

GSA: Gene Set Analysis

RNA-seq data analysis

Paulo Czarnewski

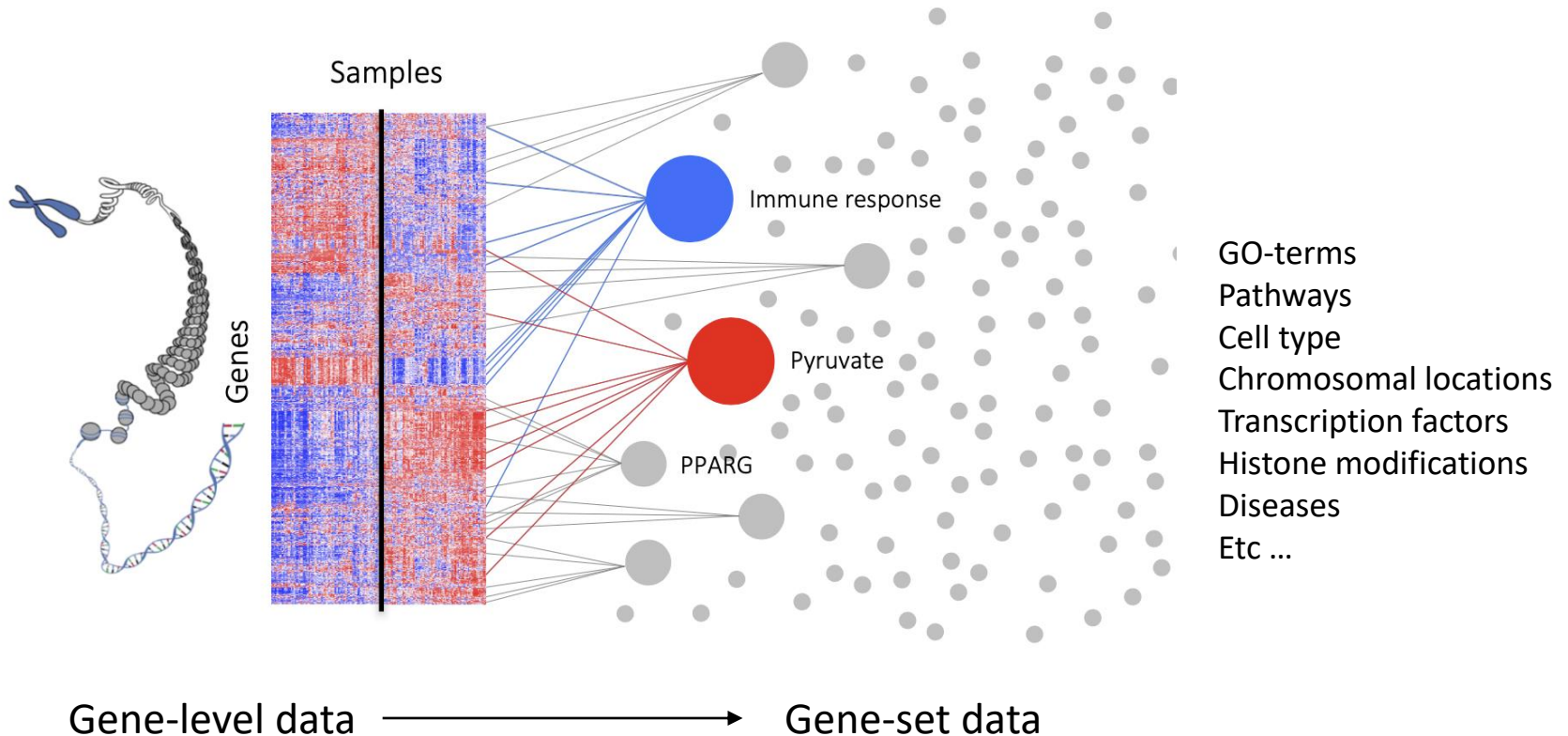
Leif Wigge

<https://czarnewski.github.io/czarnewski/index.html>



What is gene set analysis (GSA)?

WHAT is gene set analysis (GSA)?



We will focus on transcriptomics and differential expression analysis
However, GSA can in principle be used on all types of genome-wide data

WHY gene set analysis (GSA)?

