Simulation identification of differential gene expression and connectivity in inflammation, adipogenesis and cancer

Antonio Reverter*, Aaron Ingham, Sigrid A. Lehnert, Siok-Hwee Tan, Yonghong Wang, Abhirami Ratnakumar and Brian P. Dalrymple

CSIRO Livestock Industries, Queensland Bioscience Precinct, 306 Carmody Road, Brisbane, Queensland 4067, Australia

Received on May 2, 2006; revised on June 13, 2006; accepted on July 11, 2006

Advance Access publication July 24, 2006

Associate Editor: Chris Stoeckert

ABSTRACT

Motivation: Biological differences between classes are reflected in transcriptional changes which in turn affect the levels by which essential genes are individually expressed and collectively connected. The purpose of this communication is to introduce an analytical procedure to simultaneously identify genes that are differentially expressed (DE) as well as differentially connected (DC) in two or more classes of interest.

Results: Our procedure is based on a two-step approach: First, mixed-model equations are applied to obtain the normalized expression levels of each gene in each class treatment. These normalized expressions form the basis to compute a measure of (possible) DE as well as the correlation structure existing among genes. Second, a two-component mixture of bi-variate distributions is fitted to identify the component that encapsulates those genes that are DE and/or DC. We demonstrate our approach using three distinct datasets including a human systemic inflammation oligonucleotide data; a spotted cDNA data dealing with bovine in vitro adipogenesis and SAGE database on cancerous and normal tissue samples.

Contact: Tony.Reverter-Gomez@csiro.au

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

While the identification of differentially expressed (DE) genes between two or more biological states remains the most common application of high-throughput gene expression technologies, other uses including disease characterization, toxicology, developmental biology, pathway mapping and mechanism of action are emerging.

In particular, the study of the basic principles involved in the dynamics of gene networks is gaining momentum (Ramoni et al., 2002; Lucombe et al., 2004; Remondini et al., 2005). However, given (1) the potentially large number of gene to gene interactions resulting from epistatic effects; (2) the inherent stochastic nature of the individual main effects of each gene in the network and (3) the fact that most gene knockouts results in little or no phenotypic change (Smith et al., 1997; Wagner, 2000; Gu et al., 2003); it follows that successful (directed to a desired biological outcome) testing of postulated hypotheses becomes a non-trivial task and/or too complex to allow for an exact solution. To date, no likelihood-based approaches exist to infer statistical significance of differential connectivity for a gene as a result of a given perturbation.

The purpose of this communication is to present an analytical procedure to simultaneously identify genes that are DE as well as genes that are differentially connected (DC) between two or more classes of interest. DC genes are defined as genes whose correlated expression pattern differs between classes. The ability to detect DC genes may allow the identification of genes belonging to DE cohorts. For example, a gene that has few connections in one state and a greatly increased number of connections when an experimental manipulation occurs, may be involved in a coordinated biological response even though its own expression may not differ between the two states.

Our procedure is based on a two-step approach: First, mixed-model equations are applied to obtain the normalized expression of each gene in each class treatment and levels within a class (e.g. time points). These normalized expressions form the basis to compute a measure of (possible) DE as well as the correlation structure existing among genes within each treatment class. Second, a two-component mixture of bi-variate normal distributions is fitted to identify the cluster, or component in the mixture, that encapsulates those genes that are simultaneously DE and DC.

We demonstrate our approach using three datasets. The first two correspond to time-course experiments, including the inflammation oligonucleotide microarray data of Calvano et al. (2005) and a spotted cDNA microarray data from our laboratory dealing with bovine in vitro adipogenesis (Tan et al., 2006). The last dataset corresponds to the National Cancer Institute, Cancer Genome Anatomy Project, SAGE database (Boon et al., 2002).

