20.109
LABORATORY FUNDAMENTALS
IN BIOLOGICAL ENGINEERING

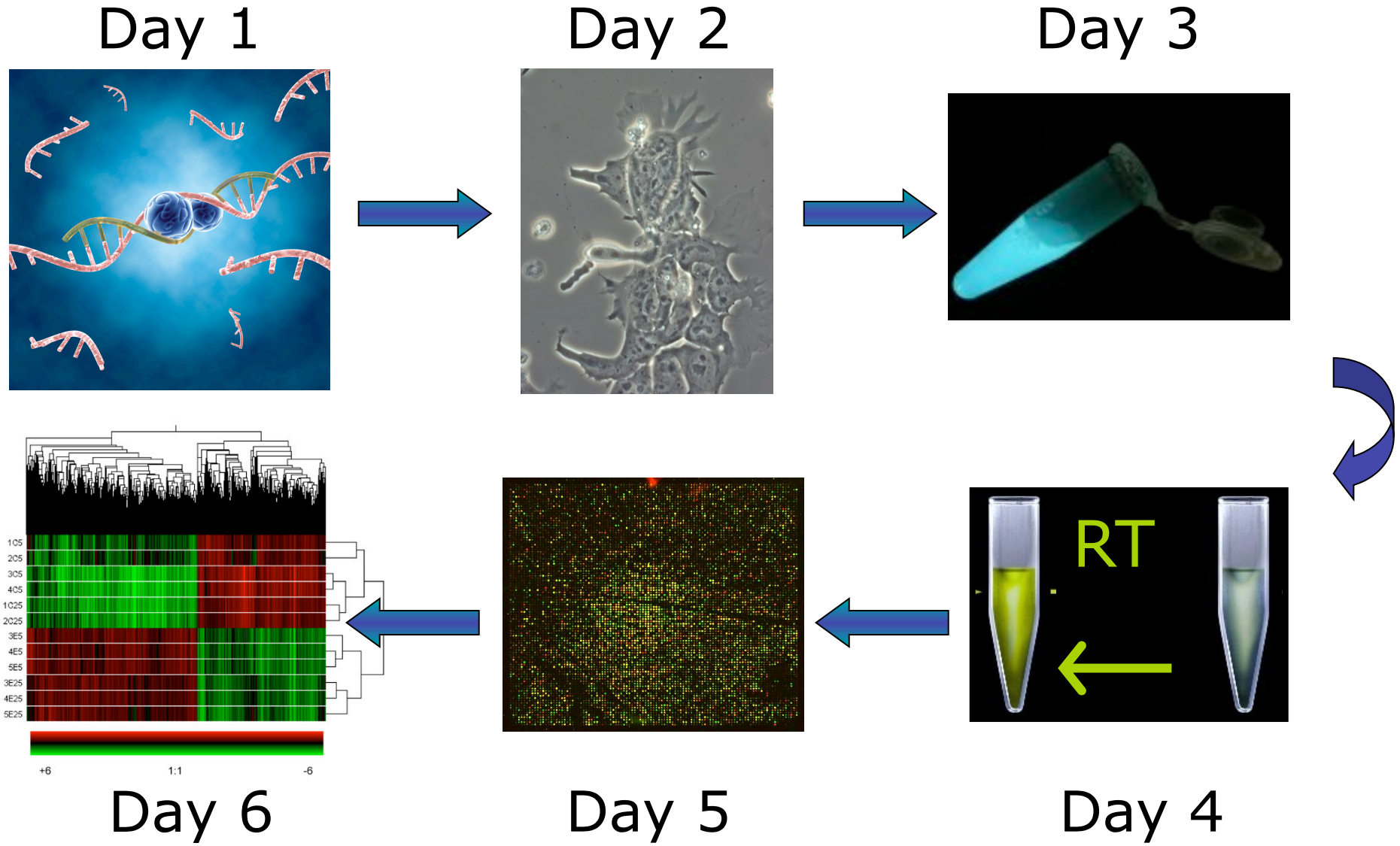
Module 2
Expression engineering
Lecture # 5: Microarray analysis

Peter Svensson
April 7th 2009

Motivation

- Course to give insights into projects conducted at MIT
- Part of a research project in the Samson lab
  - We don’t know what the results will be.
  - Human cell lines lacking these proteins are sensitive to DNA damaging agents.
  - Possibly the sensitivity is mediated through transcription

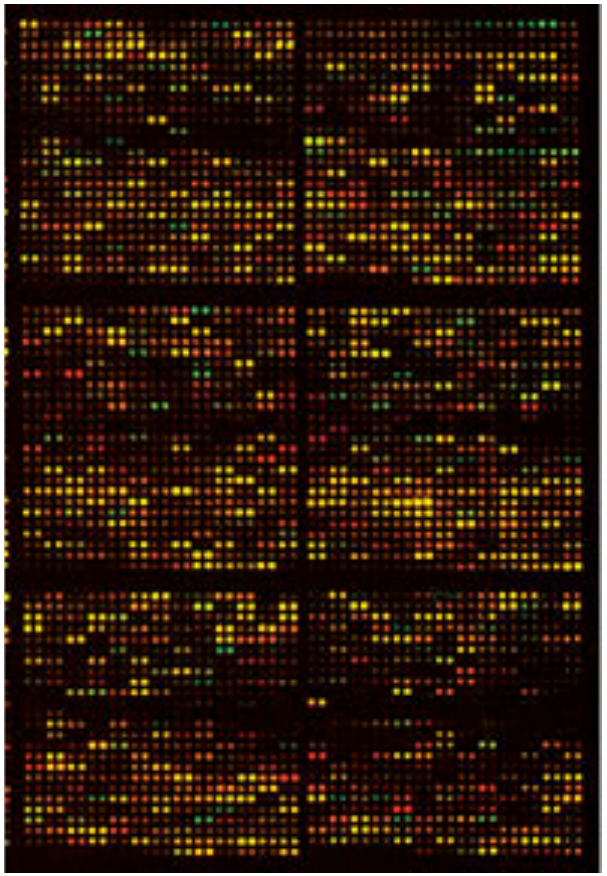
Expression Engineering Experiment



Slide adapted from Natalie Kuldell

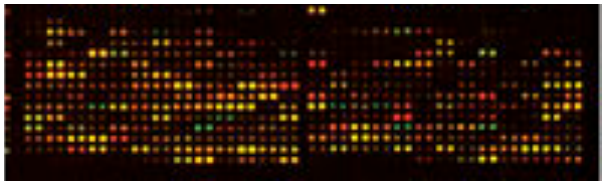
Microarray analysis

- An example of application
- Preprocessing of the data
- Identification of differentially expressed genes
  - Biological and statistical significance
- Interpretation of the data
- Communication of results

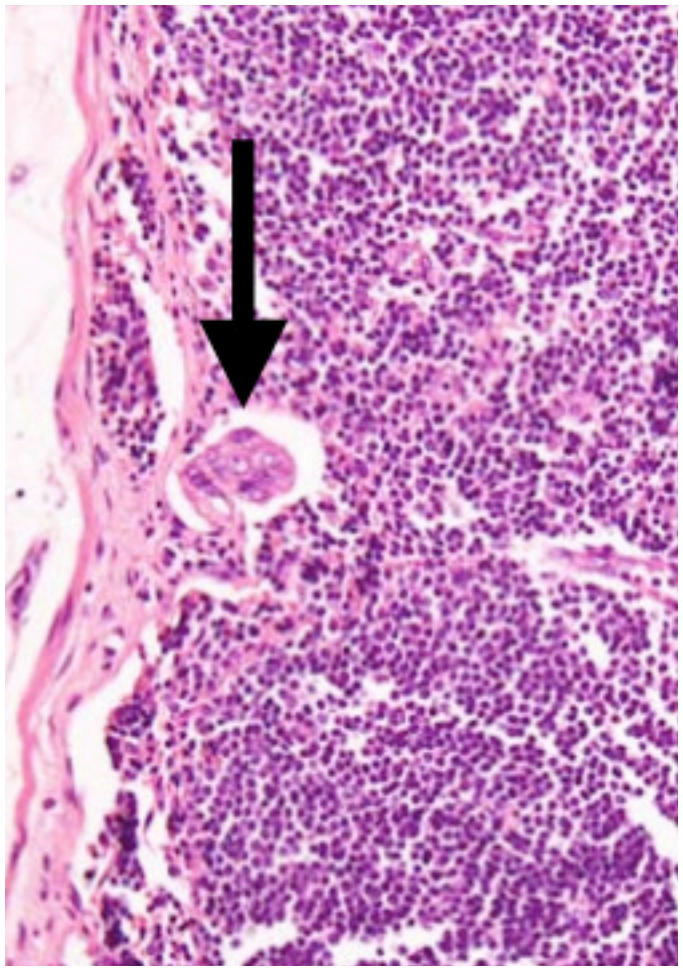


Purpose and use of microarrays

- Probe whole genome
  - Only half of the genes have a name.
  - Few thousand genes where function is known.
- Can be used to associate a function with genes
- Determine a gene expression signature for a disease state
- Classify patients/tumors.
  - Some subtypes of cancers are difficult for a pathologist to distinguish.



Diagnostic Tool: is it cancer?



Lymph node metastasis
Kondo, Cases Journal 2009

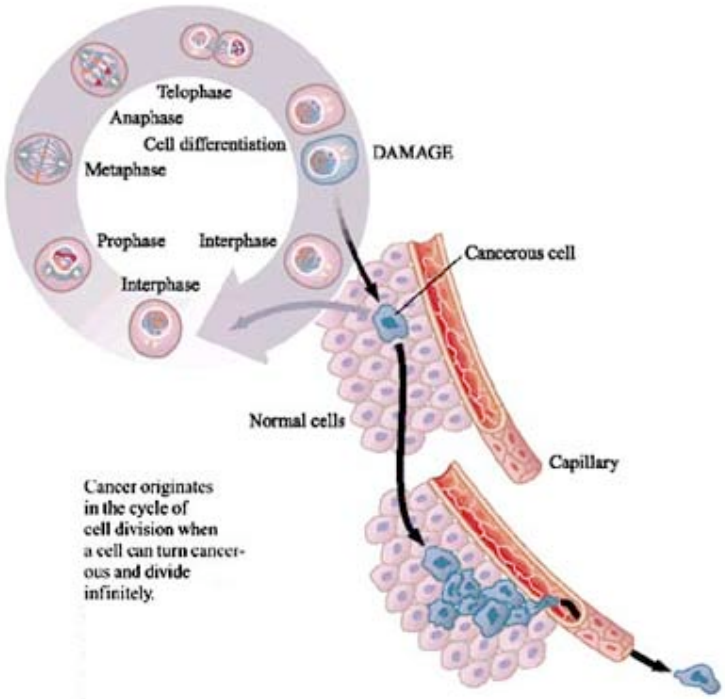
1 in 3 women develop a cancer in their lifetime
Breast cancer most prevalent,
2.5 million women in U.S.

5 to 10% are related to genetics and family history of breast cancer.

