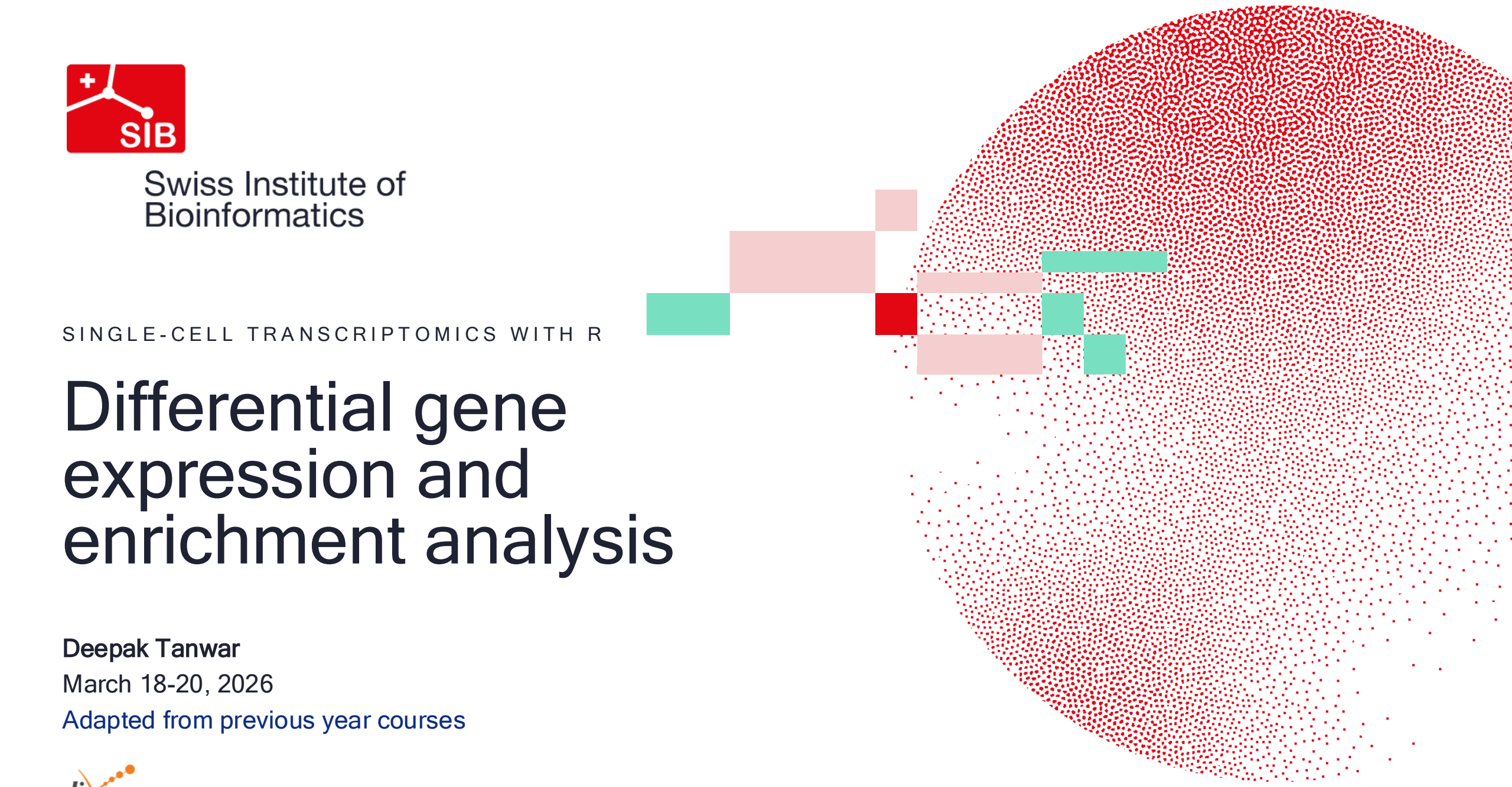


Swiss Institute of Bioinformatics

SINGLE-CELL TRANSCRIPTOMICS WITH R

Differential gene expression and enrichment analysis

Deepak Tanwar

March 18-20, 2026

Adapted from previous year courses

Learning objectives

Differential gene expression analysis

- Finding markers
- Pseudo-bulk analysis

Enrichment analysis

- What is Enrichment analysis?
- Distinguish between different ways to do it
- Challenges and Limitations of methods

What do you want to test in your data?

Differential gene expression analysis

DGE can be divided into 2 sub-groups:

a. Single-Cell-Level DGE Analysis (finding markers)

Goal: Identify genes that are differentially expressed between cell populations (e.g., cell types, clusters, or conditions) while accounting for the single-cell nature of the data (e.g., sparsity, dropout events).

Differential gene expression analysis

DGE can be divided into 2 sub-groups:

a. Single-Cell-Level DGE Analysis (finding markers)

Goal: Identify genes that are differentially expressed between cell populations (e.g., cell types, clusters, or conditions) while accounting for the single-cell nature of the data (e.g., sparsity, dropout events).

b. Pseudo-Bulk DGE Analysis (differential analysis)

Goal: Aggregate single-cell data into pseudo-bulk profiles to perform DGE analysis using bulk RNA-seq methods, reducing noise and leveraging biological replicates.

Marker gene identification

Genes overexpressed by each cell type or cluster with the dataset => It can help in cell type annotation

Marker gene identification

Methods:

- **Log-Fold Change (LFC) Analysis**: Calculate the log-fold change in expression between a target cluster and all other cells (e.g., in Seurat or Scanpy).

- **Wilcoxon Rank-Sum Test**: Test for genes with significantly higher expression in one group compared to others (e.g., in Seurat).

Marker gene identification

Use Cases:

- Annotating cell types (e.g., identifying CD3 as a marker for T cells)
- Discovering novel cell states or subpopulations

Typical Output

Gene Symbol	avg_log2FC	pct.1	pct.2	p_val	p_val_adj	cluster
CHI3L1	5.61	0.958	0.225	7.63E-255	1.97E-250	2
HLA-DRA	3.41	0.978	0.215	2.84E-253	7.32E-249	2
PTGFR	4.31	0.795	0.093	6.43E-244	1.66E-239	2
HLA-DRB5	3.54	0.818	0.097	2.10E-243	5.41E-239	2
GRIN2A	4.24	0.69	0.05	1.64E-235	4.22E-231	2
CDHR3	3.34	0.892	0.159	2.56E-229	6.59E-225	2
AKR1C3	3.16	0.955	0.254	5.75E-223	1.48E-218	2
KCNK15	3.76	0.897	0.187	4.88E-217	1.26E-212	2
HLA-DRB1	2.53	0.78	0.102	2.71E-212	6.98E-208	2
PLPP3	3.34	0.965	0.322	3.54E-210	9.11E-206	2

Marker gene identification

Challenges:

- Marker genes may not be unique to a single cell type, requiring careful validation
- Dropout events can obscure marker gene detection

Research | Open access | Published: 26 February 2024

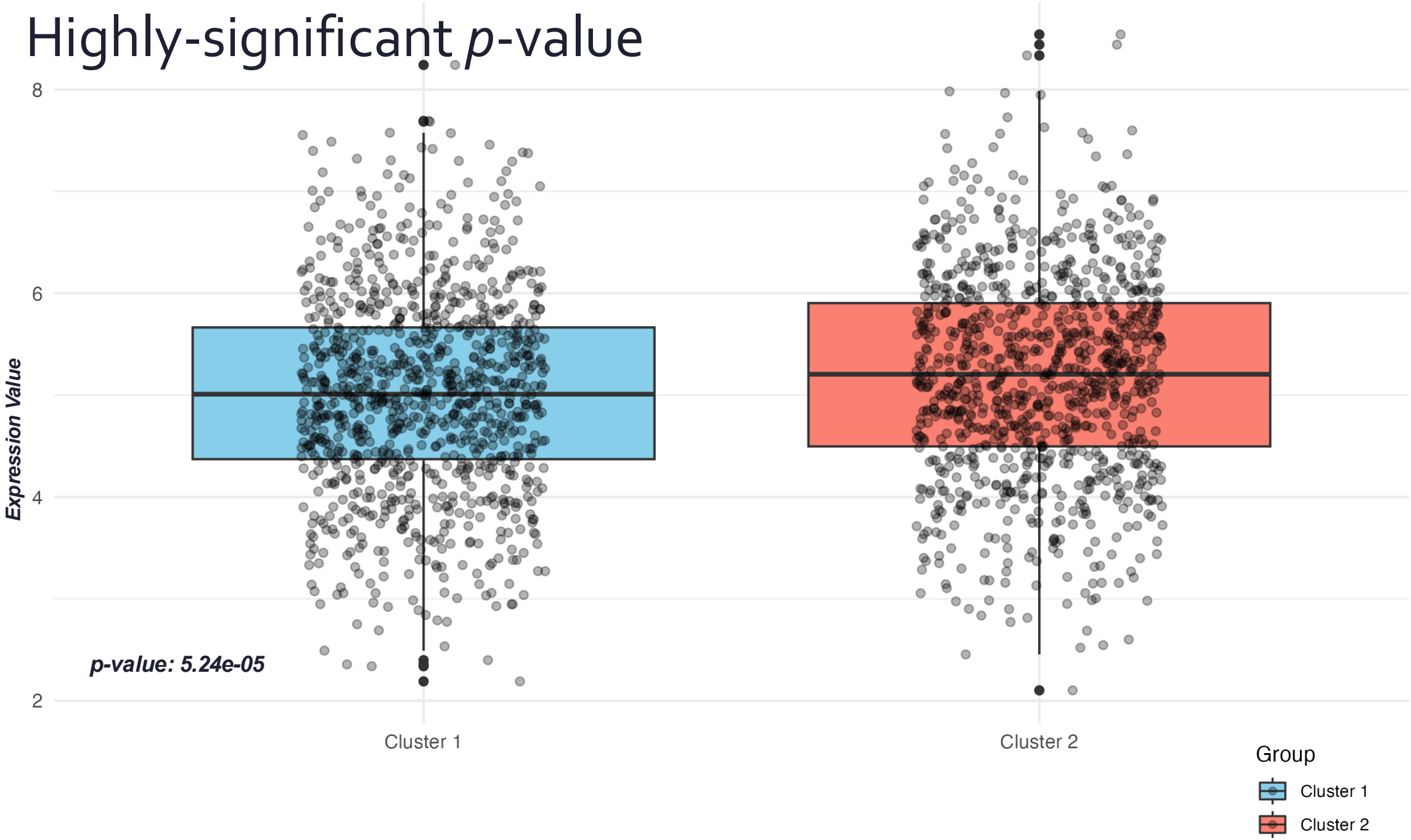
A comparison of marker gene selection methods for single-cell RNA sequencing data

