



Swiss Institute of  
Bioinformatics

BIOLOGY-INFORMED INTEGRATION AND VISUALIZATION OF  
MULTIOMICS DATA

# Enrichment analysis

Deepak Tanwar

February 17-19, 2026



# Learning objectives

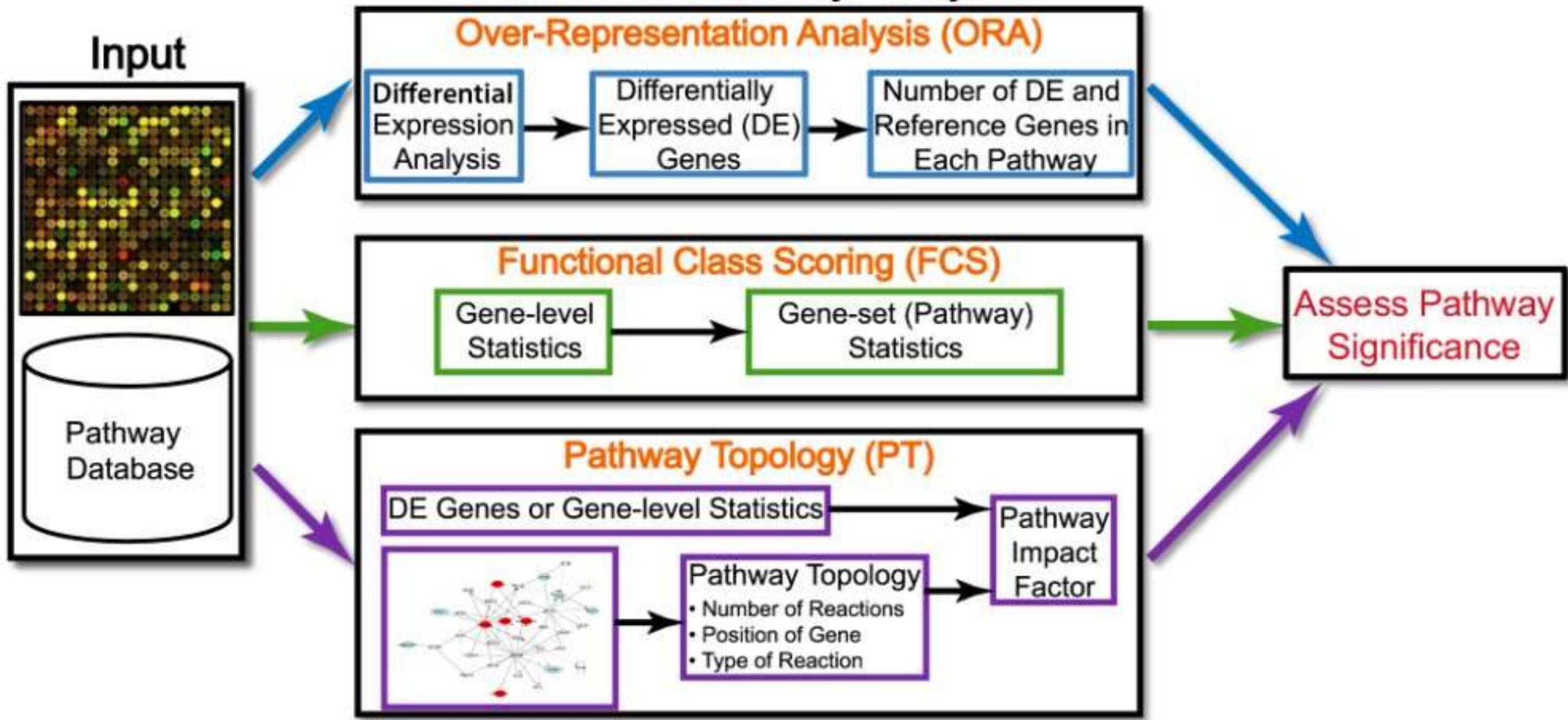
What is Enrichment analysis?

Distinguish between different ways to do it.

