AffyDEComp: towards a benchmark for differential expression methods

Richard D Pearson (richard.pearson@postgrad.manchester.ac.uk), University of Manchester

October 30, 2007

1 Introduction

The issue of method validation is of great importance to the microarray community; arguably more important than the development of new methods [Allison et al., 2006]. The microarray analyst is faced with a seemingly endless choice of methods, many of which give evidence to support their claims of being superior to other approaches, which at times can appear contradictory. Method validation is a difficult problem in microarray analysis because, for the vast majority of microarray data sets, we don't know what the "right answer" really is. For example, in a typical analysis of differential gene expression, we rarely know which genes are truly differentially expressed (DE) between different conditions.

Perhaps the most well-known and widely used benchmark for Affymetrix analysis methods is Affycomp [Cope et al., 2004]. While a very valuable tool of summarization method validation, Affycomp is not ideal for comparison of DE methods because:

1. It uses data sets which only have a small number of DE spike-in probesets.

2. It only uses fold change (FC) as a metric for DE detection, and hence cannot be used to compare other competing DE methods.

The "Golden Spike" data set of Choe et al. [2005] includes many differentially expressed spike-in probesets, making it potentially very valuable as a benchmark data set. There have, however been a number of criticisms of this data set.

2 Methods

We have identified six key stages of the analysis pipeline for the Golden Spike data where choices have to be made. By comprehensively re-analysing the Golden Spike data using all combinations of these choices, we have used this data set in a comparison of methods which is far more extensive than any previous study. We have also developed a web resource (AffyDEComp) where the effects of the different choices can be seen. All analysis has been performed using open source tools and publicly available data. We have made all our analysis scripts publicly available.

3 Results

We show that certain choices in the analysis pipeline can overcome the more serious defects in the Golden Spike data set. We have also shown how the results of these decisions can lead to the apparently contradictory results found in previous studies. We see that there is no DE method that is clearly better than other methods, but that what is important is the combination of summarization and DE method. We show that, while flawed, this data set is still a useful tool for method comparison, particularly for identifying combinations of summarization and differential expression methods that are unlikely to perform well on real data sets. Figure 1 shows the areas

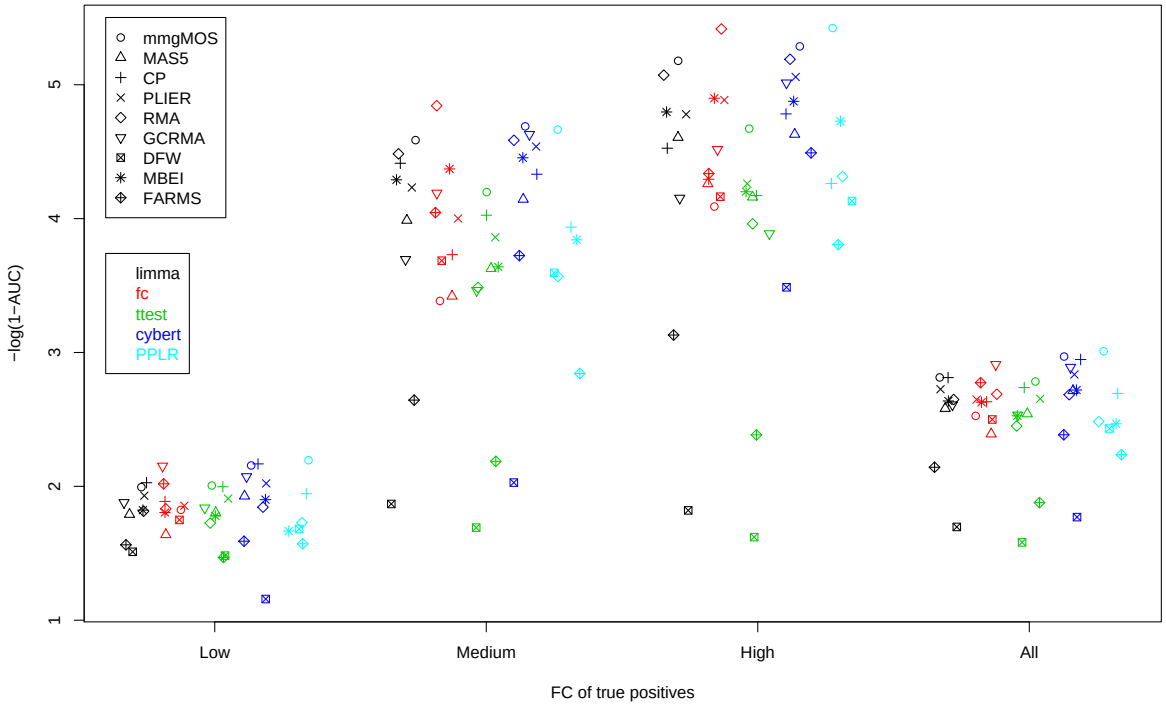

Figure 1: Areas under ROC curves of Golden Spike data using different combinations of summarization and DE detection methods, and different sets of true positives. For these charts only the equal spike-ins are used as true negatives. The chart shows probesets selected using a 1-sided test of up-regulation. The Low true positives are those spike-ins with a FC greater than 1 but less than or equal to 1.7. The Medium true positives are those spike-ins with a FC between 2 and 2.5. The High true positives are those spike-ins with a FC greater than or equal to 3. The y-axis shows -log(1-AUC) rather than AUC, as this gives a better separation between the higher AUC values, but retains the same rank order of methods. The x-axis is categorical, with points jittered to avoid placement on top of each other.

under the ROC curves (AUC) for our recommended analysis pipeline choices.

4 Conclusion

We conclude with recommendations for preferred Affymetrix analysis tools, and for the development of future spike-in data sets.

References

D. B. Allison, X. Cui, G. P. Page, and M. Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet*, 7(1):55–65, 2006.

S. E. Choe, M. Boutros, A. M. Michelson, G. M. Church, and M. S. Halfon. Preferred analysis methods for affymetrix genechips revealed by a wholly defined control dataset. *Genome Biol*, 6(2):R16, 2005.

L. M. Cope, R. A. Irizarry, H. A. Jaffee, Z. Wu, and T. P. Speed. A benchmark for affymetrix genechip expression measures. *Bioinformatics*, 20(3):323–31, 2004.