GSA: Gene Set Analysis

Workshop on RNA-Seq

Nima Rafati

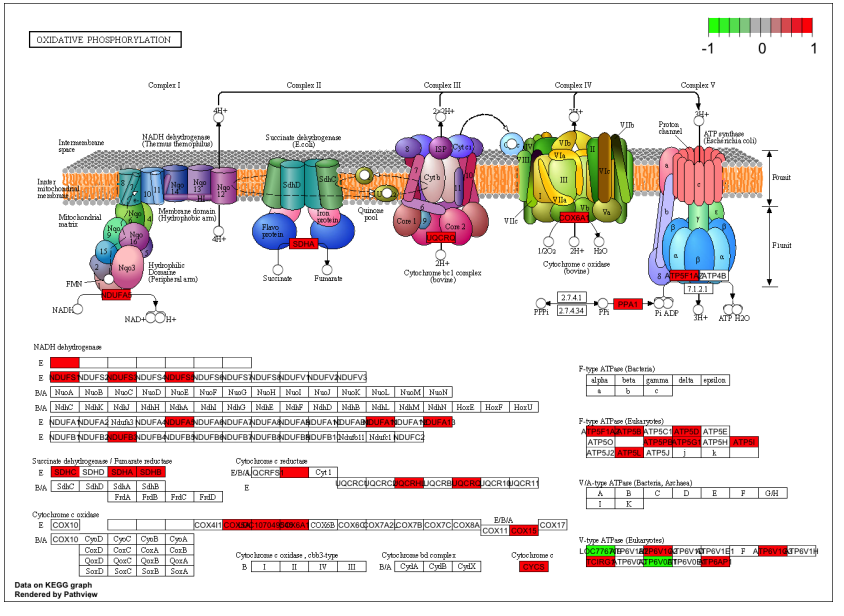
NBIS, SciLifeLab

Introduction

- What do the identified DEGs do?
- How can we link them to phenotypes/diseases/biological features we study?
- We can do that by exploring their function and in which pathways they are involved.
- While differential expression analysis identifies certain genes, is it feasible to manually explore the function of each gene?
- There are different approaches and dependent on available data we can expand it.
- At transcriptome level:
    - **Gene set analysis (GSA)**

Why GSA?

- Biological interpretation of the results; From gene list to biological insights!
- Reduce the complexity; Identifying key biological processes that are affected under the experiment or condition.
- Integrating external information.
- Cross-experiment comparisons; We can compare the results across different studies and experimental platforms.

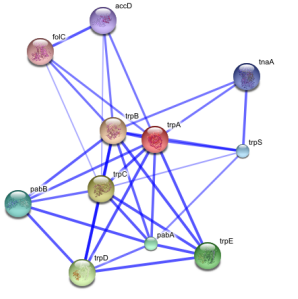


Gene set resources/databases

- **Gene Ontology (GO).**

- **Pathways (Kyoto Encyclopedia of Genes and Genomes (KEGG)).**

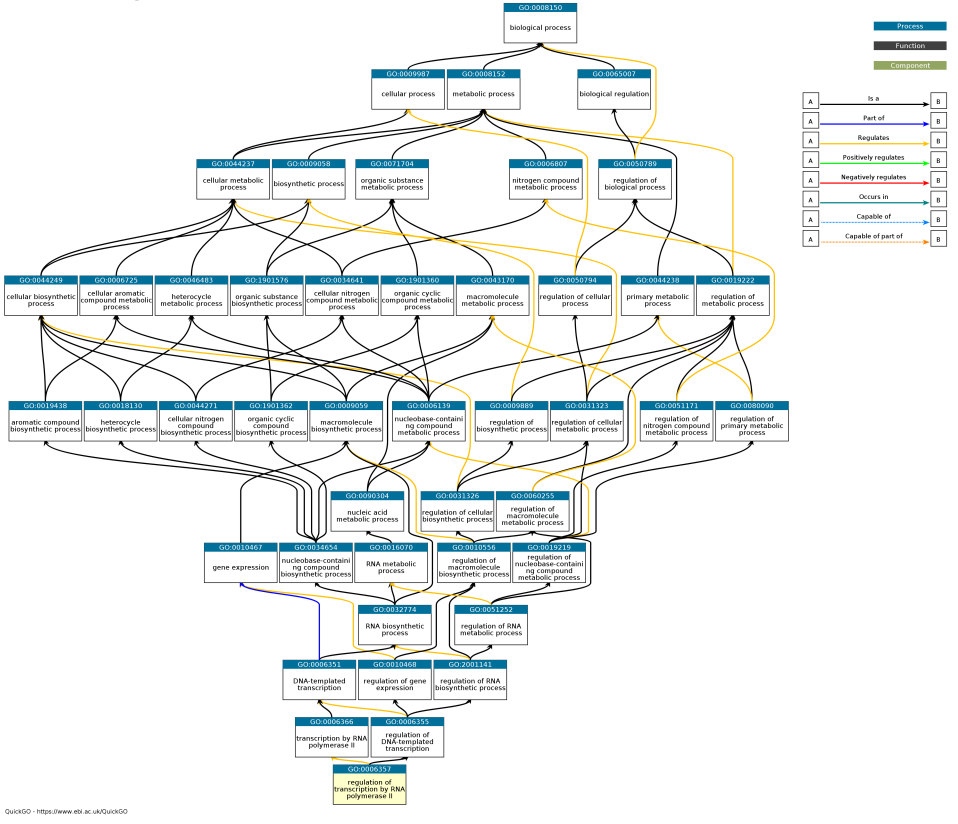
- Protein-protein interaction (PPI)



