

## Functional Class Scoring: Gene Set Enrichment Analysis (GSEA)

Lab 2c, Pathway and Network Analysis 2026

**Duration:** 1 hour (+30 min optional R section)

**Format:** Virtual, hands-on

**Prerequisites:** Completed Lab2a, (+basic R knowledge)

### Overview

In this workshop, you will:

1. Recall how to prepare a ranked gene list from DE results
2. Run GSEA using the desktop application
3. Explore and interpret GSEA output
4. Understand the enrichment plot and leading edge

**Data:** TCGA Pancreatic Adenocarcinoma (PAAD) - Moffitt Basal vs Classical subtypes.

### Required Software

- GSEA Desktop application (download from [gsea-msigdb.org](https://software.broadinstitute.org/gsea/))

### Required Files

Download to your working directory:

- I. ranked\_basal\_vs\_classical.rnk (from Lab 2a)
- II. Human\_GOBP\_AllPathways\_noPFOCR\_no\_GO\_iea\_May\_01\_2026\_symbol.gmt (from Bader Lab)

---

## Exercise 0: Create a Ranked Gene List

*Review this section to recall what you did in Lab 2a. Then start with Exercise 1.*

**Goal:** Generate a .rnk file from differential expression results

### Step 0.1: Load the DE Results

Load the DESeq2 results from the previous module.

### Step 0.2: Calculate the Ranking Metric

The ranking metric combines significance and direction:

$\text{score} = -\log_{10}(\text{pvalue}) * \text{sign}(\log_2\text{FoldChange})$

This ensures:

- Highly significant upregulated genes: large positive scores (top)
- Highly significant downregulated genes: large negative scores (bottom)
- Non-significant genes: scores near zero (middle)

### Step 0.3: Handle Problematic Values

Remove NA and Inf values that would cause GSEA errors.

### Step 0.4: Format gene identifiers

GSEA requires unique gene identifiers. Keep the entry with the highest absolute score.

Also make sure your genes are described by gene symbols.

### Step 0.5: Sort, Save, and Verify

Sort genes, save as .rnk file, and verify the format.

Important Notes about the .rnk file:

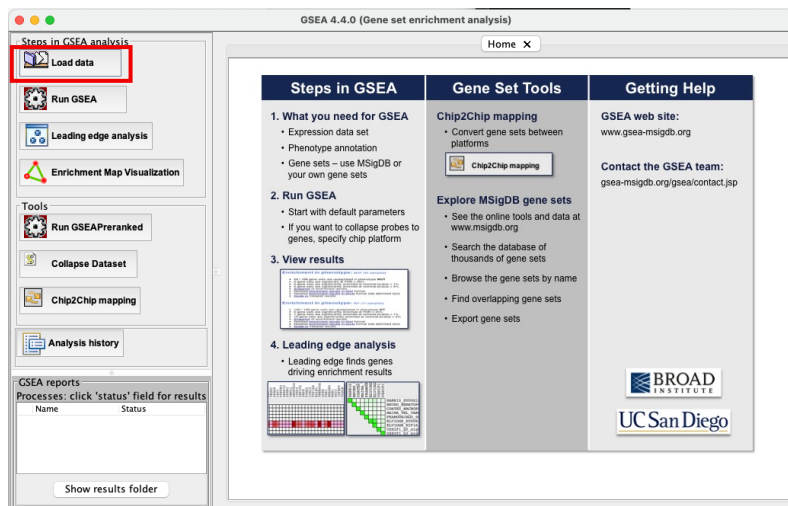
- No header row
- Tab-delimited, two columns only (gene symbol, score)
- Gene symbols (not Ensembl IDs) to match GMT file
- ALL genes included, not just significant ones