Jeffrey M. Pullin & Davis J. McCarthy

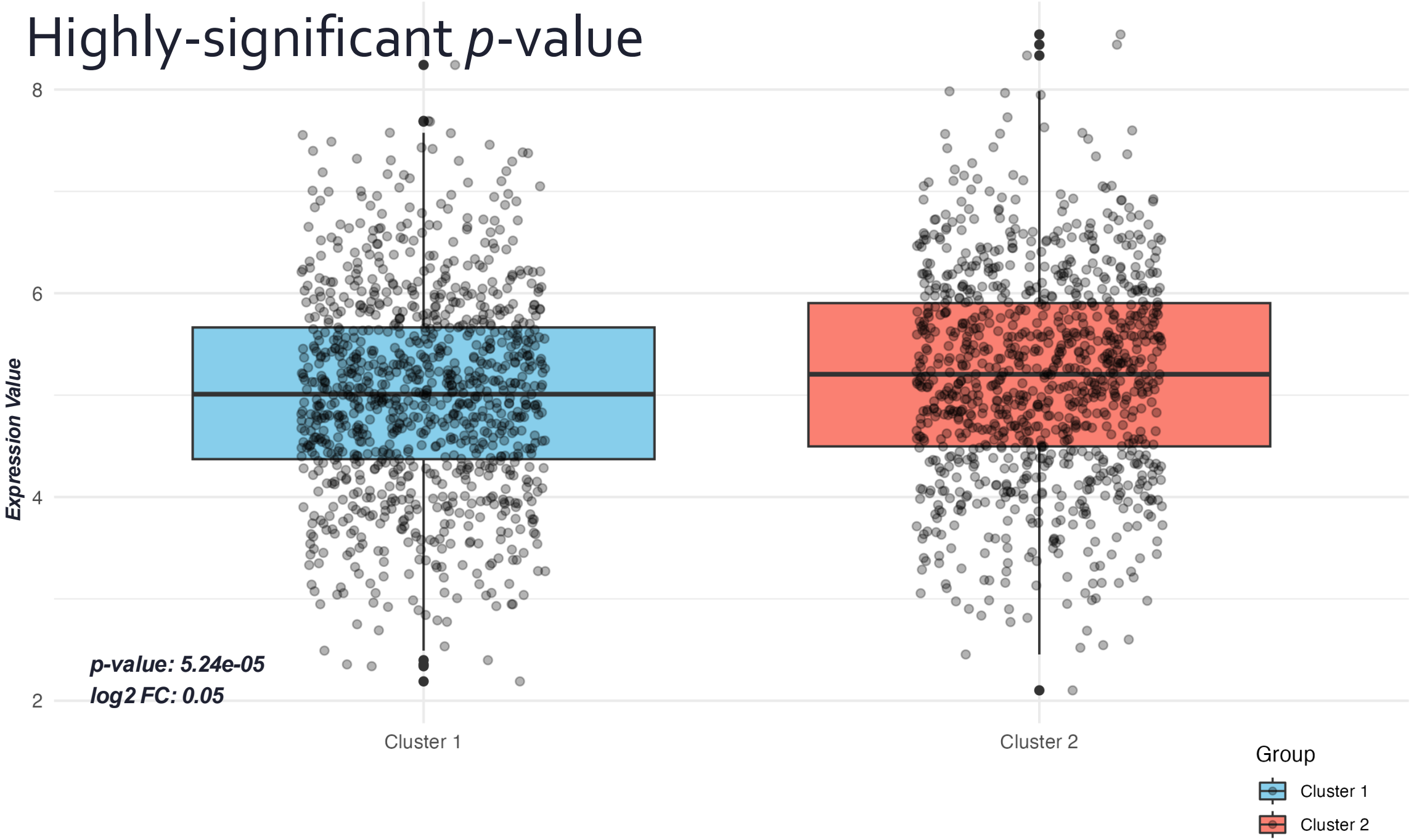
Genome Biology **25**, Article number: 56 (2024) | Cite this article

Methods based on logistic regression, Student's *t*-test and the Wilcoxon rank-sum test all have strong performance