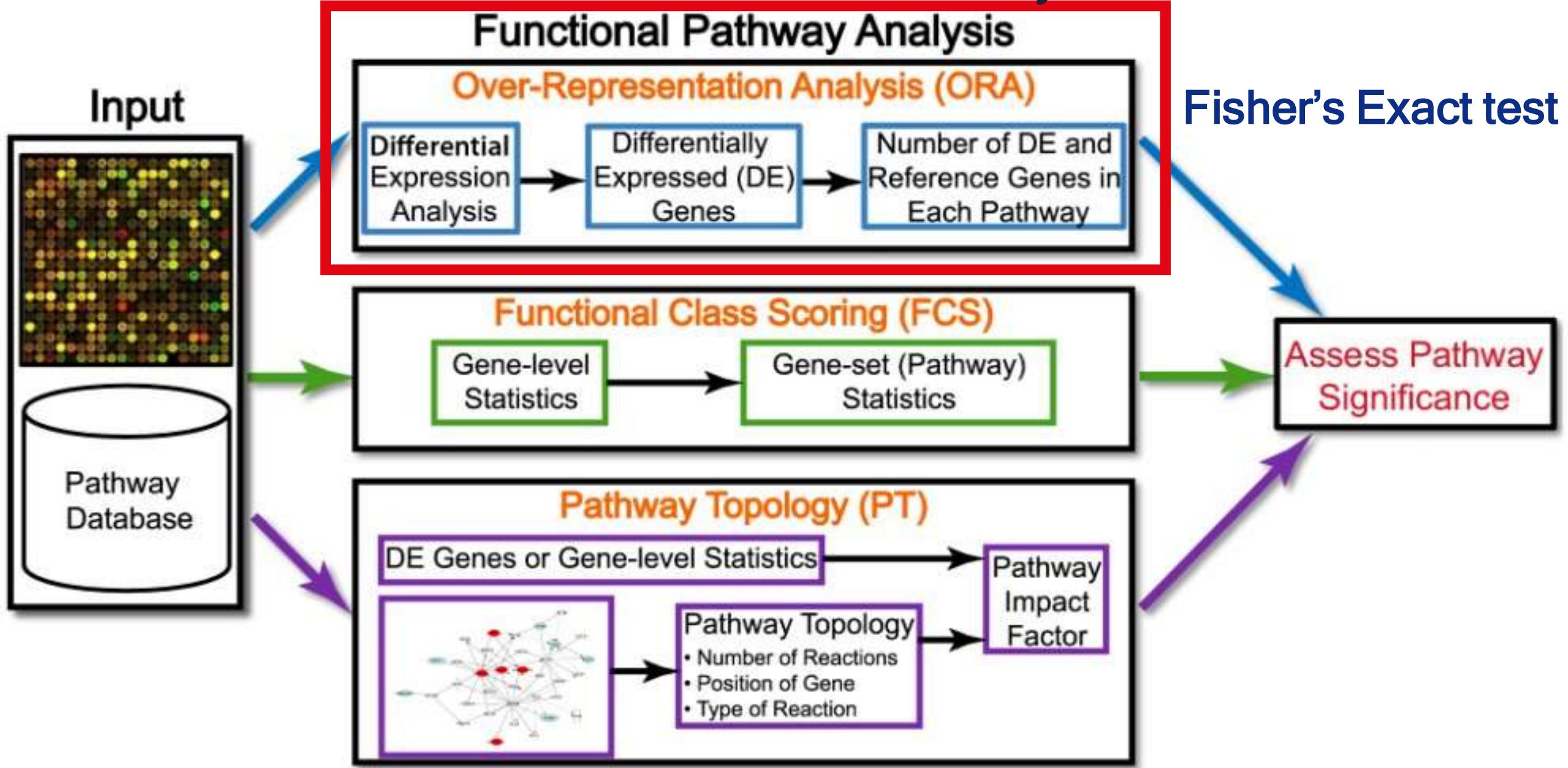
Challenges and Limitations of methods.

