Microarray Data Analysis and Gene Network Reconstruction

Petr Nazarov

petr.nazarov@crp-sante.lu

23-02-2010

OUTLINE

Part I. Microarrays

- Brief introduction to the area
- One and two-color arrays in transcriptomics
- Data processing pipeline

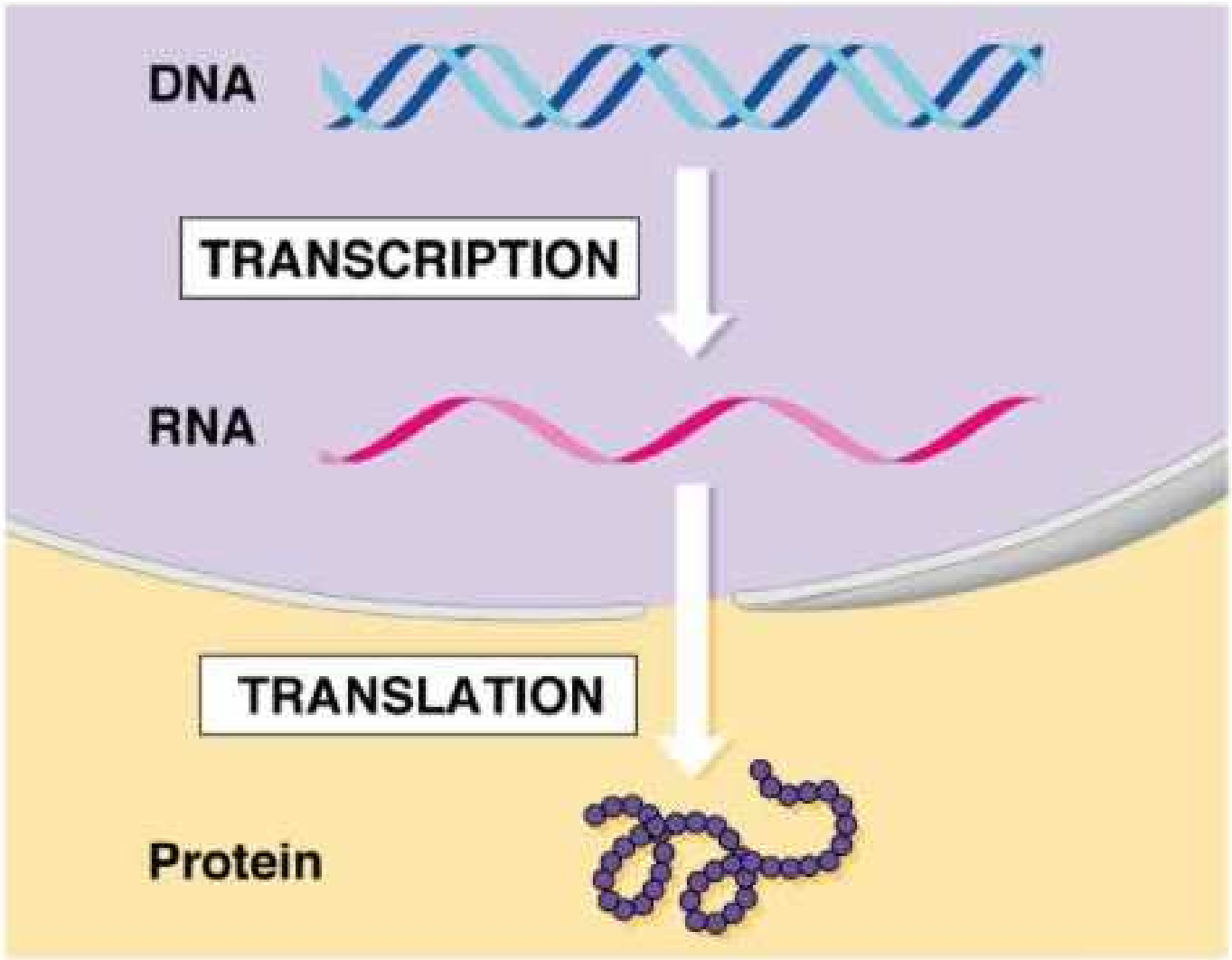
Part II. Gene networks and co-expression

- Gene networks
- Reconstruction on the basis of coexpression
- CoExpress
- Plans, ideas, open questions

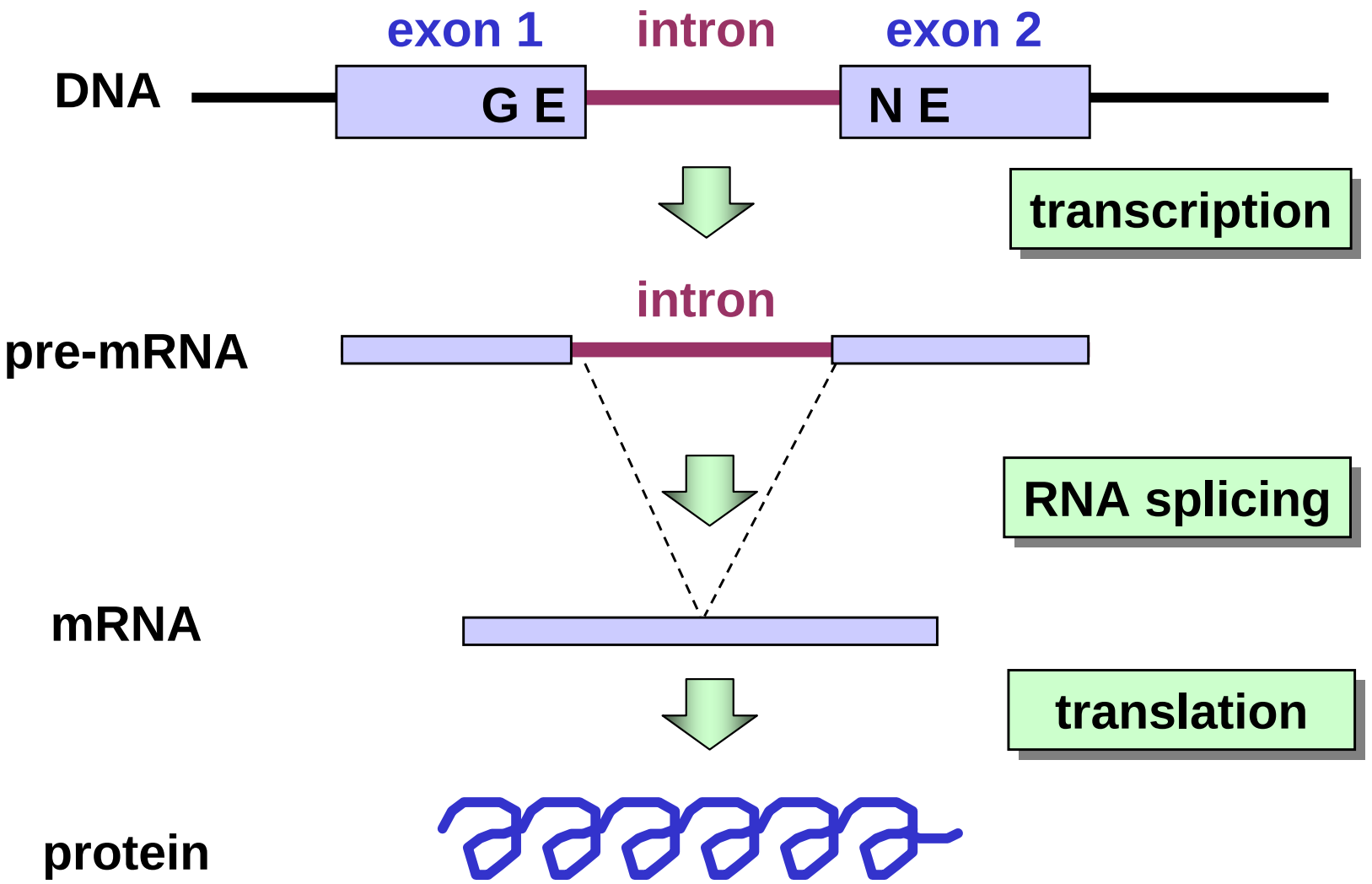
Part I

Microarrays

Basic Expression Scheme

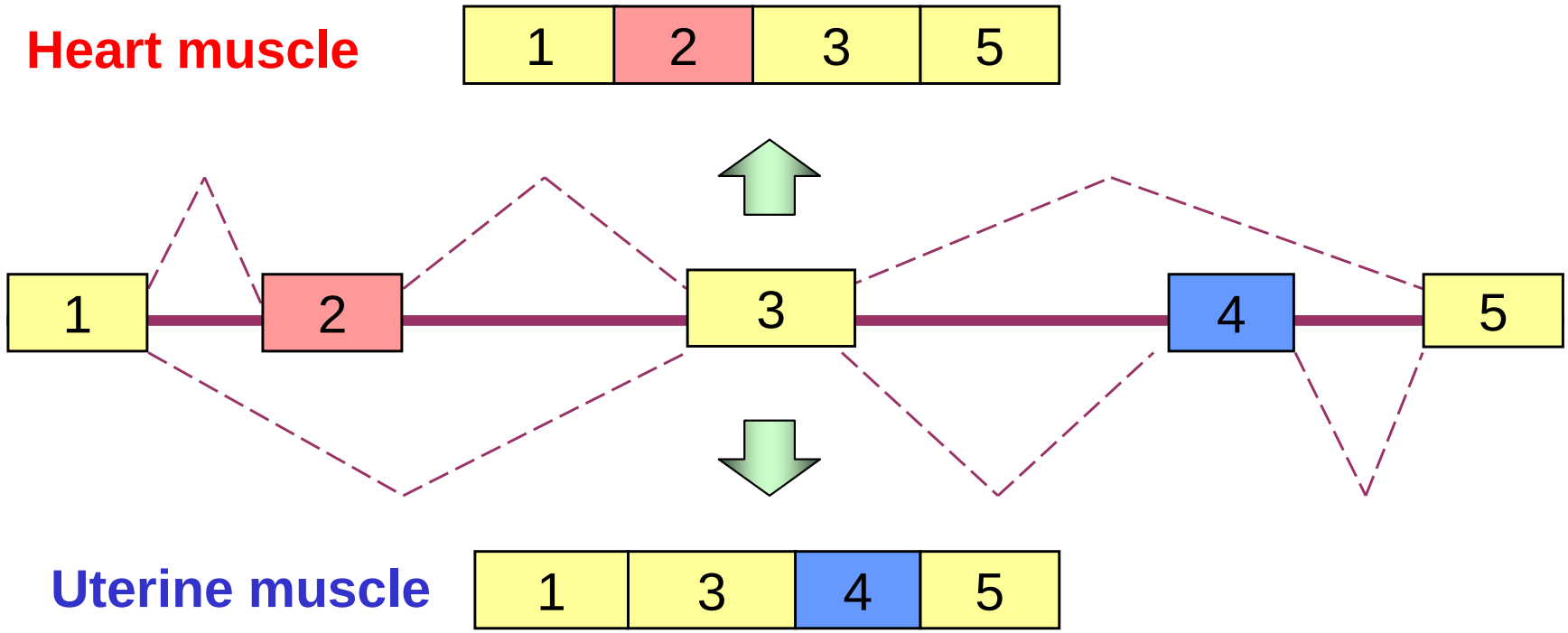


Basic Expression Scheme



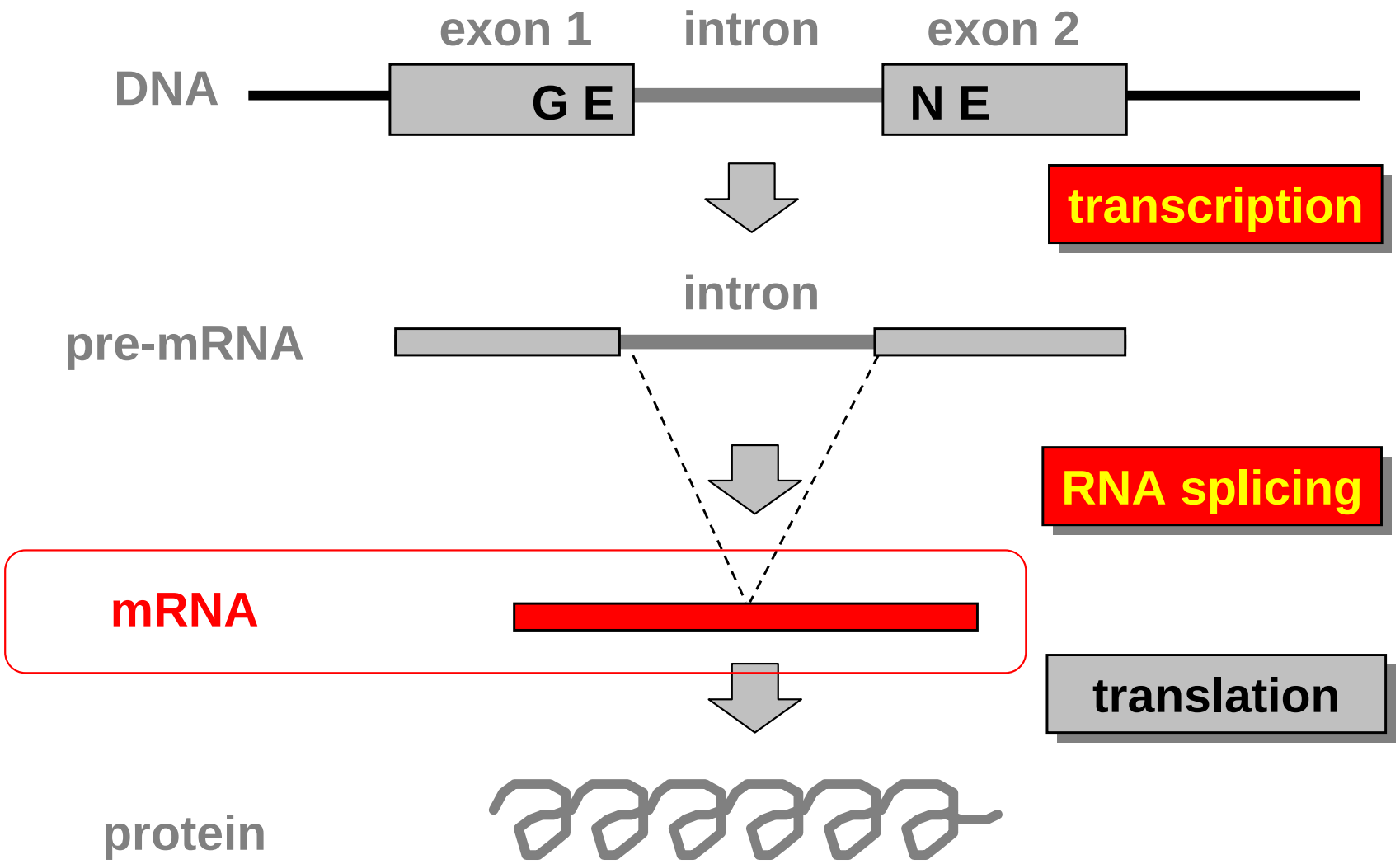
Alternative Splicing

Multiple introns may be spliced differently in different circumstances, for example in different tissues.

