

# Differential Expression

microarray.no

Kjell Petersen  
MCB course: Introduction to Integrative Bioinformatics  
November 2009

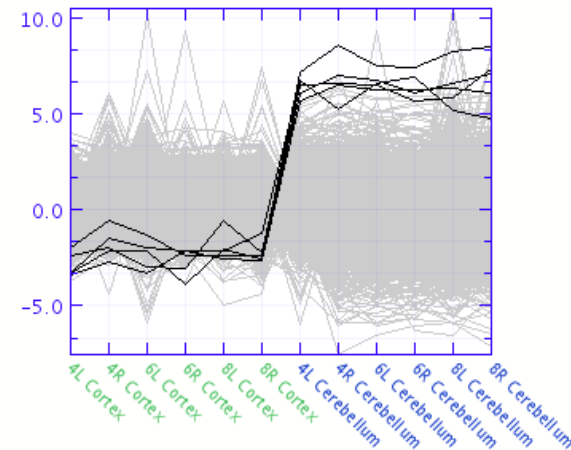
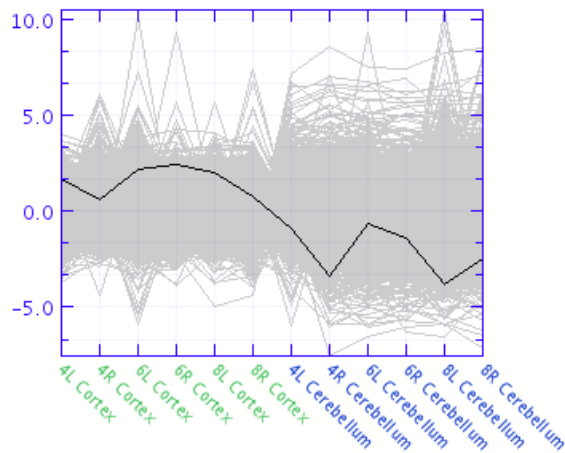
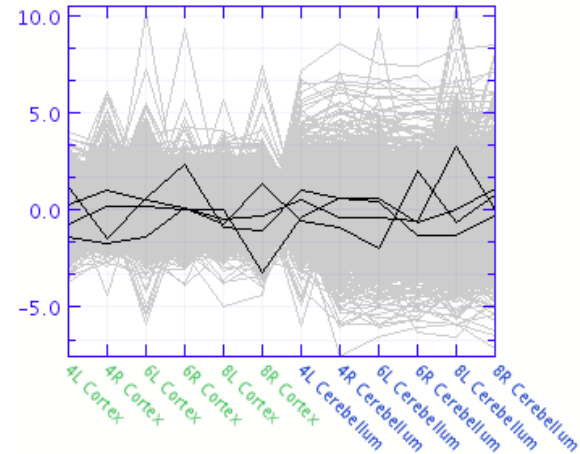
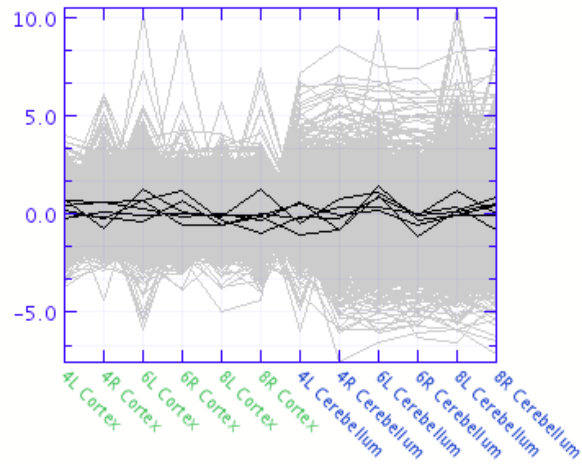
MCB



# Overview

- What is differential expression?
- Models
  - T-test
  - SAM
  - Rank Product
- Measure of significance
- We're producing lists: Cut-offs and prioritizations
- Gene Ontology and overrepresentation analysis