# Overview of functional analysis

## Functional Pathway Analysis

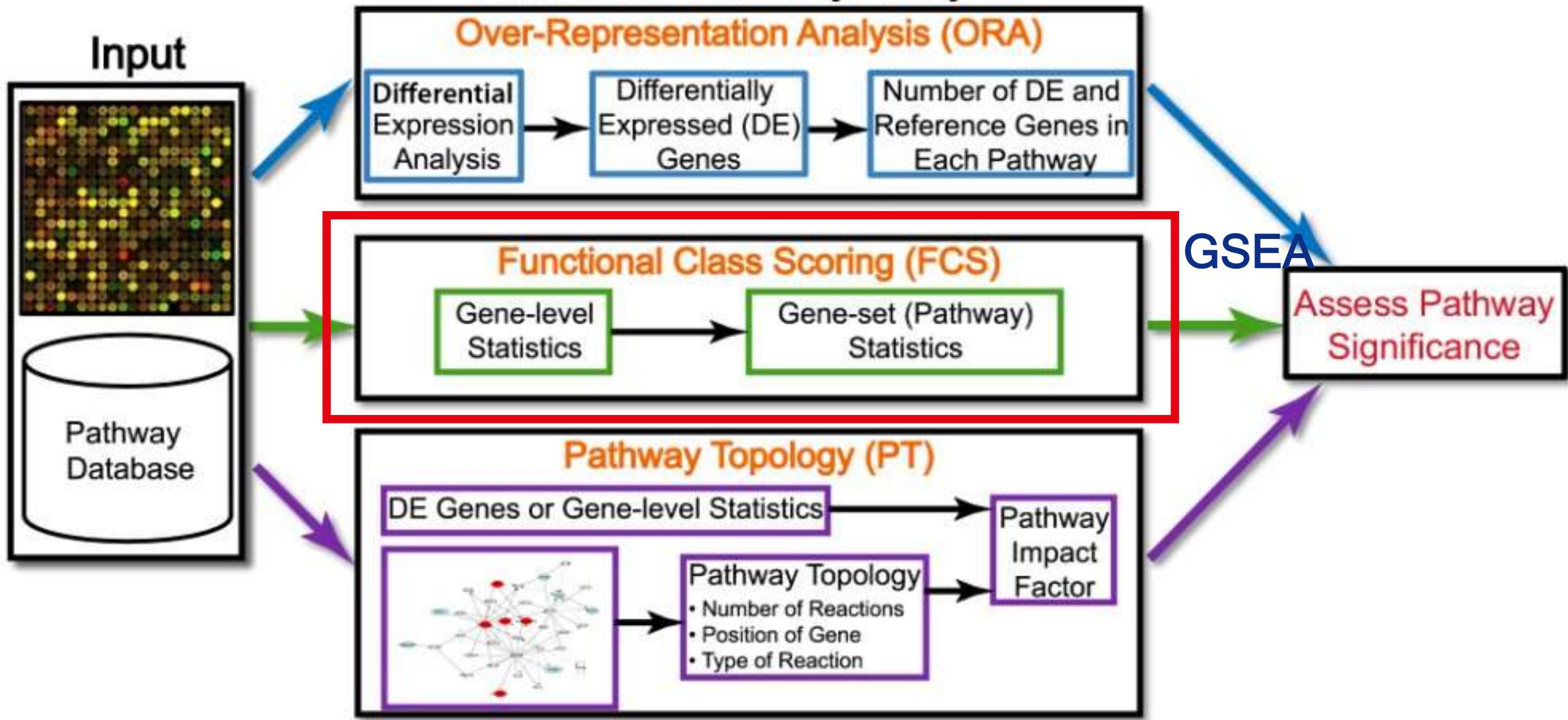


# Overview of functional analysis: ORA



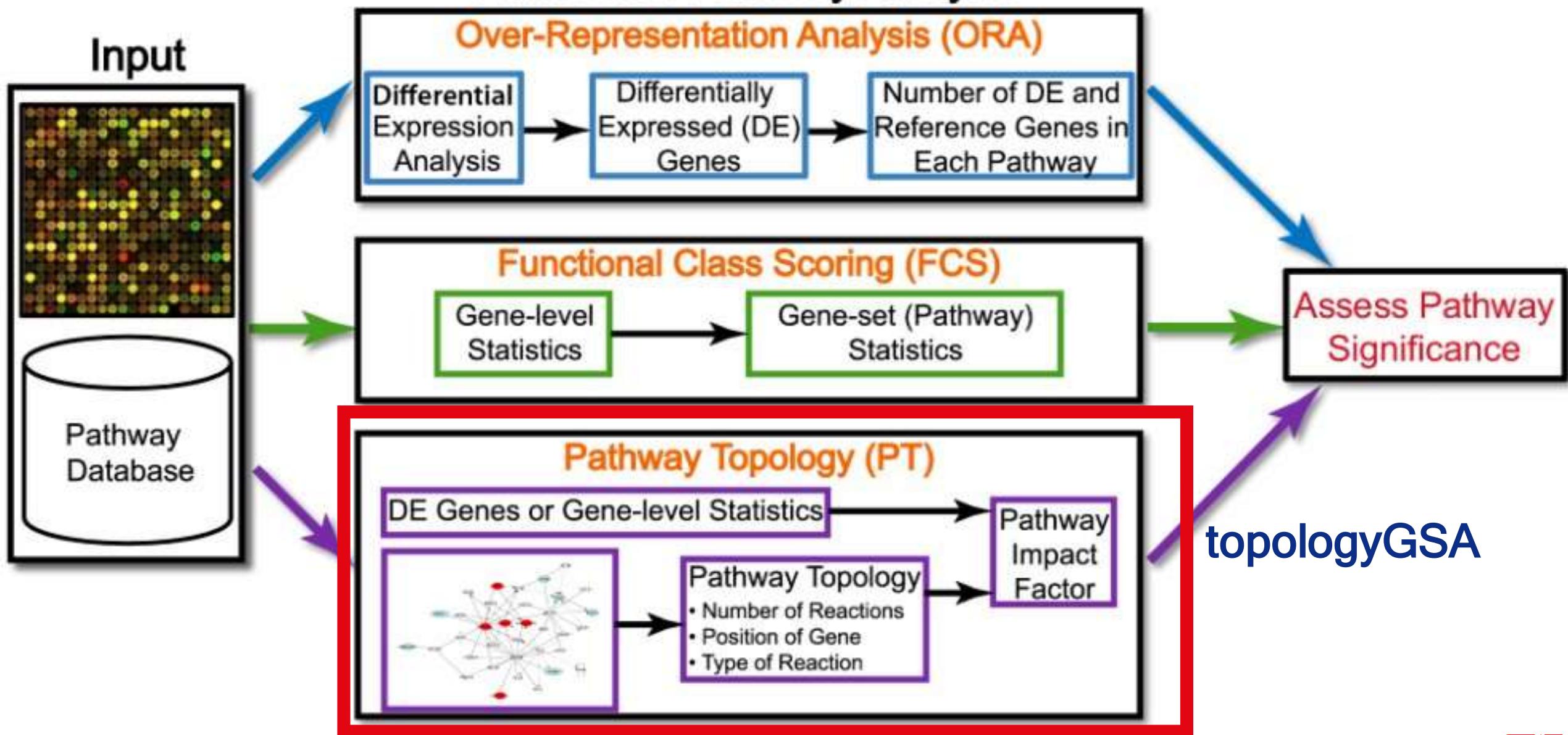
# Overview of functional analysis: FCS

## Functional Pathway Analysis

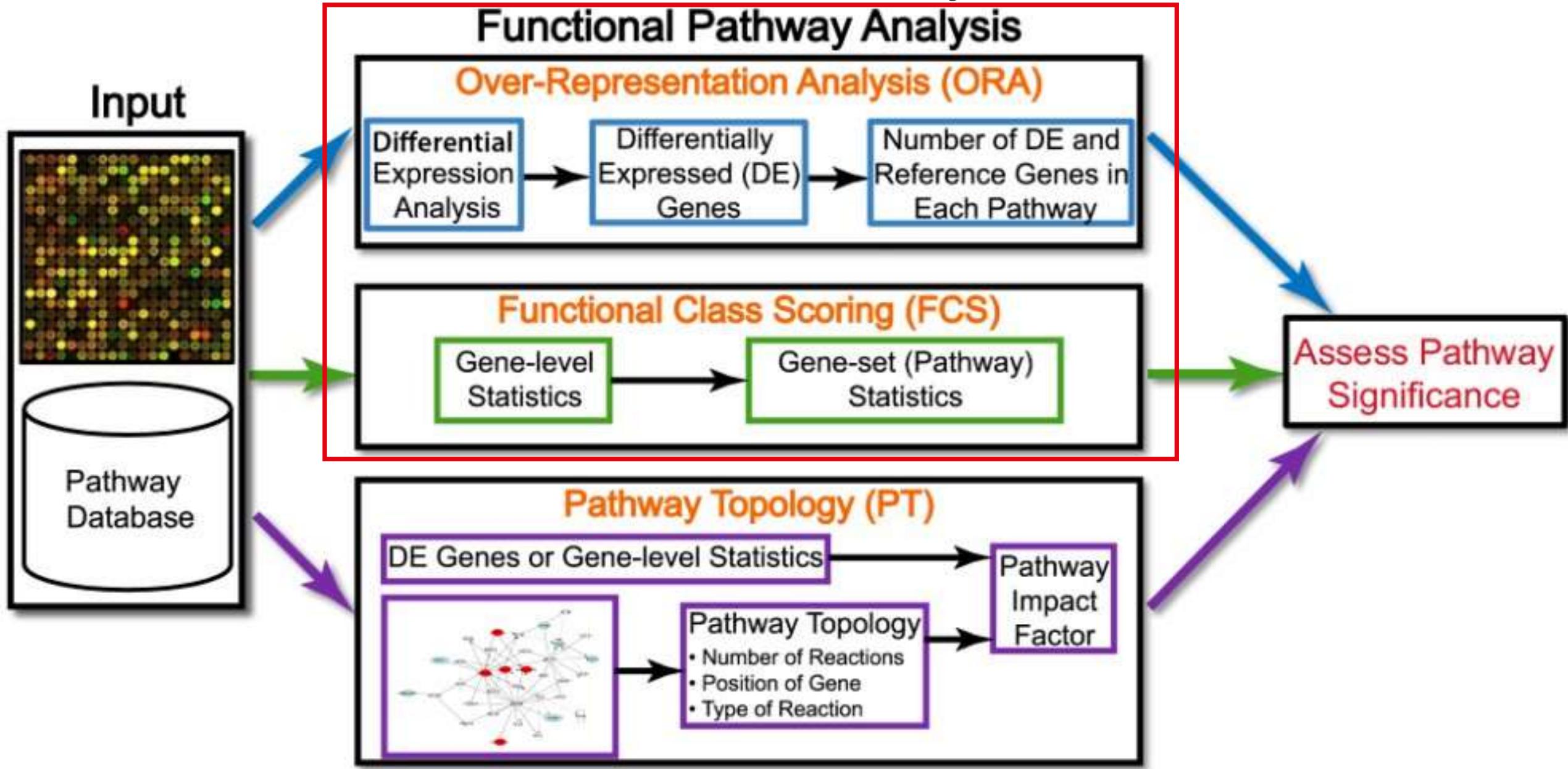


# Overview of functional analysis: PT

## Functional Pathway Analysis



# Overview of functional analysis: ORA & FCS



**Goal:** To gain biologically meaningful insights from long gene lists

# Over-representation analysis (ORA)

Statistically evaluates the fraction of genes in a particular pathway found among the set of genes showing changes in expression.

1. Select a list of genes with certain threshold ( $FDR \leq 0.05$ )
2. For each pathway, count input genes that are part of the pathway
3. Repeat for an appropriate background list of genes
4. Every pathway is tested for over- or under-representation in the list of input genes

The most commonly used tests are based on the hypergeometric, chi-square, or binomial distribution

# Over-representation analysis (ORA)

Gene1	0.051
Gene2	0.05001
Gene 3	0.049
Gene 4	0.001
Gene 5	0.023
Gene 6	0.04
Gene 7	0.01
Gene 8	0.0501
Gene 9	0.2
Gene 10	0.051
