



**bioinformatics.ca**  
bioinformaticsdotca.github.io



## CC BY-SA 4.0 DEED

Attribution-ShareAlike 4.0 International

Canonical URL: <https://creativecommons.org/licenses/by-sa/4.0/>

[See the legal code](#)


### You are free to:


**Share** — copy and redistribute the material in any medium or format for any purpose, even commercially.

**Adapt** — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

### Under the following terms:

 **Attribution** — You must give [appropriate credit](#), provide a link to the license, and [indicate if changes were made](#). You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

 **ShareAlike** — If you remix, transform, or build upon the material, you must distribute your contributions under the [same license](#) as the original.

**No additional restrictions** — You may not apply legal terms or [technological measures](#) that legally restrict others from doing anything the license permits.

### Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable [exception or limitation](#).

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as [publicity, privacy, or moral rights](#) may limit how you use the material.

# Module 2c: Ranked Gene Lists

Constance Li

Pathway and Network Analysis

May 12, 2026





# Learning Objectives

1. List differences between ORA and FCS
2. Learn how GSEA calculates **enrichment scores**
3. Understand pathway databases and GMT files
4. Interpret GSEA results
5. Identify when to use GSEA vs other methods



# DE-derived gene lists

Pathway analysis tools require different inputs depending on the method

### Defined Gene List

*For Over-Representation Analysis*

- A simple list of significant gene names
- Filtered by thresholds (FDR, FC)
- Unordered
- Often separated: up vs down
- **Tool: g:Profiler**
- **Method: Hypergeometric test (Fisher's Exact Test)**

TP53
BRCA1
MYC
...

### Ranked Gene List

*For Functional Class Scoring*

- ALL genes with a numeric score
- No arbitrary threshold cutoff
- Ranked from most up to most down
- **Tool: GSEA**
- **Method: Running sum statistic (modified Kolmogorov-Smirnov)**

TP53	14.2
BRCA1	-9.8
MYC	0.5
...	

**Functional Class Scoring** methods test whether genes in a gene set are enriched at the extremes of a ranked gene list rather than randomly distributed



# Why test enrichment in ranked gene list?

- Possible limitations of a defined gene list approach
  - No “natural” value to set thresholds
  - Different thresholds give different results
  - Possible loss of statistical power due to thresholding
    - All significant signals are weighted the same
    - Weak signals are neglected
- Key insight: If many pathway genes show modest, co-ordinated changes, that pathway may be dysregulated
  - *ie*: A pathway can be dysregulated even when no individual gene passes the significance threshold



# Some tools for ranked lists

**GSEA**  
Gene Set Enrichment Analysis

GSEA Home Downloads Molecular Signatures Database Documentation Contact Team

UC San Diego BROAD INSTITUTE

**Overview**

**Gene Set Enrichment Analysis (GSEA)** is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes).

- Download the GSEA software and additional resources to analyze, annotate and interpret enrichment results.
- Explore the **Molecular Signatures Database (MSigDB)**, a collection of annotated gene sets for use with GSEA software.
- View **documentation** describing GSEA and MSigDB.
- View guidelines for **using RNA-seq datasets with GSEA**.
- Use the **GenePattern** platform to run analyses, including **classical GSEA** and a variation designed for single-sample analysis (**ssGSEA**).

**What's New**

30-Jan-2026: MSigDB 2026.1 introduces the Human C9 collection of computational perturbation signature gene sets. The initial cohort of sets in this collection define the transcriptional signatures of various oncogene dependences. This release also provides collection updates for GO, Reactome, WikiPathways, and HPO, along with numerous new set additions for the ...

**Workflow Diagram:** Molecular Profile Data and Gene Set Database feed into Run GSEA, which produces Enriched Sets.

**License Terms**  
GSEA and MSigDB are available for use under [these license terms](#).

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

**Citing GSEA**  
To cite your use of the GSEA software, a joint project of UC San Diego and

GSEA is the most commonly used tool for FCS

**PANTHER**  
Classification System

The mission of the PANTHER knowledgebase is to support biomedical and other research by providing comprehensive information about the evolution of protein-coding gene families, particularly protein phylogeny, function and genetic variation impacting that function. [Learn more](#)

**New!** For human genes, try [PAN-GO functionome](#) recently published in [Nature](#).

search keyword  AI Go

Home About Data Version Tools API/Services Publications Workspace Downloads FAQ/Help/Tutorial Login Register Contact us

Current Release: [PANTHER 19.0](#) | [15,683](#) family phylogenetic trees | [144](#) species | [News](#) | [Whole genome function views](#)

Gene List Analysis Browse Sequence Search Genetic Variant Impact Keyword Search

Please refer to our article in [Nature Protocols](#) for detailed instructions on how to use this page.

**Help Tips**

**Steps:**

- Select list and list type to analyze
- Select Organism
- Select operation

[Using enhancer data](#)

**1. Enter ids and or select file for batch upload. Else enter ids or select file or list from workspace for comparing to a reference list.**

Enter IDs:  Supported IDs separate IDs by a space or comma

Upload IDs:  No file selected. File format

Please [login](#) to be able to select lists from your workspace.