2 METHODS

2.1 Mixed-model equations and differential (co-) expression measures

Normalized measures of differential gene (co-)expression are obtained from the optimal method among those explored by Reverter et al. (2005a). In brief, a mixed-model is fitted which, in its most parsimonious state, contains the following components:

\[ y_{ijk} = \mu + A_i + G_j + V_{ij} + e_{ijk} \]  

where \( y_{ijk} \) represents the base-2 log-intensities from the \( r \)-th replicate of the \( i \)-th gene at the \( k \)-th array chip in the \( j \)-th variety (or class treatment); \( A_i \) is the main fixed effect of the \( i \)-th gene; \( G_j \) is the main random effect of the \( j \)-th variety; \( V_{ij} \) is the two-variate effect of the \( i \)-th gene and the \( j \)-th variety; and \( e_{ijk} \) is the residual error.
Differential gene expression and connectivity

gene (or array probe); VGij is the random interaction between the i-th gene and the j-th gene and s is a residual. In what follows, it is understood that the j-th gene incorporates both the main class treatment (e.g. healthy versus disease) as well as the sub-class level (e.g. time point). That is: j = 1, 2, . . . , n for the healthy class at the n time points; and j = n+1, n+2, . . . , 2n for the disease class also at the n time points.

For the random effects in model (1), standard stochastic assumptions are Gi ~ iid N(0, σ2G), VGij ~ iid N(0, σ2VG) and εij ~ iid N(0, σ2ε), where iid denotes independently and identically distributed and N denotes the normal distribution. Variance components are between genes (σ2G), between groups within variety (σ2VG) and within genes (σ2ε). Variance components were estimated by restricted (to zero error contrast) maximum likelihood [REML; see Searle et al. (1992) for detailed formulae].

To determine which genes are DE between the two class treatments, the following t-statistic can be computed for each gene in i:

\[
DE_i = \frac{1}{n} \left( \sum_{j=1}^{n} VG_{ij} - \sum_{j=n+1}^{2n} VG_{ij} \right)
\]

(2)

This definition of DE is likely to be conservative as it is based on overall variation in expression across all time points. However, it has the advantage of dealing with irregular time intervals compared with dynamic clustering methods based on autoregressive models (Ramoni et al., 2002), where the time points have to be evenly spaced. Nevertheless, other definitions of DE could be equally valid and implemented in the present algorithm. It should also be noted that although we assumed that an equal number of time points in both treatments exists, this is not a necessary condition for the normalization process in the mixed-model or the definition of DE described above or the development of a DC measurement described next.

For the network reconstruction, and motivated by the general applicability of the guilt-by-association heuristics (Wolfe et al., 2005), a correlation-based model is assumed. In these settings, the similarity between two genes is given by the Pearson correlation coefficient between the two gene expression-level series. A link (or connection) between two genes is established if the absolute value of the correlation exceeds a fixed threshold. For the present study, a threshold of 0.99 was used. The significance of the resulting network can be verified by the scale-free property of its connectivity distribution (Barabasi and Oltvai, 2004).

For each gene, a connectivity degree is defined as the total number of genes it is connected to. To determine which genes are DC between the two class treatments, the following statistic can be computed for each gene in i:

\[
DC_i = \log_{10} \left( \frac{C_1}{C_2} \right)
\]

(3)

where C1 and C2 indicate the number of connections for the i-th gene in the class treatments 1 and 2, respectively.

2.2 Two-component bi-variate mixture model

Model-based clustering via mixture of distributions has been proposed by a number of authors to identify DE genes (Efron et al., 2001; McLachlan et al., 2002, 2006; Reverte et al., 2004). Our goal here is to apply model-based cluster analysis to the values in DEi and DCi, and see which genes have relative expression and connectivity levels far away from the majority. To this end, and for each gene, the paired data points in DEi and DCi are assumed to be independent observations from a two-component mixture model with probability density function:

\[
f_{DEi, DCi} = \pi_0 \phi_0 \left( \frac{DE_i}{DC_i}; \mu_0, V_0 \right) + \pi_1 \phi_1 \left( \frac{DE_i}{DC_i}; \mu_1, V_1 \right)
\]

(4)

where

\[
\phi_0 \left( \frac{DE_i}{DC_i}; \mu_0, V_0 \right)
\]

(5)

denotes the empirical null bi-variate normal density with two-dimensional mean vector \( \mu_0 \) (not necessarily identity) and corresponding to non-DE and non-DC genes;

\[
\phi_1 \left( \frac{DE_i}{DC_i}; \mu_1, V_1 \right)
\]

