AI LAB

4.1 – LARGE PROJECTS AND DATABASES

FRANCESCA M. BUFFA

LAB STRUCTURE

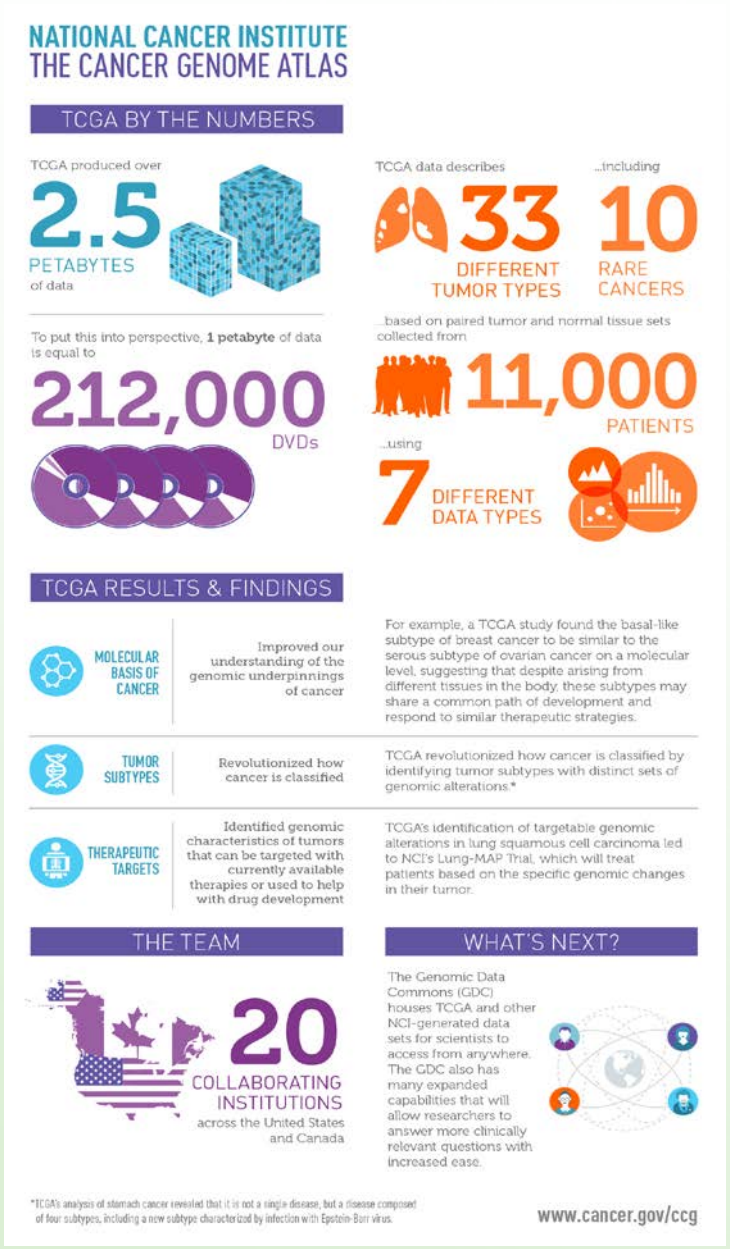
- INTRODUCTION
- THE DATA
- THE AI-LAB CHALLENGE – PART 1
- PART 1 - SHARING AND DISCUSSION

- 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



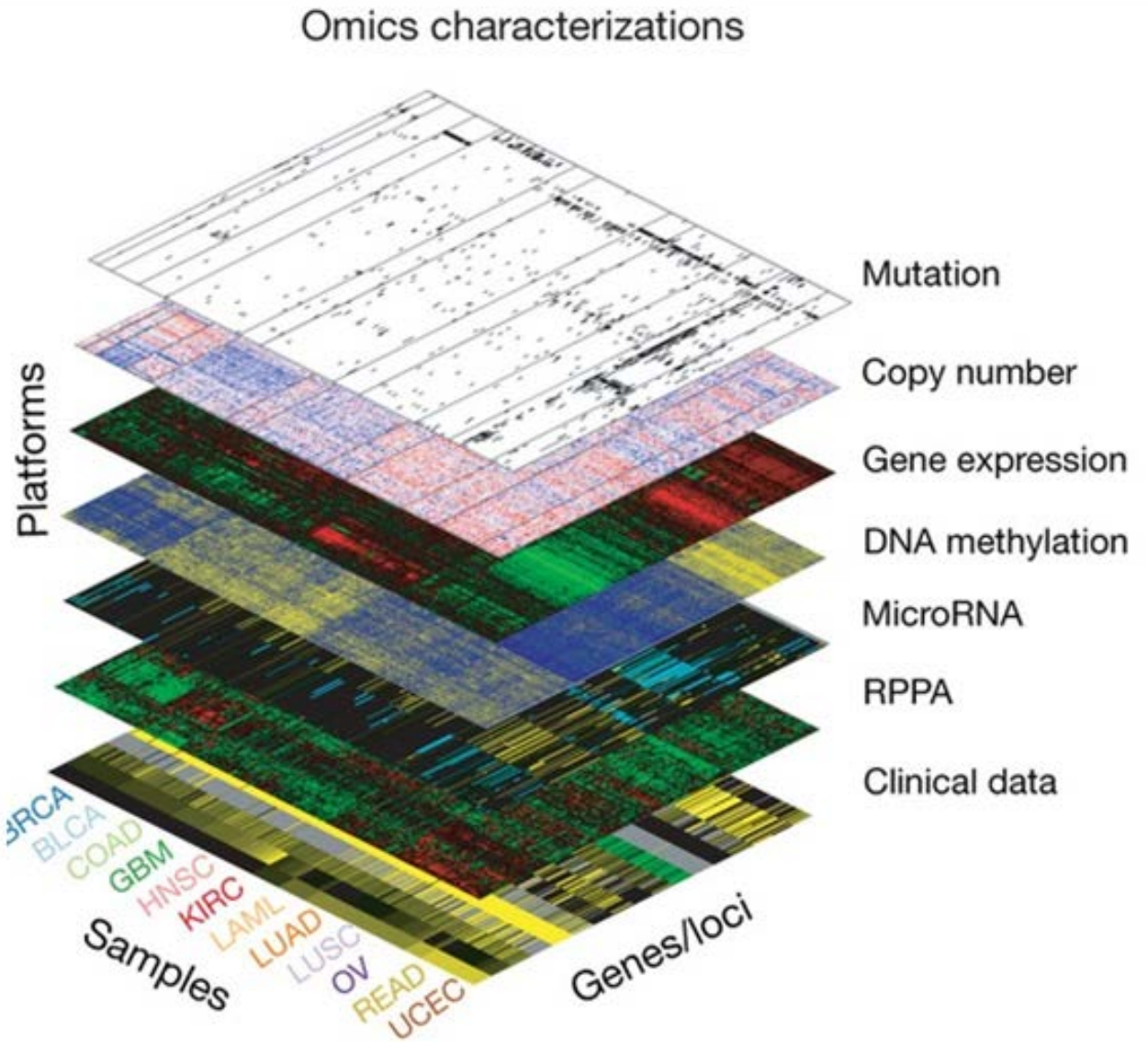


Image source: [TCGA Research Network et al. Nature Genetics 2013]

DATABASES - CELL LINES

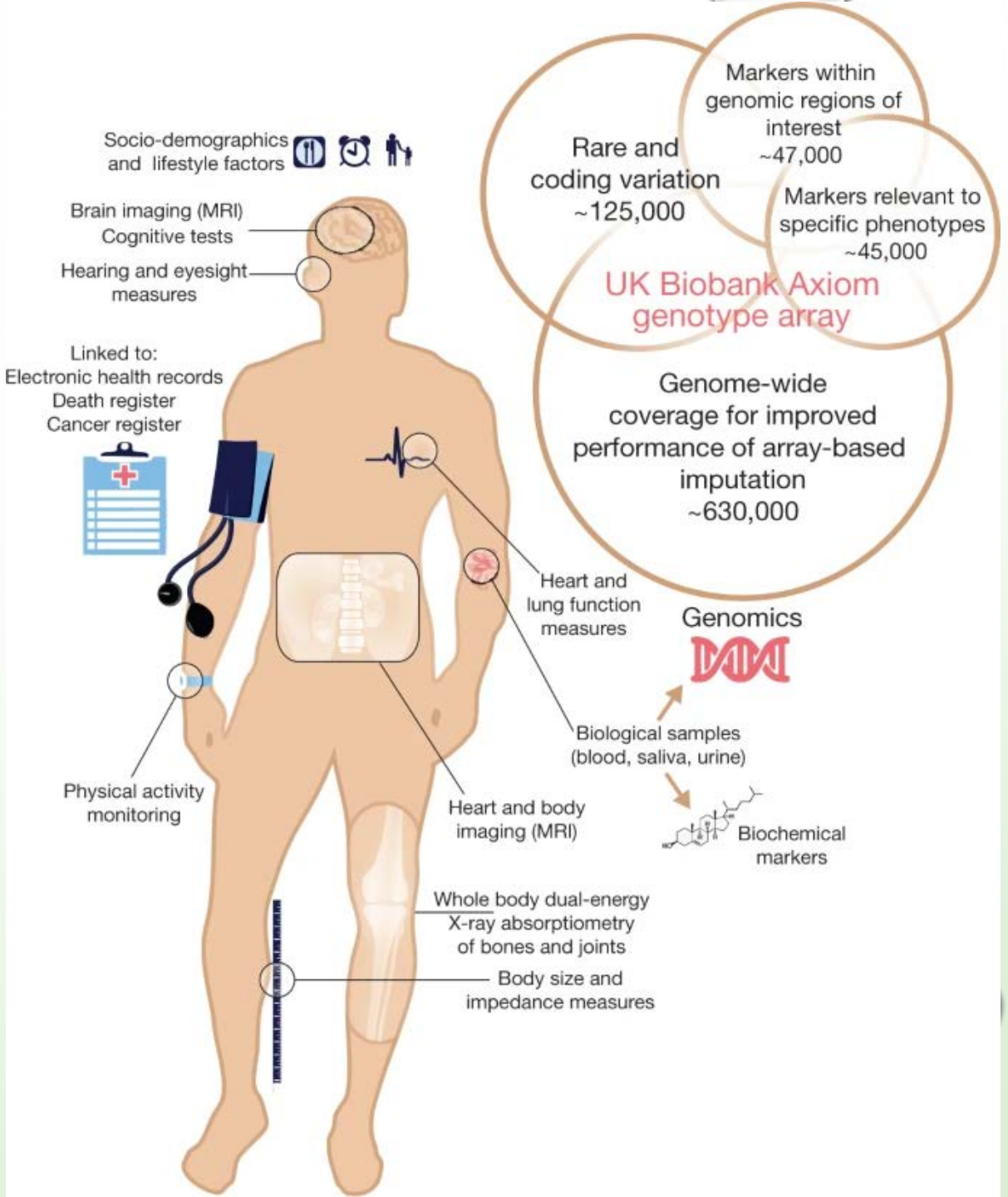
https://sites.broadinstitute.org/ccle/

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

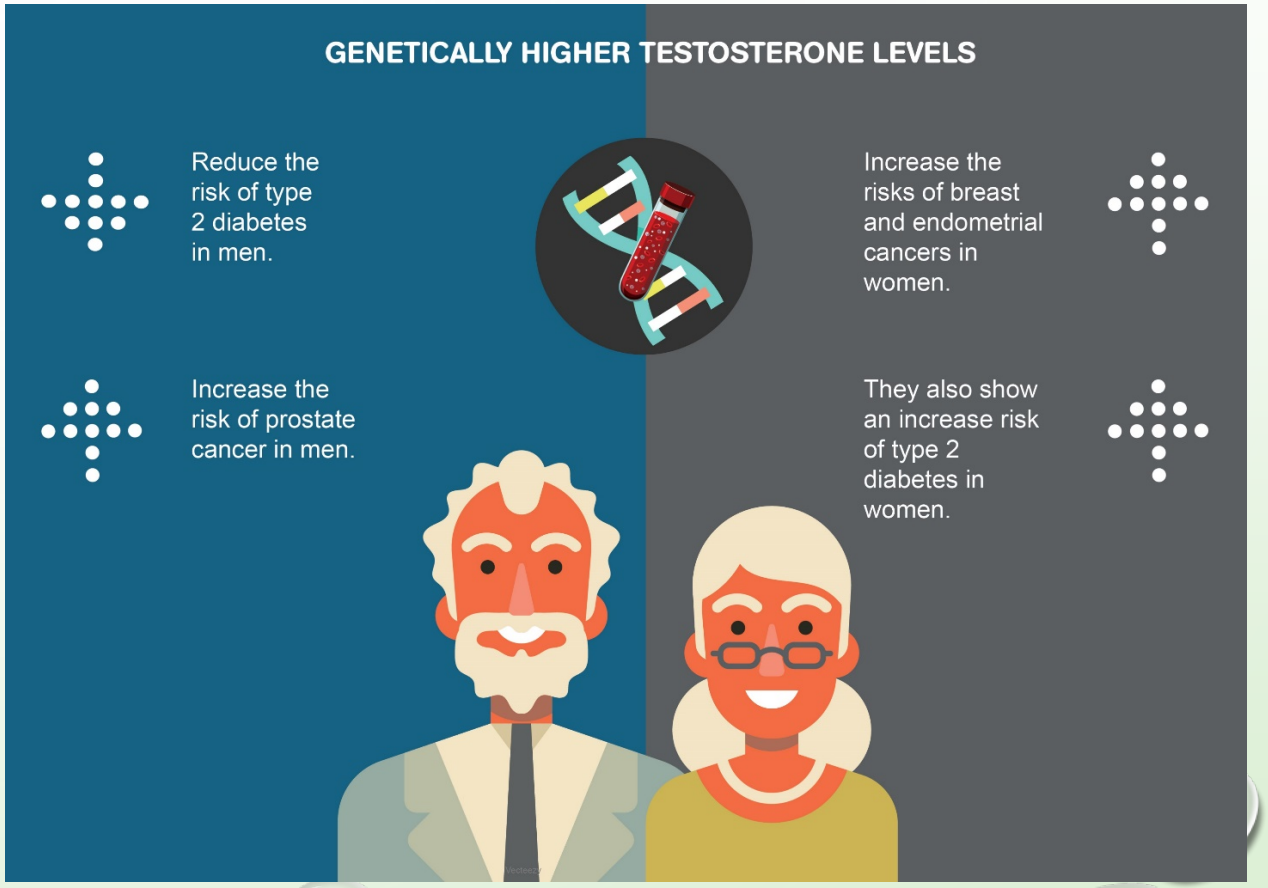
THE UK BIOBANK

- **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

https://www.nature.com/articles/s41586-018-0579-z

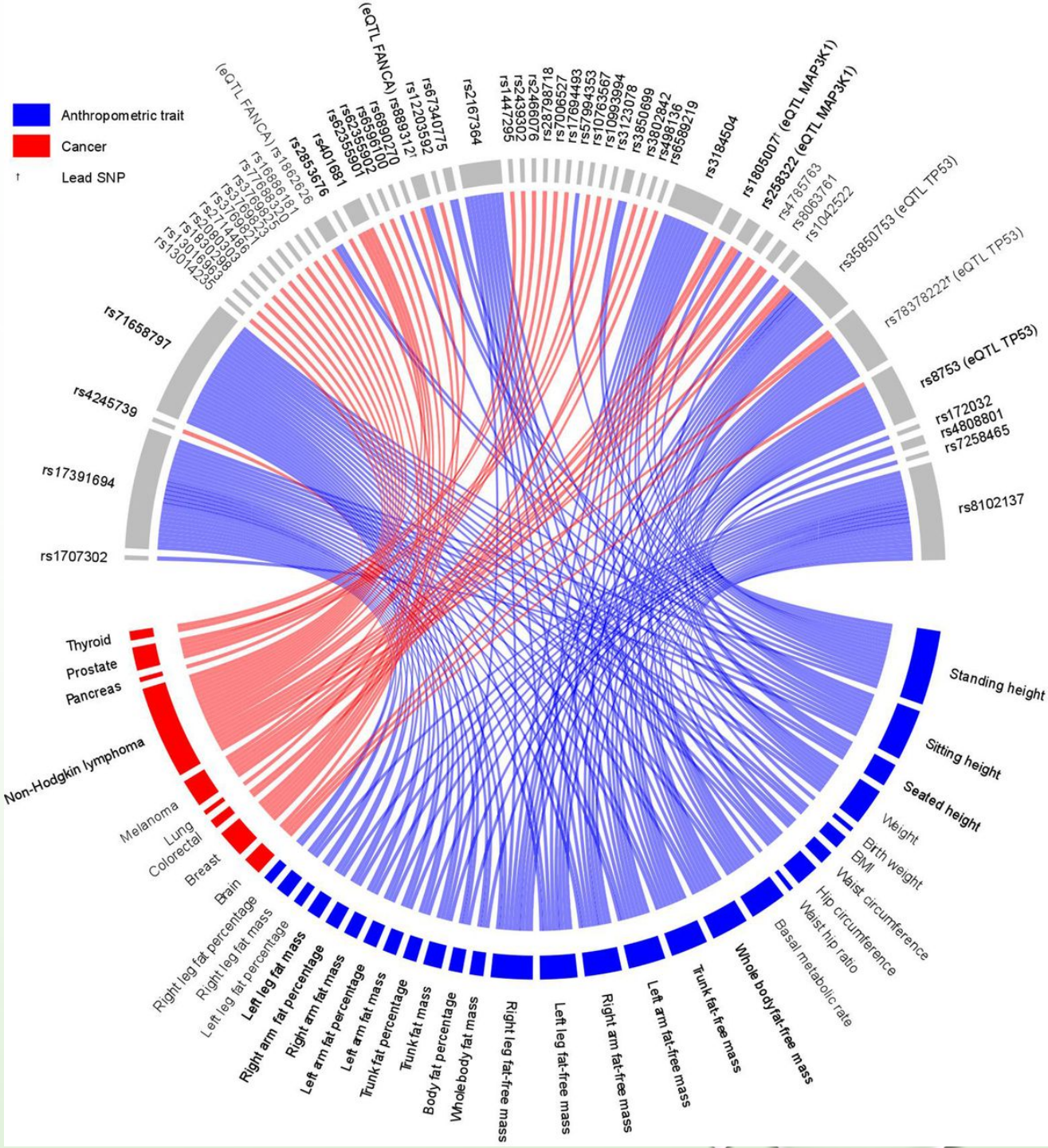


THE UK BIOBANK – EXAMPLES OF STUDIES



https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/our-impact

HERITABLE GENETIC VARIANTS LINK CANCER RISK WITH ANTHROPOMETRIC TRAITS



https://jmg.bmj.com/content/58/6/392

FOLLOW UP FROM UK BIOBANK

“META-DATABASES”