Select List Type:

- ID List
- Previously exported text search results
- Workspace list
- PANTHER Generic Mapping
- ID's from Reference Proteome Genome

Organism for id list:

VCF File Flanking region:   Search Enhancer Data

PANTHER is primarily ORA, but can accept a ranked list



# We'll focus on GSEA

- Gene Set Enrichment Analysis

The screenshot shows the GSEA website interface. At the top, there is a navigation bar with the following links: GSEA Home, Downloads, Molecular Signatures Database, Documentation, Contact, and Team. Below the navigation bar, the UC San Diego and BROAD INSTITUTE logos are displayed. The main content area is titled "Overview" and contains the following text:

**Gene Set Enrichment Analysis (GSEA)** is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes).

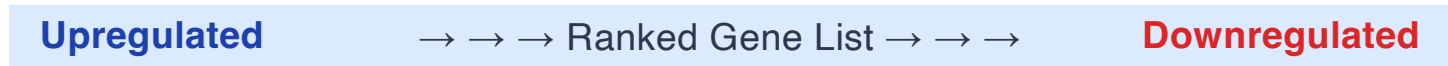
- ▶ **Download** the GSEA software and additional resources to analyze, annotate and interpret enrichment results.
- ▶ **Explore the Molecular Signatures Database (MSigDB)**, a collection of annotated gene sets for use with GSEA software.
- ▶ **View documentation** describing GSEA and MSigDB.

To the right of the text is a diagram illustrating the GSEA workflow. It shows "Molecular Profile Data" (represented by a heatmap) and "Gene Set Database" (represented by a database icon) as inputs to a central "Run GSEA" process. The output is "Enriched Sets", which is visualized as a line graph showing enrichment scores across a ranked list of genes.

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



# GSEA: The core idea



We have genes ranked from most upregulated to most downregulated

- GSEA asks: Do the genes in a given pathway cluster at the top or bottom of this list more than expected by chance?
- Three scenarios:

**Enriched at TOP**  
Pathway  
upregulated

**Random distribution**  
Pathway not  
enriched

**Enriched at BOTTOM**  
Pathway  
downregulated



# GSEA has three key steps

1. Calculate an enrichment score (ES)
2. Estimate significance level of ES
3. Adjust for multiple testing



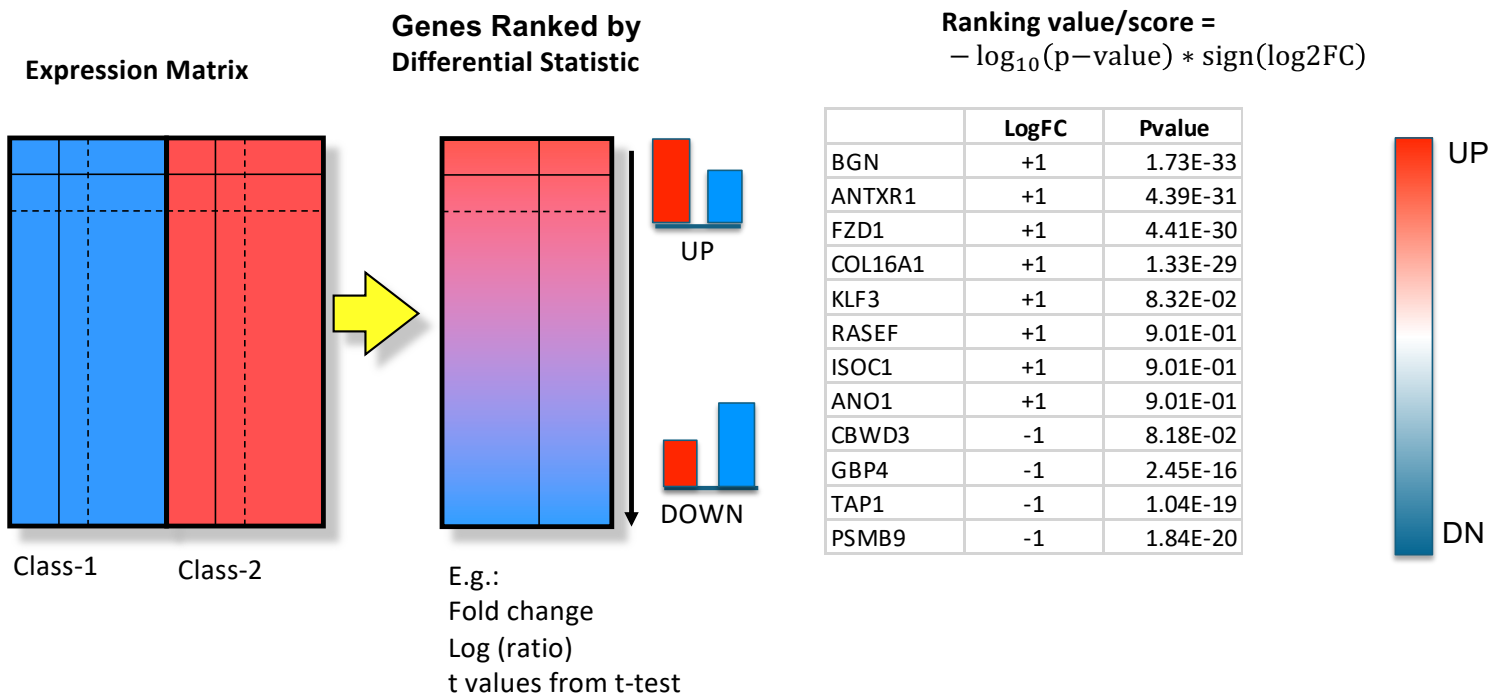
# Step 1: The GSEA enrichment score

For a given gene set  $S$  and a ranked gene list  $L$ , the ES reflects how much  $S$  is overrepresented at the extremes of  $L$

- i. Rank all genes in  $L$  by their score
  - eg. ranking value =  $-\log_{10}(\text{p-value}) * \text{sign}(\log_2\text{FC})$
- ii. Walk down  $L$ 
  - When we encounter a gene in  $S$ , increase a running sum statistic
  - When we encounter a gene not in  $S$ , decrease the running sum statistic
- iii. The ES is the maximum deviation from zero



