BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 22 2007, pages 3032–3038
doi:10.1093/bioinformatics/btm448

Gene expression

The high-level similarity of some disparate gene expression measures

Nandini Raghavan[1], An M.I.M. De Bondt[2], Willem Talloen[3], Dieder Moechars[2], Hinrich W.H. Göhlmann[2] and Dhammika Amaratunga[1,*]

[1]Nonclinical Biostatistics, Johnson & Johnson Pharmaceutical Research & Development LLC, Raritan, NJ 08869, USA, [2]Functional genomics and [3]Nonclinical Biostatistics, Johnson & Johnson Pharmaceutical Research & Development, A Division of Janssen Pharmaceutica, B-2340, Beerse, Belgium

Received on May 11, 2007; revised on August 24, 2007; accepted on August 25, 2007

Advance Access publication September 24, 2007

Associate Editor: Martin Bishop

ABSTRACT

Probe-level data from Affymetrix GeneChips can be summarized in many ways to produce probe-set level gene expression measures (GEMs). Disturbingly, the different approaches not only generate quite different measures but they could also yield very different analysis results. Here, we explore the question of how much the analysis results really do differ, first at the gene level, then at the biological process level. We demonstrate that, even though the gene level results may not necessarily match each other particularly well, as long as there is reasonably strong differentiation between the groups in the data, the various GEMs do in fact produce results that are similar to one another at the biological process level. Not only that the results are biologically relevant. As the extent of differentiation drops, the degree of concurrence weakens, although the biological relevance of findings at the biological process level may yet remain.

Contact: damaratu@prdus.jnj.com

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

1 INTRODUCTION

Complex biological processes are driven by the actions of series of interacting genes. Therefore, characterizing the roles of individual genes or gene combinations in a biological process requires an ability to elucidate genome-wide gene expression patterns. This has been made possible by the evolution of whole genome microarrays that are capable of simultaneously measuring the transcription levels of all genes in a cell. Of these, Affymetrix's GeneChips are one of the most popular.

Affymetrix GeneChips consist of 25mer oligonucleotide probes. Each probe is designed to interrogate how much of the transcript sequence complementary to its DNA sequence is present in a sample. A gene is represented by a set of 11–20 such probe pairs called a probe set. Each probe pair consists of a perfect match (PM) probe for its target sequence and a paired mismatch (MM) probe with the same 25-base sequence as the PM except for a single change in the middle nucleotide.

The presence of multiple probes in a probe set has perplexed users of this technology as it is unclear as to how best to summarize the set of probe-level values to produce a single expression measure for the probe set. A number of methods have been proposed, including Affymetrix's own Average Difference and MAS5 (Affymetrix, 2002), which are single array methods, and a plethora of multi-array methods; the most popular of which are dChip (Li and Wong, 2001), RMA (Irizarry *et al.*, 2003) and GC-RMA (Wu *et al.*, 2004). The Affycomp website (Cope *et al.*, 2004; Irizarry *et al.*, 2005) lists several dozen more.

It is clearly disconcerting to researchers that the various summarization techniques produce different gene expression measures (GEMs) (Cope *et al.*, 2004); and therefore different analysis results at least as far as individual genes are concerned (Shedden *et al.*, 2005). That this would happen is not surprising as the methods postulate different models and use different methods for model fitting. Spike-in experiments, in which known concentrations of mRNA have been added to the hybridization cocktail, suggest that RMA and GC-RMA are superior in terms of sensitivity to MAS5 and dChip while being slightly inferior in terms of specificity (see the Affycomp website and Supplementary Fig. 1 which is based on the information there). In fact, these spike-in data have been the focus of many assessments of the performance of summarization techniques at the individual gene level. Yet, these data and assessments are very simplistic in terms of correlational structure and number of genes affected, whereas the challenges in actual applications are largely due to the correlational complexity and the high dimensionality of the data. They miss the point stated in the first sentence above: that biological processes are driven by the actions of series of interacting genes.

Thus, it would be intriguing to ascertain whether the different summarization schemes would, in fact, produce high-level results that are similar to one another even though the results at the probe-set level might differ. This article explores this question in the context of an experiment involving knockout mice.

*To whom correspondence should be addressed.

2 METHODS

2.1 Experiment

It is imperative, when studying a question such as this, to have at hand an experiment for which there is available some knowledge as to the biological processes affected. To this end, we used an experiment involving knockout mice. As an added bonus, gene expression was studied at two stages of development of these mice: a newborn stage, when there is little differentiation between knockout and wild-type mice, and a later stage when there is more. This enabled us to compare the GEMs with differing amounts of differentiation in the data.

