

# Introduction to data analysis: Supervised analysis

Introduction to Microarray Technology course

May 2011

Solveig Mjelstad Olafsrud

[solveig@microarray.no](mailto:solveig@microarray.no)

*Most slides adapted/borrowed from presentations by Kjell Petersen & Anne Kristin Stavrum (NMC, Bergen)*

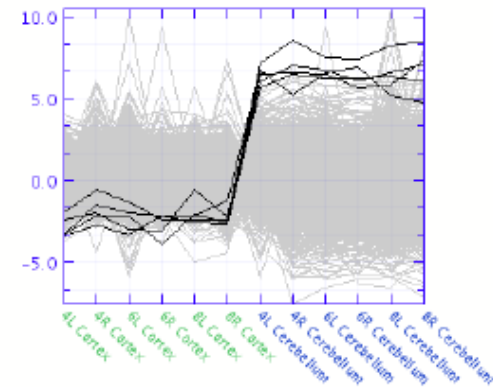
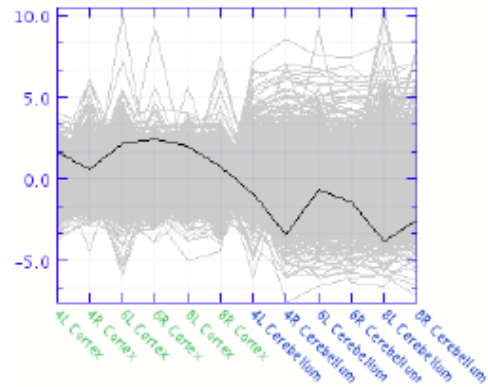
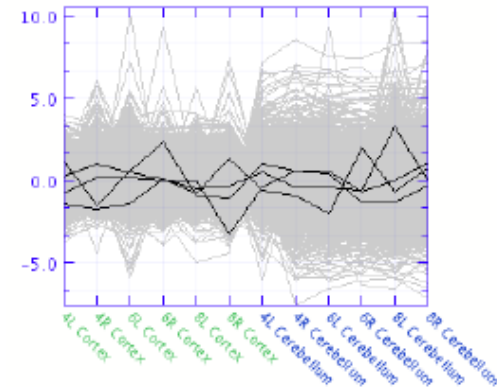
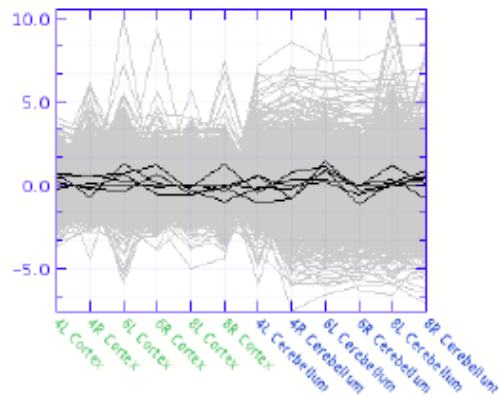


# Overview

- What is differential expression?
- Scoring a gene for differential expression
  - ✓ T-test
  - ✓ SAM
  - ✓ Rank Product
- What does the result of the analysis look like?
- Measure of significance
- We're producing lists
  - ✓ Cut-offs and prioritizations

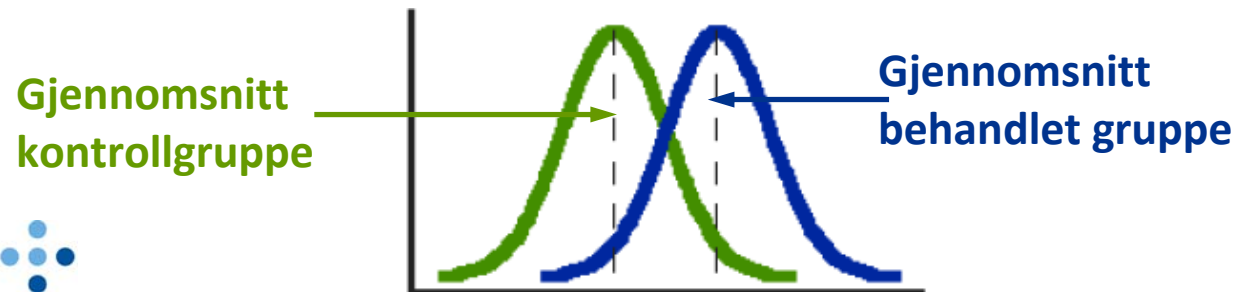


# What do they look like?



# Supervised analysis

- Goal: To find differentially expressed genes between two groups of samples.
  - ✓ Group: Different treatment, timepoint, phenotype etc.
- What is differential expression?
  - ✓ Measurements before (1.5, 0.8, 1.2) and after (2.1, 1.7, 1.5) treatment
  - ✓ We are looking for systematic differences in expression levels
  - ✓ Are the sub-groups significantly different?
    - We need a model to help us decide!



