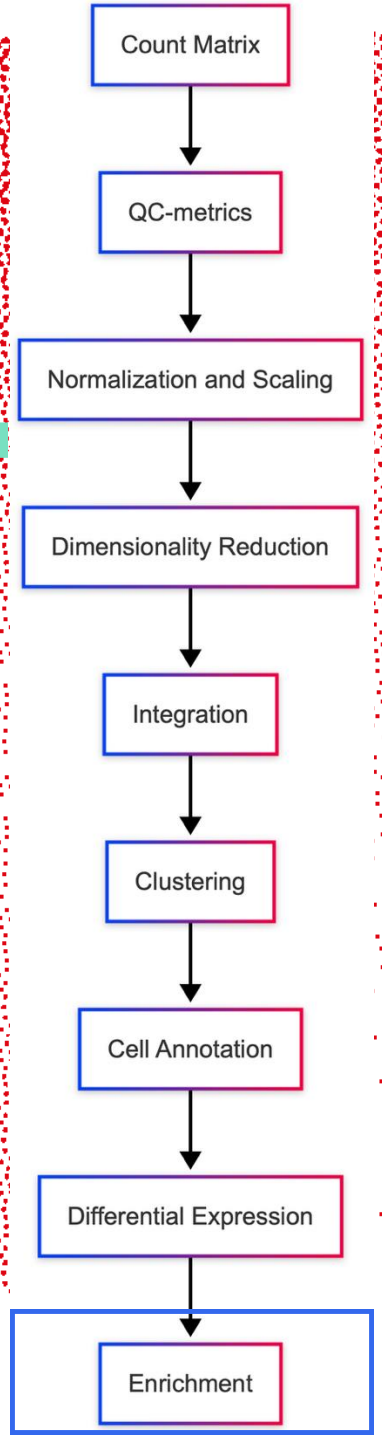
Swiss Institute of Bioinformatics

SINGLE-CELL TRANSCRIPTOMICS WITH R

Enrichment analysis



Deepak Tanwar

July 02-04 2025

Adapted from previous year courses

elixir SWITZERLAND

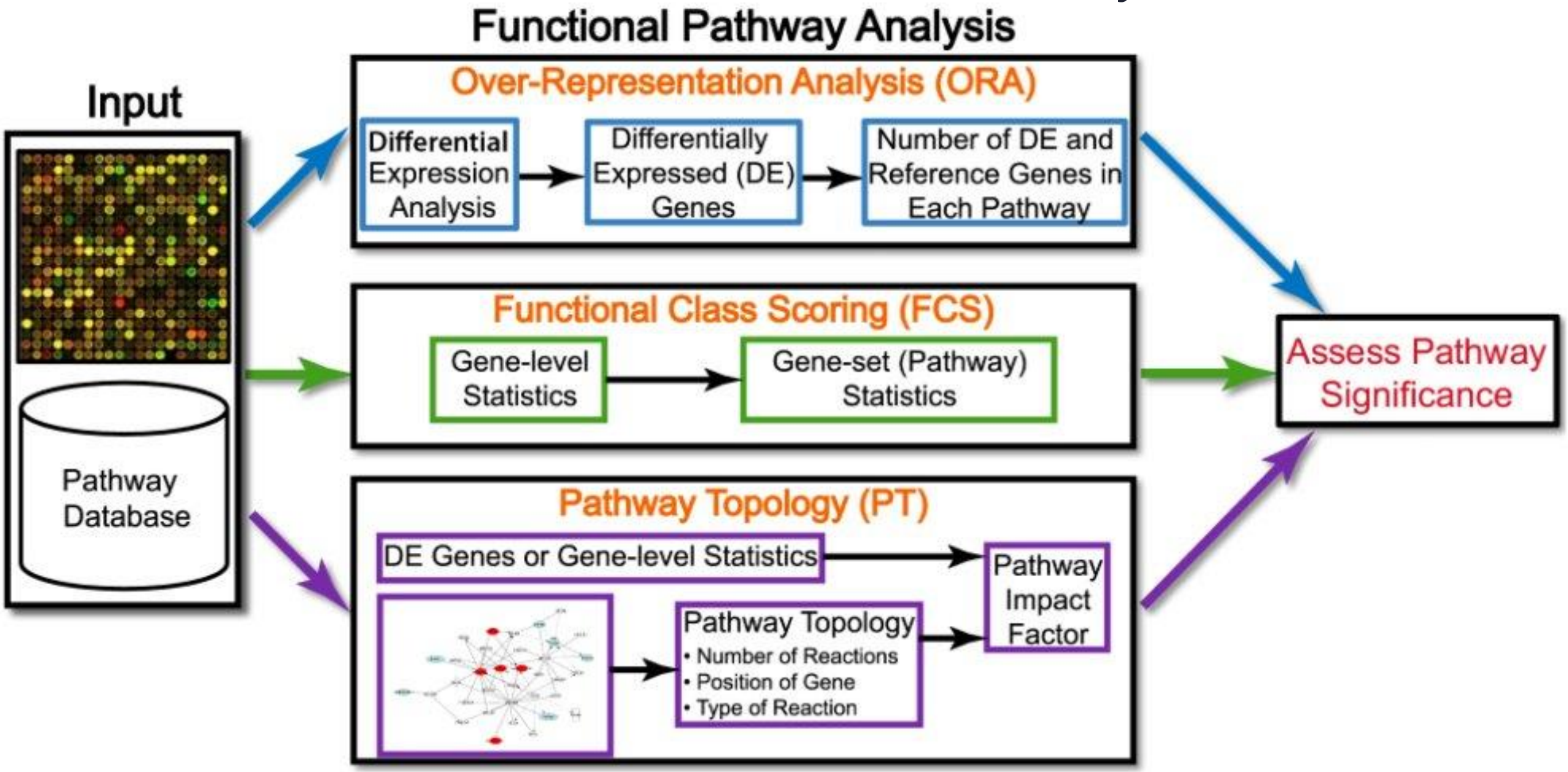
Learning objectives

What is Enrichment analysis?

