Significance Analysis for Microarrays

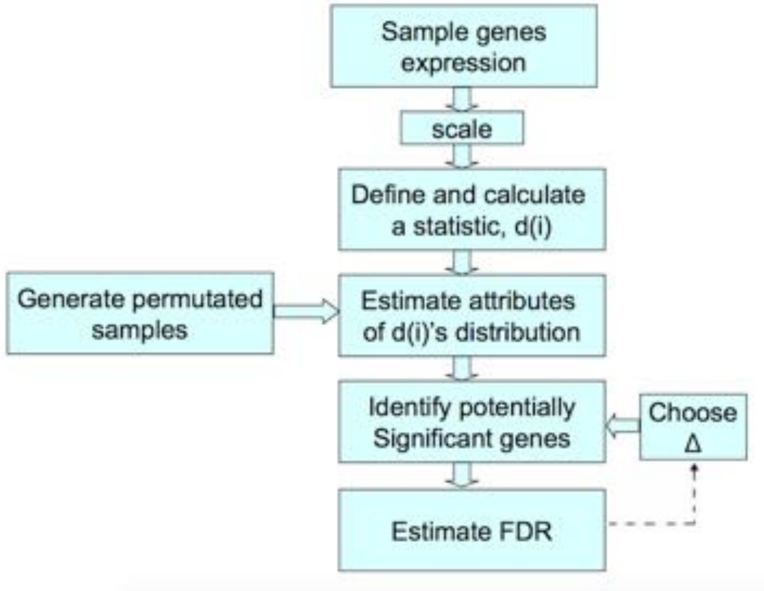
Mikhail Dozmorov

Fall 2017

Significance analysis of microarrays (SAM)

- V. G. Tusher et.al. "Significance Analysis of Microarrays Applied to the Ionizing Radiation Response" PNAS 2001 http://www.pnas.org/content/98/9/5116.long
- A clever adaptation of the t-ratio to borrow information across genes
- SAM seeks to control the proportion of false rejections among the set of rejected hypotheses (FDR).
- Permutation method is used to calculate the null distribution of the modified t-statistics.

SAM procedure



Sample genes expression

scale

Define and calculate a statistic, d(i)

Generate permutated samples

Estimate attributes of d(i)'s distribution

Identify potentially Significant genes

Choose Δ

Estimate FDR

SAM t-test

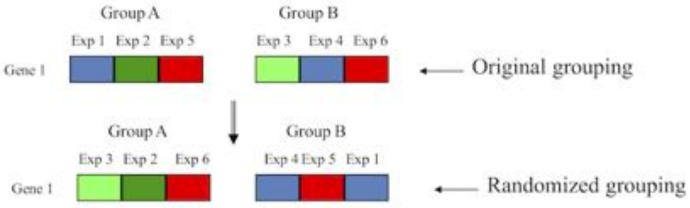
- With small sample sizes low and high variance can occur by chance
- Variance depends on expression level
- Try to remove (or minimize) the dependence of test statistics on variances (because small variance tend to lead to bigger test statistics).
- Solution: add a small constant to the denominator in calculating t statistics:

$$d_i = \frac{\bar{y_i} - \bar{x_i}}{s_i + s_0}$$

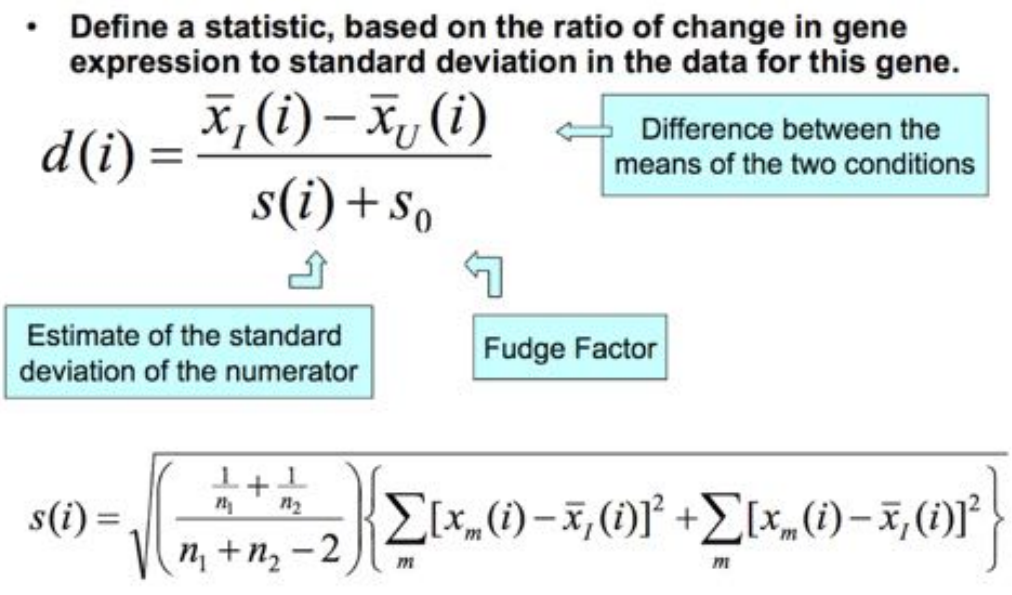
- $\bar{y_i}, \bar{x_i}$ - Means of two groups for gene i.
- s_i - Standard deviation for gene i, assuming equal variance in both groups.
- s_0 - "Exchangeability factor" estimated using all genes.

SAM two-class unpaired

- For each gene, compute the d-value (similar to a t-statistic). This is the observed d-value (d_i) for that gene.
- Randomly shuffle the expression values between groups A and B. Compute the d-value for each randomized set.
- Take the average of the randomized d-values for each gene. This is the 'expected relative difference' (d_E) of that gene. Difference between (d_i) and (d_E) is used to measure significance.
- Plot $d_{(i)}$ vs. $d_{E(i)}$
- Calculate FDR = average number of genes that exceed Δ in the permuted data.



SAM statistics



- **Define a statistic, based on the ratio of change in gene expression to standard deviation in the data for this gene.**

$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_0}$$

Difference between the means of the two conditions

Estimate of the standard deviation of the numerator

Fudge Factor

$$s(i) = \sqrt{\left(\frac{\frac{1}{n_1} + \frac{1}{n_2}}{n_1 + n_2 - 2}\right)\left\{\sum_m [x_m(i) - \bar{x}_I(i)]^2 + \sum_m [x_m(i) - \bar{x}_I(i)]^2\right\}}$$

Why s_0 ("fudge" factor)?

- Prevents $d_{(i)}$ from becoming too large when the variance is close to zero (which can lead to false positives)
- Choose one s_0 for the entire dataset, to make the coefficient of variation of $d_{(i)}$ approximately constant
- Typically, s_0 can be computed as the 90^{th} percentile of the standard errors of all genes

Estimating significance

- We have calculated a new statistics and we don’t have a parametric description of the null distribution
- Solution: generate an empirical null distribution form a set of experiments where all hypotheses should be null
- Generate permutations of data labels so no difference is expected
- For each permutation p, calculate $d_{p(i)}$.

Identifying Significant Genes

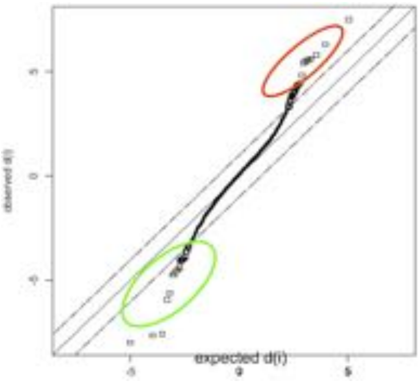
- Define a threshold Δ
- Find the smallest positive $d_{(i)}$ such that

$$|d_{(i)} - d_{E(i)}| \geq \Delta$$

- Call it t_1
- In a similar manner, find the largest negative $d_{(i)}$. Call it t_2
- For each gene i, if $d_{(i)} \geq t_1 \vee d_{(i)} \leq t_2$, call it potentially significant

Identifying Significant Genes

- Rank the original d(i)'s: $d_{(1)} \geq d_{(2)} \geq d_{(3)} \geq \ldots$
- Plot $d_{(i)}$ vs. $d_{E(i)}$
- For most of the genes, $d_{(i)} \sim d_{E(i)}$



