

Differential Expression



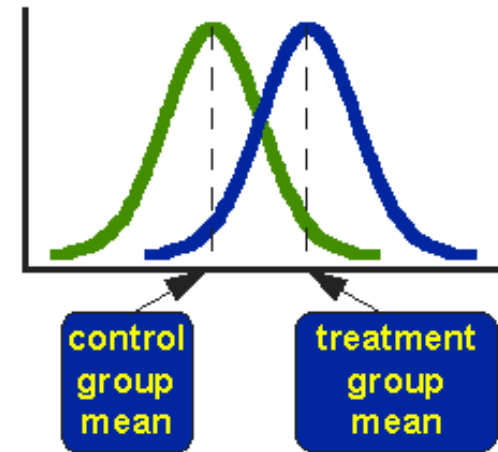
Overview

- What is differential expression?
- Models
 - T-test
 - SAM
 - Rank Product
- Measure of significance



What is differential expression?

- Measurements before and after treatment
- Before: 1.5, 0.8, 1.2
- After: 2.1, 1.7, 1.5
- Are the distributions significantly different?
 - Need a model that can help us decide



Modeling Considerations

Parametric models: need enough data to decide the distribution

Problem: with few arrays you are unwilling to make parametric assumptions about gene expression values

Non-parametric models: use of a permutation test, or similar

Problem: these models have reduced power and hence less ability to discriminate.

Aggregation across genes: one of the basic strategies used is to aggregate information across genes



T-test

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}\right)}}$$

- Assumes normal distribution of data
- Assumes student t-distribution of t-scores



Moderated / Bayesian t-tests

Rather than estimating within-group variability (denominator of t-test) over and over again for each gene, pool the information from many similar genes

Baldi, Long 2001

Tusher et al. (SAM) 2001

Lönnstedt and Speed 2002

Kendziorski et al. (Earrays) 2003

Smyth (limma) 2004



Moderated / Bayesian t-tests

- In most cases, an overall estimate of the variance, s_0^2 , is computed
- Then for each gene, an estimate of the per gene variance, s_g^2 , is computed
- The variance used is a weighted average of s_0^2 and s_g^2
- The actual method of estimating the overall variance and the method of averaging is slightly different in different contexts



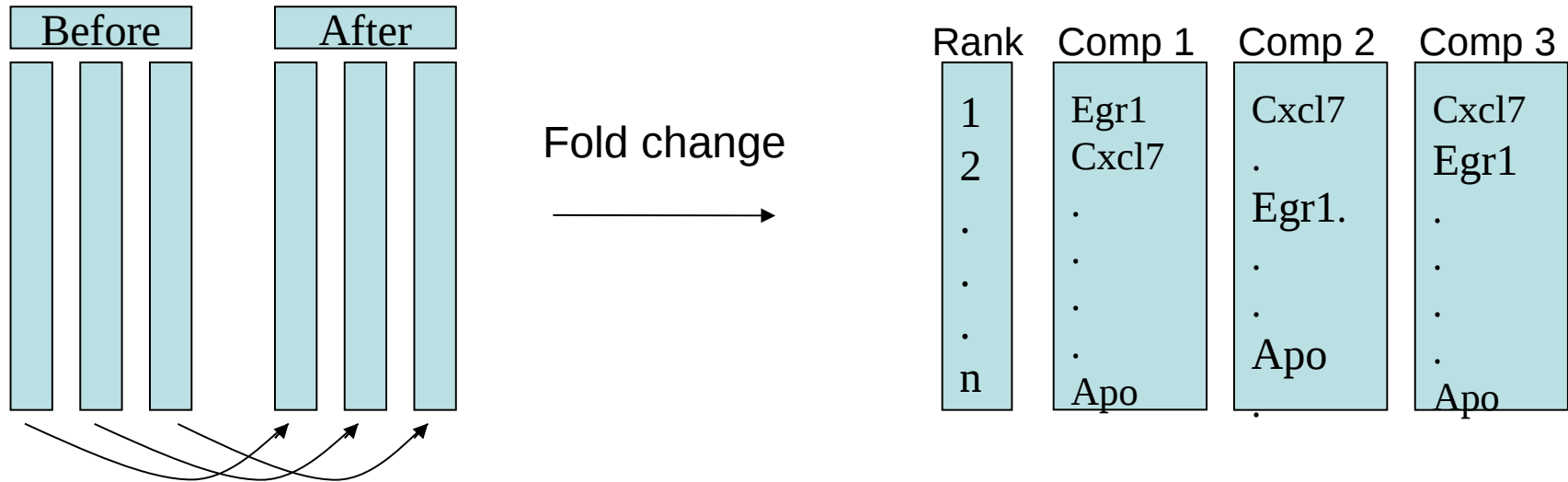
Significance Analysis of Microarrays (SAM)

