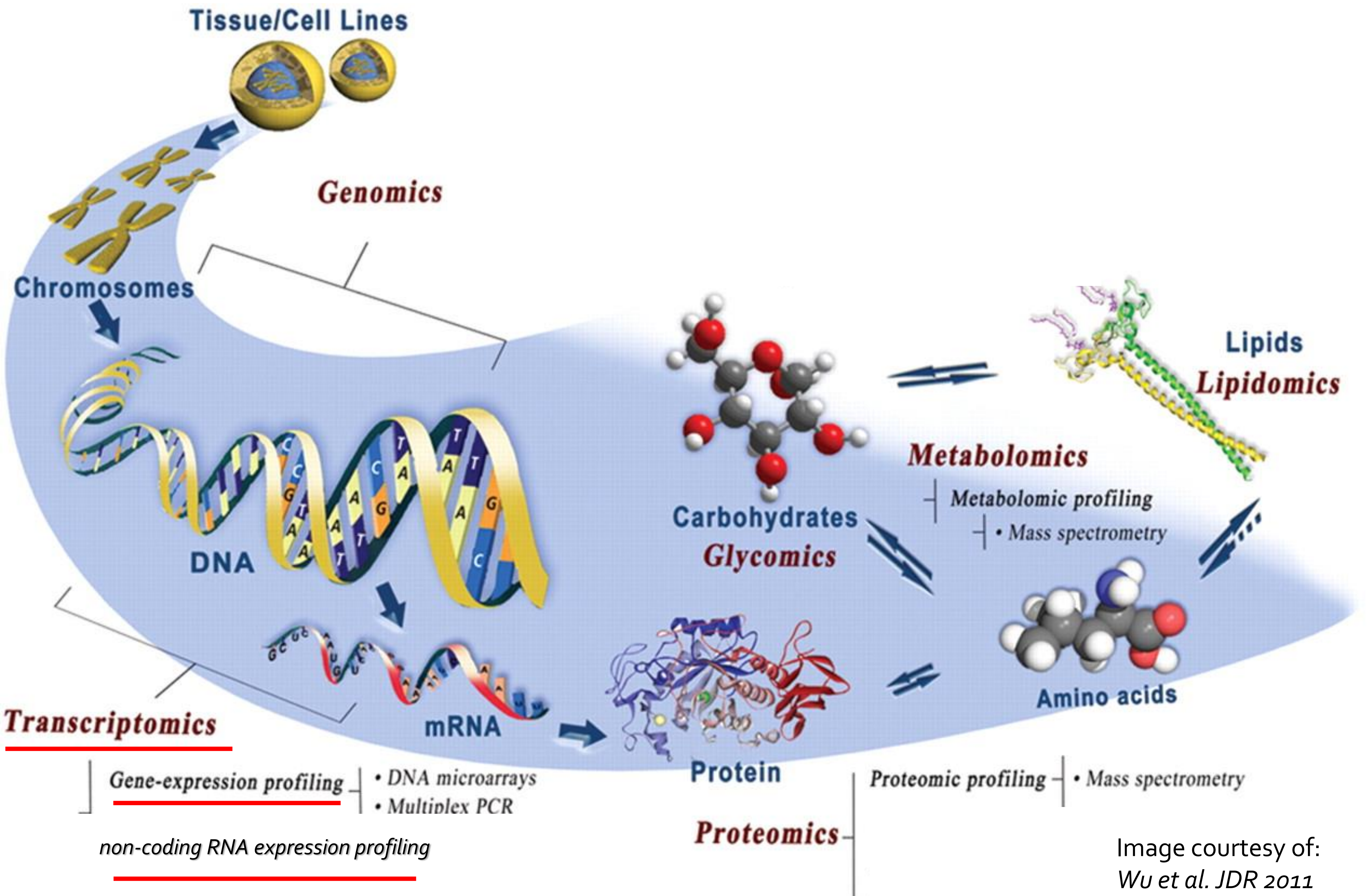
# Pathway analysis: introduction and discussion

Francesca M. Buffa

Department of Oncology
University of Oxford

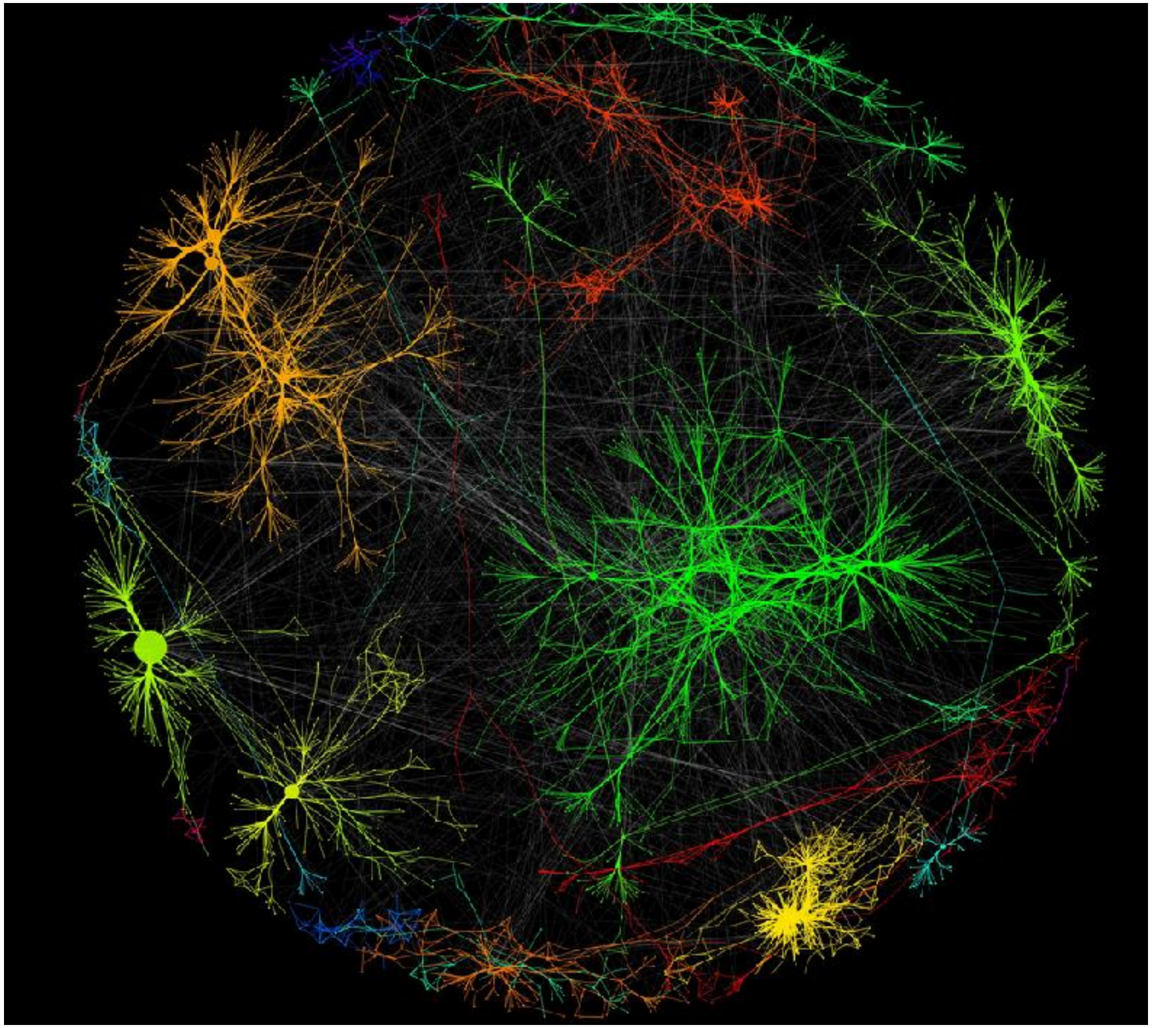
# Complex data structure

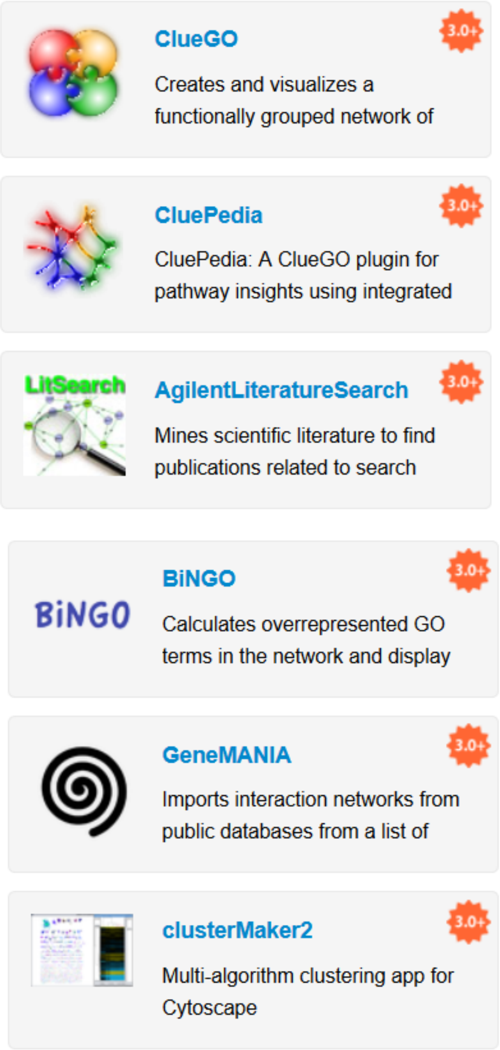


*non-coding RNA expression profiling*

Image courtesy of:
*Wu et al. JDR 2011*

# Cytoscape





**ClueGO**

Creates and visualizes a functionally grouped network of

**CluePedia**

CluePedia: A ClueGO plugin for pathway insights using integrated

**AgilentLiteratureSearch**

Mines scientific literature to find publications related to search

**BiNGO**

Calculates overrepresented GO terms in the network and display

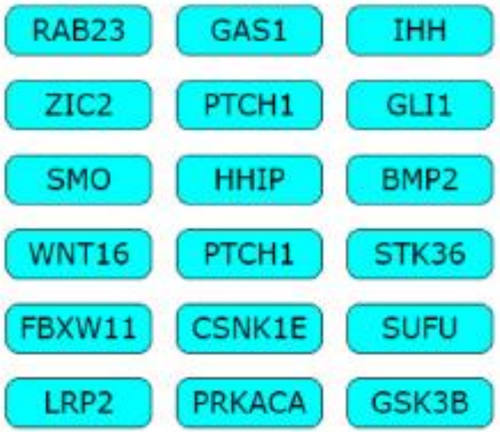
**GeneMANIA**

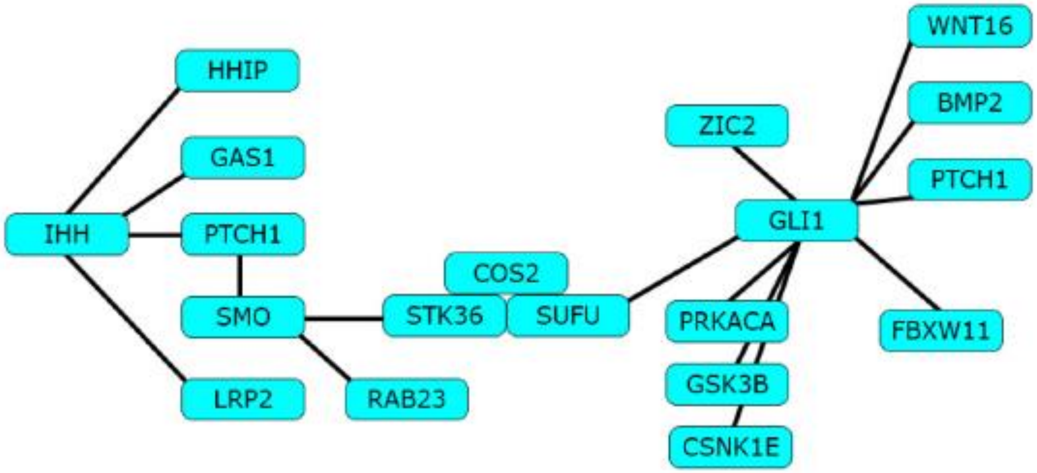
Imports interaction networks from public databases from a list of

