



BIG DATA TO KNOWLEDGE  
CENTER OF EXCELLENCE



# Knowledge-guided Algorithms in Systems Biology

KnowEnG BD2K Center

*Slides by Charles Blatti and Amin Emad*

# Summary

- Our goal in this lab is to use several pipelines of the KnowEnG platform to analyze 'omic' data sets and phenotypic spreadsheets



- We will often try both network-guided and standard modes of operation for the pipelines (if applicable)
- Other network-guided and systems biology analysis tools will also be introduced

# More Specifically

- The structure of this lab is laid out around 3 example datasets
- It is focused on topics, methods, and types of networks from lecture
- It uses browser-based analysis platforms

Data Sets	Somatic Mutations from Pan-Cancer	Drug Response in Cancer Cell Lines	ER+ Status in Breast Cancer
Topics	Sample Clustering	Gene Prioritization	Gene Expression Signatures, Gene Set Characterization
Methods	Network Based Stratification	ProGENI	GeneMANIA, DRaWR,
Networks	Integrated	Protein-Protein Interactions	Pathways, Integrated
Platforms	KnowEnG	KnowEnG	iLINCS, GeneMANIA, KnowEnG

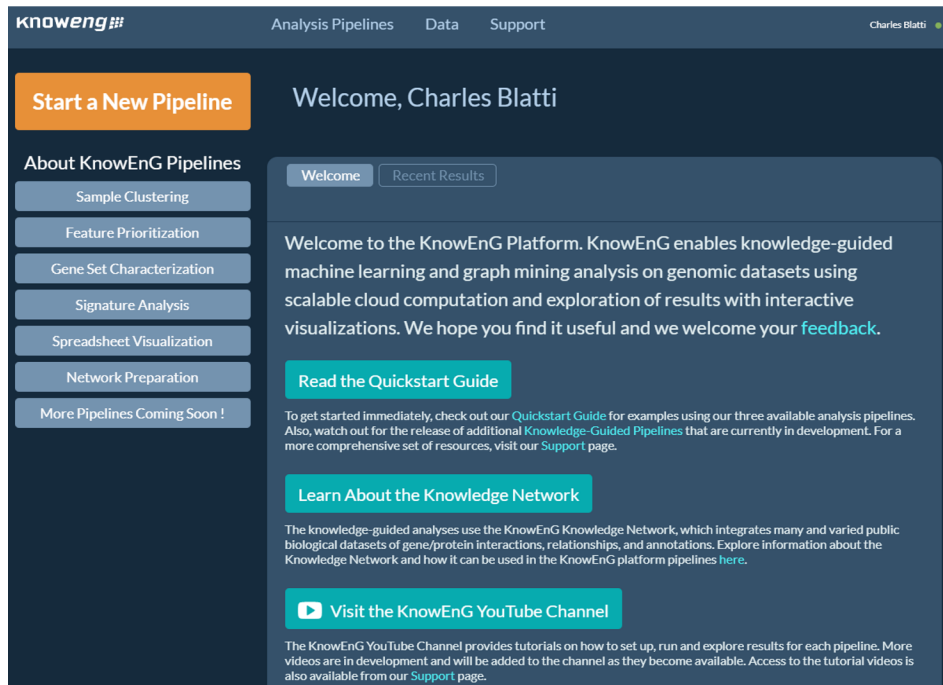
# Some Notes on the KnowEnG Platform

## Knowledge-guided analysis of "omics" data using the KnowEnG cloud platform

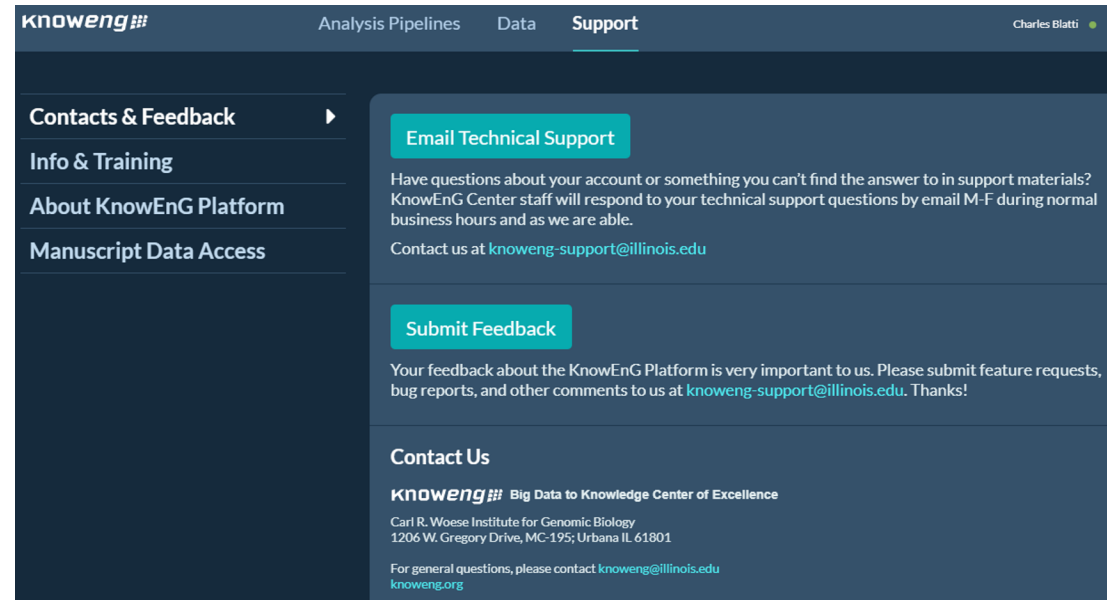
Charles Blatti III , Amin Emad , Matthew J. Berry, Lisa Gatzke, Milt Epstein, Daniel Lanier, Pramod Rizal, Jing Ge, Xiaoxia Liao, Omar Sobh, Mike Lambert, Corey S. Post, Jinfeng Xiao, [...].Saurabh Sinha   [view all]

Published: January 23, 2020 • <https://doi.org/10.1371/journal.pbio.3000583>

- The **home page** has links to many resources
- The **“Support”** tab at the top has even more resources