Gene 11	0.05
Gene 12	0.49
Gene 13	0.03
Gene 14	0.01
Gene 15	0.052
Gene 16	0.9

# Over-representation analysis (ORA)

Gene1	0.051
Gene2	0.05001
Gene 3	0.049
Gene 4	0.001
Gene 5	0.023
Gene 6	0.04
Gene 7	0.01
Gene 8	0.0501
Gene 9	0.2
Gene 10	0.051
Gene 11	0.05
Gene 12	0.49
Gene 13	0.03
Gene 14	0.01
Gene 15	0.052
Gene 16	0.9

*pvalue*  $\leq 0.05$

Gene 3	0.049
Gene 4	0.001
Gene 5	0.023
Gene 6	0.04
Gene 7	0.01

Gene 11	0.05
Gene 12	0.49
Gene 13	0.03
Gene 14	0.01

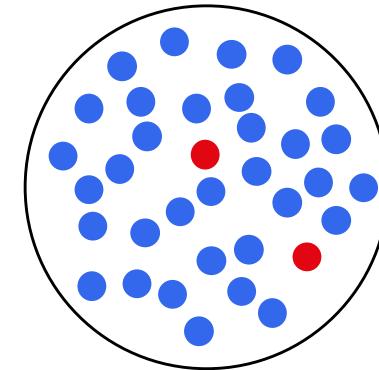
# Over-representation analysis (ORA)

Gene1	0.051
Gene2	0.05001
Gene 3	0.049
Gene 4	0.001
Gene 5	0.023
Gene 6	0.04
Gene 7	0.01
Gene 8	0.0501
Gene 9	0.2
Gene 10	0.051
Gene 11	0.05
Gene 12	0.49
Gene 13	0.03
Gene 14	0.01
Gene 15	0.052
Gene 16	0.9

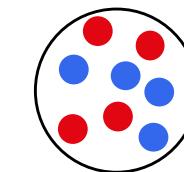
*pvalue*  $\leq 0.05$

Gene 3	0.049
Gene 4	0.001
Gene 5	0.023
Gene 6	0.04
Gene 7	0.01

Fisher's test



- In gene set
- Not in gene set



Differentially expressed

Gene 11	0.05
Gene 12	0.49
Gene 13	0.03
Gene 14	0.01

$H_0$ : The proportion of genes in the gene set is the same for both groups

$H_a$ : The proportion of genes in the gene set is higher in the differentially expressed group

# Problems with ORA

Cutoff? 0.051?