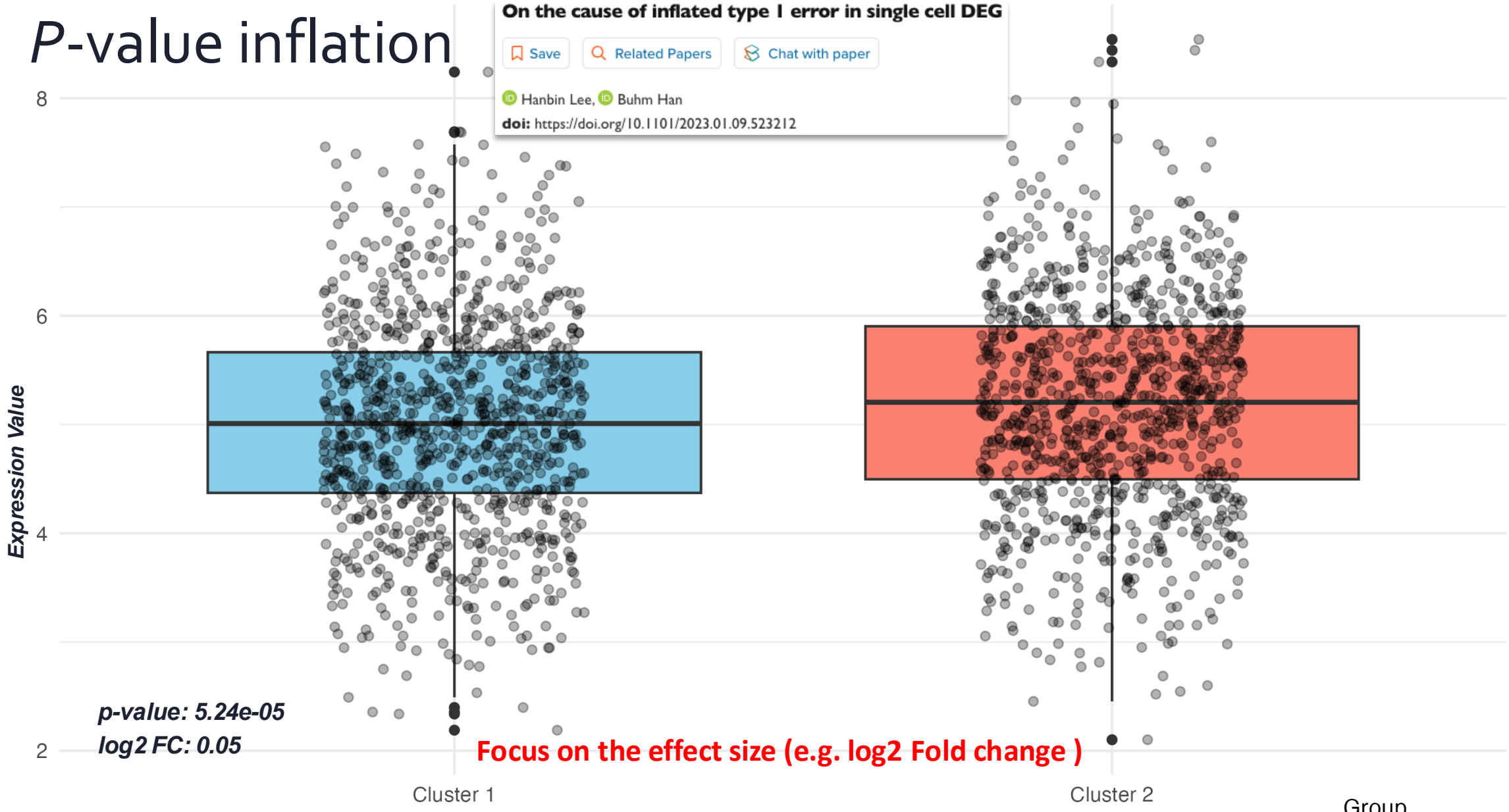
Highly-significant *p*-value



Highly-significant *p*-value



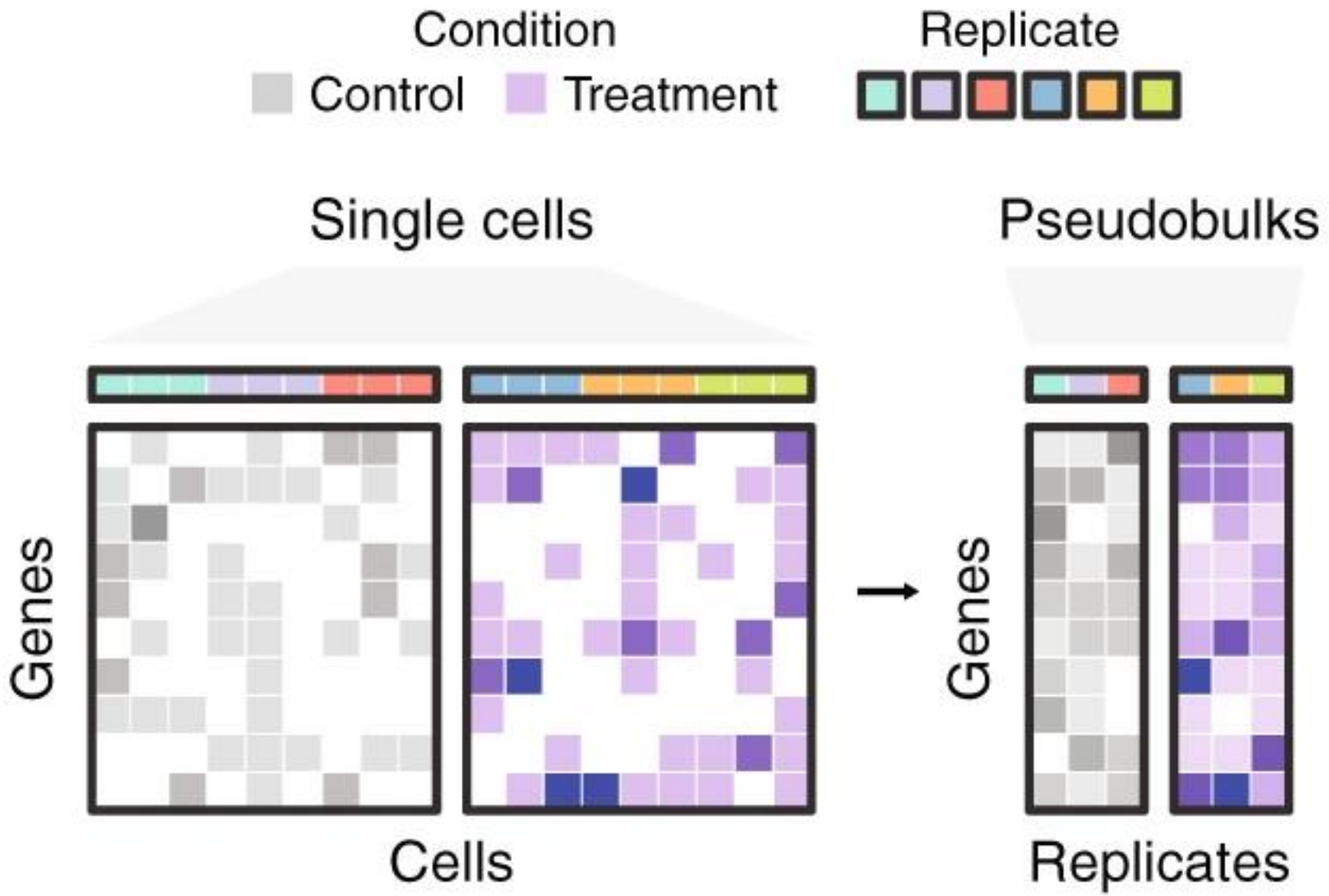
P-value inflation



The probability of observing a test statistic at least as extreme as the one observed, assuming that the null hypothesis is true.

What else can you do to avoid *P*-value inflation?

Pseudo-Bulk DGE Analysis



https://www.nature.com/articles/s41467-021-25960-2

Pseudo-Bulk DGE Analysis

Genes whose expression if modulated (up or down) by experimental condition within a cell type or cluster

What is the ideal method?

Analysis | Published: 26 February 2018

Bias, robustness and scalability in single-cell differential expression analysis

Charlotte Soneson & Mark D Robinson

Many methods have been used to determine differential gene expression from single-cell RNA (scRNA)-seq data. We evaluated 36 approaches using experimental and synthetic data and found considerable differences in the number and characteristics of the genes that are called differentially expressed. Prefiltering of lowly expressed genes has important effects, particularly for some of the methods developed for bulk RNA-seq data analysis. However, we found that bulk RNA-seq analysis methods do not generally

https://www.nature.com/articles/nmeth.4612

What is the ideal method?

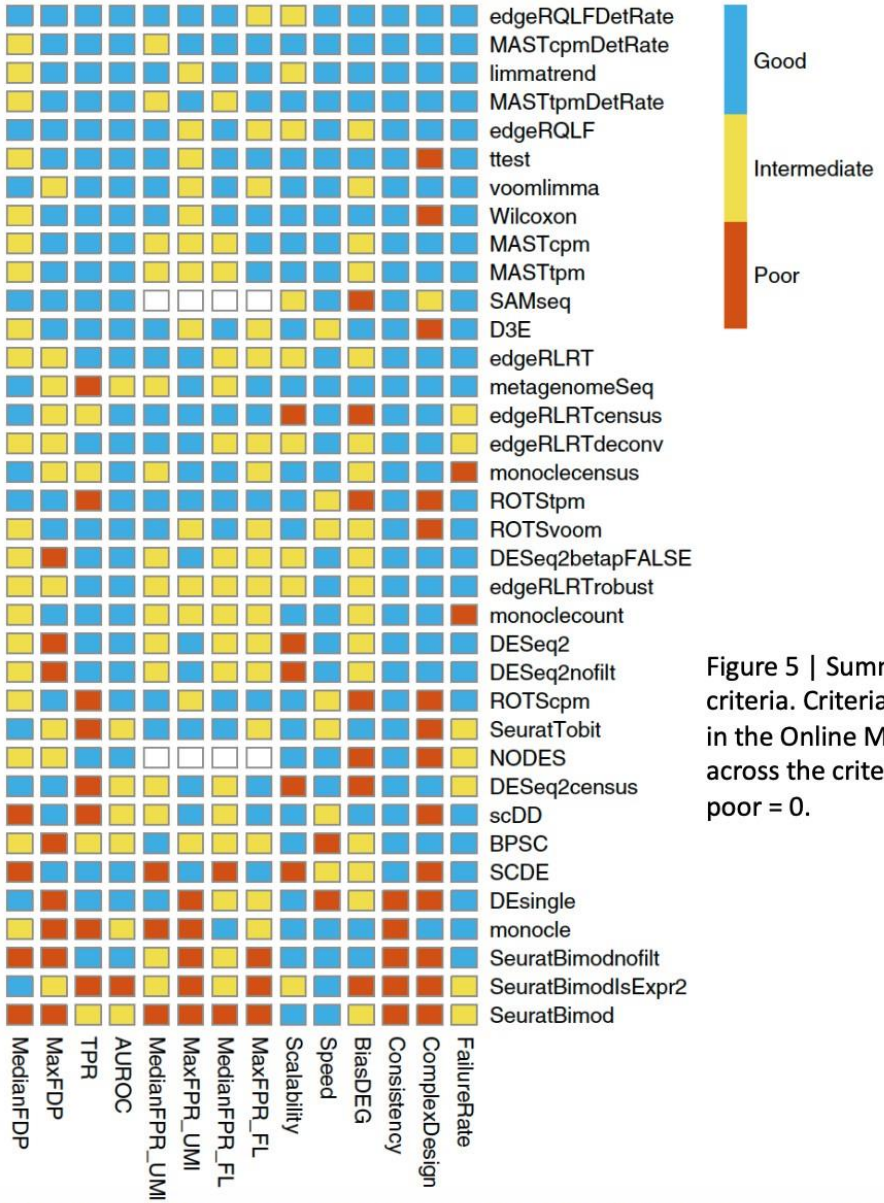


Figure 5 | Summary of DE method performance across all major evaluation criteria. Criteria and cutoff values for performance categories are available in the Online Methods. Methods are ranked by their average performance across the criteria, with the numerical encoding good = 2, intermediate = 1, poor = 0.

https://github.com/csoneson/conquer_comparison/blob/master/scripts/apply_edgeRQLFDetRate.R

Limma/edgeR: old but gold

Methods designed for microarray and bulk RNAseq analysis

- Can be used to include batch effects in model as covariates
- Compare more than 2 groups: e.g. ANOVA(F-test)

Analysis with limma and example of model with covariate:
https://ucdavis-bioinformatics-training.github.io/2018-June-RNA-Seq-Workshop/thursday/DE.html

Pseudo-Bulk DGE Analysis

Methods:

- Aggregate gene expression counts within groups (e.g., by cell type and sample) and use bulk RNA-seq tools like DESeq2, edgeR, or limma

- Tools like muscat in Bioconductor are specifically designed for pseudo-bulk DGE analysis in scRNA-seq

Pseudo-Bulk DGE Analysis

Advantages:

- Reduces noise and dropout effects
- Leverages well-validated bulk RNA-seq tools

