



AI LAB

4.1 – LARGE PROJECTS AND DATABASES

FRANCESCA M. BUFFA

LAB STRUCTURE

- INTRODUCTION
- THE DATA
- THE AI-LAB CHALLENGE – PART 1
- PART 1 - SHARING AND DISCUSSION

- UNSUPERVISED LEARNING EXAMPLES
- THE AI-LAB CHALLENGE – PART 2
- PART 2 - SHARING AND DISCUSSION

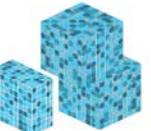
- SUPERVISED LEARNING EXAMPLES
- THE AI-LAB CHALLENGE – PART 3

- LARGE PROJECTS, DATABASES AND DATA INTERPRETATION
- DISCUSS AND PREPARE WORKSHOP PRESENTATIONS

GENOMICS BIG DATASETS – THE CANCER GENOME ATLAS

NATIONAL CANCER INSTITUTE THE CANCER GENOME ATLAS

TCGA BY THE NUMBERS

TCGA produced over
2.5 PETABYTES of data


TCGA data describes
33 DIFFERENT TUMOR TYPES ...including
10 RARE CANCERS
...based on paired tumor and normal tissue sets collected from

To put this into perspective, 1 petabyte of data is equal to
212,000 DVDs


using
11,000 PATIENTS
...using
7 DIFFERENT DATA TYPES


TCGA RESULTS & FINDINGS



MOLECULAR BASIS OF CANCER

Improved our understanding of the genomic underpinnings of cancer

For example, a TCGA study found the basal-like subtype of breast cancer to be similar to the serous subtype of ovarian cancer on a molecular level, suggesting that despite arising from different tissues in the body, these subtypes may share a common path of development and respond to similar therapeutic strategies.



TUMOR SUBTYPES

Revolutionized how cancer is classified

TCGA revolutionized how cancer is classified by identifying tumor subtypes with distinct sets of genomic alterations.*



THERAPEUTIC TARGETS

Identified genomic characteristics of tumors that can be targeted with currently available therapies or used to drug development

TCGA's identification of targetable genomic alterations in lung squamous cell carcinoma led to NCI's Lung-MAP Trial, which will treat patients based on the specific genomic changes in their tumor.

THE TEAM



20

COLLABORATING INSTITUTIONS
across the United States and Canada

WHAT'S NEXT?

The Genomic Data Commons (GDC) houses TCGA and other NCI-generated data sets for scientists to access from anywhere. The GDC also has many expanded capabilities that will allow researchers to answer more clinically relevant questions with increased ease.



www.cancer.gov/ccg

Omics characterizations

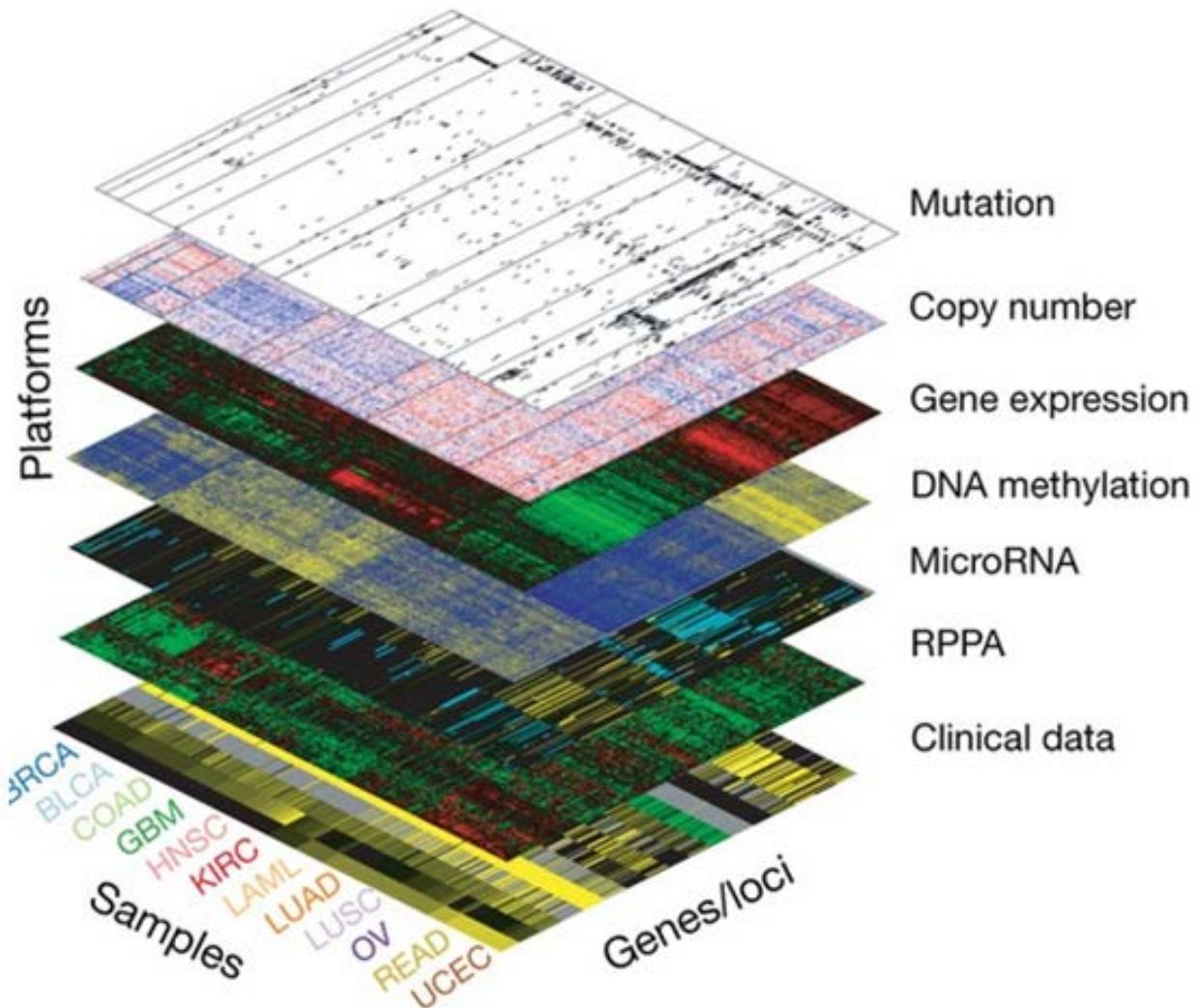


Image source: [TCGA Research Network et al. Nature Genetics 2013](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3704723/)

*TCGA's analysis of stomach cancer revealed that it is not a single disease, but a disease composed of four subtypes, including a new subtype characterized by infection with Epstein-Barr virus.

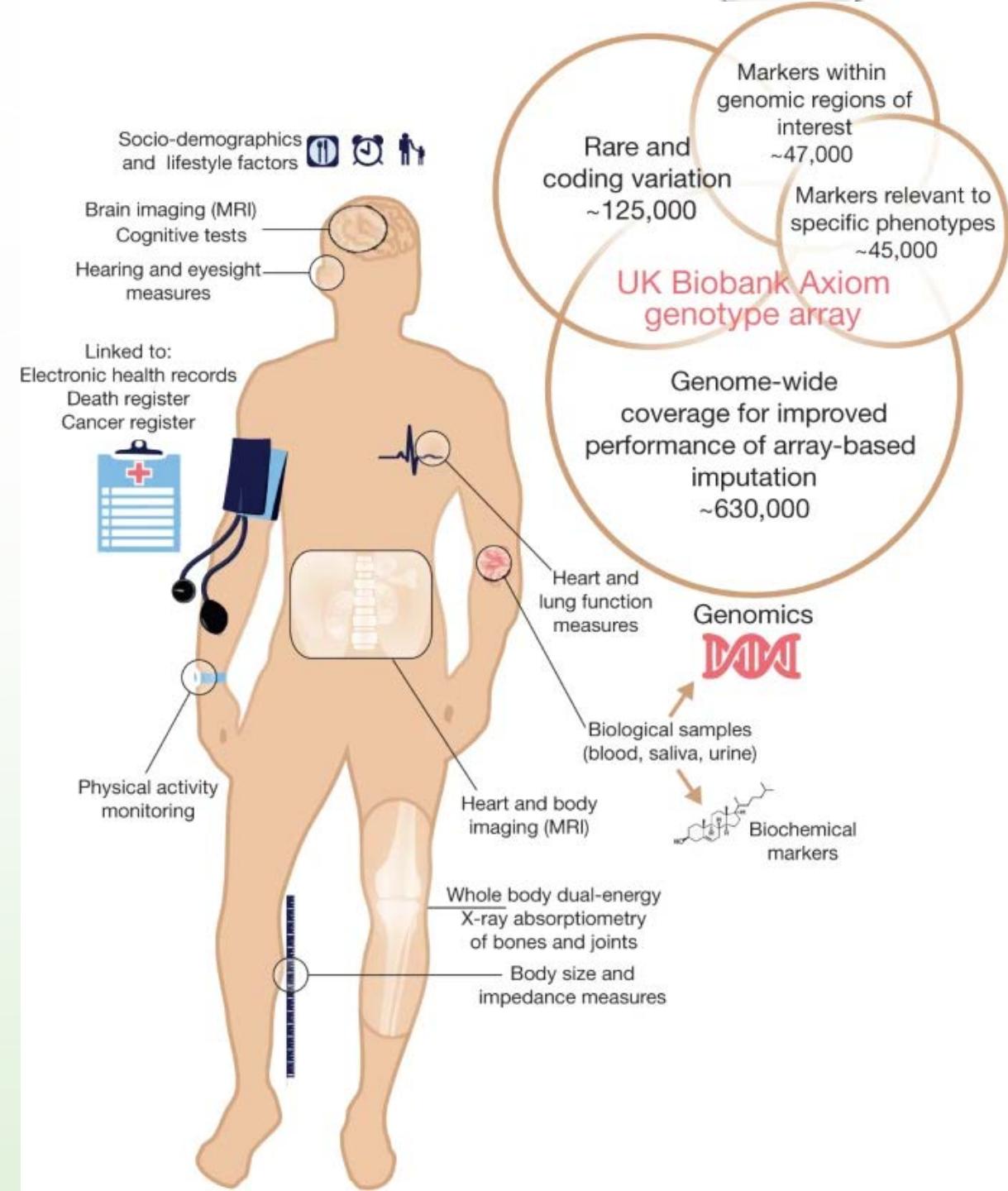
DATABASES - CELL LINES

<https://sites.broadinstitute.org/ccle/>

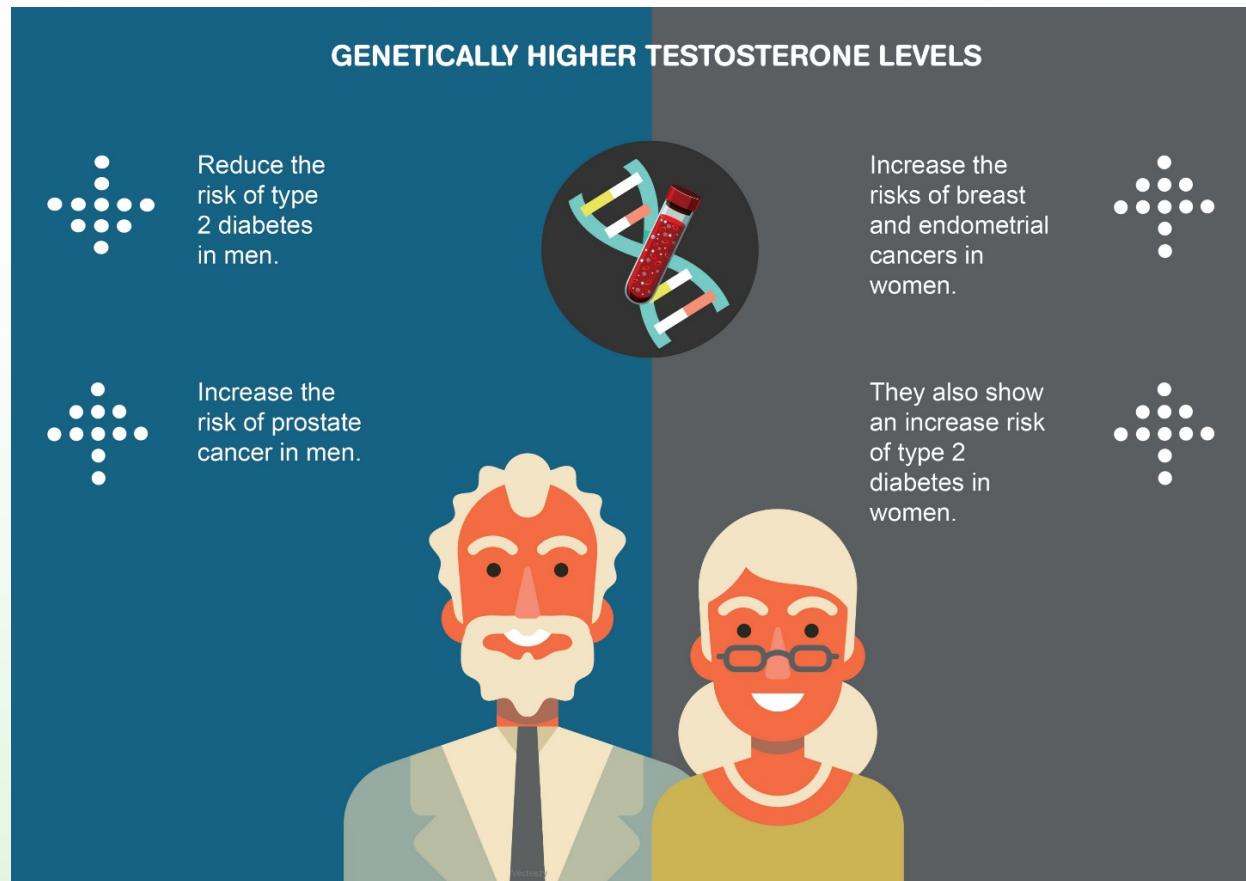
<https://www.nature.com/articles/s41586-019-1186-3>

THE UK BIOBANK

- Longitudinal study considering volunteers
- 500,000 participants
- Ages between 40-69 years in 2006-2010
- Initial assessment at a centre in UK
- Consent for long-term follow-up
- Questionnaires about their health & lifestyle
- Samples: blood, urine and saliva for long-term storage and analysis
- Physical measurements: height, weight, spirometry, blood pressure, heel bone density
- MR brain & heart imaging, activity monitoring and follow-up questionnaires (for some)
- Genetic data on all 500,000 participants.
- PS. Not representative of the general population with evidence of a 'healthy volunteer' selection bias

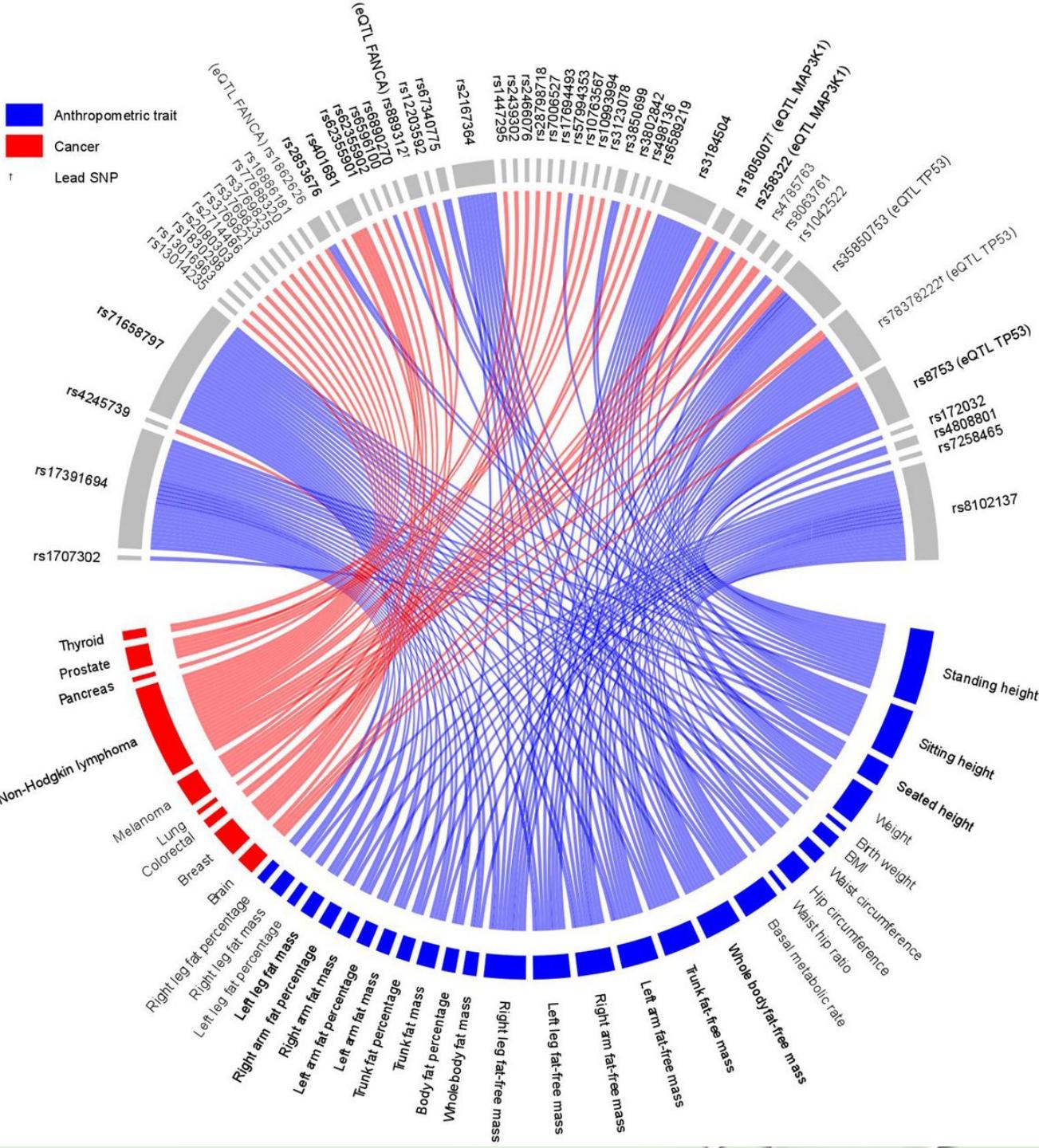


THE UK BIOBANK – EXAMPLES OF STUDIES



HERITABLE GENETIC VARIANTS LINK CANCER RISK WITH ANTHROPOMETRIC TRAITS

<https://img.bmj.com/content/58/6/392>



FOLLOW UP FROM UK BIOBANK

The infographic features the UK Biobank logo at the top left, followed by the title "The Pharma Proteomics Project". Below the title, two main points are highlighted: "Proteins circulating in our blood may play a role in the development of many life-threatening diseases." and "A greater understanding of such markers offers opportunities for more precise, targeted treatment." To the right, a large crowd of diverse people is labeled "53,000 UK Biobank participants". Above them is a test tube containing yellow protein structures, with the text "Analyse over 1,500 proteins". Below the test tube is a circular icon showing laboratory equipment, with the text "Measured by Olink". At the bottom, logos of participating pharmaceutical companies are displayed: Genentech/Biogen, AMGEN, Bristol Myers Squibb/AstraZeneca, REGENERON/gsk, Pfizer, Takeda, and Janssen.

biobank^{uk}
Enabling scientific discoveries that improve human health

The Pharma Proteomics Project

Proteins circulating in our blood may play a role in the development of many life-threatening diseases.

