# RECONOCIMIENTO DE REDES DE GENES MEDIANTE REGRESIÓN

Isabel A. Nepomuceno Chamorro

Master Bioinformática

# Outline

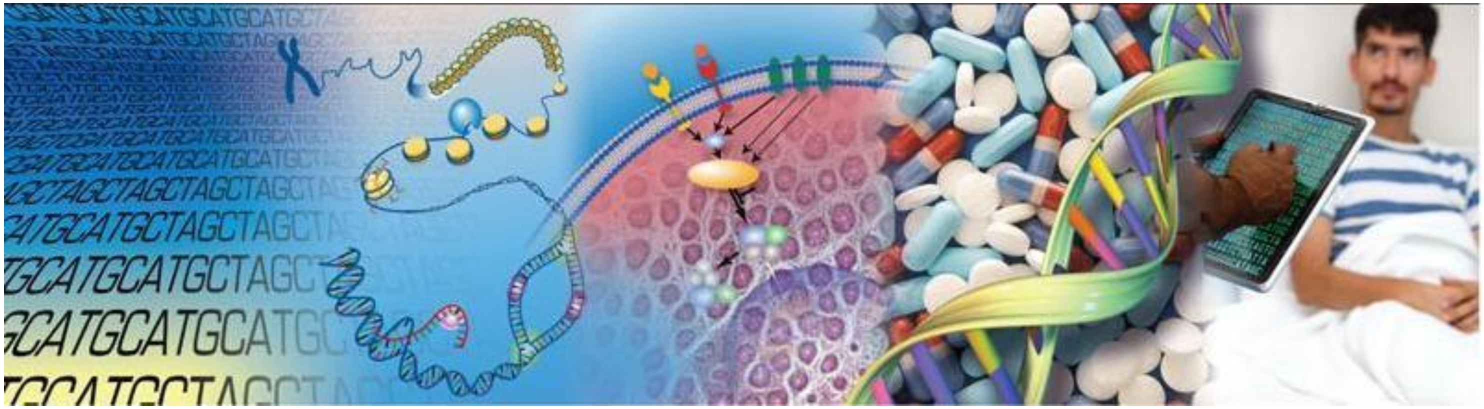
1. Introduction
2. Gene Network Inference: state-of-the art
3. Proposals & Results:
    a) RegNet: Gene Regression Networks
    b) SATuRNo: Supervised Prognostic Approach Through Regression Networks
    c) CarGene: Characterization of genes
4. Conclusions and future work
5. Project membership

# Outline

1. Introduction
2. Gene Network Inference: state-of-the art
3. Proposals & Results:
   a) RegNet: Gene Regression Networks
   b) SATuRNo: Supervised Prognostic Approach Through Regression Networks
   c) CarGene: Characterization of genes
4. Conclusions and future work
5. Project membership

# Motivation

- A cross-cutting area broadly relevant across genomic and genomic medicine:
  - Bioinformatics and Computational Biology



Understanding → the structure of genomes → the biology of genomes → the biology of disease

Advancing the science of medicine

Improving the effectiveness of healthcare

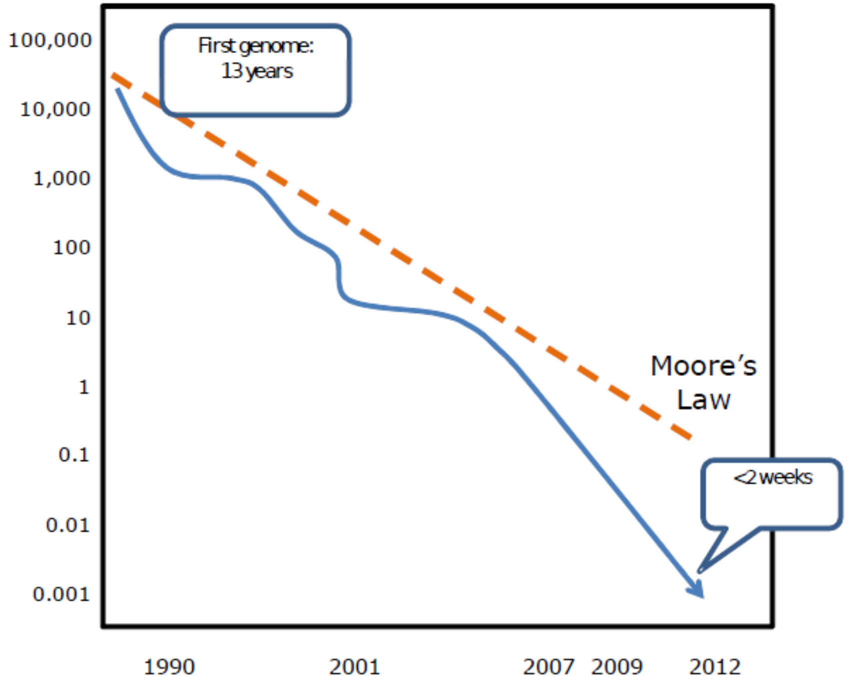
# Motivation

*"The major bottleneck in genome understanding is no longer data generation—the computational challenges around **data analysis**, display and integration are now rate limiting. New approaches and methods are required to meet these challenges"*

*(Green et al., Nature 2011)*

# Motivation

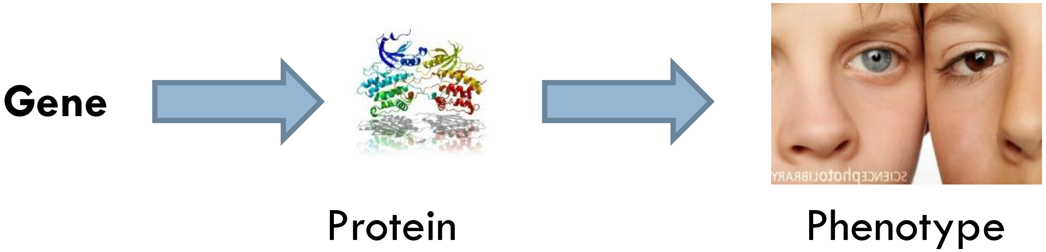
- The cost goes down, while the amount of data to manage and its complexity raise exponentially.
- As for example: generation sequencing technologies



*Dopazo: International Course of Massive Data Analysis*

# Reduccionistic vs. Holistic

- From genome to phenotype
  - Reduccionistic (Pre-genomic paradigm)

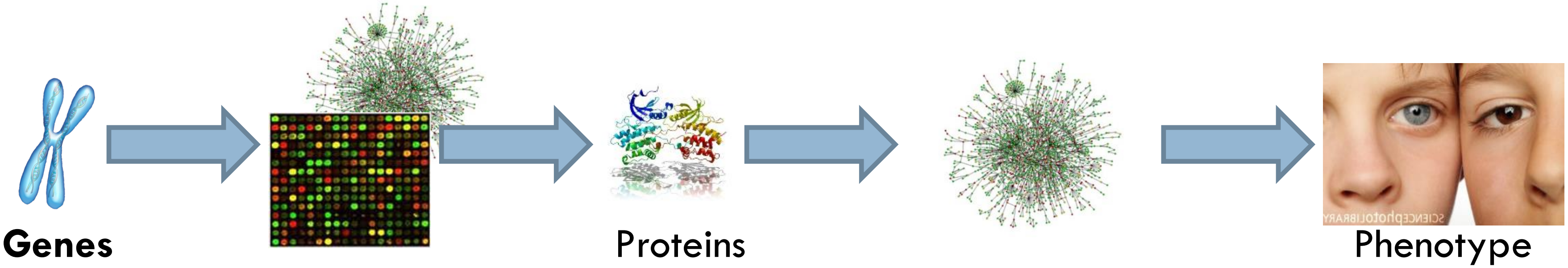


**Gene**

Protein

Phenotype

  - Holistic



**Genes**

Proteins

Phenotype

# Microarray Technology

- Los **microarrays o micromatrices** son una de las tecnologías de altas prestaciones más populares que permiten medir el nivel de expresión de todos los genes de un genoma en una muestra estimando el número de transcriptos (mRNA) de cada gen.

# Microarray Technology

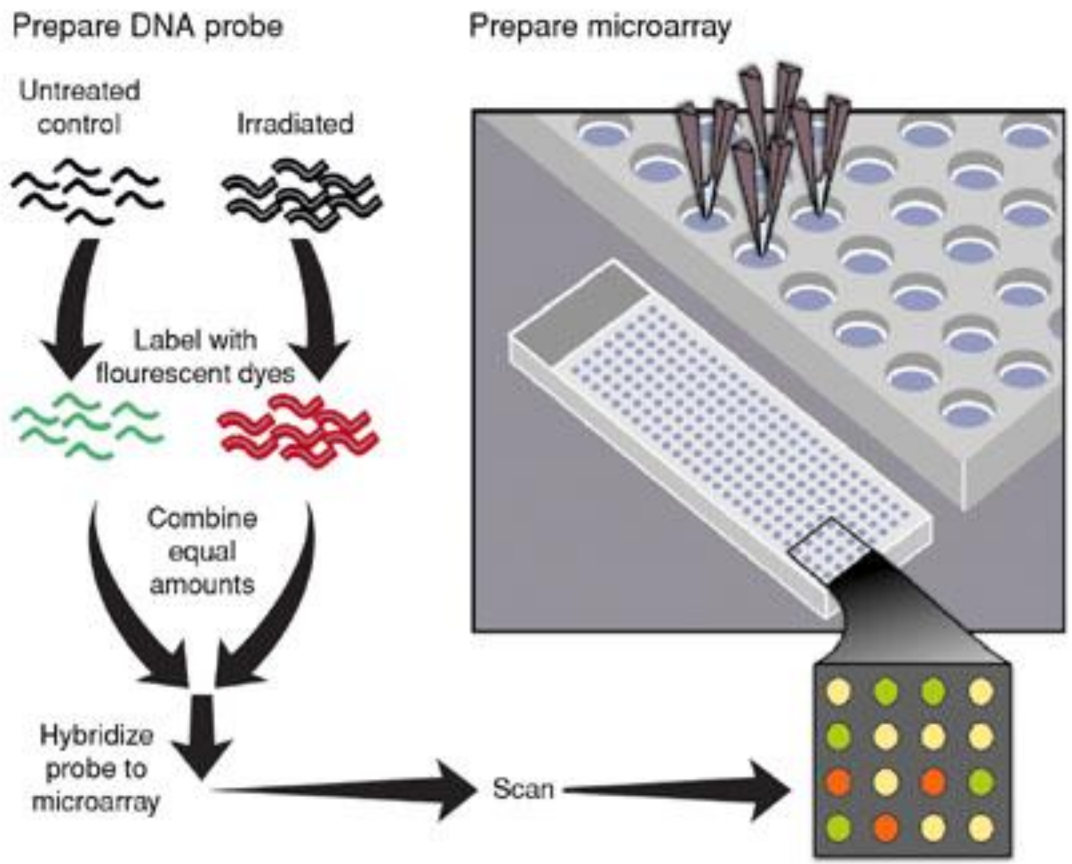
- Un **microarray** es un soporte sólido donde se disponen en forma de matriz (en una distribución regular de filas y columnas) secuencias de DNA llamadas sondas (probes) que son complementarias a las secuencias de los transcritos conocidos en una especie en particular.

# Microarray Technology

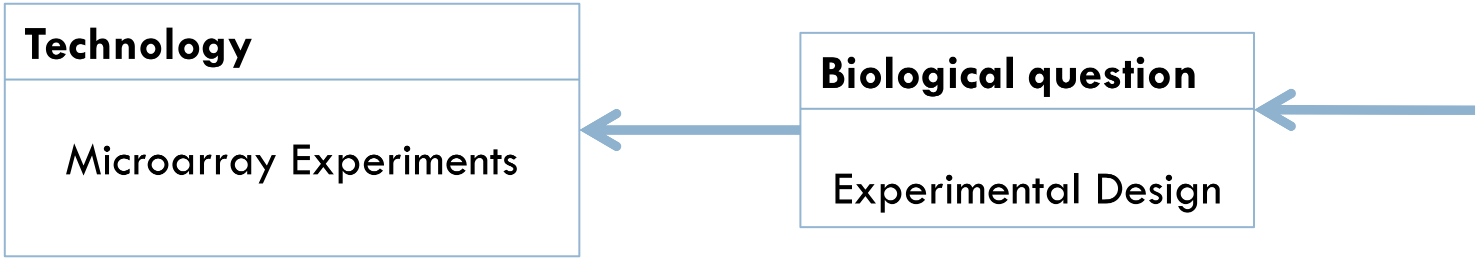
- Uso de microarray:
  - Comparar muestras de interés:
    - muestras de pacientes con un fenotipo vs muestras control.
    - Pacientes con un fenotipo y un tratamiento clínico vs pacientes del mismo fenotipo y sin dicho tratamiento
    - Etc.

# Microarray Technology

- Study of thousands of genes simultaneously



# Microarray Analysis



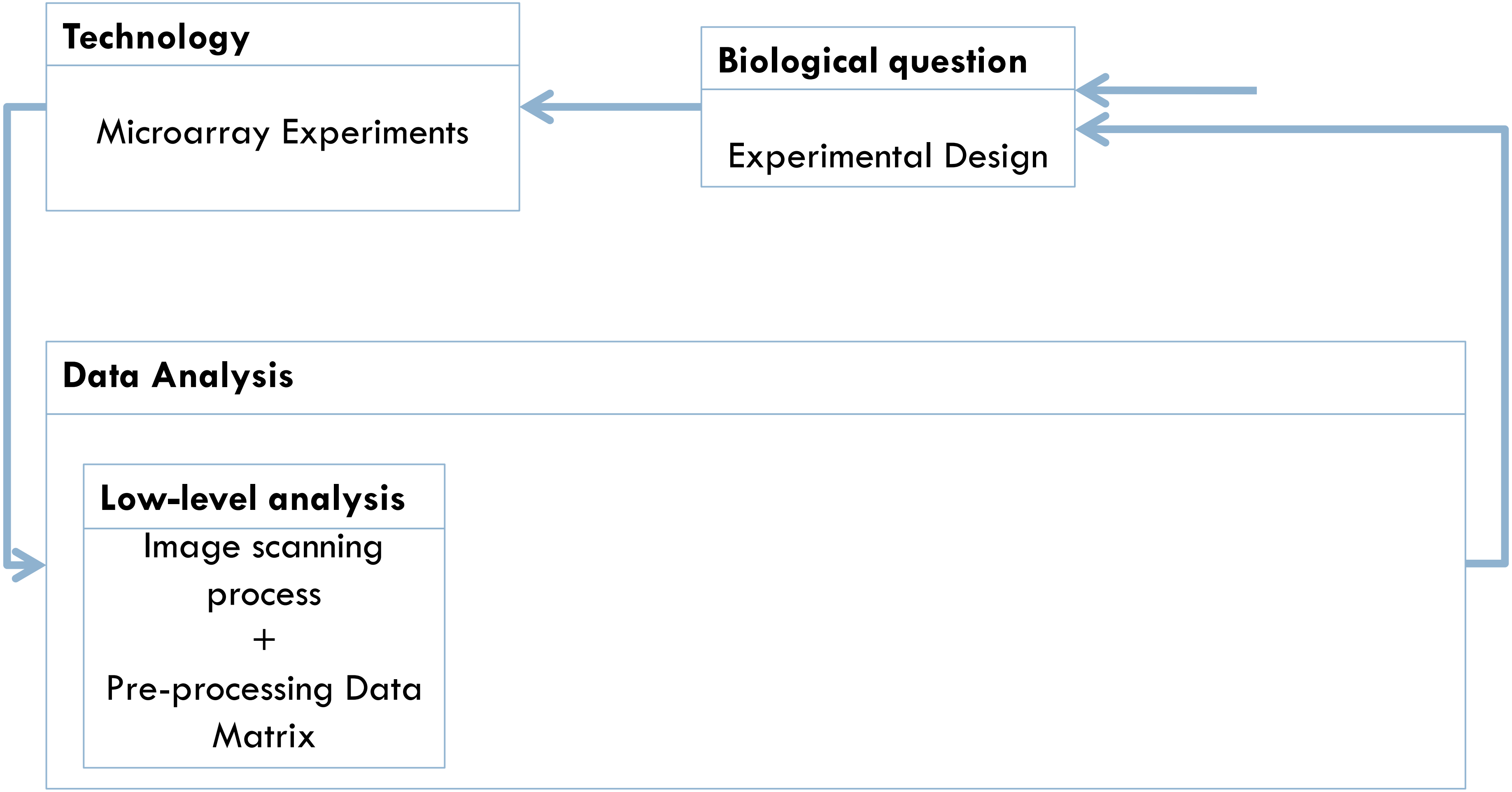
**Technology**

Microarray Experiments

**Biological question**

Experimental Design

# Microarray Analysis



**Technology**

Microarray Experiments

**Biological question**

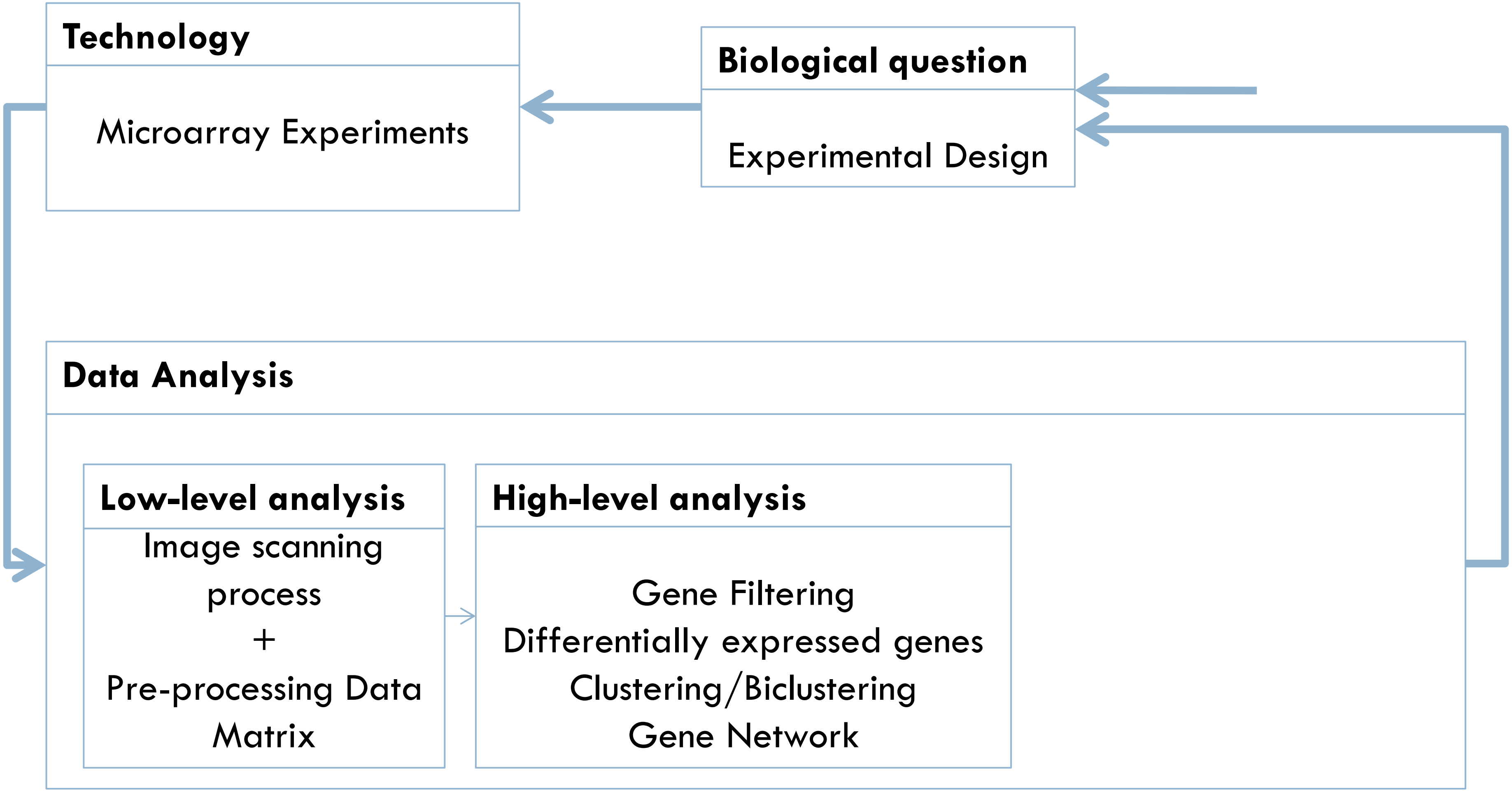
Experimental Design

**Data Analysis**

**Low-level analysis**

Image scanning process
+
Pre-processing Data Matrix

# Microarray Analysis



**Technology**

Microarray Experiments

**Biological question**

Experimental Design

**Data Analysis**

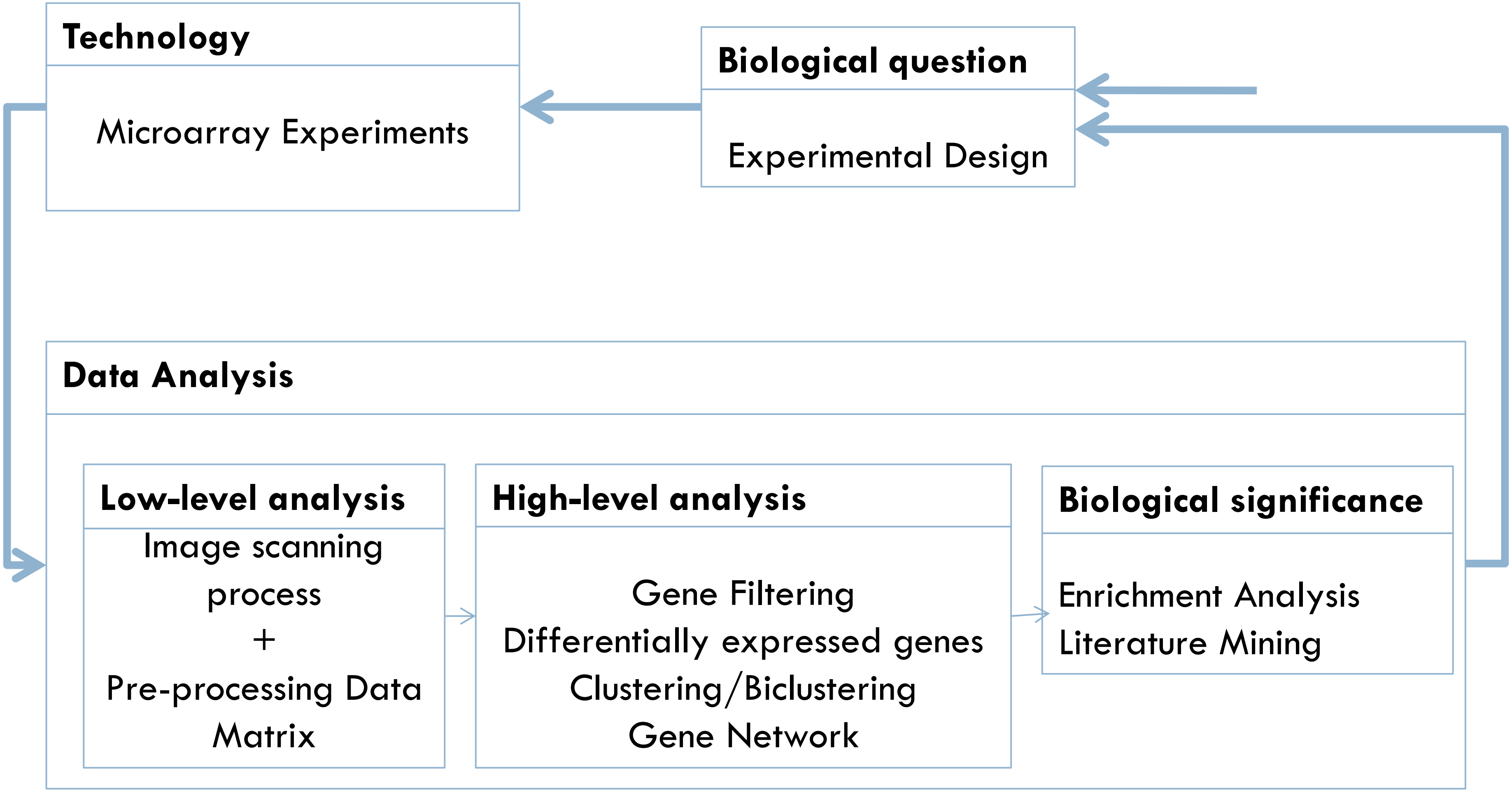
**Low-level analysis**

Image scanning process
+
Pre-processing Data Matrix

**High-level analysis**

Gene Filtering
Differentially expressed genes
Clustering/Biclustering
Gene Network