- Scalable platform using AWS cloud, but requires **some waiting**
- Right before launch, carefully **match** Job summaries to slide stills to avoid errors



The screenshot shows the KnowEnG home page. The top navigation bar includes 'Analysis Pipelines', 'Data', and 'Support'. The main content area is titled 'Welcome, Charles Blatti'. On the left, there is a sidebar with a 'Start a New Pipeline' button and a list of pipeline categories: Sample Clustering, Feature Prioritization, Gene Set Characterization, Signature Analysis, Spreadsheet Visualization, Network Preparation, and 'More Pipelines Coming Soon!'. The main content area has a 'Welcome' tab and a 'Recent Results' tab. The welcome message reads: 'Welcome to the KnowEnG Platform. KnowEnG enables knowledge-guided machine learning and graph mining analysis on genomic datasets using scalable cloud computation and exploration of results with interactive visualizations. We hope you find it useful and we welcome your feedback.' Below this, there are three buttons: 'Read the Quickstart Guide', 'Learn About the Knowledge Network', and 'Visit the KnowEnG YouTube Channel'. Each button has a corresponding paragraph of text explaining the resource.



The screenshot shows the KnowEnG Support page. The top navigation bar includes 'Analysis Pipelines', 'Data', and 'Support'. The main content area is titled 'Support'. On the left, there is a sidebar with a 'Contacts & Feedback' button and a list of categories: Info & Training, About KnowEnG Platform, and Manuscript Data Access. The main content area has three sections: 'Email Technical Support', 'Submit Feedback', and 'Contact Us'. Each section has a corresponding paragraph of text explaining the service or providing contact information.

