DNA Microarrays
Differential Expression

BCB 504: Applied Bioinformatics

Matt Settles

University of Idaho
Bioinformatics and Computational Biology Program

February 6, 2012

Outline

Gene filtering

Purpose is to remove probe sets (aka genes) that are likely to be of no use (uninteresting) when modelling the data.

non-specific Choose genes to remove (or keep) without using any phenotypic variables in the filtering process. Result can then be used with any downstream process without bias.

specific Use of phenotypic variables to choose which genes should be kept/removed. Results can be used in a descriptive nature, should not be used for statistical testing.

Affymetrix Present/Marginal Absent Calls

PMA calls - used to determine whether a probe set is reliably detected (Present), not detected (Absent), or marginally detected (Marginal) as being expressed.

- Discrimination score is first calculated for each probe pair.

$$(PM - MM)/(PM + MM) - \tau \quad (1)$$

- A One-sided Wilcoxon Signed Rank test is used to compare two related samples or repeated measurements on a single sample to assess whether their population mean ranks differ (i.e. it's a paired difference test).
- Present absent calls based on resulting p-value
  - p-value $<= \alpha_1$ = Present
  - α_1 < p-value < α_2 = Marginal
  - α_2 <= p=value = Absent
- $\tau = 0.015$, $\alpha_1 = 0.04$, $\alpha_2 = 0.06$

Other non-specific filters I

Annotation Based Filtering Can reduce be a predetermined set of genes (aka entrez ids), by Gene Ontology group, filter based on available annotation data.

Duplicate Probe Removal Probes determined by annotation to be pointing to the same gene are compared, and only the probe with the highest variance value will be retained.

Other non-specific filters II

Variance Based Filtering Perform numerical cutoff-based filtering (aka IQR). The intention is to remove uninformative probe sets, representing genes that were not expressed at all or no change is occurring. Observations have shown that unexpressed genes are detected most reliably through their low variability across samples. IQR is robust to outliers, a default cutoff value of 0.5 is motivated by the rule of thumb that in many tissues only 40% of genes are expressed.

Approaches I

LIMMA Linear Models for Microarray Data approach. Limma applies a linear model to the data and uses model contrasts to extract specific comparisons of interest. It then uses multiple testing correction based techniques (FDR, FWER, etc.) for MT corrected p-values. Can also choose a fold change parameter, to ensure that called genes change at least a pre-specified amount.

Approaches II

SAM Significance analysis of Microarrays (SAM) approach. SAM computes a statistic d_i for each gene i, measuring the strength of the relationship between gene expression and a response variable. It then uses repeated permutations of the data to determine if the expression of any genes are significantly related to the response. The cutoff 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 at least a pre-specified amount.

