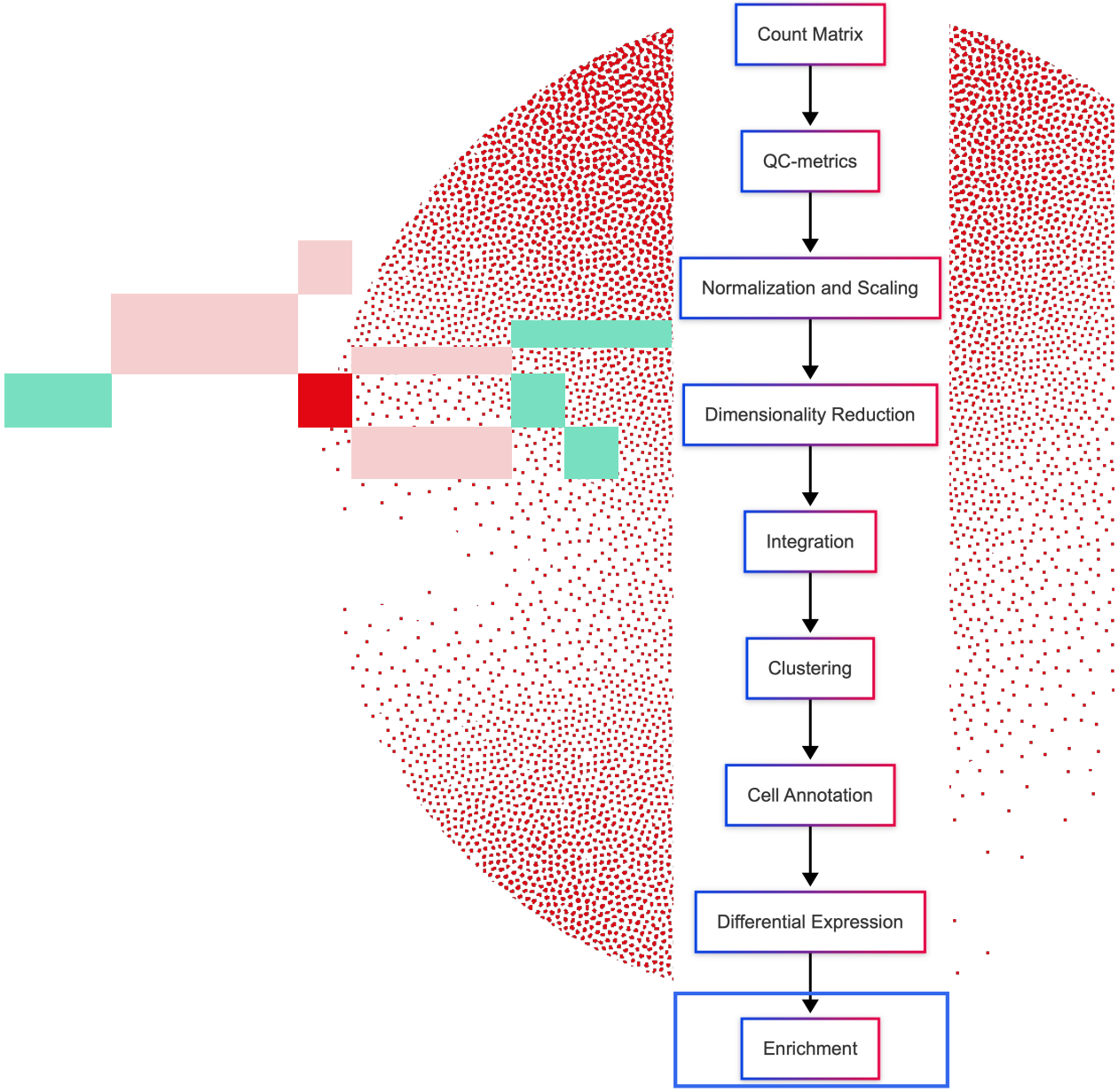
Swiss Institute of Bioinformatics

SINGLE-CELL TRANSCRIPTOMICS WITH R