Available treatments:
surgery, chemo, radiation, hormone

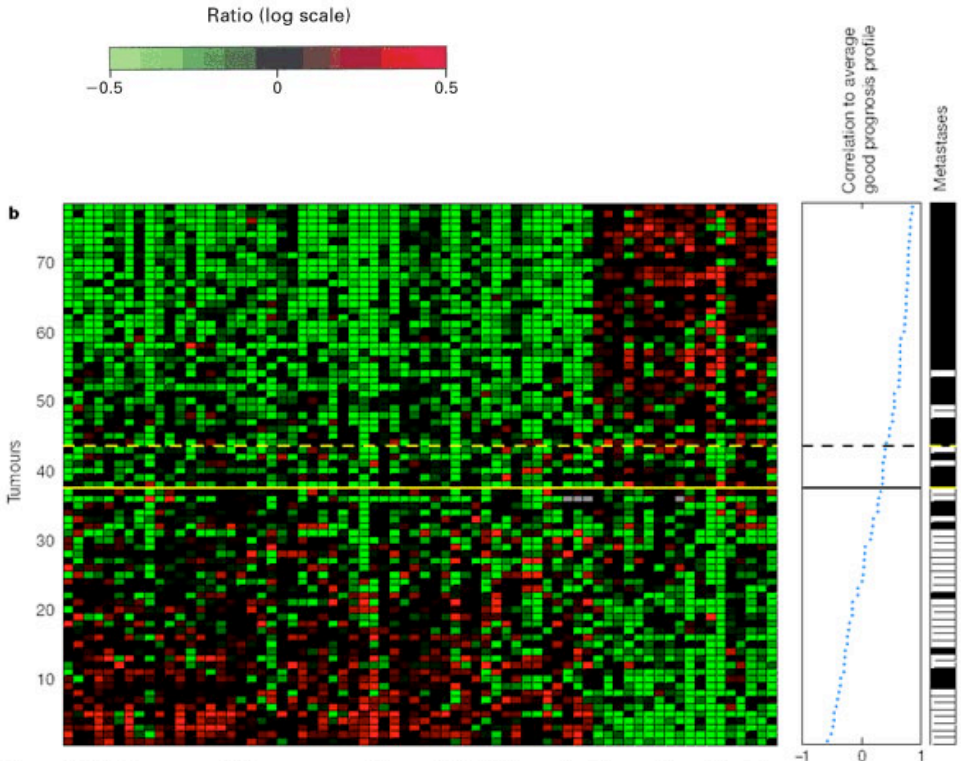
American Cancer Society

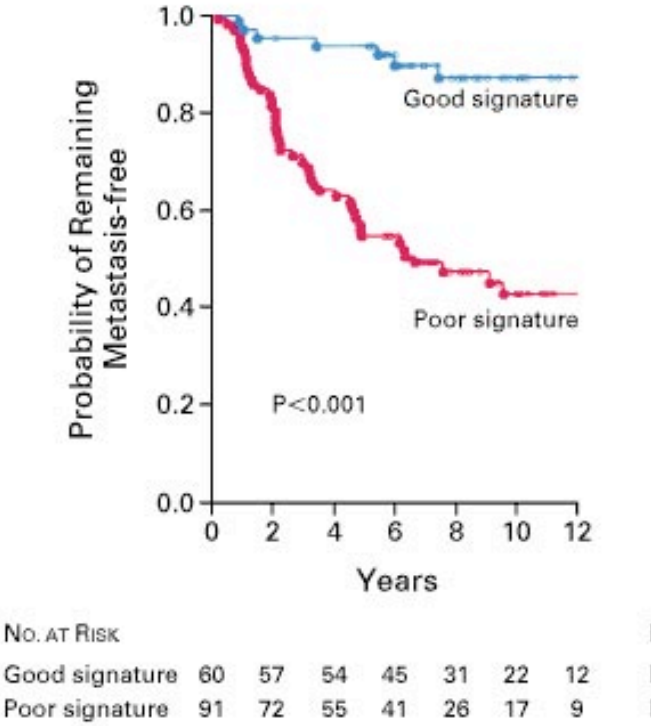
Treatment Evaluation Tool: how likely is it to spread?



- Lymph node metastasis appear in 20-30% of cases
- Breast cancer patients overtreated.
  - Everybody received painful therapy but 70-80% would survive without
- Motivation: Predict reoccurrence of breast cancer and only treat patients with risk of metastasis.

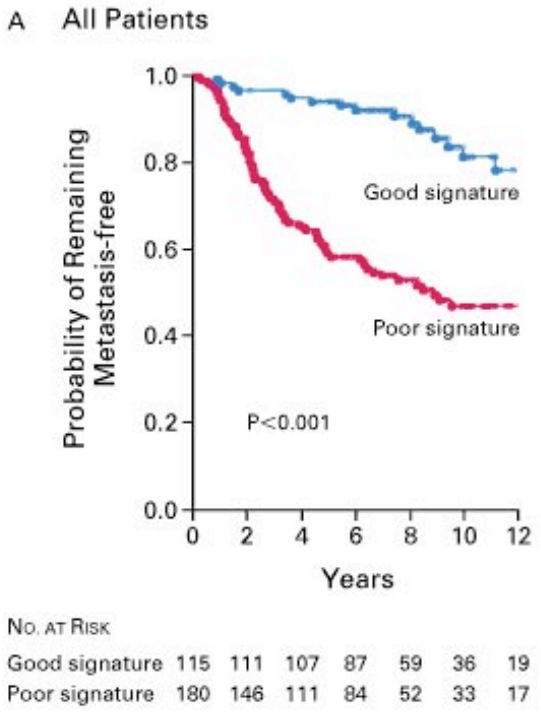
Finding 70-gene signature in set of patients