# STEP 0A: Start the VM

- Follow instructions for starting VM. (This is the Remote Desktop software.)
- The instructions are different for UIUC and Mayo participants.
- Find the instructions for this on the course website under Lab Set-up:  
<https://publish.illinois.edu/compgenomicscourse/2023-schedule/>

# Step 0: Local Files

For viewing and manipulating the files needed for this laboratory exercise, the path on the VM will be denoted as the following:

**[course\_directory]**

We will use the files found in:

**[course\_directory]\07\_Signatures\_and\_Characterization**

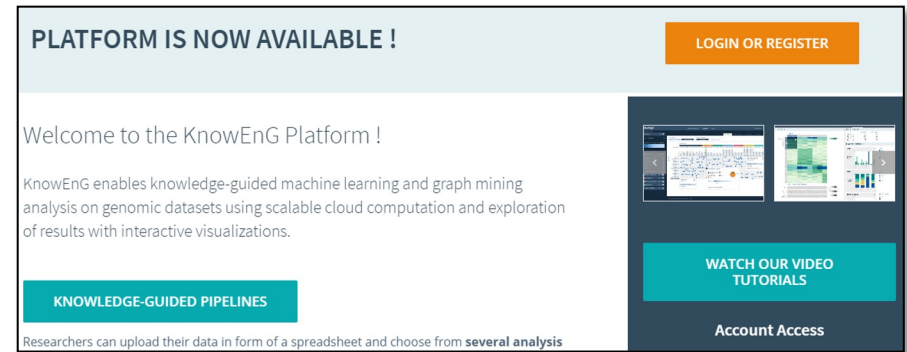
**-and-**

**[course\_directory]\08\_Clustering\_and\_Prioritization**

**[course\_directory]= Desktop\VM**

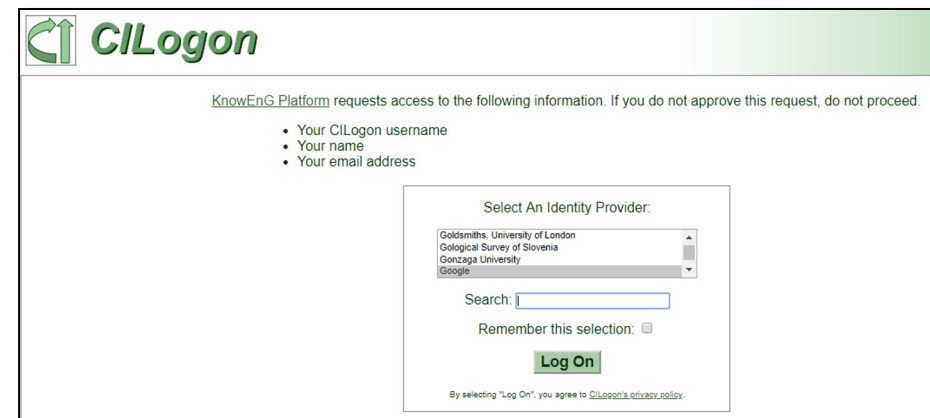
# STEP 1: Sign Into KnowEnG Platform

Go to the KnowEnG Platform:  
<https://knoweng.org/analyze/>



Click “Login or Register”

Login with **CILogon** - Login service using your other existing accounts  
Search for identity provider: **Urbana, Mayo, Google, GitHub**



# Finding Cancer Subtypes with Knowledge Guided Clustering

In this exercise, we will use a subset of somatic mutation data samples from the Cancer Genome Atlas (TCGA) and cluster them into different cancer subtypes.



# STEP 2: Sample Clustering

- We will use KnowEnG's clustering pipeline to perform both network-guided as well as standard clustering of samples
- The network-guided clustering implemented in KnowEnG is inspired by the network-based stratification approach:

[Nat Methods](#). 2013 Nov;10(11):1108-15. doi: 10.1038/nmeth.2651. Epub 2013 Sep 15.

