CRP SANTÉ
CENTRE DE RECHERCHE PUBLIC
Microarray Center

uni.lu
UNIVERSITÉ DU LUXEMBOURG

CoExpress

A Tool for Co-Regulation Analysis of mRNA and miRNA

Petr Nazarov, V.Khutko, A.Muller, N.Nicot, T.Kaoma, D.Philippidou, S.Schmitz, S.Kreis, L.Vallar

`http://bioinformatics.lu/CoExpress`

5-11-2010

CENTRE DE RECHERCHE PUBLIC SANTÉ

OUTLINE

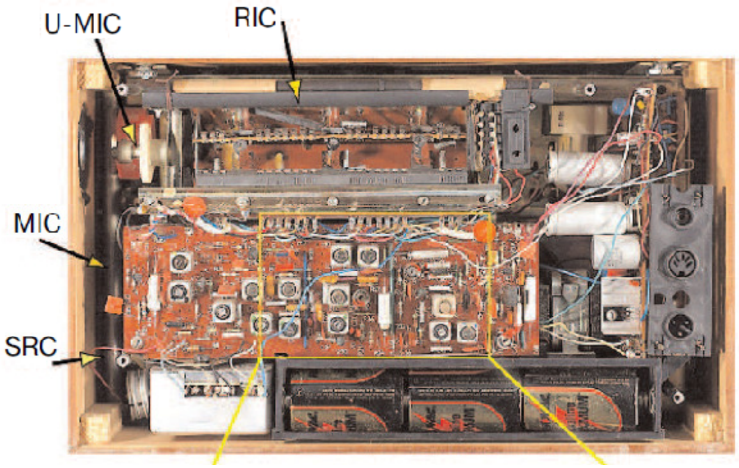
- **Introduction**
  - why gene networks
  - co-expression analysis as a tool for gene network reconstruction
- **CoExpress tool**
  - goal and tasks
  - technical details
- **mRNA - mRNA: public data analysis**
  - public data processing pipeline
  - functional validation
  - statistical validation
- **miR - mRNA: 8 cell lines**
  - data description
  - **demonstration of data analysis**
  - functional validation
  - statistical validation
- **Conclusions and future plans**

GENE NETWORKS: CAN A BIOLOGIST FIX A RADIO?

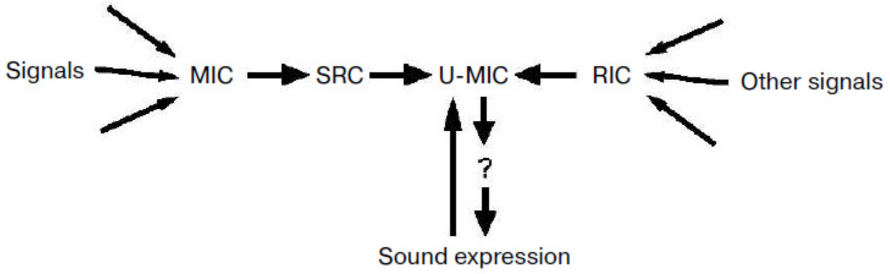
Classical Example*

Biological system: outside

Biological system: inside



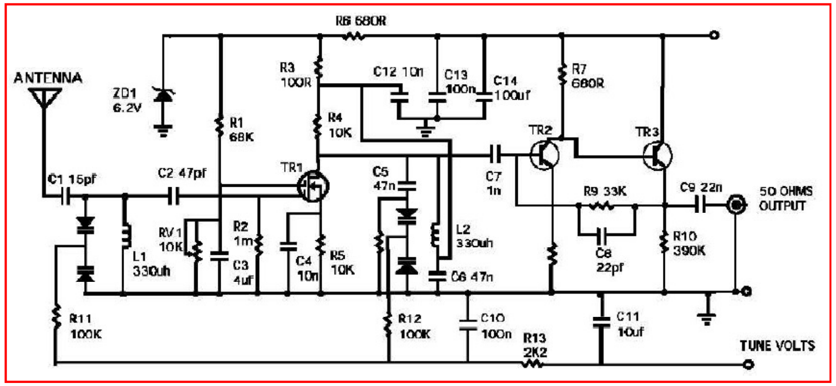
Current GN Representation



?

=

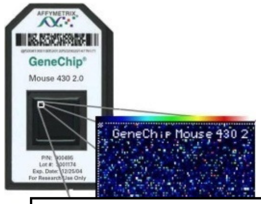
Reality



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

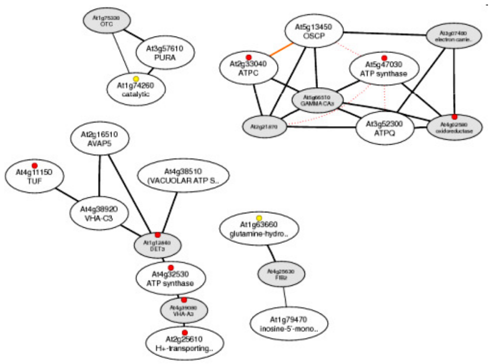
GENE CO-EXPRESSION

Experiments



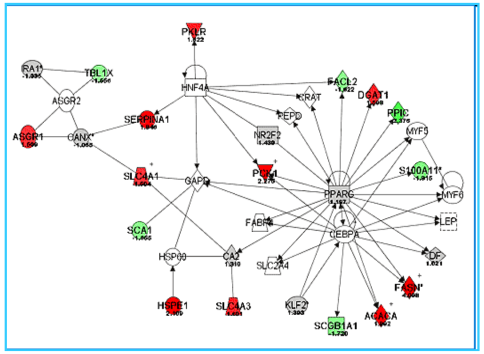
Public data

Co-expression network



Databases

Known gene networks and P-P interactions



Analysis

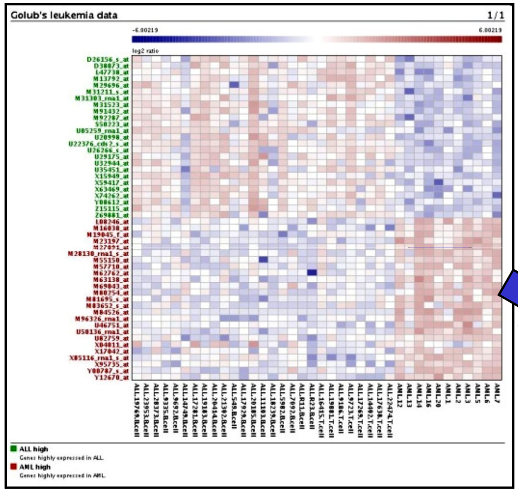
New interactions (pathways)

Better predictors (markers)

Disease-related pathways

Disease-causing genes

GENERAL “PIPELINE”



Golub's leukemia data

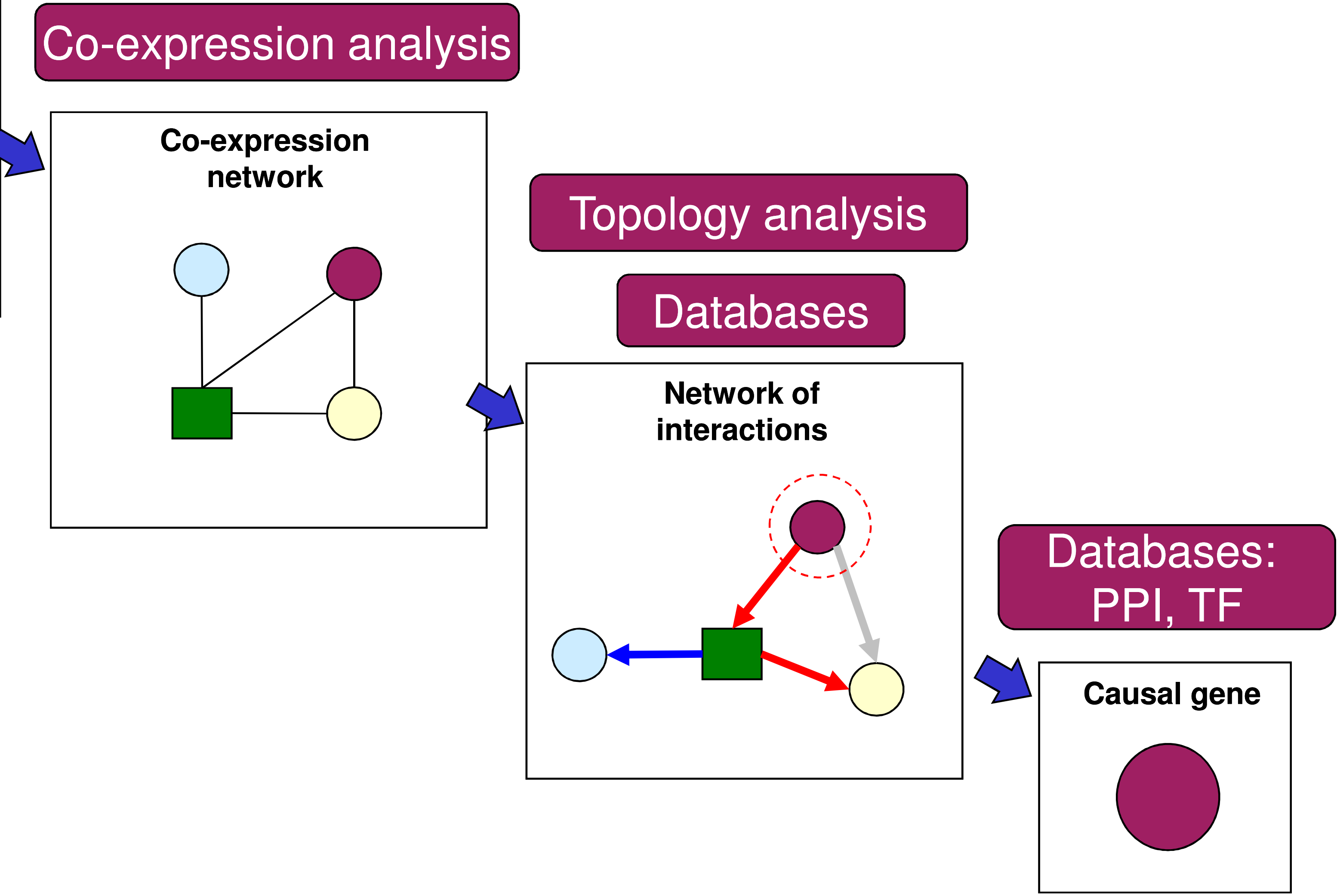
1/1

🟩 **ALL high**

Genes highly expressed in ALL.