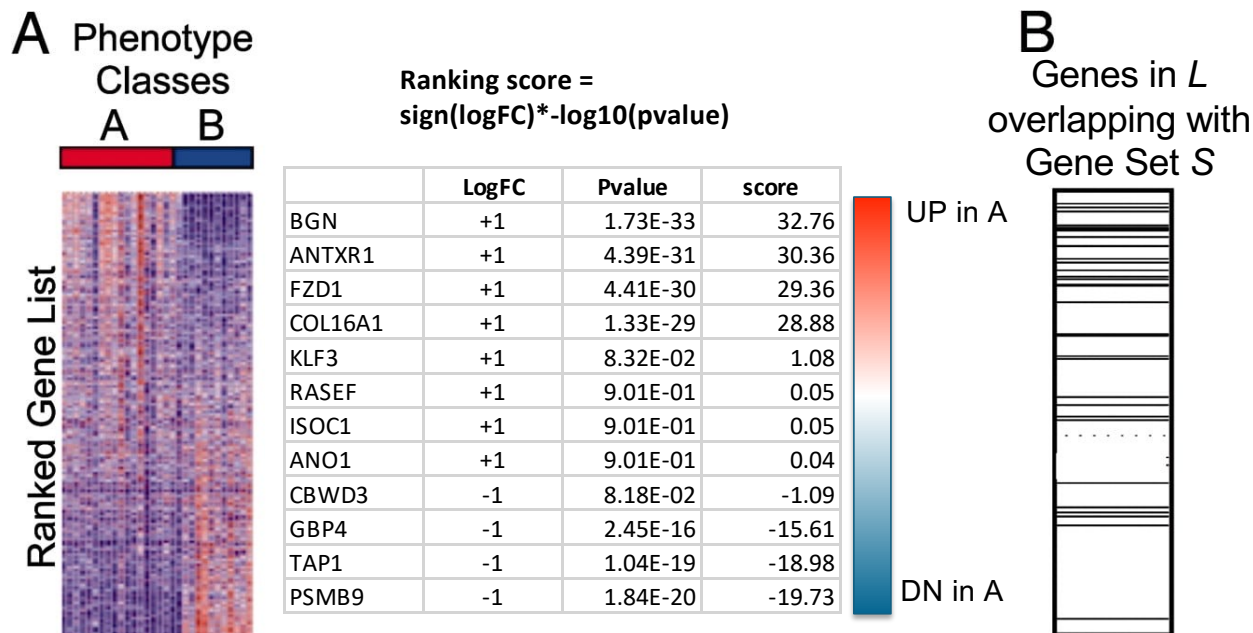
# Recall the ranked gene list





# Calculating the GSEA ES

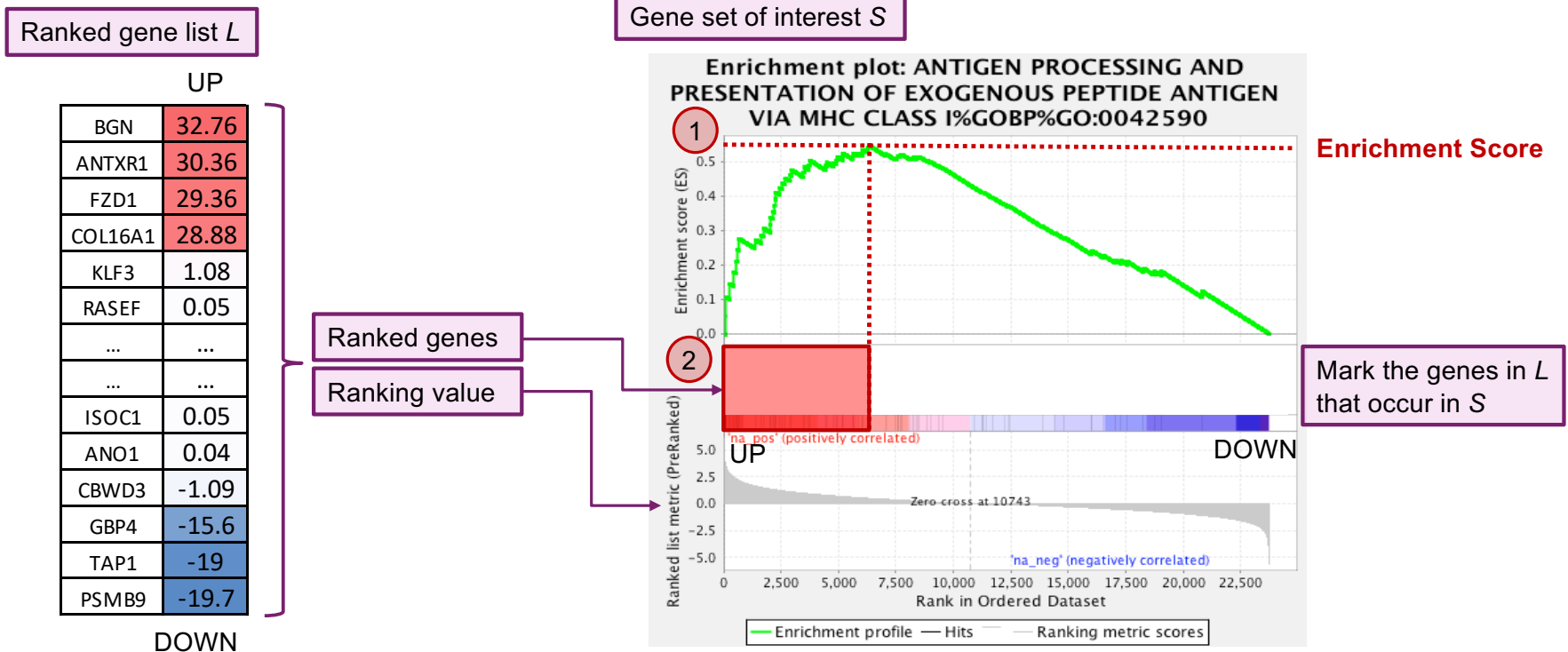
⚠ For a negative ES, the leading edge subset is the subset of members of S that appear after the peak score



Modified from <https://www.pnas.org/doi/10.1073/pnas.0506580102>



# GSEA ES walkthrough example



1. Maximum (or minimum) ES score is the final **ES score** for the gene set
2. Can define “leading edge subset” as all those genes ranked as least as high as the enriched set.

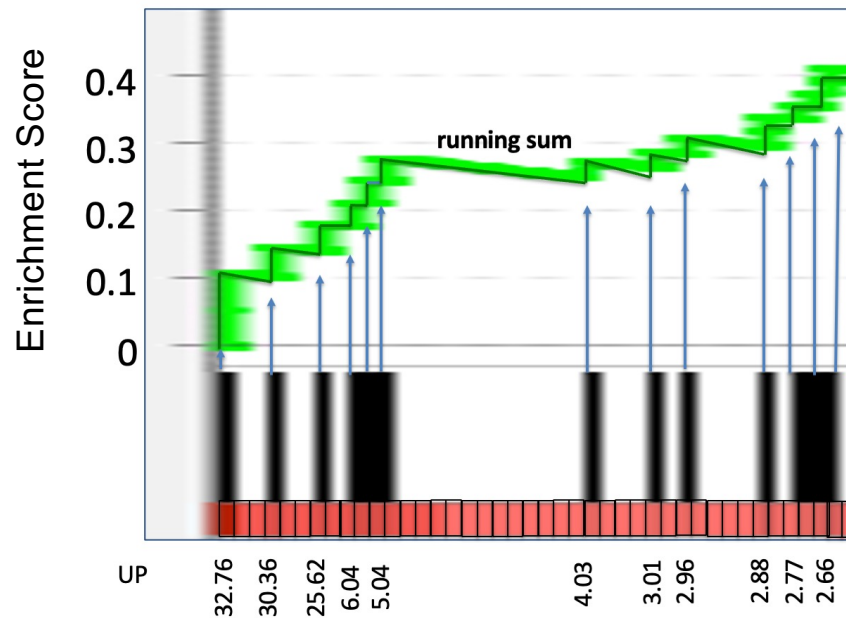


# The GSEA running sum

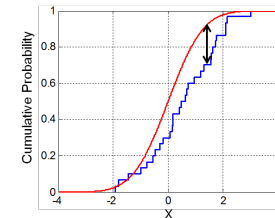
Ranked gene list  $L$

	UP
BGN	32.76
ANTXR1	30.36
FZD1	29.36
COL16A1	28.88
KLF3	1.08
RASEF	0.05
...	...
...	...
ISOC1	0.05
ANO1	0.04
CBWD3	-1.09
GBP4	-15.6
TAP1	-19
PSMB9	-19.7