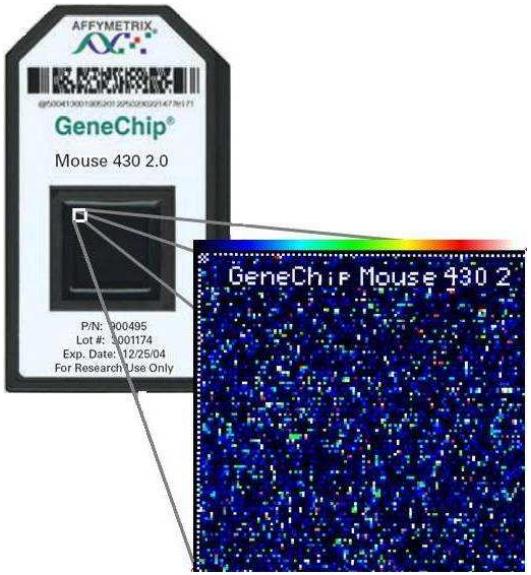


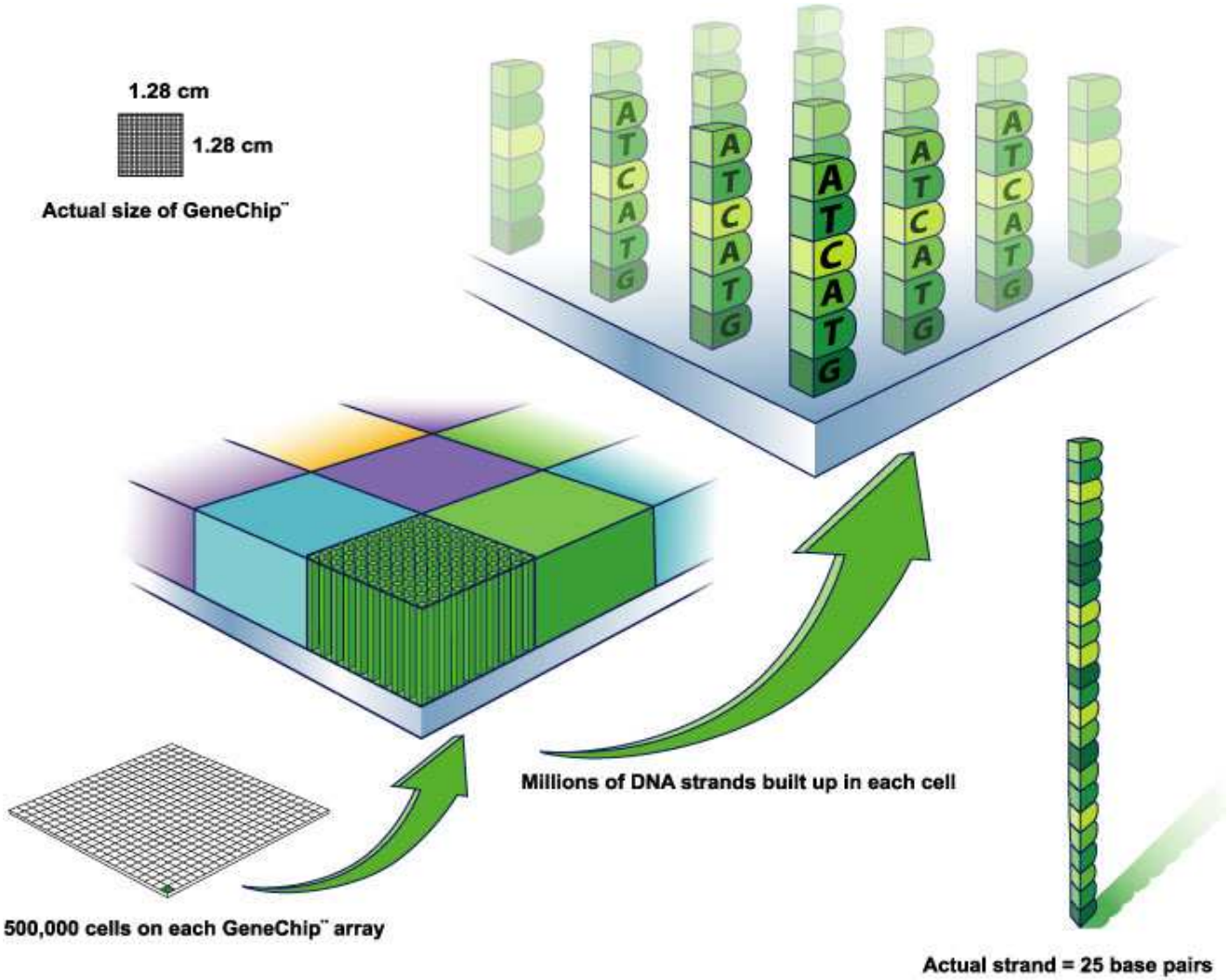
Thus one gene can encode more than one protein. The proteins are similar but not identical and may have distinct properties – an important feature for complex organisms

Basic Expression Scheme

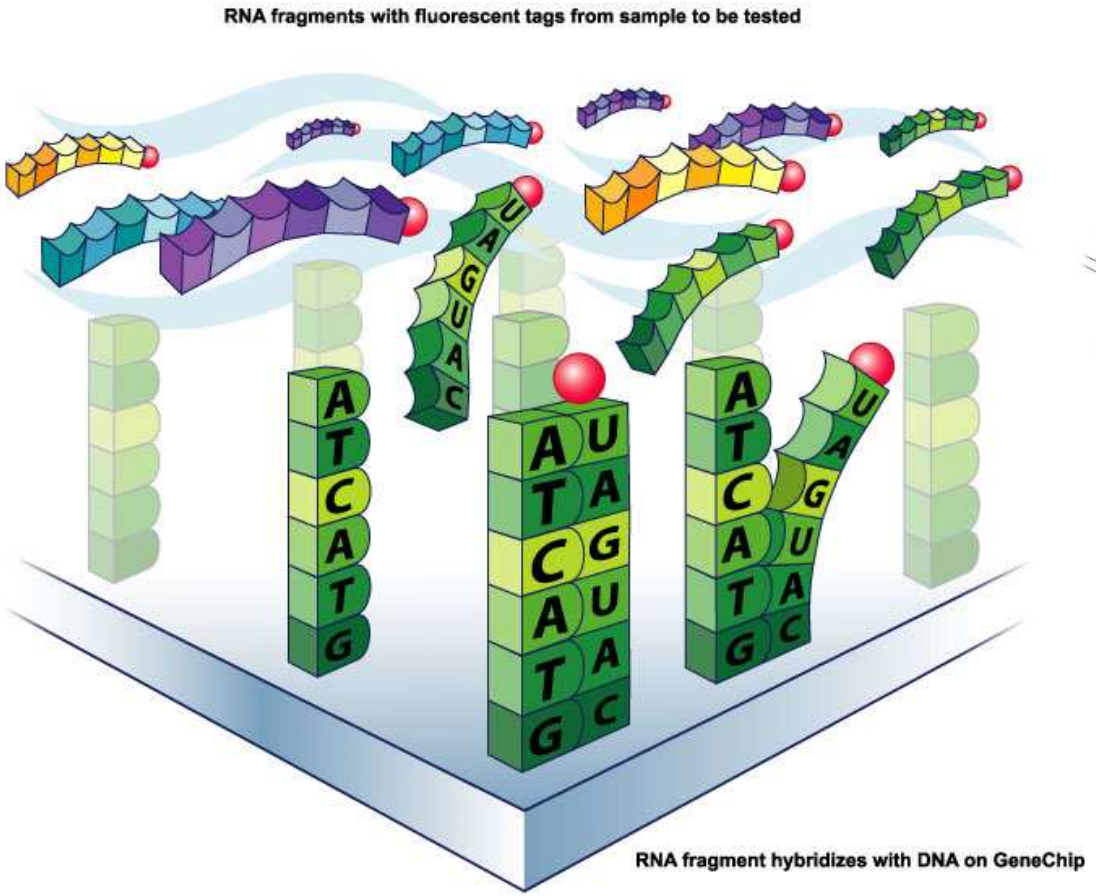


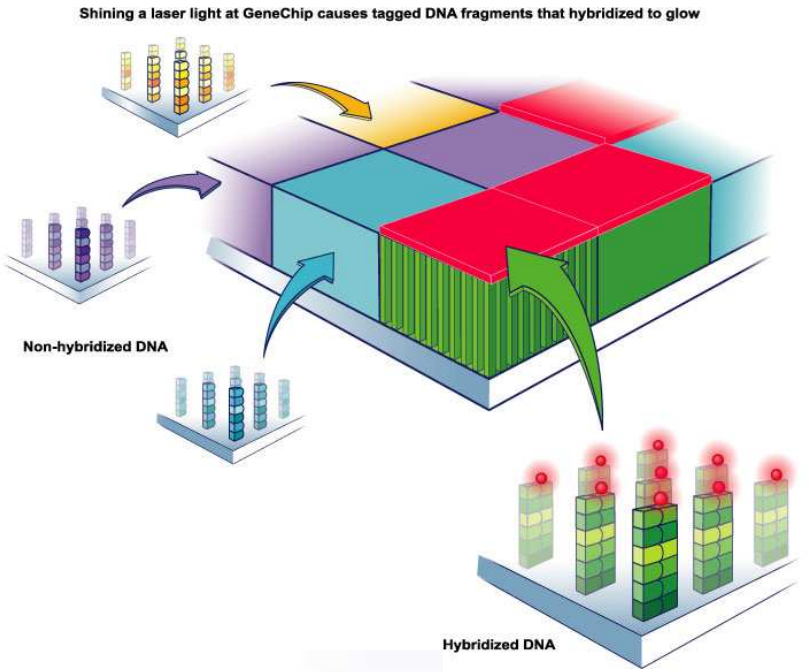
Affymetrix Microarray Design

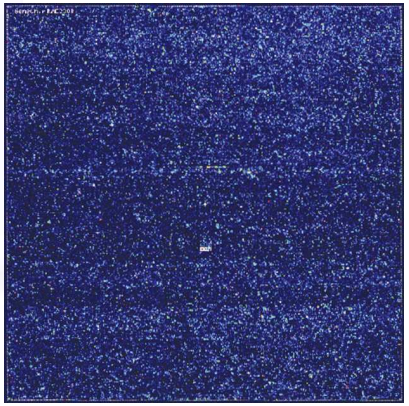




Affymetrix Microarray Design



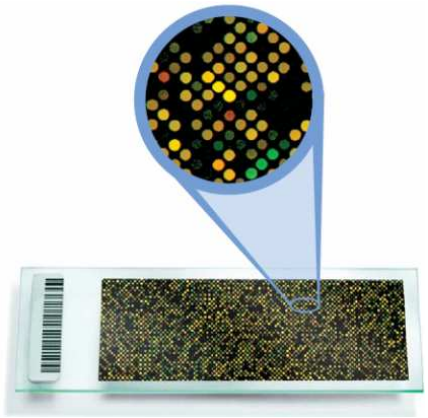




TWO BASIC TYPES OF MICROARRAYS

Two-color Arrays (2C)

- Agilent full genome
- Thematic arrays



Pro

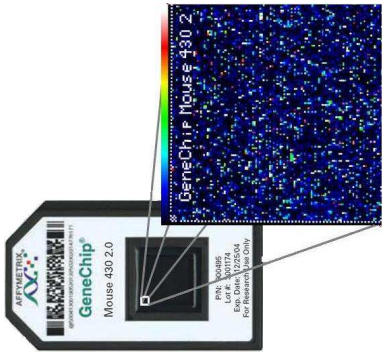
- Direct comparison
- Low price

Con

- Dye effects
- Non-flexibility in analysis

One-color Arrays (1C)

- Affymetrix GeneChip
- Affymetrix Exon
- Affymetrix mRNA



Pro

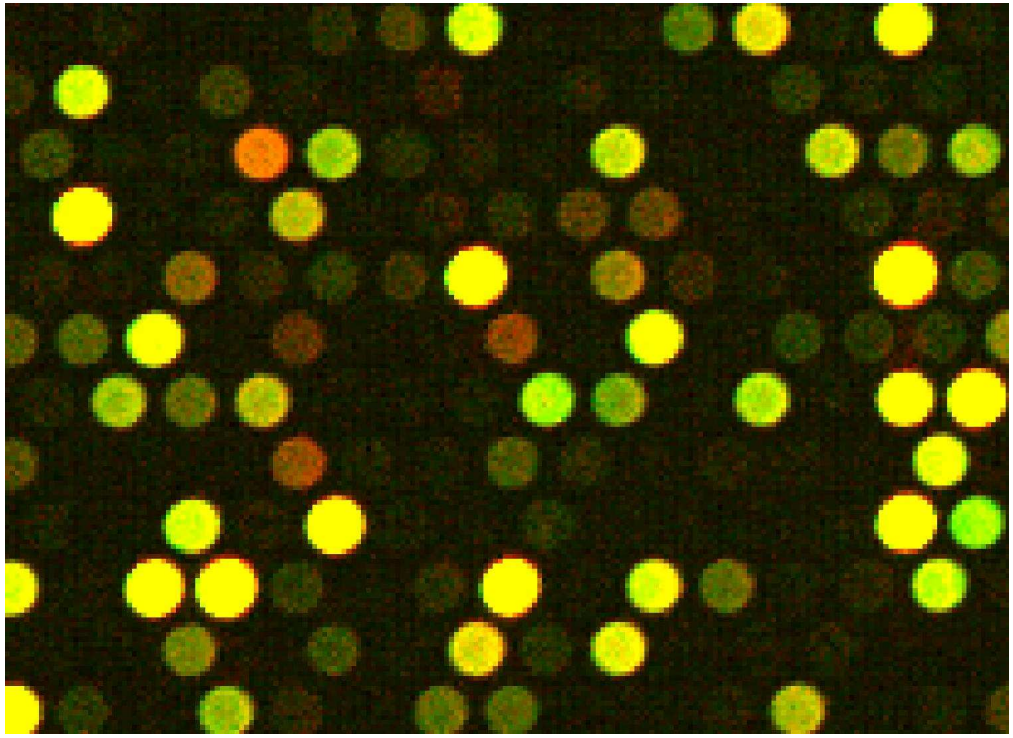
- Flexible analysis
- High level of standardization

Con

- Price

2C: ADVANCED IMAGE PROCESSING

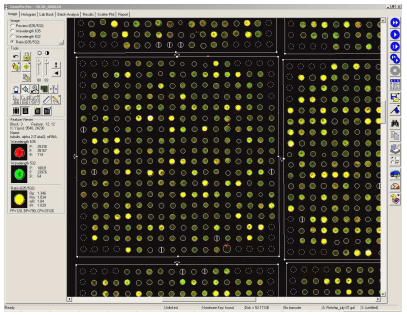
Spot Quality Control



Solution

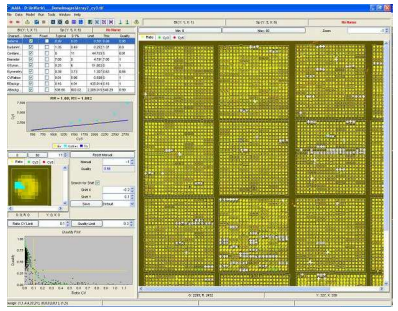
- Quantify each spot by a set of parameters
- Find an optimal rule to accept the spots
- Remove bad spots from further analysis