🟥 **AML high**

Genes highly expressed in AML.



Co-expression analysis

Co-expression network

Topology analysis

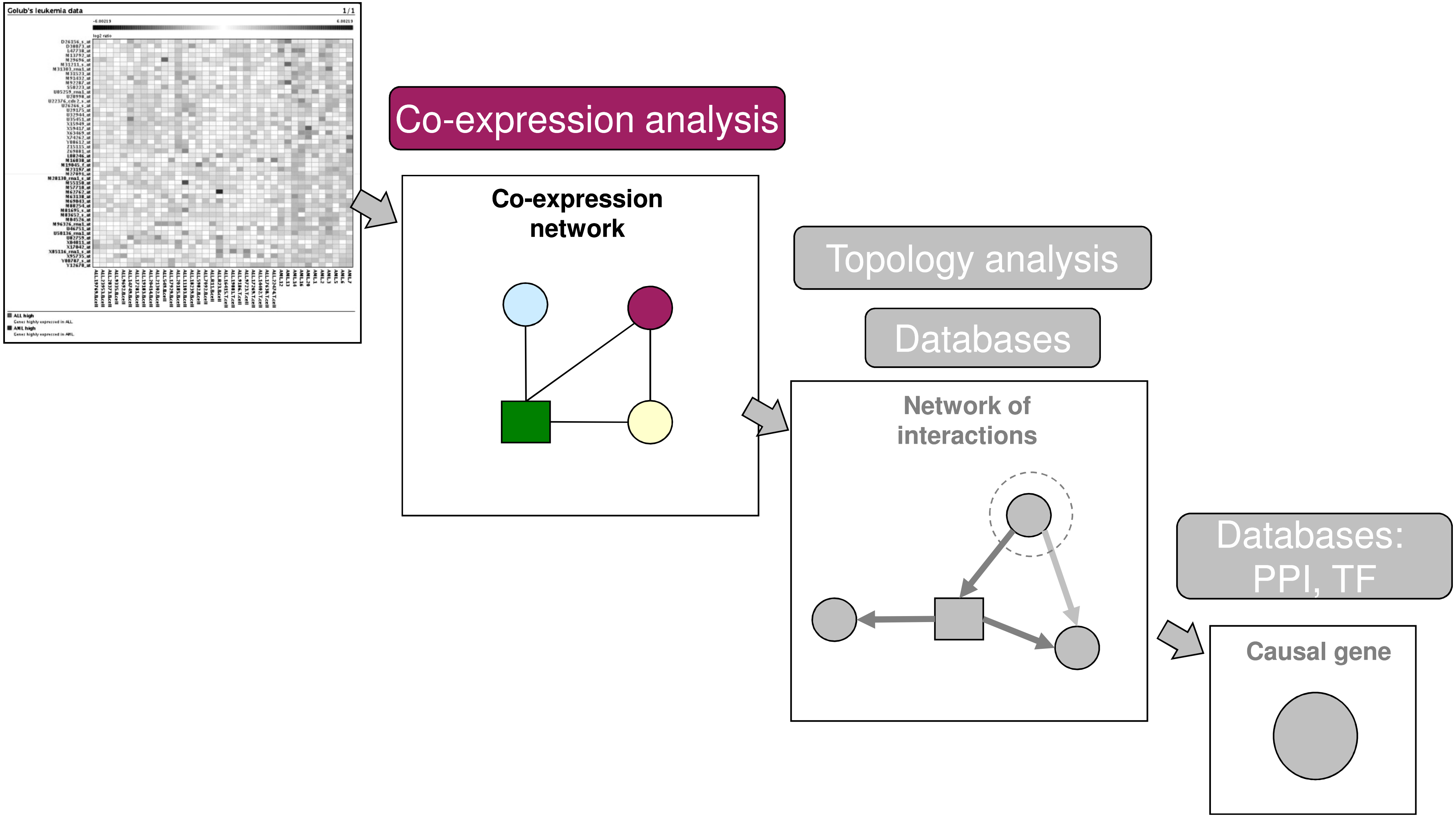
Databases

Network of interactions

Databases: PPI, TF

Causal gene

CO-EXPRESSION



Co-expression analysis

Co-expression network

Topology analysis

Databases

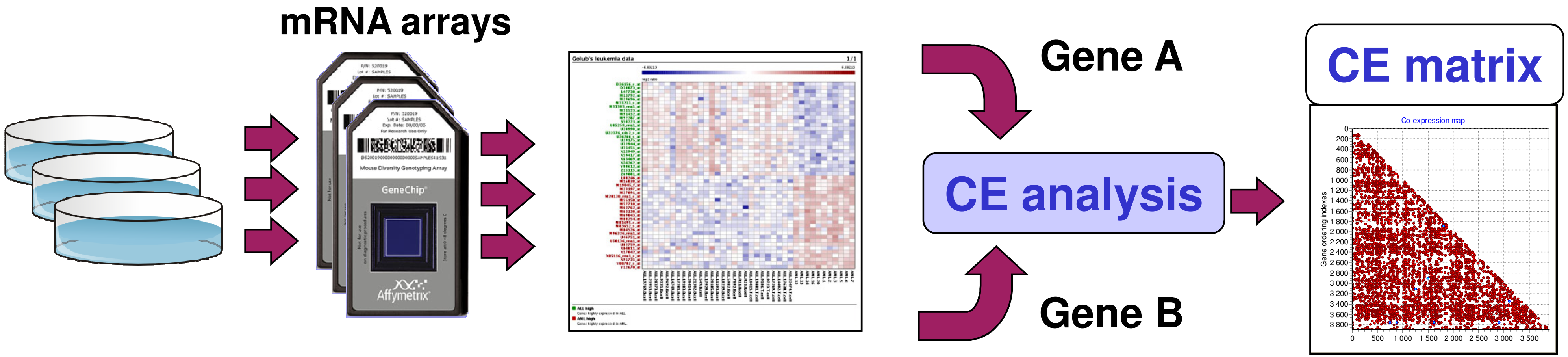
Network of interactions

Databases: PPI, TF

Causal gene

CO-EXPRESSION

Two Ways to Look on Co-Expression



mRNA arrays

Gene A

CE analysis

Gene B

CE matrix

Many tools exist, to name just a few:

- **ARACNe** method (Margolin, 2006)
- **LeMoNe**: learning modular network (Michoel, 2009;De Smet, 2010)

…but not user-friendly and easy-to-install ☺

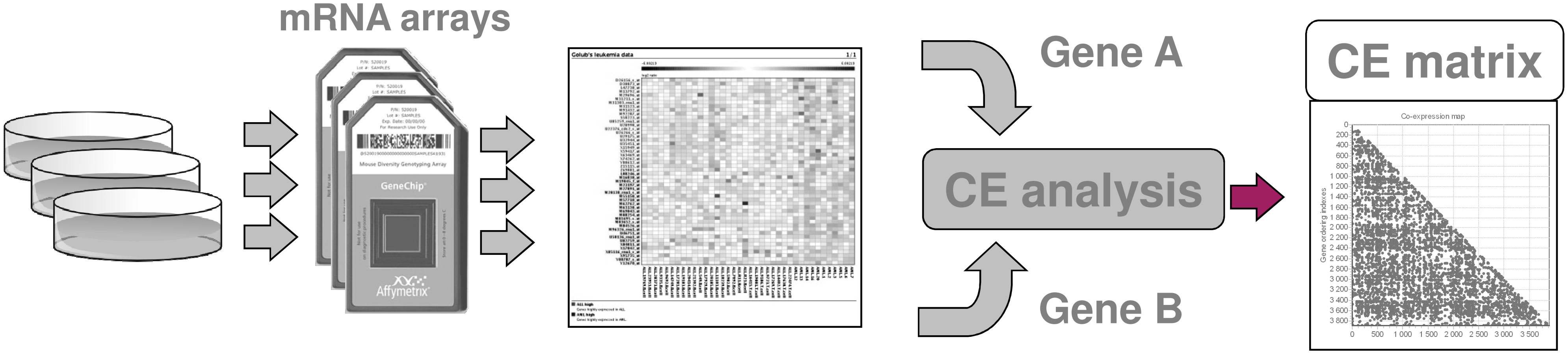
- **StarNet2**: web-based analysis using the large set of preselected public data (Jupiter, 2009)