DOWN



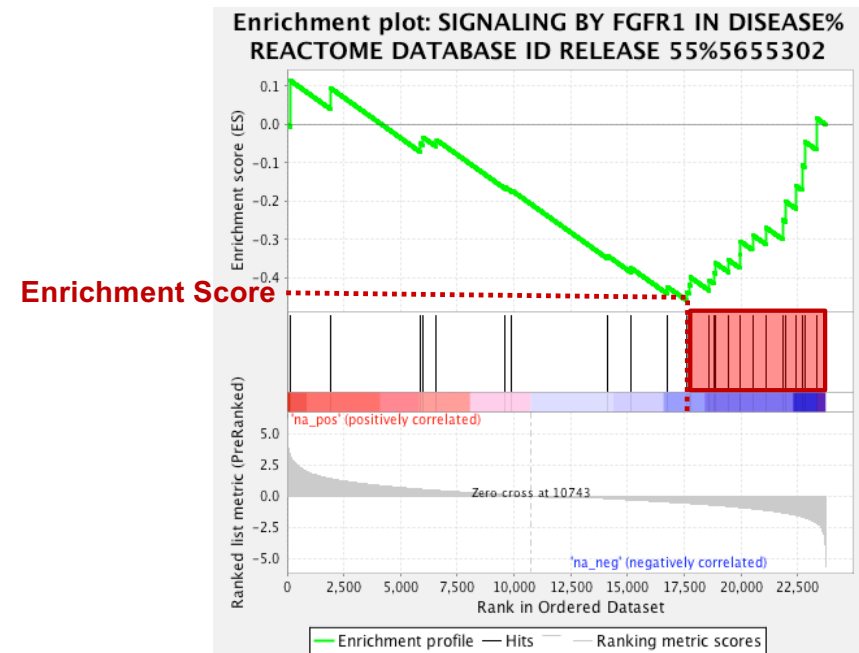
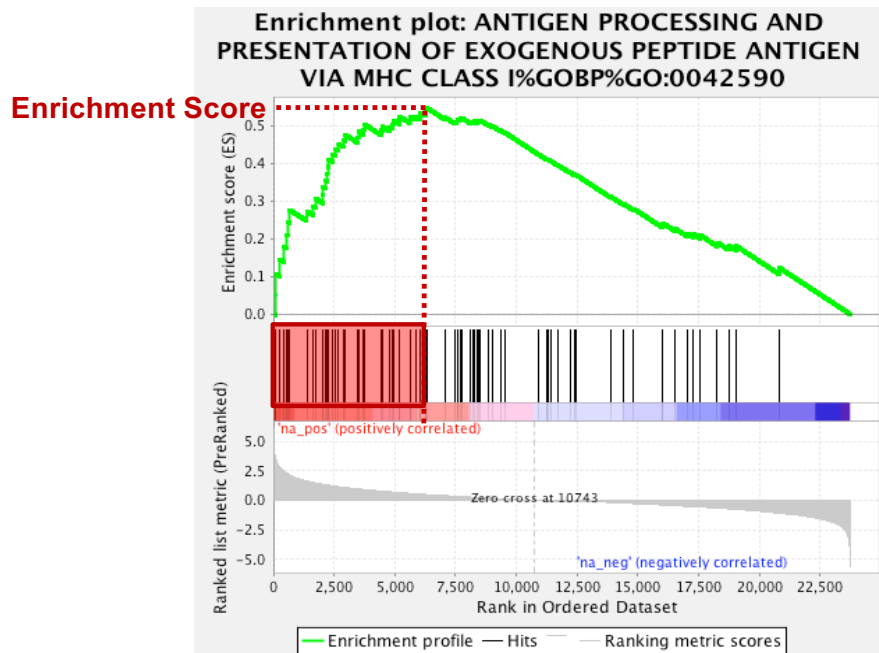
Conceptually a modified Kolmogorov–Smirnov test



- The KS test compares distributions by finding the maximum distance between two cumulative distribution functions
- The GSEA score incorporates weights into the running sum
- Running sum step sizes are unequal



# Positive and negative enrichment





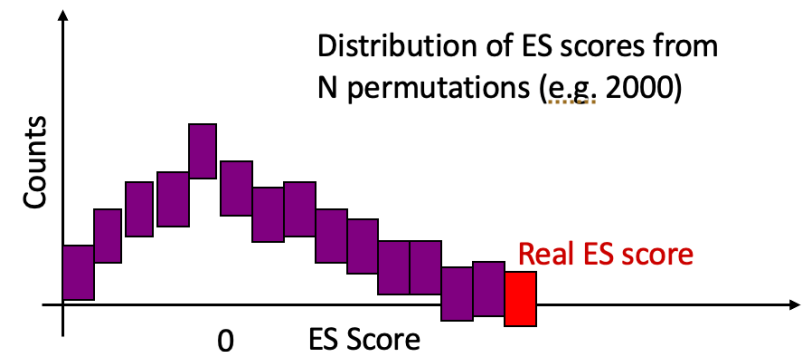
# Step 2: Estimate significance level of ES