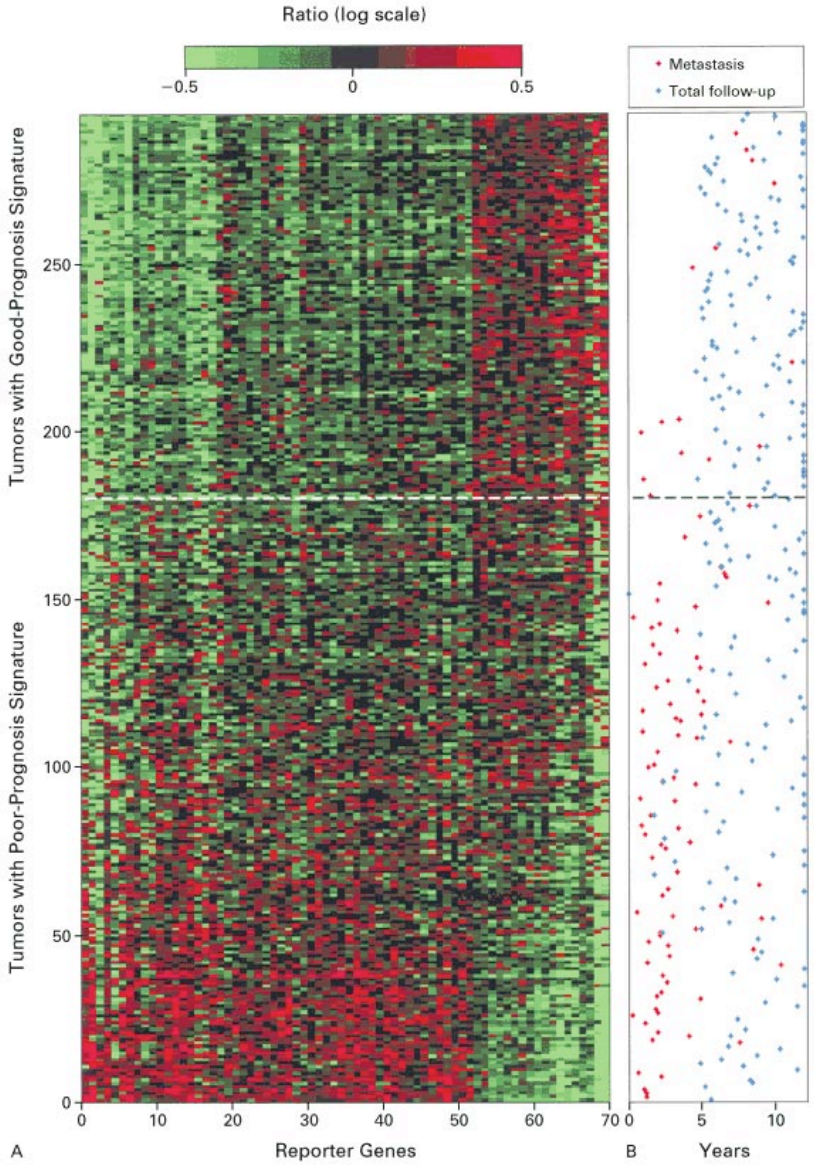




van't Veer et al. 2002 Nature

Validating signature in second set of patients

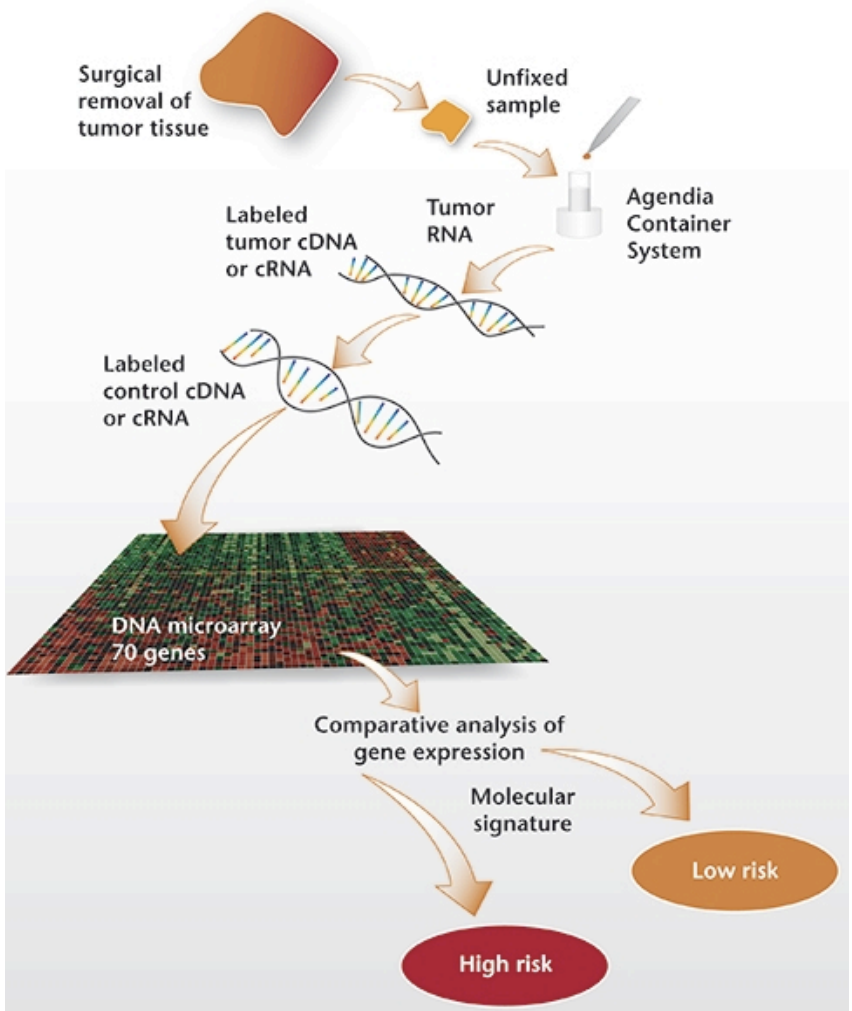




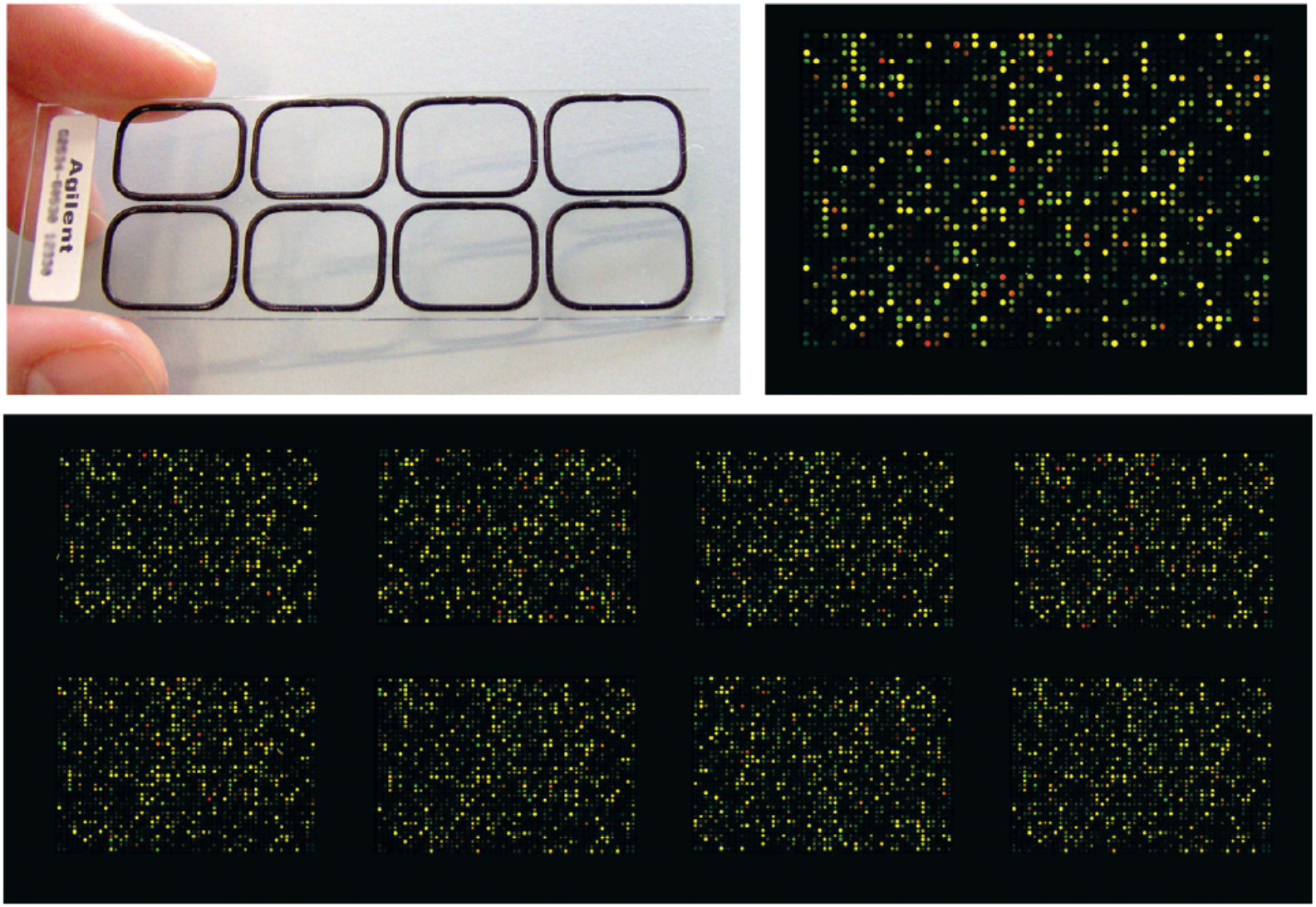
van de Vijver et al. 2002, NEJM

In clinical practice

- Breast cancer patients are routinely tested to determine treatment
- FDA approved



FDA approved "MammaPrint" assay



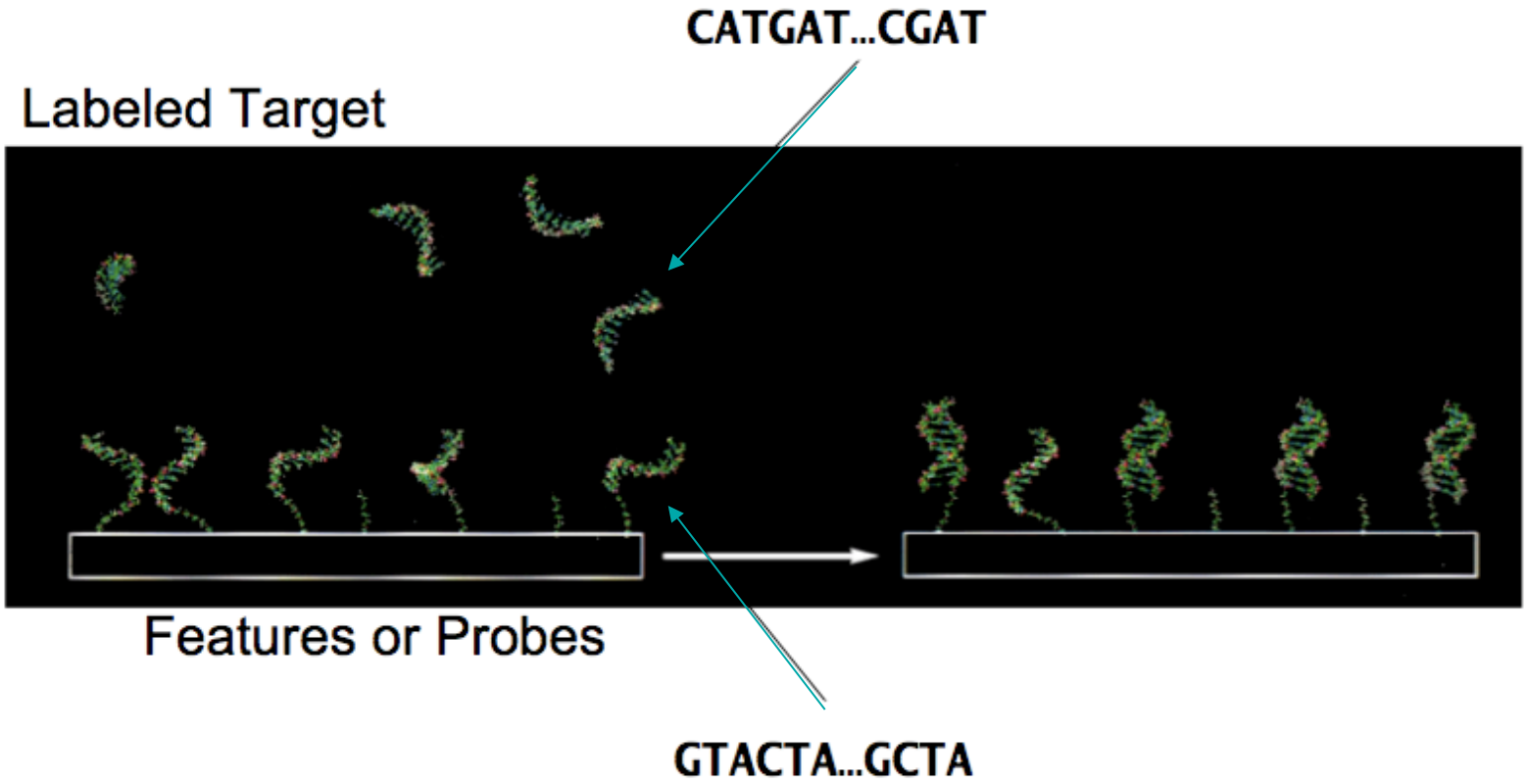
developed/marketed by Agendia

Microarray analysis

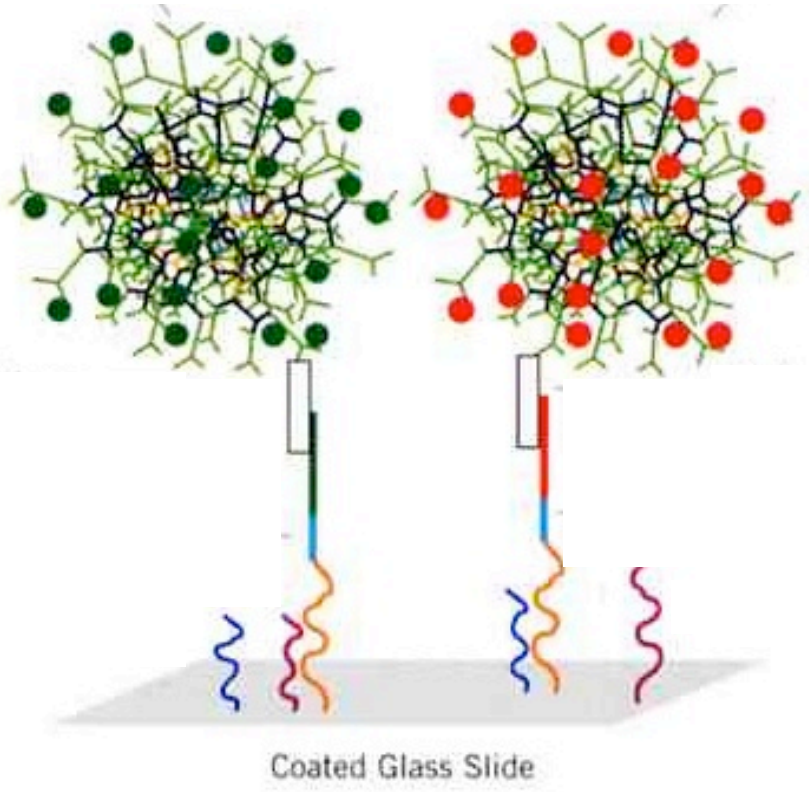
- An example of application
- Preprocessing of the data
- Identification of differentially expressed genes
  - Biological and statistical significance
- Interpretation of the data
- Communication of results

Recap microarrays

- Gene expression microarrays
  - SNP, copy number, promoter region arrays…

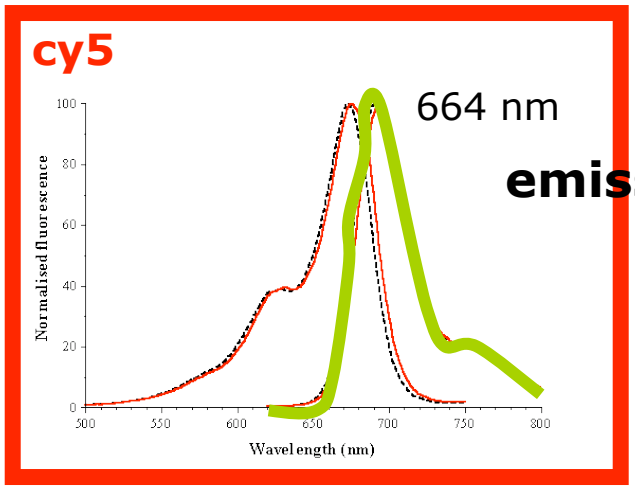


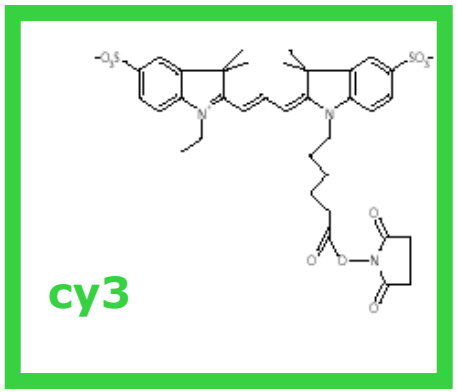
Indirect labeling of probes

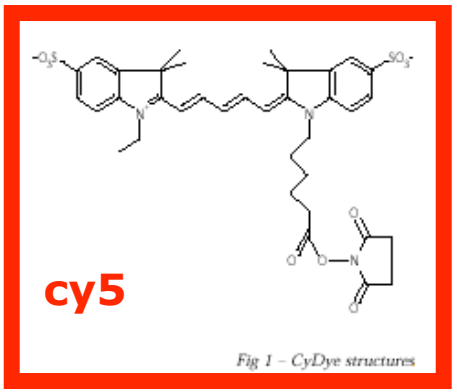


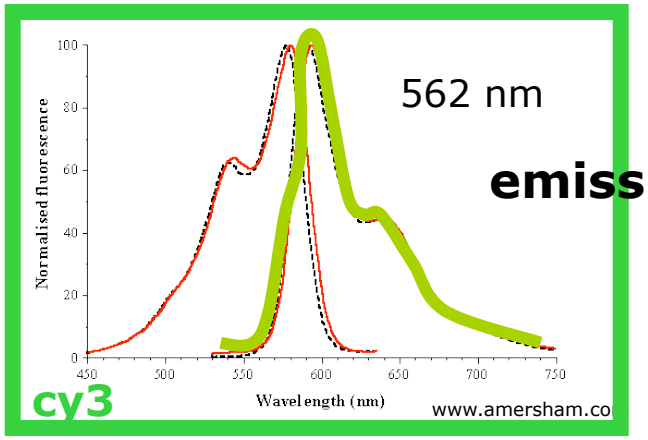
http://www.genisphere.com/about_3dna.html