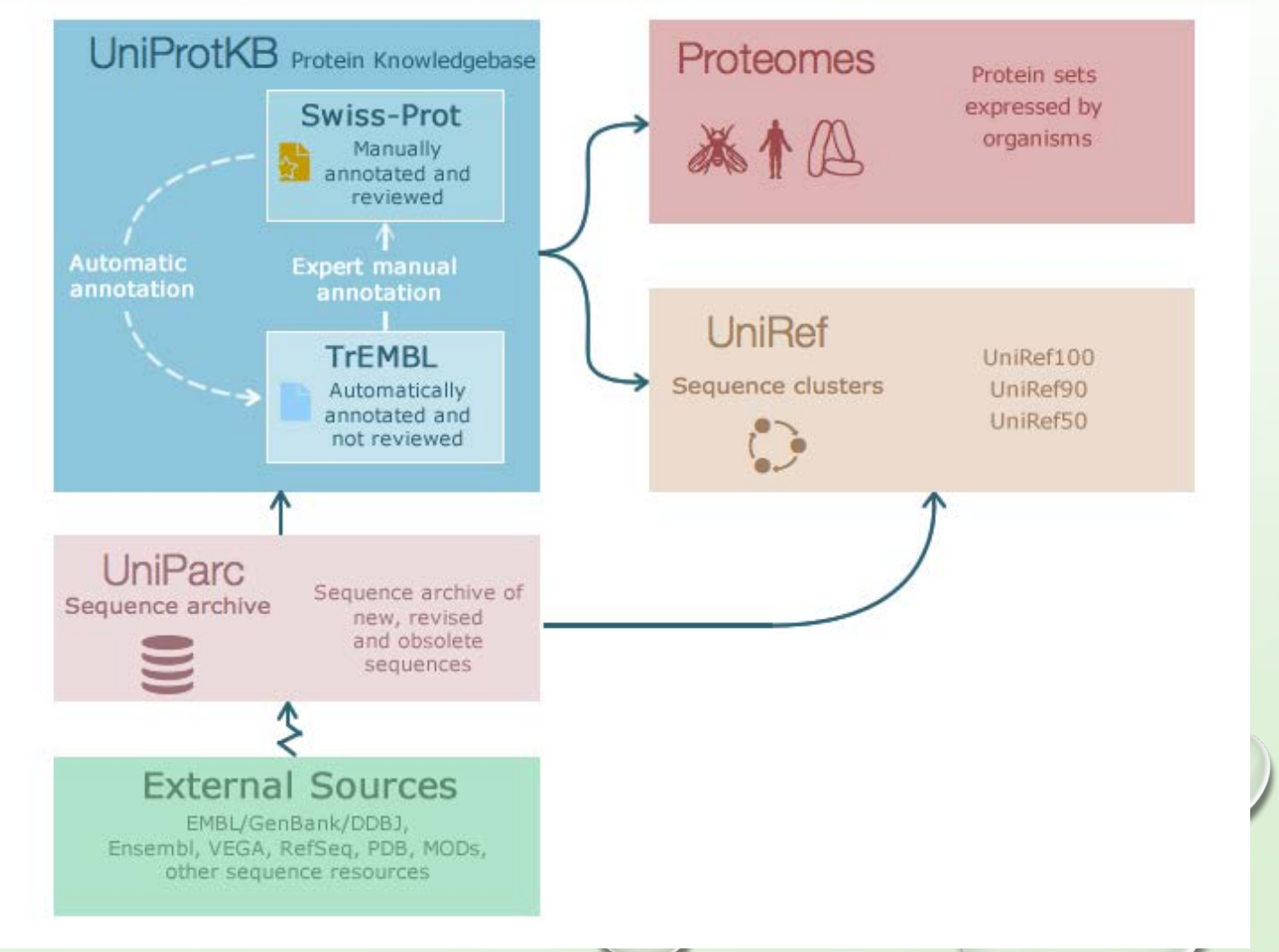
- “DATABASES OF DATABASES”
- COLLECT DATA ABOUT DATA TO GENERATE NEW DATA
- HEALTH NEEDS MULTISCIPLINARITY
- MANY USERS AND DISPARATE SCENARIOS
- NEED TO GATHER DIVERSE INFORMATION (SEARCH ENGINES)
- INFORMATION IS NEEDED FROM DIFFERENT SOURCES
- NEED TO PROCESS DIVERSE INFORMATION
- MAKE IT AVAILABLE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

BIOLOGICAL DATABASES

Libraries for biology/chemistry/medical sciences

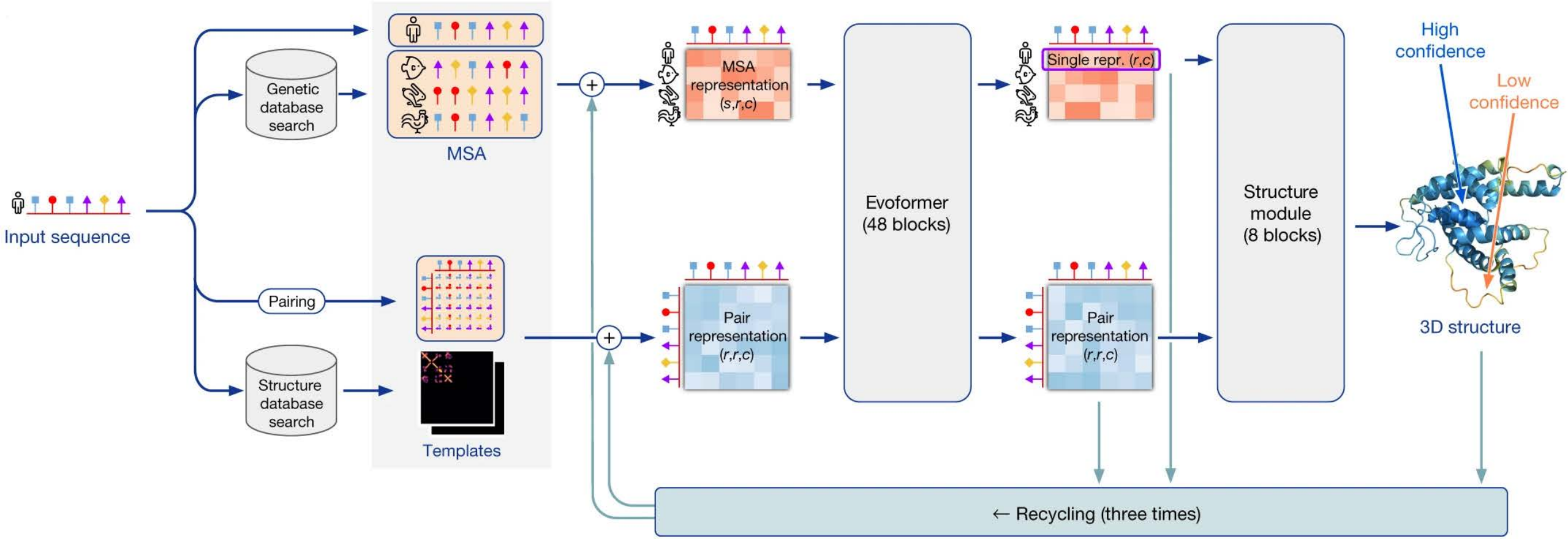
- **Scientific experiments**
- **Published literature**
- **Computational analyses**
- **…**

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

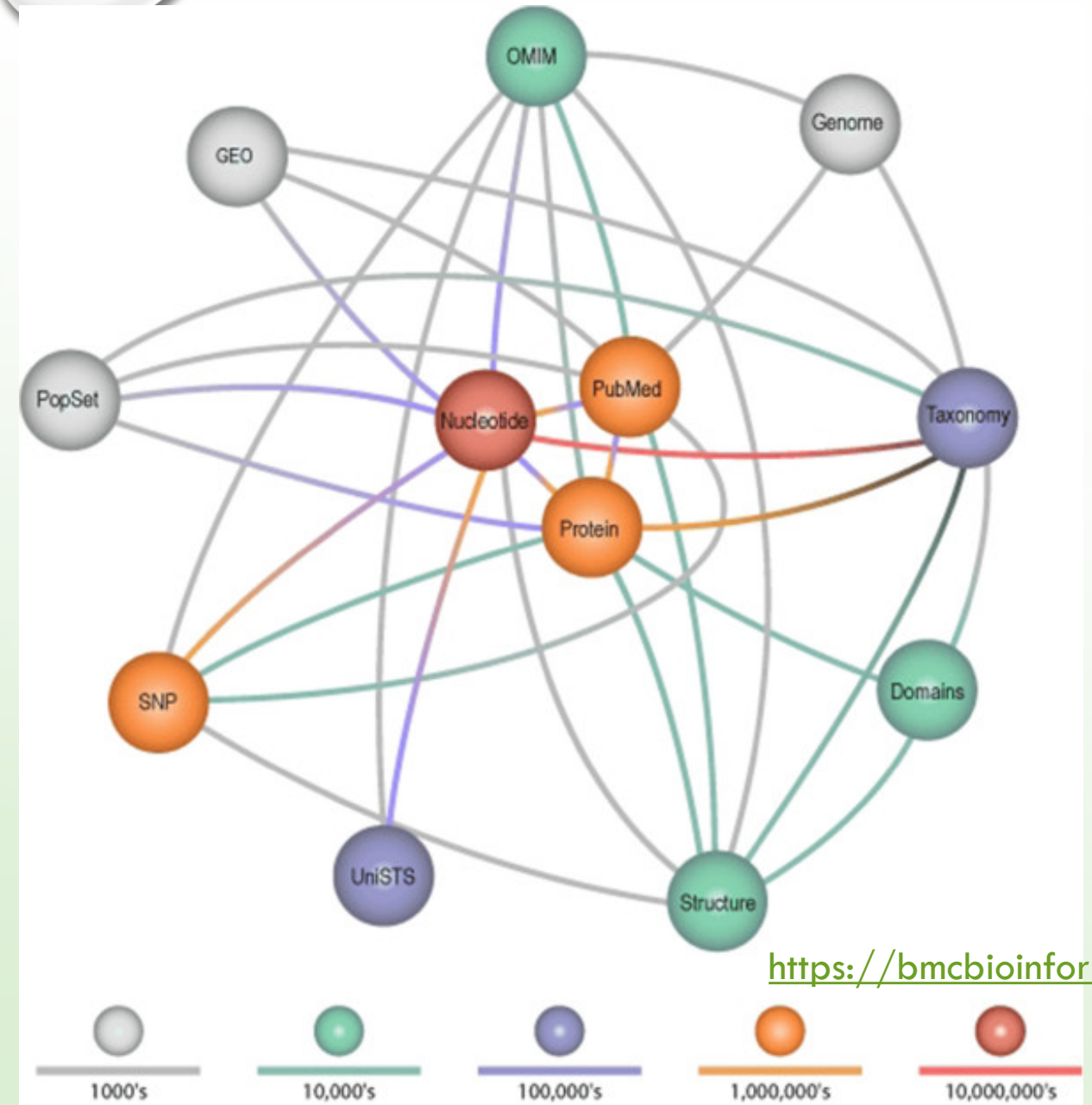
https://www.youtube.com/watch?v=B9PL__gVxLI

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

AlphaFold Protein Structure Database

Developed by DeepMind and EMBL-EBI

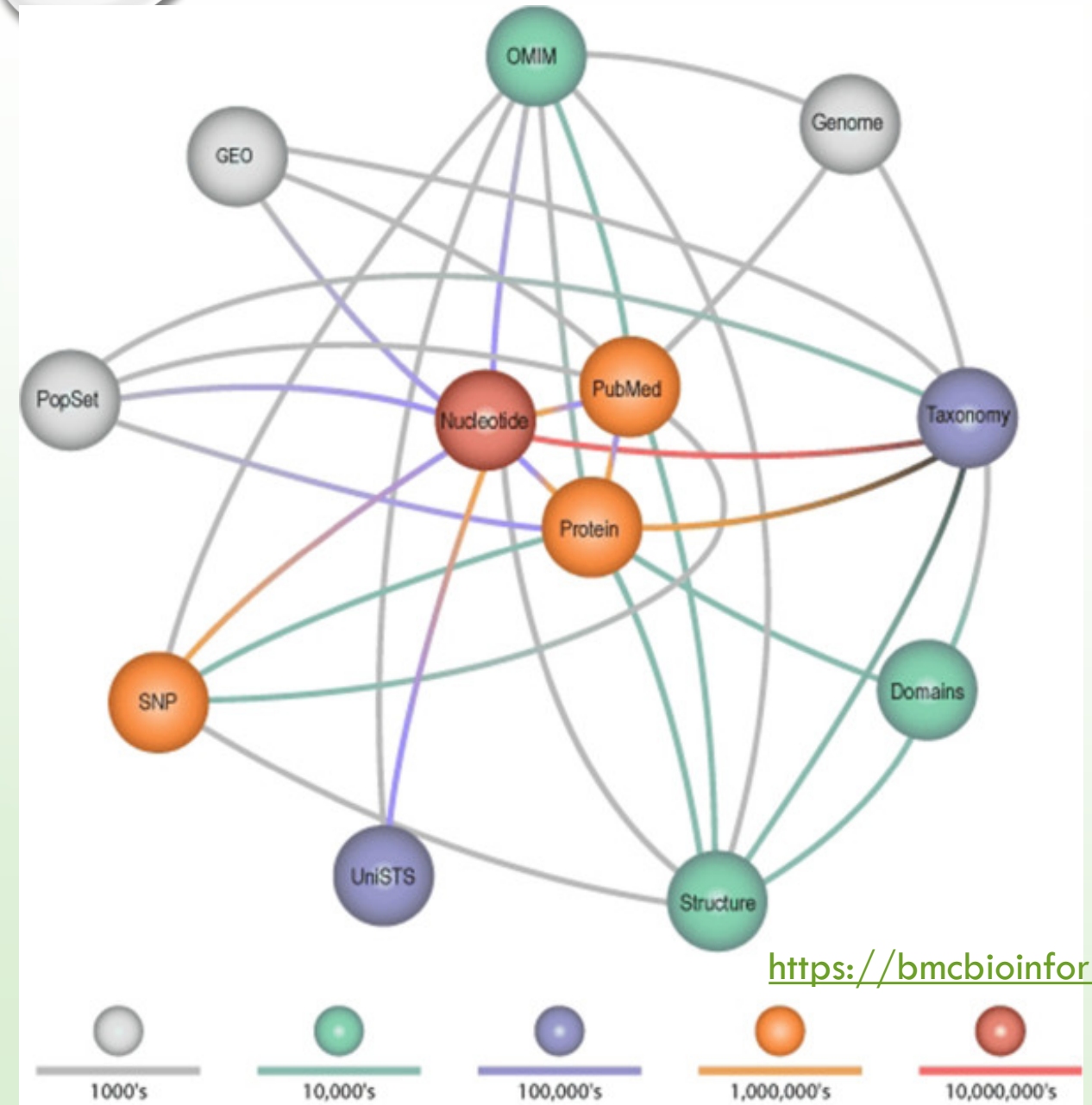
NCBI DATABASE



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

National Center for Biotechnology Information (NCBI)

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.

