SUMMARIZATION METHODS AND QUALITY PROBLEMS IN AFFYMETRIX MICROARRAYS

Cedric Gondro

The Institute for Genetics and Bioinformatics, University of New England, Armidale, NSW 2351

SUMMARY

The quality of RNA samples and slide hybridizations are paramount for gene expression studies. Here the focus is on hybridization quality in Affymetrix arrays and a comparison of the results of differential expression analysis for six summarization methods using slides of good and bad quality. The overlap of probes detected as differentially expressed across methods is very small even with good arrays. Slides of inferior quality also significantly change the generated list of differentially expressed genes. Extensive qualitative and quantitative quality control measures should be used prior to downstream analyses of data to ensure early detection and removal of problematic slides. Given the high variability of results across summarization methods it is recommended that different methods are used to conduct analyses and the intersecting results of these are used for further downstream analyses.

INTRODUCTION

Microarrays provide a simultaneous measurement of the expression level of thousands of genes from a biological sample. Their most common use is in comparative studies in which one condition is compared to another to identify relative changes between expression levels, i.e. the differentially expressed (DE) genes.

Microarrays are prone to exhibit high levels of experimental and systematic variability that are not related to the experimental contrasts. To ensure the best possible outcome it is critical that these effects are identified and adequately handled. Thus, the bulk of microarray analysis work lies in extensive pre-processing steps to determine the quality of the slides and calibration methods to remove spurious variation. Bad quality slides have unreliable intensity measures and can have a very large effect on final results. These slides should be identified and removed from the analysis. Slides deemed of adequate quality will then undergo calibration steps which generally consist of: (1) background correction to remove intensity measures that are not due to the target; (2) normalization of the probe intensities, which is achieved by adjusting the overall distribution intensities making them similar across slides (note that this step usually makes a dataset testable only within itself, if new slides are added to the experiment the entire set has to be renormalized); and (3) a summarization step which is specific to Affymetrix GeneChips, since these are unique in the use of a set of short oligos to target a transcript, usually 11 different pairs of 25mer oligos, with each pair consisting of a perfect match (PM) to the standard reference sequence and a mismatch (MM) with exactly the same sequence except for a mismatch at position 13 (in principle the MM should pick up cross hybridization noise). This *probe set* is summarized into a single intensity value for each target on each array.

Different methods have been developed for each of the above mentioned steps (Irizzary *et al.* 2006). However it is still unclear which approach is best, and it has been shown that the main source of variation between results is due to the choice of summarization method (Harrison *et al.* 2007). In this work we quantify the differences in probes detected as differentially expressed across six summarization methods using the Affymetrix GeneChip bovine genome array, and test the robustness of each method to technical hybridization problems.

MATERIALS AND METHODS

Data. Twenty Affymetrix GeneChip bovine genome arrays hybridized to RNA extracted from ovine blood samples were used. The data consists of 2 subsets of slides (taken from a larger experiment), one with 10 good quality arrays and the other with 10 bad quality arrays, each hybridized to the same RNA samples. The bad slides were due to technical hybridization problems that occurred in the original experiment while the good slides are simply a repeat with a new batch of slides using the same RNA samples. Each set is a simple control x treatment contrast with 5 slides per group.

Pre-processing. Qualitative and quantitative quality control (QC) measures were used prior to downstream analyses to detect problematic slides. These include image plots for detection of spatial effects, normalized unscaled standard errors (NUSE), relative log expression, MA plots, RNA degradation, call flags, match-mismatch intensities, slide correlations and principal component analysis (an overview of pre-processing is given in Gentleman *et al.* 2005).

Summarization methods. Six summarization methods were used to generate expression measures: MAS 5.0, RMA, GCRMA, PLIER, VSN and MBEI (methods are detailed in Gentleman *et al.* 2005).

Design. For each summarization method, differential expression of genes was tested using different combinations of ten slides. Initially analysis was conducted using only the good slides, and then successively repeated by replacing 2 good slides with their respective (same RNA source) bad slides until the analysis was run with only the bad slides (thus the number of good slides in each replicate was 10, 8, 6, 4, 2, 0). The procedure was repeated as a balanced replacement (bad slides shared equally between contrasts) and unbalanced (bad slides allocated in first instance to the control).