## Exercise 1: Run GSEA

**Goal:** Use the GSEA desktop application to perform gene set enrichment analysis

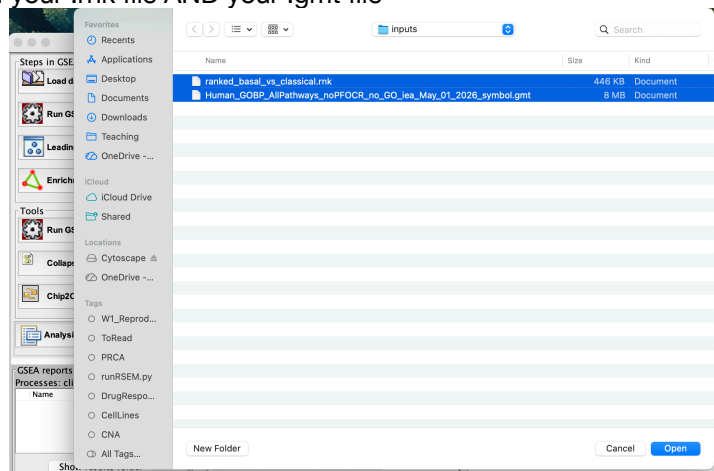
### Step 1.1: Launch GSEA

Double-click the GSEA application icon to launch it.



### Step 1.2: Load Data Files

- I. Click **Load data** in the upper left corner
- II. Click **Browse for files**
- III. Select BOTH: your .rnk file AND your .gmt file



- IV. Click **Open** and wait for the "Files loaded successfully" dialogue box, then click **OK**

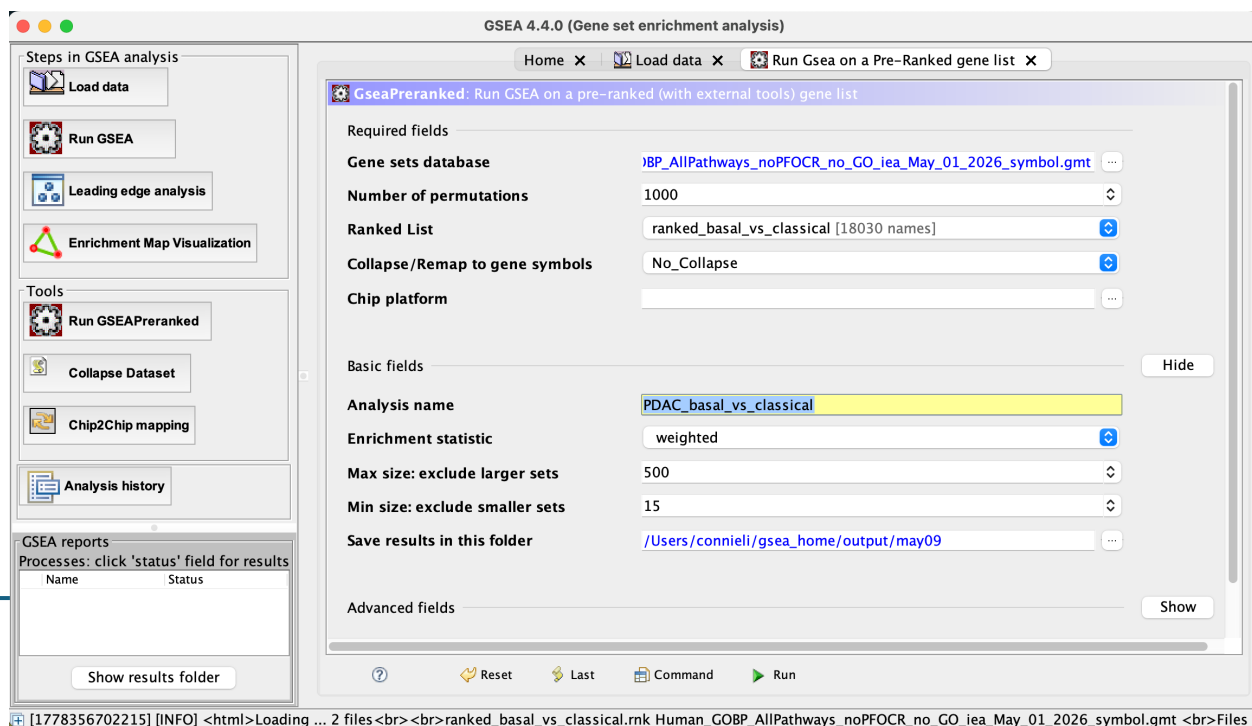
### Step 1.3: Configure GSEAPreranked

In the right sidebar, go to **Tools > Run GSEAPreranked** and set:

Parameter	Setting
Gene sets database	Select your GMT file (Local gmx/gmt tab)
Number of permutations	1000 (use 100 for testing; 1000+ for publication)
Ranked list	Select your .rnk file from dropdown

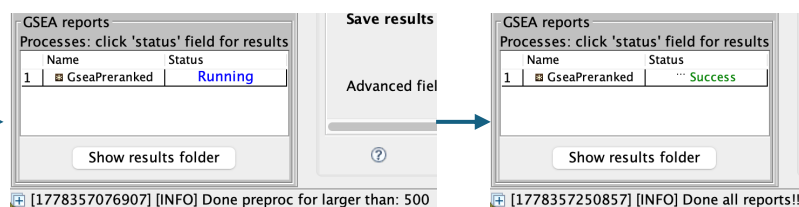
Collapse/Remap	No_Collapse
Analysis name	Basal_vs_Classical_GSEA
Max size	500 (exclude very large gene sets)
Min size	15 (exclude very small gene sets)

Hint: If you don't see the Analysis name or size parameters, make sure you're showing the Basic fields



### Step 1.4: Run the Analysis

- I. Click **Run** button (bottom right)
- II. Monitor progress in the bottom left panel (5-15 minutes)
- III. When complete, status shows "Success"



## Exercise 2: Explore GSEA Results

**Goal:** Understand and interpret the GSEA output

### Step 2.1: Open the Results Report

In the bottom left GSEA reports box, Click the "Success" link OR use your file explorer/finder to navigate to index.html in the results folder specified in your job parameters. The report will open in a web browser.

Note: It's normal to see "Enrichment in phenotype: na". Because we ran GSEA using a pre-computed ranked gene list, GSEA doesn't know what the phenotypes are.

"Enrichment in phenotype: na\_pos" = gene sets enriched at the top of your ranked list (positive end)

"Enrichment in phenotype: na\_neg" = gene sets enriched at the bottom of your ranked list (negative end)

For our lab, we ranked by Basal vs. Classical log2FC (Basal as positive), so:

na\_pos = pathways enriched in Basal

na\_neg = pathways enriched in Classical

### Step 2.2: Examine Summary Statistics

#### GSEA Report for Dataset ranked\_basal\_vs\_classical

##### Enrichment in phenotype: na

- 3924 / 6609 gene sets are upregulated in phenotype **na\_pos**
- 889 gene sets are significant at FDR < 25%
- 472 gene sets are significantly enriched at nominal pvalue < 1%
- 944 gene sets are significantly enriched at nominal pvalue < 5%
- [Snapshot](#) of enrichment results
- Detailed [enrichment results in html](#) format
- Detailed [enrichment results in TSV](#) format (tab delimited text)
- [Guide to](#) interpret results

At the top of the report, note:

- The number of Gene sets tested: Should be several thousand
- Gene sets enriched: "pos" = upregulated in Basal; "neg" = downregulated
- Statistical thresholds: Numbers of significantly enriched gene sets at FDR < 0.25 and unadjusted p-value thresholds.

Scroll down and scan the rest of the report index.

Note the Comment: **Comments**

- Timestamp used as the random seed: 1778357075532

Do your results have a different number?

What do you think this means? Why is it important?

## Step 2.3: View Detailed Results Tables

Click the pos and neg "Detailed enrichment results" links.

