Swiss Institute of Bioinformatics

Introduction to RNA-Seq: Enrichment analysis

Wandrille Duchemin

elixir SWITZERLAND

Enrichment analysis

What to do with your list of differentially expressed gene?

- Interpretation can be difficult, especially if many genes are DE

Enrichment analysis

What to do with your list of differentially expressed gene?

- Interpretation can be difficult, especially if many genes are DE

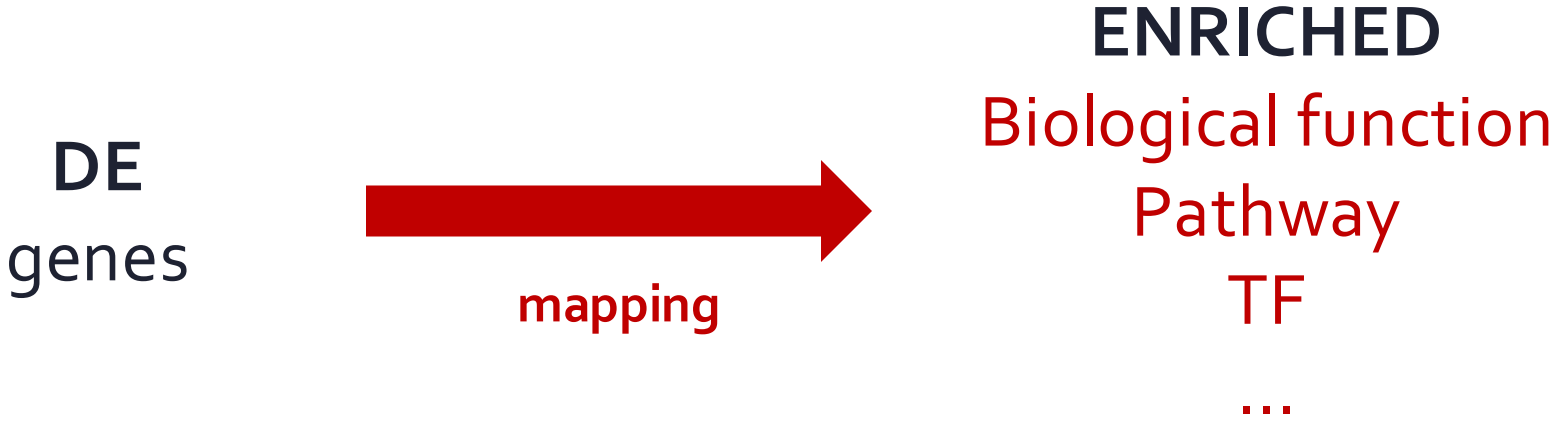
DE
genes

ENRICHED
Biological function
Pathway
TF
…

Enrichment analysis

What to do with your list of differentially expressed gene?

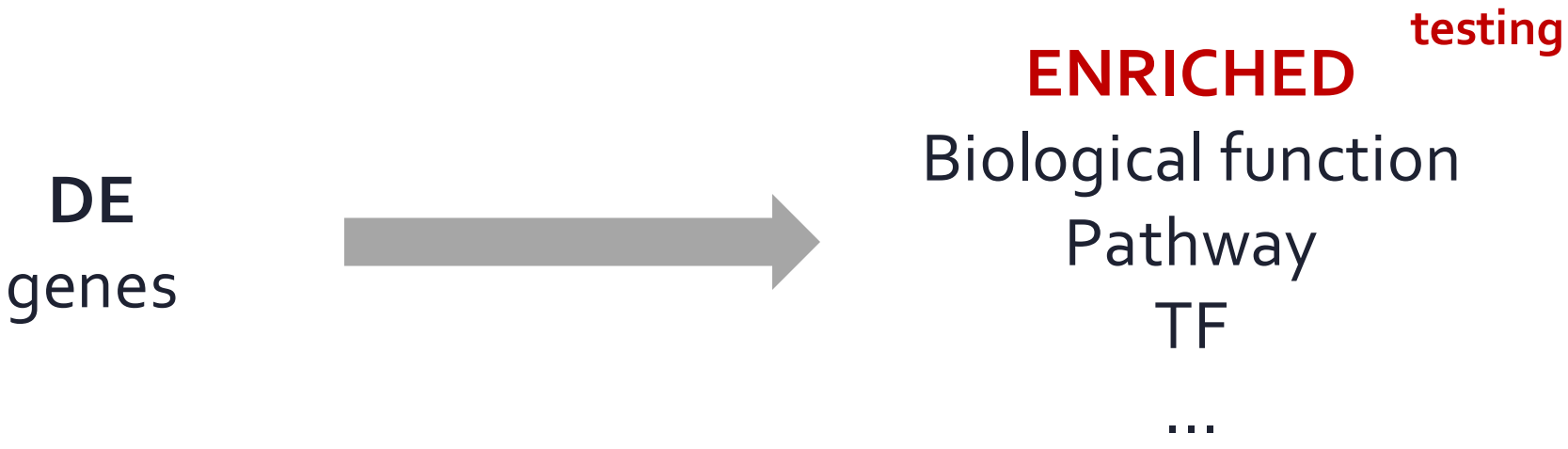
- Interpretation can be difficult, especially if many genes are DE



Enrichment analysis

What to do with your list of differentially expressed gene?