Enrichment analysis



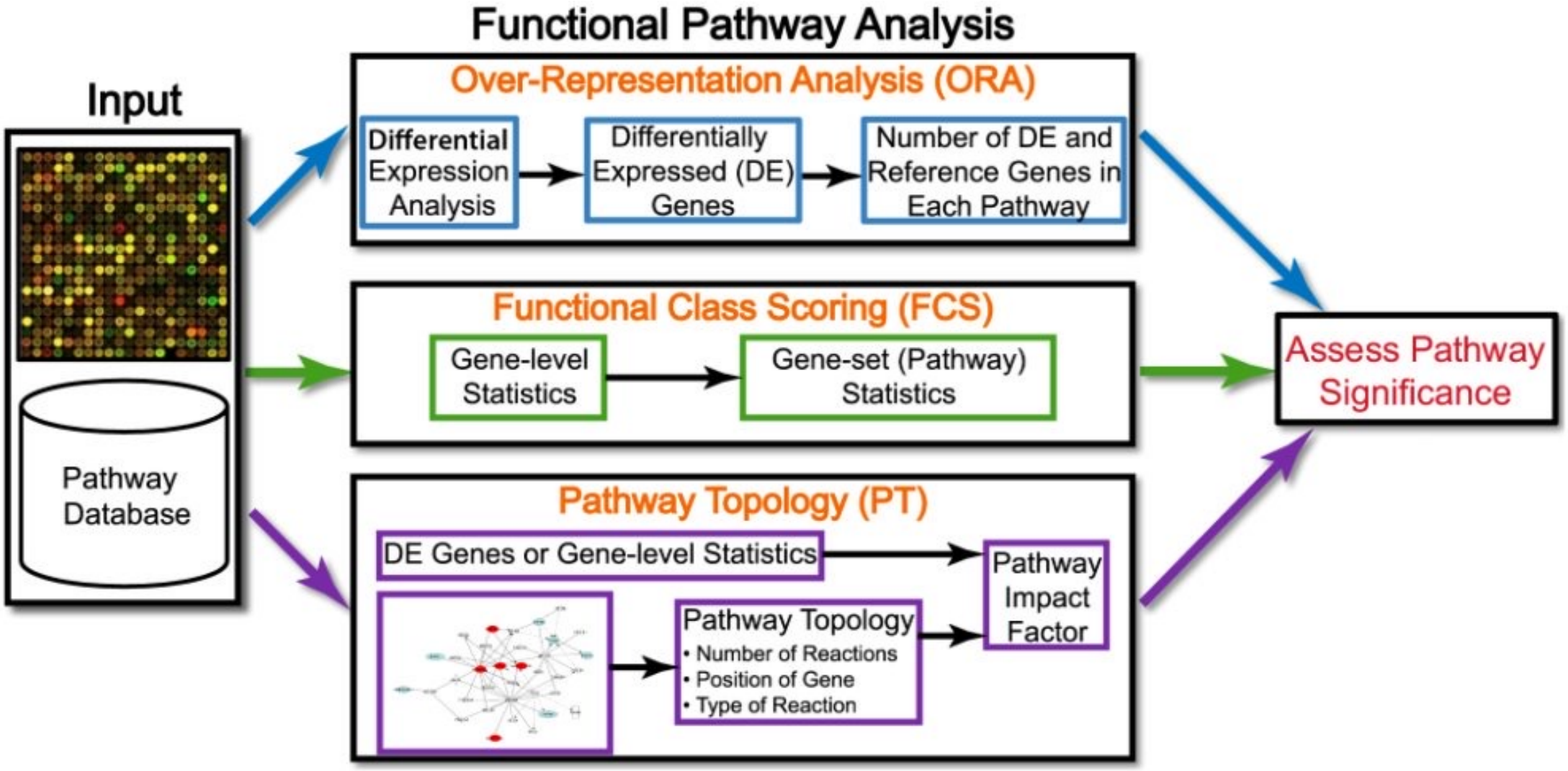
Deepak Tanwar

March 18-20, 2025

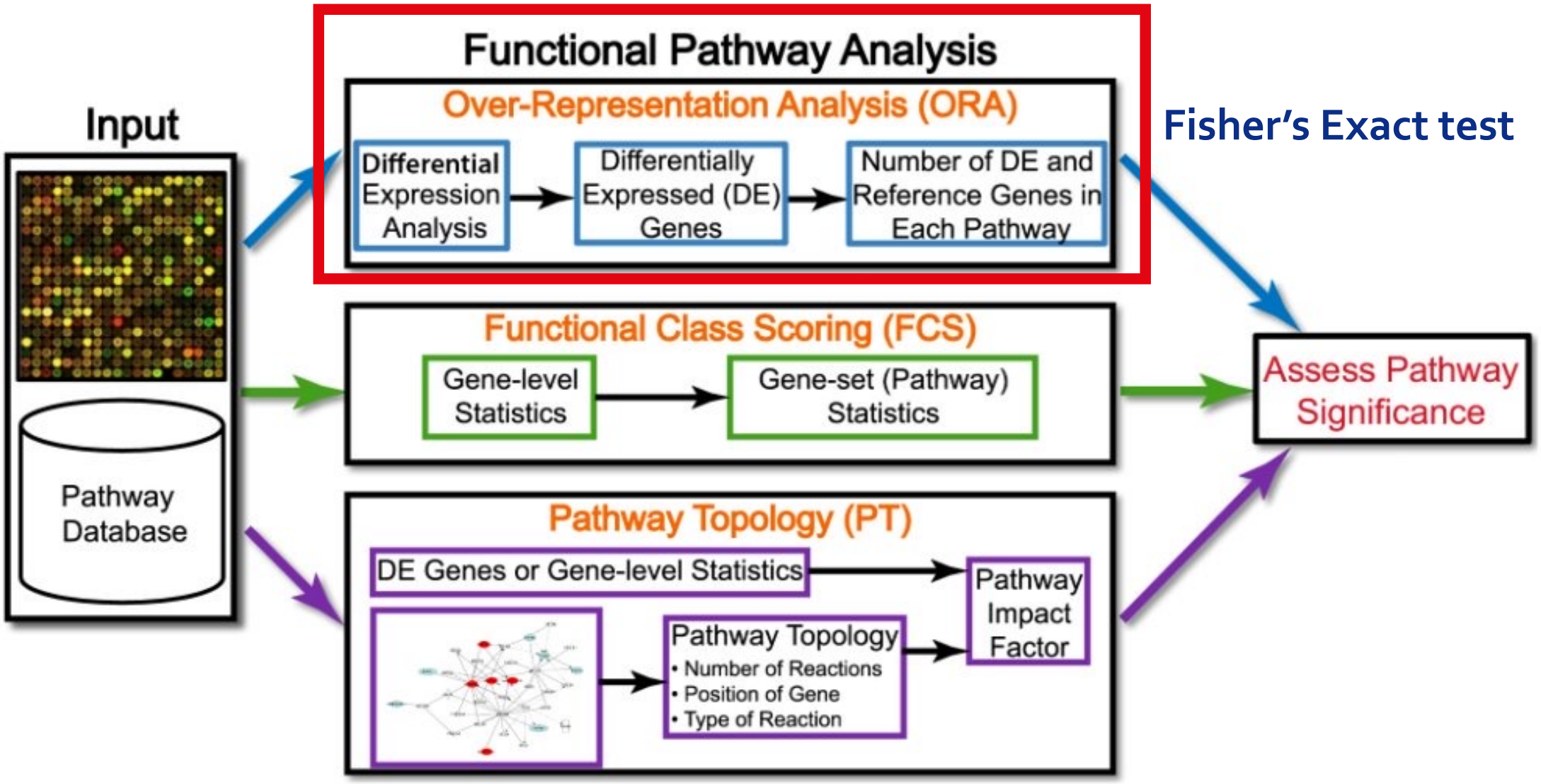