Treat all genes equally

Each gene is independent of other

Each pathway is independent of each other

# Functional class scoring (FCS)

The hypothesis of FCS is that although large changes in individual genes can have significant effects on pathways, weaker but coordinated changes in sets of functionally related genes (i.e., pathways) can also have significant effects

1. Rank the genes
2. Perform gene-level statistics in a pathway
3. Calculate pathway level-statistics: - Kolmogorov-Smirnov statistic

# Over-representation analysis (ORA)

Gene1	0.051	10
Gene2	0.05001	12
Gene 3	0.049	11
Gene 4	0.001	8
Gene 5	0.023	2
Gene 6	0.04	3
Gene 7	0.01	1
Gene 8	0.0501	3
Gene 9	0.2	-10
Gene 10	0.051	-3
Gene 11	0.05	-8
Gene 12	0.49	-19
Gene 13	0.03	-3
Gene 14	0.01	-2
Gene 15	0.052	-1
Gene 16	0.9	-4

# Over-representation analysis (ORA)

Gene1	0.051	10
Gene2	0.05001	12
Gene 3	0.049	11
Gene 4	0.001	8
Gene 5	0.023	2
Gene 6	0.04	3
Gene 7	0.01	1
Gene 8	0.0501	3
Gene 9	0.2	-10
Gene 10	0.051	-3
Gene 11	0.05	-8
Gene 12	0.49	-19
Gene 13	0.03	-3
Gene 14	0.01	-2
Gene 15	0.052	-1
Gene 16	0.9	-4

## Gene set enrichment analysis (GSEA)

Genes ranked by test statistic

or

$\log_2(\text{FC}) * t\text{-value}$



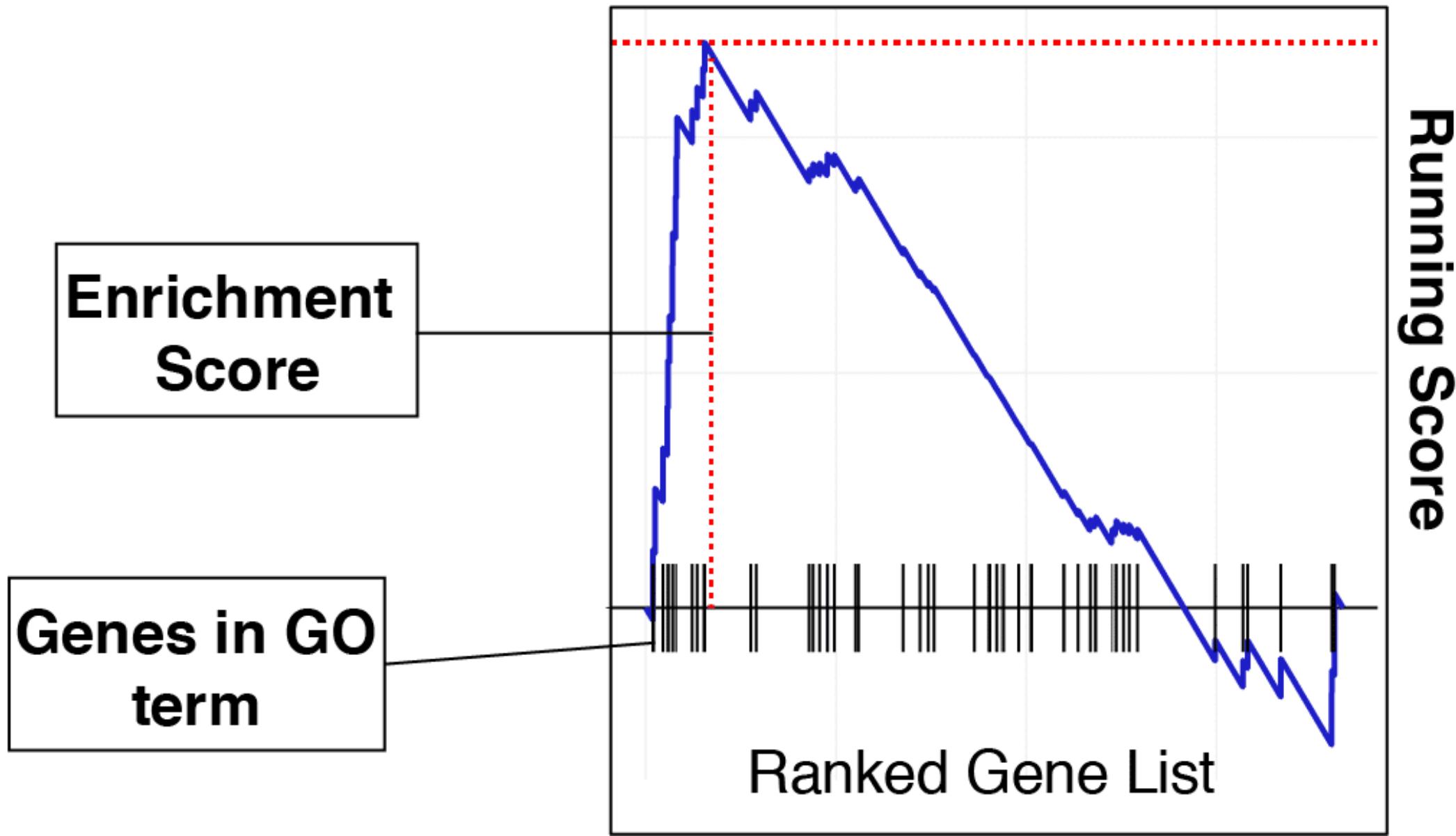
Upregulated

Downregulated

$H_0$ : Genes in set are randomly distributed over ranked list

$H_a$ : Genes in set are not randomly distributed over the ranked list

# Functional class scoring (FCS)



# Problems with FCS

Each gene is independent of other

Each pathway is independent of each other

## Databases

- GO: BP, MF, CC
- KEGG
- Reactome
- DOSE
- DisGeNET
- MSigDb
- KEGG module
- WikiPathways
- TF
- miRNA
- "user input"
- PathGuide

## Methods

- ORA
- GSEA
- SAFE
- PADOG
- ROAST
- CAMERA
- GSA
- GSVA/ssGSEA
- GlobelTest
- EBM
- MGSA
- GOSeq
- QUSAGE
- Pathview
- GOSemSim
- GGEA
- SPIA
- PathNet
- DEGraph
- TopologyGSA
- GANPA
- CePa
- NetGSA
- WGCNA