Commonly used dyes









Differential dye incorporation
cy5 less well than cy3

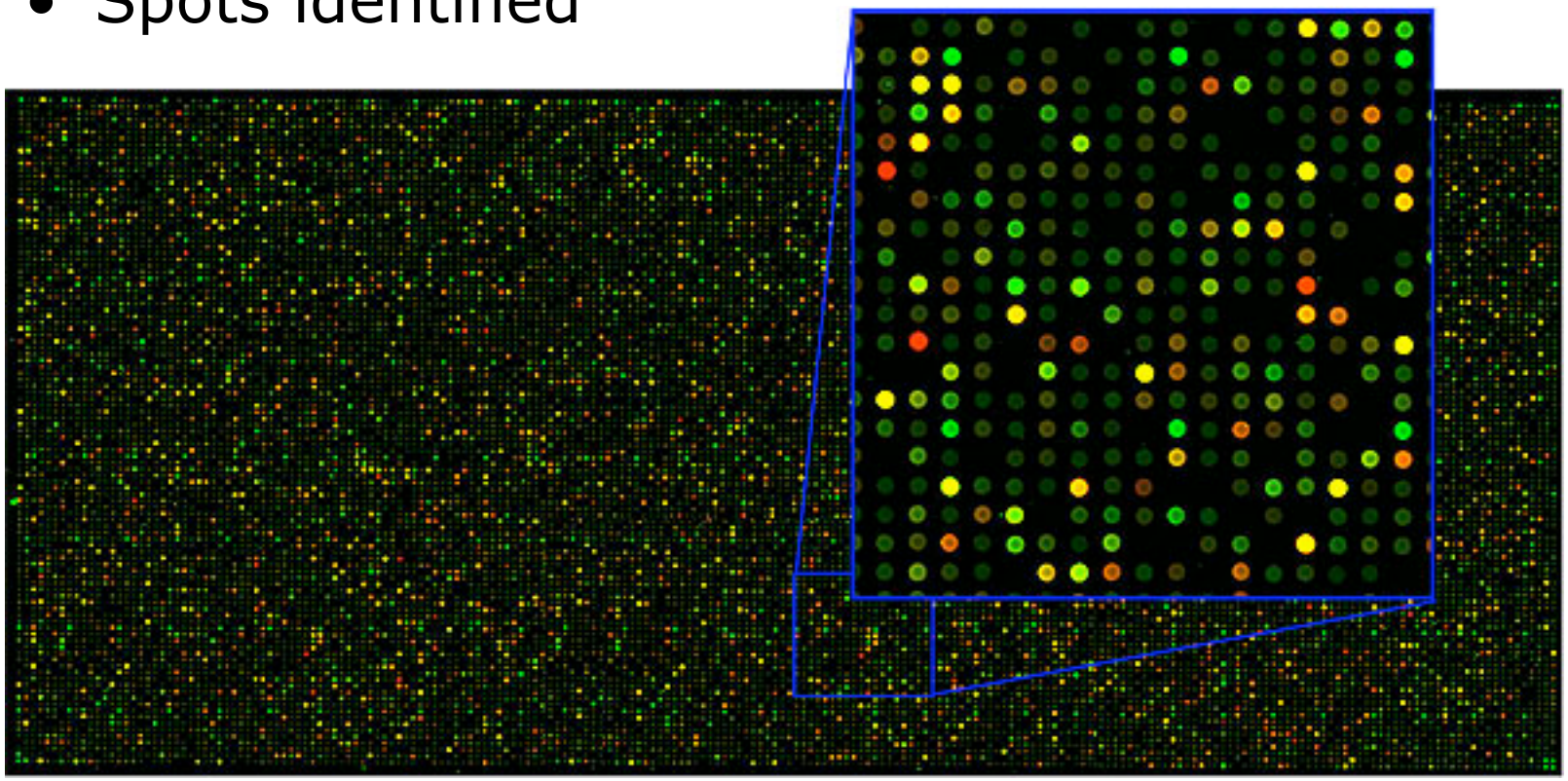
Light sensitivity: cy5 more easily degraded

For this reason, we do dye-swap experiments to rule out dye-specific effects.

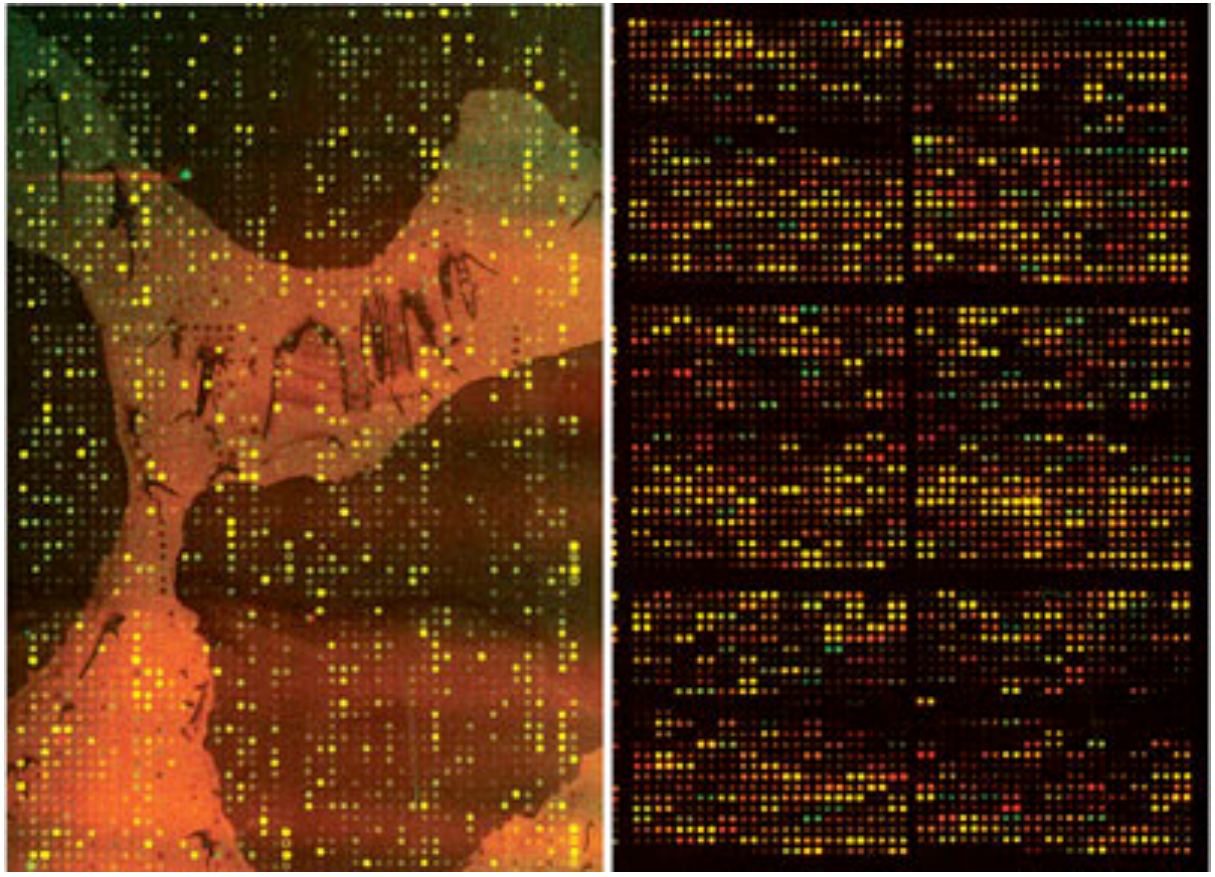
Slide adapted from Rebecca Fry

Scanning and quantification

- Arrays are scanned at two wavelengths.
- Pixel intensities quantified
- Spots identified



Quality control



Noise

- Experimental design important

- Duplicates
  - Biological/technical
- Same gene targeted twice.
  - One group label experimental RNA green
  - One group label experimental RNA red
- Get data from both

Preprocessing: Background correction

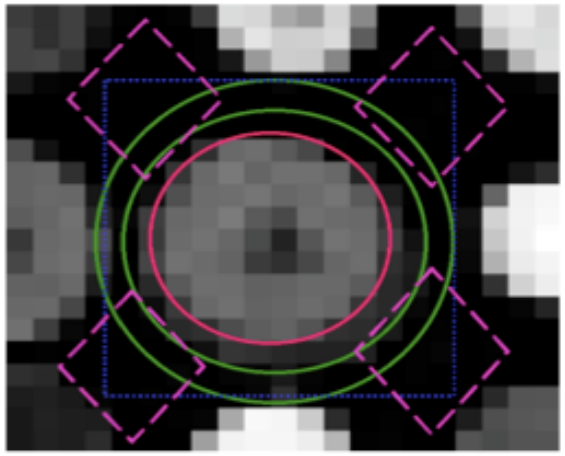
- Several causes of background
  - Uneven washing, residual dye.
  - Particles on array.

- Image analysis program will estimate background signal

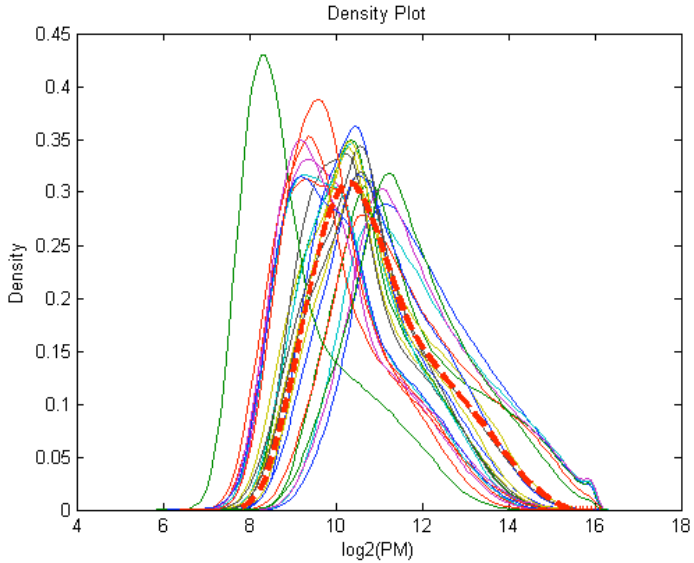
---- GenePix

---- QuantArray

---- ScanAnalyze



Preprocessing: Normalization



- Dye specific effects.
- Slide specific effects.

