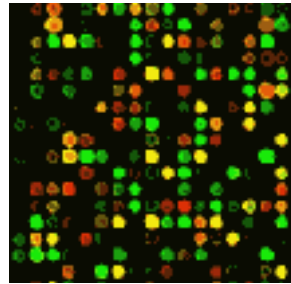


Analysis of Gene Expression Data



Rainer Breitling

r.breitling@bio.gla.ac.uk

Bioinformatics Research Centre and
Institute of Biomedical and Life Sciences
University of Glasgow

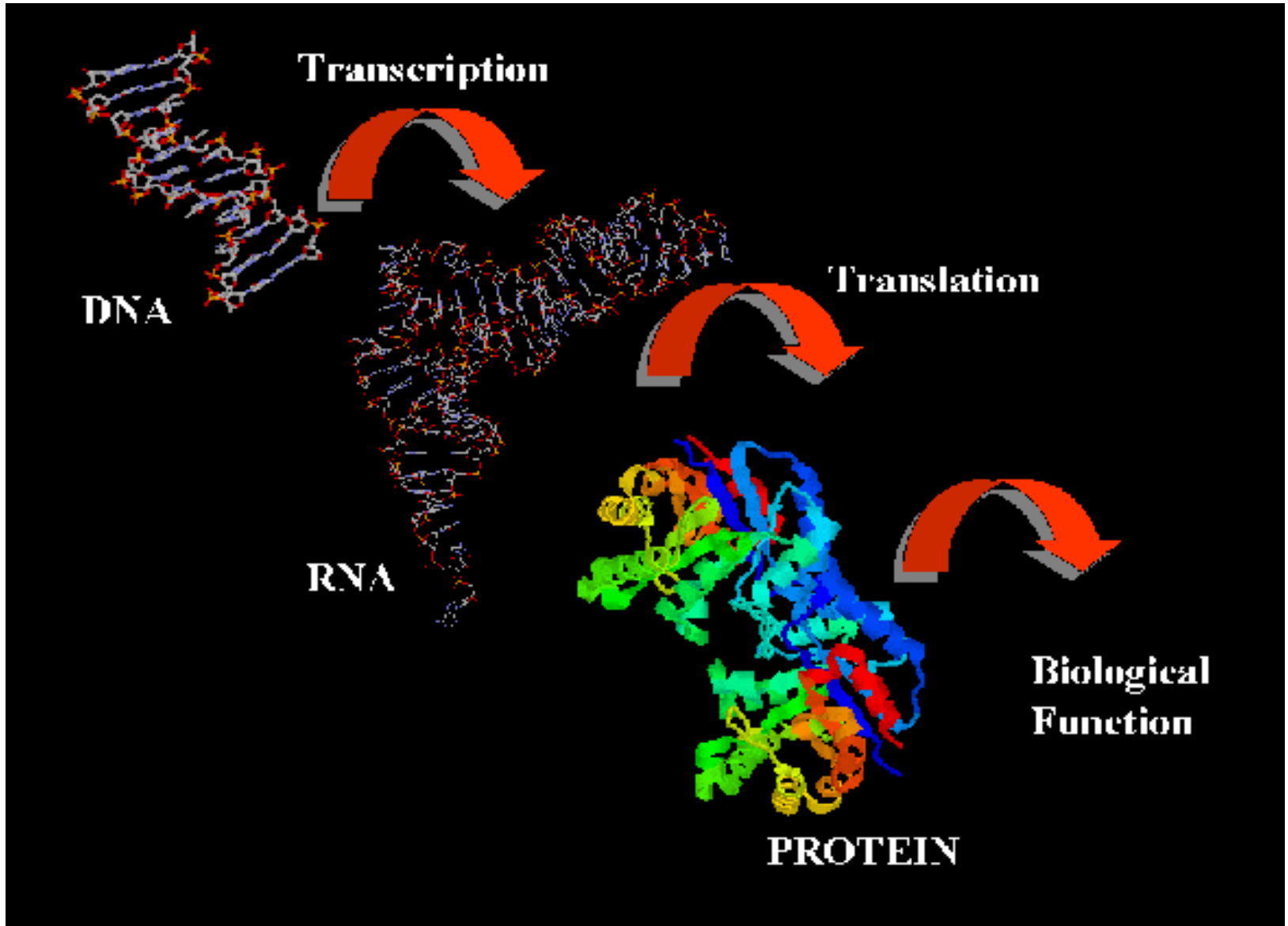
Outline

- Gene expression biology
- Measuring gene expression levels
 - two technologies: Two-color cDNA arrays and single-color Affymetrix genechips
- Finding and understanding differentially expressed genes
- Advanced analysis (clustering and classification)
- Cutting-edge uses of microarray technology

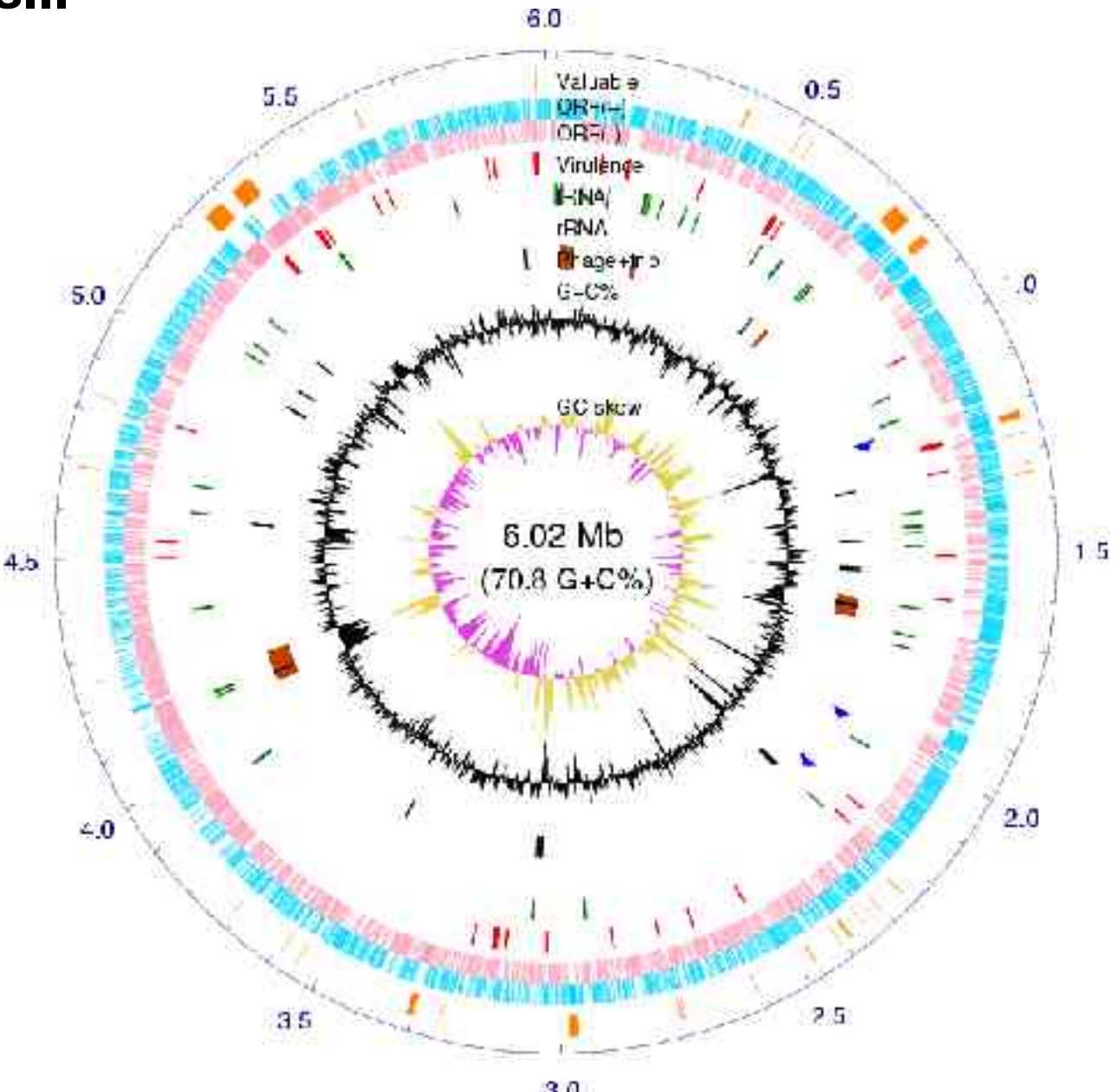


Gene expression biology

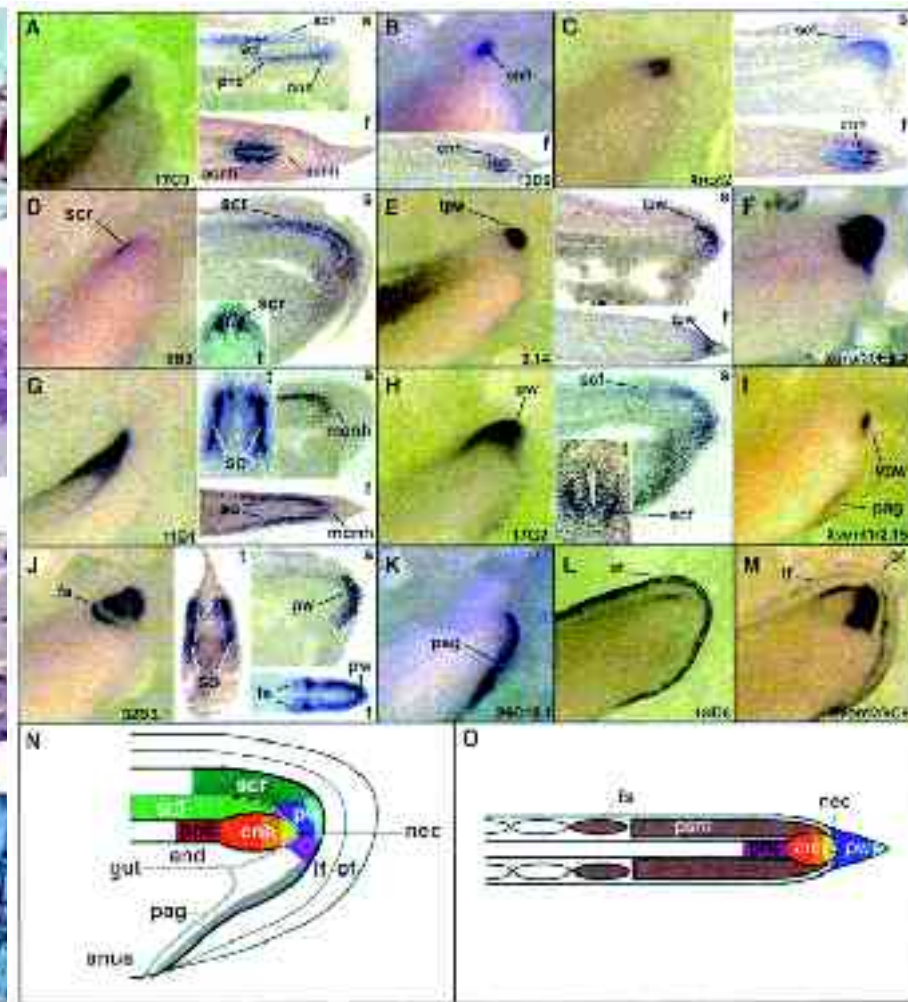
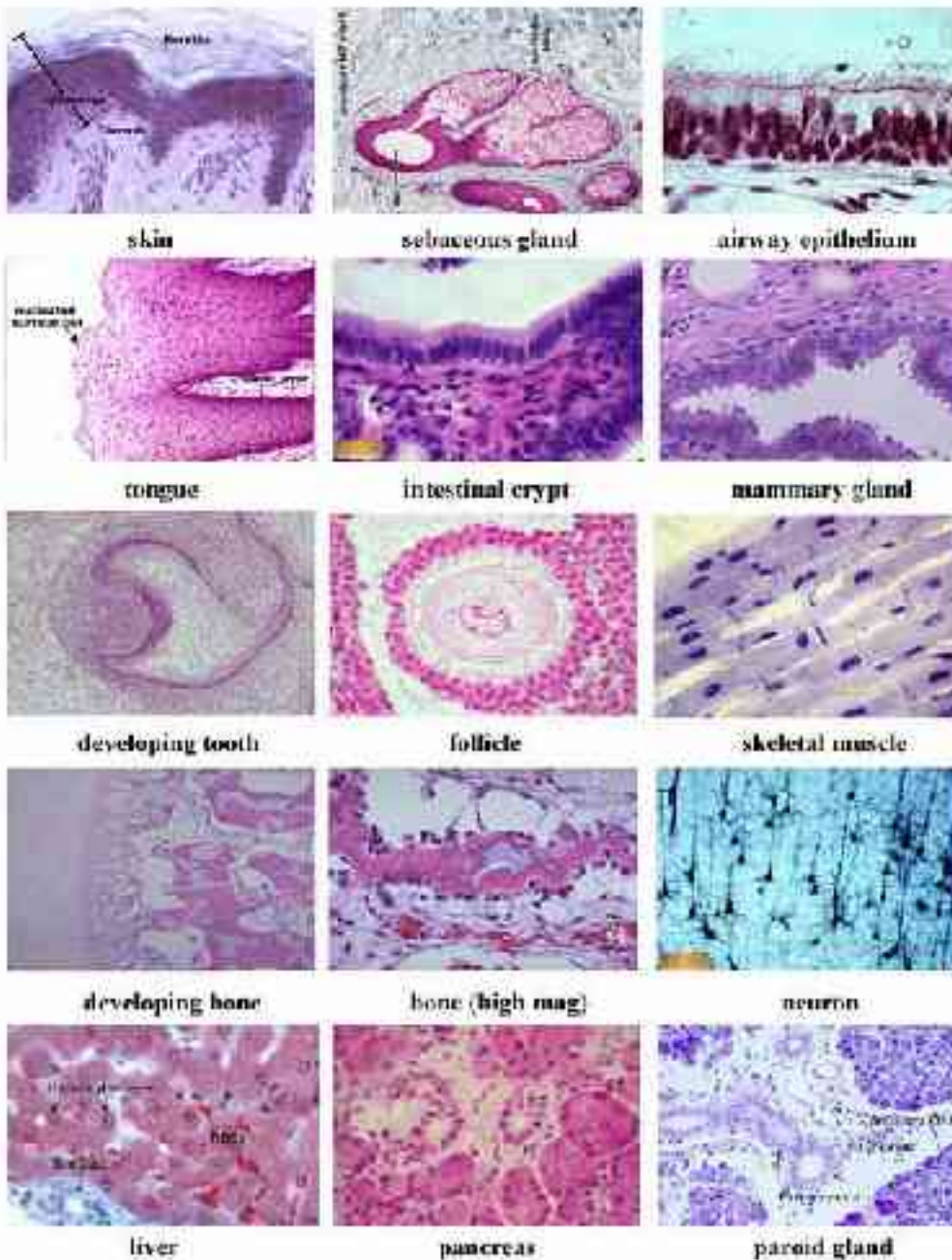
The central dogma of biology



Genome information is complete for hundreds of organisms...



...but the complexity and diversity of the resulting phenotype is challenging



whole-mount in situ hybridization of *X. laevis* tadpoles

The dramatic consequences of gene regulation in biology



Anise swallowtail, *Papilio zelicaon*

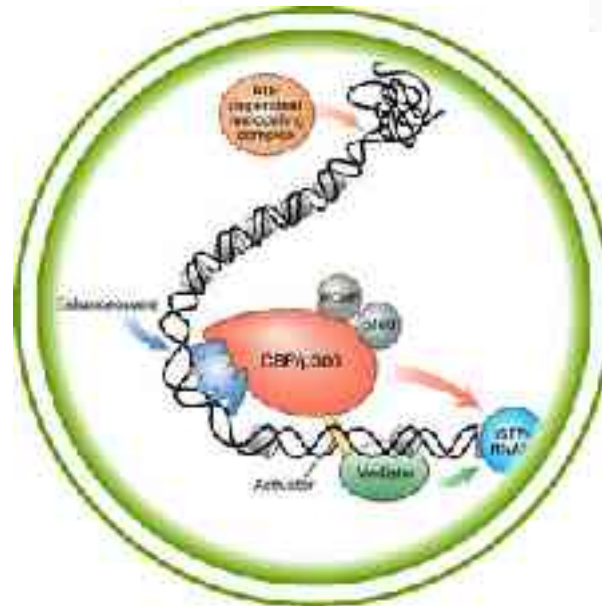
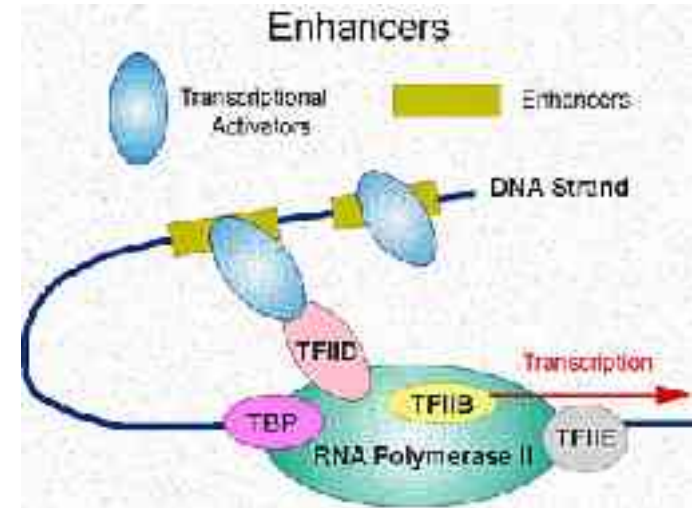
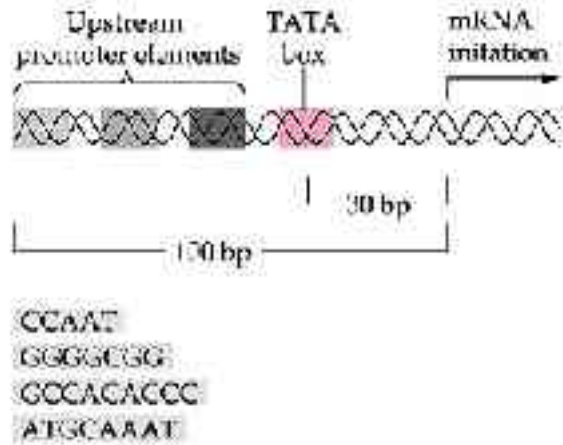