# What do they look like ?



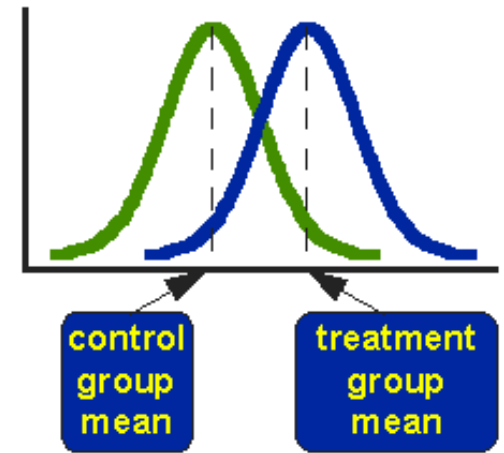
microarray.no

# Simplistic formula

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

# What is differential expression?

- Example: 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

**Nonparametric 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

Modified slide from Huber, Gentleman and Heydebreck

# 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

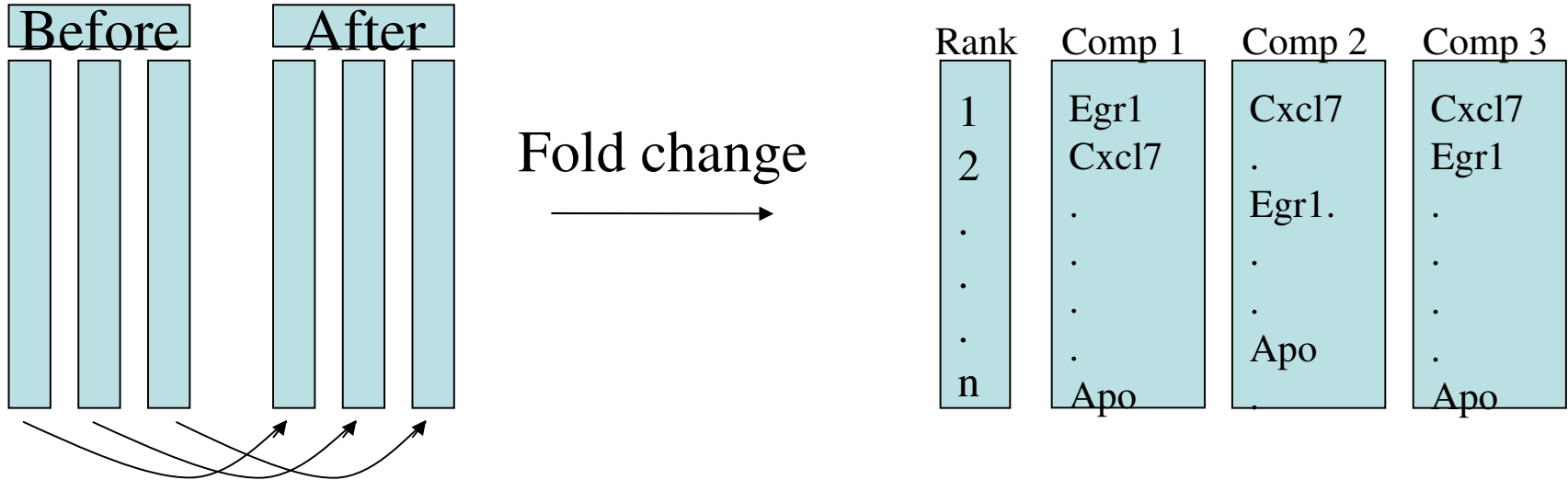
Modified slide from Huber, Gentleman and Heydebreck

# 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

1/n
2/n
./n
./n
./n
./n
n/n

Egr1
Cxcl7
.
.
.
.
Apo

1/n
2/n
./n
./n
./n
./n
n/n

Cxcl7
.
Egr1.
.
.
Apo
.