GSEA determines significance by comparing observed ES to a null distribution generated by permutation

- I. Calculate the real ES
- II. Randomly permute phenotype labels (or gene set membership)
  - Recalculate ES for each permutation
  - Repeat 1,000+ times to build the null distribution
- III. Compare real ES to null distribution to derive p-value

For positive ES:

$$\begin{aligned} \text{p-value} &= (\# \text{ of null ES} \geq \text{observed ES}) / (\# \text{ of permutations}) \\ &= 4 / 2,000 \\ &= 0.002 \end{aligned}$$



For negative ES:

$$\begin{aligned} \text{p-value} &= (\# \text{ of null ES} \leq \text{observed ES}) / (\# \text{ of permutations}) \\ &= 4 / 2,000 \\ &= 0.002 \end{aligned}$$



## Step 3: Adjust for multiple testing

When evaluating many gene sets, remember we must account for multiple hypothesis testing!



# Interpreting GSEA output

- Sample gene set enrichment: *TP53*-mut vs *TP53*-WT
  - Enrichment in *TP53*-WT

Normalized  
Enrichment  
Score

Family-wise  
Error Rate  
correction

NAME	GS	GS DESC	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val
hsp27Pathway	hsp27Pathway	Details ...	15	0.7782923	2.1820793	0	0.003337581	0.004
p53hypoxiaPathway	p53hypoxiaPathway	Details ...	20	0.67939055	2.0815823	0	0.008710083	0.013
p53Pathway	p53Pathway	Details ...	16	0.7459421	2.0595584	0	0.009290409	0.018
P53_UP	P53_UP	Details ...	40	0.5995161	1.8757944	0	0.065383784	0.195
radiation_sensitivity	radiation_sensitivity	Details ...	26	0.5673681	1.8265389	0.001689189	0.088138245	0.296
ck1Pathway	ck1Pathway	Details ...	15	0.53940195	1.6332773	0.015180266	0.4493492	0.835
inflamPathway	inflamPathway	Details ...	28	0.54343283	1.5540237	0.049069375	0.7013238	0.955
no2il12Pathway	no2il12Pathway	Details ...	17	0.62922	1.5182602	0.048824593	0.7759864	0.975
badPathway	badPathway	Details ...	21	0.47360346	1.514384	0.055658627	0.7095338	0.977
lairPathway	lairPathway	Details ...	15	0.5720317	1.497634	0.05719921	0.70921886	0.986
chemicalPathway	chemicalPathway	Details ...	21	0.45646906	1.48676	0.045289855	0.69150734	0.989
GPCRs_Class_A_Rhodopsin-like	GPCRs_Class_A_Rhodopsin-like	Details ...	111	0.42651632	1.4715267	0.062305298	0.6924941	0.991
cytokinePathway	cytokinePathway	Details ...	21	0.5414412	1.4707	0.09540636	0.64282876	0.991
p53_signalling	p53_signalling	Details ...	87	0.34070346	1.4658169	0.028037382	0.6143147	0.994
MAP00561_Glycerolipid_metabolism	MAP00561_Glycerolipid_metabolism	Details ...	43	0.41517496	1.4188069	0.072607264	0.752188	0.998
ST_Interleukin_4_Pathway	ST_Interleukin_4_Pathway	Details ...	24	0.4032152	1.3417566	0.09965035	1	1
UTERT_DOWN	UTERT_DOWN	Details ...	64	0.3437000	1.3177700	0.11056016	1	1

NES normalizes ES by the mean of the null distribution

- Allows comparison between gene sets of different sizes

$$NES = \frac{\text{observed ES}}{\text{mean}(|\text{null ES values}|)}$$



# The leading edge

- The leading edge subset contains the genes that "drive" the enrichment signal.
- Leading Edge Statistics:
  - Tags: % of pathway genes appearing before (or after) the ES peak
  - List: % of ranked list genes appearing before (or after) the ES peak
  - Signal: Enrichment signal strength combining Tags and List

## **Positive ES (Upregulated)**

Leading edge = genes BEFORE the peak  
(at the top of the ranked list, contributing to the upward climb)

## **Negative ES (Downregulated)**

Leading edge = genes AFTER the trough  
(at the bottom of the ranked list, contributing to the downward dive)



# Gene set databases

- Gene sets are stored in GMT (Gene Matrix Transposed) format:

```
PATHWAY_NAME Description GENE1 GENE2 GENE3 ...
```

- Some resources:

Resource	Description	Website
MSigDB	Broad Institute collection (Hallmarks, C2 curated, C5 GO, etc.)	<a href="https://www.gsea-msigdb.org/gsea/msigdb">https://www.gsea-msigdb.org/gsea/msigdb</a>
Bader Lab	Combined GO + pathway databases, updated monthly	<a href="https://baderlab.org/GeneSets">https://baderlab.org/GeneSets</a>
Reactome	Curated human pathway database	<a href="https://reactome.org/">https://reactome.org/</a>
KEGG	Manually drawn pathway maps	<a href="https://www.genome.jp/kegg/pathway.html">https://www.genome.jp/kegg/pathway.html</a>
WikiPathways	Open-source, community-curated biological pathways	<a href="https://www.wikipathways.org/">https://www.wikipathways.org/</a>