A greater understanding of such markers offers opportunities for more precise, targeted treatment.

53,000 UK Biobank participants

Analyse over 1,500 proteins

Measured by Olink

Genentech
Biogen

AMGEN

Bristol Myers Squibb™
AstraZeneca

REGENERON
gsk

Pfizer

Takeda

Janssen
PHARMACEUTICAL COMPANIES OF
Johnson & Johnson

“META-DATABASES”

- “DATABASES OF DATABASES”
- COLLECT DATA ABOUT DATA TO GENERATE NEW DATA
- HEALTH NEEDS MULTIDISCIPLINARITY
- MANY USERS AND DISPARATE SCENARIOS
- NEED TO GATHER DIVERSE INFORMATION (SEARCH ENGINES)
- INFORMATION IS NEEDED FROM DIFFERENT SOURCES
- NEED TO PROCESS DIVERSE INFORMATION
- MAKE IT AVAILABLE IN SUITABLE FORMAT/CONTEXT/ETC
- MODELS OF METADATA MANAGEMENT ARE CRUCIAL

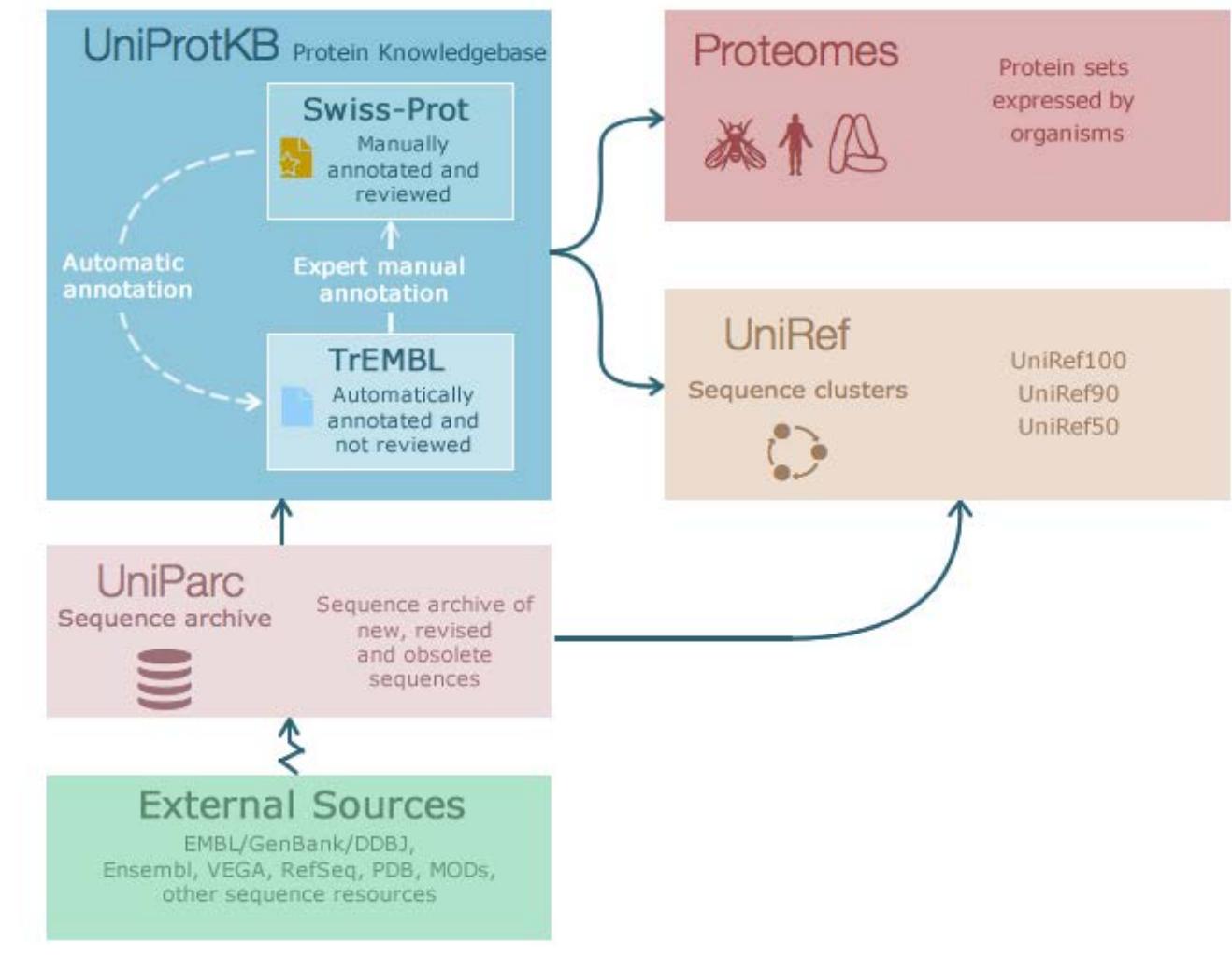
BIOLOGICAL DATABASES

Libraries for biology/chemistry/medical sciences

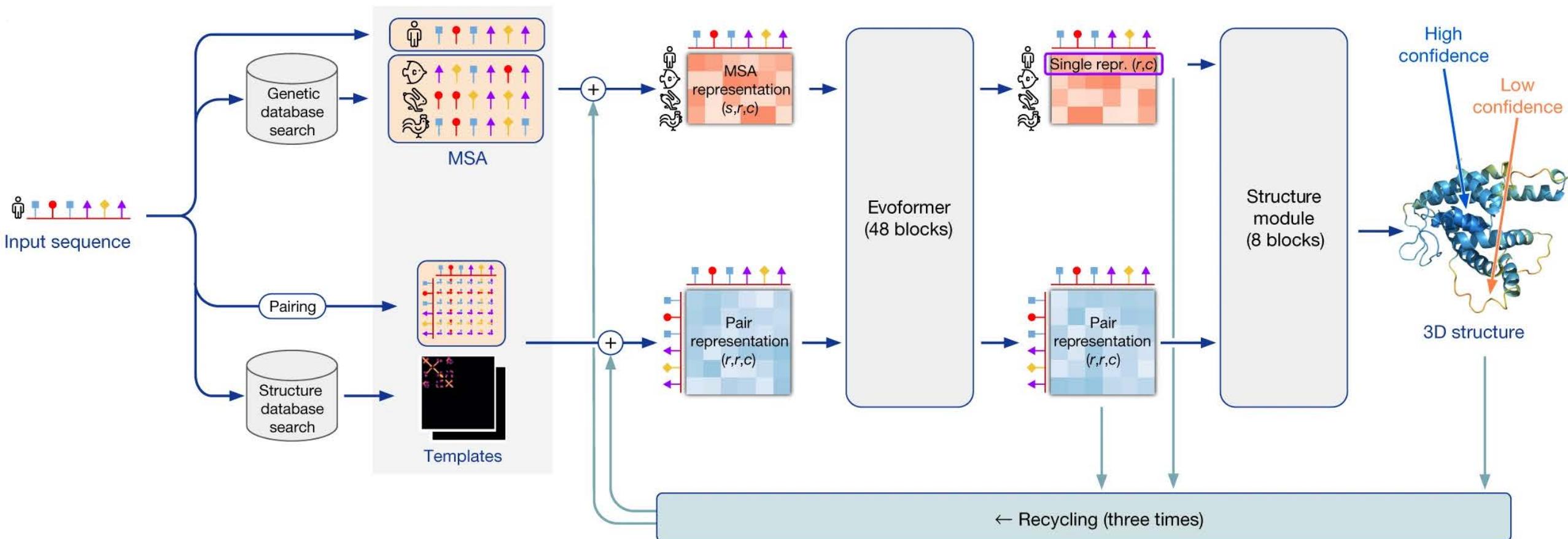
- Scientific experiments
- Published literature
- Computational analyses
- ...

The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data. The UniProt databases are the [UniProt Knowledgebase \(UniProtKB\)](#), the [UniProt Reference Clusters \(UniRef\)](#), and the [UniProt Archive \(UniParc\)](#).

<https://www.uniprot.org/>



ALPHAFOLD



<https://predictioncenter.org/casp14/index.cgi>

<https://www.nature.com/articles/s41586-021-03819-2>

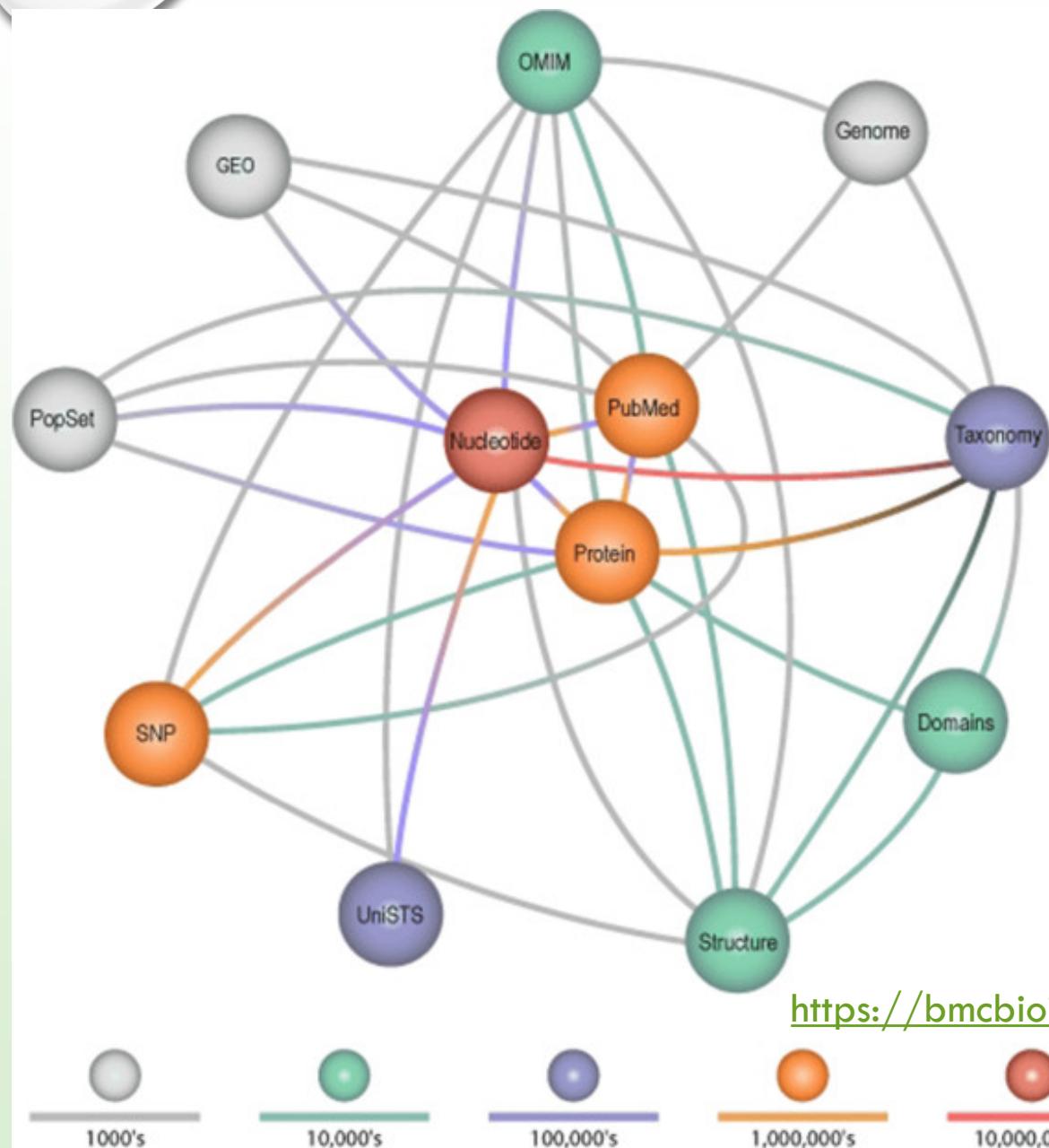
https://www.youtube.com/watch?v=B9PL_gVxLI

<https://alphafold.ebi.ac.uk/>

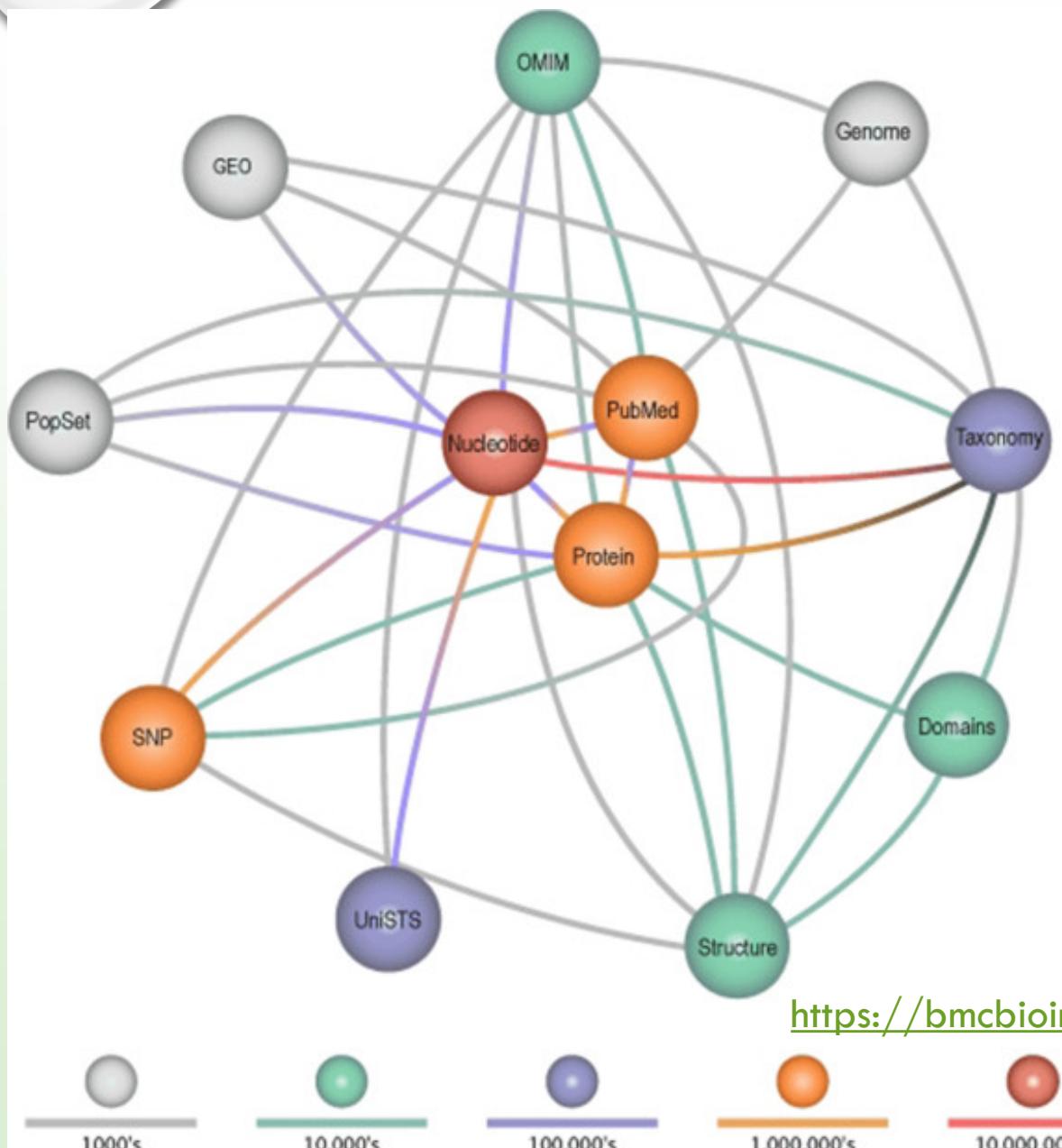
AlphaFold
Protein Structure Database

Developed by DeepMind and EMBL-EBI

NCBI DATABASE



NCBI DATABASE



<https://www.ncbi.nlm.nih.gov/Entrez/>

 NCBI

Entrez Molecular Sequence Database System

PubMed Entrez BLAST OMIM Taxonomy Structure

▶ **Introduction**

Entrez is a molecular biology database system that provides integrated access to nucleotide and protein sequence data, gene-centered and genomic mapping information, 3D structure data, PubMed MEDLINE, and more. The system is produced by the National Center for Biotechnology Information (NCBI) and is available via the Internet.