- Cell type
- Chromosomal location
- Metabolic and Signaling pathway.
- Diseases

Gene Ontology

- It is a resource to unify the representation of gene/gene products into hierarchical categories:
  - Biological Process (BP); *Cell cycle, Signal transduction*.
  - Molecular Function (MF); *Phosphorylation, DNA binding*.
  - Cellular Component (CC); *Nucleus, Cytoplasm*.
- Genes can belong to multiple GO *terms*

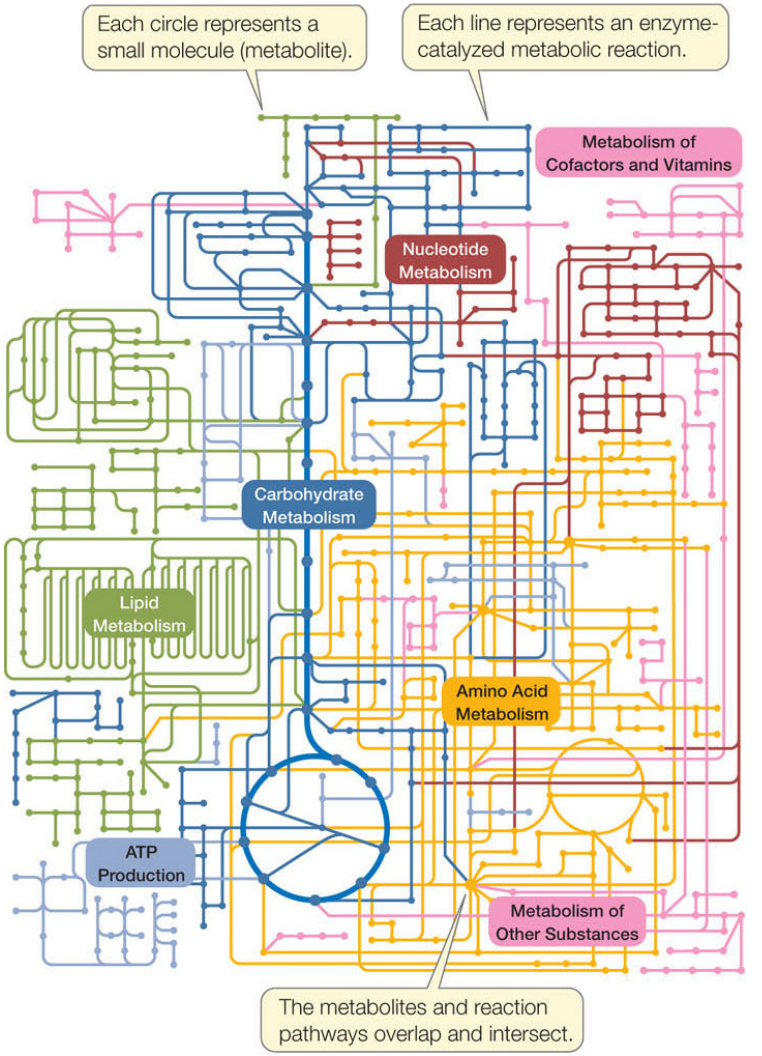


Pathway

Can you unravel the mystery of this pathway?