# Microarray Analysis



**Technology**

Microarray Experiments

**Biological question**

Experimental Design

**Data Analysis**

**Low-level analysis**

Image scanning process
+
Pre-processing Data Matrix

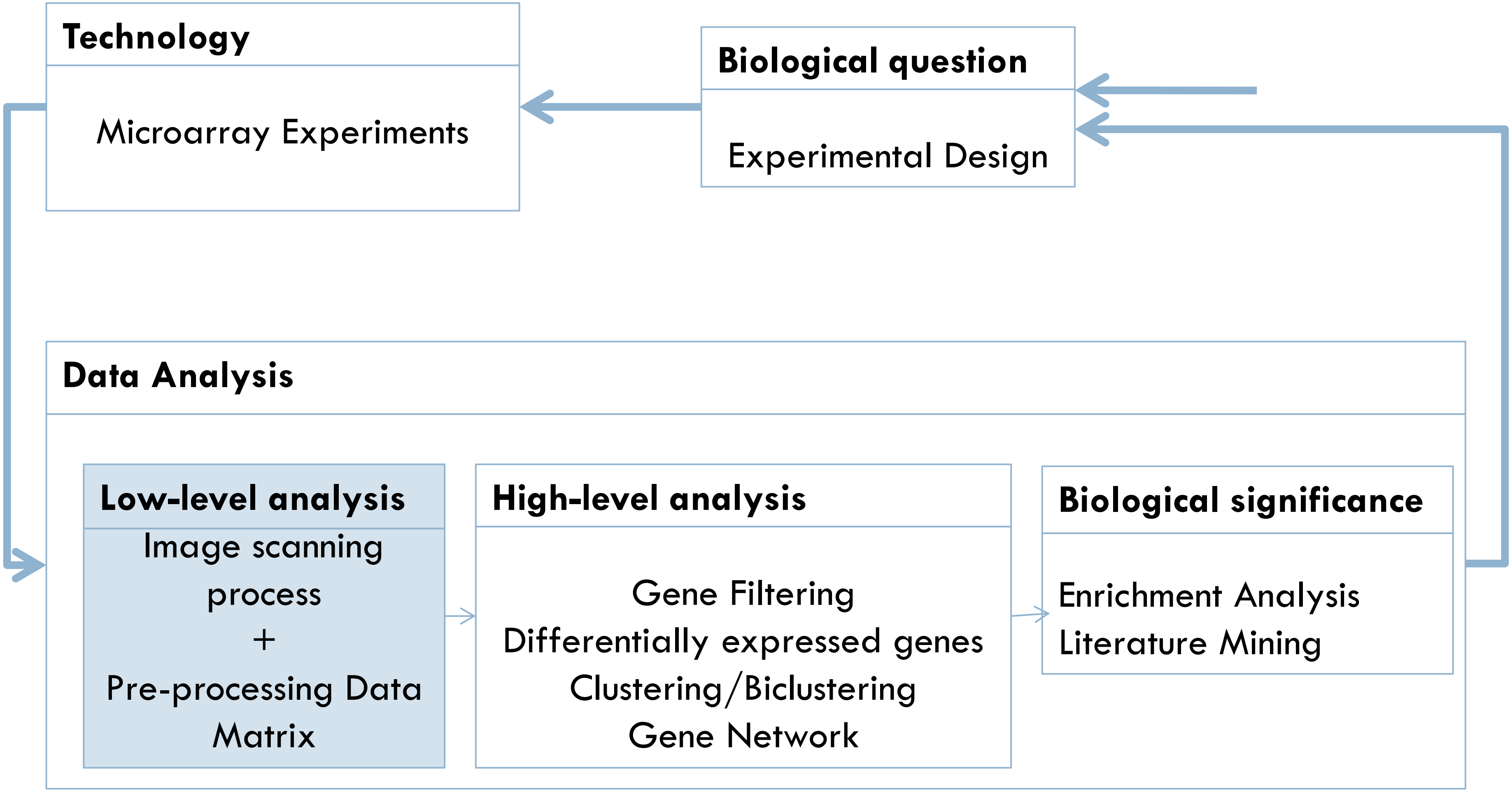
**High-level analysis**

Gene Filtering
Differentially expressed genes
Clustering/Biclustering
Gene Network

**Biological significance**

Enrichment Analysis
Literature Mining

# Microarray Analysis



**Technology**

Microarray Experiments

**Biological question**

Experimental Design

**Data Analysis**

**Low-level analysis**

Image scanning process
+
Pre-processing Data Matrix

**High-level analysis**

Gene Filtering
Differentially expressed genes
Clustering/Biclustering
Gene Network

**Biological significance**

Enrichment Analysis
Literature Mining

# Low-level analysis

- Image scanning process
- Pre-processing Data Matrix

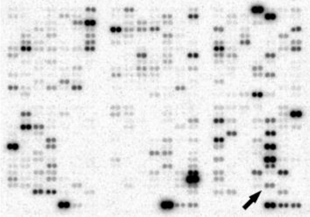
# Low-level analysis

- **Image scanning process**
- Pre-processing Data Matrix

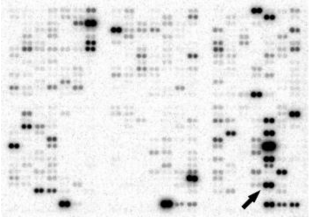
# Low-level analysis

- Proceso de escaneado de imágenes
  - Input: niveles de fluorescencia (indicador del nivel de expresión génica) o datos crudos

Muestra 1



Muestra n



  - Output: matriz de datos

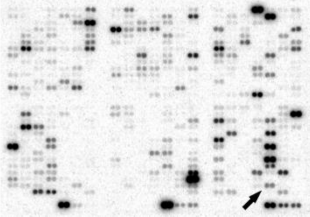
**Genes**

| | Gene1 | Gene2 | Gene3 … |
|---|---|---|---|
| **Muestra1** | 4,19 | 4,48 | 3,83 |
| **Muestra2** | 4,72 | 4,93 | 4,53 |
| **…** | 5,56 | 6,27 | 5,16 |

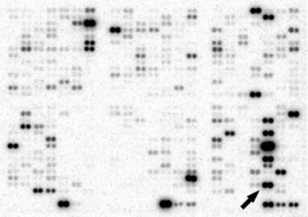
# Low-level analysis

- Proceso de escaneado de imágenes
  - Input: niveles de fluorescencia (indicador del nivel de expresión génica) o datos crudos

Muestra 1



Muestra n



  - Output: matriz de datos

**Genes**

| | Gene1 | Gene2 | Gene3 … |
|---|---|---|---|
| **Muestra1** | 4,19 | 4,48 | 3,83 |
| **Muestra2** | 4,72 | 4,93 | 4,53 |
| **…** | 5,56 | 6,27 | 5,16 |

  - How to:

BABELOMICS 4
gene expression and functional profiling analysis suite

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

# Low-level analysis

- Image scanning process
- **Pre-processing Data Matrix**

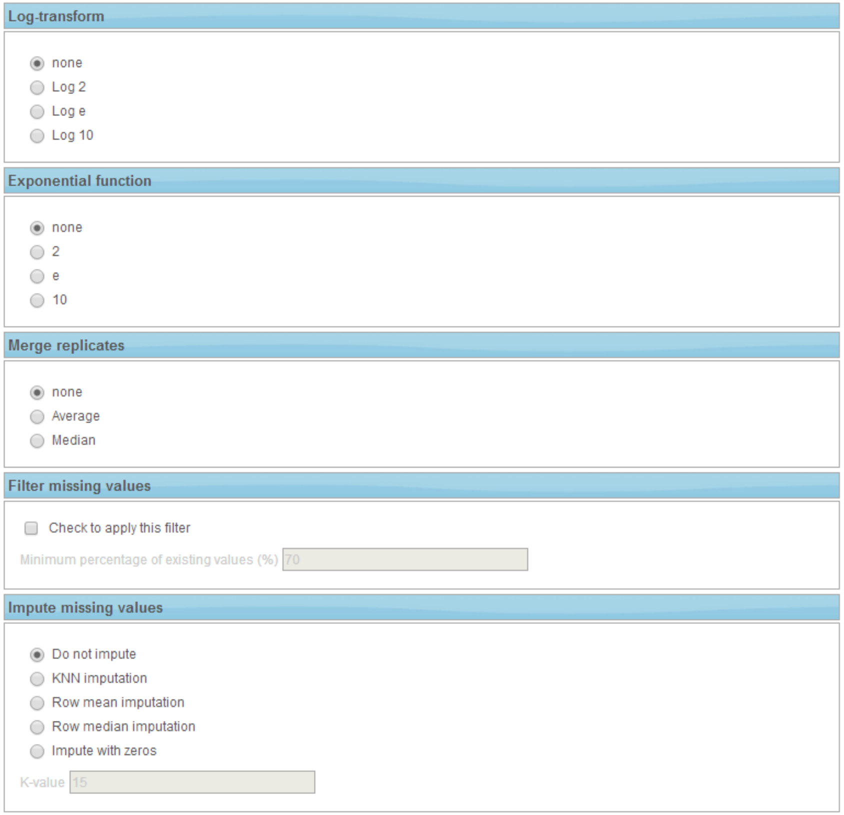
# Low-level analysis

- Preprocesado de la matriz de datos
  - Objetivo: realizar transformaciones básicas
    - Modificar nomenclatura de nombre de genes
    - Merge replicates: mean, median
    - Tratar valores perdidos: mean imputation, median imputation, input with zeros, …
    - Transformación logarítmica
    - …

# Low-level analysis

- Preprocesado de la matriz de datos
  - How to:

BABELOMICS 4
gene expression and functional profiling analysis suite



**Log-transform**

- ● none
- ○ Log 2
- ○ Log e
- ○ Log 10

**Exponential function**

- ● none
- ○ 2
- ○ e
- ○ 10

**Merge replicates**

- ● none
- ○ Average
- ○ Median

**Filter missing values**

☐ Check to apply this filter

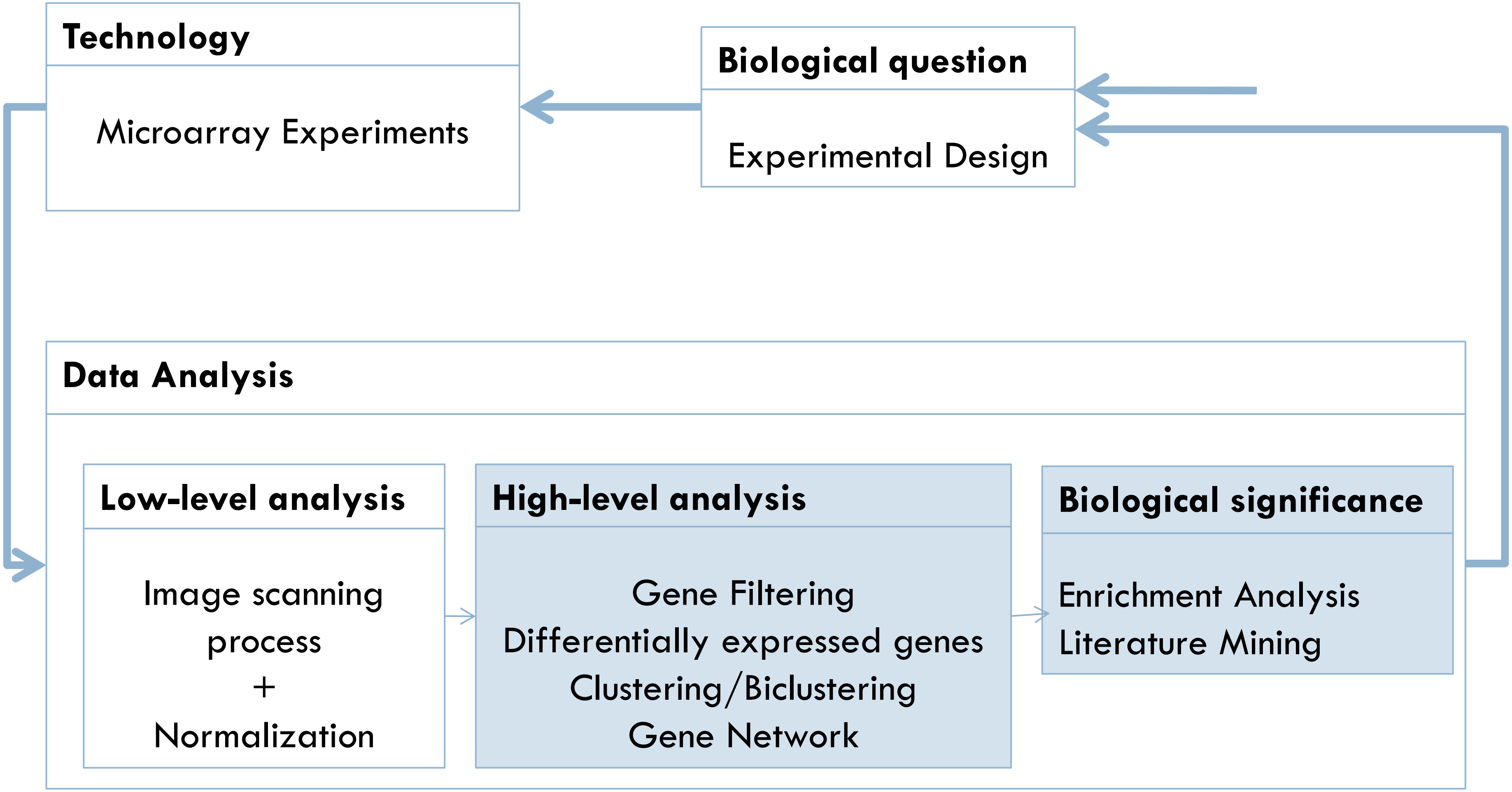
Minimum percentage of existing values (%) 70

**Impute missing values**

- ● Do not impute
- ○ KNN imputation
- ○ Row mean imputation
- ○ Row median imputation
- ○ Impute with zeros

K-value 15

# Microarray Analysis



**Technology**

Microarray Experiments

**Biological question**

Experimental Design

**Data Analysis**

**Low-level analysis**

Image scanning process
+
Normalization

**High-level analysis**

Gene Filtering
Differentially expressed genes
Clustering/Biclustering
Gene Network

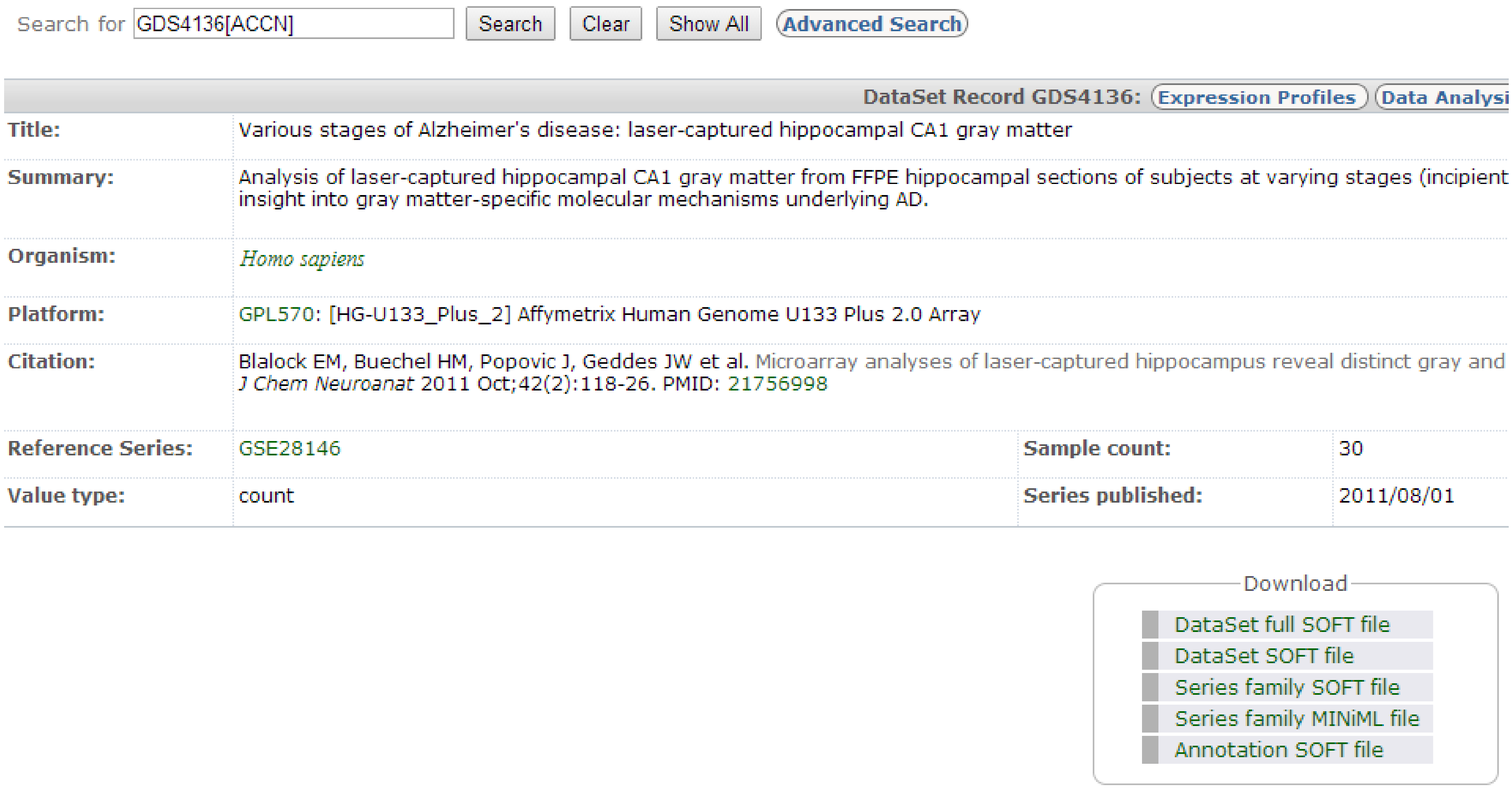
**Biological significance**

Enrichment Analysis
Literature Mining

# How to start

Gene Expression Omnibus

- Repositorio de datos de Microarray públicos:



Search for GDS4136[ACCN]

**DataSet Record GDS4136:**

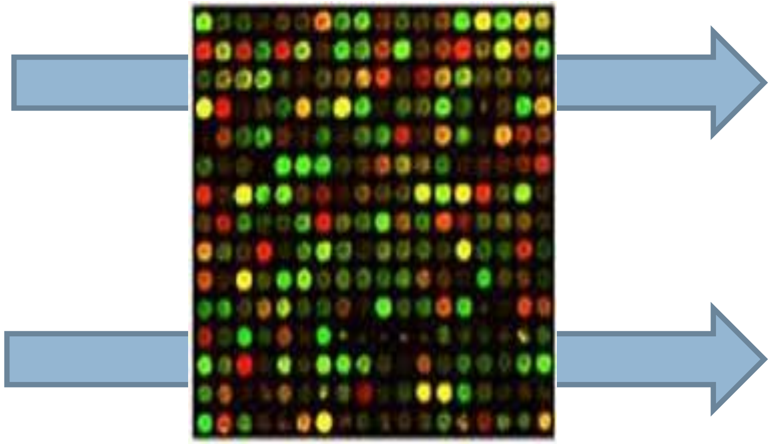
| | | | |
|---|---|---|---|
| **Title:** | Various stages of Alzheimer's disease: laser-captured hippocampal CA1 gray matter | | |
| **Summary:** | Analysis of laser-captured hippocampal CA1 gray matter from FFPE hippocampal sections of subjects at varying stages (incipient<br>insight into gray matter-specific molecular mechanisms underlying AD. | | |
| **Organism:** | *Homo sapiens* | | |
| **Platform:** | GPL570: [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array | | |
| **Citation:** | Blalock EM, Buechel HM, Popovic J, Geddes JW et al. Microarray analyses of laser-captured hippocampus reveal distinct gray and<br>*J Chem Neuroanat* 2011 Oct;42(2):118-26. PMID: 21756998 | | |
| **Reference Series:** | GSE28146 | **Sample count:** | 30 |
| **Value type:** | count | **Series published:** | 2011/08/01 |

Download

- DataSet full SOFT file
- DataSet SOFT file
- Series family SOFT file
- Series family MINiML file
- Annotation SOFT file

# Our Aim

- <u>Our aim</u>: in High Level Analysis



Coexpressing genes

Classification of samples

What do genes have in common?
How can genes interact?

What genes are responsible for?

