated Region.Transcript effects for OCTX versus FCTX

Figure 2C shows the Q–Q plot for the fitted mixture model for HYPO: while not a perfect fit, they fit the expression effects reasonably well, indicating that a two-component mixture model is adequate. (Two transcripts are identified on that plot as being even more extreme: one upregulated and one downregulated). Figure 2D shows the distribution of the probabilities of transcripts being DE in HYPO. What is apparent for each region is a cluster of transcripts with extremely high probabilities of being DE, i.e. DE transcripts. Transcripts with a probability $\tau > 0.8$ can be selected as being DE. The average of the values $(1 - \tau)$ for transcripts declared DE is the FDR: for the current dataset, this is calculated as 4.3% [although it varies across regions, from 3.1 (HIPP) to 5.7% (CRBL)].

## 5 SUBSEQUENT FUNCTIONAL ANALYSIS OF DE GENES

Fitting the LMM results in the production of a 'dataset' of Transcript.Region effects (19 215 rows × 12 columns). Fitting the finite mixture model within each region identified 7786 transcripts that are DE in one or more regions. These can be extracted to form a reduced file (7786 rows × 12 columns) and apply standard bioinformatic approaches for downstream analysis. For example, a heatmap of this dataset (using the heatmap function in *R*) confirms the similar patterns of expression across the three cortex regions, and distinct separation of CRBL and WHMT from other regions (Fig. 3). It also demonstrates two 'super-clusters' of transcripts.

A simple means of clustering transcripts with similar expression profiles is to group together transcripts with the same patterns of DE (upregulation, +; downregulation; –) or non-DE (.) across the 12 regions (Fig. 4). While Figure 4A shows a simple cluster (upregulated in CRBL only, 381 transcripts), Figure 4B shows a far more complex cluster of gene expression showing up- and downregulation in 6 of the 12 regions (82 transcripts). With

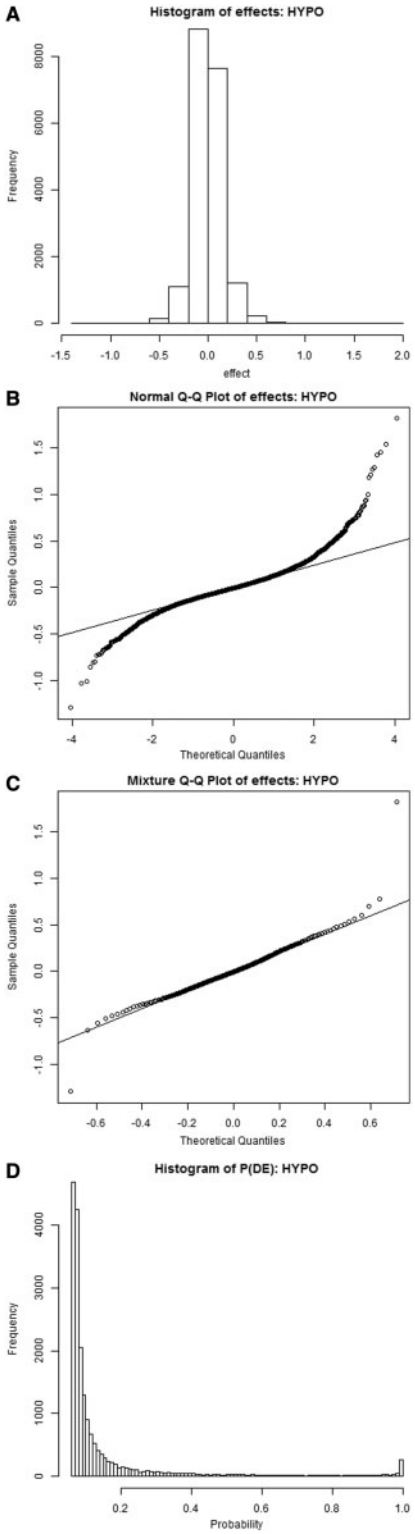


**Fig. 2.** Analysis of expression data from the hypothalamus (HYPO). The four panels show (**A**) histogram of the estimates of the Transcript.Region effects in that region; (**B**) normal Q–Q plot of these effects demonstrating the extreme tails of this distribution; (**C**) mixture model Q–Q plot (DE and non-DE) of these effects indicating a satisfactory model fit; and (**D**) histogram of the posterior probability of the transcripts being DE; note the subset of transcripts with a high probability of being DE