<https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-73>

1000's

10,000's

100,000's

1,000,000's

10,000,000's

National Center for Biotechnology Information (NCBI)

A GENE-FINDING MODEL

5' untranslated region (UTR)
of a messenger RNA
(mRNA): directly upstream
from the initiation codon

Input: genomic DNA sequence

GCCTGGGAAAACCCTCAACTT... .

Gene model

Internal
exon

Internal
intron

Final
exon

Initial
exon

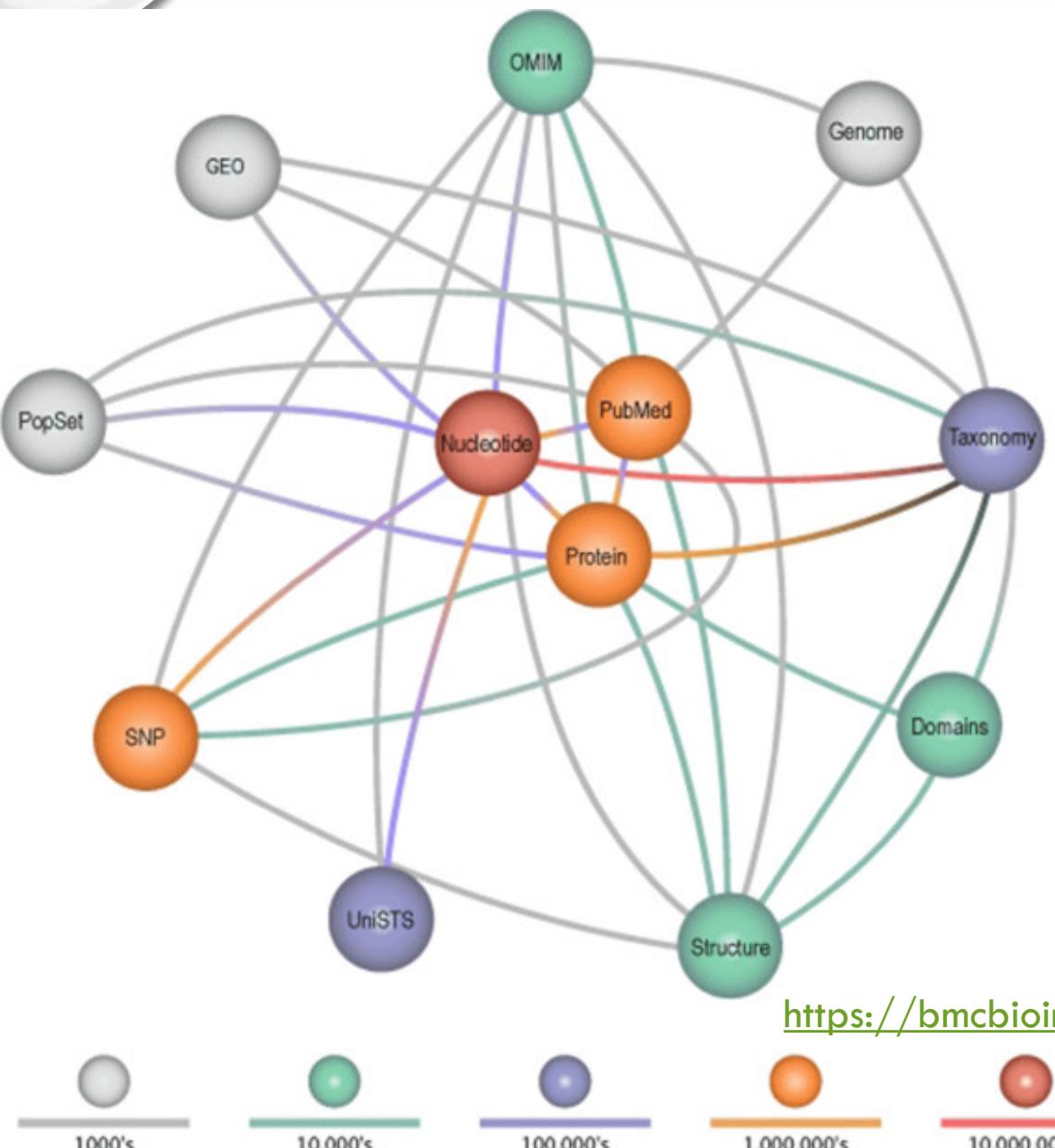
Non-
coding

3' untranslated region (UTR)
of a messenger RNA
(mRNA): immediately follows
the translation termination
codon

Output: gene annotation



NCBI DATABASE



 NCBI

Entrez Molecular Sequence Database System

PubMed Entrez BLAST OMIM Taxonomy Structure

▶ **Introduction**

Entrez is a molecular biology database system that provides integrated access to nucleotide and protein sequence data, gene-centered and genomic mapping information, 3D structure data, PubMed MEDLINE, and more. The system is produced by the National Center for Biotechnology Information (NCBI) and is available via the Internet.

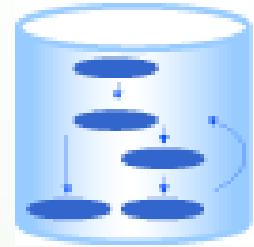
WHY: INTERPRETING RESULTS FROM OMICS ANALYSES

gene.name	accession	unigene	Fold Change	Adjusted p
Adamdec1	NM_021475	Mm.36742	27.31	0.00
Vcam1	BB250384	Mm.76649	26.58	0.02
Itgax	NM_021334	Mm.22378	17.06	0.03
Vcam1	BB250384	Mm.76649	13.57	0.00
5830443L24Rik	NM_029509	Mm.425261	11.76	0.00
Gna12	AV238106	Mm.370185	11.64	0.00
Vcam1	L08431	Mm.76649	9.35	0.00
Ptgs1	AA833146	Mm.275434	8.01	0.00
Acp5	AK008391	Mm.46354	7.78	0.03
Cd4	NM_013488	Mm.2209	7.53	0.004
Ptgs1	BB520073	Mm.275434	7.44	0.005

HOW DO WE REPRESENT A PATHWAY

RAB23	GAS1	IHH
ZIC2	PTCH1	GLI1
SMO	HHIP	BMP2
WNT16	PTCH1	STK36
FBXW11	CSNK1E	SUFU
LRP2	PRKACA	GSK3B

A list or set of genes



MSigDB

Molecular Signatures Database

Gene sets

List of genes acting in a pathway/determine a phenotypes

<https://www.gsea-msigdb.org/gsea/msigdb>

Collections

The MSigDB gene sets are divided into 8 major collections:

H **hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

C1 **positional gene sets** for each human chromosome and cytogenetic band.

C2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

C3 **motif gene sets** based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.

C4 **computational gene sets** defined by mining large collections of cancer-oriented microarray data.

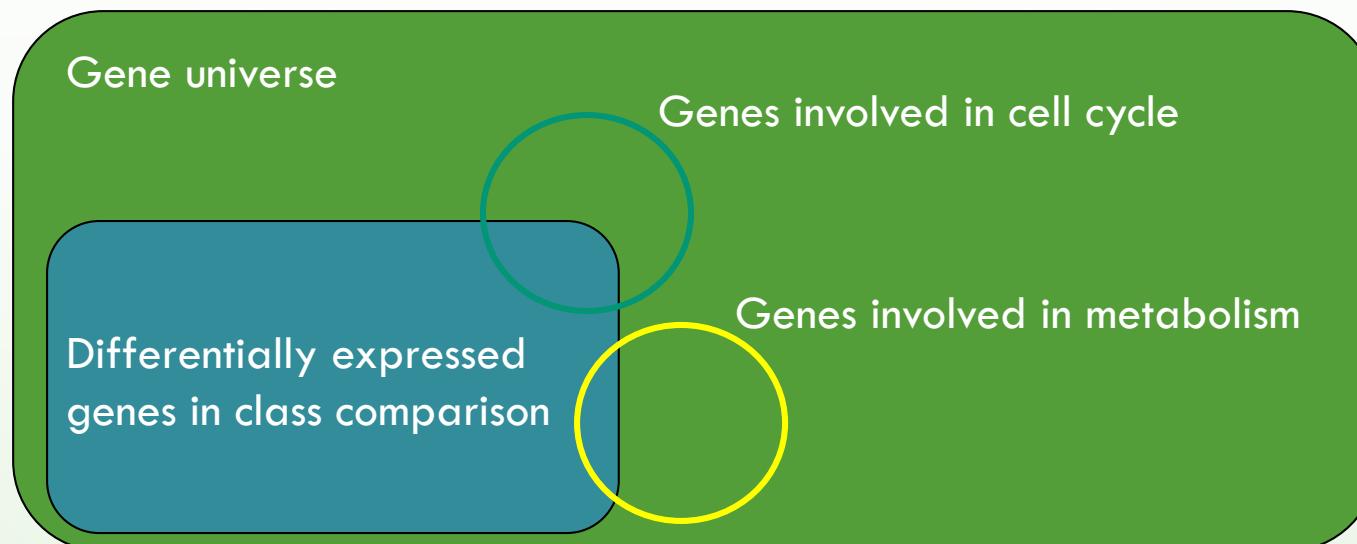
C5 **GO gene sets** consist of genes annotated by the same GO terms.

C6 **oncogenic gene sets** defined directly from microarray gene expression data from cancer gene perturbations.

C7 **immunologic gene sets** defined directly from microarray gene expression data from immunologic studies.

Over-representation analysis

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



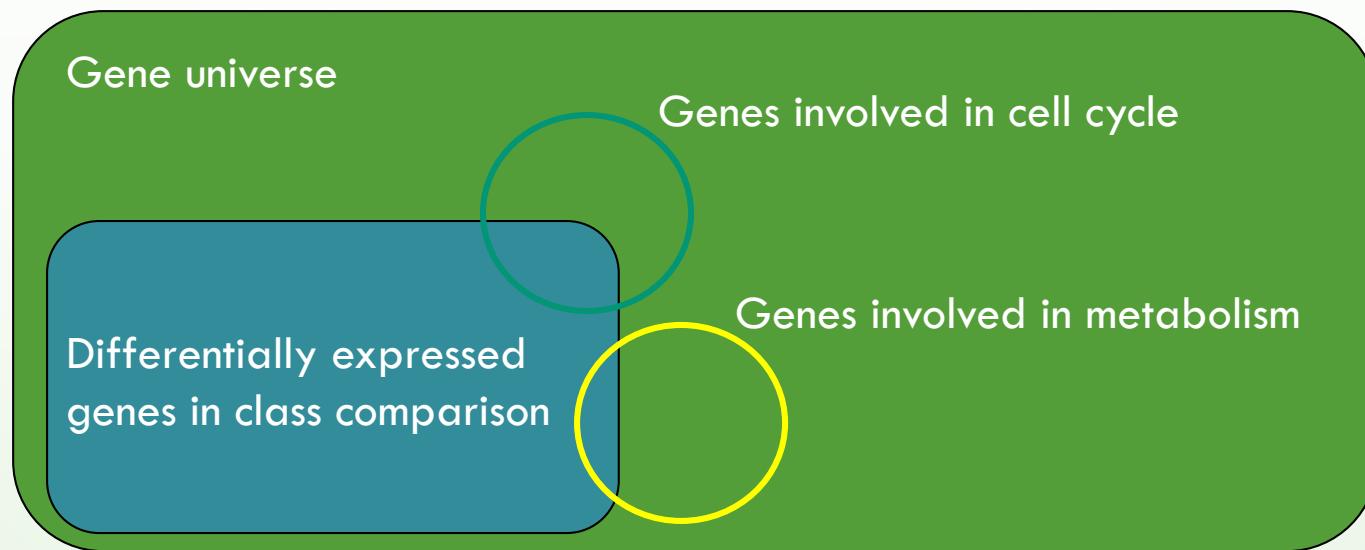
Two-way table:

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

$$\text{Fold enrichment} = (22/50) / (7500/30000) = 45\% / 25\% = 1.8$$

Over-representation analysis

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



"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

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

LIMITATIONS

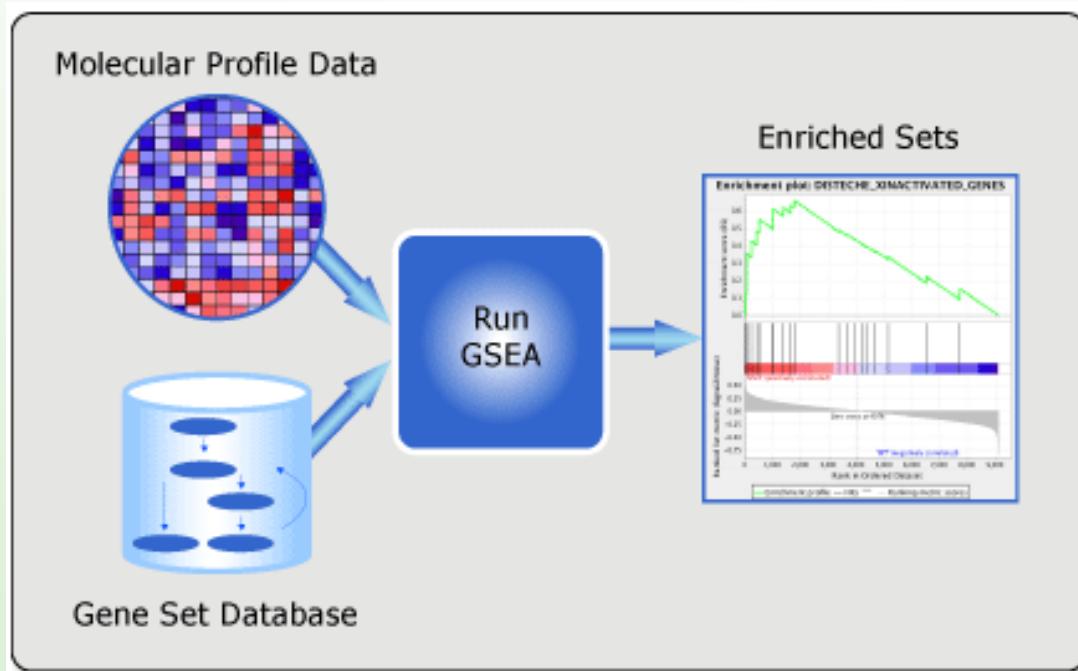
- Base on gene sets/lists
- Genes are assumed to be independent and the correlation structure is ignored
- Not always clear how to define the universe
- The over-representation analysis is independent of the changes measured. All genes are treated equally.
- Only the most significant genes are used =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

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.