- Interpretation of genome-wide results
- Gene-sets are (typically) fewer than all the genes and have more descriptive names
- Difficult to manage a long list of significant genes
- Detect patterns that would be difficult to discern simply by manually going through *e.g. the list of differentially expressed genes*
- Top genes might not be the interesting ones, several coordinated smaller changes
- Integrates external information into the analysis
- Less prone to false-positives on the gene-level

Gene sets

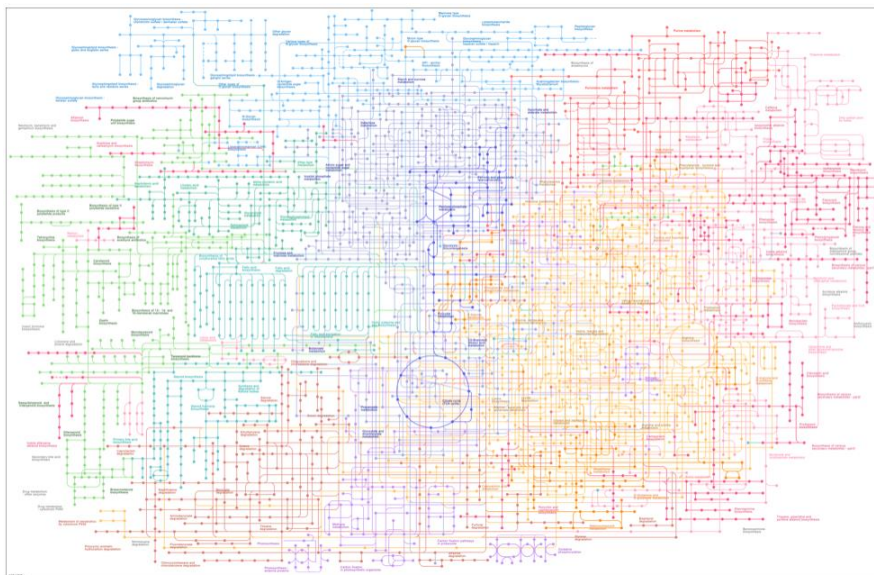
Which gene sets should I use?

- Depends on the research question
- Several databases/resources available providing gene-set collections
e.g. MSigDB, Enrichr, Panther
- Included directly in some analysis tools
- GO-terms are probably one of the most widely used gene-sets

GO-terms
Pathways
Cell type
Chromosomal locations
Transcription factors
Histone modifications
Diseases
Etc ...

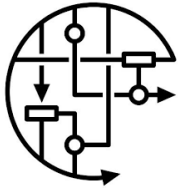


Metabolic Pathways



Category	Database	Content	Color
Systems information	KEGG PATHWAY	KEGG pathway maps	KEGG
	KEGG BRITE	BRITE hierarchies and tables	
	KEGG MODULE	KEGG modules and reaction modules	
Genomic information	KEGG ORTHOLOGY (KO)	Functional orthologs	KEGG
	KEGG GENOME	KEGG organisms and viruses	
	KEGG GENES	Genes and proteins	
	KEGG SSDB	GENES sequence similarity	
Chemical information	KEGG COMPOUND	Small molecules	KEGG
	KEGG GLYCAN	Glycans	
	KEGG REACTION / RCLASS	Reactions and reaction class	
Health information	KEGG ENZYME	Enzyme nomenclature	KEGG
	KEGG NETWORK	Disease-related network variations	
	KEGG VARIANT	Human gene variants	
	KEGG DISEASE	Human diseases	
	KEGG DRUG / DGROUP	Drugs and drug groups	
	KEGG ENVIRON	Health-related substances	

KEGG is an integrated database resource consisting of eighteen databases (including computationally generated SSDB) shown below. They are broadly categorized into systems information, genomic information, chemical information and health information, which are distinguished by color coding of web pages.



page discussion view source history Log in / create account

Share your pathway knowledge in the fight against COVID-19

ACCESS the rapidly growing [collection of COVID-19 pathways](#), **CONTRIBUTE** your time and domain knowledge about pathway biology as a [pathway author](#), and **USE** these pathways in [your research](#).

Welcome to WikiPathways

WikiPathways is a database of biological pathways maintained by and for the scientific community.

Read about our 12-year journey so far and [official exit from beta](#).

Find Pathways

Search

Search

You can search by:

- Pathway name (*Apoptosis*)
- Gene or protein name (*p53*)
- Any page content (*cancer*)

Browse

Browse pathways

Browse by species and category

Get Pathways

Today's Featured Pathway

Host-pathogen interaction of human corona viruses - Interferon induction (Homo sapiens)

