BIOINFORMATION open access

Supplementary material:

Microarray gene expression experiment

The sheep experiment consisted of twenty time-mated (synchronized with progesterone sponges and then artificially inseminated) pregnant Merino ewes that were allocated to 4 equally sized treatment groups receiving daily intramuscular injections of a control or metyrapone between day 55 and 65 of gestation. Ewes were killed in humane manner and midside foetal skin samples (2cm) were collected from the 16 single pregnancies at either day 60 or 67 of gestation. RNA was extracted and hybridised to Affymetrix GeneChip® Genome Arrays.

Microarray Data Quality Control and Exploratory Analyses

Microarray data was explored and analysed using R 2.13.0 and BioConductor. Many quality control (QC) plots were explored **(Figure 1)** including using methods available in the following BioConductor packages: *affyPLM*, *affy*, *simpleaffy*, *affycoretools*, *made4* and *vsn*. Many of the QC plots were performed on both raw and normalised data. Data were normalised using gcRMA background correction, quantile normalisation and expression values computed using median polish. The identification of differentially expressed (DE) genes was achieved using the *limma* package while *GOEAST* was used to identify gene ontology (GO) terms enriched in a list of DE genes. We identified possible abundance of genes linked to muscle related GO terms (due to contamination with muscle tissue during biopsy of fetal skin tissues) and hence were removed from the skin network analyses.

No RNA or hybridisation (**Figure 1 top left**) quality issues were detected. PCA analysis of the normalised data showed a clear separation of two groups of samples on PC1 **(Figure 1 top middle)** and was believed to be linked to the possible contamination issue. GO enrichment analysis of the 334 significantly DE genes identified by a contrast between samples thought to be contaminated and not **(Figure 1 top right),** revealed a high abundance of genes linked to muscle related GO terms (**Figure 1 bottom**).

