

## Integrative Analysis: ReactomeFI and GeneMANIA

Lab 4, Pathway and Network Analysis 2026

**Duration:** 1 hour

**Format:** Virtual, hands-on

**Prerequisites:** Completed Labs 2a, 2b, 2c, 3

### Overview

In previous modules, you identified differentially expressed genes between Basal and Classical PAAD subtypes (DESeq2) and discovered enriched pathways using ranked gene analysis (GSEA). Now we'll add a second data dimension: [somatic mutations](#).

### The key question:

Do the pathways disrupted by mutations overlap with those showing expression changes?

**Data:** TCGA Pancreatic Adenocarcinoma (PAAD) - Moffitt Basal vs Classical subtypes and recurrently mutated gene list

### Required Software

- Cytoscape with EnrichmentMap, ReactomeFIPlugin, and GeneMANIA applications

### Required Files

- I. PAAD\_MutatedCGC\_genes.txt
- II. Your GSEA results from Lab 2c

---

## Exercise 0: Review Results from Lab 2c

From DESeq2:

- Basal subtype shows distinct gene expression from Classical
- Thousands of differentially expressed genes ( $FDR < 0.05$ ,  $|\log_2FC| > 1$ )

From GSEA:

- Basal-enriched (positive NES): EMT, ECM organization, cell migration, TGF- $\beta$  signaling
- Classical-enriched (negative NES): Metabolic processes, pancreatic secretion, digestion

Consider:

PAAD is driven by mutations in KRAS (93%), TP53 (64%), CDKN2A, SMAD4, and other genes. Do these mutations:

1. Cluster into functional modules?
2. Target the same pathways showing expression changes?
3. Connect to each other through known protein interactions?

Exercises 1-3 will walk us through an integrative analysis to address these questions

## Exercise 1: ReactomeFI — Mutation Network Analysis

**Goal:** Build a functional interaction network based on PAAD mutation data to see if mutated genes cluster into connected modules that affect the same biological pathways.

### Background

Unlike expression data, mutation data presents unique challenges:

- Sparse: Each gene is mutated in only a fraction of patients
- Heterogeneous: Different patients mutate different genes
- No natural ranking: A gene is mutated or not (binary)

We use ReactomeFI to build a *de novo* network of functional interactions. If mutations cluster into connected modules, they likely affect the same biological processes, even if the specific genes differ between patients.

### Data Source

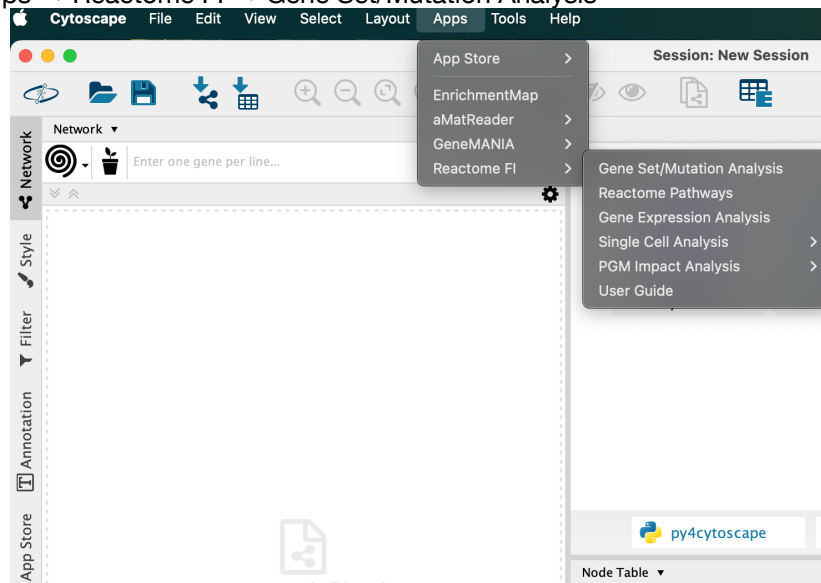
We'll use genes from **Supplementary Table S1** of the TCGA PAAD paper (Cancer Cell 2017). Specifically, we extracted genes from the "Mutated CGC Genes" column, which lists Cancer Gene Census genes that are somatically mutated in TCGA-PAAD samples. The `PAAD_MutatedCGC_genes.txt` file contains 30 Cancer Gene Census genes that are recurrently mutated in PAAD, including key drivers like *KRAS* (93%), *TP53* (64%), *CDKN2A*, *SMAD4*, and chromatin modifiers like *ARID1A* and *PBRM1*.