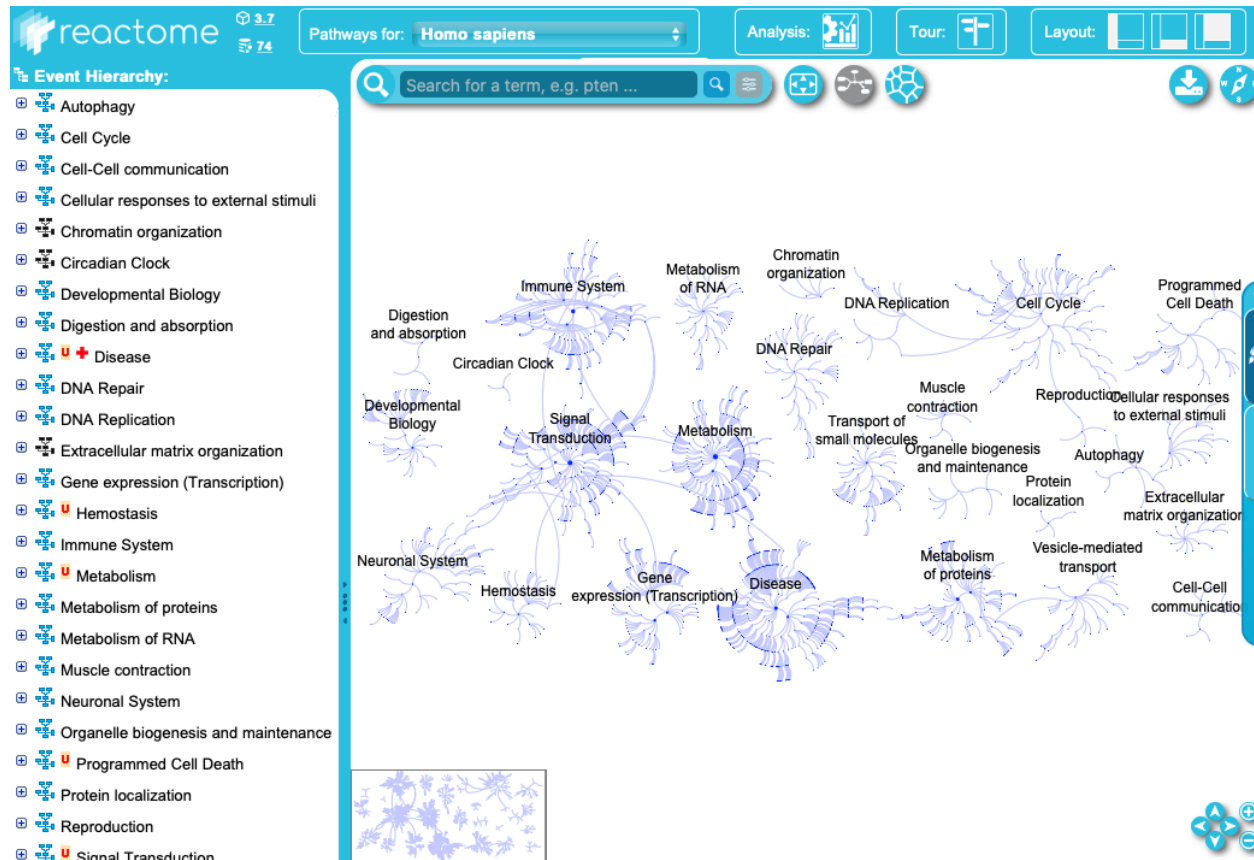
Host-pathogen interaction of human corona viruses - Interferon induction

Curator of the Week

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

Building on the same MediaWiki software that powers Wikipedia, we added a custom graphical pathway editing tool and integrated databases covering major gene, protein, and small-molecule systems.