More on the experiment is described as follows. Defects in the metabolism of sialic acid are known to be responsible for the so-called sialic acid storage diseases. These are autosomal recessive neurodegenerative disorders that may present as a severe infantile form (ISSD or infantile sialic acid storage disease) or a slowly progressive adult form (Salla disease). Both forms of sialic acid storage disease are caused by mutations in Slc17A5, which encodes the protein sialin, named such because of its relation to sialic acid storage diseases. Mutant proteins have been shown to be impaired in the natural function of sialin, which acts as a transporter to export sialic acid out of the lysosome. Genetic deletion of sialin function in a mouse was reported before to mimic ISSD (Moechars *et al*., 2005; Verheijen *et al*., 1999), with excretion of large amounts of free sialic acid in the urine, an accumulation of sialic acid in lysosomes, hypo-myelination, growth and neuromotor retardation, coarse facial features, fast deterioration and early death. To gain insight into the pathogenic mechanism resulting in this severe phenotype, newborn mice were studied for their gene expression in the brain at two time points: (1) a post-natal time point (day 0) that preceded the occurrence of obvious phenotypic traits but impacts organic acid transport on a molecular level and (2) a post-natal time point (day 18) where impaired sialic acid transport led to observable morphological alterations such as defects in myelination (Moechars *et al*., 2005). The experiment was conducted using Slc17A5 knockout mice as a model for ISSD with the overall objective of identifying the earliest genes and biological processes affected at the transcript level in young knockout mice before the disease is manifested.

To perform the experiment, RNA samples from total brain were derived from newborn and 18-day-old mice for each of two groups: Slc17A5 knockout ('KO') and wild type ('WT'). There were six biological samples in each group. Microarray experiments were performed on the RNA samples using Affymetrix Mouse430_2 GeneChips. The PM and MM values were recorded and GEMs were calculated using four different summarization procedures: MAS5, RMA, GCRMA and dChip. The calculations were done in R (R Development Core Team, 2006) using the BioConductor (Gentleman *et al*., 2004) version 1.9 packages affy and gcrma with the default settings. Each set of GEMs was quantile-normalized (Amaratunga and Cabrera, 2004) and the MAS5 and dChip values were also log-transformed.

2.2 Topset concurrence

The goal now is to measure agreement among the various GEMs with respect to their ability to detect differential expression between the two groups: WT and KO. One way to do this would be to compare *t*-test *P*-values. However, the more common way to utilize microarray data is to use them to rank the genes according to *P*-value and to flag for further study the 'topset', the set of most significant genes. Therefore, it seems more pertinent to judge similarity by generating topsets using the various GEMs and then seeing how similar these topsets are to one another. We do this using 'Topset Concurrence' as the basic construct for assessing agreement.

Topset concurrence scores are computed as follows. In the following, g indexes the probe sets ($g = 1, \ldots, G$), j indexes the GEMs (the possible values for j are MAS5, RMA, GCRMA, dChip) and k is a given integer (such as $k = 10$ or $k = 100$).

(1) Calculate a *P*-value for each probe set (e.g. using the two-sample *t*-test).

(2) For the jth GEM, determine its 'topset', the set S_j of probe sets with the k smallest *P*-values; in other words, the topset S_j is the list of probe sets deemed the k-most significant using the jth GEM.

(3) For each probe-set i in S_j, count the number of other topsets (S_l such that $l \neq j$) that also contains probe-set i. Call this number A_{ij}. Clearly A_{ij} will be 0, 1, 2 or 3.

(4) The topset concurrence TC(k, j) for GEM j is defined as the mean of the A_{ij} over i.

A high topset concurrence (i.e. TC near 3) would imply that the four GEMs give almost identical topsets. A low topset concurrence (i.e. TC below 2) would imply that the different GEMs produce different topsets. To see how the degree of agreement varies with k, concurrence scores can be calculated for different values of k and the resulting topset concurrence scores TC(k, j) can be plotted against k to produce a 'topset concurrence plot' (or 'TC plot'). Clearly, this plot will asymptote towards 3 as $k \rightarrow G$.