- Interpretation can be difficult, especially if many genes are DE



Enrichment analysis : mapping

Regrouping genes together in meaningful sets

- Same pathway
- Same location in the cell
- Same molecular function
- …

Enrichment analysis : mapping

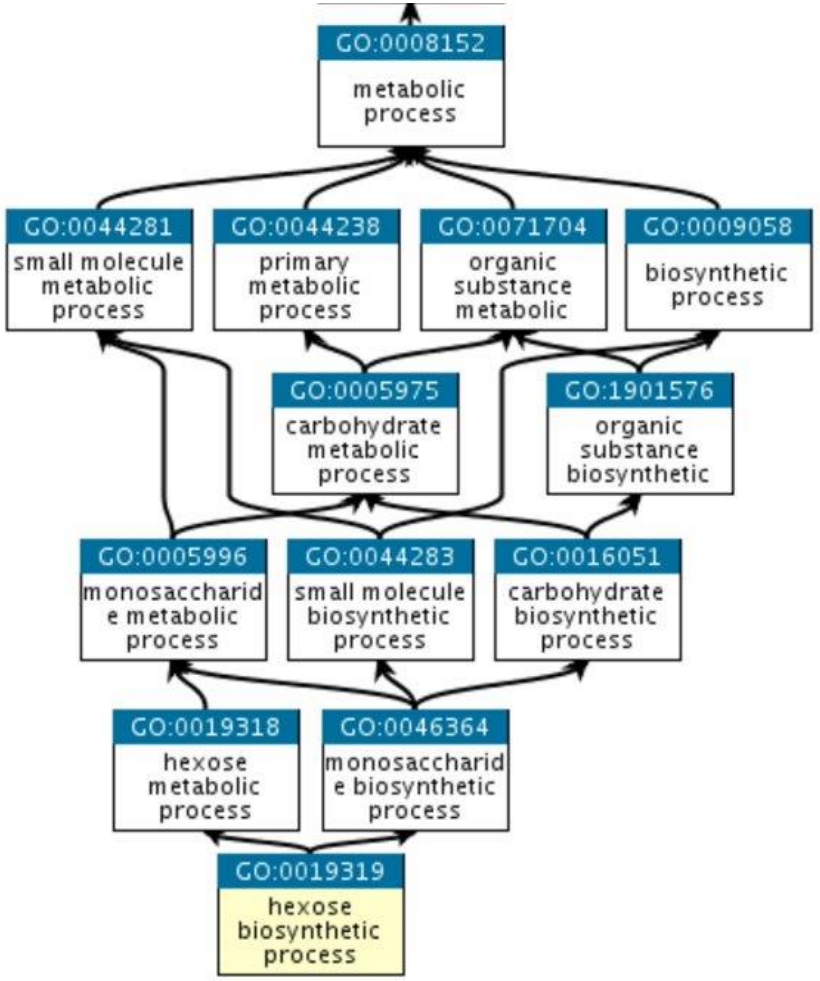
- Gene Ontologies
- Reactome
- KEGG
- MSigDB
- Custom set
- …

Enrichment analysis : mapping

Gene Ontologies geneontology.org

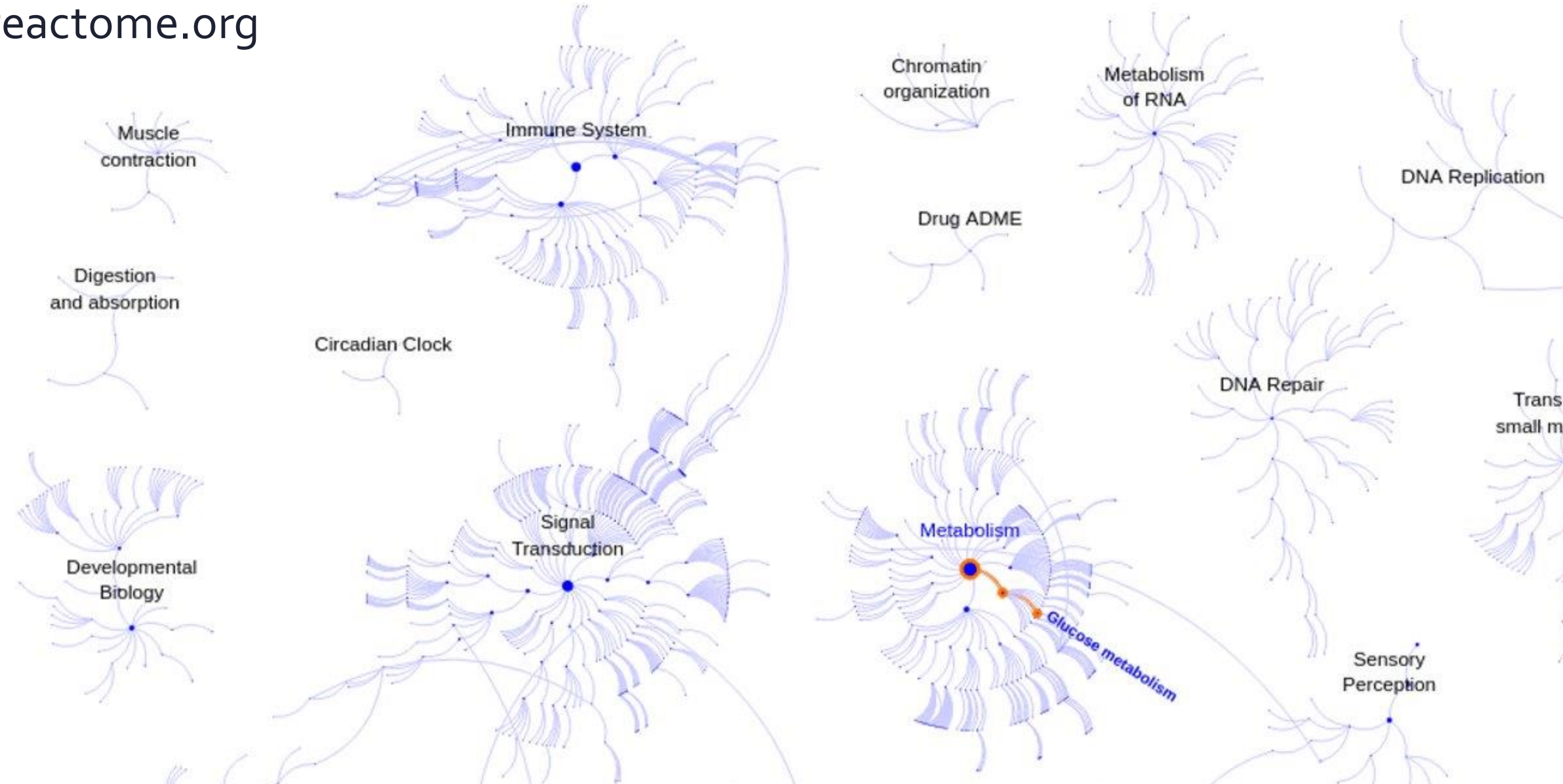
3 domains of nested terms:

- Molecular Function
- Cellular Component
- Biological Process



Enrichment analysis : mapping

reactome.org



Enrichment analysis : mapping

MSigDB : http://www.gsea-msigdb.org/gsea/msigdb/index.jsp

Human, mouse, and rat only

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 **regulatory target gene sets** based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.

C4 **computational gene sets** defined by mining large collections of cancer-oriented microarray data.

C5 **ontology gene sets** consist of genes annotated by the same ontology term.

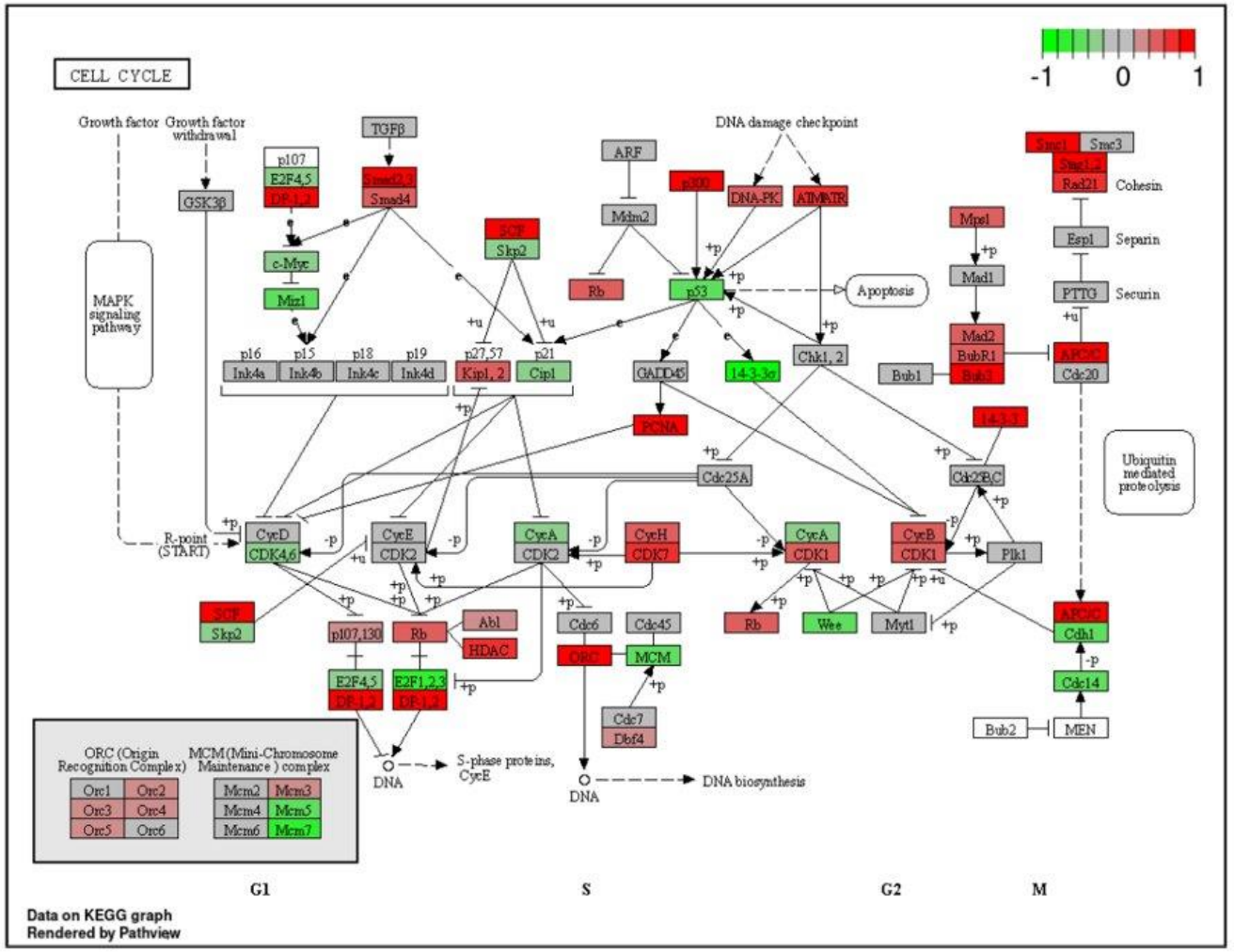
C6 **oncogenic signature gene sets** defined directly from microarray gene expression data from cancer gene perturbations.