Distinguish between different ways to do it.

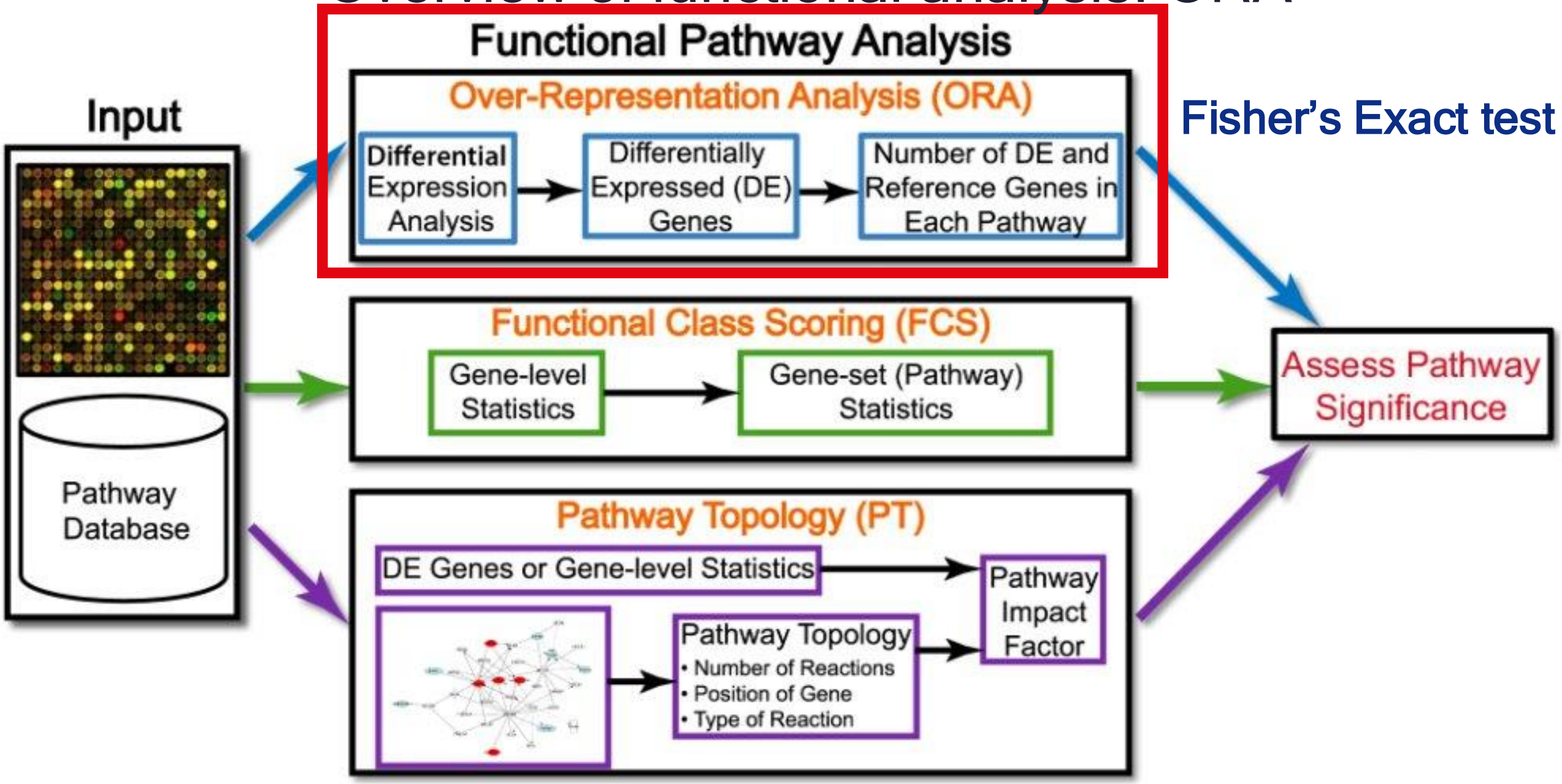
Challenges and Limitations of methods.