Same genome

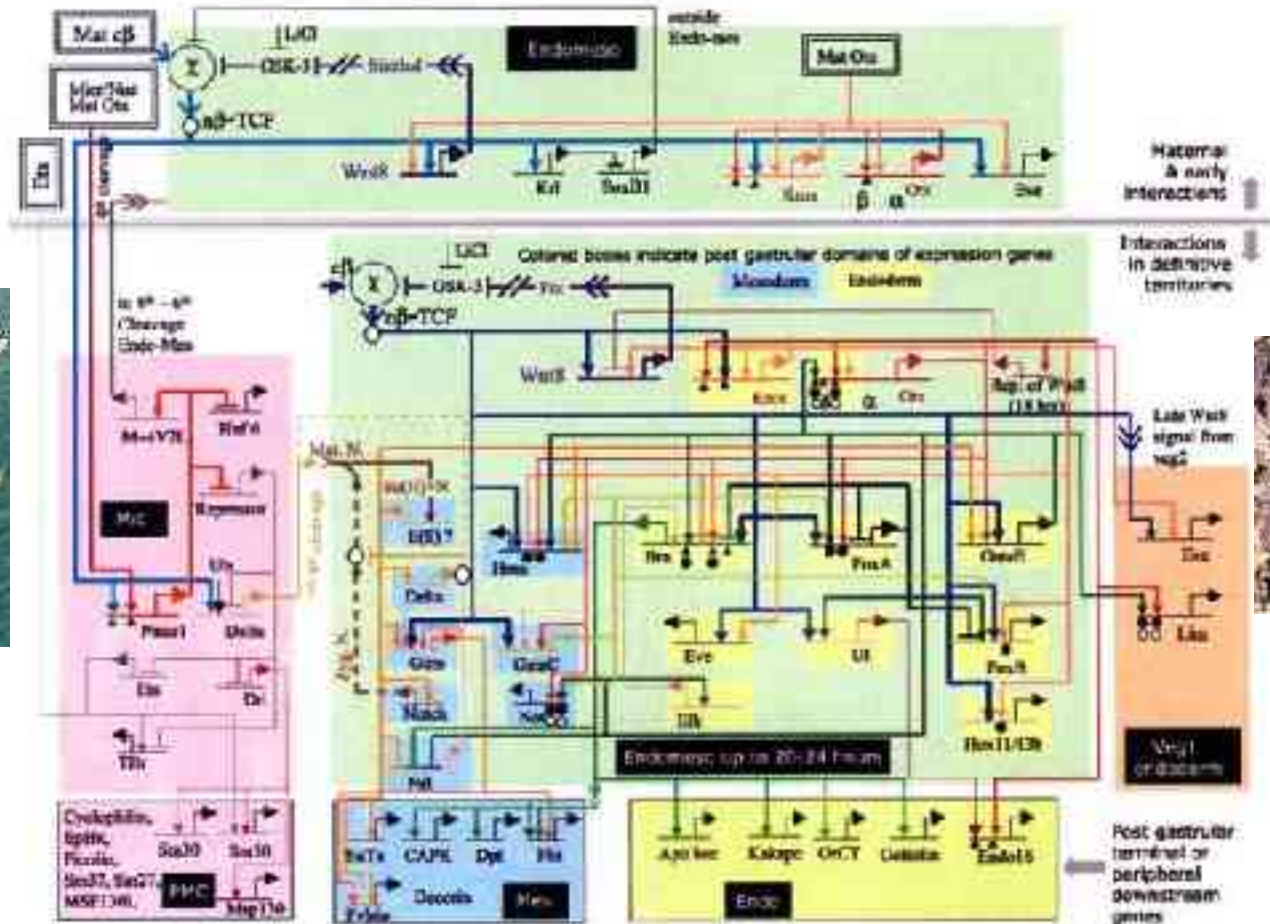
Different tissues

- Different physiology
- Different proteome
- Different expression pattern

The complexity of eukaryotic gene expression regulation



Regulatory Networks – integrating it all together



Genetic regulatory network controlling the development of the body plan of the sea urchin embryo Davidson *et al.*, *Science*, 295(5560):1669-1678.



Gene expression distinguishes...

- ...physiological status (nutrition, environment)
- ...sex and age
- ...various tissues and cell types
- ...response to stimuli (drugs, signals, toxins)
- ...health and disease
 - underlying pathogenic diversity
 - progression and response to treatment
 - patient classes of varying prospects

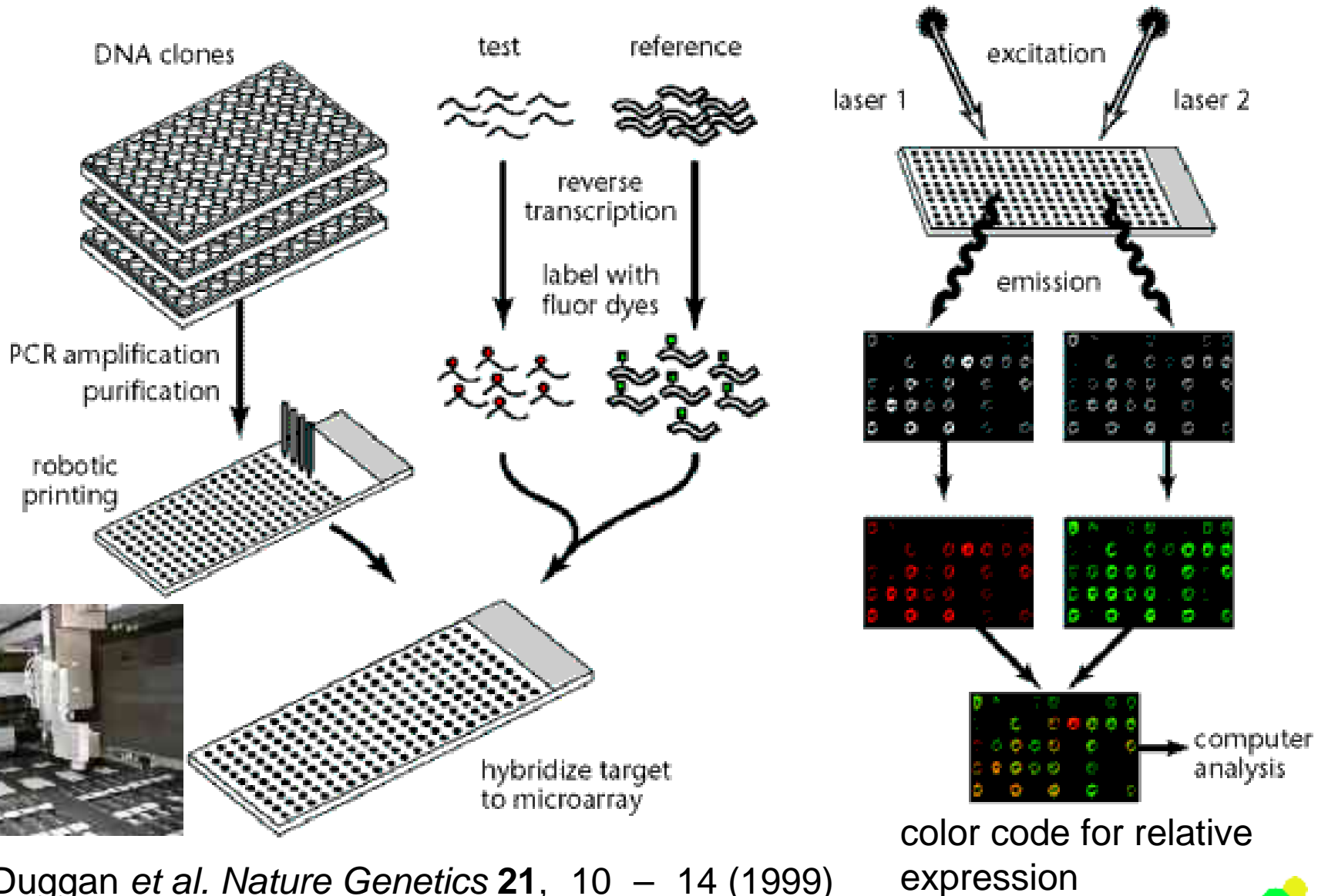


Measuring gene expression levels

1. total amount of mRNA = optical density at appropriate (UV) wavelength
2. mass separation and specific probing, one gene at a time = Northern blot
3. comprehensive “molecular sorting” = microarray technology
 1. two-color cDNA or oligo arrays
 2. single-color Affymetrix genechips



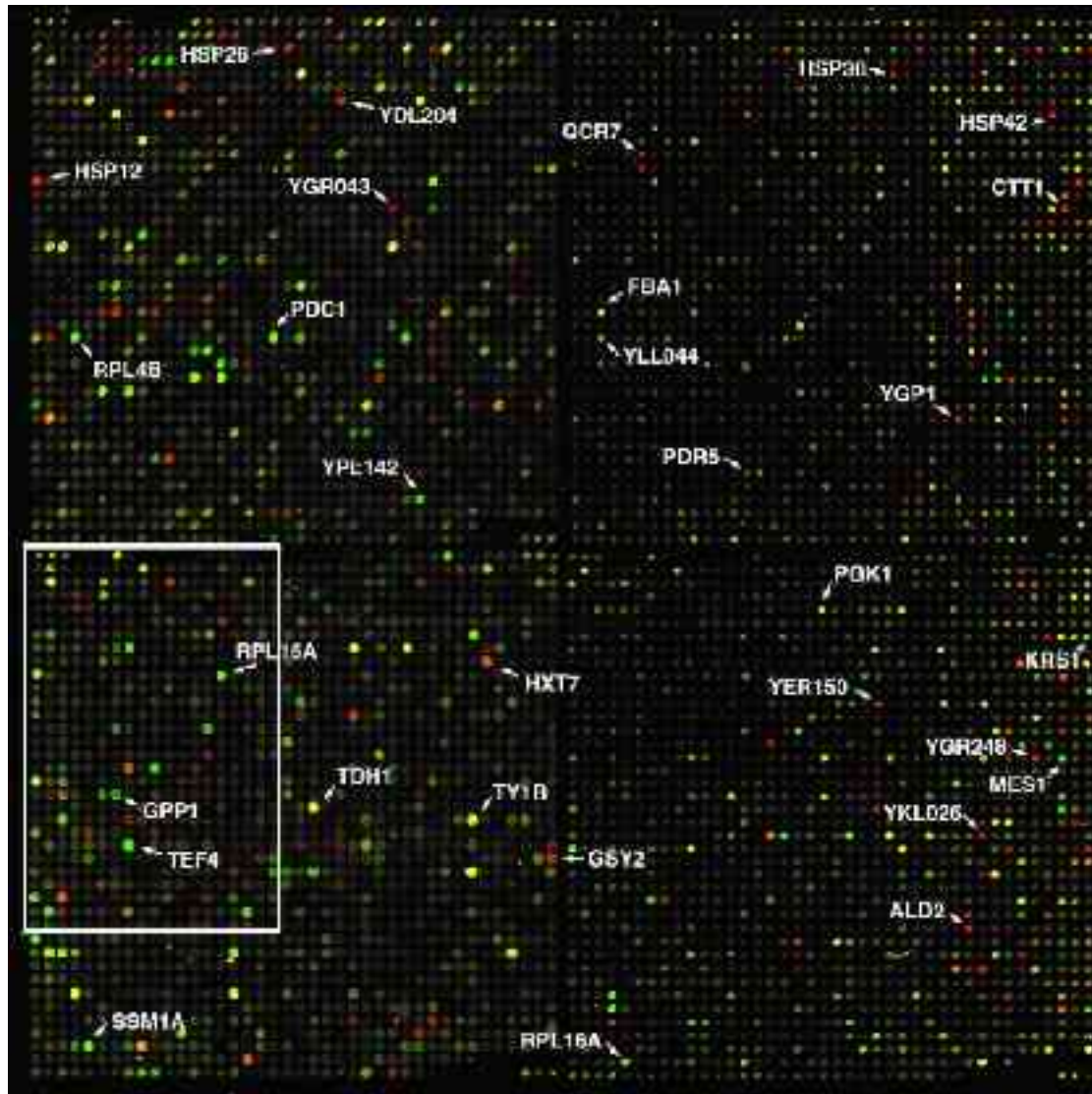
cDNA microarray schema



From Duggan *et al. Nature Genetics* **21**, 10 – 14 (1999)



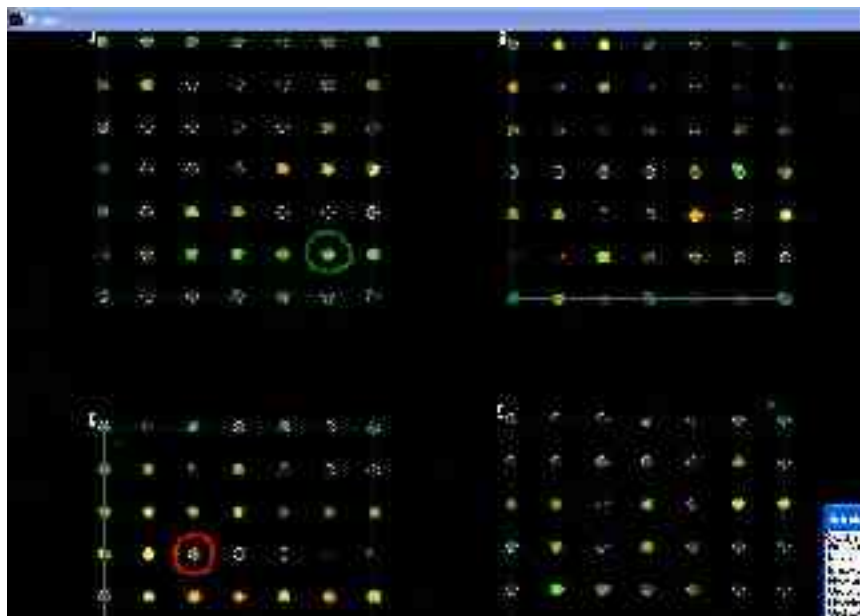
cDNA microarray raw data



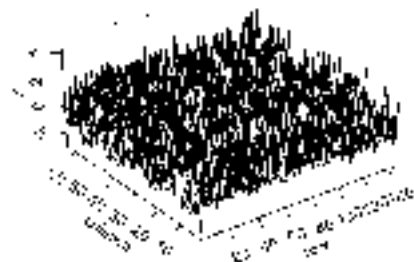
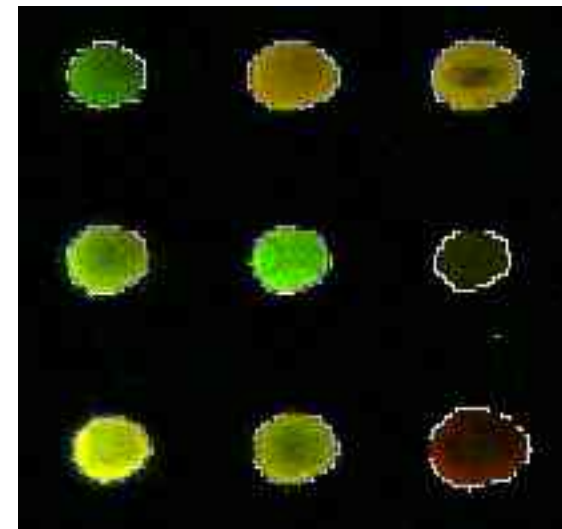
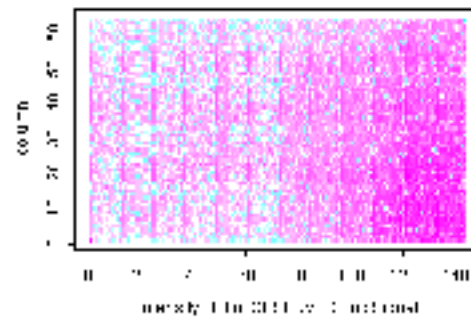
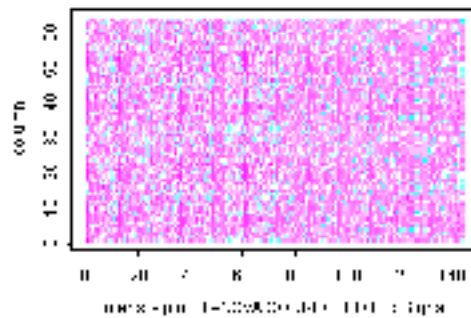
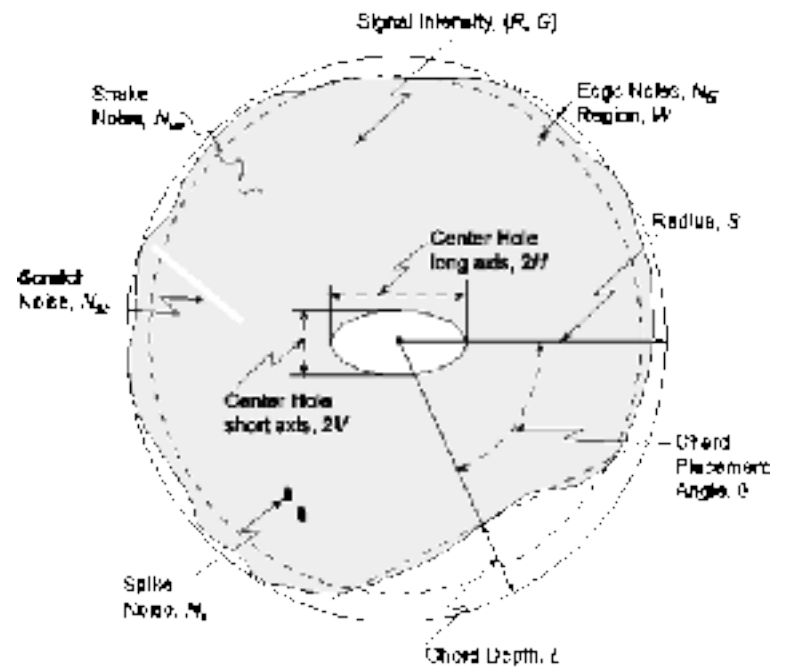
- can be custom-made in the laboratory
- always compares two samples
- relatively cheap
- up to about 20,000 mRNAs measured per array
- probes about 50 to a few hundred nucleotides

Yeast genome microarray. The actual size of the microarray is 18 mm by 18 mm. (DeRisi, Iyer & Brown, *Science*, 268: 680-687, 1997)