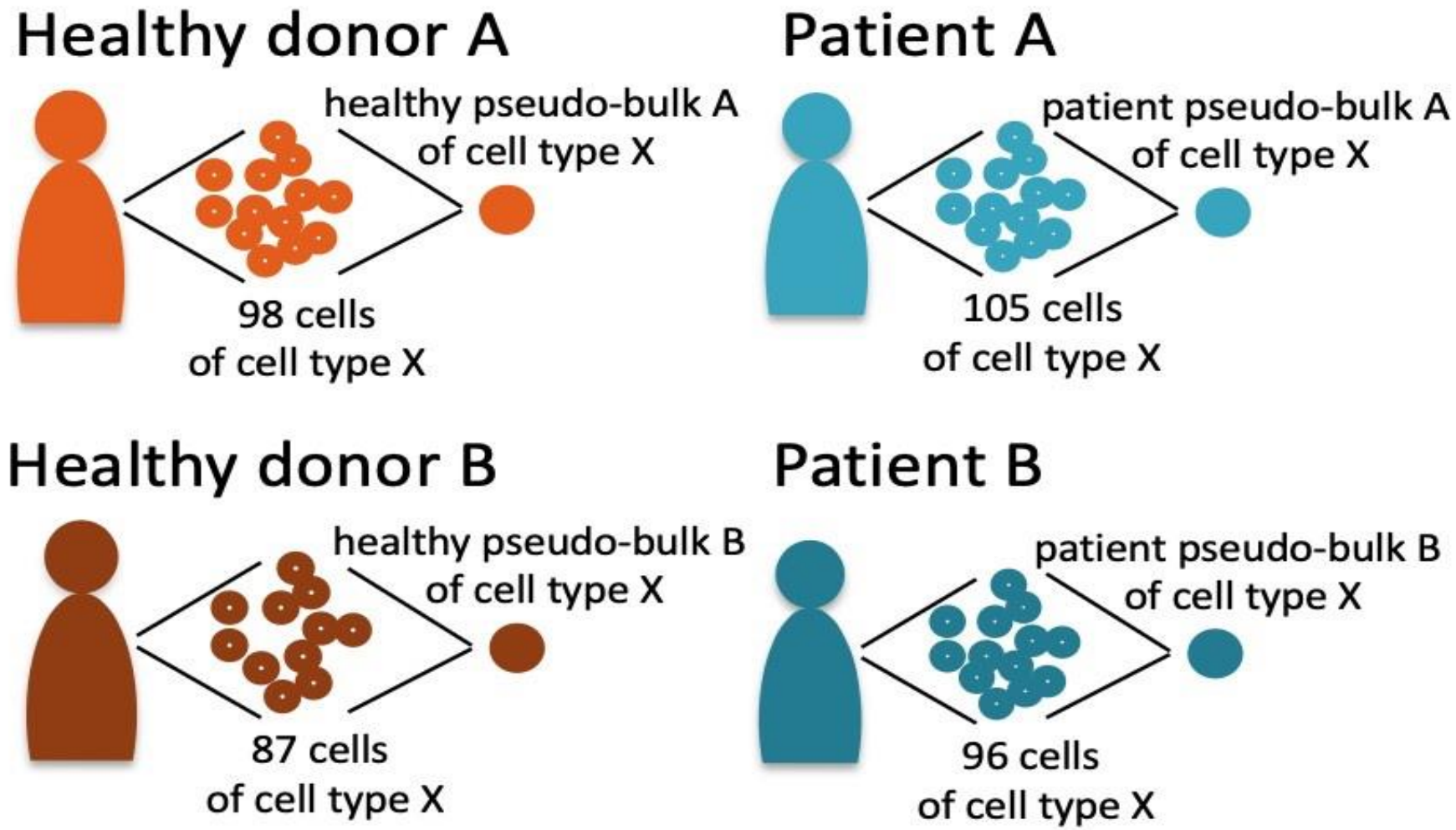
Pseudo-Bulk DGE Analysis

Limitations: Loses single-cell resolution and cannot detect cell-to-cell variability

Single Cell-Level: Finding markers

Pseudo-Bulk: Differential expression

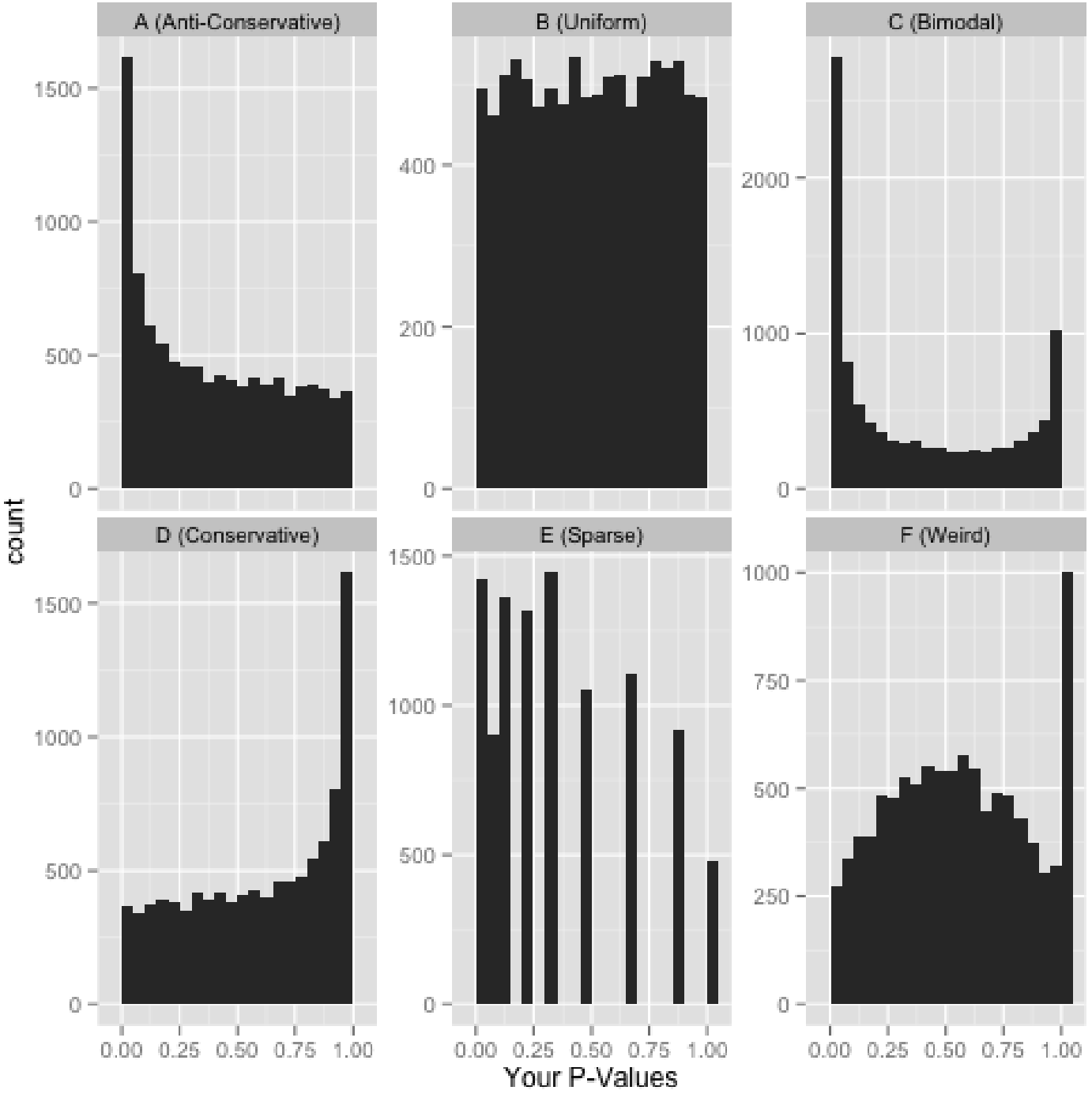
Pseudo-bulk DE analysis: muscat



https://www.bioconductor.org/packages/release/bioc/vignettes/muscat/inst/doc/analysis.html

What to do if you have no significant *p*-values?

Check the distribution of *p*-values



Quiz

What is a key limitation of pseudo-bulk analysis compared to single cell-level analysis?

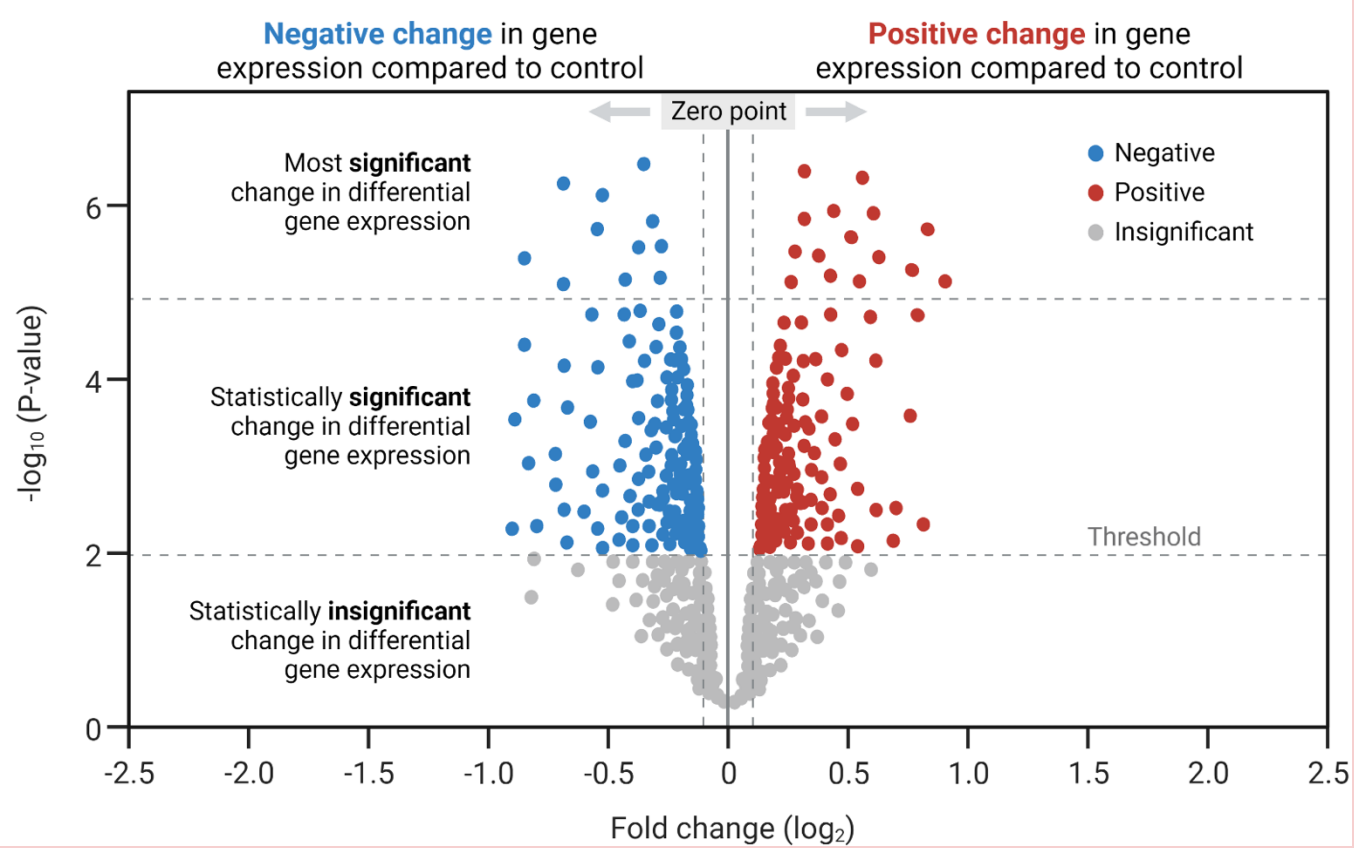
A) It is computationally more intensive.