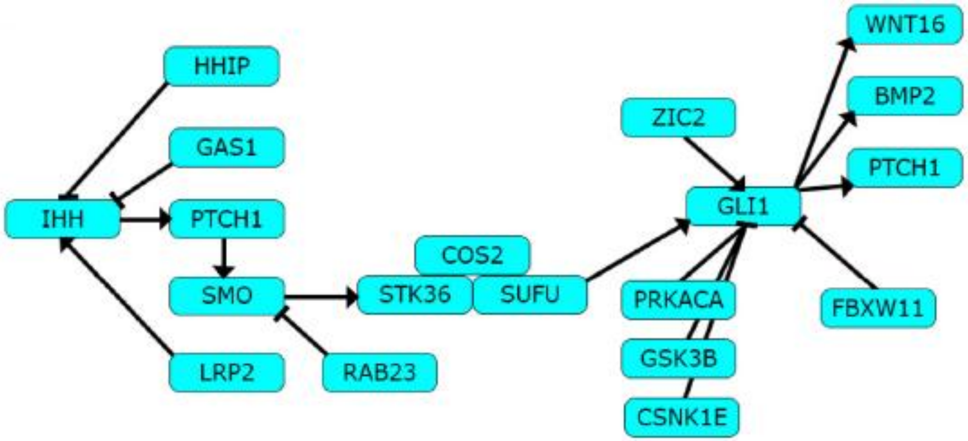
**clusterMaker2**

Multi-algorithm clustering app for Cytoscape

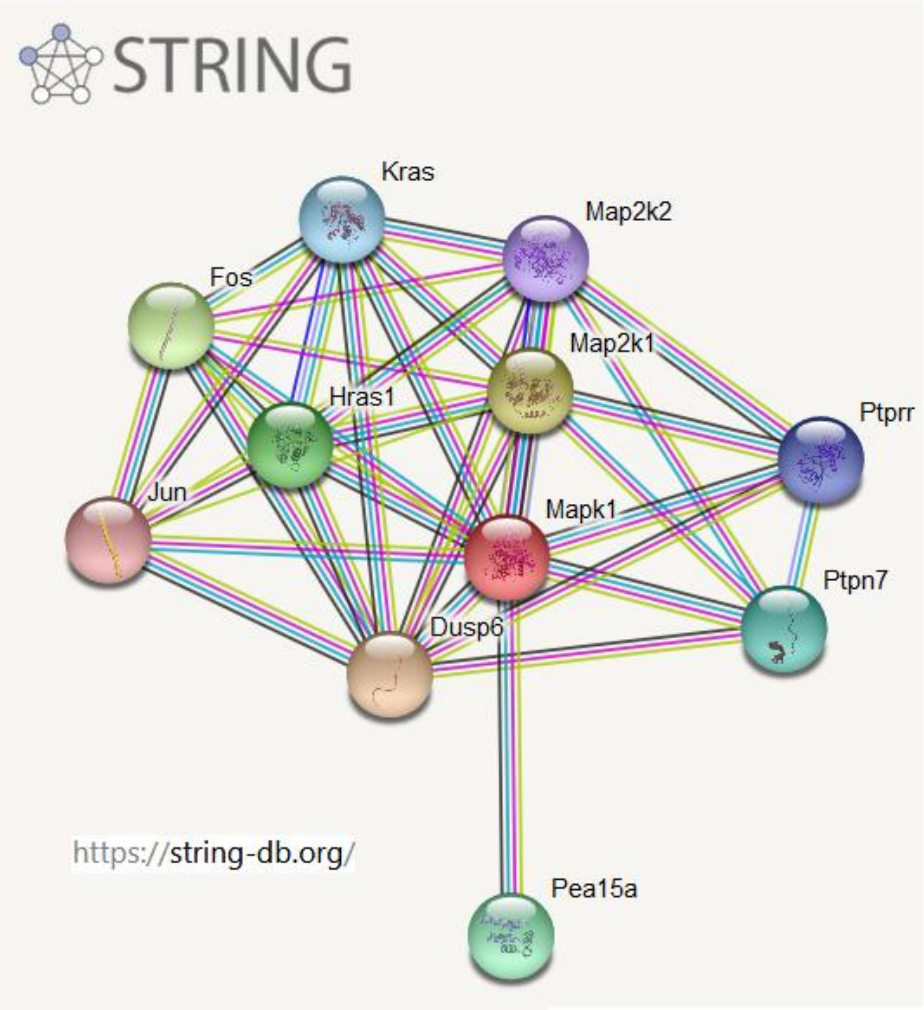
# How do we represent a pathway



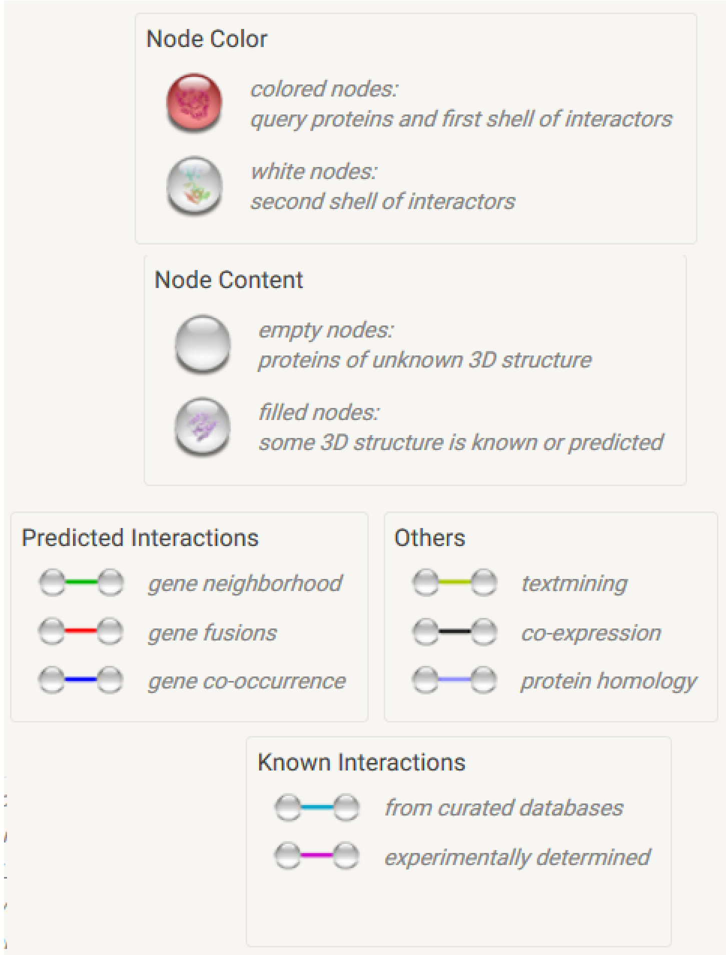




# STRING



https://string-db.org/



Node Color

*colored nodes:*
*query proteins and first shell of interactors*

*white nodes:*
*second shell of interactors*

Node Content

*empty nodes:*
*proteins of unknown 3D structure*

*filled nodes:*
*some 3D structure is known or predicted*

Predicted Interactions

*gene neighborhood*

*gene fusions*

*gene co-occurrence*

Others

*textmining*

*co-expression*

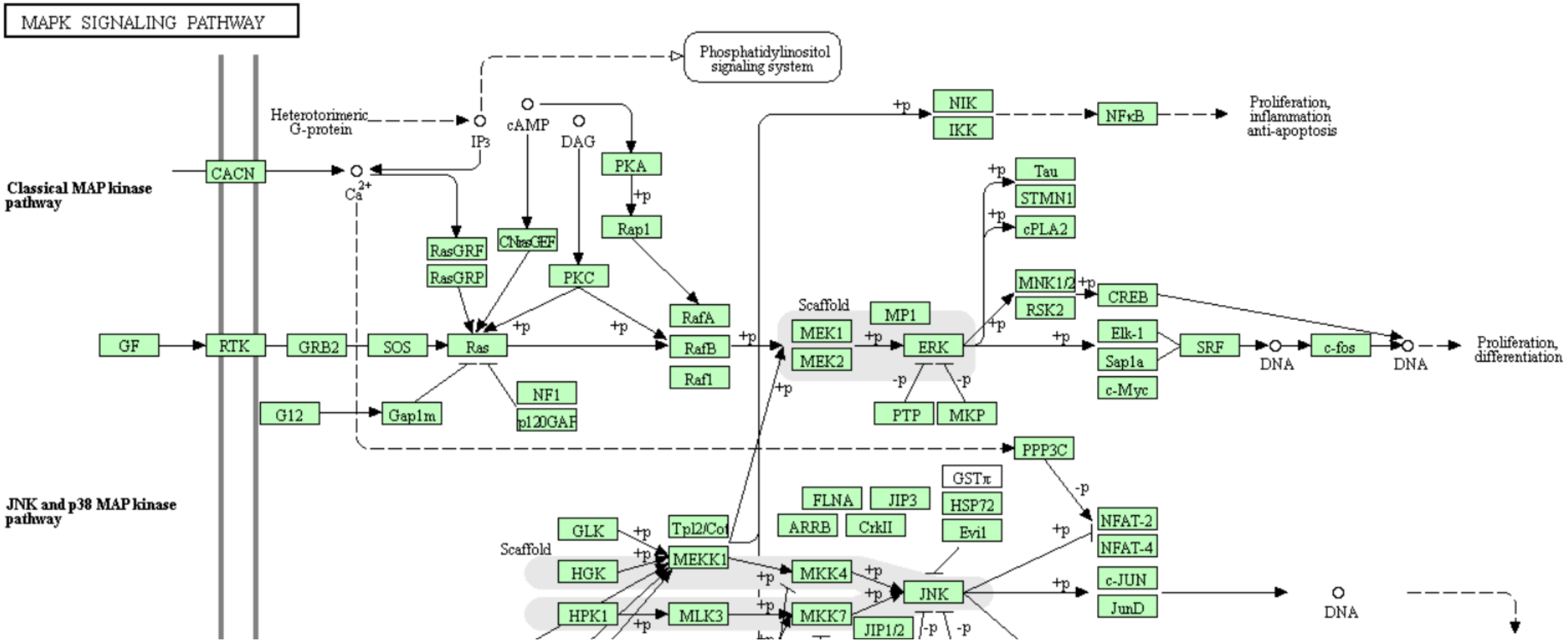
*protein homology*

Known Interactions

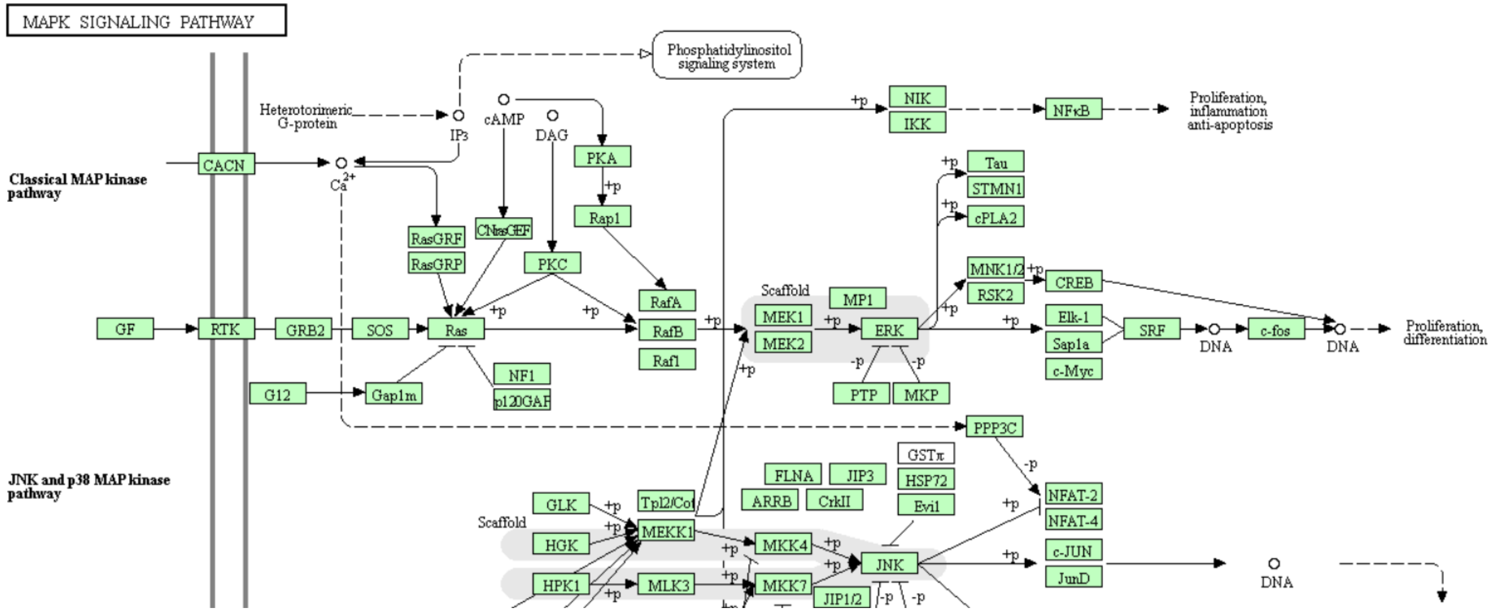
*from curated databases*

*experimentally determined*

# Kegg pathways



# Kegg pathways



**Nodes**

- Genes
- Group of genes
- Compounds
- Other networks

**Edges**

Activation/Inhibition

Expression

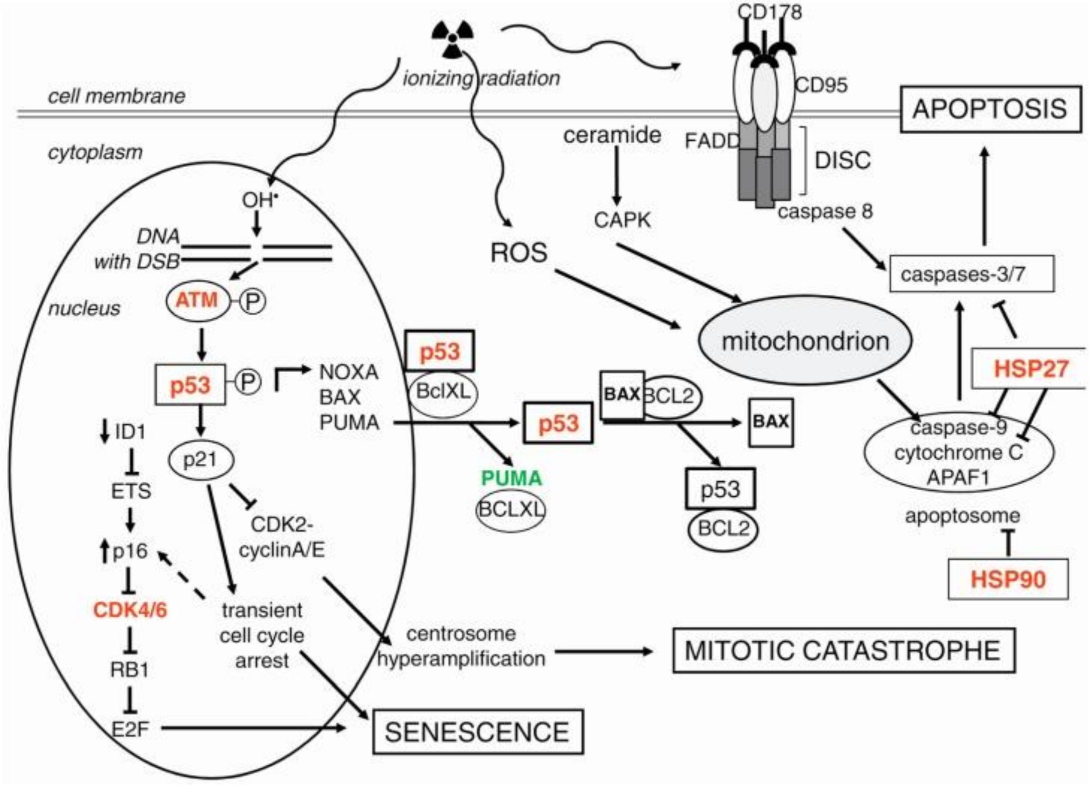
Indirect

Phosphorylation/Diphosphorilation

Ubiquination

Association/Dissocation

# Many repositories of biological pathways



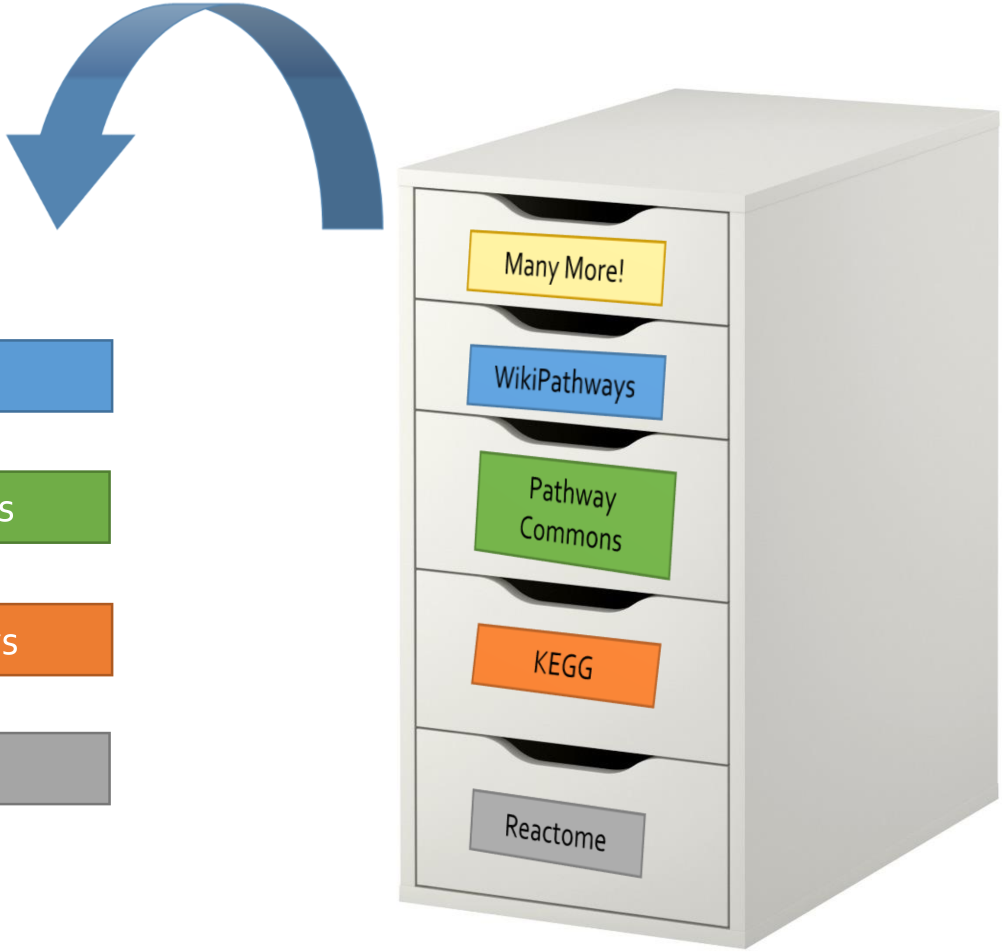
# Many repositories of biological pathways

2471 Pathways

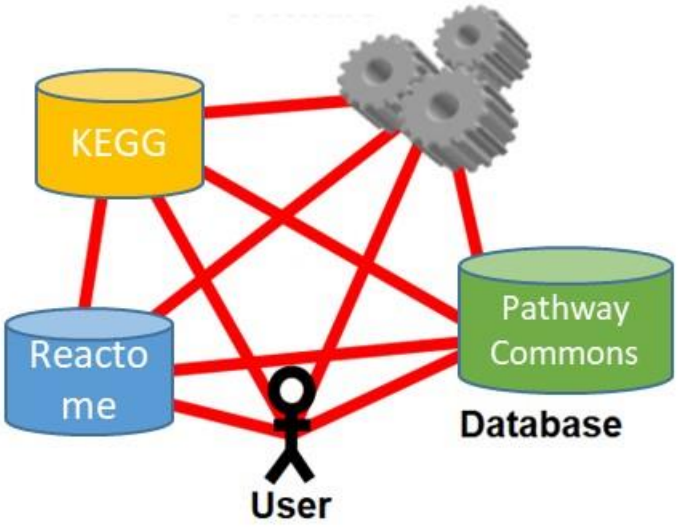
>42 000 Pathways

505, 700 Pathways

2132 (H. sapiens)