1/n
2/n
./n
./n
./n
./n
n/n

Cxcl7
Egr1
.
.
.
.
Apo

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

# 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 **ranked gene list** that is expected to be false positive
- If the p-value of gene nr 100 on a **ranked 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	1.005	5.181

## q-value

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

microarray.no

MCB

S

## 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	

## 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

MCB

S

# 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

MCP

S

## 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 interpretation of the data/the biological picture you see
  - Do you believe it?
    - Enough to do follow-up experiments?
- We're working with lists, don't chop of the top and forget the rest.
  - Distribution of related genes in the whole list

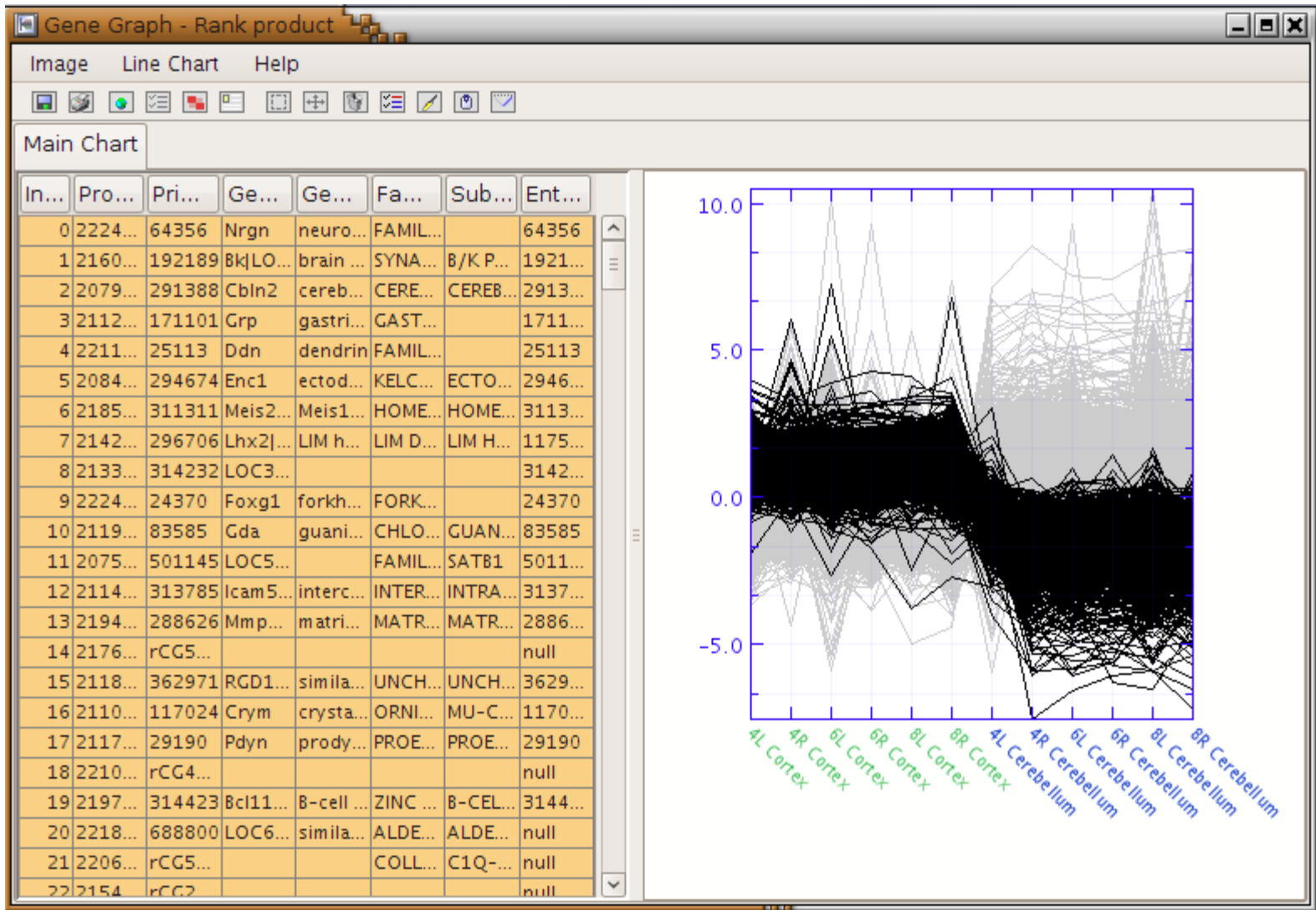
# Questions so far?