**Table 2.** Parameter estimates of the mixture model fitted separately to each brain region

| Region | CRBL | FCTX | HIPP | HYPO | MEDU | OCTX |
|---|---|---|---|---|---|---|
| $\hat{\pi}_1$ | 0.338 | 0.380 | 0.185 | 0.155 | 0.294 | 0.451 |
| $\hat{\sigma}_0$ | 0.559 | 0.225 | 0.161 | 0.112 | 0.160 | 0.203 |
| $\hat{\sigma}_1$ | 1.306 | 0.685 | 0.688 | 0.286 | 0.597 | 0.689 |
| Region | PUTM | SNIG | SPCO | TCTX | THAL | WHMT |
| $\hat{\pi}_1$ | 0.225 | 0.201 | 0.347 | 0.340 | 0.153 | 0.382 |
| $\hat{\sigma}_0$ | 0.181 | 0.179 | 0.223 | 0.272 | 0.203 | 0.322 |
| $\hat{\sigma}_1$ | 0.792 | 0.642 | 0.572 | 0.735 | 0.774 | 0.932 |

$\hat{\pi}_1$ = probability that a transcript is DE; $\hat{\sigma}_0$ = SD of non-DE transcripts; $\hat{\sigma}_1$ = SD of DE transcripts.

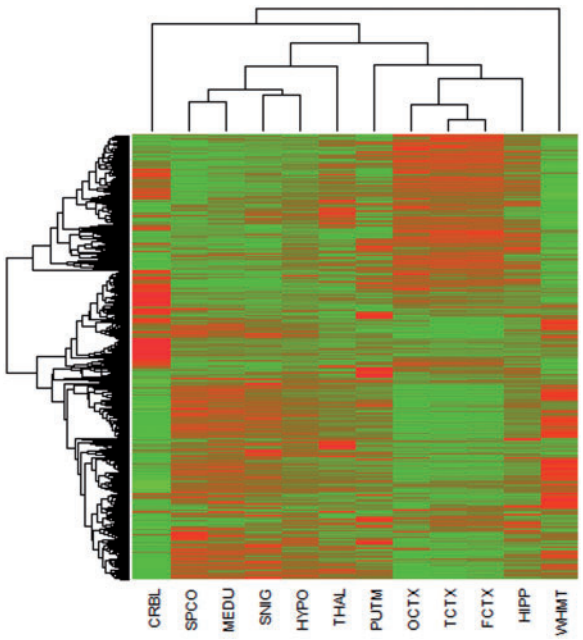


**Fig. 3.** Heatmap of estimates of Transcript.Region effects for transcripts identified as being DE in any of the 12 regions

the exception of CRBL, all transcripts in this cluster show a similar pattern, whether it be DE or non-DE in a particular region. Other clusters using this method are shown in Supplementary Figure S10.

Naturally, more complex methods for analysing this reduced dataset are appropriate. For example, the weighted gene co-expression network analysis method (Zhang and Horvath, 2005) may also be applied to identify network connectivity modules in specific regions or between all regions (Johnson *et al.*, 2009), as well as gene annotation and functional analysis such as DAVID to identify enriched biological themes such as functional and clustering classification using gene ontology (GO) (Huang da, *et al.*, 2009). Singh *et al.* (2013) demonstrated the use of DAVID, in conjunction with the current LMM/finite mixture model approach. Identification of over-represented transcription factor binding sites in targeted modules (Ho Sui *et al.*, 2005) and