Overview of functional analysis



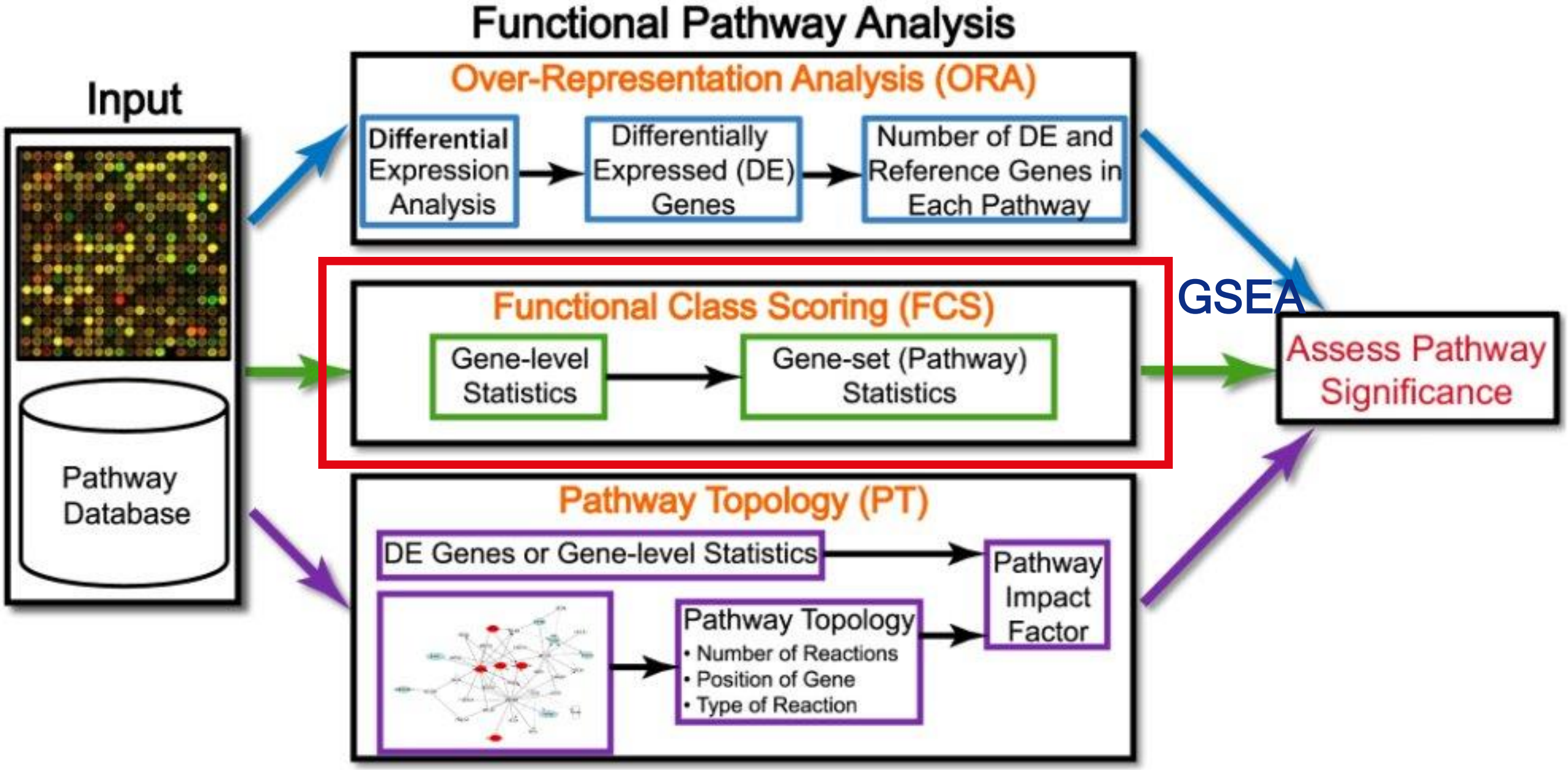
https://doi.org/10.1371/journal.pcbi.1002375