B) It loses single-cell resolution and cannot detect cell-to-cell variability.

C) It cannot be used for differential expression analysis.

D) It is less robust to dropout events.

Many differentially expressed genes. What's next?



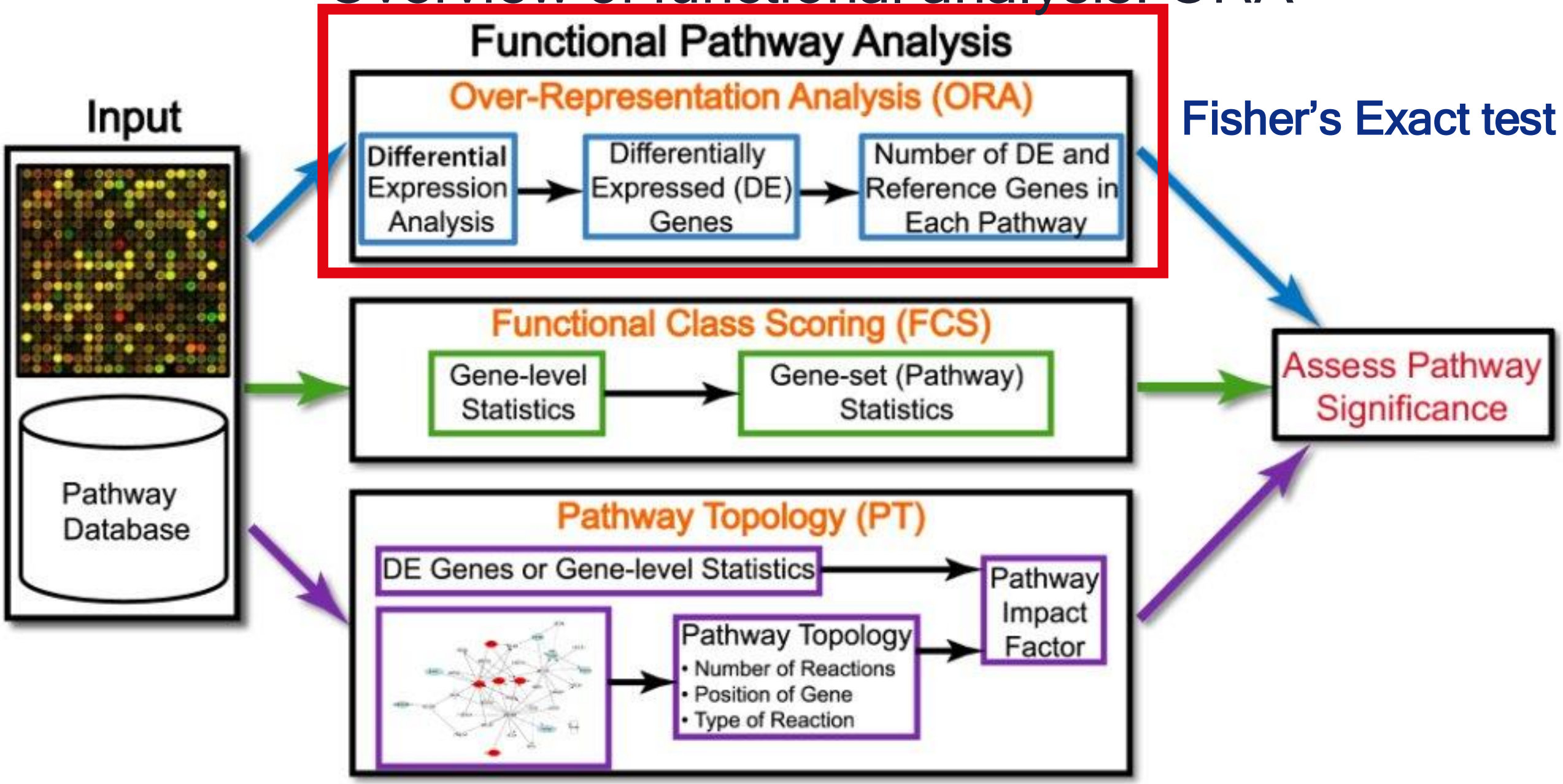
Many differentially expressed genes. What's next?

Manually check > 100 (often > 1000) genes

Check for enriched sets of genes based on e.g. pathway, function, location

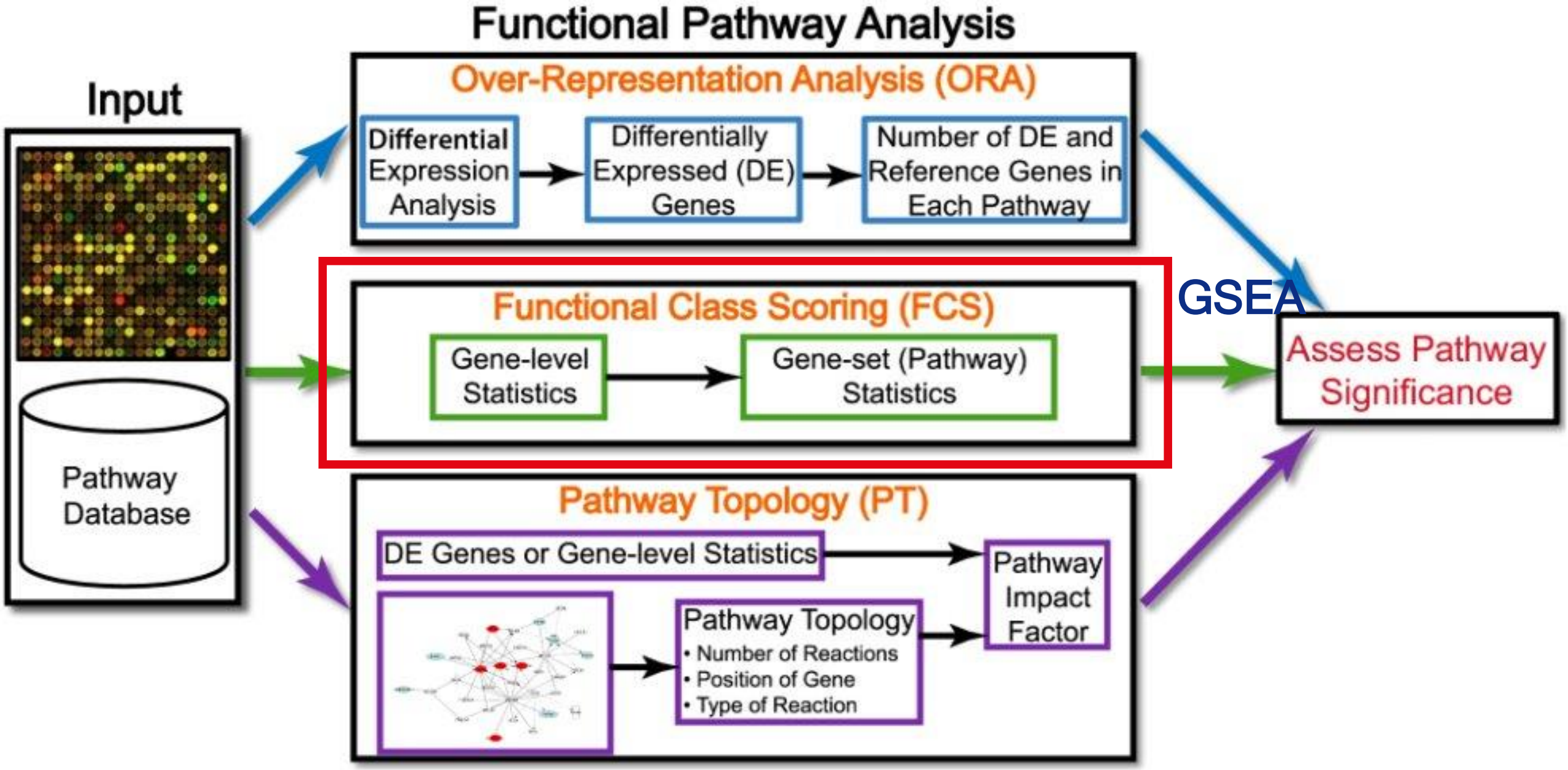
Goal: To gain biologically meaningful insights from long gene lists