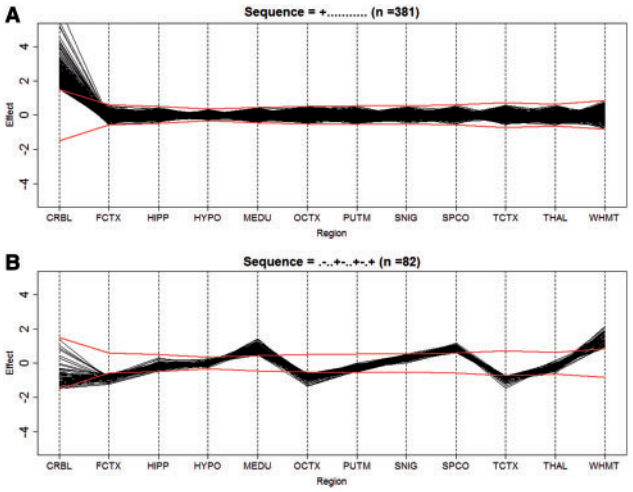


**Fig. 4.** Clusters of transcripts with similar pattern of being DE (up- and downregulated) on non-DE across the 12 brain regions. Top (**A**) shows the most common cluster (381 transcripts) that was upregulated in CRBL but non-DE in the other regions. Bottom (**B**) shows a more complex pattern of co-regional expression for a cluster (82) of transcripts

the use of a bioinformatics database, Encyclopaedia of DNA Elements (ENCODE) consortium, are also applicable.

## 6 DISCUSSION AND CONCLUSIONS

With ever-increasing volumes of data generated from expression arrays or RNA sequencing, there is a need to increase the efficiency and sensitivity in analysing large biological datasets. There is also a need to provide methods that are accessible to researchers without a strong bioinformatic and statistical background. The methods outlined here are aimed at addressing these goals: the method described should be able to be adapted to other situations, and some *R* code has been provided to assist with this (see Supplementary File Section S1). The use of LMMs as a means for analysing all expression data simultaneously is intuitively appealing, as there is much common information across the entire genome that can be used simultaneously to estimate genetic parameters and obtain predicted effects of individual transcripts. The analysis outlined here demonstrates how the standard LMM can be extended to include variance heterogeneity in the genetic effects but also in residual effects. Using recent developments in mixed model theory, we can incorporate a spatial correlation structure over the array.

The method as described here, and also as illustrated using the brain expression data, assumes that the expression data have been normalized. However, even in the absence of normalization, results are fairly similar, but with some indications of more transcripts being flagged as being DE (Supplementary File, Section S4). Although not suggesting that expression data should not be normalized, it does indicate that a low-level form of normalization would be expected to return similar results to the RMA method used here.

Another important consideration is the software used to fit the large-scale LMM to the expression data. For the analyses presented here, the commercial package ASReml-R has been used

because of the flexibility of model specification, in particular the ability to specify a heterogeneous variance model, all conducted within the *R* environment. However, for this dataset, the impact of fitting a homogeneous variance model was not found to have serious consequences (see Supplementary File, Section S1.5), allowing the freely available lme4 or nlme packages to be used.

Naturally, we are aware that using any analysis method to detect DE transcripts will have its own limitations. For example, there might be a functionally important DE transcript missed out as part of the detection above background process used to reduce some of the background 'noise'. Or alternatively, it may fail the test for being DE (threshold $\tau = 0.8$). However, such a transcript still may be detected using one of the clustering methods mentioned here, which only requires information on the effect size, and not the DE status. Note that the LMM/finite mixture model method proposed here produced a more targeted list of DE transcripts, the list becoming more accurate with increasing number of arrays. This is a point of contrast to the 'gene by gene' approach in limma (Smyth, 2004): with increasing sample size (number of arrays), the limma approach will flag the majority of transcripts as DE, but these will not necessarily be biologically important. Some comparisons of analysis using limma and the current method are included in Supplementary File Section S5.

In the method proposed here, there are two separate procedures, namely, (i) fitting the LMM and then (ii) fitting a finite mixed model to the BLUPs of these effects, separately in each state (brain region, in the current example). Ideally, the two processes would be combined, fitting a mixture model to capture DE/non-DE within the LMM. However, with a large number of states ($s$), the theoretical number of possible DE/non-DE patterns becomes immense ($2^s = 4096$), and fitting a mixture model with this many components would be infeasible. However, the LMM is reasonably robust against model misspecification. As an illustration, the correlation between the BLUPs of the Transcript.Region effects obtained from models with and without variance heterogeneity included was 0.98 (see Supplementary Table S2), despite the obvious between-region variance heterogeneity. Nonetheless, further research in combining these two stages is warranted.