The SAM plot

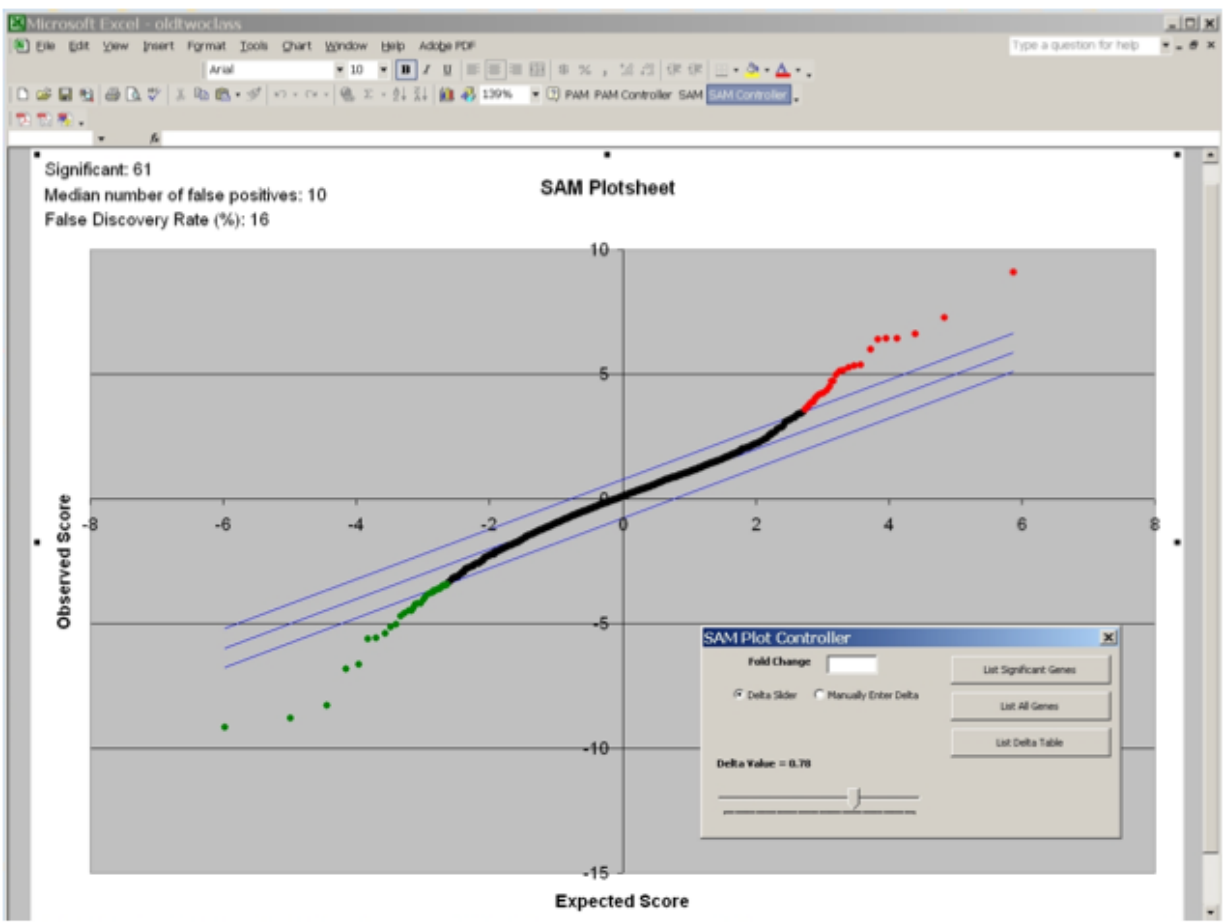

Figure 3: The SAM Plot

The SAM plot

SAM table - gives details about different delta values

(A)			
Δ	#false pos	# called	FDR
0.3	11.7	100	0.117
0.4	9.3	76	0.122
0.5	5.9	65	0.091
0.6	4.4	39	0.113
0.7	3.5	33	0.106
0.8	2.1	29	0.072
0.9	1.6	17	0.094
1.0	1.3	16	0.081

Variance Corrections

- When doing statistical test, we estimate the variance for each gene individually. This is fine, *if* we have enough replicates, but with few replicages (say 2-5 per group), these variances are highly variable.
- In a moderated statistic, the estimated gene-specific variance s_g^2 is replaced by a weighted average of s_g^2 and s_0^2, which is a global variance estimator obtained from pooling all genes.
- This produces an interpolation between the t-test and a fold-change criterion.

Both SAM and LIMMA, perform Variance Correction, SAM bins genes with common fold change, LIMMA performs an empirical Bayes estimate for s_0^2.

Multiple Testing Correction

Aim: For a given type I error rate α, use a procedure to select a set of "significant" genes that guarantees a type I error rate less than or equal to α.

Bonferonni Basically multiple raw p-value by the number of tests. **Unrealistic** given the number of genes tested. Many methods available in the bioconductor packages *multtest* and *qvalue*

Benjimini and Hochberg (BH) Controls the false discovery rate. Works for independent test statistics and for some types of dependence. Tends to be conservative if many genes are differentially expressed.

q-value the q-value of a gene is defined as the minimal estimated FDR at which it appears significant.