Overview of functional analysis: ORA



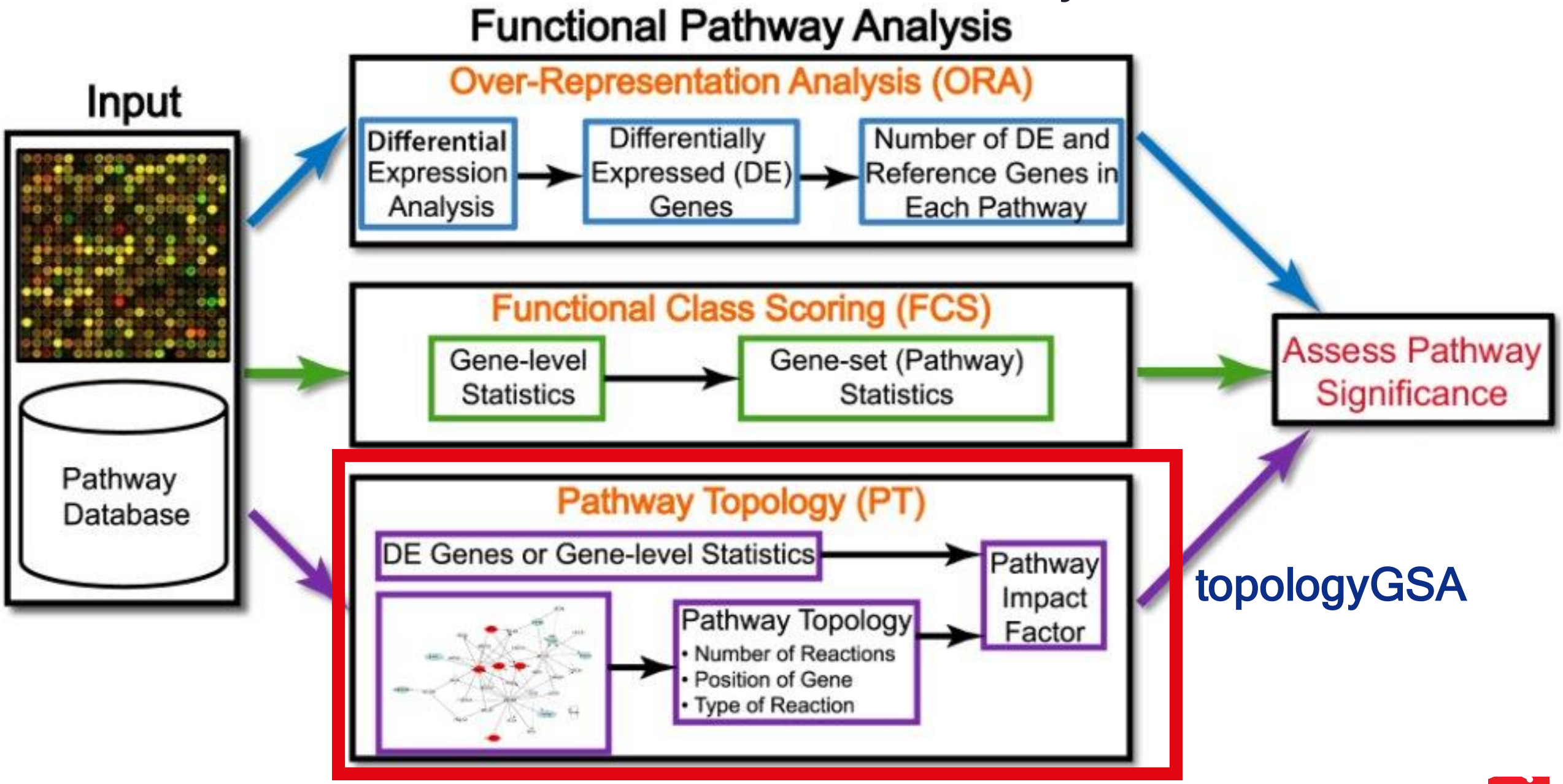
https://doi.org/10.1371/journal.pcbi.1002375

Overview of functional analysis: FCS



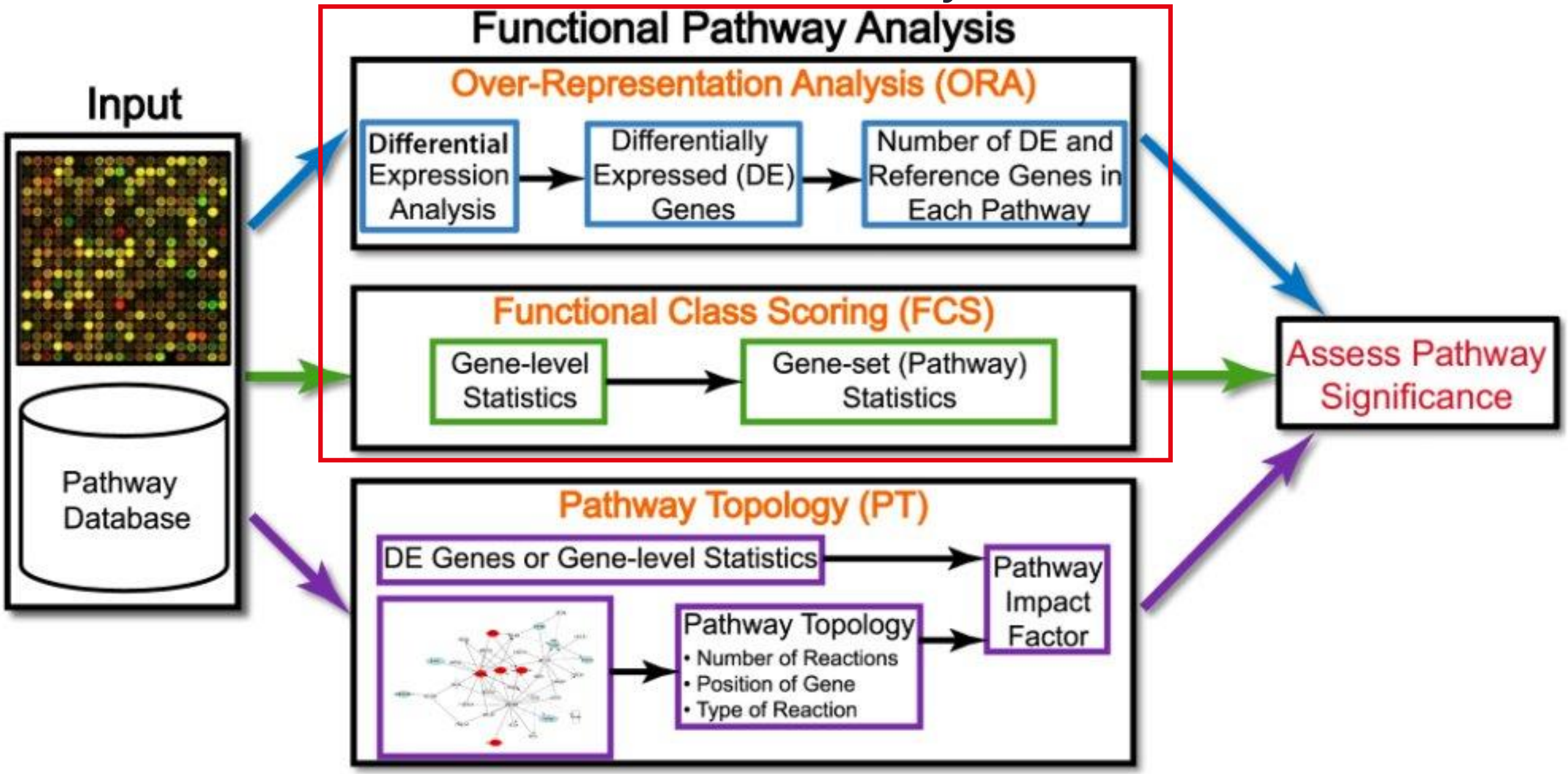
https://doi.org/10.1371/journal.pcbi.1002375