(6)

denotes the bi-variate normal density function corresponding to genes that are DE and/or DC (DE/DC); finally, the mixing proportions \( \pi_0 \) and \( \pi_1 \) are constrained to be non-negative and sum to unity. In the present study, parameters of the mixture model (4) were estimated using the EMMLIX-GENE software (McLachlan et al., 2002). Once the parameters have been estimated, the posterior probability that the i-th gene is DE/DC [i.e. belongs to the second component of the mixture in (4)] is given by

\[
\tau_i = \frac{\pi_1 \phi_1 \left( \frac{DE_i}{DC_i}; \mu_1, V_1 \right)}{f \left( \frac{DE_i}{DC_i} \right)}
\]

(7)

Equivalently, the posterior probability that the i-th gene is not DE and/or DC (i.e. belongs to the first component of the mixture in (4)) is given by \( \tau_i = 1 - \tau_i \). For the present study, the i-th gene was assessed as DE and/or DC if its \( \tau_i \) was >0.99. Finally, an estimate of false discovery rate (FDR) is obtained from the arithmetic average of \( \tau_i \) from all DE/DC elements (McLachlan et al., 2006).

2.3 Example 1: inflammation data

For our first example, we consider the data from the study of Calvano et al. (2005) which examined gene expression profiles in whole blood leukocytes immediately before and at 2, 4, 6, 9 and 24 h after intravenous administration of bacterial lipopolysaccharide (LPS) endotoxin to four healthy human subjects. For the control (placebo) data, four additional subjects were studied under identical conditions but without LPS administration. Individual RNA samples were hybridized onto Hu133A and Hu133B oligonucleotide Gene-Chip arrays (Affymetrix, Santa Clara, CA). Because no records were available for one placebo individual at times 4 and 6 h, expression data from a total of 92 arrays were used in this study.

Data acquisition criteria were as follows: First, probe sets identified as absent on all arrays were removed; Second, expression signals <128 were deemed as inaccurate and not included; Finally, only probe sets providing signals at the six time points and for the two class treatments (infected and placebo) were included in the analyses. These criteria resulted in 510 672 expression signals from 12 192 probe sets that were jointly normalized by fitting a mixed-model equivalent to model (1) with the additional random effect of probe by individual person interaction.

2.4 Example 2: adipogenesis data

For our second example, we consider the data from Tan et al. (2006), which studied the changes in the expression of genes during adipogenic conversion of bovine bone marrow stromal cells. Cells that were treated with adipogenic stimulants and untreated cells were collected at six time points (0 h, 6 h, 24 h, 3 days, 6 days and 14 days post-differentiation), and gene transcript differences between treated and untreated cells determined from a total of 15 hybridizations using a muscle and fat cDNA microarray platform (Lehner et al., 2004).

A total of 236 671 intensity signals from 5418 genes (or cDNA clones) were background-corrected, base-2 log-transformed and analyzed by fitting a mixed model that included the fixed effect of comparison group (defined as the group of signals from the same array slide, printing block and dye channel), and the random effects of gene, gene × array, gene × dye, gene × time-point-treatment and residual.

2.5 Example 3: cancer SAGE data

For our final example, we consider the SAGE database (Boon et al. 2002; http://cgap.nci.nih.gov/SAGE) of 234 libraries across 41 tissues (23 normal and 18 cancerous). For each library, the abundance of all 10 nt SAGE tags...
was normalized to a value of 200,000 total tags and log-transformed. When more than one tag corresponded to the same gene, the reading from the most abundant tag was assigned to that gene. In total, there were 1,648,973 expression readings from 18,138 genes. For the present study, we removed genes not observed in the two conditions, normal and cancer (but not necessarily for the same tissue) and genes expressing in at least 20 of the 41 tissues for which at least 7 ‘organs’ are represented for the 2 conditions (thus pairwise gene correlations can be computed for each condition, cancer and normal). Finally, and in order to avoid high-leverage points, which could distort the correlation estimates, tissue-specific genes were removed. These were defined as genes having an expression in the given tissue greater than their average expression across all tissues plus three standard deviations.

The edited dataset contained 1,271,464 readings on 8,323 genes that were analyzed fitting a mixed-model that included the following five effects: The fixed effect of SAGE library to account for systematic non-genetic effects such as GC content bias (Margulies et al., 2001), sequencing errors (Stern et al., 2003), and the possibility of non-unique tags; the random effect of gene; the random gene x condition interaction (aimed at capturing DE genes between cancer and normal tissues); the random gene x tissue interaction (aimed at capturing differences due to specific gene-tissue combinations) and the random error.