$$s = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}\right)} + B}$$

- Assumes normal distribution of data
- Makes no assumption about the distribution of the score
- **Advantages:**
 - eliminate occurrence of accidentally large t-values due to accidentally small within-group variance
 - effectively introduce a ‘fold-change’ criterion



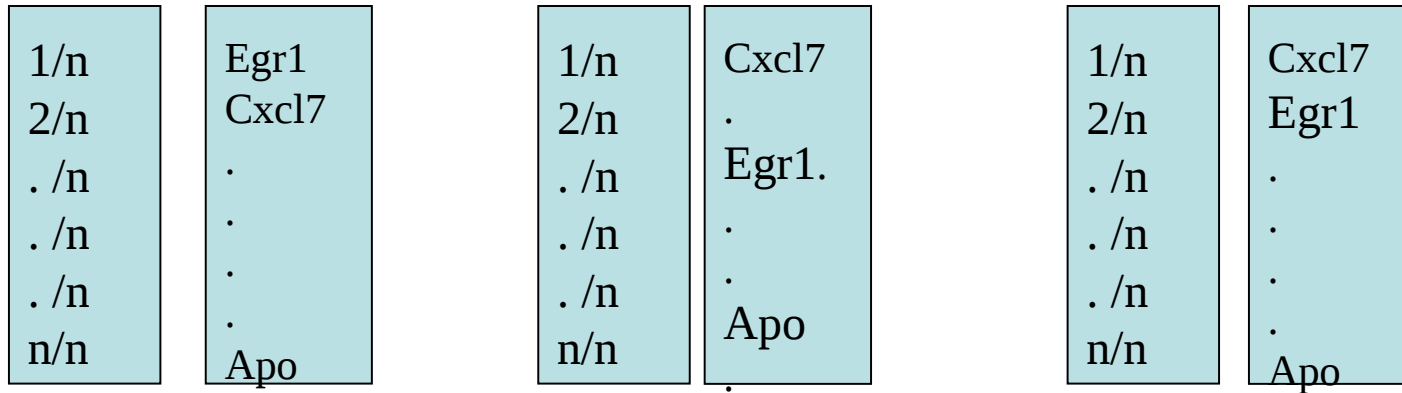
Rank Product



- Makes no assumption about distribution of the data
- No calculation of variance across samples



Rank Product



- $RP (Cxcl7) = 2/n * 1/n * 1/n$



Significance of scores

- P-value is defined as the probability of a gene obtaining the score by chance
 - Assuming only one gene has been tested
 - Does not take into account that when multiple genes are tested, the probability of randomly obtaining a high score increases
- The p-value should therefore be corrected for multiple testing
 - E.g Bonferroni correction



False Discovery Rate

- FDR refers to the number of genes on a gene list that is expected to be false positive
- If the p-value of gene nr 100 on a gene list is 0.001 and we have analysed 20 000 genes
 - Expect $20\ 000 \times 0.001 = 20$ genes to be false positives among the top 100 genes
- $FDR = FP/Rank \times 100\%$