C7 **immunologic signature gene sets** represent cell states and perturbations within the immune system.

C8 **cell type signature gene sets** curated from cluster markers identified in single-cell sequencing studies of human tissue.

Enrichment analysis : mapping

KEGG : http://www.genome.jp/keggKEGG PATHWAY



Projection of DE data
onto a KEGG pathway map
With R package pathview

https://bioconductor.org/packages/release/bioc/html/pathview.html

Enrichment analysis : mapping

Custom gene sets

- Derived from specialized litterature
- Tentative annotation

Very important in non-model organisms

Enrichment analysis : computing enrichment

2 main approaches

- Over Representation Analysis (ORA)
- Gene Set Enrichment Analysis (GSEA)

Enrichment analysis : computing enrichment

Over Representation Analysis (ORA)

- Basically a Fisher's exact test with p-value correction

	DE	Not DE
in gene set	A	B
not in gene set	C	D

N = A+B+C+D **# total genes**
M = A+B **# genes in set**

n = A+C **# DE genes**
k = A **# DE genes in set**

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

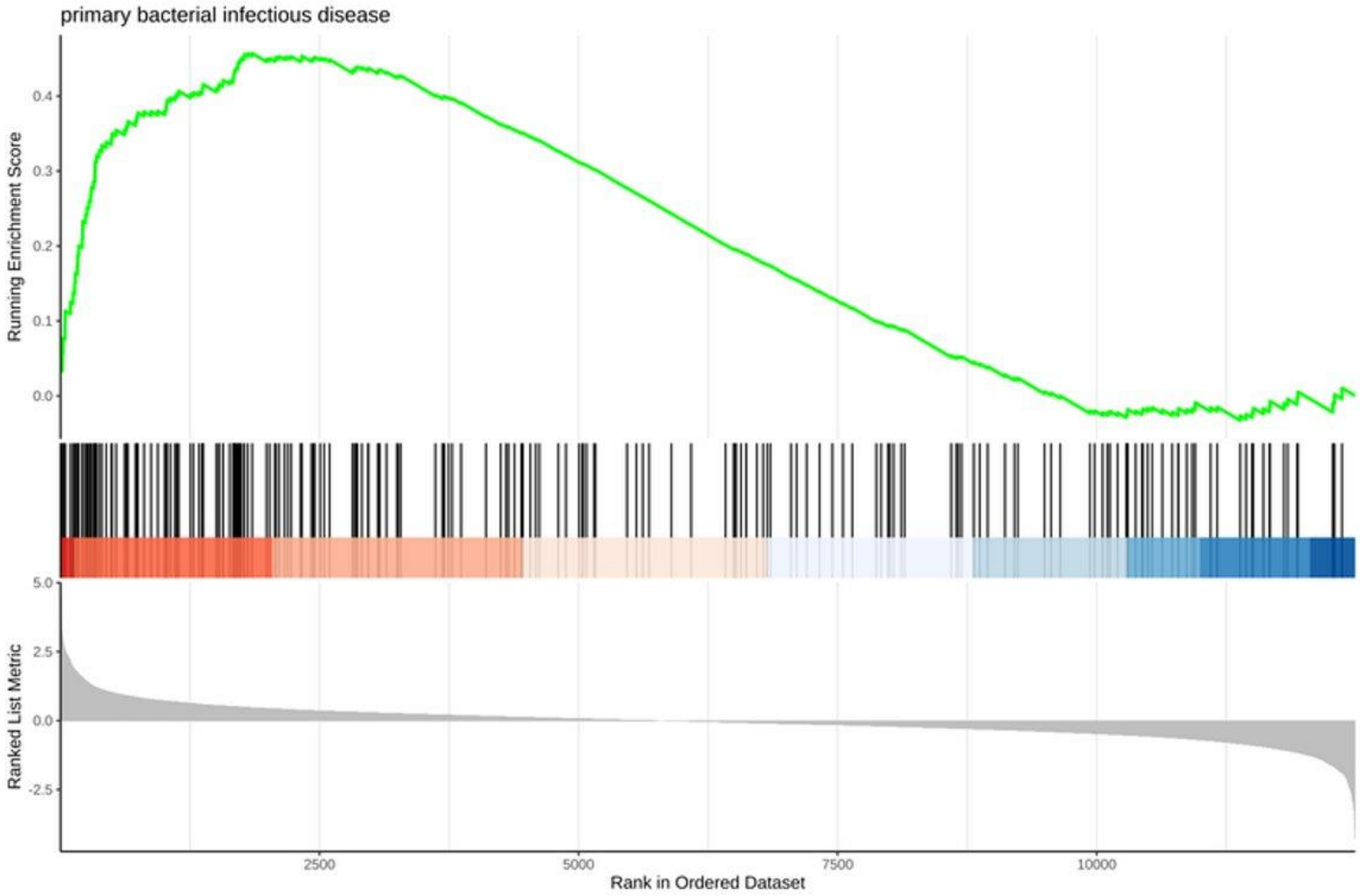
Enrichment analysis : computing enrichment