Databases and methods

## Databases

- GO: BP, MF, CC
- KEGG
- Reactome
- DOSE
- DisGeNET
- MSigDb
- KEGG module
- WikiPathways
- TF
- miRNA
- "user input"
- PathGuide

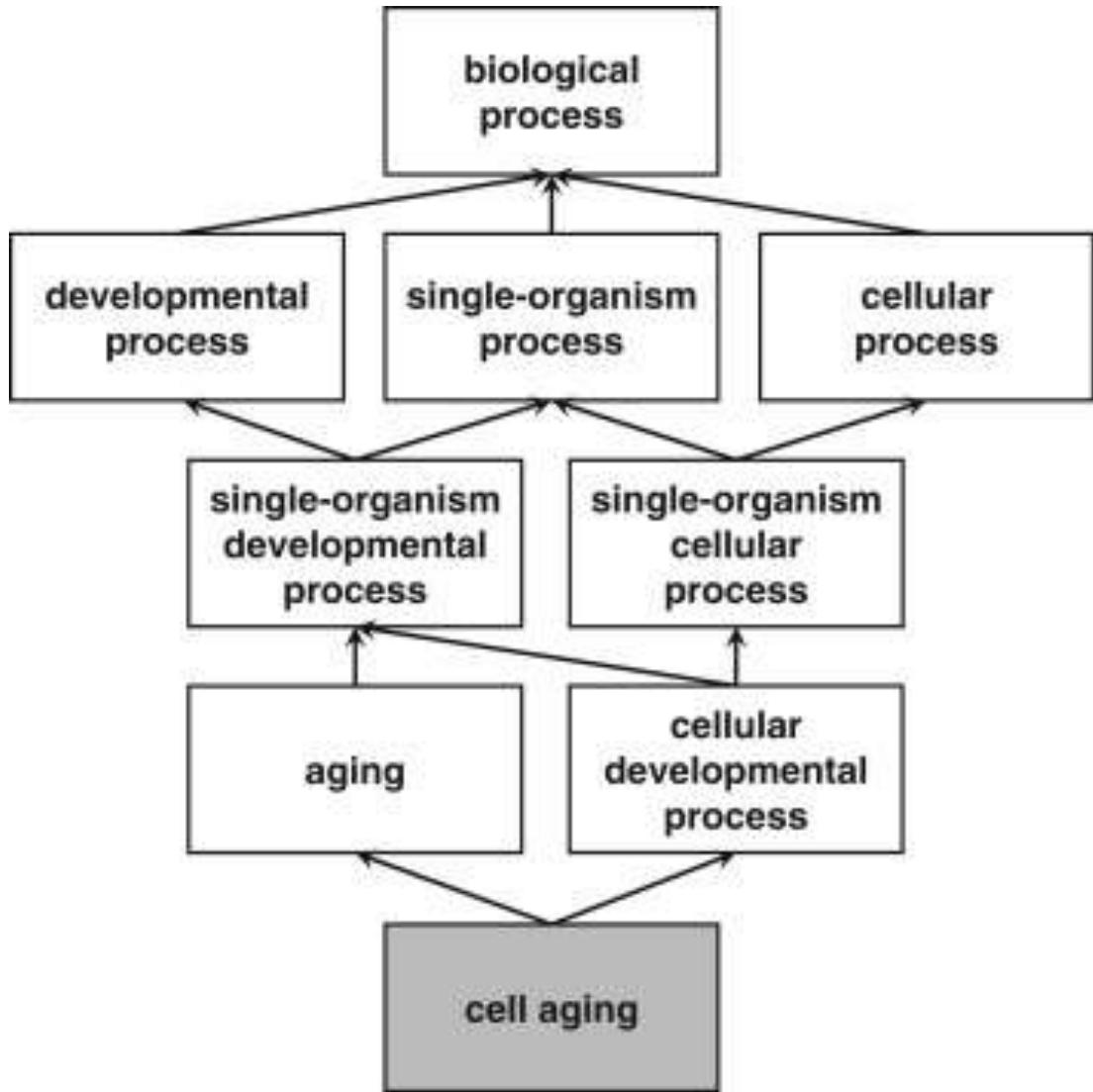
## Methods

- ORA
- GSEA
- SAFE
- PADOG
- ROAST
- CAMERA
- GSA
- GSVA/ssGSEA
- GlobelTest
- EBM
- MGSA
- GOSeq
- QUSAGE
- Pathview
- GOSemSim
- GGEA
- SPIA
- PathNet
- DEGraph
- TopologyGSA
- GANPA
- CePa
- NetGSA
- WGCNA