The analysis thus far is for individual genes. Next, it is pertinent to assess the degree of agreement at the biological process level. In other words, if the genes are categorized according to the biological process they are involved in, do the gene categories called significant by the different GEMs agree? To study this, the probe sets were categorized according to their Gene Ontology (GO) Biological Process annotations (Gene Ontology Consortium, 2000). The format of this data is such that genes are assigned only to the most specific level of the branch they belong to in the hierarchical GO tree; they are not assigned to nodes higher up in the tree which are less specific. Because of the structure of the GO hierarchy, if a gene belongs to a certain node, it also belongs to its parent's node as well as to all its ancestor's nodes. This gene propagation from a specific level to a more general level was performed using a specially developed Java program. A total of 1335 GO terms were ultimately represented in the data. When a gene is represented by multiple probe sets on the array, the one with the smallest *P*-value was used to represent the gene. After this, for each GO term, its MLP statistic = mean(−log *P*-value) was computed (Pavlidis *et al*., 2003, 2004; Raghavan *et al*., 2006) and its significance determined using the permutation procedure described in Raghavan *et al* (2006). The resulting *P*-values were ranked and a topset concurrence score was evaluated for each method using the above-mentioned procedure.

These analyses were done on both the newborn and the 18-day datasets. In addition, to establish a null baseline, an artificial dataset, 'Scramb' was created by permuting the samples of the newborn data.

2.3 Variations

Many choices are possible at every stage of a microarray data analysis starting with the choice of summarization procedure and continuing on to the choice of chip definition file (e.g. Affymetrix CDF or EntrezGene CDF), gene-level test statistic (e.g. *t*-test or Limma), test statistic for finding differences at the biological process level (e.g. MLP or hypergeometric) and annotation database [e.g. GO or (KEGG)]. Our study incorporated several of these variations to determine whether any of them would substantially affect the degree of concurrence.

Topset concurrence scores can be derived with any test. We also generated them using *P*-values from Limma (Smyth *et al*., 2004), a popular alternative to the *t*-test that employs a hierarchical model to

borrow strength across genes to render the analysis more stable for experiments with small sample sizes.

It has been argued that alternative probe-set annotations (CDFs), e.g. the alternative probe mappings based on the genes contained in the EntrezGene database, produced better and more interpretable results (Dai *et al*., 2005; Sandberg and Larsson, 2007). Therefore, we repeated the above-mentioned procedures with these alternative CDFs.

Over-representation analysis using a hypergeometric test (Hosack *et al*., 2003) is a popular method for identifying significant GO terms. We also repeated the above-mentioned procedures with this test.

In certain instances, it is more useful to categorize genes based on which pathway they are involved in rather than by which GO term they belong to. Therefore, we also repeated the above-mentioned procedures with genes grouped according to their KEGG pathways (Ogata *et al*., 1999).

3 RESULTS

Spectral maps (Wouters *et al*., 2003) of the data (Fig. 1) show clear separation between the two groups (WT and KO) for 18-day mice, marginal separation for newborns and none for the Scramb dataset, thus illustrating the differing amounts of information regarding the differences between KO and WT in the three datasets.

Next, individual gene *t*-tests and topset concurrences were computed. The results are shown in Figure 2: the left column shows the False Discovery Rate (FDR) (Benjamini and Hochberg, 1995) for the top k genes versus k, the center column shows the TC plots for the data at the individual gene level and the right column shows the TC plots for the data at the GO term level. For each dataset, all GO terms which are in the top 15 for any of the GEMs with TC = 3 across the GEMs are listed in Table 1.

With the 18-day data, where there is a strong separation between the two groups, FDR < 0.001 for all methods for $k < 58$ (Fig. 2a), indicating, for instance, that the gene-level topset of size 50 for any method will contain mostly genes for which there is strong evidence of differential expression. Yet, topset concurrence at the gene level is modest (Fig. 2b), meaning that, for example, were one to select the top 50 genes, the different GEMs would produce different topsets with only a little overlap. On the other hand, the GO terms display remarkably strong topset concurrence (Fig. 2c), meaning that, for example, all the GEMs produce the same list of top 10 GO terms. The maximum topset concurrence score of 3 is achieved in the range of 7–13 GO terms. All 13 TC = 3 GO terms (i.e. those listed in Table 1) correspond to processes expected to be affected (Supplementary Table 1 shows a short list of independent biological processes that were expected to be affected). Results for the EntrezGene annotations (Supplementary Fig. S2) are similar; in fact, topset concurrence there appears even slightly stronger. Topset concurrences using Limma instead of a *t*-test are also similarly strong (Supplementary Figs S3 and S4).