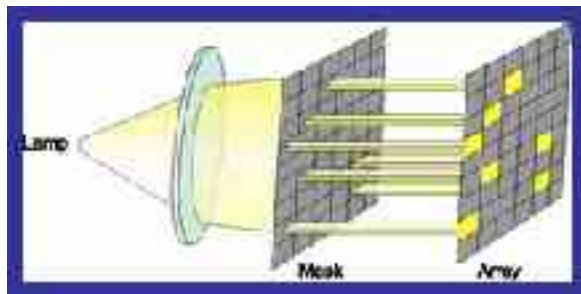
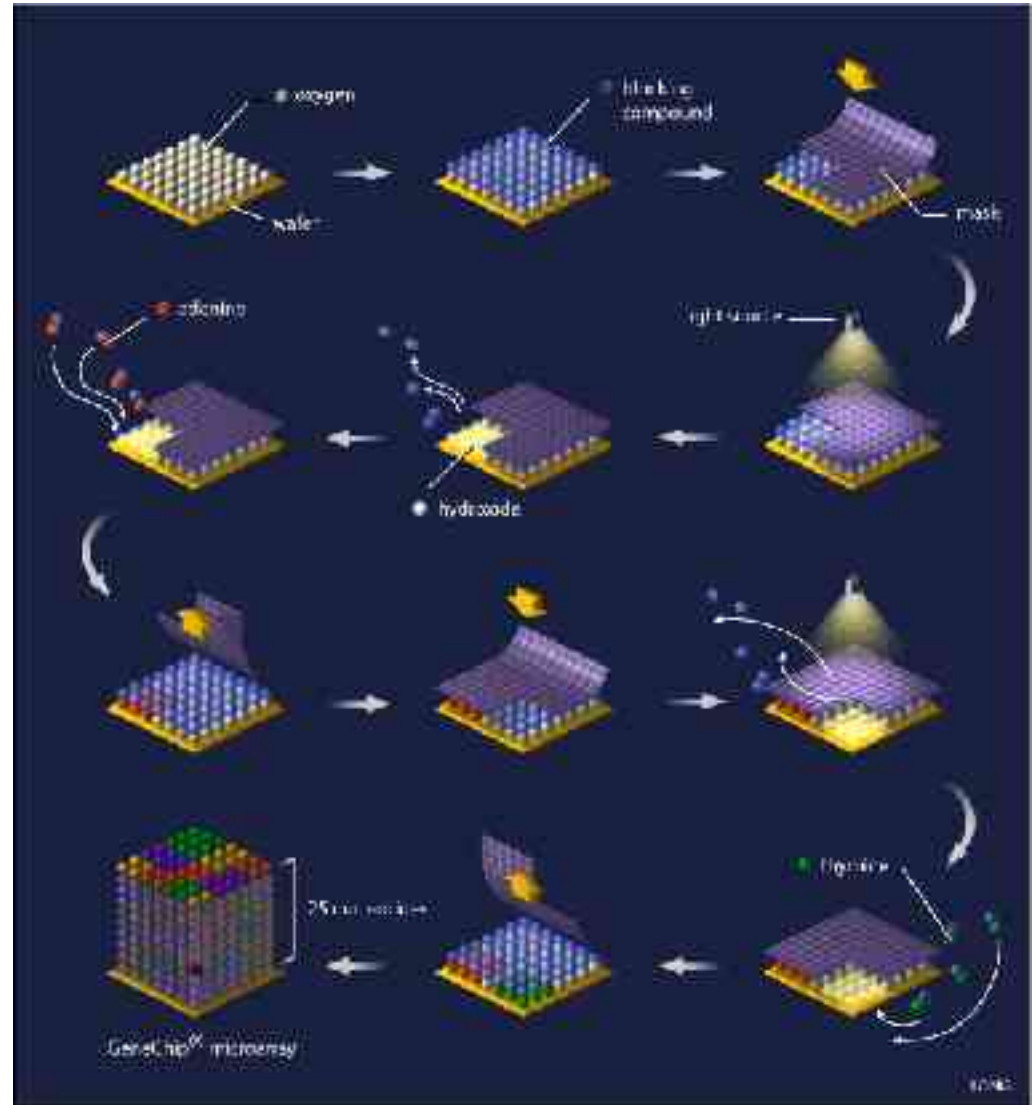




-HGMP 2b Cy 5 -- Hec94

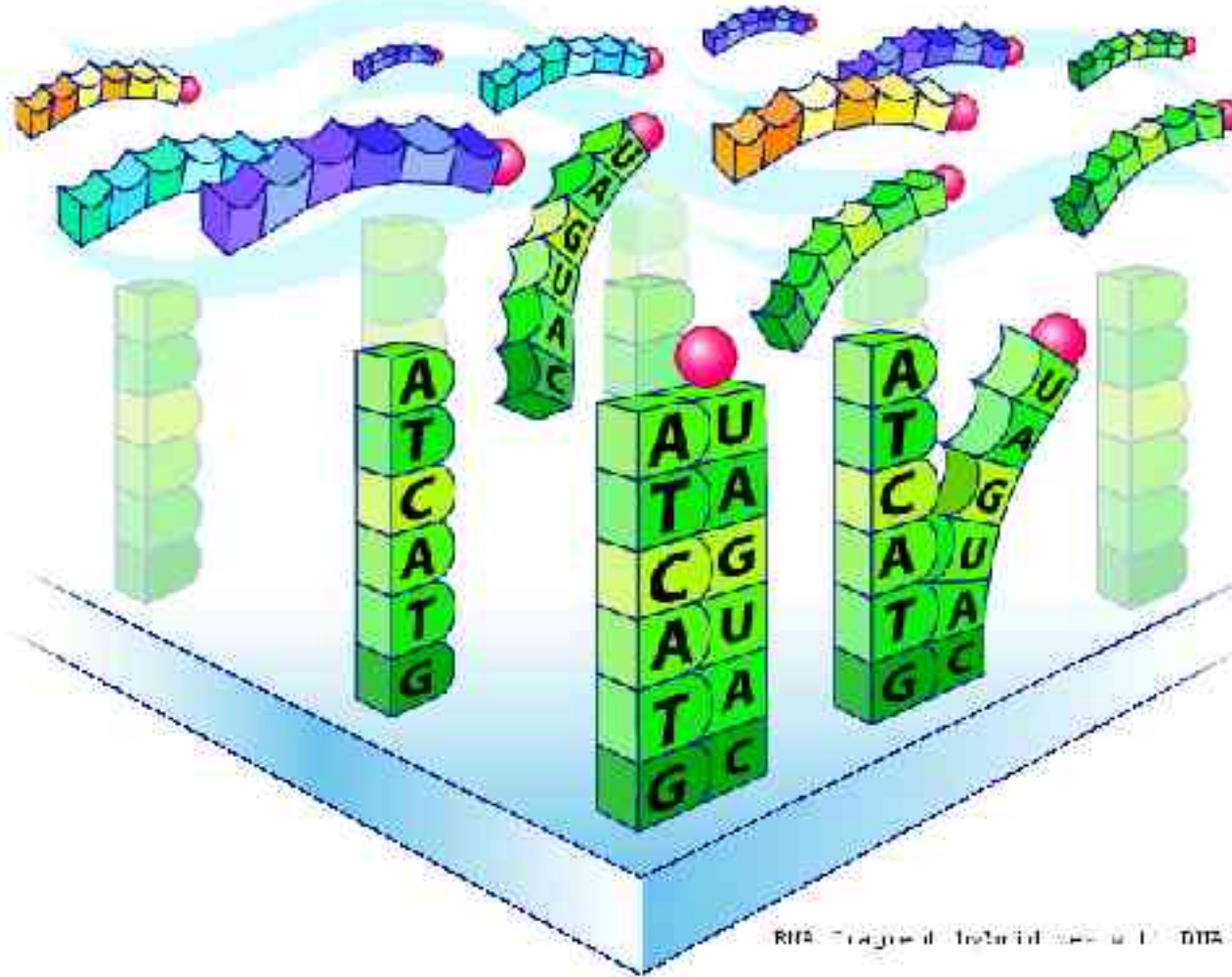


GeneChip® Affymetrix



GeneChip® Hybridization

RNA fragments with fluorochrome labels form stable hybrids



RNA fragments hybridize to DNA probes on chip

Image courtesy of Affymetrix.

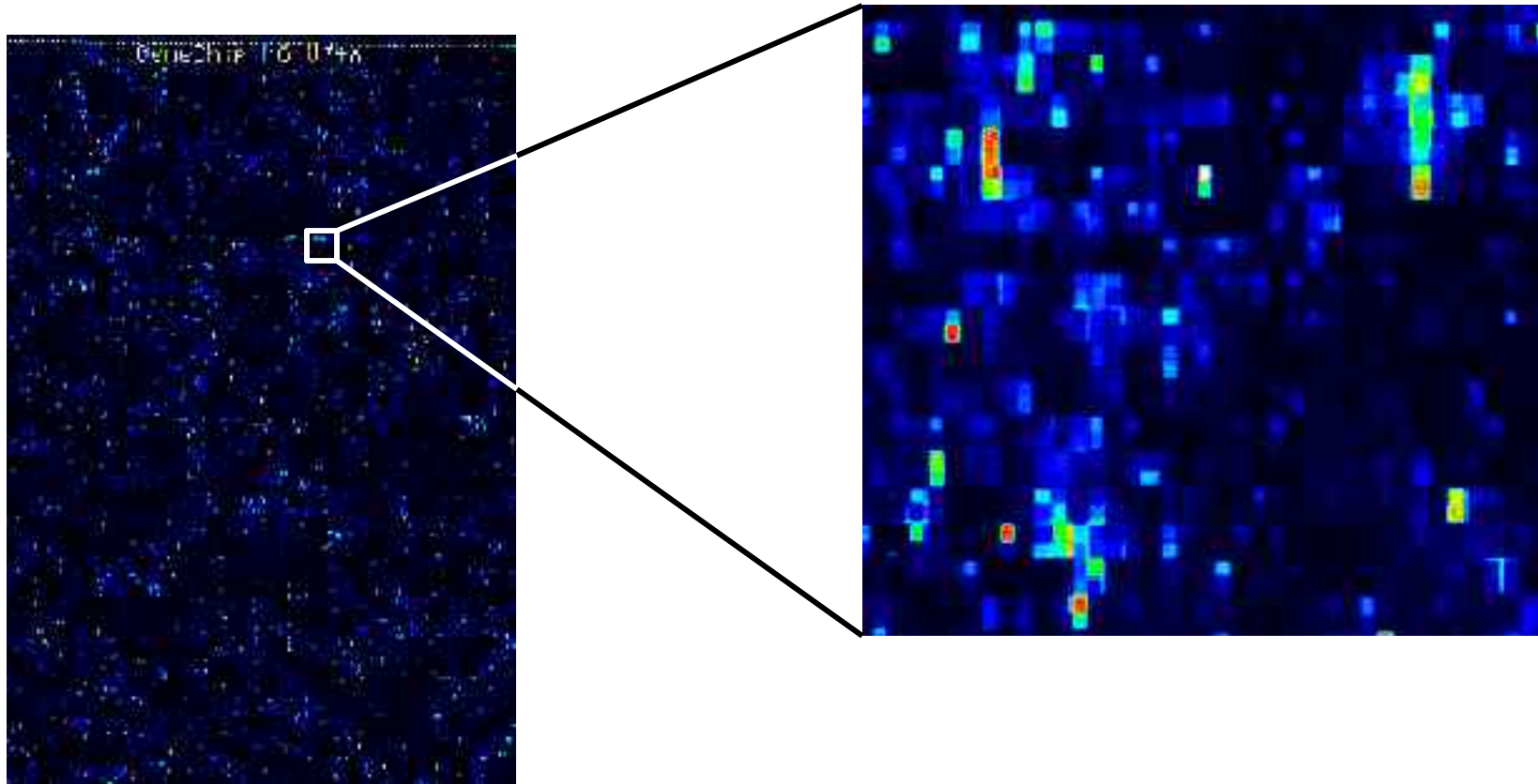


Affymetrix genearrays

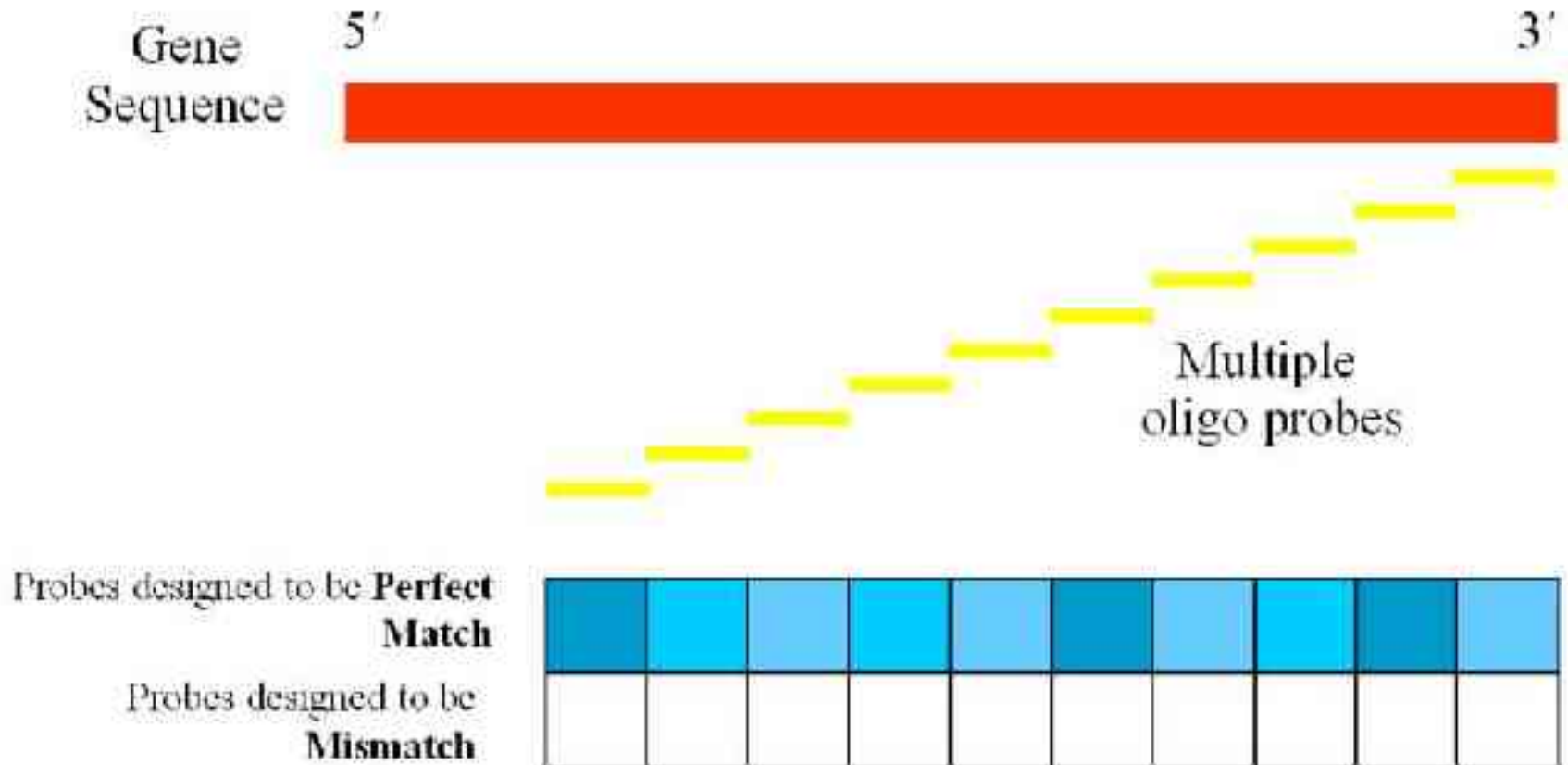
single color (color code indicates only hybridization intensity)

high density, perfectly addressable probes

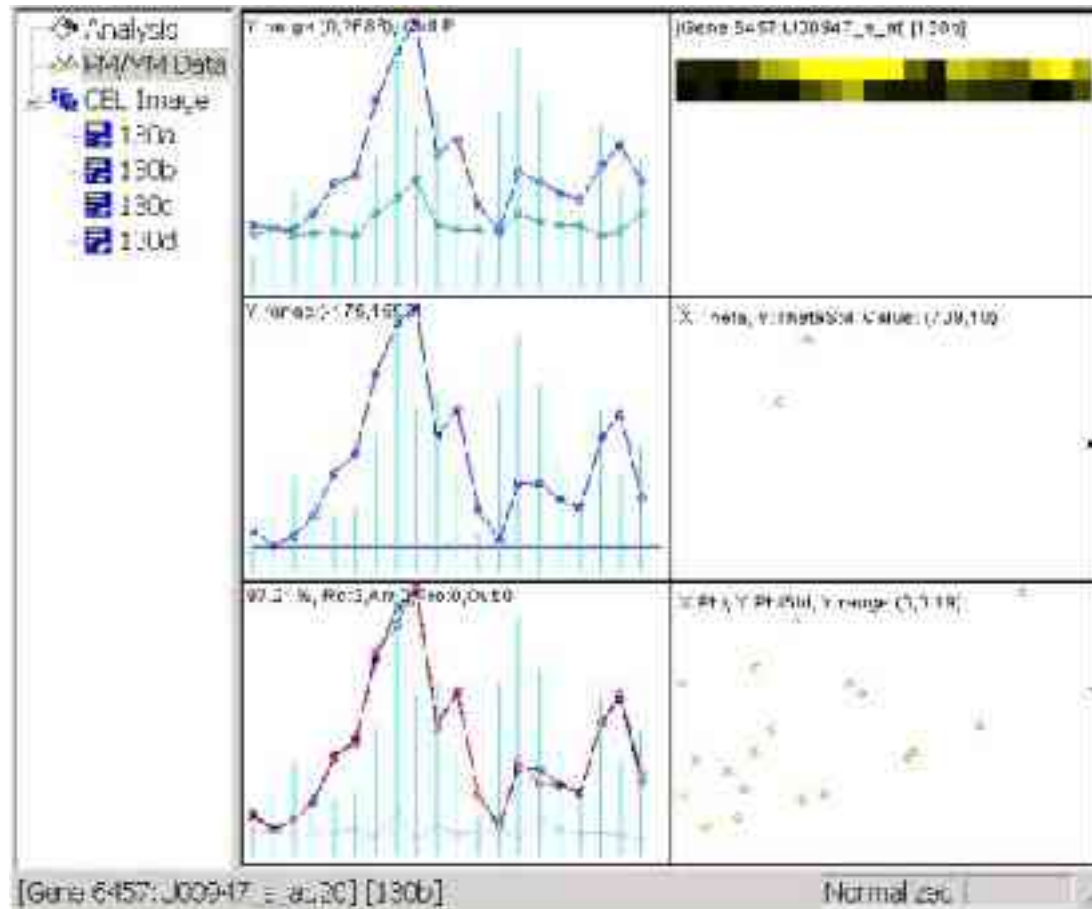
multiple probes per gene/mRNA



Affymetrix genechips contain “probe sets” instead of single probes per gene
better reliability of the results (each probe is [almost] an independent test)



Mismatch probes allow present/absence calls for every single probe set



PM probes
MM probes

Wilcoxon Signed Rank Test : non-parametric test; Take the paired observations (PM-MM), calculate the differences, and rank them from smallest to largest by **absolute value**. Add all the ranks associated with **positive** differences, giving the T_+ statistic. Finally, the **p-value** associated with this statistic is found from an appropriate table. (MathWorld)



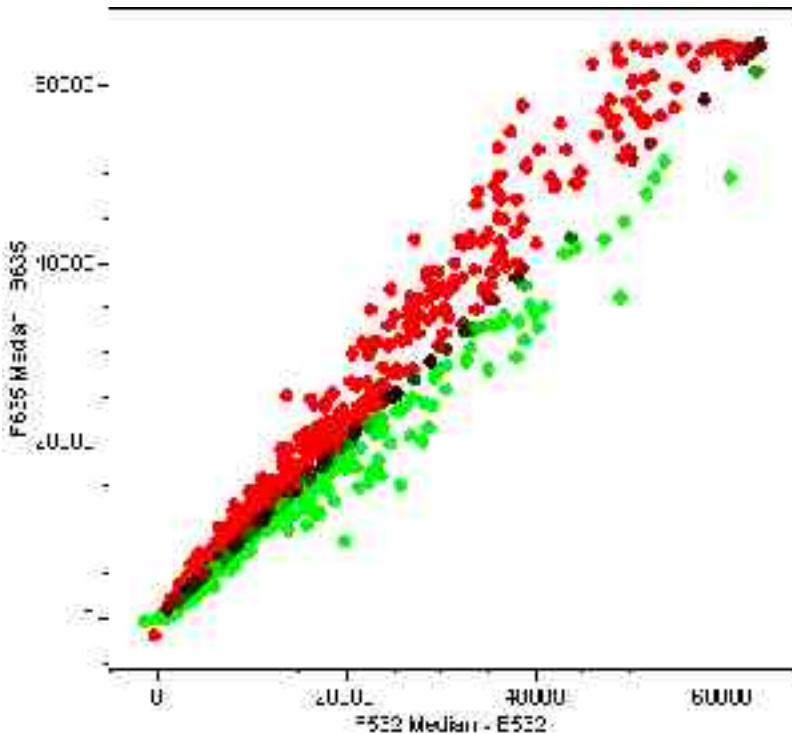
Finding and understanding differentially expressed genes

