

Effects of Different Normalisation and Analysis Procedures on Illumina Gene Expression Microarray Data

Dan Johnstone^{1,2}
Daniel.Johnstone@newcastle.edu.au

Carlos Riveros^{2,3}
Carlos.Riveros@newcastle.edu.au

Pablo Moscato^{2,3}
Pablo.Moscato@newcastle.edu.au

Liz Milward^{1,2}
Liz.Milward@newcastle.edu.au

¹ School of Biomedical Sciences MSB and ² Priority Research Centre for Bioinformatics, Biomarkers and Informatics Based Medicine (CIBM), University of Newcastle, Callaghan NSW 2308, Australia

³ School of Computer Engineering, University of Newcastle, Callaghan NSW 2308, Australia

Keywords: microarray, normalisation

1 Introduction

The investigation of microarray expression data for large numbers of genes poses a number of challenges. Firstly, there is the challenge of transforming raw data to make it comparable across arrays and reduce technical variation while conserving true biological effects. Technical variation (eg cRNA loading, scanning and hybridisation inconsistency) can impact substantially on raw signal intensity. Normalisation has been examined for some platforms but little information is available for Illumina users. Background correction has been examined by Dunning et al. [1] but there is little literature on subsequent normalization. Secondly, statistical analysis of the normalised data poses a challenge due to the large number of simultaneous comparisons. The challenge lies in reducing false positive results while avoiding false negative results. Conservative multiple testing corrections (eg Bonferroni) can miss real changes.

Here we investigate the effects of applying different normalisation and analysis techniques to a microarray dataset obtained from a study of the effects of high dietary iron on heart gene expression in mice. We also compare our results to a similar study in the literature to illustrate variation due to experimental design.

2 Method and Results

Total RNA was extracted from hearts of biological replicates ($n \geq 3$) of 9 week-old AKR mice fed either normal chow or a high-iron diet (normal chow supplemented with 2% carbonyl iron for 3 weeks). Differential expression was assessed using Illumina Sentrix MouseRef-8 (v1.1) BeadChip arrays. As outlined in Fig. 1, raw data were normalised using BeadStudio v3 (Illumina), which has five normalisation options: background, average, rank invariant, cubic spline or no normalisation. The output from each of these options was then either considered directly or subjected to additional normalisation with GeneSpring GX 7.3 software (Agilent) (set values < 0.01 to 0.01 , normalise to chip median, normalise to gene median). Statistical analysis using one-way ANOVA identified probes detecting transcripts with differential expression ($p < 0.05$). The different normalisation methods showed relatively little concordance in the probes identified in this way. Cubic spline normalisation generated probe sets that had the highest average overlap with other normalisation method (935 genes, 44%), irrespective of whether or not subsequent normalization with GeneSpring was applied, and was applied prior to all subsequent analysis. A total of 314 probes were common to the probe lists from all five different BeadStudio normalisation methods.

To investigate variation in results due to different data analysis methods, three different approaches were used to calculate differential expression following cubic spline normalisation in BeadStudio. These were: BeadStudio differential expression (error model Illumina Custom, difference score $> |13|$), GeneSpring (one-way ANOVA, $p < 0.05$) and an in-house algorithm that discretises the data and selects probes based on their ability to discriminate the experimental test and control groups [2]. Probes that gave signals below a raw intensity threshold of 20, potentially attributable to background noise, were eliminated.

Analysis using BeadStudio differential expression gave a list of 908 probes, GeneSpring gave 1299 probes and the in-house algorithm gave 1823 probes. Of the total probes identified by each of Beadstudio, GeneSpring and the in-house algorithm, 122 (13%), 160 (12%) and 484 (27%) respectively did not overlap

with the probes identified by any other method. The in-house algorithm yielded the highest number of both concordant and non-concordant probes. A total of 524 probes had differential expression by all three methods. Of these 524 probes, 148 were also identified by all different BeadStudio normalisation algorithms.