## **Network-based stratification of tumor mutations.**

[Hofree M](#)<sup>1</sup>, [Shen JP](#), [Carter H](#), [Gross A](#), [Ideker T](#).

---

- We will use some of the samples from the TCGA pancan12 dataset

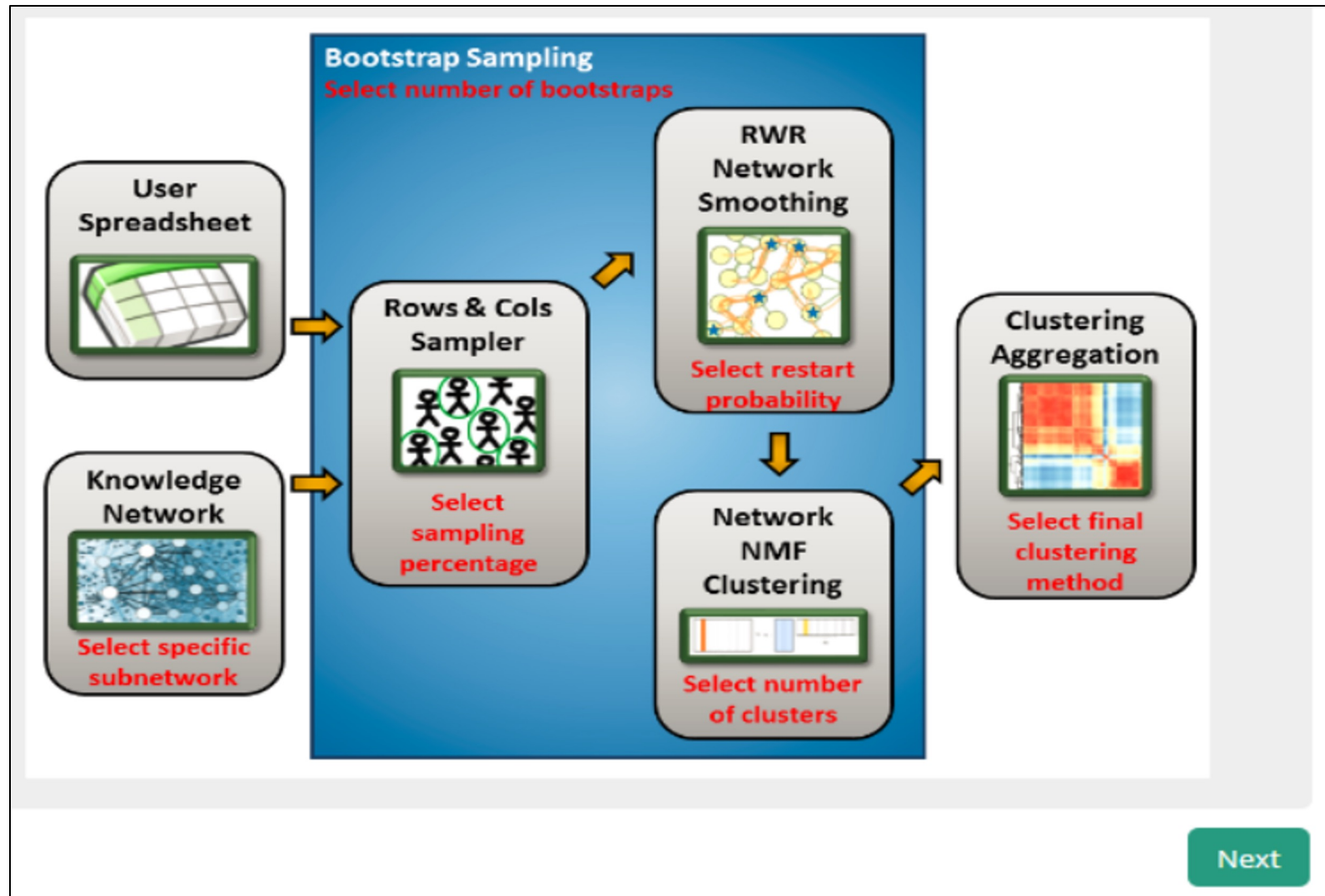
[Cell](#). 2014 Aug 14;158(4):929-944. doi: 10.1016/j.cell.2014.06.049. Epub 2014 Aug 7.