Gene Set Enrichment Analysis (GSEA)

- Does not rely on 0/1 DE
- Use ranking along a continuous measure (eg. log2FC)

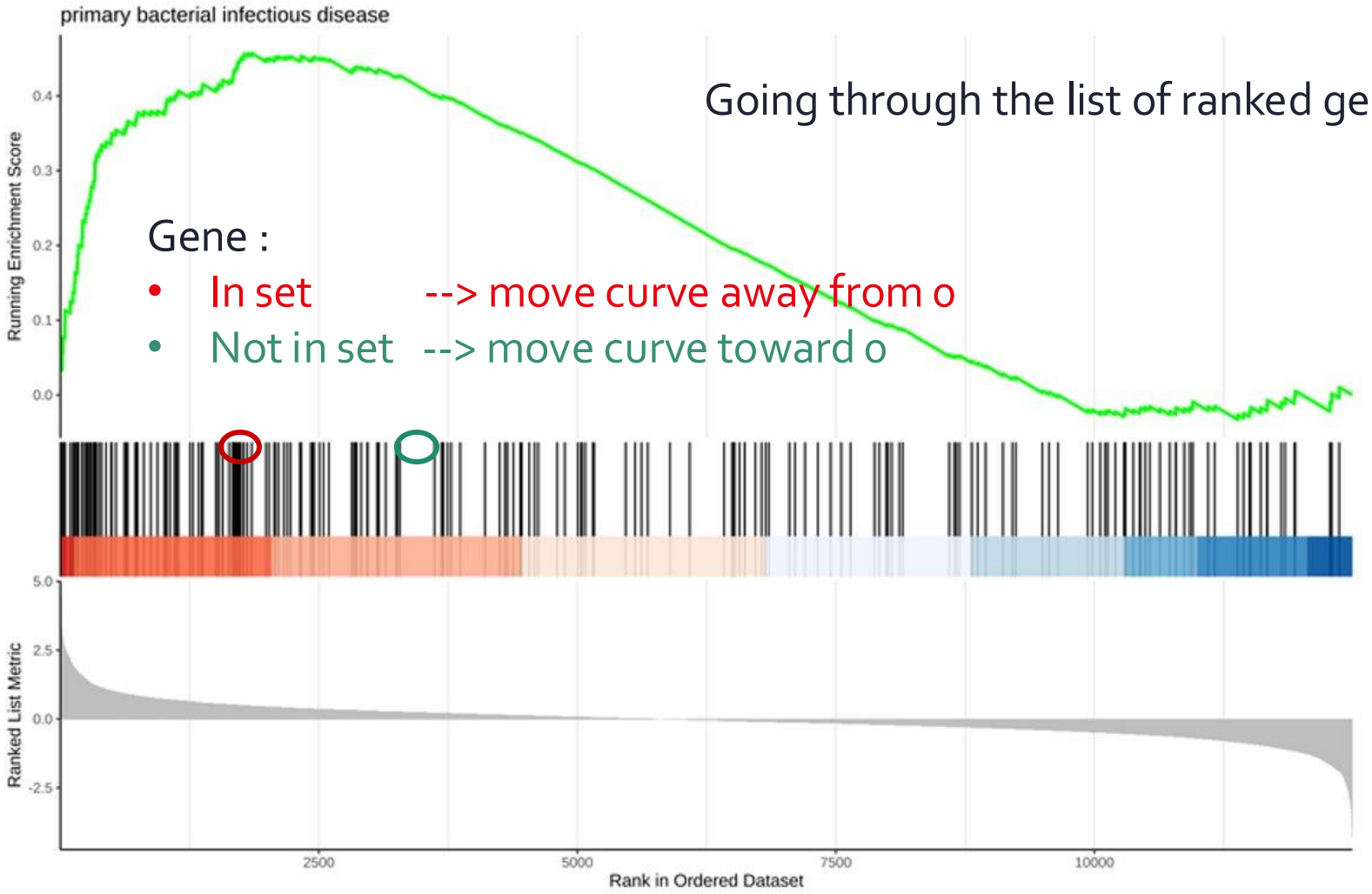
- Compute Enrichment Score
- Estimate significance with a permutation scheme