The main endpoint of the methods outlined here is to produce a table of effects of gene transcripts in different states, the table being restricted to transcripts that are DE in any one of the $s$ states: following this, further analysis is of course required to identify clusters and the functional pathways using different advanced tools as mentioned in the previous section. This provides a way forward to obtain more reliable information about transcriptional regulation in the human CNS to understand the effect of genetic disorders targeting specific human CNS regions.

## ACKNOWLEDGEMENTS

UKBEC Brain expression data scientists are John Hardy, Michael E. Weale, Daniah Trabzuni, Mina Ryten, Colin Smith, Robert Walker and Adaikalavan Ramasamy. The authors thank AROS Applied Biotechnology AS company laboratories. They also specially thank Jane Ramsey, Geoff Scopes and Wilson Lew (Affymetrix) for their valuable input. (PMID: 21848658, 22723018, 23967090).

*Funding*: This study was performed using the UK Brain Expression Consortium (UKBEC), which was supported by the MRC through the MRC Sudden Death Brain Bank (C.S.), by a project grant (G0901254 to J.H. and M.W.). D.T. was supported by the King Faisal Specialist Hospital and Research Centre, Saudi Arabia. Computing facilities used at King's College London were supported by the National Institute for Health Research (NIHR) Biomedical Research Centre at King's College London.

*Conflict of Interest*: none declared.

## REFERENCES

Auer,P.L. *et al.* (2012) Differential expression—the next generation and beyond. *Brief. Funct. Genomics*, **11**, 57–62.

Butler,D. *et al.* (2007) *ASReml-R Reference Manual*. Queensland Department of Primary Industries and Fisheries, Brisbane.

Geschwind,D.H. and Konopka,G. (2009) Neuroscience in the era of functional genomics and systems biology. *Nature*, **461**, 908–915.

Ho Sui,S.J. *et al.* (2005) oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic Acids Res*, **33**, 3154–3164.

Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

Irizarry,R.A. *et al.* (2003) Exploration, normalization and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **2**, 249–264.

Ji,H. and Liu,X.S. (2010) Analyzing 'omics data using hierarchical models. *Nat. Biotechnol.*, **28**, 337–340.

Johnson,M.B. *et al.* (2009) Functional and evolutionary insights into human brain development through global transcriptome analysis. *Neuron*, **62**, 494–509.

Kerr,M.K. *et al.* (2000) Analysis of variance for gene expression microarray data. *J. Comp. Biol.*, **7**, 819–837.

Kumar,A. *et al.* (2013) Age-associated changes in gene expression in human brain and isolated neurons. *Neurobiol. Aging*, **34**, 1199–1209.

McLachlan,G.J. *et al.* (2002) A mixture model-based approach to the clustering of gene expression data. *Bioinformatics*, **18**, 413–422.

McLachlan,G.J. and Krishnan,T. (1997) *The EM Algorithm and Extensions*. Wiley, New York.

Millar,T. *et al.* (2007) Tissue and organ donation for research in forensic pathology: the MRC sudden death brain and tissue bank. *J. Pathol.*, **213**, 369–375.

Pinheiro,J.C. and Bates,D.M. (2000) *Mixed Effects Models in S and S-PLUS (Statistics and Computing)*. Springer, New York.

Singh,M. *et al.* (2013) Comparative transcriptome analyses reveal conserved and distinct mechanisms in ovine and bovine lactation. *Funct. Integr. Genomics*, **13**, 1–17.

Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Molec. Biol.*, **3**, Article 3.

Trabzuni,D. *et al.* (2011) Quality control parameters on a large dataset of regionally dissected human control brains for whole genome expression studies. *J. Neurochem.*, **119**, 275–282.

Trabzuni,D. *et al.* (2013) Widespread sex differences in gene expression and splicing in the adult human brain. *Nat. Commun.*, **4**, Article 2771.

Zhang,B. and Horvath,S. (2005) A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, **4**, Article 17.