3 RESULTS

3.1 Mixed-model equations and DE and DC

Goodness of fit, as measured by $R^2$ (the proportion of the total variation accounted for by the mixed-model), was 96.1, 97.9 and 52.8% for the inflammation, adipogenesis and cancer datasets, respectively. The smaller $R^2$ found for the cancer dataset was attributed to the discrete nature of the SAGE readings and to the heterogeneity of tissues being considered. The majority of the variation was attributed to the main effect of gene, while the proportions attributed to $V_{G(i)}$ were 6.1 and 2.1% for the inflammation and adipogenesis datasets, respectively. For the cancer dataset, the gene x condition and the gene x tissue interactions accounted for 0.2 and 9.9% of the total variation, respectively.

Table 1 presents summary statistics for the measures of DE and DC. As seen, DE measures in all datasets, as well as DC measures in the adipogenesis data, were well-centered around zero, while DC measurements in the inflammation data were skewed towards positive values, indicating an increase in the network connectivity from placebo to infected states. Differences in network connectivity are further detailed in Table 2.

For the inflammation data, a move from placebo to infected resulted in a 10-fold increase in the connectivity and clustering coefficient of the network. The clustering coefficient, computed from the number of existing connections divided by the total number of possible connections, captures how many of a gene’s neighbors are connected to each other. The connectivity distribution of the resulting networks is further explored in Supplementary Figure 3. A strong linear relationship in the log-log scale for the distribution of links was found indicating that the resulting networks were highly non-random with a significant enrichment of highly connected genes.

3.2 Mixture models and DE/DC elements

The following mixture models were estimated:

1. for the inflammation data,
$$f \left( \frac{DE}{DC} \right) = 0.579 \times N \left( \begin{pmatrix} -0.008 \\ 0.101 \end{pmatrix}, \begin{pmatrix} 0.022 & 0.004 \\ 0.004 & 0.180 \end{pmatrix} \right) + 0.421 \times N \left( \begin{pmatrix} 0.011 \\ 1.193 \end{pmatrix}, \begin{pmatrix} 0.111 & -0.026 \\ -0.026 & 0.205 \end{pmatrix} \right)$$

2. for the adipogenesis data,
$$f \left( \frac{DE}{DC} \right) = 0.639 \times N \left( \begin{pmatrix} -0.012 \\ 0.043 \end{pmatrix}, \begin{pmatrix} 0.003 & 0.001 \\ 0.001 & 0.089 \end{pmatrix} \right) + 0.361 \times N \left( \begin{pmatrix} 0.021 \\ 0.264 \end{pmatrix}, \begin{pmatrix} 0.023 & 0.004 \\ 0.004 & 0.435 \end{pmatrix} \right)$$

3. and for the cancer data,
$$f \left( \frac{DE}{DC} \right) = 0.584 \times N \left( \begin{pmatrix} 0.000 \\ 0.009 \end{pmatrix}, \begin{pmatrix} 0.001 & 0.000 \\ 0.000 & 0.269 \end{pmatrix} \right) + 0.416 \times N \left( \begin{pmatrix} -0.000 \\ -0.267 \end{pmatrix}, \begin{pmatrix} 0.000 & 0.000 \\ 0.000 & 0.322 \end{pmatrix} \right)$$

In the three datasets and mixture models, the second component (or cluster) with smaller mixing proportion, non-zero means and higher variances had a higher probability to encapsulate genes with extreme DE/DC values. Figure 1 shows the two-component normal mixture models weighted by the mixing proportions and for each dataset. The empirical histograms of the observed DE and DC values are superimposed in Figure 1 to illustrate both the (possible) deviations from normality of the original values as well as the goodness of fit of the mixture models. Supplementary Figure 1 illustrates the estimates of FDR and the proportion of selected genes that would result as a function of $\tau$.