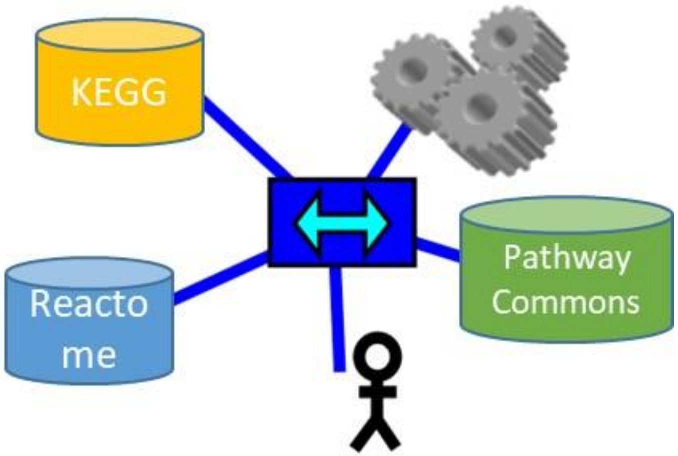


# Biological Pathway Exchange (BioPAX)



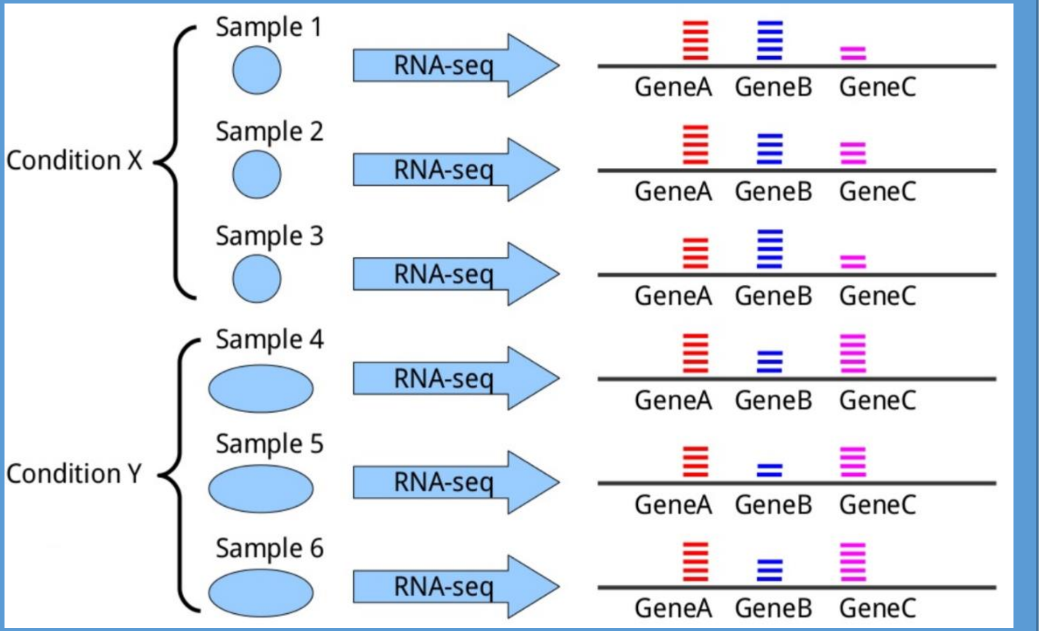
Before BioPAX

>100 DBs and tools
Tower of Babel



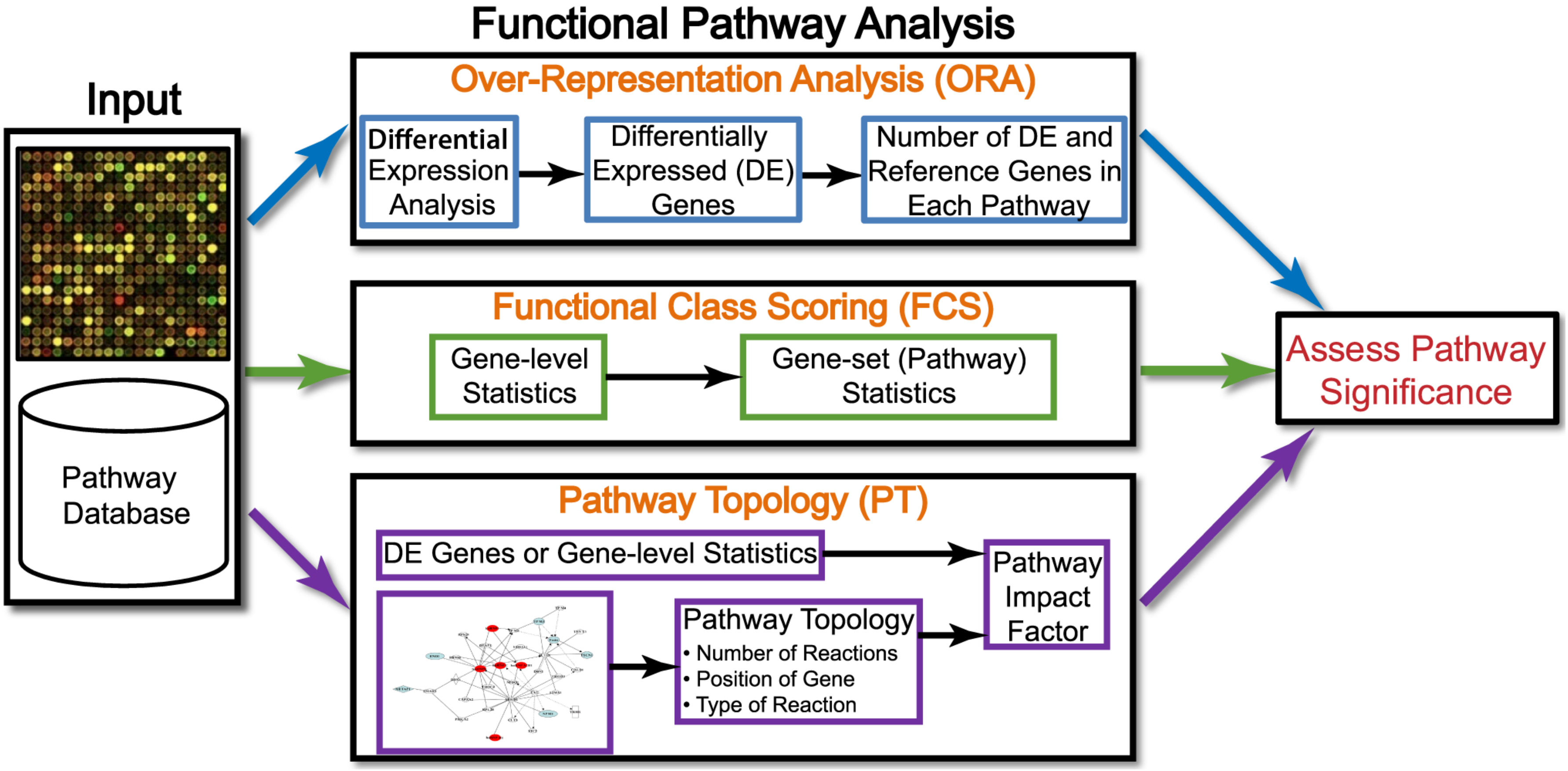
After BioPAX
Unifying language

# Pathway analysis



| affyID | gene.name | accession | unigene | FC | pfp |
|---|---|---|---|---|---|
| 1419476_at | Adamdec1 | NM_021475 | Mm.36742 | 27.31 | 0 |
| 1448162_at | Vcam1 | BB250384 | Mm.76649 | 26.58 | 2.44 |
| 1419128_at | Itgax | NM_021334 | Mm.22378 | 17.06 | 3.83 |
| 1415989_at | Vcam1 | BB250384 | Mm.76649 | 13.57 | 0 |
| 1418776_at | 5830443L24Rik | NM_029509 | Mm.425261 | 11.76 | 0 |
| | | | Mm.37018 | | |

# Many ways to approach pathway analysis



Functional Pathway Analysis

Input

Pathway Database

Over-Representation Analysis (ORA)

Differential Expression Analysis

Differentially Expressed (DE) Genes

Number of DE and Reference Genes in Each Pathway

Functional Class Scoring (FCS)

Gene-level Statistics

Gene-set (Pathway) Statistics

Pathway Topology (PT)

DE Genes or Gene-level Statistics

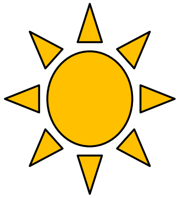
Pathway Topology
- Number of Reactions
- Position of Gene
- Type of Reaction

Pathway Impact Factor

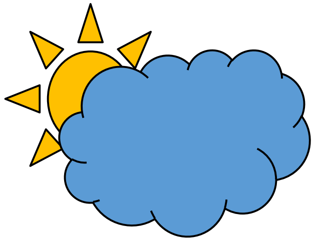
Assess Pathway Significance

Khatri P, Sirota M, Butte AJ (2012) Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Comput Biol 8(2): e1002375. doi:10.1371/journal.pcbi.1002375
http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1002375

# Over-representation analysis



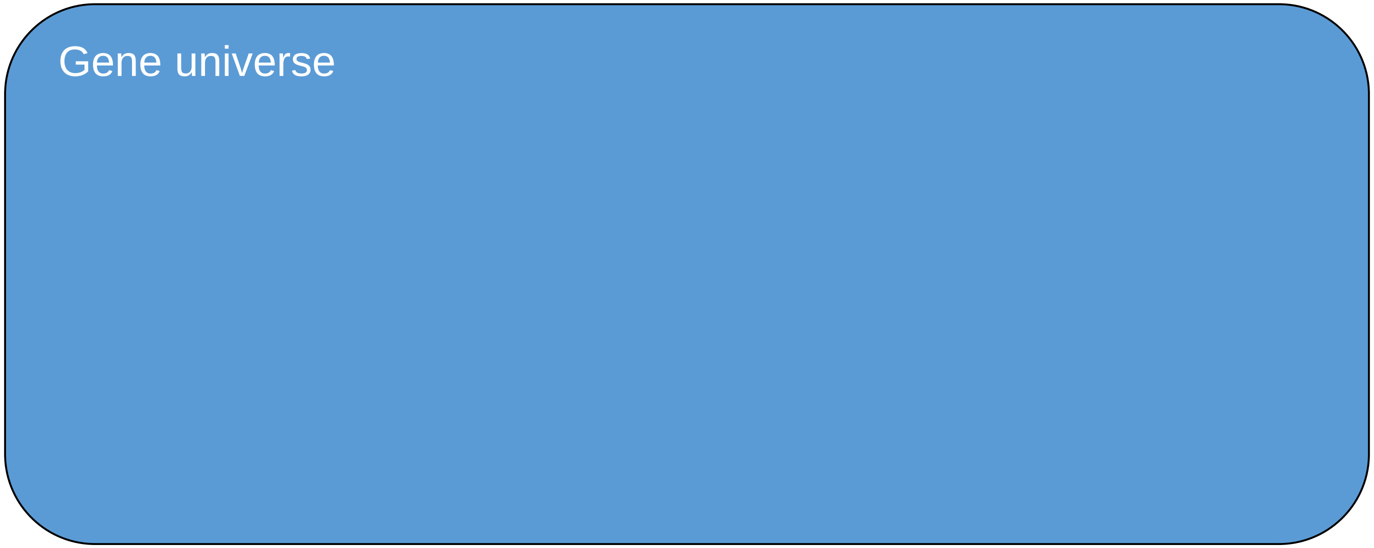
We have mapped our significantly differentially expressed genes to pathways. So we can start to interpret our results.



How likely is it that if we consider a random set of genes we will observe these pathways?

# Over-representation analysis

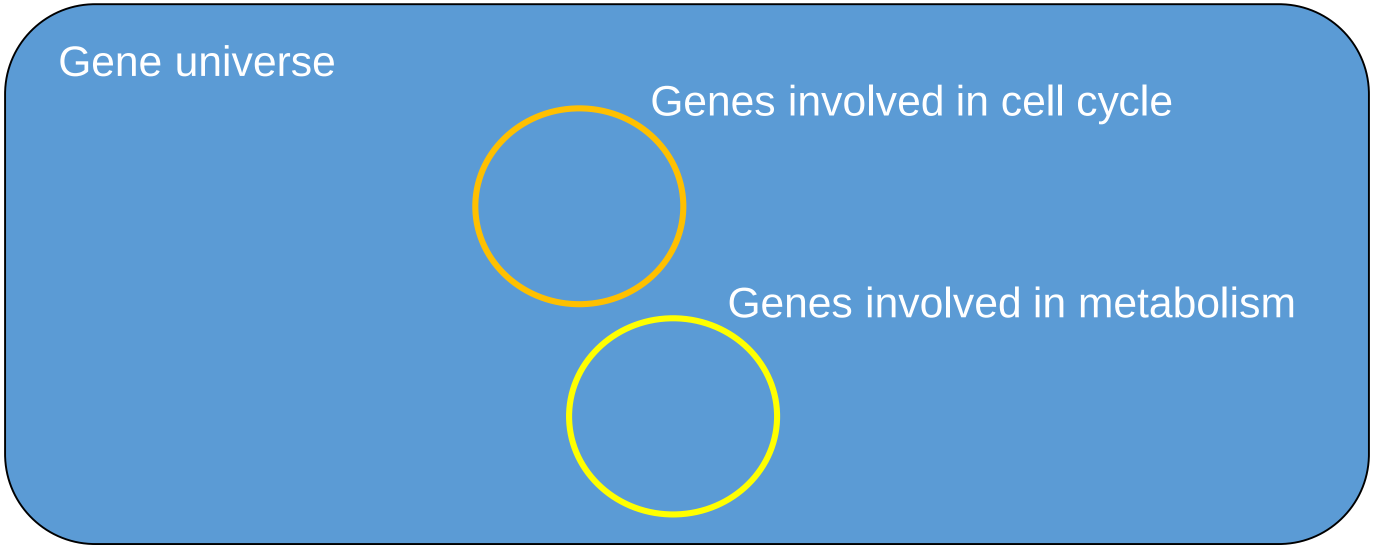
Are there any pathways that have a larger than expected subset of our selected genes in their annotation list?



Gene universe

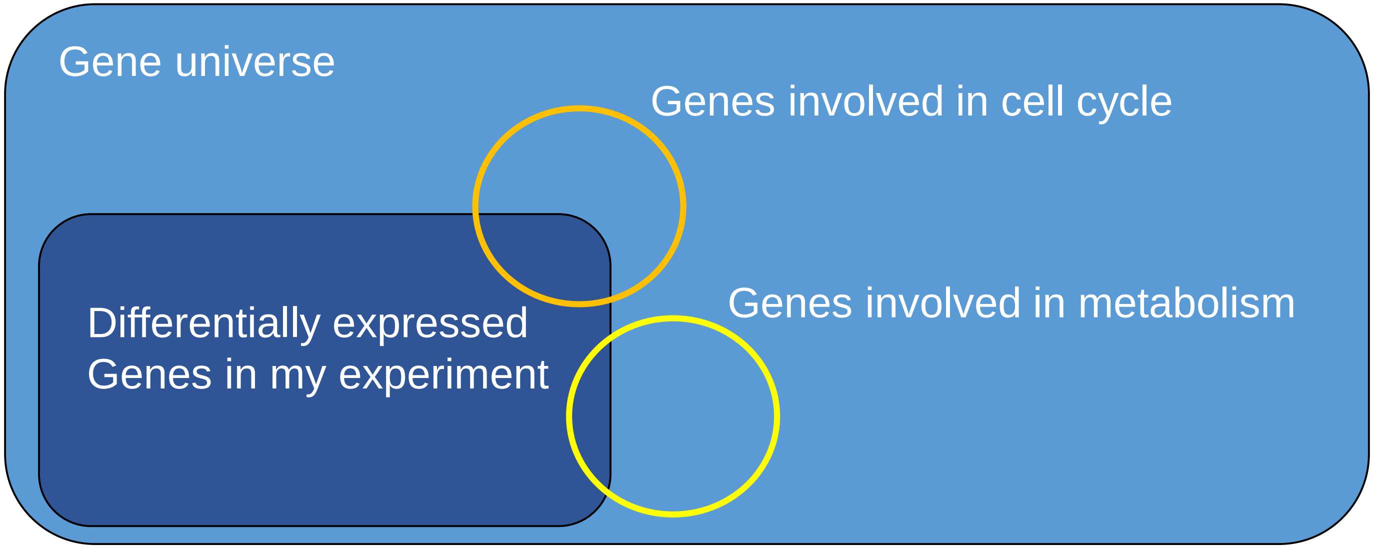
# Over-representation analysis

Are there any pathways that have a larger than expected subset of our selected genes in their annotation list?



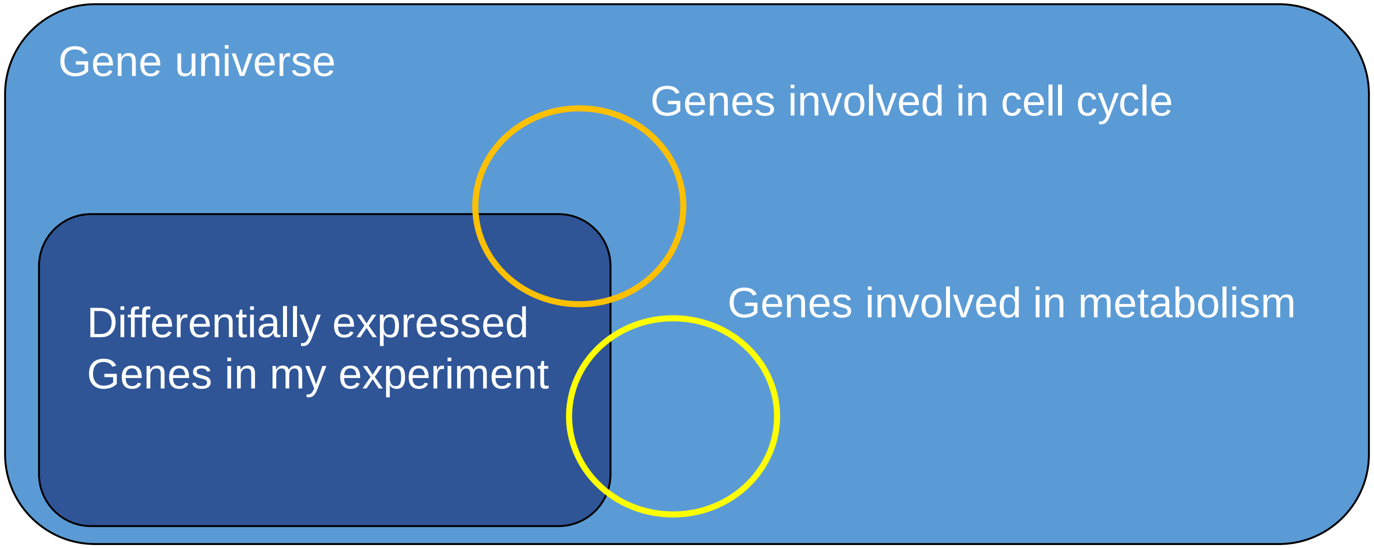
# Over-representation analysis

Are there any pathways that have a larger than expected subset of our selected genes in their annotation list?



# Over-representation analysis

Are there any pathways that have a larger than expected subset of our selected genes in their annotation list?

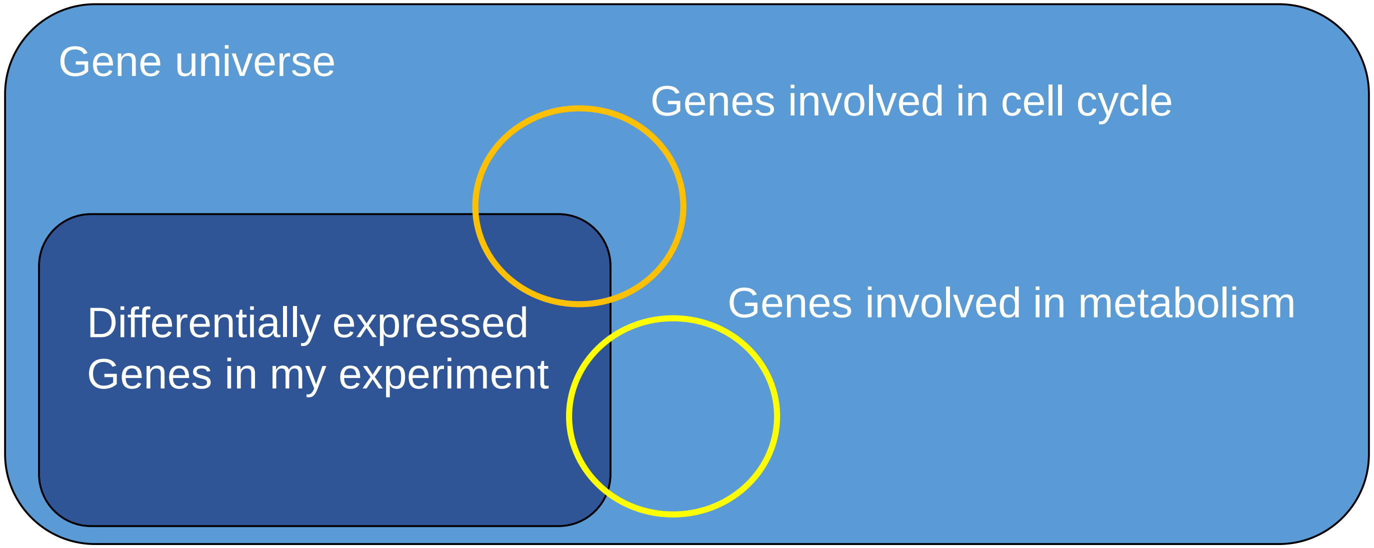

Gene universe
Genes involved in cell cycle
Differentially expressed Genes in my experiment
Genes involved in metabolism

Two-way table:

| | Selected | Universe |
|---|---|---|
| In Pathway | ? | ? |
| Not In Pathway | ? | ? |
| Total | ? | ? |

# Over-representation analysis

Are there any pathways that have a larger than expected subset of our selected genes in their annotation list?