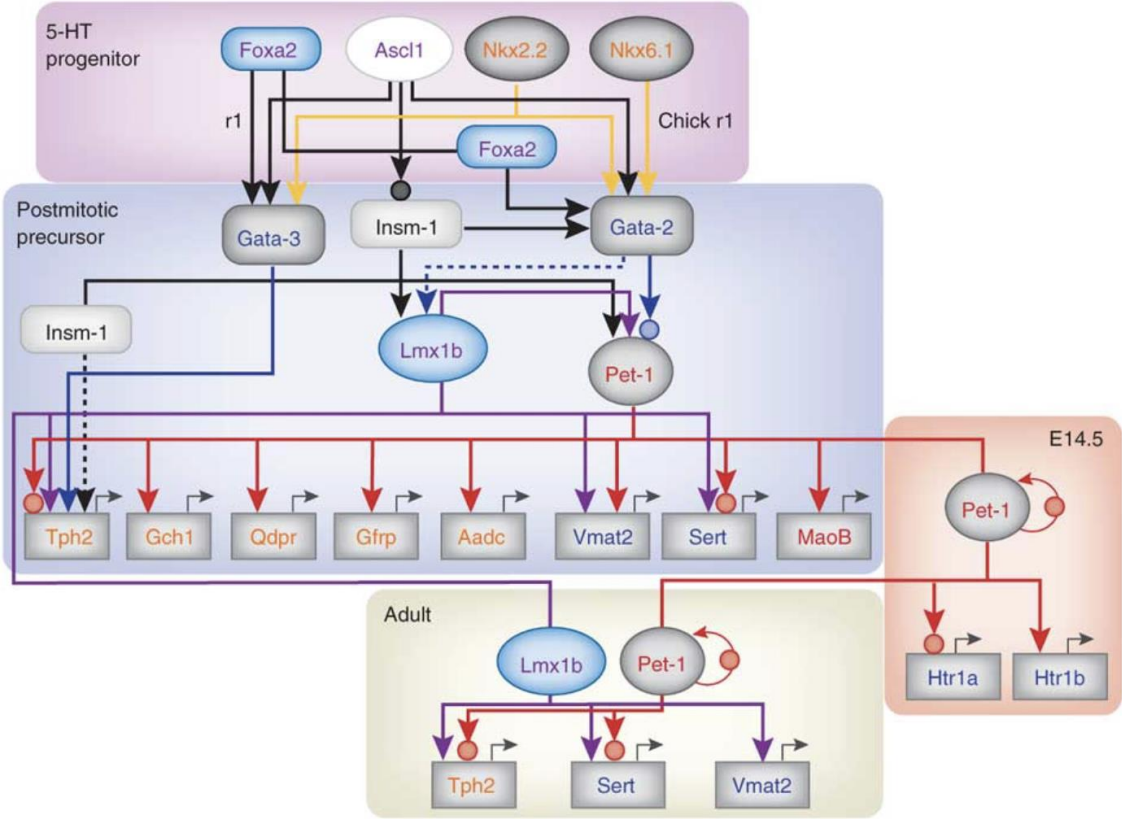
Slenter et al (2018) *Nucleic Acid Research*

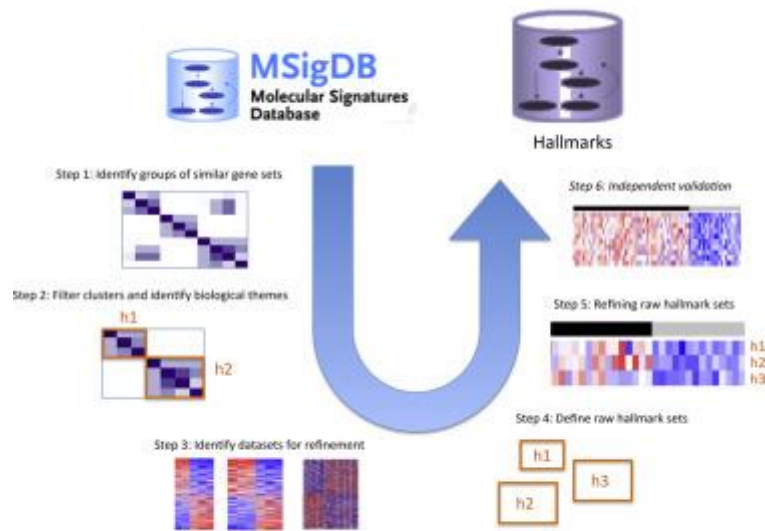


Reactome is a free, open-source, curated and peer-reviewed pathway database. Our goal is to provide intuitive bioinformatics tools for the visualization, interpretation and analysis of pathway knowledge to support basic research, genome analysis, modeling, systems biology and education.

Jassal et al (2020) *Nucleic Acid Research*

Transcription Factor





- Each hallmark in this collection consists of a “refined” gene set, derived from multiple “founder” sets, that conveys a specific biological state or process and displays coherent expression.
- The hallmarks effectively summarize most of the relevant information of the original founder sets and, by reducing both variation and redundancy, provide more refined and concise inputs for gene set enrichment analysis.

Where to get gene set collections?



Molecular Signatures Database v7.2

Molecular Signatures Database

Overview

The Molecular Signatures Database (MSigDB) is a collection of annotated gene sets for use with GSEA software. From this web site, you can

- ▶ **Search** for gene sets by keyword.
- ▶ **Browse** gene sets by name or collection.
- ▶ **Examine** a gene set and its annotations. See, for example, the HALLMARK_APOPTOSIS gene set page.
- ▶ **Download** gene sets.
- ▶ **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**

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

MSigDB database v7.2 updated September 2020. [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 al. (2005, PNAS 102, 15545-15550) and one or more of the following as appropriate: Liberzon, et al. (2011, Bioinformatics), Liberzon, et al. (2015, Cell Systems), and also the source for the gene set as listed on the gene set page.

Collections

The MSigDB gene sets are divided into 9 major 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** 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** defined directly from microarray gene expression data from immunologic studies.