With the newborn data, where there is only weak separation between the two groups, FDR is generally quite high (except FDR < 0.1 for $k < 15$ for dChip) (Fig. 2d), indicating that gene-level topsets will include many genes with only a little evidence of differential expression. Thus, the genes show poor topset concurrence (Fig. 2e). So do the GO terms (Fig. 2f); there is only one TC = 3 GO term compared to 13 with the 18-day data.

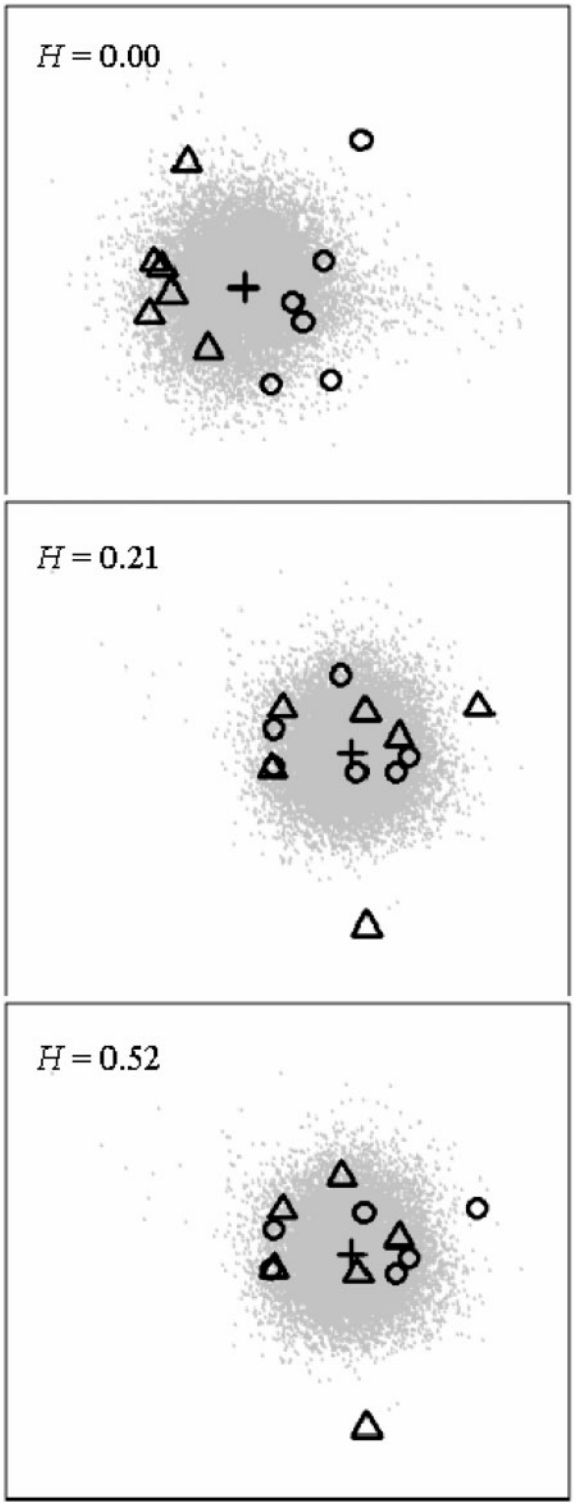

Fig. 1. Spectral maps of the 18-day data (top), Newborn data (middle) and Scramb data (bottom), with the WT mice denoted by circles, the KO mice by triangles and the individual genes by dots. The number H on each is the permutation-based P-value for a Hotelling's test on the top 100 genes selected via MAS5 *t*-tests; this number is a rough measure of the separation between the two groups; the lower the value of H the greater the separation.