Overview of functional analysis: ORA



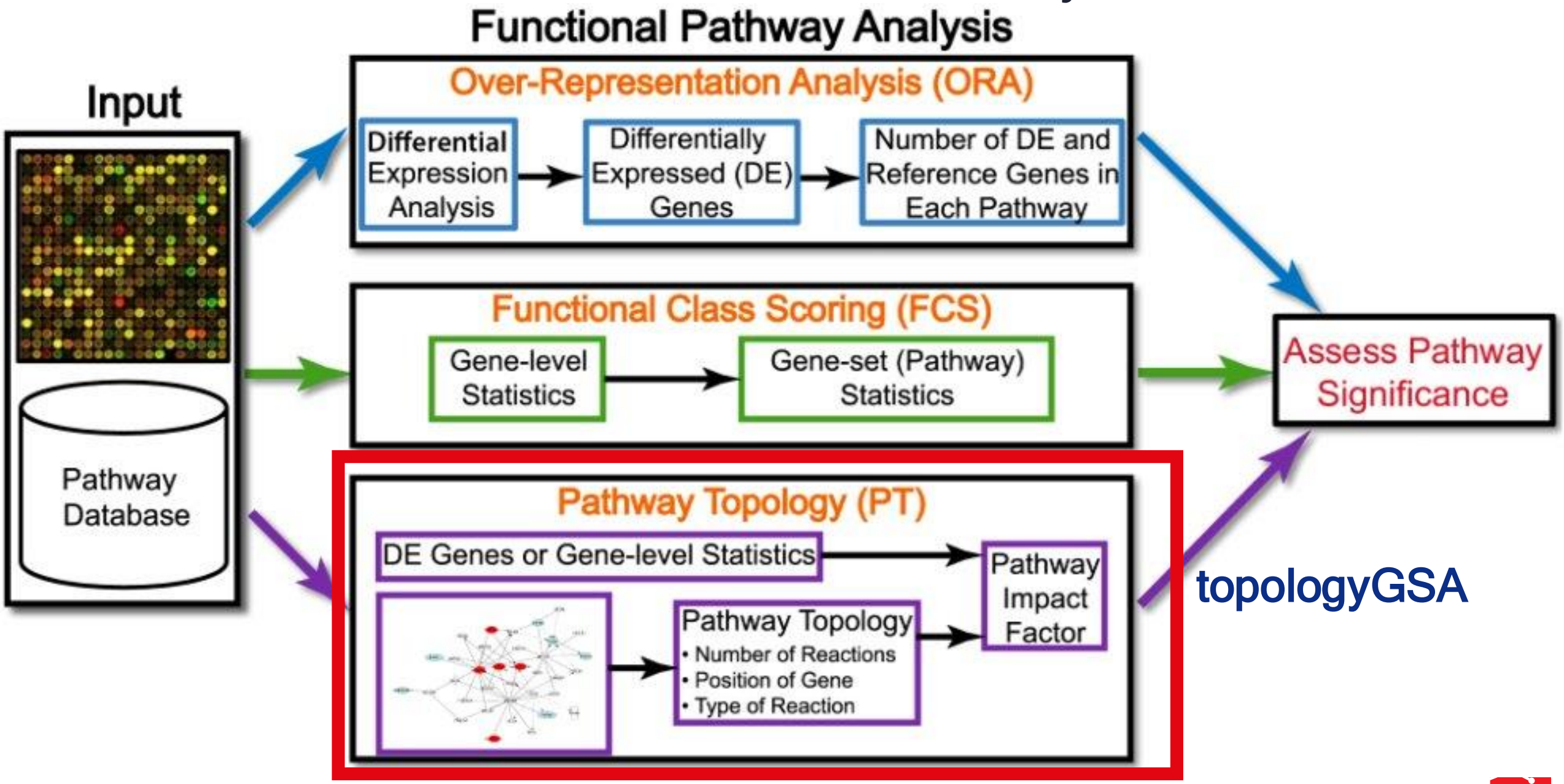
https://doi.org/10.1371/journal.pcbi.1002375

Overview of functional analysis: FCS



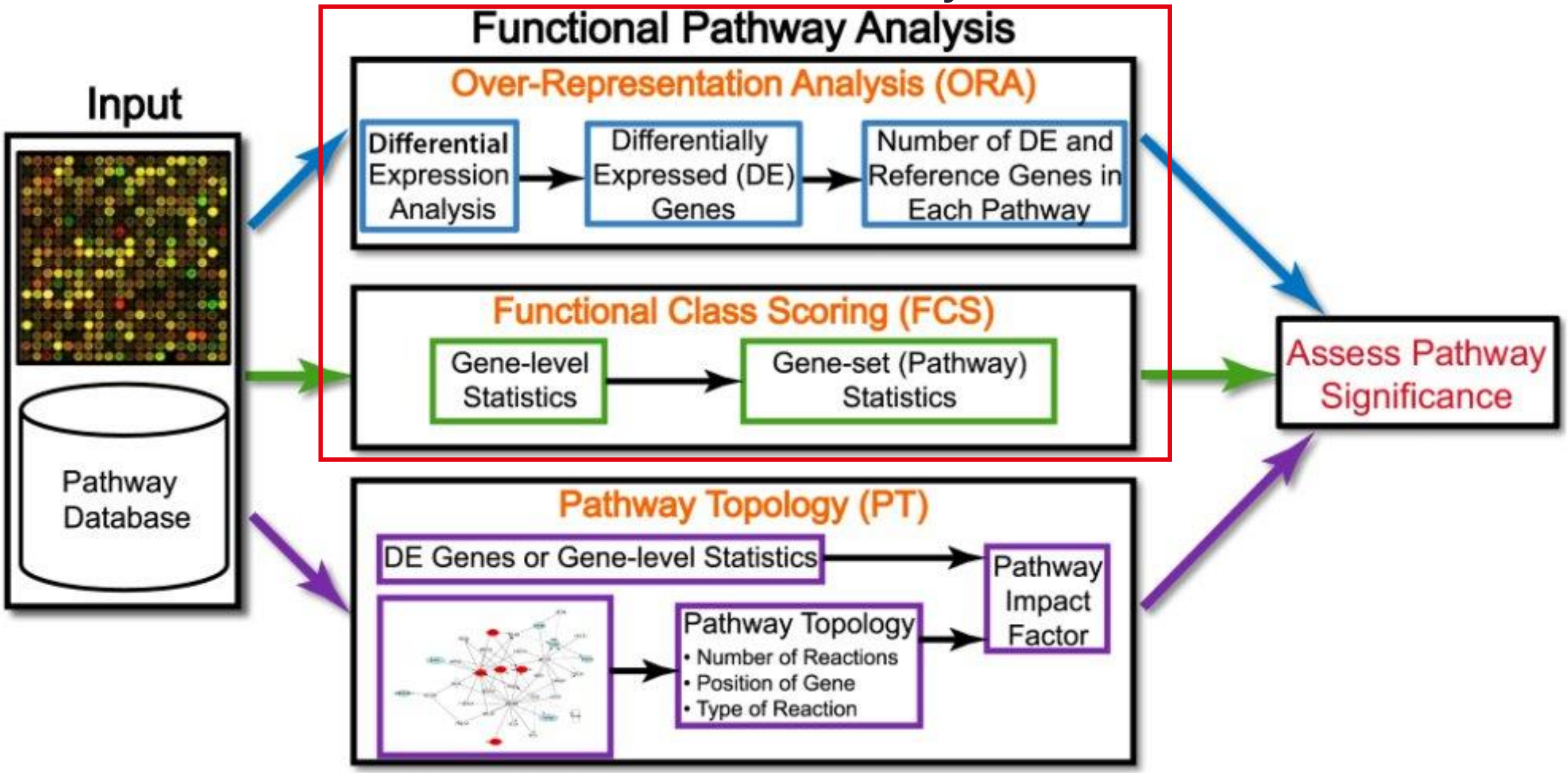
https://doi.org/10.1371/journal.pcbi.1002375

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

What is a gene set?

A gene set is an unordered collection of genes that are functionally related.

- Genes located in the same compartment in a cell (e.g. all proteins located in the cell nucleus)
- Proteins that are all regulated by a same transcription factor
- Custom gene list that comes from a publication and that are down-regulated in a mutant
- List of genes that contain SNPs associated with a disease
- ...etc!
- Several gene sets are grouped into Knowledge bases
- A pathway can be interpreted as a gene set by ignoring functional relationships among 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.

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)

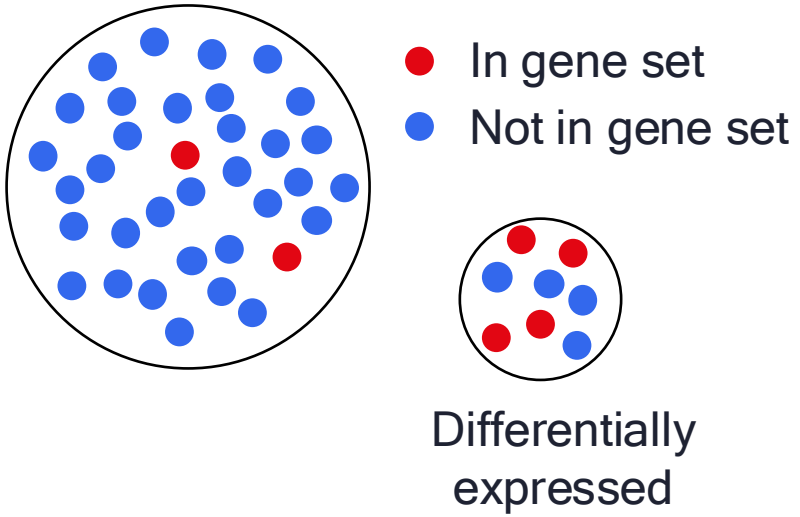
Gene1	0.051
Gene2	0.05001
Gene 3	0.049
Gene 4	0.001
Gene 5	0.023
Gene 6	0.04
Gene 7	0.01
Gene 8	0.0501
Gene 9	0.2
Gene 10	0.051
Gene 11	0.05
Gene 12	0.49
Gene 13	0.03
Gene 14	0.01
Gene 15	0.052
Gene 16	0.9

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

Functional class scoring (FCS)

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

Functional class scoring (FCS)

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

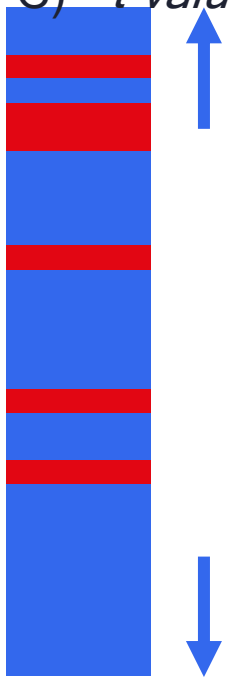
Research Article | Open access | Published: 12 May 2017

Ranking metrics in gene set enrichment analysis: do they matter?

