

Microarray Quality Control: Identifying and Dealing with less than Perfect Data

Nathan S. Watson-Haigh¹
nathan.watson-haigh@csiro.au

Haja N. Kadarmideen¹
haja.kadarmideen@csiro.au

Greg Natrass²
natrass.greg@saugov.sa.gov.au

Melanie McDowall³
melanie.mcdowall@adelaide.edu.au

¹ CSIRO Livestock Industries, J M Rendel Laboratory, Rockhampton, QLD 4701, Australia.

² South Australian Research & Development Institute (SARDI), Livestock and Farming Systems, Roseworthy 5371, Australia

³ Discipline of Agricultural and Animal Science, The University of Adelaide, Roseworthy Campus, Roseworthy 5371 Australia

Abstract

It is important that meta-data, observations and issues relating to samples, including their handling, is recorded accurately by the experimentalist. This not only allows for effective microarray analysis, but for the easy identification and explanation of systematic differences in gene expression. We show here, how a simple, but quite easily overlooked observation made at the time of sample preparation is used to identify sample contamination using a suite of quality control plots which should be used routinely in microarray analyses. Including contamination in the model of gene expression we identified 334 differentially expressed genes due to the contamination which showed enrichment for muscle related GO terms.

Keywords: sheep, bovine, microarray, quality control, contamination, GO enrichment

1 Introduction

Microarray experiments were performed on foetal skin samples from lambs exposed to metyrapone (an inhibitor of cortisol synthesis) in an attempt to elucidate the genes that are responsible for initiating primary wool follicles (between days 55-65 of gestation). However, at the time of sample collection there were concerns over possible contamination of the skin samples by underlying tissues.

Using several publicly available tools in BioConductor, we demonstrate how microarray data quality control should be an integral part and prerequisite of any microarray data analysis. In addition we highlight the need for close collaboration between the experimentalist and the bioinformatician responsible for data analysis in order that potential problems are identified quickly and solutions formulated.

2 Method and Results

Ewes were treated daily with metyrapone or control from day 55 of gestation until they were humanely killed by captive bolt and exsanguination at either day 60 or 67 of gestation in order to collect the foetal skin samples. RNA was extracted from each sample and hybridised to Affymetrix Bovine Genome arrays. Table 1 shows the treatment groups with the number of biological replicates.

Table 1: Details of the arrayed foetal skin samples collected from control and metyrapone treated ewes.

Group	Treatment	Sample Taken (Gestation day)	Biological Replicates
1	Control	60	4
2	Metyrapone	60	5
3	Control	67	4
4	Metyrapone	67	3

Microarray data was explored and analysed using R 2.7.0 and BioConductor 2.2. Many quality control (QC) plots were explored (Fig. 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

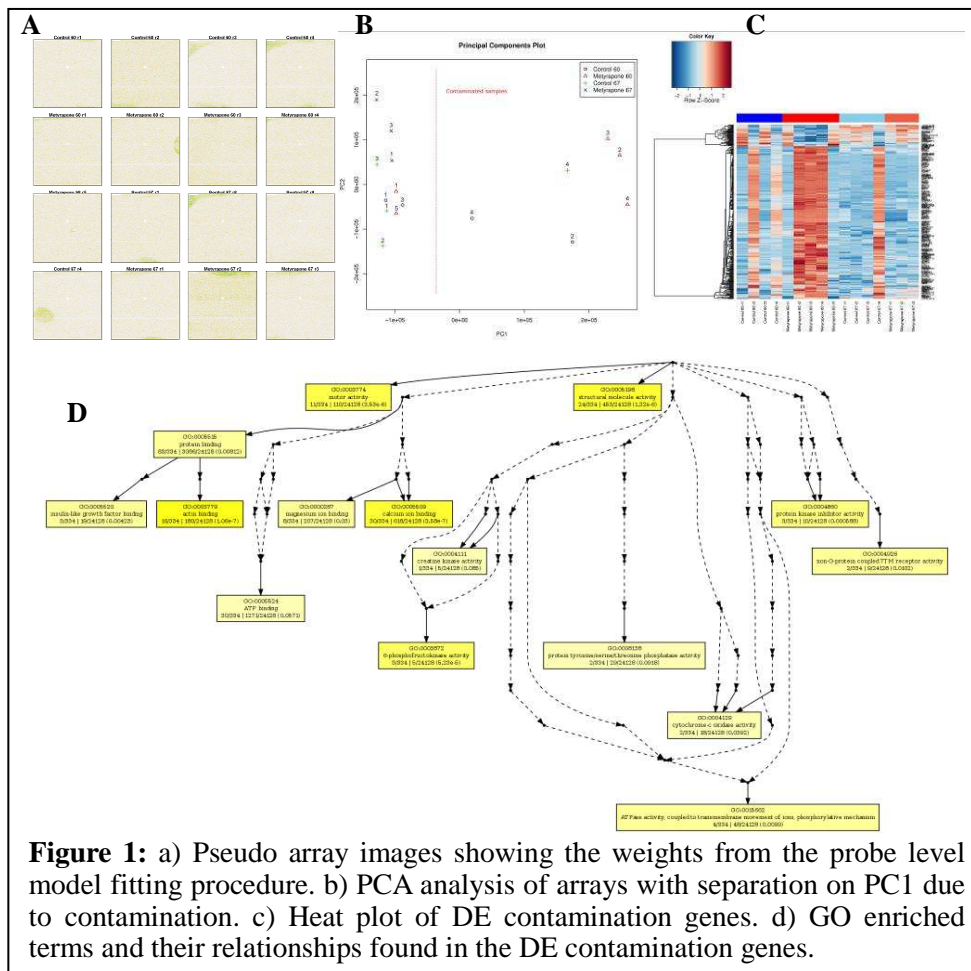


Figure 1: a) Pseudo array images showing the weights from the probe level model fitting procedure. b) PCA analysis of arrays with separation on PC1 due to contamination. c) Heat plot of DE contamination genes. d) GO enriched terms and their relationships found in the DE contamination genes.

was achieved using the limma package while GOEAST [1] was used to identify gene ontology (GO) terms enriched in a list of DE genes.

No RNA or hybridisation (Fig. 1a) quality issues were detected. PCA analysis of the normalised data showed a clear separation of two groups of samples on PC1 (Fig. 1b) 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 (Fig. 1c), revealed a high abundance of genes linked to muscle related GO terms (Fig. 1d).

3 Discussions

Through the use of existing tools it was possible to identify the skin samples that were thought, *a priori*, to be contaminated with underlying tissues. This highlights the need for experimentalists to note observations and keep good records of meta-data that may later help identify, explain and provide sufficient reason to remove problem samples encountered during the data analysis. The inclusion of such meta-data in the model for gene expression will allow the more accurate detection of DE genes of interest.

References

[1] Q. Zheng and X. Wang, "GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis," *Nucl. Acids Res.*, vol. 36, Jul. 2008, pp. W358-363.