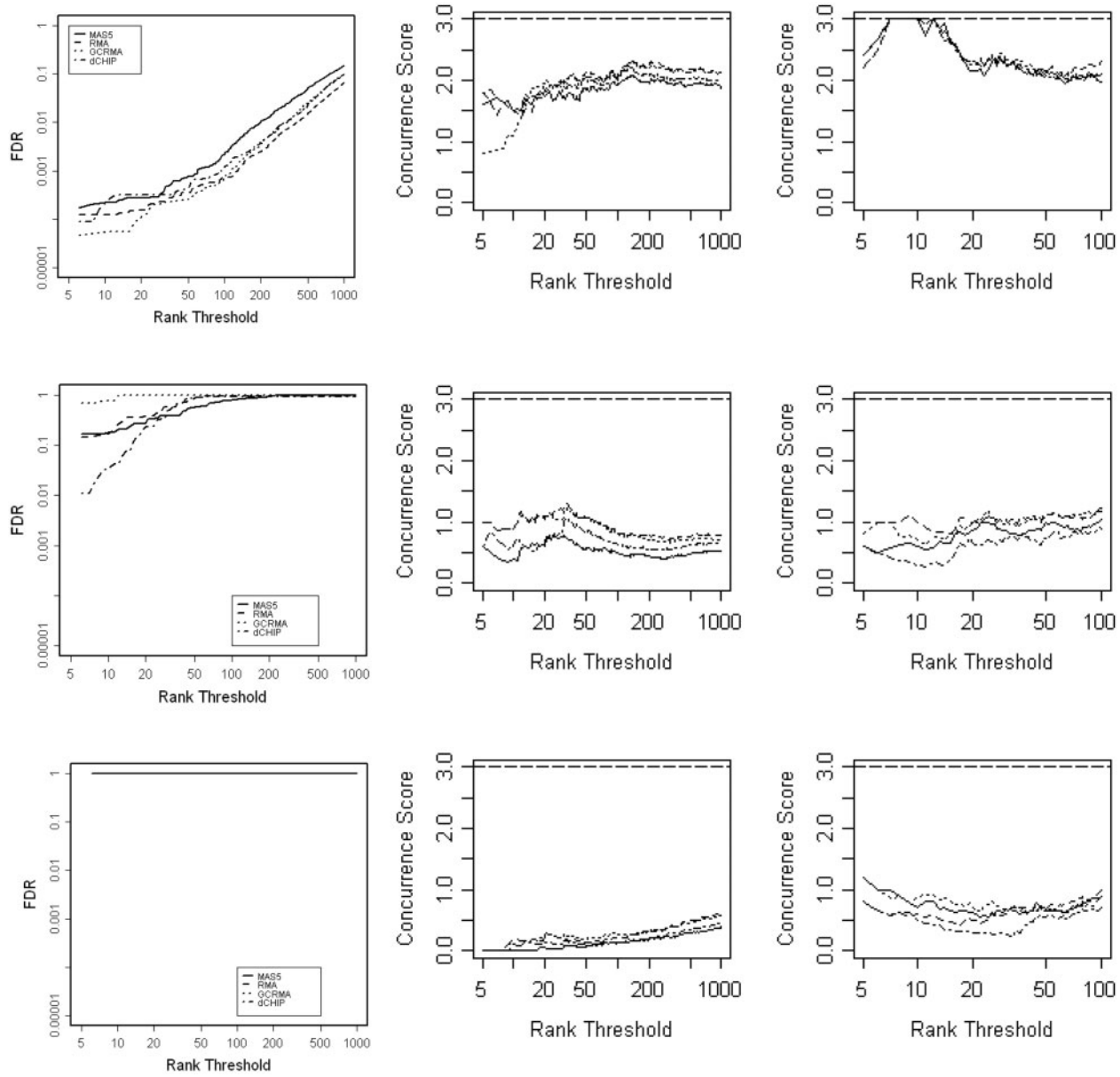

Fig. 2. Results of the comparisons across groups. The top row shows the results for the 18-day data, the middle row for the newborn data and the bottom row for the scramb data. The left column shows FDR for the top k genes versus k, the middle column shows the TC plots for assessing the concurrence of the GEMs at the individual gene level, and the right column shows the TC plots for assessing the concurrence of the GEMs at the GO term level. For this display, annotations provided by Affymetrix and t-tests were used.

However, importantly, this GO term (monocarboxylic acid transport which includes the knocked-out gene Slc17A5) and many of the other GO terms appearing in the different GO term topsets are associated with biological processes known to be affected, indicating a certain degree of reliability in the accumulated results at the biological process level even though the separation between groups is small.

Not surprisingly, the Scramb dataset, which has no true separation between the groups, shows extremely high FDR levels (Fig. 2g), hardly any concurrence at either level (Figs 2h and i) and neither the top genes nor the top GO terms have any relevance to the situation at hand.