- We expect average expression values and variation to be the same for all arrays.
  - If one array has lower intensities: it will seem like most genes are down-regulated
  - If variation is larger on one array: many genes will seem changed.

- Other ways of normalization
  - Add RNA controls of known concentrations prior to hybridization
  - Normalize to ‘house-keeping’ genes

Large matrices of data

- Software to analyze
  - Excel will suffice for our experiments

Image Analysis: Spotted arrays
What information do we see?

Information:

Spot statistics generated

Gene name
Spot position
Cy3 intensity
Cy5 intensity
Background signal
Variation in spot

	A	B	C	D	E	F	G	H	I	J							
1	FeatureNum	ProbeName	GeneName	SystematicNa	Description	LogRatio	gMedianSign	rMedianSigna	gBGMedianSi	rBGMedianSi						EEP	gB
2	12	A_52_P61635	Ccr1	NM_009912	Mus musculu	0.00E+00	72	86	68	85							
3	13	A_52_P58058	Nppa	NM_008725	Mus musculu	0.00E+00	81	91	73	80							
4	14	A_52_P40340	Aqp7	NM_007473	Mus musculu	0.00E+00	80	90	75	79							
5	15	A_52_P81915	AK046412	AK046412	Mus musculu	0.00E+00	78	82	76	78							
6	16	A_51_P33183	Hvcn1	NM_0010424	Mus musculu	0.00E+00	80	88	75	81							
7	18	A_51_P43063	Gpr33	NM_008159	Mus musculu	0.00E+00	75	86	75	80							
8	19	A_52_P50235	C230086J09F	AK084122	Mus musculu	0.00E+00	78	91	77	81							
9	20	A_52_P29996	LOC434369	BC053388	Mus musculu	0.00E+00	82	96	78	81							
10	21	A_51_P35638	A330106F07F	AK039774	Mus musculu	0.00E+00	92	90	83	81							
11	22	A_52_P68440	Ptdss2	NM_013782	Mus musculu	-1.80E-01	101.5	91.5	91	83							
12	23	A_51_P41420	1110014K05I	NM_028622	Mus musculu	-3.48E-01	1159	545	109	84						4715	
13	24	A_51_P28091	Itfg1	AK141948	Mus musculu	-3.40E-01	107.5	89	113	82							
14	25	A_52_P61368	Elmo1	NM_198093	Mus musculu	0.00E+00	96	95.5	109	78							
15	26	A_52_P25819	Crtac1	NM_145123	Mus musculu	-2.90E-01	99	94	104	78							
16	27	A_52_P22927	Pnpt1	NM_027869	Mus musculu	3.30E-01	95	93	98	78							
17	28	A_52_P21463	Sox9	NM_011448	Mus musculu	0.00E+00	90	86	98	79							
18	29	A_52_P57951	Tmem144	NM_027495	Mus musculu	0.00E+00	91	87	91	76							
19	30	A_52_P97999	AK039768	AK039768	Mus musculu	0.00E+00	84.5	82.5	87	76	-2.5	6.5		7.15		0	
20	31	A_52_P45386	Syne1	NM_153399	Mus musculu	0.00E+00	78	78	86	77	-8	1		1.1		0	
21	32	A_52_P65584	Ank1	NM_031158	Mus musculu	0.00E+00	82	81	85	75	-3	6		6.6		0	
22	33	A_51_P28237	H2-M10.5	NM_177637	Mus musculu	0.00E+00	82.5	93	85	77	-2.5	16		17.6		0	
23	34	A_52_P17601	Isgf3g	NM_008394	Mus musculu	0.00E+00	83.5	82	86	77	-2.5	5		5.5		0	
24	35	A_51_P12117	BC056923	NM_173395	Mus musculu	0.00E+00	78	85	85	75	-7	10		11		0	
25	36	A_51_P44841	Pusl1	AK151452	Mus musculu	-2.61E-01	156.5	108.5	85	77	71.5	31.5	2.26984127	34.65	0.48461538	-1.04508789	
26	37	A_51_P21503	5330410G16	NM_182991	Mus musculu	4.74E-01	84	96	90	77	-6	19		20.9		0	
27	38	A_52_P42702	Ldlr	NM_010700	Mus musculu	4.05E-01	116	185	91	79	25	106	0.23584906	116.6	4.664	2.22156779	
28	39	A_52_P66914	Hps5	NM_0010052	Mus musculu	-8.22E-02	94	85	90	77	4	8		8.8		0	
29	40	A_52_P65218	Thoc1	AK042548	Mus musculu	0.00E+00	84	80	90	76	-6	4		4.4		0	
30	41	A_51_P31856	Wnk1	NM_198703	Mus musculu	2.98E-01	130	167	90	76	40	91	0.43956044	100.1	2.5025	1.32337007	
31	42	A_52_P12371	AK032795	AK032795	Mus musculu	0.00E+00	84	85.5	93	77	-9	8.5		9.35		0	
32	43	A_52_P68238	Scd1	NM_009127	Mus musculu	6.58E-02	390	475	96	77	294	398	0.73869347	437.8	1.48911565	0.5744558	
33	44	A_52_P57176	Kif13b	AK031482	Mus musculu	-1.72E-01	95.5	84.5	97	77	-1.5	7.5		8.25		0	
34	46	A_52_P26754	BC013529	NM_145418	Mus musculu	-6.61E-02	101	90	111	79	-10	11		12.1		0	
35	47	A_52_P17208	BE956575	BE956575	UI-M-BG2-bb	-1.00E-01	102	89	126	77	-24	12		13.2		0	
36	48	A_51_P14420	BG963265	BG963265	BG963265 60	-1.83E-01	106	93	128	77	-22	16		17.6		0	
37	49	A_51_P11472	Hao3	NM_019545	Mus musculu	0.00E+00	100	81	120	76	-20	5		5.5		0	
38	50	A_52_P36865	Zswim4	NM_172503	Mus musculu	-7.17E-02	221	173	108	76.5	113	96.5	1.17098446	106.15	0.93938053	-0.0902184	
39	51	A_52_P93159	AK051556	AK051556	Mus musculu	0.00E+00	88	85	100	77	-12	8		8.8		0	
40	52	A_51_P41466	2210010C04	NM_023333	Mus musculu	0.00E+00	81	79	94	76.5	-13	2.5		2.75		0	
41	53	A_52_P29037	2810011L19F	BC059025	Mus musculu	2.85E-01	89	93	90	77	-1	16		17.6		0	
42	54	A_52_P10362	NAP103812-1	NAP103812-1	Unknown	2.72E-01	861	1521	87	77	774	1444	0.53601108	1588.4	2.05219638	1.03716879	
43	55	A_51_P29133	Ssrp1	NM_182990	Mus musculu	-1.83E-01	2030	1331	86.5	77	1943.5	1254	1.54984051	1379.4	0.70975045	-0.49461624	

Analysis: Spotted arrays
What information do we need?

Summarize the information in one value/gene.

1) Which genes are changed?
2) How much are they changed?

Excel Analysis: Spotted arrays
What information do we see?

For each gene:
Use median signal for green and red (less sensitive to outliers).

Subtract the background signal for each channel.

Normalize the two colors (same average signal).

Filter for expressed genes (typically keep genes where signal>100)

Take the log ratio of each gene: `=log(signal/signal,[base=]2)`

Excel Analysis: Spotted arrays
Optional: Model the error of the measurements

For each gene:
Calculate X

$$X = (a_2 - a_1)/[s_1^2 + s_2^2 + f^2 (a_1^2 + a_2^2)]$$

Where a_1, a_2 are median signal for the green and red channel, and s_1, s_2 are standard deviations of the background intensities.

X-scores follow a normal distribution, with mean≈0
P-value for significance from X-score

This only concerns statistical significance of the measurements of this array
For the experimental error, you need to consider duplicates.

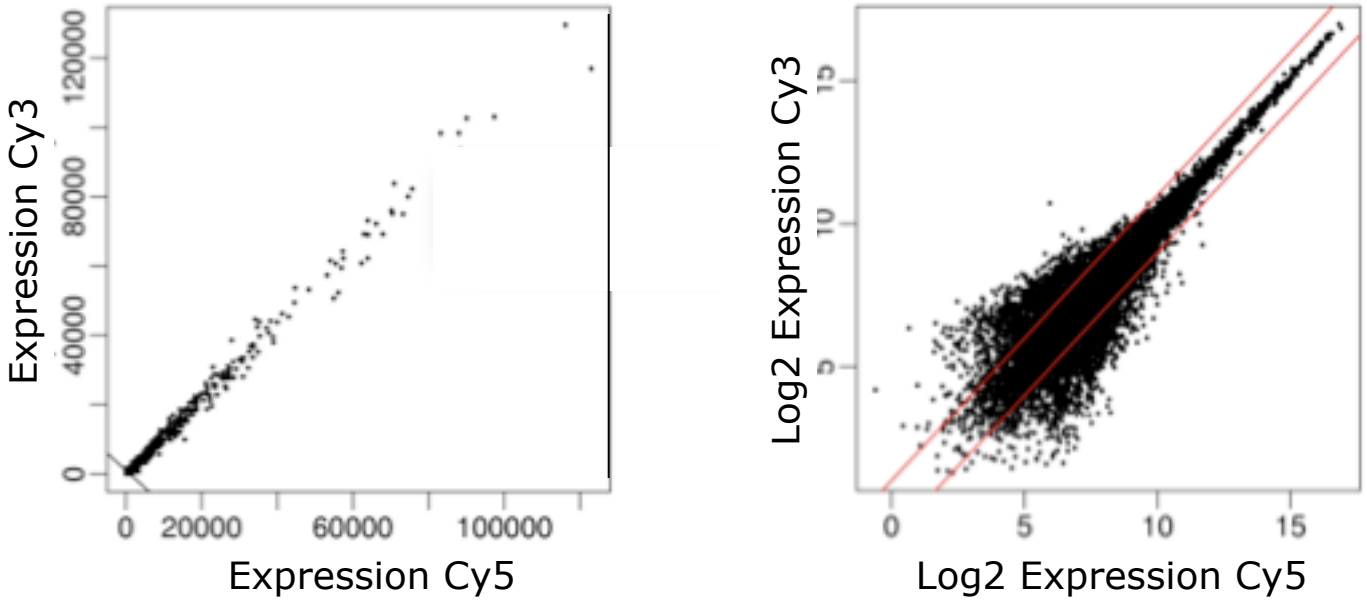
Platform specificity ends here

- Compared to classical one gene techniques, any genome-wide technique (array techniques, sequencing, genomic phenotyping…) need some special attention.
  - Controls
  - Multiple testing problematic