microarray.no

MCB



# So here you are



microarray.no

## Gene lists

	A	B	C
1	probe	sym	qValue
2	5050241	ABHD3	0.03372434
3	60609	AFG3L1	0.0339233
4	70239	AGGF1	0.03621495
5	1820044	ALDH9A1	0.00847458
6	6900674	ANKRD10	0.0075188
7	4880132	ARGLU1	0.03443526
8	160561	ARHGAP17	0.01052632
9	4250008	NA	0.00793651
10	1070189	AXIN2	0.0106383
11	460711	BAT2D1	0.03982301
12	4230528	BTBD1	0.00763359
13	4730369	BTG1	0.03324808
14	1260575	C13orf23	0.01156069
15	6660711	C13orf23	0.01442308
16	3870112	TSEN15	0.02075472
17	6280626	C21orf66	0.03402367
18	3400138	CASC4	0.02
19	5810750	CBR4	0.02640845
20	3940451	CCDC131	0.04246285
21	70603	CCDC50	0.01898734
22	2370204	CCNC	0.02787456
23	2190671	CCR3	0.01627907
24	2120451	CD46	0.03473945
25	3130356	CDV3	0.04334038
26	5720746	CDV3	0.02881356
27	1340470	CECR5	0.04255319
28	4490154	CEP192	0.02768166
29	7040280	CLINT1	0.01930502
30	4640484	CNOT8	0.03144654
31	670026	CRLF3	0.04270833
32	7210241	CSNK1A1	0.03811659
33	4210064	CSNK1G3	0.0328125
34	770424	CTGLF3	0.02669039
35	2230215	ALG13	0

- Long list of differentially expressed genes
- Possibly hundreds of papers describing the functions of the genes
- Misleading names
- Different names in different organisms

# Goal of GO Consortium



(<http://www.geneontology.org/>)

- Produce a controlled vocabulary describing aspects of molecular biology, that could be applied to all organism.
- Describe gene products using vocabulary terms (annotation).
- Develop tools:
  - to query and modify the vocabularies and annotations