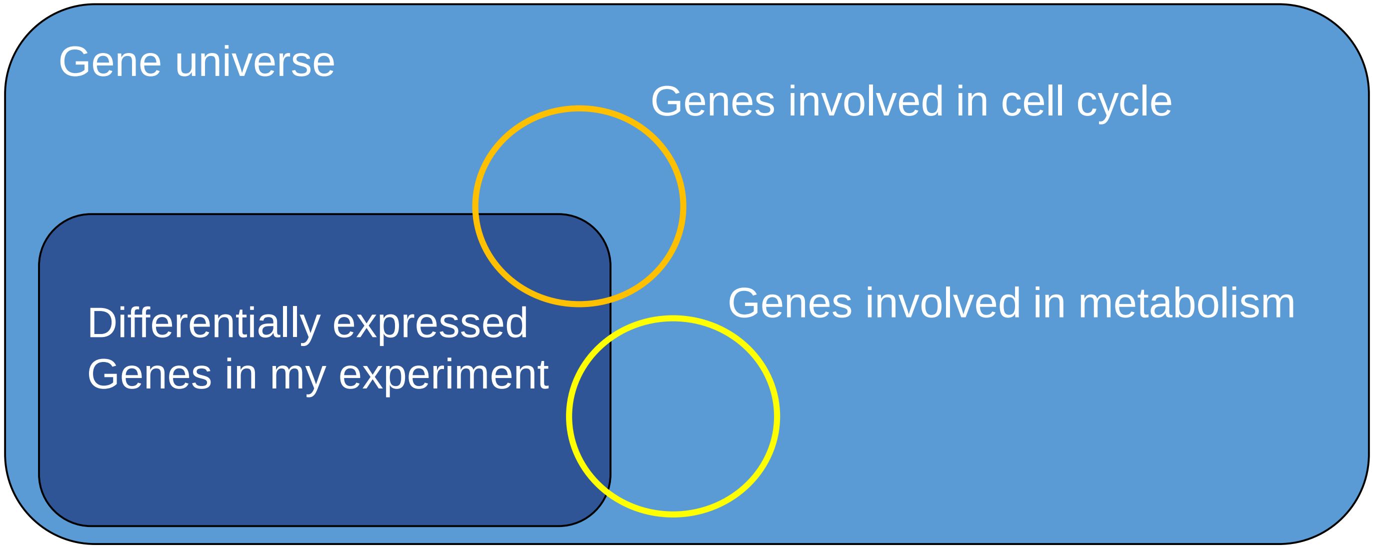

Gene universe
Genes involved in cell cycle
Differentially expressed Genes in my experiment
Genes involved in metabolism

Two-way table:

| | Selected | Universe |
|---|---|---|
| In Pathway | 22 | 7500 |
| Not In Pathway | 28 | 22500 |
| Total | 50 | 30000 |

# Over-representation analysis

Are there any pathways that have a larger than expected subset of our selected genes in their annotation list?

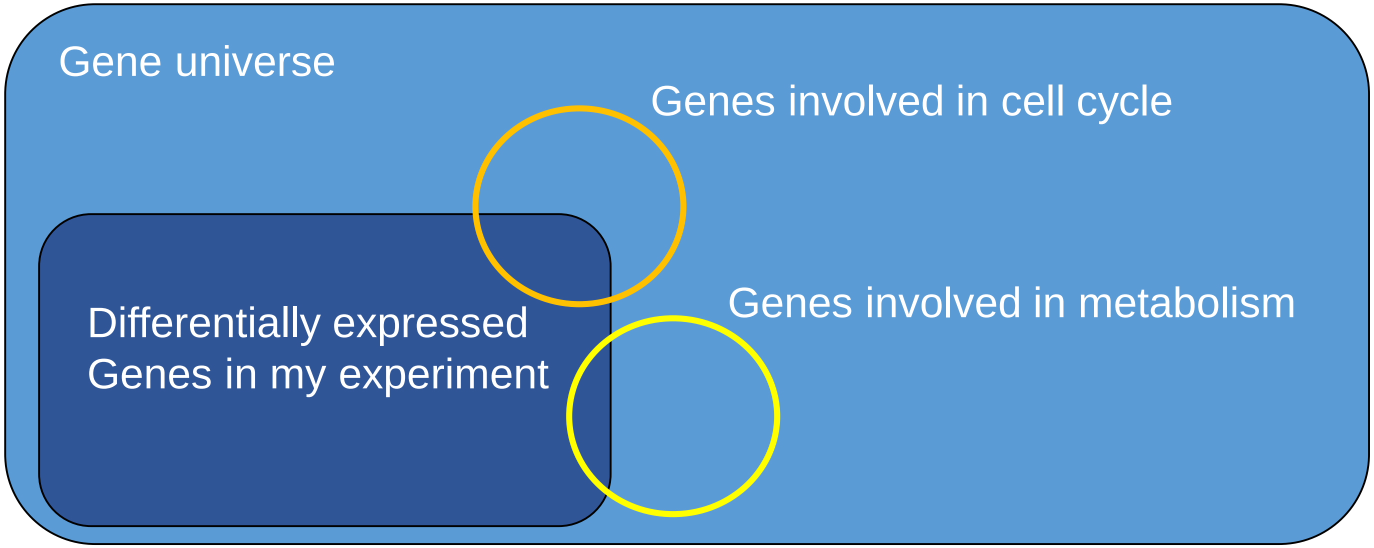

Gene universe
Genes involved in cell cycle
Differentially expressed Genes in my experiment
Genes involved in metabolism

Two-way table:

| | Selected | Universe |
|---|---|---|
| In Pathway | 22 | 7500 |
| Not In Pathway | 28 | 22500 |
| Total | 50 | 30000 |

Fold enrichment = (22/50) / (7500/30000) = 45% / 25% = 1.8

# The Hypergeometric test



"The probability of drawing up to *x* of a possible *K* items in *N* drawings without replacement from a group of *M* objects"

X = the number of differentially expressed genes belonging to the pathway

K = the number of genes belonging to the pathway

N = the differentially expressed genes (or selected genes)

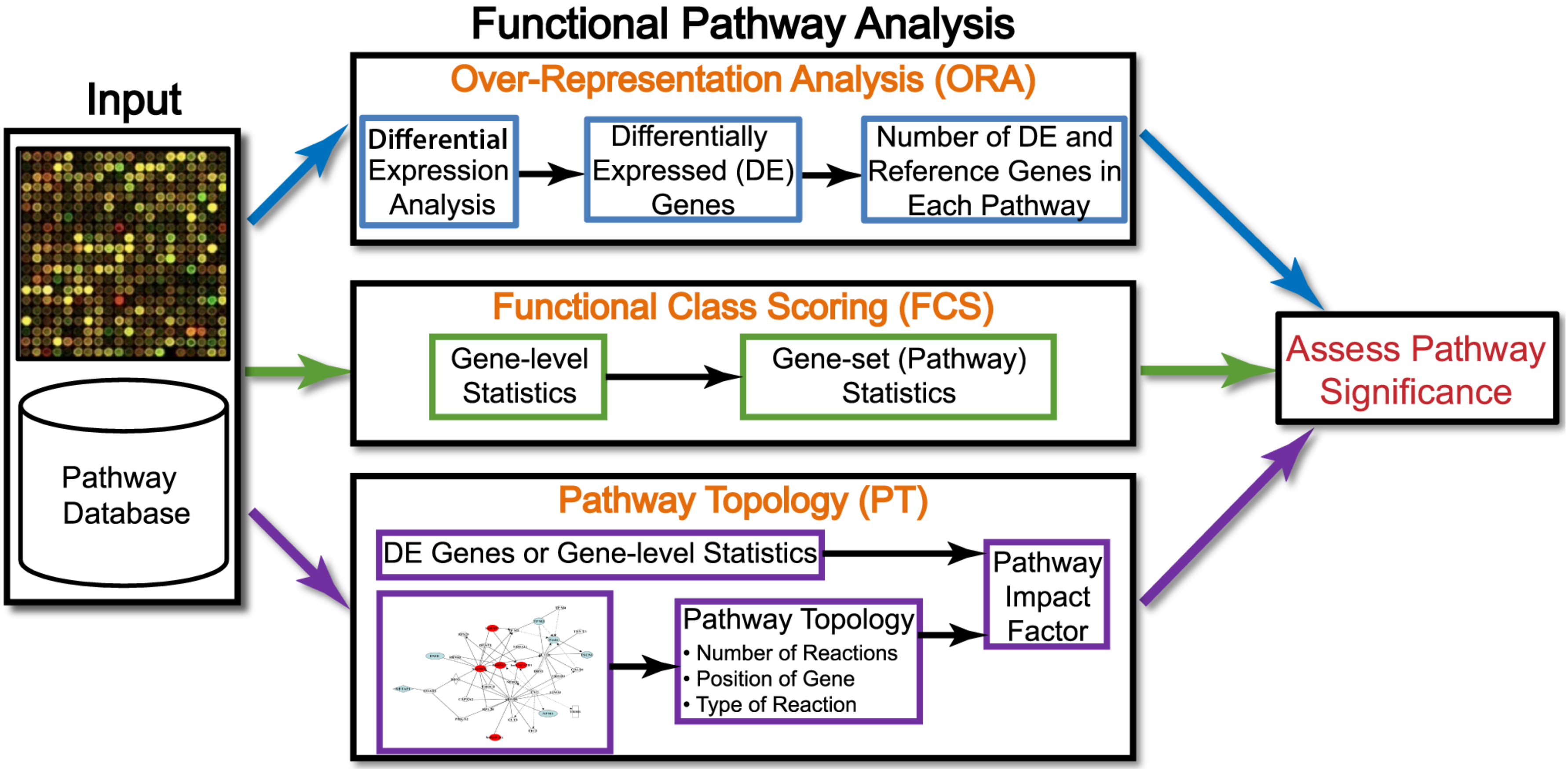
M = the universe

$$p = F(x \mid M, K, N) = \sum_{i=0}^{x} \frac{\binom{K}{i}\binom{M-K}{N-i}}{\binom{M}{N}}$$

# Limitations

- Not always clear how to define the universe
- The over-representation analysis is independent of the changes measured. All genes are treated equally.
- Only the most significant genes are used which causes information loss
- Genes are assumed to be independent and the correlation structure is ignored
- Pathways are assumed to be independent

# Many ways to approach pathway analysis

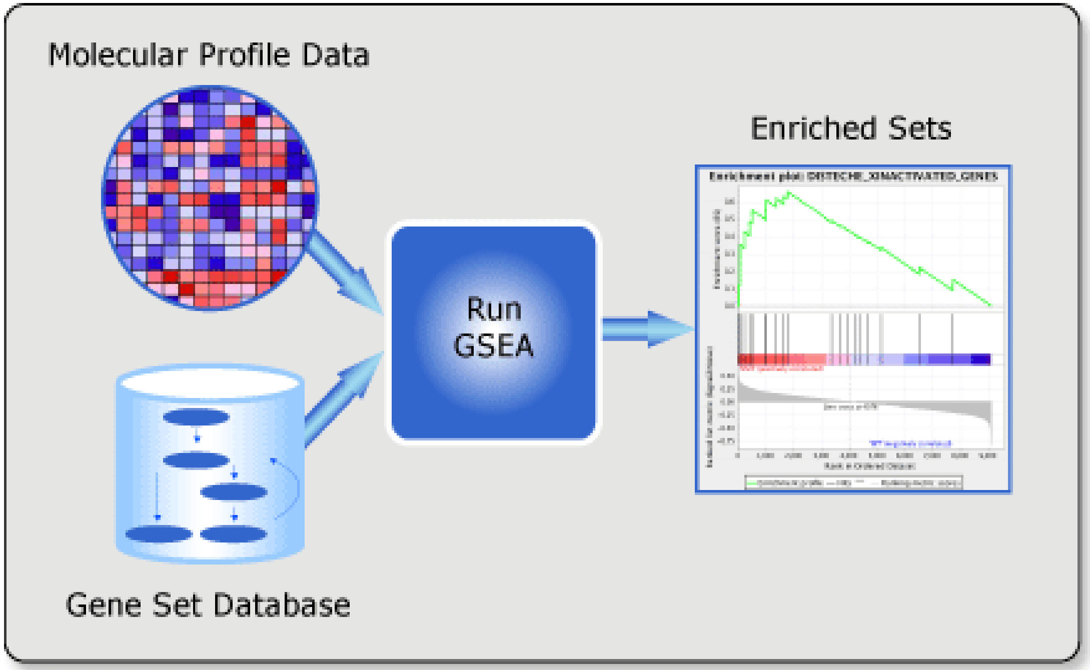

Functional Pathway Analysis
Input
Pathway Database
Over-Representation Analysis (ORA)
Differential Expression Analysis
Differentially Expressed (DE) Genes
Number of DE and Reference Genes in Each Pathway
Functional Class Scoring (FCS)
Gene-level Statistics
Gene-set (Pathway) Statistics
Pathway Topology (PT)
DE Genes or Gene-level Statistics
Pathway Topology
• Number of Reactions
• Position of Gene
• Type of Reaction
Pathway Impact Factor
Assess Pathway Significance

Khatri P, Sirota M, Butte AJ (2012) Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Comput Biol 8(2): e1002375. doi:10.1371/journal.pcbi.1002375
http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1002375

# Gene Set Enrichment Analysis

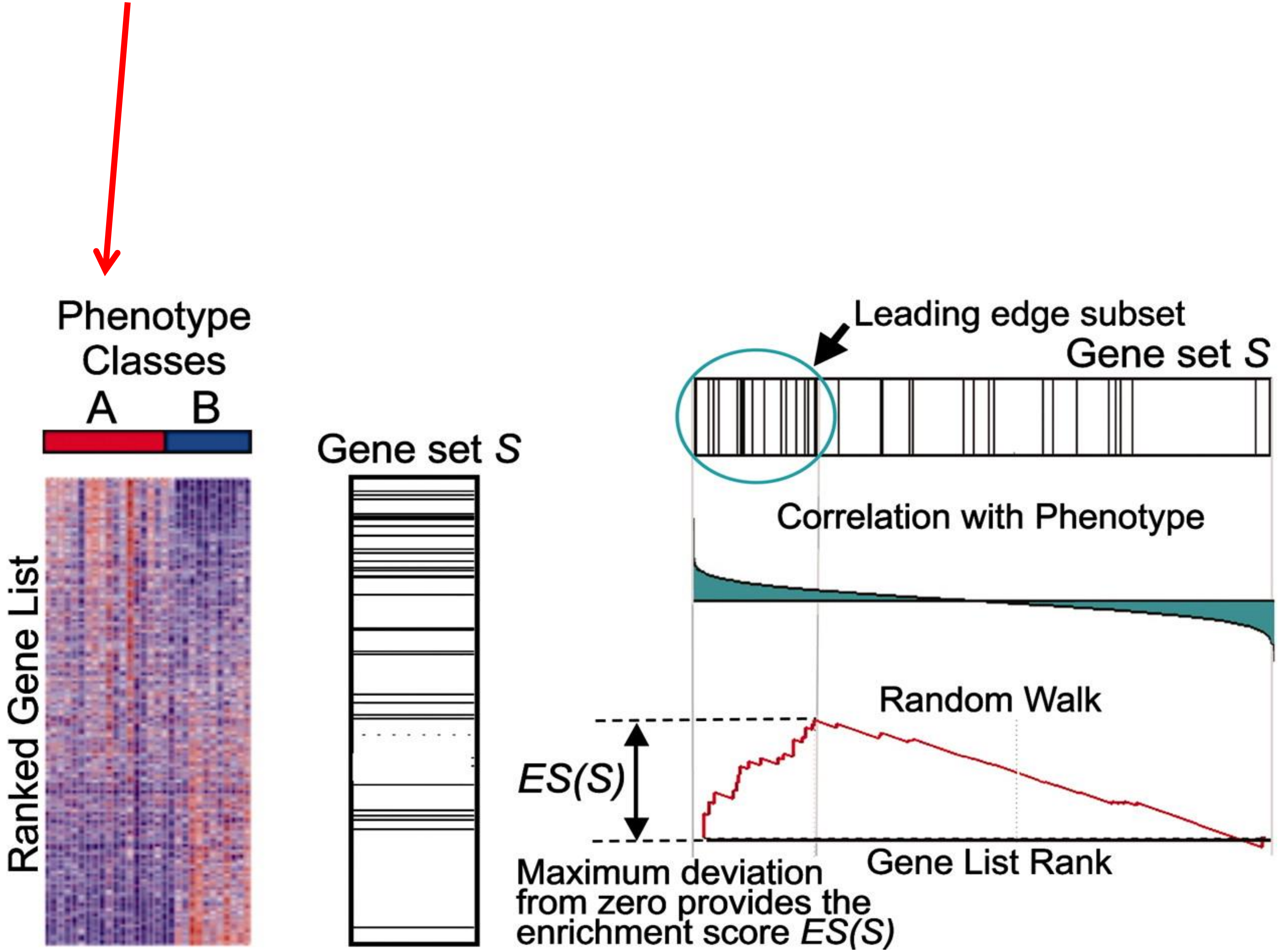
Tests whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes)

Hypothesis:
Pathway regulation can be detected either by looking at large changes in individual genes or by looking at coordinated changes in sets of functionally related genes.