The Panther classification system (<http://www.pantherdb.org>) was used to identify over- or under-represented pathways and ontologies. Analysis of the 148 probes showed an overrepresentation of genes involved in the biological processes of muscle development and muscle contraction ($p=0.006$ for both, Bonferroni corrected), both of strong probable biological relevance in the context of heart disease. There was also an enrichment of cytoskeletal protein genes ($p=0.006$).

Results of the GeneSpring analysis, the more concordant of the two commercially available programs, were compared with a similar published study by Rodriguez *et al.* [3] using Illumina Mouse-6 chips to investigate differential expression in heart of 16-18 week-old C57BL/6 mice fed a diet supplemented with 2% carbonyl iron for 6 weeks. The data were normalised with Inforsense KDE v2.0.4 (London, UK) using quantile normalisation. The study reported all fold changes of 40% or more (i.e. 1.4 or more and -1.4 or less) following normalisation but did not take into account probability values and statistical comparisons.

Filtering our GeneSpring output data by the same fold change used by Rodriguez and colleagues [3] identified 351 upregulated and 413 downregulated genes. This was in contrast to the gene sets generated by Rodriguez and colleagues (35 upregulated and 40 downregulated genes) which were much smaller and had only 6 upregulated and 7 downregulated genes in common with our dataset. As for our analysis, genes involved in muscle contraction were significantly overrepresented in the Rodriguez gene set, however the latter identified only 4 muscle contraction-related genes in total compared to the 15 muscle contraction-related genes found to alter in our analysis. No other pathway identified by the Rodriguez gene set (protein folding; immunity & defense; blood clotting) matched those identified in our analysis (lipids & fatty acids; intracellular protein traffic; cell structure & motility; developmental processes).

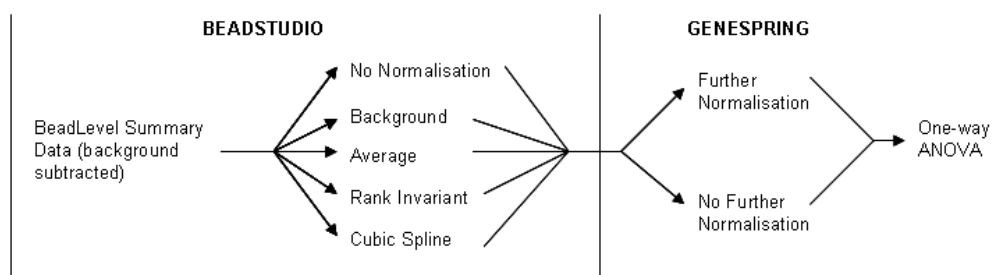


Figure 1: Schematic illustrating the different normalisation procedures

3 Discussion

Choice of normalisation and data analysis algorithms, including whether or not statistical testing is applied and how, has a profound effect on which genes are identified as having differential expression. While different normalization and analysis approaches can often be valuable in revealing different data features, our results suggest that when only one or two approaches are used, there is the danger of bias and large numbers of false positives or negatives. As shown here, this can have important consequences for subsequent ontology and pathway investigations. We propose that data from microarray experiments should be subjected to a range of normalisation and analysis procedures and then comparisons made between these to obtain more robust gene lists. We also advise caution when extrapolating from single published reports without comparing alternative normalisation and analyses.

References

- [1] Dunning, M., Barbosa-Morais, N., Lynch, A., Tavare, S., Ritchie, M., Statistical issues in the analysis of Illumina data, *BMC Bioinformatics*, 9:85, 2008.
- [2] Berretta, R., Costa, W., Moscato, P., Combinatorial optimization models for finding genetic signatures from gene expression datasets, *Methods Mol Biol*, 453: 363-77, 2008
- [3] Rodriguez, A., Hilvo, M., Kytomaki, L., Fleming, R., Britton, R., Bacon, B., Parkkila, S., Effects of iron loading on muscle: genome-wide mRNA expression profiling in the mouse, *BMC Genomics*, 8:379, 2007.