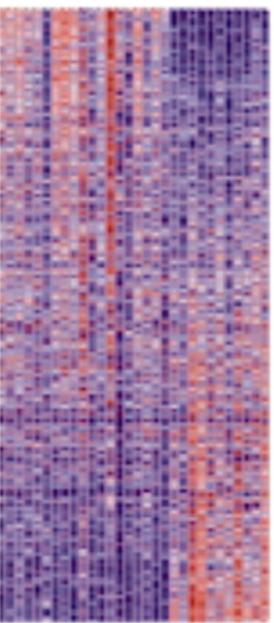
<http://www.broadinstitute.org/gsea/>

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

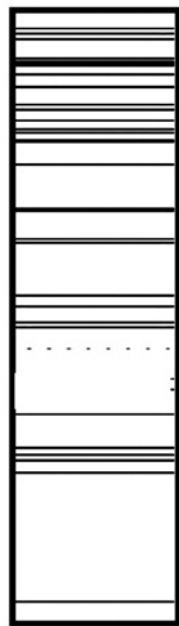


Phenotype
Classes
A B

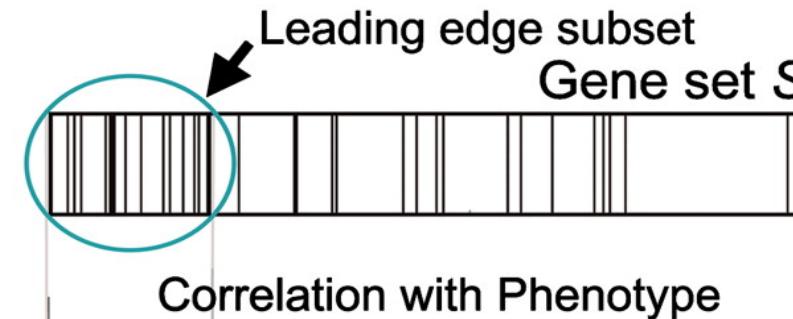
Ranked Gene List



Gene set S



Leading edge subset
Gene set S



Correlation with Phenotype

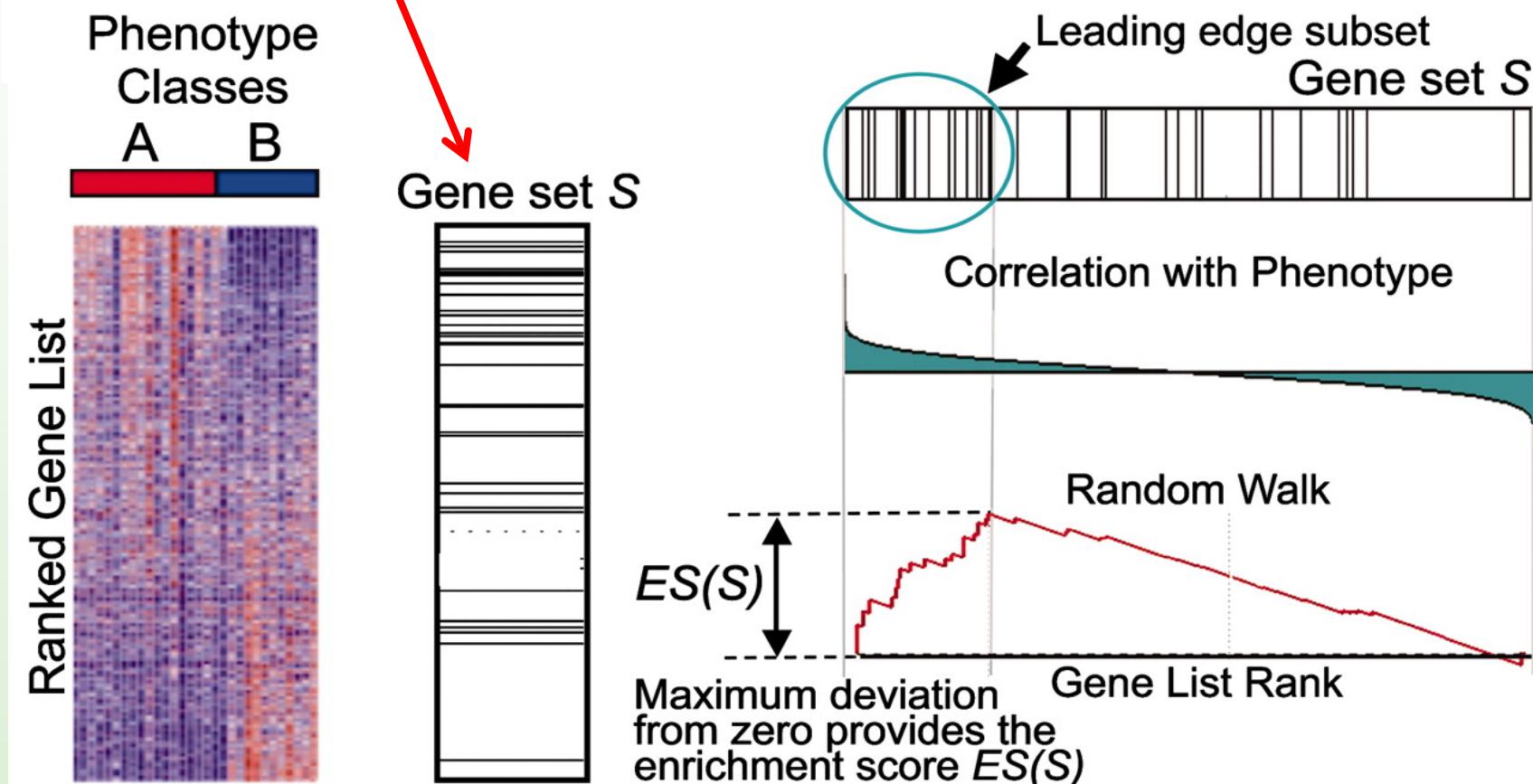
Random Walk

$ES(S)$

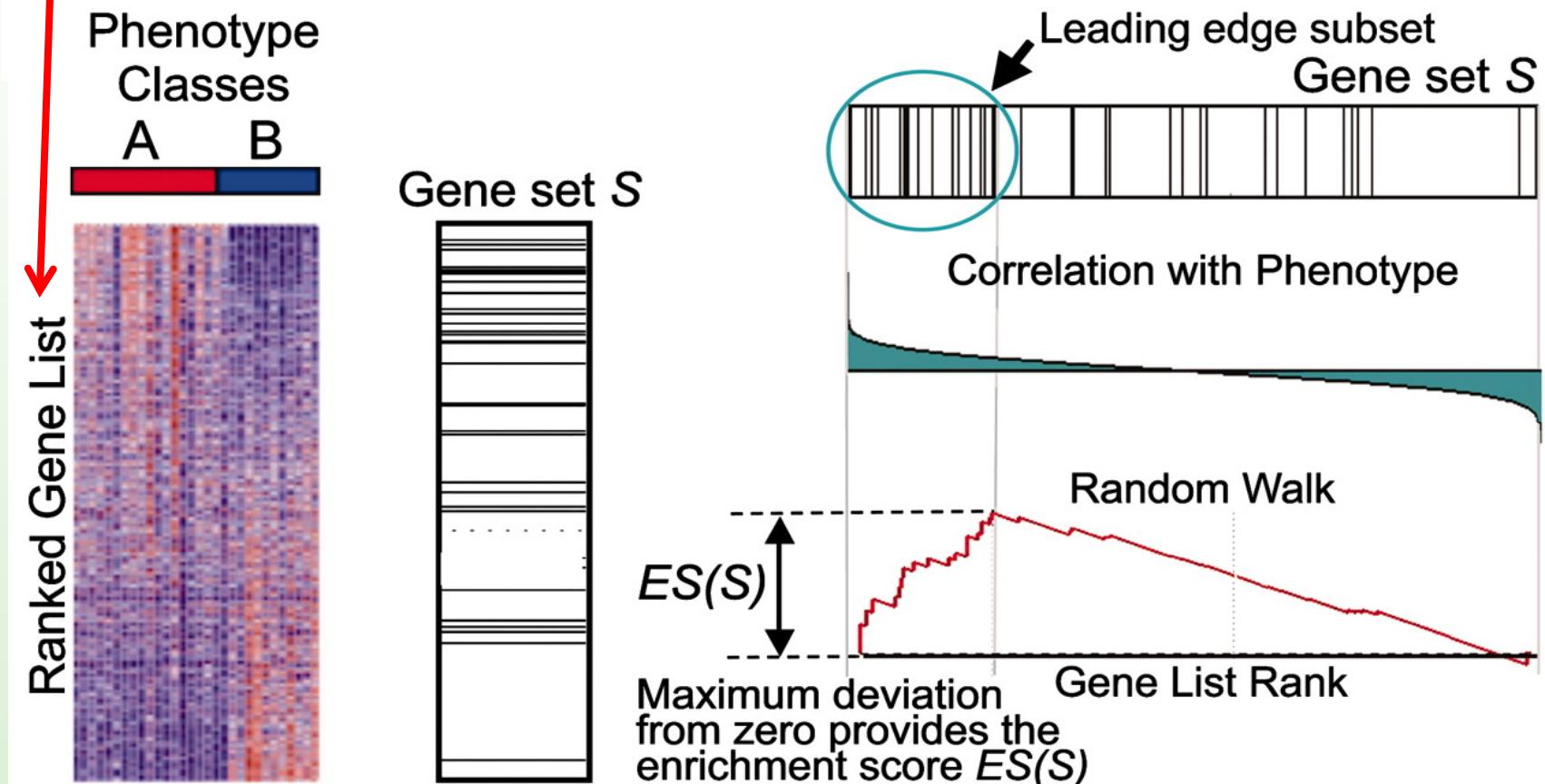
Maximum deviation
from zero provides the
enrichment score $ES(S)$

Gene List Rank

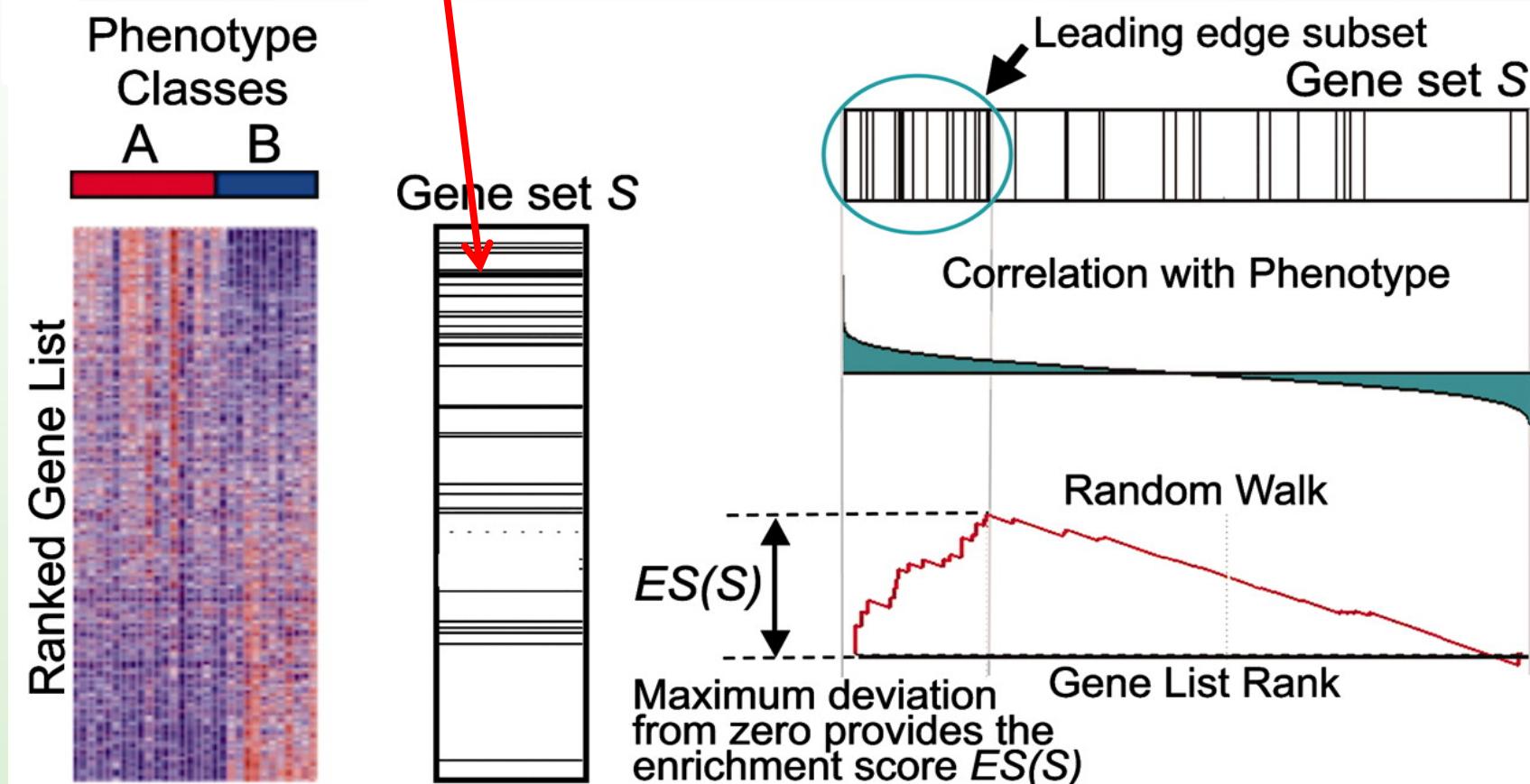
- 1) Define phenotype classes (e.g. Cells in Hypoxia or Normoxia)
- 2) Define a gene set of interest (e.g. Genes in the HIF1 α pathway)



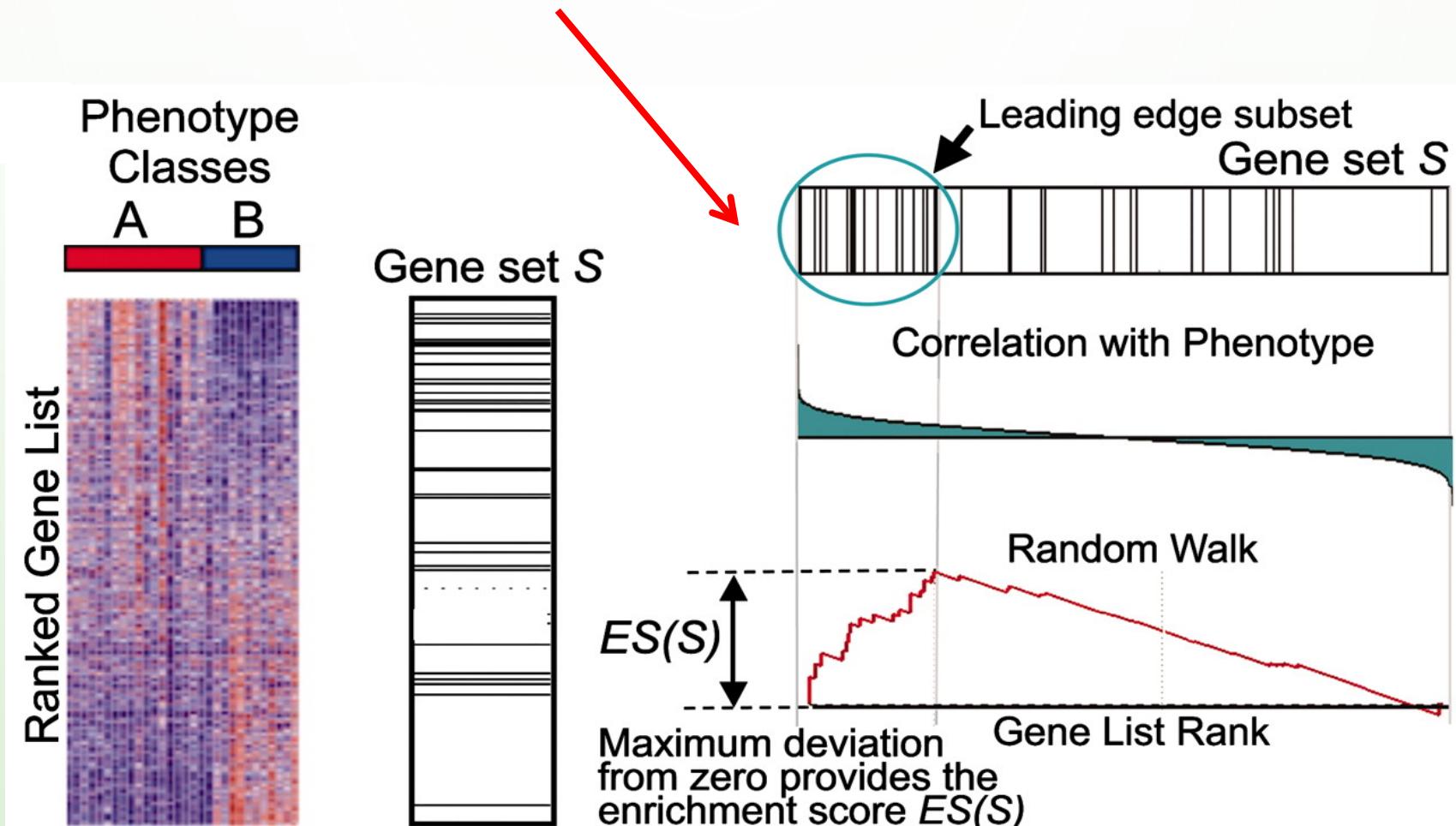
- 1) Define phenotype classes (e.g. Cells in Hypoxia or Normoxia)
- 2) Define a gene set of interest (e.g. Genes in the HIF1 α pathway)
- 3) Sort genes based on their differential expression between classes



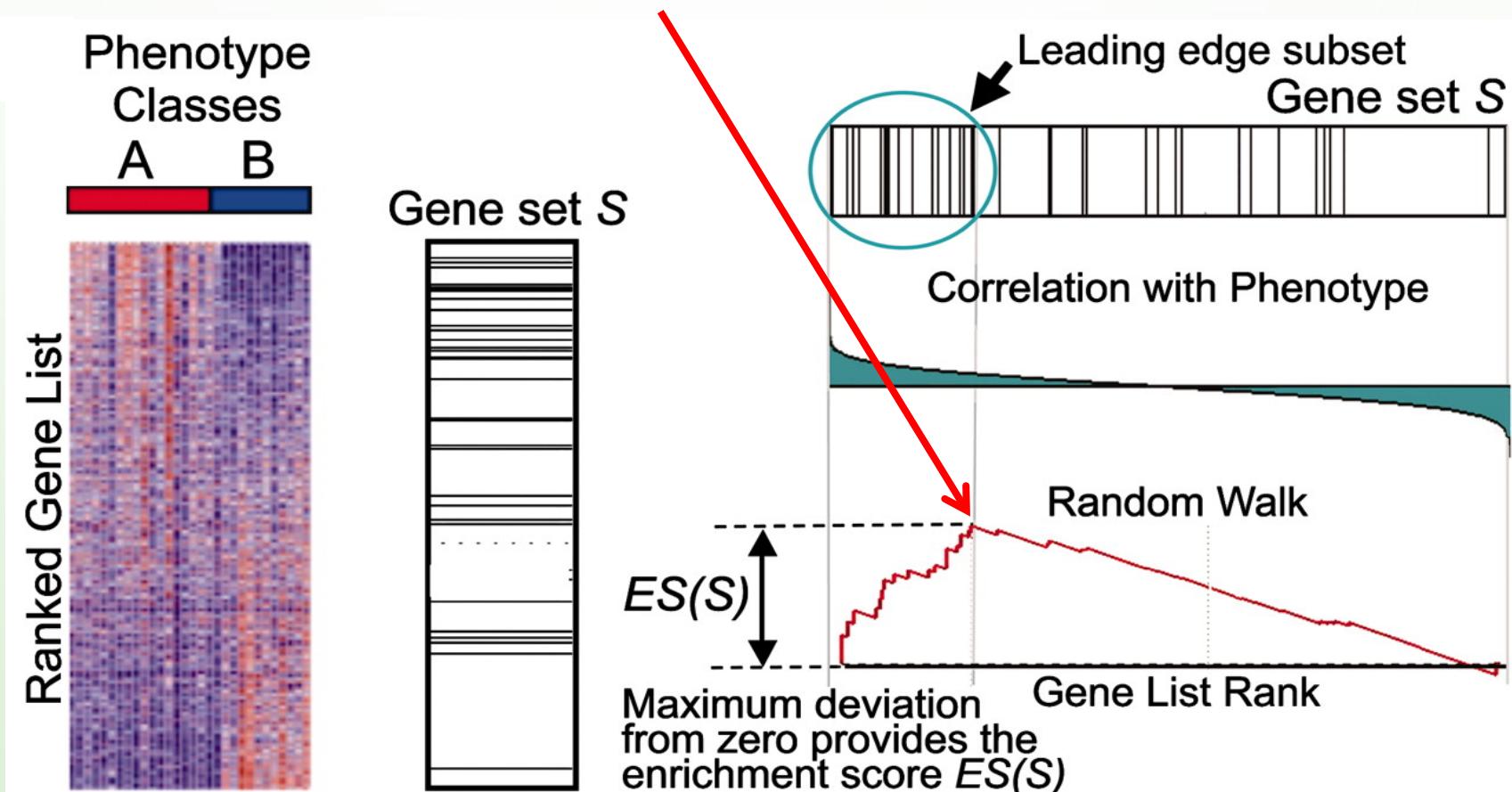
- 1) Define phenotype classes (e.g. Cells in Hypoxia or Normoxia)
- 2) Define a gene set of interest (e.g. Genes in the HIF1 α pathway)
- 3) Sort genes based on their differential expression between classes
- 4) Tag genes from the set S within the sorted list