## **Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin.**

[Hoadley KA](#)<sup>1</sup>, [Yau C](#)<sup>2</sup>, [Wolf DM](#)<sup>3</sup>, [Cherniack AD](#)<sup>4</sup>, [Tamborero D](#)<sup>5</sup>, [Ng S](#)<sup>6</sup>, [Leiserson MDM](#)<sup>7</sup>, [Niu B](#)<sup>8</sup>, [McLellan MD](#)<sup>8</sup>, [Uzunangelov V](#)<sup>6</sup>, [Zhang J](#)<sup>9</sup>, [Kandoth C](#)<sup>8</sup>, [Akban R](#)<sup>10</sup>, [Shen H](#)<sup>11</sup>, [Omberg L](#)<sup>12</sup>, [Chu A](#)<sup>13</sup>, [Margolin AA](#)<sup>12</sup>, [Van't Veer LJ](#)<sup>3</sup>, [Lopez-Bigas N](#)<sup>14</sup>, [Laird PW](#)<sup>11</sup>, [Raphael BJ](#)<sup>7</sup>, [Ding L](#)<sup>8</sup>, [Robertson AG](#)<sup>13</sup>, [Byers LA](#)<sup>10</sup>, [Mills GB](#)<sup>10</sup>, [Weinstein JN](#)<sup>10</sup>, [Van Waes C](#)<sup>15</sup>, [Chen Z](#)<sup>16</sup>, [Collisson EA](#)<sup>17</sup>; Cancer Genome Atlas Research Network, [Benz CC](#)<sup>18</sup>, [Perou CM](#)<sup>19</sup>, [Stuart JM](#)<sup>20</sup>.

# STEP 2: Sample Clustering

- Overview of KnowEnG's Network-based Stratification for Samples:



# STEP 2: Sample Clustering

Find the files in this slide under [\[course\\_directory\]/08\\_Clustering\\_and\\_Prioritization](#)

- Dataset characteristics:

Name	Description
Demo2_Mutation_pancan12_30	A matrix of (gene x samples) containing the somatic mutation status of ~15k protein coding genes in 360 tumor samples from 12 cancer types.
Demo2_Clinical_pancan12_30	A matrix of (samples x clinical phenotypes) including primary disease, PANCAN consensus cluster, survival years, etc.

# STEP 2A: Sample Clustering (standard)

## Select the pipeline:

- Once logged into the KnowEnG Platform
- Select “**Analysis Pipelines**” at the top of the page
- Select “**Sample Clustering**” and Click on “**Start Pipeline**”