Overview of functional analysis: PT



https://doi.org/10.1371/journal.pcbi.1002375

Overview of functional analysis: ORA & FCS



https://doi.org/10.1371/journal.pcbi.1002375

Goal: To gain biologically meaningful insights from long gene lists

Over-representation analysis (ORA)

Over-representation analysis (ORA)

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

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 <= 0.05)

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 <= 0.05)
2. For each pathway, count input genes that are part of the pathway

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 <= 0.05)
2. For each pathway, count input genes that are part of the pathway
3. Repeat for an appropriate background list of genes

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 <= 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

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 <= 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 <= 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)

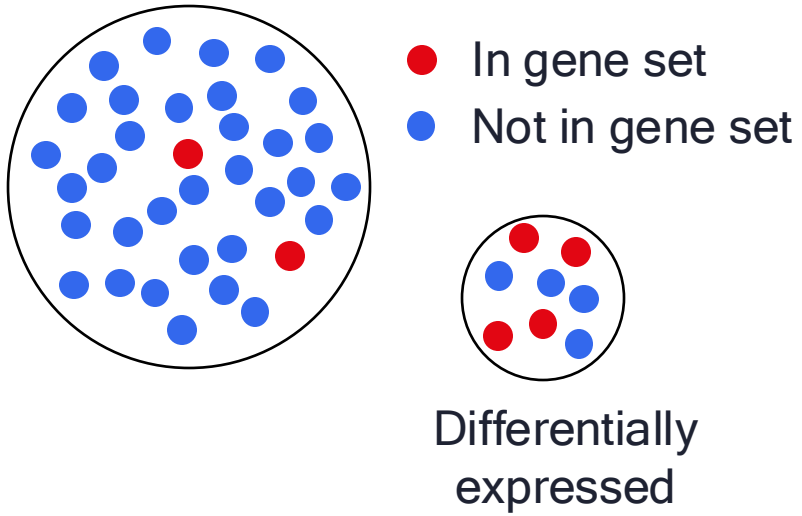
Gene1	0.051
Gene2	0.05001
Gene 3	0.049
Gene 4	0.001
Gene 5	0.023
Gene 6	0.04
Gene 7	0.01
Gene 8	0.0501
Gene 9	0.2
Gene 10	0.051
Gene 11	0.05
Gene 12	0.49
Gene 13	0.03
Gene 14	0.01
Gene 15	0.052
Gene 16	0.9

pvalue <= 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

Fisher's test



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?

Problems with ORA

Cutoff? 0.051?

Treat all genes equally

Problems with ORA

Cutoff? 0.051?

Treat all genes equally

Each gene is independent of other

Problems with ORA

Cutoff? 0.051?

Treat all genes equally

Each gene is independent of other

Each pathway is independent of each other

Over-representation analysis (ORA)

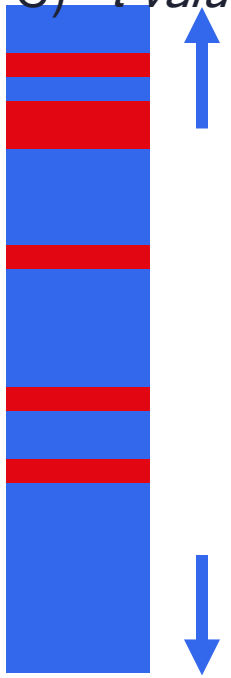
Gene1	0.051	10
Gene2	0.05001	12
Gene 3	0.049	11
Gene 4	0.001	8
Gene 5	0.023	2
Gene 6	0.04	3
Gene 7	0.01	1
Gene 8	0.0501	3
Gene 9	0.2	-10
Gene 10	0.051	-3
Gene 11	0.05	-8
Gene 12	0.49	-19
Gene 13	0.03	-3
Gene 14	0.01	-2
Gene 15	0.052	-1
Gene 16	0.9	-4