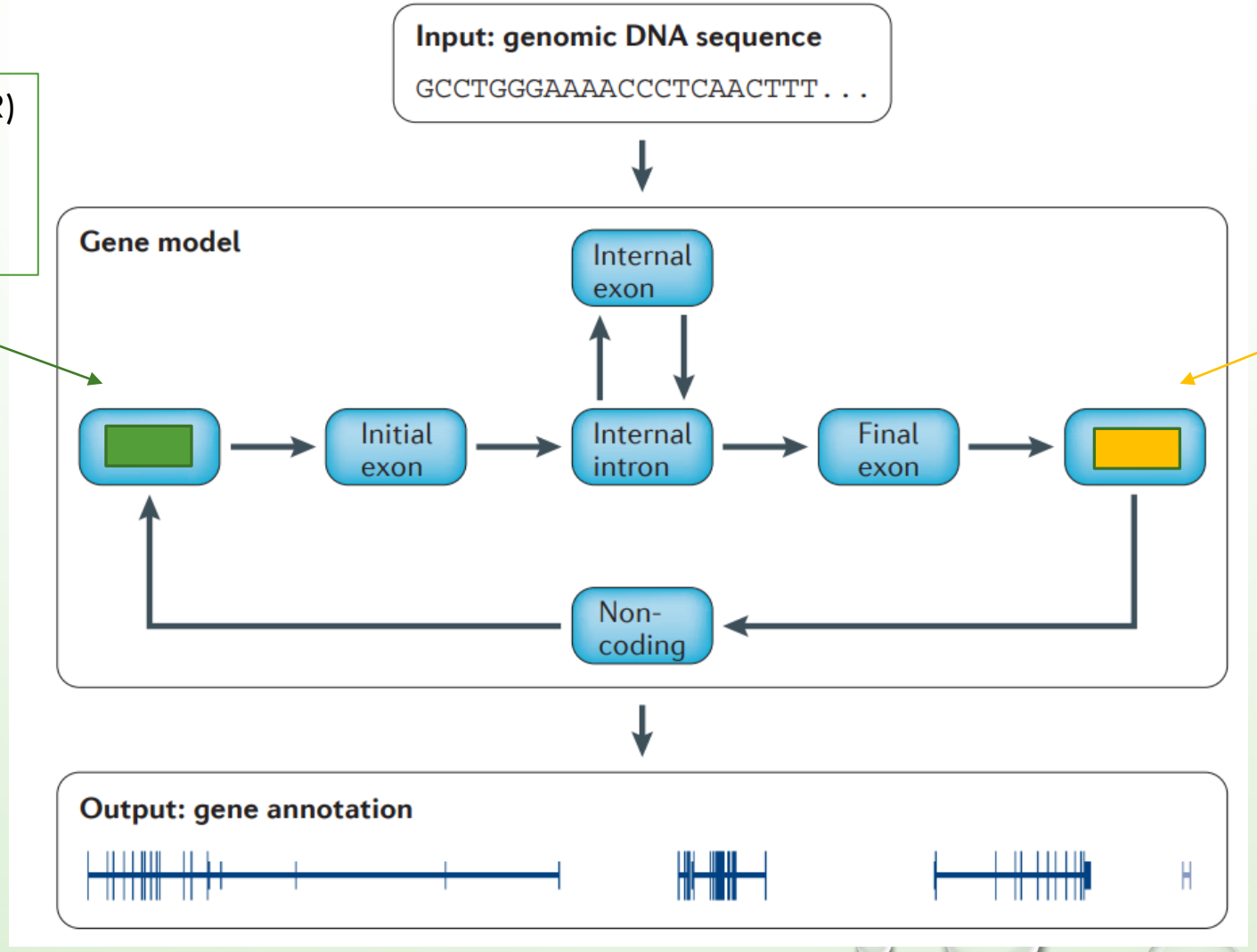
https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-73

National Center for Biotechnology Information (NCBI)

A GENE-FINDING MODEL

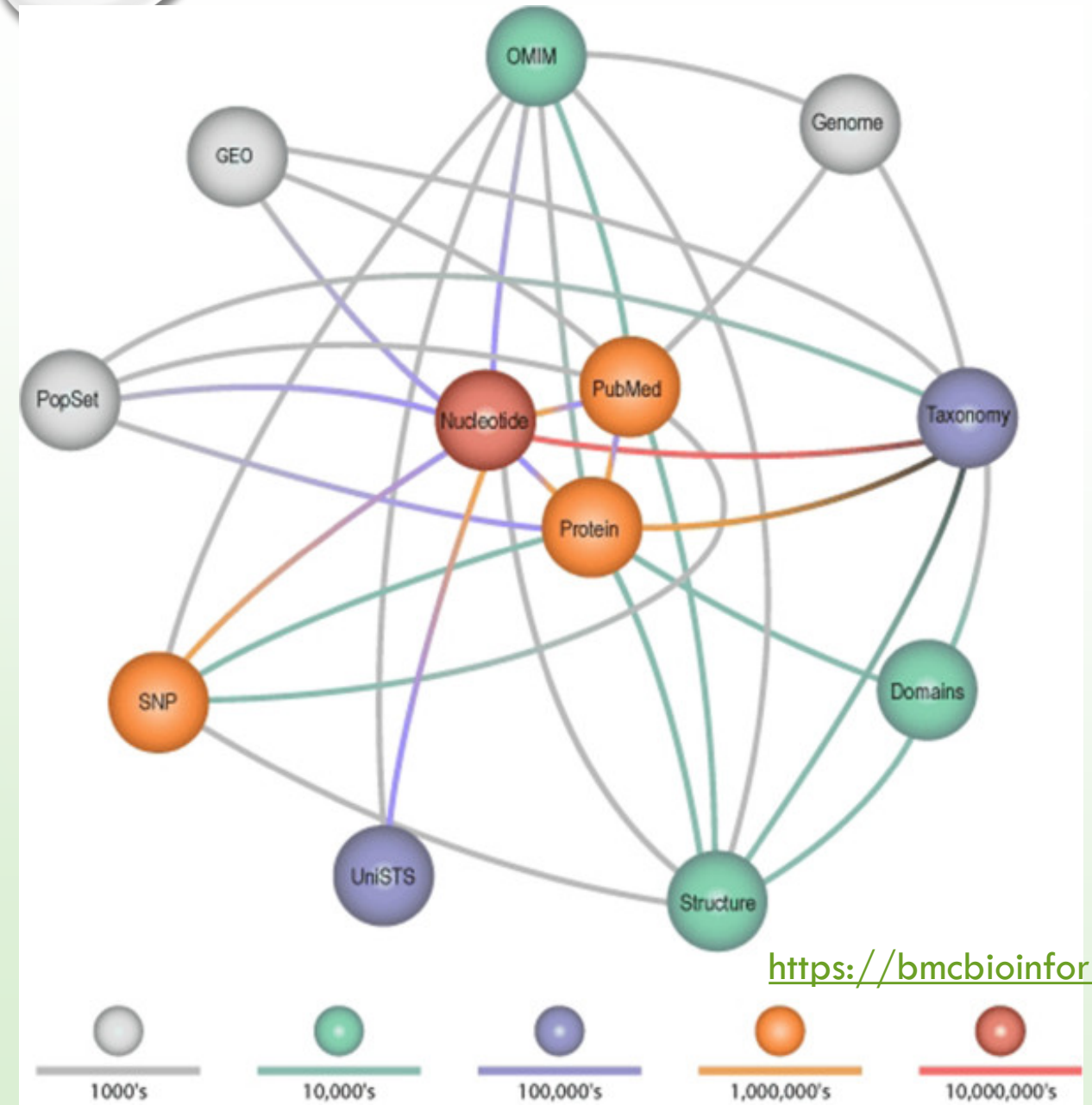
5' untranslated region (UTR) of a messenger RNA (mRNA): directly upstream from the initiation codon

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



https://www.nature.com/articles/nrg3920

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://pypi.org/project/entrezpy/

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

National Center for Biotechnology Information (NCBI)

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

A list or set of genes

Gene sets

List of genes acting in a pathway/determine a phenotypes

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

Molecular Signatures Database v6.2

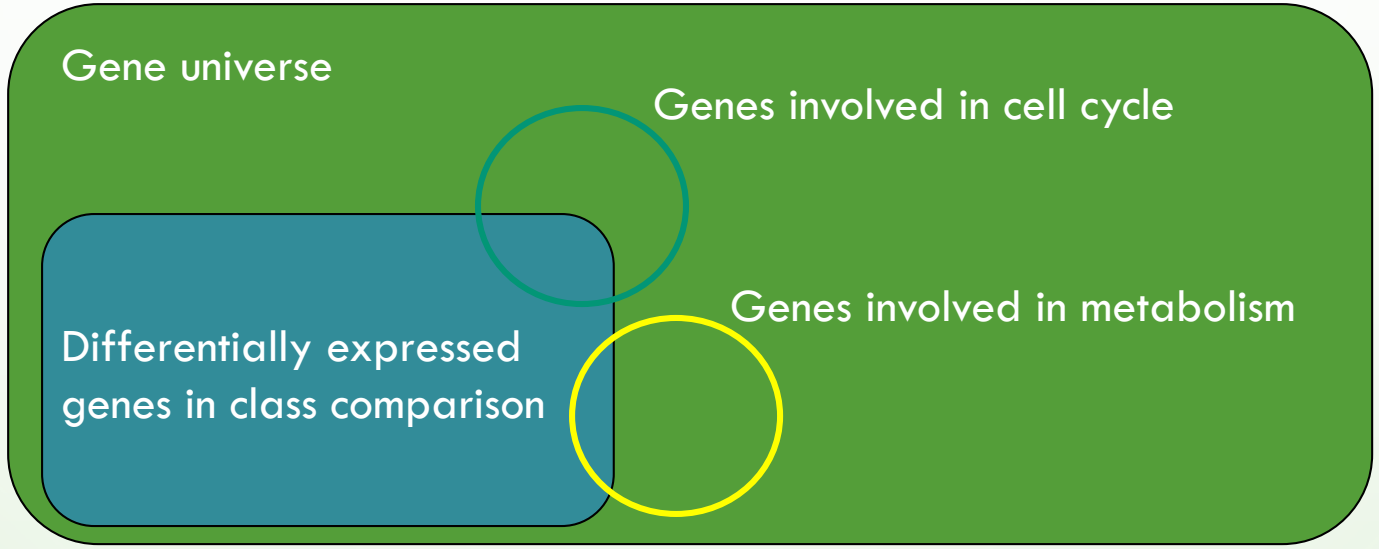
Collections

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?



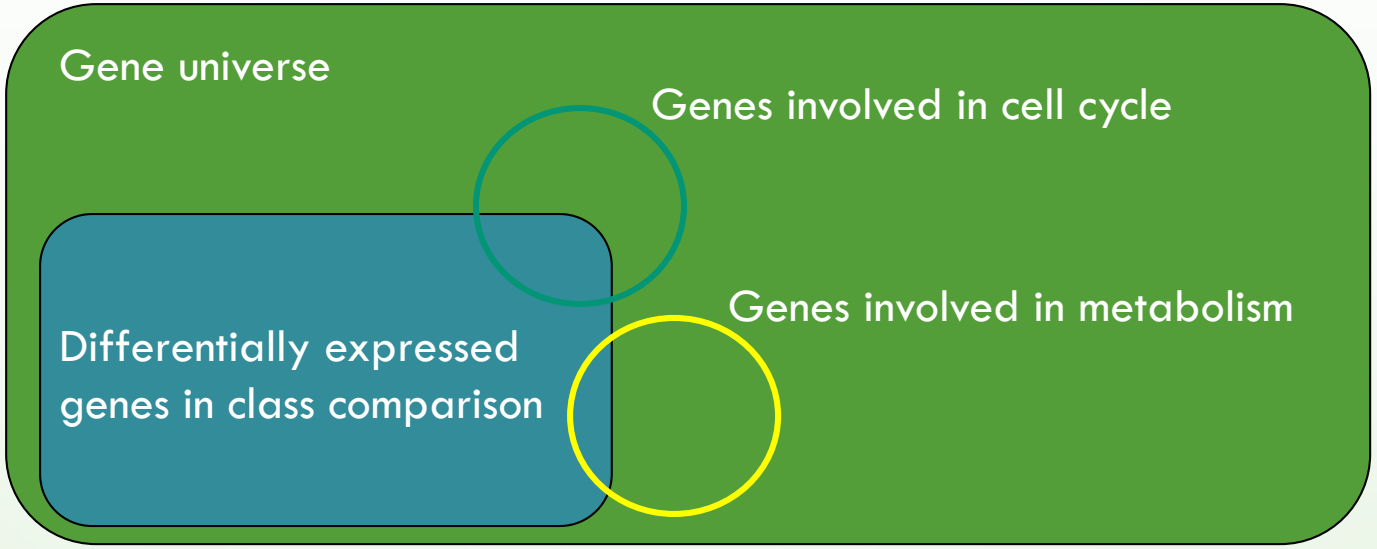
Two-way table:

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

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

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

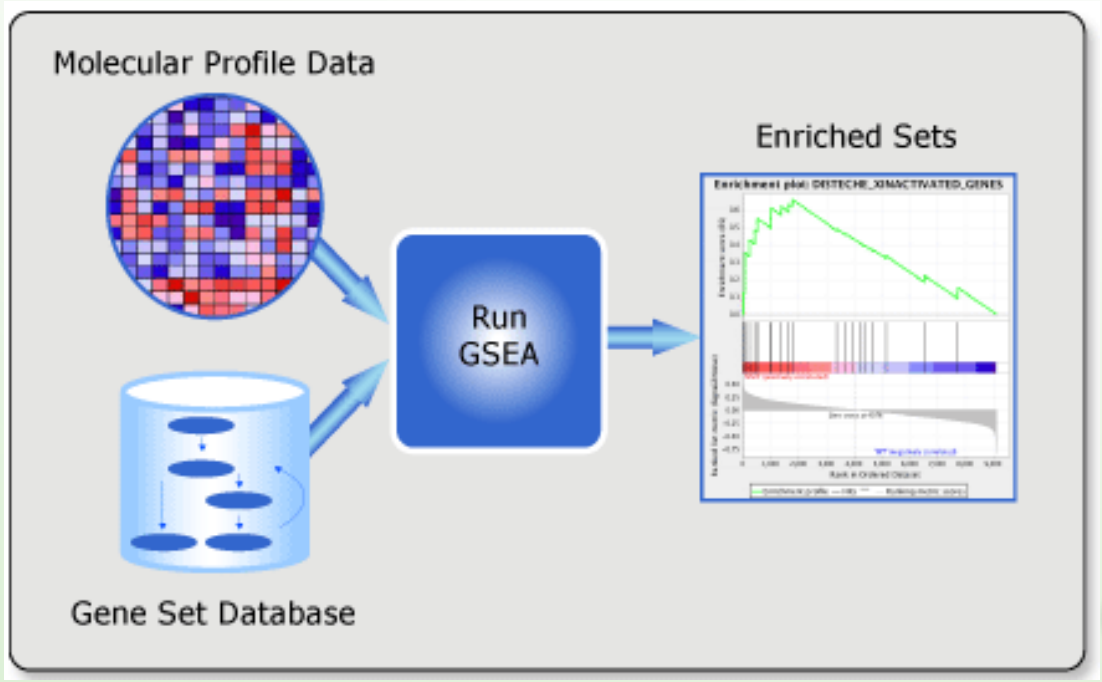
LIMITATIONS

- 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

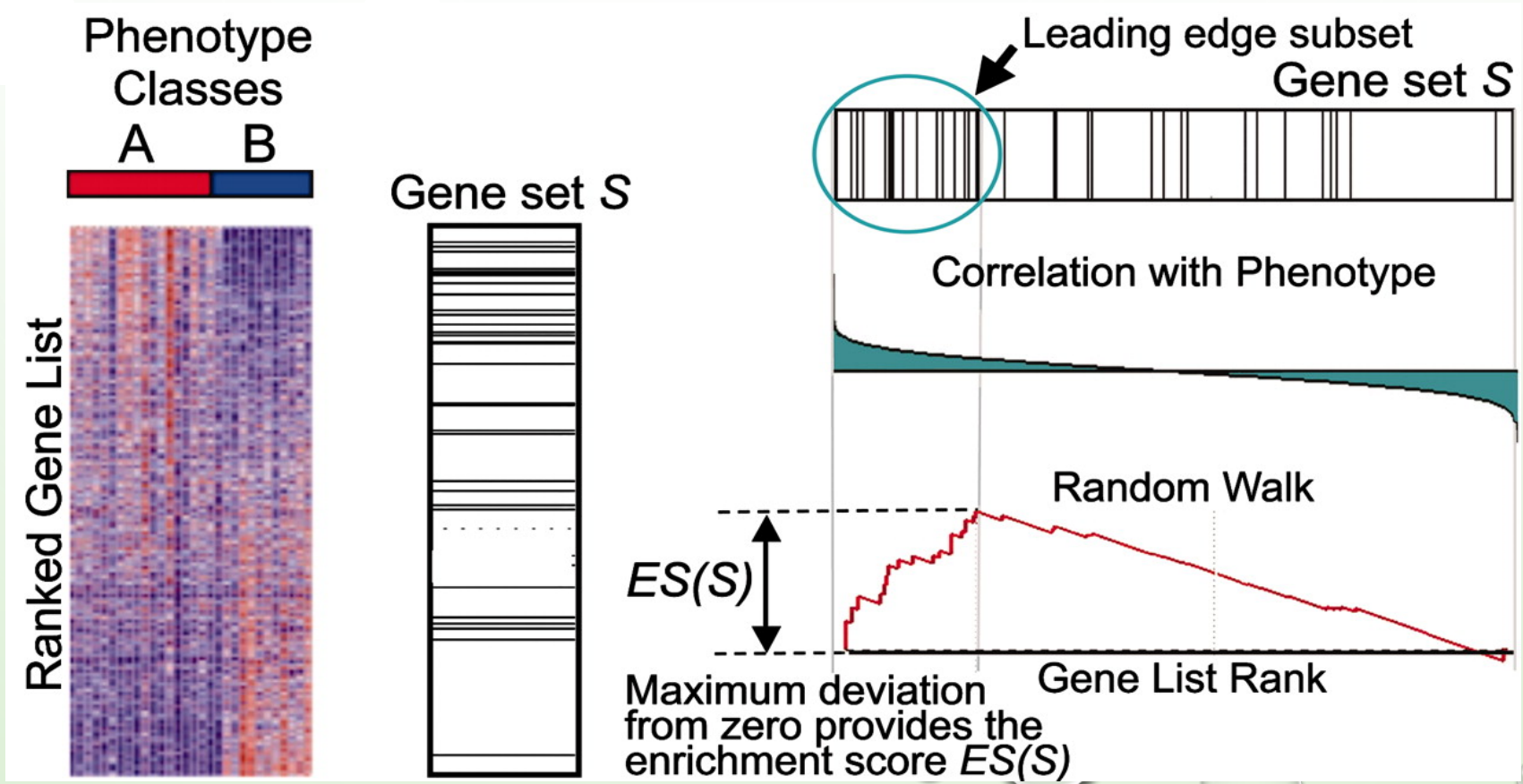
Tests whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes)

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.



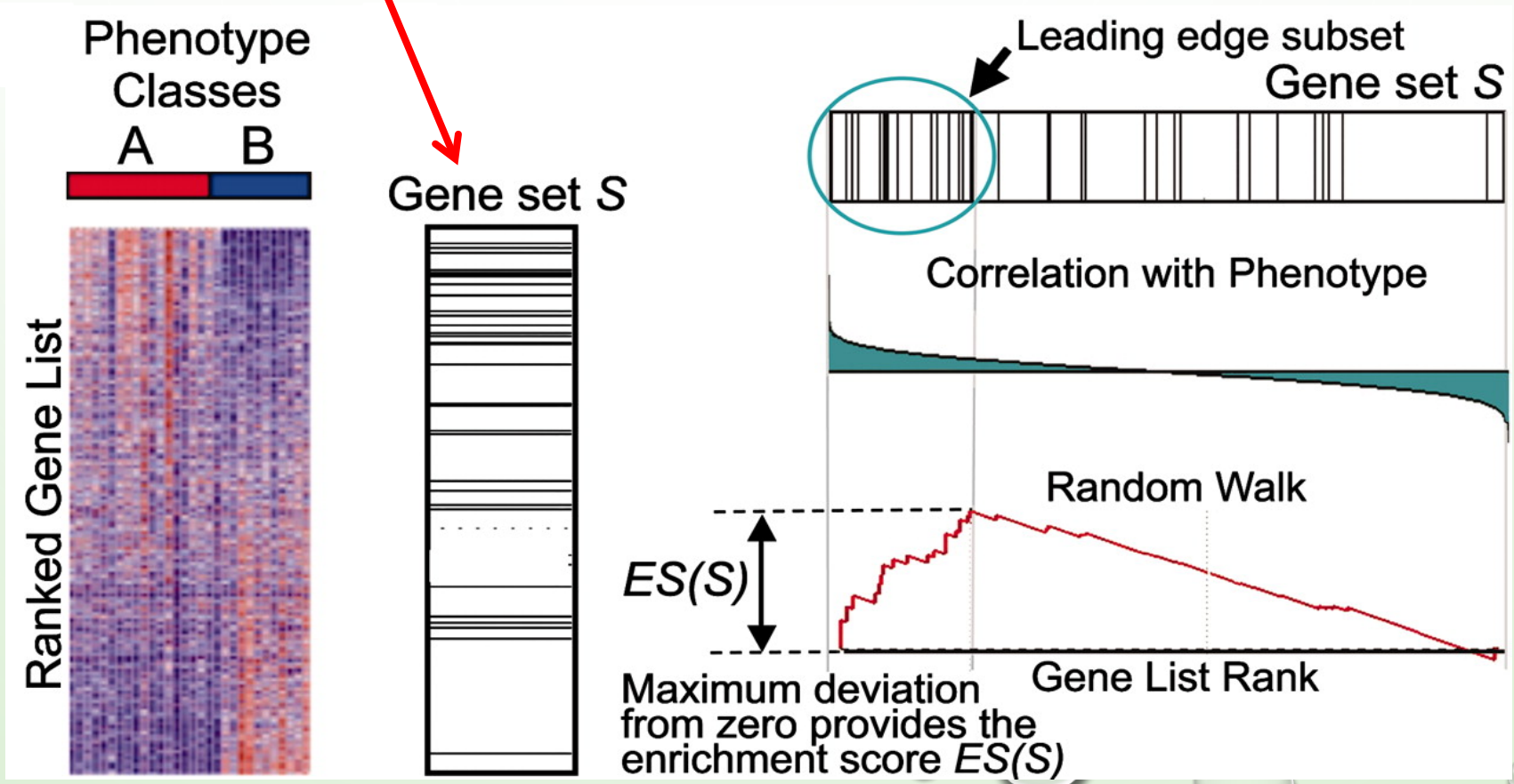
http://www.broadinstitute.org/gsea/

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



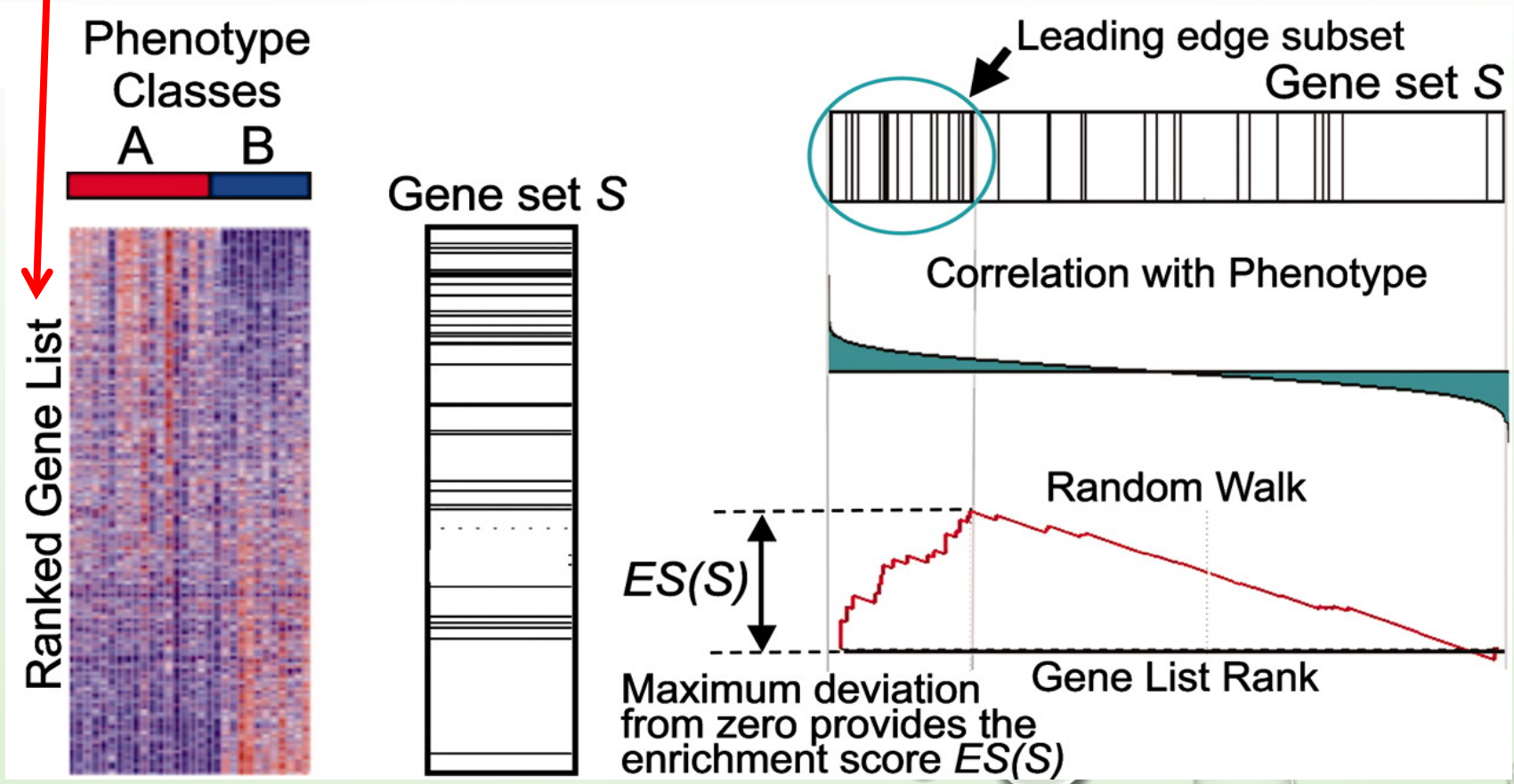
Subramanian et al, PNAS, 2005

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)