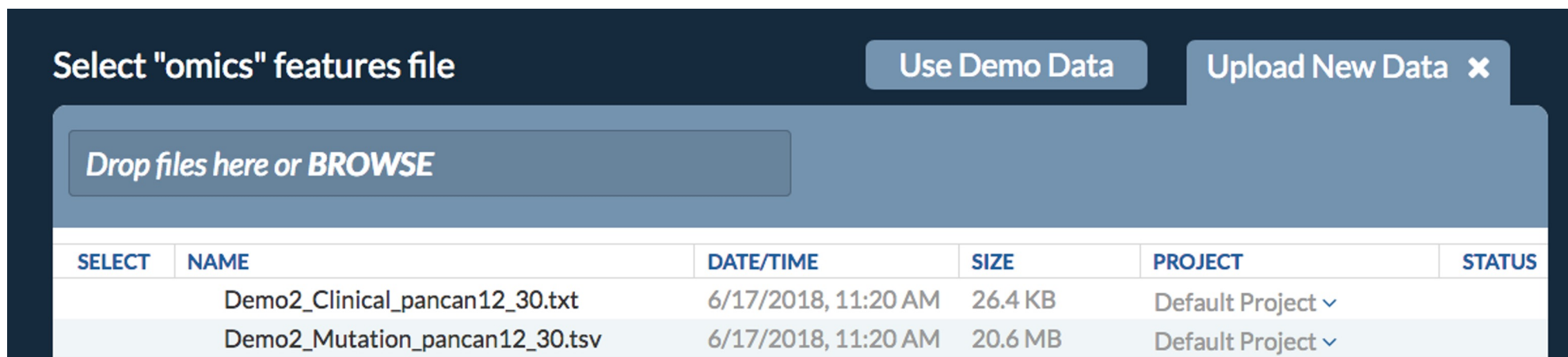
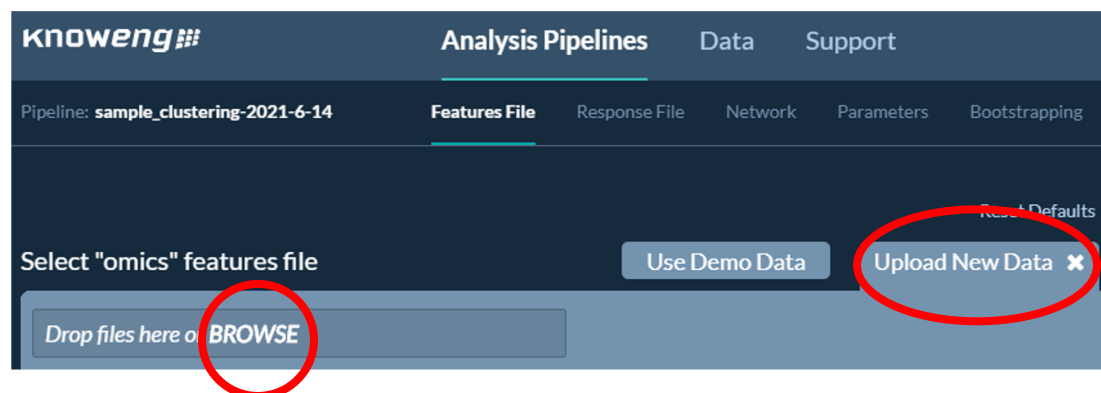
The screenshot displays the KnowEnG platform interface. At the top, the 'knoweng' logo is on the left, and navigation links for 'Analysis Pipelines', 'Data', and 'Support' are on the right. The 'Analysis Pipelines' link is circled in red. Below this, a secondary navigation bar shows 'Analysis Pipelines', 'Data', and 'Support' with 'Analysis Pipelines' underlined. The main content area is titled 'SELECT A PIPELINE' in teal. A list of pipeline options is shown: 'Sample Clustering', 'Feature Prioritization', 'Gene Set Characterization', 'Signature Analysis', and 'Spreadsheet Visualization'. The 'Sample Clustering' option is highlighted with a red oval, and a teal 'Start Pipeline' button is visible to its right.

Find the files for this under [course\_directory]/08\_Clustering\_and\_Prioritization