Table: Gene sets enriched in phenotype na [plain text format]

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
1	HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION%MSIGDBHALLMARK% <a href="#">HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION</a>	<a href="#">Details ...</a>	156	0.74	2.53	0.000	0.000	0.000	3218	tags=62%, list=18%, signal=75%
2	SKIN DEVELOPMENT%GOBP%GO:0043588	<a href="#">Details ...</a>	118	0.72	2.40	0.000	0.000	0.000	1560	tags=36%, list=9%, signal=40%
3	INTERMEDIATE FILAMENT ORGANIZATION%GOBP%GO:0045109	<a href="#">Details ...</a>	57	0.81	2.39	0.000	0.000	0.000	419	tags=26%, list=2%, signal=27%
4	KERATINOCYTE DIFFERENTIATION%GOBP%GO:0030216	<a href="#">Details ...</a>	66	0.80	2.38	0.000	0.000	0.000	1064	tags=35%, list=6%, signal=37%
5	INTERMEDIATE FILAMENT-BASED PROCESS%GOBP%GO:0045103	<a href="#">Details ...</a>	69	0.78	2.38	0.000	0.000	0.000	453	tags=25%, list=3%, signal=25%
6	INTERMEDIATE FILAMENT CYTOSKELETON ORGANIZATION%GOBP%GO:0045104	<a href="#">Details ...</a>	68	0.78	2.35	0.000	0.000	0.000	453	tags=25%, list=3%, signal=25%

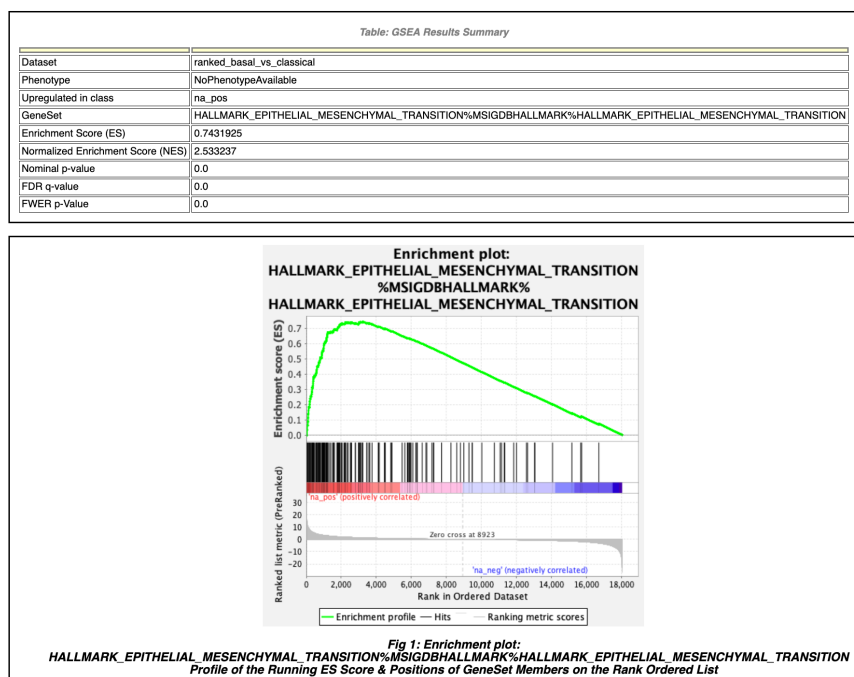
Here are some some key columns:

Column	Description
NES	Normalized Enrichment Score (use for comparison)
FDR q-val	False Discovery Rate - USE THIS for significance
RANK AT MAX	Where ES peak occurs in ranked list
LEADING EDGE	Tags/List/Signal statistics describing key genes

GSEA provides a handy guide to interpret results. Can you find the link to it in your report?