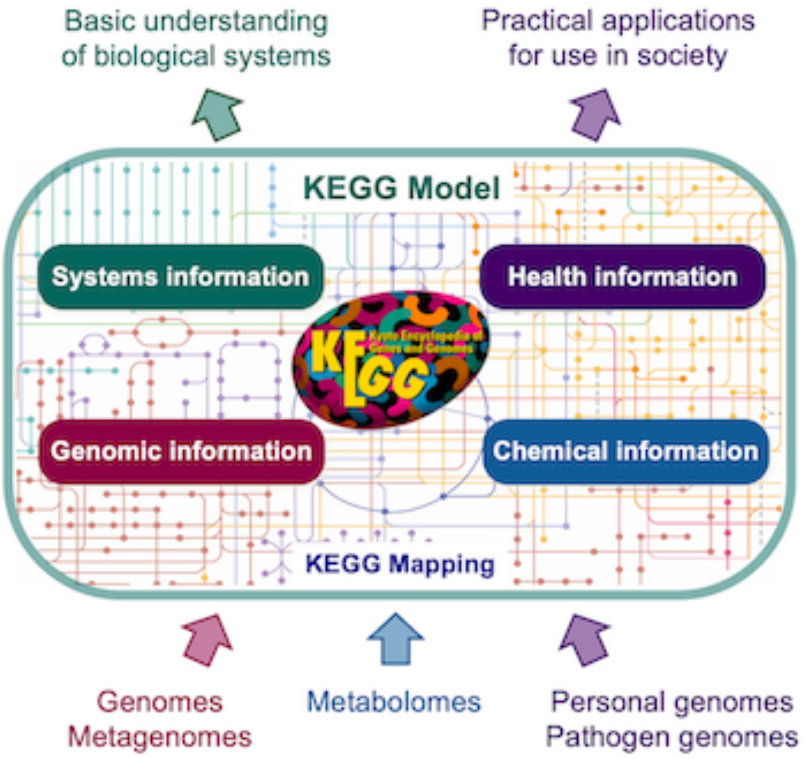
Pathway

- Biology is complex but has an organized structure.



KEGG

- KEGG is a comprehensive database resource that integrates genomic, chemical, and systemic functional information. It provides data on biological pathways, genomes, diseases, drugs, and chemical substances. KEGG is widely used for bioinformatics research, including the study of gene functions and networks.