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
log2(FC) * *t-value*



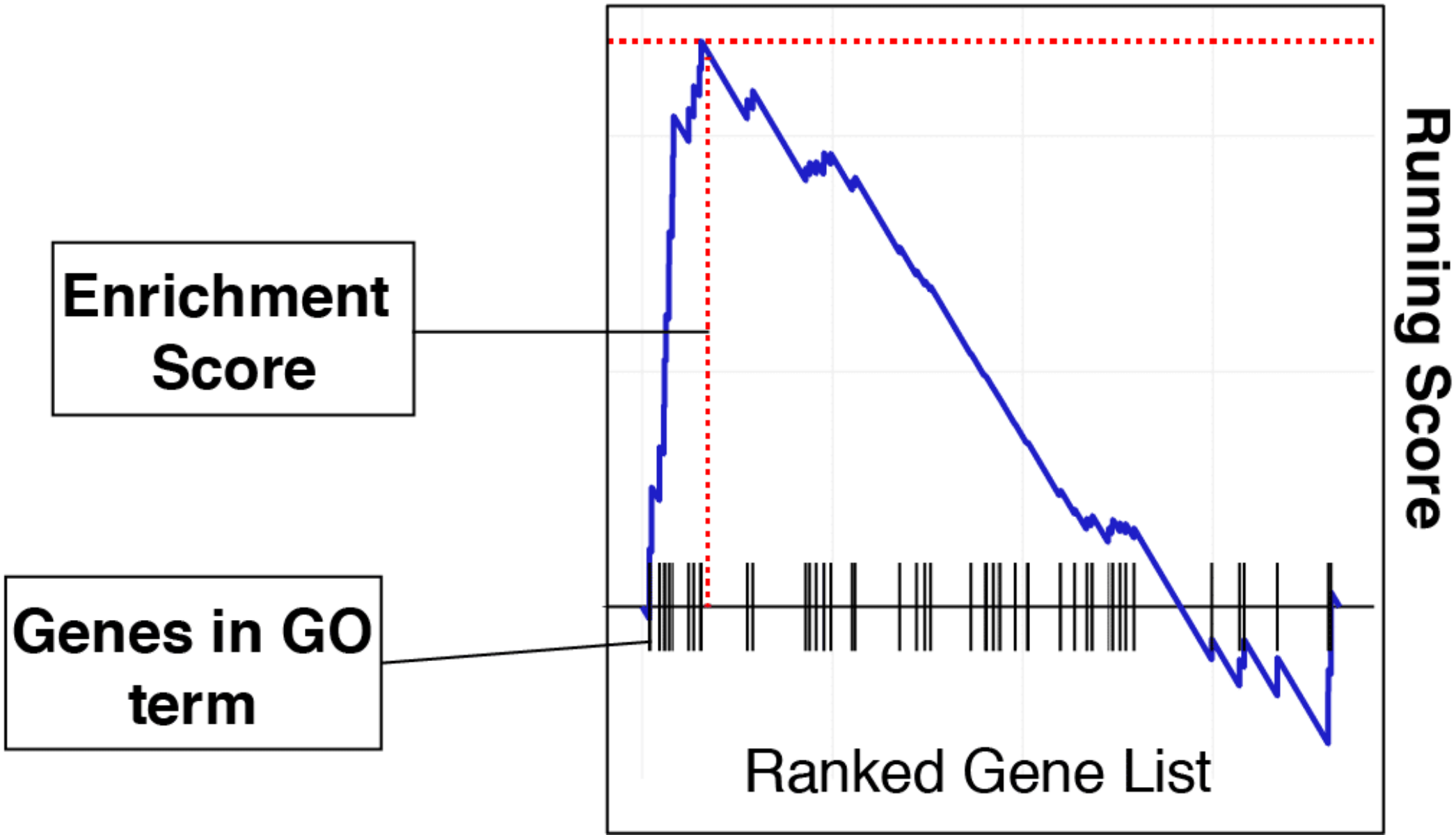
Upregulated

H_0: Genes in set are randomly distributed over ranked list

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

Downregulated

Functional class scoring (FCS)