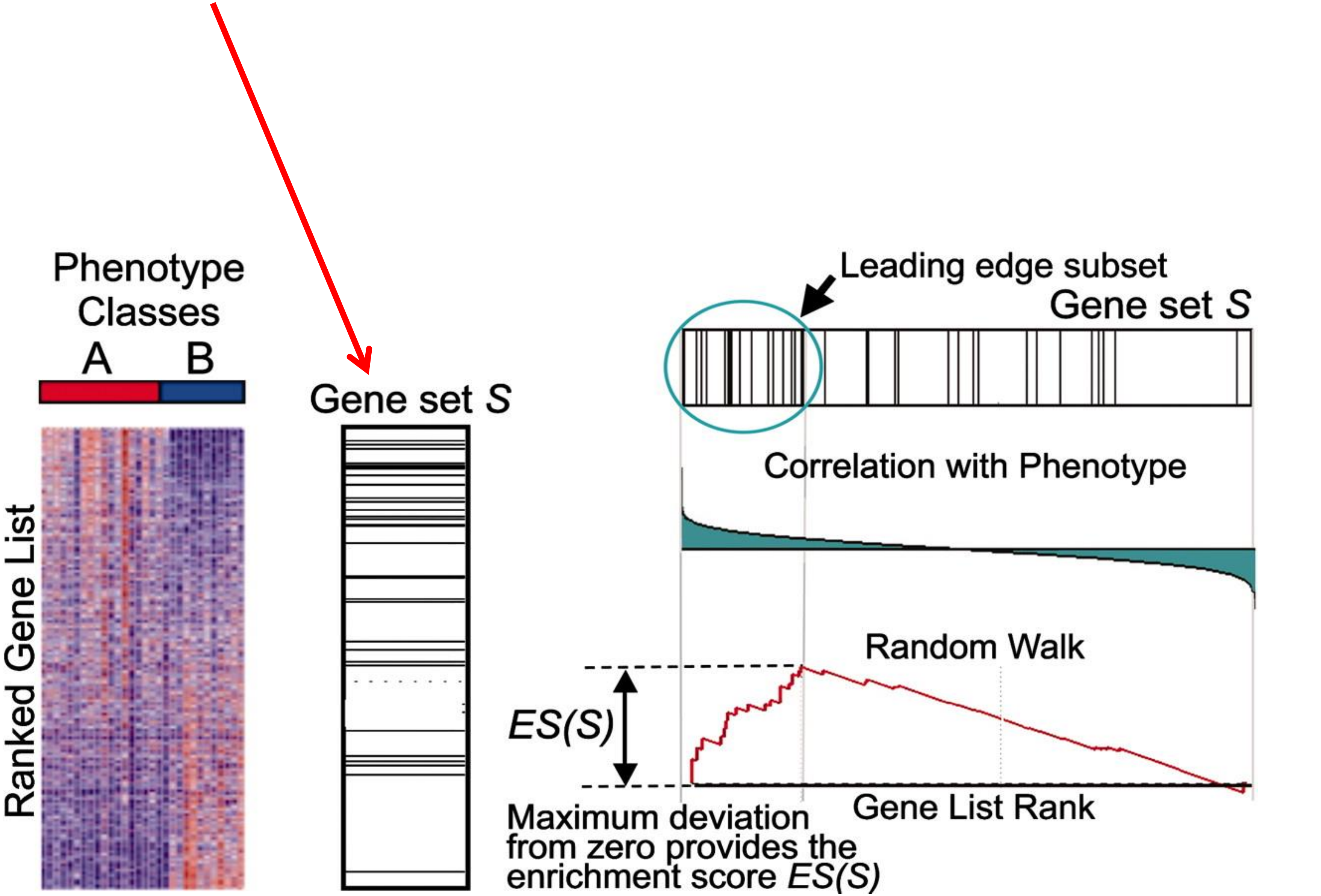


http://www.broadinstitute.org/gsea/

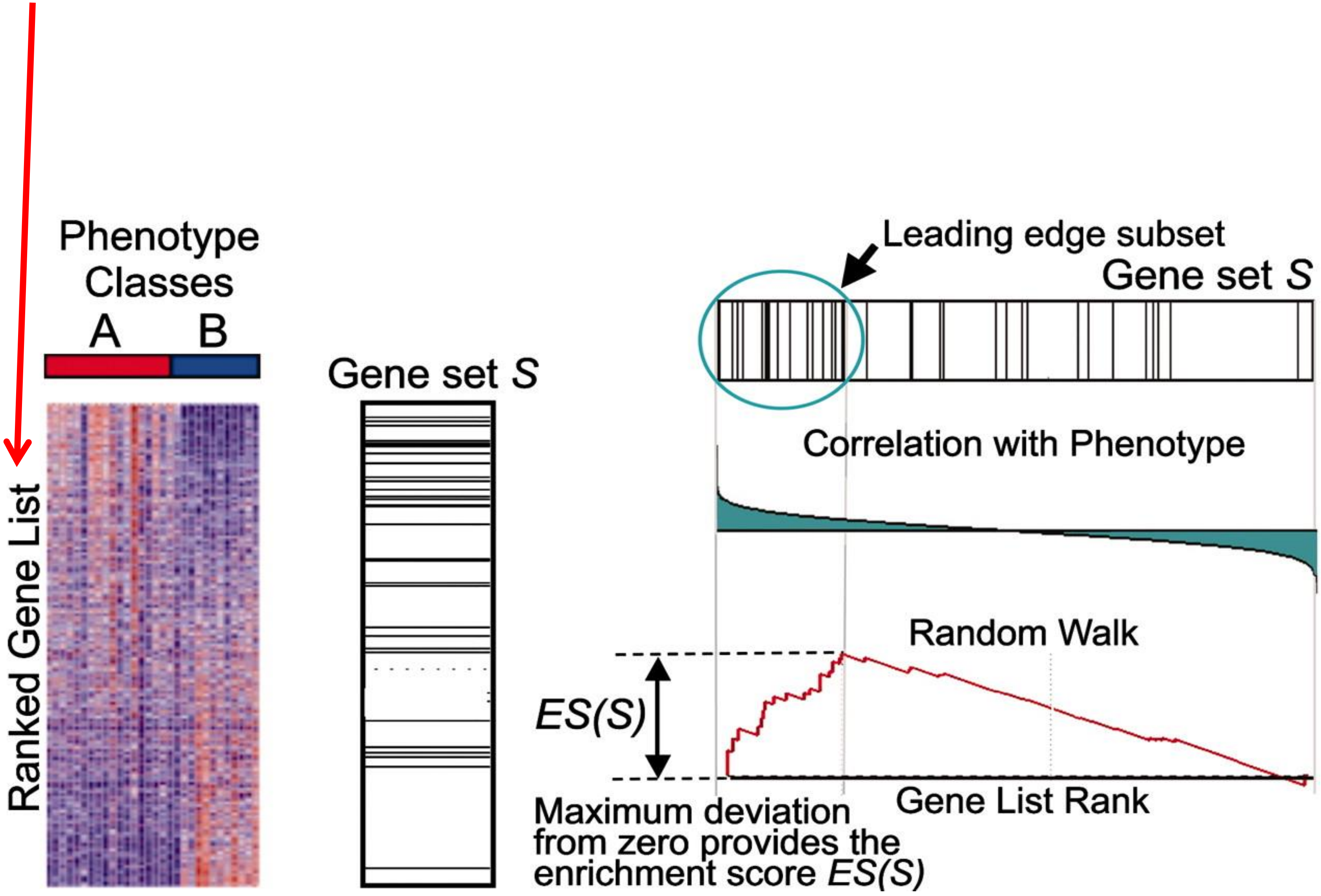
1) Define phenotype classes (e.g. Cells in Hypoxia or Normoxia)



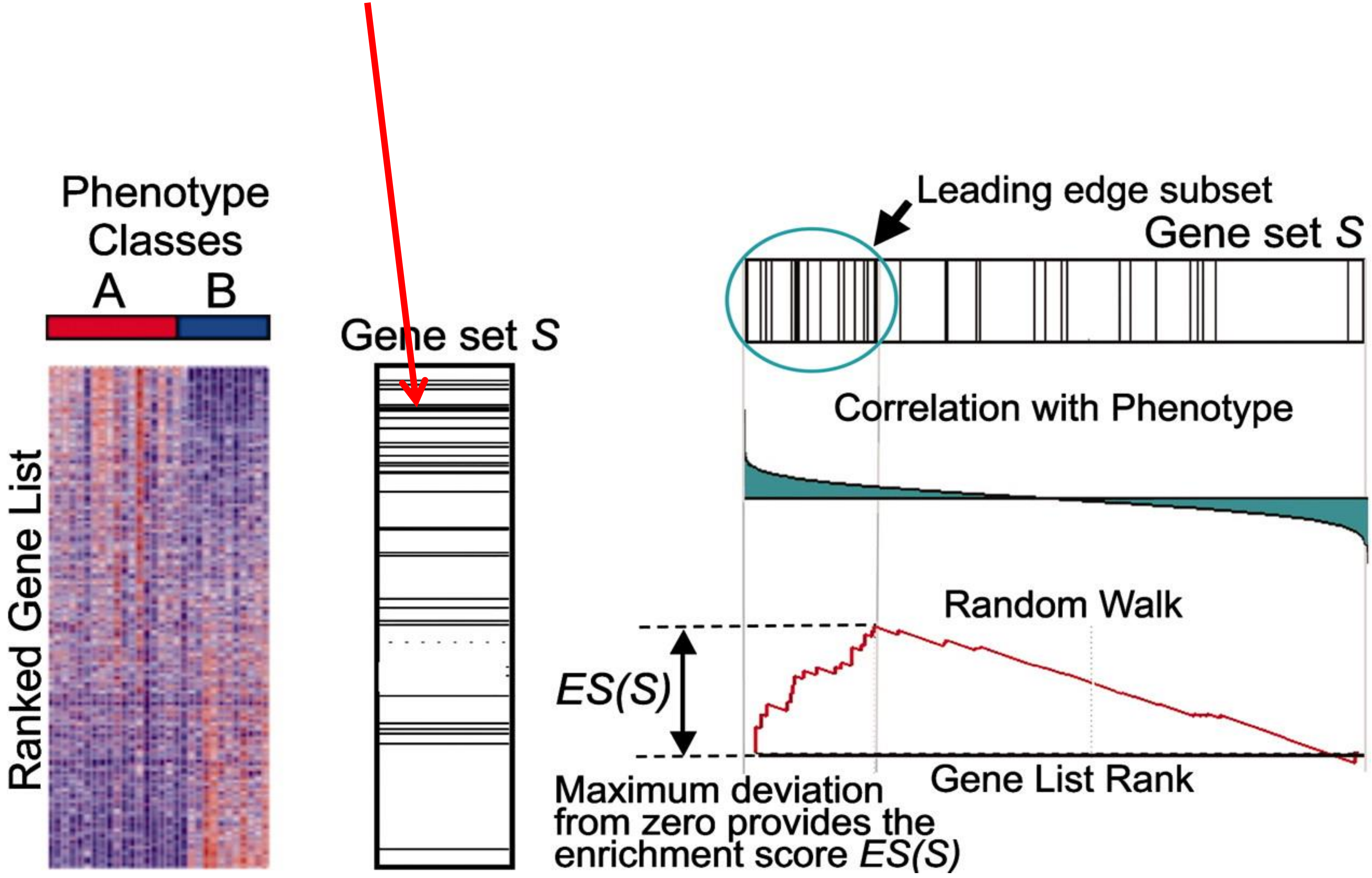
1) Define phenotype classes (e.g. Cells in Hypoxia or Normoxia)
2) Define a gene set of interest (e.g. Genes in the HIF1a pathway)



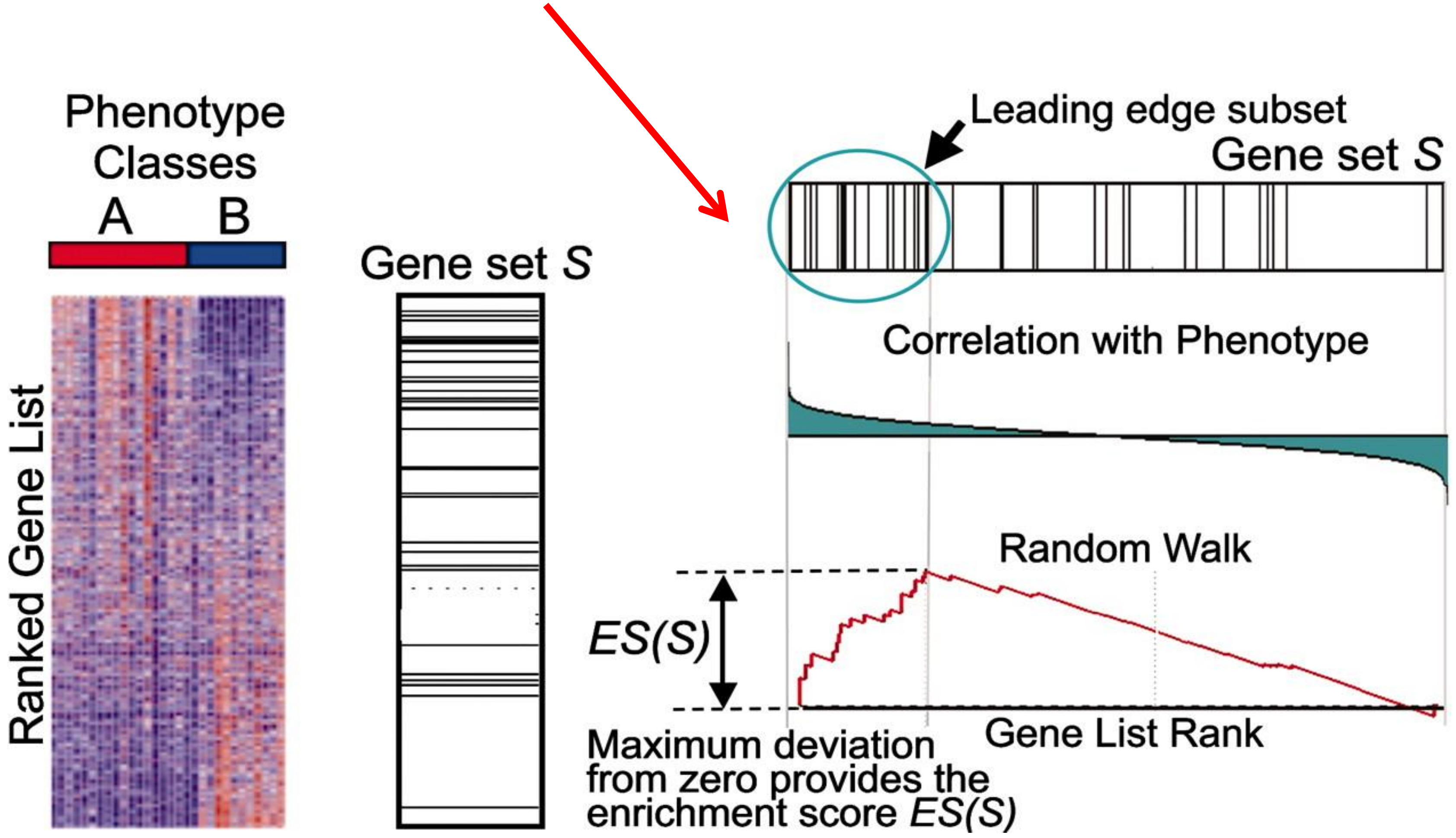
1) Define phenotype classes (e.g. Cells in Hypoxia or Normoxia)
2) Define a gene set of interest (e.g. Genes in the HIF1a pathway)
3) Sort genes based on their differential expression between classes



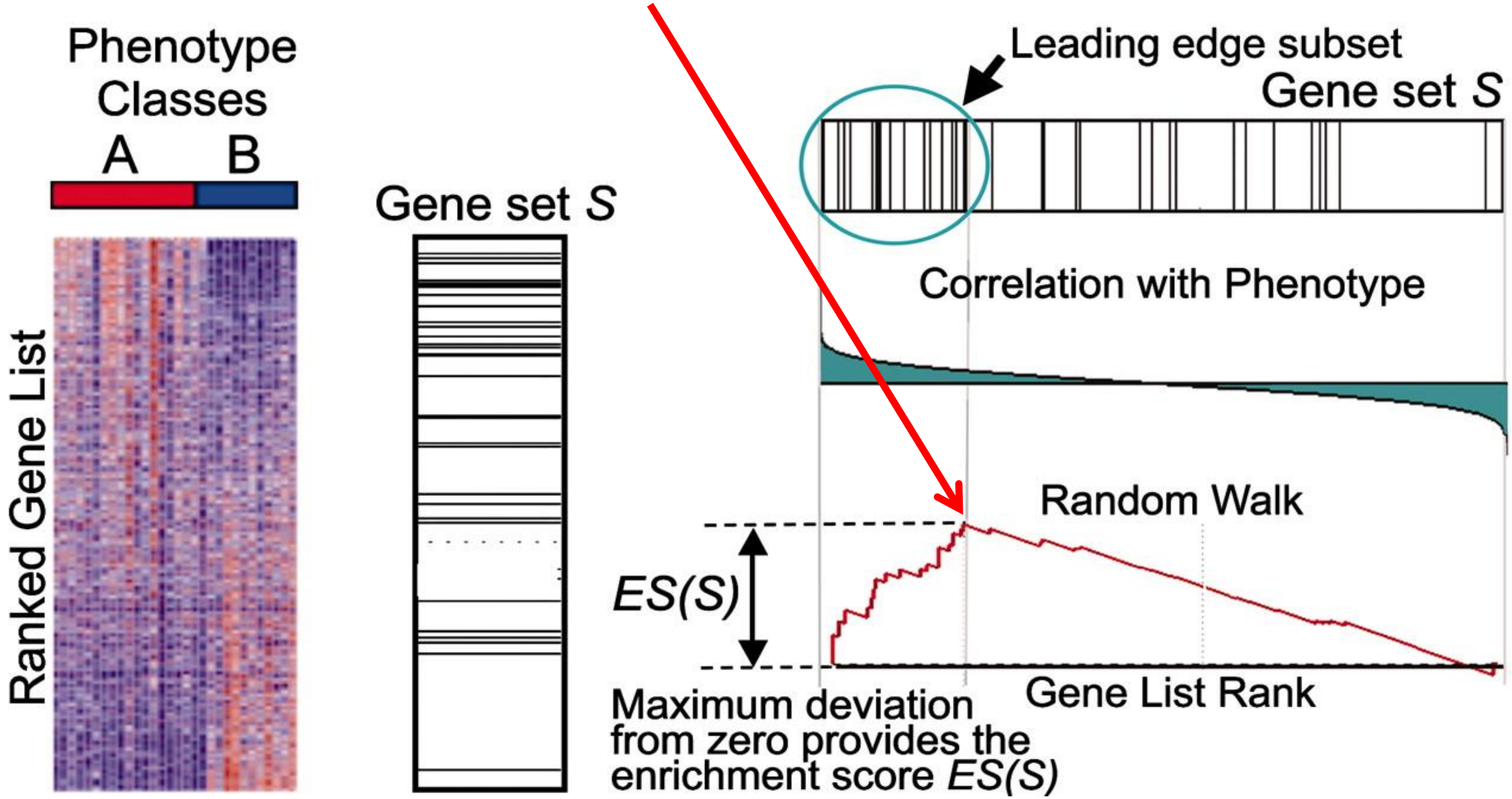
1) Define phenotype classes (e.g. Cells in Hypoxia or Normoxia)
2) Define a gene set of interest (e.g. Genes in the HIF1a pathway)
3) Sort genes based on their differential expression between classes
4) Tag genes from the set S within the sorted list



1) Define phenotype classes (e.g. Cells in Hypoxia or Normoxia)
2) Define a gene set of interest (e.g. Genes in the HIF1a pathway)
3) Sort genes based on their differential expression between classes
4) Tag genes from the set S within the sorted list
5) Walk down the list, for each gene: if gene is in S running-sum statistic up, if not down. (The magnitude of the increment depends on FC)



1) Define phenotype classes (e.g. Cells in Hypoxia or Normoxia)
2) Define a gene set of interest (e.g. Genes in the HIF1a pathway)
3) Sort genes based on their differential expression between classes
4) Tag genes from the set S within the sorted list
5) Walk down the list, for each gene: if gene is in S running-sum statistic up, if not down. (The magnitude of the increment depends on FC)
6) ES is the maximum deviation from zero of this random walk



# Improvement on over-representation

- No need to define an arbitrary threshold for selecting significant genes

- The molecular measurements of the actual changes are not ignored but used in order to detect coordinated changes in the expression of genes in the same pathway.