The Cancer Gene Census (CGC) is a curated catalog of genes causally implicated in cancer. By focusing on CGC genes mutated in PAAD, we ensure biological relevance while keeping the gene list manageable for network analysis.

Note that in the given input `PAAD_MutatedCGC_genes.txt` the header lines beginning with '#' and are ignored by Cytoscape.

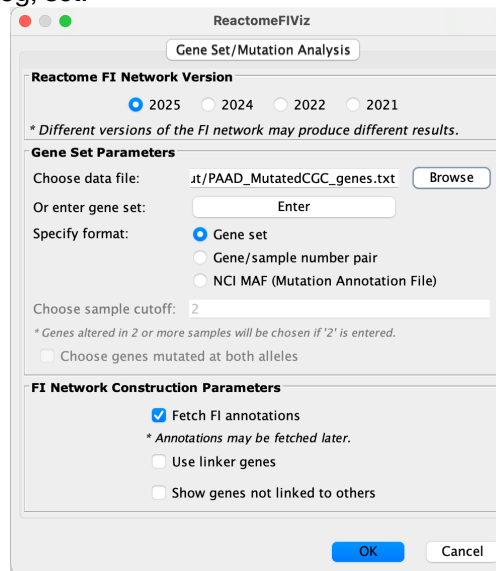
### Step 1.1: Load ReactomeFI in Cytoscape

- I. Open Cytoscape (ensure version 3.9+ with ReactomeFI app installed)
- II. Go to Apps → Reactome FI → Gene Set/Mutation Analysis



## Step 1.2: Load PAAD Mutation Data

- I. In the ReactomeFI dialog, set:



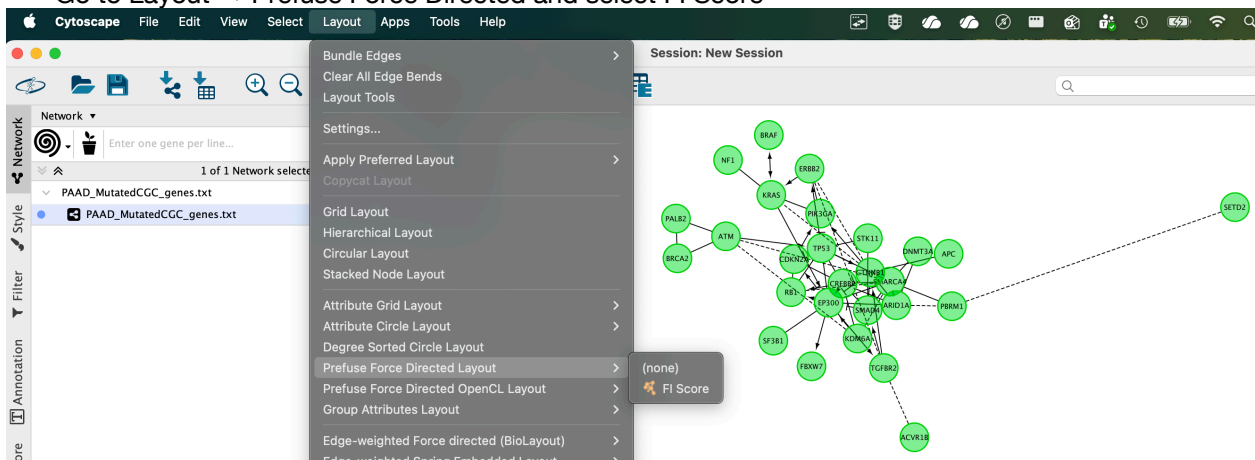
- Gene/protein file: Browse to `PAAD\_MutatedCGC\_genes.txt`
- Specify format: Gene set (one gene per line)
- 2025 Reactome FI Network Version (or other version, for your exploration)
- Do not include linker genes. Linker genes are added to connect otherwise disconnected nodes. For mutation analysis, we want to see which mutated genes are *directly* connected through known functional interactions.