# Gene set size considerations

- The size of gene sets affects statistical power and interpretability:

## Too Small (<15)

- Low statistical power
- High variance in ES
- Sensitive to noise

## Recommended

(15-500 genes)

- Good statistical power
- Interpretable pathways
- Robust ES estimates

## Too Large (>500)

- Overly broad terms
- Less specific biology
- Harder to interpret

## GSEA Settings

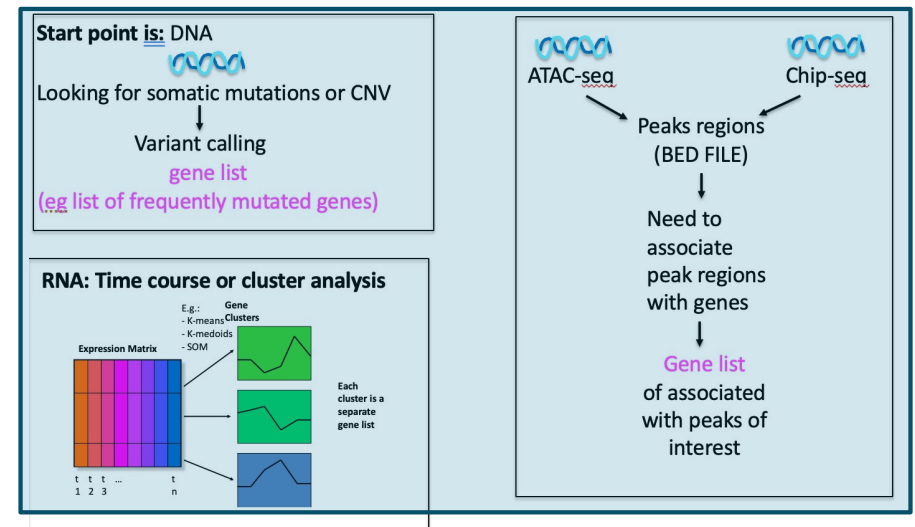
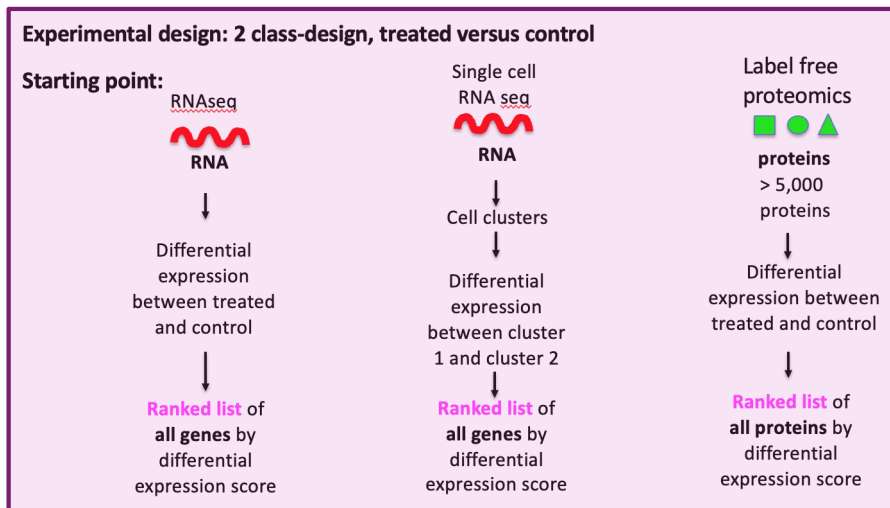
**Min size:** Default 15, can lower to 10 for exploratory analysis

**Max size:** Default 500, consider 200-250 for more specific results



# ORA or FCS?

- First, consider your gene set: Is it **ranked** or **not ranked**?