## STEP 2A: Sample Clustering (standard)

### Upload the data:

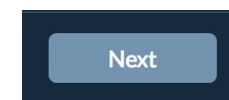
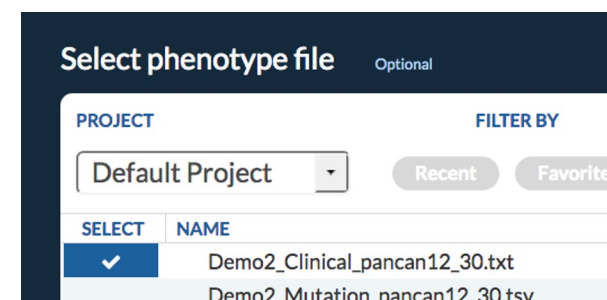
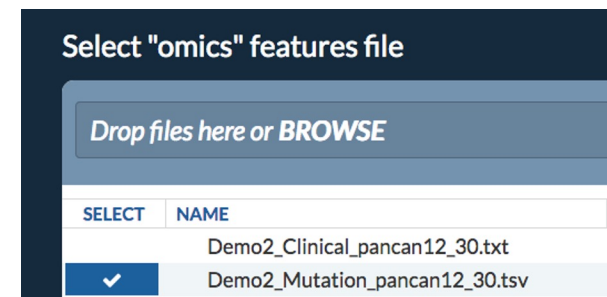
- Click on “**Upload New Data**”
- Click “**BROWSE**” and find the files to upload:
  - Demo2\_Clinical\_pancan12\_30
  - Demo2\_Mutation\_pancan12\_30.tsv



# STEP 2A: Sample Clustering (standard)

## Configure the pipeline:

- For the “**omics**” file select:
  - Demo2\_Mutation\_pancan12\_30.tsv
- Click “**Next**” at the bottom right corner
- For the “**phenotype**” file select:
  - Demo2\_Clinical\_pancan12\_30.txt
- Click “**Next**” at the bottom right corner



# STEP 2A: Sample Clustering (standard)

- Select “**No**” in response to using the knowledge network:
  - *This allows us to perform standard clustering on the data*
- Click on “**Next**” at the bottom right corner
- Choose **8** as number of clusters
  - *This is what was found as optimal in the TCGA paper*
- We will use the default “**K-Means**” clustering algorithm
- Click on “**Next**” at the bottom right corner

Do you want to use the Knowledge Network?

Yes No

1 Enter the number of clusters you wish the analysis to return

8

2 Select the clustering algorithm to use

K-Means default

Next

# STEP 2A: Sample Clustering (standard)

- Select “**Yes**” in response to using bootstrap sampling:
  - *This allows us to obtain a more robust final clustering*
- Choose **5** as number of bootstraps.
  - *This is unusually low for the purposes of quicker completion*
- We will use the default **80%** rate to sample the data in each bootstrap
- Click on “**Next**” at the bottom right corner

Do you want to use bootstrapping?

1 Enter the number of bootstraps to use in your analysis

2 Enter the bootstrap sample percent

% default



# STEP 2A: Sample Clustering (standard)

- Review the summary of the job and change the default “**Job Name**” to easily recognize later

Features File	Demo2_Mutation_pancan12_30.tsv	Job Name	SC_nonet_clust8
Response File	Demo2_Clinical_pancan12_30.txt	Project	Default Project
Use Network	No	Notes	
# Clusters	8		
Method	K-Means		
Use Bootstrapping	Yes		
# Bootstraps	5		
Sample %	80%		

- **Submit the job**
  - *The job will run for several minutes, so we will return to the results after launching the next job*

# STEP 2B: Sample Clustering (network-guided)

*Now we are going to repeat the analysis using a knowledge network to provide richer information about the similarity between the sparse mutation samples. Nearly all steps will be the same as before.*

## Select the pipeline:

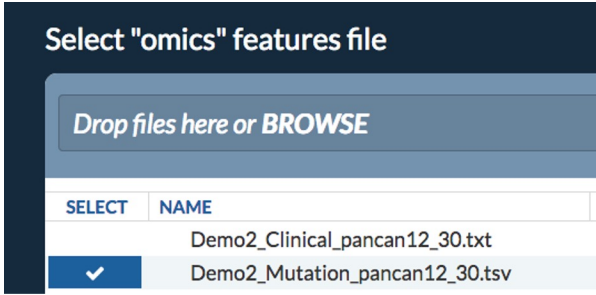
- Select “**Analysis Pipelines**” at the top of the page on the same webpage
- Select “**Sample Clustering**” and Click on “**Start Pipeline**”