GenePix Pro

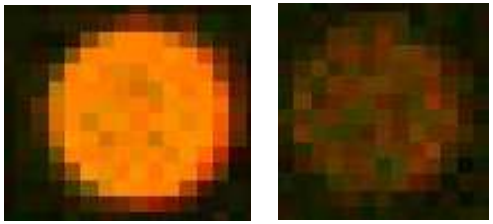


✓ Commercial and well established software

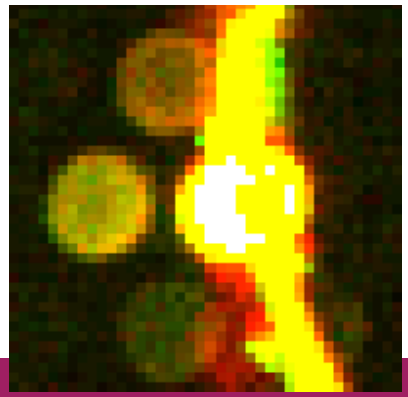
MAIA



✓ Freeware, academic and high quality spot evaluation

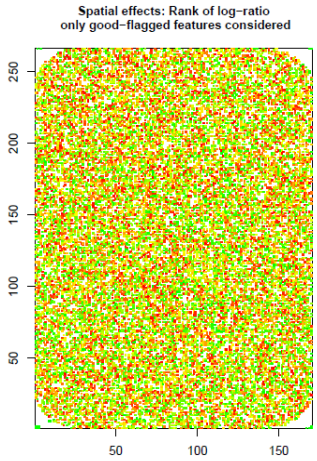


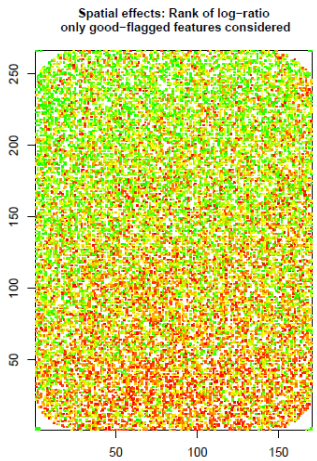
LogFC ≈ 2



Yatskou M., et al,
BMC Res Notes. 2008; 1:80

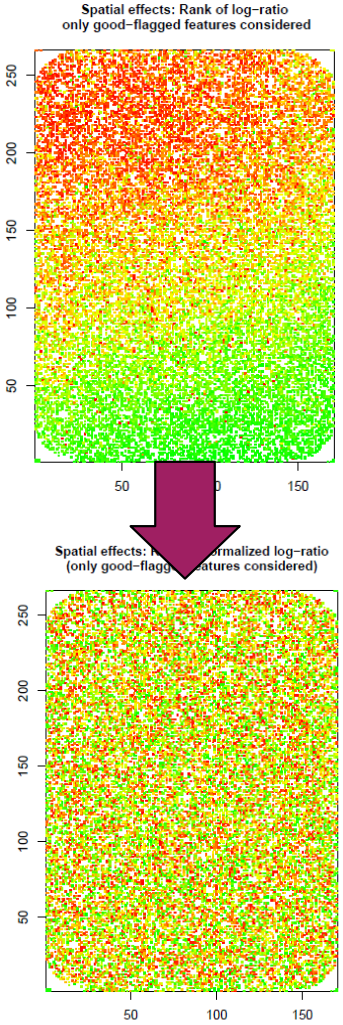
- Ideal picture (rare)
- Standard picture

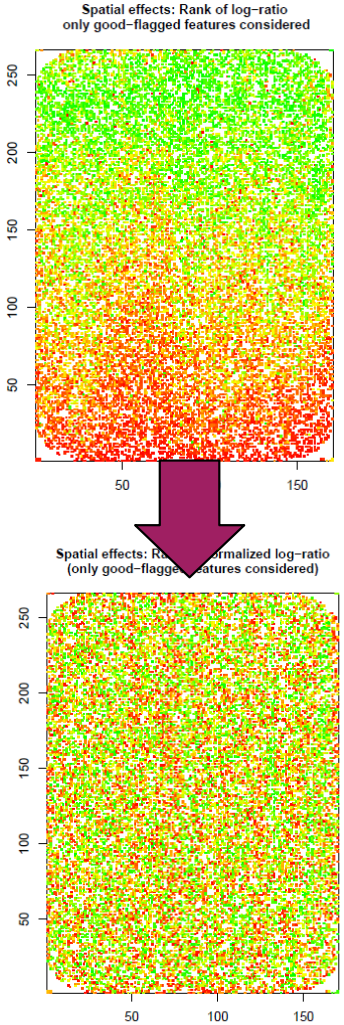




Here you see the “false” rank-based colors

- Bad luck…
- Or even worse…



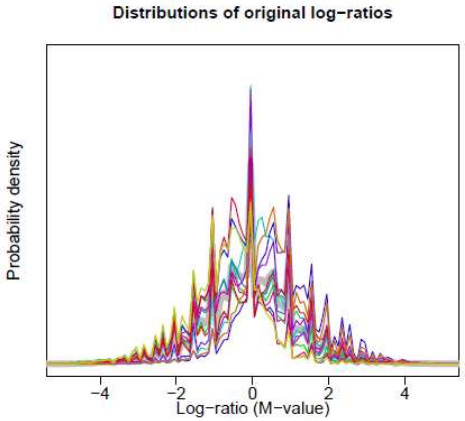


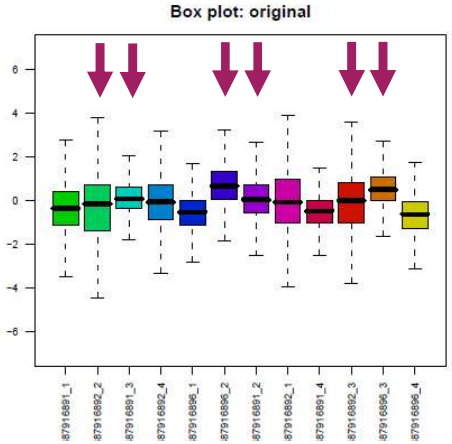
Solution. Spatial normalization

- Using spikes (Agilent)
- Using numerical methods

Linear dye-effect

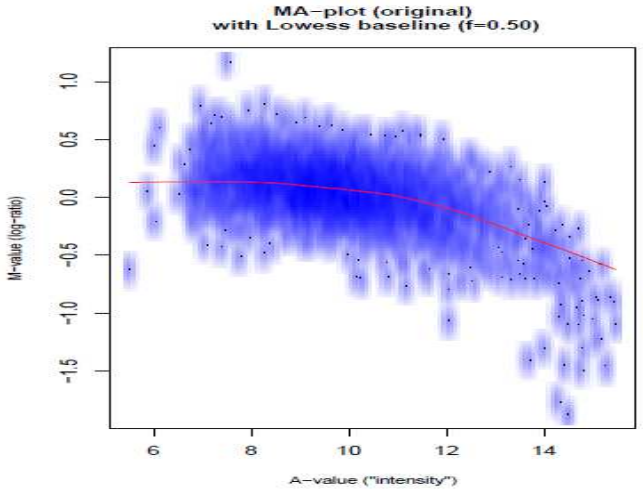
due to difference in extinction coefficient and quantum yield of the flourophore



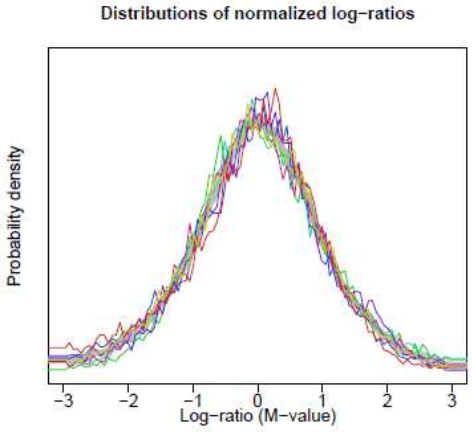


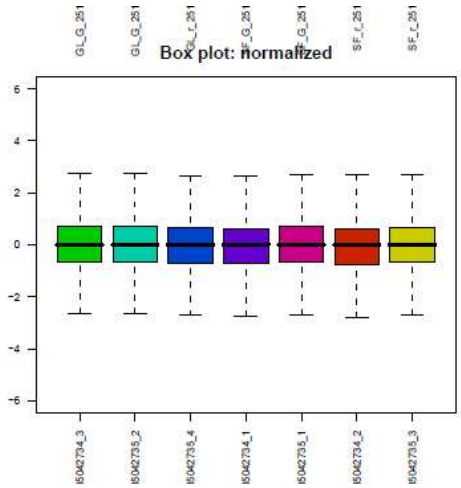
Non-linear dye-effect

photodegradation, radiationless energy transfer, quenching

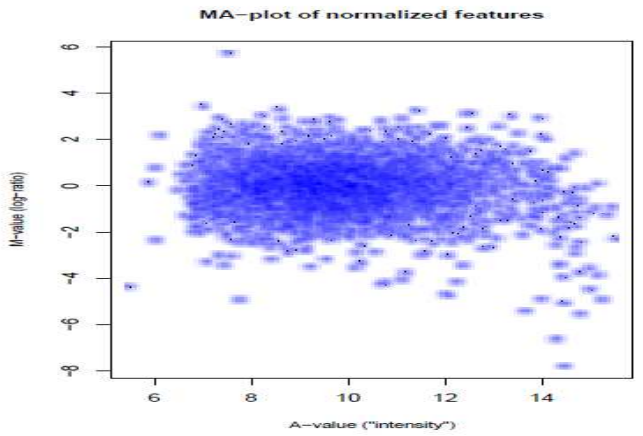


Solution. Linear normalization



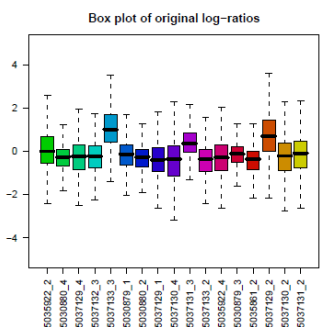


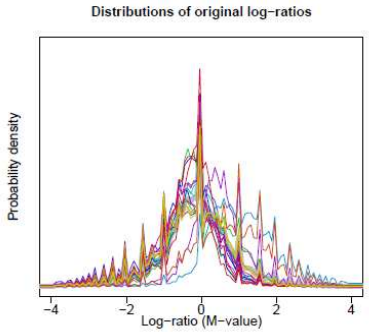
Solution. Lowess (loess) correction

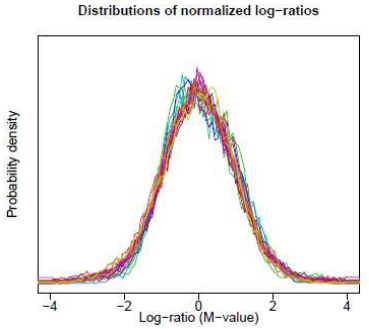


2C: QUALITY CONTROL / QUALITY ASSESSMENT

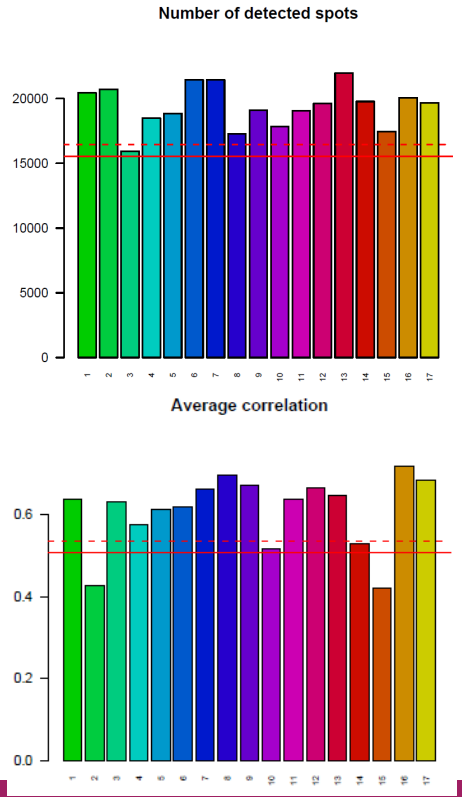
How to Remove Bad Microarrays?

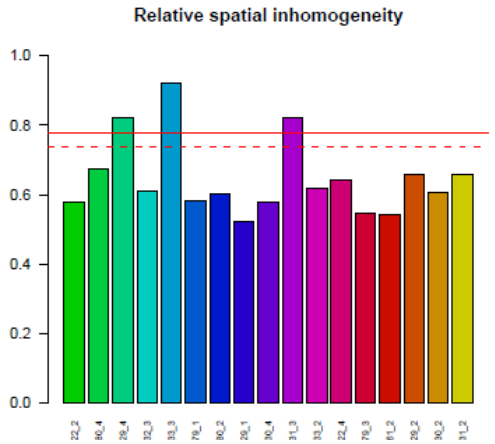


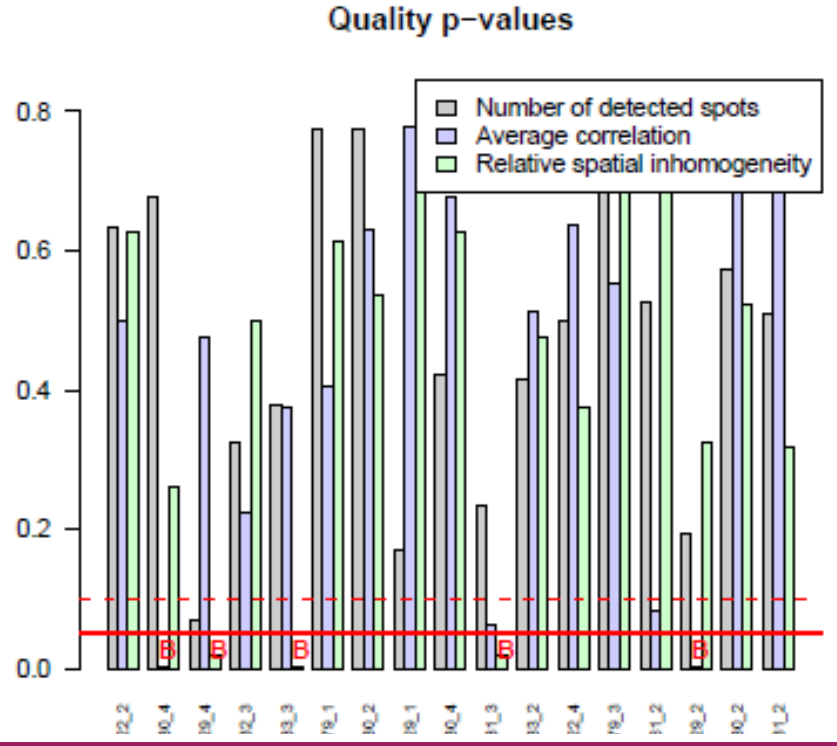




Solution. Introduce several measures and develop rules for outliers







2C: STATISTICAL ANALYSIS

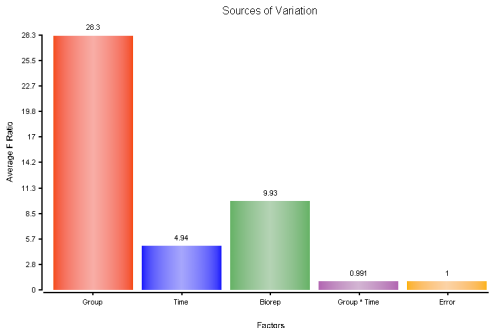
Detection of the Significant Genes

SAM

- Non-parametric

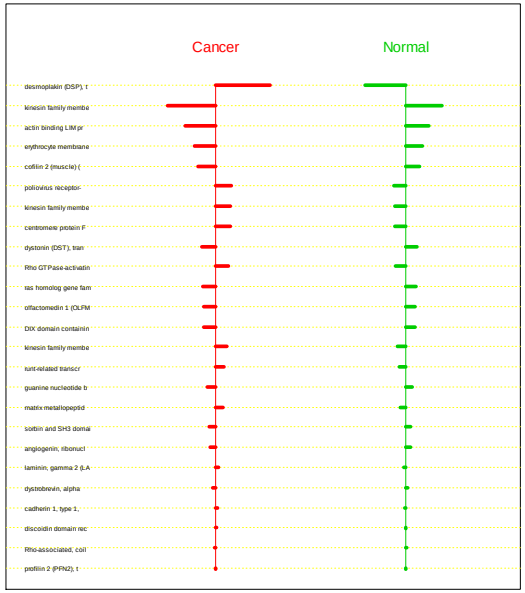
ANOVA

- Multifactor
- Complex comparisons



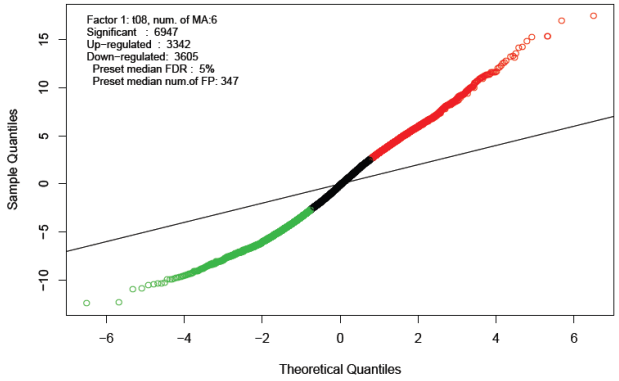
PAM

- Detect biomarkers



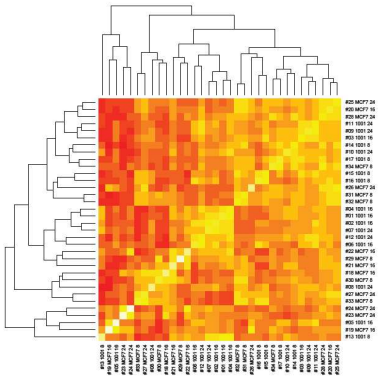
EBS (limma)