Enrichment analysis : computing enrichment

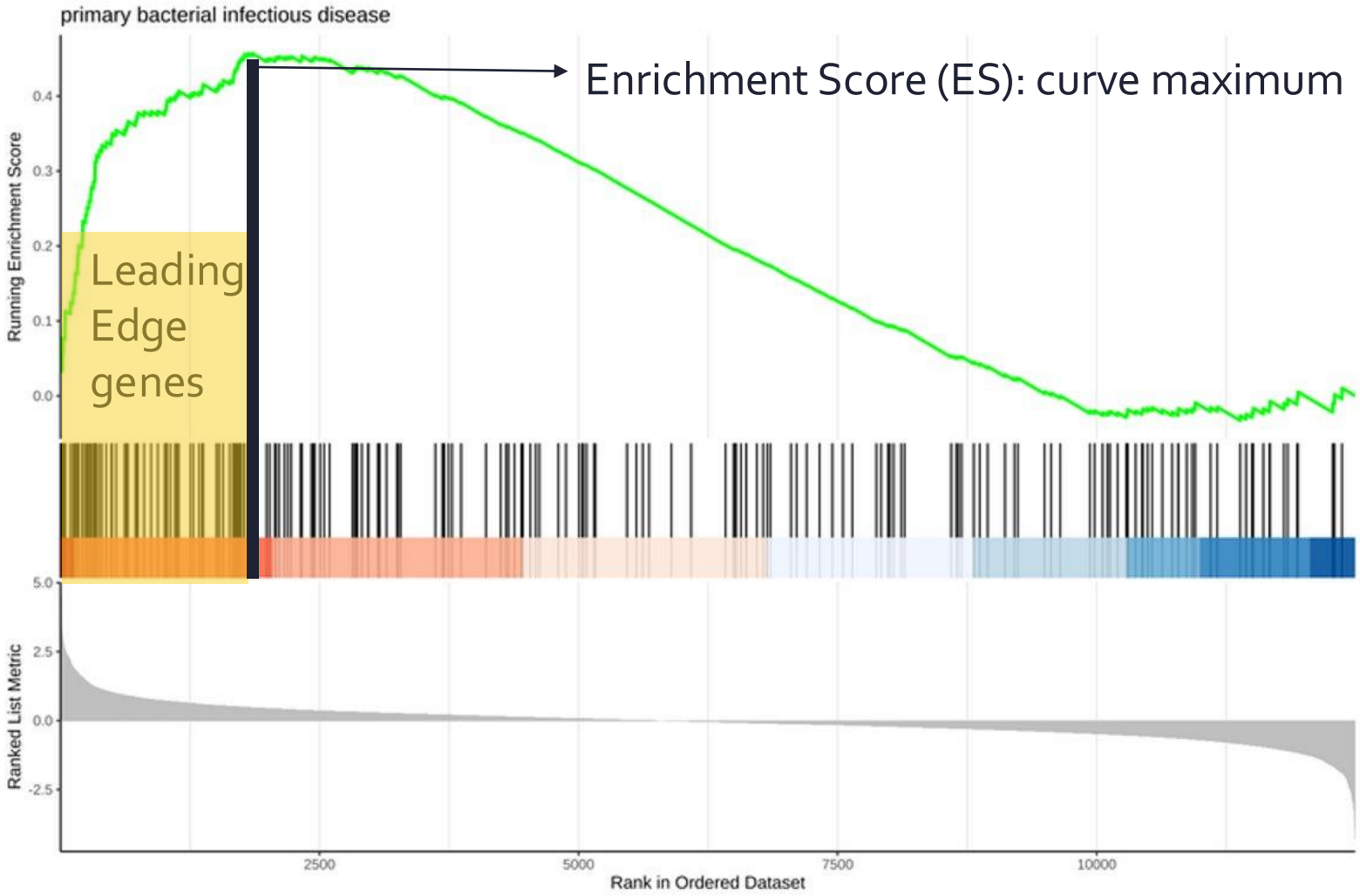


Enrichment Score visualized using functions from the R package enrichplot

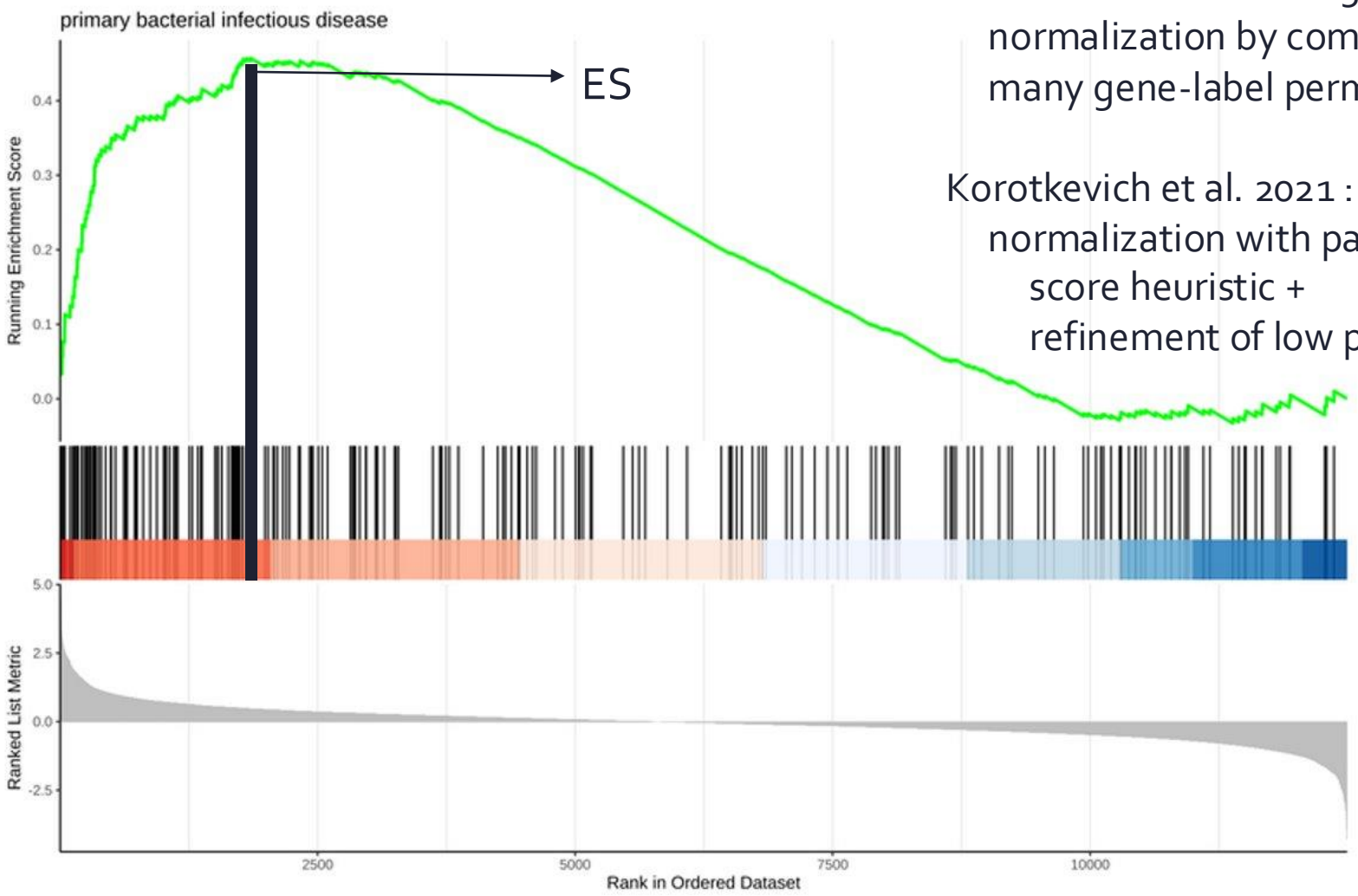