Biological validation or significance

# Objectives

1. RegNet
   - Regression Networks
2. SATuRNo
   - Supervised Approach Through Regression Networks
3. CarGene
   - Characterization of set of genes

# Outline

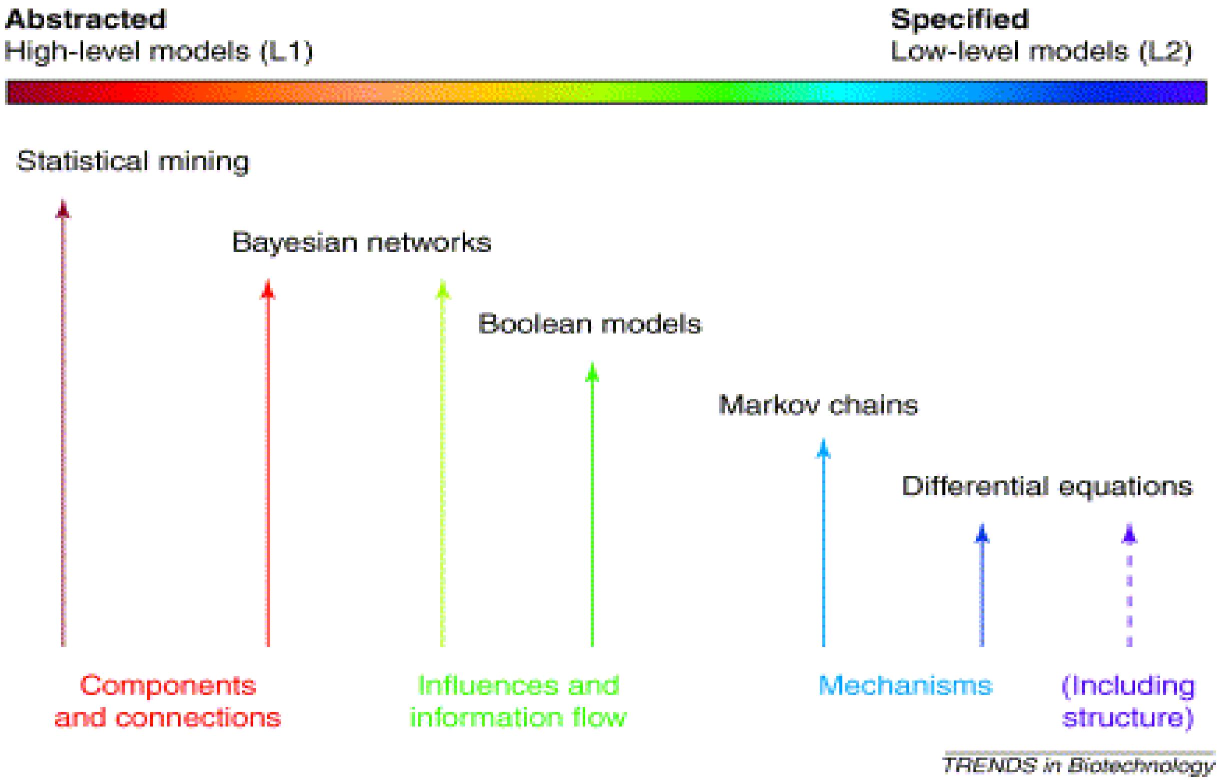
1. Introduction
2. Gene Network Inference: state-of-the art
   a) Classification methods
   b) We focus on…
   c) Conditional Dependence Models
3. Proposals & Results:
   a) RegNet: Gene Regression Networks
   b) SATuRNo: Supervised Prognostic Approach Through Regression Networks
   c) CarGene: Characterization of genes
4. Conclusions and future work
5. Project membership

# Classification methods I

I. Based on the level of abstraction
    I. (Ideker and Lauffenburger 2003)
II. Based on two general strategies
    I. (Gardner and Faith 2005)

# Classification methods I

I. Based on the level of abstraction (Ideker and Lauffenburger 2003)



*TRENDS in Biotechnology*

# Classification methods II

I. Based on two general strategies (Gardner et al. 2005)

**Physical modeling**

- to identify the molecules that physically control RNA synthesis

**Influence modeling**

- to model relationships between RNA transcripts

# Classification methods II

I. Based on two general strategies (Gardner et al. 2005)

**Physical modeling** ✚ **G. Sequence Data**

- to identify the molecules that physically control RNA synthesis

**Influence modeling**

- to model relationships between RNA transcripts

# We focus on…

- From steady state microarray datasets (non time-course)

High Level Models

Influence modeling strategy

# State of Art: Conditional Independence models

- Models based on corr. or statistical dependence
  - Marginal dependencies: Coexpression Networks

    $$X_i \perp\!\!\!\perp X_j \mid X_z \quad , \quad z = \emptyset$$

  - Full conditional models: Markov networks

    $$X_i \perp\!\!\!\perp X_j \mid X_z \quad , \quad z = Rest\backslash\{i,j\}$$

  - Low-order conditional models

    $$X_i \perp\!\!\!\perp X_j \mid X_z \quad , \quad z = \text{for all } k \in Rest\backslash\{i,j\}$$

  - Bayesian Networks

    $$X_i \perp\!\!\!\perp X_j \mid X_S \quad , \quad S = \text{for all subset} \in Rest\backslash\{i,j\}$$

# State of Art: Conditional Independence models

- Models based on corr. or statis
  - Marginal dependencies: Coexpres Networks

    $$X_i \perp\!\!\!\perp X_j \mid X_z \quad , \quad z = \emptyset$$

    **Global similarity into pairs of genes**

  - Full conditional models: Markov networks

    $$X_i \perp\!\!\!\perp X_j \mid X_z \quad , \quad z = \mathrm{Rest}\setminus\{i,j\}$$

  - Low-order conditional models

    $$X_i \perp\!\!\!\perp X_j \mid X_z \quad , \quad z = \text{for all } k \in \mathrm{Rest}\setminus\{i,j\}$$

  - Bayesian Networks

    $$X_i \perp\!\!\!\perp X_j \mid X_S \quad , \quad S = \text{for all subset} \in \mathrm{Rest}\setminus\{i,j\}$$

# State of Art: Conditional Independence models

- Models based on corr. or statis
  - Marginal dependencies: Coexpres Networks

$$X_i \not\perp X_j \mid X_z \quad , \quad z = \emptyset$$

Global similarity into pairs of genes

# State of Art: Conditional Independence models

- Models based on corr. or statis
  - Marginal dependencies: Coexpressi Networks

$$X_i \perp\!\!\!\perp X_j \mid X_z \quad , \quad z = \emptyset$$

**Global similarity into pairs of genes**

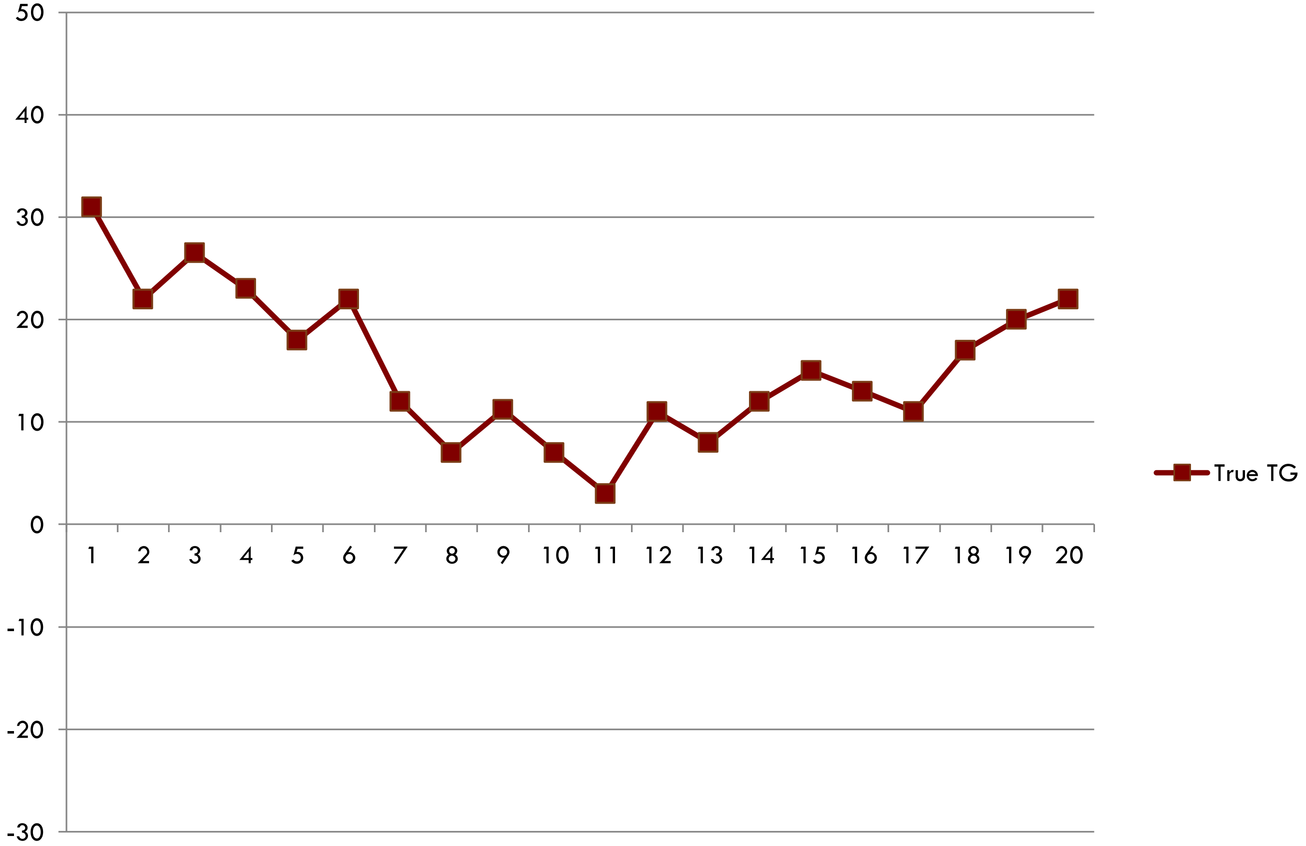
- Drawbacks of Coexpression Networks:
  - Correlation favors global similarity over more localized similarities
  - Many pairs of genes show similar behavior in gene expression profiles by chance even though they are not biologically related

# Drawback of Coexpression Networks

**Genes**

**Patients**

| TG | GR1 | GR2 |
|---|---|---|
| 31 | 0 | 40 |
| 22 | 0 | 30 |
| 27 | 5 | 35 |
| 23 | 5 | 33 |
| 18 | 5 | 26 |
| 22 | 5 | 30 |
| 12 | 2 | 20 |
| 7 | 4 | 15 |
| 11 | 10 | 18 |
| 7 | 12 | 14 |
| 3 | 10 | 10 |
| 11 | 20 | -5 |
| 8 | 15 | 5 |
| 12 | 25 | -10 |
| 15 | 30 | -12 |
| 13 | 25 | -15 |
| 11 | 20 | -10 |
| 17 | 35 | -15 |
| 20 | 38 | -20 |
| 22 | 40 | -25 |

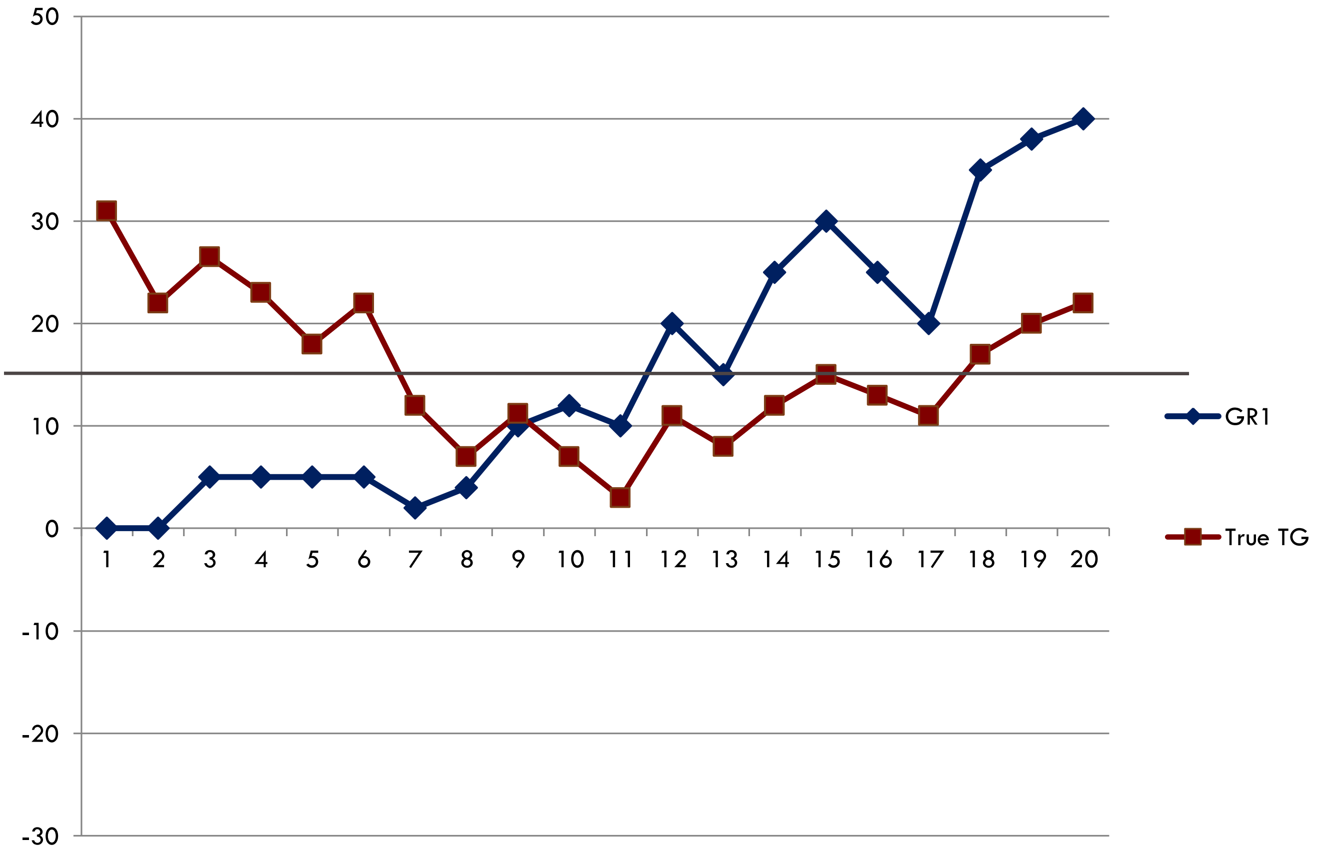


# Drawback of Coexpression Networks

**Genes**

**Patients**

| TG | GR1 | GR2 |
|---|---|---|
| 31 | 0 | 40 |
| 22 | 0 | 30 |
| 27 | 5 | 35 |
| 23 | 5 | 33 |
| 18 | 5 | 26 |
| 22 | 5 | 30 |
| 12 | 2 | 20 |
| 7 | 4 | 15 |
| 11 | 10 | 18 |
| 7 | 12 | 14 |
| 3 | 10 | 10 |
| 11 | 20 | -5 |
| 8 | 15 | 5 |
| 12 | 25 | -10 |
| 15 | 30 | -12 |
| 13 | 25 | -15 |
| 11 | 20 | -10 |
| 17 | 35 | -15 |
| 20 | 38 | -20 |
| 22 | 40 | -25 |

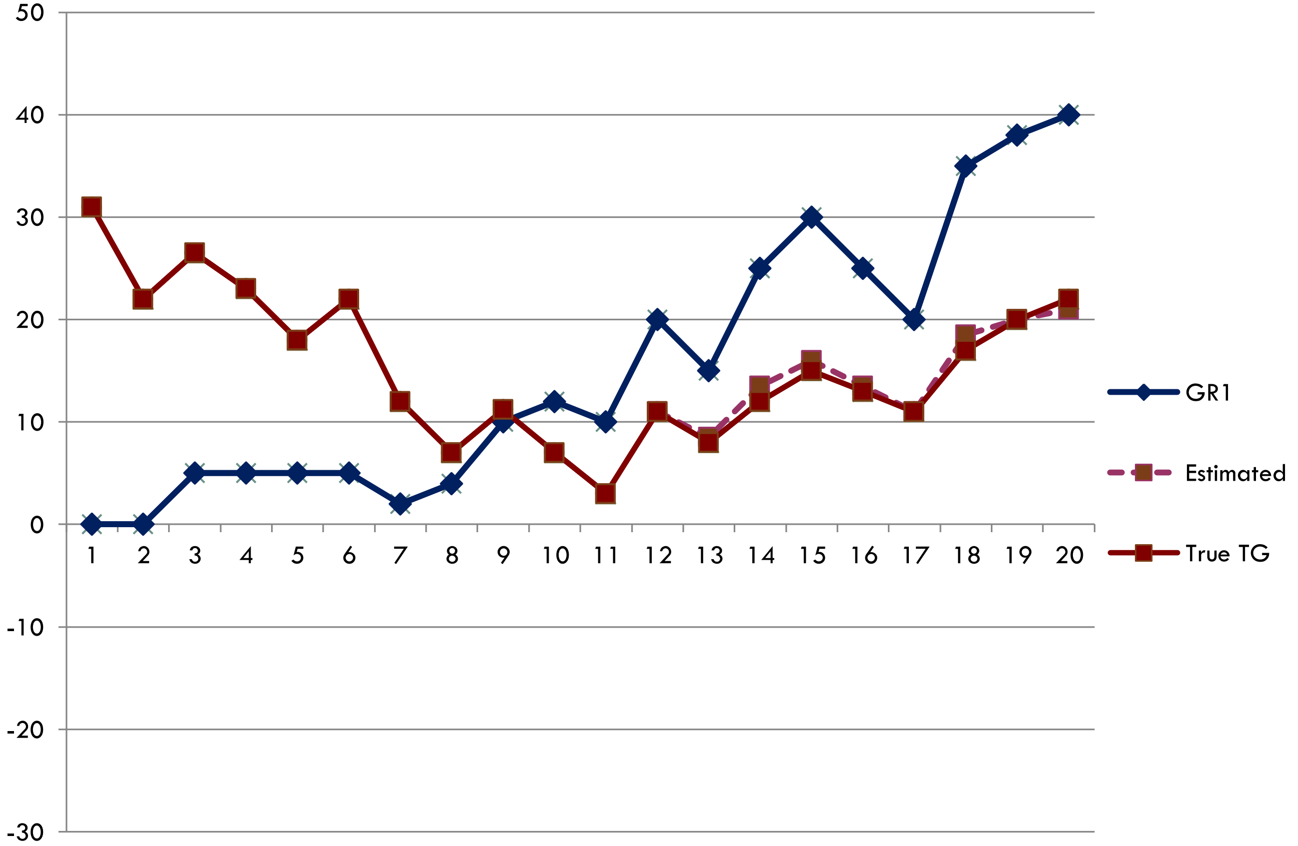


# Drawback of Coexpression Networks

**Genes**

**Patients**

| TG | GR1 | GR2 |
|---|---|---|
| 31 | 0 | 40 |
| 22 | 0 | 30 |
| 27 | 5 | 35 |
| 23 | 5 | 33 |
| 18 | 5 | 26 |
| 22 | 5 | 30 |
| 12 | 2 | 20 |
| 7 | 4 | 15 |
| 11 | 10 | 18 |
| 7 | 12 | 14 |
| 3 | 10 | 10 |
| 11 | 20 | -5 |
| 8 | 15 | 5 |
| 12 | 25 | -10 |
| 15 | 30 | -12 |
| 13 | 25 | -15 |
| 11 | 20 | -10 |
| 17 | 35 | -15 |
| 20 | 38 | -20 |
| 22 | 40 | -25 |



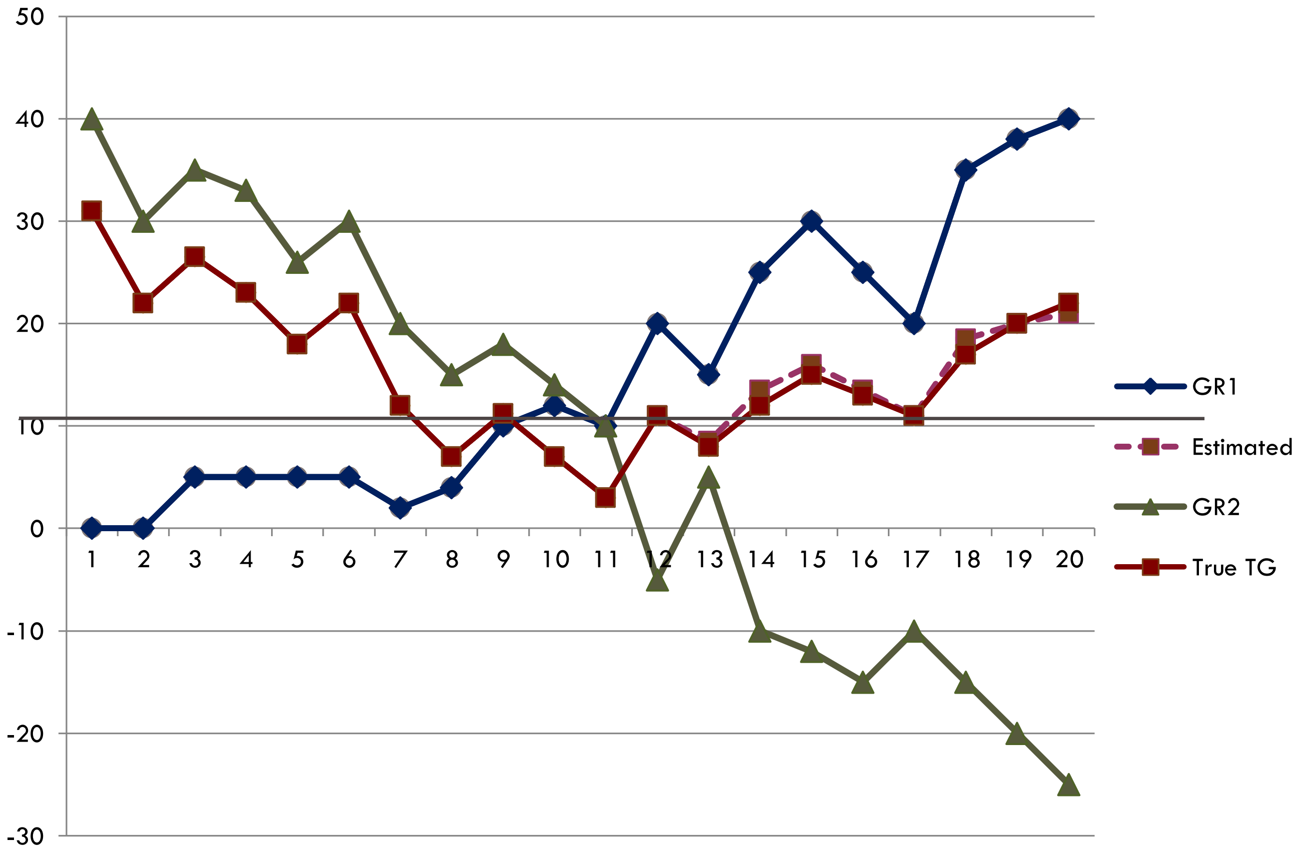
**Corr(TG,GR1) = -0.09 IF GR1>10 then Estimated-TG = 0.5 * GR1 + 1**