Microarray analysis

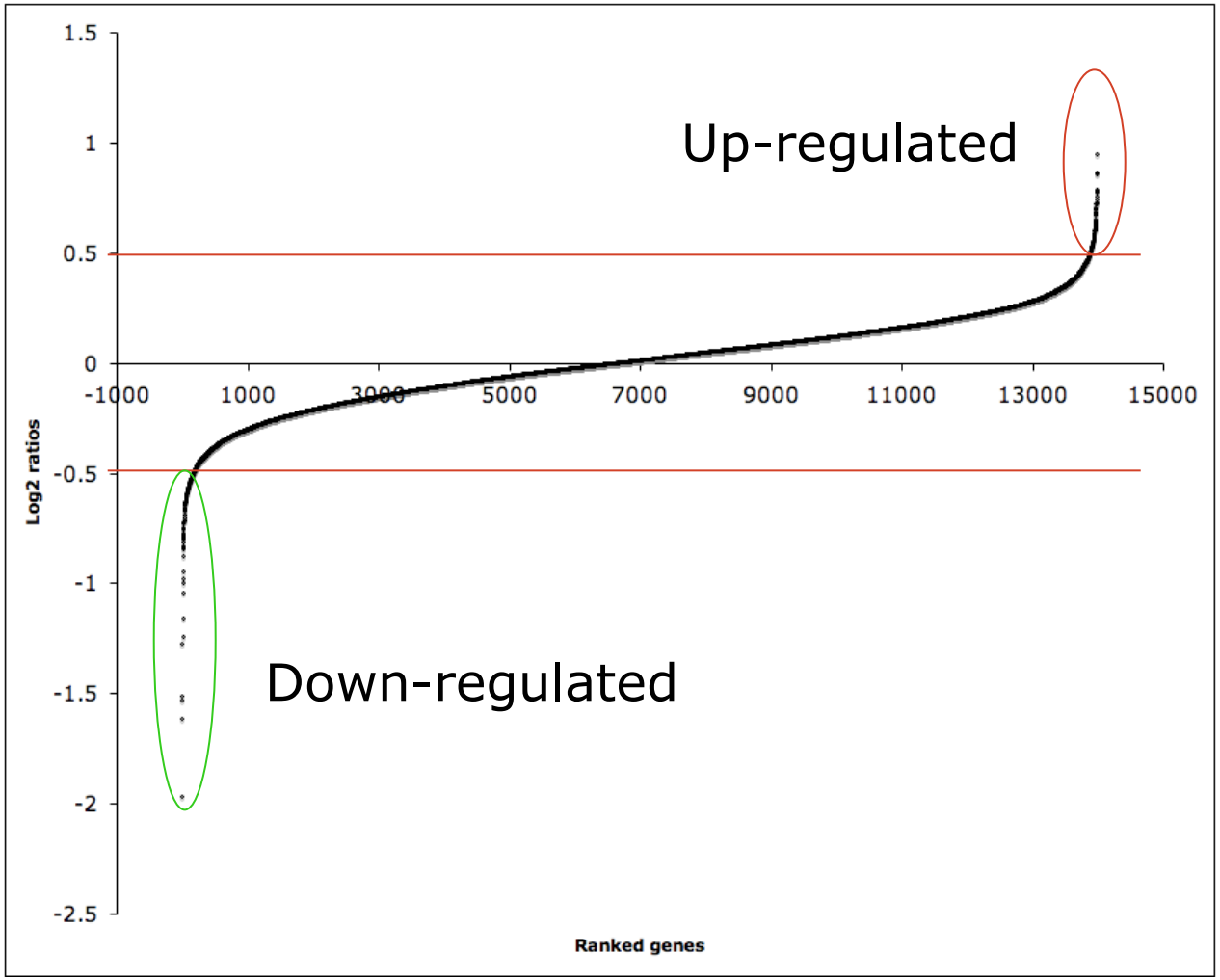
- An example of application
- Preprocessing of the data
- Identification of differentially expressed genes
  - Biological and statistical significance
- Interpretation of the data
- Communication of results

Scatter plots to visualize results



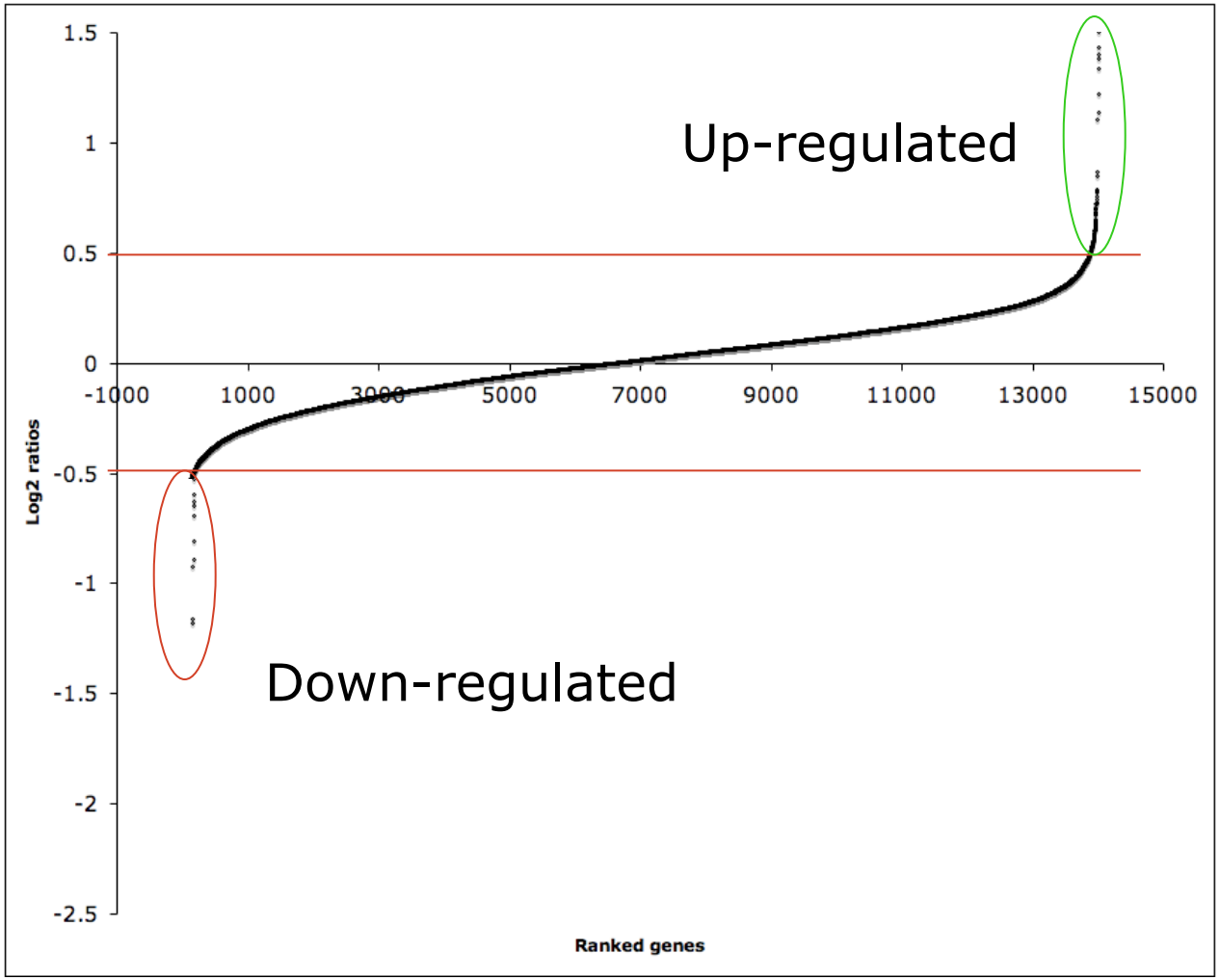
- Set threshold for fold-change
  - E.g. log2 ratio > 0.5

Differential gene expression



- How much is experimental noise?
- How many of these genes were also found in the dye-swap experiment?
- Second group replicate?

Biological/statistical significance

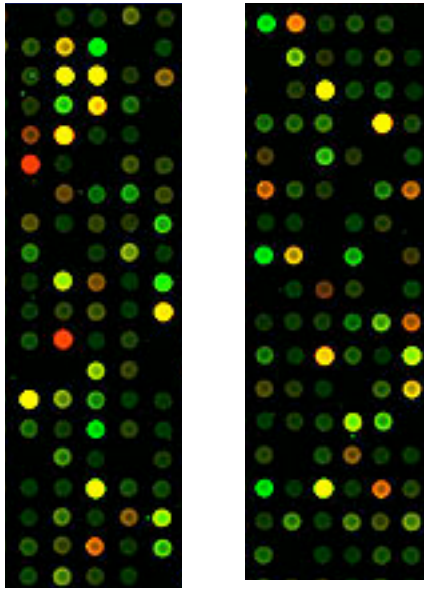


- Relevant changes should be both biologically and statistically significant
  - Detectable fold-change
  - Significant p-value

Classical statistics

- One null hypothesis tested once. H_0: no change
  - H_0: *Renilla* luciferase in not affected by siRNA
  - Several measurements of *Renilla* luciferase with scrambled siRNA
  - Several measurements of *Renilla* luciferase with luciferase siRNA
  - Reject if t-test shows $p<0.05$, probability of observed results is 5% to occur by chance.

Classical statistics: genome-wide experiments



- One null hypothesis tested multiple times. H_0: no change
  - H_0: gene is not affected by siRNA.
  - Few measurements of each gene with scrambled siRNA
  - Few measurements of each gene with targeted siRNA
  - Reject if t-test shows $p<0.05$, probability of observed results is 5% to occur by chance.
  - If testing 44k features, $p<0.05$ will identify 2,200 genes by chance!

Multiple testing

- Several ways to find significant results:
- Easiest: set the p-value lower.
  - Selecting genes associated with $p<0.00001$, <1 gene will be selected by chance.
  - Conservative
  - (Bonferroni correction $p_{adjusted}=\min\{p_{initial}*N,1\}$)

- Popular way is to look at False Discovery Rate
  - How many of the selected genes are false positives?
  - Based on distribution of p-values.

Expectations

- Primary hit: siRNA target downregulated
  - If not, transfection might not have worked.
  - For this experiment, we have used validated siRNAs.
  - Use experiment to look for microarray technical variation.
- Secondary hits: Genes interacting with primary target
- Tertiary hits: Genes changed because of changed cellular environment

siRNA off-targets

- If other genes have been targeted by the siRNA, they will show up together with their secondary effects.

- This experiment cannot distinguish different hits

Results

- Lists of genes (going up/going down)
  - With number of false positives estimated

- What then?
- Read and make conclusion from literature.
- How to make more sense of the data?