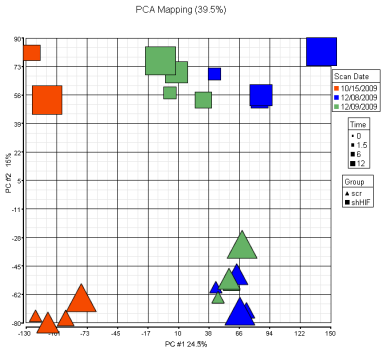
- Parametric
- Complex comparison



PCA and Clustering

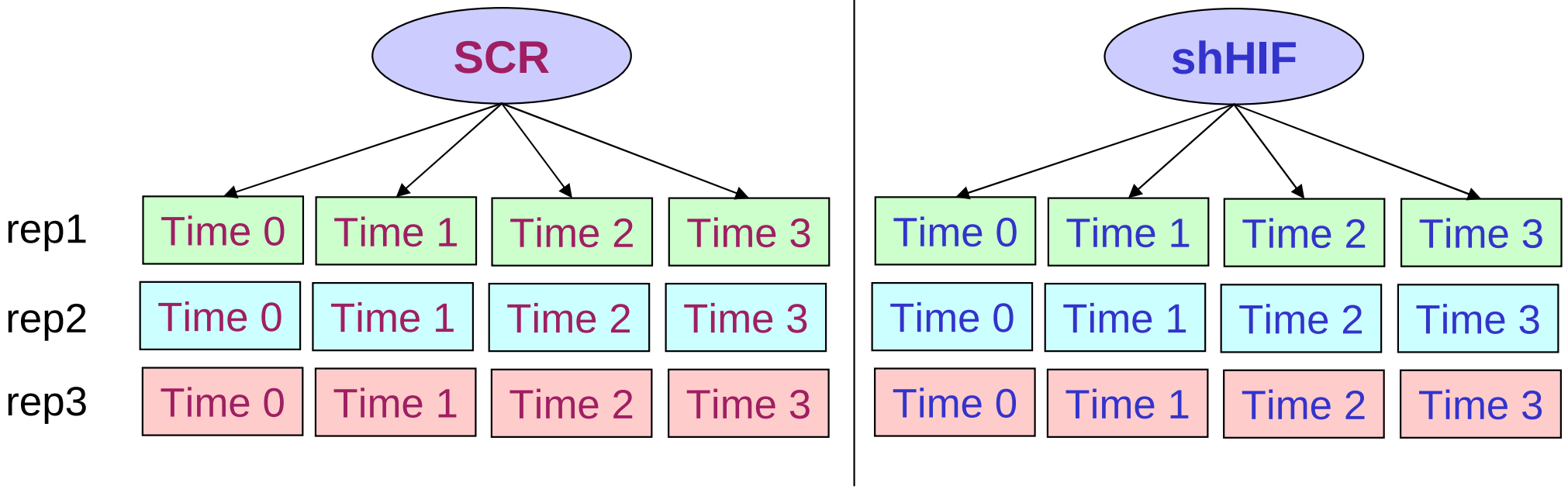
- Detect similarities





EXAMPLE OF DATA ANALYSIS

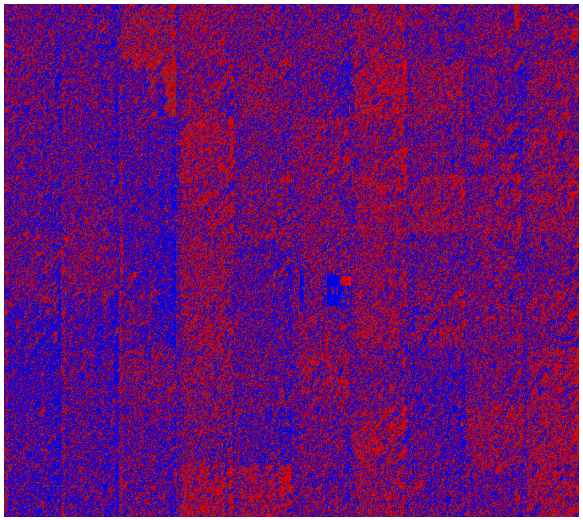
Data Coming from dr. Iris Behrmann



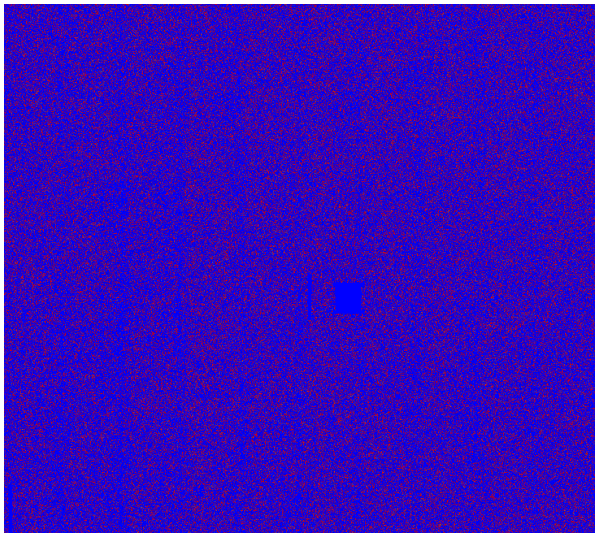
12 MAs used - platform: Affymetrix Human Gene 1.0 ST Array