Adapted from previous year courses

Feedback from Geert van Geest

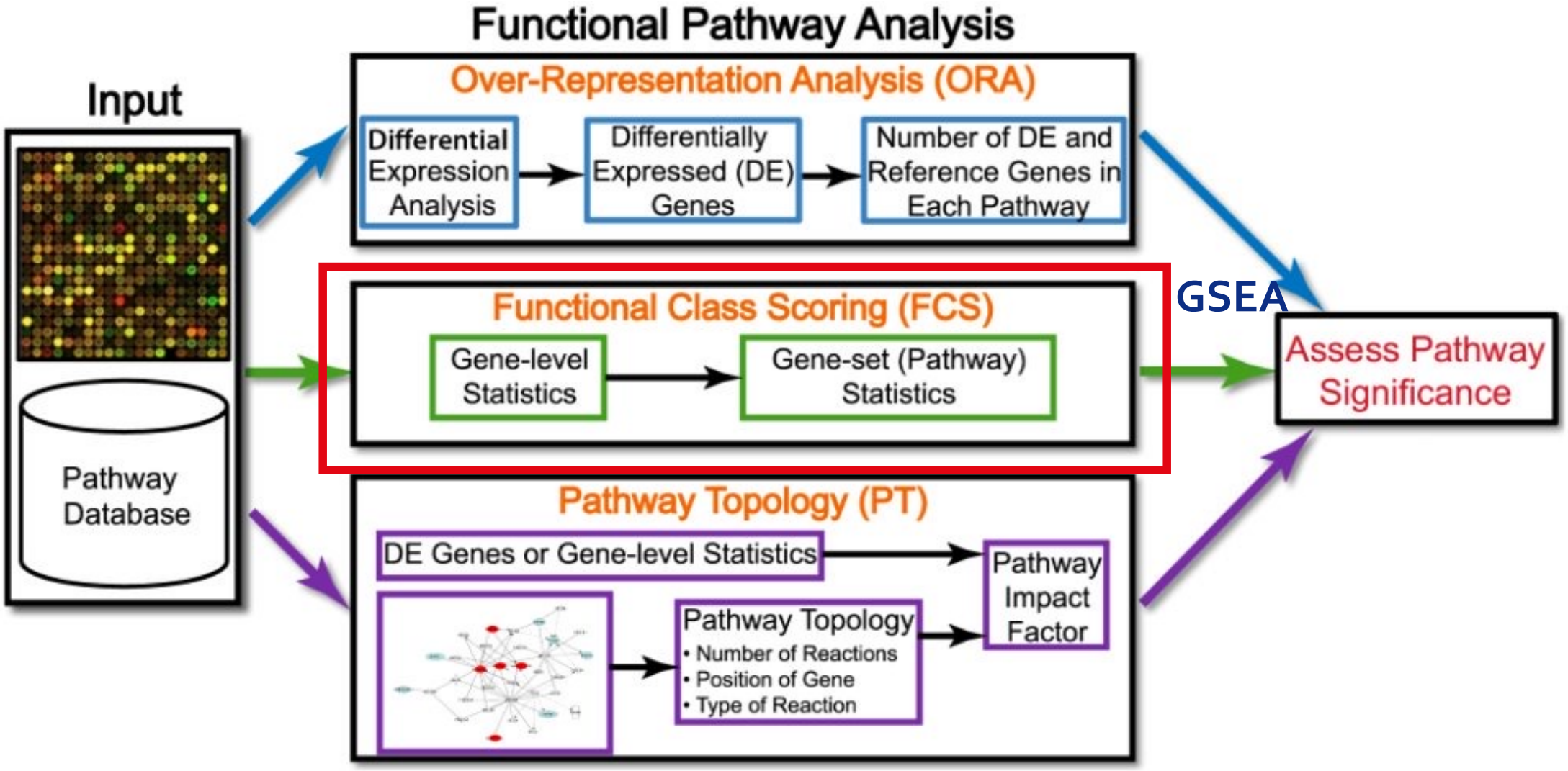