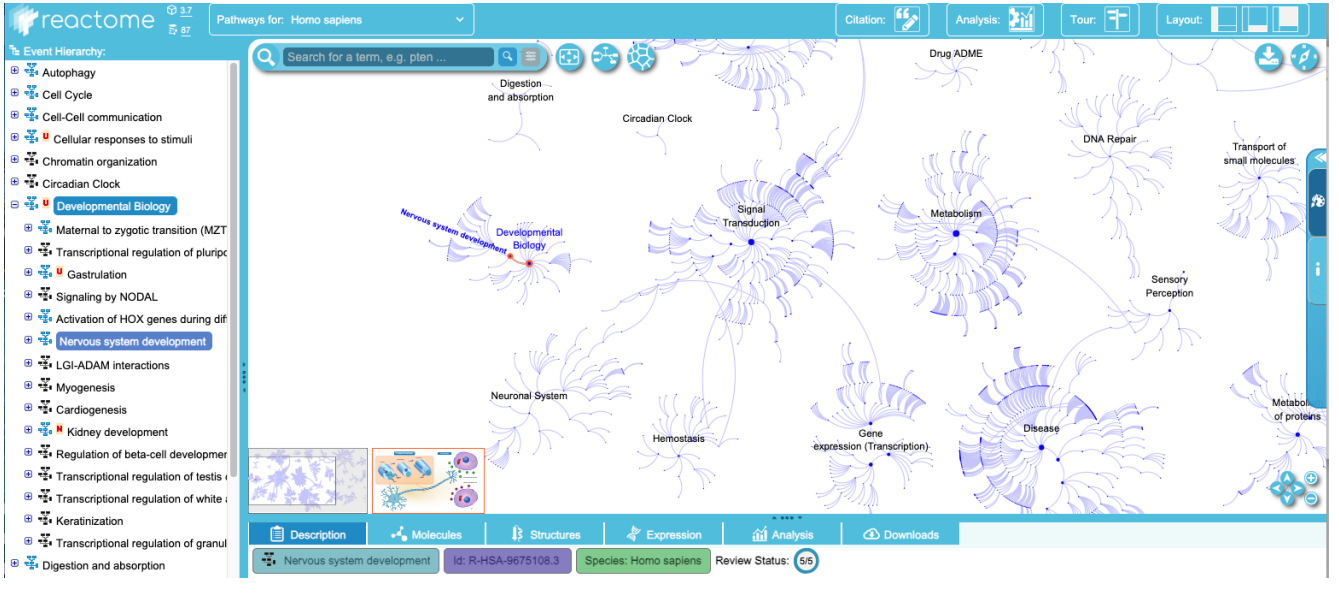
Wikipahtway

- WikiPathways is an open, collaborative platform dedicated to the curation of biological pathways.

- It allows scientists from various fields to contribute to and edit pathway information, offering a wide range of pathways for research and education purposes.
- The database facilitates the visualization and analysis of pathway information to support understanding of complex biological processes.

Reactome

- Reactome is a curated database of pathways and reactions in human biology.
- It covers various aspects of human biology, including metabolism, signaling, molecular transport, and cellular processes.
- Reactome provides tools for visualization, interpretation, and analysis of pathway data, making it a valuable resource for researchers in genomics and systems biology.

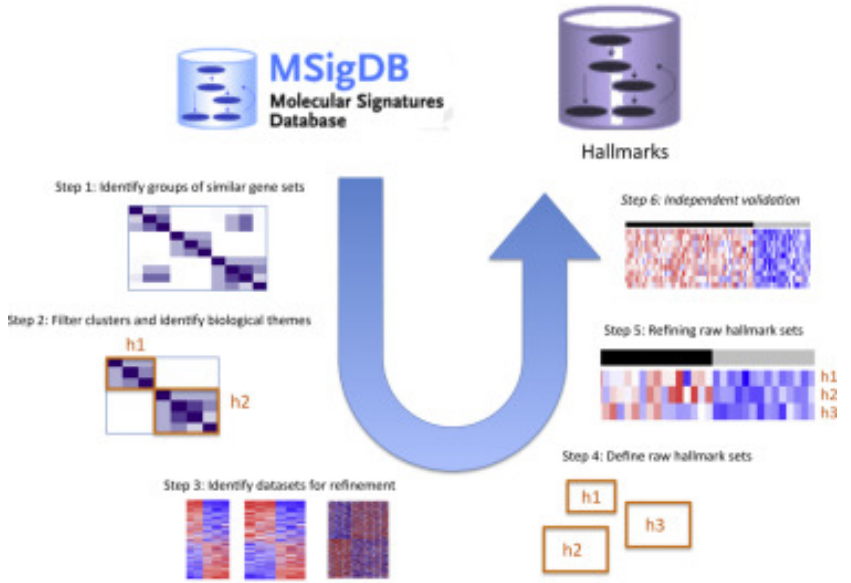


Transcription Factor databases

- There are different databases compile information about TF, their DNA binding sites and regulatory network they form:
    - TRANSFAC.
    - JASPAR.
    - ENCODE.

Hallmark Gene Set

- The Hallmark gene set is part of the Molecular Signatures Database (MSigDB), which is a collection of annotated gene sets for use with GSEA (Gene Set Enrichment Analysis) software.
- The Hallmark gene set distills complex gene signatures into a concise set of gene sets that represent specific and well-defined biological states or processes.
- These gene sets are designed to be universally applicable for annotating gene expression patterns in a wide variety of biological contexts.



Where to get gene sets for the analyses?