When the genes were categorized according to their KEGG pathways (Fig. 3), the pattern of topset concurrence at 18 days is similar to that with GO terms but weaker. However, this was expected as KEGG pathways do not fit as well with the known phenotype as GO's biological processes in the context of this experiment. Deletion of Slc17A5 was found to result in dysmyelinization due to impaired oligodendrocyte functionality, which translates well in GO terms such as 'nerve

Table 1. For each dataset, the top 15 GO terms with TC = 3

18-Day	GO Term description
1508	Regulation of action potential
6694	Steroid biosynthesis
6695	Cholesterol biosynthesis
6911	Phagocytosis, engulfment
7272	Ionic insulation of neurons by glial cells
8203	Cholesterol metabolism
8366	Nerve ensheathment
16 125	Sterol metabolism
16 126	Sterol biosynthesis
42 551	Neuron maturation
42 552	Myelination
42 553	Cellular nerve ensheathment
48 469	Cell maturation
Newborn	**GO Term description**
15 718	Monocarboxylic acid transport
Scramb	**GO Term description**
6376	mRNA splice site selection

ensheathment', 'myelination', 'cellular nerve ensheathment', 'ionic insulation of neurons by glial cells' and 'regulation of action potential' and GO terms that relate to lipid metabolism such as 'cholesterol biosynthesis', 'cholesterol metabolism', 'sterol biosynthesis', 'membrane lipid biosynthesis' and 'membrane lipid metabolism' as oligodendrocytes are the major source of lipid biosynthesis in the brain (listed in Supplementary Table 1). Deletion of Slc17A5, a transporter of sialic acid would not immediately affect the biochemical pathways that constitute the KEGG database; therefore it is not unexpected that the concurrence with KEGG pathways is weaker. There is hardly any topset concurrence with the newborn data. However, again this was unsurprising as at birth no phenotypic difference was observed between the mutant mice and their wild-type littermates.

When the MLP approach is replaced by the hypergeometric test (Fig. 4), the pattern of topset concurrence is somewhat similar to that seen in Figure 2. In fact, of the 13 GO terms in Table 1, 10 are also identified as being among the top 15 GO terms with TC = 3 using the hypergeometric approach. However, the concurrence is weaker. A notable example is GO 6911, phagocytosis engulfment, a process that is expected to be significant because the biologically observed phenotype of defective myelination involves a degradation of the oligodendrocytes and subsequent debris clearance by phagocytosis. It appears in Table 1, but is only picked up by MAS5 as being among the top 15 GO terms when using the hypergeometric approach. The concurrence is weaker because the hypergeometric approach itself is weaker in the sense that it tends to have less statistical power than the MLP approach (Raghavan *et al*., 2006). This underscores the importance of employing adequately powered statistical tests for data analysis.