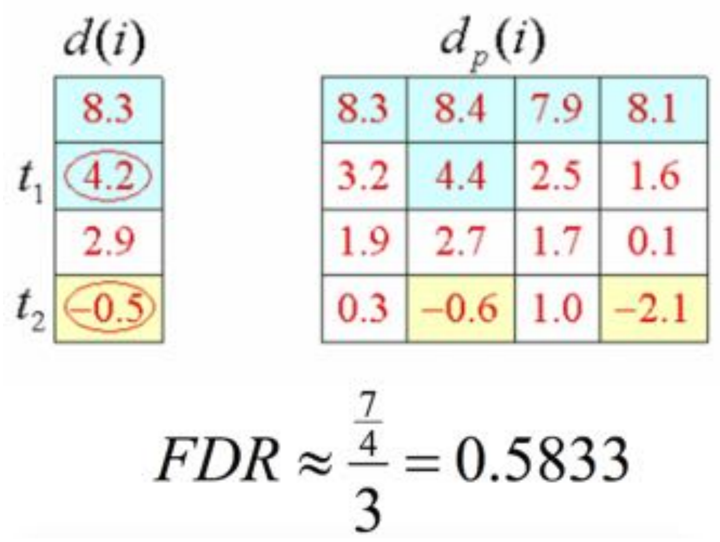
Estimate FDR

- t_1 and t_2 will be used as cutoffs
- Calculate the average number of genes that exceed these values in the permutations.
- Estimate the number of falsely significant genes, under H_0:

$$\frac{1}{n.perm}\sum_{p=1}^{n.perm} number\{d_{p(i)} \geq t_1 \vee d_{p(i)} \leq t_2\}$$

- Divide by the number of genes called significant

Estimate FDR example

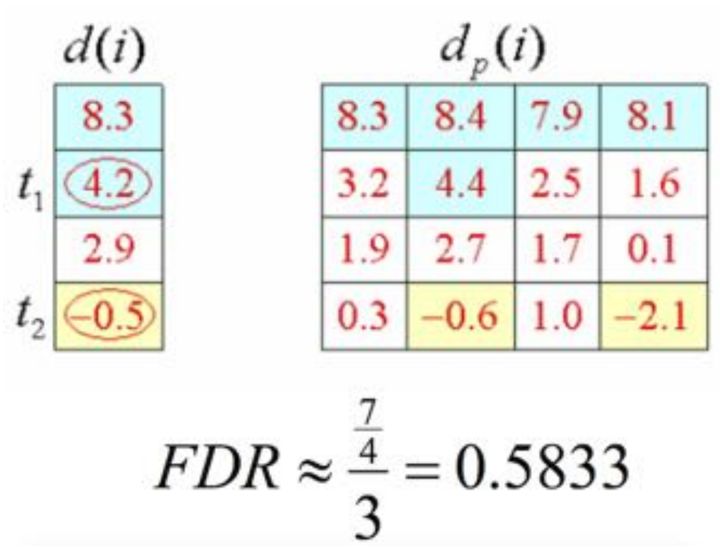


	$d(i)$
	8.3
t_1	4.2
	2.9
t_2	−0.5

$d_p(i)$			
8.3	8.4	7.9	8.1
3.2	4.4	2.5	1.6
1.9	2.7	1.7	0.1
0.3	−0.6	1.0	−2.1

$$FDR \approx \frac{\frac{7}{4}}{3} = 0.5833$$

Estimate FDR from the reference distribution d



	$d(i)$
	8.3
t_1	4.2
	2.9
t_2	−0.5

$d_p(i)$

8.3	8.4	7.9	8.1
3.2	4.4	2.5	1.6
1.9	2.7	1.7	0.1
0.3	−0.6	1.0	−2.1

$$FDR \approx \frac{\frac{7}{4}}{3} = 0.5833$$

Delta Δ is the half-width of the bar around the 45-degree line

Other applications of SAM

- More than two groups
- Paired data
- Survival data, with censored response

SAM summary

- Highly cited (>7000 citations), http://www-stat.stanford.edu/~tibs/SAM/.
- Implemented as Bioconductor package `siggenes`, and Excel plugin.
- Follow-up work: SAMSeq on RNA-seq DE test.
- Limitations: solutions for s_0 often sensitive to data.

Summary on two-sample DE test

- Try to alleviate the “small sample variance” problem.
- Combine information from all genes.
- Many other variations of the model.
- In practice SAM and limma performs similarly.