Enrichr

Login | Register

68,685,026 sets analyzed
494,081 terms
225 libraries

Analyze | What's new? | Libraries | Gene search | Term search | About | Help

Gene-set Library	Terms	Gene Coverage	Genes per Term
Achilles_fitness_decrease	216	4271	128
Achilles_fitness_increase	216	4320	129
Aging_Perturbations_from_GEO_down	286	16129	292
Aging_Perturbations_from_GEO_up	286	15309	308
Allen_Brain_Atlas_10x_scRNA_2021	766	12361	124
Allen_Brain_Atlas_down	2192	13877	304
Allen_Brain_Atlas_up	2192	13121	305
ARCHS4_Cell-lines	125	23601	2395
ARCHS4_IDG_Coexp	352	20883	299
ARCHS4_Kinases_Coexp	498	19612	299
ARCHS4_TFs_Coexp	1724	25983	299
ARCHS4_Tissues	108	21809	2316
Azimuth_2023	1425	3712	9
Azimuth_Cell_Types_2021	341	1683	10
BioCarta_2013	249	1295	18
BioCarta_2015	239	1678	21
BioCarta_2016	237	1348	19
BioPlanet_2019	1510	9813	49
BioPlex_2017	3915	10271	22
Cancer_Cell_Line_Encyclopedia	967	15797	176
CCLE_Proteomics_2020	378	11851	586
CellMarker_Augmented_2021	1097	14167	80
ChEA_2013	353	47172	1370
ChEA_2015	395	48230	1429
ChEA_2016	645	49238	1550
ChEA_2022	757	18365	1214
Chromosome_Location	386	32740	85

MSigDB

Molecular Signatures Database

Molecular Signatures Database

Overview

The Molecular Signatures Database (MSigDB) is a resource of tens of thousands of annotated gene sets for use with GSEA software, divided into Human and Mouse collections. From this web site, you can

- **Examine** a gene set and its annotations. See, for example, the HALLMARK_APOPTOSIS human gene set page.
- **Browse** gene sets by name or collection.
- **Search** for gene sets by keyword.
- **Investigate** gene sets:
  - **Compute overlaps** between your gene set and gene sets in MSigDB.
  - **Categorize** members of a gene set by gene families.
  - **View the expression profile** of a gene set in a provided public expression compendia.
  - Investigate the gene set in the online **biological network repository NDEx**
- **Download** gene sets.

License Terms

GSEA and MSigDB are available for use under these license terms.

Please register to download the GSEA software and the MSigDB gene sets, and to use our web tools. After registering, you can log in at any time using your email address. Registration is free. Its only purpose is to help us track usage for reports to our funding agencies.

Current Version

Human MSigDB v2023.2.Hs updated October 2023. Release notes.

Mouse MSigDB v2023.2.Mm updated October 2023. Release notes.

Citing the MSigDB

To cite your use of the Molecular Signatures Database (MSigDB), a joint project of UC San Diego and Broad Institute, please reference Subramanian, Tamayo, et

Human Collections

- **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** corresponding to human chromosome cytogenetic bands.
- **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 expression 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.

Mouse Collections

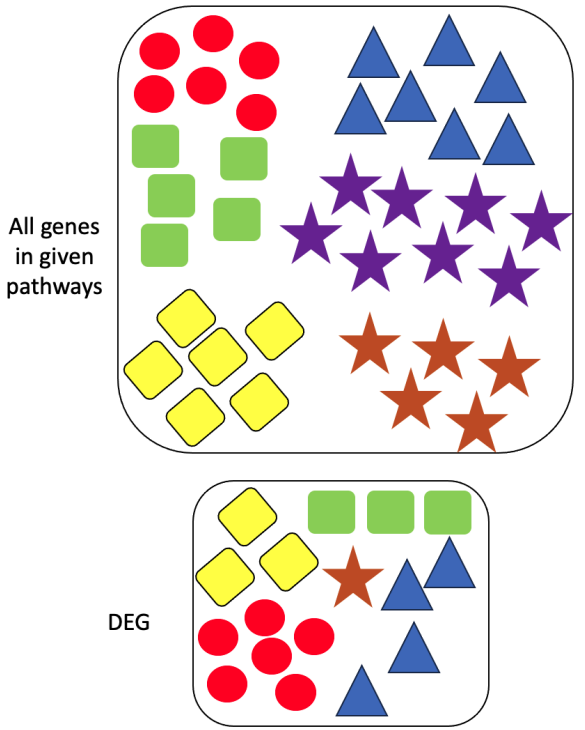
- **MH** **mouse-ortholog hallmark gene sets** are versions of gene sets in the MSigDB Hallmarks collection mapped to their mouse orthologs.
- **M1** **positional gene sets** corresponding to mouse chromosome cytogenetic bands.
- **M2** **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.
- **M3** **regulatory target gene sets** based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.
- **M5** **ontology gene sets** consist of genes annotated by the same ontology term.
- **M8** **cell type signature gene sets** curated from cluster markers identified in single-cell sequencing studies of mouse tissue.