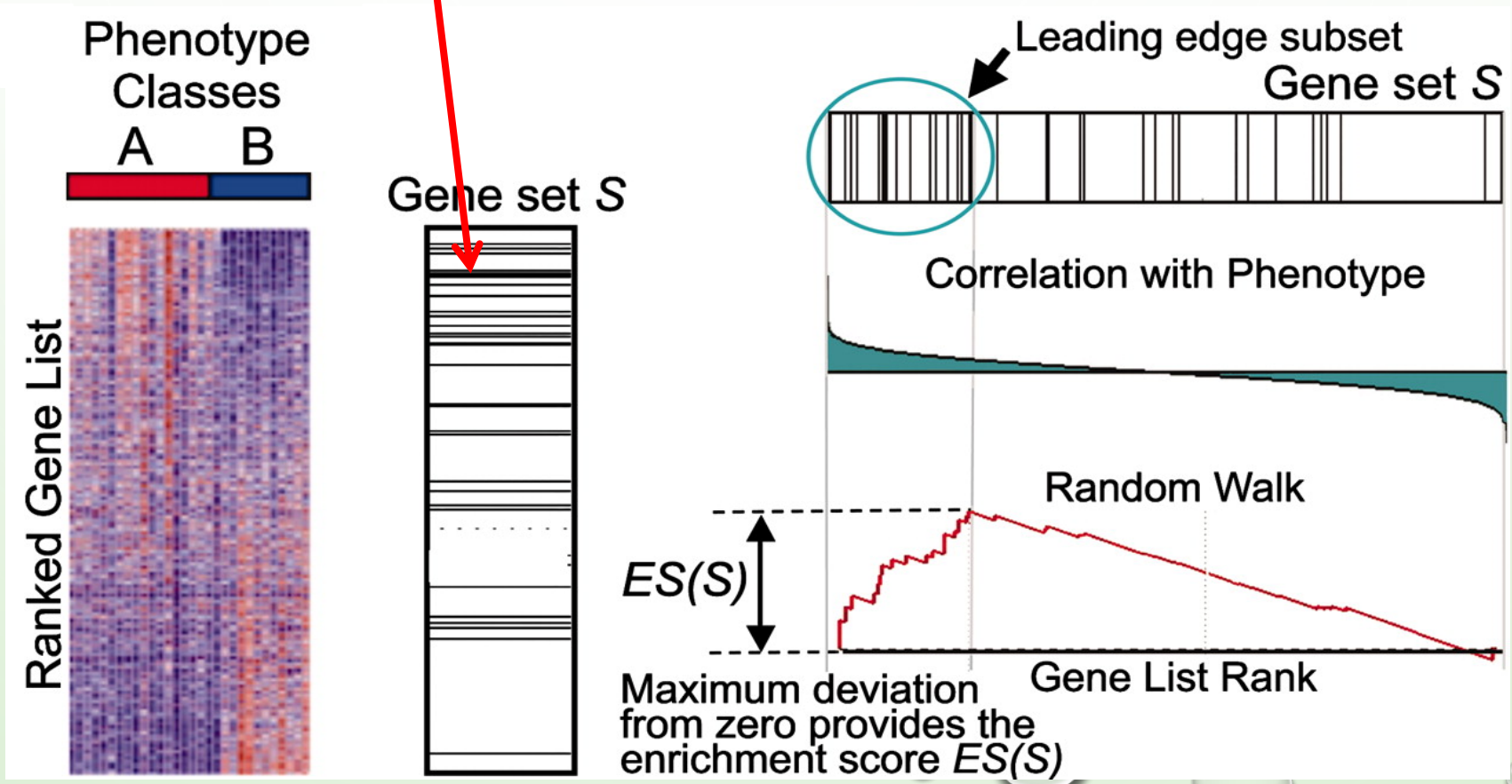
Subramanian et al, PNAS, 2005

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



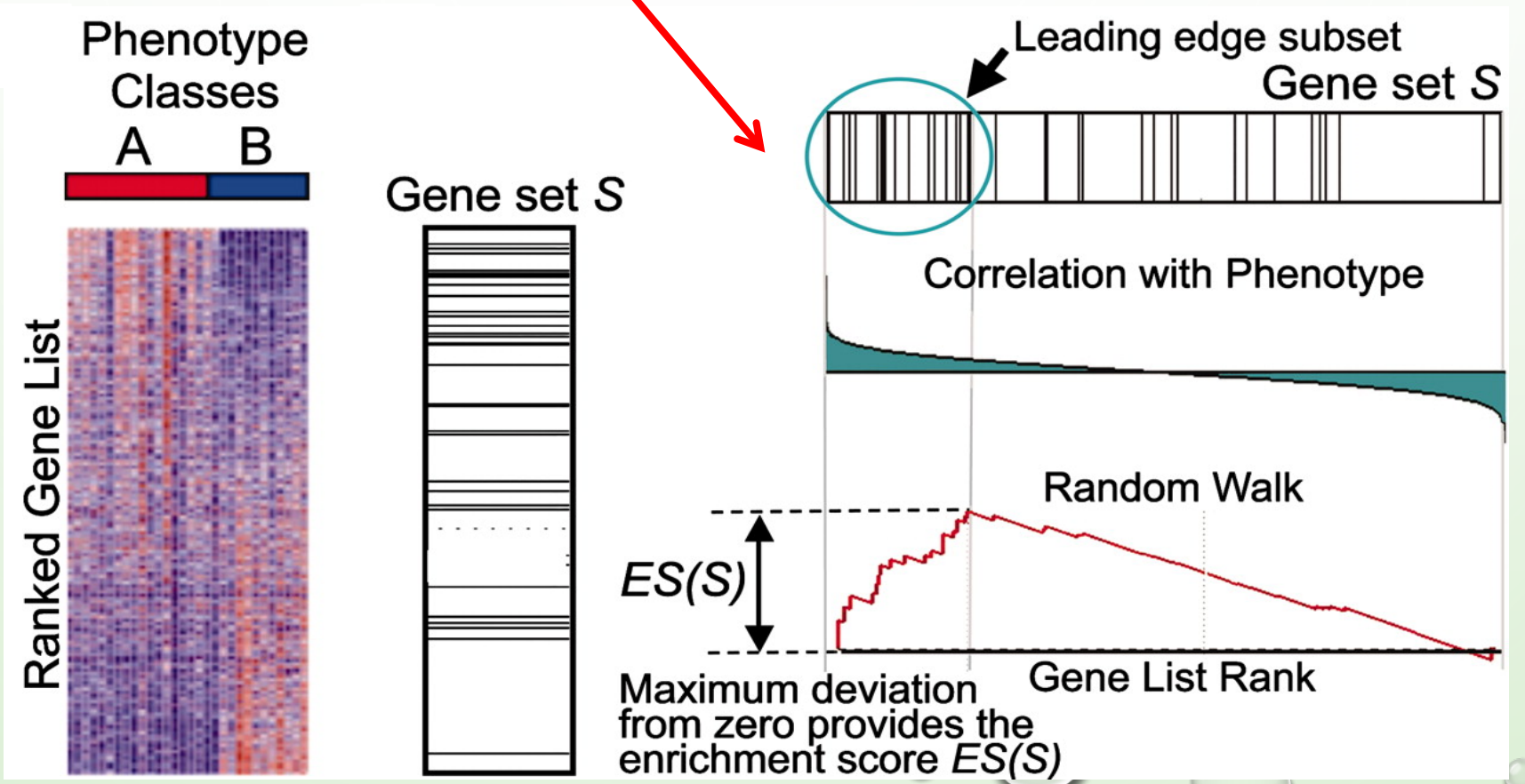
Subramanian et al, PNAS, 2005

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



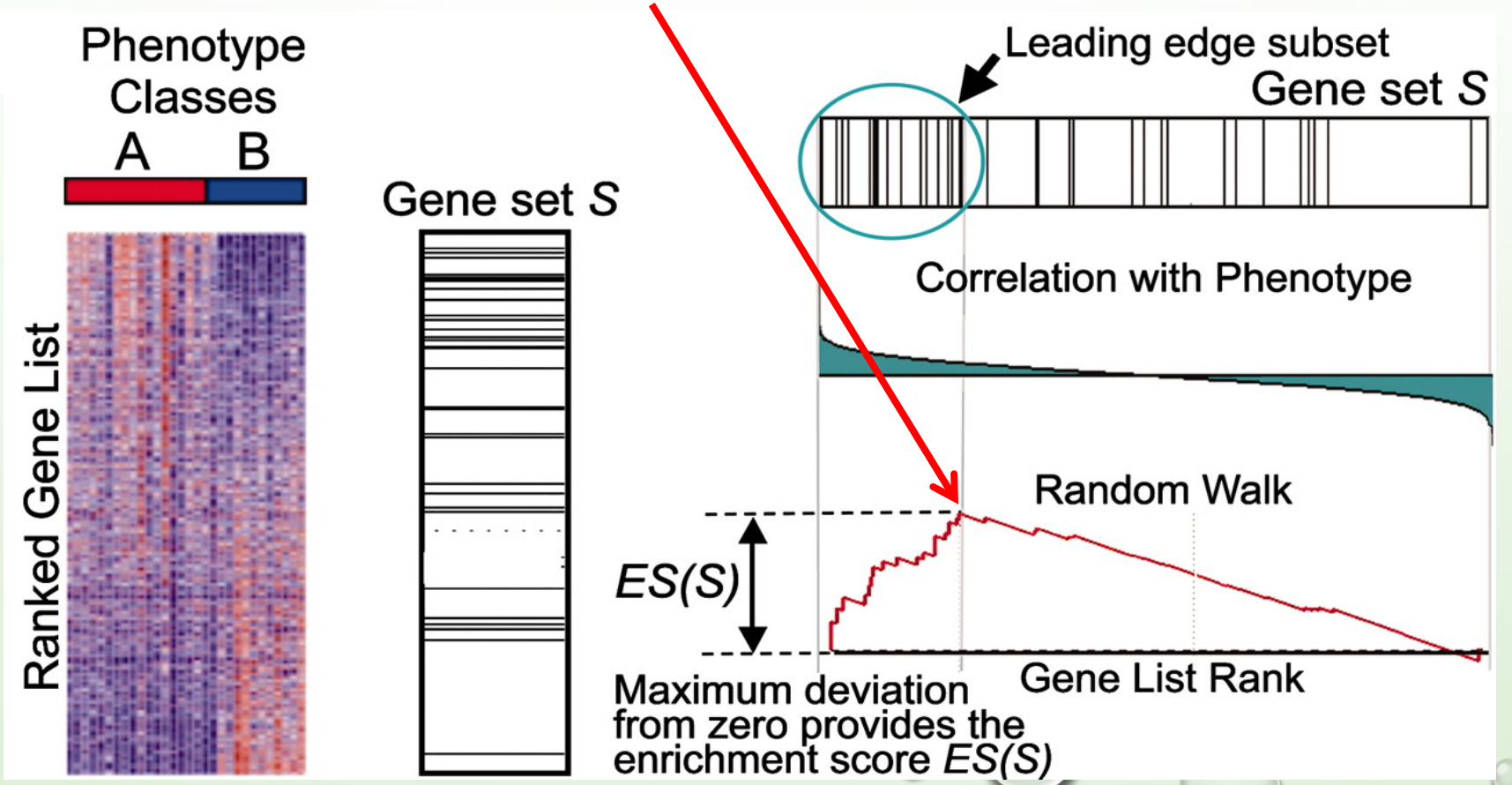
Subramanian et al, PNAS, 2005

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)



Subramanian et al, PNAS, 2005

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



Subramanian et al, PNAS, 2005

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

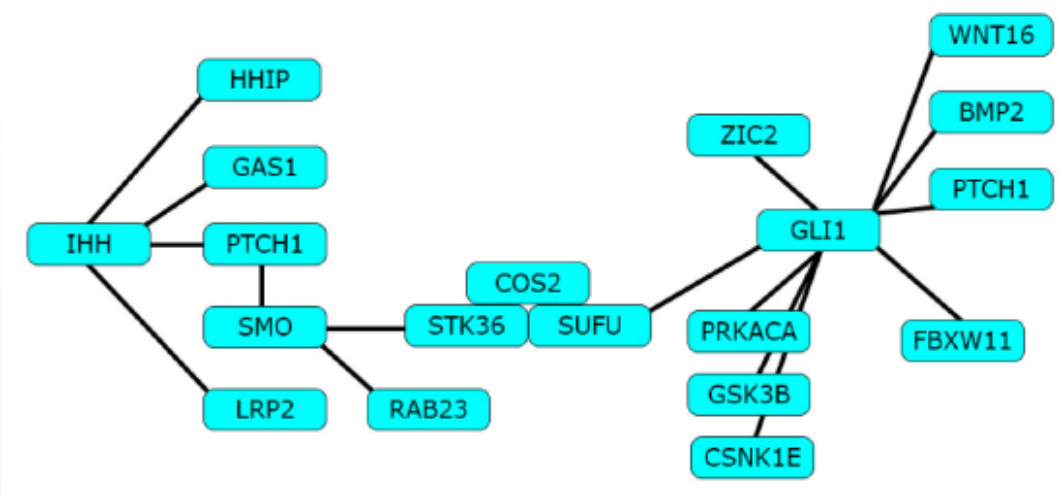
- Pathway are still considered independent. However a gene can function in more than one pathway, meaning that pathways gene sets overlap.

- 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)

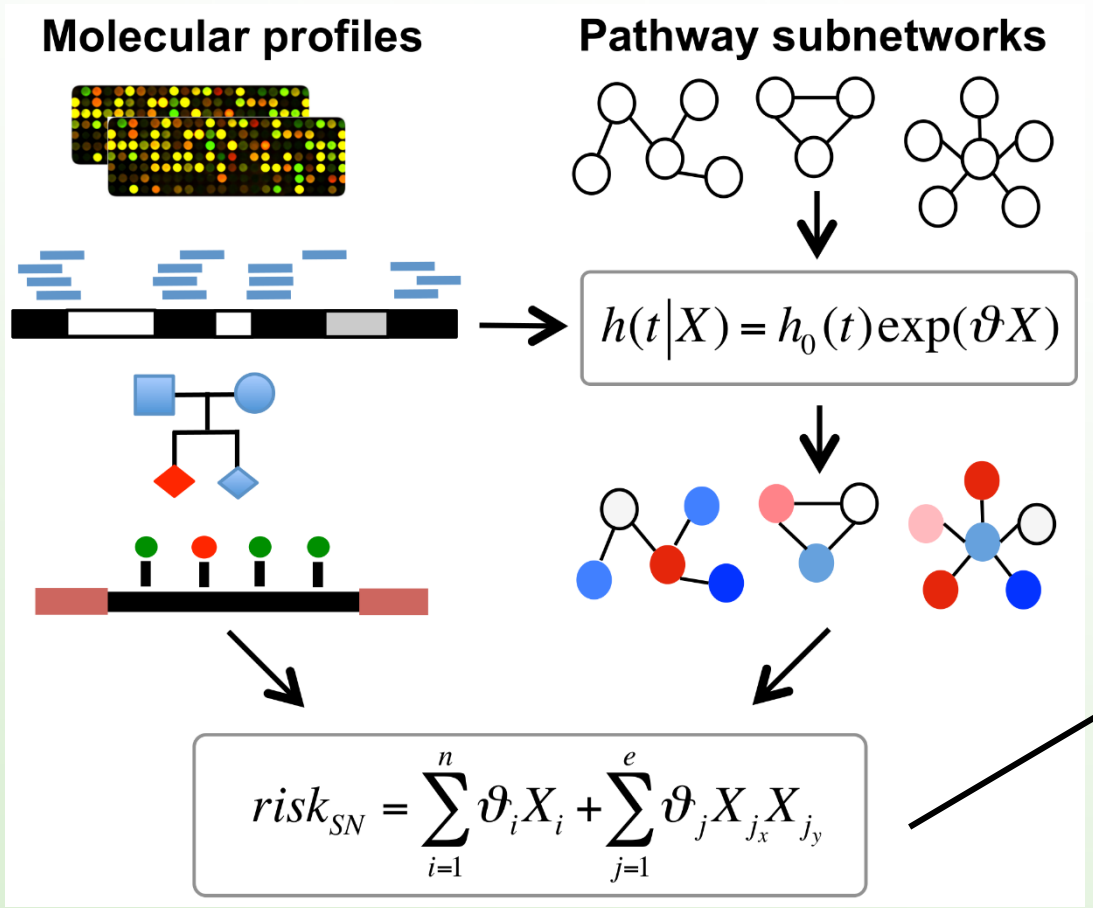
HOW DO WE REPRESENT A PATHWAY

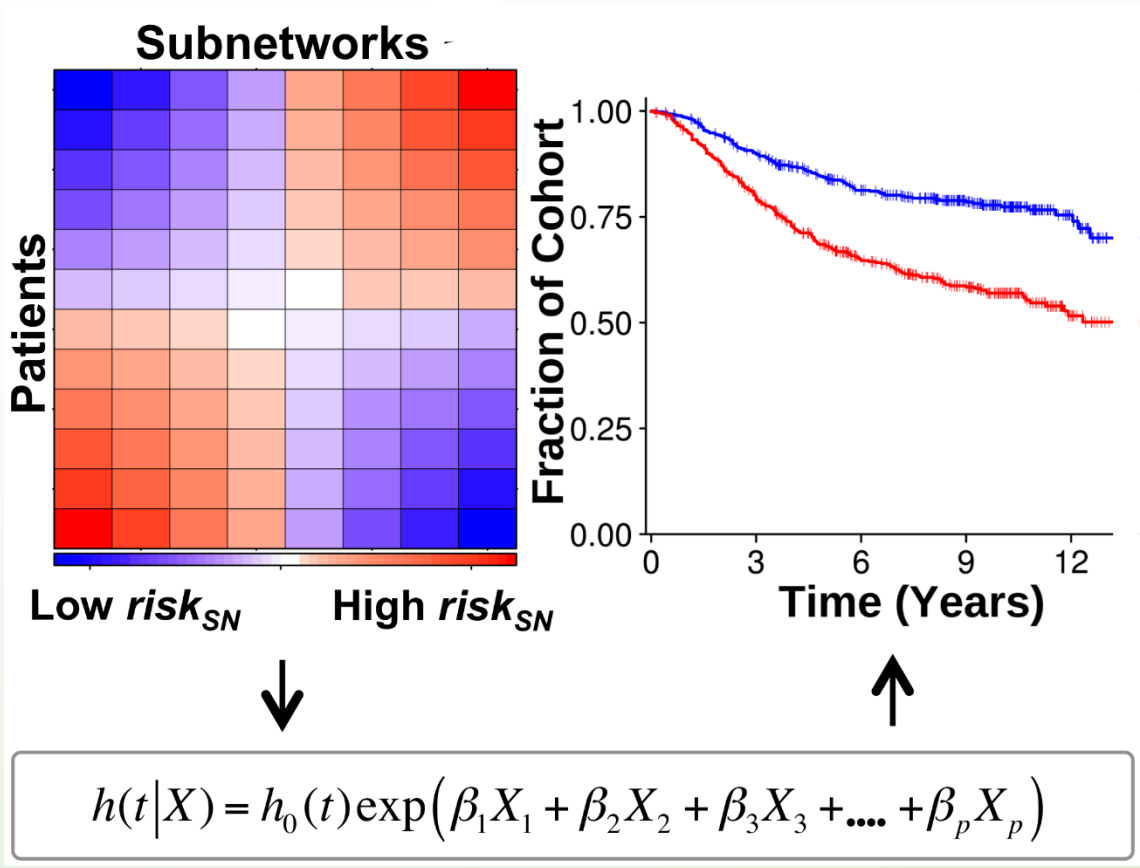
RAB23	GAS1	IHH
ZIC2	PTCH1	GLI1
SMO	HHIP	BMP2
WNT16	PTCH1	STK36
FBXW11	CSNK1E	SUFU
LRP2	PRKACA	GSK3B

