Technical Note

UHN Microarray Centre

Increasing the number of technical replicates will increase the number of significantly differentially expressed genes

- Data from an in-house experiment indicates that by simply increasing the number of technical replicates, the number of significantly differentially expressed genes will also increase
- In this example, Significance Analysis of Microarrays (SAM) was the statistical technique used for determining the number of statistically significant genes

Introduction

When designing microarray experiments, the number of technical and biological replicates is an important consideration that is often dictated by the project's budget. The goal of this technical note is to provide an example of how increasing the number of replicate microarray experiments, from 3 replicates to 6 replicates (and even to 9 replicates) can drastically increase the number of genes which are found to be significantly differentially expressed.

Significance Analysis of Microarrays (SAM) is a statistical technique for finding significant genes in a set of microarray experiments[1]. The input to SAM is a set of gene expression measurements (the normalised $\log_2$ ratio for each gene) from a set of microarray hybridisations and a response variable from each experiment. The response variable may be a grouping like untreated, treated (unpaired or paired), multiclass, a quantitative variable (like blood pressure) or a time course. SAM computes a statistic di for each gene i, measuring the strength of the relationship between gene expression and the response variable[1]. It uses repeated permutations of the data to determine if the expression of any genes is significantly related to the response. The cut-off for significance is determined by a tuning parameter delta, chosen by the user based on the false positive rate. One can also choose a fold change parameter, to ensure that called genes change by a pre-specified amount (1).

Method & Results

The data presented in this technical note was obtained from an in-house study that looked at the fidelity and reproducibility of several RNA amplification methods. SAM version 2.21 was used to analyse the data, however, SAM version 3.0, released on January 23, 2007, offers Gene Set Analysis, a variation on the Gene Set Enrichment Analysis technique.

Three microarray experiments were performed (technical replicates of 10 µg HeLa RNA, indirectly labelled with Cy5, co-hybridised with 10 µg UHRR, indirectly labelled with Cy3) on three separate occasions, for a total of 9 hybridisations. On each occasion, a master mix was used. The samples were labelled and hybridised following the standard UHNMAC Indirect (Aminoallyl) Labelling protocol and hybridised to Human19K6 cDNA arrays.

In the one-class response variable, SAM tests whether the mean gene expression differs from zero. Instead of a p-value, the median number of false significant is indicated and in this example, the median number of false significant was always zero. One-class SAM was performed on each set of three replicates to determine the number of significantly differentially transcribed genes between the two samples. Then, one-class SAM was performed on various combinations of six replicates (Sets 1 and 2, Sets 1 and 3, and Sets 2 and 3) to see if the increased number of replicates had an

effect on the number of significant genes found. Finally, one-class SAM was performed on all 9 replicates (Sets 1, 2, and 3) to determine if the number of significant genes found continued to increase (Table 1).

The number of positive significant genes indicates transcripts that appear to be more abundant in the HeLa sample compared with the UHRR sample, and the number of negative significant genes indicates the number of transcripts that are less abundant in the HeLa sample compared with the UHRR sample.

Discussion & Conclusion

When each set of three hybridisations were analysed using SAM (one-class analysis), the number of significantly differentially expressed genes was less than when the two sets of triplicate hybridisations (total of six hybridisations) were analysed together, and considerably less than when three sets of triplicate hybridisations were analysed together. The number of significant genes from each set of three hybridisations, and various combinations of the sets (six hybs and 9hybs) is outlined in the table below (Table 1). The number of positive significant genes indicates transcripts that appear to be more abundant in the HeLa sample compared with the UHRR sample, and the number of negative significant genes indicates the number of transcripts that are less abundant in the HeLa sample compared with the UHRR sample. It is interesting to note that proportionally, there are very few positive significant genes when considering only three hybridisations together, but many more positive significant genes are found when all nine hybridisations are analysed together.

A set of Venn diagrams (Figure 1) illustrates the overlap among the significant genes from the three sets of replicate data. It would appear that Sets 1 and 2 had the best overlap (highest number of common genes found by one-class SAM analysis). The actual number of common elements on the array was actually higher due to the redundancy of some genes. A cluster image has been generated for the 4737 significant genes found to be differentially transcribed (one-class SAM on all 9 hybridisations) (data not shown).

Regardless of the data analysis method chosen, this study, which used one-class SAM analysis to determine the number of significantly differentially transcribed genes, illustrates the importance of designing microarray experiments with an appropriate number of replicates.

References:
1. SAM Manual, available online: http://www-stat.stanford.edu/~tibs/SAM/sam.pdf

Table 1. The number of statistically significant genes found by one-class SAM analysis. The number of significant genes increased as the number of hybridisations analysed by SAM increased from three to nine hybridisations. This finding suggests that the number of significantly differentially expressed genes can be increased as more replicates are performed.