…but not with user-data

And almost no tools for the moment works with the second approach:

CO-EXPRESSION

Two Ways to Look on Co-Expression



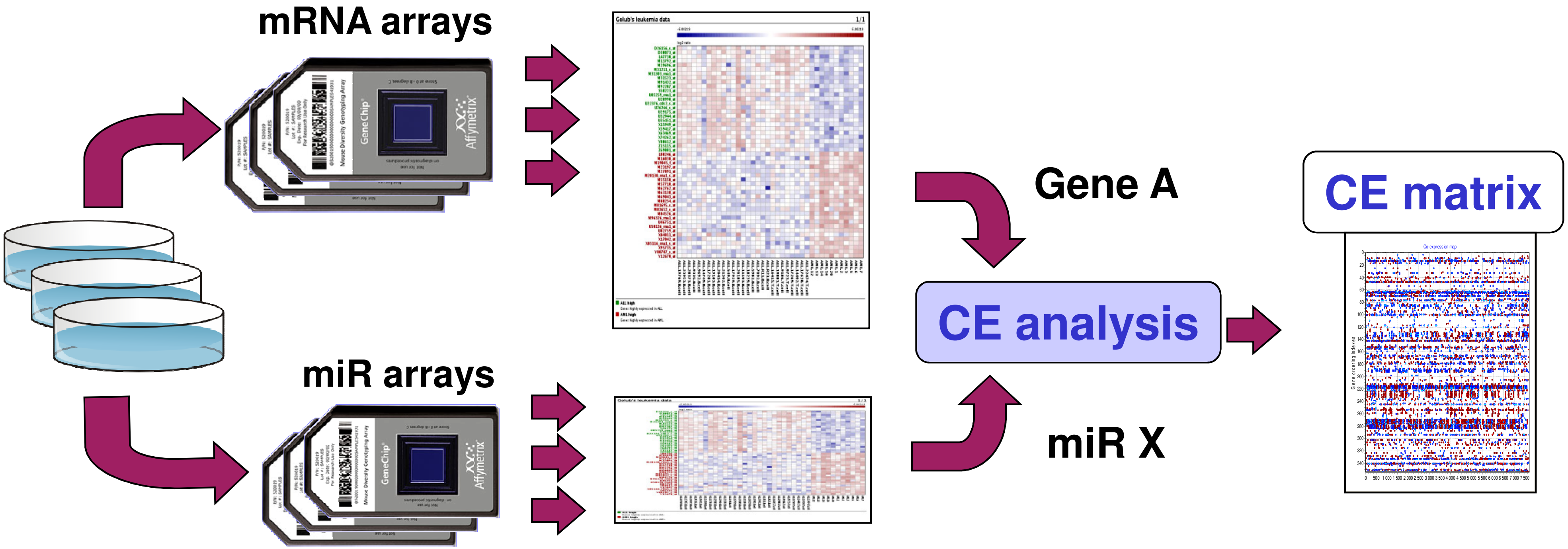
mRNA arrays

Gene A

CE analysis

Gene B

CE matrix



mRNA arrays

miR arrays

Gene A

CE analysis

miR X

CE matrix

“CoExpress” SOFTWARE TOOL

Goal and Features

Goal: develop a friendly tool for the interactive analysis of gene-gene, gene-miR (and even gene-protein) co-regulations

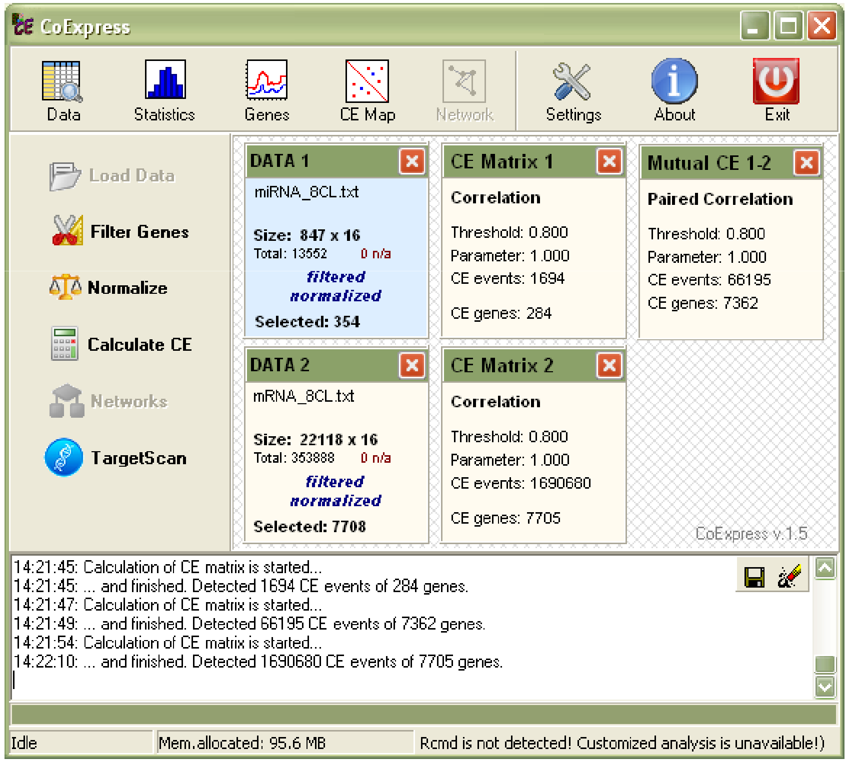
Features of the current version

- User friendly (Windows version), multithread (Linux version)
- Fast linear normalization or customized R-based preprocessing (optionally)
- Interactive comparison of gene expression profiles
- Building and visualization of CE matrix using Pearson correlation or mutual information metrics;
- Data export into SIF format of Cytoscape for network visualization
- Comparison of detected miR-mRNA pairs with TargetScan predictions

crp SANTÉ
CENTRE DE RECHERCHE PUBLIC

“CoExpress” SOFTWARE TOOL

CoExpress: freely available at *bioinformatics.lu*



DATA 1
miRNA_8CL.txt
Size: 847 x 16
Total: 13552 0 n/a
filtered
normalized
Selected: 354

CE Matrix 1
Correlation
Threshold: 0.800
Parameter: 1.000
CE events: 1694
CE genes: 284

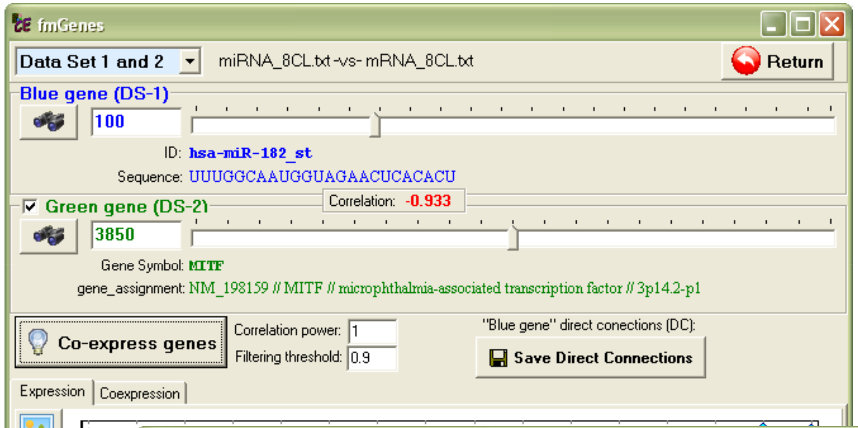
Mutual CE 1-2
Paired Correlation
Threshold: 0.800
Parameter: 1.000
CE events: 66195
CE genes: 7362

DATA 2
mRNA_8CL.txt
Size: 22118 x 16
Total: 353888 0 n/a
filtered
normalized
Selected: 7708

CE Matrix 2
Correlation
Threshold: 0.800
Parameter: 1.000
CE events: 1690680
CE genes: 7705

```
14:21:45: Calculation of CE matrix is started...
14:21:45: ... and finished. Detected 1694 CE events of 284 genes.
14:21:47: Calculation of CE matrix is started...
14:21:49: ... and finished. Detected 66195 CE events of 7362 genes.
14:21:54: Calculation of CE matrix is started...
14:22:10: ... and finished. Detected 1690680 CE events of 7705 genes.
```

Idle | Mem.allocated: 95.6 MB | Rcmd is not detected! Customized analysis is unavailable!)