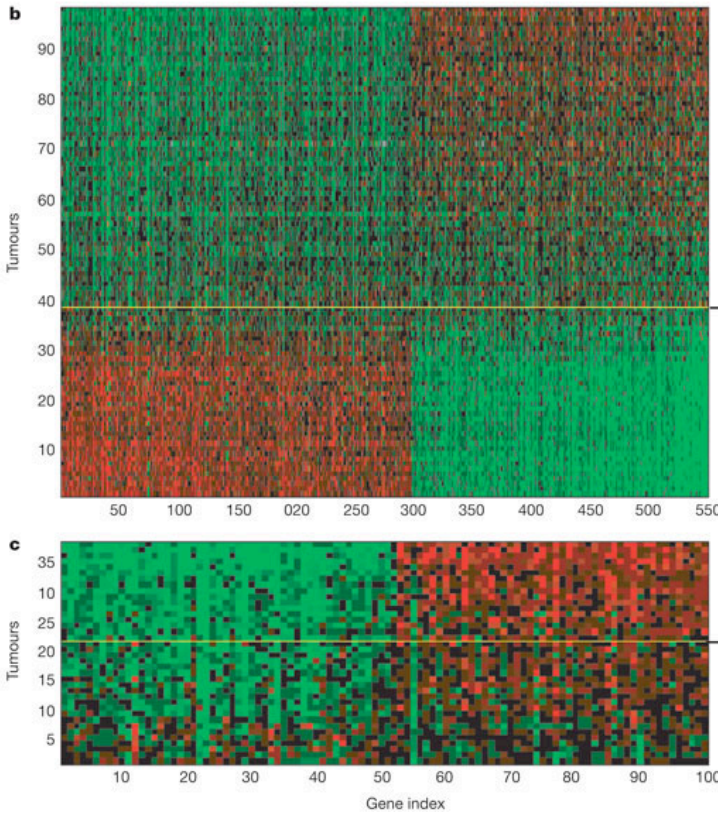
Microarray analysis

- An example of application
- Preprocessing of the data
- Identification of differentially expressed genes
  - Biological and statistical significance
- Interpretation of the data
- Communication of results

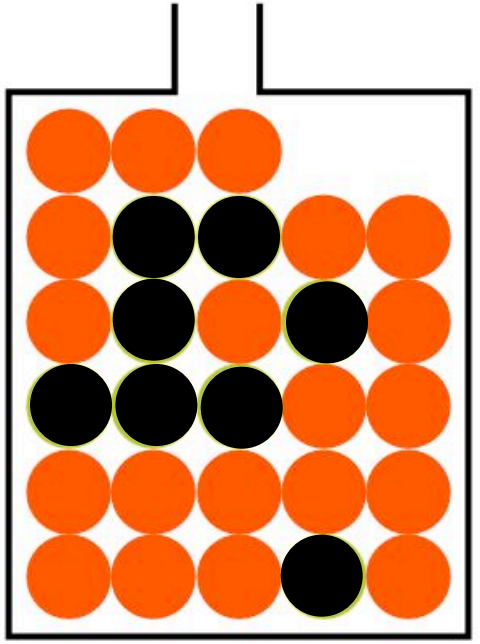
Clustering

- Heatmaps / clustering

- To visualize patterns in data.
- Similarity between genes by Euclidean distance, Pearson's correlation coefficient…
- Genes/samples ordered based on similarity.



Enrichment



- Determine enrichment by Hypergeometric distribution

- Example 1: Urn with 28 balls where 8 are black.
- You draw 6 balls,
  - Expect 4 orange and 2 black
  - If you get 6 black balls, there is a significant deviation.

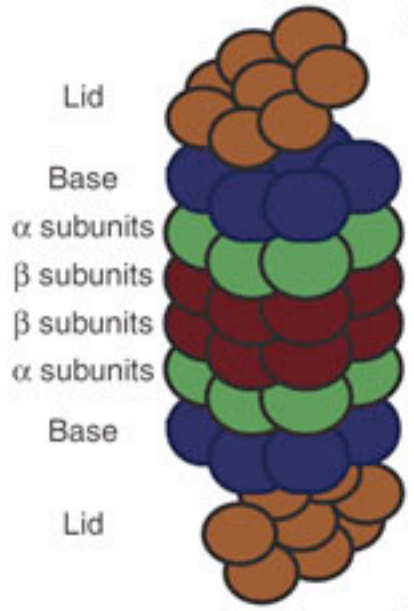
- Example 2: 10,000 expressed genes on the array
- 500 of the 10,000 genes are involved in DNA repair
- 300 are changed
  - expected number of DNA repair genes: 15
  - observed: 200 out of the 300 selected genes are involved in DNA repair
  - There is a significant enrichment of DNA repair genes among the changed genes.

- Need pre-defined groups of genes.

Gene groups

- Protein complexes
  - E.g. proteasome to break down proteins

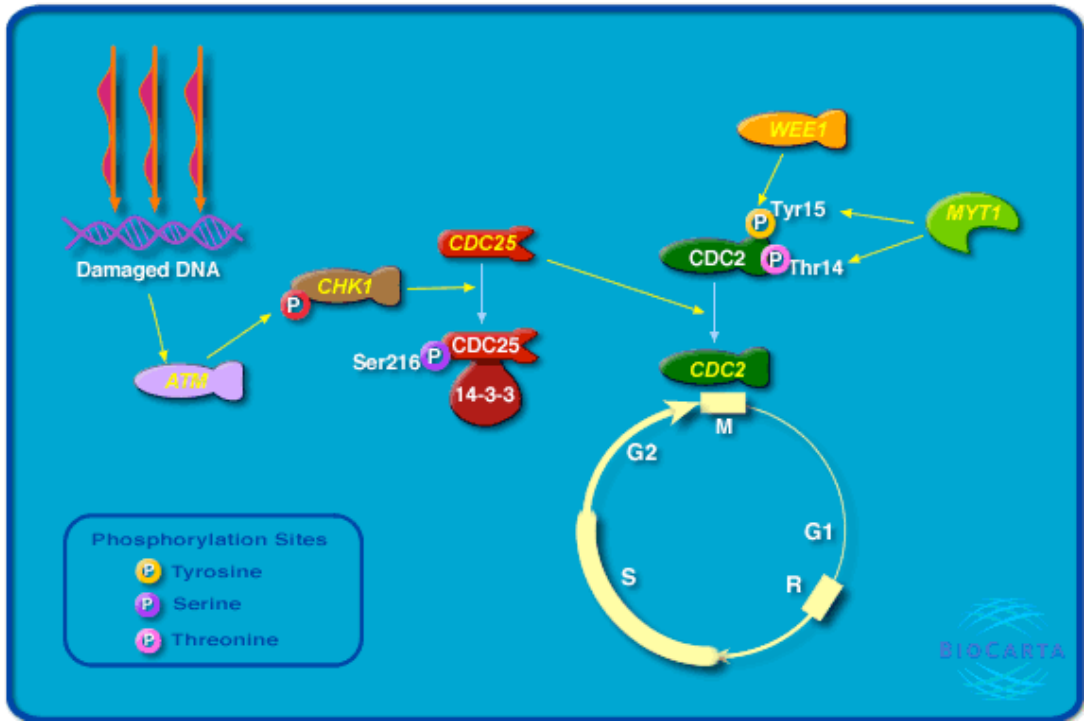
- Pathways
- Gene ontology:Structured vocabulary



Breslow et al. 2008 Nature Methods

Gene groups

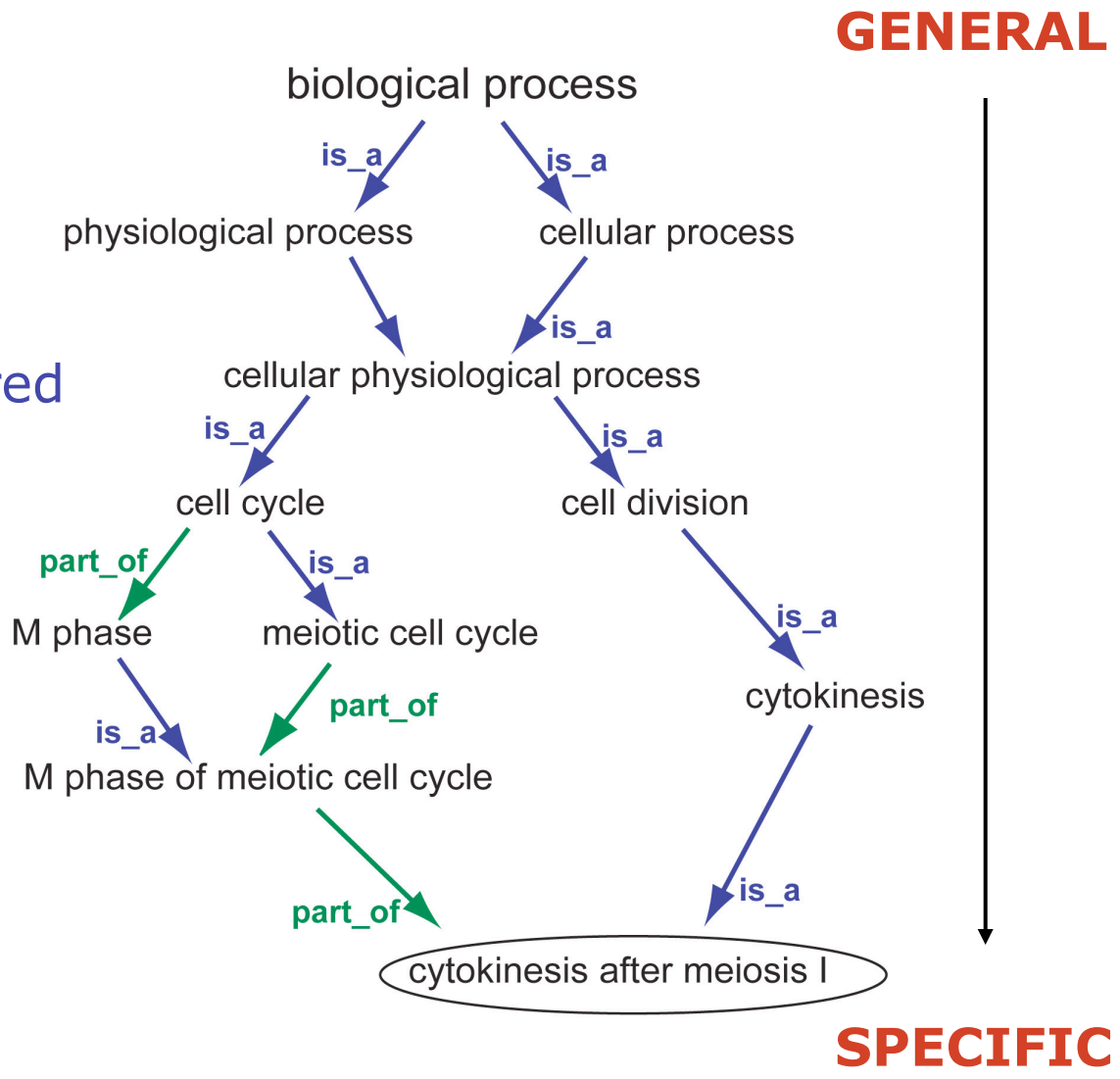
- Protein complexes
- Pathways
  - E.g. Regulatory pathwa in response to DNA damage
  - Biocarta, KEGG
- Gene ontology