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

Other Gene Set Resources

- ▶ **Signatures of post-translational modification (PTM) sites** from the Proteomics group at the Broad Institute
- ▶ **Miscellaneous gene sets** from community contributors.

<https://www.gsea-msigdb.org/gsea/msigdb>



Enrichr

[Login](#) | [Register](#)

29,520,225 lists analyzed
338,361 terms
170 libraries

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

Gene-set Library	Terms	Gene Coverage	Genes per Term	
Genes_Associated_with_NIH_Grants	32876	15886	9	↓
Cancer_Cell_Line_Encyclopedia	967	15797	176	↓
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_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	↓
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	↓
CLE_Proteomics_2020	378	11851	586	↓
ChEA_2013	353	47172	1370	↓
ChEA_2015	395	48230	1429	↓
ChEA_2016	645	49238	1550	↓
Chromosome_Location	386	32740	85	↓
Chromosome_Location_hg19	36	27360	802	↓
ClinVar_2019	182	1397	13	↓
CORUM	1658	2741	5	↓
COVID-19_Related_Gene_Sets	205	16979	295	↓
Data_Acquisition_Method_Most_Popular_Genes	12	1073	100	↓
dbGaP	345	5613	36	↓
DepMap_WG_CRISPR_Screens_Broad_CellLines_2019	558	7744	363	↓
DepMap_WG_CRISPR_Screens_Sanger_CellLines_2019	325	6204	387	↓
Disease_Perturbations_from_GEO_down	839	23939	293	↓
Disease_Perturbations_from_GEO_up	839	23561	307	↓
Disease_Signatures_from_GEO_down_2014	142	15406	300	↓

<https://maayanlab.cloud/Enrichr/#stats>

Gene set analysis methods

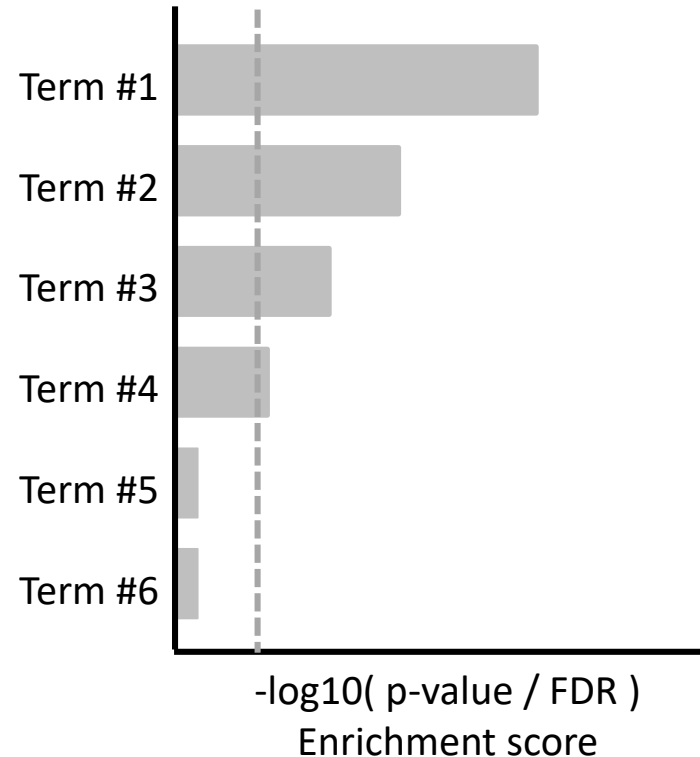
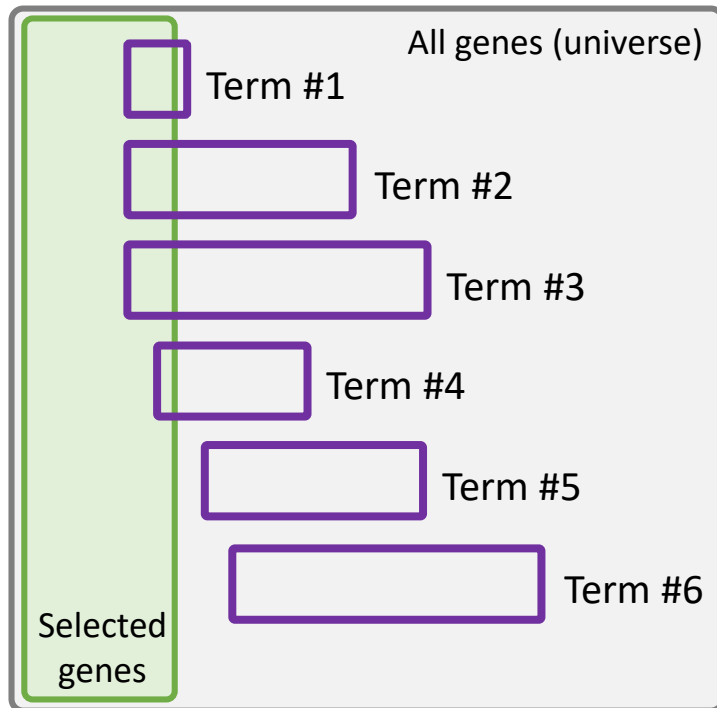
Overrepresentation analysis