# Problems with databases: Low resolution

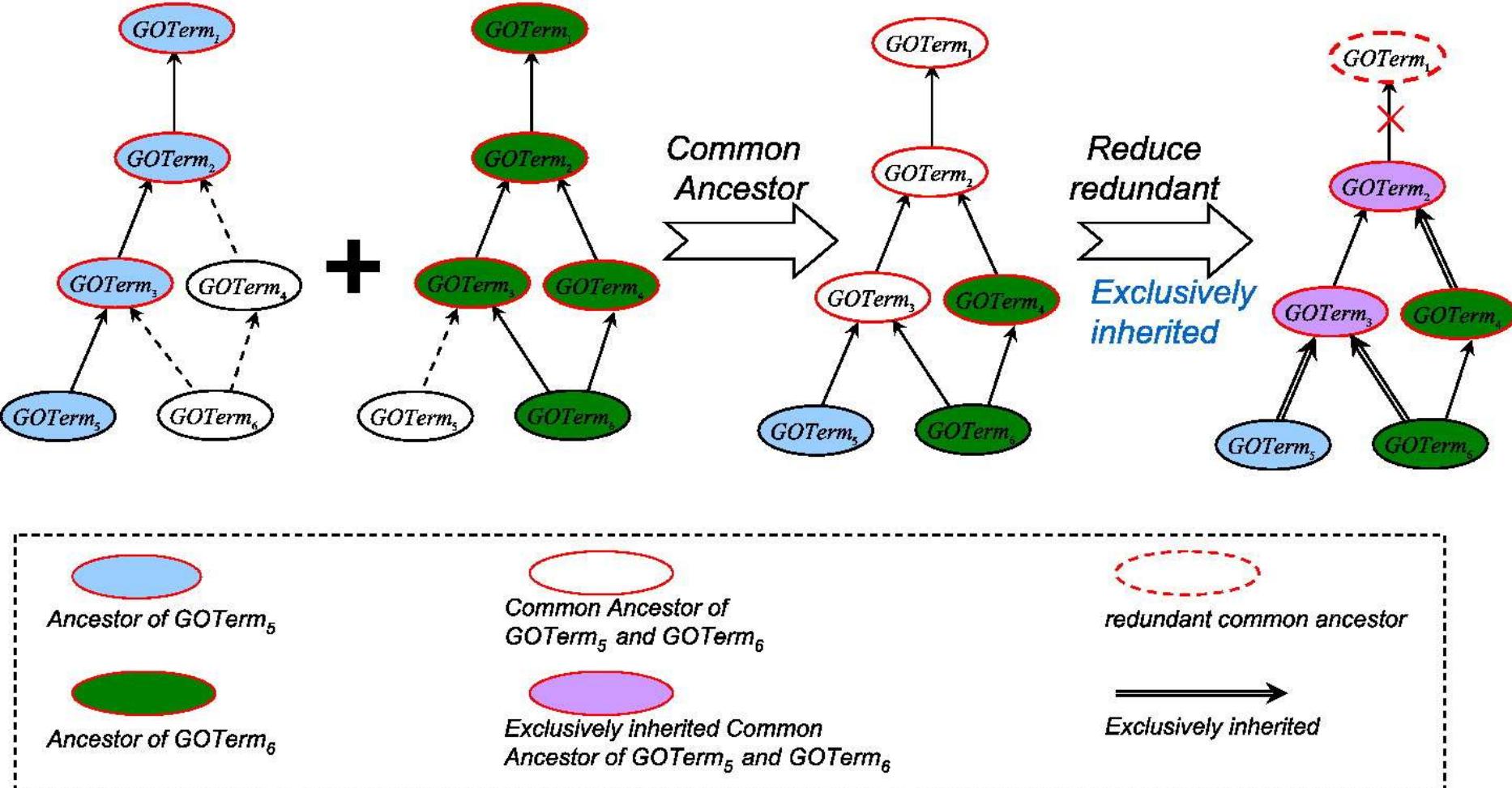
Databases and methods

# Gene Ontology: the world's largest source of information on the functions of genes



The GO contains many terms that are highly similar or overlapping in meaning (e.g., "cell cycle" and "mitosis").

# Semantic Similarity Measurement Based on *Exclusively Inherited* Shared Information for Gene Ontology



"exclusively inherited" refers to the subset of shared information that is **unique** to the two terms being compared ( $GOTerm_5$  and  $GOTerm_6$ ) and **not** inherited by other unrelated terms.

Illustration of Semantic Similarity Measurement for Gene Ontology Terms Using Exclusively Inherited Shared Information

# Making your own database

database\_seeds

\$paper1\_day1

Gene1, Gene2, Gene3, Gene4

\$paper2\_day2

Gene3, Gene4, Gene5, Gene6

# GREAT

## nature biotechnology

---

[nature](#) > [nature biotechnology](#) > [analyses](#) > [article](#)

Analysis | Published: 02 May 2010

### GREAT improves functional interpretation of *cis*-regulatory regions



Save



Related Papers



Chat with paper

[Cory Y McLean](#), [Dave Bristor](#), [Michael Hiller](#), [Shoa L Clarke](#), [Bruce T Schaar](#), [Craig B Lowe](#), [Aaron M Wenger](#) & [Gill Bejerano](#)

*Nature Biotechnology* **28**, 495–501 (2010)

**25k** Accesses | **30** Altmetric | [Metrics](#)

<https://www.nature.com/articles/nbt.1630>



# GREAT

- GREAT helps determine whether these regions are linked to **gene regulation**
- **Handles distal regulatory elements:** GREAT accounts for **long-range gene regulation**, making it more effective for studying enhancers and other non-coding regions
- **Chromatin conformation capture techniques** (e.g., Hi-C, ChIA-PET) that reveal **long-range interactions** between enhancers and promoters.
- **Epigenetic markers** such as histone modifications (e.g., H3K27ac) that indicate active regulatory regions.

# GREAT

It mostly works with the mouse and human genome

Annotations are not open access

## **rGREAT: an R/bioconductor package for functional enrichment on genomic regions**

[!\[\]\(a2f9594c2c856a03df90ec4016df4a10\_img.jpg\) Save](#)[!\[\]\(c7774dea93eb10ead3ed0542c77a8534\_img.jpg\) Related Papers](#)[!\[\]\(f15da8627380db409bac161a6cb03047\_img.jpg\) Chat with paper](#)