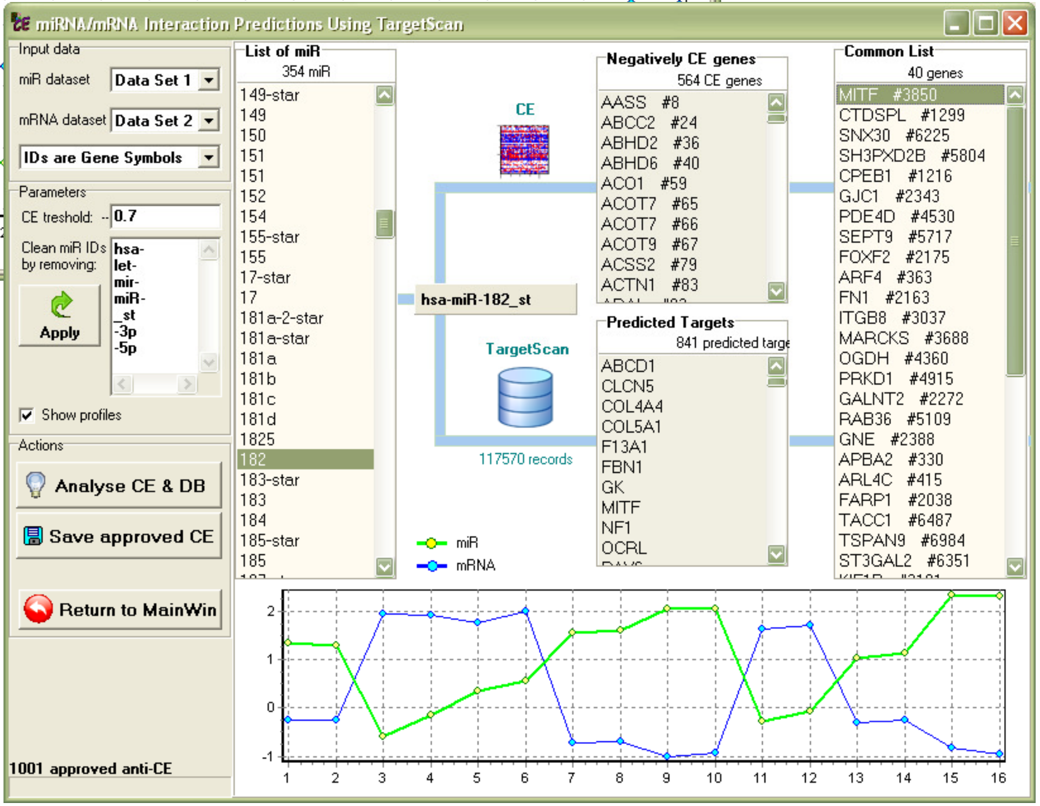


Data Set 1 and 2 miRNA_8CL.txt -vs- mRNA_8CL.txt

Blue gene (DS-1): 100
ID: hsa-miR-182_st
Sequence: UUUGGCAAUGGUAGAACUCACACU

☑ Green gene (DS-2): 3850
Correlation: -0.933
Gene Symbol: MITF
gene_assignment: NM_198159 // MITF // microphthalmia-associated transcription factor // 3p14.2-p1

Correlation power: 1
Filtering threshold: 0.9
"Blue gene" direct conections (DC):



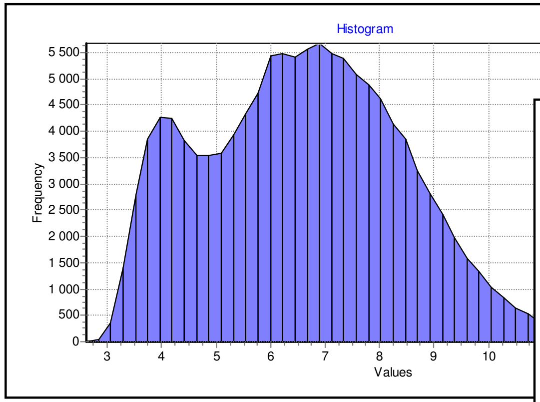
Input data
miR dataset: Data Set 1
mRNA dataset: Data Set 2
IDs are Gene Symbols
Parameters
CE treshold: 0.7
Clean miR IDs by removing: hsa- let- mir- miR- _st -3p -5p
☑ Show profiles
Actions
1001 approved anti-CE

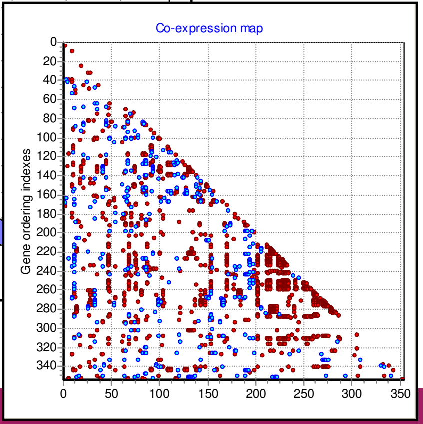
List of miR
354 miR
149-star
149
150
151
151
152
154
155-star
155
17-star
17
181a-2-star
181a-star
181a
181b
181c
181d
1825
182
183-star
183
184
185-star
185

Negatively CE genes
564 CE genes
AASS #8
ABCC2 #24
ABHD2 #36
ABHD6 #40
ACO1 #59
ACOT7 #65
ACOT7 #66
ACOT9 #67
ACSS2 #79
ACTN1 #83

Predicted Targets
841 predicted targe
ABCD1
CLCN5
COL4A4
COL5A1
F13A1
FBN1
GK
MITF
NF1
OCRL

Common List
40 genes
MITF #3850
CTDSPL #1299
SNX30 #6225
SH3PXD2B #5804
CPEB1 #1216
GJC1 #2343
PDE4D #4530
SEPT9 #5717
FOXF2 #2175
ARF4 #363
FN1 #2163
ITGB8 #3037
MARCKS #3688
OGDH #4360
PRKD1 #4915
GALNT2 #2272
RAB36 #5109
GNE #2388
APBA2 #330
ARL4C #415
FARP1 #2038
TACC1 #6487
TSPAN9 #6984
ST3GAL2 #6351





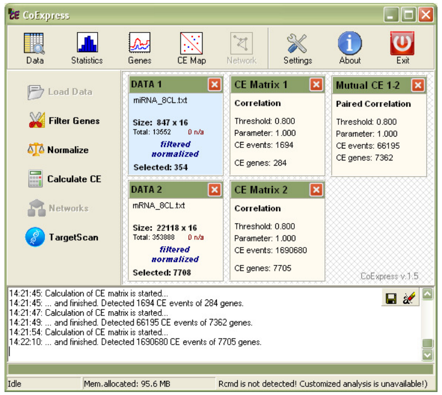
LuciLinx 2010 www.bioinformatics.lu

"CoExpress" SOFTWARE TOOL

Technical Notes

CoExpress

Windows-based:
graphical, user friendly tool



DATA 1
miRNA_8CL.txt
Size: 847 x 16
Total: 13552 0 n/a
filtered
normalized
Selected: 354

CE Matrix 1
Correlation
Threshold: 0.800
Parameter: 1.000
CE events: 1694
CE genes: 284

Mutual CE 1-2
Paired Correlation
Threshold: 0.800
Parameter: 1.000
CE events: 66195
CE genes: 7362

DATA 2
mRNA_8CL.txt
Size: 22118 x 16
Total: 353888 0 n/a
filtered
normalized
Selected: 7708

CE Matrix 2
Correlation
Threshold: 0.800
Parameter: 1.000
CE events: 1690680
CE genes: 7705

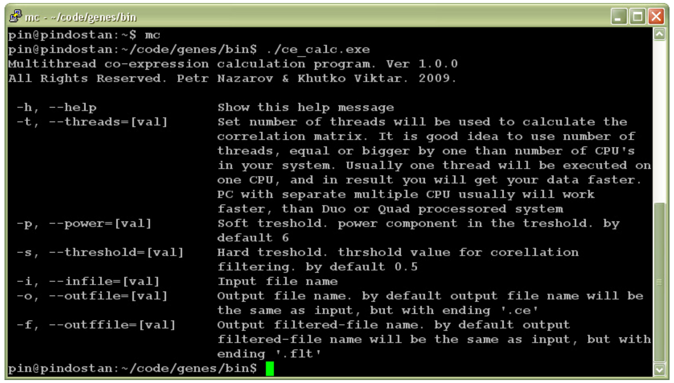
```
14:21:45: Calculation of CE matrix is started...
14:21:45: ... and finished. Detected 1694 CE events of 284 genes.
14:21:47: Calculation of CE matrix is started...
14:21:49: ... and finished. Detected 66195 CE events of 7362 genes.
14:21:54: Calculation of CE matrix is started...
14:22:10: ... and finished. Detected 1690680 CE events of 7705 genes.
```

Idle | Mem.allocated: 95.6 MB | Rcmd is not detected! Customized analysis is unavailable!)

Borland C++ Builder
(aka CodeGear)

Interactive Analysis

Linux-based:
fast, multithread command-line tool



```
pin@pindostan:~$ mc
pin@pindostan:~/code/genes/bin$ ./ce_calc.exe
Multithread co-expression calculation program. Ver 1.0.0
All Rights Reserved. Petr Nazarov & Khutko Viktar. 2009.

 -h, --help                Show this help message
 -t, --threads=[val]       Set number of threads will be used to calculate the
                           correlation matrix. It is good idea to use number of
                           threads, equal or bigger by one than number of CPU's
                           in your system. Usually one thread will be executed on
                           one CPU, and in result you will get your data faster.
                           PC with separate multiple CPU usually will work
                           faster, than Duo or Quad processored system
 -p, --power=[val]         Soft treshold. power component in the treshold. by
                           default 6
 -s, --threshold=[val]     Hard treshold. thrshold value for corellation
                           filtering. by default 0.5
 -i, --infile=[val]        Input file name
 -o, --outfile=[val]       Output file name. by default output file name will be
                           the same as input, but with ending '.ce'
 -f, --outffile=[val]      Output filtered-file name. by default output
                           filtered-file name will be the same as input, but with
                           ending '.flt'
pin@pindostan:~/code/genes/bin$
```