- 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

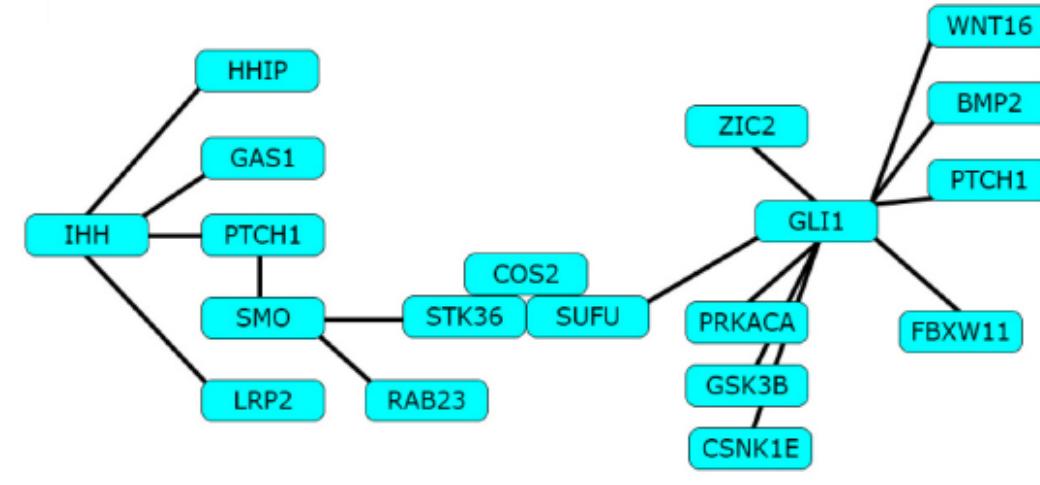
LIMITATIONS

- 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 are used
- Pathway are still considered independent. However a gene can function in more than one pathway, meaning that pathways gene sets 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)

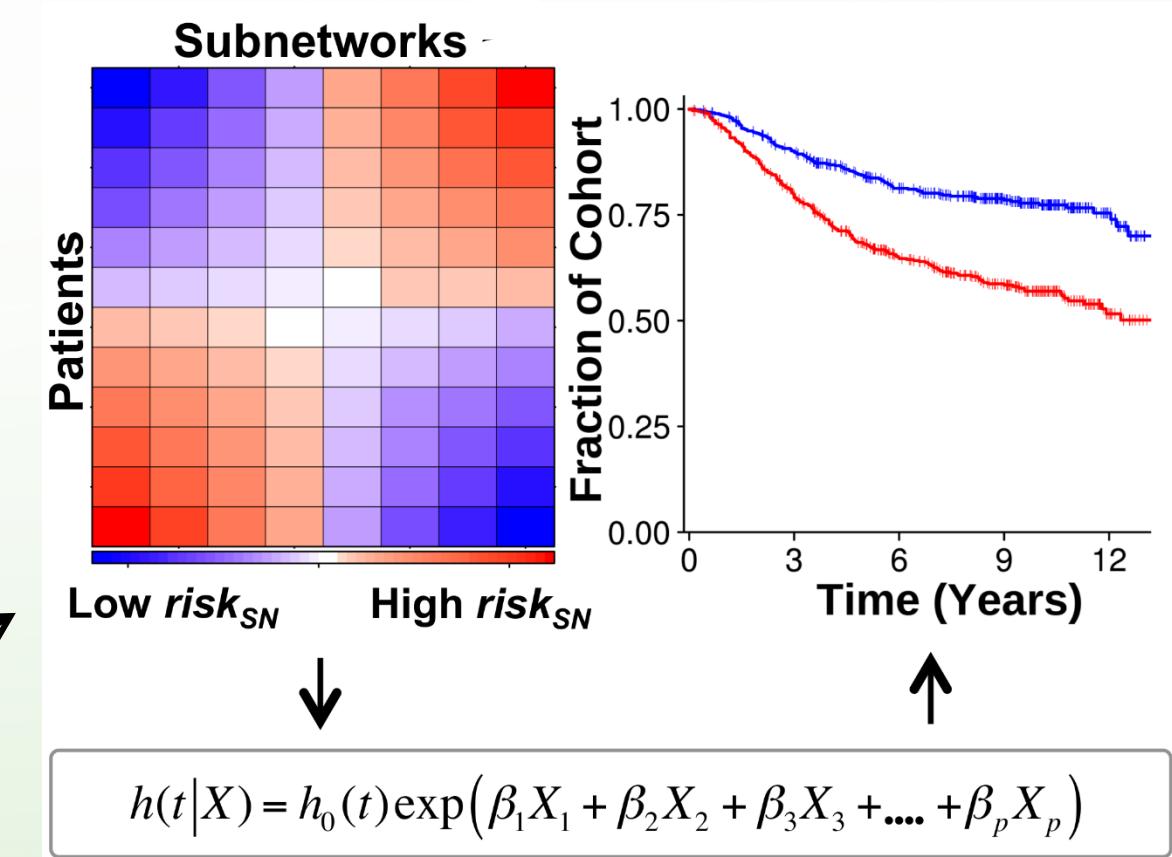
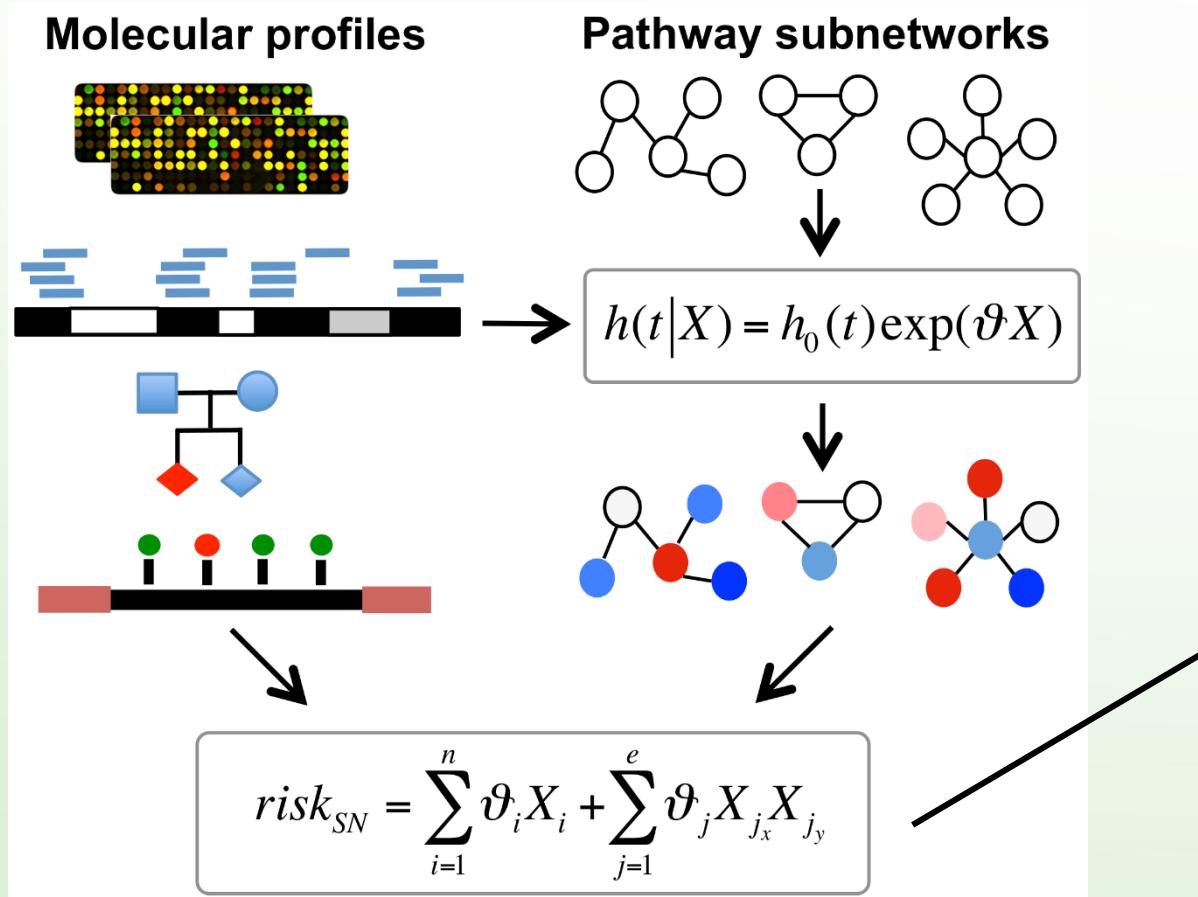
HOW DO WE REPRESENT A PATHWAY

RAB23	GAS1	IHH
ZIC2	PTCH1	GLI1
SMO	HHIP	BMP2
WNT16	PTCH1	STK36
FBXW11	CSNK1E	SUFU
LRP2	PRKACA	GSK3B

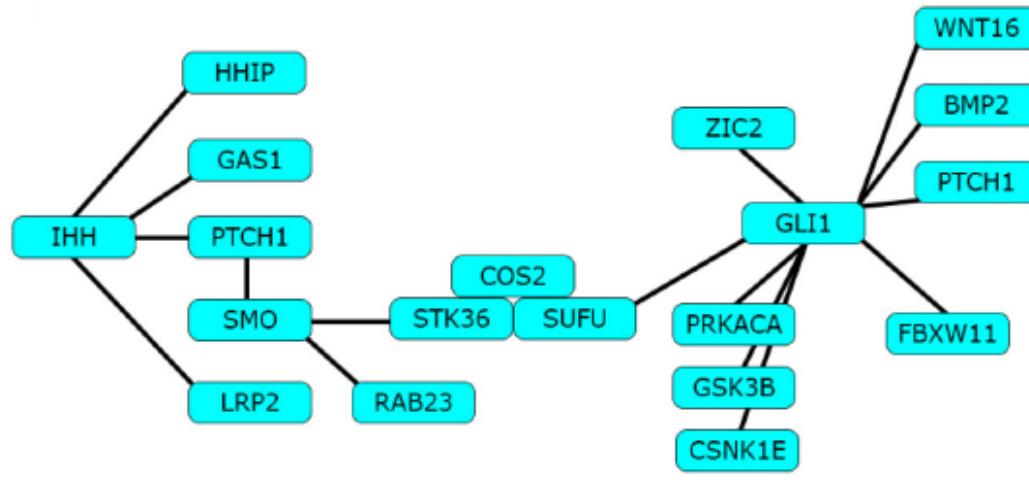
A network of associated genes



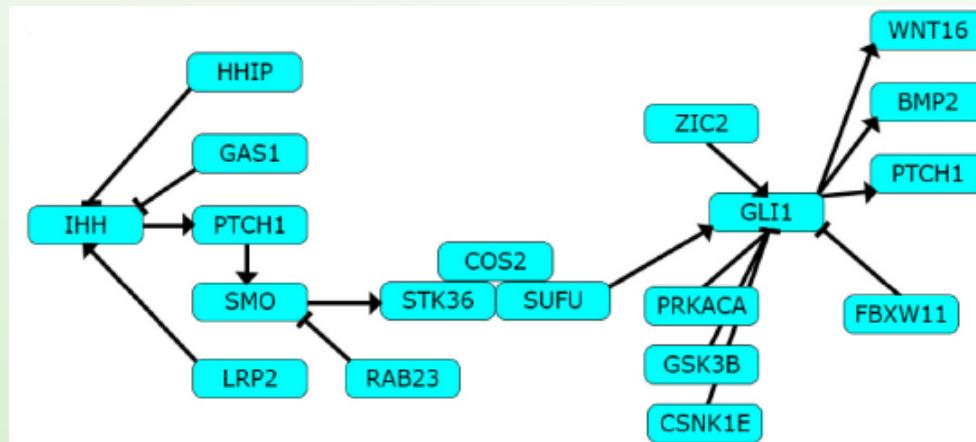
PATHWAY-BASED SUBNETWORKS FOR BIOMARKER DISCOVERY



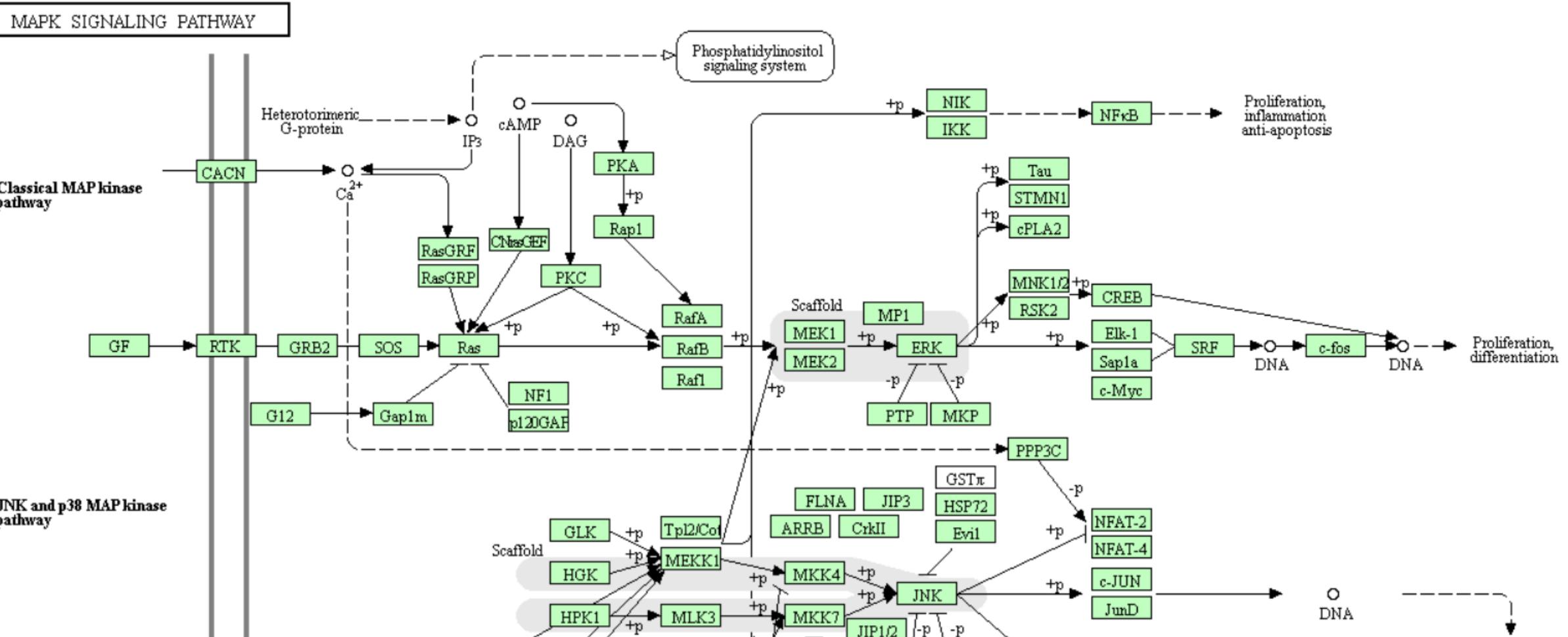
HOW DO WE REPRESENT A PATHWAY



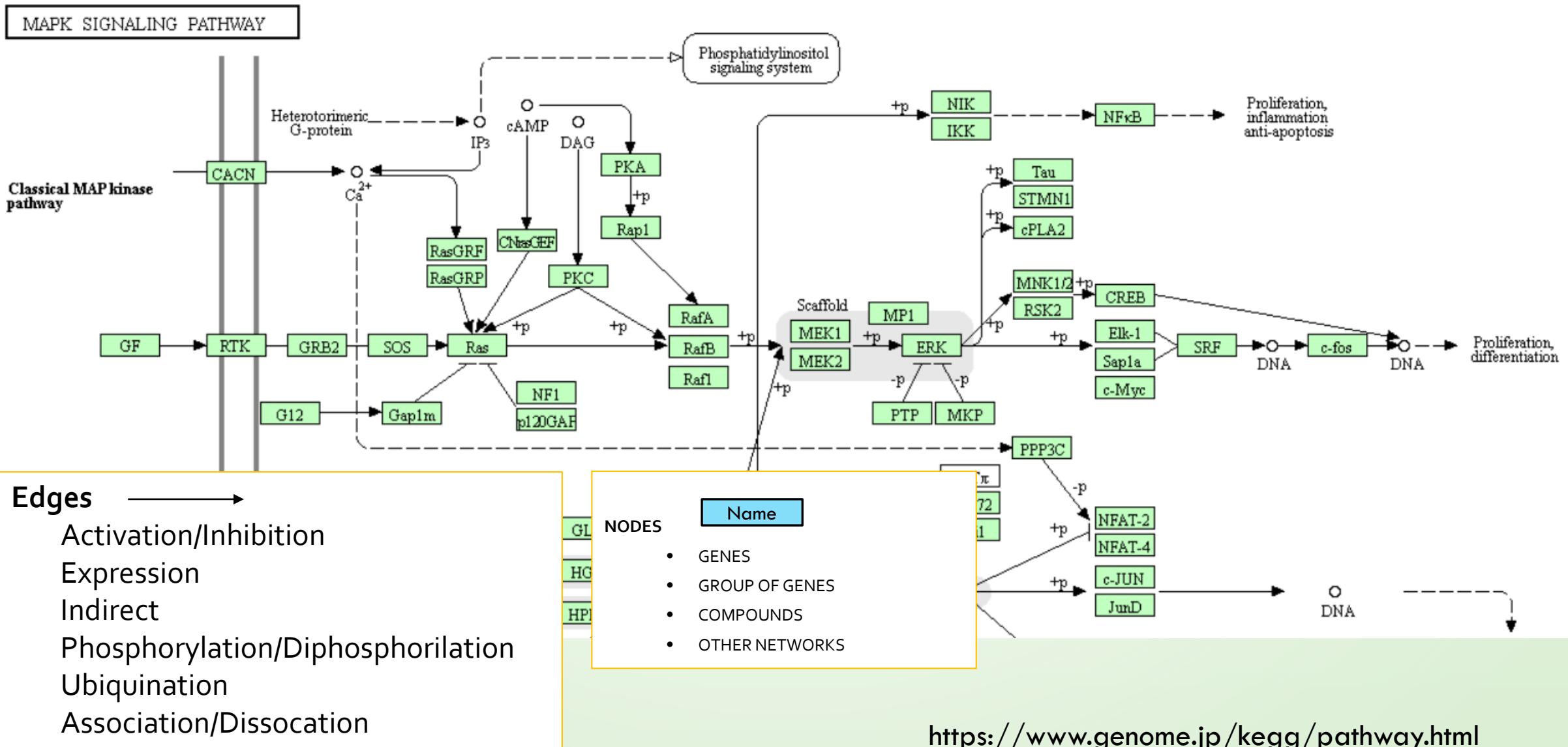
A directed network of genes



KEGG PATHWAYS



Kegg pathways



MINEPATH

www.minepath.org

Apps Training Tools Other bookmarks

Input

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?