A network of associated genes



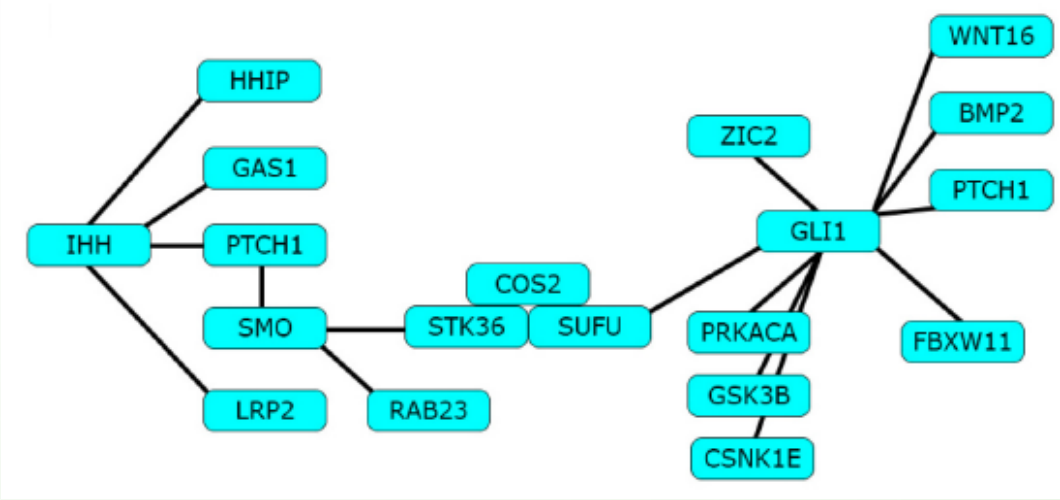
PATHWAY-BASED SUBNETWORKS FOR BIOMARKER DISCOVERY



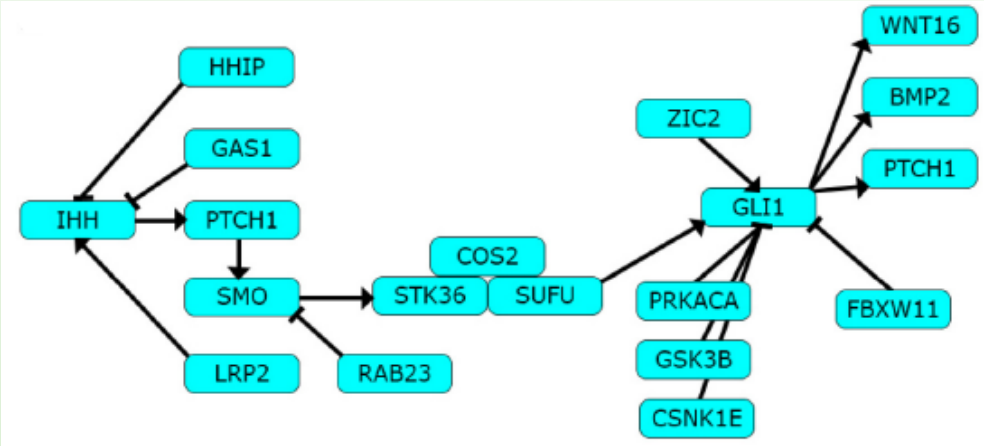


Haider et al, Nat Commun ***9****, 4746 (2018)*

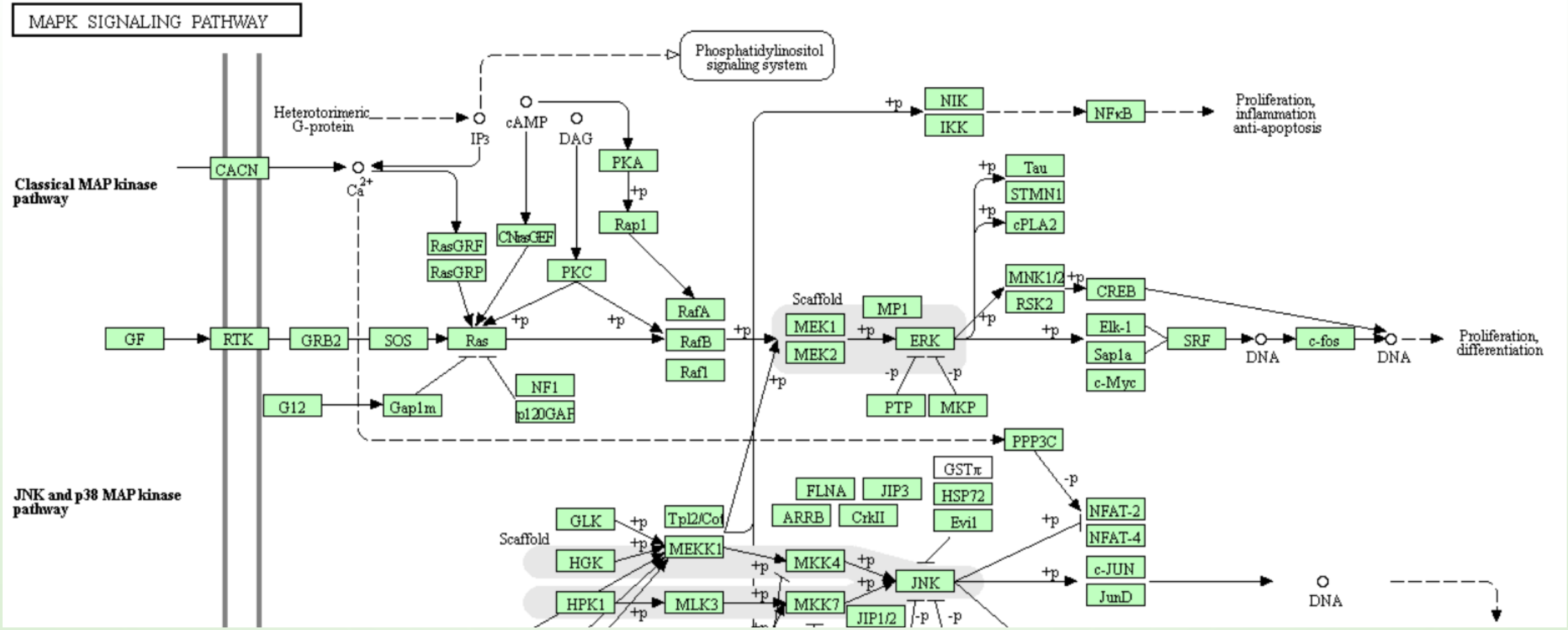
HOW DO WE REPRESENT A PATHWAY



A directed network of genes

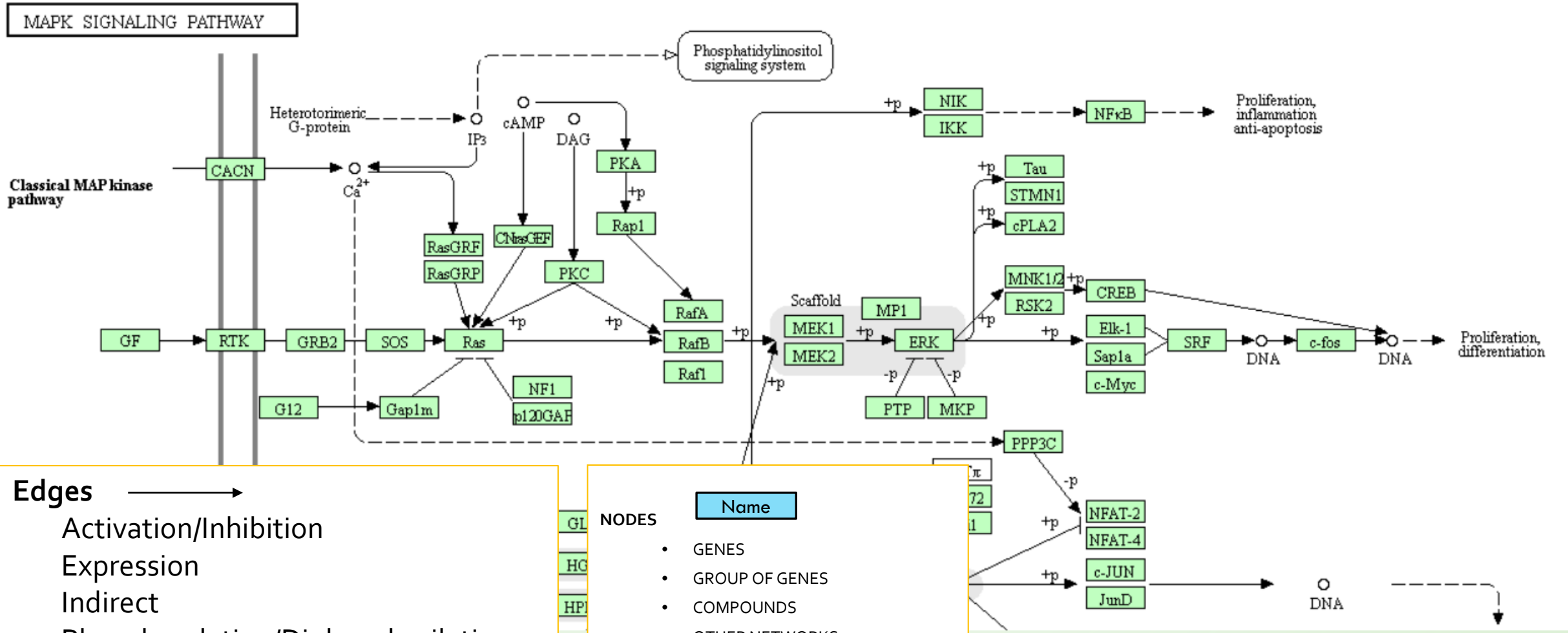


KEGG PATHWAYS



https://www.genome.jp/kegg/pathway.html

Kegg pathways



Edges ⟶

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

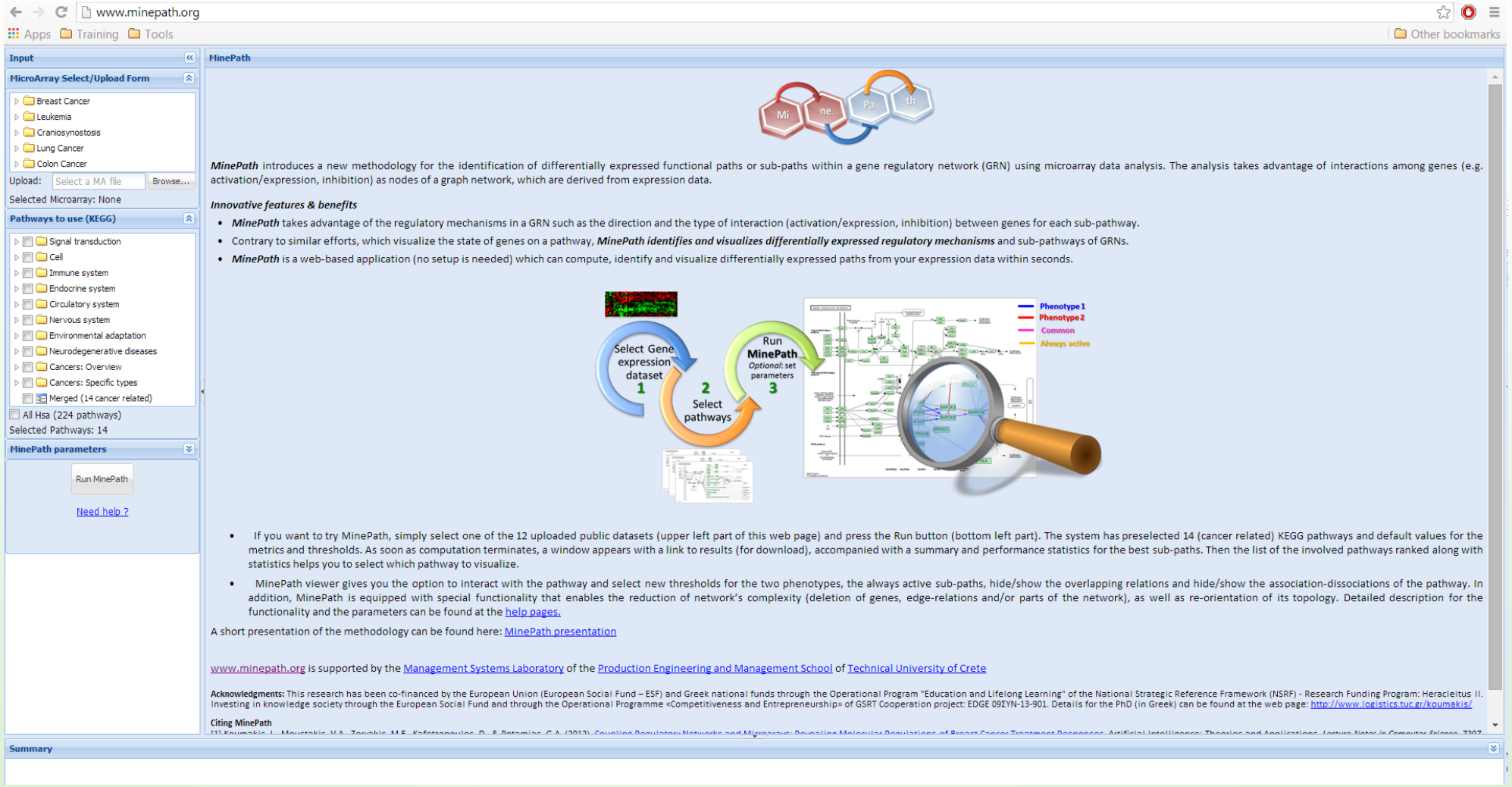
NODES

Name

- GENES
- GROUP OF GENES
- COMPOUNDS
- OTHER NETWORKS

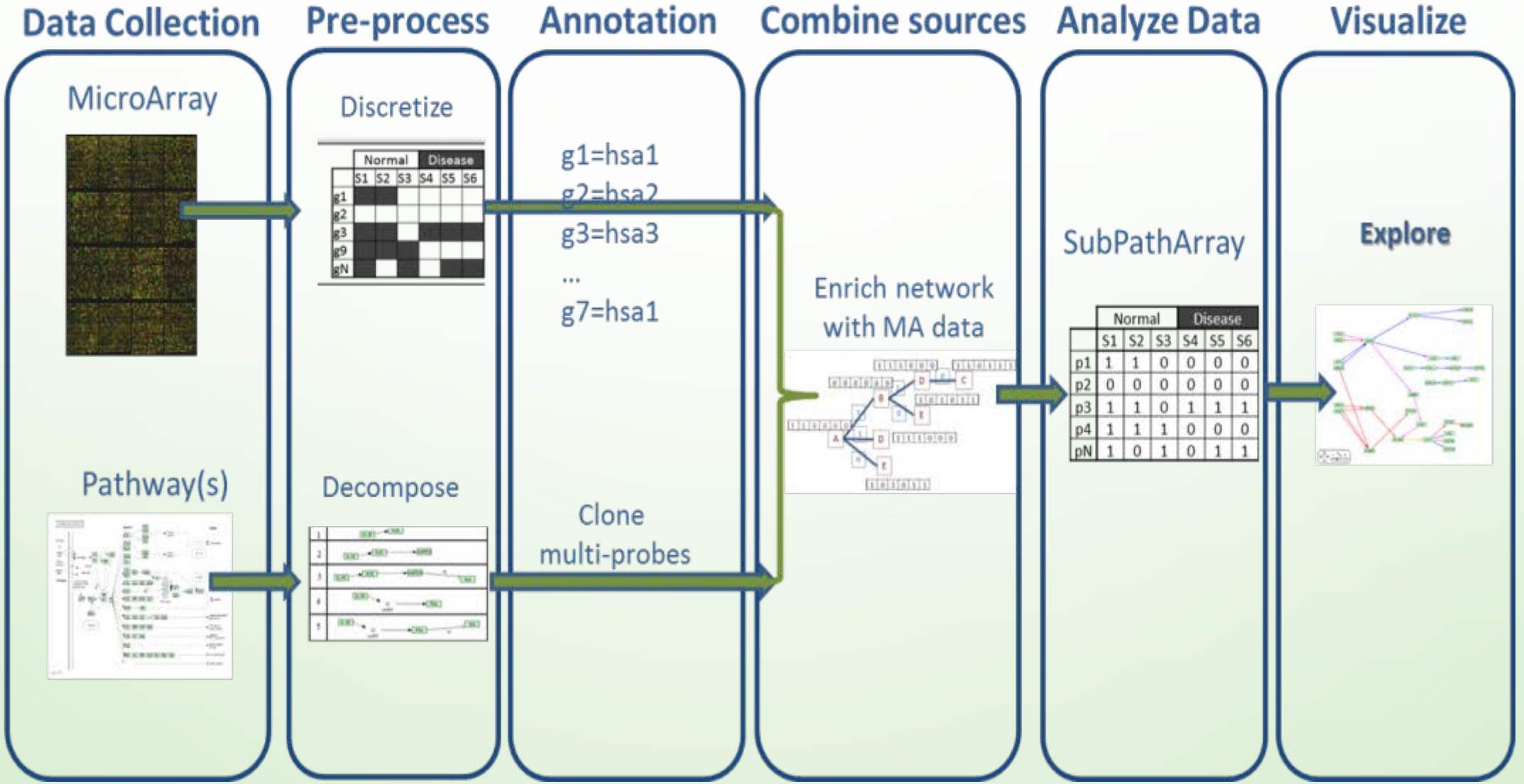
https://www.genome.jp/kegg/pathway.html

MINEPATH

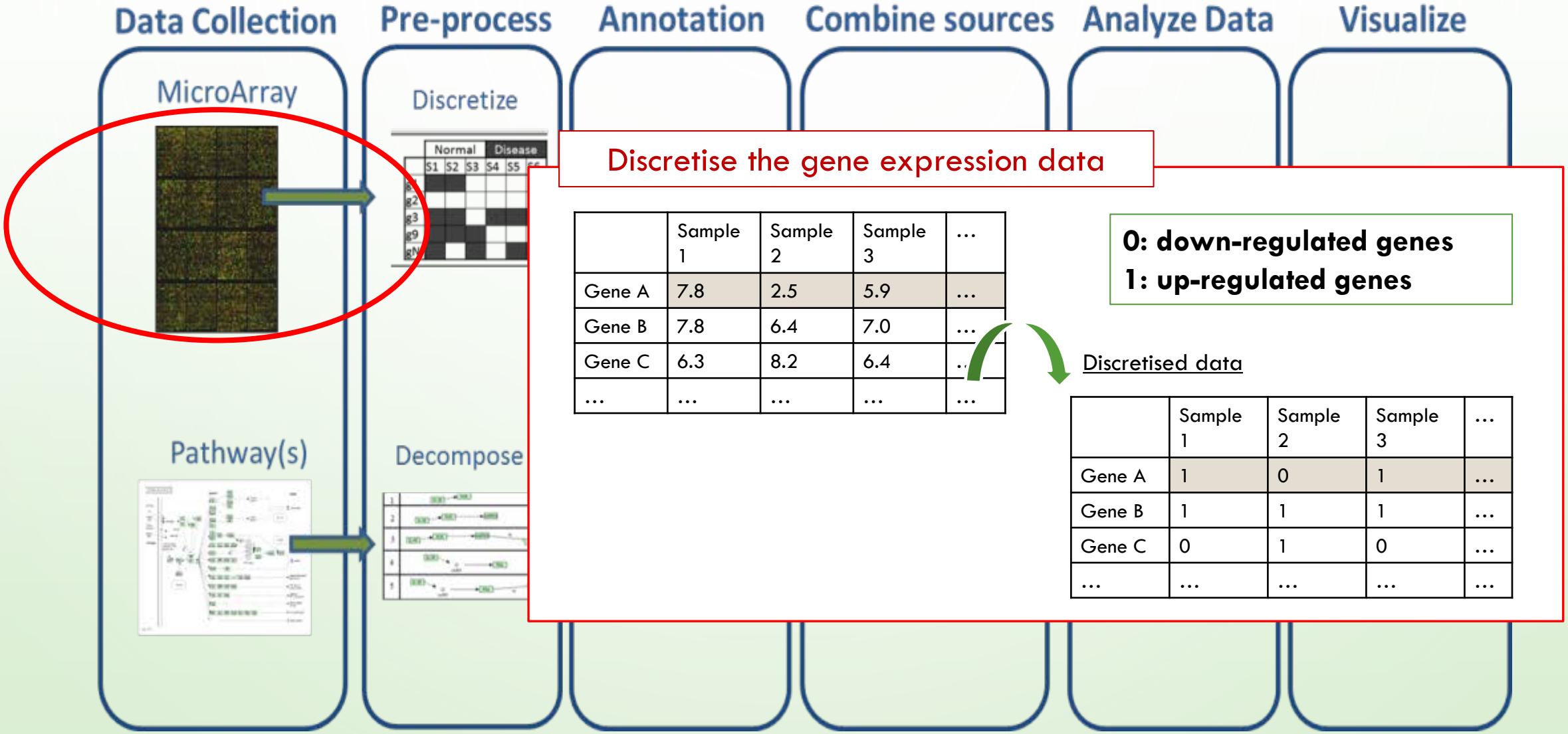


MINEPATH.ORG (KOUMAKIS ET AL., 2012)