- Coordinate changes are considered: the dependence between genes in a pathway is accounted for

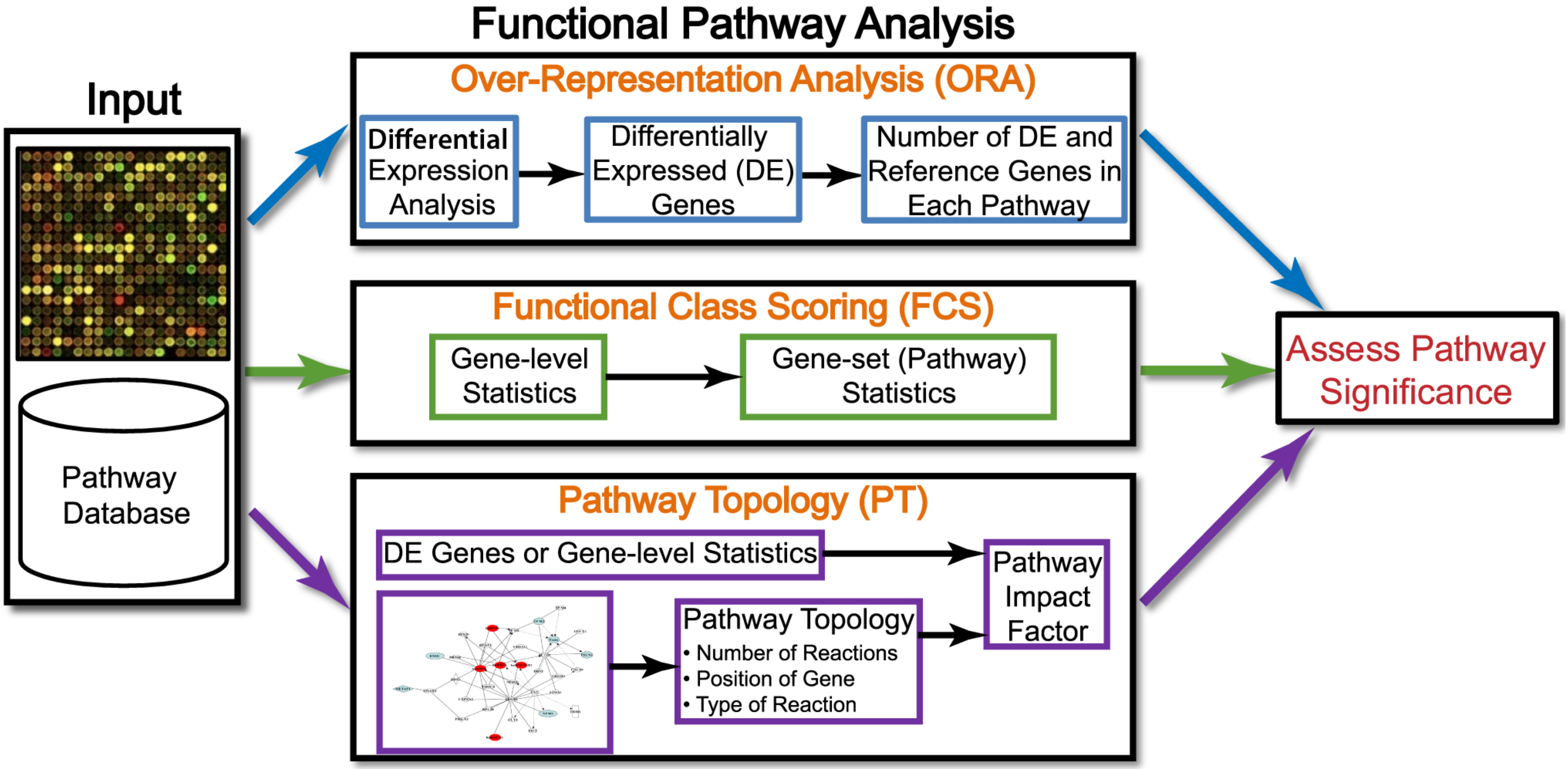
# Limitations

- Pathway are considered independent. However a gene can function in more than one pathway, meaning that pathways can cross and overlap.

- Most methods use ranks instead of the actual changes (exception exist: gene set analysis http://statweb.stanford.edu/~tibs/GSA/ but only available as R function at the moment)

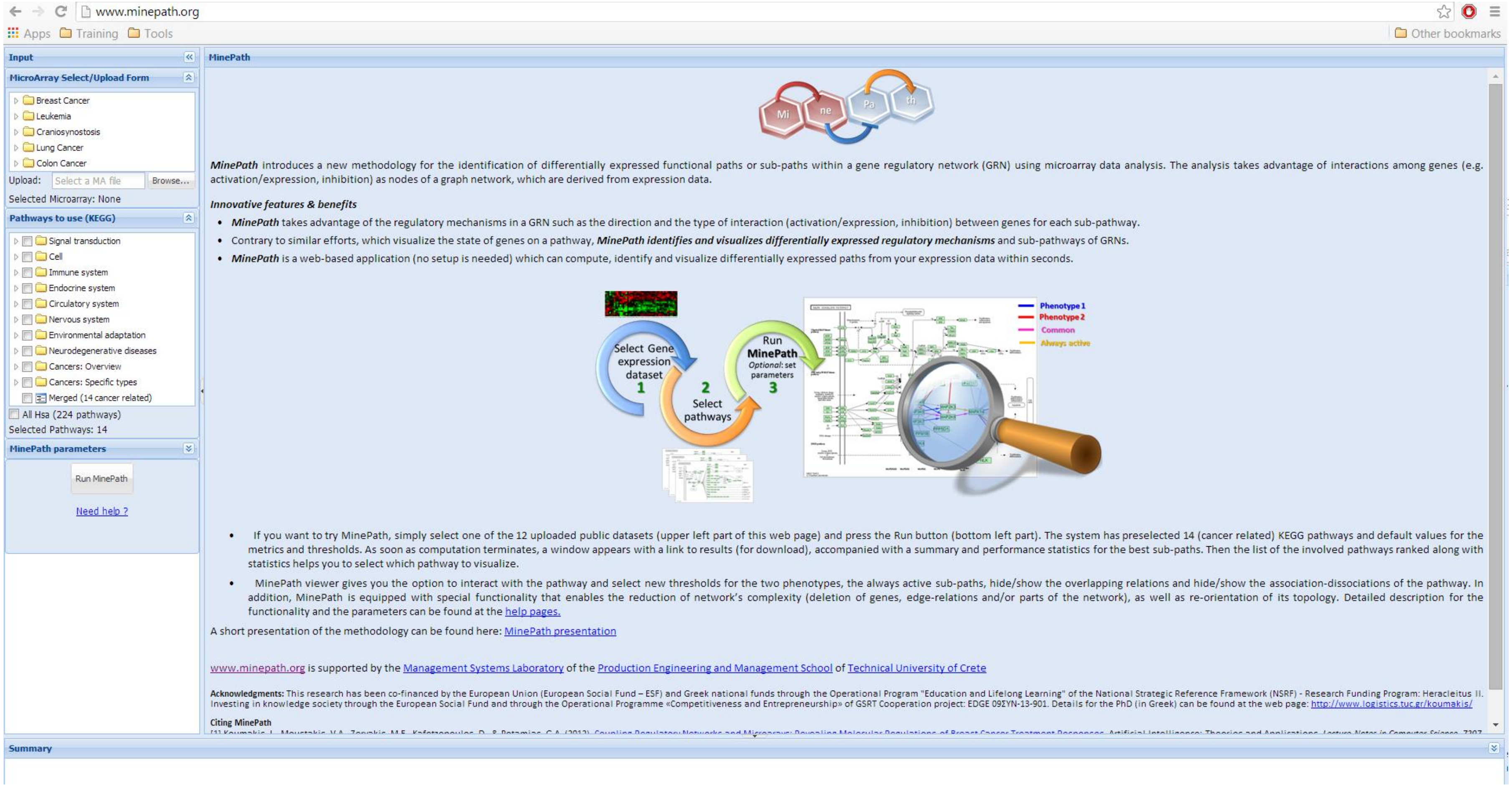
- The nature of the functional link between genes , the strength of the evidence for this link, the role of the genes in the pathway are not considered, only the list of genes in a pathway is used

# Many ways to approach pathway analysis


Functional Pathway Analysis
Input
Pathway Database
Over-Representation Analysis (ORA)
Differential Expression Analysis
Differentially Expressed (DE) Genes
Number of DE and Reference Genes in Each Pathway
Functional Class Scoring (FCS)
Gene-level Statistics
Gene-set (Pathway) Statistics
Pathway Topology (PT)
DE Genes or Gene-level Statistics
Pathway Topology
• Number of Reactions
• Position of Gene
• Type of Reaction
Pathway Impact Factor
Assess Pathway Significance

Khatri P, Sirota M, Butte AJ (2012) Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Comput Biol 8(2): e1002375. doi:10.1371/journal.pcbi.1002375
http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1002375

# MinePATH



**MicroArray Select/Upload Form**

- Breast Cancer
- Leukemia
- Craniosynostosis
- Lung Cancer
- Colon Cancer

Upload: Select a MA file Browse...

Selected Microarray: None

**Pathways to use (KEGG)**

- □ Signal transduction
- □ Cell
- □ Immune system
- □ Endocrine system
- □ Circulatory system
- □ Nervous system
- □ Environmental adaptation
- □ Neurodegenerative diseases
- □ Cancers: Overview
- □ Cancers: Specific types
- □ Merged (14 cancer related)

□ All Hsa (224 pathways)

Selected Pathways: 14

**MinePath parameters**

Run MinePath

Need help ?

*MinePath* introduces a new methodology for the identification of differentially expressed functional paths or sub-paths within a gene regulatory network (GRN) using microarray data analysis. The analysis takes advantage of interactions among genes (e.g. activation/expression, inhibition) as nodes of a graph network, which are derived from expression data.

***Innovative features & benefits***

- *MinePath* takes advantage of the regulatory mechanisms in a GRN such as the direction and the type of interaction (activation/expression, inhibition) between genes for each sub-pathway.
- Contrary to similar efforts, which visualize the state of genes on a pathway, ***MinePath identifies and visualizes differentially expressed regulatory mechanisms*** and sub-pathways of GRNs.
- *MinePath* is a web-based application (no setup is needed) which can compute, identify and visualize differentially expressed paths from your expression data within seconds.

- If you want to try MinePath, simply select one of the 12 uploaded public datasets (upper left part of this web page) and press the Run button (bottom left part). The system has preselected 14 (cancer related) KEGG pathways and default values for the metrics and thresholds. As soon as computation terminates, a window appears with a link to results (for download), accompanied with a summary and performance statistics for the best sub-paths. Then the list of the involved pathways ranked along with statistics helps you to select which pathway to visualize.
- MinePath viewer gives you the option to interact with the pathway and select new thresholds for the two phenotypes, the always active sub-paths, hide/show the overlapping relations and hide/show the association-dissociations of the pathway. In addition, MinePath is equipped with special functionality that enables the reduction of network's complexity (deletion of genes, edge-relations and/or parts of the network), as well as re-orientation of its topology. Detailed description for the functionality and the parameters can be found at the help pages.

A short presentation of the methodology can be found here: MinePath presentation

www.minepath.org is supported by the Management Systems Laboratory of the Production Engineering and Management School of Technical University of Crete

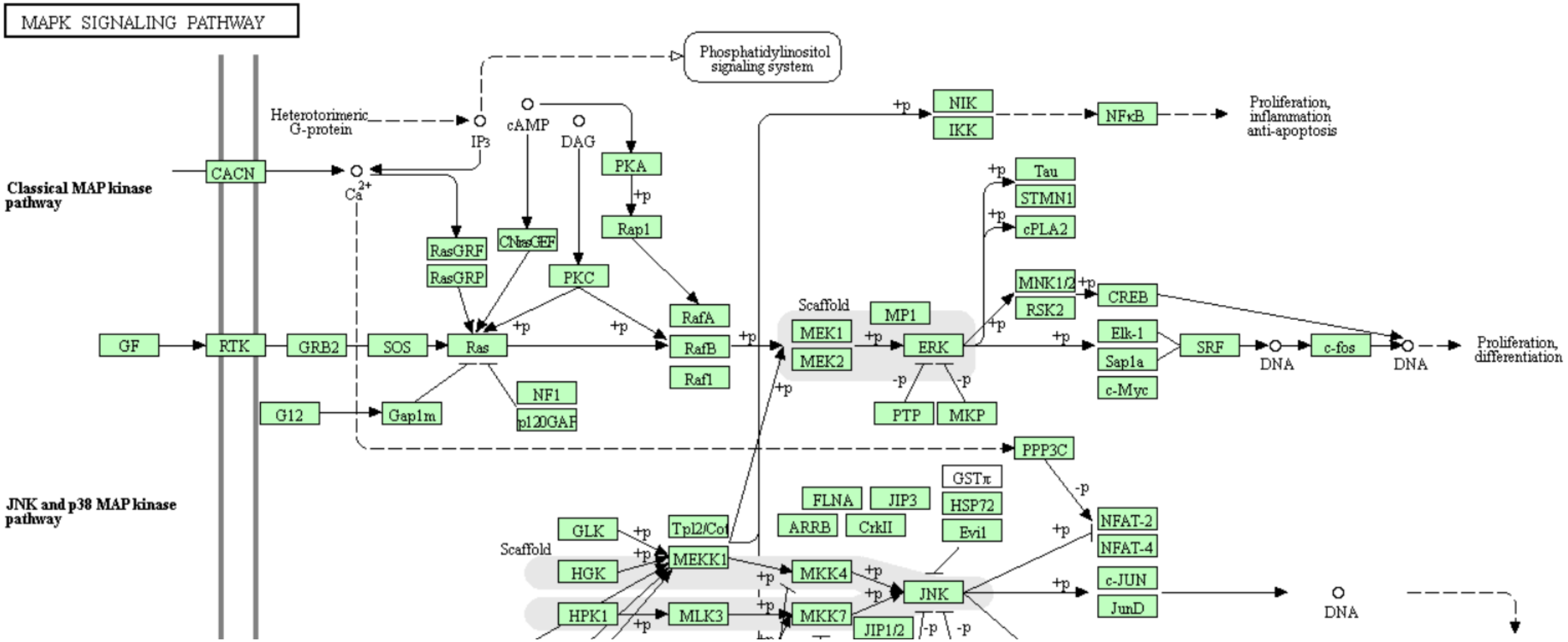
**Acknowledgments:** This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund and through the Operational Programme «Competitiveness and Entrepreneurship» of GSRT Cooperation project: EDGE 09ΣΥΝ-13-901. Details for the PhD (in Greek) can be found at the web page: http://www.logistics.tuc.gr/koumakis/

**Citing MinePath**

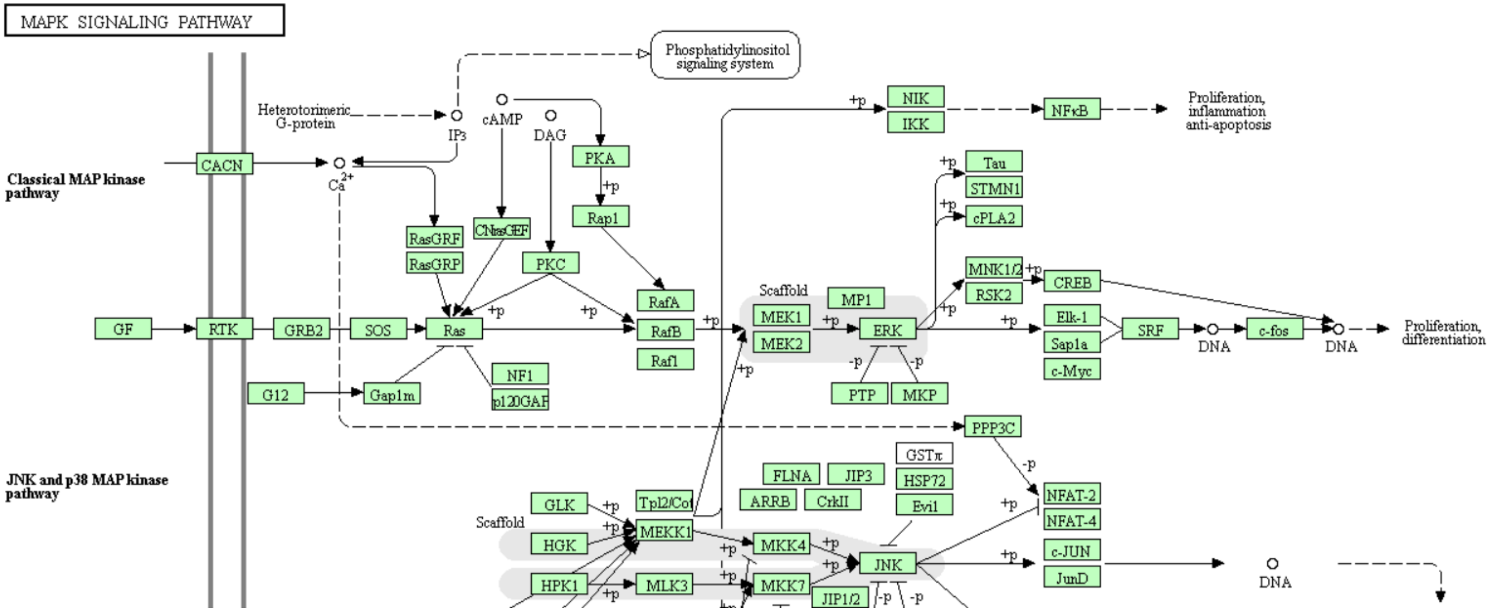
[1] Koumakis, L., Moustakis, V.A., Zervakis, M.E., Kafetzopoulos, D., & Potamias, G.A. (2012). Coupling Regulatory Networks and Microarrays: Revealing Molecular Regulations of Breast Cancer Treatment Responses. Artificial Intelligence: Theories and Applications. *Lecture Notes in Computer Science*, 7297

Available at minepath.org (Koumakis et al., 2012)