elixir SWITZERLAND



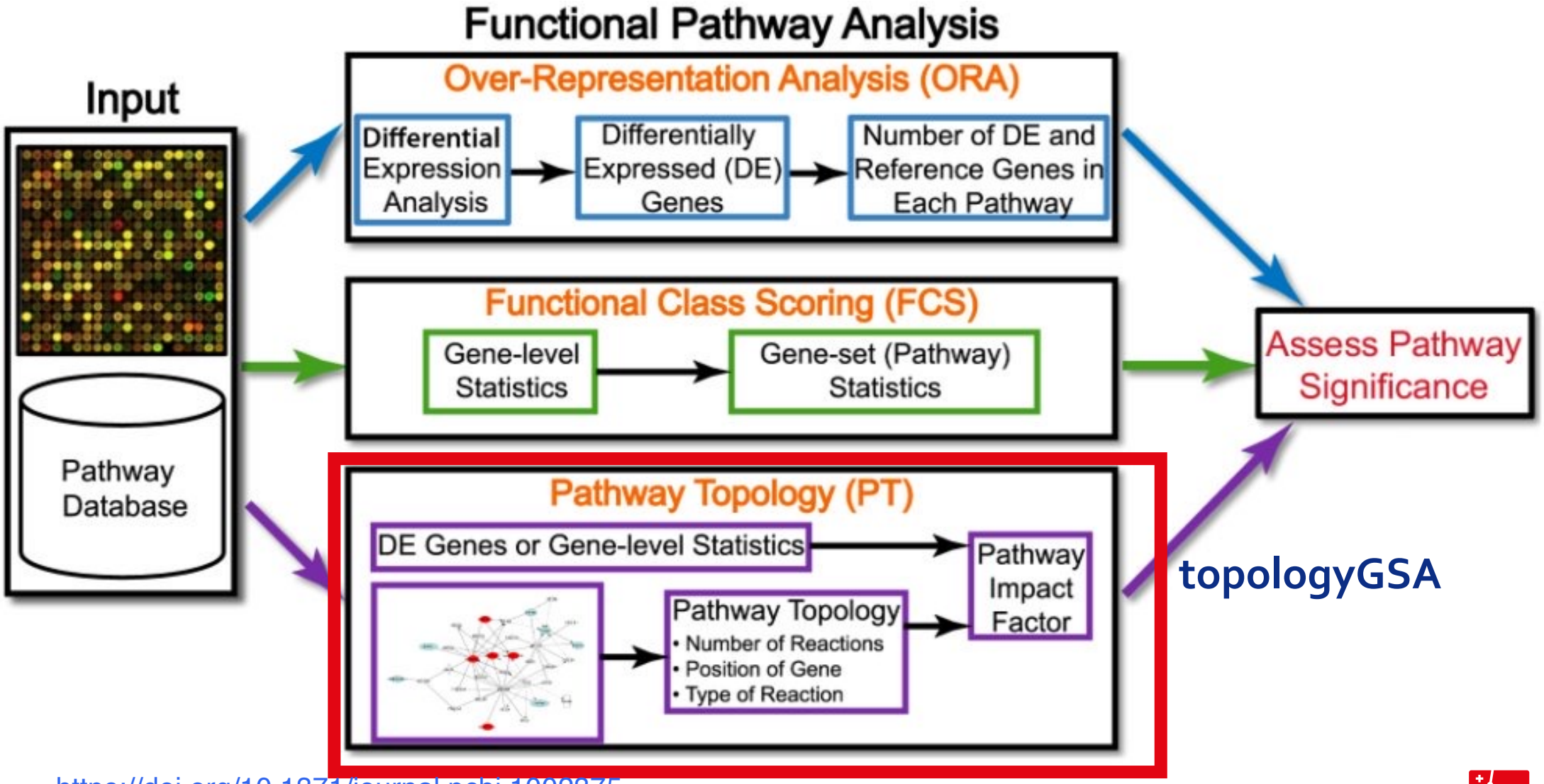
https://doi.org/10.1371/journal.pcbi.1002375