# How does GO work?

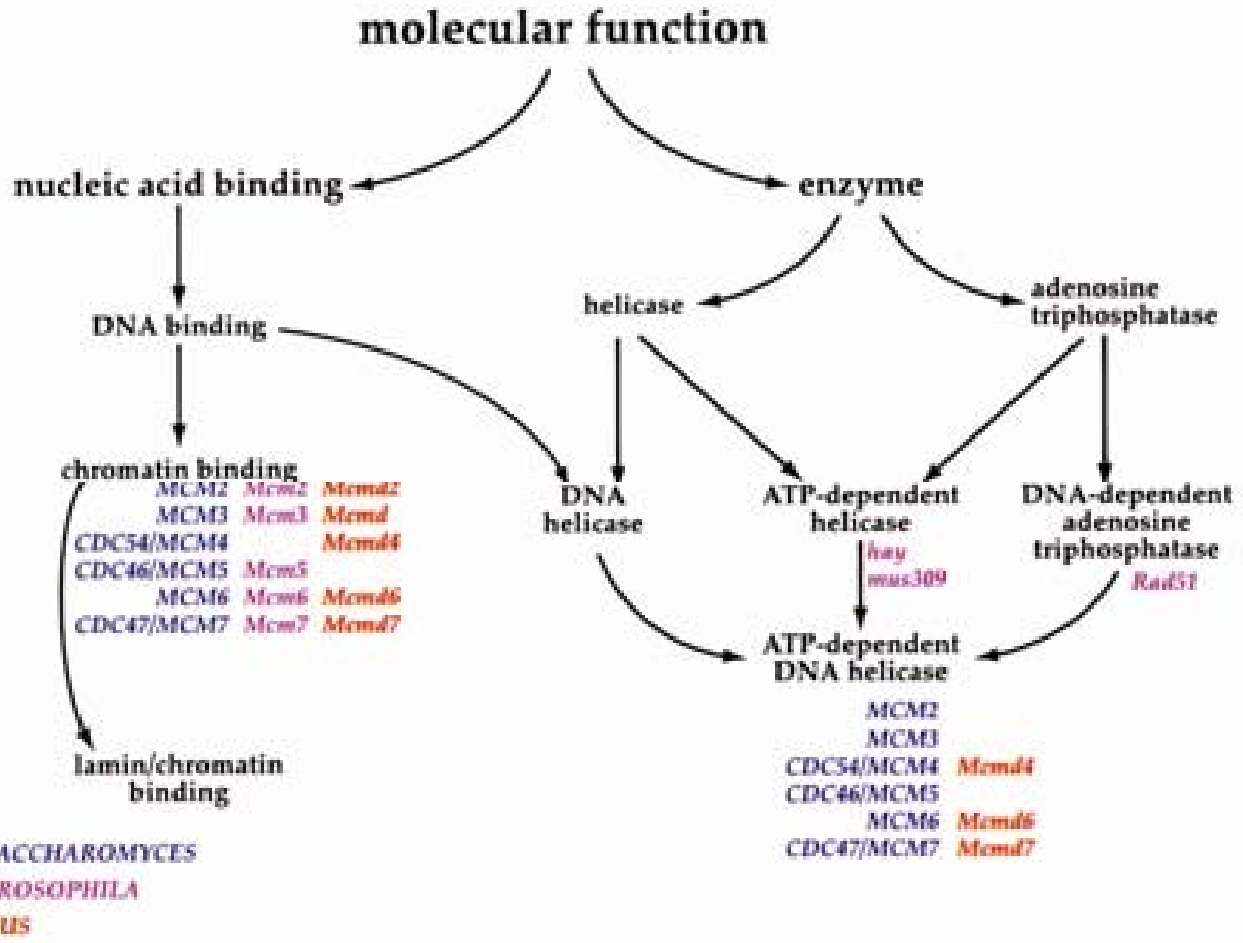
What information might we want to capture about a gene product?

- **What** does the gene product do?
- **Why** does it perform these activities?
- **Where** does it act?

# The Gene Ontology (GO)

- Molecular function:
  - Gene product at biochemical level.
- Biological process:
  - Cellular events to which the gene product contributes.
- Cellular component:
  - Location or complex of gene/protein.

b



# Browsing GO in J-Express

**GO DAG SpotPix Data**

File View Help

- chromosome segregation 2 2 0
- cell cycle 98 149 51
- meiotic cell cycle 0 7 7
- endomitotic cell cycle 0 0 0
- abortive mitotic cell cycle 0 0 0
- mitotic cell cycle 2 69 67
- regulation of ubiquitin ligase activity during mitosis 0 0 0
- interphase of mitotic cell cycle 0 25 25

Locate:

Name: mitotic cell cycle  
GO: GO:000278

**Selection**

Automatic Selection Update  
 Recursive Selection  
Maximum Members:

**Mapping File**

Data Identifier Code:   
 Use Synonyms

**Tags**

Parameter	Value
xref_analog	Reactome:212083
xref_analog	Reactome:221841
xref_analog	Reactome:230408

**Data Set**

Index	ID	Name	Replicat...	Used re...	Groups
2359	244767	barren ...	1	1,1,1,0,...	
4197	165921	centro...	1	1,1,1,0,...	