This data set is a “ready” academic example

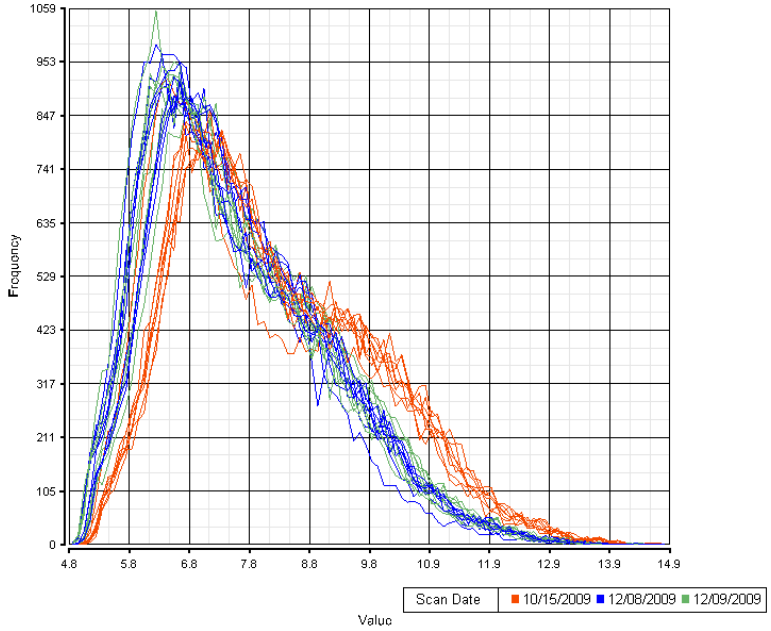
SUMMARIZATION AND NORMALIZATION

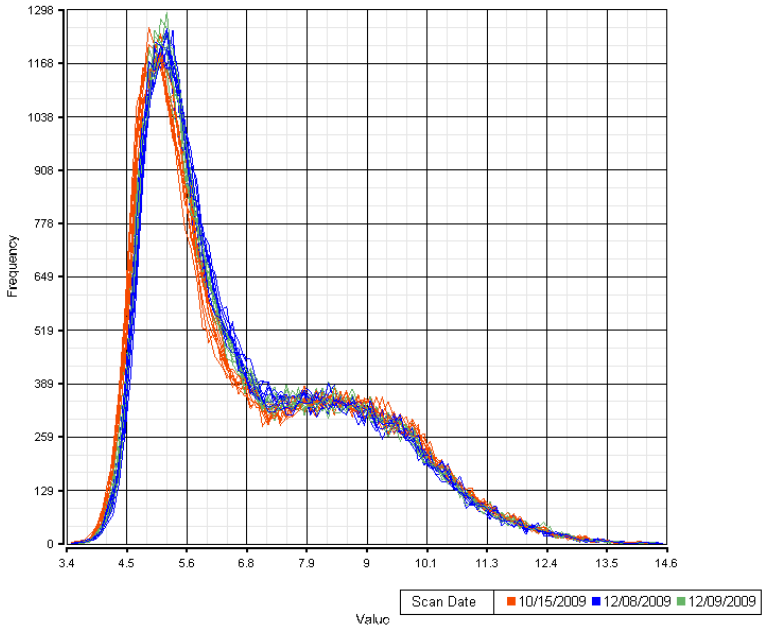


All Rows of 1



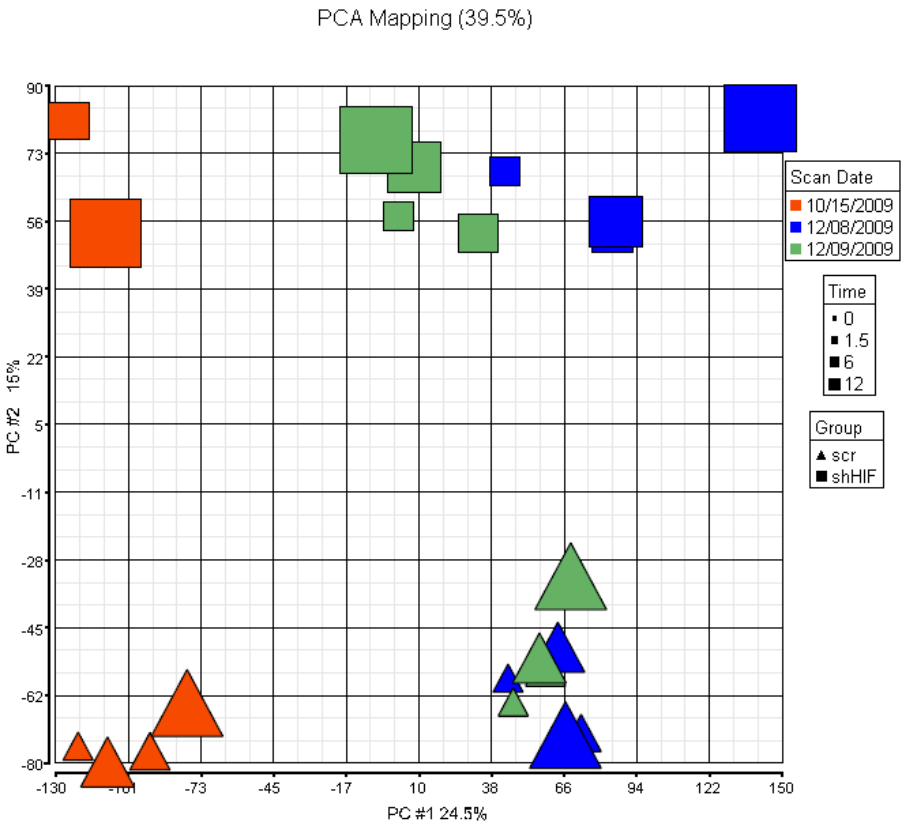
All Rows of 1

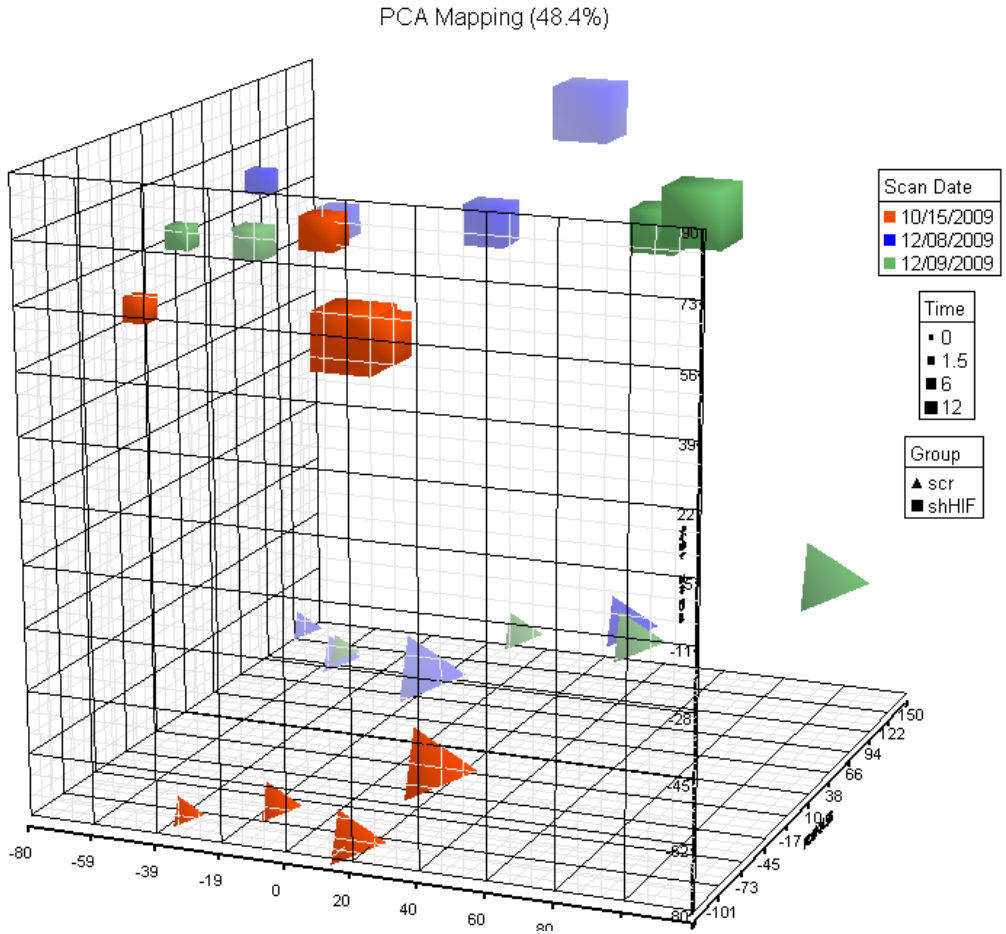




QC AND DETECTION OF THE EFFECTS

PCA for Detection of Important Factors

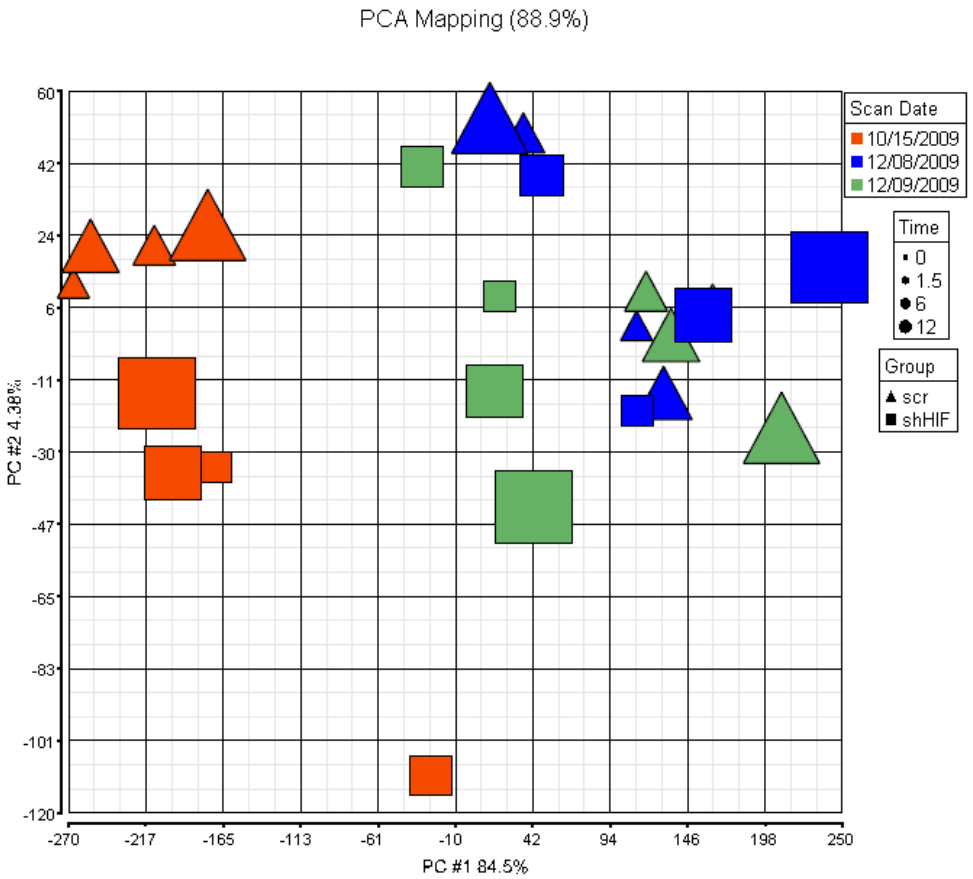




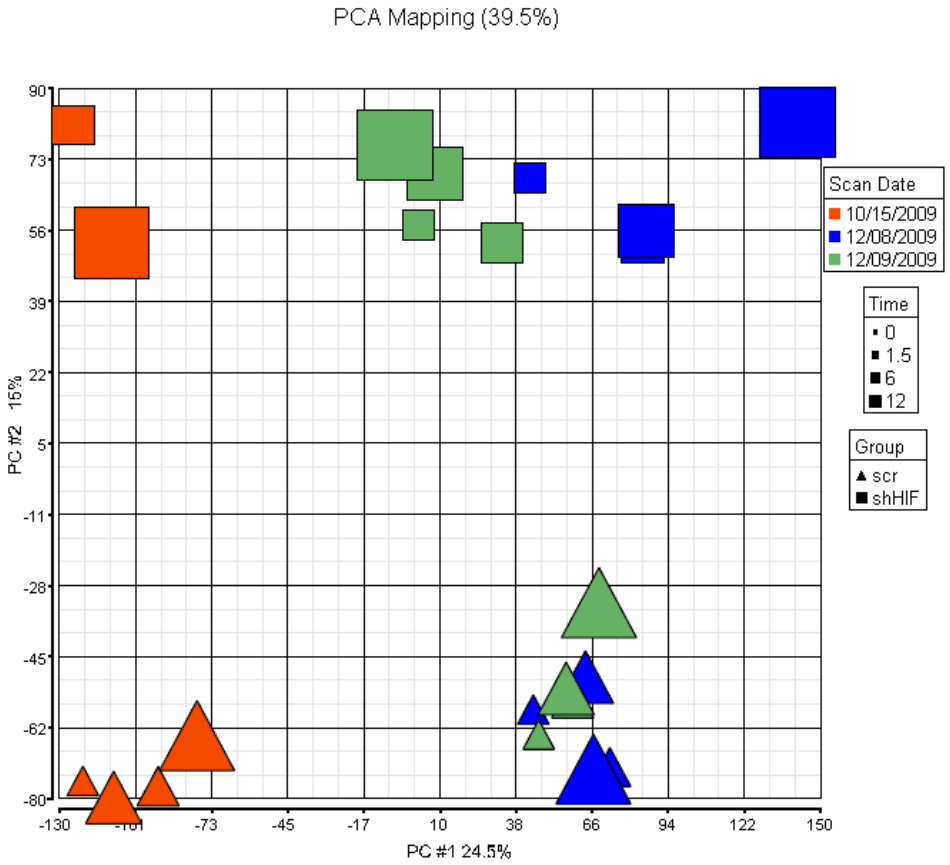
QC AND DETECTION OF THE EFFECTS

Efficiency of Normalization

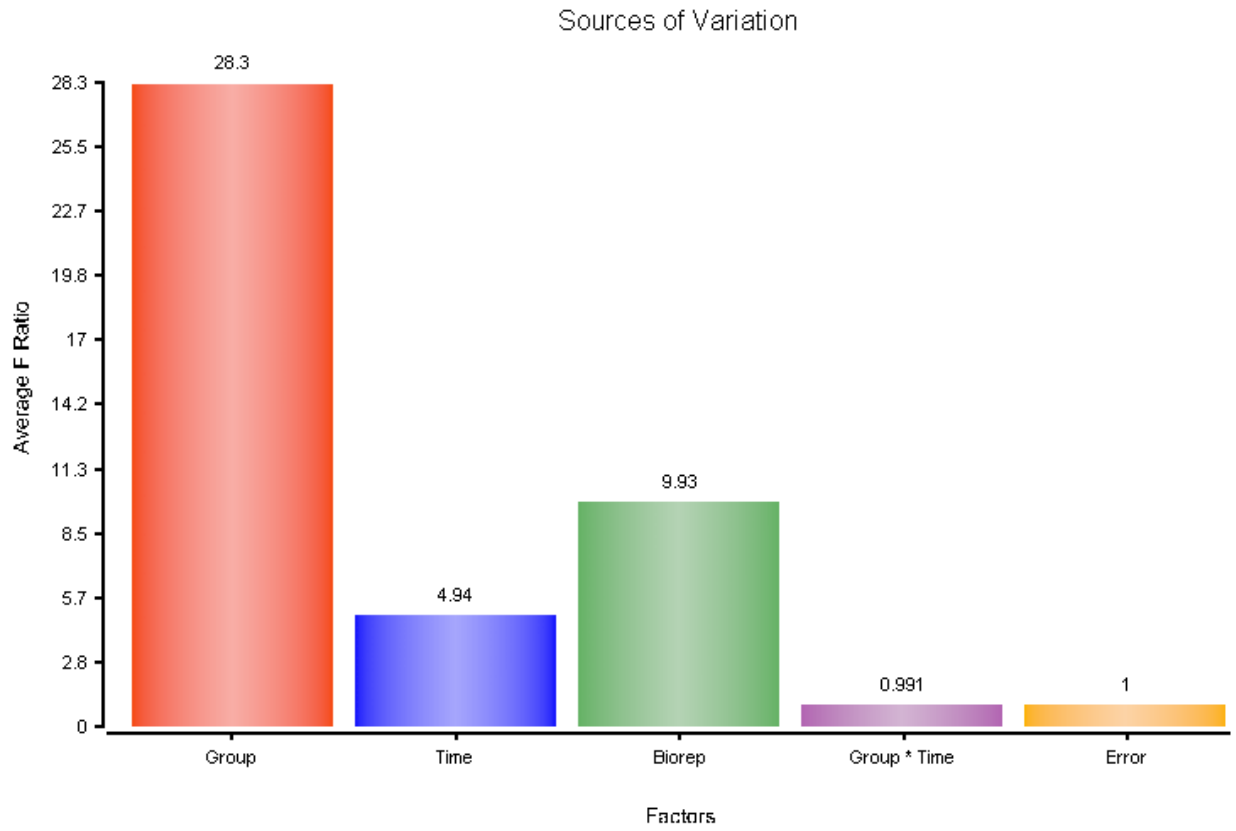
Original



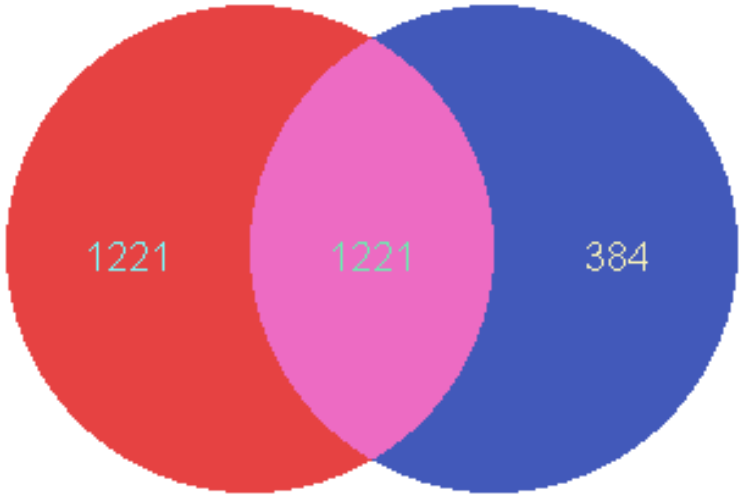
Normalized



STATISTICAL ANALYSIS
Factors and Their Effects



STATISTICAL ANALYSIS
FDR-based List Generation

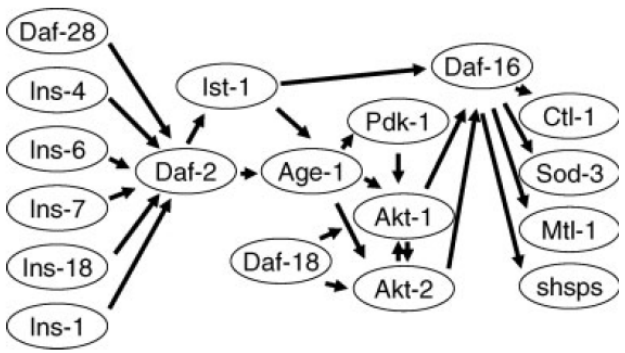


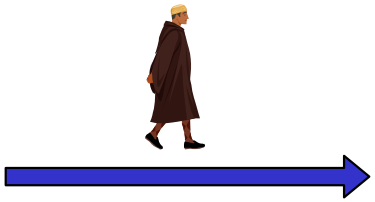
Probeset ID	gene_assignment	Gene Symbol	RefSeq	p-value 12/0 in SCR
8097910	NM_005141 // FGB // fibrinogen beta chain // 4q28 // 2244 /// ENS	FGB	NM_005141	9.66E-14
8126839	NM_014452 // TNFRSF21 // tumor necrosis factor receptor superf	TNFRSF21	NM_014452	1.14E-13
8066214	NM_004613 // TGM2 // transglutaminase 2 (C polypeptide, protein	TGM2	NM_004613	3.04E-12
8006005	NM_016518 // PIPOX // pipecolic acid oxidase // 17q11.2 // 51268	PIPOX	NM_016518	4.53E-12
8103311	NM_000508 // FGA // fibrinogen alpha chain // 4q28 // 2243 /// NM	FGA	NM_000508	2.29E-11
7976496	NM_001085 // SERPINA3 // serpin peptidase inhibitor, clade A (alp	SERPINA3	NM_001085	3.60E-11
8032834	NM_052972 // LRG1 // leucine-rich alpha-2-glycoprotein 1 // 19p13	LRG1	NM_052972	5.75E-11
8123598	NM_030666 // SERPINB1 // serpin peptidase inhibitor, clade B (ov	SERPINB1	NM_030666	1.49E-10
7932985	NM_003873 // NRP1 // neuropilin 1 // 10p12 // 8829 /// NM_00102	NRP1	NM_003873	1.50E-10
7997188	NM_005143 // HP // haptoglobin // 16q22.1 // 3240 /// NM_001126	HP	NM_005143	1.60E-10
8043995	NM_000877 // IL1R1 // interleukin 1 receptor, type I // 2q12 // 3554	IL1R1	NM_000877	1.65E-10
8120552	NM_001105531 // FAM135A // family with sequence similarity 135	FAM135A	NM_001105531	3.83E-10
8105040	NM_003999 // OSMR // oncostatin M receptor // 5p13.1 // 9180 ///	OSMR	NM_003999	5.43E-10
8096301	NM_001040058 // SPP1 // secreted phosphoprotein 1 // 4q21-q25	SPP1	NM_001040058	5.51E-10
8091411	NM_014220 // TM4SF1 // transmembrane 4 L six family member 1	TM4SF1	NM_014220	6.70E-10
8105229	NM_015946 // PELO // pelota homolog (Drosophila) // 5q11.2 // 53	PELO	NM_015946	7.49E-10
8015607	NM_139276 // STAT3 // signal transducer and activator of transcri	STAT3	NM_139276	8.47E-10
8089112	NM_182909 // FILIP1L // filamin A interacting protein 1-like // 3q12	FILIP1L	NM_182909	1.10E-09
8098581	NM_031953 // SNX25 // sorting nexin 25 // 4q35.1 // 83891 /// ENS	SNX25	NM_031953	1.21E-09
8099326	NM_020041 // SLC2A9 // solute carrier family 2 (facilitated glucose	SLC2A9	NM_020041	1.33E-09
8102800	NM_014331 // SLC7A11 // solute carrier family 7, (cationic amino a	SLC7A11	NM_014331	1.40E-09
7952601	NM_001143820 // ETS1 // v-ets erythroblastosis virus E26 oncoge	ETS1	NM_001143820	1.64E-09
8136940	NM_001130025 // FAM115C // family with sequence similarity 115	FAM115C	NM_001130025	2.02E-09
8150592	NM_005195 // CEBPD // CCAAT/enhancer binding protein (C/EBP	CEBPD	NM_005195	2.14E-09
7910134	NM_031944 // MIXL1 // Mix1 homeobox-like 1 (Xenopus laevis) //	MIXL1	NM_031944	2.16E-09
7988767	NM_031226 // CYP19A1 // cytochrome P450, family 19, subfamily	CYP19A1	NM_031226	2.17E-09
8021563	---		---	2.55E-09
8068202	NM_058187 // C21orf63 // chromosome 21 open reading frame 63	C21orf63	NM_058187	2.61E-09
7973352	NM_014045 // LRP10 // low density lipoprotein receptor-related pr	LRP10	NM_014045	2.76E-09
8094679	NM_175737 // KLB // klotho beta // 4p14 // 152831 /// ENST00000	KLB	NM_175737	3.26E-09
7994280	NM_000418 // IL4R // interleukin 4 receptor // 16p12.1-p11.2 // 35	IL4R	NM_000418	3.28E-09
7998063	NM_006086 // TUBB3 // tubulin, beta 3 // 16q24.3 // 10381 /// NM_	TUBB3	NM_006086	3.58E-09
8035351	NM_000215 // JAK3 // Janus kinase 3 // 19p13.1 // 3718 /// ENST	JAK3	NM_000215	3.80E-09
7904244	NM_152367 // C1orf161 // chromosome 1 open reading frame 161	C1orf161	NM_152367	3.83E-09
8103326	NM_021870 // FGG // fibrinogen gamma chain // 4q28 // 2266 /// N	FGG	NM_021870	3.93E-09
8054135	NM_012214 // MGAT4A // mannosyl (alpha-1,3-)-glycoprotein beta	MGAT4A	NM_012214	4.27E-09