Gene set analysis methods

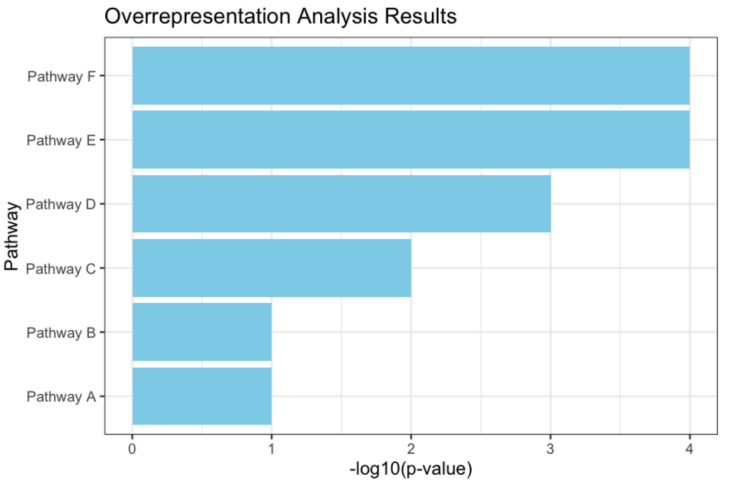
- Overrepresentation analysis (ORA):
  - A statistical method for identifying terms (e.g. GO terms or pahtways) that are more represented in a given gene/protein set than expected by chance.
- Gene Set Enrichment Analysis (GSEA):
  - A statistical method for evaluating the distribution of genes across a ranked list of genes showing the same signature (upregulated or downregulated) which happen to be involved in a given category (e.g. pathway).

ORA

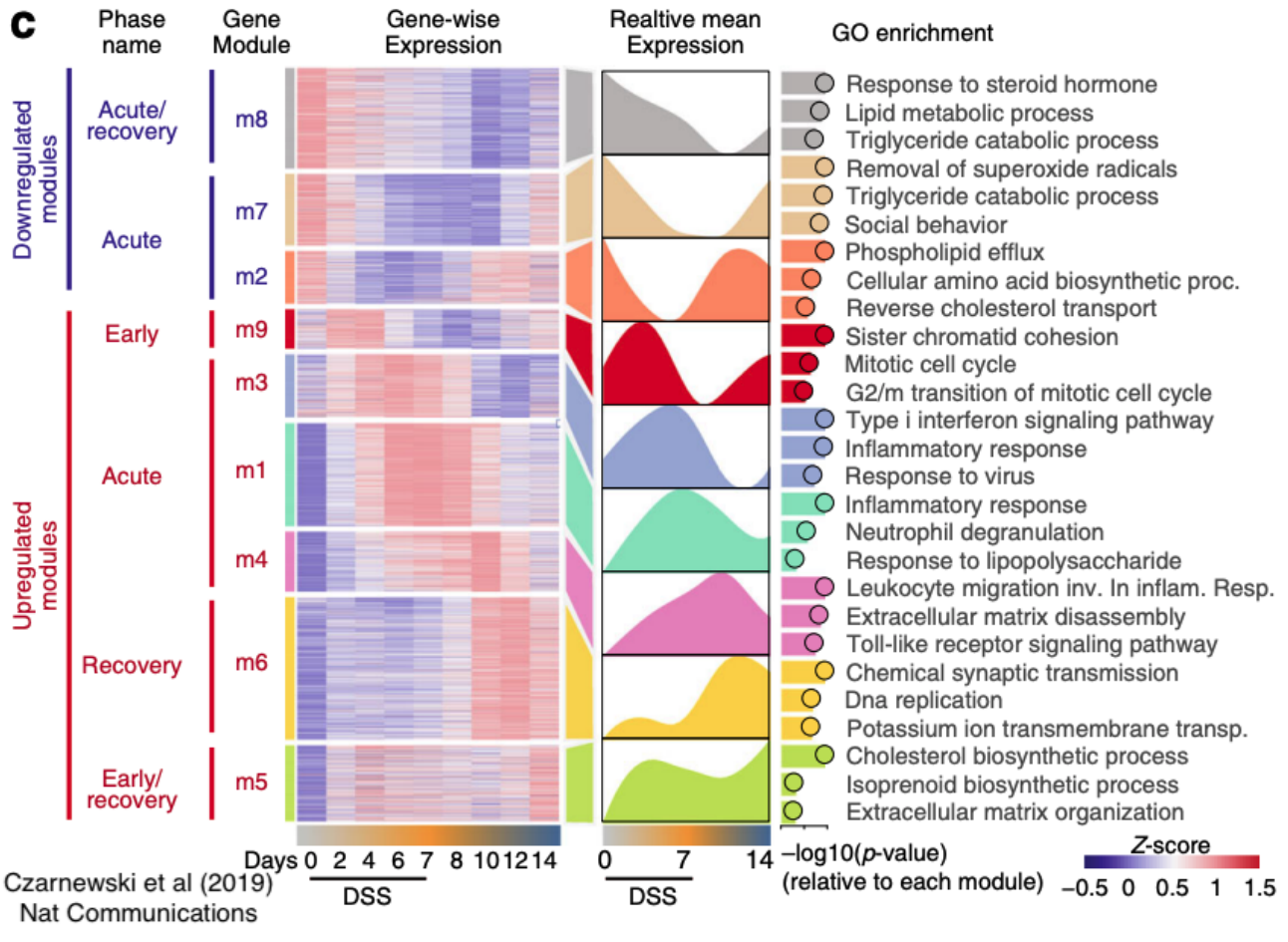
- It is a hypergeometric test (Fisher's exact test)
- Selected genes are differentially expressed genes (DEG: Up or Down)
- Category can be GO, Pathway,....