**Gene Graph - SpotPix Data**

Image Line Chart Help

**Main Chart**

Ind...	Na...	ID	Re...	Us...	Gr...
41...	ce...	16...	1	1...	
41...	U...	46...	1	1...	
41...	F...	50...	1	1...	
42...	pr...	42...	2	2...	
42...	W...	37...	1	1...	
42...	ad...	24...	1	1...	
42...	co...	70...	1	1...	
42...	Tr...	70...	1	1...	
42...	de...	42...	1	1...	
42...	ce...	51...	1	1...	
42...	nu...	24...	1	1...	
42...	pr...	50...	1	1...	
42...	Tr...	50...	1	1...	
42...	cy...	84...	1	1...	
42...	In...	42...	1	1...	
42...	22...	22...	1	1...	
42...	hy...	37...	1	1...	
42...	tr...	82...	1	1...	
42...	U...	H...	5	5...	
42...	ne...	13...	1	1...	
42...	ch...	51...	1	1...	
42...	R...	28...	1	1...	

daughter cells. In some variant cell cycles nuclear replication or nuclear division may not be followed by cell division, or G1 and G2 phases may be absent." [GOC:mah, ISBN:0815316194, Reactome:69278]

# Overrepresentation of GO terms

- We have a subset of genes
  - List of differentially expressed genes
  - List of genes that cluster together
- Which biological processes do these genes take part in?
- Is there an over-representation of the number of genes belonging to a particular biological process, compared to what could be expected?

## Question

- If we look at the dataset containing all of our genes and see that 10% of these belong to cell cycle. We then do a differentially expressed genes analysis and get a list of genes we believe are significantly changed.
- How many of the genes in the gene list do you expect belong to cell cycle?

# Setup

- We name
- And the
- we extract
- want to c

The screenshot shows a software window titled "Project". Inside, there is a tree view with the following structure:

- New Project
  - Breast cancer (with a box labeled "Reference data" pointing to it)
  - SAM (with a box labeled "Test data" pointing to it)

At the bottom of the window, there is a summary table:

Rows	5494
Columns	11
<input type="checkbox"/> Scale Relative To Parents	

for test data  
s, the dataset  
nd that we  
nce data

# Gene Ontology Analysis

Project

- New Project
- Breast cancer
- SAM

Reference data

Test data

Statistical comparison between the two GO components

GO Comparison result

GO term	Exact test p value	Genes in selecti...	Genes in reference	Enrichment
mitotic cell cycle	4.782E-13	15	135	14.305
mitosis	1.182E-10	11	82	17.27
M phase of mitotic cell c...	1.677E-10	11	85	16.661
cytokinesis	1.88E-10	11	86	16.467
cell division	1.88E-10	11	86	16.467
cell cycle	2.375E-9	18	397	5.837
spindle	2.444E-7	6	29	26.636
microtubule cytoskeleton	1.853E-6	7	69	13.061
mitotic spindle elongation	8.741E-6	3	3	128.743
mitotic spindle organizat...	2.42E-5	3	5	77.246
regulation of cyclin depe...	5.091E-5	4	23	22.39
G2/M transition of mitoti...	7.81E-5	4	26	19.807
cyclin-dependent protei...	1.202E-4	3	10	38.623
kinetochore	2.314E-4	3	13	29.71
		16	16	24.139
		3	3	85.828
		21	21	18.392
		21	21	18.392
		56	56	9.196
		5	5	51.497
		101	101	6.373
		276	276	3.732
		7	7	36.784
		4801	4801	1.399
		1419	1419	1.815
		50	50	7.725
		16	16	16.093
		5	5	4.049
		2	2	15.146
		3	3	6.997

Rows: 5494

GO DAG SAM

- endomitotic c...
- centrosome c...
- M phase 0 0 1.0
- mitotic cell cycle 1 15 14 4.782E-13
- regulation of ubiquitin ligase activity during mitotic cell cycle 0 0 1.0
- interphase of mitotic cell cycle 0 4 0.001
- regulation of progression through mitotic cell cycle 0 0 1.0
- M phase of mitotic cell cycle 0 11 11 1.677E-10
- regulation of progression through cell cycle 3 4 1.0 0.3

- M phase 2 2 0
- mitotic cell cycle 2 135 133
- regulation of ubiquitin ligase activity during mit...
- interphase of mitotic cell cycle 0 56 56
- regulation of progression through mitotic cell cycle 0 0 0
- M phase of mitotic cell cycle 5 85 80
- regulation of progression through cell cycle 137 14

mitotic chromosome co...	0.008	2	16	16.093
cell proliferation	0.009	5	159	4.049
regulation of mitosis	0.009	2	17	15.146
wound healing	0.011	3	56	6.997