Table 2. Summary statistics for the connections in the two networks obtained in each of the three datasets

<table>
<thead>
<tr>
<th></th>
<th>Inflammation (12,192 probes)</th>
<th>Adipogenesis (5418 probes)</th>
<th>Cancer (8323 genes)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Infected Placebo</td>
<td>Treated Untreated</td>
<td>Cancer Normal</td>
</tr>
<tr>
<td>N</td>
<td>126,543</td>
<td>1668</td>
<td>10,484</td>
</tr>
<tr>
<td>Mean</td>
<td>20.7</td>
<td>2.1</td>
<td>3.9</td>
</tr>
<tr>
<td>Min.</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Max.</td>
<td>244</td>
<td>13</td>
<td>57</td>
</tr>
<tr>
<td>C(k)*</td>
<td>1.7E−3</td>
<td>1.7E−4</td>
<td>7.1E−4</td>
</tr>
</tbody>
</table>

*C(k) = Clustering coefficient.*
When the criterion of \( t_1 > 99\% \) was applied, a total of 2048 (FDR < 0.20%), 496 (FDR < 0.17%) and 429 (FDR < 0.19%) probes were found to be DE/DC in the inflammation, adipogenesis and cancer datasets, respectively. The sensitivity, as measured by the minimal concentration of transcript that was reliably measured, and estimated using the approach of Reverter et al. (2005b) was estimated at 21.4, 112.2 and 56.2 transcripts per million in the inflammation, adipogenesis and cancer experiments, respectively. Supplementary Figure 2 illustrates the sensitivity analysis for each experiment based on the distribution of all and DE/DC probes. These sensitivities compare favorably with other microarray and tag-based expression profiling experiments (Reverter et al., 2005b).

The location of DE/DC elements at \( t_1 > 99\% \) are highlighted in the scatter of Figure 2. These elements are deemed to enhance the differences between the networks for each class treatment. Hence, the way in which they interact with each other deserves further scrutiny. Figure 3 illustrates the histograms of the correlation coefficients between each pairwise DE/DC probes and for each network. For the treated class of the inflammation dataset, the number of coefficients close to +1 or −1 increased significantly, indicating that many of the 2048 genes that were affected over time by the LPS infection became either strongly correlated or anticorrelated. These findings are in complete agreement with those from Remondini et al. (2005) using rat cell lines and targeting c-Myc-activated genes.

The same comparison applied to the adipogenesis data resulted in more extreme correlations being observed in the sample treated with the adipogenic stimulant. In contrast, the distributions of the correlation coefficients among the 429 genes found to be DE/DC in the SAGE data did not vary that much between the cancer and the normal states. This result was attributed to the SAGE dataset originating from not so well defined system (with a variety of tissues and libraries being represented) as in the previous two experiments.

In addition, by means of the gene × tissue interaction, the analysis of the cancer data allowed for the computation of gene
3.3 Relevance of results from the inflammation data

Calvano et al. (2005) have reported the sequential activation of genes during the 24 h post endotoxin challenge using differential expression of genes between time points as a measure of significance. The authors treated the infected and control datasets separately. While the expression of 5093 probe sets, representing 3714 unique genes, changed significantly in response to endotoxin, no temporal changes of the probe sets reached the significance level (FDR < 0.1% according to the text, and FDR < 1% according to the Supplementary Methods 1 available from the online version of their manuscript).

Table 3. Measures of differential expression (DE) and differential connection (DC) for genes involved in the TLR signaling pathway

<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>DE</th>
<th>DC</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCL4</td>
<td>0.721</td>
<td>0.221</td>
</tr>
<tr>
<td>CCL5</td>
<td>-0.952</td>
<td>1.061</td>
</tr>
<tr>
<td>FOS</td>
<td>0.472</td>
<td>1.079</td>
</tr>
<tr>
<td>IKBKG</td>
<td>0.092</td>
<td>1.732</td>
</tr>
<tr>
<td>IL1B</td>
<td>0.669</td>
<td>1.580</td>
</tr>
<tr>
<td>IRF3</td>
<td>-0.049</td>
<td>1.613</td>
</tr>
<tr>
<td>MAP2K6</td>
<td>0.499</td>
<td>1.176</td>
</tr>
<tr>
<td>MAP3K71P2</td>
<td>0.640</td>
<td>1.322</td>
</tr>
<tr>
<td>MAPK14</td>
<td>0.416</td>
<td>1.452</td>
</tr>
<tr>
<td>MYD88</td>
<td>0.336</td>
<td>1.477</td>
</tr>
<tr>
<td>NFKB1</td>
<td>0.553</td>
<td>1.000</td>
</tr>
<tr>
<td>NFKB2</td>
<td>0.711</td>
<td>0.653</td>
</tr>
<tr>
<td>PIK3C2A</td>
<td>-0.282</td>
<td>1.602</td>
</tr>
<tr>
<td>PIK3R1</td>
<td>-0.255</td>
<td>1.568</td>
</tr>
<tr>
<td>PIK3R5</td>
<td>0.368</td>
<td>1.255</td>
</tr>
<tr>
<td>RAC2</td>
<td>0.627</td>
<td>1.230</td>
</tr>
<tr>
<td>TLR1</td>
<td>0.332</td>
<td>1.653</td>
</tr>
<tr>
<td>TLR4</td>
<td>0.538</td>
<td>1.518</td>
</tr>
</tbody>
</table>