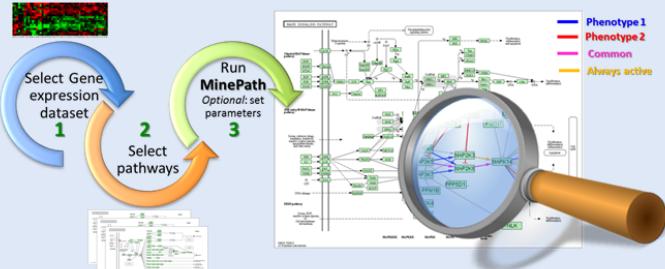
MinePath



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](#)

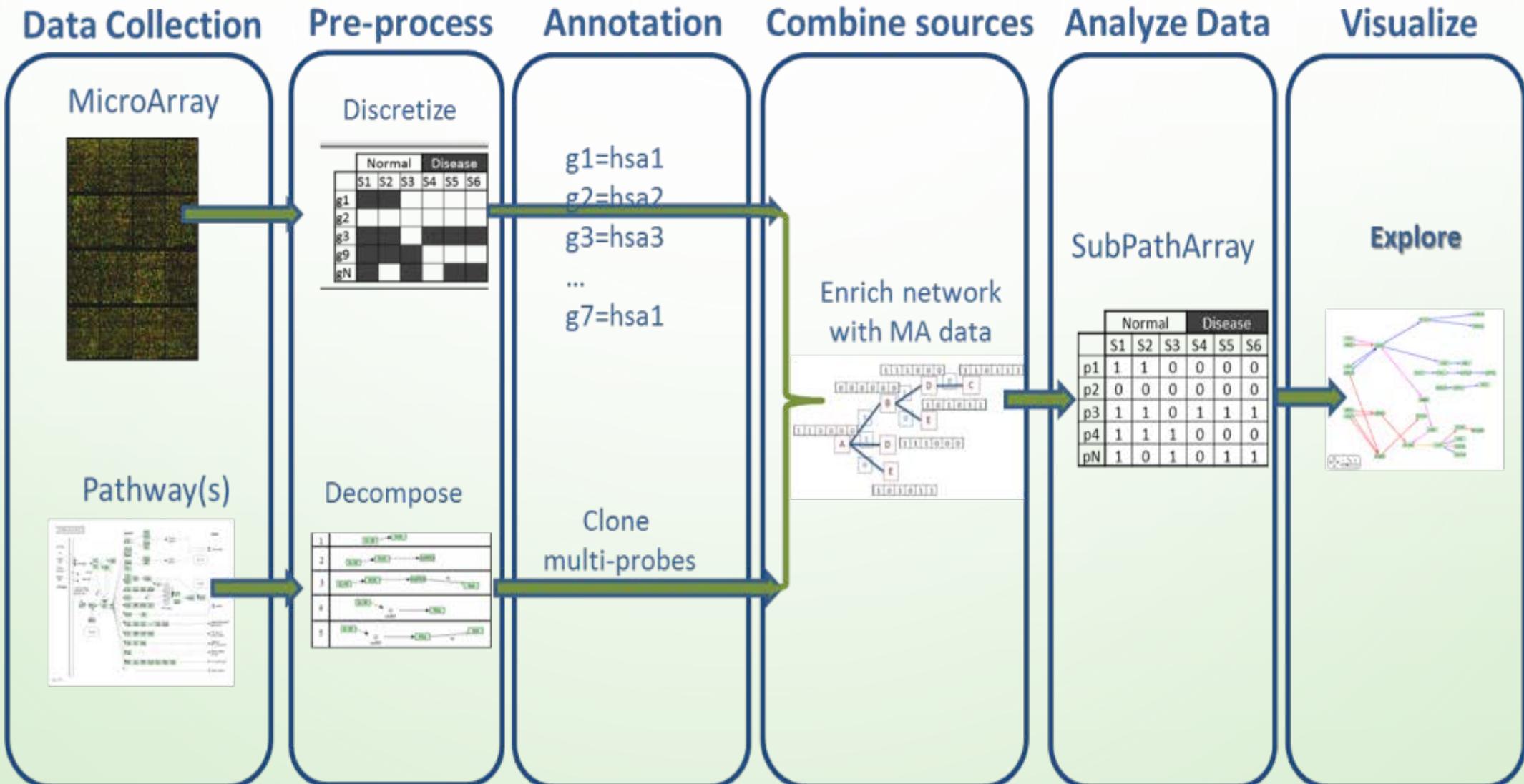
Acknowledgments: This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund and through the Operational Programme «Competitiveness and Entrepreneurship» of GSRT Cooperation project: EDGE 092YN-13-901. Details for the PhD (in Greek) can be found at the web page: <http://www.logistics.tuc.gr/koumakis/>

Citing MinePath

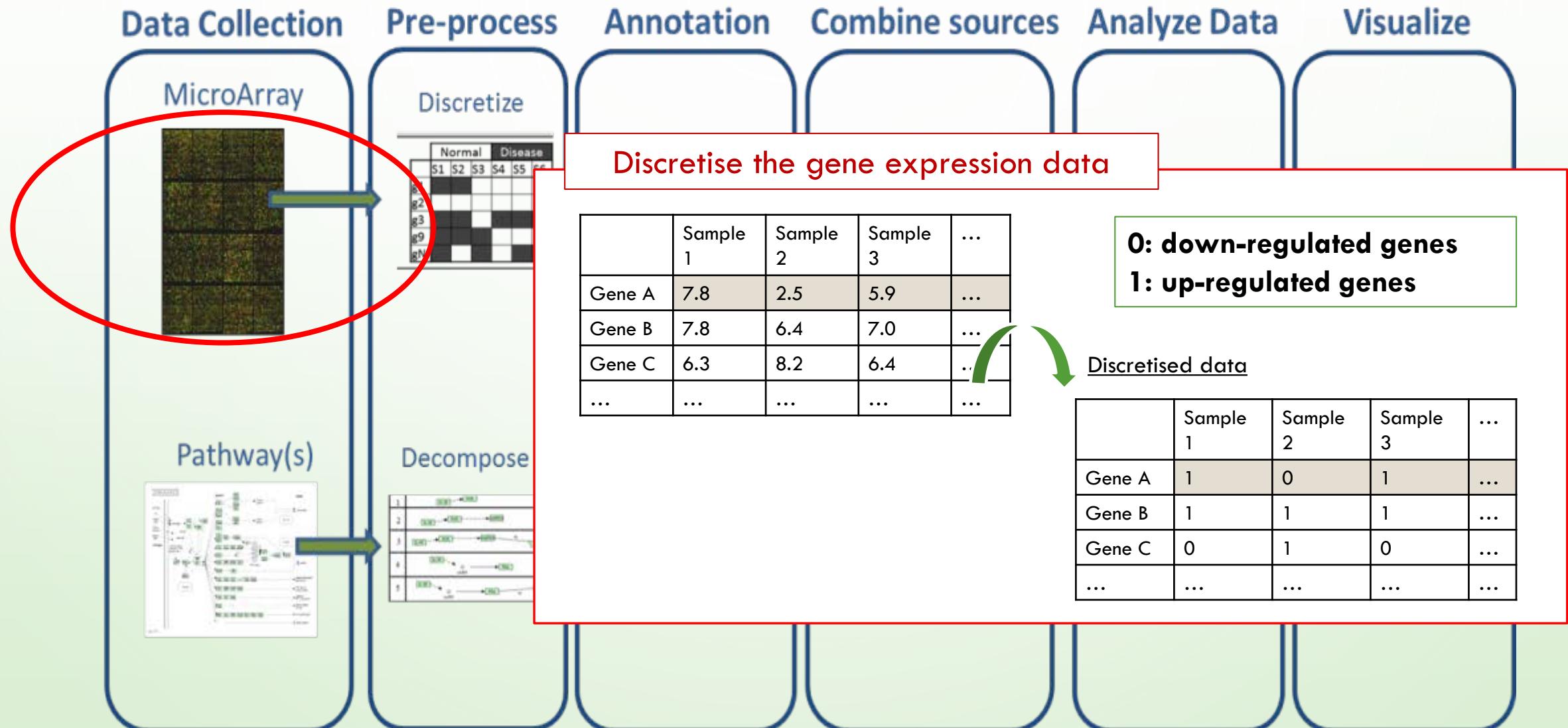
111 Koumakis I, Mavroukeli IA, Tsaknis MC, Kalafatisou D, S. Paraskevopoulou C, A. (2012). Coupling Regulatory Networks and Microarray Data Using Molecular Oscillations of Breast Cancer Treatment Responses. Artificial Intelligence Theories and Applications. Lecture Notes in Computer Science, 7307.