Hypergeometric test (Fisher's exact test)

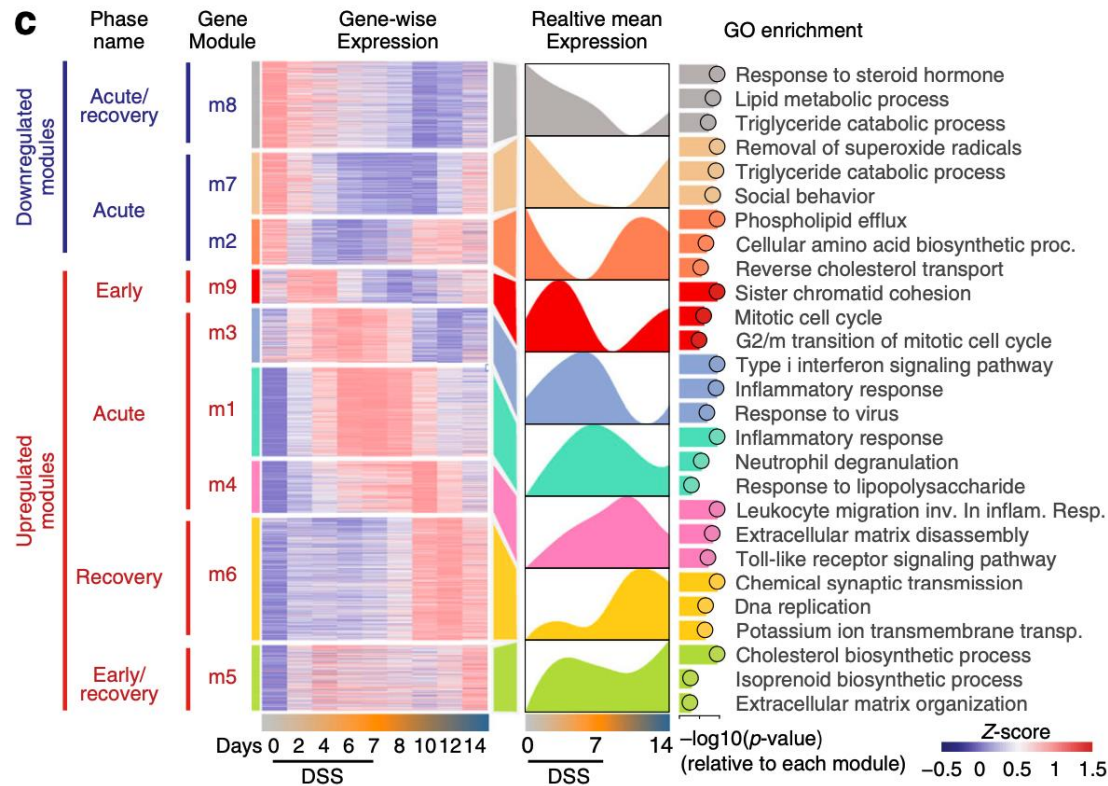
Uses a list of genes:

- Differentially expressed genes (UP or DOWN)
- List of genes in a cluster / module