SYSTEM OVERVIEW



SYSTEM OVERVIEW



Discretise the gene expression data

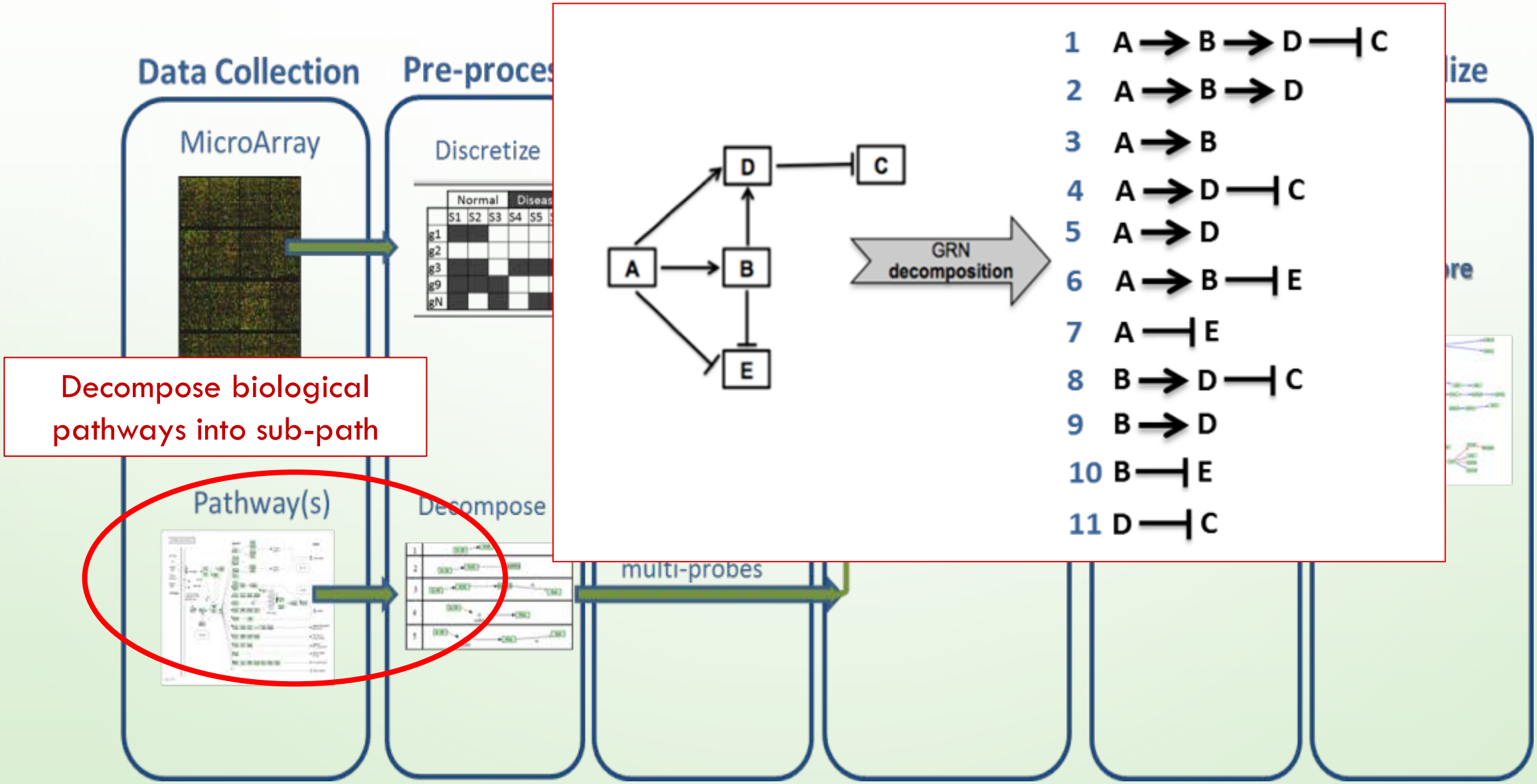
	Sample 1	Sample 2	Sample 3	…
Gene A	7.8	2.5	5.9	…
Gene B	7.8	6.4	7.0	…
Gene C	6.3	8.2	6.4	…
…	…	…	…	…

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

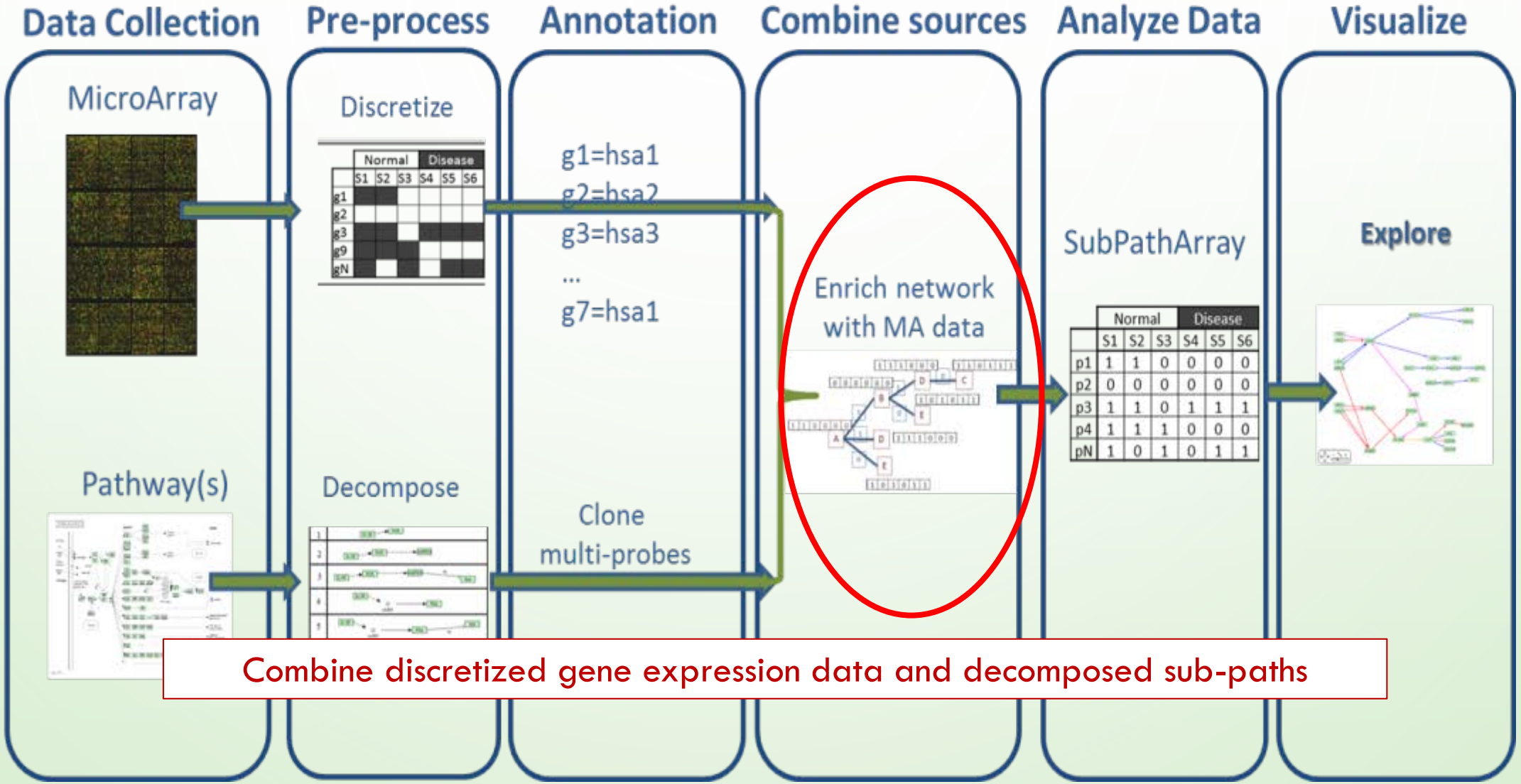
Discretised data

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

SYSTEM OVERVIEW

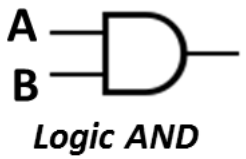


SYSTEM OVERVIEW



Combine discretized gene expression data and decomposed sub-paths

Activation



A → B

		B: ON	B: 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	…
…	…	…	…	…

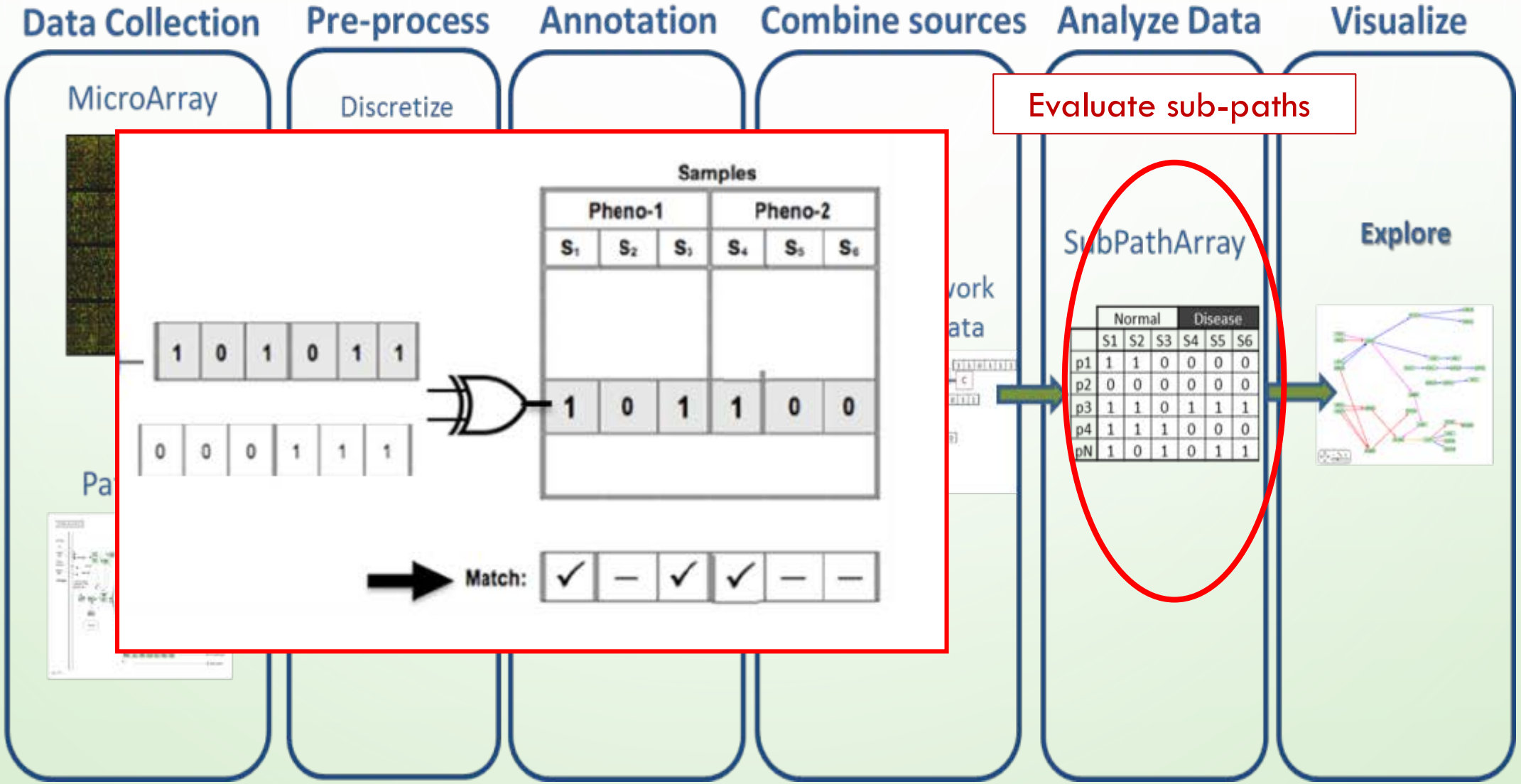
Inhibition

A ⊣ B

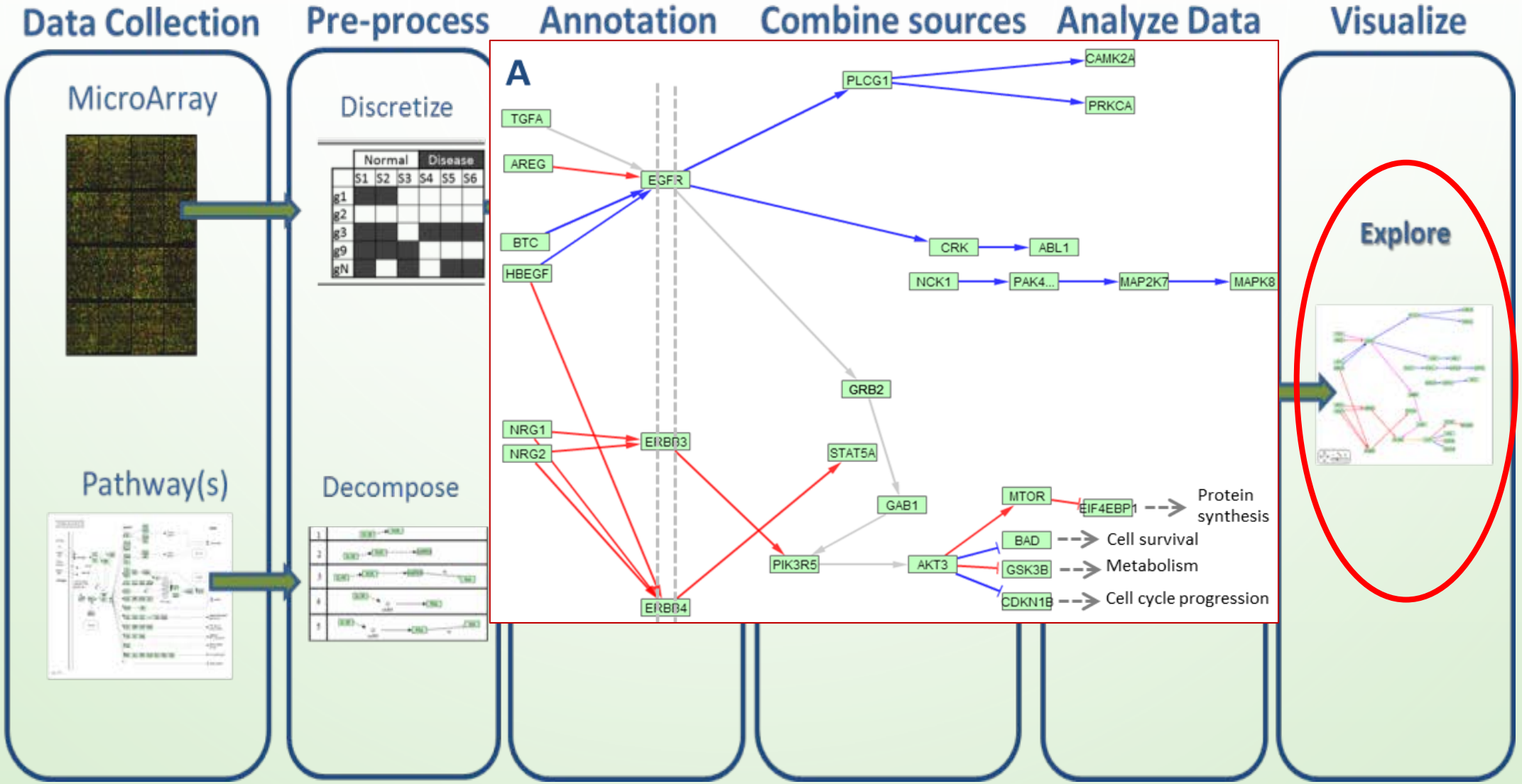
		B: ON	B: OFF
A	ON	×	✓
	OFF	✓	×

INTERACTIONS AS LOGICAL OPERATORS

SYSTEM OVERVIEW



SYSTEM OVERVIEW



Many repositories of biological pathways

2471 Pathways

>42 000 Pathways

505, 700 Pathways

2132 (H. sapiens)