Here, in the belief that the addition of a DC value would enhance our ability to detect members of biologically relevant pathways, we have focused on identifying genes that are both DE/DC at any time point, treating both data sets simultaneously. As a result, a set of 2048 probes representing 1589 unique genes were found to be DE/DC (FDR < 0.2%). These genes were processed using the Pathway-Express (Khatri et al., 2005) facility of Onto-Tools (Draghici et al., 2003). Three biological pathways were found to be over-represented in our selected set of genes including toll-like receptor (TLR) signaling with 18 genes (P < 1.0E−10), MAPK signaling with 37 genes (P < 1.9E−9) and cytokine–cytokine receptor interactions with 23 genes (P < 7.35E−7).

Listed in Table 3 are the DE/DC values for 18 genes from the TLR signaling pathway. Importantly, an analysis based only on DE would have failed to capture many of these genes, including IKBKG, IRF3, MYD88, PIK3C2A, PIK3R1, PIK3R5 and TLR1, for which |DEi| < 0.4. Similarly, this method identified an additional 15 genes in the MAPK signaling pathway and 11 in the Cytokine–cytokine receptor pathway. All 11 additional genes involved in cytokine response (CCR1, IFNGR1, IL10RB, IL13RA1, IL17R, IL18RAP, IL1R1, IL4R, TNFRSF7, TNFRSF10C, TNFRSF25) function as receptors, suggesting that they differ little in expression following LPS challenge however, their connectivity changes dramatically. It is hypothesized that the addition of a connectivity value might be particularly useful in identifying receptors.

The three identified pathways have been previously implicated in the response to LPS challenge. The TLR signaling pathway is involved in the primary detection of LPS (Poltorak et al., 1998) and once stimulated initiates a series of intracellular signaling cascades, including the MAPK signaling pathway (Kawai et al., 1999; Zhang et al., 1999). Activation of these pathways leads to the activation of transcriptional regulators and culminates in increased expression of a number of genes that mediate the early stages of the inflammation response.
response including members of the cytokine–cytokine receptor pathway identified here (Bowie and O’Neill, 2000).

Cluster analysis showed the 0 and 24 h profiles are most similar. There is ample evidence indicating that an inflammatory response is already initiated shortly after exposure to LPS (Brownstein et al., 2006 and Talwar et al., 2006). Our result highlights the rapid nature of the response and the ability of the immune system to return to homeostasis, and hence gene expression to return to baseline values, once the challenge has been effectively dealt with.

An in-depth comparison of results revealed that among the 2048 DE/DC probes identified here, 1356 of them (or 66.2%) appeared in the 5093 DE list of Calvano et al. (2005). Both lists of significant probes represented 4280 unique genes of which 608 and 2065 appeared only in ours and in the previously reported list, respectively. MAPK signaling and Cytokine–cytokine receptor were among the significant pathways (hypergeometric \( P < 0.01 \)) over-represented in our list only. Significant pathways over-represented in the previous list only included Phosphatidylinositol signaling system, Cell cycle, Focal adhesion and Apoptosis.

3.4 Relevance of results from the adipogenesis data

The mean number of connections between genes in the treated and untreated classes only changed slightly in this experiment, compared with the other two examples (Table 2 and Fig. 2, middle panel). This could be owing to the smaller number of genes on the array or it may reflect the fact that the difference between the two biological classes was not as dramatic as in the other examples.