GNU C++ Compiler:
gcc

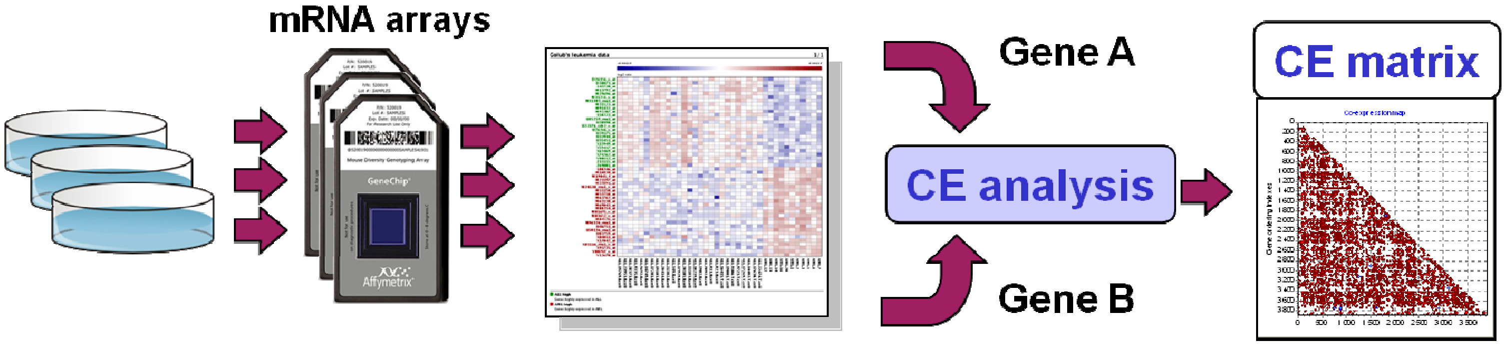
Data export

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

Application I

mRNA – mRNA

Analysis of a Big Public Dataset



mRNA arrays

Gene A

CE analysis

Gene B

CE matrix

ANALYSIS OF THE PUBLIC DATA

Public Data: Pipeline

EBI
ArrayExpress

human
Affy133plus2

blood,
plasma

>2000 MA

remove cancer and not annotated samples

APT: RMA summarization, quantile normalization.

R: QC, summarization by GeneSymbol

1354 MA
20233 genes

CoExpress:
filtering, linear normalization

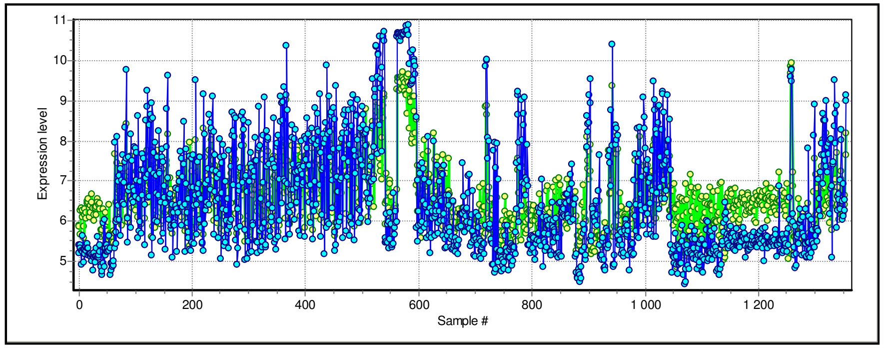
max LI > 6
st.dev > 0.8
remove “ORF”

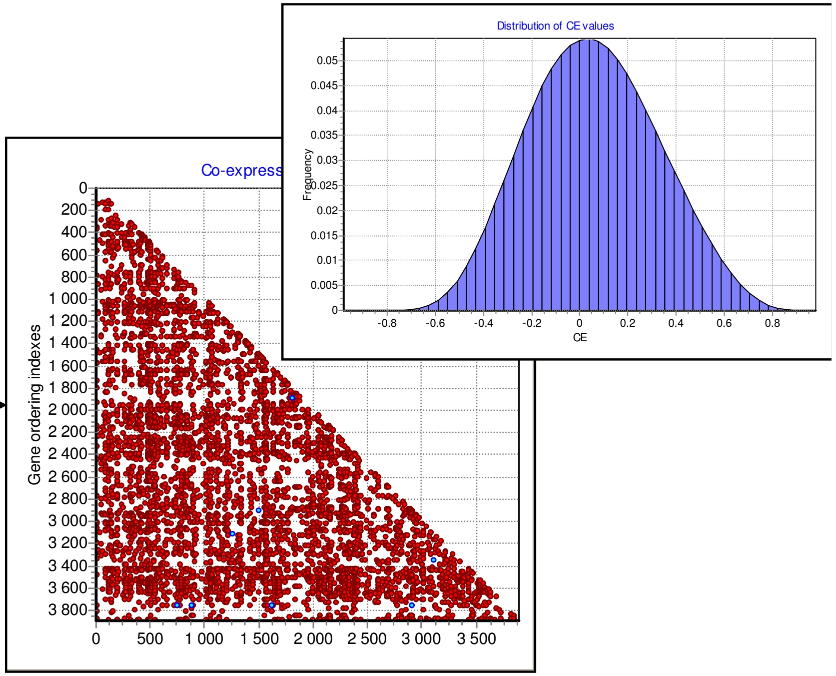
1354 MA
3893 genes

CoExpress:
Pearson correlation

$|r| > 0.8$

Example: CDC2 – BIRC5, $r = 0.86$

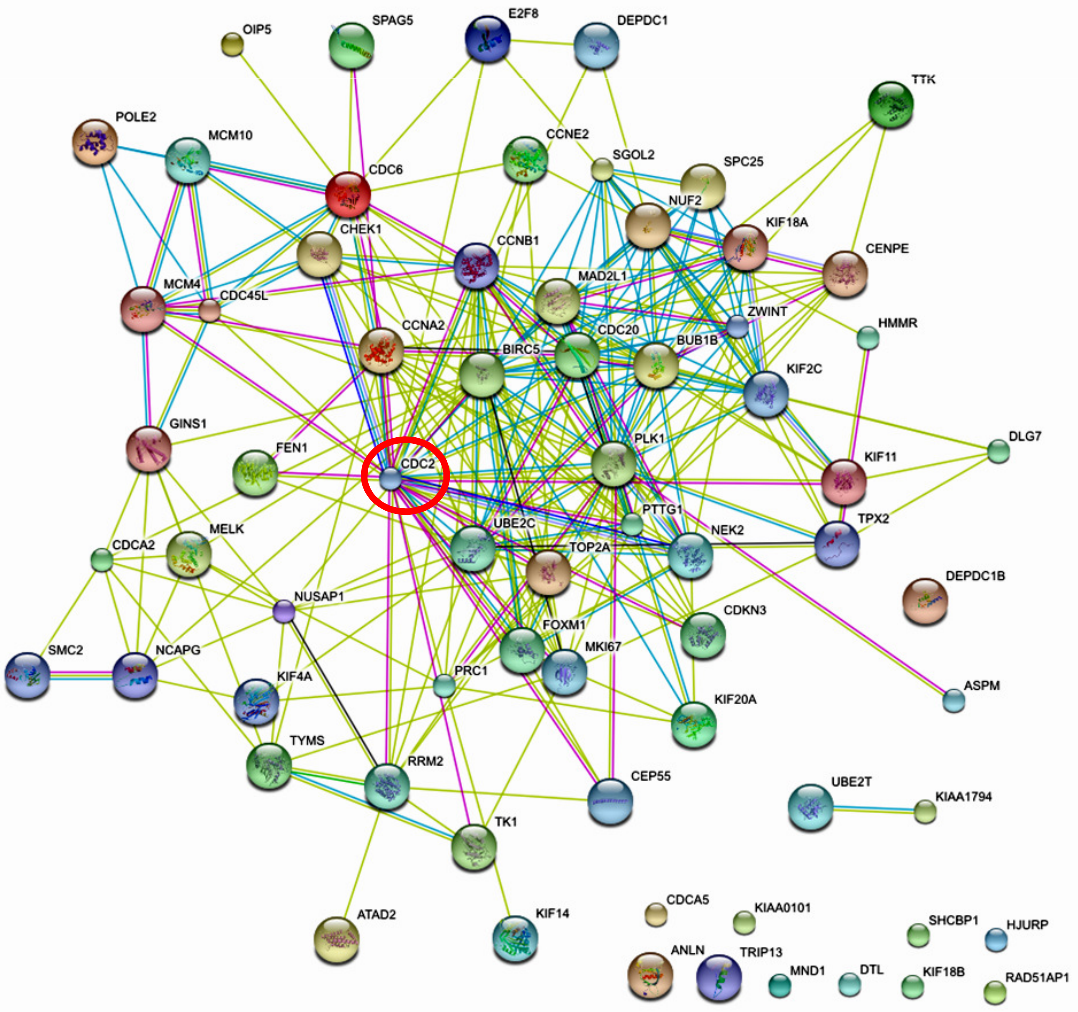




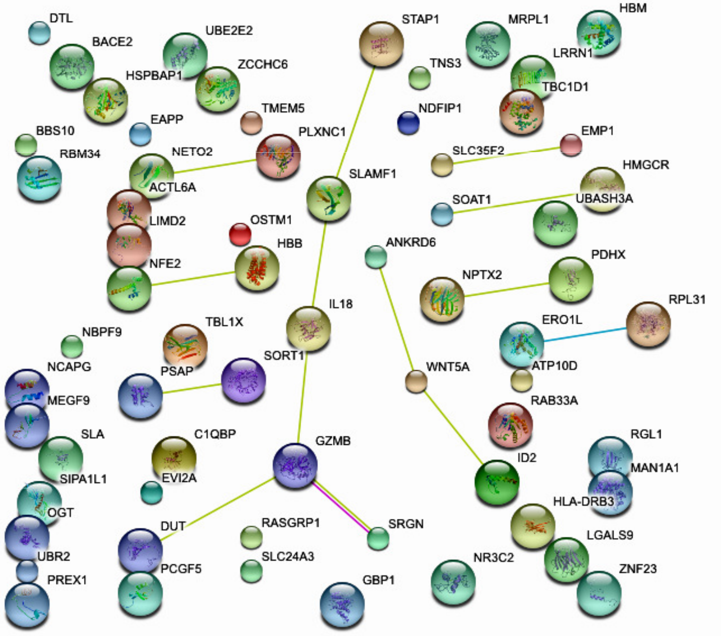
ANALYSIS OF THE PUBLIC DATA

Validation with STRING 8.3 (string-db.org)