Summary

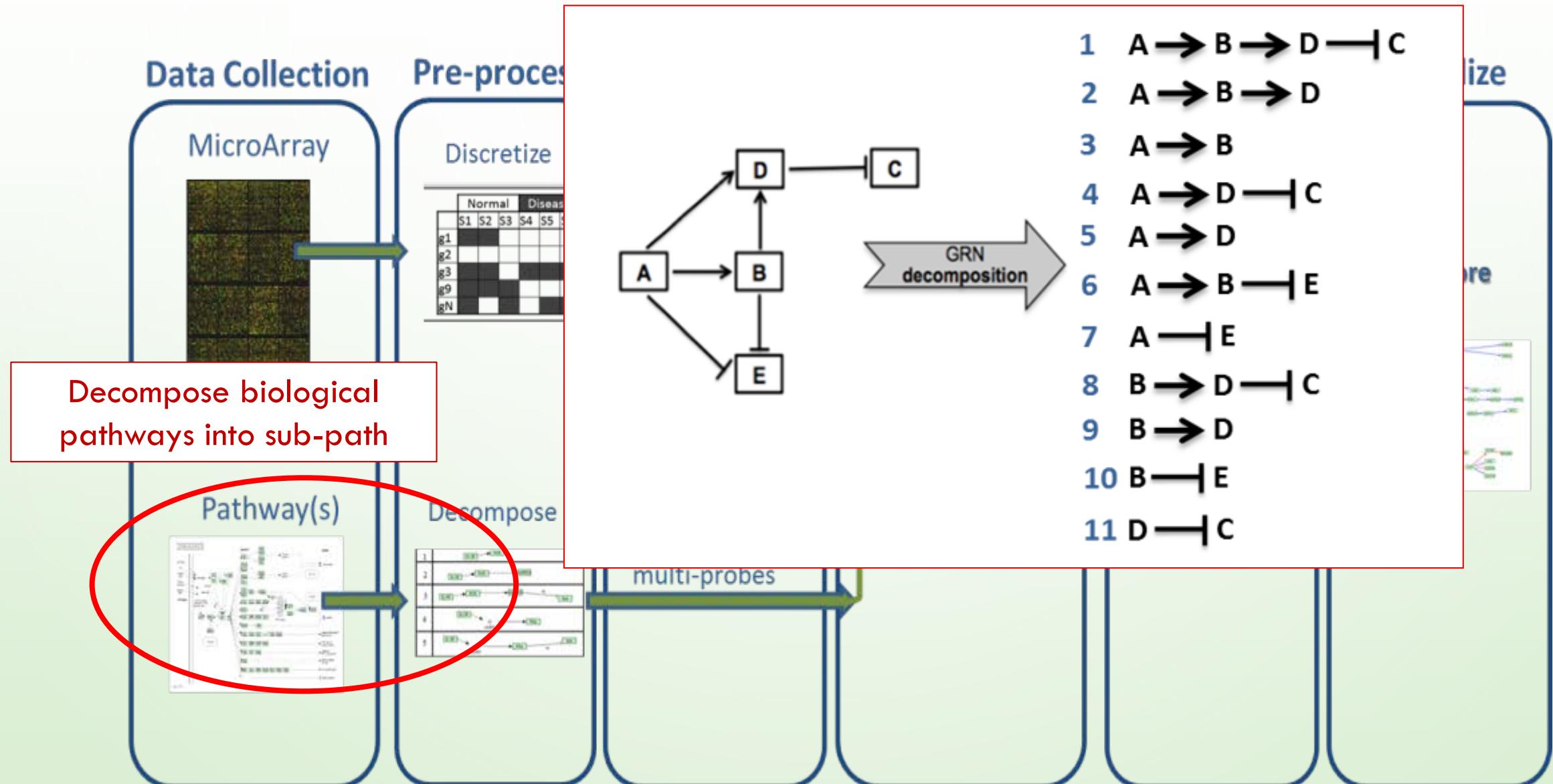
SYSTEM OVERVIEW



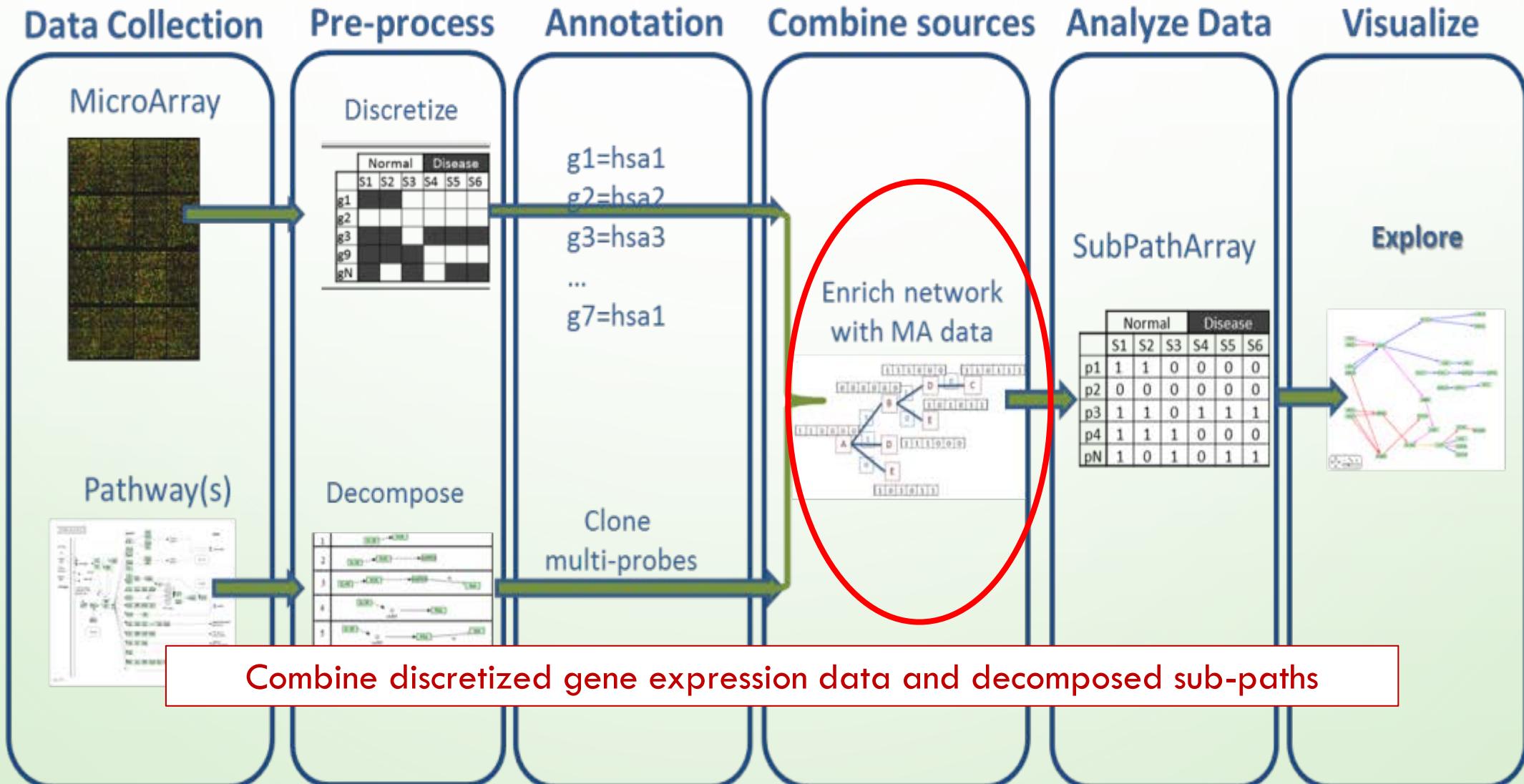
SYSTEM OVERVIEW



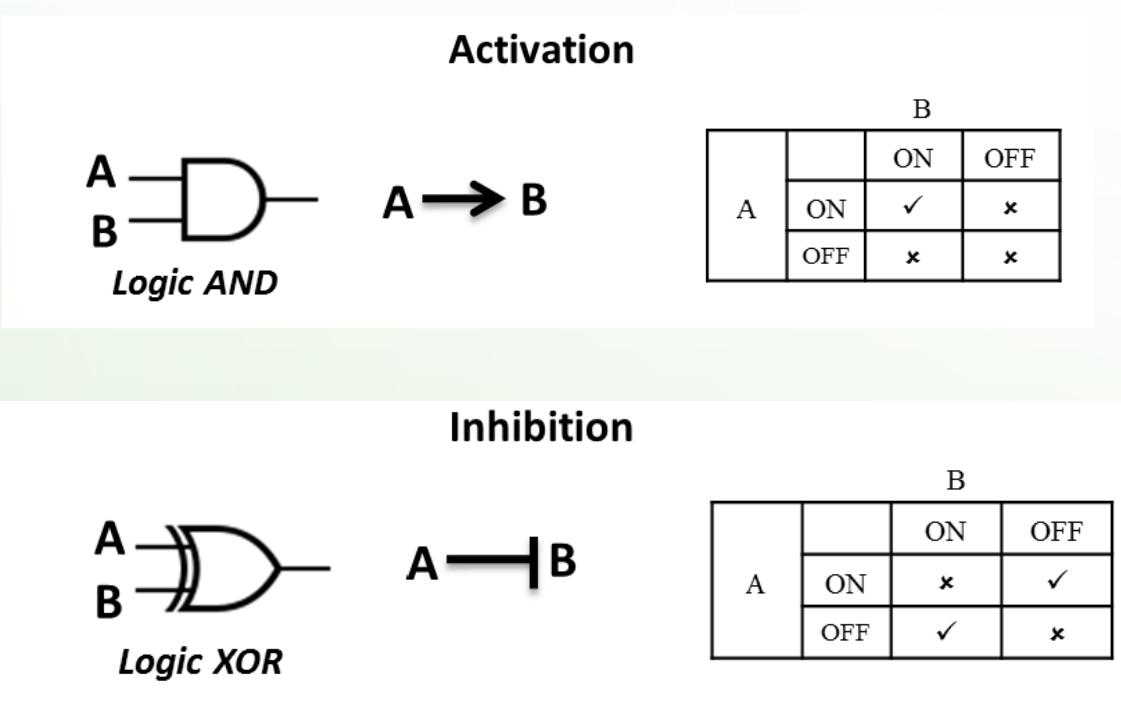
SYSTEM OVERVIEW



SYSTEM OVERVIEW



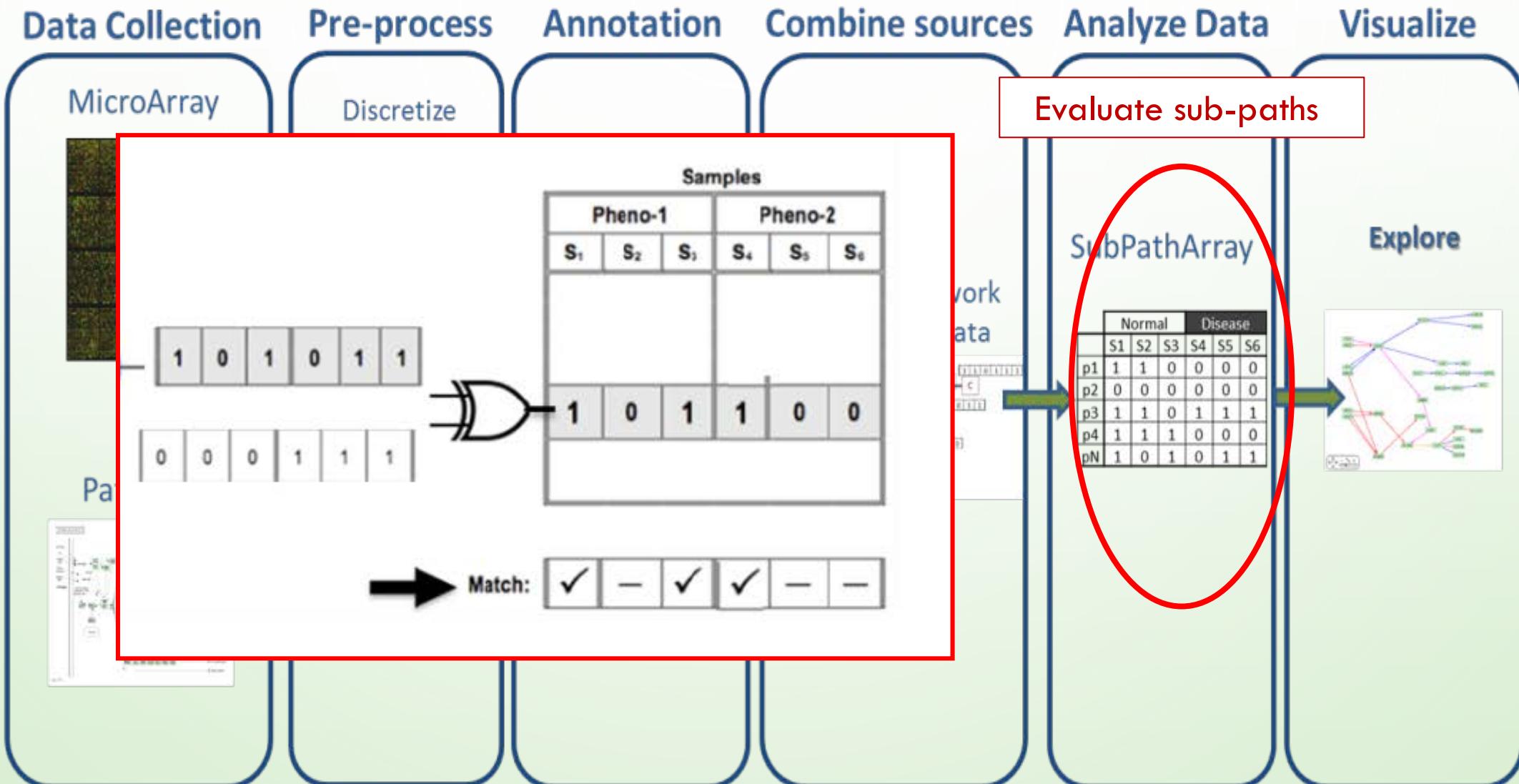
Combine discretized gene expression data and decomposed sub-paths



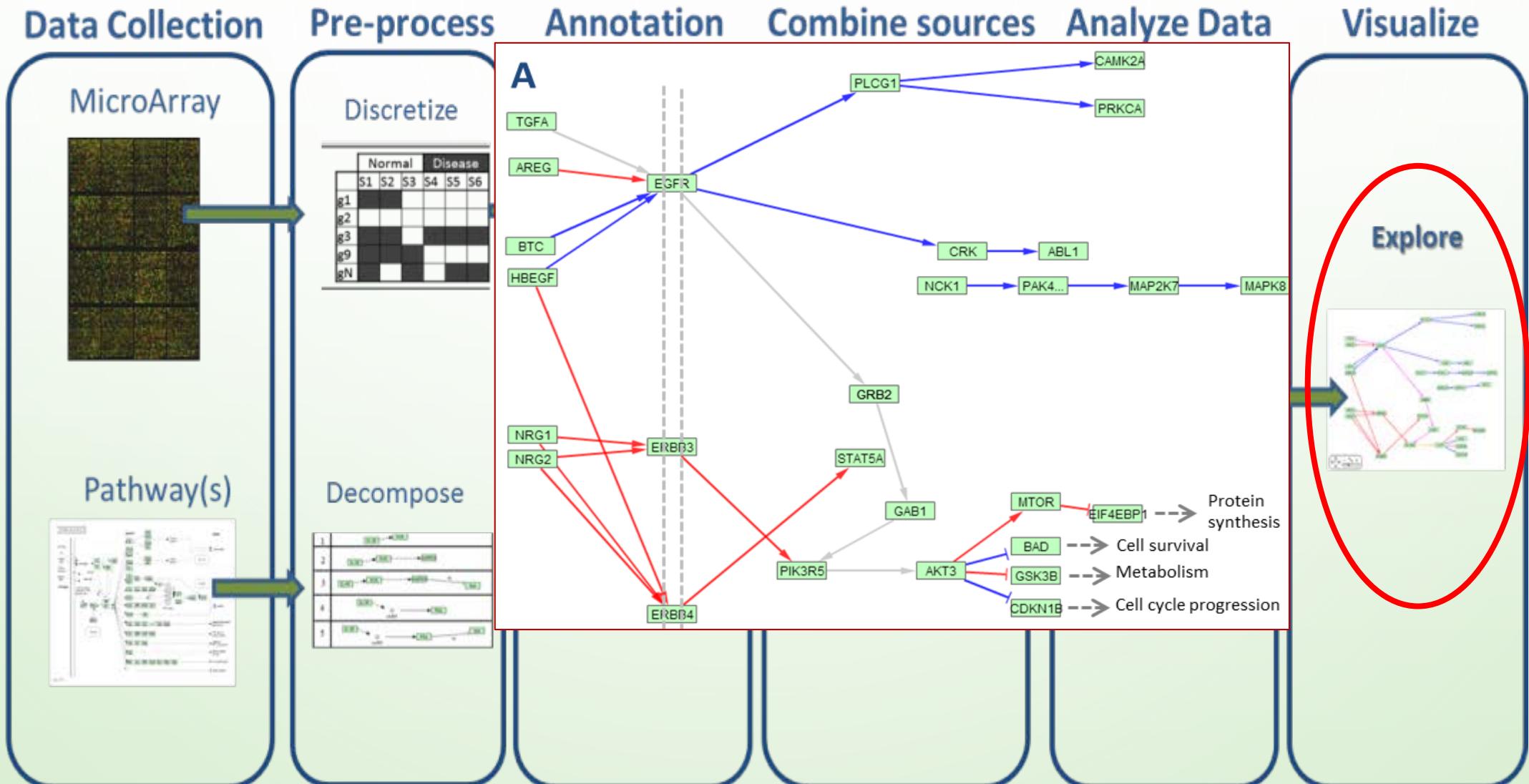
	Sample 1	Sample 2	Sample 3	...
Gene A	1	0	1	...
Gene B	1	1	1	...
Gene C	0	1	0	...
...

INTERACTIONS AS LOGICAL OPERATORS

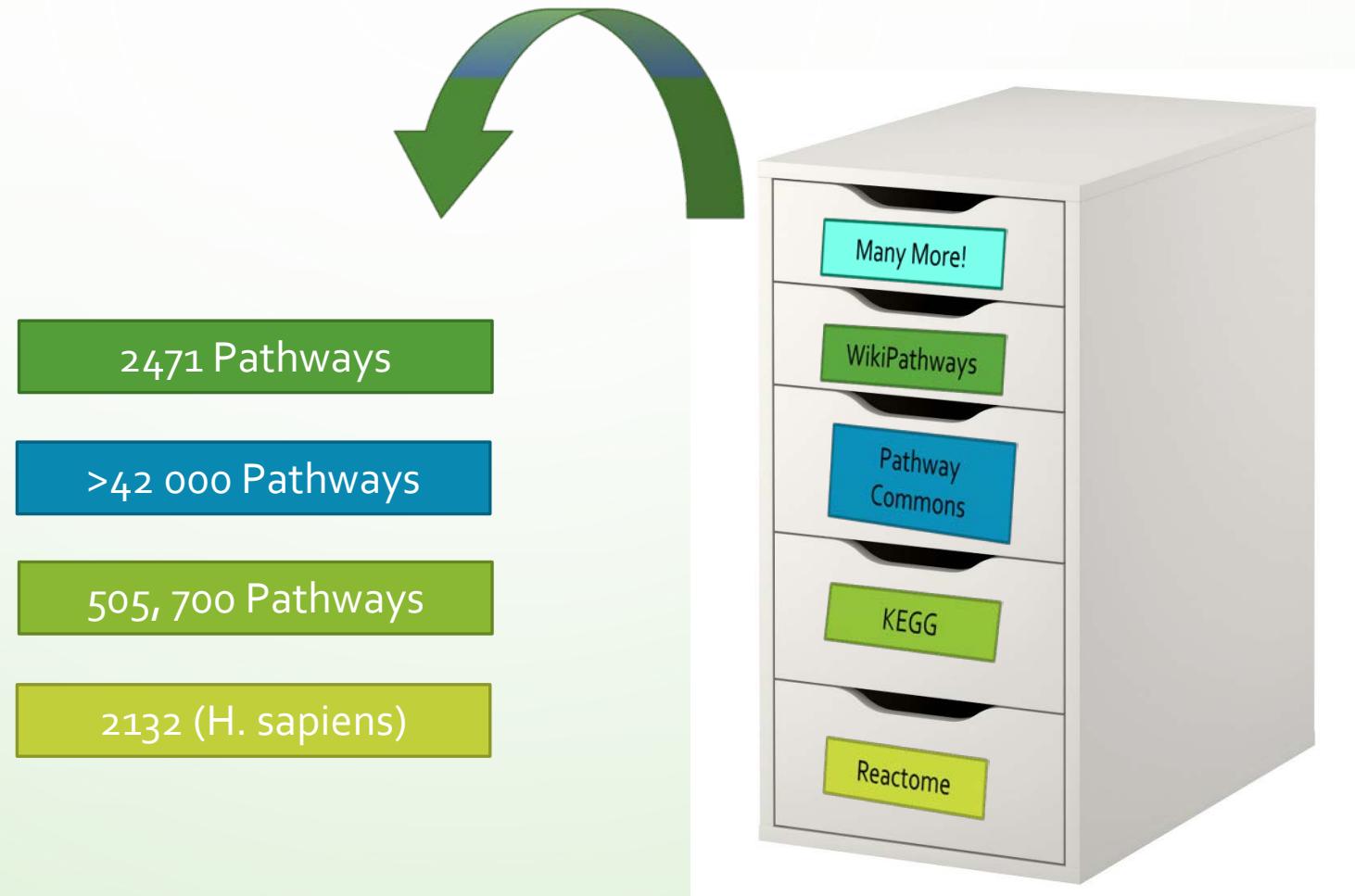
SYSTEM OVERVIEW



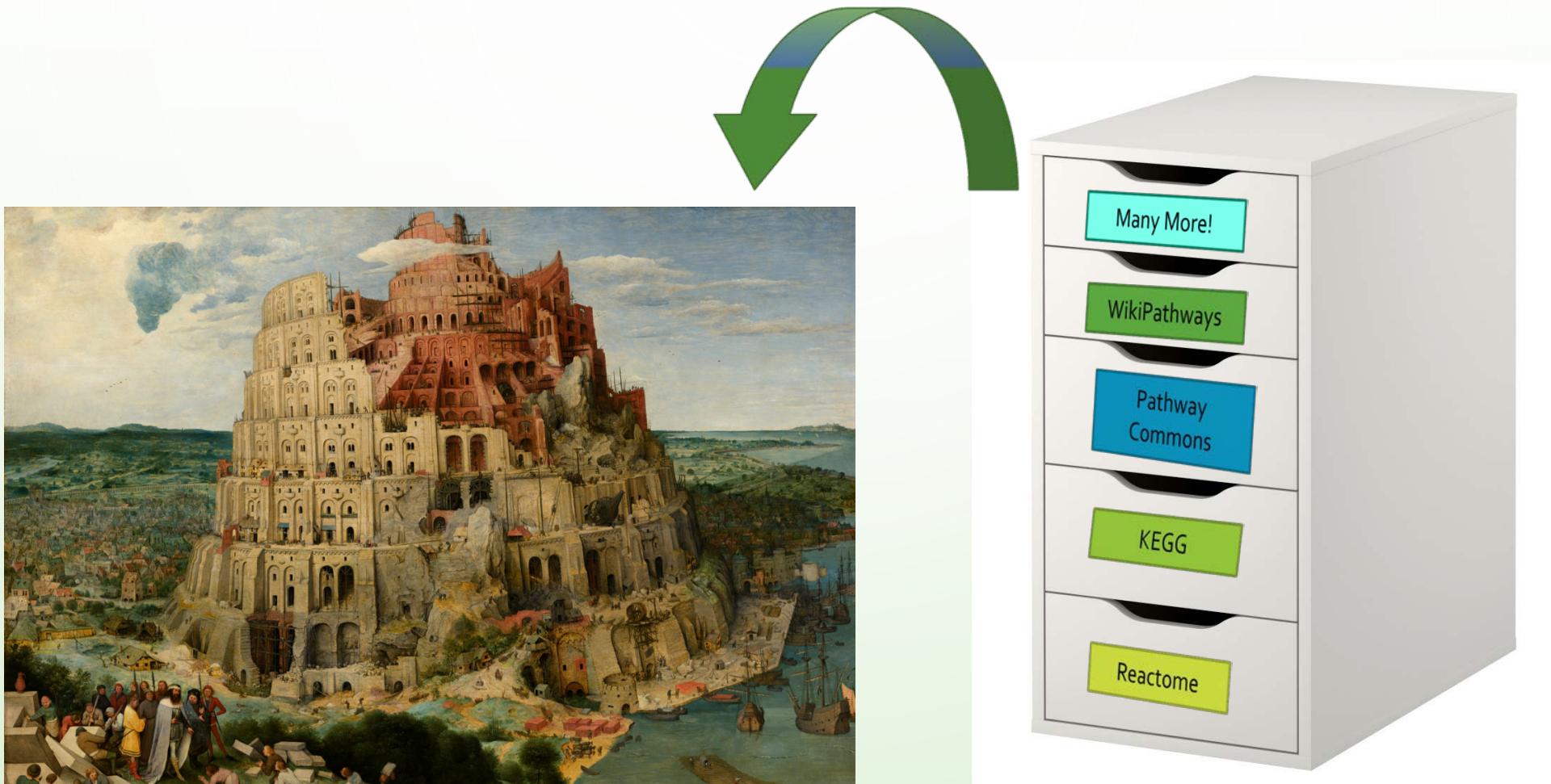
SYSTEM OVERVIEW



Many repositories of biological pathways



Many repositories of biological pathways



Pieter Brueghel the Elder, *The Tower of Babel*, c. 1563, Kunsthistorisches Museum, Vienna