Gene groups

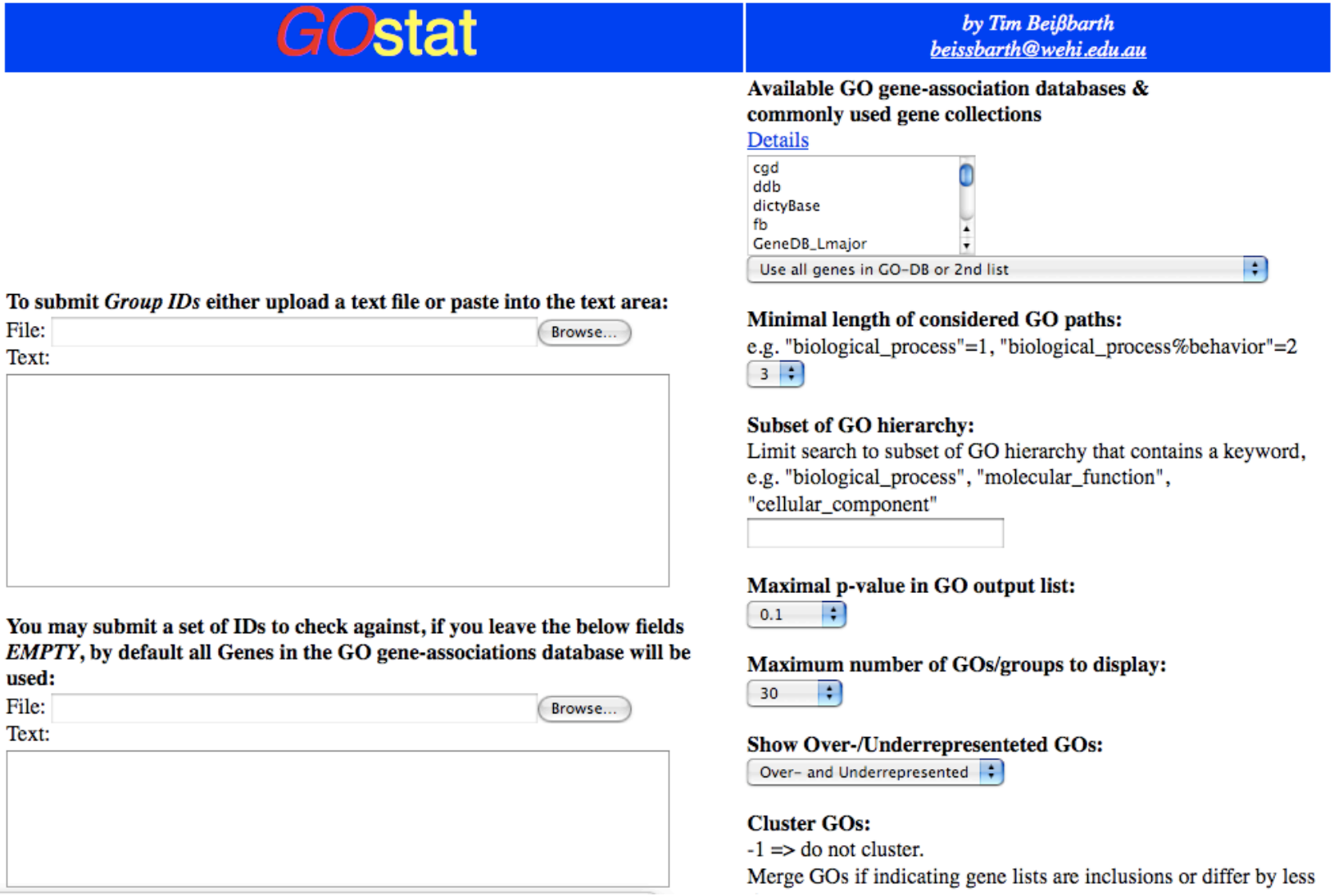
- Protein complexes
- Pathways

- Gene ontology:Structured vocabulary



Web-tools

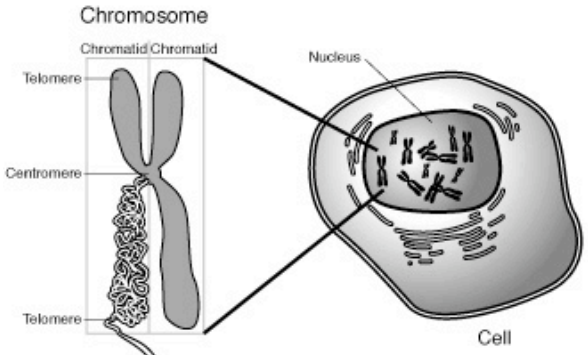
http://gostat.wehi.edu.au/

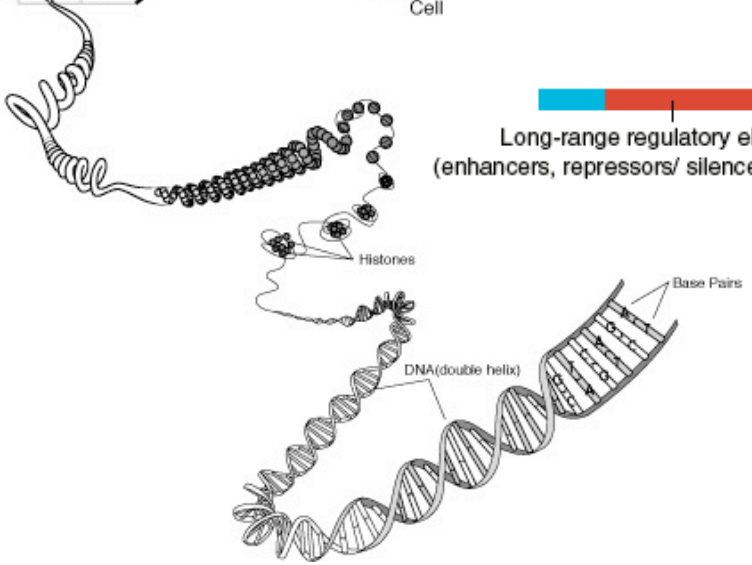


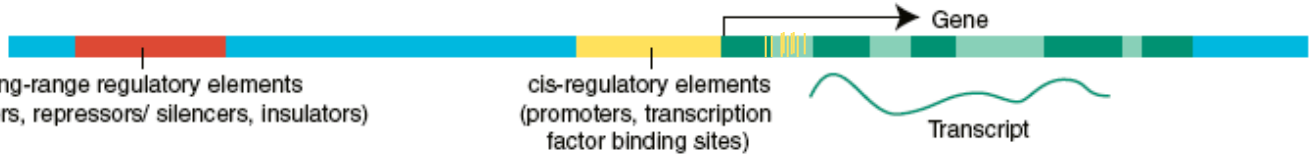
Genes used in course

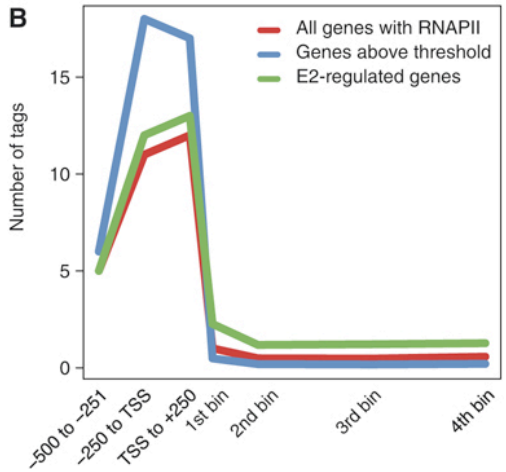
- Mpg - glycosylase (enzyme)
- p53 - cell death/cell cycle arrest inducer
    - Normally with transcriptional activity, but not in ES cells
- Polr2d - subunit of RNA polymerase
- Tada2l - transcriptional activator
- Nanog - ES specific transcription factor

Transcription factor binding sites





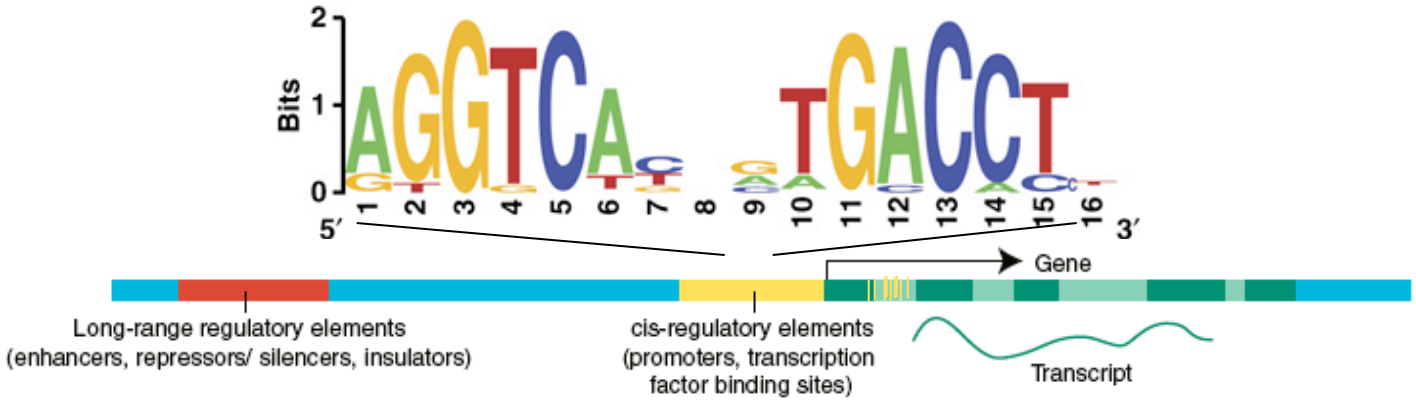




Welboren et al. 2009 EMBO J

Transcription factor binding sites

- Often palindromic motifs found near Transcription Start Site (±1000bp)
- Databases with DNA binding proteins and the sequence they bind
  - TRANCFAC, JASPAR, TESS
  - Can be queried through PRIMA (http://acgt.cs.tau.ac.il/prima/PRIMA.htm)



Transcription factor binding sites

- For the genes with transcriptional activity (Nanog, p53, Tada2l and Polr2d)

  we expect the list of down-regulated genes to be enriched for the respective binding sequence in regulatory region

Microarray analysis

- An example of application
- Preprocessing of the data
- Identification of differentially expressed genes
  - Biological and statistical significance
- Interpretation of the data
- Communication of results

NCBI

All Databases | PubMed | Nucleotide | Protein

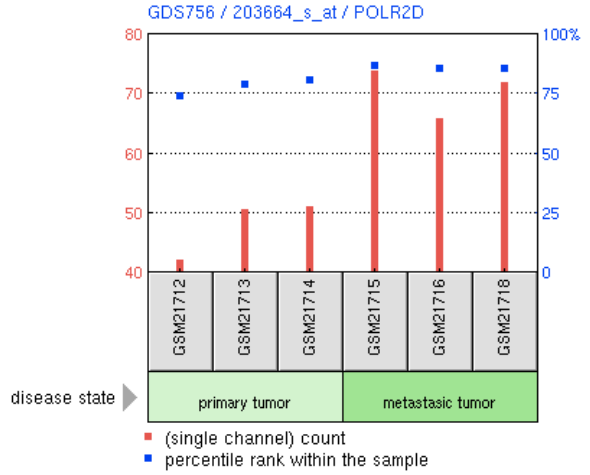
GEO PROFILES Gene Expression Omnibus

Communicating results

- Peer review. Other people should see data to challenge it.

- Data deposited at databases (NCBI or EBI), together with information on the samples
  - Array type, protocol, sample identification (species, treatment, time points…)

Title: GDS756 / 203664_s_at / POLR2D / Homo sapiens
Summary: Comparison of gene expression in SW480, a primary tumor colon cancer cell line, to that in SW620, an isogenic metastatic colon cancer cell line. Cell lines derived from one individual. Results provide insight the progression of cancer from primary tumor growth to metastasis.



MIAME

Minimal Information About a Microarray Exp't

Provide:

1. Raw data for each hybridization
2. Final processed data for the set of hybridizations
3. Experimental factors and their values (e.g., compound and dose in a dose response experiment)
4. Experimental sample relationships (e.g., which raw data file relates to which sample, which hybridisations are technical, which are biological replicates)
5. Array annotation (e.g., commercial array catalog number)
6. Data processing protocols (e.g., normalization method used)

Recap: Motivation

- Course to give insights into projects conducted at MIT
- Part of a research project in the Samson lab
  - We don't know what the results will be.
  - Human cell lines lacking these proteins are sensitive to DNA damaging agents.
  - Possibly the sensitivity is mediated through transcription

- Bigger picture:
  - By modulating sensitivity to DNA damaging agents, we might be able to sensitize tumors / de-sensitizes patients.