# Drawback of Coexpression Networks

**Genes**

**Patients**

| TG | GR1 | GR2 |
|---|---|---|
| 31 | 0 | 40 |
| 22 | 0 | 30 |
| 27 | 5 | 35 |
| 23 | 5 | 33 |
| 18 | 5 | 26 |
| 22 | 5 | 30 |
| 12 | 2 | 20 |
| 7 | 4 | 15 |
| 11 | 10 | 18 |
| 7 | 12 | 14 |
| 3 | 10 | 10 |
| 11 | 20 | -5 |
| 8 | 15 | 5 |
| 12 | 25 | -10 |
| 15 | 30 | -12 |
| 13 | 25 | -15 |
| 11 | 20 | -10 |
| 17 | 35 | -15 |
| 20 | 38 | -20 |
| 22 | 40 | -25 |



**Corr(TG,GR2) = 0.35**

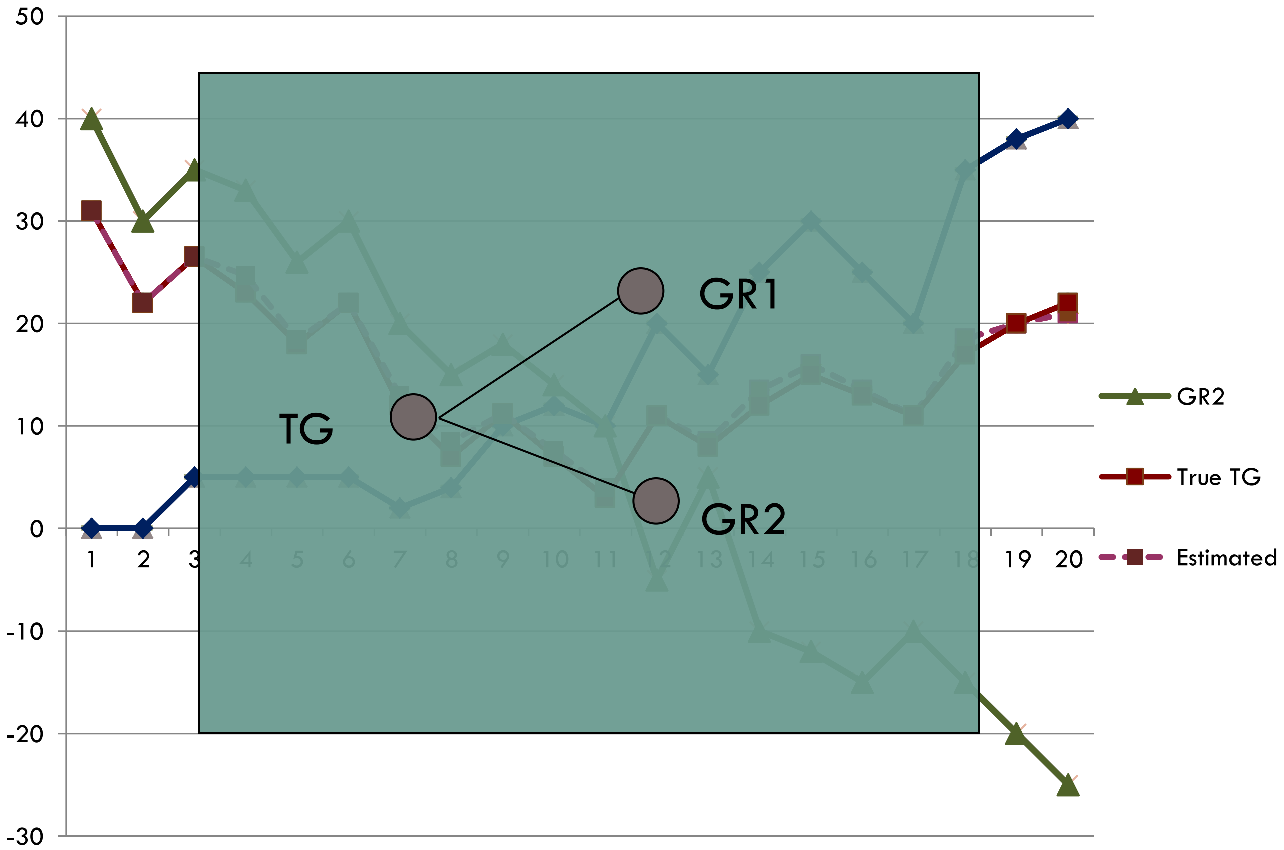
**IF GR2>10 then Estimated-TG = 0.9 * GR2 -5**

# Drawback of Coexpression Networks

**Genes**

**Patients**

| TG | GR1 | GR2 |
|---|---|---|
| 31 | 0 | 40 |
| 22 | 0 | 30 |
| 27 | 5 | 35 |
| 23 | 5 | 33 |
| 18 | 5 | 26 |
| 22 | 5 | 30 |
| 12 | 2 | 20 |
| 7 | 4 | 15 |
| 11 | 10 | 18 |
| 7 | 12 | 14 |
| 3 | 10 | 10 |
| 11 | 20 | -5 |
| 8 | 15 | 5 |
| 12 | 25 | -10 |
| 15 | 30 | -12 |
| 13 | 25 | -15 |
| 11 | 20 | -10 |
| 17 | 35 | -15 |
| 20 | 38 | -20 |
| 22 | 40 | -25 |



# Outline

1. Introduction
2. Gene Network Inference: state-of-the art
3. Proposals & Results:
    a) RegNet: Gene Regression Networks
    b) SATuRNo: Supervised Prognostic Approach Through Regression Networks
    c) CarGene: Characterization of genes
4. Conclusions and future work
5. Project membership

# Outline

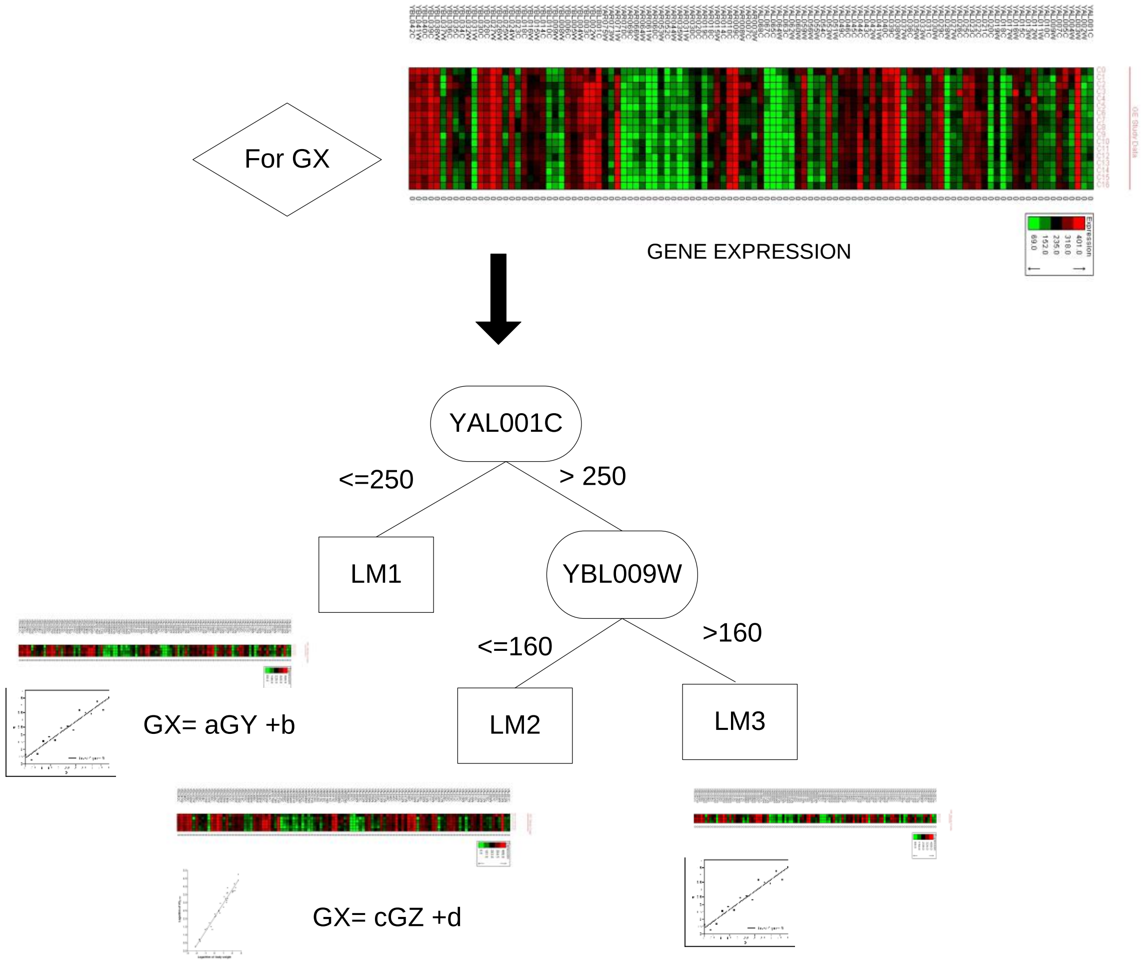
1. Introduction
2. Gene Network Inference: state-of-the art
3. Proposals & Results:
   a) RegNet: Gene Regression Networks
   b) SATuRNo: Supervised Prognostic Approach Through Regression Networks
   c) CarGene: Characterization of genes
4. Conclusions and future work
5. Project membership

# I RegNet

- REGNET
  - Method to infer Gene Regression Network
  - It is based on Model Tree (M5’)
  - Favoring to infer localized similarities over a more global similarity

# I RegNet…is based on

- ☐ Model trees: M5' algorithm



For GX

GENE EXPRESSION

YAL001C

<=250

> 250

LM1

YBL009W

<=160

>160

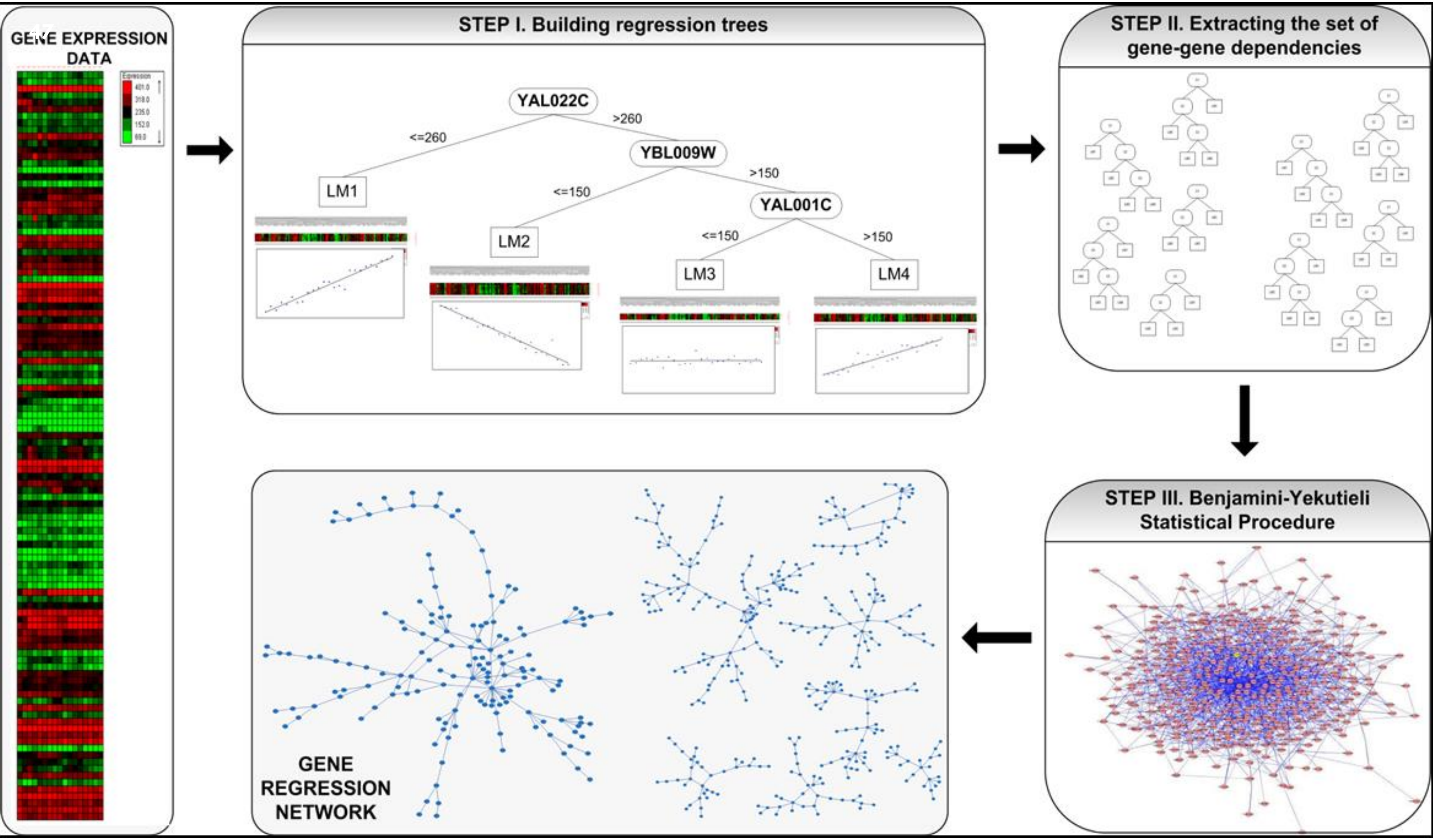
LM2

LM3

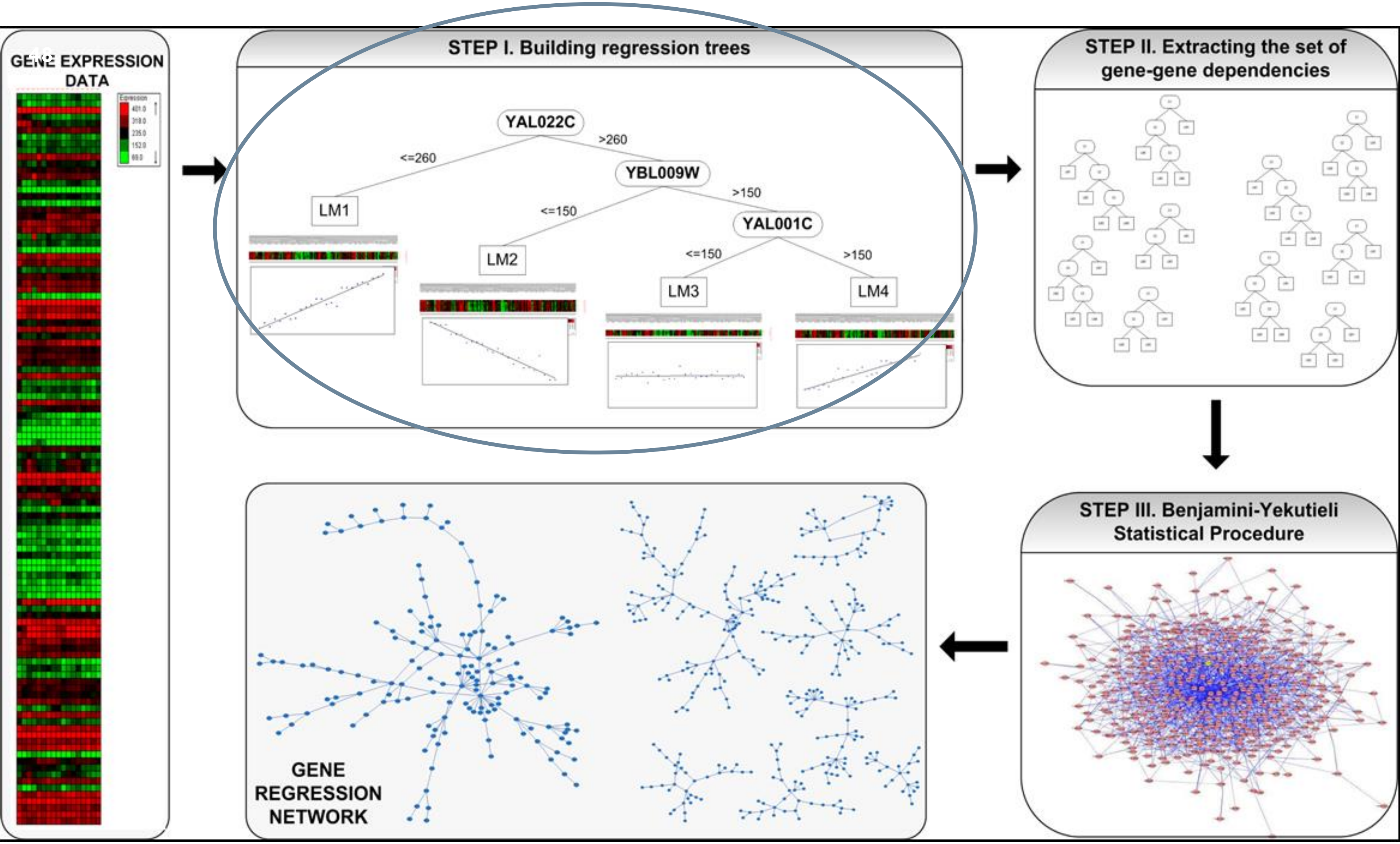
GX= aGY +b

GX= cGZ +d

# I RegNet


GENE EXPRESSION DATA
Expression
401.0
318.0
235.0
152.0
69.0
STEP I. Building regression trees
YAL022C
<=260
>260
YBL009W
LM1
<=150
>150
LM2
YAL001C
<=150
>150
LM3
LM4
STEP II. Extracting the set of gene-gene dependencies
STEP III. Benjamini-Yekutieli Statistical Procedure
GENE REGRESSION NETWORK