- II. Click OK to build the network

## Step 1.3: Analyze the Network

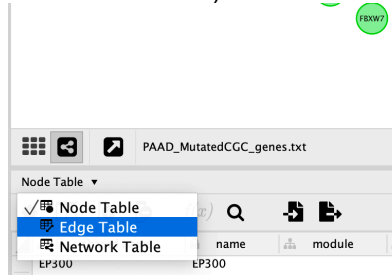
- I. You may notice that all your nodes are stacked on top of each other. To spread out your notes, we'll apply a layout

- Go to Layout → Prefuse Force Directed and select FI Score



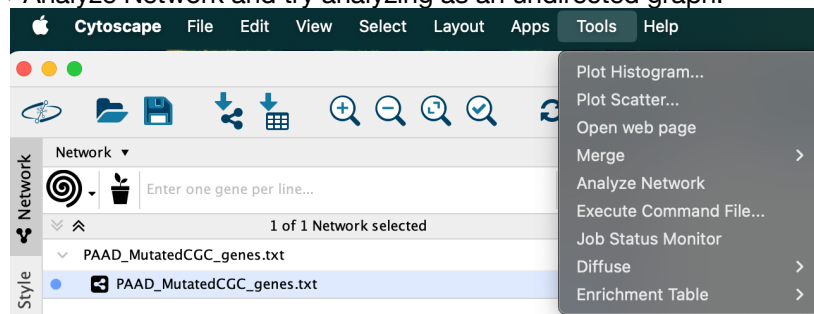
- You can try other layouts like yFiles Organic Layout (if installed)

- II. Observe the resulting network:
  - Nodes = mutated CGC genes
  - Edges = functional interactions (from Reactome, curated pathways, protein-protein interactions)
- III. Explore edge annotations
  - Click on an edge (e.g., between *KRAS* and *BRAF*)

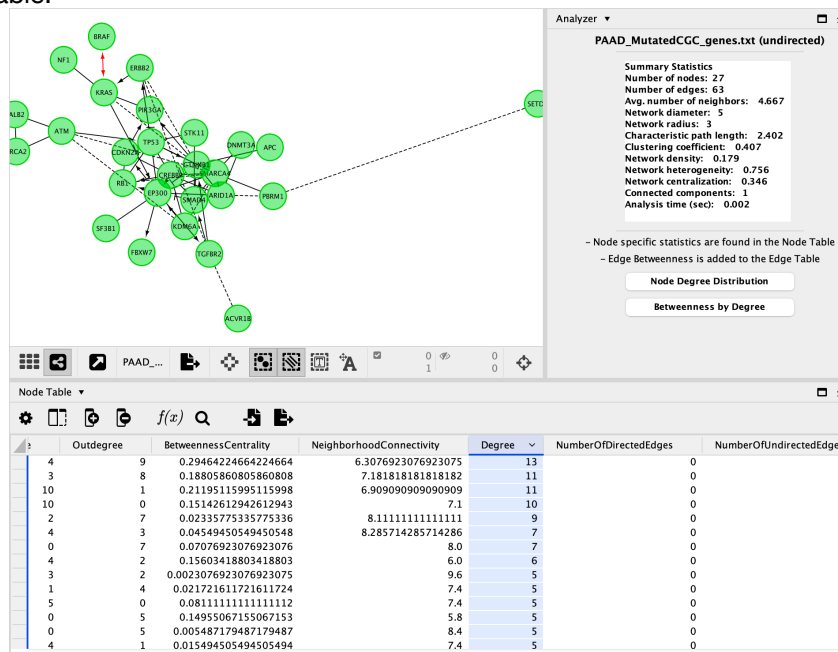


- Select the Edge Table panel in the drop down menu. Make sure your edge of interest is still selected. Look at the "FI Annotation" column
- Note the interaction type: "activate," "inhibit," "complex," etc.

- IV. Identify hub genes:
  - Go to Tools → Analyze Network and try analyzing as an undirected graph.



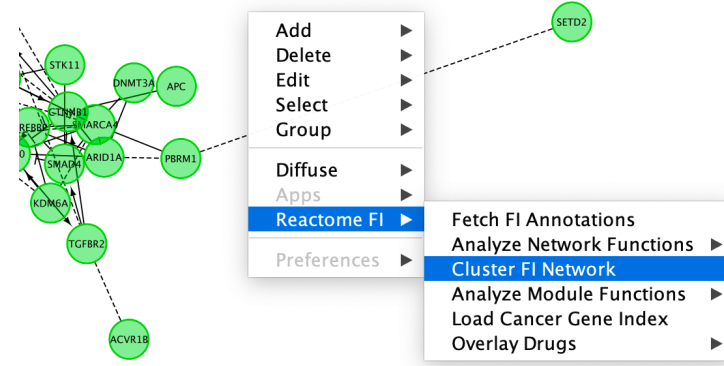
- Toggle back to the Node Table, sort by Degree (number of connections). You may have to scroll across the table.



