

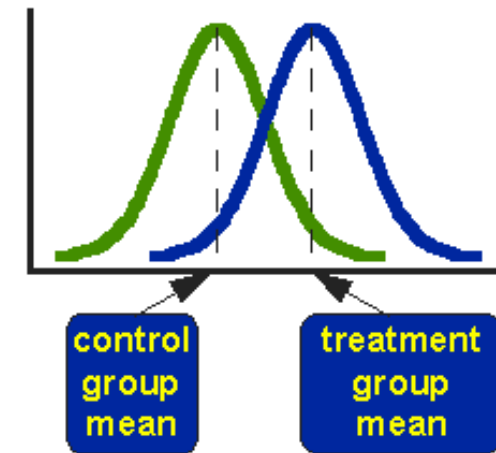
# Differential Expression

# Overview

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

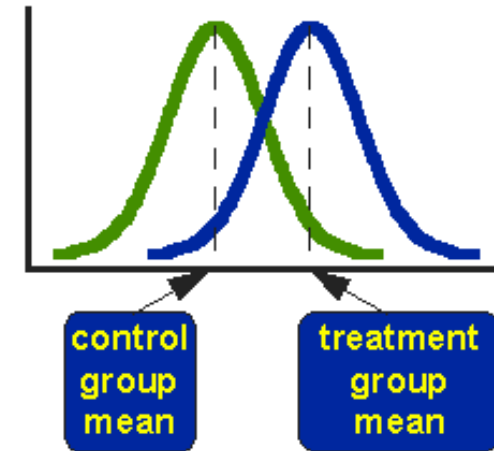
# What is differential expression?

- Measurements before and after treatment
- Before: 1.5, 0.8, 1.2
- After: 2.1, 1.7, 1.5



# What is differential expression?

- Distribution is defined by mean and stdv
- Significant change in mean of measurements
- Typical methods:
  - T-test
  - SAM
  - Rank Product



# How do the methods work?

- Most methods look at each gene by itself and tries to determine whether it has changed significantly between two states
- Some methods assume a certain distribution, while others don't
- They all provide p-values or equivalent that can help us say something about the significance of the results

# 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

# Significance of t-score

- P-value is often given as a statistical measure of the significance of the t-score
- P-value is defined as the probability of a gene obtaining the score by chance

# Significance Analysis of Microarrays (SAM)

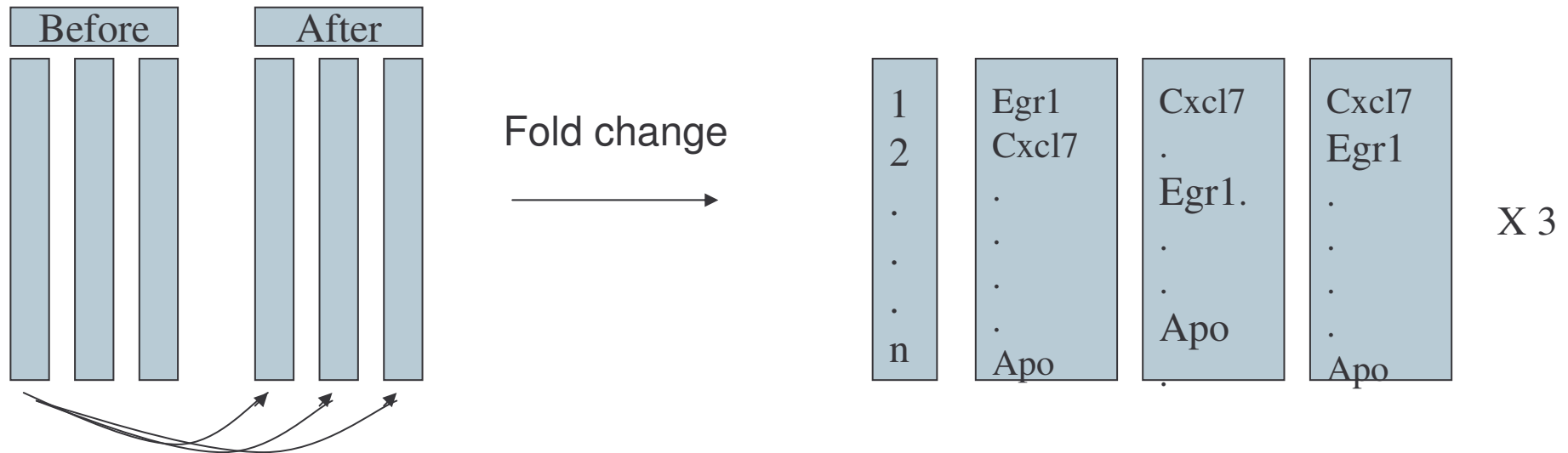
$$s = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}\right)} + B}$$