# I RegNet



GENE EXPRESSION DATA

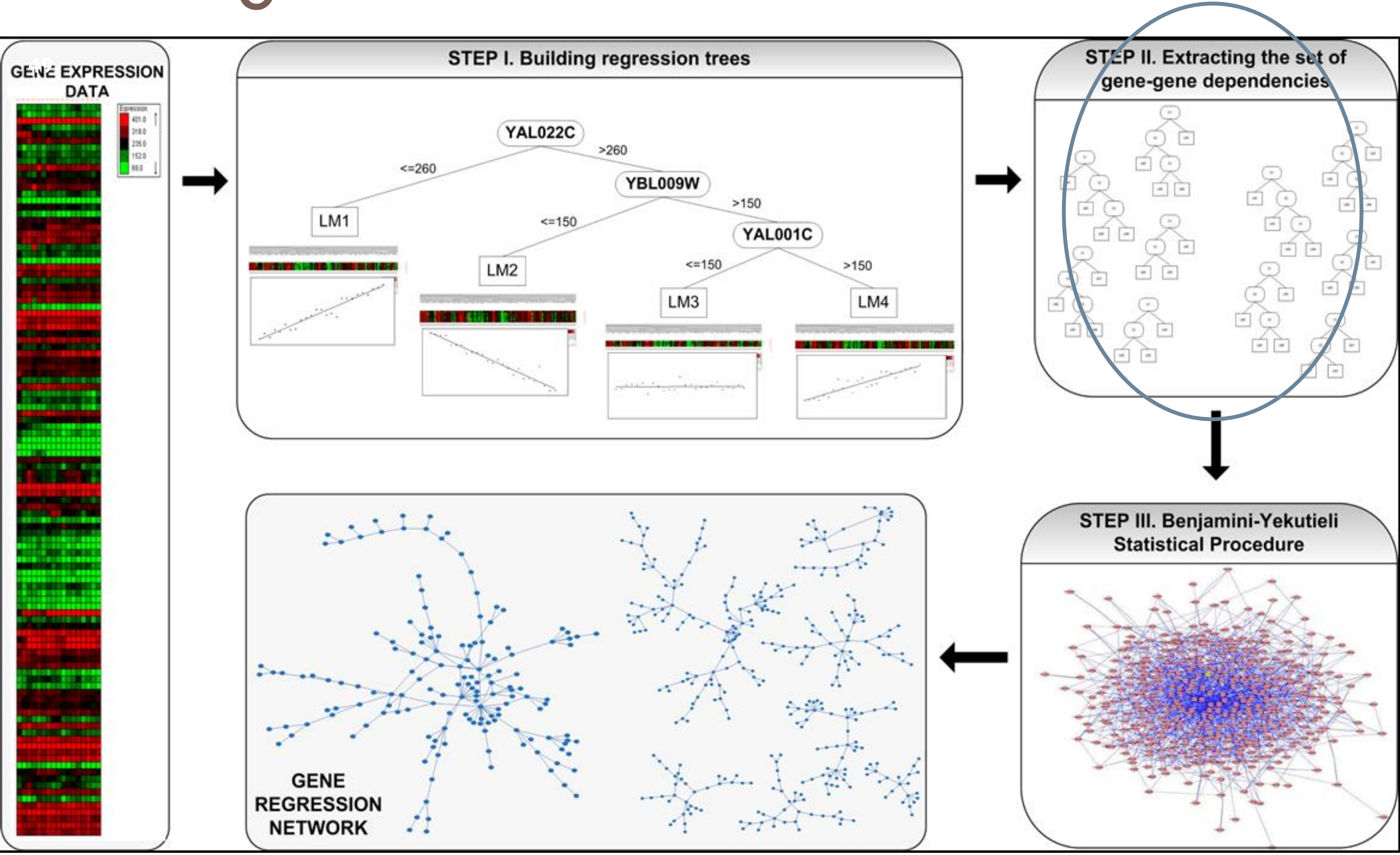
STEP I. Building regression trees

STEP II. Extracting the set of gene-gene dependencies

STEP III. Benjamini-Yekutieli Statistical Procedure

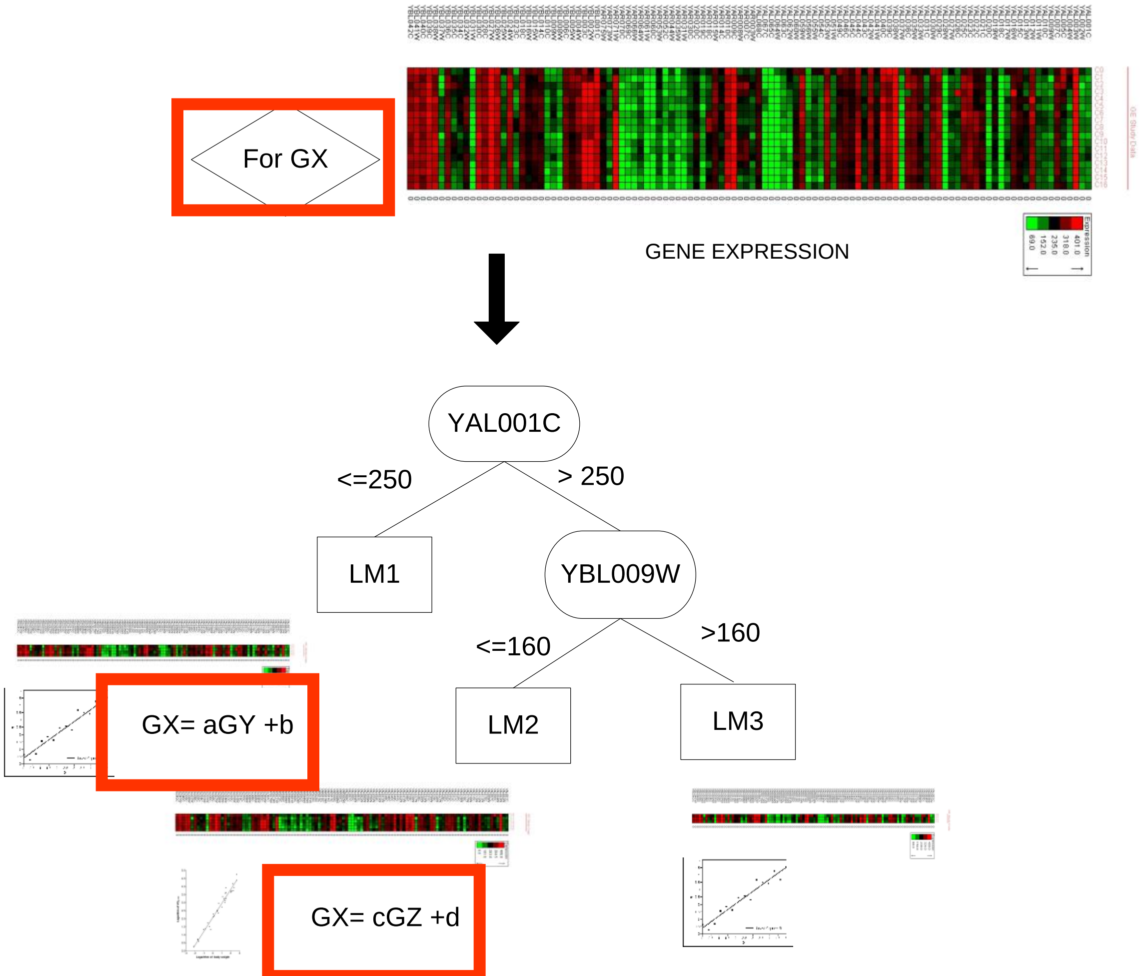
GENE REGRESSION NETWORK

# I RegNet


GENE EXPRESSION DATA
Expression
401.0
318.0
235.0
152.0
69.0
STEP I. Building regression trees
YAL022C
<=260
>260
YBL009W
LM1
<=150
>150
YAL001C
LM2
<=150
>150
LM3
LM4
STEP II. Extracting the set of gene-gene dependencies
STEP III. Benjamini-Yekutieli Statistical Procedure
GENE REGRESSION NETWORK

# I RegNet (Step II)

- Step II…that means



For GX

GENE EXPRESSION

YAL001C

<=250

> 250

LM1

YBL009W

<=160

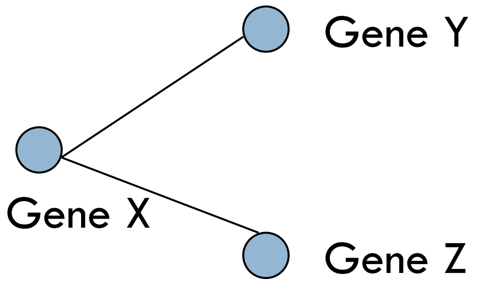
>160

LM2

LM3

GX= aGY +b

GX= cGZ +d



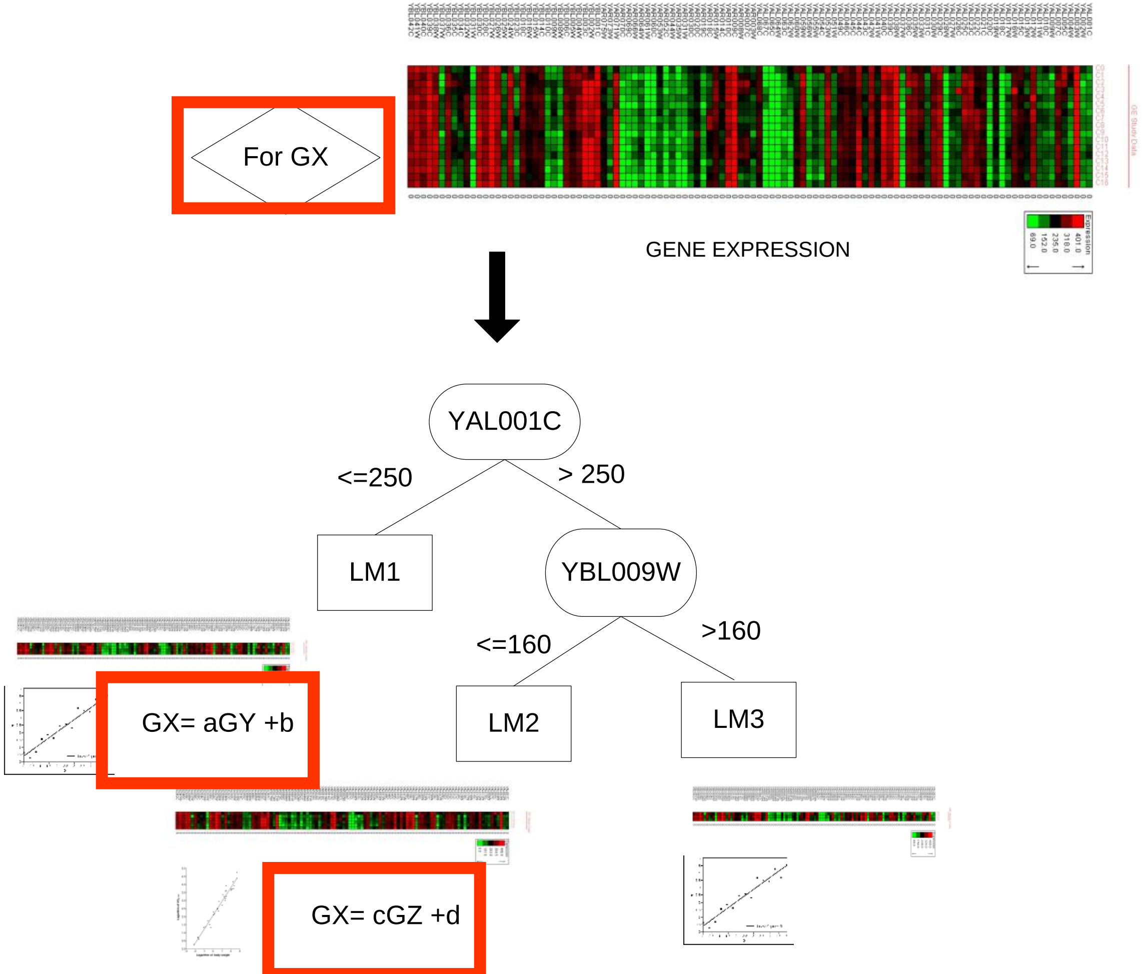
Gene Y

Gene X

Gene Z

# I RegNet (Step II)

- ☐ Step II…that means if and only if ε of $MT_{GX} < \theta$



For GX

GENE EXPRESSION

YAL001C

<=250

> 250

LM1

YBL009W

<=160

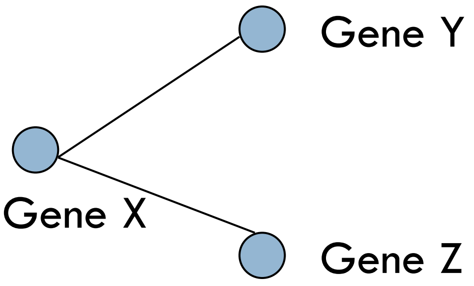
>160

LM2

LM3

GX= aGY +b

GX= cGZ +d

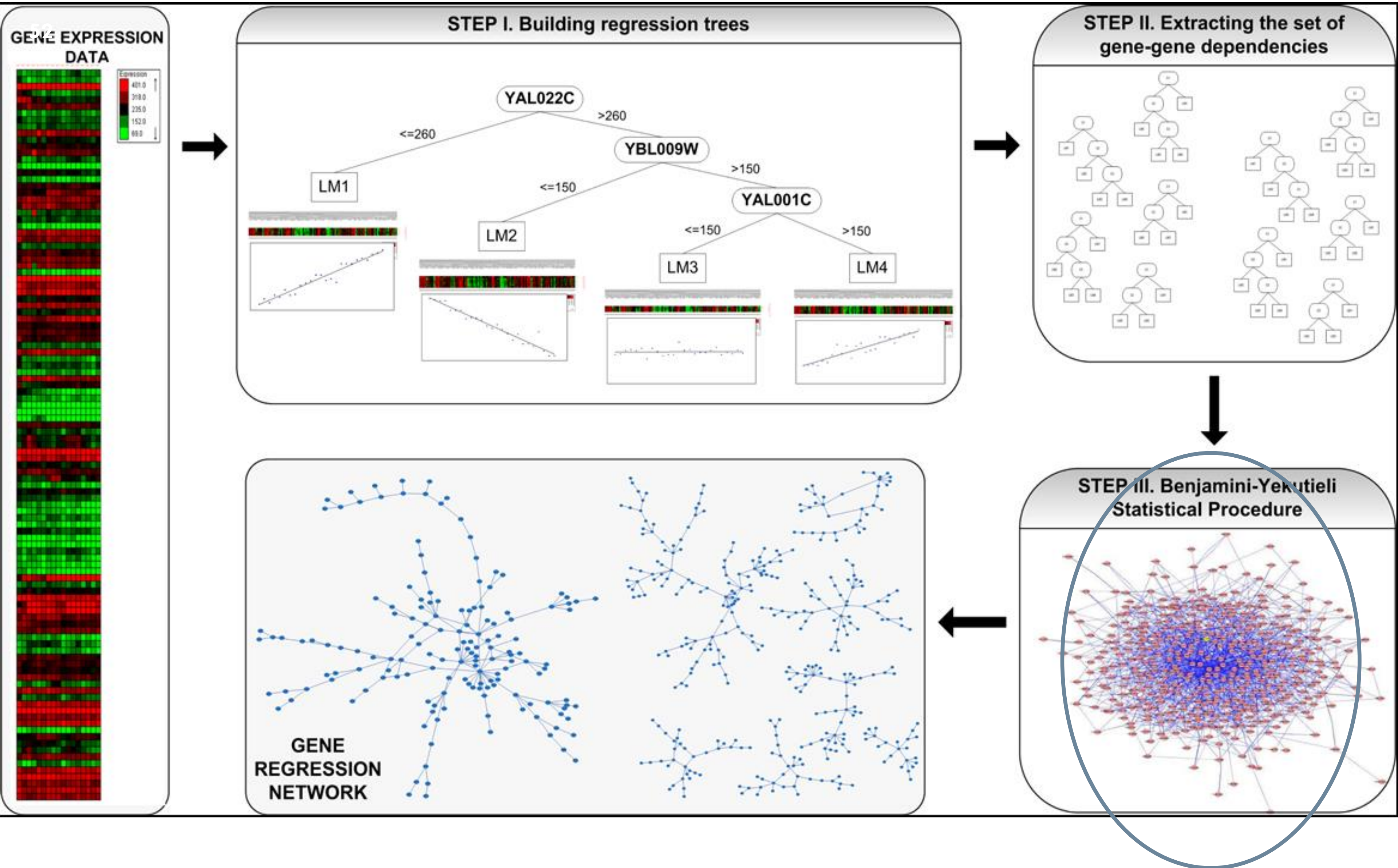


Gene Y

Gene X

Gene Z

# I RegNet


GENE EXPRESSION DATA
Expression
401.0
318.0
235.0
152.0
69.0
STEP I. Building regression trees
YAL022C
<=260
>260
LM1
YBL009W
<=150
>150
LM2
YAL001C
<=150
>150
LM3
LM4
STEP II. Extracting the set of gene-gene dependencies
STEP III. Benjamini-Yekutieli Statistical Procedure
GENE REGRESSION NETWORK

# I RegNet (Step III)

- A family of hypotheses is tested simultaneously
- Deciding which dependency is considered as a discovery
  - Benjamini-Yekutieli procedure

$$H_0^1, H_0^2, ..., H_0^m \text{ and } p_1, ..., p_m$$

$$\text{Let } p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)} \text{ be the ordered p-values}$$

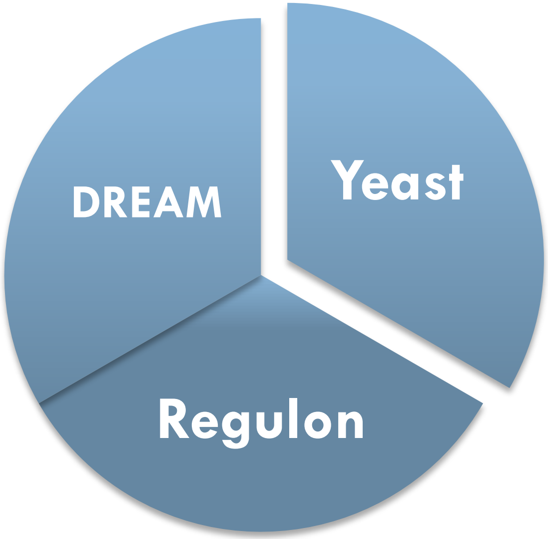
$$\text{Rejects all hypothesis } H_0^1, H_0^2, ..., H_0^k$$

$$k = max\{i : p_{(i)} \frac{m}{i} \sum_{k=1}^{m} \frac{1}{k} \leq \alpha\}$$

# I RegNet. Results

- <u>Performance Evaluation</u>
  - A. In-silico benchmark suite of datasets (***DREAM***)
  - B. True benchmark network
    - Escherichia Coli Dataset
    - ***Regulon DB***

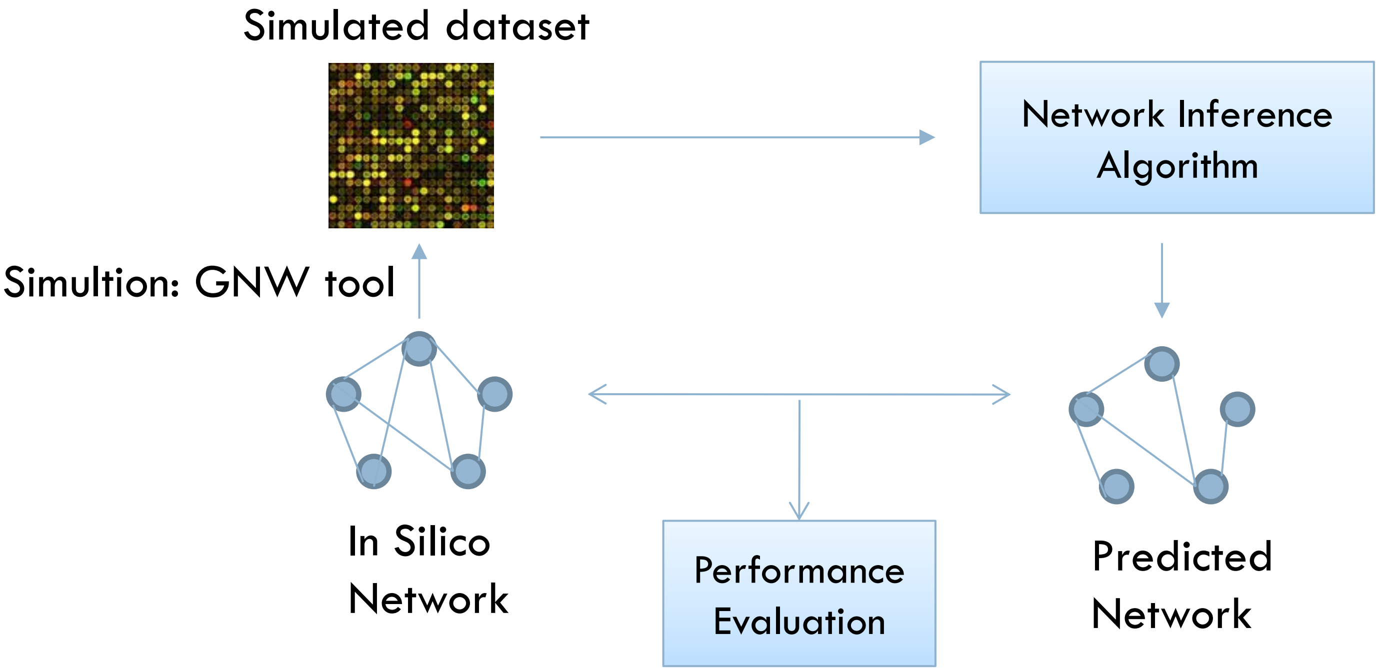
- <u>Data Analysis</u>
  - C. Saccharomyces Cerevisiae Dataset



# I RegNet. In silico DS

## A) Performance Evaluation

- In silico benchmark suite of datasets (DREAM4 Challenge) (PNAS 2010)
- Blind performance test



Simulated dataset

Network Inference Algorithm

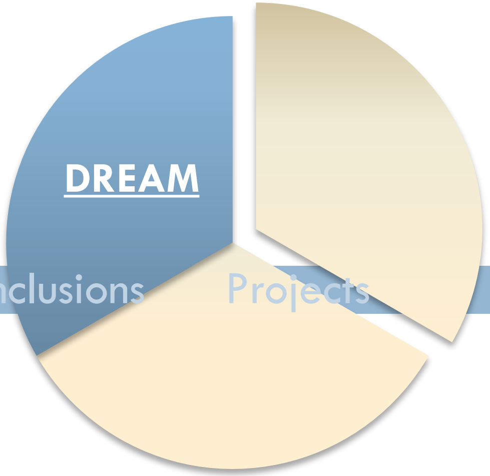
Simultion: GNW tool

In Silico Network

Performance Evaluation

Predicted Network

# I RegNet. In silico DS



## A) Performance Evaluation

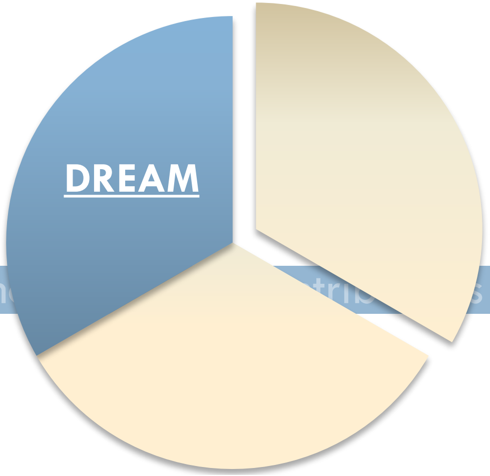
- In silico benchmark suite of datasets:
  - 5 nets (size 100) hidden in 15 microarray experiments

| Type of Microarray Experiment | N° of DataSets |
|---|---|
| Knockout | 5 |
| Knockdown | 5 |
| Multifactorial | 5 |

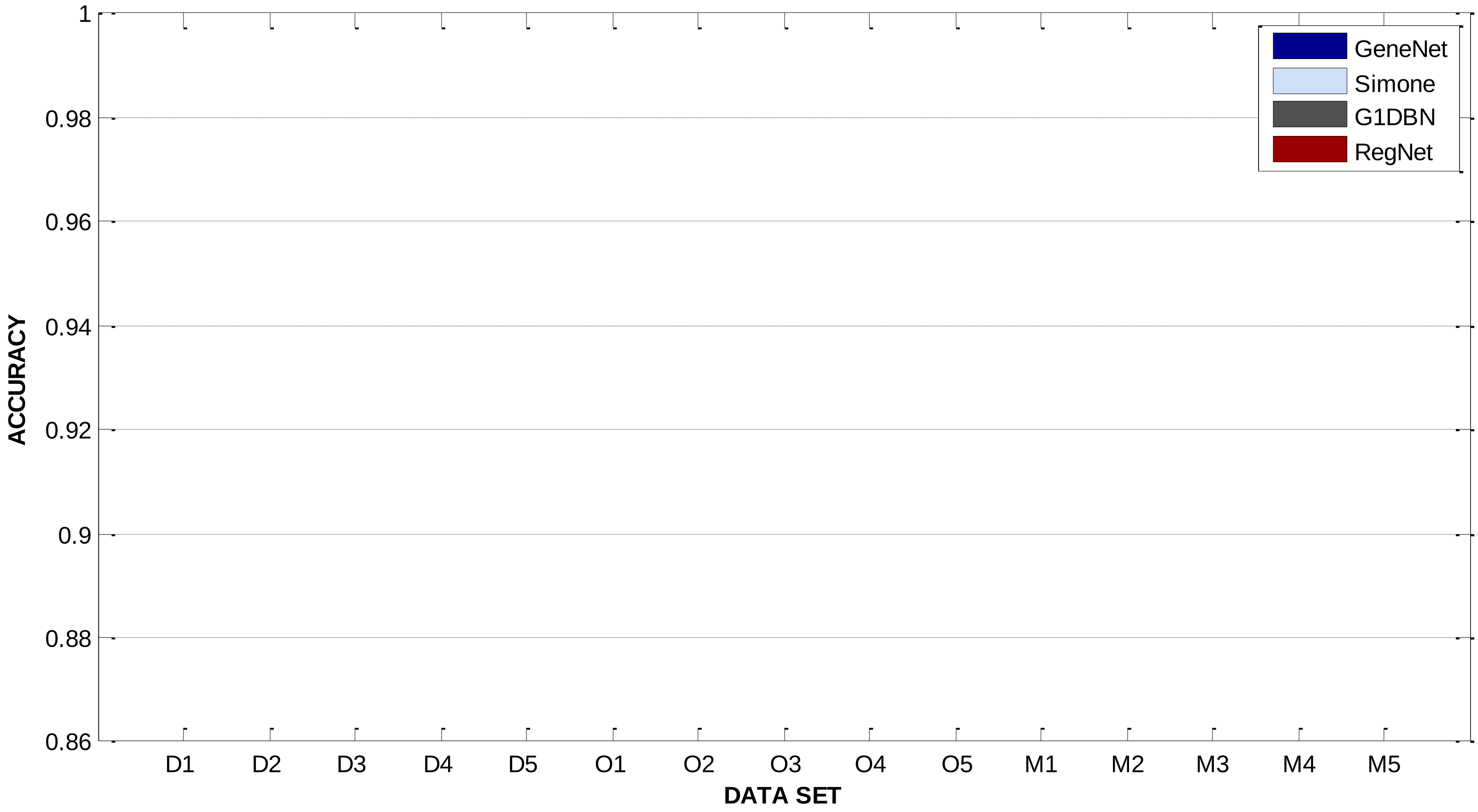
- Benchmark methods

| Benchmark Approaches |
|---|
| GeneNet (Partial Correlations) |
| Simone (Weighted-LASSO) |
| G1DBN (Bayesian Net) |