Q1. Which mutated genes have the most interactions? Is it an oncogene or tumor suppressor gene?

### Step 1.4: Cluster the Network

- I. Right-click on any whitespace in the Network View Panel (not on the network itself). Under ReactomeFI, select Analyze Network Functions → Cluster FI Network

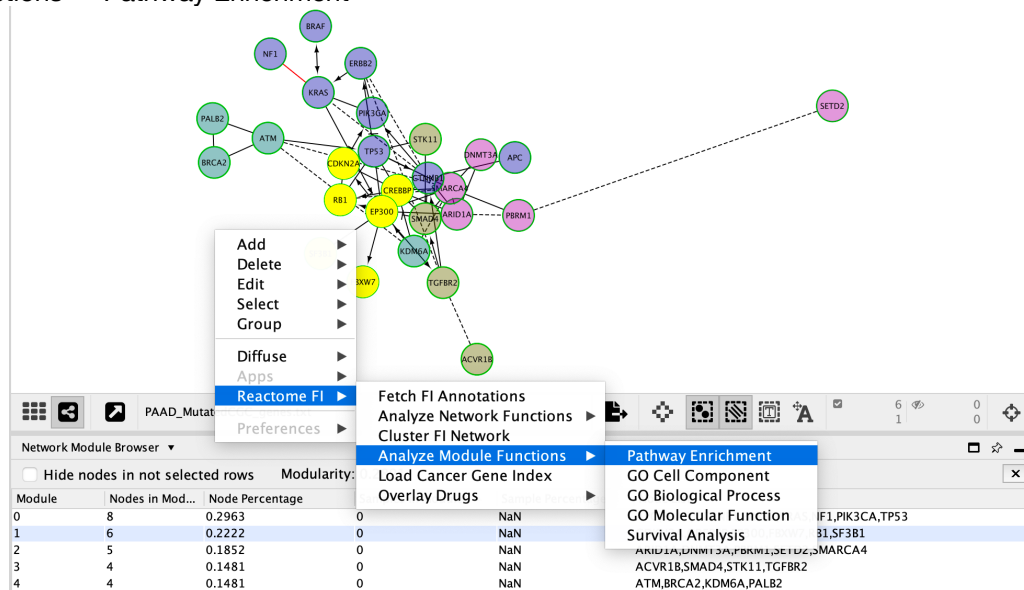


- II. Observe the resulting modules (coloured by cluster). The bottom table should now show the Network Module Browser. If not, select it in the drop-down menu. Try clicking on nodes in the network, and then rows in the table. What happens?

Q2: How many modules were identified? Which module contains *KRAS*?

### Step 1.5: Perform Pathway Enrichment on Modules

- I. Right-click on whitespace in the Network View Panel → ReactomeFI → Analyze Module Functions → Pathway Enrichment



- II. Cytoscape will ask you to Choose a Module Size. Use a size of 4 or 5. Very small modules don't contain enough genes for reliable pathway enrichment; any "enrichment" could be due to chance. A cutoff of 5 ensures each module has sufficient genes for meaningful statistics while

still capturing smaller biologically relevant clusters. For larger input gene lists (100+), you might increase this to 10.

- III. In the results table, examine the top pathways for each module. Note that you must scroll down to see the pathways enriched in other modules.

Q3: What pathways are enriched in the *KRAS*-containing module?

Q4: Do you see any overlap with the GSEA results from the expression analysis?

### Step 1.6: Note Key Results

1. Note the genes in the largest module containing *KRAS* (you'll need these for Part 3)
2. Take a screenshot of the clustered network for your records

## Exercise 2: Extract Leading Edge Genes from GSEA Results

**Goal:** This section bridges your GSEA results to network analysis. You'll extract leading edge genes from your GSEA output to use in Part 3 for integration with mutation data.