Sl. No.	Description	Unit	Quantity	Rate	Amount
1	Concrete	m ³	1.00	10000	10000
2	Reinforcement	kg	100	1000	100000
3	Formwork	m ²	100	1000	100000
4	Labour	man-days	100	1000	100000
5	Transport	km	100	1000	100000
6	Water	m ³	100	1000	100000
7	Electricity	kWh	100	1000	100000
8	Other	kg	100	1000	100000
9	Profit	%			
10	Total				
11	Concrete	m ³	1.00	10000	10000
12	Reinforcement	kg	100	1000	100000
13	Formwork	m ²	100	1000	100000
14	Labour	man-days	100	1000	100000
15	Transport	km	100	1000	100000
16	Water	m ³	100	1000	100000
17	Electricity	kWh	100	1000	100000
18	Other	kg	100	1000	100000
19	Profit	%			
20	Total				
21	Concrete	m ³	1.00	10000	10000
22	Reinforcement	kg	100	1000	100000
23	Formwork	m ²	100	1000	100000
24	Labour	man-days	100	1000	100000
25	Transport	km	100	1000	100000
26	Water	m ³	100	1000	100000
27	Electricity	kWh	100	1000	100000
28	Other	kg	100	1000	100000
29	Profit	%			
30	Total				
31	Concrete	m ³	1.00	10000	10000
32	Reinforcement	kg	100	1000	100000
33	Formwork	m ²	100	1000	100000
34	Labour	man-days	100	1000	100000
35	Transport	km	100	1000	100000
36	Water	m ³	100	1000	100000
37	Electricity	kWh	100	1000	100000
38	Other	kg	100	1000	100000
39	Profit	%			
40	Total				
41	Concrete	m ³	1.00	10000	10000
42	Reinforcement	kg	100	1000	100000
43	Formwork	m ²	100	1000	100000
44	Labour	man-days	100	1000	100000
45	Transport	km	100	1000	100000
46	Water	m ³	100	1000	100000
47	Electricity	kWh	100	1000	100000
48	Other	kg	100	1000	100000
49	Profit	%			
50	Total				
51	Concrete	m ³	1.00	10000	10000
52	Reinforcement	kg	100	1000	100000
53	Formwork	m ²	100	1000	100000
54	Labour	man-days	100	1000	100000
55	Transport	km	100	1000	100000
56	Water	m ³	100	1000	100000
57	Electricity	kWh	100	1000	100000
58	Other	kg	100	1000	100000
59	Profit	%			
60	Total				
61	Concrete	m ³	1.00	10000	10000
62	Reinforcement	kg	100	1000	100000
63	Formwork	m ²	100	1000	100000
64	Labour	man-days	100	1000	100000
65	Transport	km	100	1000	100000
66	Water	m ³	100	1000	100000
67	Electricity	kWh	100	1000	100000
68	Other	kg	100	1000	100000
69	Profit	%			
70	Total				
71	Concrete	m ³	1.00	10000	10000
72	Reinforcement	kg	100	1000	100000
73	Formwork	m ²	100	1000	100000
74	Labour	man-days	100	1000	100000
75	Transport	km	100	1000	100000
76	Water	m ³	100	1000	100000
77	Electricity	kWh	100	1000	100000
78	Other	kg	100	1000	100000
79	Profit	%			
80	Total				
81	Concrete	m ³	1.00	10000	10000
82	Reinforcement	kg	100	1000	100000
83	Formwork	m ²	100	1000	100000
84	Labour	man-days	100	1000	100000
85	Transport	km	100	1000	100000
86	Water	m ³	100	1000	100000
87	Electricity	kWh	100	1000	100000
88	Other	kg	100	1000	100000
89	Profit	%			
90	Total				
91	Concrete	m ³	1.00	10000	10000
92	Reinforcement	kg	100	1000	100000
93	Formwork	m ²	100	1000	100000
94	Labour	man-days	100	1000	100000
95	Transport	km	100	1000	100000
96	Water	m ³	100	1000	100000
97	Electricity	kWh	100	1000	100000
98	Other	kg	100	1000	100000
99	Profit	%			
100	Total				

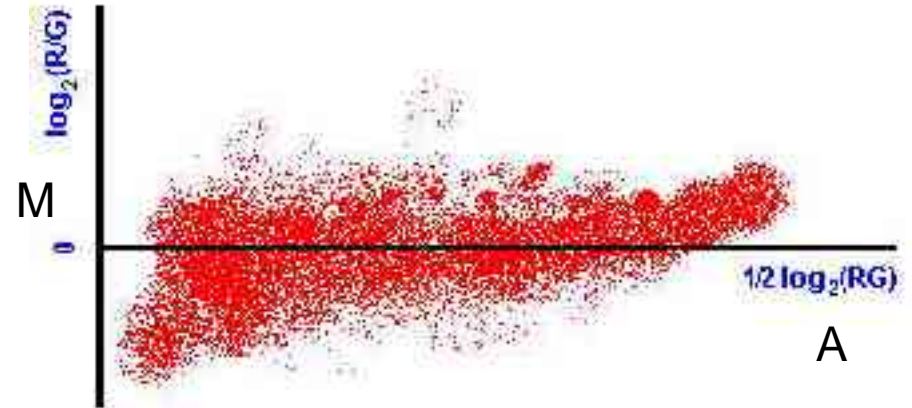


Scatter plots

classical scatter plot



M-A plot for microarray analysis



$$M = \log_2 \left(\frac{R}{G} \right)$$

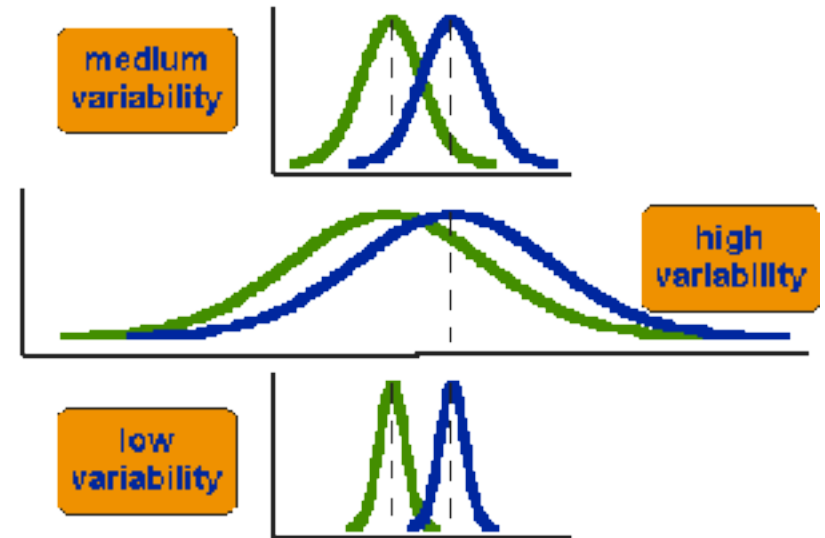
$$A = \log_2 \sqrt{RG} \quad \text{OR} \quad \frac{1}{2} \log_2 RG$$

Differentially expressed genes are higher (or lower) in one of the samples

Use an appropriate cut-off ('distance' from diagonal) to select relevant genes **highly arbitrary!**

t-test = statistical significance of observed difference

- requires independent experimental replication
- assumes the data are identically normally distributed

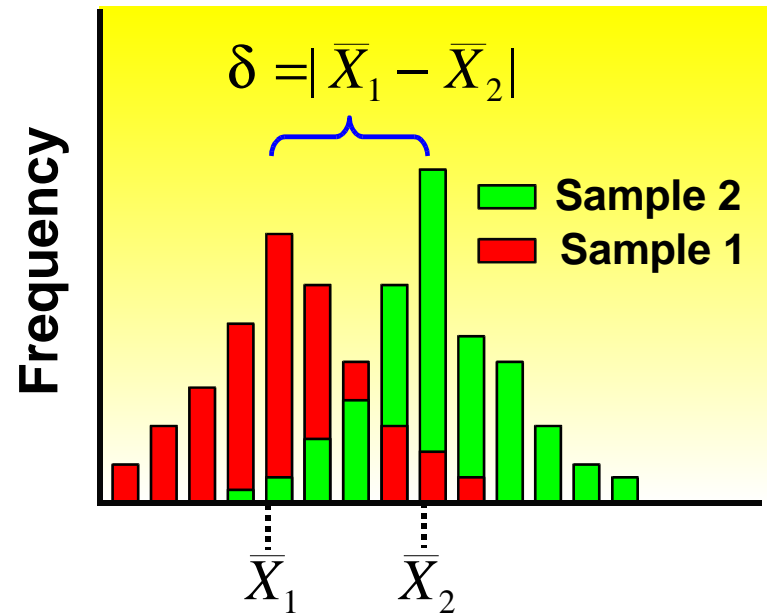


$$t = \frac{\text{difference of means}}{\text{variability}}$$

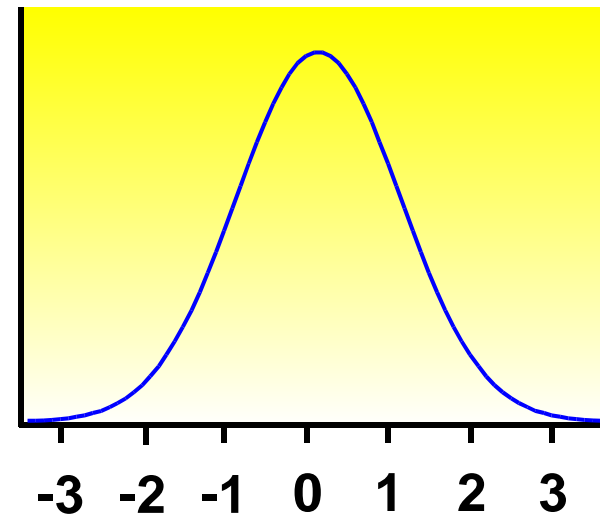
$$t = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{\text{var}_T}{n_T} + \frac{\text{var}_C}{n_C}}}$$

Testing an intrinsic hypothesis

- Two samples (1, 2) with mean expression that differ by some amount δ .
- If $H_0 : \delta = 0$ is true, then the expected distribution of the test statistic t is

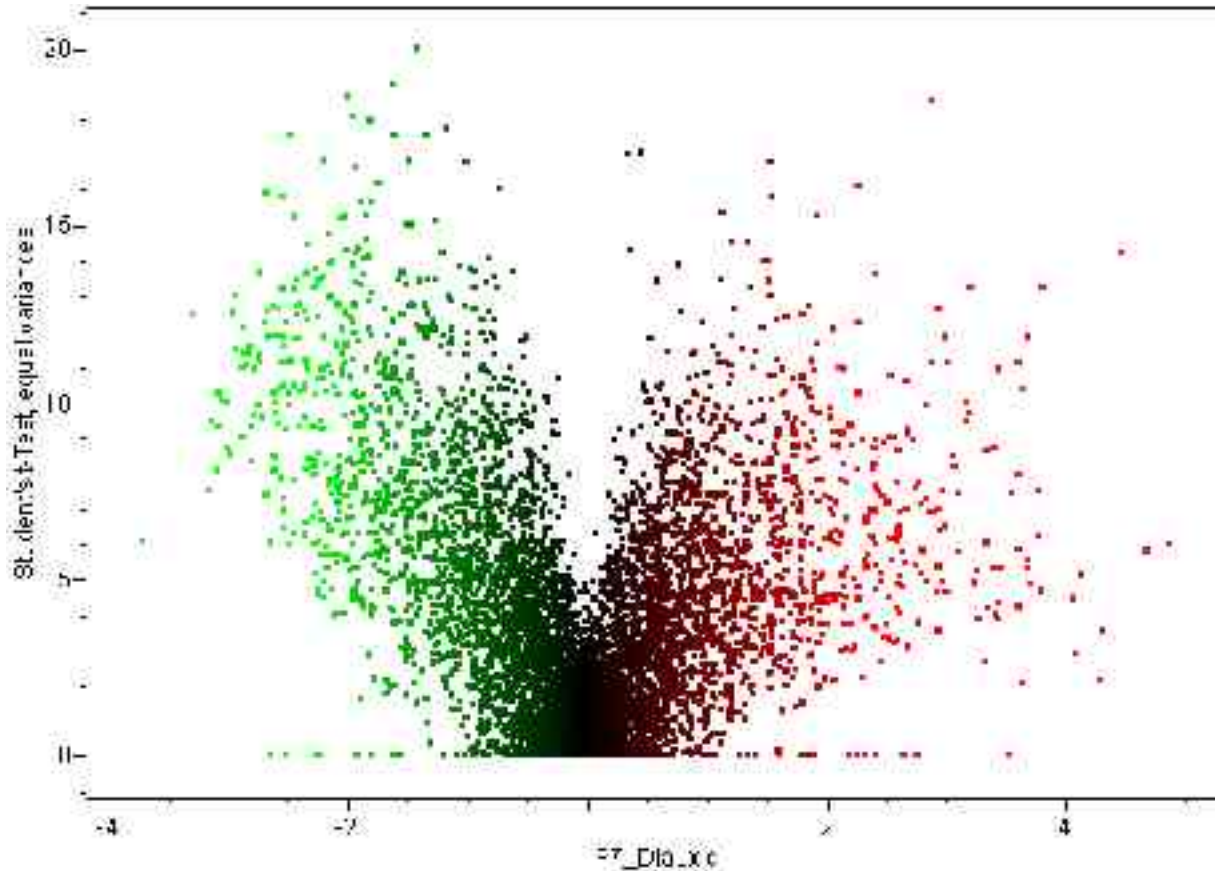


Probability



$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}$$

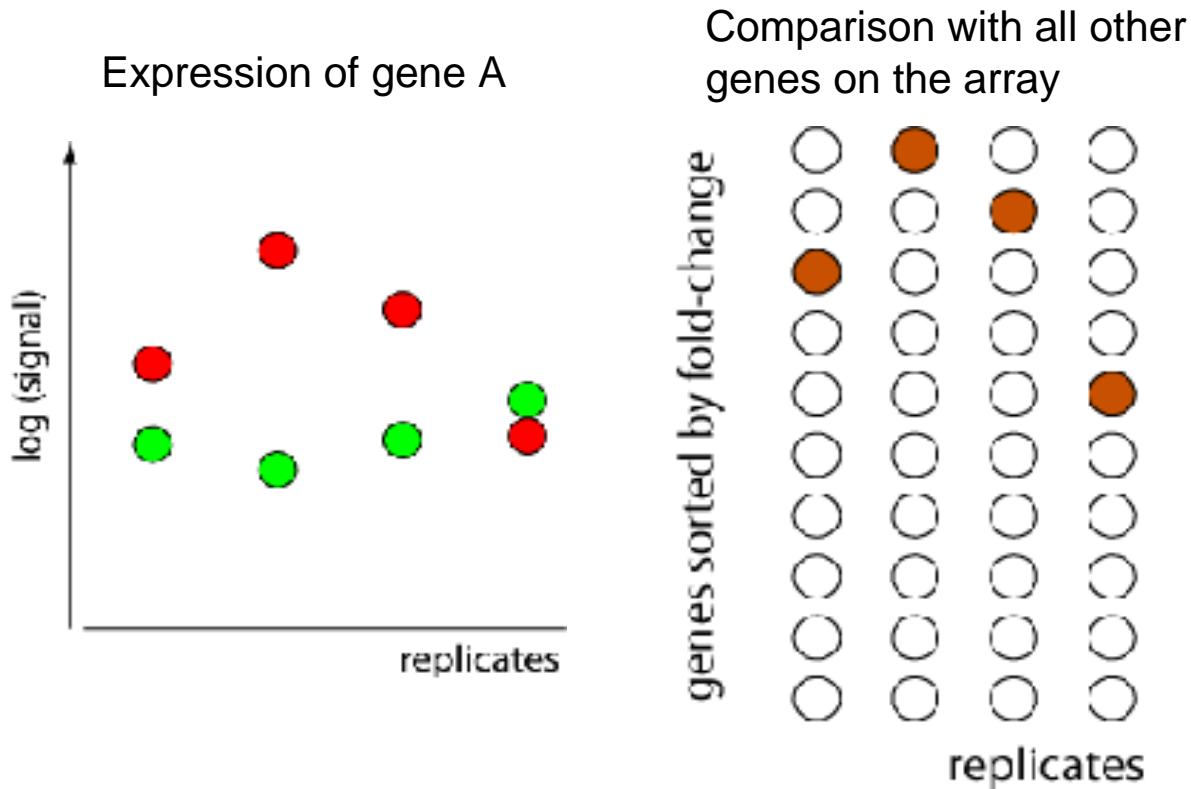
Volcano plot



Scatter plot of $-\log(p\text{-value})$ from a t-test vs. log ratio. Visualises fold-change and statistical significance at the same time: Find genes that are significant and have large fold change, **and** genes that are significant but have small fold change.



Is this gene changed?



Rank Product:

$$RP = (3/10) * (1/10) * (2/10) * (5/10)$$

- intuitive
- non-parametric, powerful test statistic
- more reliable detection of changed genes in noisy data with few replicates

Significance estimate based on random permutations:

Probability that gene A shows such an effect by chance: $p \leq 0.03$

Expectation to see any gene (out of 10) with such a effect: E-value ≈ 0.5

Multiple Testing Problem

- microarrays measure expression of $>10,000$ genes at the same time many thousands of statistical tests are performed
- type 1-error: Calling a gene significantly changed, even if it's just by chance protect yourself by Bonferroni correction
- type 2-error: Missing a significantly changed gene reduce this problem by Benjamini-Hochberg false-discovery rate procedure



Multiple Testing Problem

Bonferroni correction. n independent tests, control the probability that a spurious result passes the test at significance level α adjust acceptance level for each individual test as:

$$P(\text{some } T_i \text{ passes} | H_0) \leq \alpha,$$

$$P(T_i \text{ passes} | H_0) \leq \frac{\alpha}{n}$$

Benjamini-Hochberg False Discovery Rate. Control the number of false positives (N_{10}) among the top R genes at the significance level α .

$$\text{FDR} = \begin{cases} \frac{N_{10}}{R} & \text{if } R > 0, \\ 0, & \text{if } R = 0. \end{cases}$$

0. Select $0 < \alpha < 1$.

1. Define $P_{(0)} \equiv 0$ and

$$R_{\text{BH}} = \max \left\{ 0 \leq i \leq m: P_{(i)} \leq \alpha \frac{i}{m} \right\} \rightarrow E(\text{FDR}) \leq \frac{M_0}{m} \alpha \leq \alpha$$

2. Reject H_0 for every test where $P_j \leq P_{(R_{\text{BH}})}$.



The result of “differential expression” statistical analysis a long list of genes!