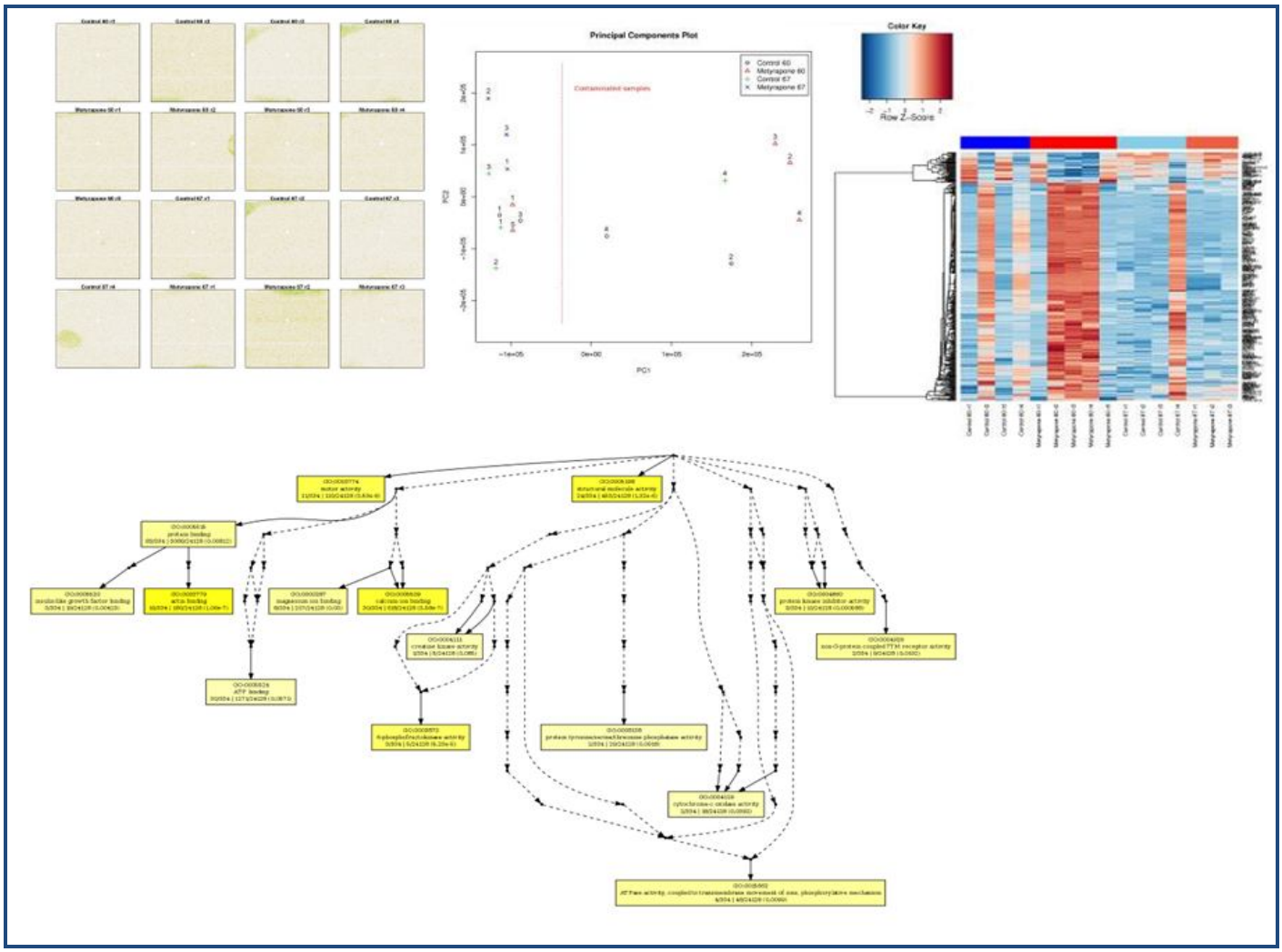

Figure 1: Top left: Pseudo array images showing the weights from the probe level model fitting procedure. **Top middle:** PCA analysis of arrays with separation on PC1 due to contamination. **Top right:** Heat plot of DE contamination genes. Bottom: GO enriched terms and their relationships found in the DE contamination genes

Weighted Gene Co-expression Network Analysis (WGCNA)

Details of the WGCNA method and algorithms are thoroughly discussed in the original paper of Zhang and Horvath **[7]**, an R package is also available for performing these analyses **[11].** Since its first publication **[7]**, the WGCNA method has been refined, standardized and now widely used in the construction of gene co-expression networks including our own previous work **[5, 12].** Hence, we only briefly describe the method here. As with most co-expression networks, the Pearson correlation coefficients (ρij) calculated from the expression values for all pairs (i and j) of transcripts are used to define the edge weights. Typically, a hard threshold would result in an adjacency value (aij) between a pair of nodes as either 1 or 0 as:

$$a_{i,j} = signum(\rho_{i,j}, \theta) \equiv \begin{cases} 1 & if \ \rho_{i,j} \geq \theta \\ 0 & if \ \rho_{i,j} < \theta \end{cases}$$

where, θ is a hard threshold (with a range 0 to 1).

Rather than applying a ''hard'' threshold to define an unweighted adjacency matrix (network), WGCNA applies the power adjacency function to the absolute Pearson correlation matrix to defining a weighted adjacency matrix as:

$$a_{i,j} = power(\rho_{i,j}, \beta) \equiv |\rho_{i,j}|^{\beta}$$

The value of the power function exponent (β) is chosen using the scale-free topology criterion, which is biologically motivated **[7]**. A high β maintains high adjacencies but pushes lower adjacencies towards zero. A linear regression model fitting index R2 between log10 p(k) and log10(k), where k is the measure of connectivity, is used to determine how well a network fits the scale-free topology criterion. There is a trade-off between maximizing model fit (R2) and maintaining a high mean number of connections.

PCIT

PCIT is a method used to identify spurious edges for removal and is a data driven approach. Full details of the PCIT algorithm are provided in Reverter and Chan [**13**], so we only briefly describe it here, and an R package implementing the algorithm is also available [**14**].

For any given edge in a gene co-expression network it's weight, derived from a Pearson correlation coefficient, may only be present due to high correlations with a third node in the network. For example, let us consider a trio of genes (A, B and C). If there is a strong correlation between AC and BC, it follows that there is likely to be a strong correlation between AB **(Figure 2)**. This confounding of direct and indirect associations leads to a spurious edge forming between AB and is likely to cause problems when it comes to identifying and interpreting gene modules.

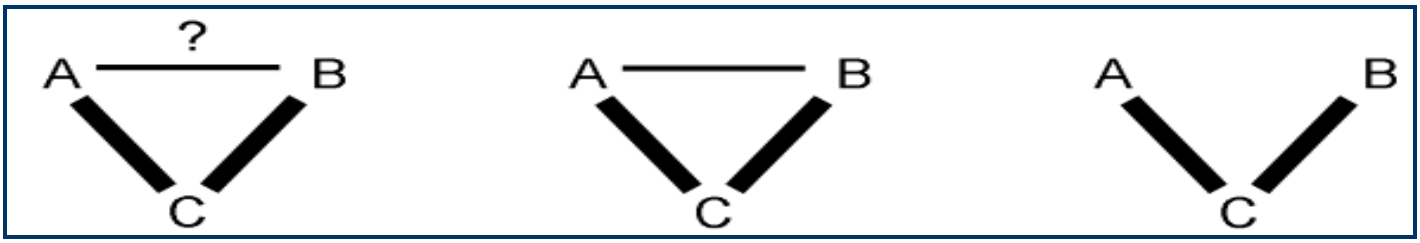

Figure 2: Correlations between a trio of genes A, B and C. The strength of correlation between pairs of genes is indicated by line width. PCIT determines if the correlation between AB is independent of the strong correlations between AC and BC **(left)**. If the correlation between AB is independent of C, the edge is retained **(middle)**. If the edge is found to be dependent on C, the edge is removed **(right)**.