**Recall:** Why Leading Edge Genes? In GSEA, the leading edge is the subset of genes that contribute most to the enrichment signal. These genes:

- Are at the "front" of the ranked list (for positively enriched gene sets; at the "back" for negatively enriched sets)
- Represent the most biologically relevant genes for downstream integration

We will analyze the HALLMARK\_EPITHELIAL\_MESENCHYMAL\_TRANSITION together. You can repeat this analysis for any pathway of your choosing.

### Step 2.1: Open the gene set results table

- I. Navigate to your GSEA output folder
- II. In your GSEA output folder, find the `.tsv` file for your gene set
  - Example: `HALLMARK\_EPITHELIAL\_MESENCHYMAL\_TRANSITION.tsv`
- III. Open in Excel. Alternatively, you can do the Step 2.2 programmatically.

### Step 2.2: Record Your Leading Edge Genes

- I. Filter the CORE ENRICHMENT column for "Yes"
- II. Copy the SYMBOL column
- III. Past them into a new text file and save as .txt. You can compare your gene list with ours in GSEA\_HallmarkEMT\_leadingEdgeGenes.txt

### Exercise 3: GeneMANIA — Integrating GSEA and Mutation Data

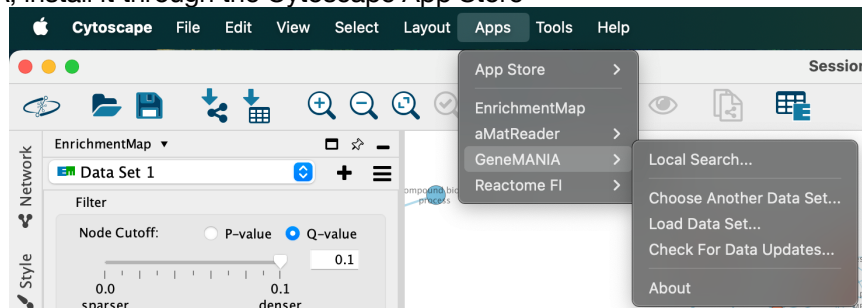
**Goal:** We now have two gene lists from different data types. We want to find out if these independently-derived gene lists connect through known biological interactions

**Recall:** Our gene lists are

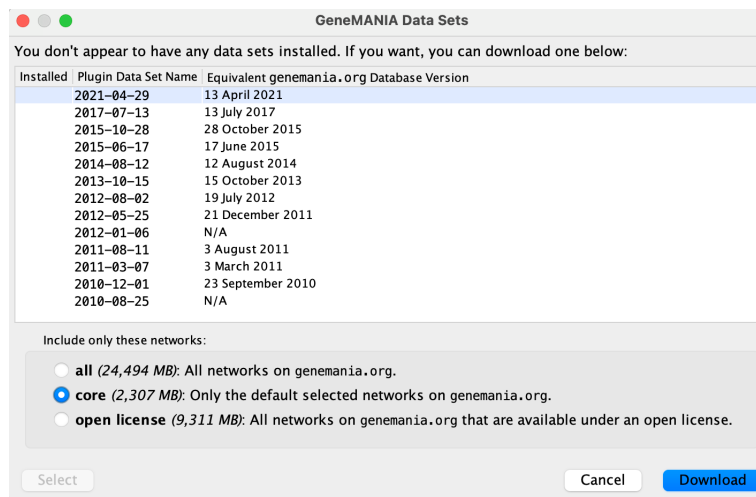
1. GSEA-derived (expression): Leading edge genes from Basal-enriched pathways (EMT, ECM, TGF- $\beta$ )
2. Mutation-derived: Cancer Gene Census genes mutated in PAAD

#### Step 3.1: Launch GeneMANIA

- I. In Cytoscape, find GeneMANIA in the App menu and select “Local Search”. If you do not see GeneMANIA, install it through the Cytoscape App Store



- II. If it's your first time running GeneMANIA, you'll be asked to download some datasets. Choose the most recent version, choose “core”, and then hit Download.