- Assumes normal distribution of data
- Tolerates more noise in data

# Significance of SAM

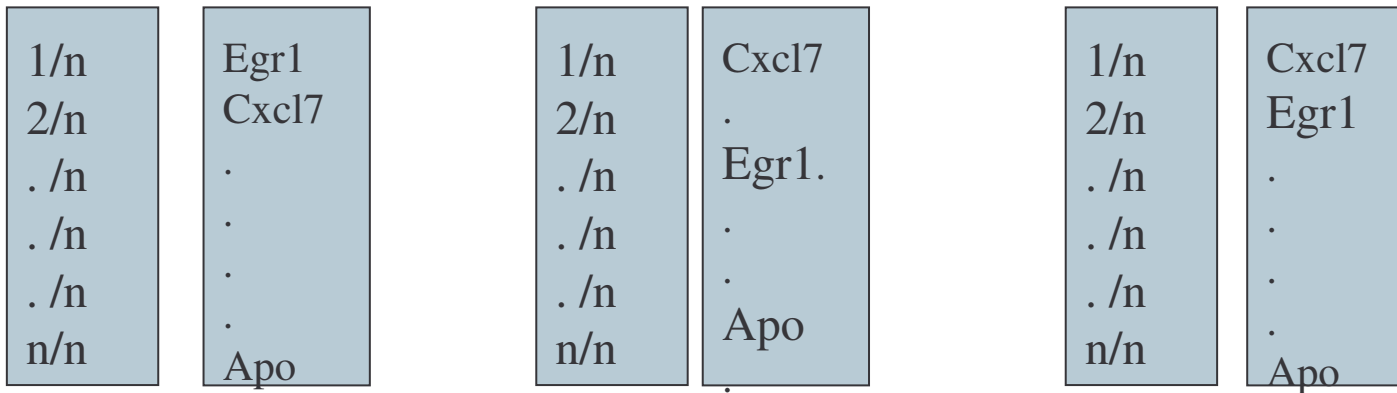
- False Discovery Rate (FDR) or q-value are commonly used for statistical significance
- FDR is the number of false positives you can expect to find in your gene list
- Q-value is the best FDR value seen for all of the possible gene lists a gene can be a part of

# Rank Product



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

# Rank Product



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

# Significance of Rank Product

- Q-value is used as the statistical value to say something about significance of the Rank Product values.

# How do we get significance values?

- If the distribution is known we can read it from a precalculated table
- If the distribution is unknown we must estimate it
  - Randomise data (referred to as permutation or resampling), repeat test and compare result to the original one.
  - Repeat permutation a number of times and count the number of times we get a better result than the original one.

# False Discovery Rate

- 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
- $\text{FDR} = \text{FP}/\text{Rank} \times 100\%$
- FDR refers to the number of genes on a gene list that is expected to be false positive

# Q-value

- FDR is not strictly increasing the further down on a gene list you
- The smallest FDR value that is seen for all the gene lists a gene is a member of is referred to as a q-value.
- 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	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

# 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