Integration and Interpretation of Microarray Data – Explorations and Ideas

Suraj Menon[1], Peter Giles[1], Hui-Sun Leong[1], Ian Brewis[2] and David Kipling[1]

[1] *Department of Pathology, and* [2] *Department of Medical Biochemistry and Immunology, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK.* Email: menons1@cardiff.ac.uk

Analytical objective 1 – How does one find links between experiments?

The end product of most microarray experiments is a list of differentially expressed genes, and extraction of biological meaning from these genelists presents a major challenge. One popular method is over-representation analysis (ORA), for example of Gene Ontology (GO) terms or pathways within a genelist. Here we explore whether a similar approach can be used to identify novel biological links between different microarray experiments, by using ORA and the large amounts of microarray data available in the public domain to detect, within one genelist, enrichment of genelists from other experiments.

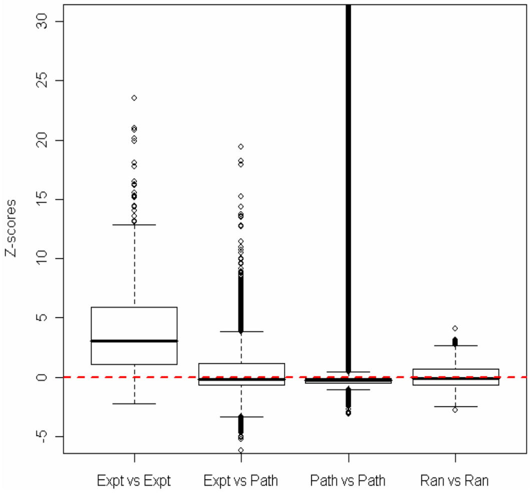

Figure 1

Results The hypergeometric distribution (commonly used for ORA), was initially used to evaluate the overlap between genelists (**Figure 1** and **Table 1**; data based on analysis of 32 genelists from experiments using the Affymetrix HGU-133A array). Overlap simply due to chance would give a Z score distribution with median ~0 and MAD of ~1. However, as shown in Figure 1, there is a positive bias in the Z score distribution. These data suggest that this initial methodology gives an unexpectedly high (and perhaps biologically implausible) number of pairwise comparison reported as statistically significant. This effect is also seen with other chip types and species (data not shown).

Comparison	Median Z-score	MAD of Z-scores	% significant p<0.05(BC)
Expt vs Expt	3.02	3.27	35.88%
Expt vs Path	-0.17	1.07	1.41%
Path vs Path	-0.28	0.27	6.28%
Ran vs Ran	-0.07	1.01	0.00%

Table 1

Discussion For practical considerations, one might argue that a useable metric to find links between experiments should result in the bulk of experimentally-derived genelist pairs having Z scores following a distribution consistent with them not being significant. Interesting matches between experiments should then appear as occasional outliers, and detectable as such (as is indeed observed in the pathway-pathway and experiment-pathways comparisons in Figure 1). While apparently the large number of significant matches that we see could indeed represent real biological similarities, a more plausible hypothesis is that these are a side-effect of the presence on the chip of genes which can never be selected into a genelist. There may be several reasons for this - genes never subject to differential regulation in any situation or experiment (e.g. house-keeping genes), or more prosaically where the probe for that gene may be unable to detect the transcript efficiently and so it would never be scored as differentially expressed. Such phenomena would violate the assumptions of the hypergeometric distribution, and are difficult to quantify and thus compensate for.

We are currently exploring whether we can identify links between genelist pairs by detecting enrichment of GO terms. This is analogous to carrying out ORA of GO terms on genelists, in that the “chip” is represented by one of the pair of genelists, and the “genelist” is now represented by the overlapping genes. As well as curtailing the effects described above, this method can in principle find links between genelists where the overlap does not show a statistically significant overlap at the individual gene level. Links significant at the p<0.05 level (Bonferroni corrected) were found in ~17% of pair-wise comparisons.

Analytical objective 2 – Using clustering metrics to find potentially important biological themes

Biologists often wish to ask whether a certain biological theme (e.g. as might be defined as all the genes within a certain pathway, or with the same GO term) is differentially regulated in a particular dataset. One approach is to first define a set (or ordered list) of differentially expressed genes (DEGs), and then apply ORA to assess whether there is any enrichment of that theme within the DEGs. Here we explore a complementary approach, which is to ask whether assessment of a heatmap generated by hierarchical clustering of the subset of expression data for that particular theme can be used to identify particular themes that appear to contain biological information. Our specific objective was to ask whether a metric could be devised that would could be used to rank the relative degree of informativeness of different themes in a dataset without the need to define sets of DEGs in advance.