Zuguang Gu , Daniel Hübschmann 

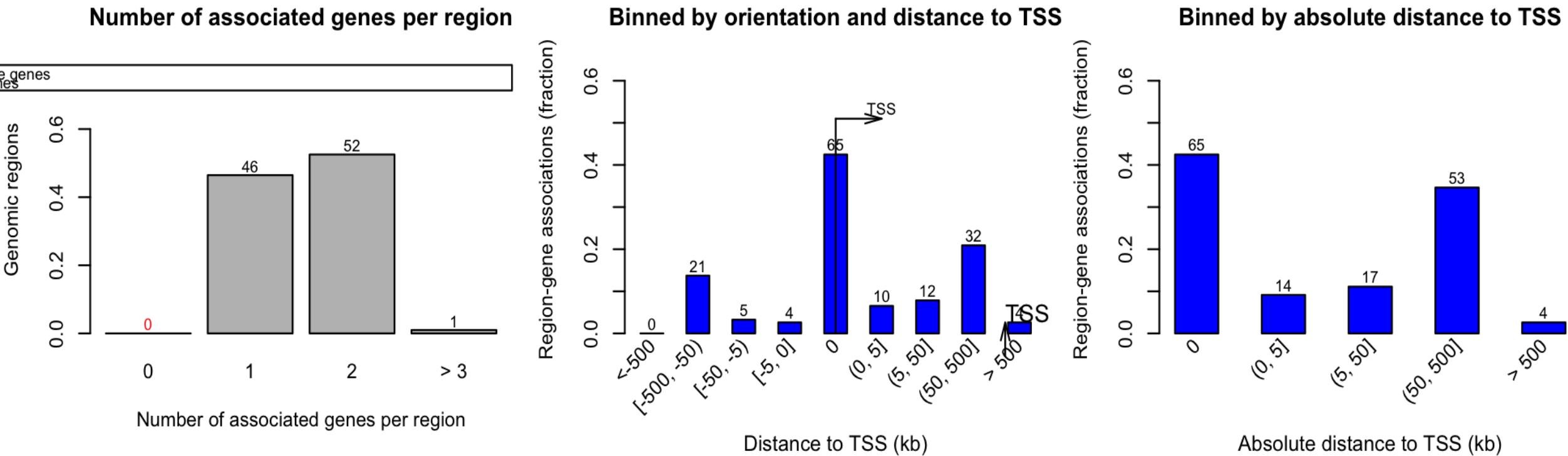
*Bioinformatics*, Volume 39, Issue 1, January 2023, btac745,

<https://doi.org/10.1093/bioinformatics/btac745>

**Published:** 17 November 2022

**Article history ▾**

# Example plot



# Summary

Three types of methods for enrichment analysis:

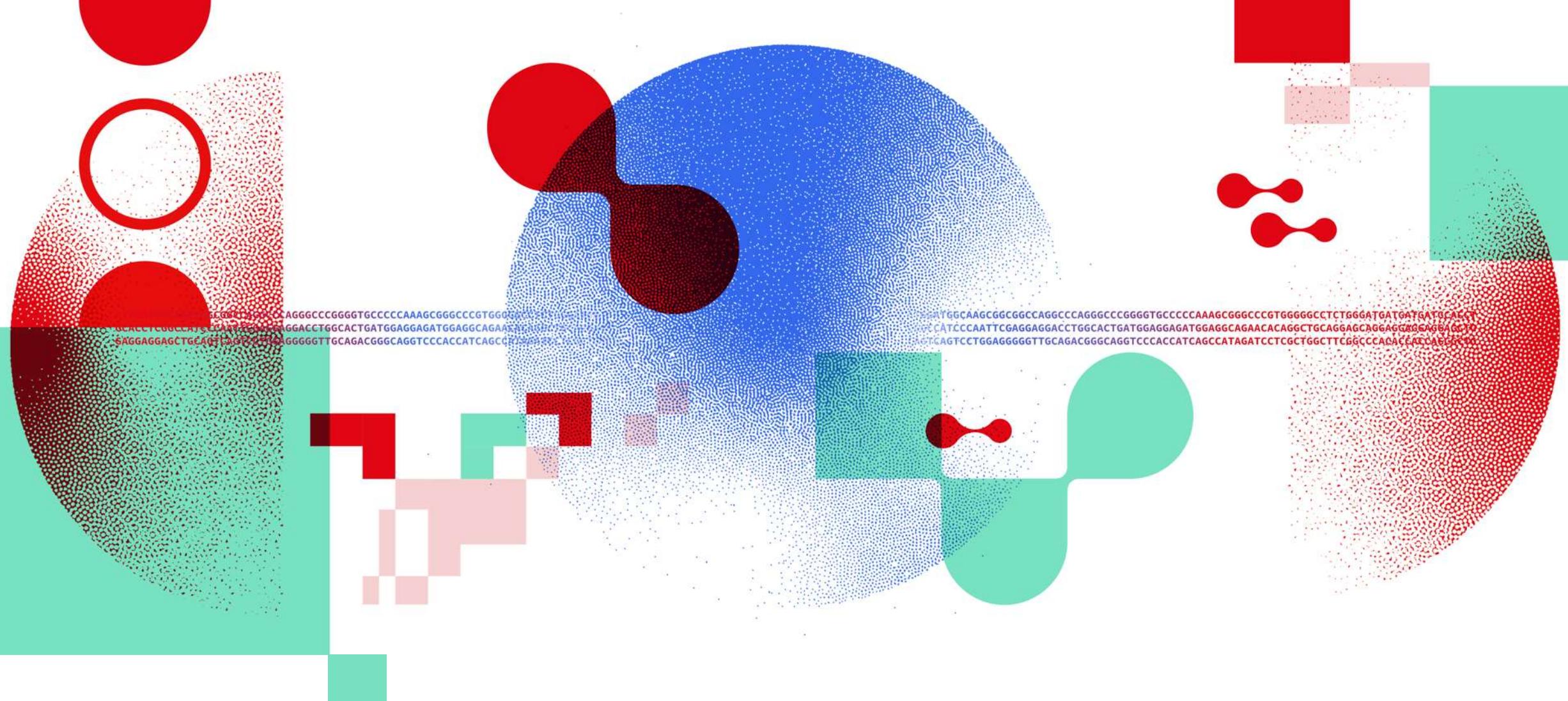
1. ORA
2. FCS
3. Pathway Topology

Databases problem

GO semantic similarity

GREAT/ rGREAT for analysis of Genomic Ranges

# Exercise 7



# Thank you

DATA SCIENTISTS FOR LIFE  
[sib.swiss](http://sib.swiss)