	selected	not selected
in GO-term	8	2
not in GO-term	92	19768



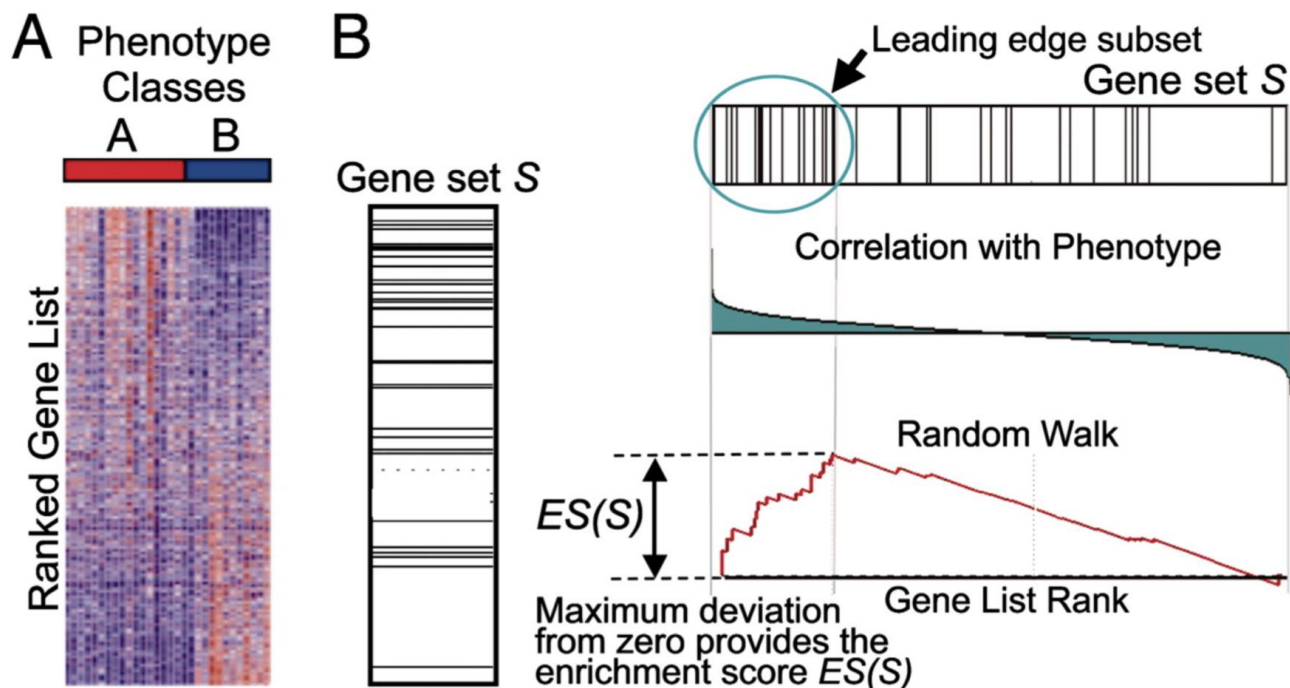
Overrepresentation analysis



Czarnecki et al (2019) Nat Communications

Gene set enrichment analysis

2 sample comparison



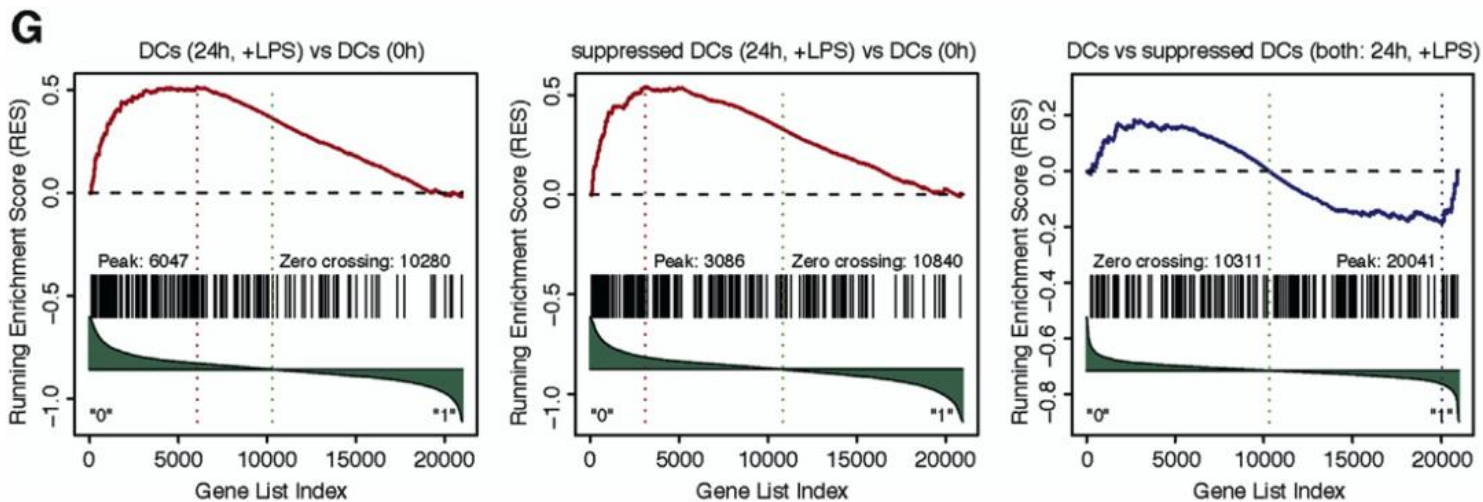
Mootha et al (2003) Nature Genetics
Subramanian et al (2005) PNAS