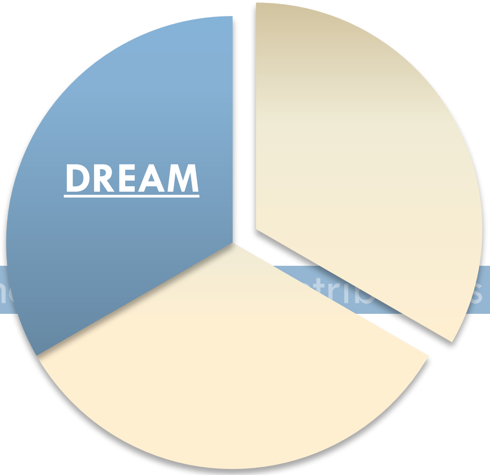
# I RegNet. In silico DS



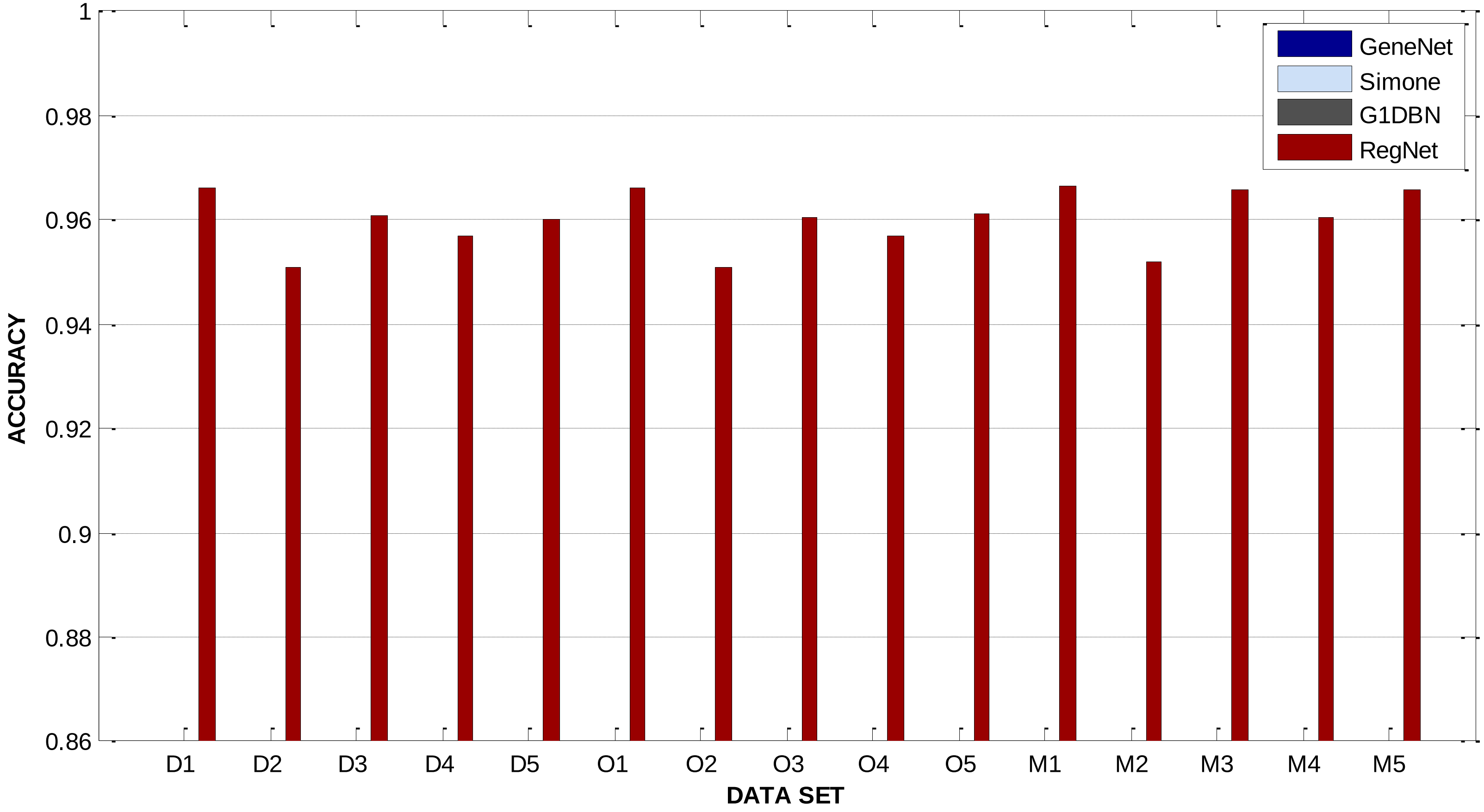
Acc = (TP+TN) / (TP+FP+FN+TN)



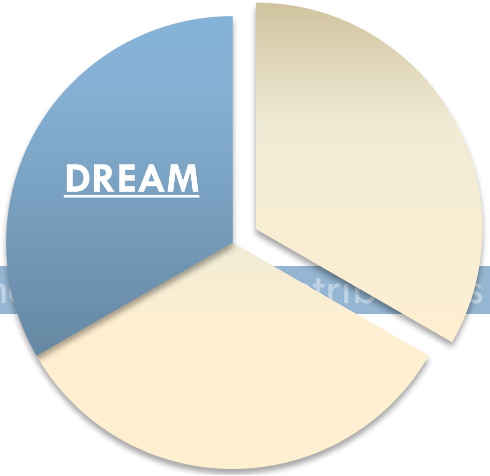
# I RegNet. In silico DS



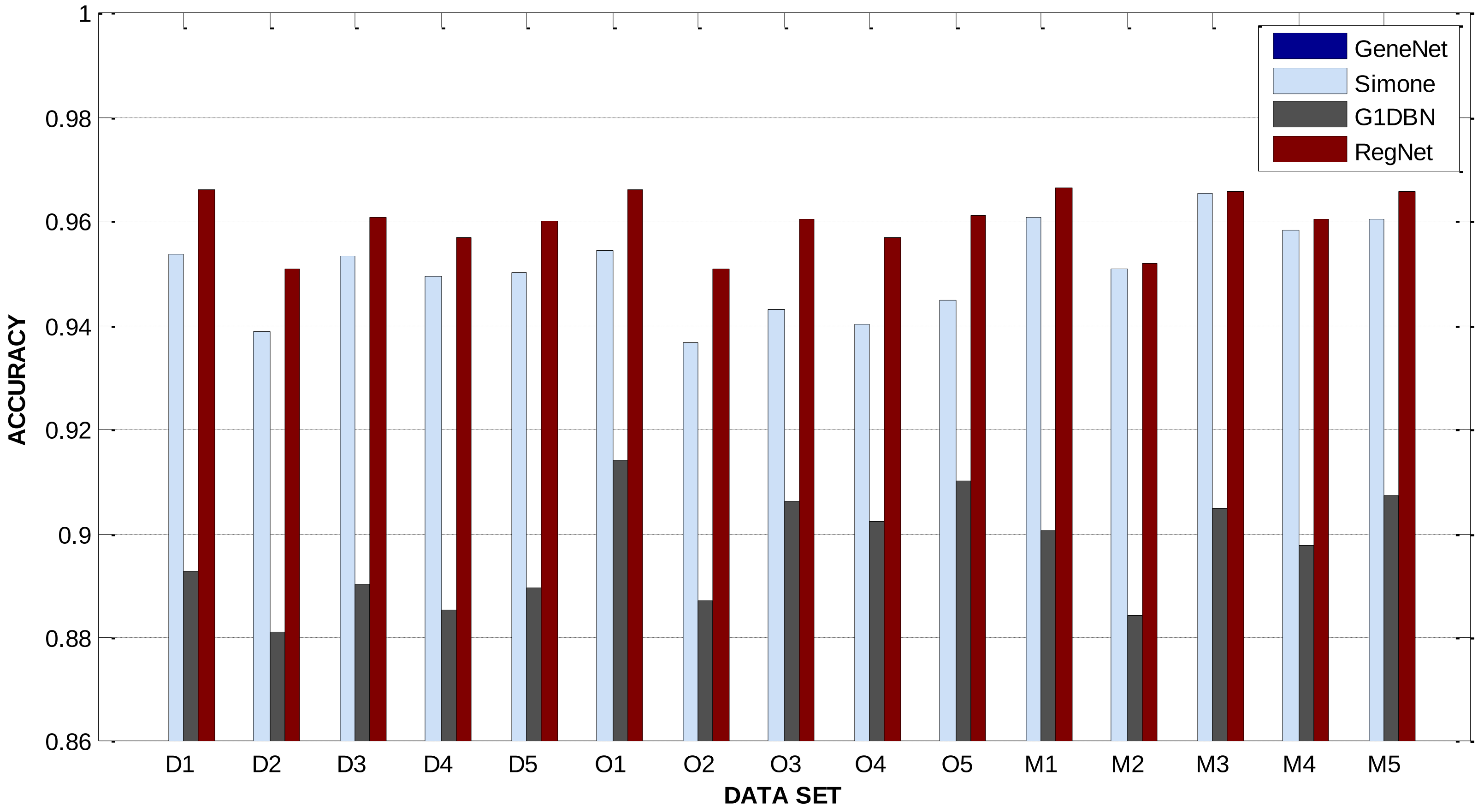
Acc = (TP+TN) / (TP+FP+FN+TN)



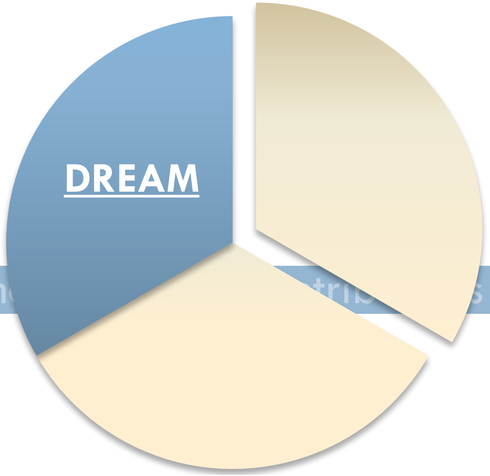
# I RegNet. In silico DS



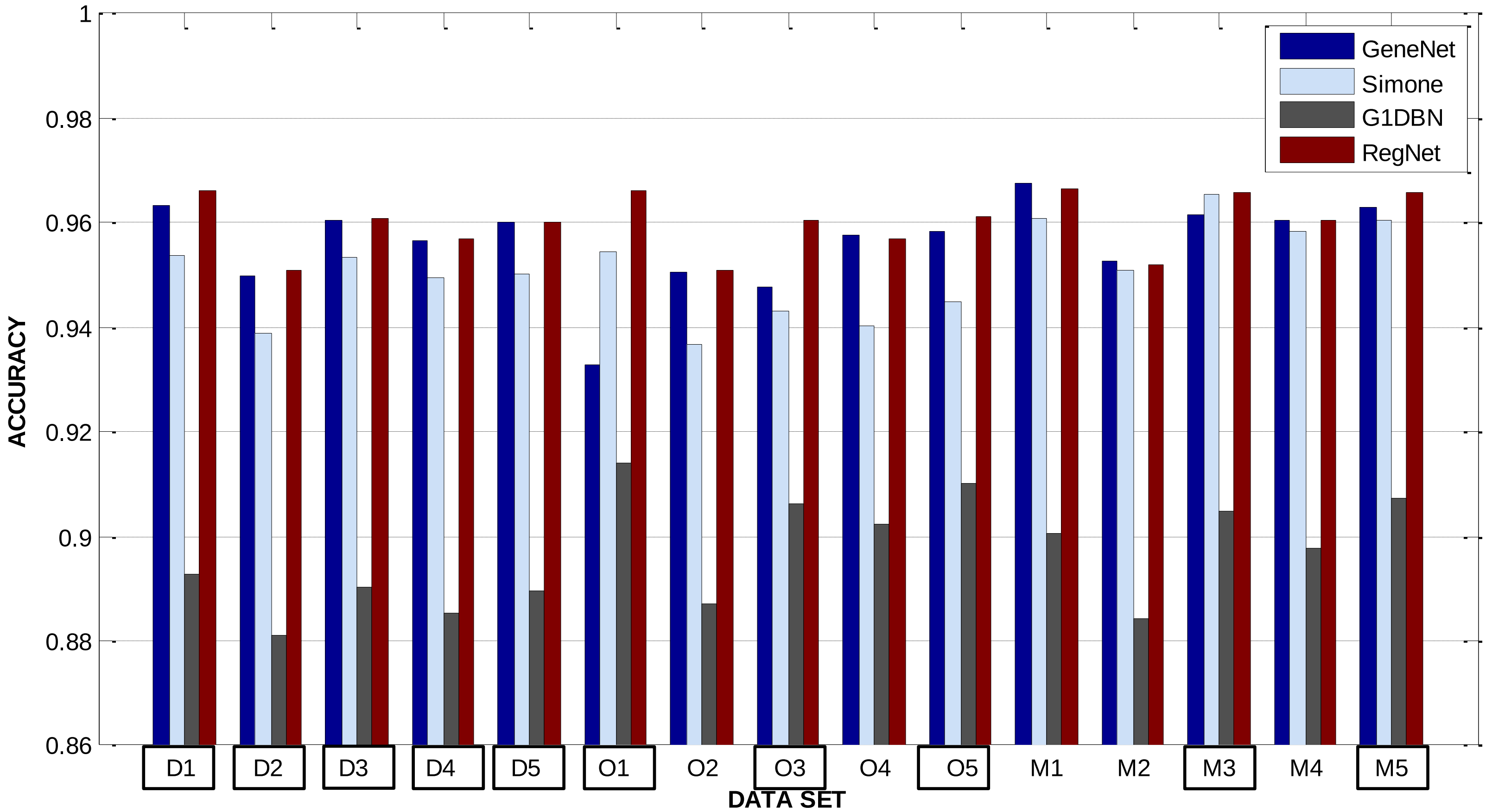
Acc = (TP+TN) / (TP+FP+FN+TN)



# I RegNet. In silico DS



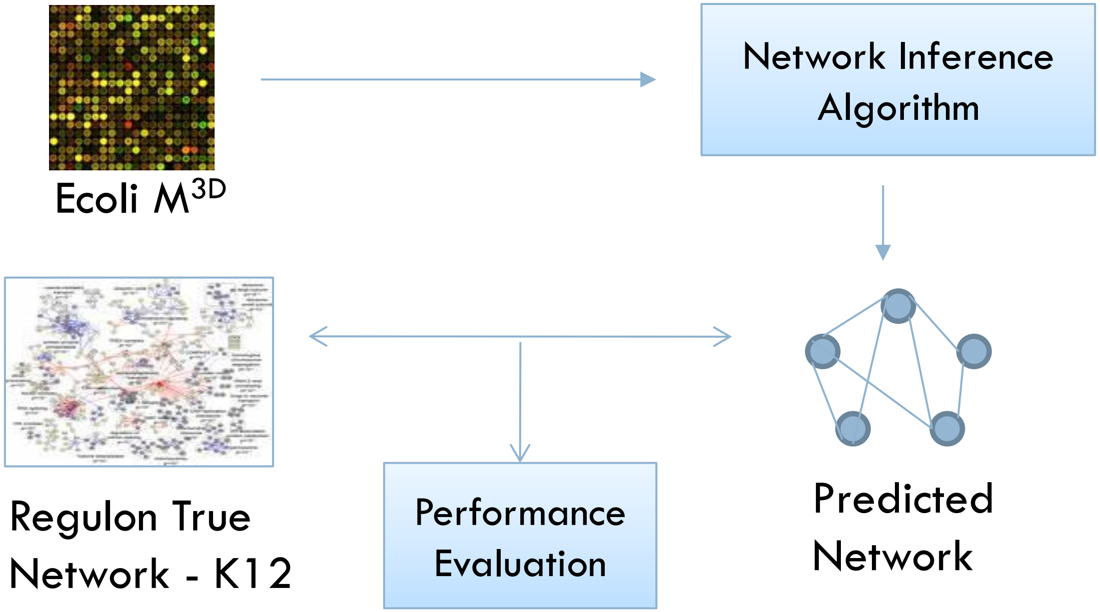
$$Acc = (TP+TN) / (TP+FP+FN+TN)$$



# I RegNet. True DS

## Experiment B)

- Dataset: from Gardner Lab (4292 genes)
- True Network: E.coli K12 transcriptional network


Ecoli $M^{3D}$

Network Inference Algorithm

Predicted Network

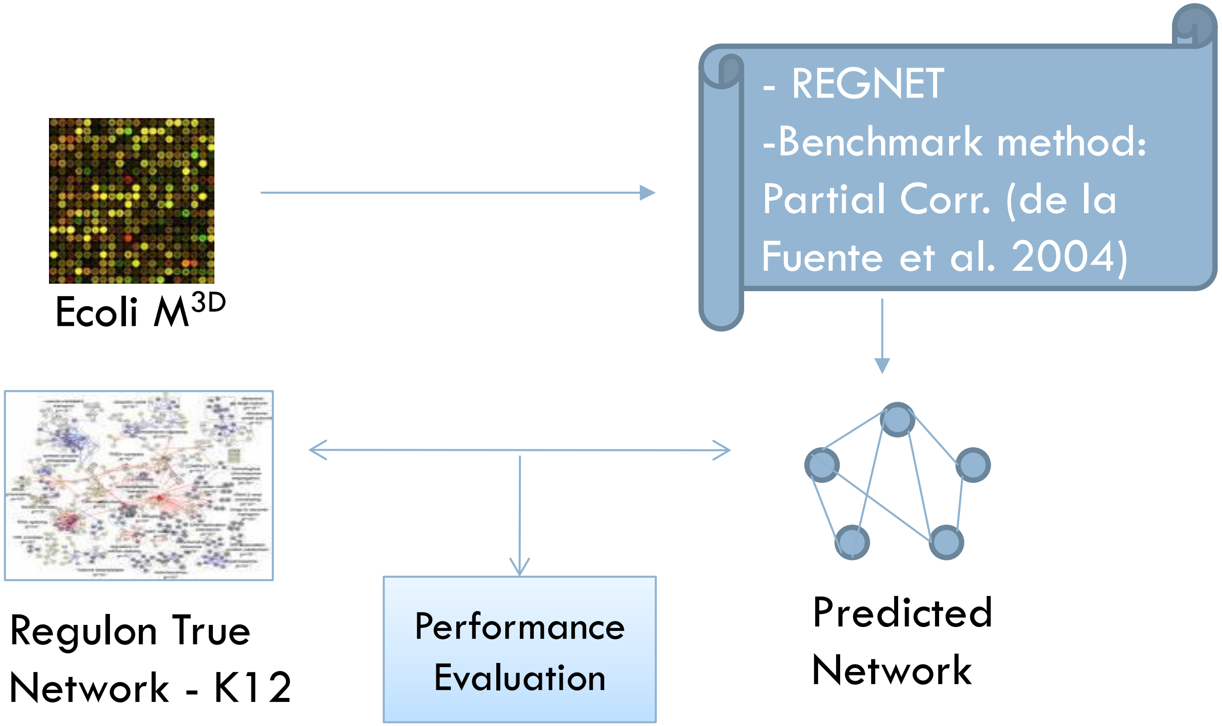
Regulon True Network - K12

Performance Evaluation

# I RegNet. True DS

## Experiment B)

- Dataset: from Gardner Lab (4292 genes)
- True Network: E.coli K12 transcriptional network



Ecoli $M^{3D}$

- REGNET
-Benchmark method: Partial Corr. (de la Fuente et al. 2004)