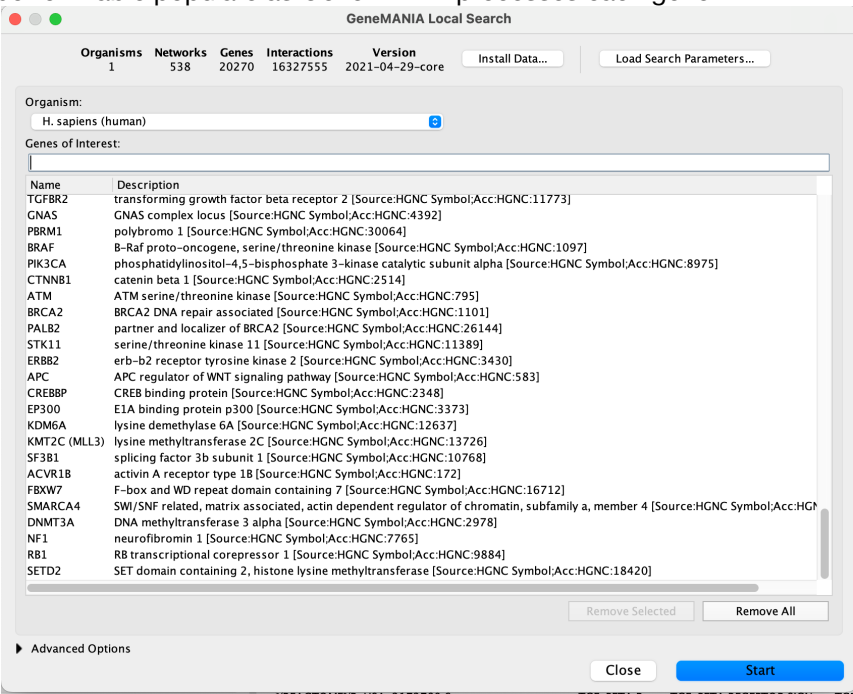


#### Step 3.2: Prepare Your Combined Gene List and Configure the GeneMANIA Search

- I. Combine genes from both analyses. You can simply copy the cosmic mutated genes (PAAD\_MutatedCGC\_genes.txt) and the GSEA Leading Edge genes for your pathway of interest. Remember that in this assignment, we are focusing on the EMT Hall mark genes (GSEA\_HallmarkEMT\_leadingEdgeGenes.txt).



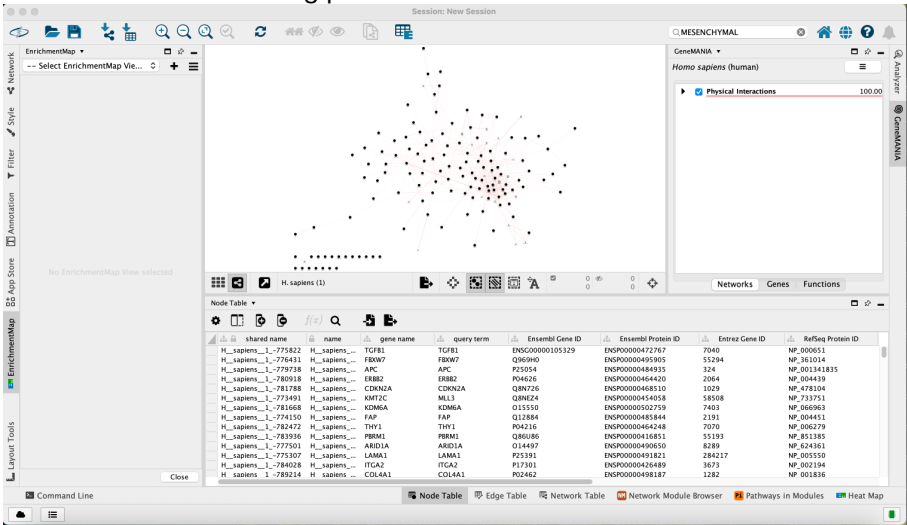
- II. Paste both gene lists into the “Genes of Interest” search bar in the GeneMANIA dialogue. You’ll see the bottom table populate as GeneMANIA processes each gene.