	Fold-Change	Gene Symbol	Gene Title
1	26.45	TNFAIP6	tumor necrosis factor, alpha-induced protein 6
2	25.79	THBS1	thrombospondin 1
3	23.08	SERPINE2	serine (or cysteine) proteinase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 2
4	21.5	PTX3	entaxin-related gene, rapidly induced by IL-1 beta
	8.82	HBS1	hrombospondin 1
	6.68	XCL10	hemokine (C-X-C motif) ligand 10
7	18.23	CCL4	chemokine (C-C motif) ligand 4
8	14.85	SOD2	superoxide dismutase 2, mitochondrial
9	13.62	IL1B	interleukin 1, beta
10	11.53	CCL20	chemokine (C-C motif) ligand 20
11	11.82	CCL3	chemokine (C-C motif) ligand 3
12	11.27	SOD2	uperoxide dismutase 2, mitochondrial
13	10.89	GCH1	GTP cyclohydrolase 1 (dopa-responsive dystonia)
14	10.73	IL8	interleukin 8
15	9.98	ICAM1	intercellular adhesion molecule 1 (CD54), human rhinovirus receptor
16	9.97	SLC2A6	solute carrier family 2 (facilitated glucose transporter), member 6
17	8.36	BCL2A1	BCL2-related protein A1
18	7.33	TNFAIP2	tumor necrosis factor, alpha-induced protein 2
19	6.97	SERPINB2	serine (or cysteine) proteinase inhibitor, clade B (ovalbumin), member 2
20	6.69	MAFB	v-maf musculoaponeurotic fibrosarcoma oncogene homolog B (avian)

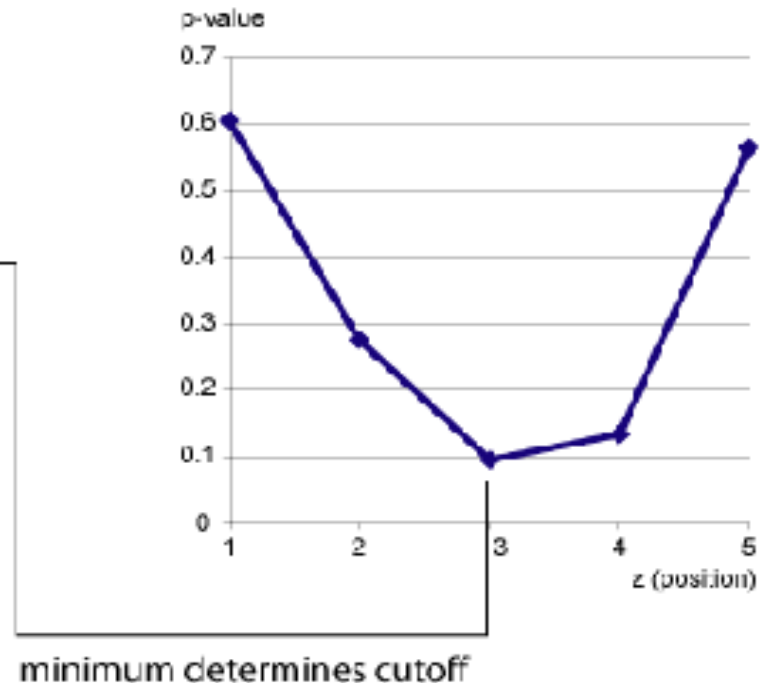
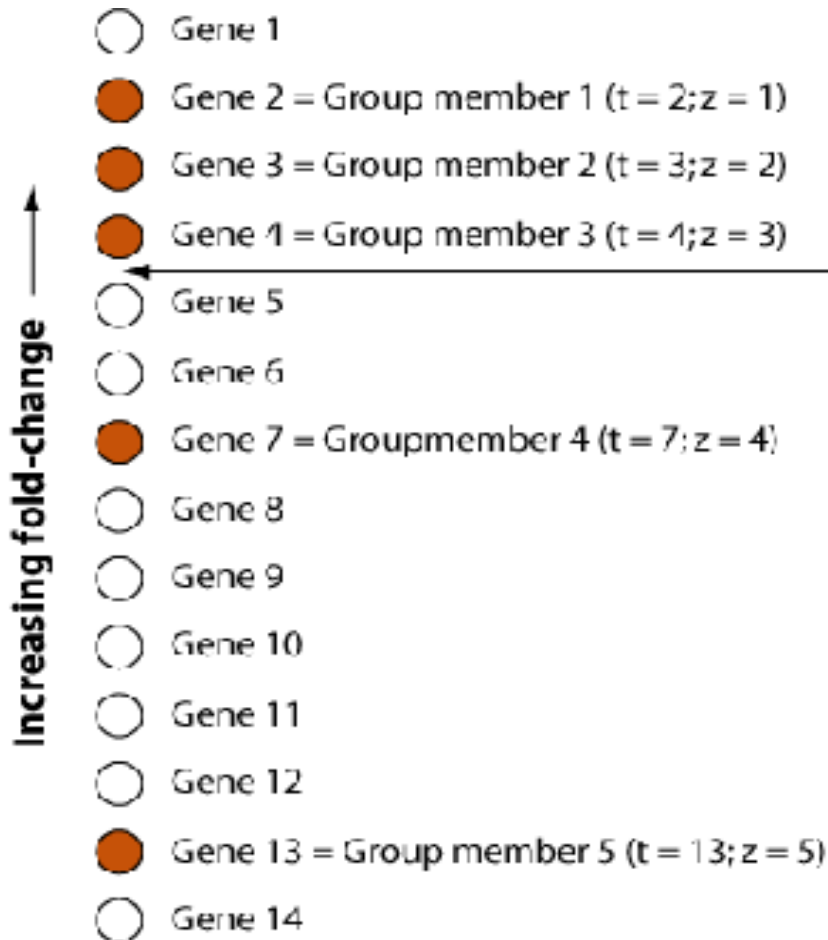


Biological Interpretation Strategy

- Are certain types of genes more common at the top of the list and is that significant?
- Challenges:
 - Some types of genes are more common in the genome/on the array
 - The list of genes usually stops at an arbitrary cut-off (“significantly changed genes”)
 - Classifying genes according to “gene type” is a tedious task
 - Expectations and focused expertise might bias the interpretation
 - Early discoveries might restrict further analysis
- Solution: Automated procedure using available annotations



iterative Group Analysis (iGA)



total number of genes $n = 14$
group members $x = 5$

iGA uses a simple hypergeometric distribution to obtain p-values
 Breitling et al. (2004), *BMC Bioinformatics*, 5:34.



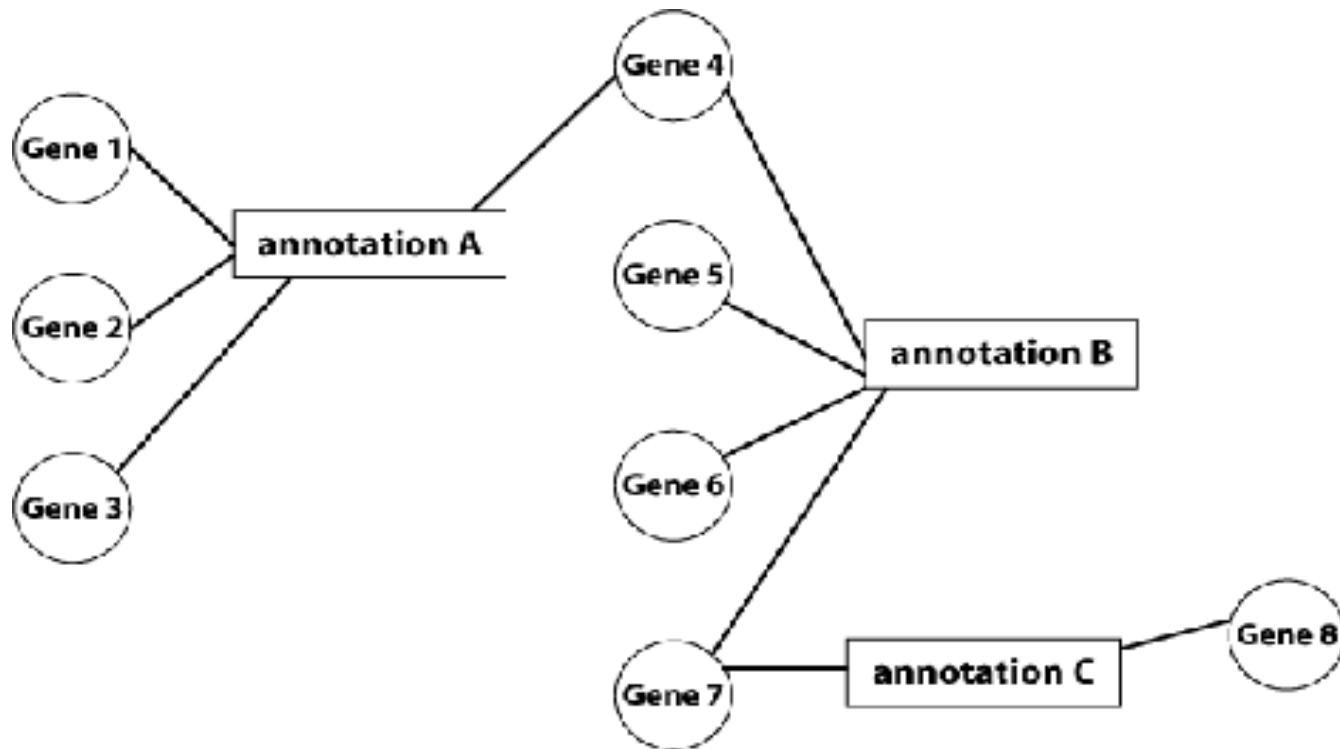
Possible sources of classification

- adjacency in metabolic networks
- shared biological processes
- co-expression in microarray experiments
- co-occurrence in the biomedical literature
- gene ontology annotations (shared terms from a controlled vocabulary)



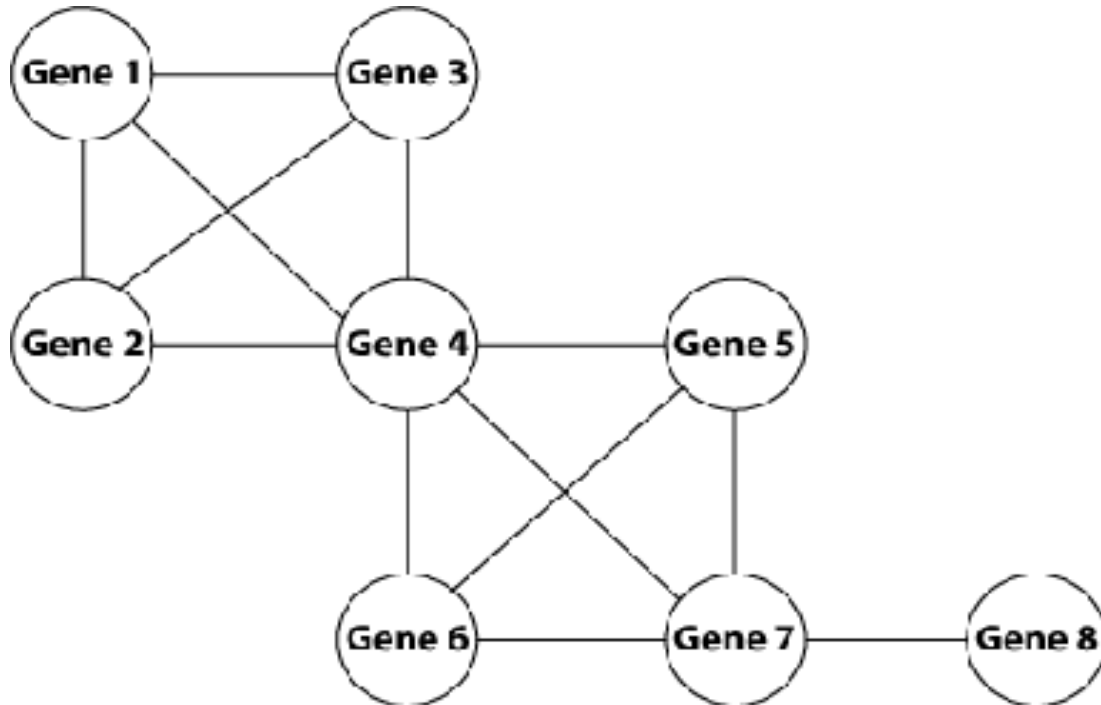
Graph-based iGA

exploits the overlap of annotations to produce a comprehensive picture of the microarray results



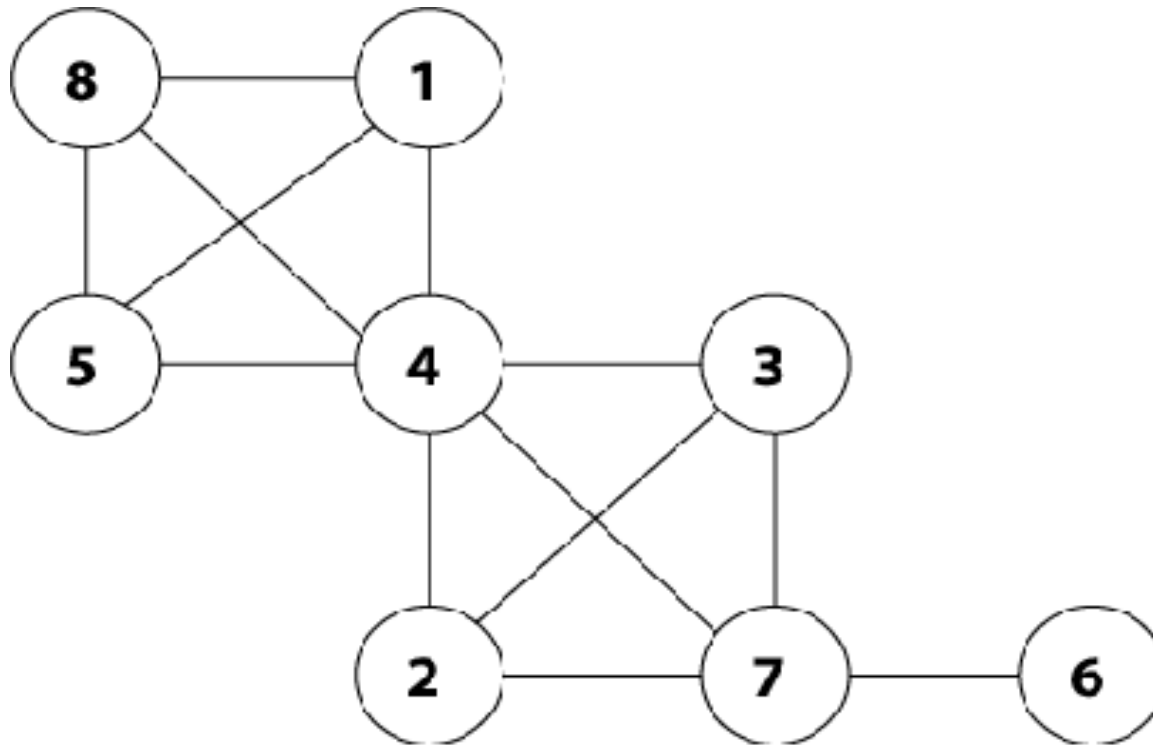
Graph-based iGA

1. step: build the network



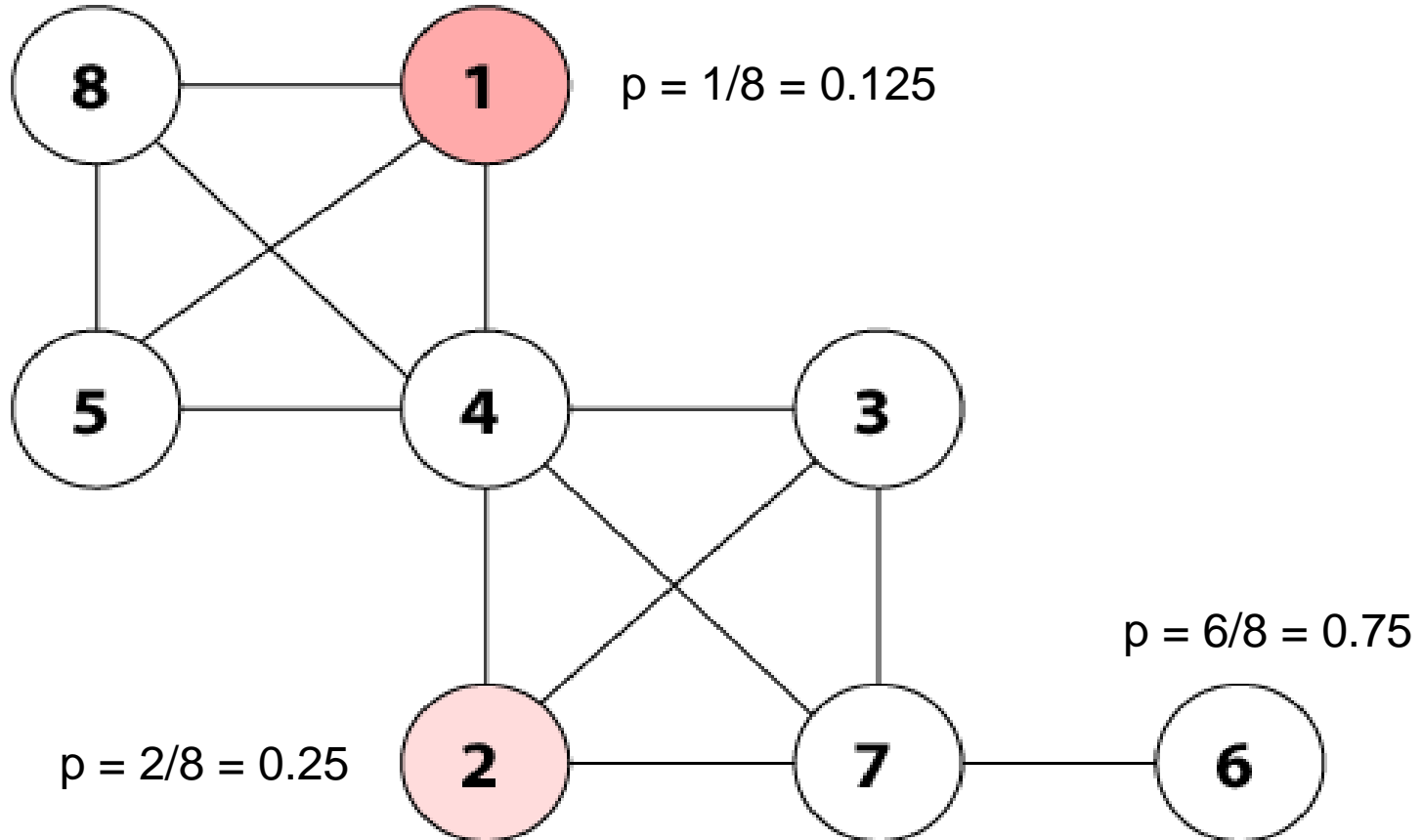
Graph-based iGA

2. step: assign experimentally determined ranks to genes



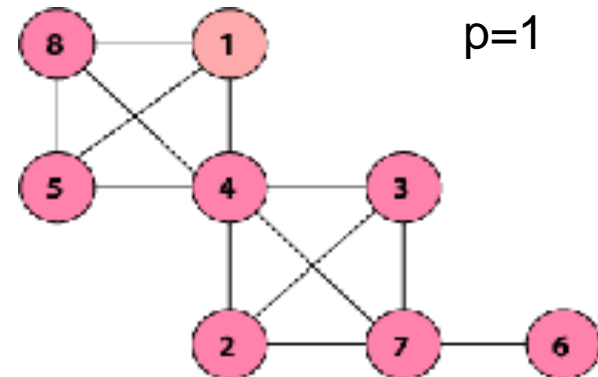
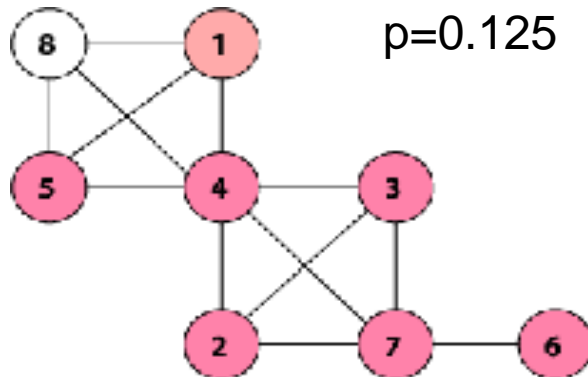
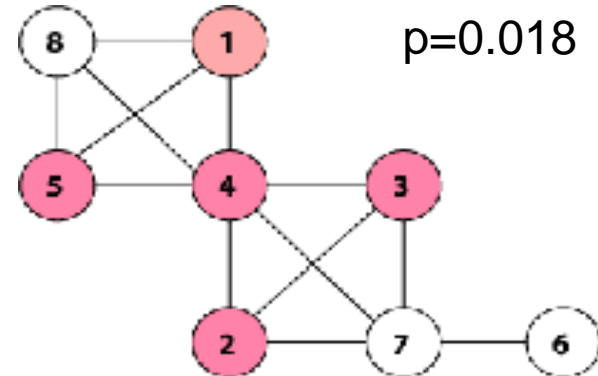
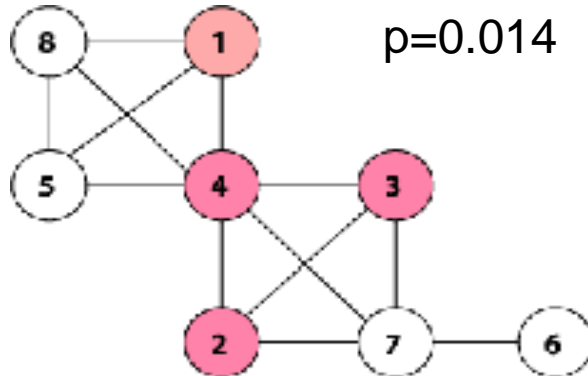
Graph-based iGA