Regulon True Network - K12

Performance Evaluation

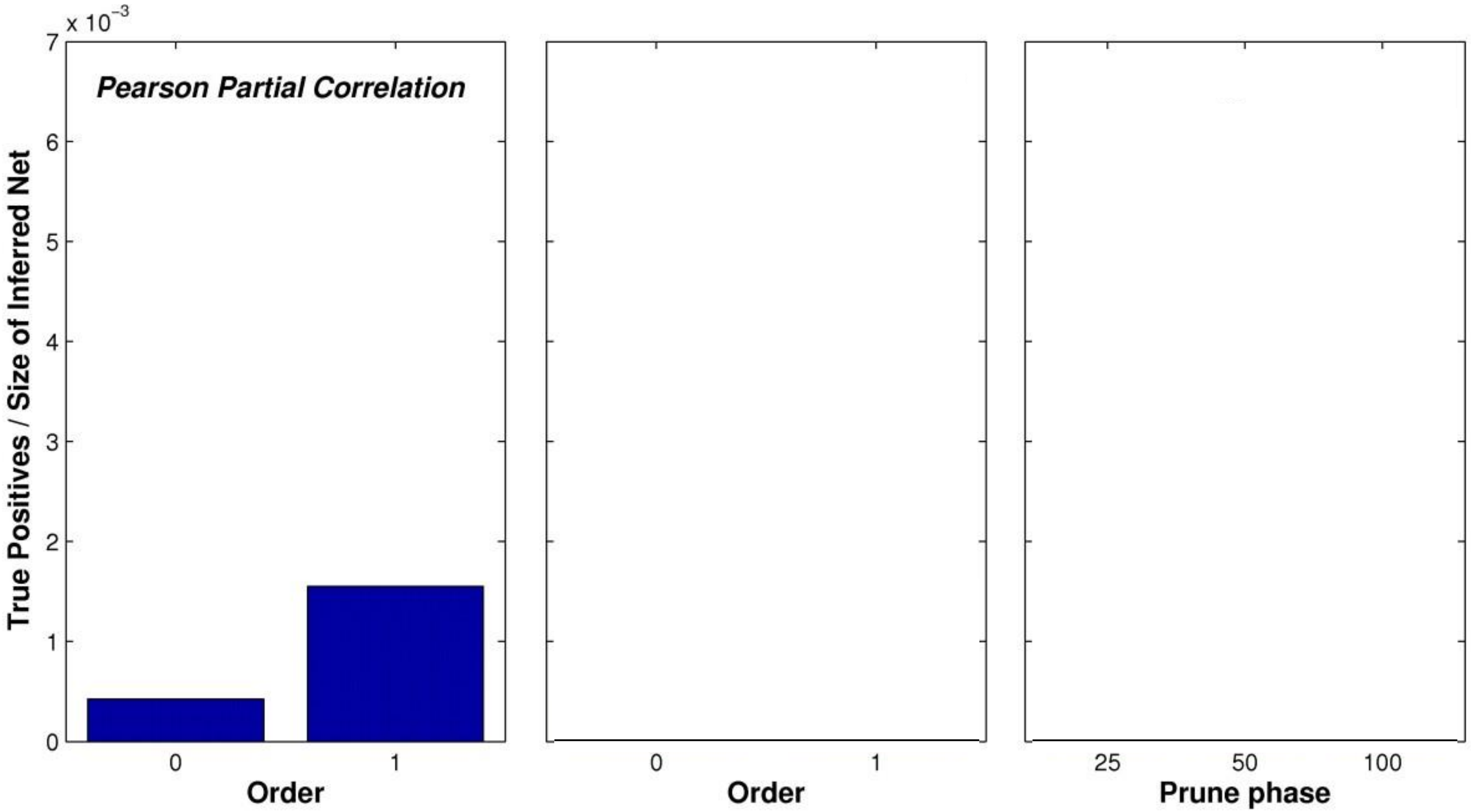
Predicted Network

# I RegNet. True DS

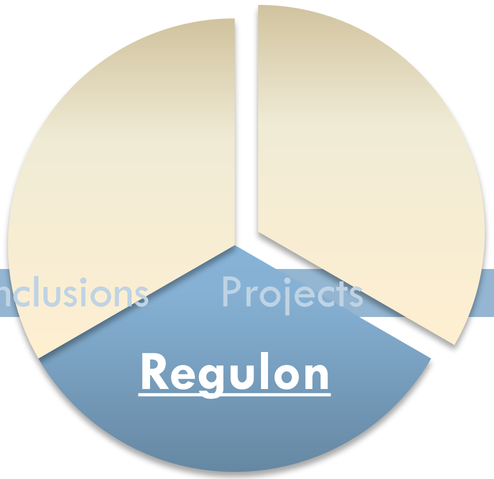
## B) Performance Evaluation

# I RegNet. True DS

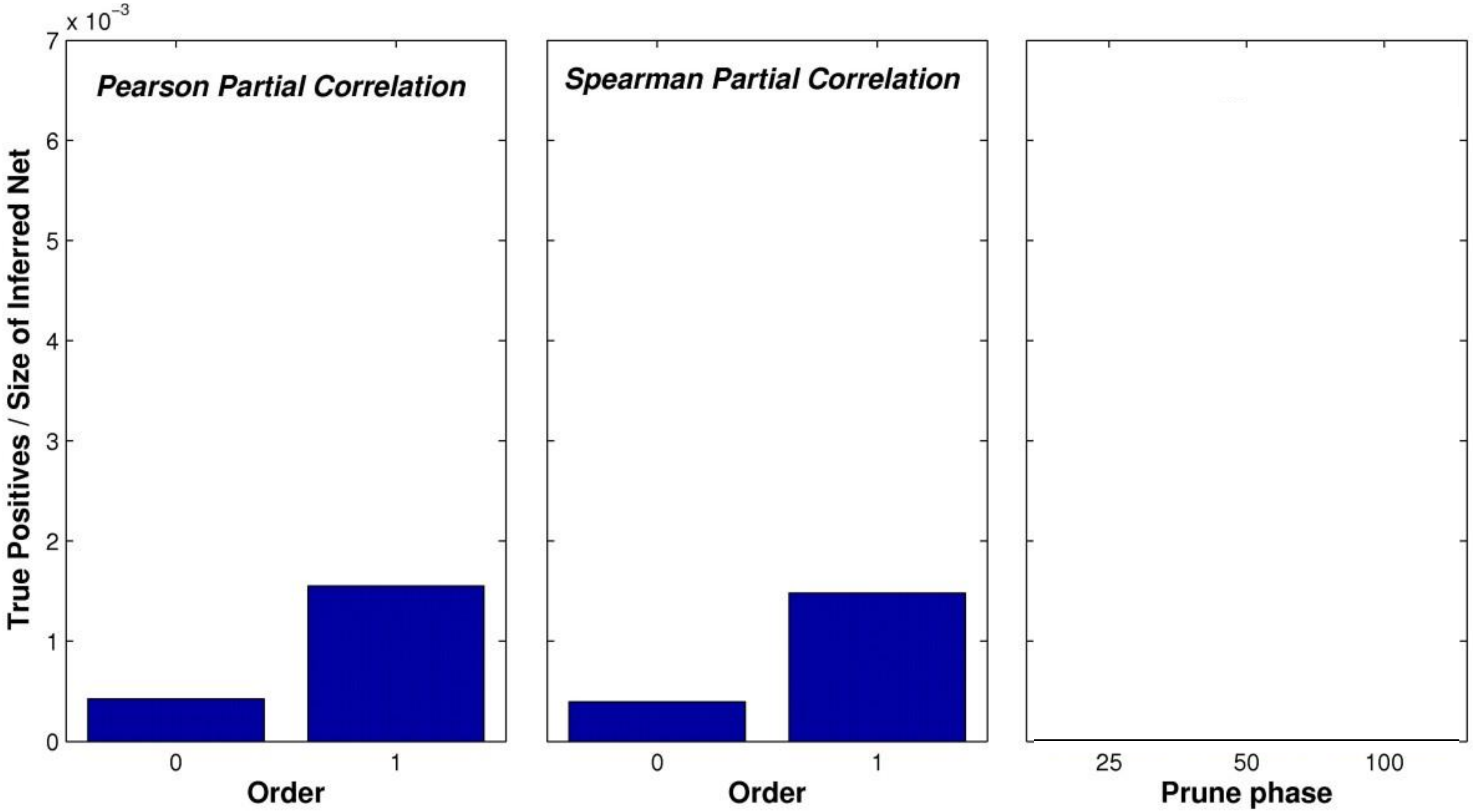
## B) Performance Evaluation



# I RegNet. True DS

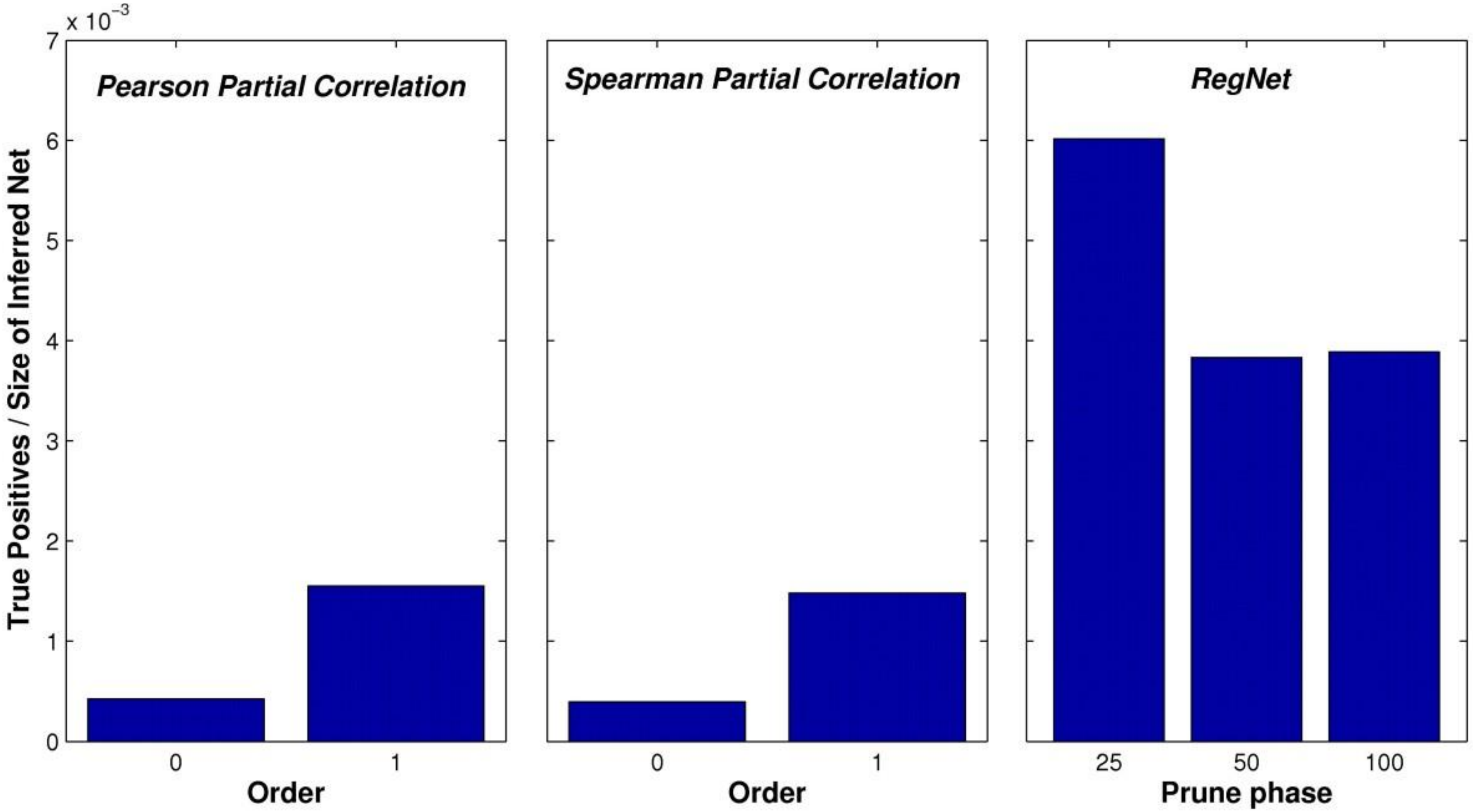


## B) Performance Evaluation



# I RegNet. True DS

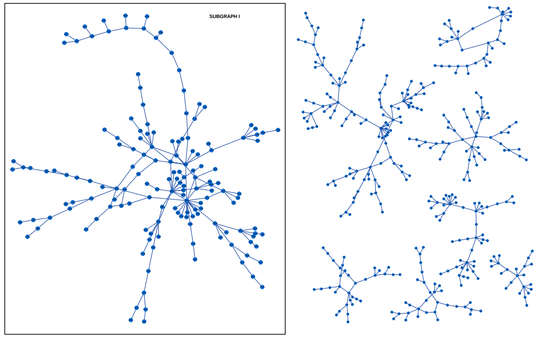
## B) Performance Evaluation



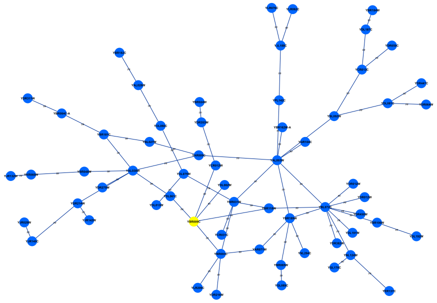
# I RegNet. Data analysis

## C) Saccharomyces Cerevisiae cell cycle

- 2884 genes and 17 experimental conditions
- Executions ($\alpha=0.05$):
  - Non-prunning phase: all model trees

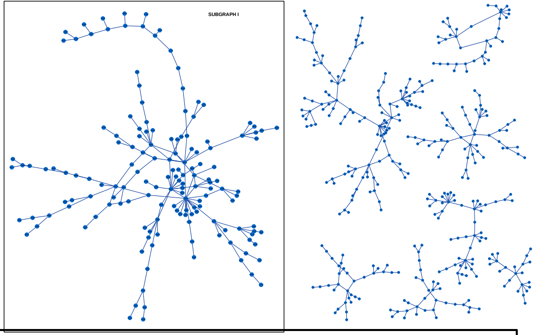


  - Prunning phase: model tree with ε<50%

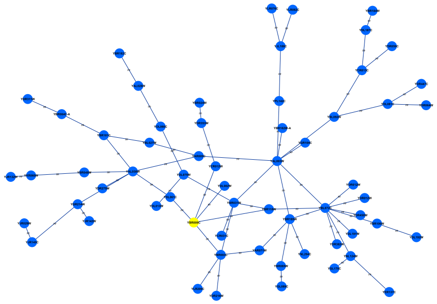


# I RegNet. Data analysis

| P-adj | GO Attribute |
|---|---|
| < 0.001 | 0005830: cytosolic ribosome |
| < 0.001 | 0005840: ribosome |
| < 0.001 | 0003735: structural constituent of ribosomal protein |
| < 0.001 | 0030529: ribonucleoprotein complex |
| < 0.001 | 0015935: small ribosomal subunit |
| < 0.001 | 0015934: large ribosomal subunit |



Non-Prunning phase: all model trees



model trees ε<50

| P-adj | GO Attribute |
|---|---|
| < 0.001 | 0005830: cytosolic ribosome |
| < 0.001 | 0005840: ribosome |
| < 0.001 | 0003735: structural constituent of ribosomal protein |
| < 0.001 | 0005198: structural molecule activity |
| < 0.001 | 0030529: ribonucleoprotein complex |
| < 0.001 | 0005843: cytosolic small ribosomal subunit …. |

# Outline

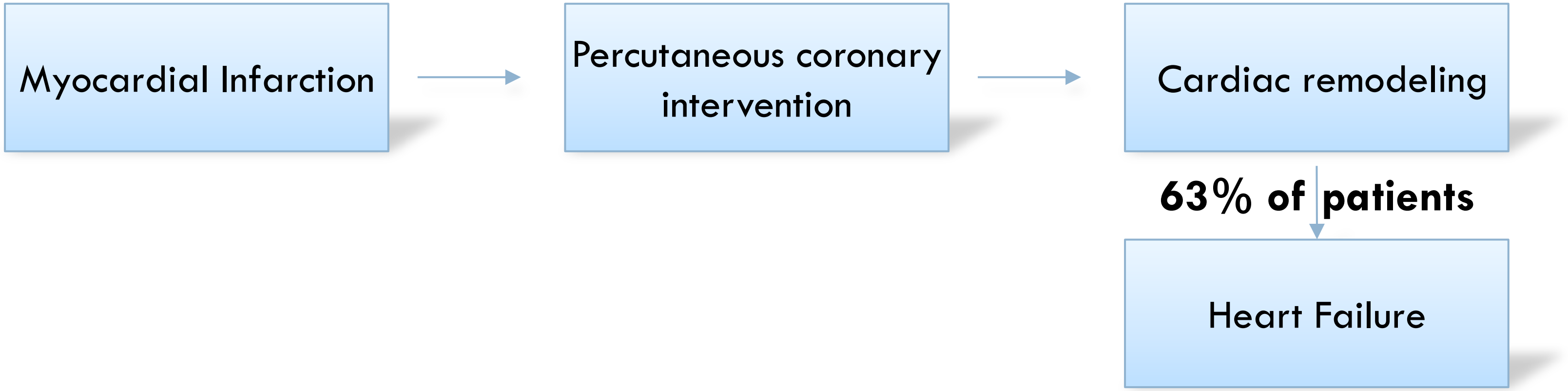
1. Introduction
2. Gene Network Inference: state-of-the art
3. Proposals & Results:
   a) RegNet: Gene Regression Networks
   b) SATuRNo: Supervised Prognostic Approach Through Regression Networks
   c) CarGene: Characterization of genes
4. Conclusions and future work
5. Project membership

# II SATuRNo

- Idea
  - Network-based approach for prognostic model
- Resulted method
  - SATuRNo (Supervised prognostic Approach Through Regression Networks)
- In collaboration with
  - Dr. Francisco Azuaje (Laboratory of Cardiovascular Research in Luxembourg)

# II SATuRNo. Motivation

- Scenario: 63% develop Heart F(Eur Heart J 2008)



Myocardial Infarction

Percutaneous coronary intervention

Cardiac remodeling

**63% of patients**

Heart Failure

# II SATuRNo. Motivation

- Scenario: 63% develop Heart F(Eur Heart J 2008)

Myocardial Infarction → Percutaneous coronary intervention → Cardiac remodeling

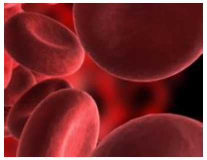
**63% of patients**

↓

Heart Failure

- Desired Scenario

Myocardial Infarction → Percutaneous coronary intervention



↓

BIOSIGNATURE →

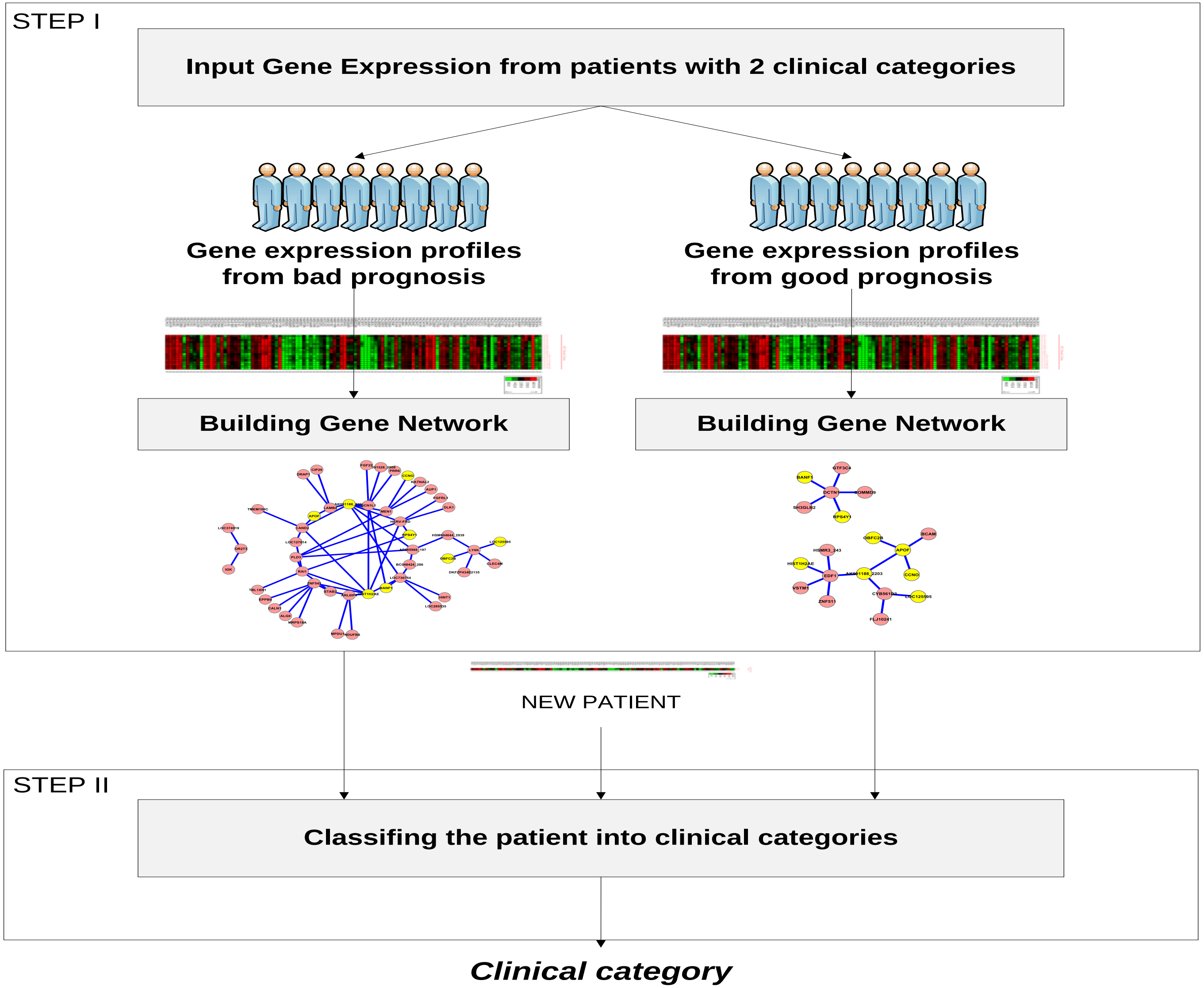
→ Personalized therapy

# Il SATuRNo. AIMS

- To investigate:
  - mechanism driving HF in post-MI patients
  - early identification of patients at risk of HF

- To infer:
  - clinically relevant gene association networks
  - potential prognosis biomarkers, based on gene association networks

# II SATuRNo. The method



STEP I

**Input Gene Expression from patients with 2 clinical categories**

**Gene expression profiles from bad prognosis**

**Gene expression profiles from good prognosis**

**Building Gene Network**

**Building Gene Network**

NEW PATIENT

STEP II

**Classifing the patient into clinical categories**

***Clinical category***

# II SATuRNo. The method

- Step I: Building Networks
  - Adaptation of REGNET
    - without  Proc. Benjamini Yekutieli

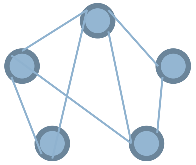
- Step II: Classifing post-MI patients

# II SATuRNo. The method

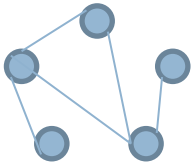
- ☐ Step II: Classifing post-MI patients

**New Patient (Real values)**

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **G1** | **G2** | **G3** | **G4** | **G5** | **G6** | **G7** | **G8** | **…** | **…** | **…** | **…** | **…** | **…** |



Good



Bad

**Target genes: estimated values**

| | | | |
|---|---|---|---|
| **G3’** | **G12’** | **G123’** | **G247’** |

**Target genes: estimated values**

| | | | |
|---|---|---|---|
| **G3’** | **G180’** | **G256’** | **G258’** |

Class of new patient = the class of the network with less difference between the true and predicted expression value