## Step 2.4: Examine Individual Enrichment Plots

In a "Detailed enrichment results" table, click on the GS Details for a gene set to see its enrichment plot:



Review the interpretations of these plot panels, and interpret the results you have

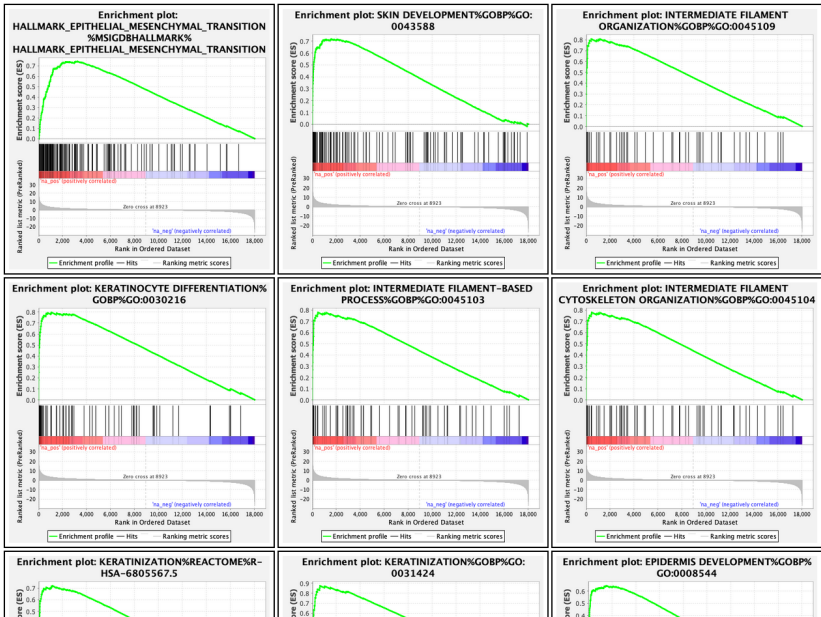
- **Top Panel:** Running enrichment score (green line); peak = ES
- **Middle Panel:** Gene hits (black lines); clustering = enrichment

- **Bottom Panel:** Ranking metric distribution
- **Leading Edge:** Genes before peak (pos ES) or after (neg ES) that drive enrichment

Step 2.5: Explore the Snapshot View

Return to your main report page.  
Click "Snapshots" to see top 20 enriched gene sets in each direction.

Table: Snapshot of enrichment results



Exercise 3: Examine Leading Edge Genes

Goal: Identify the genes driving pathway enrichment

Step 3.1: Find the Leading Edge Subset

From your “Snapshot” or a "Detailed enrichment results" page, click on a significant gene set and scroll down to the GSEA details tables. This table contains your leading edge genes marked in the CORE ENRICHMENT column.

Table: GSEA details [plain text format]

	SYMBOL	RANK IN GENE LIST	RANK METRIC SCORE	RUNNING ES	CORE ENRICHMENT
1	KRT6C	2	27.564	0.0603	Yes
2	KRT6A	3	25.354	0.1159	Yes
3	KRT5	8	20.392	0.1603	Yes
4	IVL	16	18.005	0.1994	Yes
5	KRT14	17	17.907	0.2387	Yes
6	ANXA1	23	16.440	0.2744	Yes
7	KRT9	28	15.318	0.3078	Yes
8	KRT4	29	15.148	0.3410	Yes
9	KRT17	50	12.827	0.3679	Yes
10	TP63	63	11.603	0.3927	Yes
11	SCEL	73	11.087	0.4165	Yes
12	TGFB2	74	11.043	0.4407	Yes
13	KRT7	84	10.583	0.4634	Yes
14	KRT16	101	9.999	0.4844	Yes
15	KRT81	109	9.779	0.5055	Yes
16	ITGA3	117	9.568	0.5260	Yes
17	EREG	149	8.692	0.5434	Yes
18	TGM1	168	8.339	0.5606	Yes
19	WNT10A	200	7.692	0.5758	Yes
20	KRT78	210	7.606	0.5010	Yes

To explore the gene list to file, you can click the “plain text format” link at the top of the table.