3. step: find local minima



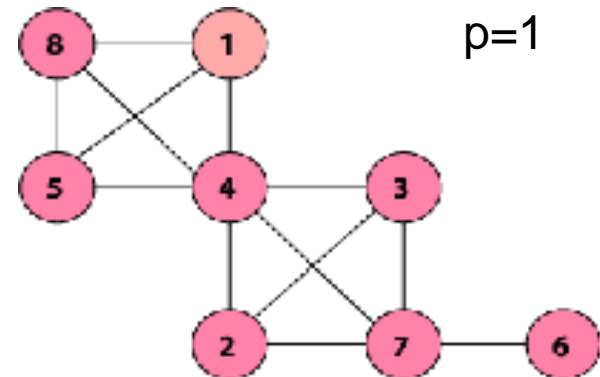
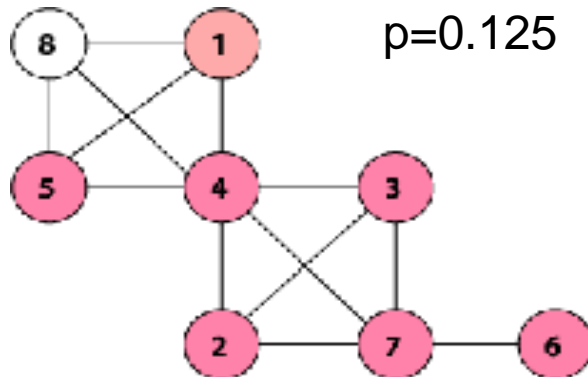
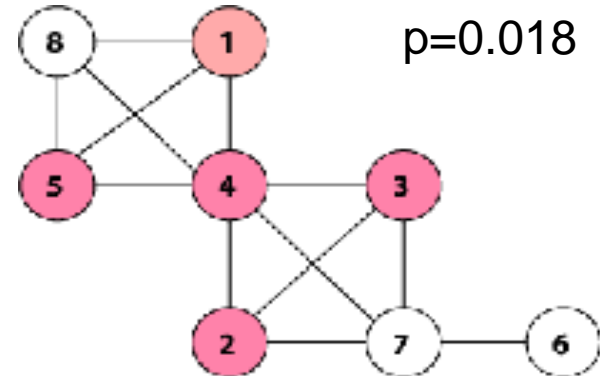
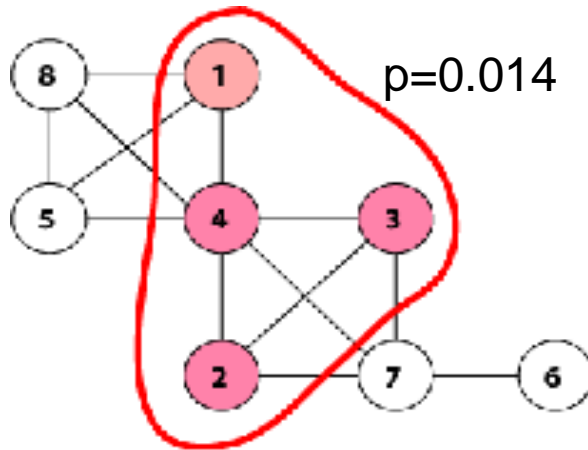
Graph-based iGA

4. step: extend subgraph from minima

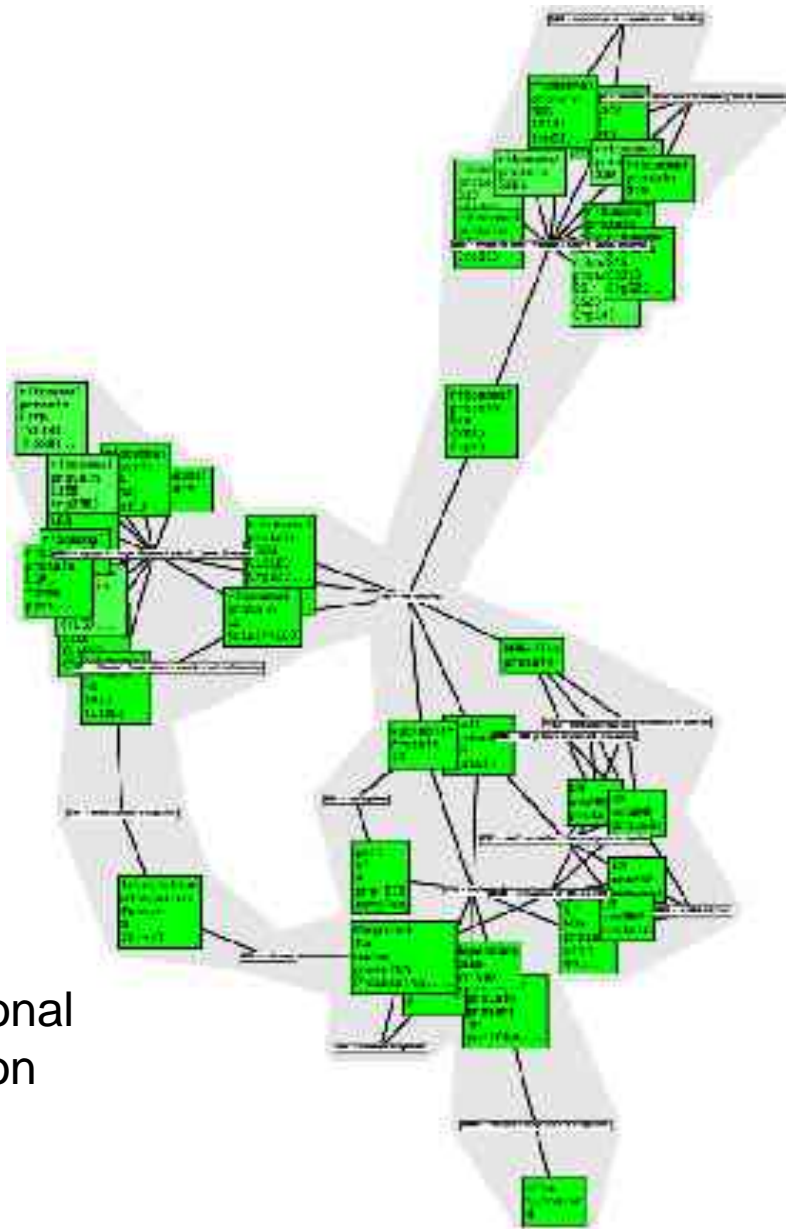


Graph-based iGA

5. step: select p-value minimum



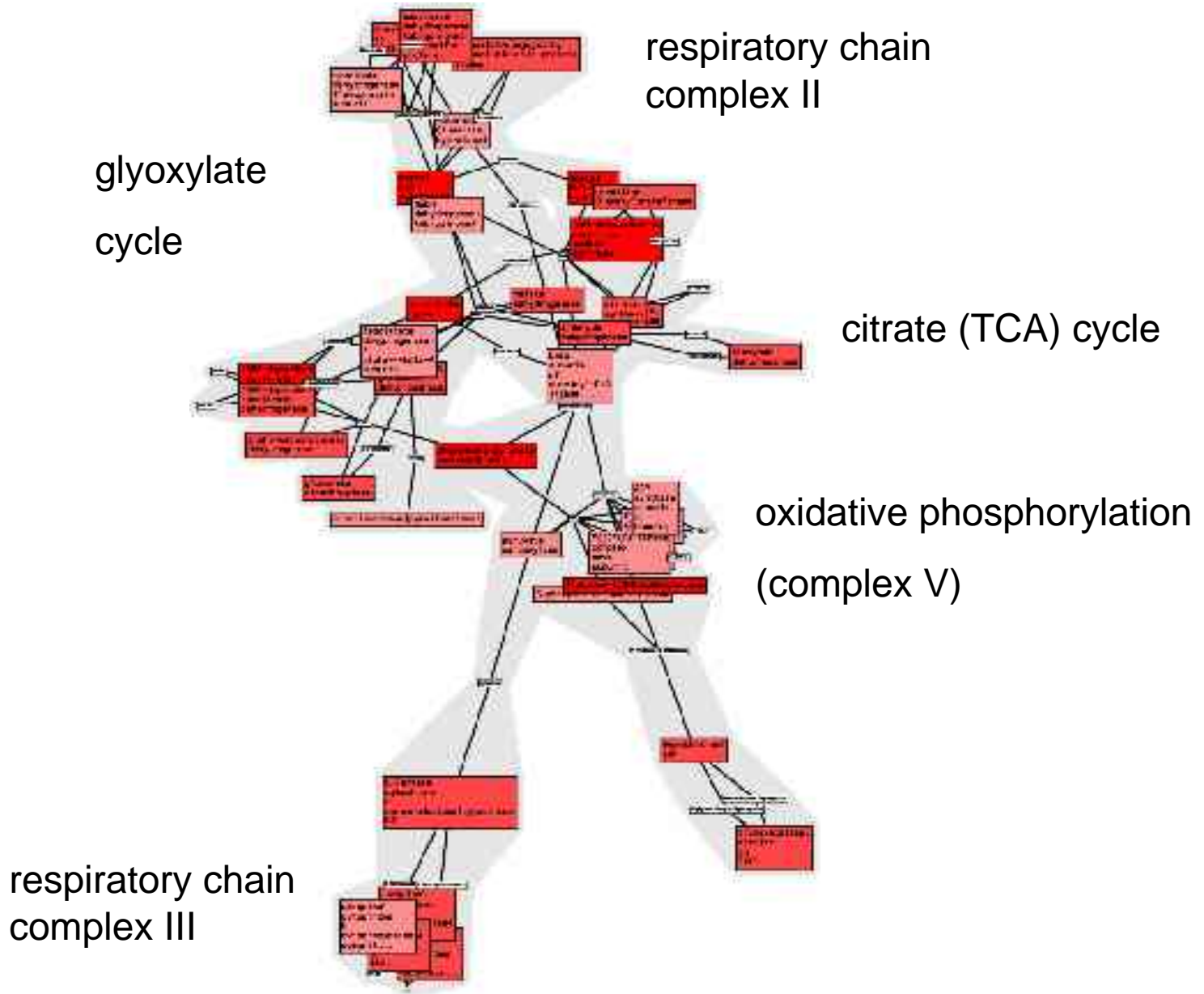
large
ribosomal
subunit



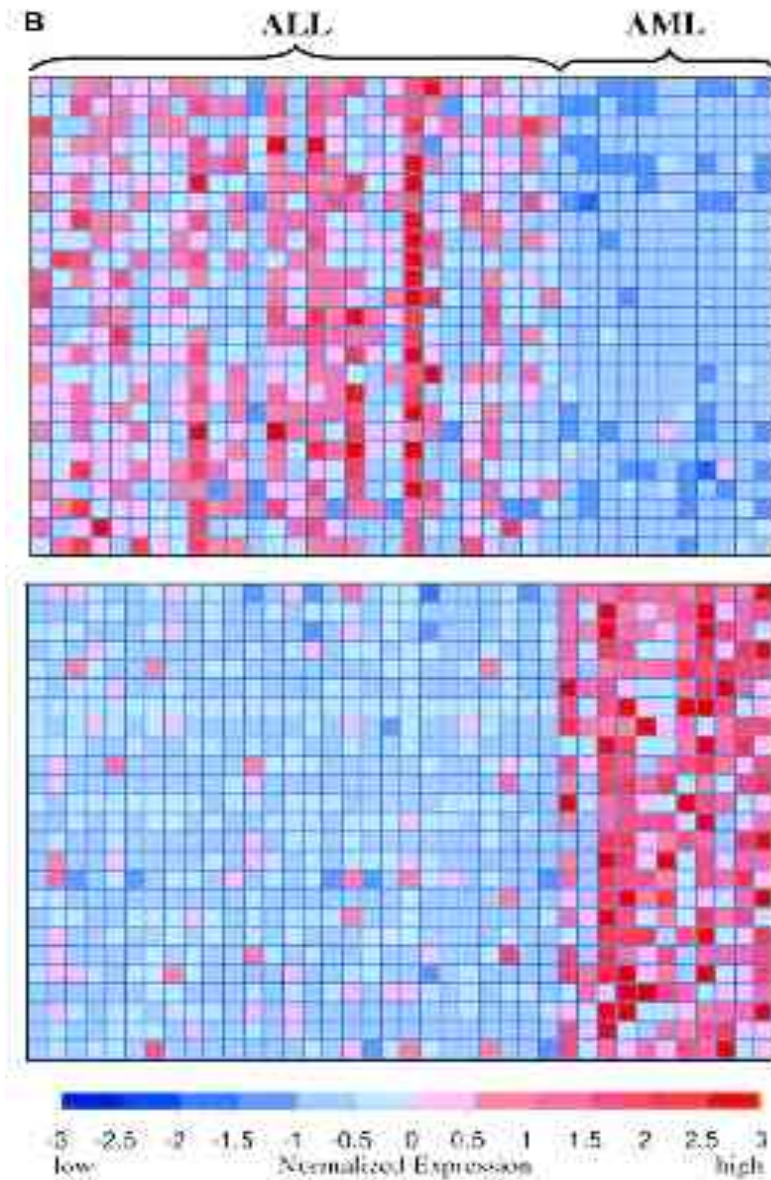
small
ribosomal
subunit

nucleolar
rRNA
processing

translational
elongation



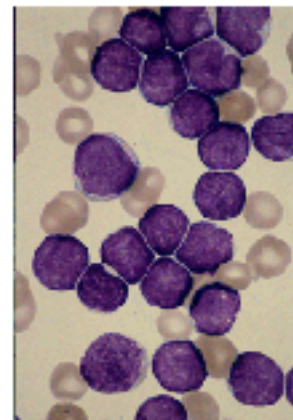
Advanced analysis (clustering and classification)



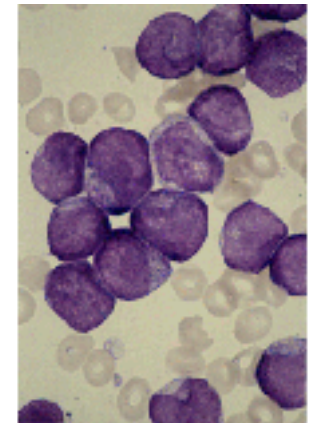
Classical study of cancer subtypes

Golub et al. (1999)

identification of diagnostic genes



ALL
acute lymphoblastic leukemia
(lymphoid precursors)



AML
acute myeloid leukemia
(myeloid precursor)

Similarity between microarray experiments or expression patterns

distance between points in high dimensional space

Pearson correlation

(looks for similarity in shape of the response profile, not the absolute values)

$$r = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \left(\frac{y_i - \bar{y}}{\sigma_y} \right)$$

Euclidean distance

(shortest direct path),
takes absolute expression level into account

$$d = \sum_{i=1}^n (x_i - y_i)^2$$

Manhattan (or city-block) distance

$$d = \sum_{i=1}^n |x_i - y_i|$$



Gene expression data analysis

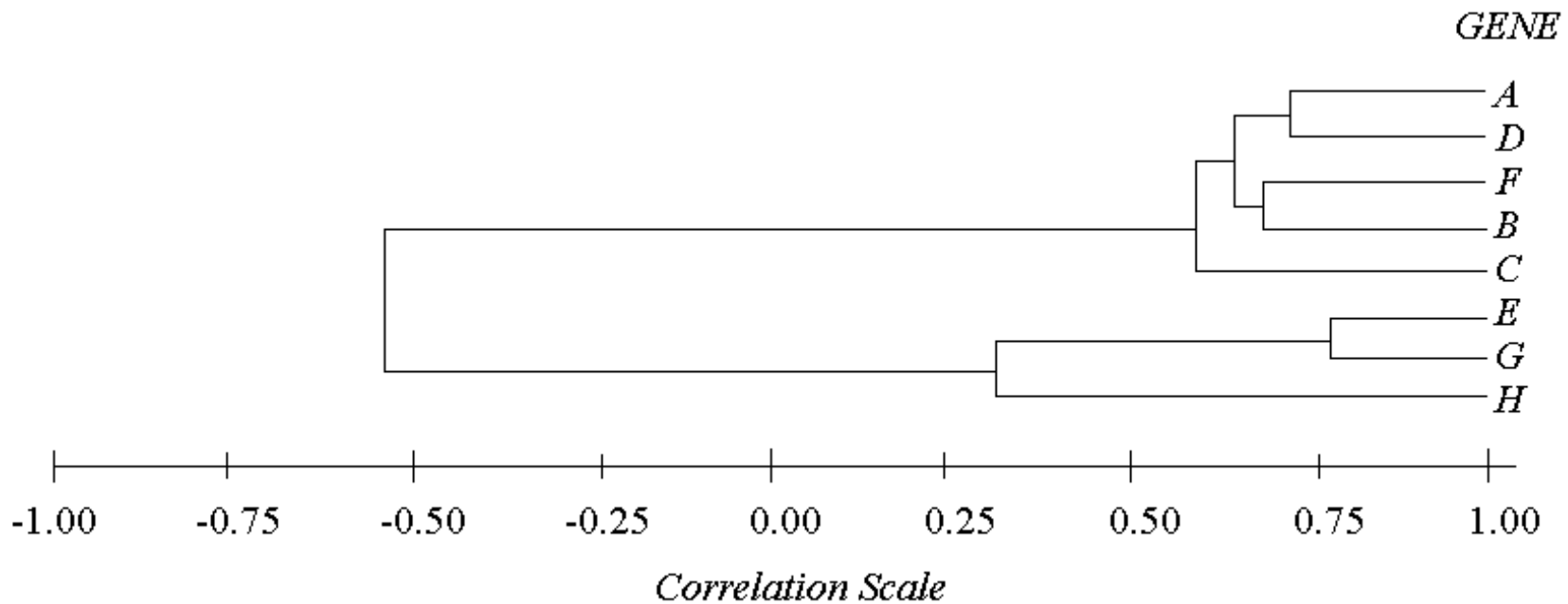


(Ramaswamy and Golub 2002)