# ORA or FCS (specifically GSEA)?

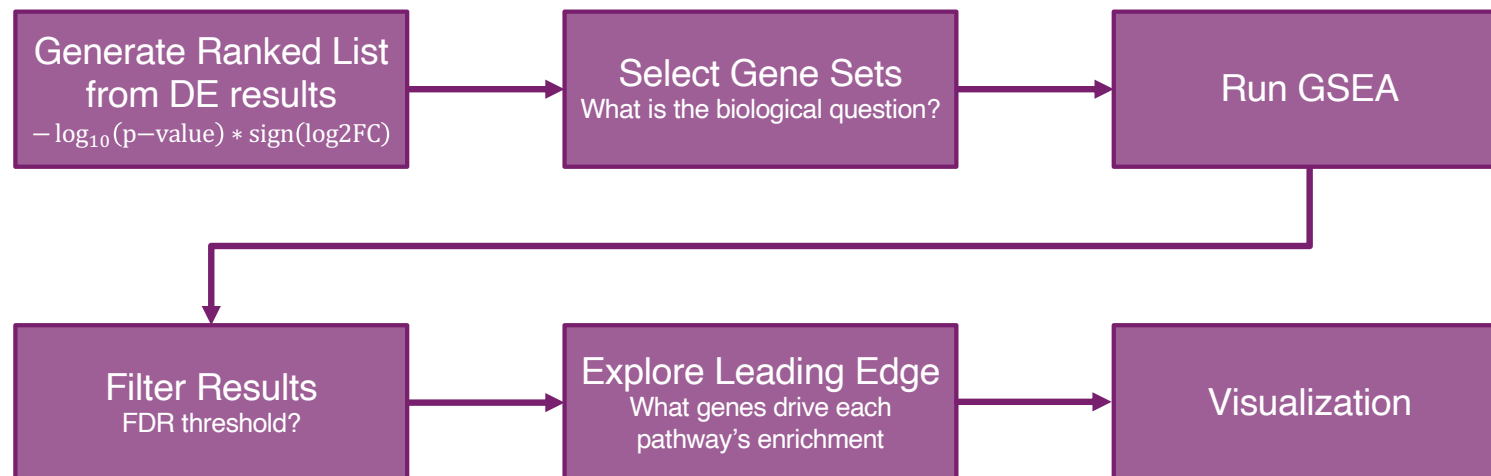
- Also consider characteristics of the method

	<b>ORA</b>	<b>GSEA</b>
Input	Defined gene list (threshold-based)	Ranked gene list (all genes)
Information used	Binary (in list or not)	Continuous (ranking scores)
Threshold dependence	High - results change with cutoff	None - uses full ranking
Subtle effects	May miss coordinated changes	Detects subtle, consistent changes
Speed	Very fast	Slower (permutations)
Interpretation	Straightforward	Richer (leading edge, NES)

- Suggestion: Try both! ORA for quick exploration, GSEA for deeper investigation



# GSEA workflow summary





# Summary

- **Functional Class Scoring** methods are appropriate for ranked gene lists
  - Uses ALL genes, not just significant ones (no arbitrary threshold)
  - GSEA is the most commonly used tool
- GSEA detects coordinated changes in gene sets
- Enrichment Score reflects clustering at list extremes
  - NES allows comparison across gene sets of different sizes
- Leading edge genes drive the enrichment signal



# Some other ranked list methods

Method	Algorithm	Key features
GSAV	Gene set variation analysis	Per-sample enrichment scores; good for clustering
ssGSEA	Single-sample GSEA	Variant of GSEA that scores each sample individually
CAMERA	Correlation-adjusted mean rank	Accounts for inter-gene correlation; parametric
GAGE	Generally Applicable Gene-set Enrichment	Two-sample t-test on gene sets; directional



# Time to try it!

- Download the lab materials:
  - [https://bioinformaticsdotca.github.io/PNA\\_CalSask-2605/module-2.html](https://bioinformaticsdotca.github.io/PNA_CalSask-2605/module-2.html)
  - Module 2c: Lab protocol pdf
    - You can use our input .rnk or use the one you generated in Lab 2a
- You'll need to register and download GSEA

# Prep for Lab

- Go to <https://www.gsea-msigdb.org>
  - Register to download
  - Then Download
- Go to [https://download.baderlab.org/EM\\_Genesets/current\\_releases/Human/symbol/Human\\_GOBP\\_AllPathways\\_noPFOCR\\_no\\_GO\\_iaa\\_May\\_01\\_2026\\_symbol.gmt](https://download.baderlab.org/EM_Genesets/current_releases/Human/symbol/Human_GOBP_AllPathways_noPFOCR_no_GO_iaa_May_01_2026_symbol.gmt)
  - Download the geneset

**GSEA**  
Gene Set Enrichment Analysis

GSEA Home Downloads Molecular Signatures Database Documentation Contact Team

UC San Diego BROAD INSTITUTE

**Overview**

**Gene Set Enrichment Analysis (GSEA)** is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes).

**Download** the GSEA software and additional resources to analyze, annotate and interpret enrichment results.

- ▶ **Explore the Molecular Signatures Database (MSigDB)**, a collection of annotated gene sets for use with GSEA software.
- ▶ **View documentation** describing GSEA and MSigDB.
- ▶ View guidelines for **using RNA-seq datasets with GSEA**.
- ▶ Use the **GenePattern** platform to run analyses, including **classical GSEA** and a variation designed for single-sample analysis (**ssGSEA**).

**What's New**

30-Jan-2026: MSigDB 2026.1 introduces the Human C9 collection of computational perturbation signature gene sets. The initial cohort of sets in

**Molecular Profile Data**

**Enriched Sets**

Run GSEA

Gene Set Database

**License Terms**

GSEA and MSigDB are available for use under [these license terms](#).

Please **register** to download the GSEA software, access our web tools, and view the MSigDB gene sets. 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.



# GSEA Lab: Continue our PAAD exploration