Remember that the tags, list, and signal metrics are in the "Detailed enrichment results" table.

- **Tags:** % of gene set genes in leading edge
- **List:** % of ranked list before/after ES peak
- **Signal:** Combined enrichment signal strength

Example: tags=62%, list=18%, signal=75% means 62% of the gene set members appear in the top 18% of the ranked list, giving a strong 75% signal.

### Step 3.2: Multi-Gene-Set Leading Edge Analysis

GSEA helps us compare leading edge overlap across multiple gene sets.

1. Return to the GSEA application. In the left sidebar, click **Leading edge analysis**
2. If your results aren't pre-loaded from your session, or if you want to load other GSEA results, use “Locate the GSEA result folder from your file system” and “Open” the desired folder. Then click the “Load GSEA Results” buttons and wait for the table to be populated.

GSEA 4.4.0 (Gene set enrichment analysis)

Home x **Leading edge analysis** x Run Gsea x

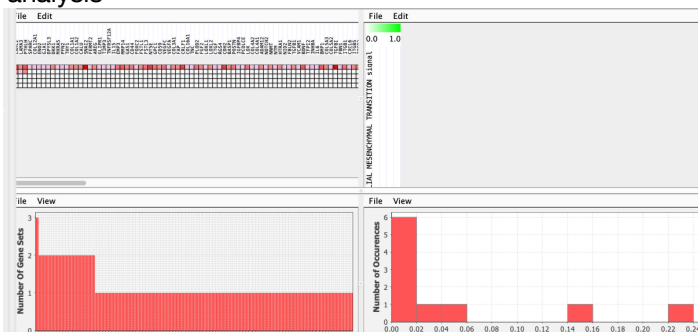
Select a GSEA result from the application cache

[ OR ] Locate a GSEA result folder from the file system

GSEA Results x

Gene Set	Size	ES	NES	NOM p-val	FDR q-val	FWER p-val	Rank at Max	Leading Edge
XENOBIOTIC METABOLIC PROCESS%GOBP%...	124	-0.732	-2.294	0	0	0	0	1659 tags=36%, list=9%, signal=40%
CELLULAR RESPONSE TO XENOBIOTIC STIM...	147	-0.694	-2.232	0	0	0	0	1747 tags=34%, list=10%, signal=37%
DIGESTION%GOBP%GO:0007586	68	-0.738	-2.144	0	0	0	0	1854 tags=46%, list=10%, signal=51%
ESTROGEN METABOLIC PROCESS%GOBP%G...	36	-0.815	-2.113	0	0	0	0	1971 tags=47%, list=11%, signal=53%
BIOLOGICAL OXIDATIONS%REACTOME%R-H...	194	-0.634	-2.115	0	0	0	0	1689 tags=31%, list=9%, signal=34%
ASPIRIN ADME%REACTOME%R-HSA-97496...	37	-0.788	-2.051	0	0	0.004	0	1971 tags=46%, list=11%, signal=51%
XENOBIOTICS%REACTOME DATABASE ID RE...	21	-0.877	-2.019	0	0.001	0.007	0	912 tags=48%, list=5%, signal=50%
DIGESTIVE SYSTEM PROCESS%GOBP%GO:00...	50	-0.738	-2.006	0	0.001	0.011	0	1854 tags=46%, list=10%, signal=51%
RESPONSE TO XENOBIOTIC STIMULUS%GOB...	187	-0.61	-2.004	0	0.001	0.012	0	1659 tags=29%, list=9%, signal=31%
PHASE I - FUNCTIONALIZATION OF COMPO...	94	-0.637	-1.936	0	0.006	0.08	0	1031 tags=27%, list=6%, signal=28%
PID_HNF3B_PATHWAY%MSIGDB_C2%PID_HN...	34	-0.753	-1.933	0	0.006	0.085	0	1624 tags=41%, list=9%, signal=45%
PHASE II - CONJUGATION OF COMPOUNDS%	94	-0.627	-1.916	0	0.008	0.125	0	1971 tags=34%, list=11%, signal=38%
SODIUM ION HOMEOSTASIS%GOBP%GO:005...	30	-0.759	-1.915	0	0.007	0.129	0	826 tags=30%, list=5%, signal=31%
INTESTINAL ABSORPTION%GOBP%GO:0050...	24	-0.806	-1.908	0	0.008	0.151	0	1854 tags=50%, list=10%, signal=56%
SULFUR COMPOUND METABOLIC PROCESS%	211	-0.571	-1.903	0	0.008	0.168	0	3128 tags=39%, list=17%, signal=46%
EPOXYGENASE P450 PATHWAY%GOBP%GO...	16	-0.855	-1.902	0	0.008	0.168	0	912 tags=50%, list=5%, signal=53%
ICOPOLYMER METABOLIC PROCESS%GOBP%	61	-0.636	-1.899	0	0.009	0.169	0	2818 tags=44%, list=16%, signal=52%