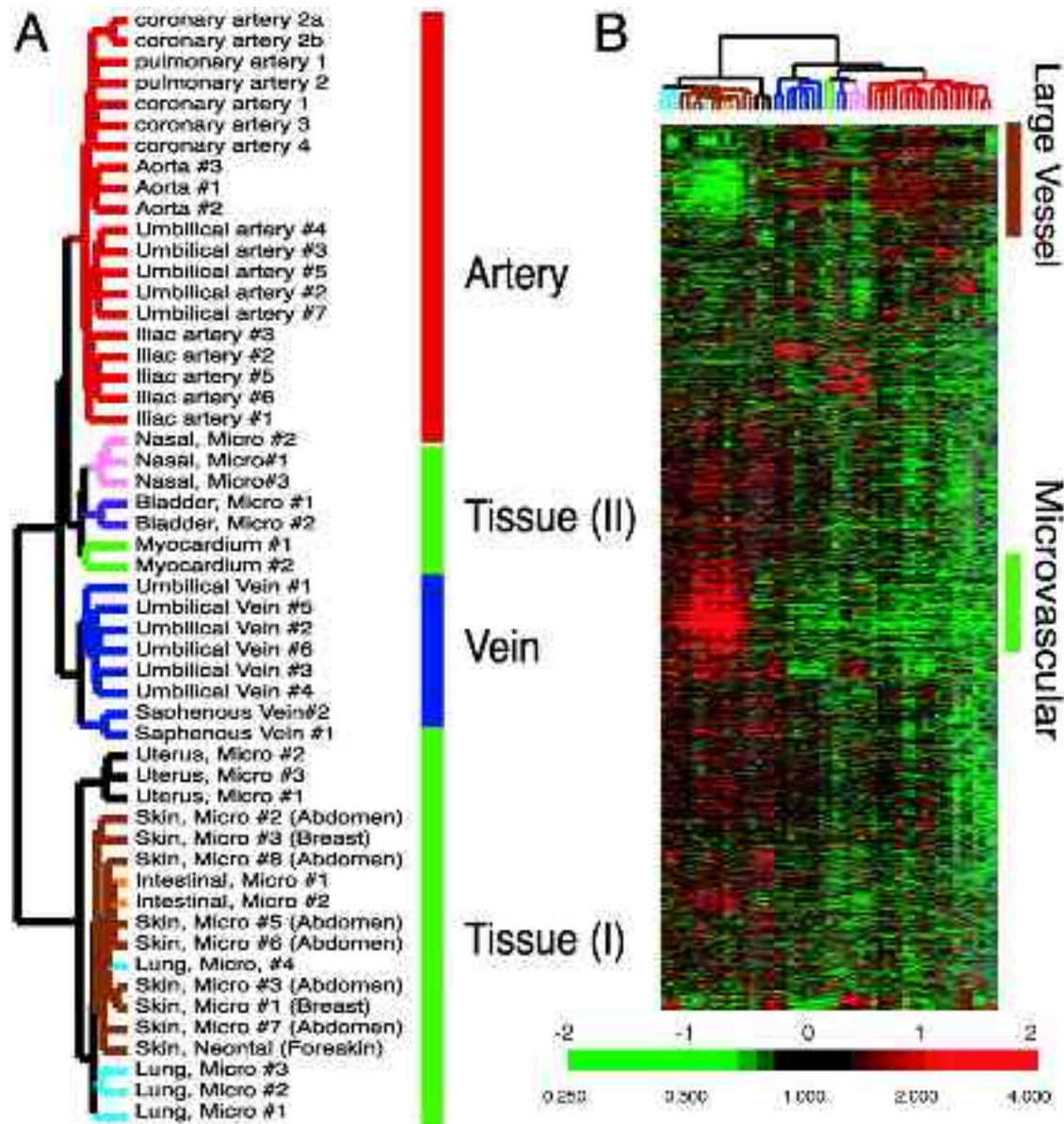


Hierarchical clustering

- Combine most similar genes into agglomerative clusters, build tree of genes
- Do the same procedure along the second dimension to cluster samples
- Display the sorted expression values as a heatmap



Hierarchical clustering results



Chi et al., PNAS | September 16, 2003 | vol. 100 | no. 19 | 10623-10628

“Endothelial cell diversity revealed by global expression profiling”



Biologically Valid Linear Factor Models of Gene Expression

$$e_{ga} = \sum_{p=1}^{\mathcal{P}} A_{pa} \theta_{gp} + n_{ga}.$$

e_{ga}

expression level of gene g in array a

θ_{gp}

expression level of gene x in hypothetical process p

A_{pa}

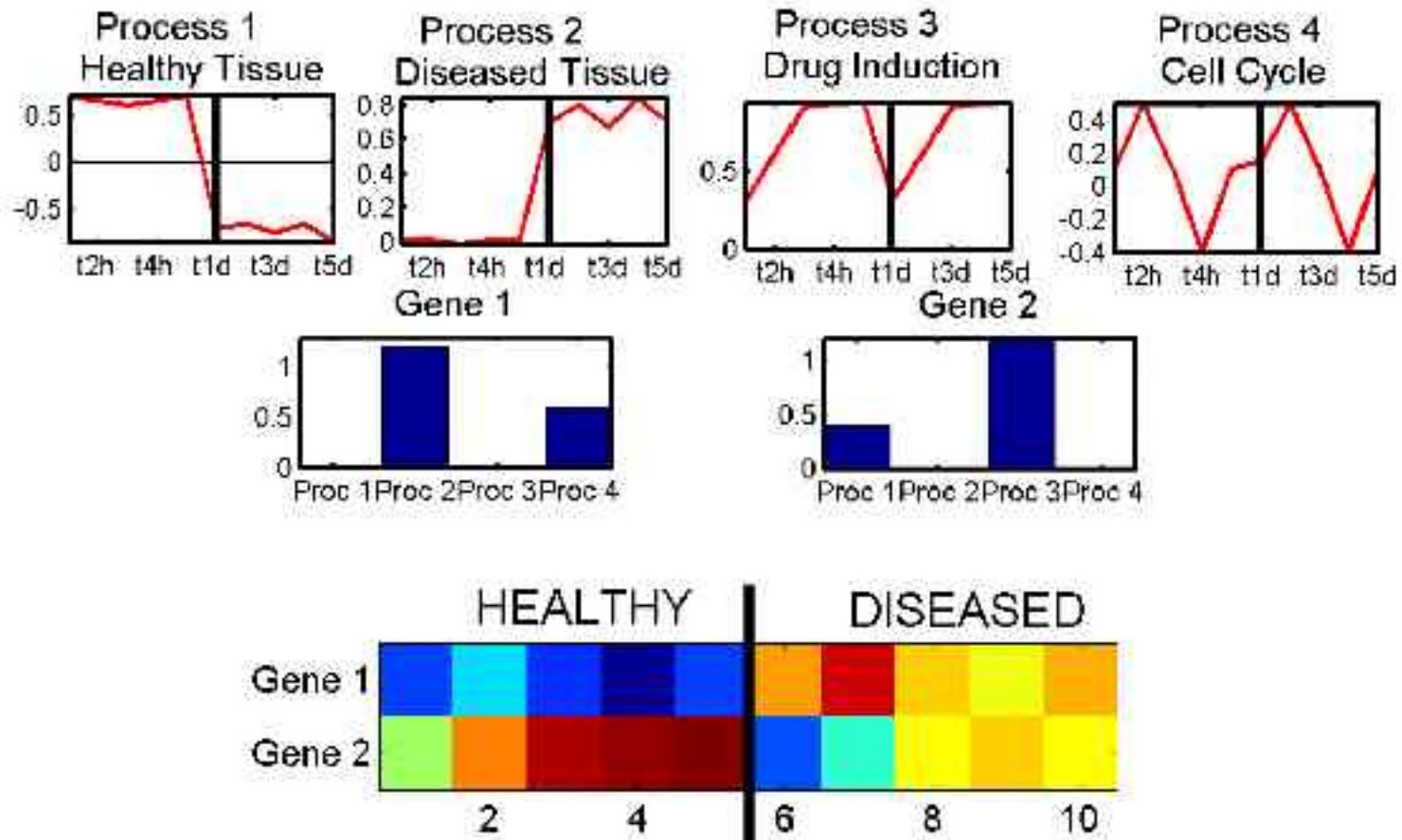
contribution of process p to expression pattern in array a

n_{ga}

experiment- and gene-specific noise

$$\mathbf{E} = \mathbf{\Theta} \mathbf{A} + \mathbf{N}.$$

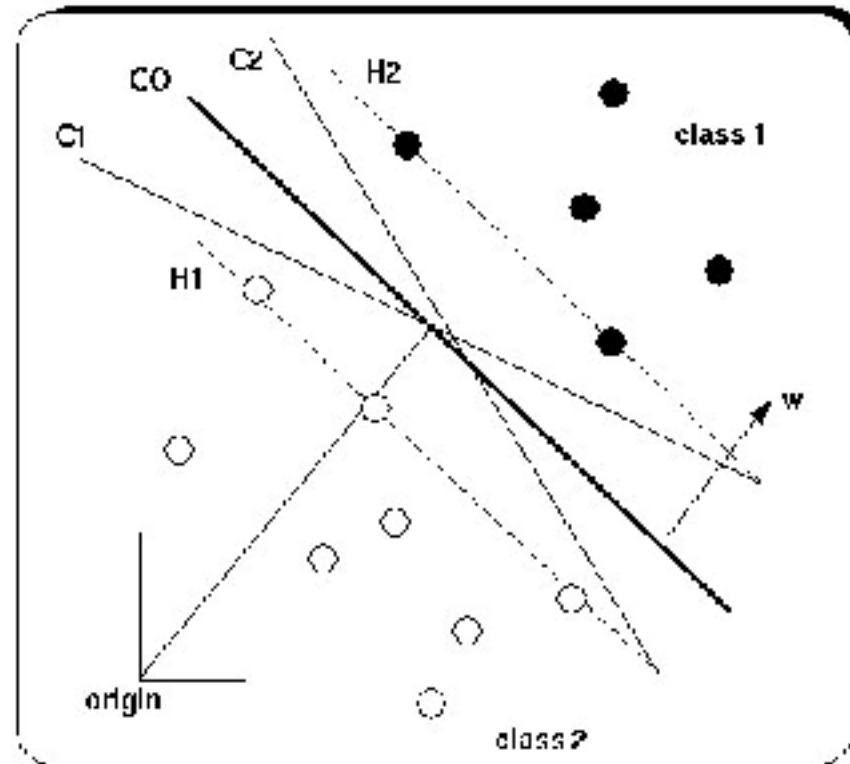
Biologically Valid Linear Factor Models of Gene Expression



M. Girolami & R. Breitling (2004), *Bioinformatics*, 20(17):3021-33

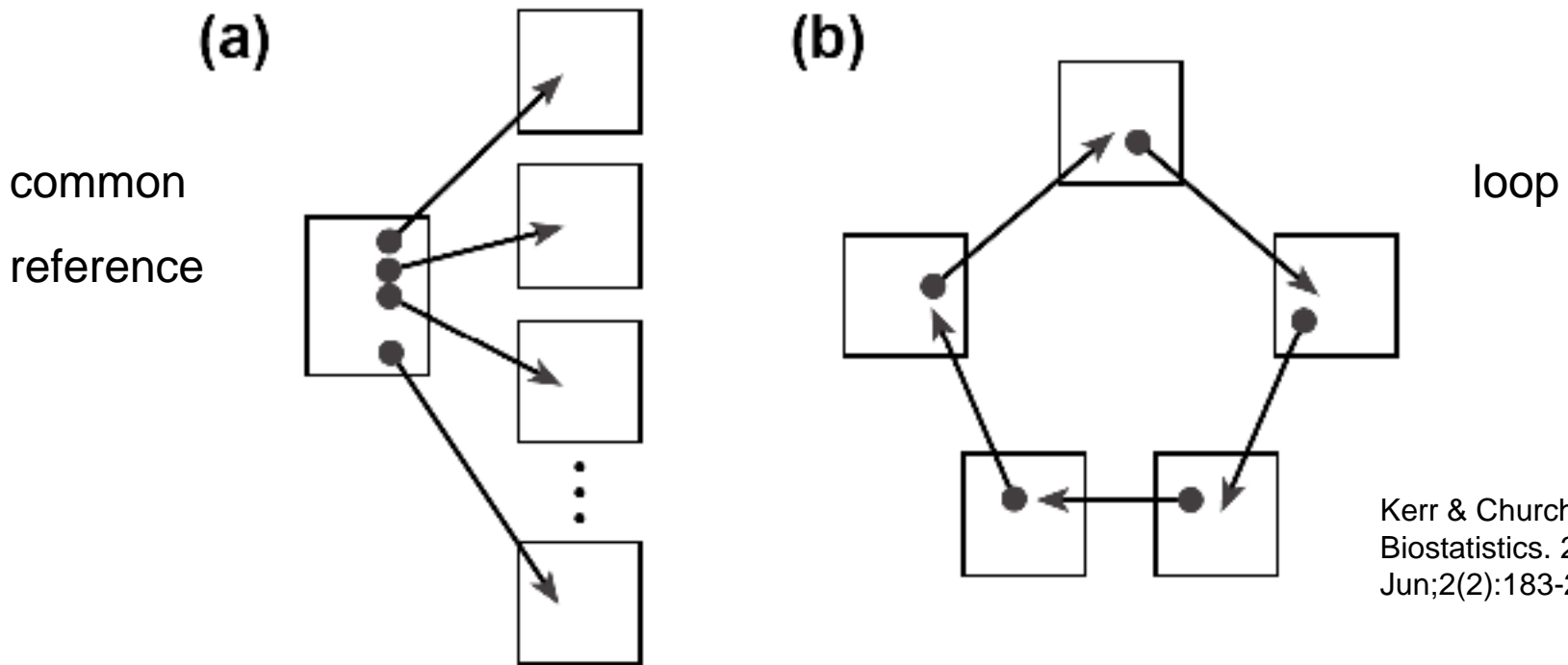


Support Vector Machines (SVM) for **supervised** classification



Find separating hyperplane that maximizes the margin between the two classes use this to classify new samples (e.g. in a microarray-based diagnostic test)

Excursus: Experimental design

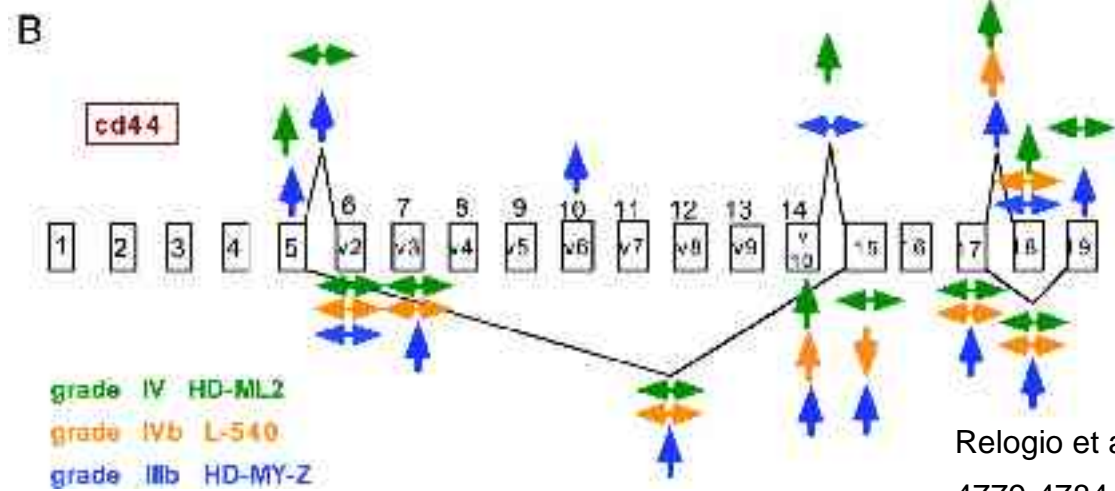
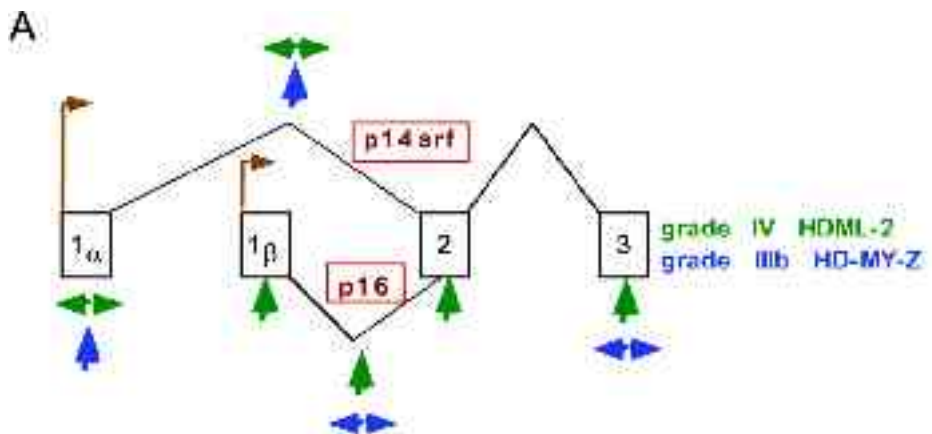
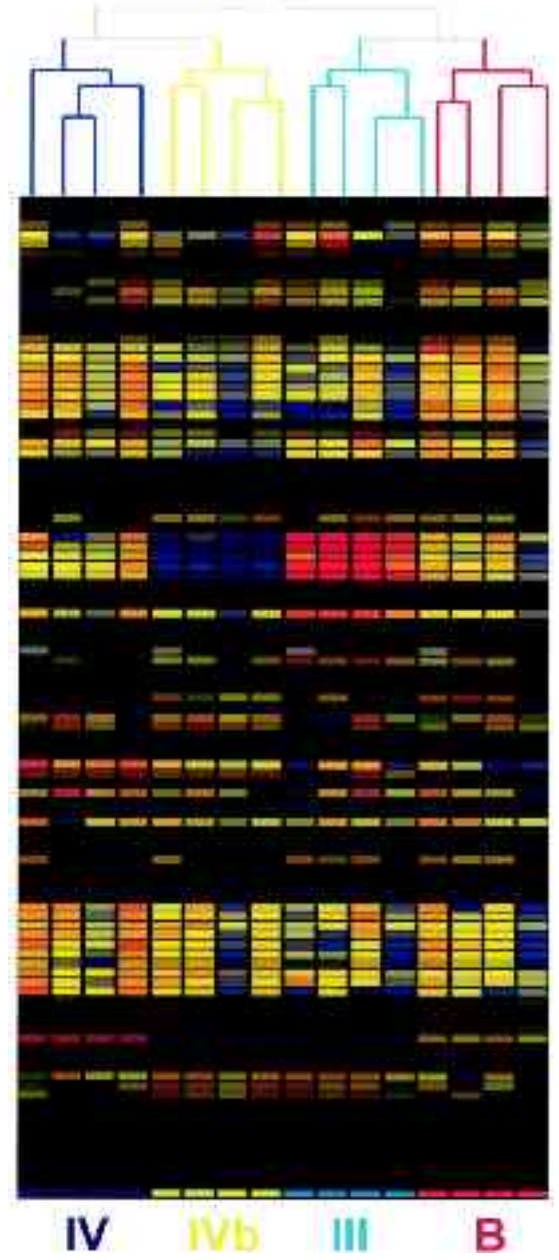
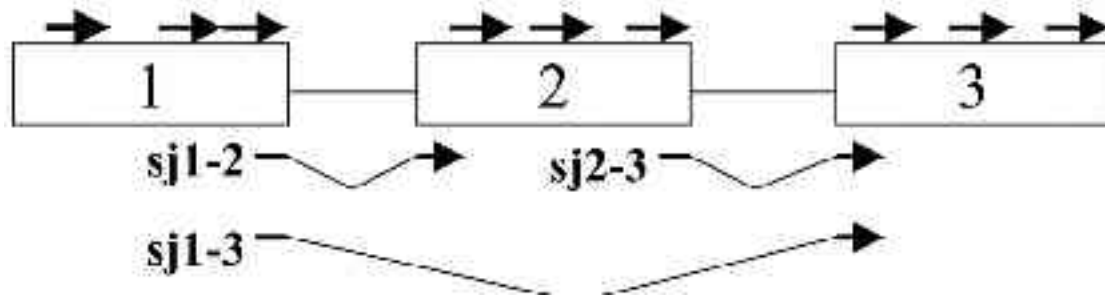


$$y_{ijk g} = \mu + A_i + D_j + V_k + G_g + (VG)_{kg} + (AG)_{ig} + (DG)_{jg} + \epsilon_{ijk g}.$$