Enrichment analysis : computing enrichment



Enrichment analysis : computing enrichment



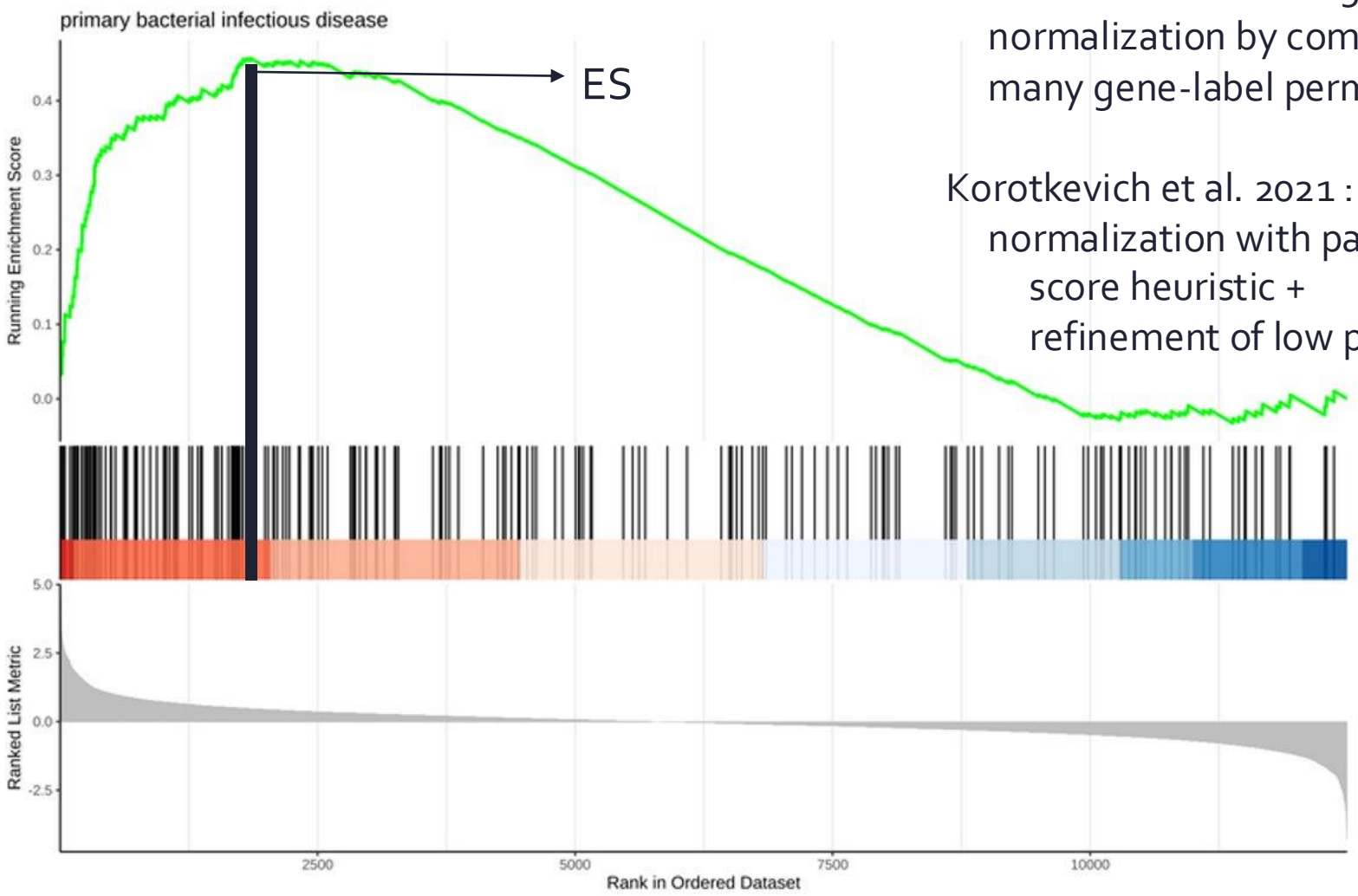
Enrichment analysis : computing enrichment