# Kegg pathways



# Kegg pathways



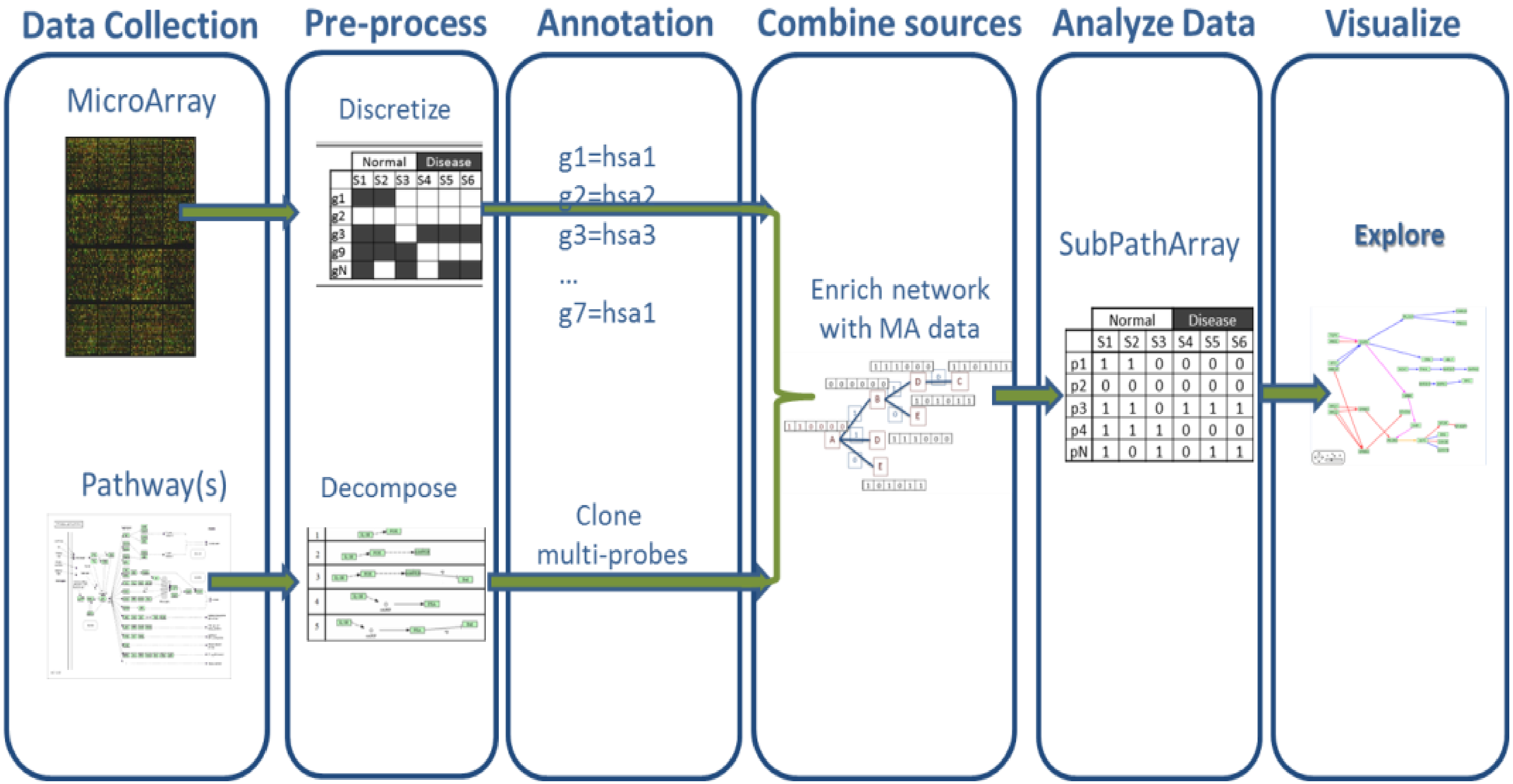
**Nodes**

- Genes
- Group of genes
- Compounds
- Other networks

**Edges**

Activation/Inhibition
Expression
Indirect
Phosphorylation/Diphosphorilation
Ubiquination
Association/Dissocation

# System overview



**Data Collection**

MicroArray

Pathway(s)

**Pre-process**

Discretize

|  | Normal |  |  | Disease |  |  |
|---|---|---|---|---|---|---|
|  | S1 | S2 | S3 | S4 | S5 | S6 |
| g1 | ■ | ■ | □ | □ | □ | □ |
| g2 | □ | □ | □ | □ | □ | □ |
| g3 | ■ | ■ | □ | ■ | ■ | ■ |
| g9 | ■ | ■ | ■ | □ | □ | □ |
| gN | ■ | □ | ■ | □ | ■ | ■ |

Decompose

**Annotation**

g1=hsa1
g2=hsa2
g3=hsa3
...
g7=hsa1

Clone
multi-probes

**Combine sources**

Enrich network
with MA data

**Analyze Data**

SubPathArray

|  | Normal |  |  | Disease |  |  |
|---|---|---|---|---|---|---|
|  | S1 | S2 | S3 | S4 | S5 | S6 |
| p1 | 1 | 1 | 0 | 0 | 0 | 0 |
| p2 | 0 | 0 | 0 | 0 | 0 | 0 |
| p3 | 1 | 1 | 0 | 1 | 1 | 1 |
| p4 | 1 | 1 | 1 | 0 | 0 | 0 |
| pN | 1 | 0 | 1 | 0 | 1 | 1 |

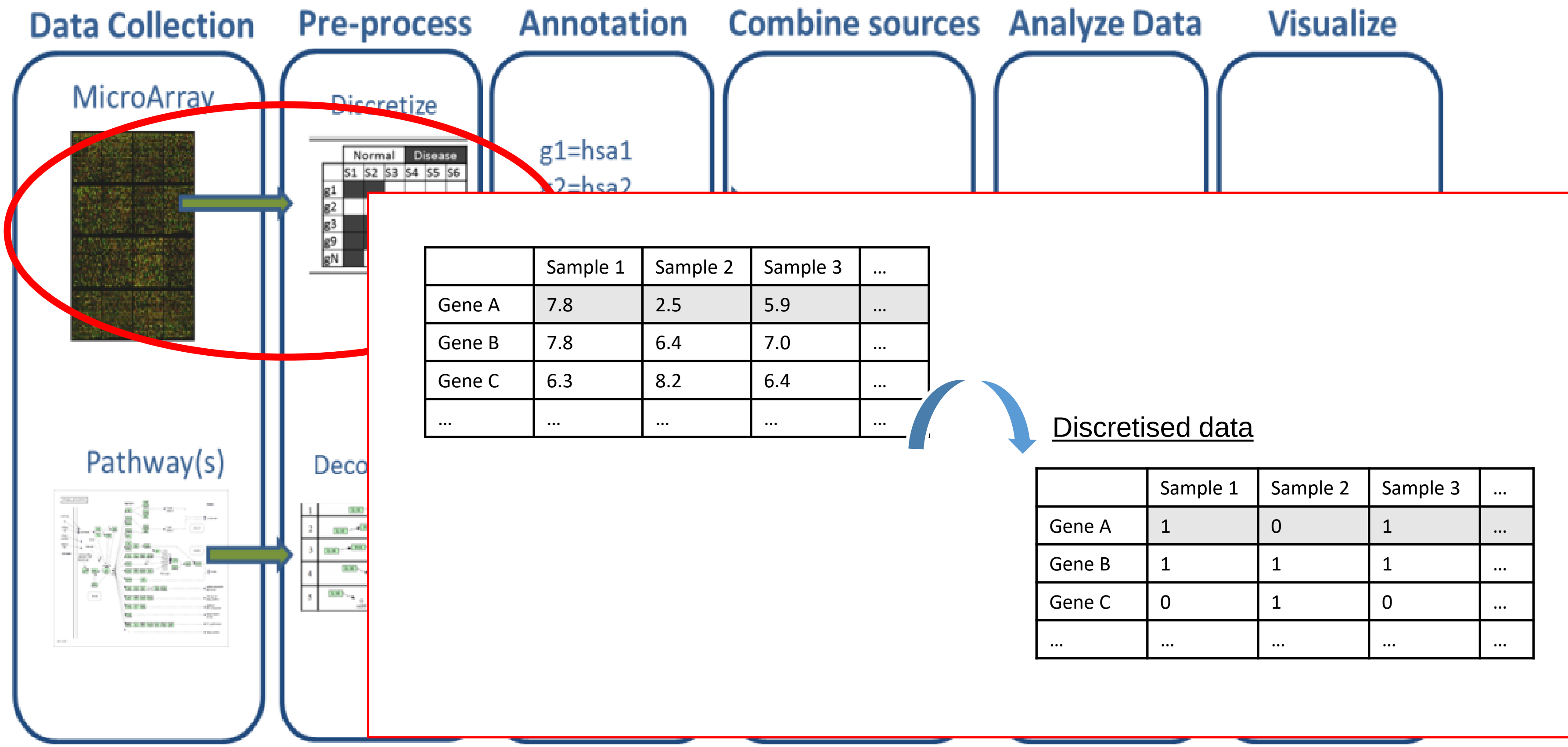
**Visualize**

Explore

# System overview

Discretise the gene expression data

**0: down-regulated genes**
**1: up-regulated genes**



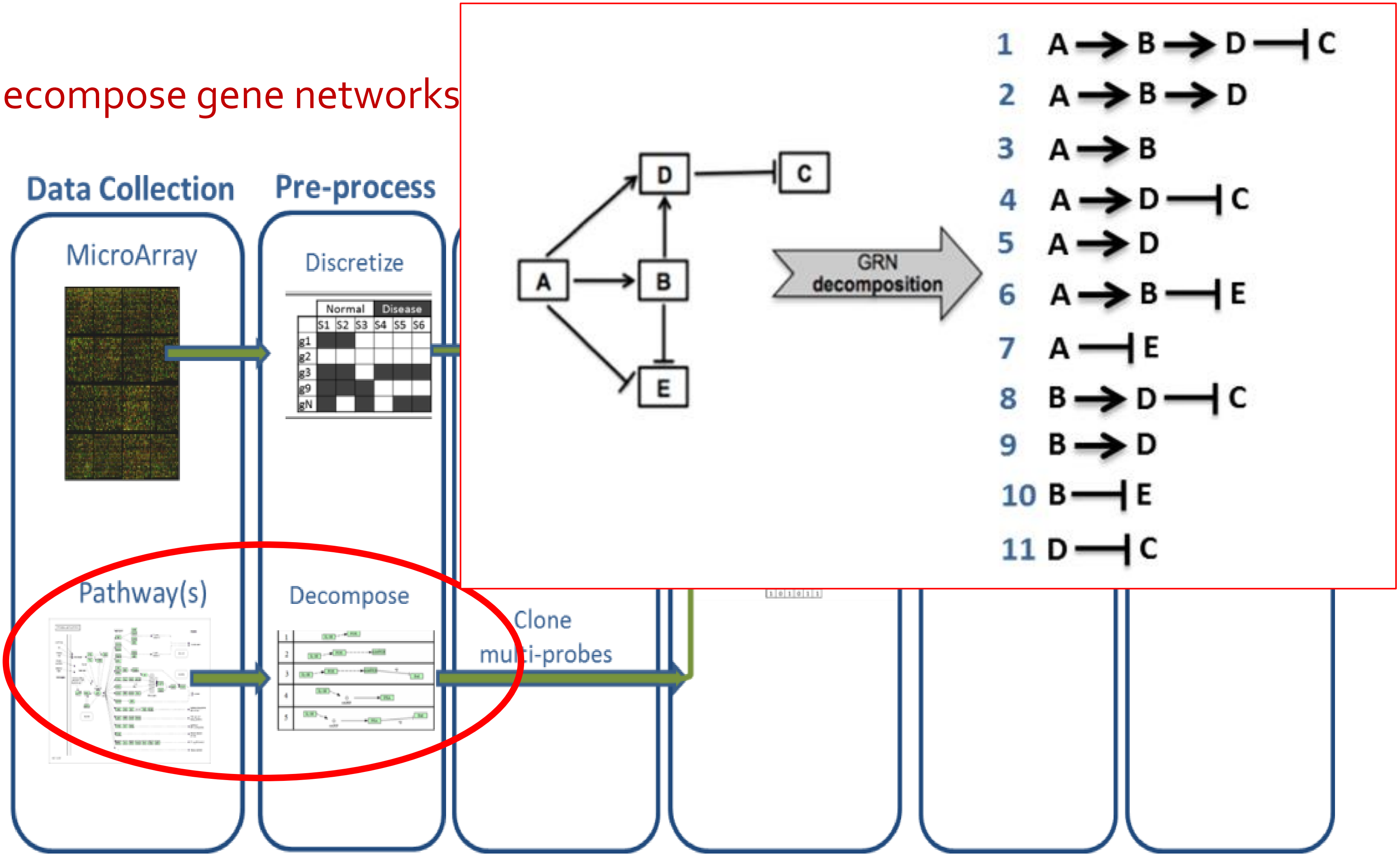
| | Sample 1 | Sample 2 | Sample 3 | … |
|---|---|---|---|---|
| Gene A | 7.8 | 2.5 | 5.9 | … |
| Gene B | 7.8 | 6.4 | 7.0 | … |
| Gene C | 6.3 | 8.2 | 6.4 | … |
| … | … | … | … | … |

Discretised data

| | Sample 1 | Sample 2 | Sample 3 | … |
|---|---|---|---|---|
| Gene A | 1 | 0 | 1 | … |
| Gene B | 1 | 1 | 1 | … |
| Gene C | 0 | 1 | 0 | … |
| … | … | … | … | … |

# System overview

Decompose gene networks



1 A → B → D ⊣ C

2 A → B → D

3 A → B

4 A → D ⊣ C

5 A → D

6 A → B ⊣ E

7 A ⊣ E

8 B → D ⊣ C

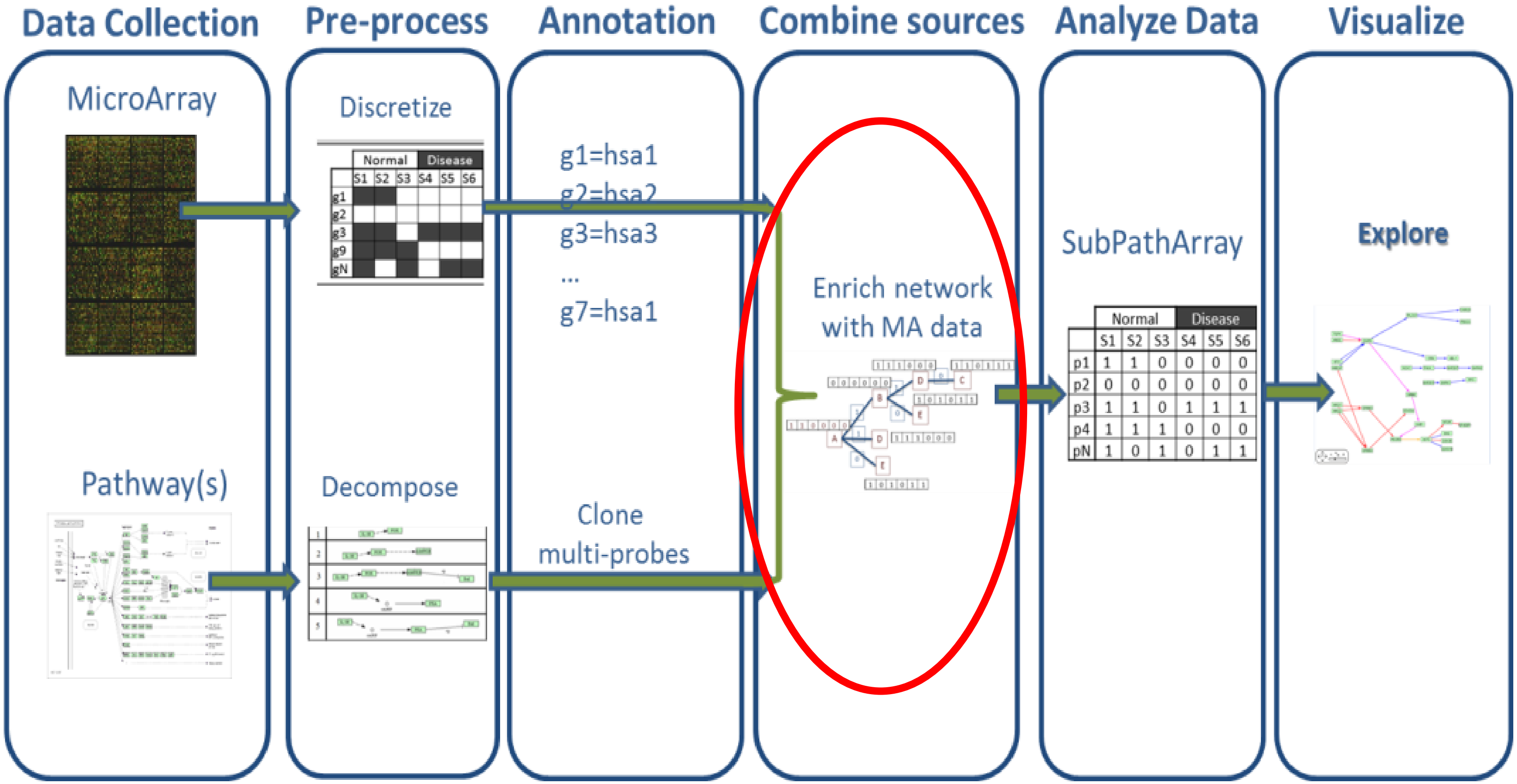
9 B → D

10 B ⊣ E

11 D ⊣ C

# System overview

Combine discretized gene expression data and decomposed sub-paths

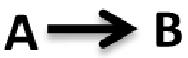


|  | Normal |  |  | Disease |  |  |
|---|---|---|---|---|---|---|
|  | S1 | S2 | S3 | S4 | S5 | S6 |
| g1 | ■ | ■ | □ | □ | □ | □ |
| g2 | □ | □ | □ | □ | □ | □ |
| g3 | ■ | ■ | □ | ■ | ■ | ■ |
| g9 | ■ | ■ | ■ | □ | □ | □ |
| gN | ■ | □ | ■ | □ | ■ | ■ |

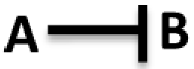
|  | Normal |  |  | Disease |  |  |
|---|---|---|---|---|---|---|
|  | S1 | S2 | S3 | S4 | S5 | S6 |
| p1 | 1 | 1 | 0 | 0 | 0 | 0 |
| p2 | 0 | 0 | 0 | 0 | 0 | 0 |
| p3 | 1 | 1 | 0 | 1 | 1 | 1 |
| p4 | 1 | 1 | 1 | 0 | 0 | 0 |
| pN | 1 | 0 | 1 | 0 | 1 | 1 |