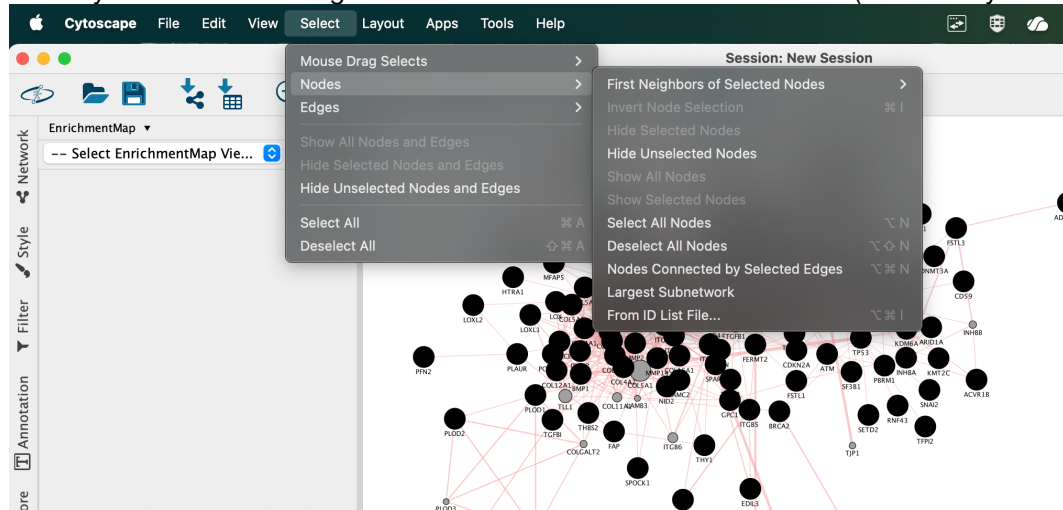
- III. Now configure the GeneMANIA search. Remember you may have to expand the Advanced Options.
- Organism: H. sapiens
  - Interaction Networks: For this analysis, select Physical interactions only. We’ll do this so simplify the analysis and decrease runtime. Physical interactions will also narrow our search down to genes that physically bind to each other. This provides mechanistic evidence that expression changes and mutations converge on the same protein complexes.
- IV. Click Start

### Step 3.4: Explore the Network

The network will load in the main viewing panel:



- I. Remember, you can Apply a layout if nodes are clustered together:
  - Go to Layout → Perforce Force Directed
- II. Understand the network:
  - Nodes = Your input genes
  - Node color = Indicate gene function (check legend in Results Panel). If you only see one colour, you'll notice they all indicate Physical Interactions.
  - Edges = Physical interactions between proteins
  - Edge thickness = Interaction confidence/weight
- III. Identify your two gene sources visually:
  - Select your GSEA-derived genes: Select → Nodes → From ID List File (or manually select)



- You should now notice some grey nodes in your network. Those are the genes you just input from your File. Note that you can use Style panel to control their appearance.

Q5: Which genes connect the mutation-derived genes to the expression-derived genes? These are potential "hub" genes.

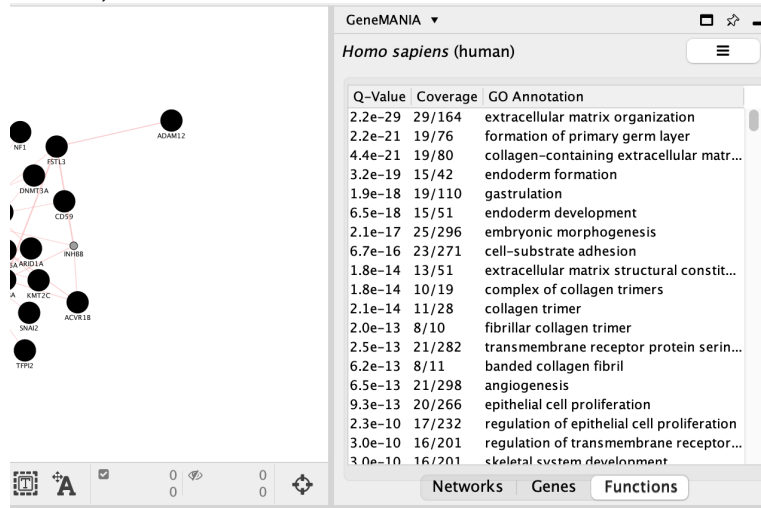
### Step 3.5: Identify Hub Genes

- I. Examine the node degree (number of connections) for each gene. Remember how to get the Degree column? Hint: Revisit Step 1.3.
- II. Find high-degree nodes:
  - Using the Node Table (bottom panel), find out which genes have the highest degree. Genes with the most connections are your "hub" genes
- III. Identify integration hubs:
  - Look for genes that connect BOTH mutation-derived AND expression-derived genes
  - These are the key integration points between your two data types

Q6: List the top 3 hub genes and their connection patterns.