Many repositories of biological pathways



<http://vcell.org/biopax/sbpax.html>

2471 Pathways

>42 000 Pathways

505, 700 Pathways

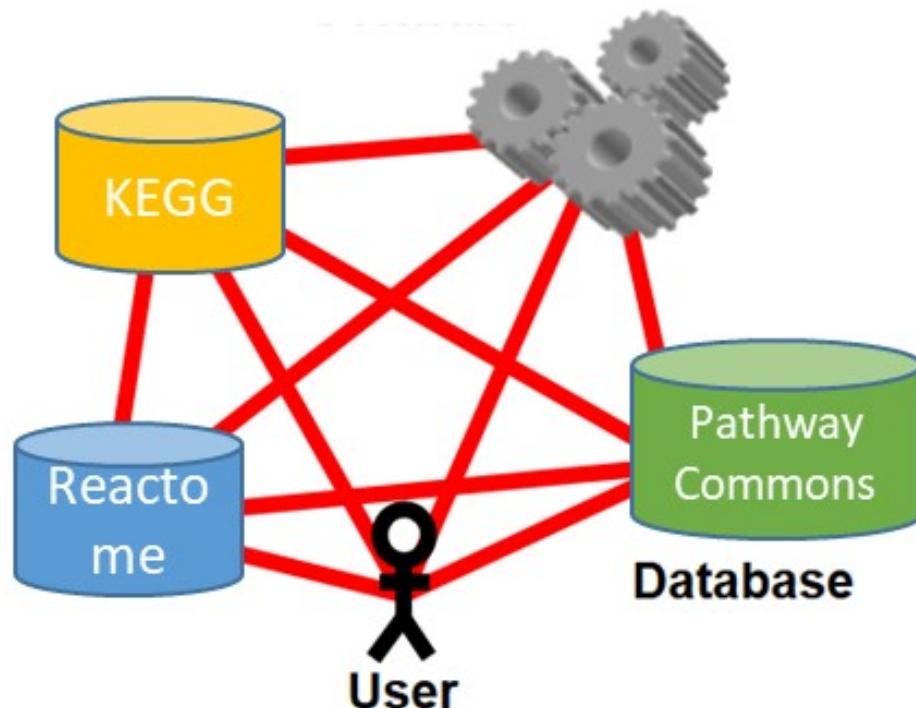
2132 (H. sapiens)



<https://omnipathdb.org/>

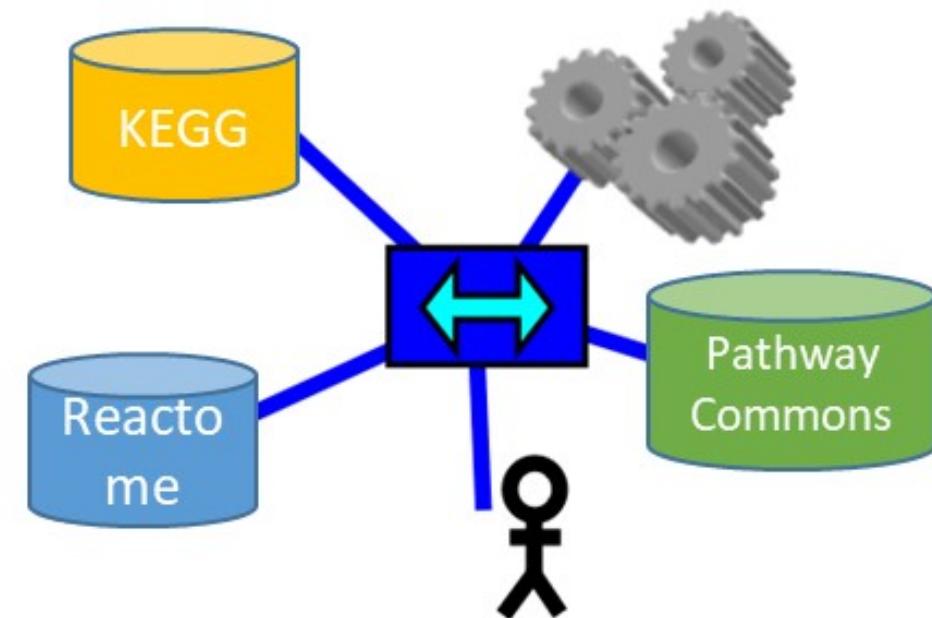


Biological Pathway Exchange (BioPAX)



Before BioPAX

>100 DBs and tools
Tower of Babel



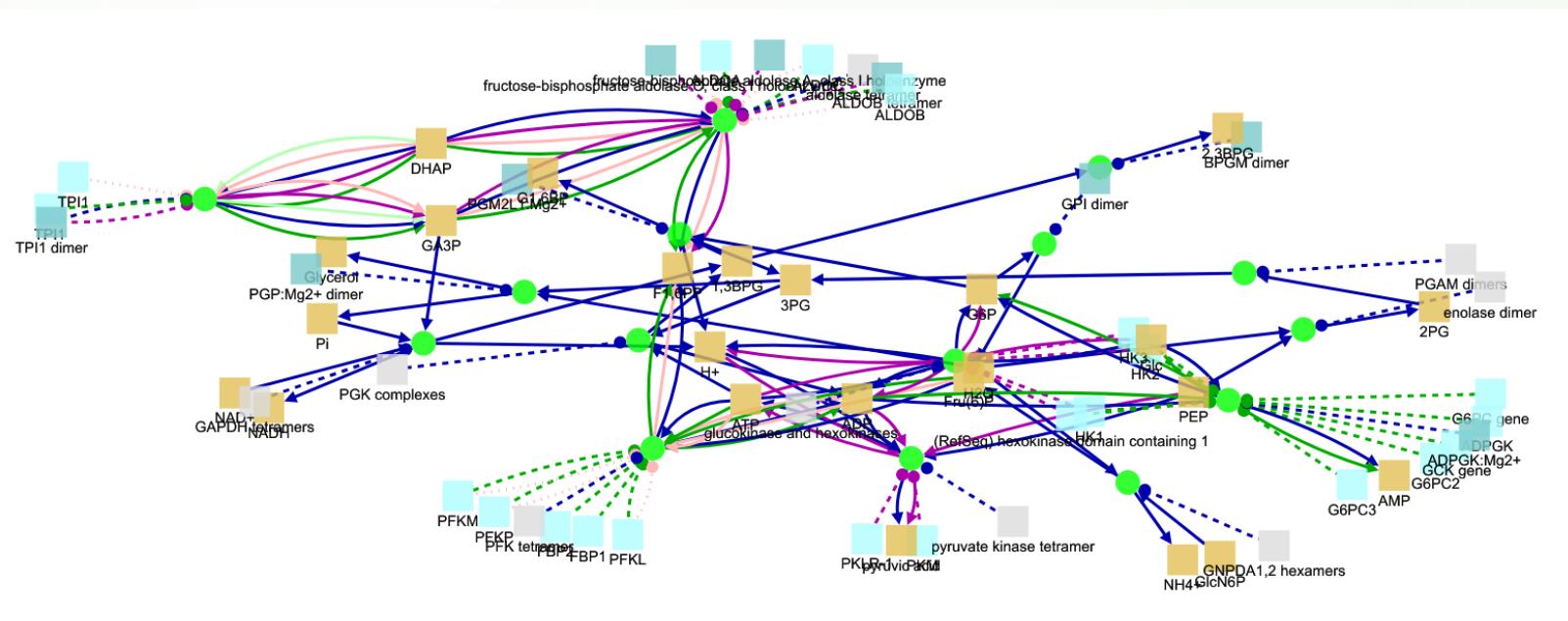
After BioPAX
Unifying language

CONSENSUSPATH DATABASE

Integrated databases:

	name	protein interactions	signalling reactions	metabolic reactions	gene regulations	genetic interactions	drug-target interactions	biochemical pathways
	BIND	✓	✗	✗	✓	✗	✗	✗
	BioCarta	✗	✓	✗	✓	✗	✗	✓
	BioGRID	✓	✗	✗	✗	✓	✗	✗
	mips	✓	✗	✗	✗	✗	✗	✗
	ChEMBL	✗	✗	✗	✗	✗	✓	✗
	DIP	✓	✗	✗	✗	✗	✗	✗
	EHMN	✗	✗	✓	✗	✗	✓	✗
	HPRD	✓	✗	✗	✗	✗	✗	✗
	HumanCyc	✗	✗	✓	✗	✗	✓	✓
	INOH	✗	✓	✓	✗	✗	✗	✓
	InnateDB	✓	✓	✗	✓	✗	✗	✗
	IntAct	✓	✗	✗	✗	✗	✗	✗
	KEGG	✗	✓	✓	✗	✗	✓	✓
	MINT	✓	✗	✗	✗	✗	✓	✗
	mips	✓	✗	✗	✗	✗	✗	✗
	MatrixDB	✓	✗	✗	✗	✗	✗	✗
	NetPath	✓	✓	✗	✗	✗	✓	✓
	PDB	✓	✗	✗	✗	✗	✗	✗
	PDZBase	✓	✗	✗	✗	✗	✗	✗
	PID	✗	✓	✗	✓	✗	✓	✓
	PIG	✓	✗	✗	✗	✗	✗	✗
	PiNdb	✓	✗	✗	✗	✗	✓	✗
	PharmGKB	✗	✗	✗	✗	✗	✓	✓
	PhosphoPOINT	✓	✓	✗	✗	✗	✗	✗
	PhosphoSitePlus	✗	✓	✗	✗	✗	✗	✗
	Reactome	✓	✓	✓	✗	✗	✓	✓
	aMPDB	✗	✗	✗	✗	✗	✗	✓
	SignalLink	✗	✗	✗	✗	✗	✓	✓
	SPIKE	✓	✓	✗	✓	✗	✗	✗
	TTD	✗	✗	✗	✗	✗	✓	✗
	Wikipathways	✓	✓	✓	✗	✗	✓	✓
	ConsensusPathDB	✓	✓	✓	✓	✓	✓	✓

[HTTP://CPDB.MOLGEN.MPG.DE/](http://CPDB.MOLGEN.MPG.DE/)

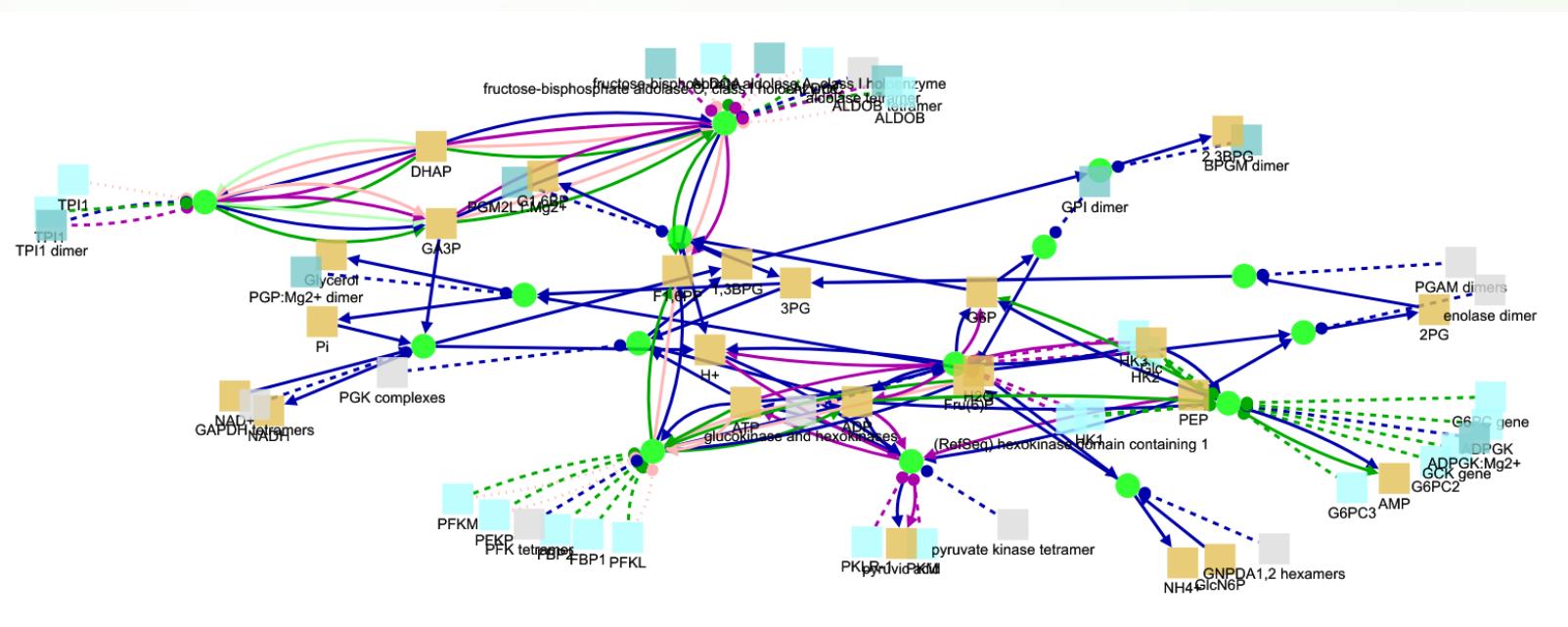


CONSENSUSPATH DATABASE

Integrated databases:

	name	protein interactions	signalling reactions	metabolic reactions	gene regulations	genetic interactions	drug-target interactions	biochemical pathways
	BIND	✓	✗	✗	✓	✗	✗	✗
	BioCarta	✗	✓	✗	✓	✗	✗	✓
	BioGRID	✓	✗	✗	✗	✓	✗	✗
	mips	✓	✗	✗	✗	✗	✗	✗
	ChEMBL	✗	✗	✗	✗	✗	✓	✗
	DIP	✓	✗	✗	✗	✗	✗	✗
	EHMN	✗	✗	✓	✗	✗	✓	✗
	HPRD	✓	✗	✗	✗	✗	✗	✗
	HumanCyc	✗	✗	✓	✗	✗	✓	✓
	INOH	✗	✓	✓	✗	✗	✗	✓
	InnateDB	✓	✓	✗	✓	✗	✗	✗
	IntAct	✓	✗	✗	✗	✗	✗	✗
	KEGG	✗	✓	✓	✗	✗	✓	✓
	MINT	✓	✗	✗	✗	✗	✓	✗
	MIPS	✓	✗	✗	✗	✗	✗	✗
	MatrixDB	✓	✗	✗	✗	✗	✗	✗
	NetPath	✓	✓	✗	✗	✗	✓	✓
	PDB	✓	✗	✗	✗	✗	✗	✗
	PDZBase	✓	✗	✗	✗	✗	✗	✗
	PID	✗	✓	✗	✓	✗	✓	✓
	PIG	✓	✗	✗	✗	✗	✓	✗
	PINdb	✓	✗	✗	✗	✗	✓	✗
	PharmGKB	✗	✗	✗	✗	✗	✓	✓
	PhosphoPOINT	✓	✓	✗	✗	✗	✗	✗
	PhosphoSitePlus	✗	✓	✗	✗	✗	✗	✗
	Reactome	✓	✓	✓	✗	✗	✓	✓
	tMPDB	✗	✗	✗	✗	✗	✗	✓
	SignalLink	✗	✗	✗	✗	✗	✓	✓
	SPIKE	✓	✓	✗	✓	✗	✗	✗
	TTD	✗	✗	✗	✗	✗	✓	✗
	Wikipathways	✓	✓	✓	✗	✗	✓	✓
	ConsensusPathDB	✓	✓	✓	✓	✓	✓	✓

[HTTP://CPDB.MOLGEN.MPG.DE/](http://cpdb.molgen.mpg.de/)



<https://gist.github.com/ngopal/9164149>

MORE LEARNING RESOURCES

- [HTTPS://WWW.EBI.AC.UK/TRAINING/](https://www.ebi.ac.uk/training/)
- [HTTPS://WWW.NCBI.NLM.NIH.GOV/HOME/LEARN/](https://www.ncbi.nlm.nih.gov/home/learn/)