PCIT uses partial correlation and information theory approaches to identify and remove such edges, thus only edges are retained if they are there on their own merit. The algorithm first builds partial correlations for every trio of genes A, B and C; the three first-order partial correlation coefficients are computed by:

$$r_{AB,C} = \frac{r_{AB} - r_{AC}.r_{BC}}{\sqrt{(1-r_{AC}^2)(1-r_{BC}^2)}}$$ and like wise for $r_{AC,B}$ and $r_{BC,A}$

The partial correlation coefficient between *A* and *B* given *C* (here denoted by $r_{AB.C}$) indicates the strength of the linear relationship between *A* and *B* that is independent of (uncorrelated with) *C*.

In the context of network reconstruction, a connection between genes *A* and *B* is discarded if

$$|r_{AB}| \leq |\varepsilon r_{AC}| \ and \ |r_{AB}| \leq |\varepsilon r_{BC}|$$

where ε is the local threshold and is the average of ratios of 3 partial to direct correlations. Otherwise, the association is defined as significant, and a connection between the pair of genes is established.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 8(18): 855-861 (2012)

Because PCIT is a completely data-driven approach, it is deemed to be a soft-thresholding approach to edge removal. The network generated following PCIT edge deletion has several attractive features: 1) many edges are removed resulting in a much sparser network which is easier to analyse; 2) the ability to treat remaining edges as unweighted, thus opening up these networks to unweighted network analysis algorithms; 3) the knowledge that all remaining edges are present in their own right i.e. independent.

Highly differentially ranked (HDR) nodes

We defined highly differentially ranked (HDR) nodes based on the following formulation. First, the connectivity (k) of the i^{th} gene (k_i) is the sum of the adjacencies between the i^{th} gene and all other genes in the network:

$$k_i = \sum_{j=1}^{n} a_{ij}$$

The connectivities of nodes cannot be easily compared between the networks due to the use of different algorithms and different coefficients of β in the WGCNA derived networks. Therefore we compare the ranks of the node connectivities (coded in ascending order as 1,2,3,…) to identify those which are highly differentially ranked (HDR) between WGCNA and PCIT derived networks.

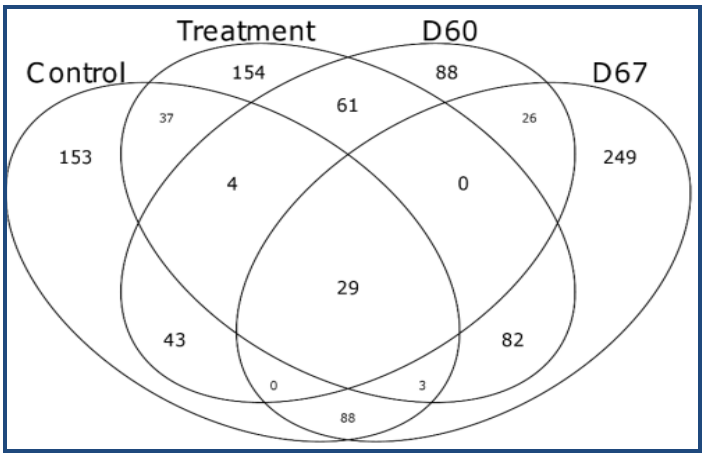

Figure 3: Venn diagram of the highly differentially ranked (HDR) nodes identified in the Control, Treatment, D60 and D67 networks. A total of 1,017 HDR nodes were identified across 1 or more of these networks.

Table 1: Microarray experimental design showing treatment groups of 16 pregnant merino ewes in drug challenge experiment

Group	Treatment	Treatment period (day of gestation)	Sample collected (day of gestation)	Number of single pregnancies
1	Control	55-59	60	4
2	Metyrapone	55-59	60	5
3	Control	55-65	67	4
4	Metyrapone	55-65	67	3

Table 2: GOEAST analyses of "greenyellow" module from WGCNA analyses. It consisted of 267 genes including those identified through traditional differential gene expression analysis in limma, showing biologically relevant genes for wool / hair development

GOID	Definition	No. of genes	P-value
GO:0051056	regulation of small GTPase mediated signal transduction	9	0.004
GO:0007389	pattern specification process	3	0.018
GO:0010646	regulation of cell communication	11	0.018
GO:0001763	morphogenesis of a branching structure	2	0.028
GO:0048754	branching morphogenesis of a tube	2	0.028
GO:0030509	BMP signaling pathway	1	0.051
GO:0001569	patterning of blood vessels	1	0.051
GO:0009880	embryonic pattern specification	1	0.051
GO:0035239	tube morphogenesis	2	0.056
GO:0009799	determination of symmetry	1	0.070