Joanna Zyla, Michal Marczyk, January Weiner & Joanna Polanska

BMC Bioinformatics **18**, Article number: 256 (2017) | Cite this article

Genes ranked by test statistic
or
log2(FC) * *t-value*



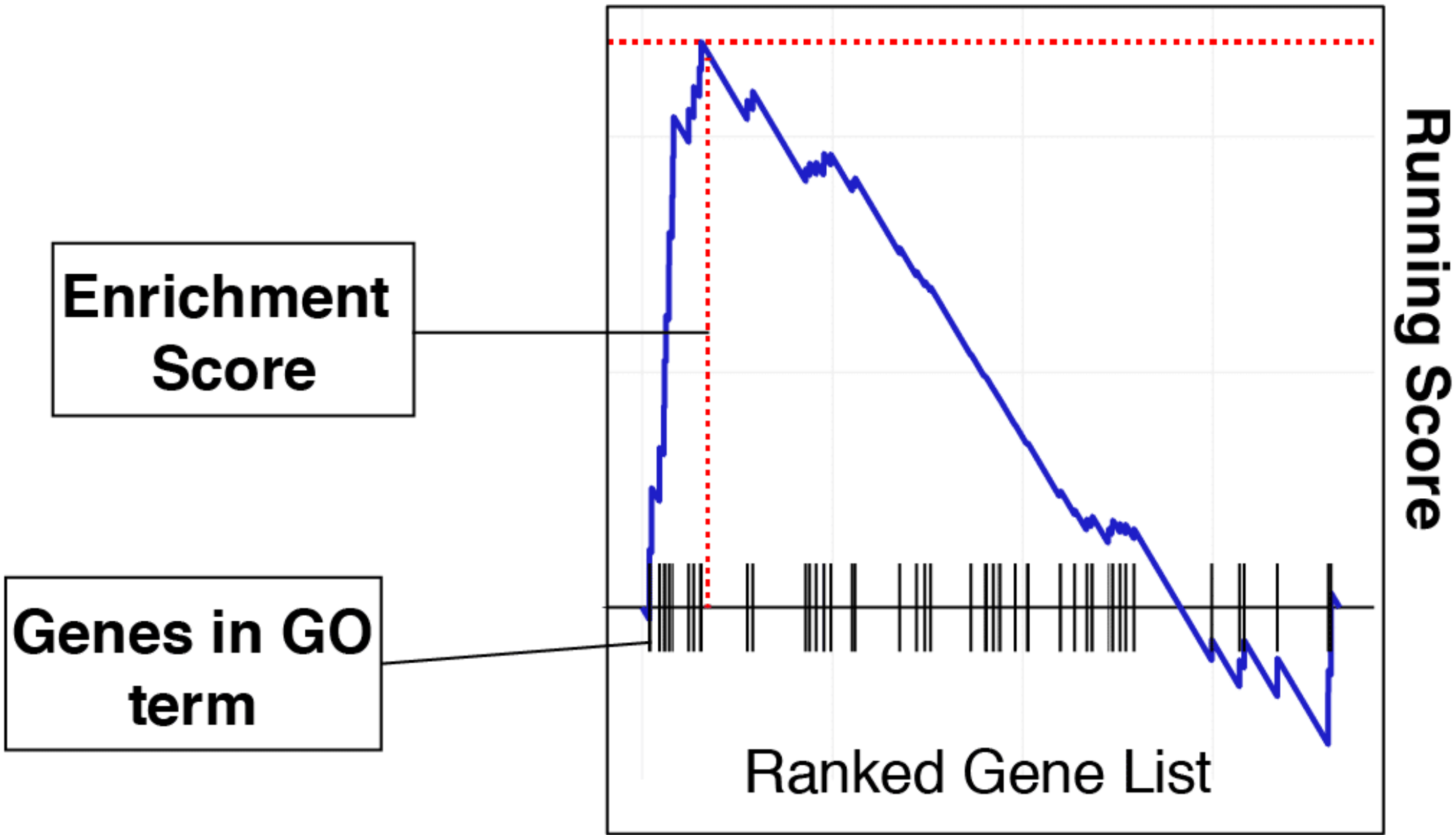
Upregulated

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

Gene set enrichment analysis (GSEA)



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

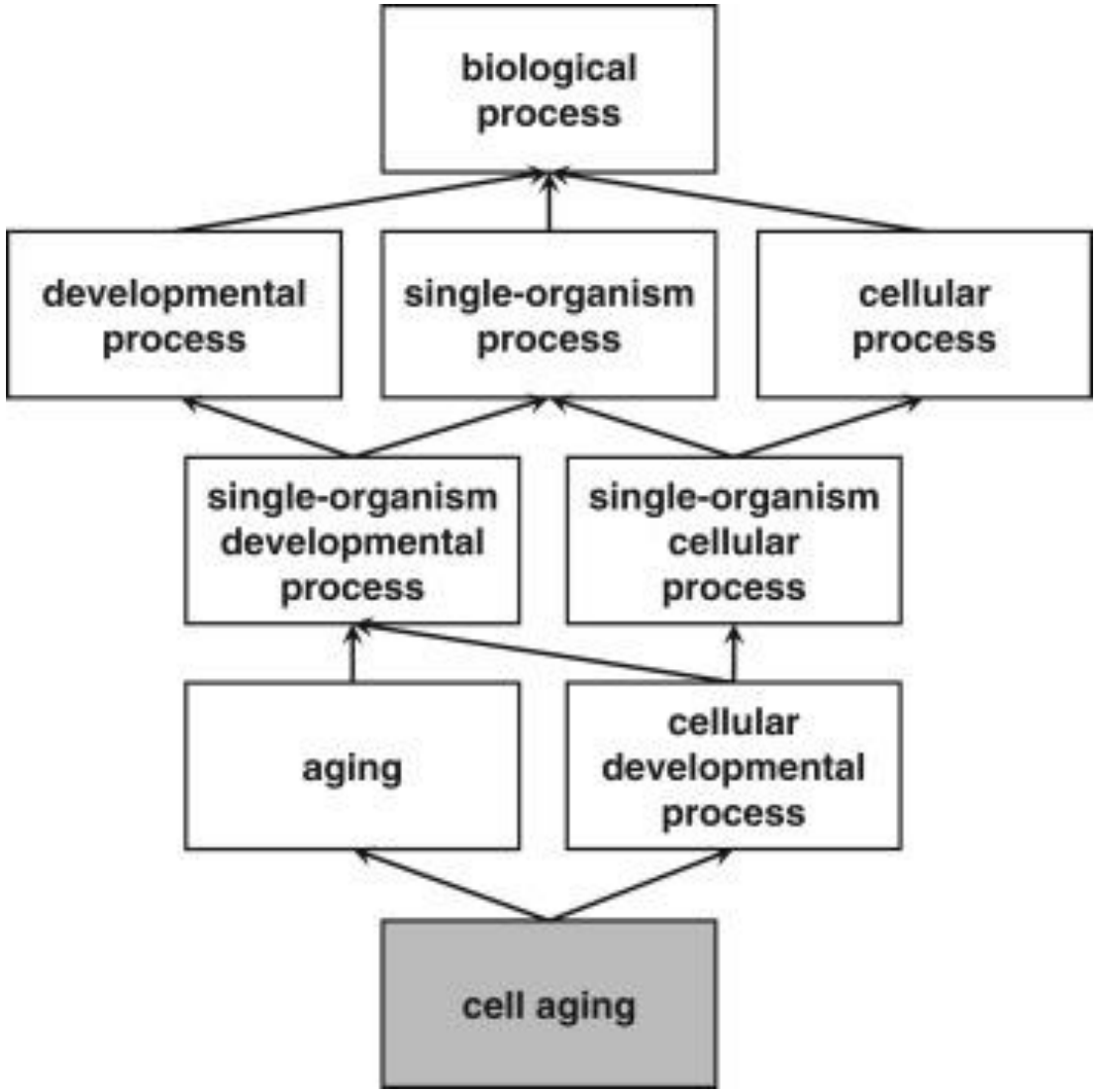
- ORA
- GSEA
- SAFE
- PADOG
- ROAST
- CAMERA
- GSA
- GSVA/ssGSEA
- GlobelTest
- EBM
- MGSA
- GOSeq
- QUSAGE
- Pathview
- GOSemSim
- GGEA
- SPIA
- PathNet
- DEGraph
- TopologyGSA
- GANPA
- CePa
- NetGSA
- WGCNA

Problems with databases:

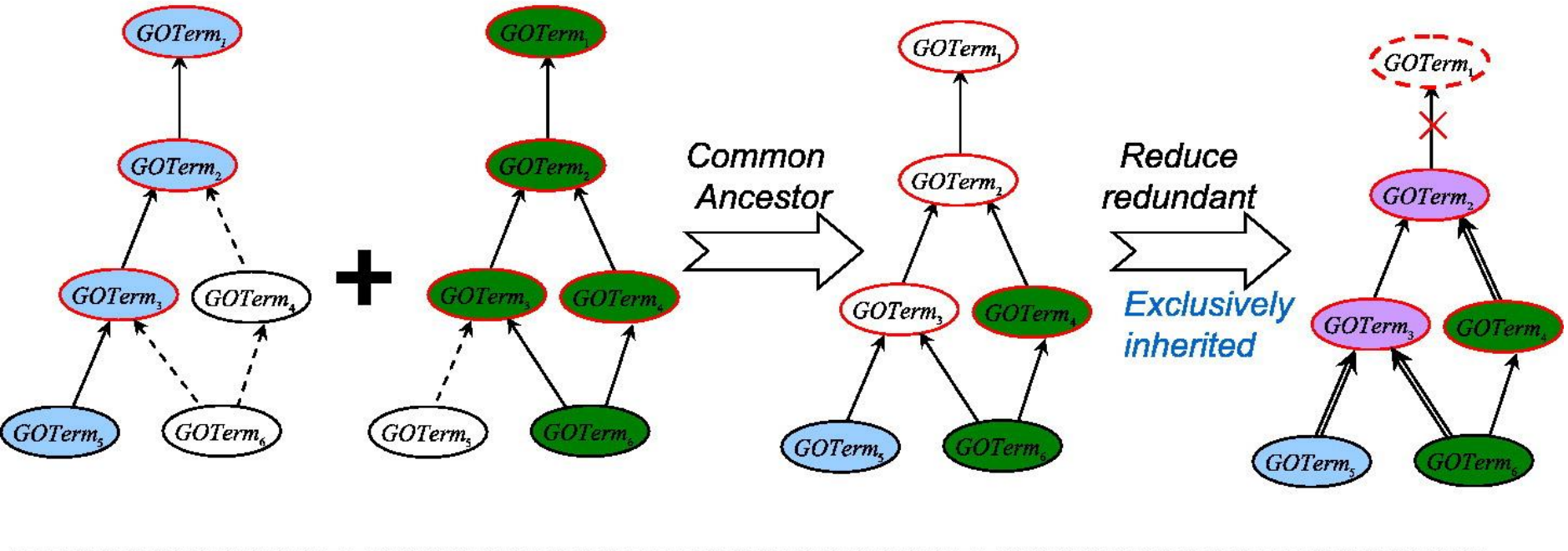
Low resolution

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

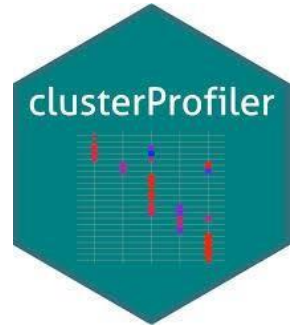
database_seeds

$paper1_day1
Gene1, Gene2, Gene3, Gene4

$paper2_day2
Gene3, Gene4, Gene5, Gene6

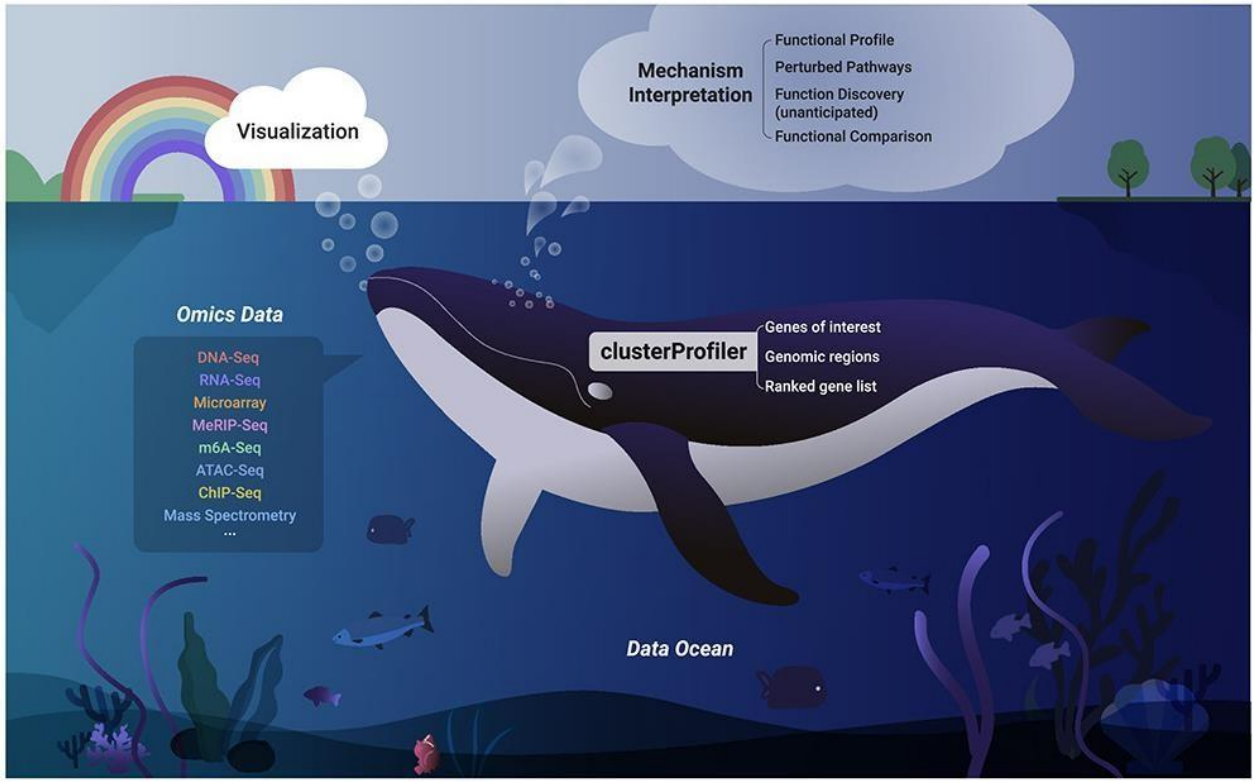
clusterProfiler

A universal enrichment tool for interpreting omics data



platforms all | rank 36 / 2300 | support 1 5 / 1 8 | in Bioc 13.5 years | build ok | updated < 3 months | dependencies 132

DOI: 10.18129/B9.bioc.clusterProfiler



Wu T, et al. (2021). "clusterProfiler 4.0: A universal enrichment tool for interpreting omics data." *The Innovation*, **2**(3), 100141. doi:10.1016/j.xinn.2021.100141.

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

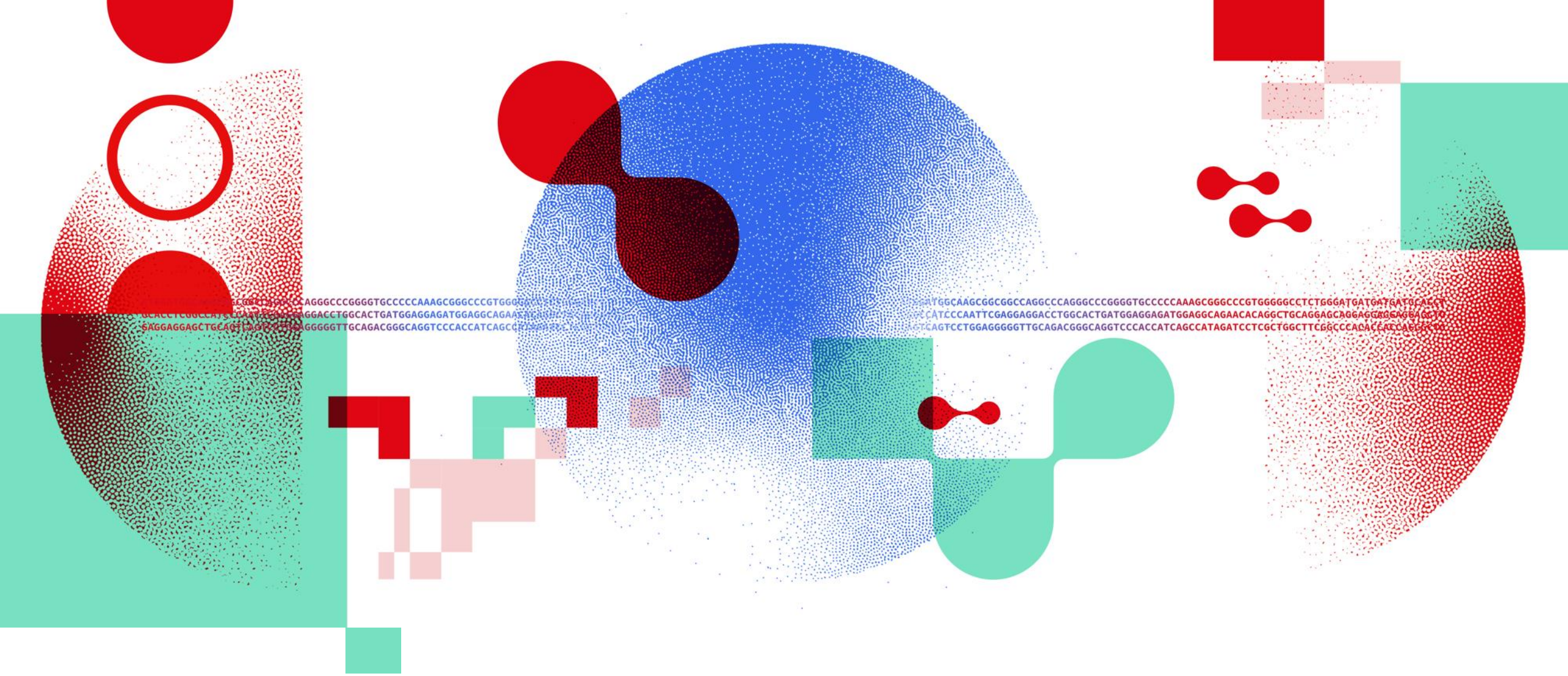
Summary

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