Analysis. Differential expression was tested on log2 expression intensities for the 6 summarization methods for each of the 10 datasets using a moderated t-statistic (Smyth 2004) after removing control probes and probes detected as marginal or absent across all arrays (probes with low intensities). Probes were selected as differentially expressed for a p-value of 0.01.

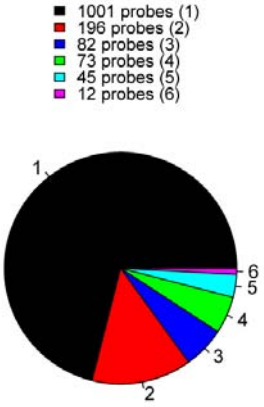

Figure 1. Intersect of DE probes across summarization methods for good slides. Parentheses indicate number of methods in which the probes were detected.

RESULTS AND DISCUSSION

The number of DE probes detected is summarized in table 1. Of immediate notice, and concern, is that the intersect of probes across all methods is extremely low irrespective of the quality of the slides. The intersect is moderately better for the good quality slides but still only 3.2%, that is only 12 probes out of the average 371DE probes per method (the total number of probes across methods was 2228 with 1409 unique). Most probes are detected in only one method (71%, see Figure 1). Even though some methods are methodologically close (e.g. same normalization is used), results tend to bear limited replicable correlation with different datasets. Hence if for a given analysis RMA and GCRMA yield similar results while MAS and PLIER are similar between them but further apart from the first two, this cannot be used to make decisions on a choice of summarization

method since in another experiment this order may be completely reversed. Figure 2 illustrates this issue with the good and bad arrays. In terms of numbers of DE probes there is a considerable spread across methods with GCRMA consistently showing the lowest numbers of DE probes and MBEI the highest, up to 10-fold differences can be seen in Table 1.

Table 1. Differentially expressed probes per summarization method. Numbers in parenthesis refer to the DE probes detected in the unbalanced designs. The last column shows the number of probes in common across all methods

arrays	mas	rma	gcrma	plier	vsn	mbei	intersect
good slides	292	347	126	386	306	771	12
8 good	158 (124)	127 (74)	52 (37)	58 (101)	55 (60)	300 (198)	2 (1)
6 good	65 (231)	20 (170)	19 (43)	22 (321)	18 (270)	38 (411)	1 (0)
4 good	40 (127)	25 (103)	22 (50)	9 (94)	17 (70)	41 (356)	0 (1)
2 good	91 (65)	143 (64)	31 (35)	32 (20)	43 (31)	75 (99)	0 (0)
bad slides	213	430	164	439	766	315	1

Taking the DE probes using the good slides as a *gold standard*, Figure 3 shows how even small numbers of bad slides can result in very different lists of DE probes. With two bad slides (one in each treatment) on average around 28% of the original DE probes are still detected. MAS is somewhat more robust at 43% overlap, whilst PLIER at the lower end shows only 16%. With four or more bad slides the numbers fall under the 20% line for all methods. Even with the lower numbers of DE probes in good/bad combinations there are still many new probes being detected which are just noise due to the poor slide quality. As would be expected the unbalanced designs show even greater disparities of results (Figure 3, right pane). The somewhat greater robustness of MAS is due to the method not normalizing between arrays but on a targeted predefined mean value. The downside of MAS is that it tends to overestimate the effects at low intensities.

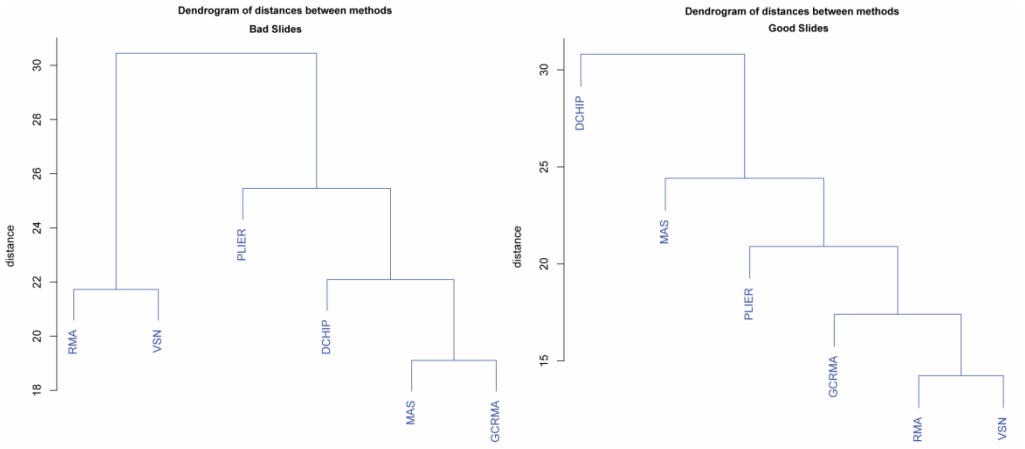

Figure 2. Distance plots of summarization methods for the bad and good slides.