Subramanian et al. 2005:
normalization by comparison with many gene-label permuted sets

Korotkevich et al. 2021 : fGSEA
normalization with parallel evaluation +
score heuristic +
refinement of low p-values

Enrichment analysis : computing enrichment



Subramanian et al. 2005:
normalization by comparison with many gene-label permuted sets

Korotkevich et al. 2021 : fGSEA
normalization with parallel evaluation +
score heuristic +
refinement of low p-values

~100x faster
Default
in clusterProfiler

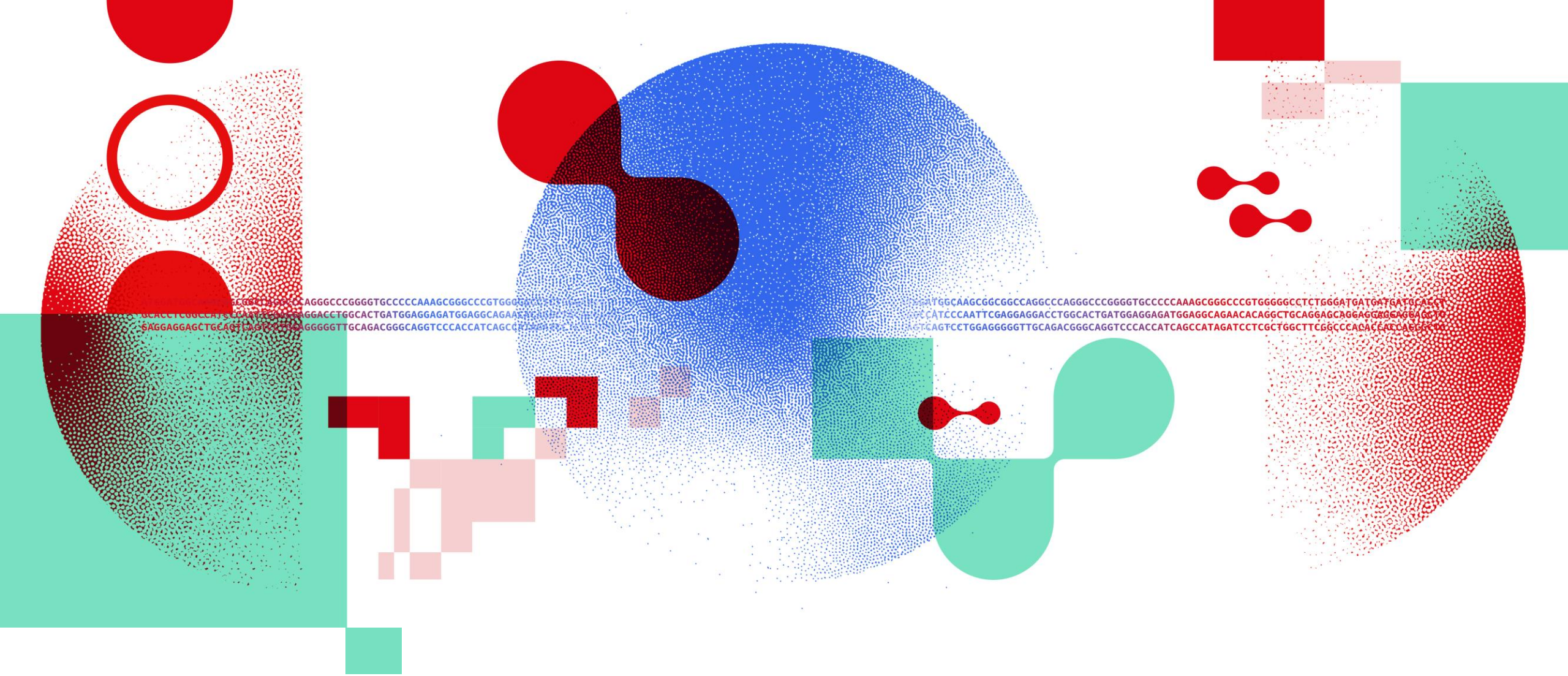
Enrichment analysis : computing enrichment

Many more methods or implementations

- Signaling Pathway Impact Analysis
  https://bioconductor.org/packages/release/bioc/html/SPIA.html

- ISMARA (TF-based)
  ismara.unibas.ch/mara

Practical



Thank you

elixir SWITZERLAND

SIB