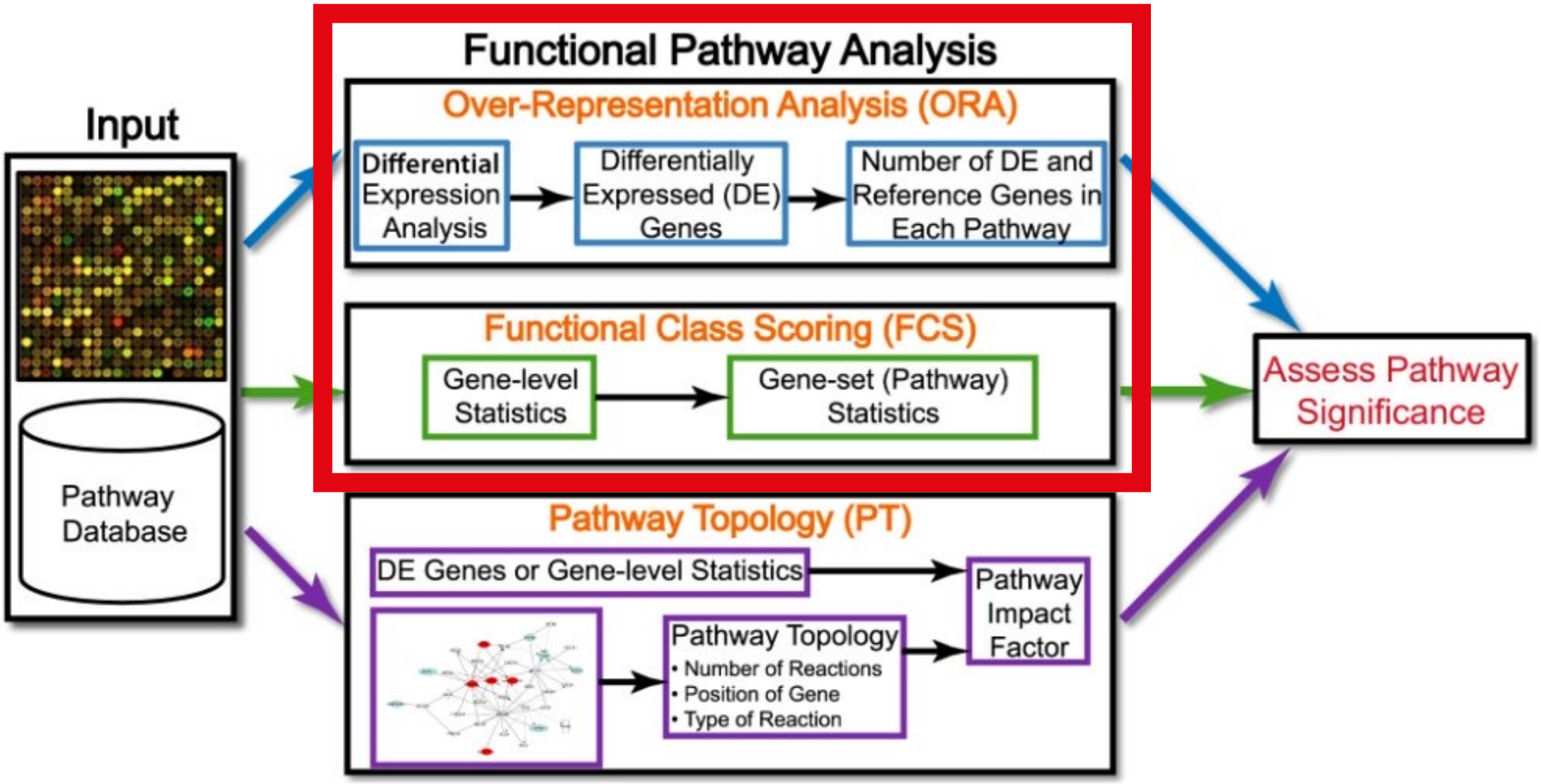
https://doi.org/10.1371/journal.pcbi.1002375



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



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



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

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 <= 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)

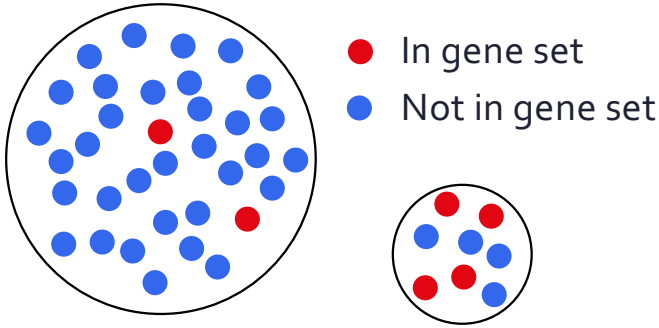
Gene1	0.051
Gene2	0.05001
Gene 3	0.049
Gene 4	0.001
Gene 5	0.023
Gene 6	0.04
Gene 7	0.01
Gene 8	0.0501
Gene 9	0.2
Gene 10	0.051
Gene 11	0.05
Gene 12	0.49
Gene 13	0.03
Gene 14	0.01
Gene 15	0.052
Gene 16	0.9

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_o: 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)