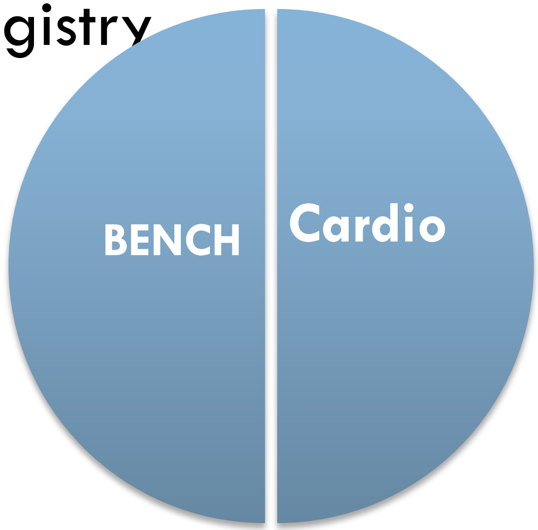
# II SATuRNo. Results

A. <u>Performance Evaluation</u>

- Benchmark methods
- Benchmark dataset

B. <u>Data Analysis</u>

- Luxembourg Myocardial Infarction Registry



# II SATuRNo. Performance

## A) Performance Evaluation

- Benchmark Dataset (Dunkley et al. 2006)
  - 13 control
  - 20 disease
- Benchmark Methods
  - Network inference: correlation-based method
  - Classification tasks: IB1, C4.5 and Naive Bayes

# II SATuRNo. Performance

## A) Performance Evaluation

| | **SATuRNo** | **PC-based method** | | |
|---|---|---|---|---|
| | | IB1 | C4.5 | NB |
| Genes | 34 | 29 | | |
| Acc. | **90.9%** | 87.87% | 78.78% | 87.87% |

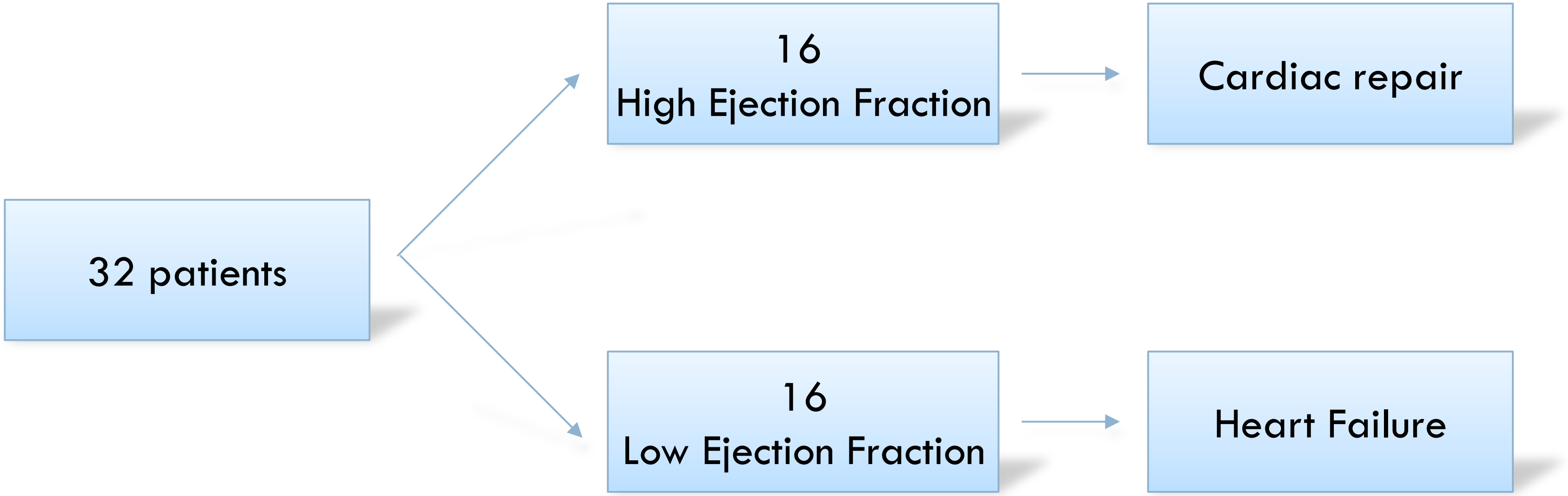
# II SATuRNo. Data Analysis

## B) Data Analysis

- National Registry:
  - Laboratory of Cardiovascular Research
  - Centre Hospitalier of Luxembourg
- Patients with acute MI undergoing primary PCI
- Blood samples for RNA isolation taken at the time of PCI (day 0)
- Left ventricular function evaluated with echocardiography at day 30. (Ejection Fraction EF)

# II SATuRNo. Data Analysis

- Data Analysis


32 patients

16
High Ejection Fraction

Cardiac repair

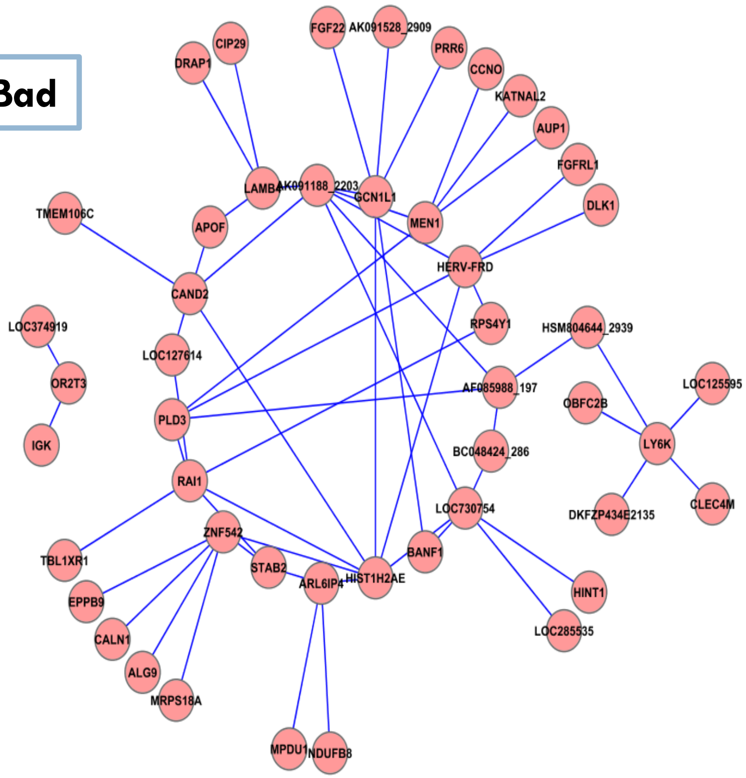
16
Low Ejection Fraction

Heart Failure

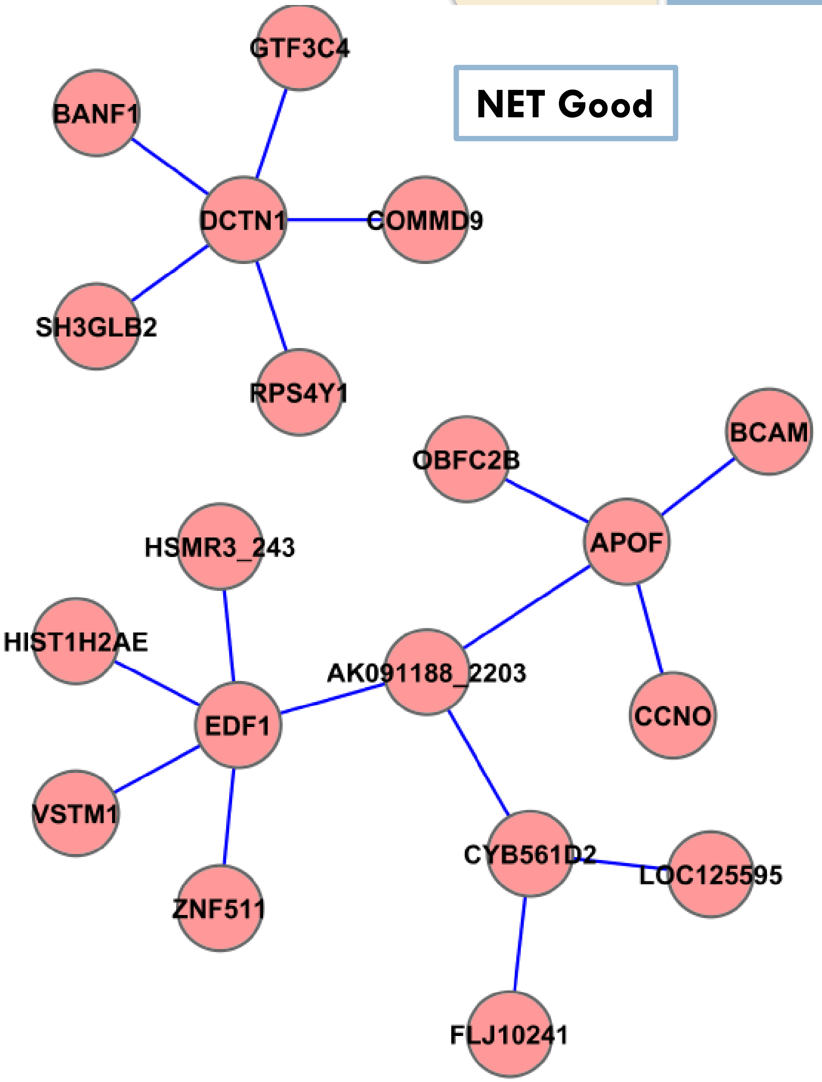
# II SATuRNo. Data Analysis

- 72.145%

**NET Bad**



**NET Good**



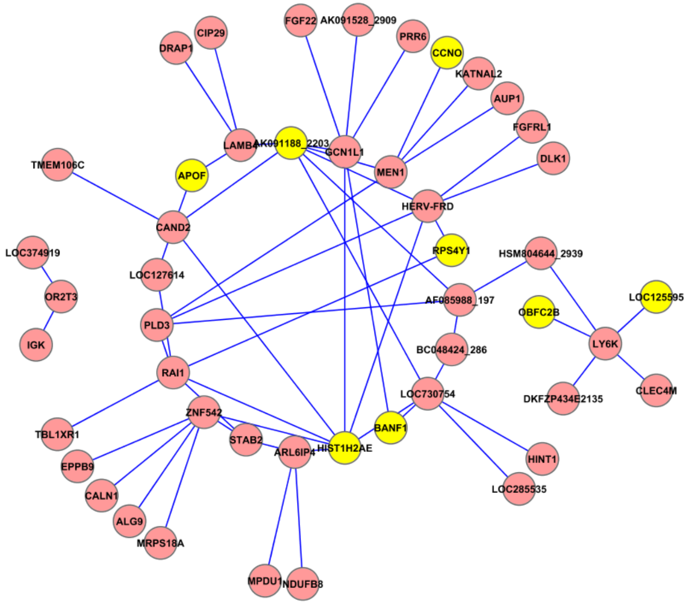
# II SATuRNo. Data Analysis

**Genes in common:**

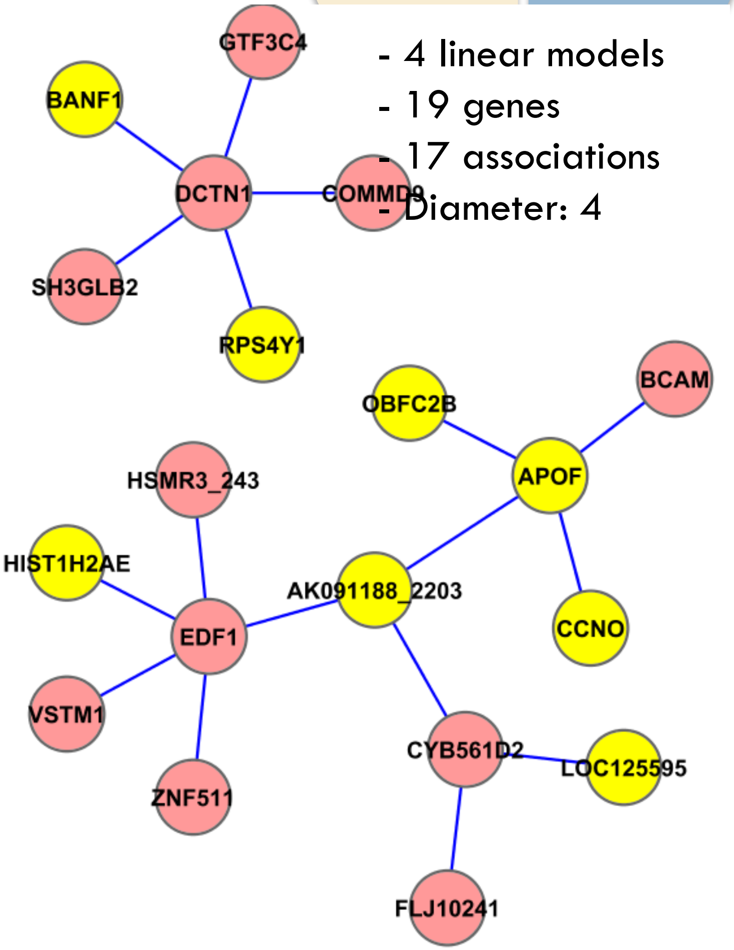
AK091188_2203
APOF
BANF1
CCNO
HIST1H2AE
LOC125595
OBFC2B
RPS4Y1

- 12 linear models
- 48 genes
- 59 associations
- Diameter: 8

**NET Bad**



- 4 linear models
- 19 genes
- 17 associations
- Diameter: 4

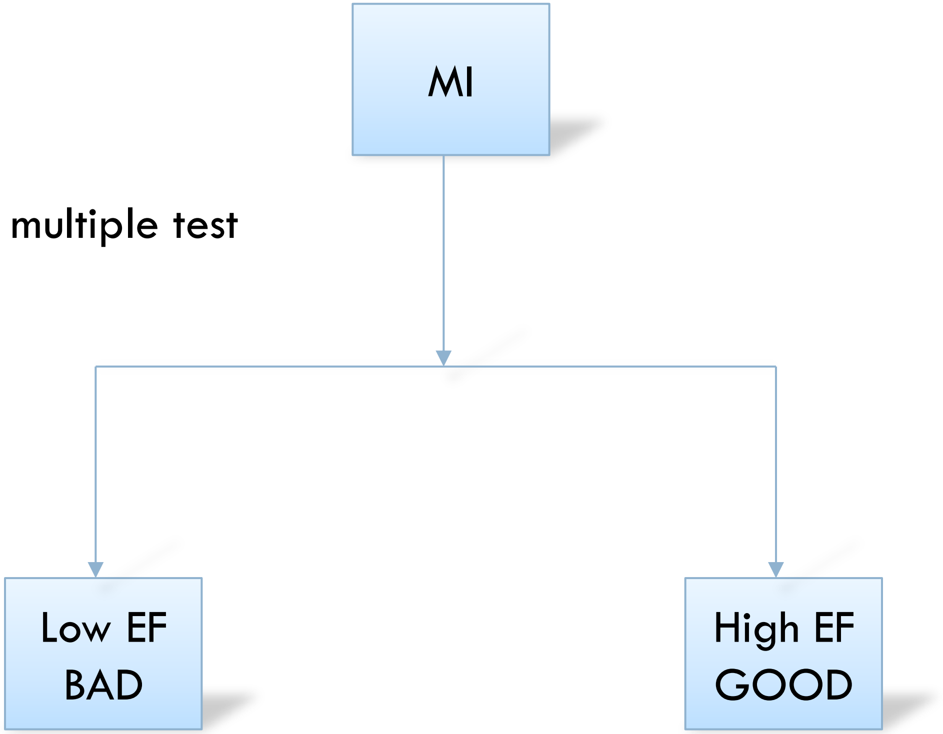


**NET Good**

# II SATuRNo. Data Analysis

## Functional Characterization

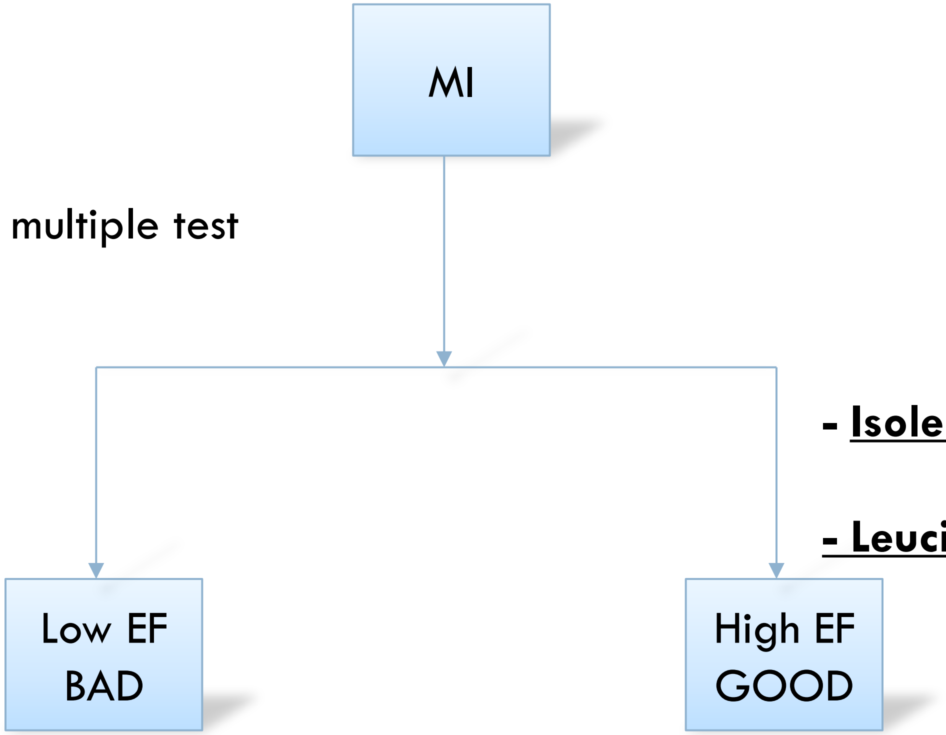
GO analysis:
-Fisher’s exact test
-Benjamin-Hochberg multiple test correction



# II SATuRNo. Data Analysis

## Functional Characterization

GO analysis:
-Fisher's exact test
-Benjamin-Hochberg multiple test correction



**- Isoleucine catabolic process**

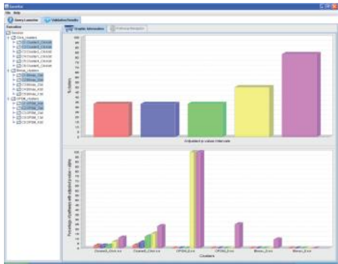
**- Leucine biosynthetic process**

# Outline

1. Introduction
2. Gene Network Inference: state-of-the art
3. Proposals & Results:
    a) RegNet: Gene Regression Networks
    b) SATuRNo: Supervised Prognostic Approach Through Regression Networks
    c) CarGene: Characterization of genes
4. Conclusions and future work
5. Project membership

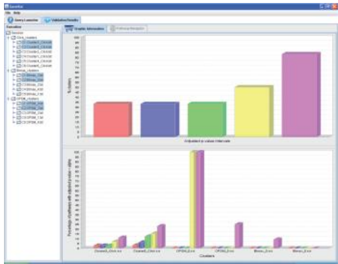
# II CARGENE

- CARGENE:
  - Characterisation set of Genes using Kegg
- AIM
  - Enrichment Analysis of Kegg Pathways
- Software Tool
  - 3 components
- In collaboration with
  - Domingo S. Rodriguez-Baena
  - Norberto Díaz-Díaz



# II CARGENE

- Software Tool
  1. Web service interface
  2. Multihread and Visualization
  3. Statistical software components
     - Fisher's exact test
     - Bonferroni correction
     - Westfall and Young using Monte Carlo simulations



# Outline

1. Introduction
2. Gene Network Inference: state-of-the art
3. Proposals & Results:
    a) RegNet: Gene Regression Networks
    b) SATuRNo: Supervised Prognostic Approach Through Regression Networks
    c) CarGene: Characterization of genes
4. Conclusions and future work
5. CV & Contributions

# Conclusions

RegNet

- Generates new hypothesis of interactions among genes from gene expression data
- Favours localized similarities (drawbacks of correlation- based methods)
- Experimental results show good
- In general, REGNET is a powerful method to hypothesize on unknown relationships

# Conclusions

SATuRNo

- Network-based prognostic approach
  - Based on the discovery of clinically relevant transcriptional association networks
- Can provide insights into the interplay of genes and their association with clinical phenotypes
- Insights into underlying molecular mechanisms to characterize and possibly treat the development of ventricular dysfunction.

# Conclusions

- CarGene
  - Characterisation set of Genes using KEGG
- Future work
  - Network Enrichment Analysis

# Future works

- RegNet
  - Regression Networks from time series datasets
  - Step III: statistical procedure
- RegNet + SATuRNo
  - Integration of different types of omic data to our approches RegNet and SATuRNo: DNA sequencing or pp interactions
  - Apply our methods to other biomedical applications as Alzheimer Disease

# Outline

1. Introduction
2. Gene Network Inference: state-of-the art
3. Proposals & Results:
    a) RegNet: Gene Regression Networks
    b) SATuRNo: Supervised Prognostic Approach Through Regression Networks
    c) CarGene: Characterization of genes
4. Conclusions and future work
5. Project membership

# Project membership

- Modelos Avanzados en Minería de Datos: Escalabilidad y Aplicación Biológica (07/12)
    - IP: Dr Jesús S. Aguilar Ruiz
- Heurísticas escalables para la extracción de conocimiento en grandes volúmenes de información (07/10)
    - IP: Dr José Riquelme Santos
- Red Española de Minería de Datos (05/06)
    - IP: Dr José Riquelme Santos

MINISTERIO
DE EDUCACIÓN
Y CIENCIA

MINISTERIO
DE EDUCACIÓN
Y CIENCIA

MINISTERIO
DE EDUCACIÓN
Y CIENCIA

# Project membership

- MINERVA: Técnicas emergentes de minería de datos para la extracción de conocimiento en grandes volúmenes de información: aplicación a datos científicos e industriales (06/07)
  - IP: Dr José Riquelme Santos

MINISTERIO
DE EDUCACIÓN
Y CIENCIA

# Project membership

- Knowledge Discovery Network of Excellence (06)
  - Dr. Michael May and Dr. Codrina Lauth
- TIC134 (Plan Andaluz de Investigaicón)
- Mindat-Plus: Minería de datos para los usuarios en diferentes áreas de aplicación (06/08)
  - IP: Dr Francisco Herrera Triguero

# Funded

- Stay at the Laboratory of Cardiovascular research
  - (August-September 2009)
  - (May-June 2010)
- Funded by:
  - Junta de Andalucía
  - Plan Propio (Universidad de Sevilla)
  - Fonds National de la Recherche Luxembourg

# THANK YOU
# FOR YOUR ATTENTION