# Combine discretized gene expression data and decomposed sub-paths

**Activation**



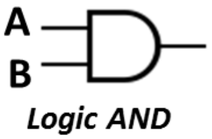
**Inhibition**

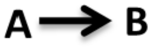


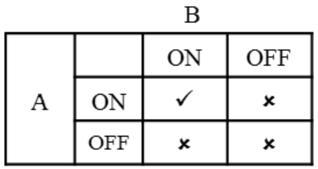
| | Sample 1 | Sample 2 | Sample 3 | … |
|---|---|---|---|---|
| Gene A | 1 | 0 | 1 | … |
| Gene B | 1 | 1 | 1 | … |
| Gene C | 0 | 1 | 0 | … |
| … | … | … | … | … |

# Combine discretized gene expression data and decomposed sub-paths

**Activation**

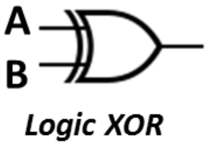


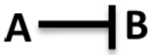


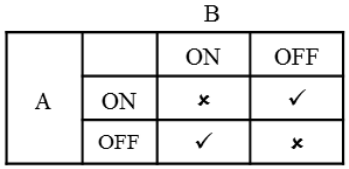


| | | B | |
|---|---|---|---|
| | | ON | OFF |
| A | ON | ✓ | × |
| | OFF | × | × |

**Inhibition**





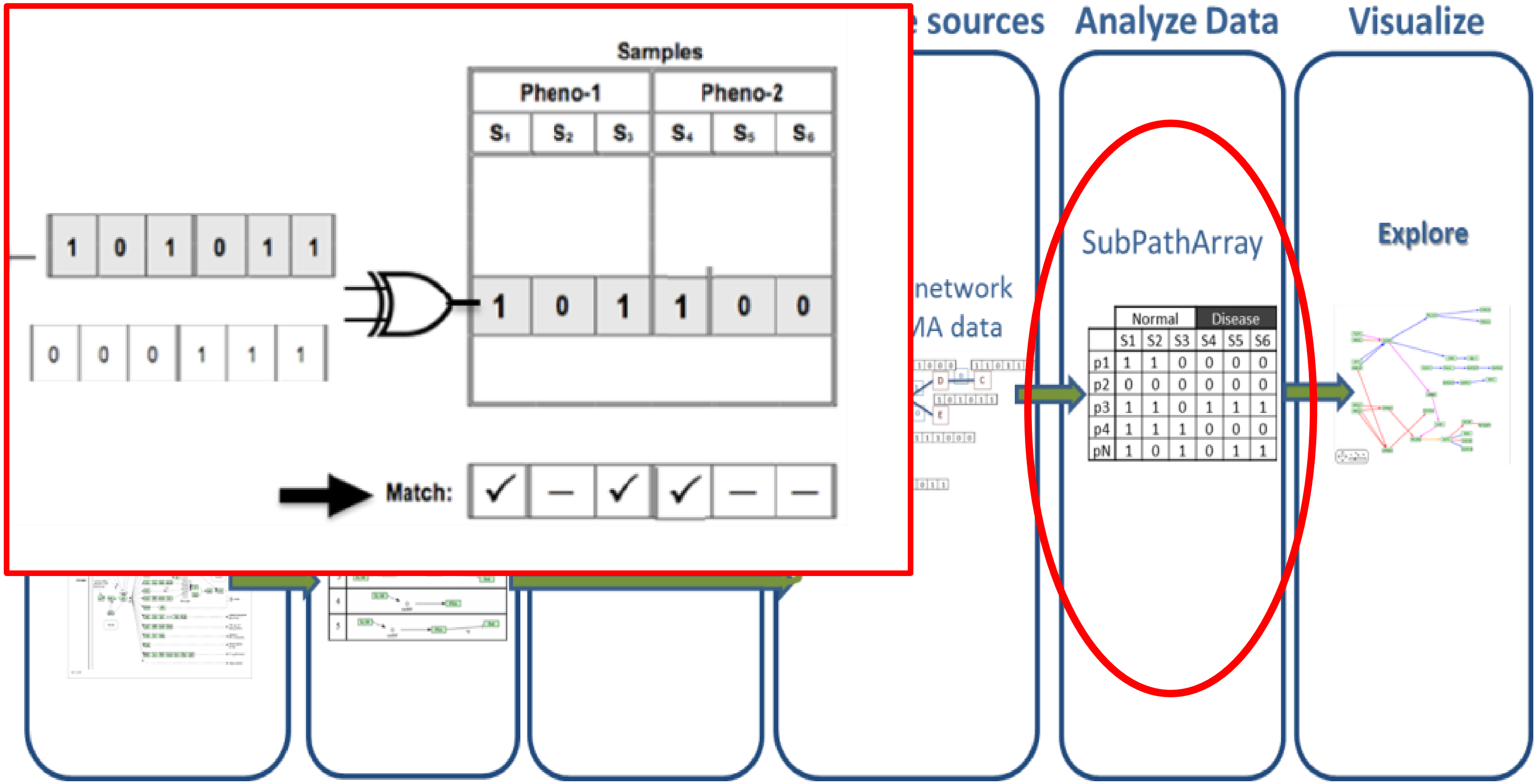


| | | B | |
|---|---|---|---|
| | | ON | OFF |
| A | ON | × | ✓ |
| | OFF | ✓ | × |

| | Sample 1 | Sample 2 | Sample 3 | … |
|---|---|---|---|---|
| Gene A | 1 | 0 | 1 | … |
| Gene B | 1 | 1 | 1 | … |
| Gene C | 0 | 1 | 0 | … |
| … | … | … | … | … |

Interactions as logical operators

# System overview



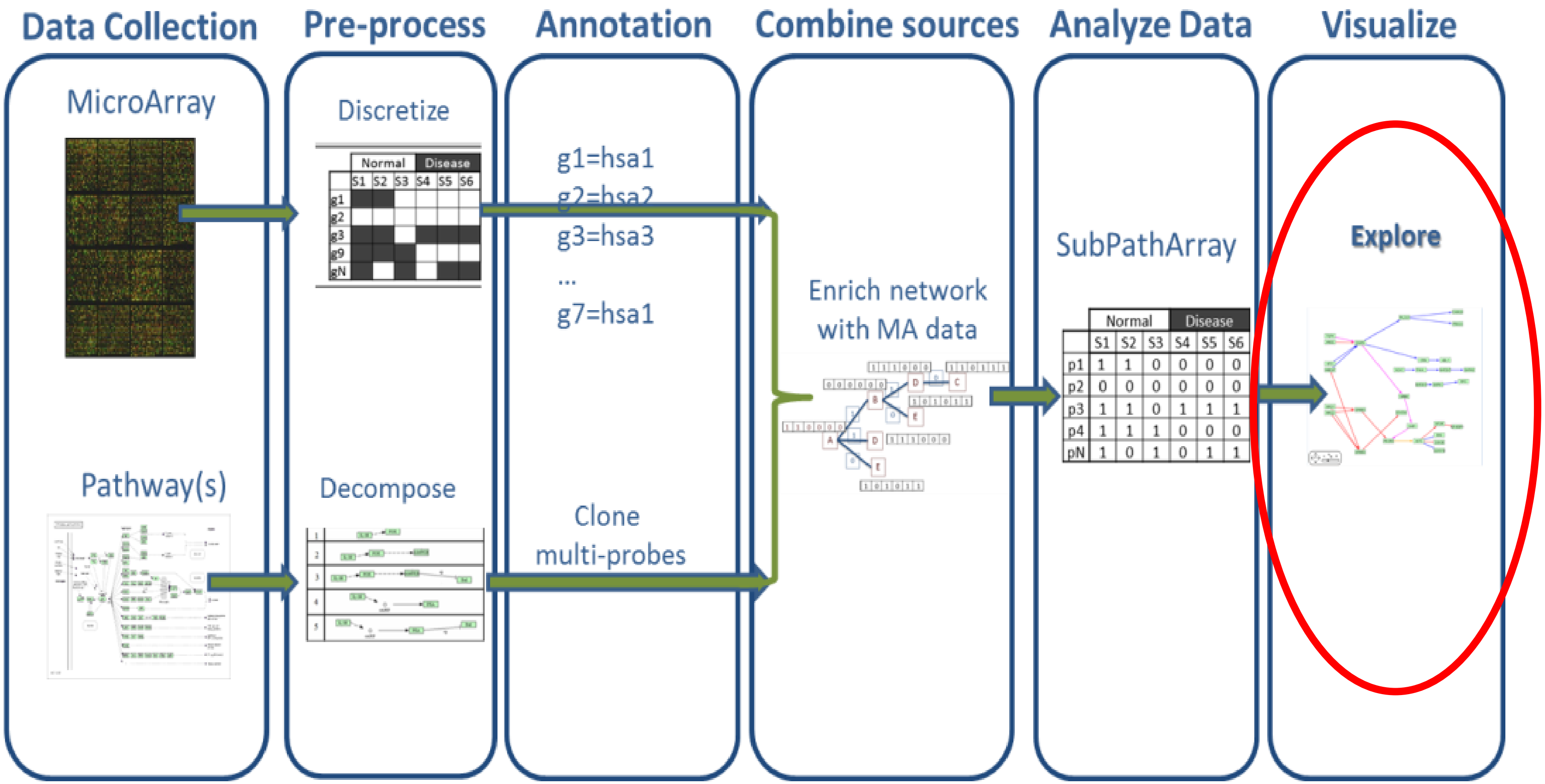
Evaluate sub-paths

|  | Samples |  |  |  |  |  |
|---|---|---|---|---|---|---|
|  | Pheno-1 |  |  | Pheno-2 |  |  |
|  | $S_1$ | $S_2$ | $S_3$ | $S_4$ | $S_5$ | $S_6$ |
|  | 1 | 0 | 1 | 1 | 0 | 0 |
| Match: | ✓ | – | ✓ | ✓ | – | – |

|  | Normal |  |  | Disease |  |  |
|---|---|---|---|---|---|---|
|  | S1 | S2 | S3 | S4 | S5 | S6 |
| p1 | 1 | 1 | 0 | 0 | 0 | 0 |
| p2 | 0 | 0 | 0 | 0 | 0 | 0 |
| p3 | 1 | 1 | 0 | 1 | 1 | 1 |
| p4 | 1 | 1 | 1 | 0 | 0 | 0 |
| pN | 1 | 0 | 1 | 0 | 1 | 1 |

# System overview

Visualize the results



Data Collection

MicroArray

Pathway(s)

Pre-process

Discretize

|  | Normal |  |  | Disease |  |  |
|---|---|---|---|---|---|---|
|  | S1 | S2 | S3 | S4 | S5 | S6 |
| g1 | ■ | ■ | □ | □ | □ | □ |
| g2 | □ | □ | □ | □ | □ | □ |
| g3 | ■ | ■ | □ | ■ | ■ | ■ |
| g9 | ■ | ■ | ■ | □ | □ | □ |
| gN | ■ | □ | ■ | □ | ■ | ■ |

Decompose

Annotation

g1=hsa1

g2=hsa2

g3=hsa3

...

g7=hsa1

Clone multi-probes

Combine sources

Enrich network with MA data

Analyze Data

SubPathArray

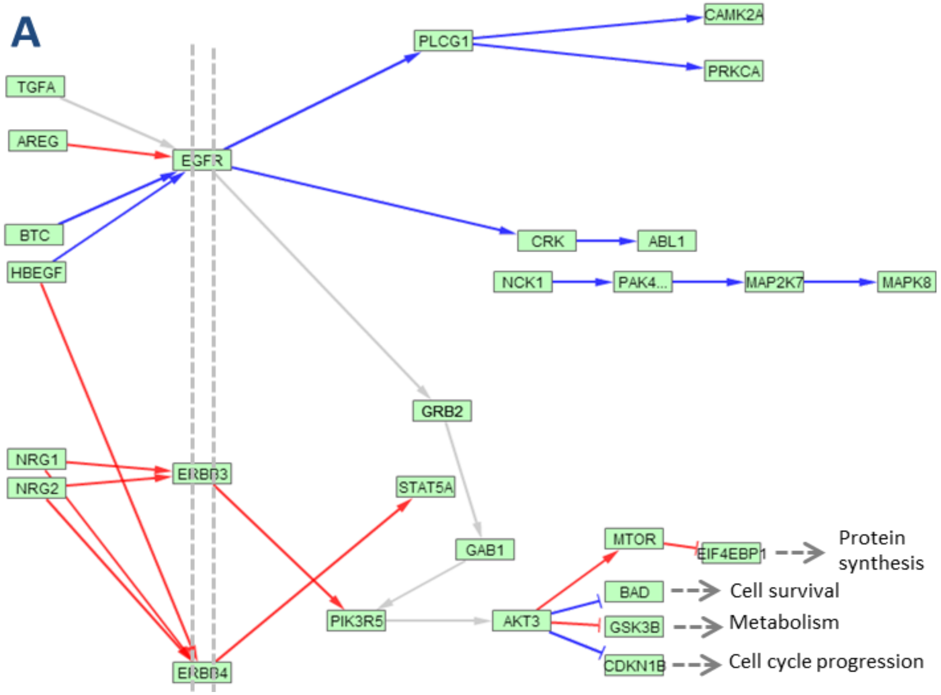
|  | Normal |  |  | Disease |  |  |
|---|---|---|---|---|---|---|
|  | S1 | S2 | S3 | S4 | S5 | S6 |
| p1 | 1 | 1 | 0 | 0 | 0 | 0 |
| p2 | 0 | 0 | 0 | 0 | 0 | 0 |
| p3 | 1 | 1 | 0 | 1 | 1 | 1 |
| p4 | 1 | 1 | 1 | 0 | 0 | 0 |
| pN | 1 | 0 | 1 | 0 | 1 | 1 |

Visualize

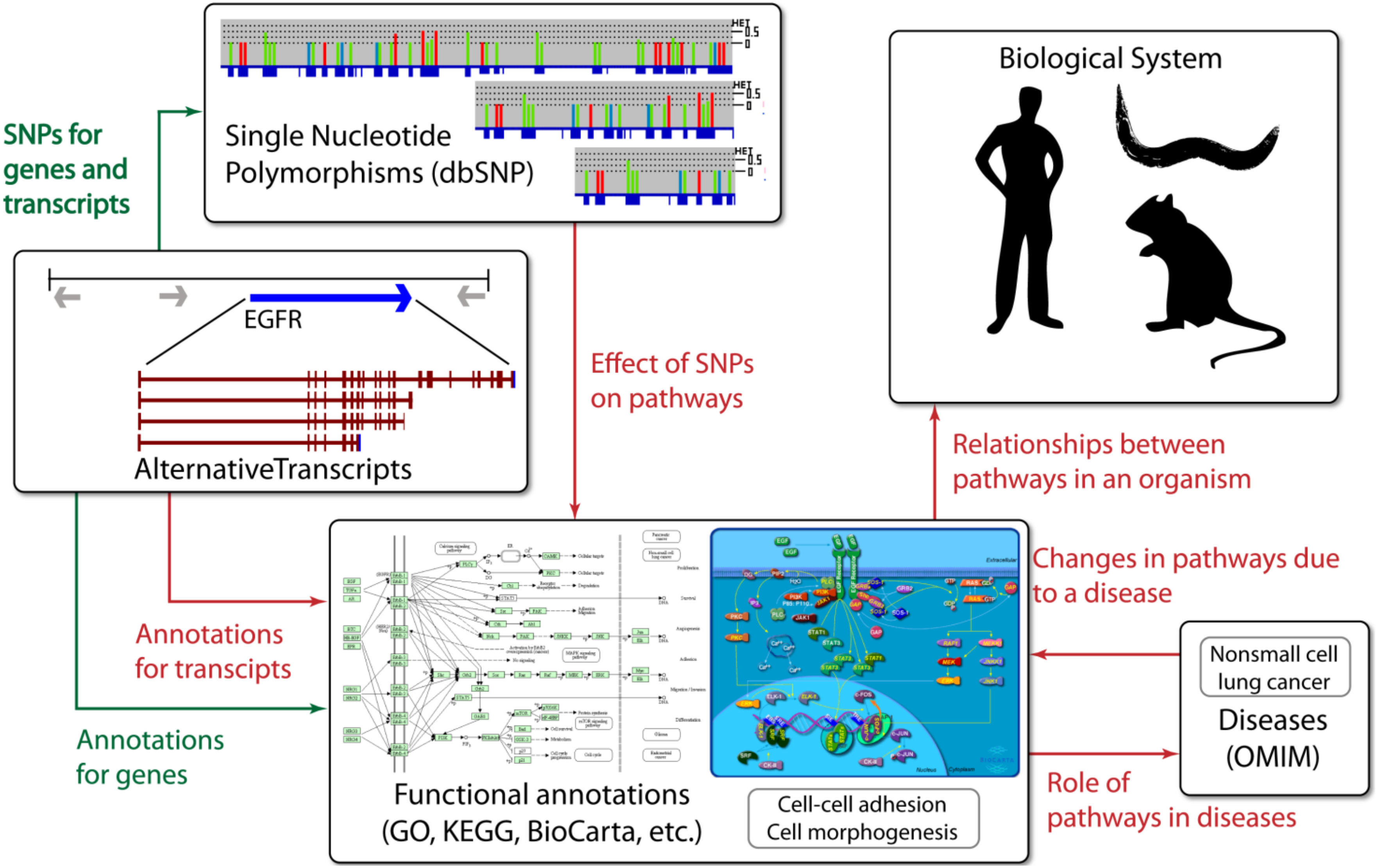
Explore

# Visualizing the best pathways

- '**Red**' is used to encode sub-path relations that are active for phenotype 1 (Class 1)
- '**Blue**' for relations that are active for phenotype 2 (Class 2)
- '**Magenta**' for relations holding for both phenotypes
- '**Orange**' for relations that are "always-active"
- "**Yellow**" for the association/disassociation relations
- '**Grey**' for inactive relations.



# There are still many gaps



SNPs for genes and transcripts

Single Nucleotide Polymorphisms (dbSNP)

AlternativeTranscripts

Effect of SNPs on pathways

Biological System

Relationships between pathways in an organism

Annotations for transcipts

Annotations for genes

Functional annotations (GO, KEGG, BioCarta, etc.)

Cell-cell adhesion
Cell morphogenesis

Changes in pathways due to a disease

Nonsmall cell lung cancer

Diseases (OMIM)

Role of pathways in diseases