Figure 4 presents the hierarchical cluster analysis, obtained using the PermutMatrix software (Caraux and Pinloche, 2005), of selected genes found to be DE/DC. Four clusters are clearly visible in Figure 4. A number of genes were identified that showed large increases in connectivity, but were not identified in an earlier study that looked at DE genes alone (Tan et al., 2006). Importantly, the study of Tan et al. (2006) assessed DE by exploring the changes in expression of each gene at each time point, relative to its expression at time zero (hence, five DE measures were tested for each gene). The new genes identified here belonged to the functional classes of protein synthesis (RPLP0, RPL8, RPL18, RPS12), molecular chaperones (HSPA8, HSPB8, DNAJ4) and extracellular matrix (CILP). Other genes from these classes had all been reported to be DE, but three others (BAT1, IFRD1, STAT3), with a role in transcriptional regulation during bovine adipogenesis, are identified here for the first time. Interestingly, STAT3 has previously been found to be associated with human and murine adipogenesis (Deng et al., 2000; Harp et al., 2001). While this transcription regulator has not been previously found to be DE, the analysis performed here has suggested its possible role in bovine adipogenesis. Again, as in the case of the inflammation data, this finding highlights the importance of the incorporation of DC measurements for detecting specific types of genes.

3.5 Relevance of results from the cancer data

The identity of the top 50 DE/DC genes in cancerous tissues is given in Supplementary Table 1, and Supplementary Figure 5 presents the heat map obtained from the normalized expression across the 16 tissues for which data from both conditions, cancer and normal, were available. As expected, none of the genes were uniformly DE/DC in all tissues. Instead, their ranking reflects a combination of the magnitude of DE and the breadth of tissue representation. While many of the under- and most of the over-expressed genes have been reported in one or more of the many studies in the recent literature (see last column of Supplementary Table 1), most studies have concentrated on a single type of cancer.

In addition, the simultaneous identification of DE and DC genes has allowed the detection of many key processes and some of the key effector proteins involved in neoplasia. The protein products of a set of DE/DC genes underlying neoplasia is schematically...
represented in relation to their cellular localization and functional assignment in Figure 5. The largest group of genes encodes proteins involved in the deposition and degradation of extracellular matrix (ECM). Smaller groups encode proteins involved in control of cell cycle progression, lipid metabolism and the glutatione/glutathione/oxidative stress pathway. Genes encoding proteins that stimulate the cell cycle are up-regulated, whilst genes that encode proteins that inhibit cell cycle progression are down-regulated. This is consistent with the fundamental definition of neoplasia: as new or abnormal growth of tissue. Components of the actin cytoskeleton, intermediate filaments and microtubules are involved in cell shape, adhesion and motility. In the case of the actin cytoskeleton the two down-regulated genes encode proteins involved in anchoring actin polymers to membranes, whilst the two up-regulated genes encode proteins that disrupt actin polymerization. TMSB10 binds oligomers to inhibit polymerization, and CAPG caps actin polymers, leading to shorter unanchored chains. Thus, all four of the observed changes have a similar outcome, reduced actin cytoskeleton, but via different mechanisms.

However, expression of the two up-regulated genes is not correlated, and neither is the expression of the two down-regulated genes. There is also a non-significant correlation between any of the up- or down-regulated genes in this group. Thus, whilst the loss of actin cytoskeleton is clearly associated with cancer, several different mechanisms appear to be able to deliver equivalent outcomes. A similar lack of correlated gene expression is observed for the genes encoding two other nuclear proteins TOP2B and YPELS. TOP2B is a well-validated target of anticancer drugs (Gao et al., 2003).

In contrast, highly correlated changes in gene expression are observed where the protein products are known to bind to each other in complexes, examples include: the collagens; the two keratins KRT18 and KRT19; and S100A8 and S100A9. Interestingly, the expression of the complement component C1R is also highly positively correlated with most of the collagen subunits suggesting that in neoplasia, C1R is intimately involved in the ECM. The expression of genes encoding proteins involved in the degradation of the ECM, such as PLAU and HSPCAL3 was not correlated with ECM components, or with each other.

Matrix remodeling is a feature of many mechanistic models of cancer progression or metastasis, where aspects of cellular adhesion, proliferation and migration are related to turnover of specific matrix macromolecules. In this regard, over-expression and over-connectivity of collagens I, III and VI seem consistent with the literature (Choi et al., 2005). The appearance of collagen XV amongst the down-regulated genes probably reflects the non-structural, and regulatory roles attributed to this and other minor collagens in basement membranes (Amenta et al., 2000).