Hybridisation set ID	Number of hybridisations	Number of significant genes (SAM, one-class)	Positive significant	Negative significant
Set 1	3	2619	5	2614
Set 2	3	2107	6	2101
Set 3	3	1847	5	1842
Sets 1 & 2	6	3379	465	2914
Sets 1 & 3	6	4422	791	3631
Sets 2 & 3	6	3248	359	2889
Sets 1, 2 & 3	9	4737	1166	3571

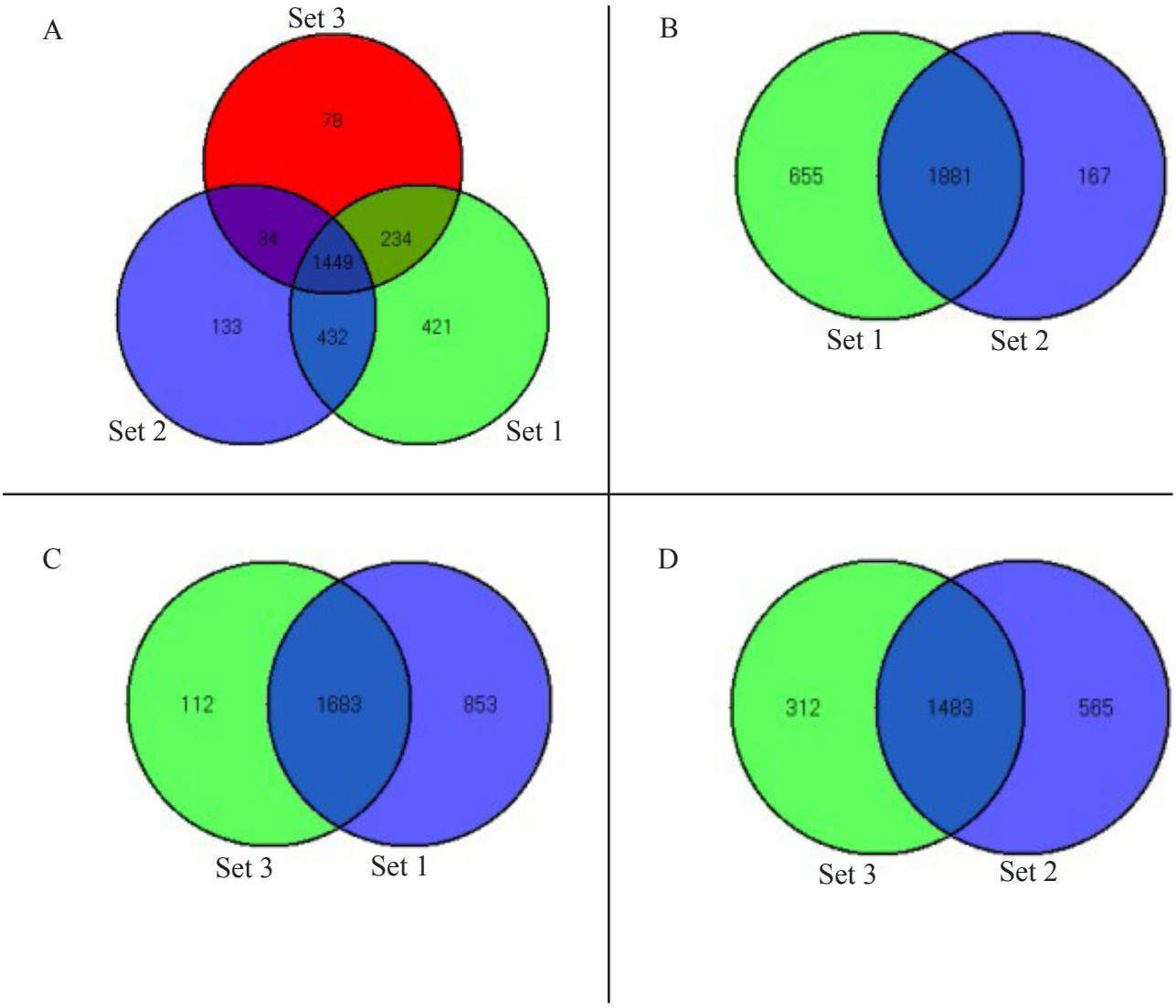

Figure 1. Venn diagrams illustrating the overlap in differentially expressed genes found to be statistically significant using one-class SAM analysis. Each set represents 3 replicates (data outlined in Table 1).