Part II

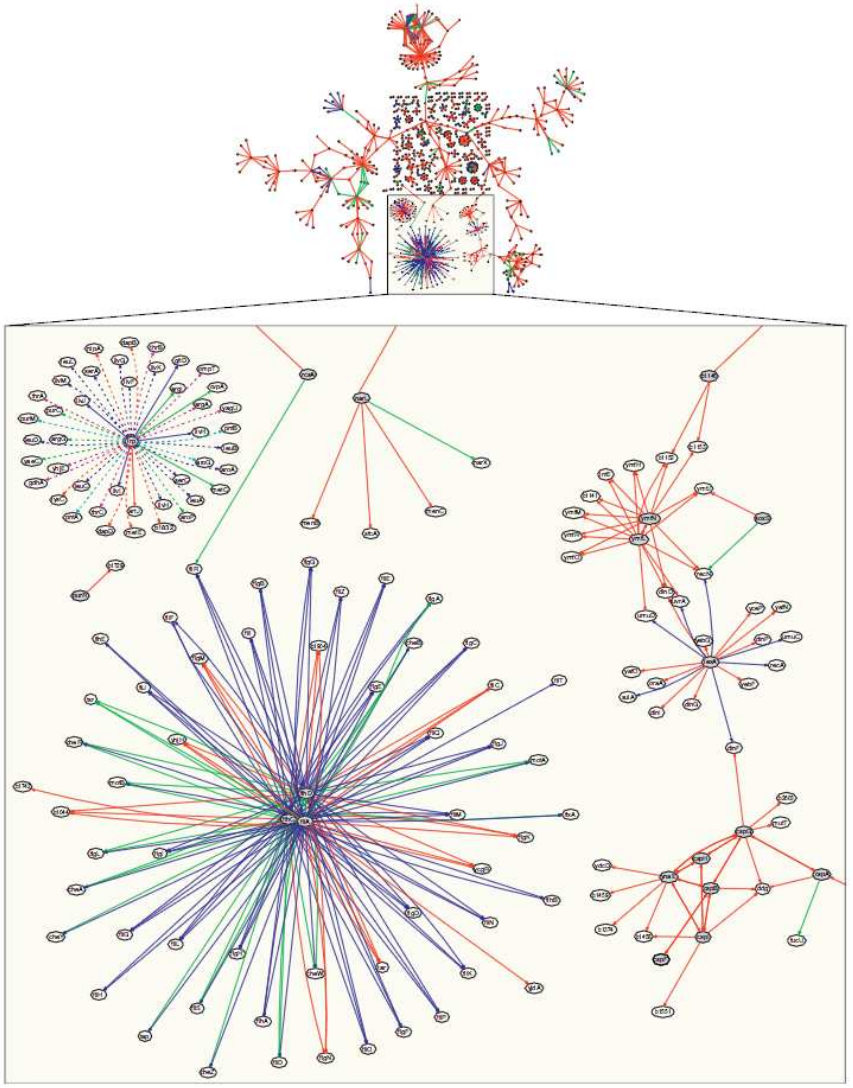
Gene Networks and Co-expression

Single pathway





Genome-wide pathways

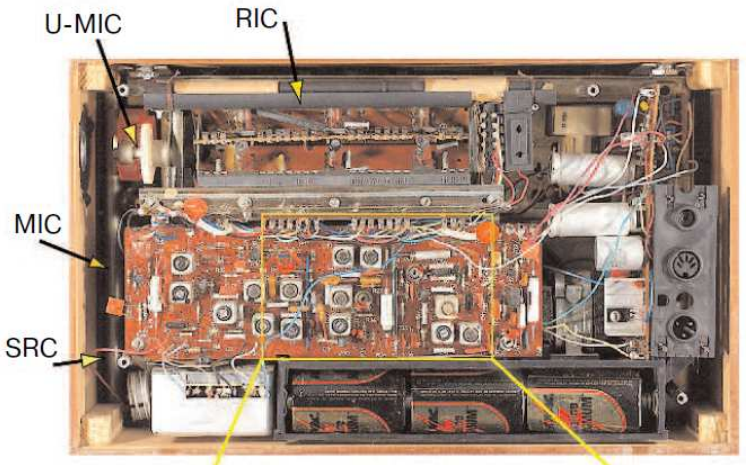


- How to transform experimental data into networks?
- How to validate the hypotheses about networks?
- How to use the knowledge about the networks?

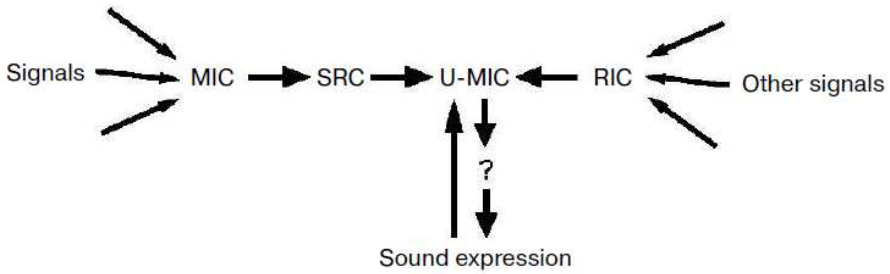
GENE NETWORKS: CAN A BIOLOGIST FIX A RADIO?

Biological system: outside

Biological system: inside

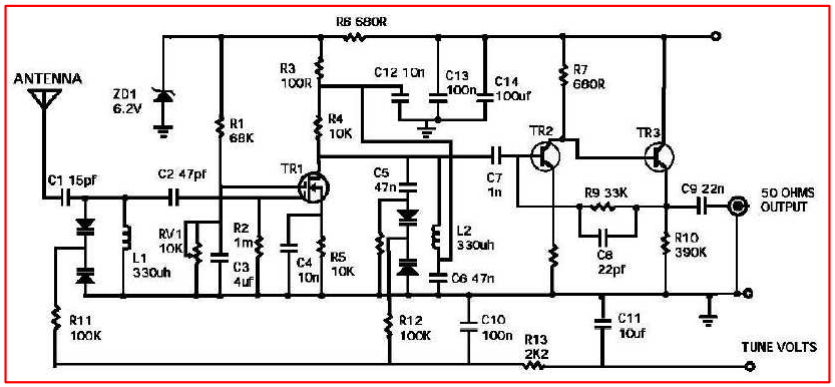


Current GN Representation



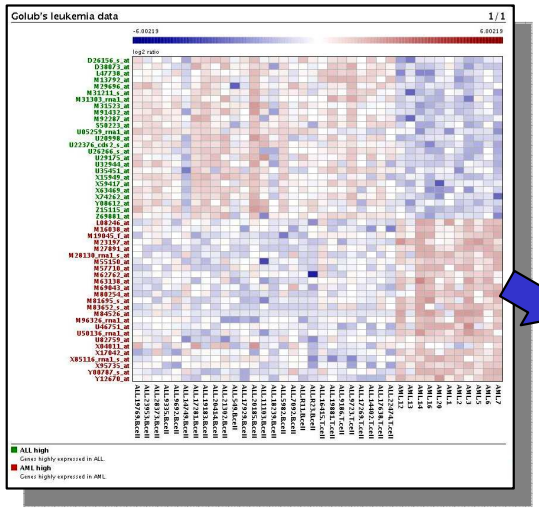
?

Reality

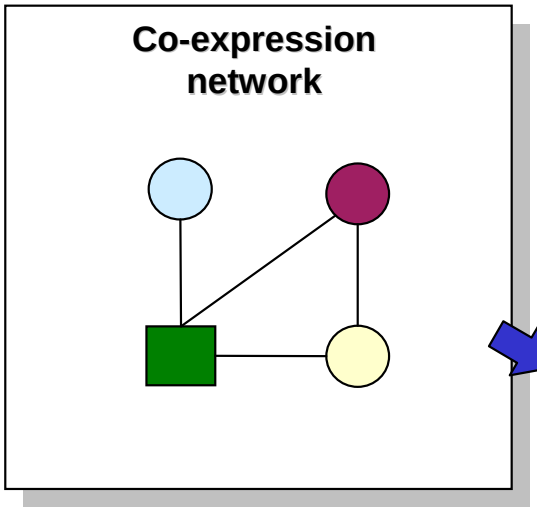


Y. Lazebnik, Cancer Cell, 2002, 2, 172-182

GENERAL “PIPELINE”

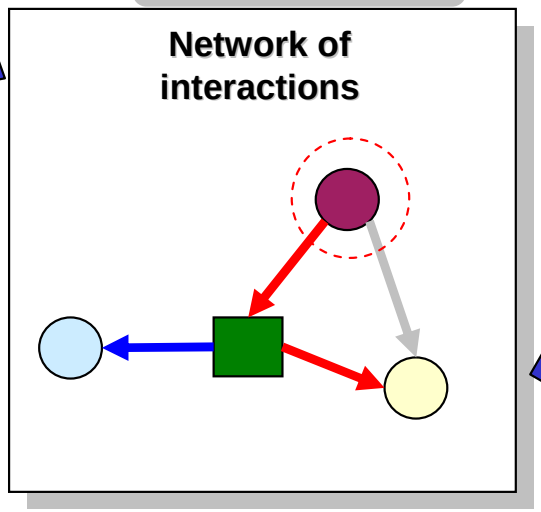


Co-expression analysis

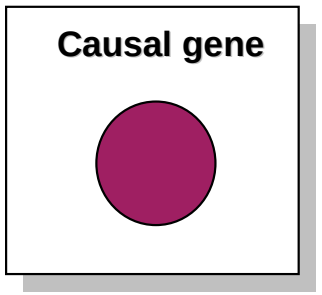


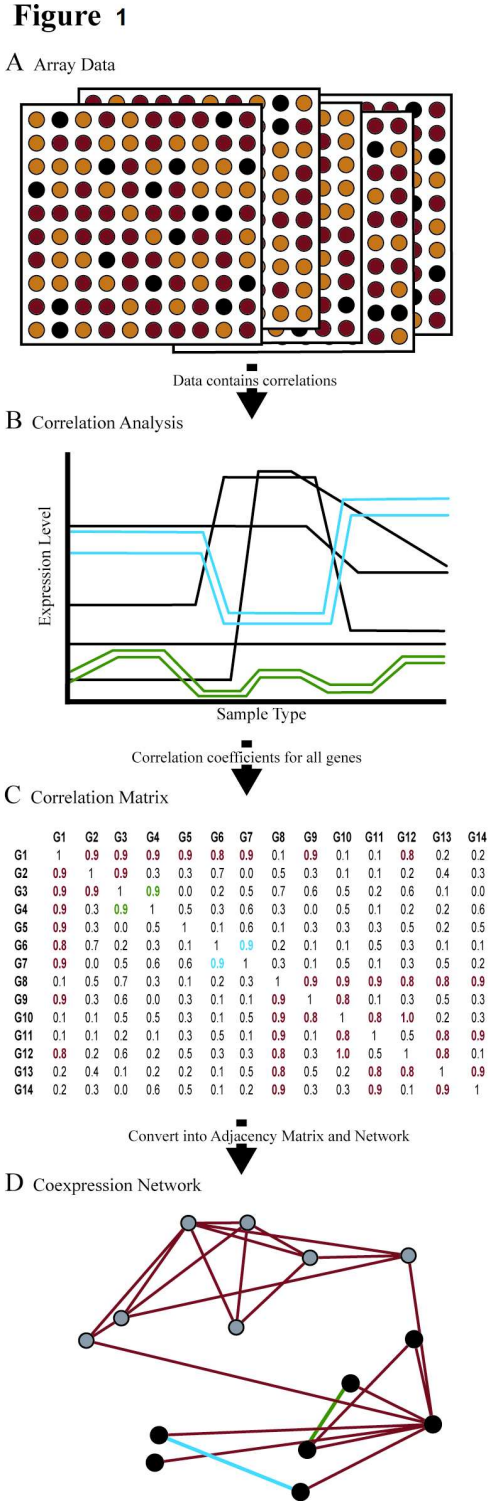
Topology analysis

Databases



Databases: PPI, TF





	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14
G1	1	0.9	0.9	0.9	0.9	0.8	0.9	0.1	0.9	0.1	0.1	0.8	0.2	0.2
G2	0.9	1	0.9	0.3	0.3	0.7	0.0	0.5	0.3	0.1	0.1	0.2	0.4	0.3
G3	0.9	0.9	1	0.9	0.0	0.2	0.5	0.7	0.6	0.5	0.2	0.6	0.1	0.0
G4	0.9	0.3	0.9	1	0.5	0.3	0.6	0.3	0.0	0.5	0.1	0.2	0.2	0.6
G5	0.9	0.3	0.0	0.5	1	0.1	0.6	0.1	0.3	0.3	0.3	0.5	0.2	0.5
G6	0.8	0.7	0.2	0.3	0.1	1	0.9	0.2	0.1	0.1	0.5	0.3	0.1	0.1
G7	0.9	0.0	0.5	0.6	0.6	0.9	1	0.3	0.1	0.5	0.1	0.3	0.5	0.2
G8	0.1	0.5	0.7	0.3	0.1	0.2	0.3	1	0.9	0.9	0.9	0.8	0.8	0.9
G9	0.9	0.3	0.6	0.0	0.3	0.1	0.1	0.9	1	0.8	0.1	0.3	0.5	0.3
G10	0.1	0.1	0.5	0.5	0.3	0.1	0.5	0.9	0.8	1	0.8	1.0	0.2	0.3
G11	0.1	0.1	0.2	0.1	0.3	0.5	0.1	0.9	0.1	0.8	1	0.5	0.8	0.9
G12	0.8	0.2	0.6	0.2	0.5	0.3	0.3	0.8	0.3	1.0	0.5	1	0.8	0.1
G13	0.2	0.4	0.1	0.2	0.2	0.1	0.5	0.8	0.5	0.2	0.8	0.8	1	0.9
G14	0.2	0.3	0.0	0.6	0.5	0.1	0.2	0.9	0.3	0.3	0.9	0.1	0.9	1

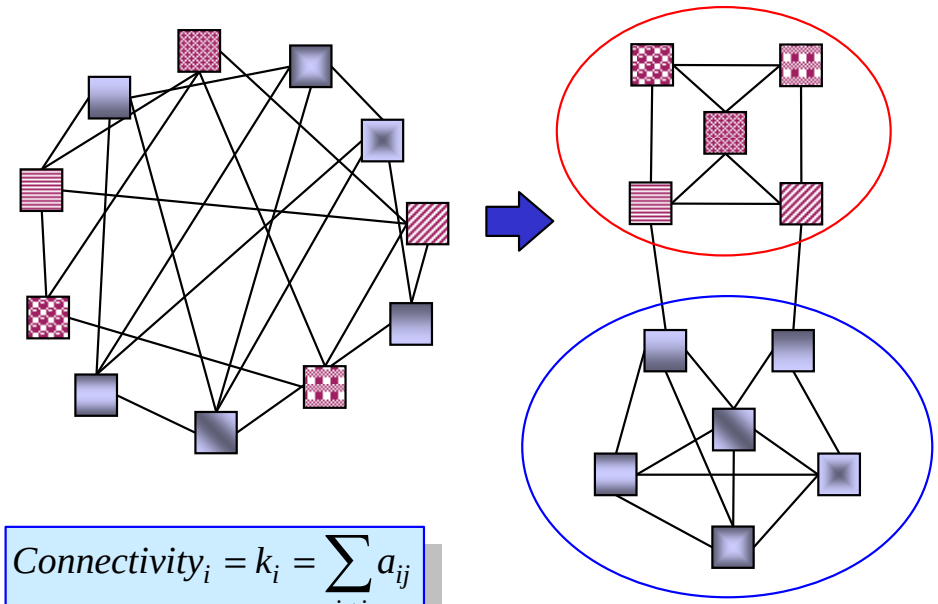
Co-expression Network Reconstruction

- Microarray transcriptomic data (A)
  - 2-color with a common reference or 1-color data or
  - Data normalization to remove batch and slide effects
- Concordance of gene expression (co-expression) (B)
  - Pearson correlation (linear interaction)
  - Spearman correlation
  - Mutual information (non-linear interaction)
  - etc… (> 10 various methods can be found)
- Transformation of concordance matrix (C)
  - CM can be dichotomized → unweighted network
  - or transformed continuously → weighted network
- Building and topology analysis of the co-expression network (*math*.: *undirected graph*) (D)
  - modules detection
  - connectivity analysis
  - optimal visualization