	Selected	Not selected
In a category	10	2
Not In a category	90	14000

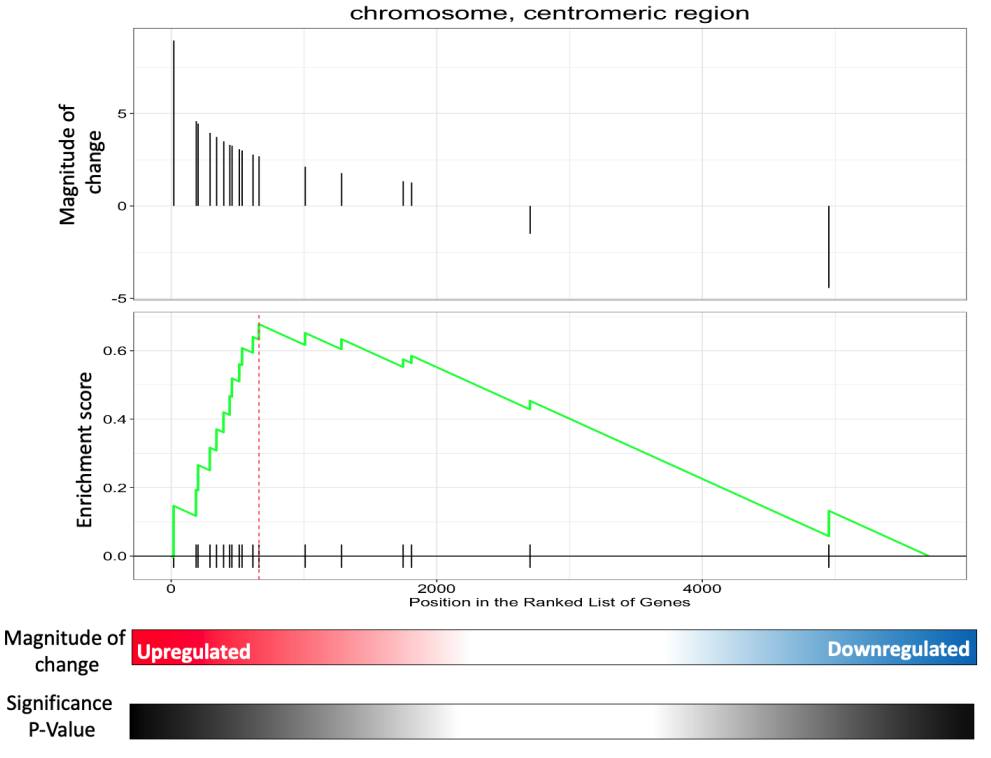


ORA


c
Phase name
Gene Module
Gene-wise Expression
Realtive mean Expression
GO enrichment
Downregulated modules
Upregulated modules
Acute/ recovery
Acute
Early
Acute
Recovery
Early/ recovery
m8
m7
m2
m9
m3
m1
m4
m6
m5
Response to steroid hormone
Lipid metabolic process
Triglyceride catabolic process
Removal of superoxide radicals
Triglyceride catabolic process
Social behavior
Phospholipid efflux
Cellular amino acid biosynthetic proc.
Reverse cholesterol transport
Sister chromatid cohesion
Mitotic cell cycle
G2/m transition of mitotic cell cycle
Type i interferon signaling pathway
Inflammatory response
Response to virus
Inflammatory response
Neutrophil degranulation
Response to lipopolysaccharide
Leukocyte migration inv. In inflam. Resp.
Extracellular matrix disassembly
Toll-like receptor signaling pathway
Chemical synaptic transmission
Dna replication
Potassium ion transmembrane transp.
Cholesterol biosynthetic process
Isoprenoid biosynthetic process