Gene set enrichment analysis

enriched

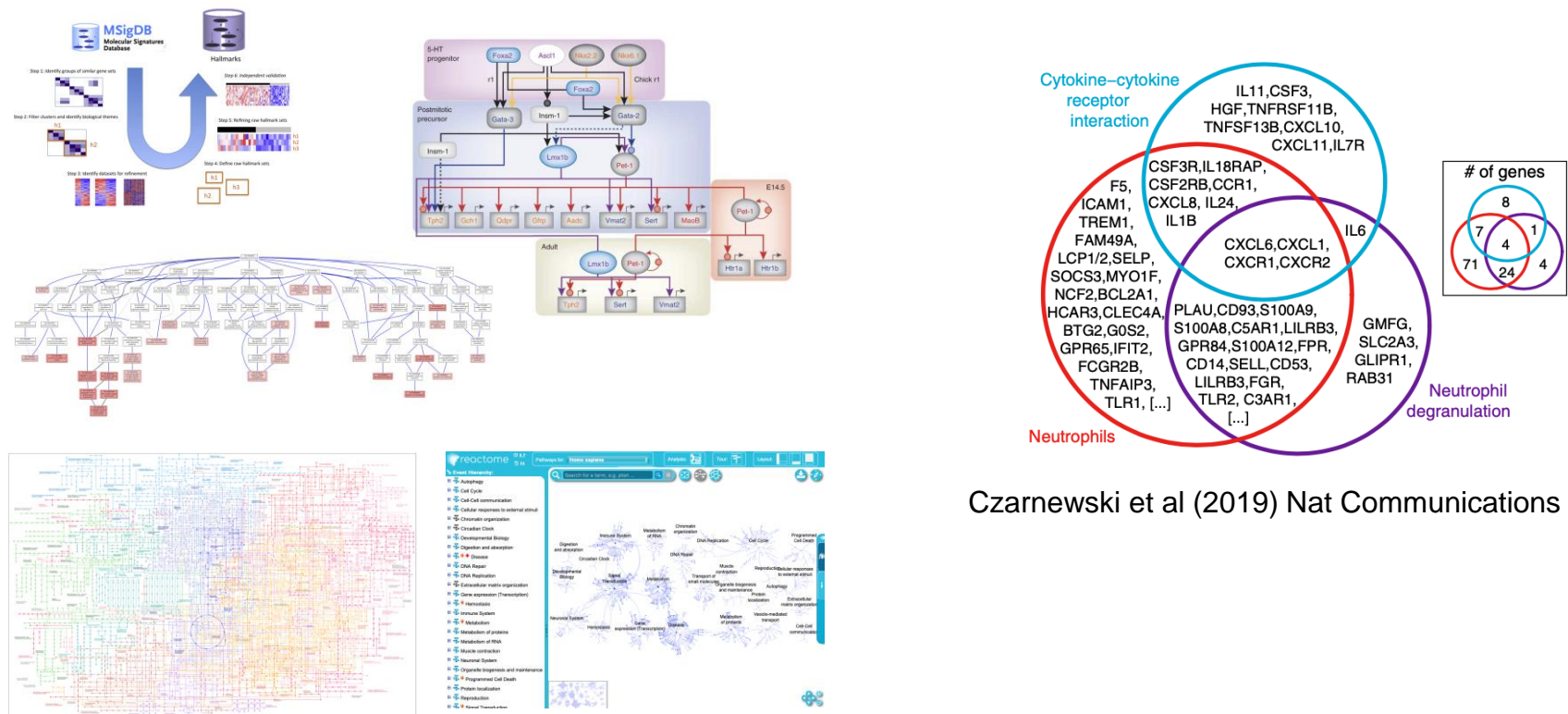
enriched

NOT enriched



Seitz et al (2018) Journal of Autoimmunity

- High number of very overlapping gene-sets (representing a similar biological theme) can bias interpretation and take attention from other biological themes that are represented by fewer gene-sets.
- Can be valuable to take gene-set interaction into account



Czarnecki et al (2019) Nat Communications

- Bias in gene-set collections (popular domains, multifunctional genes, ...)
- Gene-set names can be misleading (revisit the genes!)
- Consider the gene-set size, i.e. number of genes (specific or general)
- Positive and negative association between genes and gene-sets makes gene-level fold-changes tricky to interpret correctly
- (Typically) binary association to gene-sets, does not take into account varying levels of influence from individual genes on the process that is represented by the gene-sets
- Remember to revisit the gene-level data! Are the genes significant? Are they correctly assigned to the specific gene-set?



Thank you. Questions?

Paulo Czarnewski