Problems with FCS

Each gene is independent of other

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

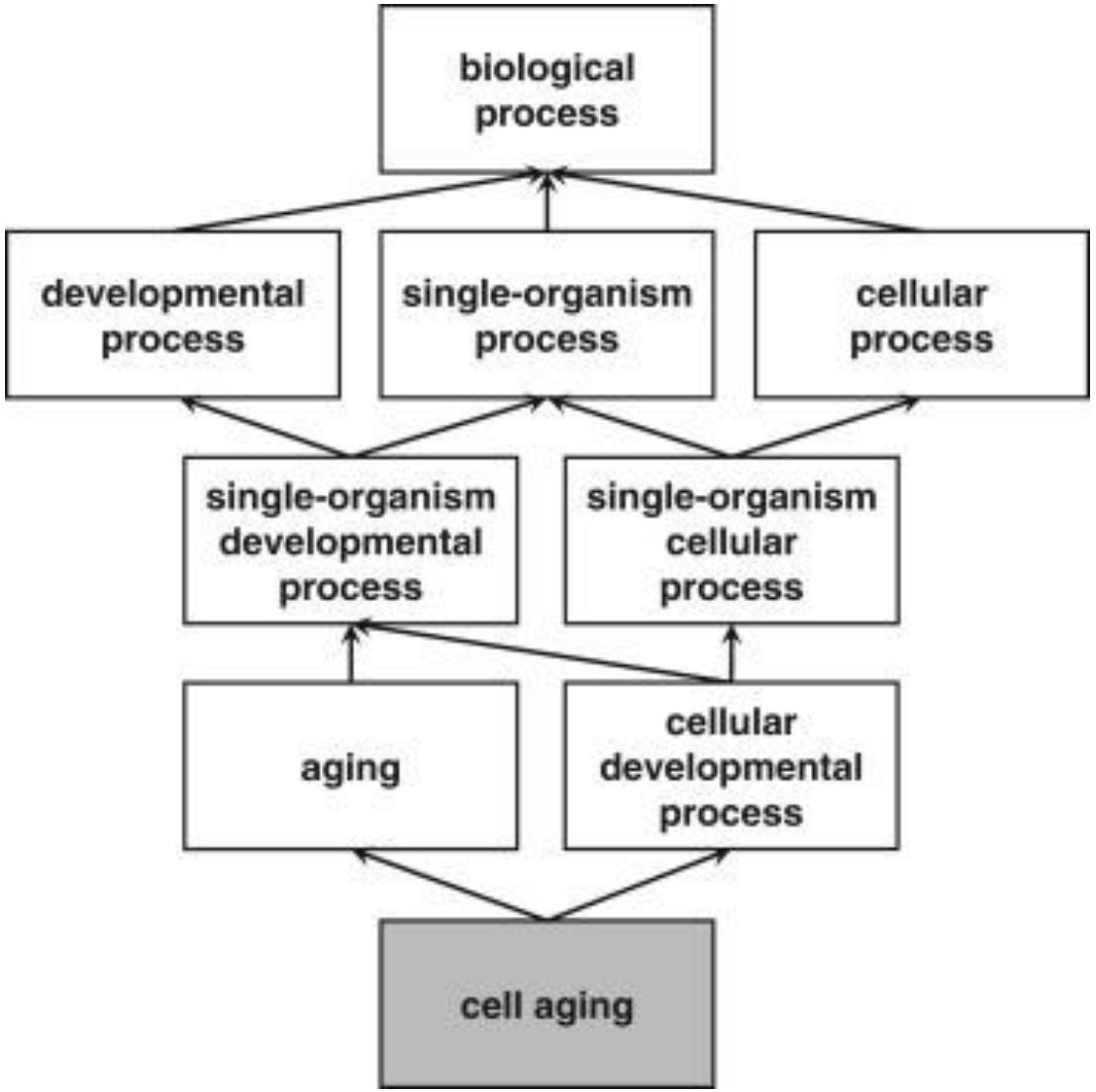
- ORA
- GSEA
- SAFE
- PADOG
- ROAST
- CAMERA
- GSA
- GSVA/ssGSEA
- GlobelTest
- EBM
- MGSA
- GOSeq
- QUSAGE
- Pathview
- GOSemSim
- GGEA
- SPIA
- PathNet
- DEGraph
- TopologyGSA
- GANPA
- CePa
- NetGSA
- WGCNA

Problems with databases:

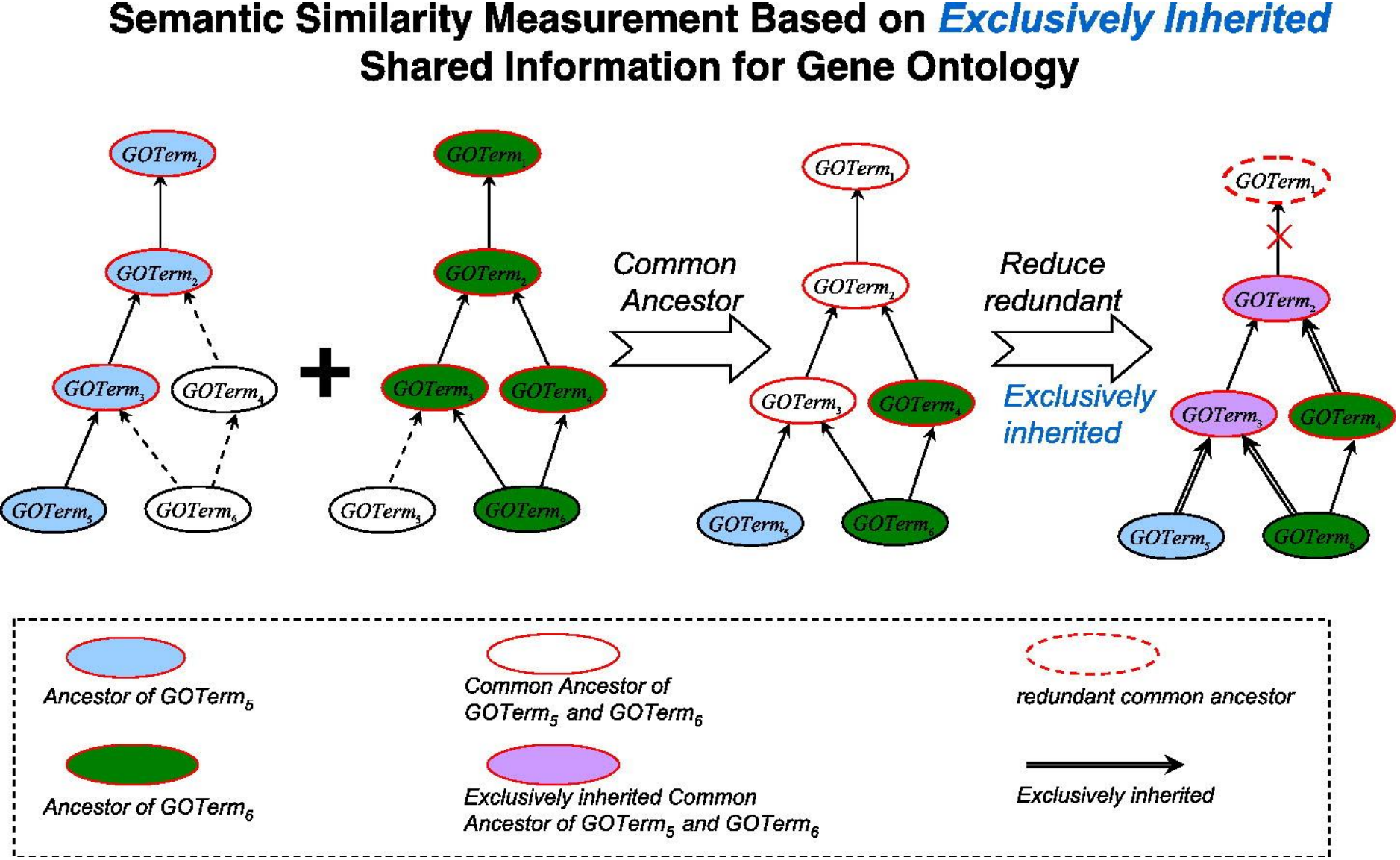
Low resolution

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").



"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

Quiz

1. Single cell-level pathway analysis can provide insights into cell-to-cell variability in pathway activity, while pseudo-bulk analysis cannot.

A) True

B) False

2. Using "exclusively inherited" shared information in semantic similarity calculations helps reduce the impact of redundant GO terms.

A) True

B) False

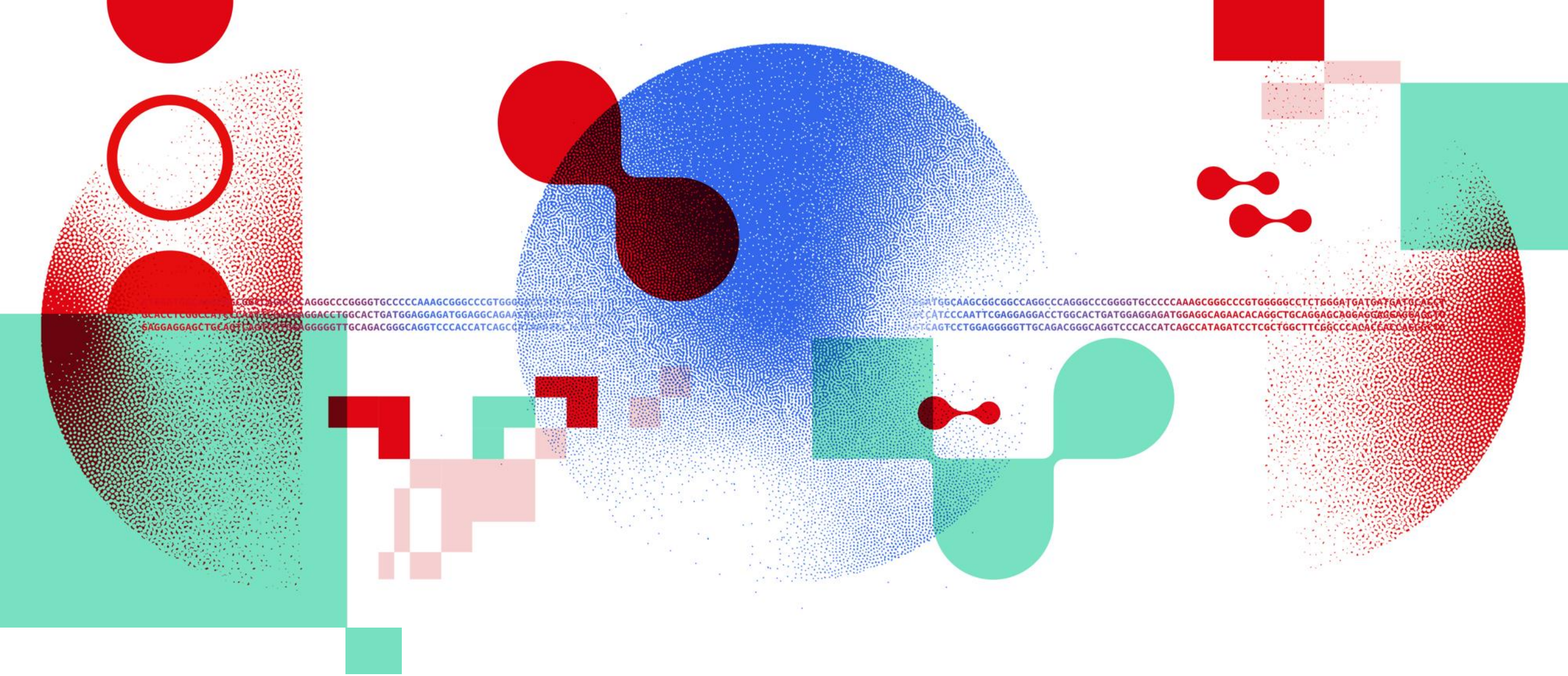
Summary

Three types of methods for enrichment analysis:

1. ORA
2. FCS
3. Pathway Topology

Databases problem

GO semantic similarity



Thank you

DATA SCIENTISTS FOR LIFE

sib.swiss