Extracellular matrix organization
Days 0 2 4 6 7 8 10 12 14
DSS
0 7 14
DSS
−log10(p-value)
(relative to each module)
Z-score
−0.5 0 0.5 1 1.5
Czarnewski et al (2019)
Nat Communications

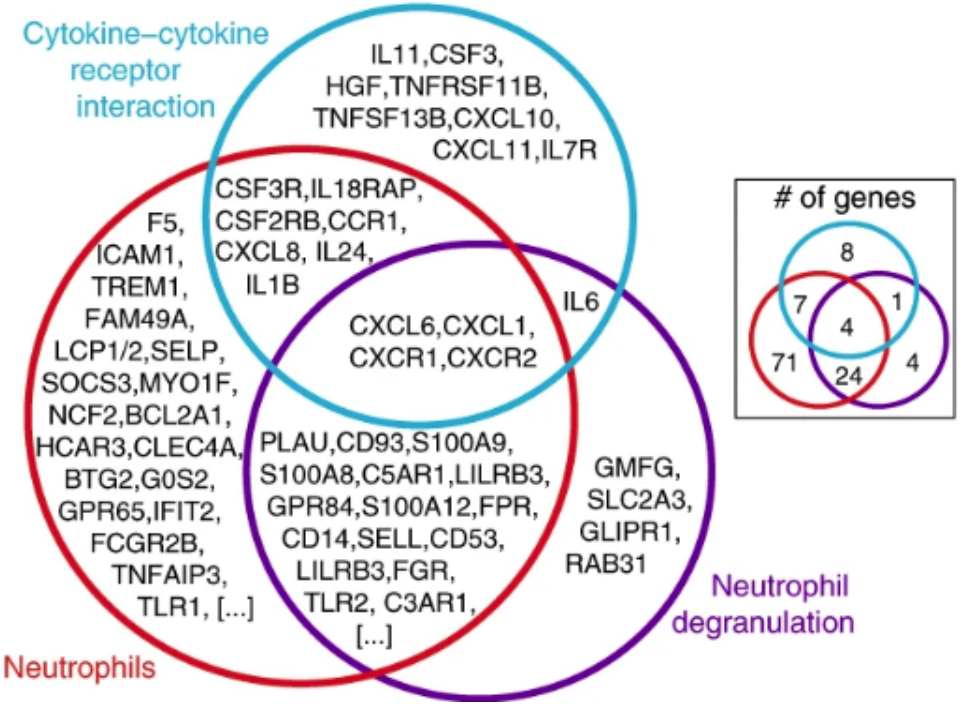
GSEA

- In GSEA we do not have any prior selection of the genes (such as DEG)
- Genes are listed by logFC and their distribution is tested with a statistical test adapted from Kolmogrov-smirinov test. This test calculates an enrichment score (ES) for each predefined gene set which reflects the degree to which the genes in the set are overrepresented at the extremes (top or bottom) of the ranked list. In other words, it tries to identify maximum deviation form zero.



GSEA

- Few notes:
  - The ES differ among tested pathways/terms. Thus use Normalized Enrichment Score (NES).
  - Some genes may be involved in different pathways and thus can bias interpretation.
  - As an alternative, topology-based method has been introduced which takes gene-set interaction into account. (Ma et al., 2019).



Thank you. Questions?