3. Select two or more gene sets of interest. You can use the column headers to sort the table. Click “Run leading edge analysis”



4. Scroll across the heat map showing which leading edge genes appear in multiple gene sets. Explore the other figure panels describing the sizes of overlap.

Note: The multi-gene-set Leading Edge Analysis requires selecting at least 2 gene sets. This is because its purpose is to find overlap — which genes are shared across pathways. For a single gene set, use the CORE ENRICHMENT column in the gene details table instead.

## Questions for Exploration

Answer these questions based on your GSEA results:

1. How many gene sets were tested? How many passed  $FDR < 0.25$ ?
2. What are the top 3 positively enriched pathways (upregulated in Basal)? What is their NES and FDR?
3. What are the top 3 negatively enriched pathways? What biological processes do they represent?
4. For one top pathway, examine the enrichment plot. Is the gene clustering clear? What genes are in the leading edge?
5. Do you see related pathways that might represent the same biology?

## Troubleshooting

Problem	Solution
Very few gene sets in results	Check gene ID format - must match GMT
All FDR = 1.0	Increase permutations; verify ranking scores
Java Heap Space error	Download GSEA version with more memory
Duplicate gene error	Remove duplicates (keep highest  score )
Wrong enrichment direction	Verify log2FC sign convention

## Additional Resources

- **GSEA User Guide:** [https://docs.gsea-msigdb.org/GSEA/GSEA\\_User\\_Guide/](https://docs.gsea-msigdb.org/GSEA/GSEA_User_Guide/)
- **GSEA Paper:** Subramanian et al., PNAS 2005. PMID: 16199517
- **Bader Lab Gene Sets:** [download.baderlab.org/EM\\_Genesets/](http://download.baderlab.org/EM_Genesets/)
- **MSigDB:** [gsea-msigdb.org/gsea/msigdb](http://gsea-msigdb.org/gsea/msigdb)

## Linking to EnrichmentMap (Cytoscape)

To visualize and analyze your GSEA results in Cytoscape, you'll need some files in your GSEA output folder. Remember where you ask GSEA to write these results!

- gsea\_report\_for\_na\_pos\_\*.tsv - Positive enrichment results
- gsea\_report\_for\_na\_neg\_\*.tsv - Negative enrichment results
- Your GMT file
- Your RNK file (optional, for expression overlay)

## Optional Exercise: Running GSEA in R

The fgsea package provides a fast R implementation of GSEA. See `Lab2cExtra_fgSEA.R` for code.

### Advantages of fgsea:

- Faster than GSEA desktop
- Reproducible and scriptable
- Easy to integrate into R pipelines
- Same algorithm, equivalent results

### Steps covered in the R code:

- Install and load fgsea, data.table, ggplot2
- Prepare ranked gene list as named vector
- Load GMT file with gmtPathways()
- Run fgseaMultilevel()
- Explore and visualize results
- Extract leading edge genes
- Export for EnrichmentMap