Figures are adopted and adapted from Hovath et al.
http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/

TOPOLOGY ANALYSIS

Raw network **Processed network**



$$Connectivity_i = k_i = \sum_{j \neq i} a_{ij}$$

$$Density = \frac{\sum_i \sum_{j \neq i} a_{ij}}{n(n-1)} = \frac{S_1(k)}{n(n-1)} = \frac{mean(k)}{n-1}$$

where n is the number of network nodes.

$$Centralization = \frac{n}{n-2}\left(\frac{\max(k)}{n-1} - Density\right) \approx \frac{\max(k)}{n-1} - Density$$

$$ClusterCoef_i = \frac{\sum_{l \neq i} \sum_{m \neq i,l} a_{il} a_{lm} a_{mi}}{\left(\sum_{l \neq i} a_{il}\right)^2 - \sum_{l \neq i} a_{il}^2}$$

Main aims of the topology analysis

- Determine the sub-networks or **modules** (usually highly connected)
- Perform the **classification** of the network topologies
  - Distinguish between network of healthy and ill cells
- Optimal **visualization**

Methods

- Use graph theory (mathematics)
- Introduce network concepts (indices):
  - connectivity (how node is connected)
  - density (mean adjacency)
  - centralization (determines centers)
  - clustering coefficient, etc…
- Use weighted edges when possible → better results

"CoExpress" SOFTWARE TOOL

Goal and Features

Goal: obtain and study experimentally-derived gene regulatory networks

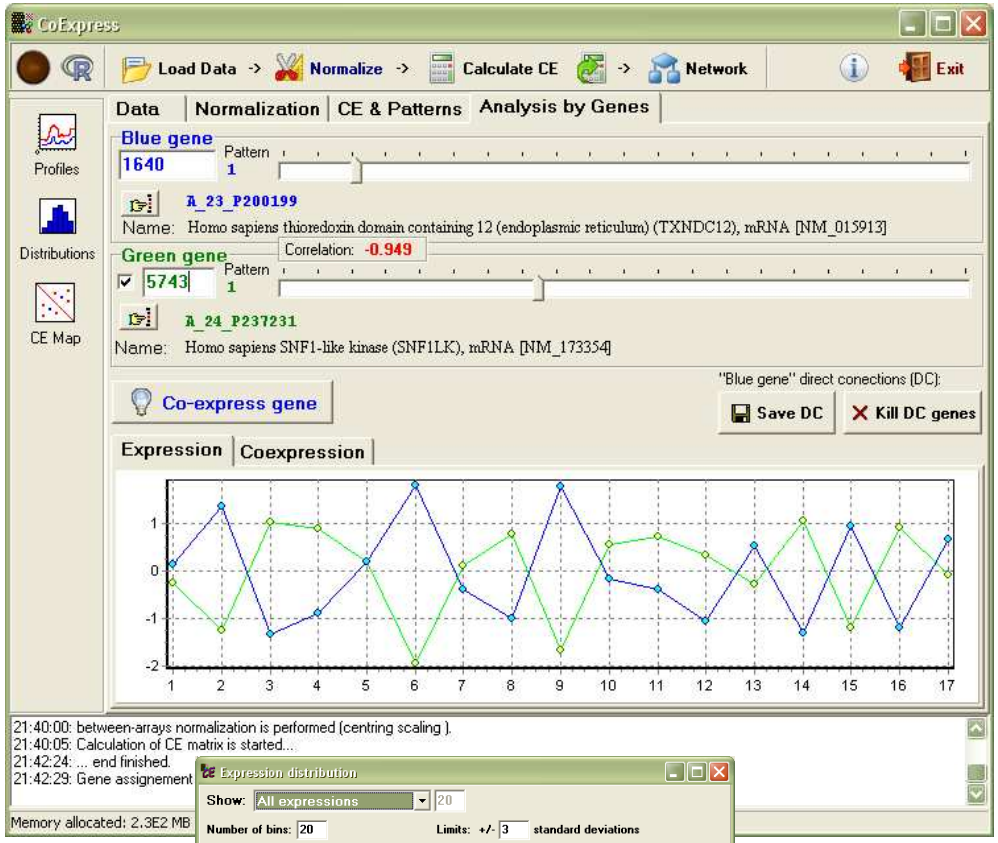
Features of the current versions

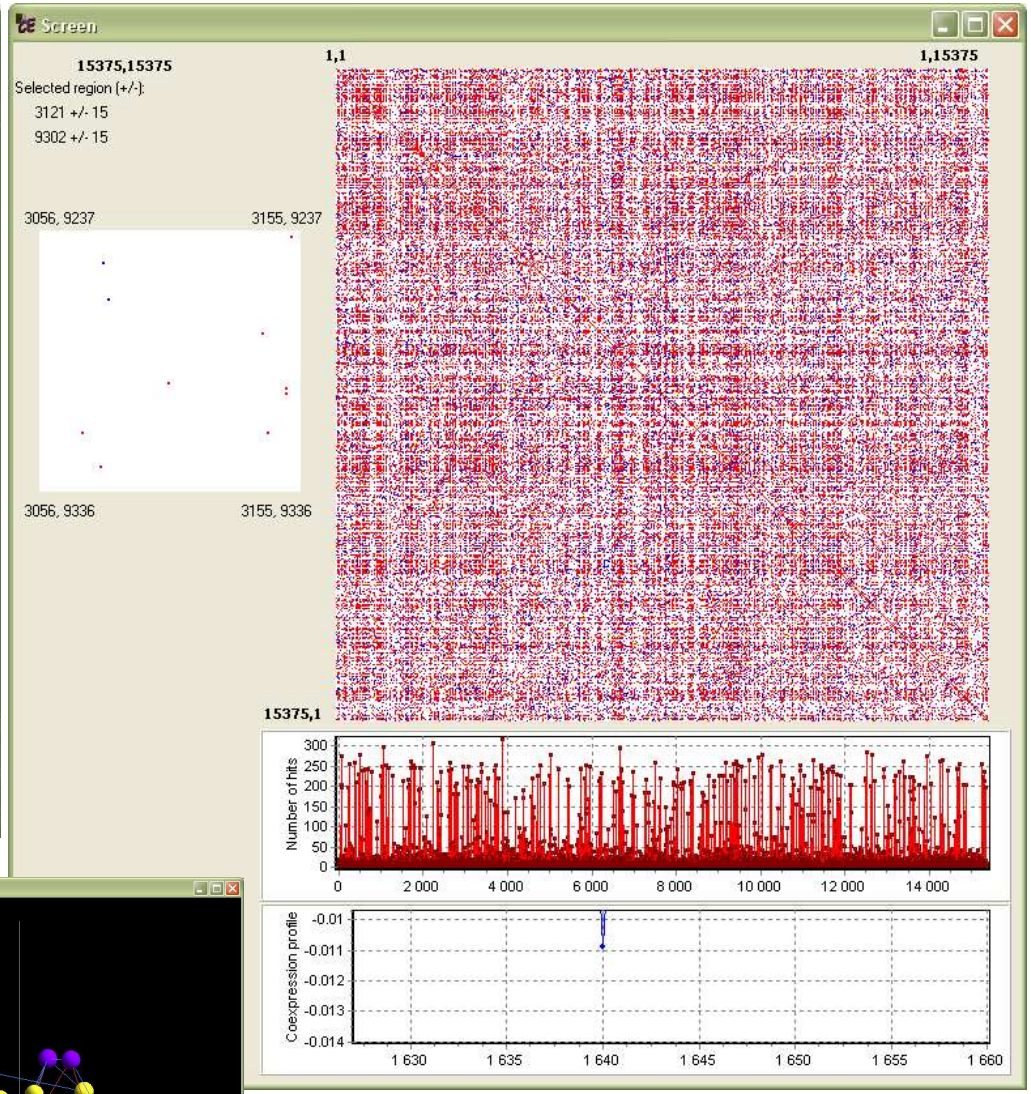
- User friendly (Windows version), multithread (Linux version)
- expression data linear normalization or R-based preprocessing (optionally)
- building and visualization of CE matrix using correlation or mutual information metrics;
- clustering, visualization and filtering of CE profiles
- visualization of co-expression networks for genes of interest

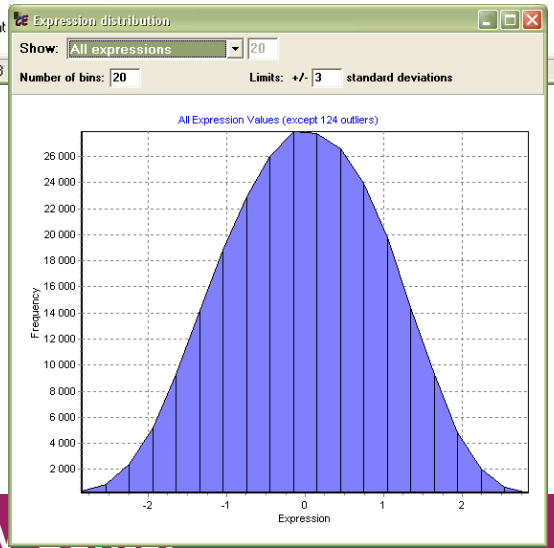
CENTRE DE RECHERCHE PUBLIC

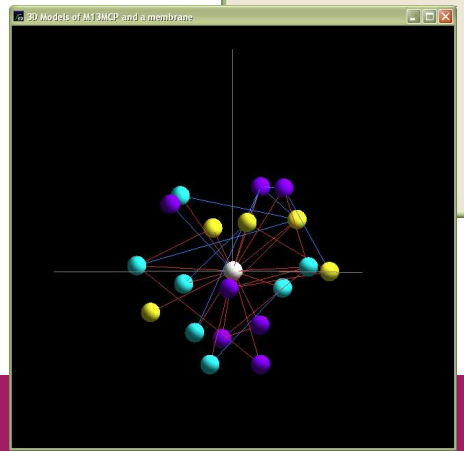
"CoExpress" SOFTWARE TOOL

CoExpress: freely available at *sablab.net/coexpress*





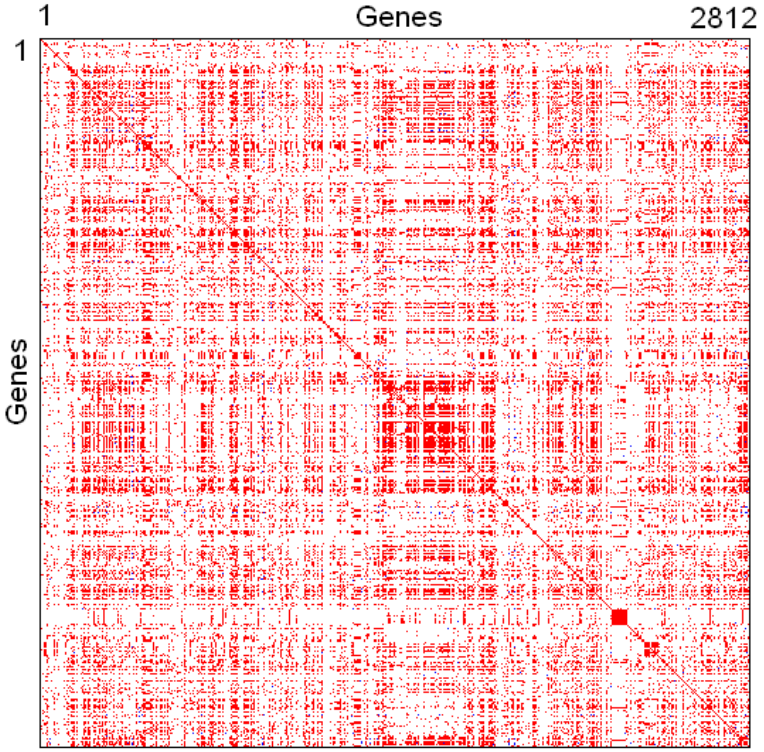




- Because of high cross-platform variability here we used only data from **2428 Affymetrix HGU133plus2 array** experiments.

- Data were normalized using RMA and then summarized, using gene symbols as indexes. The resulting data matrix, containing measurements for **19894 unique gene symbols**, were analyzed using the multi-thread Linux version of CoExpress.

- The analysis revealed that **2812 genes are co-expressed** (each has at least one other gene with the absolute correlation $|r| \geq 0.8$).