	Liposarcoma	Thyroid
Pathways selected by human (A)	37	47
Pathways selected by method (B)	126	93
Overlap between A and B	29	37
Expected overlap	10.57	9.91
Fold Change	2.74	3.73
Hypergeometric Z-score	6.99	10.24
Hypergeometric P-value	6.23e-11	1.32e-19
%A found by B	78.38%	78.72%
%B not in A	76.98%	60.22%

Table 2

Results We took two in-house Affymetrix datasets and produced subsets of the data reflecting all the KEGG, Biocarta and GenMapp pathways. These subsets (441 for each dataset) were then subjected to hierarchical clustering and heatmap visualization. The sample, gene and pathway identifiers were removed and the heatmaps (plus sample dendrograms) were scored in a blind fashion by a biologist as to whether the subset seemed to contain substantial numbers of genes that were expressed in a potentially interesting fashion. These produced a set of pathways ("A" in Table 2). Each subset was then analysed using a metric that scored the degree of variation within the distance matrix used to create the sample clustering, standardized for genelist length and converted to a Z score. This is based on the hypothesis that distance matrices of the most discriminatory clusters would contain considerable numbers of both large inter-group distances and small intra-group distances, which would therefore increase the overall variance of the distance matrix. This metric produced the method-derived set of genelists ("B" in Table 2).

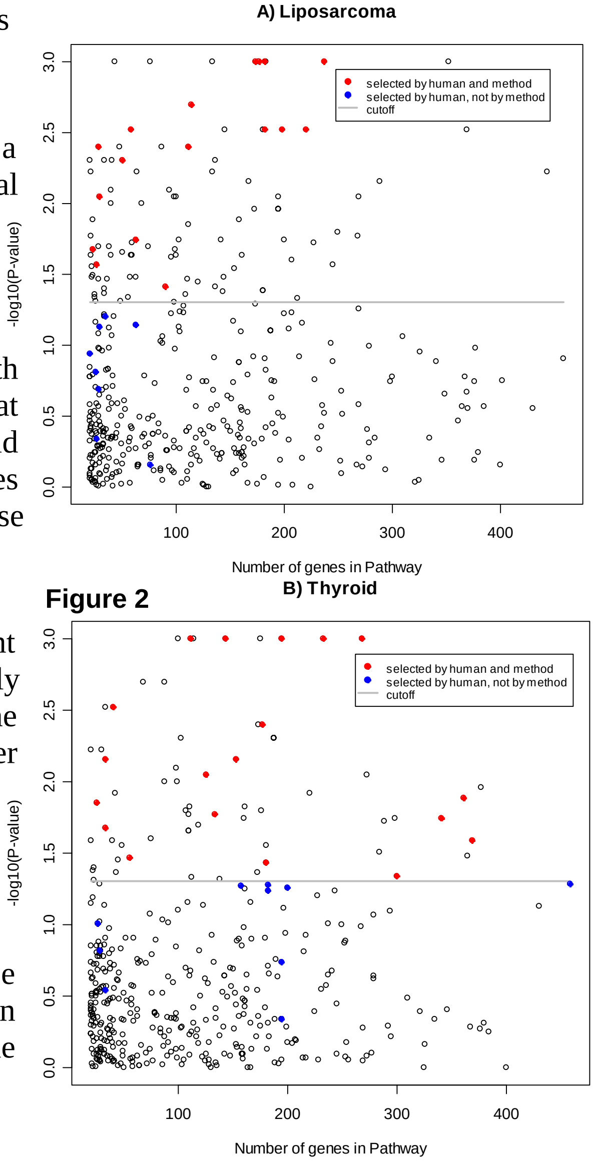

Figure 2

Discussion In both datasets there was an extremely significant overlap in the pathway themes determined as "potentially interesting" based on visual inspection by the biologist, and the pathways selected by this metric (Table 2). Indeed, closer inspection reveals that many of the A-list genelists are close to being called as significant by the metric method (**Figure 2A,B**). The exceptions are several small genelists that are called as significant by the human observer but not by the metric method, which may reflect issues of shape perception in heatmaps based on small numbers of genes. These explorations suggest that such metrics can have utility in ordering biological themes for potential relevance without the need for formal designation of DEGs.