FDR-value

	score	Fold change	FDR
rCG34061	4.655	1.421	0
56227	4.552	2.168	6.476
313504	4.525	3.182	5.667
rCG48508	4.411	1.788	4.318
315095	4.343	4.724	3.778
307947	4.2	1.515	6.476
rCG22278	4.196	1.673	6.253
rCG26536	4.186	2.56	6.045
304092	4.167	2.47	5.495
359725	4.14	1.443	5.333
360415	4.117	1.995	5.181

q-value

- FDR is not strictly increasing the further down on a gene list you
- The q-value is the smallest FDR value that is seen for a particular gene list
- q-value is an FDR value and it is strictly increasing the further down the gene list you get



q-value

	score	Fold change	FDR	q-value
rCG34061	4.655	1.421	0	
56227	4.552	2.168	6.476	
313504	4.525	3.182	5.667	
rCG48508	4.411	1.788	4.318	
315095	4.343	4.724	3.778	
307947	4.2	1.515	6.476	
rCG22278	4.196	1.673	6.253	
rCG26536	4.186	2.56	6.045	
304092	4.167	2.47	5.495	
359725	4.14	1.443	5.333	
360415	4.117	1.995	5.181	

q-value

	score	Fold change	FDR	q-value
rCG34061	4.655	1.421	0	0
56227	4.552	2.168	6.476	
313504	4.525	3.182	5.667	
rCG48508	4.411	1.788	4.318	
315095	4.343	4.724	3.778	
307947	4.2	1.515	6.476	
rCG22278	4.196	1.673	6.253	
rCG26536	4.186	2.56	6.045	
304092	4.167	2.47	5.495	
359725	4.14	1.443	5.333	
360415	4.117	1.995	5.181	

q-value

	score	Fold change	FDR	q-value
rCG34061	4.655	1.421	0	0
56227	4.552	2.168	6.476	
313504	4.525	3.182	5.667	
rCG48508	4.411	1.788	4.318	
315095	4.343	4.724	3.778	
307947	4.2	1.515	6.476	
rCG22278	4.196	1.673	6.253	
rCG26536	4.186	2.56	6.045	
304092	4.167	2.47	5.495	
359725	4.14	1.443	5.333	
360415	4.117	1.995	5.181	

q-value

	score	Fold change	FDR	q-value
rCG34061	4.655	1.421	0	0
56227	4.552	2.168	6.476	3.778
313504	4.525	3.182	5.667	3.778
rCG48508	4.411	1.788	4.318	3.778
315095	4.343	4.724	3.778	3.778
307947	4.2	1.515	6.476	
rCG22278	4.196	1.673	6.253	
rCG26536	4.186	2.56	6.045	
304092	4.167	2.47	5.495	
359725	4.14	1.443	5.333	
360415	4.117	1.995	5.181	

microarray.no



q-value

	score	Fold change	FDR	q-value
rCG34061	4.655	1.421	0	0
56227	4.552	2.168	6.476	3.778
313504	4.525	3.182	5.667	3.778
rCG48508	4.411	1.788	4.318	3.778
315095	4.343	4.724	3.778	3.778
307947	4.2	1.515	6.476	
rCG22278	4.196	1.673	6.253	
rCG26536	4.186	2.56	6.045	
304092	4.167	2.47	5.495	
359725	4.14	1.443	5.333	
360415	4.117	1.995	5.181	

microarray.no



q-value

	score	Fold change	FDR	q-value
rCG34061	4.655	1.421	0	0
56227	4.552	2.168	6.476	3.778
313504	4.525	3.182	5.667	3.778
rCG48508	4.411	1.788	4.318	3.778
315095	4.343	4.724	3.778	3.778
307947	4.2	1.515	6.476	5.181
rCG22278	4.196	1.673	6.253	5.181
rCG26536	4.186	2.56	6.045	5.181
304092	4.167	2.47	5.495	5.181
359725	4.14	1.443	5.333	5.181
360415	4.117	1.995	5.181	5.181

microarray.no



Where do we cut ?

- Commonly used strategies:
 - Sufficiently large fold change
 - Suitably small p-value or q-value
- Any strategy results in a random cut-off
 - There is no perfect cut-off where every gene above the cut-off is truly differentially expressed, while every gene below the cut-off is not differentially expressed