A-Optimality = minimize $\frac{1}{\binom{v}{2}} \sum_{k_1 \neq k_2} \text{var} \left((\widehat{VG})_{k_1 g} - (\widehat{VG})_{k_2 g} \right)$

Cutting-edge uses of microarray technology

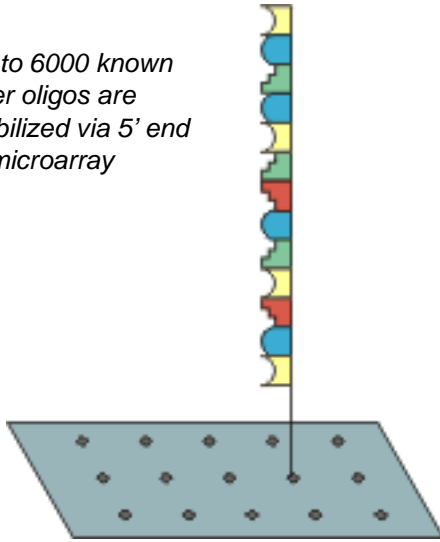
Alternative splicing on microarrays



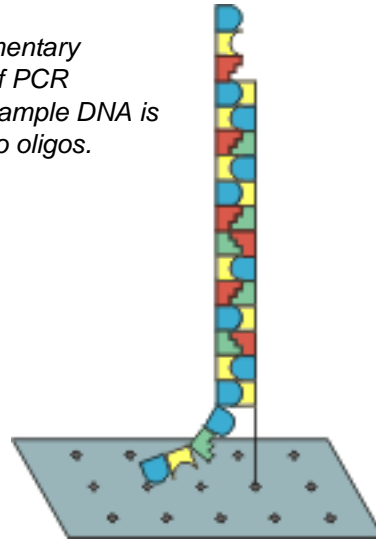
Religio et al., J. Biol. Chem., Vol. 280, Issue 6, 4779-4784, February 11, 2005



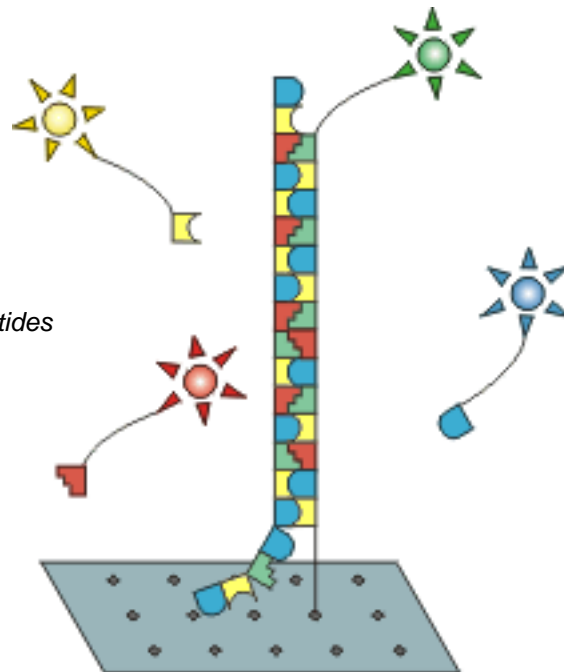
1. Up to 6000 known 25-mer oligos are immobilized via 5' end on a microarray



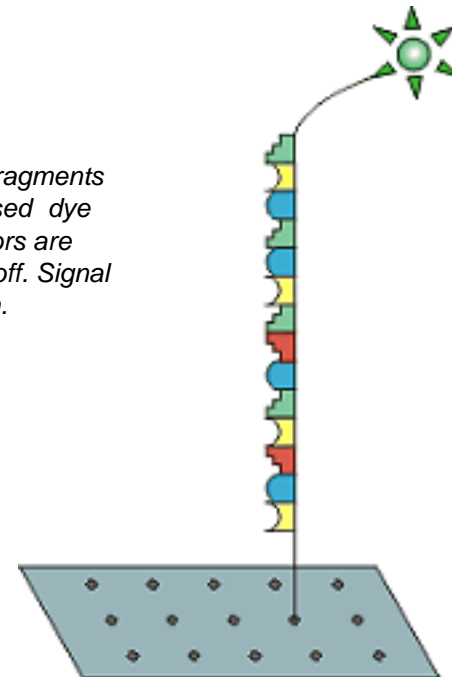
2. Complementary fragment of PCR amplified sample DNA is annealed to oligos.



3. Template dependent single nucleotide extension by DNA polymerase. Terminator nucleotides are labelled with 4 different fluorescent dyes.



4. DNA fragments and unused dye terminators are washed off. Signal detection.

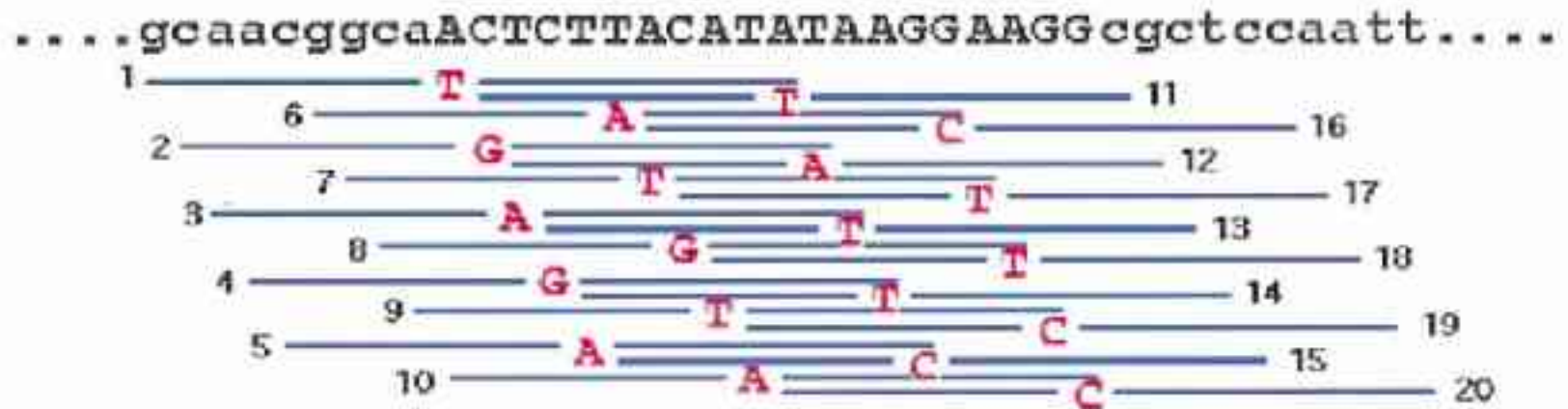
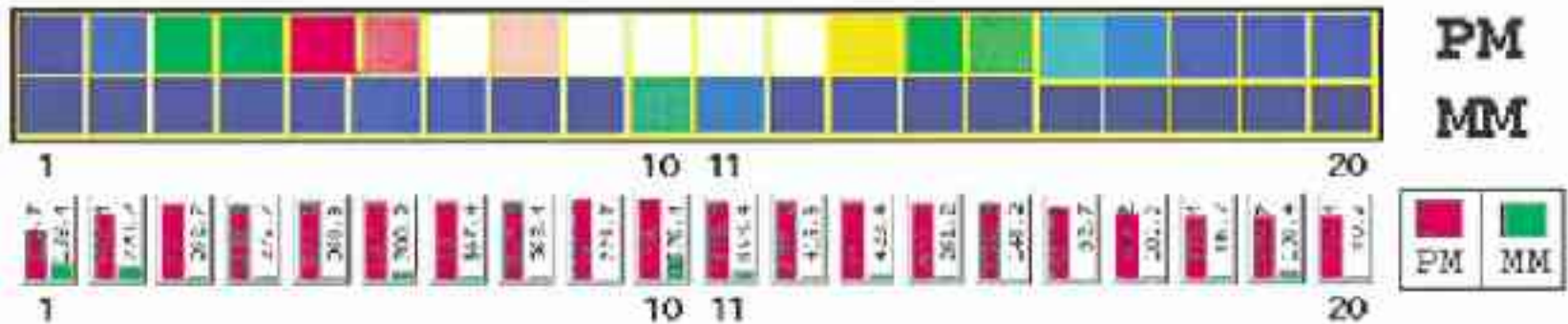


Customised detection of genetic polymorphisms in human patients

individual genotype
personalised medicine

example: **ARRAYED PRIMER EXTENSION (APEX)**

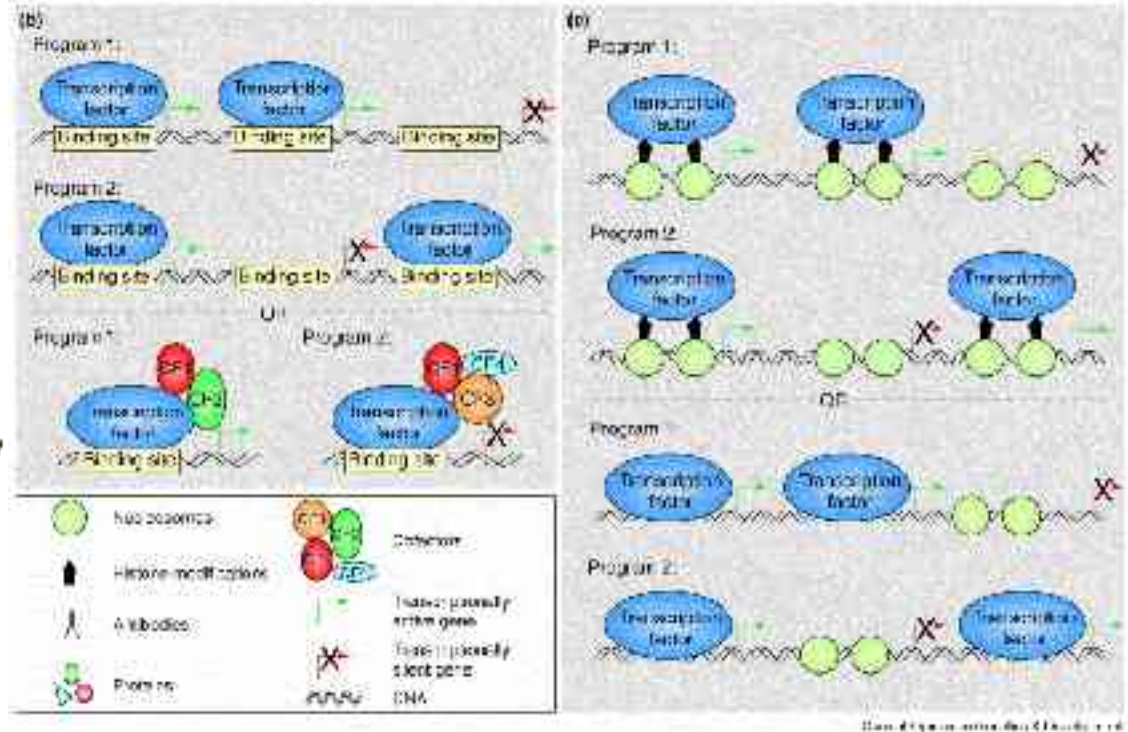
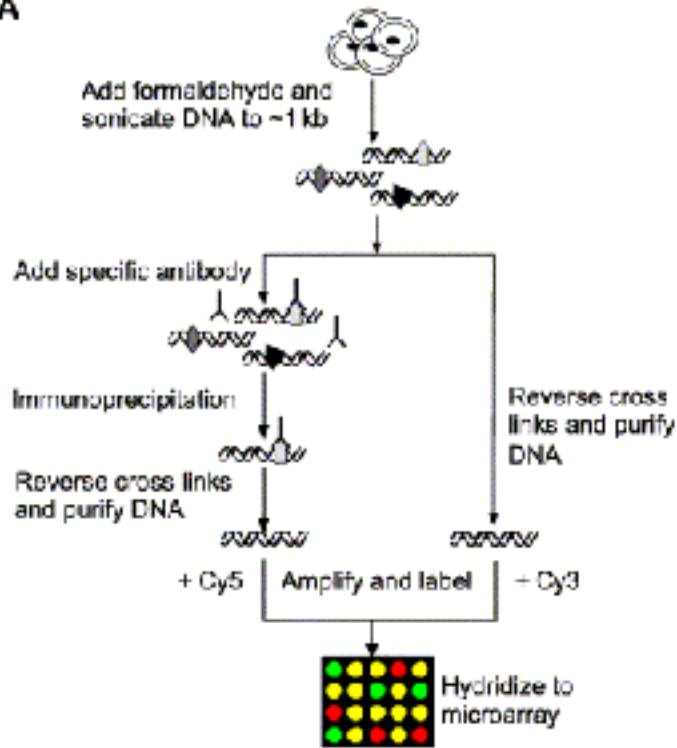
Identification of pathogens in environmental (patient) samples – Sequencing by hybridization



between 3 and 10 probe sets per species, each containing a few hundred probes
 sensitivity about 500fg pathogen genomic DNA per sample

Global identification of transcription factor target sites using chromatin immunoprecipitation plus whole-genome tiling microarrays (ChIP-chip)

A



preferably the array should provide continuous genome coverage, not just ORFs

Inference of gene regulatory networks from gene expression data (indirect method, in contrast to the direct ChIP-chip approach)

A: Gene Expression Matrix E

Gene	G1	G2	G3	G4
Wild type	1	1	1	1
G4 disruptant	1	1	2.5	-
G3 disruptant	1	1	-	0.2
G2 disruptant	1	-	0.4	3.0
G1 disruptant	-	0.1	0.1	0.3

B: Binary relation R

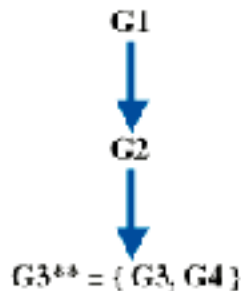
G4 affects G3
 G3 affects G4
 G2 affects G3
 G2 affects G4
 G1 affects G4
 G1 affects G3
 G1 affects G2

C: Adjacency matrix A

	G1	G2	G3	G4
G1	0	1	1	1
G2	0	0	1	1
G3	0	0	0	1
G4	0	0	1	0

remove ambiguous relationships

F: Multi-level digraph



E: Skeleton matrix S
(remove indirect connections)

	G1	G2	G3**
G1	0	1	0
G2	0	0	1
G3**	0	0	0

D: Partition genes into equivalence sets

	G1	G2	G3**
G1	0	1	1
G2	0	0	1
G3**	0	0	0

G3** = {G3, G4}

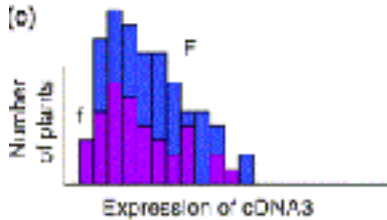
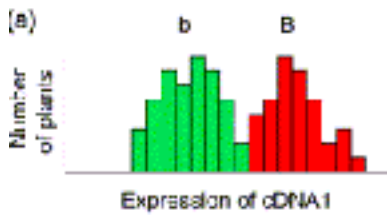
Directed graph of regulatory influences – gene network



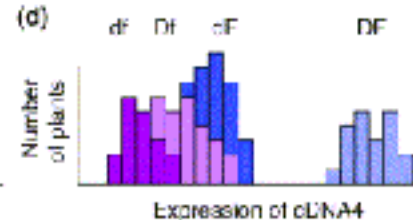
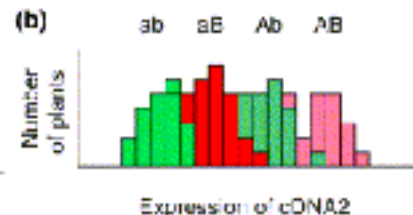
Genetical genomics

gene expression as a Quantitative Trait

qualitative expression

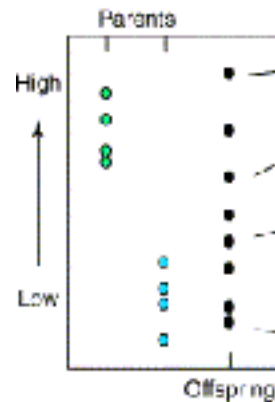


quantitative expression

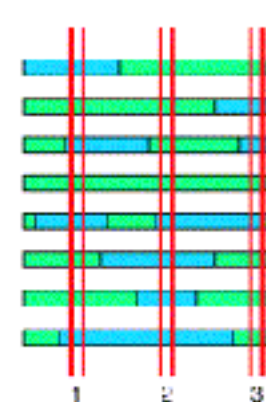


epistatic interaction

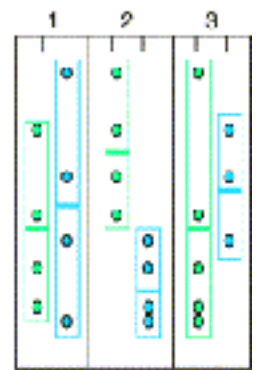
(a) Observed expression



(b) Genetic fingerprint



(c) Differential expression analysis



TRENDS in Genetics

the combination of genotype and expression information can identify cis- and trans-regulatory sites

Jansen & Nap, Trends Genet. 2001 Jul;17(7):388-91 and

Jansen & Nap, Trends Genet. 2004 May;20(5):223-5.



Further reading

- **Kerr MK, Churchill GA.** Genet Res. 2001; 77: **Statistical design and the analysis of gene expression microarray data.**
- **Eisen MB, Spellman PT, Brown PO, Botstein D.** Proc Natl Acad Sci U S A. 1998; 95: **Cluster analysis and display of genome-wide expression patterns.**
- **Hughes TR, Marton MJ, Jones AR, Roberts CJ, et al.** Cell. 2000; 102: **Functional discovery via a compendium of expression profiles.**
- **Wit E, McClure J.** 2005: **Statistics for Microarrays – Design, Analysis and Inference**



Conclusions

- microarrays measure gene expression globally
new post-genomic biology
- two principal technologies: one-color (Affymetrix)
and two-color (cDNA arrays)
- multiple measurements pose particular statistical
challenges
- interpretation requires combination with previous
knowledge
- creative application of microarrays opens new
avenues for biological insight