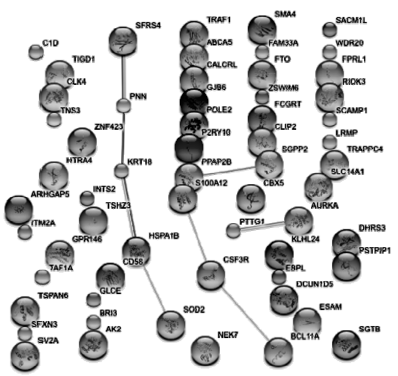
69 genes co-expressed genes (top subnetwork)



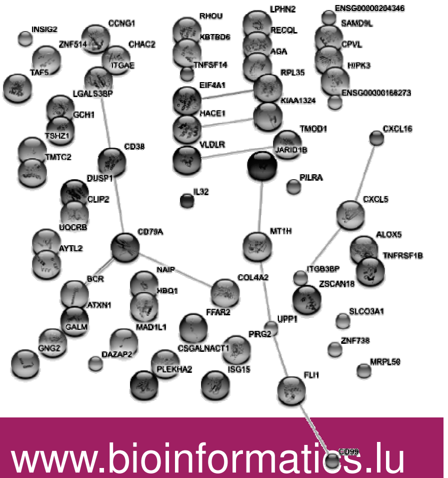
69 random genes



2nd set of random genes



and 3rd set

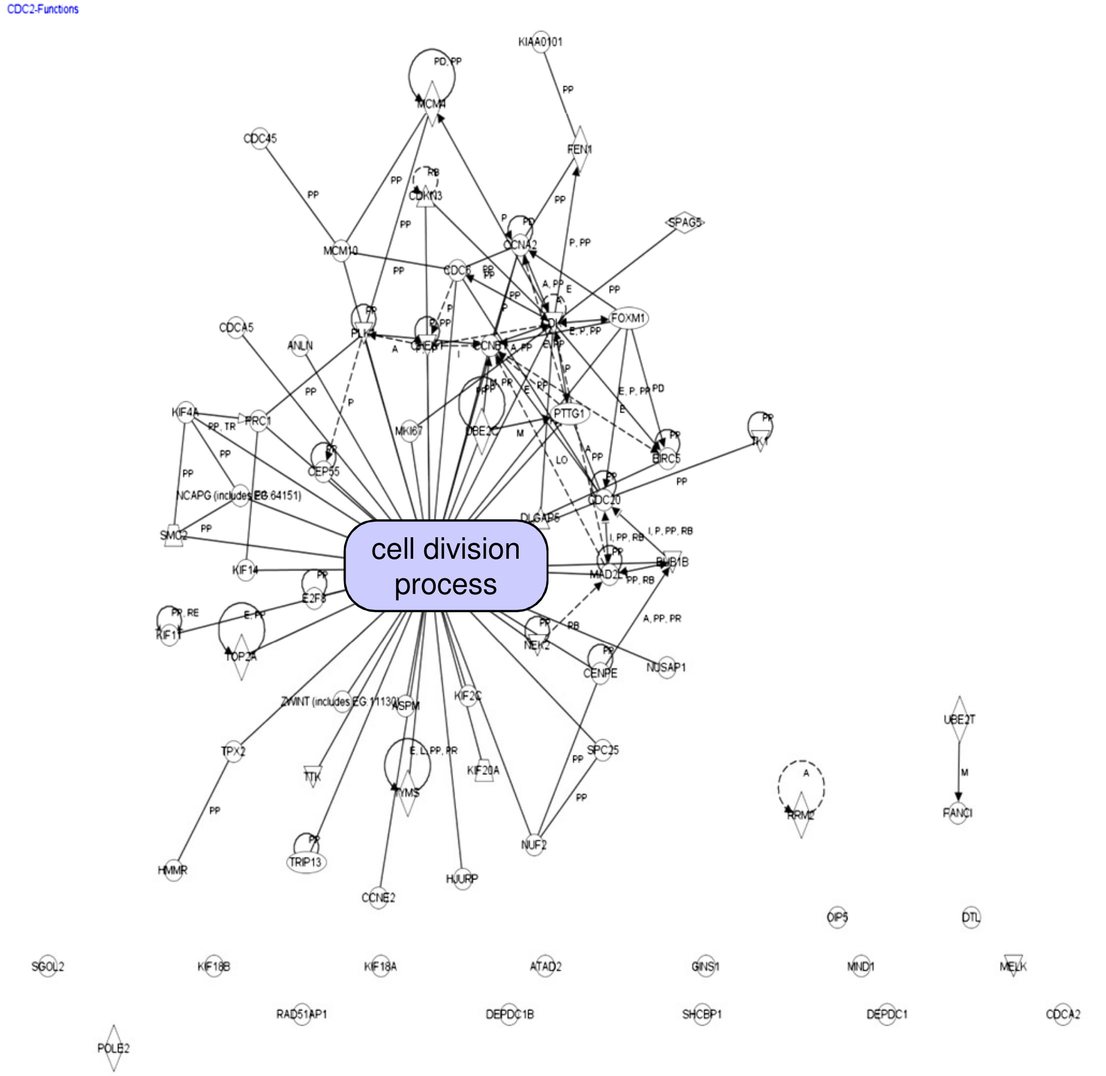


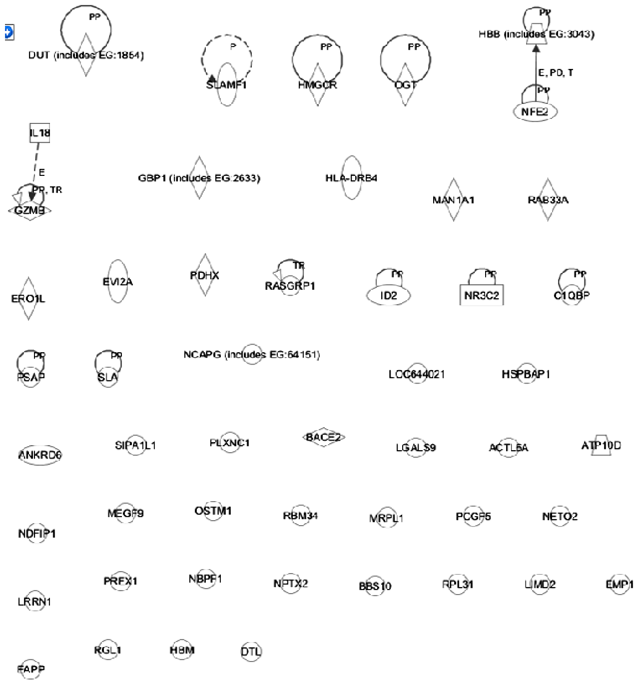
ANALYSIS OF THE PUBLIC DATA

Validation with IPA®

69 genes co-expressed genes (top subnetwork)

69 random genes





Ingenuity Pathway Analysis

ANALYSIS OF THE PUBLIC DATA

Statistical Validation

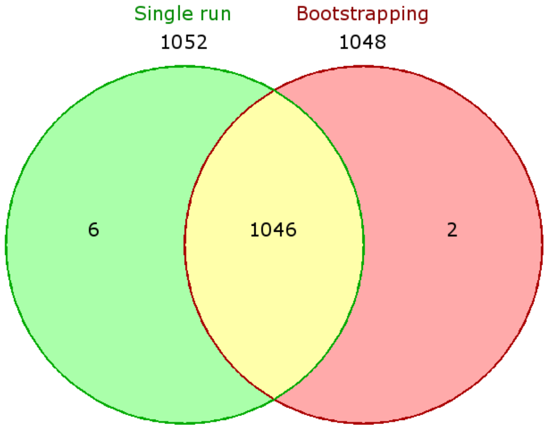
Bootstrapping

- ➢ 10% of the randomly selected microarrays are removed at each iteration.
- ➢ At iteration CE events with $r < 0.7$ are penalized (set to zero)
- ➢ Only CE events with average $| r | > 0.8$ are taken.