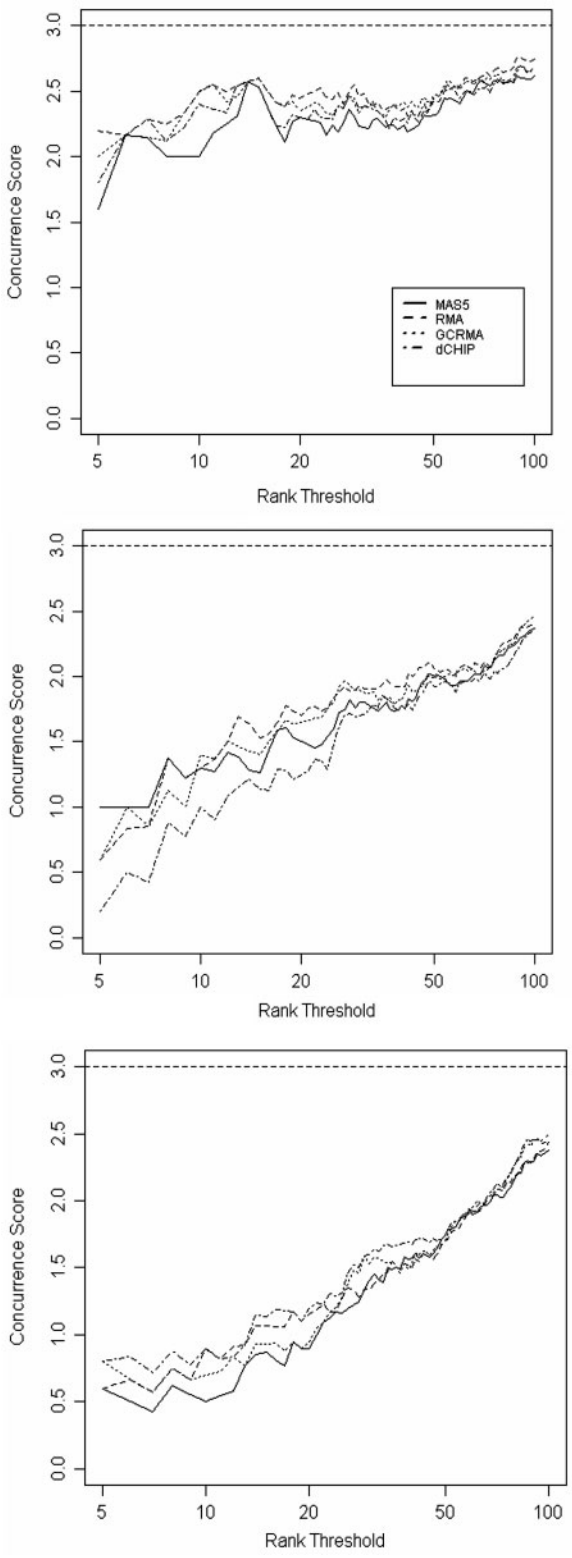

Fig. 3. Topset concurrence plots for KEGG pathways with the 18-day data (top), newborn data (middle) and scrambled data (bottom). As there were only 144 KEGG pathways represented in the data, the tendency of TC to trend towards its upper limit of 3 is clearly visible in these plots.

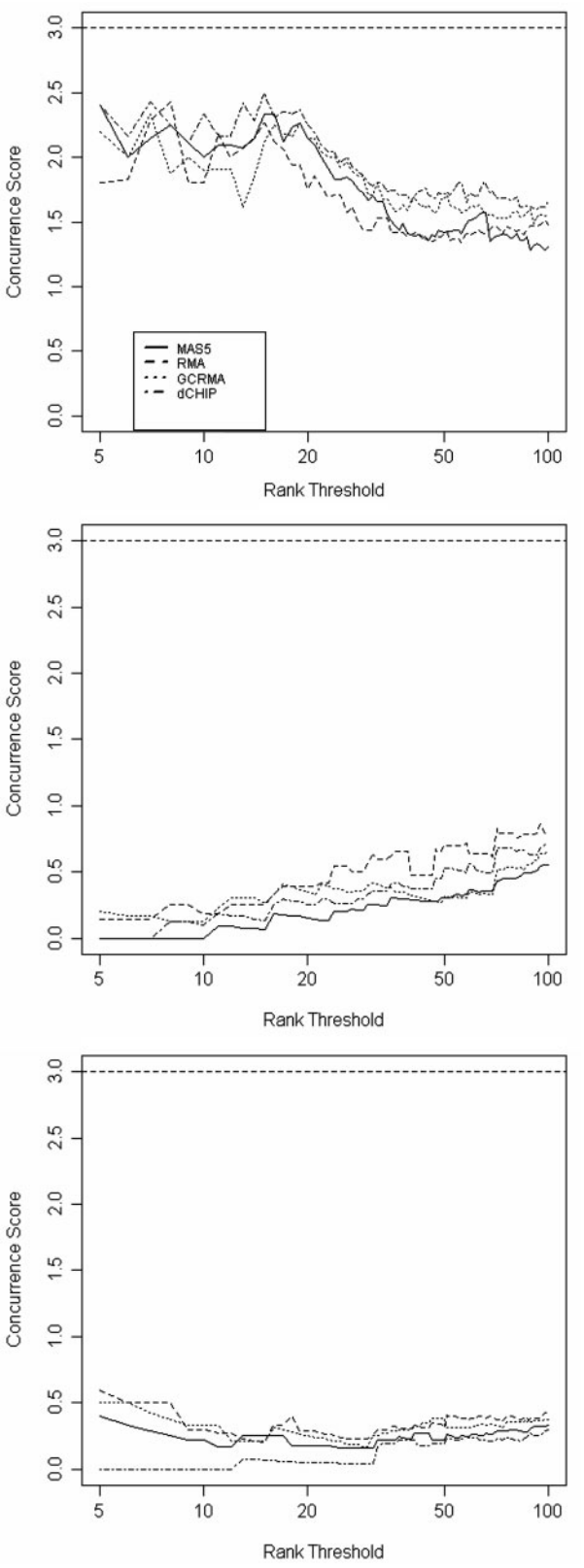

Fig. 4. Topset concurrence plots for the hypergeometric test for the GO terms with the 18-day data (top), newborn data (middle) and scrambled data (bottom).

Table 2. For each dataset, the ranks of the Slc17A5 gene

	18 DAY	Newborn	Scramb
MAS5	9	3	38 492
RMA	109	1	43 537
GCRMA	3	1	34 811
DCHIP	6	2	42 962

Table 3. For the 18-day data, the ranks of GO 6911 (phagocytosis engulfment) and the range of ranks for the probesets involved in GO 6911

	MAS5	RMA	GCRMA	DCHIP
GO 6911 ranks	13	13	14	15
Probeset rank ranges	152–45 038	63–44 637	85–41 241	116–42 871

One final point worth noting is that the Slc17A5 gene that was knocked out turns up with high ranks at the gene level on both day 18 and newborn (Table 2).