The 12 intersect probes detected across all summarizations using the good slides are more robust to slide quality. With two bad slides the average across methods is close to 55% (between

25% for PLIER and 67% for MAS, RMA and MBEI) and 39% for unbalanced designs. Even using only bad slides the average overlap is still over 26%.

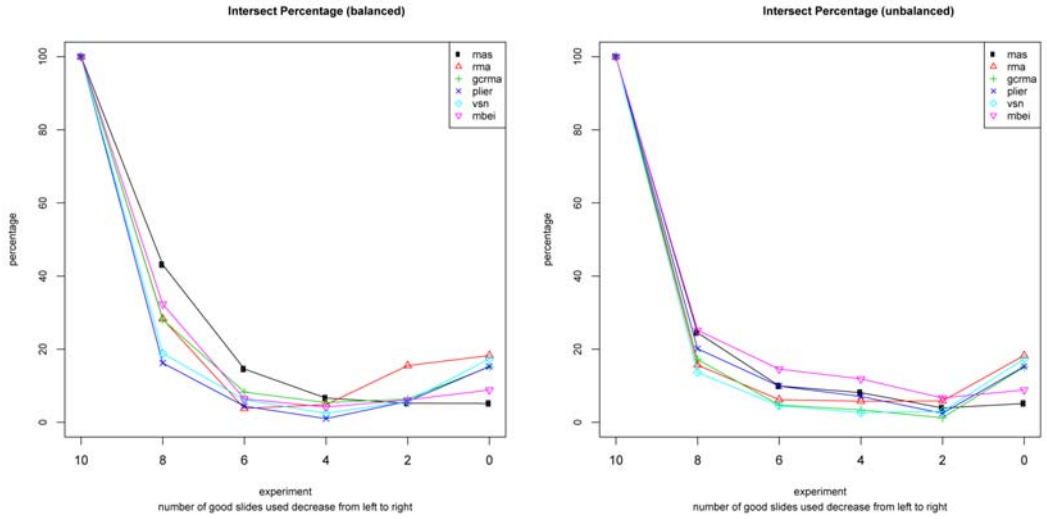

Figure 3. Similarity of DE probes for each summarization in relation to 10 good slides caused by replacing good slides with bad slides. Left pane balanced, right pane unbalanced.

CONCLUSIONS

The dataset used in this work was only available because of extensive quality control test which helped flag problems with the original hybridizations in the larger experiment. If these checks had not been performed the results would have been very different for whichever analysis methodology had been used. The choice of summarization method also has a large impact on final results and there is very little overlap between them. No individual method is highly robust to experimental noise but different summarizations combined can be more robust.

For positive outcomes from array experiments it is recommended that extensive quality control checks, such as those previously mentioned, are performed. If in doubt an array should be discarded or at least the analysis should be run with and without a questionable slide to quantify the effect it is having on results.

More than one summarization method should be used for the analyses. The intersect of results can help identify a more stable subset of results and in effect also help to correct for multiple testing problems.

ACKNOWLEDGEMENTS

This research was financially supported through SheepGenomics by Meat and Livestock Australia and Australian Wool Innovation Limited.

REFERENCES

Gentleman, R., Carey, V.J., Huber, W., Irizarry, R.A. and Dudoit, S (eds.) (2005). "Bioinformatics and Computational Biology Solutions Using R and Bioconductor" Springer, The Netherlands.

Harrison, A., Johnston, C. and Orengo, C. (2007). *BMC Bioinformatics* **8**:195.

Irizarry, R., Wu, Z. and Jaffee, H. (2006). *Bioinformatics* **22:**789.

Smyth, G.K. (2004). *Stat Appl Genet Mol Biol*. **3**:Article3.