4 DISCUSSION

As the number of molecular profiling experiments grows and the need to extract more information from these complex data grows as well, it is of interest to develop methods capable of teasing apart the changes in the transcriptional network that result between different biological conditions and that go beyond the standard methods commonly applied today based solely on differential expression. In this paper, we formulate an analytical approach to simultaneously identify genes that are DE and DC between two (or more) classes of interest. Our method incorporates the power of mixed-model equations in normalizing gene expression data by accommodating covariance structures in a rather general manner, with the sensitivity of mixture models in identifying a cluster (or component in the mixture) that encapsulates the extreme values. We show that this approach is able to be successfully applied to a variety of experimental layouts from time series to cross-sectional expression datasets.

For simplicity, measures of DE used in this study were based on overall variation in expression across all biological conditions. Notably, our approach is not limited by the definition of DE. In any given experiment, alternate measures of DE could be equally biologically meaningful. Similarly, to reconstruct the co-expression network, a threshold of 0.99 for the correlation coefficient was used to establish that two genes are connected. While equally high thresholds have been used (e.g., 0.98 in Remondini et al., 2005; 0.85 for a FDR < 0.5% in Reverter et al., 2005a; the top 0.5% of gene pairs for 8542 genes and a FDR < 0.1% in Choi et al., 2005), the scale-free property of the reconstructed network is likely to hold with a lower threshold.

However, when considering the strong dependencies observed in gene expression data, the use of a high threshold is warranted in studies with few classes (or time points) because lowering the threshold leads to the inclusion of random connections and an exponential degree distribution with a smaller clustering coefficient (van Noort et al., 2003). Ultimately, a likelihood-based approach that takes into account the degrees of freedom used in computing the correlation coefficients, along with the study of the connectivity distribution of the resulting network, will be critical in the identification of the optimal threshold.

In spite of these constraints, namely simplicity of DE measurements and a high threshold for correlations, our method has captured a whole set of interesting genes that were missed in previous studies.

Fig. 5. Location of proteins and processes for a selection of the up- and down-regulated cancer genes. Up-regulated genes are shown in capitals and down-regulated genes are shown in capital italics. Protein–protein interactions are shown by double-headed arrows, stimulation of processes are shown “+” and inhibition is shown “−”. Degradation actions are shown by a single-headed arrow. Indirect action is shown by dotted lines.
that focused on DE only. This was particularly evident for the case of receptors and regulators which, while differing little in their expression across biological conditions, their connectivity changed dramatically. This finding is consistent with the Trace-back algorithm proposed by Luscombe et al. (2004) to reconstruct the dynamics of regulatory networks. In brief, the condition of DE is applied to target genes only, while transcription factors are marked as being present in a condition if they have a sufficiently high expression level.

Similarly, our definition of DC as the log of the ratio of the connectivity of a gene between two conditions has a lot of intuitive appeal. However, we acknowledge that the development of validation procedures for DC gene discovery will be challenging, compared with DE gene validation. The DE status of a gene can be confirmed by measuring its expression using a quantitative method such as qRT–PCR. For DC genes, large datasets of correlated gene expression are required to make the measurements, which are hard to generate by any other method than microarray profiling.

The DC status of a gene would be validated most effectively by a demonstration that the gene in question does indeed play a part in a biological process, despite not dramatically altering its expression. These demonstrations may be achieved, for example, by RNAi-mediated gene knockout (Chatterjee-Kishore and Miller, 2005) in a cellular model system. Another possible avenue for biological validation would be to demonstrate regulatory connections between DC genes and the cohort of DE genes they are correlated to. In some instances, assays that explore DNA and protein binding interactions, such as ChIP-on-chip technology (Odom et al., 2004; Blais and Dynlacht, 2005), may be able to demonstrate changes in connectivity.

Finally, robustness to the model assumption is critical for the control of FDR. Significance analysis of microarrays (SAM; Tusher et al., 2000) is a powerful approach; however, its non-parametric nature does not take care of the correlation and heteroskedastic structure within and across arrays. The combined approach of mixed-model with mixtures of distributions proposed here overcomes these drawbacks; no significance level needs to be specified before identifying DE/DC genes, FDR is controlled from here overcomes these drawbacks; no significance level needs to be specified before identifying DE/DC genes, FDR is controlled from

**ACKNOWLEDGEMENTS**

The authors would like to thank Prof. Nicholas Sangster, Dr Evgeny Glazov and an anonymous reviewer who by suggestions and advises contributed to the significant improvement of this manuscript.

**Conflict of Interest:** none declared.

**REFERENCES**