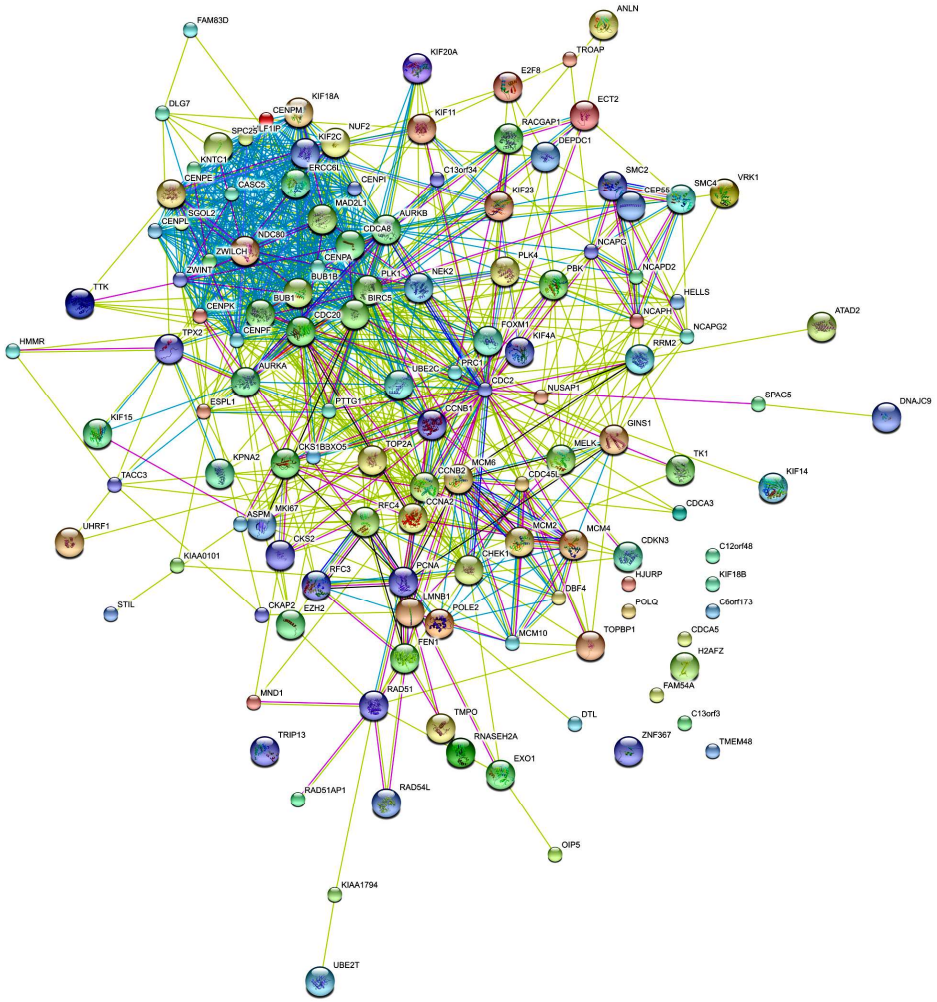


ANALYSIS OF THE PUBLIC DATA

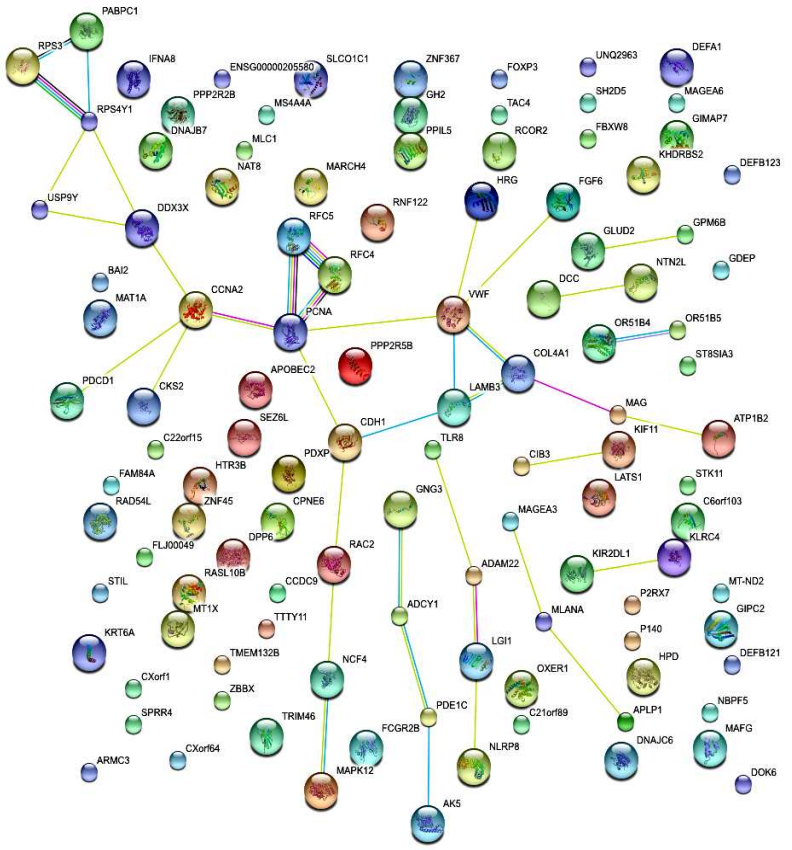
Validation II

➢ The validation of the obtained co-expression networks were performed using **STRING 8.2** (http://string.embl.de) – a service, public database and web resource dedicated to protein-protein interaction.

127 co-expressed genes



127 random genes



CONCLUDING REMARKS AND PLANS

- CoExpress is an actively-developed tool
  - Freely available at **http://sablab.net/coexpress/**
  - Often updated
  - Needs users for improving its qualities
- To be done soon:
  - Optimal visualization of the network
  - Topology analysis
  - Introduction correlation with miRNA expression profiles
- Next step: try to combine with exon-level analysis, to detect correlation between AS and gene expression.

- CoExpress
  - Algorithms for the network topology analysis
  - PCA in C/C++
  - Introduction optimal mRNA-vs-miRNA coexpression
- Simulation of microarray data/experiments
  - Simulation of the gene expression (ANOVA model can be used)
  - Simulation of gene networks and resulting MA data
- User friendly multifactor ANOVA

MICROARRAY CENTER

Created: 2005
2007 – Oncology Department
2010 – Department of Technology

microarray.lu
bioinformatics.lu

Head of the Center

- **Laurent Vallar, PhD**

Experimental Unit

- **Christelle Ghoneim, PhD** *(biology)*
- **Nathalie Nicot, Eng.** *(MA, biology)*
- **Francois Bernardin,** *(MA, biology)*

Bioinformatics and Biostatistics Unit

- **Petr Nazarov, PhD** *(biostatics)*
- **Arnaud Muller, Eng.** *(bioinformatics)*
- **Tony Kaoma, Eng.** *(bioinformatics)*

Students

- **Joana Corte-Real**
  (MSc, systems biology)
- **Stephanie Schmitz**
  (MSc, systems biology)

Thank you for your attention

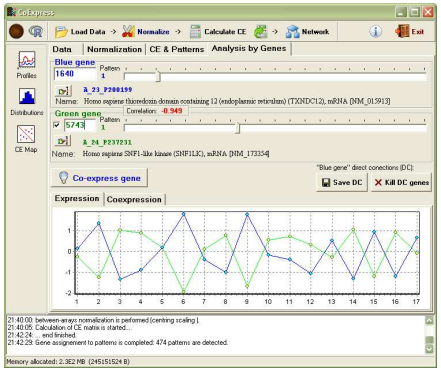
STORAGE

"CoExpress" SOFTWARE TOOL

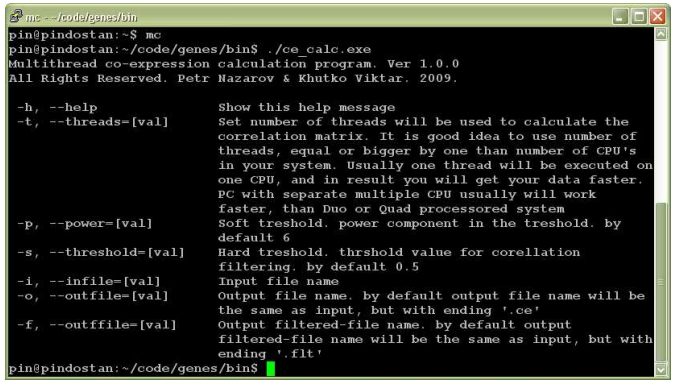
Technical Notes

CoExpress

Windows-based:
graphical, user friendly tool



Linux-based:
fast, multithread command-line tool



Data export

Further analysis:
in CoExpress, R, Excel, etc.

Parameters	Windows	Linux
Maximum genes	~ 30 000	> 60 000
Maximum arrays	< 1 000	> 1 000
Multi-CPU support	-	+
Graphical User Inerface	+	-
Compiller	bcc32	gcc
Time for CE calculation on a big dataset*		
1 CPU	3h 45m	55 m
8 CPU	n/a	7m
Time for CE calculation on a small dataset**		
1 CPU	1m 26s	1m 13s

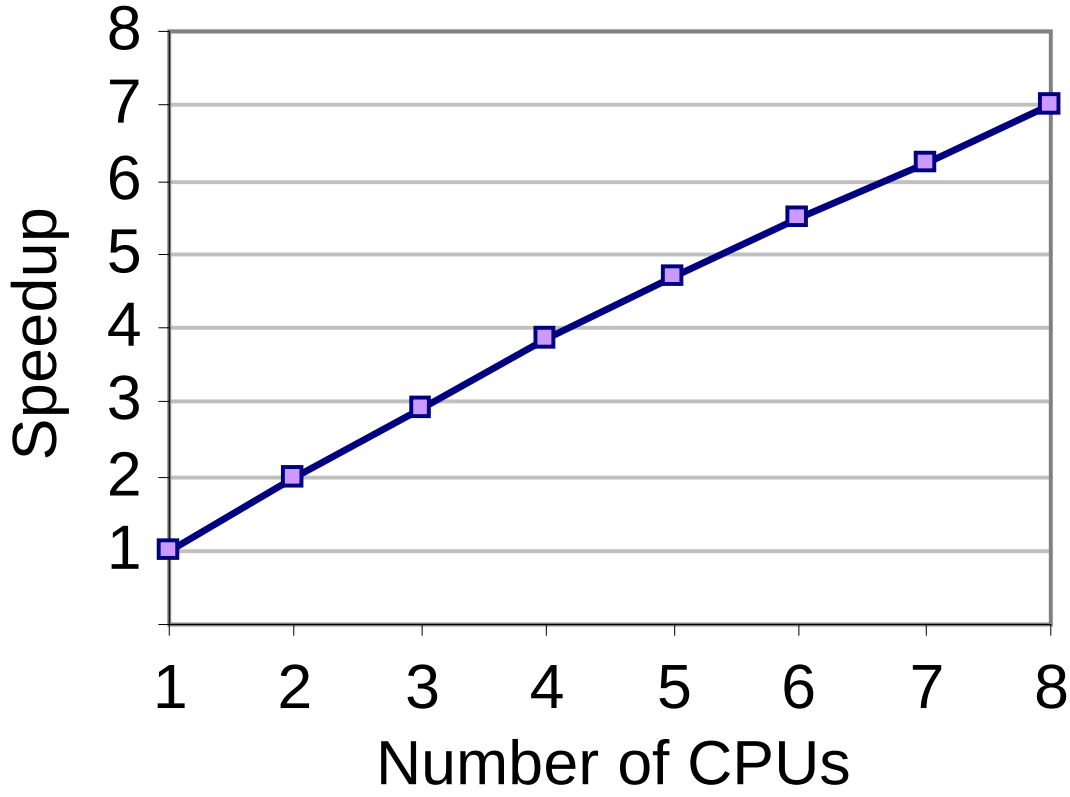
(*) 2428 Affymetrix arrays with 19894 genes were used

(**) 17 Agilent two-color arrays with 15375 genes were used

Efficiency of parallelization

Speedup with the increase of number of CPUs

on a 2428 MA and 20K genes

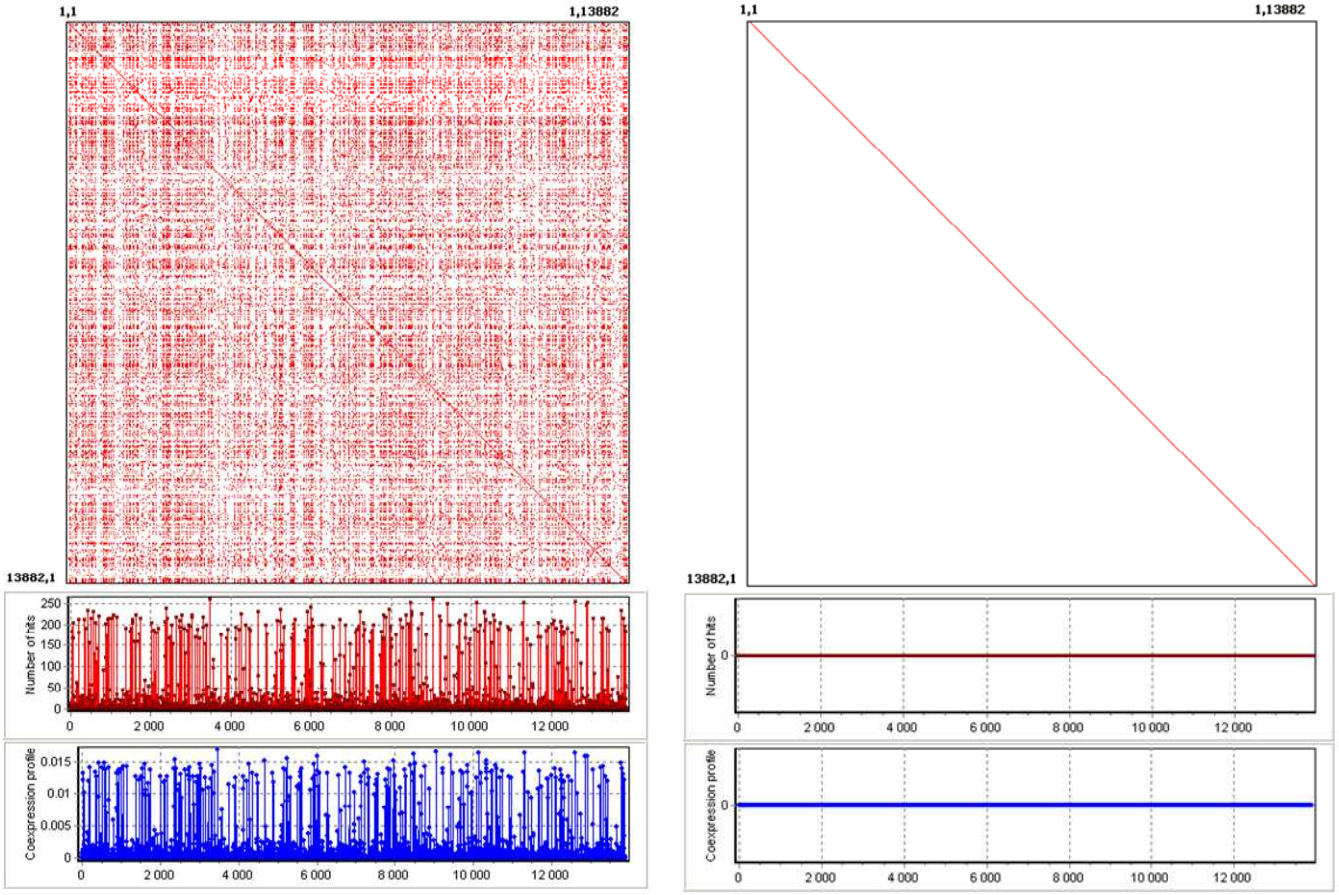


Linear part of Amdahl's law. Approximately 97% of the calculations can be parallelized *(value for correlation measure)*

Can co-expression pattern be just due to a coincidence?

Experiment. Comparison CE matrixes for experimental and simulated datasets. The filtering criterion selected for this experiment was

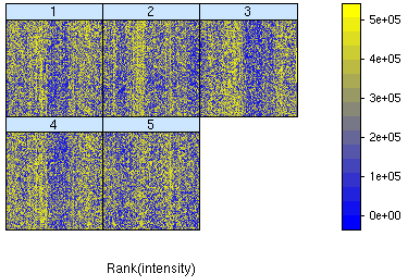
$r_{i,j}^2 > 0.8$, i.e. $|r_{i,j}| > 0.894$



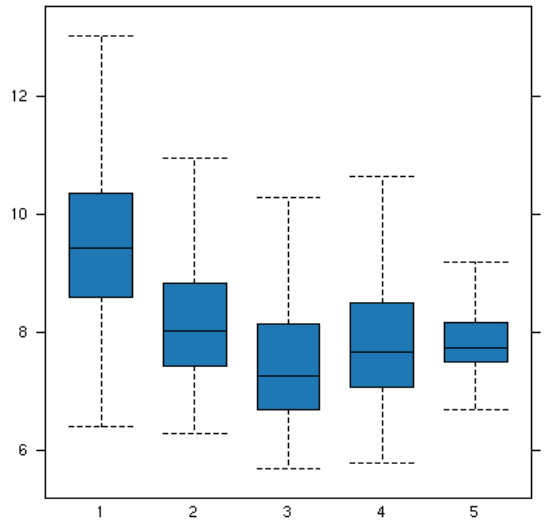
(a) (b)

Figure 1. Gene co-expression patterns for the best 13882 genes obtained on 34 experimental (a) and randomized (b) MAs.

- Data an Meta-data come from Geo (**ncbi**) and Array Express (**ebi**)
- Downloading of approximately 10 000 files with 200 pathologies and 300 tissues from 5 five types of platforms (hgu133a,hgu133b,hgu133plus2,hgu95a,hgu95av2)
- **Meta-Data**
  - Manually currated
  - We focus on the localization, pathology and tissues
- **Data**
  - Quality assessment (R scripts)
    - Box plot
    - Spatial distribution
    - ….
  - Quality control
    - We remove all the Data (and Meta-Data) for arrays of poor quality



Spatial distribution



Density plot

~7000 files after quality control