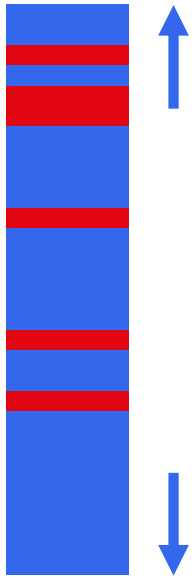
Gene1	0.051	10
Gene2	0.05001	12
Gene 3	0.049	11
Gene 4	0.001	8
Gene 5	0.023	2
Gene 6	0.04	3
Gene 7	0.01	1
Gene 8	0.0501	3
Gene 9	0.2	-10
Gene 10	0.051	-3
Gene 11	0.05	-8
Gene 12	0.49	-19
Gene 13	0.03	-3
Gene 14	0.01	-2
Gene 15	0.052	-1
Gene 16	0.9	-4

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 log2(FC) * *t-value*



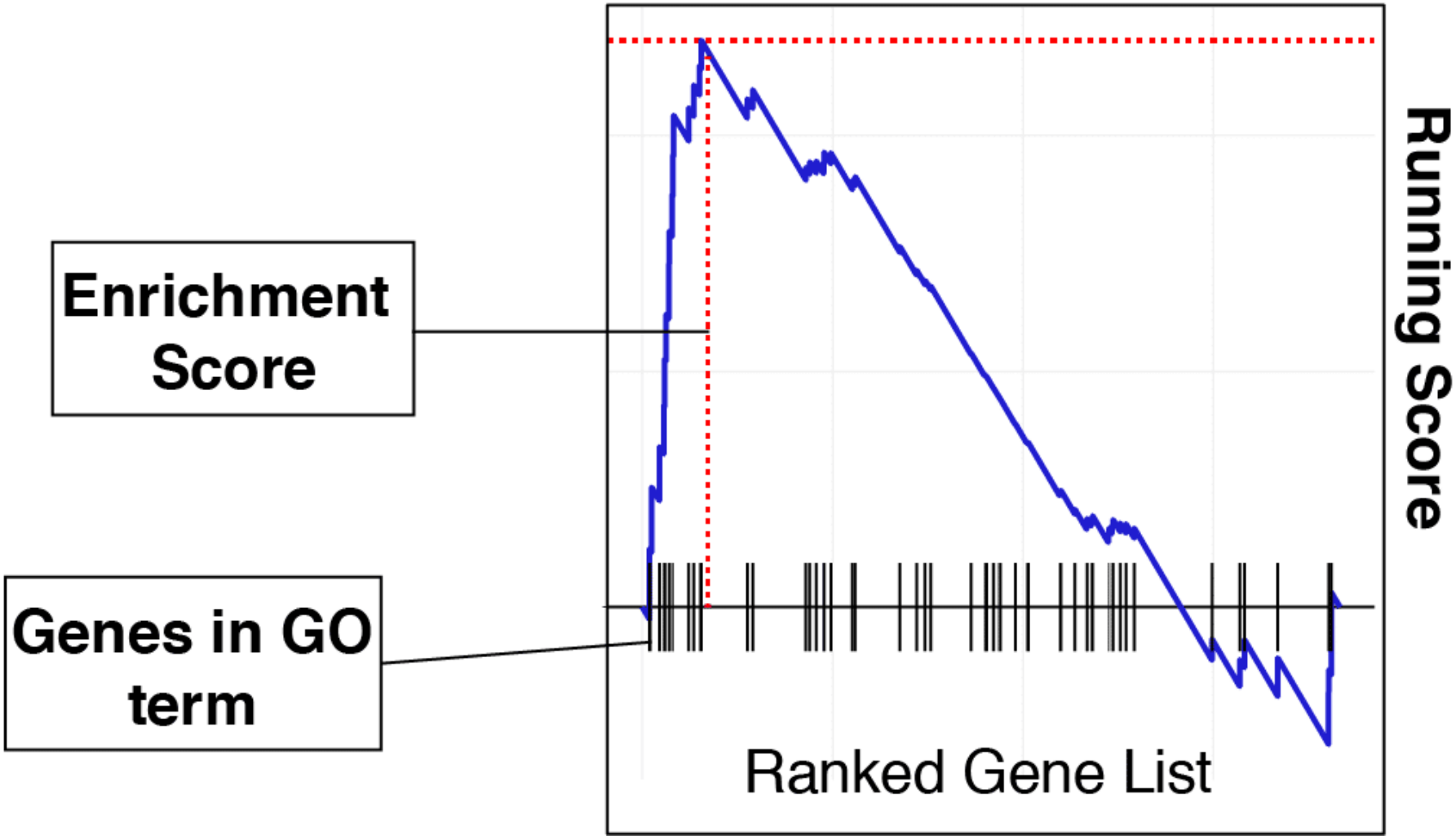
Upregulated

Downregulated

H_o: 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

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

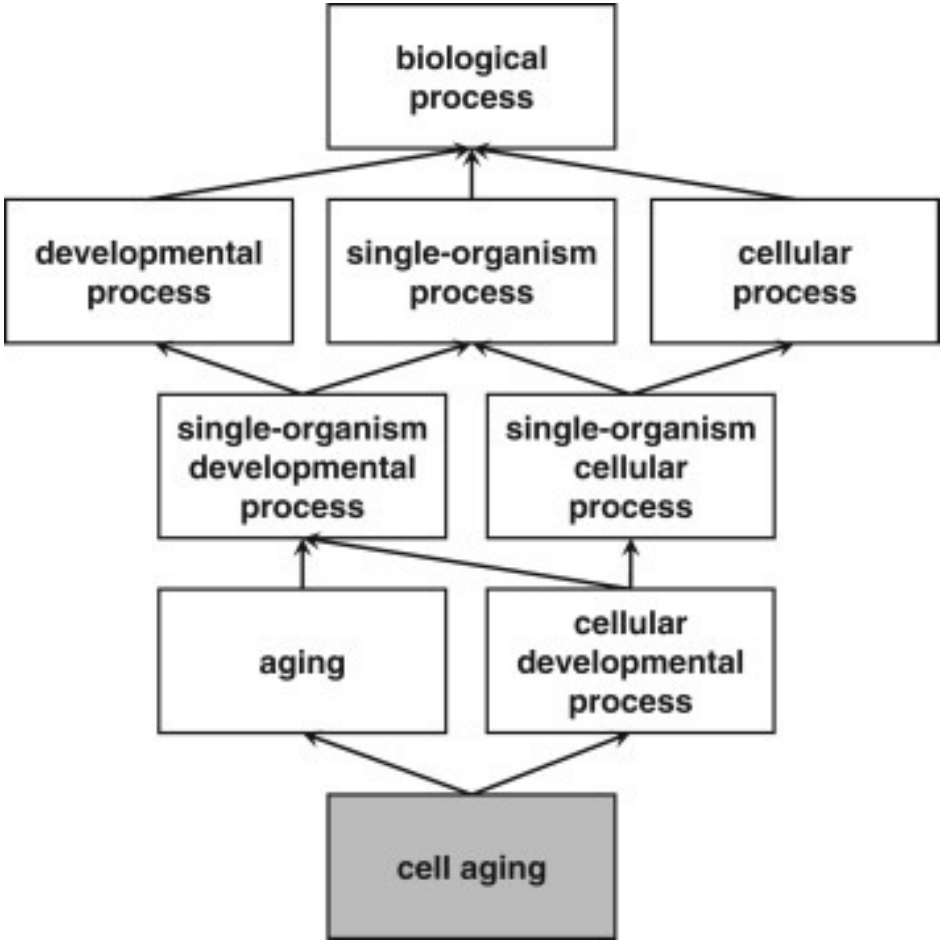
Methods

- 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

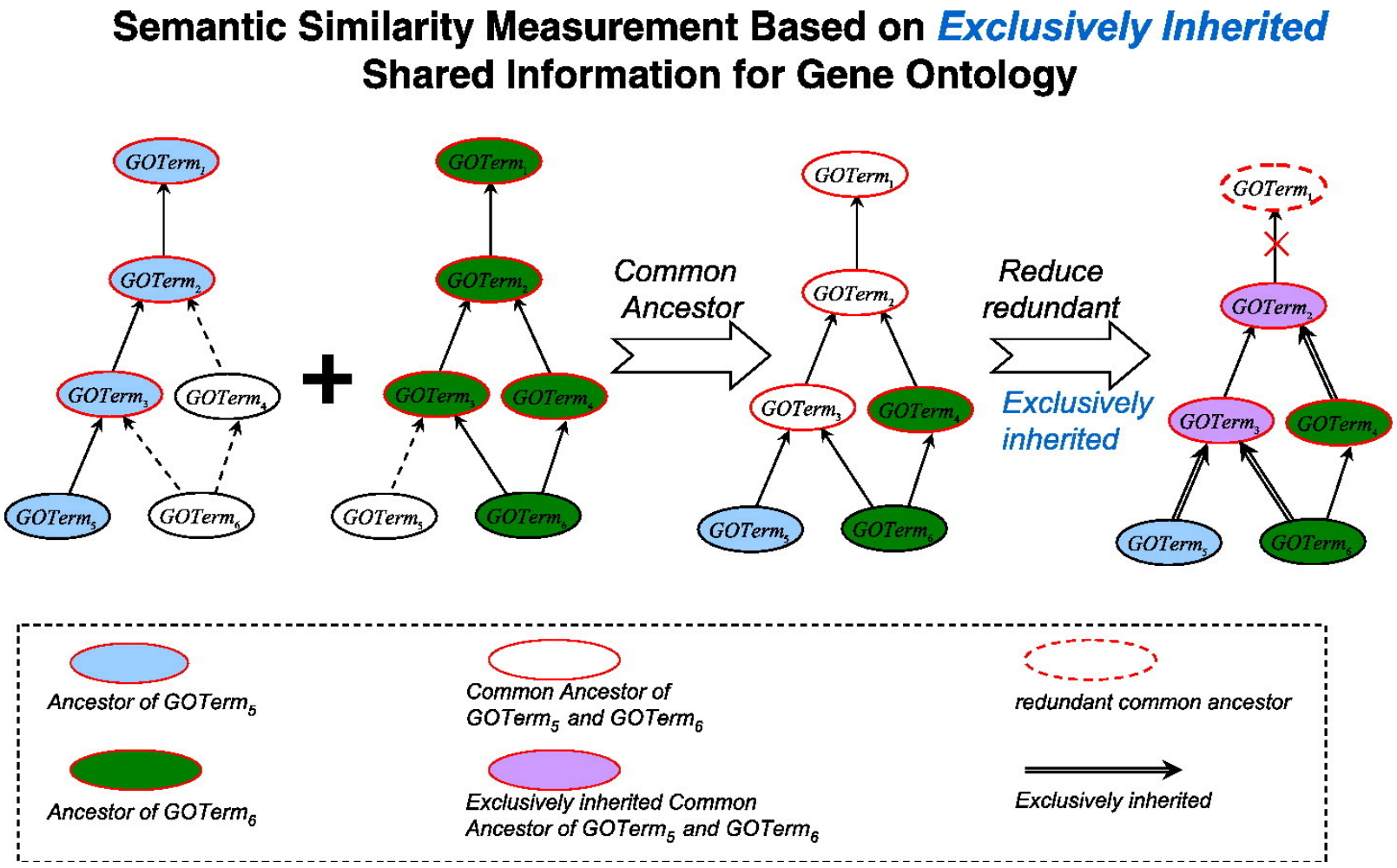
Problems with databases:

Low resolution

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.

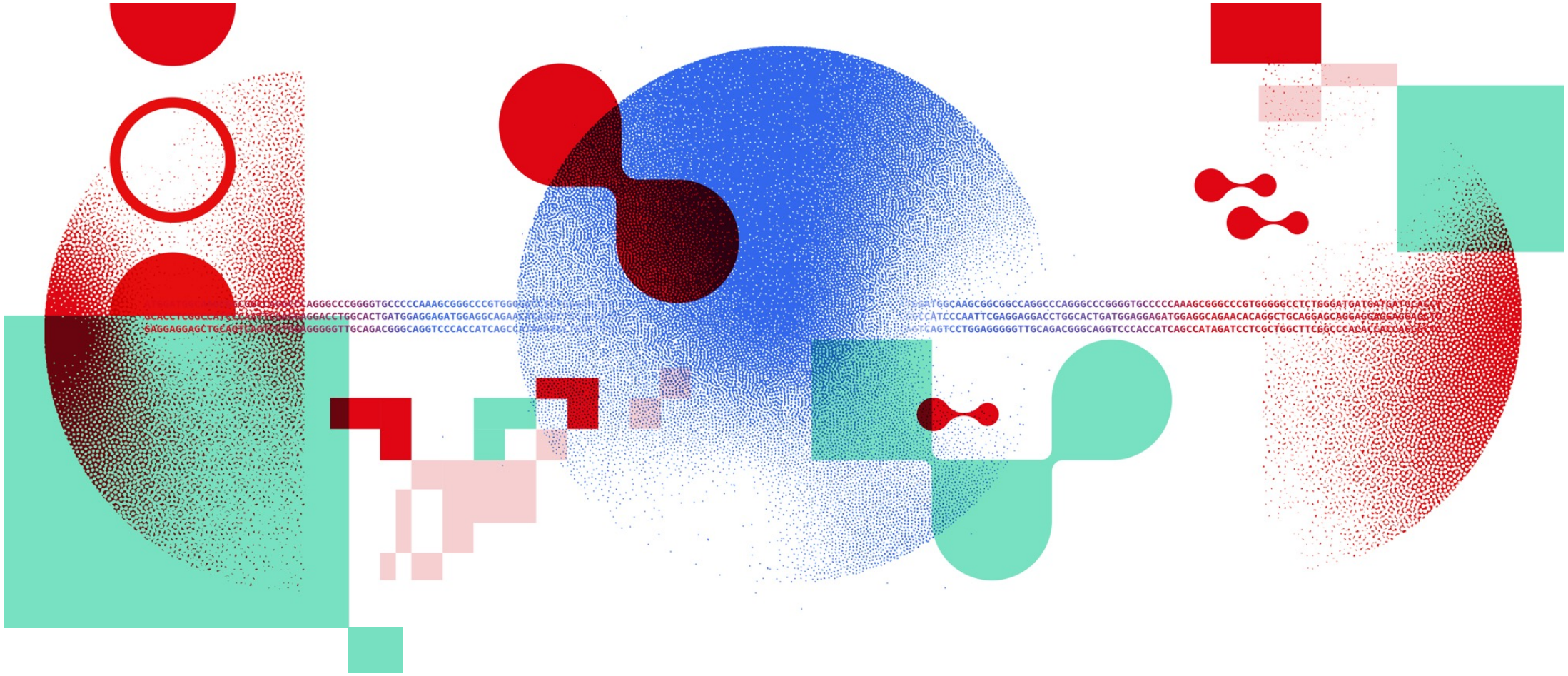
A) True

B) False

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

A) True

B) False



Thank you

DATA SCIENTISTS FOR LIFE

sib.swiss