Permutation-based FDR

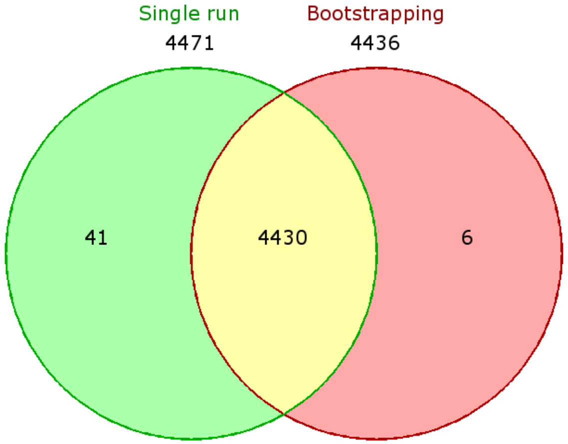
- ➢ Values for each gene are permutated: assigned to random microarrays.
- ➢ FDR can be calculated from average number of genes and CE events detected for such permutated data.

Genes (nodes)

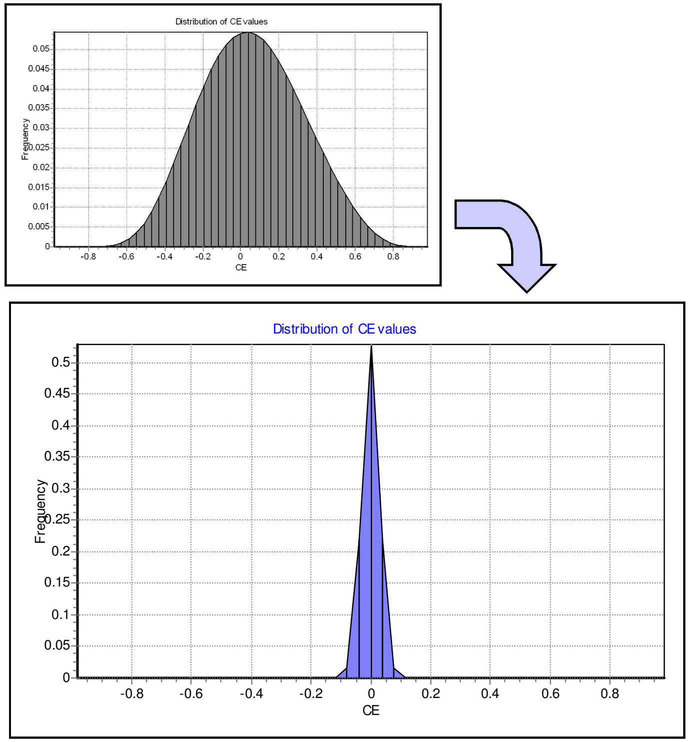


Lists contain 1054 unique elements

CE events (links/edges)



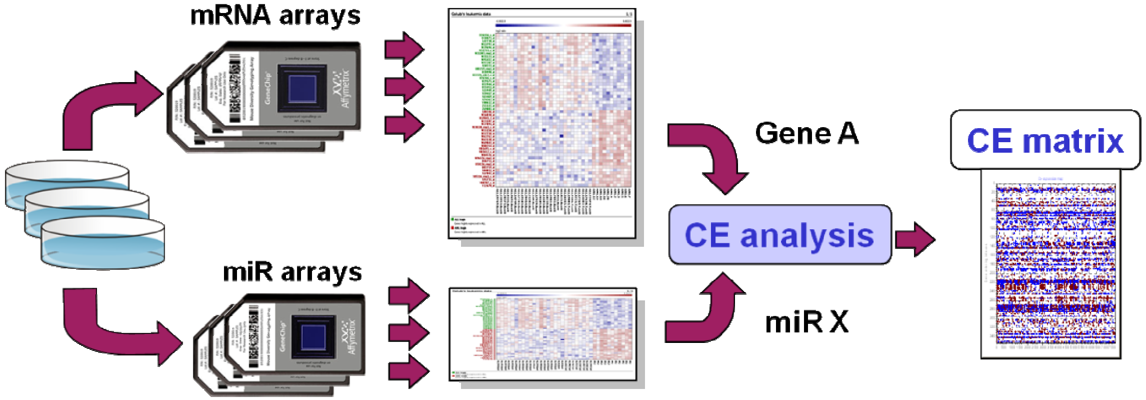
Lists contain 4477 unique elements



Application II

miR – mRNA

Demonstration of CoExpress on our Results



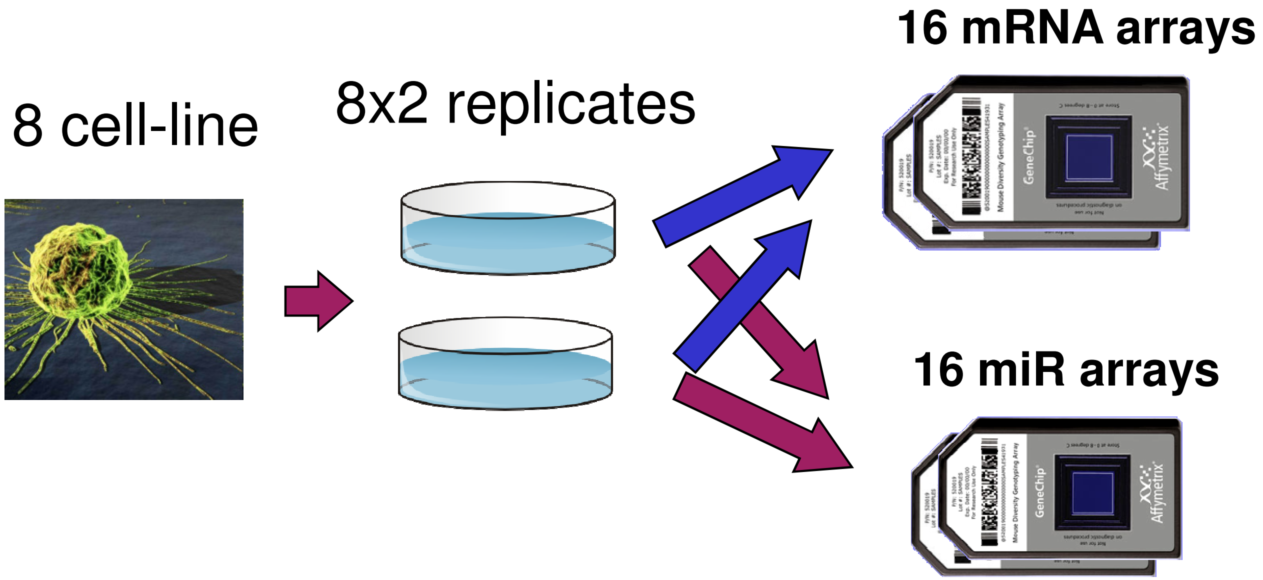
miR – mRNA CO-REGULATION ANALYSIS

Data Coming from 8 Cell-Lines

We have selected 8 cell-lines:

A375, FM55M1, FM55P, HL60, HTC15, IGR37, IGR39, MCF7

and take 2 samples (biological replicates) for each resulting in 16 samples in total



8 cell-line

8x2 replicates

16 mRNA arrays

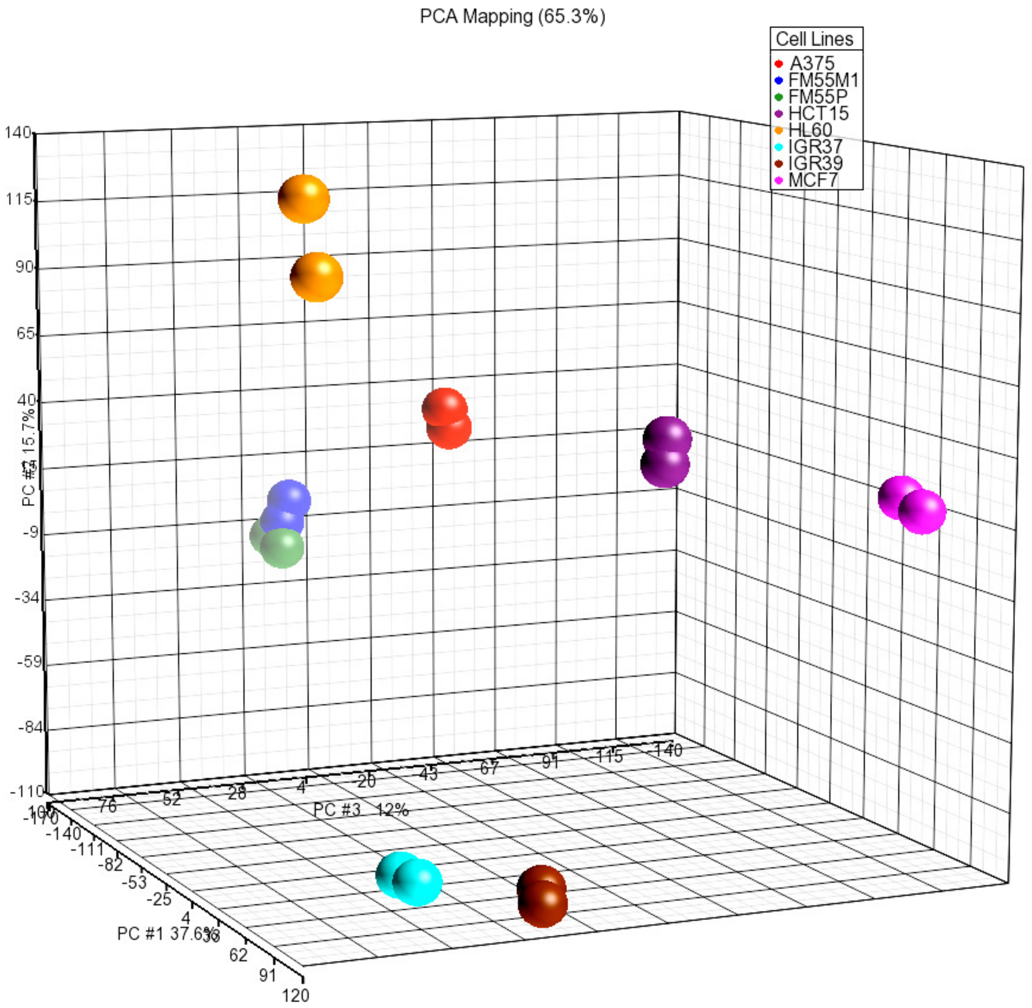
16 miR arrays

Samples obtained from S.Kreis, microarray experiments performed by N.Nicot at Microarray Center, CRP-Sante