4 DISCUSSION

The results demonstrate that, when there is strong differentiation between the groups in the data, the various GEMs, although quite disparate methodologically, produce results that are reasonably similar at a high level, e.g. at the biological process level, even though the gene level results may not necessarily match very well. This concurrence holds even when different methods are used to process the data (e.g. using Limma instead of a *t*-test) as long as the overall procedure has adequate statistical power (e.g. as evidenced by the fact that the concurrence drops when the hypergeometric test is used to test for differences at the biological process level rather than the MLP approach). However, when the differentiation between the groups in the data is subtle, the degree of agreement is substantially reduced with no preference evident for any one GEM.

Above all, one key fact highlighted by the results above is the importance of, whenever possible, interpreting microarray results at the biological process level rather than at the gene level. This is exemplified by the 18-day results for GO 6911. All GEMs identify GO 6911 within the top 15 GO terms (Table 3). However, hardly any of the 31 probe sets associated with this GO term lie within the top 100 probe sets of any GEM (Table 3). In fact, the GO terms identified as being significant and concurrent tend to consist of probe sets whose significance levels are mixed (Supplementary Fig. S5 shows the *P*-value distribution within each of these GO terms), which reflects the fact that the significance of these GO terms and their

concurrence is being driven by the concerted action of a combination of several affected genes, not just a few.

Results such as this and the fact that, despite the weakness of the separation in the newborn data, several of the top GO terms are associated with biological processes known to be affected, suggest that, cumulatively, modest differential expression among many genes involved in a process can still capture the true underlying signal, an important point to consider when interpreting microarray data.

ACKNOWLEDGEMENT

The authors would like to thank James J. Colaianne Jr for his support during this project.

Conflict of Interest: none declared.

REFERENCES

Affymetrix. (2002) Statistical algorithms description document. Technical report. http://www.affymetrix.com/support/technical/whitepapers/sadd whitepaper.pdf

Amaratunga,D. and Cabrera,J. (2004) *Exploration and Analysis of DNA Microarray and Protein Array Data*. John Wiley, New York.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser B*, **57**, 289–300.

Cope,L. *et al.* (2004) A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, **20**, 323–331.

Dai,M. *et al.* (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.*, **33**, e175.

Gene Ontology Consortium. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Hosack,D.A. *et al.* (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol.*, **4**, R70. doi: 10.1186/gb-2003-4-10-r70.

Irizarry,R.A. *et al.* (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.*, **31**, e15.

Irizarry,R.A. *et al.* (2005) Comparison of Affymetrix GeneChip expression measures. *Bioinformatics*, **22**, 789–794.

Li,C. and Wong,W. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.

Moechars,D. *et al.* (2005) *Sialin-deficient Mice: A Novel Animal Model for Infantile Free Sialic Acid Storage Disease (ISSD)*. Society for Neuroscience 35th Annual Meeting, Washington, USA.

Ogata,H. *et al.* (1999) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **27**, 29–34.

Pavlidis,P. *et al.* (2003) *Statistical Analysis of Gene Ontology Classes as Tools for Understanding Gene Expression Changes in the Brain*. Society for Neuroscience Annual Meeting Abstract 758.12.

Pavlidis,P. *et al.* (2004) Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochem. Res.*, **29**, 1213–1222.

Raghavan,N. *et al.* (2006) On methods for gene function scoring as a means of facilitating the interpretation of microarray results. *J. Comput. Biol.*, **13**, 798–809.

R Development Core Team. (2006) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sandberg,R. and Larsson,O. (2007) Improved precision and accuracy for microarrays using updated probe set definitions. *BMC Bioinformatics*, **8**, 48.

Shedden,K. *et al.* (2005) Comparison of seven methods for producing Affymetrix expression scores based on false discovery rates in disease profiling data. *BMC Bioinformatics*, **6**, 26.

Smyth,G. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article 3. doi: 10.2202/1544-6115.1027.

Verheijen,F.W. *et al.* (1999) A new gene, encoding an anion transporter, is mutated in sialic acid storage diseases. *Nat. Genet.*, **23**, 462–465.

Wouters,L. *et al.* (2003) Graphical exploration of gene expression data: a comparative study of three multivariate methods. *Biometrics*, **59**, 1133–1141.

Wu,Z. *et al.* (2004) A model-based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc.*, **99**, 909–917.