Locate: Regular Expression

Name: mitotic cell cycle  
GO: GO:0000278

Selection:  Automatic Selection Update,  Recursive Selection, Maximum Members: 100

Mapping File: course.txt, Data Identifier Column: ID,  Use Synonyms, Map Data Set

Tags: xref\_analog Reactome:69278

Definition: "Progression through the phases of the mitotic cell cycle, the most common eukaryotic cell cycle, in which a cell is duplicated without changing ploidy, comprises four successive phases called G1, S, G2, and M." [ISBN:0815316194, SGD:mah]

Index	ID	Name	Replica...	Used re...	Groups
2	25k_7...	Hs.344...	1	1,1,1,1...	I
5	25k_8...	MAD2...	1	1,1,0,1...	I
7	452363	kinesin...	1	1,1,1,1...	I
12	435076	centro...	1	1,1,1,1...	I
13	129865	serine...	1	1,1,1,1...	I
19	744047	polo-lik...	1	1,1,1,1...	I
21	769921	ubiquiti...	1	1,1,1,1...	I

Locate: Regular Expression

Name: mitotic cell cycle  
GO: GO:0000278

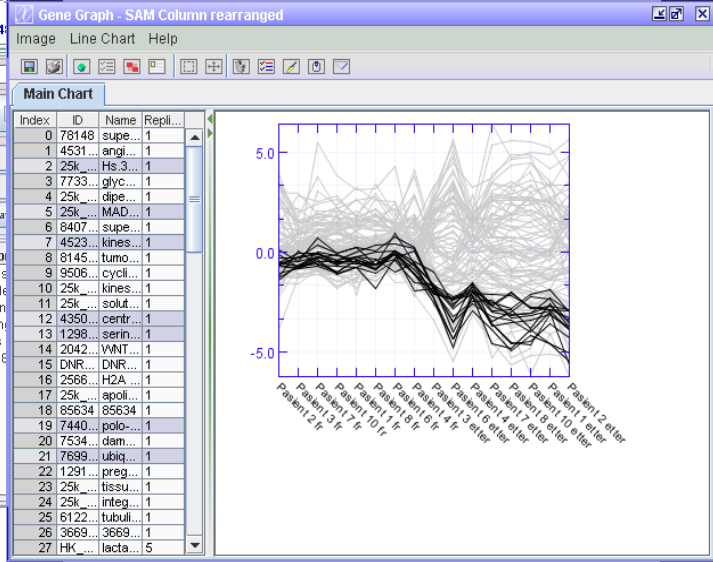
Selection:  Automatic Selection Update,  Recursive Selection, Maximum Members: 100

Mapping File: course.txt, Data Identifier Column: ID,  Use Synonyms, Map Data Set

Tags: xref\_analog Reactome:60278

Definition: "Progression through the phases of the mitotic cell cycle, the most common eukaryotic cell cycle, in which a cell is duplicated without changing ploidy, comprises four successive phases called G1, S, G2, and M." [ISBN:0815316194, SGD:mah]

Index	ID	Name	Used re...	LLID	Groups
245	784744	M pha...	1,1,1,1...	10200	I
267	25k_8...	NIMA (...)	1,1,1,1...		I
287	115443	In multi...	1,1,1,1...	In multi...	I
302	853066	853066	1,1,1,1...	9918	I
559	25k_7...	BUB1 b...	1,1,1,1...		I
736	25k_1...	serine...	1,1,1,1...		I
891	824897	CDK5 r...	1,1,1,1...	In multi...	I
903	358736	G1 to S...	1,1,1,1...	23708	I
908	897642	v-abl A...	1,1,1,1...	25	I
930	295985	cyclin...	1,1,1,1...	1021	I



# Acknowledgements

- Most slides adapted from Anne-Kristin Stavrum

# Questions?

microarray.no

MCB