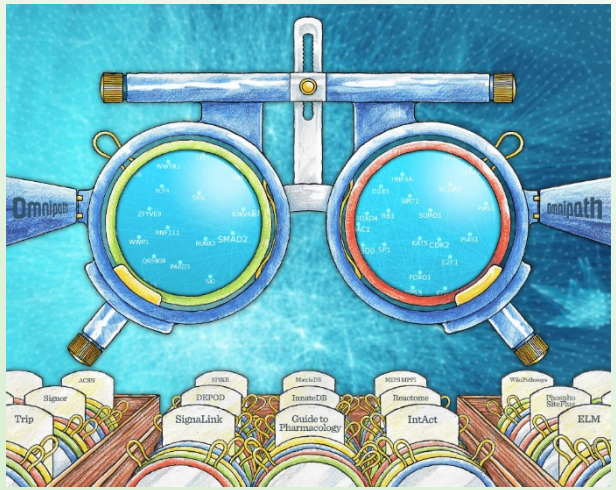
Many repositories of biological pathways

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

Many repositories of biological pathways

BioPAX@ VCELL
Biological Pathways Exchange

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



https://omnipathdb.org/

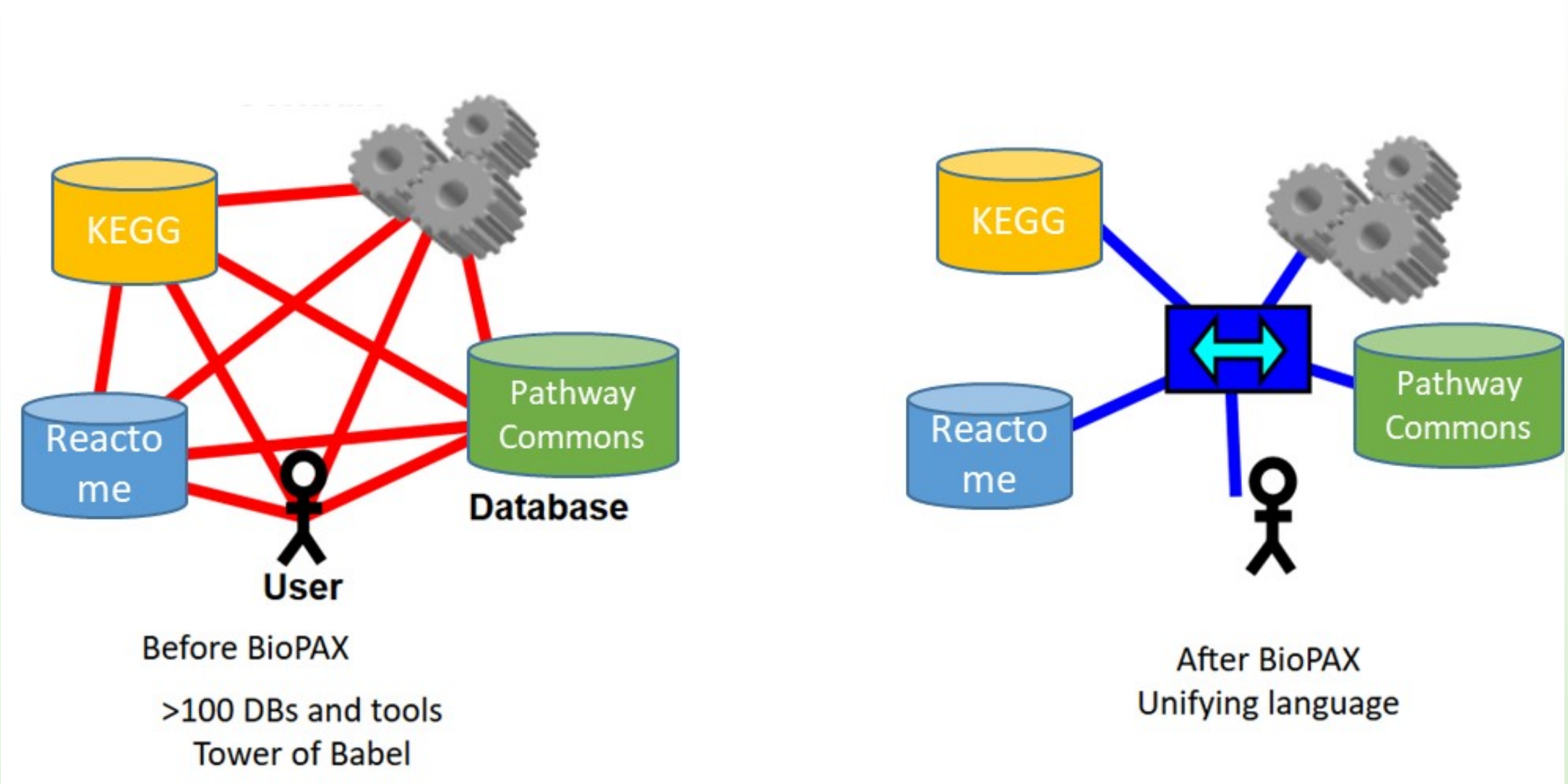
2471 Pathways

>42 000 Pathways

505, 700 Pathways

2132 (H. sapiens)

Biological Pathway Exchange (BioPAX)



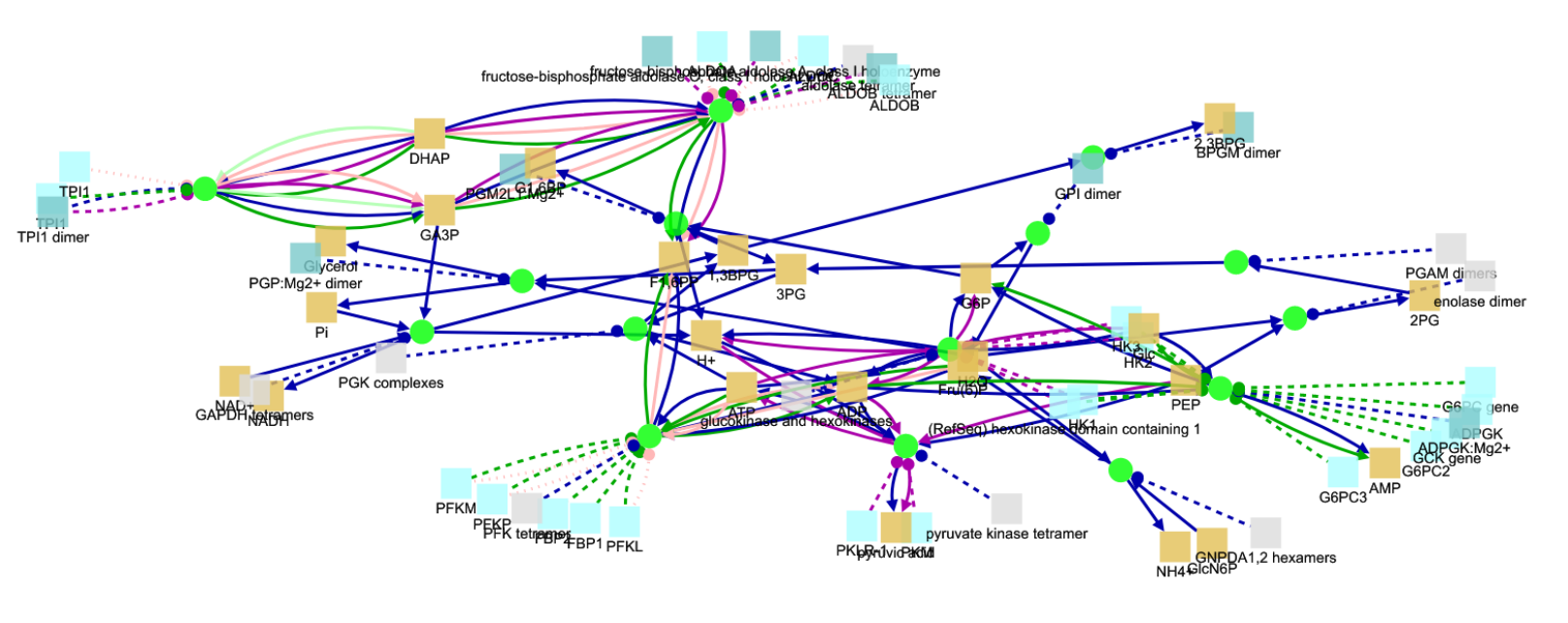
Demir E et al *Nat Biotechnol.* 2010;28(9):935-942

CONSENSUSPATH DATABASE

Integrated databases:

	name	protein interactions	signalling reactions	metabolic reactions	gene regulations	genetic interactions	drug-target interactions	biochemical pathways
	BIND	✓	✗	✗	✓	✗	✗	✗
	BioCarta	✗	✓	✗	✓	✗	✗	✓
BioGRID	Biogrid	✓	✗	✗	✗	✓	✗	✗
mips	CORUM	✓	✗	✗	✗	✗	✗	✗
CHEMBL	Chembl	✗	✗	✗	✗	✗	✓	✗
	DIP	✓	✗	✗	✗	✗	✗	✗
EHMN	EHMN	✗	✗	✓	✗	✗	✗	✓
	HPRD	✓	✗	✗	✗	✗	✗	✗
BioCyc	HumanCyc	✗	✗	✓	✗	✗	✗	✓
INOH	INOH	✗	✓	✓	✗	✗	✗	✓
InnateDB	InnateDB	✓	✓	✗	✓	✗	✗	✗
IntAct	IntAct	✓	✗	✗	✗	✗	✗	✗
KEGG	KEGG	✗	✓	✓	✗	✗	✓	✓
MINT	MINT	✓	✗	✗	✗	✗	✗	✗
mips	MIP 8-MPPI	✓	✗	✗	✗	✗	✗	✗
MatrixDB	MatrixDB	✓	✗	✗	✗	✗	✗	✗
NetPath	NetPath	✓	✓	✗	✗	✗	✗	✓
PDB	PDB	✓	✗	✗	✗	✗	✗	✗
PDZBase	PDZBase	✓	✗	✗	✗	✗	✗	✗
PID	PID	✗	✓	✗	✓	✗	✗	✓
PIG	PIG	✓	✗	✗	✗	✗	✗	✗
PINdb	PINdb	✓	✗	✗	✗	✗	✗	✗
PharmGKB	PharmGKB	✗	✗	✗	✗	✗	✓	✓
PhosphoPOINT	PhosphoPOINT	✓	✓	✗	✗	✗	✗	✗
PhosphoSitePlus	PhosphoSitePlus	✗	✓	✗	✗	✗	✗	✗
Reactome	Reactome	✓	✓	✓	✗	✗	✗	✓
SMPDB	SMPDB	✗	✗	✗	✗	✗	✗	✓
SignaLink	SignaLink	✗	✗	✗	✗	✗	✗	✓
SPIKE	Spike	✓	✓	✗	✓	✗	✗	✗
Therapeutic Targets Database	TTD	✗	✗	✗	✗	✗	✓	✗
WikiPathways	Wikipathways	✗	✓	✓	✗	✗	✗	✓
CPDB	ConsensusPathDB	✓	✓	✓	✓	✓	✓	✓

HTTP://CPDB.MOLGEN.MPG.DE/

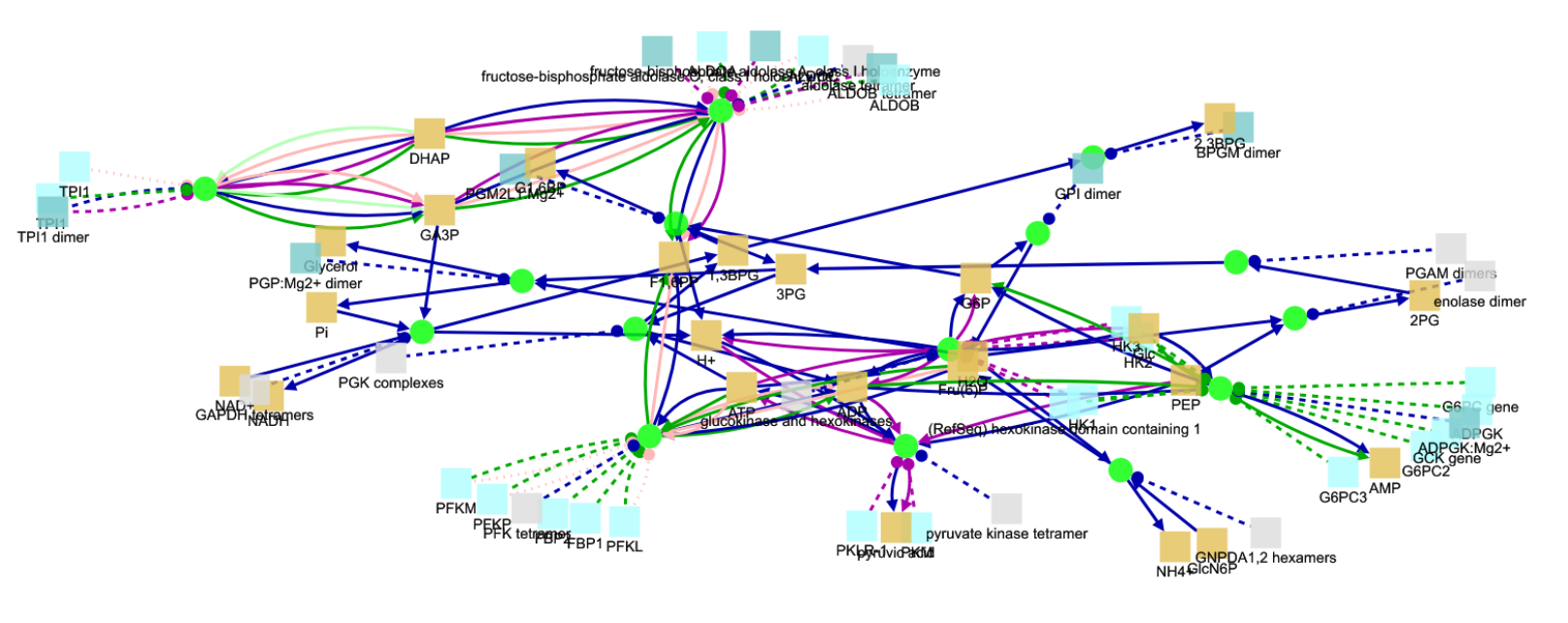


CONSENSUSPATH DATABASE

Integrated databases:

name	protein interactions	signalling reactions	metabolic reactions	gene regulations	genetic interactions	drug-target interactions	biochemical pathways
BIND	✓	✗	✗	✓	✗	✗	✗
BioCarta	✗	✓	✗	✓	✗	✗	✓
Biogrid	✓	✗	✗	✗	✓	✗	✗
CORUM	✓	✗	✗	✗	✗	✗	✗
Chembl	✗	✗	✗	✗	✗	✓	✗
DIP	✓	✗	✗	✗	✗	✗	✗
EHMN	✗	✗	✓	✗	✗	✗	✓
HPRD	✓	✗	✗	✗	✗	✗	✗
HumanCyc	✗	✗	✓	✗	✗	✗	✓
INOH	✗	✓	✓	✗	✗	✗	✓
InnateDB	✓	✓	✗	✓	✗	✗	✗
IntAct	✓	✗	✗	✗	✗	✗	✗
KEGG	✗	✓	✓	✗	✗	✓	✓
MINT	✓	✗	✗	✗	✗	✗	✗
MIP 8-MPPI	✓	✗	✗	✗	✗	✗	✗
MatrixDB	✓	✗	✗	✗	✗	✗	✗
NetPath	✓	✓	✗	✗	✗	✗	✓
PDB	✓	✗	✗	✗	✗	✗	✗
PDZBase	✓	✗	✗	✗	✗	✗	✗
PID	✗	✓	✗	✓	✗	✗	✓
PIG	✓	✗	✗	✗	✗	✗	✗
PINdb	✓	✗	✗	✗	✗	✗	✗
PharmGKB	✗	✗	✗	✗	✗	✓	✓
PhosphoPOINT	✓	✓	✗	✗	✗	✗	✗
PhosphoSitePlus	✗	✓	✗	✗	✗	✗	✗
Reactome	✓	✓	✓	✗	✗	✗	✓
SMPDB	✗	✗	✗	✗	✗	✗	✓
Signalink	✗	✗	✗	✗	✗	✗	✓
Spike	✓	✓	✗	✓	✗	✗	✗
TTD	✗	✗	✗	✗	✗	✓	✗
Wikipathways	✗	✓	✓	✗	✗	✗	✓
ConsensusPathDB	✓	✓	✓	✓	✓	✓	✓

HTTP://CPDB.MOLGEN.MPG.DE/



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

MORE LEARNING RESOURCES

- HTTPS://WWW.EBI.AC.UK/TRAINING/
- HTTPS://WWW.NCBI.NLM.NIH.GOV/HOME/LEARN/