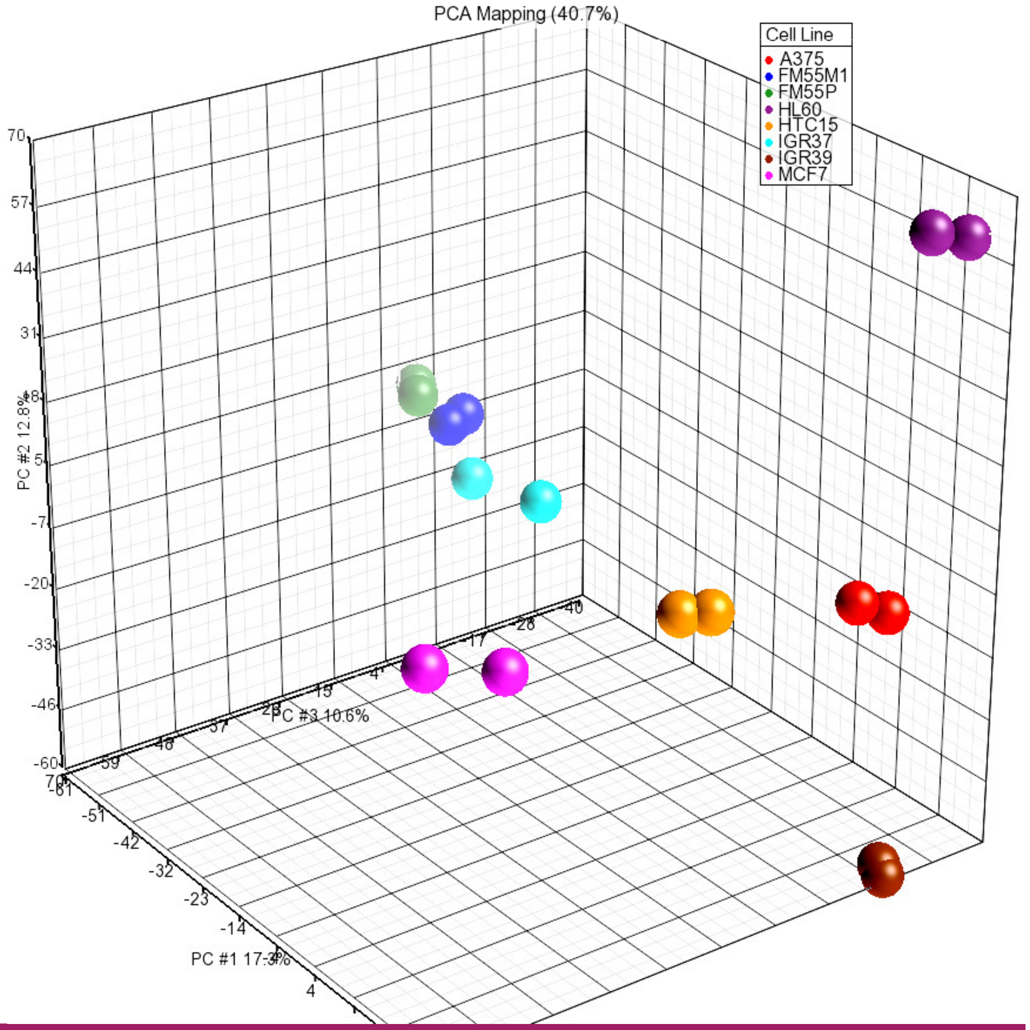
miR – mRNA CO-REGULATION ANALYSIS

Data Coming from 8 Cell-Lines

mRNA arrays



miRNA arrays



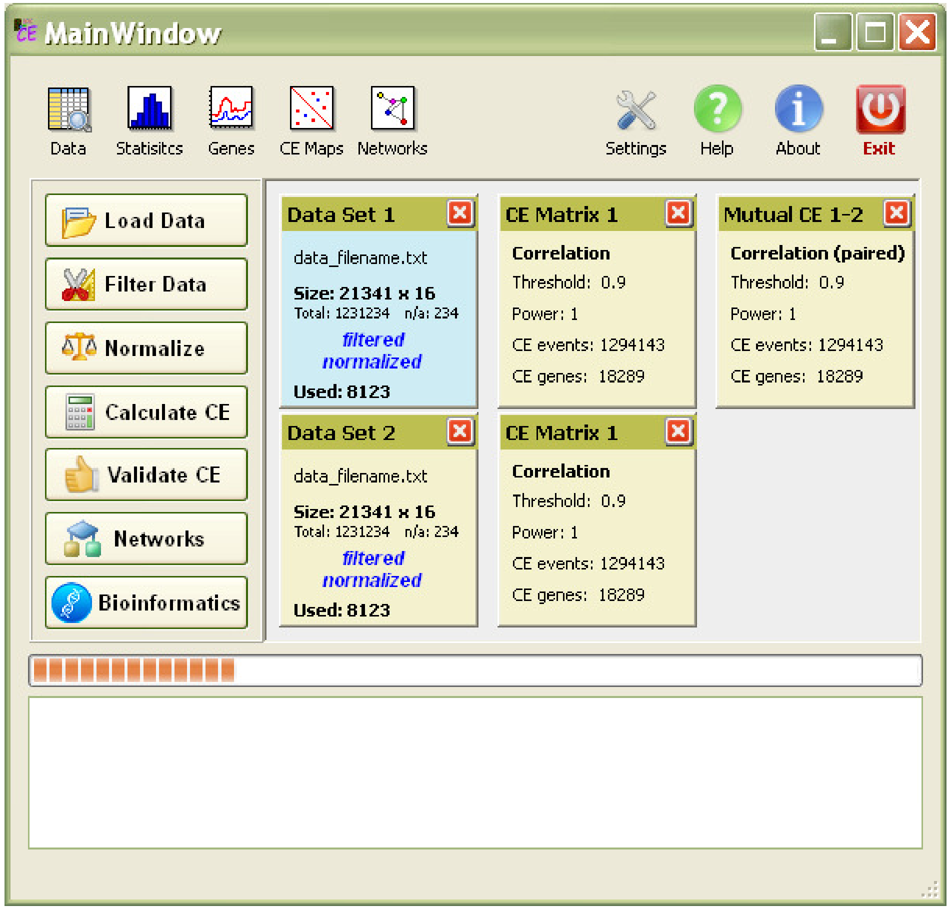
DEMONSTRATION

- **Load and preprocess the data for 8 cell-lines**
  - table and statistics
  - filtering
  - normalization and R-based preprocessing
- **Gene view**
  - miR-miR, mRNA-mRNA
  - miR-mRNA: mir-182 and MITF (Xu, Biol.Chem., 2007)
- **Calculate CE**
  - miR-miR, mRNA-mRNA (start)
  - miR-mRNA
  - data export to text and Cytoscape
- **CE map viewer**
  - selection the top co-expressed genes
  - manual check with gene view
- **Comparison with TargetScan**
  - total number of proved predictions
  - mir-by-mir search

Run CoExpress

FUTURE PLANS

- **Finalize manual and get a publication ☺**
- **Develop a cross-platform version**
  - use **Qt SDK** (Nokia) →
  - use **pthread** library for multi-CPU calculations in Win32 *(done in command-line version)*
  - use **boost** library to speed up *(done in command-line version)*
- **Enhance the network analysis part (topology analysis)**
- **Introduce more advanced bioinformatical analysis**
  - interaction with public DB
  - miR-gene-TF networks



Data Set 1
data_filename.txt
Size: 21341 x 16
Total: 1231234 n/a: 234
filtered
normalized
Used: 8123

CE Matrix 1
Correlation
Threshold: 0.9
Power: 1
CE events: 1294143
CE genes: 18289

Mutual CE 1-2
Correlation (paired)
Threshold: 0.9
Power: 1
CE events: 1294143
CE genes: 18289

Data Set 2
data_filename.txt
Size: 21341 x 16
Total: 1231234 n/a: 234
filtered
normalized
Used: 8123

CE Matrix 1
Correlation
Threshold: 0.9
Power: 1
CE events: 1294143
CE genes: 18289

Interface of the currently developed version (in Qt SDK)

`http://bioinformatics.lu/CoExpress`

ACKNOWLEGEMENT

CENTRE DE RECHERCHE PUBLIC

Microarray Center

Arnaud Muller

Tony Kaoma

Nathalie Nicot

Christelle Ghoneim

Francois Bernardin

Laurent Vallar

UNIVERSITÉ DU LUXEMBOURG

Demetra Philippidou

Martina Schmitt

Stephanie Schmitz

Christiane Margue

Stephanie Kreis

Viktar Khutko, ing. programmer