# Differential expression

Basic strategy:

- Find mean expression level for each gene within each of the two groups – calculate fold change
- Find the variance of each gene within each of the two groups
- Use a statistical test for two classes to decide whether expression is different in the two groups

**Simplistic  
formula:**

$$\text{Diff Expr} = \frac{\text{Fold Change between groups}}{\text{Variance within groups}}$$

# 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
- Problem: Only valid with univariate tests
  - ✓ We are testing thousands of genes at the same time – multiple testing
  - ✓ Probably not student t-distribution of scores



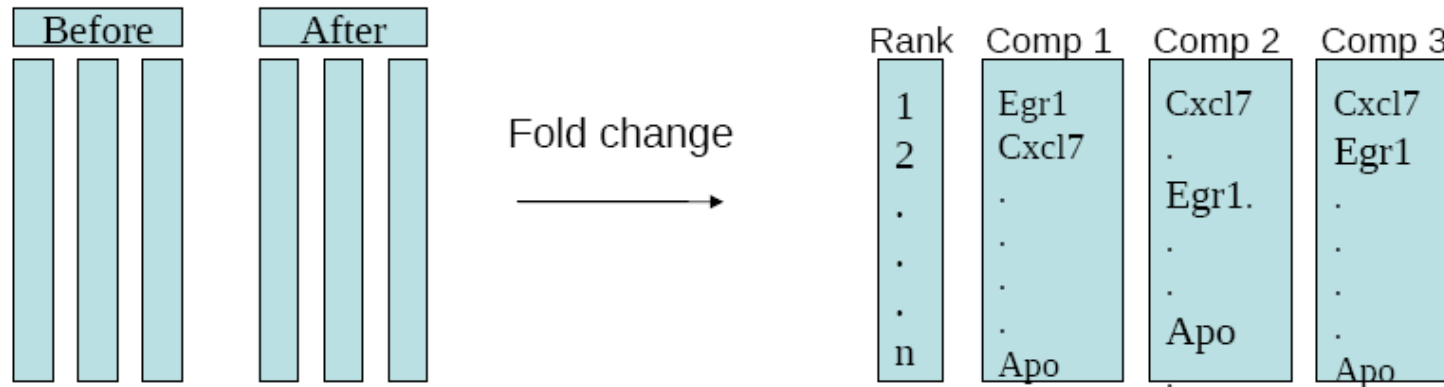
# Significance analysis of Microarrays (SAM)

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

- Assumes normal distribution of data
- Makes no assumption about the distribution of the scores
- **Advantages:**
  - ✓ Eliminate occurrence of accidentally large t-values due to accidentally small within-group variance
  - ✓ Effectively introduces a "fold-change" criterion



# Rank product



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

- Does no assumption on distribution of data
- No calculation of variance across array
- Scores genes according to their rank in multiple comparisons

# Significance of scores

- The 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 for any of the tested genes increases
- The p-value should therefore be corrected for multiple testing
  - ✓ E.g. Bonferroni correction (very strict)
  - ✓ Many other methods to correct for multiple testing, but it is hard to select the right one case by case
  - ✓ Common to work with significance of **lists** instead of single genes



# What does the results look like?

Gen_ID	Score	Fold Change	FDR	q-verdi
rCG34061	4.655	1.421	0	0
56227	4.552	2.168	6.476	3.778
313504	4.525	3.182	5.66	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



# False discovery rate (FDR)

- FDR refers to the number of genes on a ranked gene list that is expected to be false positives
- If the p-value of gene #200 in a ranked list is 0.001 and we have analysed 20 000 genes,
  - ✓ Expect  $20\ 000 * 0.001 = 20$  genes to be false positives
- $FDR = \text{False positives} / \text{Rank} * 100\% = 20/200 * 100\% = 10\%$



# FDR-value and q-value

Gen_ID	Score	Fold Change	FDR	q-verdi
rCG34061	4.655	1.421	0	0
56227	4.552	2.168	6.476	3.778
313504	4.525	3.182	5.66	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



# 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

# Don't use absolute cut-offs

- Use q-values (alternatively FDR or corrected p-values)
  - ✓ To guide your work:
    - FDR estimates is acceptable because we want to screen and search for biological knowledge, looking for emerging pictures/trends
- Never use statistics alone, only as input together with your interpretations of the data/teh biological picture you see
  - ✓ Do you believe it?
    - Enough to do follow up experiments?
- We're working on lists, don't chop off of the top and forget the rest.
  - ✓ Distribution of related genes in the whole list



# Questions?