### Step 3.6: Examine Functions

- I. In the Results Panel, click the Functions tab



- II. Review the predicted functions based on your gene set. Try clicking the terms.
- III. Note the top enriched GO terms and their FDR values

Q7: What functions does GeneMANIA predict for your gene set? Do they align with your GSEA results?

### Step 3.7: Export Your Network

- I. Save the session: File → Save Session As → 'PAAD\_GeneMANIA\_integration.cys'
- II. Export an image: File → Export → Network to Image
- III. Export the network table: File → Export → Table to File (select Node Table)

## Part 4: Synthesis and Interpretation (5 minutes)

You've now analyzed PAAD from four angles:

1. Expression (ORA): DE genes show over-representation of ECM and cell adhesion pathways
2. Expression (GSEA): Basal tumors show EMT, ECM, TGF- $\beta$  pathway activation
3. Mutations (ReactomeFI): KRAS, SMAD4, TGFBR2 cluster into signaling modules
4. Integration (GeneMANIA): Physical interactions link mutated genes to EMT expression signatures

### Consider the following

Q8: What is the biological story? How do mutations in KRAS and TGF- $\beta$  pathway genes lead to EMT expression changes?

Q9: SMAD4 is both mutated (loss-of-function) AND its target genes show expression changes. Is this consistent with what you know about TGF- $\beta$  signaling in cancer?

Q10: If you were designing a therapeutic strategy for Basal PAAD, which pathway would you target and why?

### The Big Picture

Our integrative analysis suggests that in PAAD:

- KRAS activation → drives proliferation and EMT
- SMAD4/TGFBR2 loss → removes growth inhibition from TGF- $\beta$
- Together → aggressive, mesenchymal phenotype (Basal subtype)

This multi-omics integration reveals convergent biology: multiple mutations disrupting the same pathway, leading to consistent transcriptional changes.

### Extra Content: Expanding Leading Edge Gene Extraction

In Part 2, you extracted leading edge genes from a single gene set (EMT). For a more comprehensive analysis, you can expand your gene list by extracting leading edge genes from multiple related pathways. Here are two options; the appropriate approach depends on your analysis goals. The outputs for both options are gene sets that you can use in GeneMANIA analysis.

#### Option 1: Manual Extraction from Multiple Gene Sets

Repeat Steps 2.1-2.2 for additional Basal-enriched gene sets:

1. You can choose any combinations of gene sets. For example,
  - HALLMARK\_EPITHELIAL\_MESENCHYMAL\_TRANSITION
  - GOBP\_EXTRACELLULAR\_MATRIX\_ORGANIZATION
  - GOBP\_COLLAGEN\_FIBRIL\_ORGANIZATION
  - REACTOME\_COLLAGEN\_FORMATION
  - KEGG\_TGF\_BETA\_SIGNALING\_PATHWAY
  - GOBP\_CELL\_SUBSTRATE\_ADHESION
2. Extract the Leading Edge Genes for each gene set:
  - Open the table containing your enrichment results
  - Copy genes with CORE ENRICHMENT = Yes
  - Paste or otherwise combine all Leading Edge Genes in one file.
3. Deduplicate:
  - Make sure to remove duplicate genes in your combined gene set
  - This gives you a broader "Basal signature" (~50-100 genes)

#### Option 2: GSEA Leading Edge Analysis Tool

GSEA has a built-in tool specifically designed to compare leading edge genes across multiple gene sets. We touched on this tool in Lab 2c, Step 3.2.

1. Launch Leading Edge Analysis:
  - In GSEA, go to Tools → Leading Edge Analysis
  - 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.
2. Select at least 2 gene sets:
  - Choose multiple significant gene sets from your results
  - For example, you could select all EMT/ECM/TGF- $\beta$  related pathways
3. Run the analysis:
  - Click \*\*Run\*\*
  - GSEA generates a heat map showing which genes appear in which gene sets
4. Identify hub genes:
  - Genes appearing in multiple gene sets are "transcriptional hubs"
  - These drive enrichment across multiple related pathways
  - Prioritize these for integration with mutation data

5. Export the gene list:
  - Click Gene Set Matrix to see the full matrix
  - Export genes that appear in your desired overlap of gene sets

For this workshop, the single gene set approach we covered in Part 2 is sufficient: EMT pathways already capture the core biology. In practice, you can use Option 2 above to systematically identify genes that drive enrichment across multiple related pathways.