The screenshot shows the Knoweng website interface. At the top, there is a dark blue navigation bar with the Knoweng logo on the left and three menu items: 'Analysis Pipelines', 'Data', and 'Support'. The 'Analysis Pipelines' menu item is circled in red. Below this is a secondary navigation bar with the same three items, where 'Analysis Pipelines' is underlined. The main content area is titled 'SELECT A PIPELINE' in light blue. Below this title, there is a list of pipeline options: 'Sample Clustering', 'Feature Prioritization', 'Gene Set Characterization', 'Signature Analysis', and 'Spreadsheet Visualization'. The 'Sample Clustering' option is highlighted with a red oval, and a 'Start Pipeline' button is visible to its right, also circled in red.

# STEP 2B: Sample Clustering (network-guided)

## Configure the pipeline:

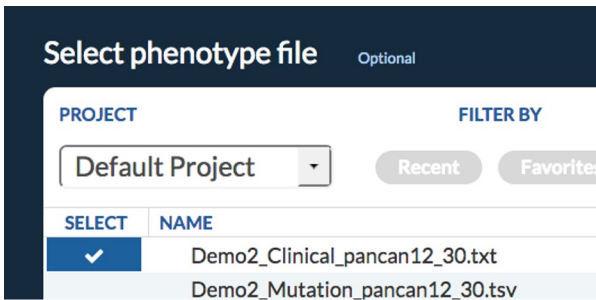
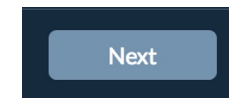
- For the “**omics**” file select:
  - Demo2\_Mutation\_pancan12\_30.tsv
- Click “**Next**” at the bottom right corner
- For the “**phenotype**” file select:
  - Demo2\_Clinical\_pancan12\_30
- Click “**Next**” at the bottom right corner



Select "omics" features file

Drop files here or BROWSE

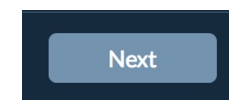
SELECT	NAME
	Demo2_Clinical_pancan12_30.txt
<input checked="" type="checkbox"/>	Demo2_Mutation_pancan12_30.tsv



Select phenotype file Optional

PROJECT: Default Project FILTER BY: Recent Favorite

SELECT	NAME
<input checked="" type="checkbox"/>	Demo2_Clinical_pancan12_30.txt
	Demo2_Mutation_pancan12_30.tsv



# STEP 2B: Sample Clustering (network-guided)

*This is different from the previous run.*

- Select **“Yes”** in response to using the knowledge network:
  - *This allows us to perform network-guided clustering*
- Keep the species as **“Human”**
- Select **“HumanNet Integrated Network”** as the network
  - *This is a network that creates scores pairwise interactions of gene by combining many different types of gene relationships*
- Keep network smoothing at **50%** and **click Next:**
  - *This controls how much importance is put on network connections instead of the somatic mutations*

Do you want to use the Knowledge Network?

Yes  No

1 Select species

Human (Hsap) default

2 Select Interaction Network for analysis

HumanNet Integrated Network

3 Choose the amount of network smoothing

50 % default

Next

# STEP 2B: Sample Clustering (network-guided)

- Choose **8** as number of clusters and click **Next**
- Select “**Yes**” in response to using bootstrap sampling:
- Choose **5** as number of bootstraps
- We will use the default **80%** rate to sample the data in each bootstrap
- Click “**Next**”

Enter the number of clusters you wish the analysis to return

Do you want to use bootstrapping?

Yes

No

1 Enter the number of bootstraps to use in your analysis

2 Enter the bootstrap sample percent

%

*default*

# STEP 2B: Sample Clustering (network-guided)

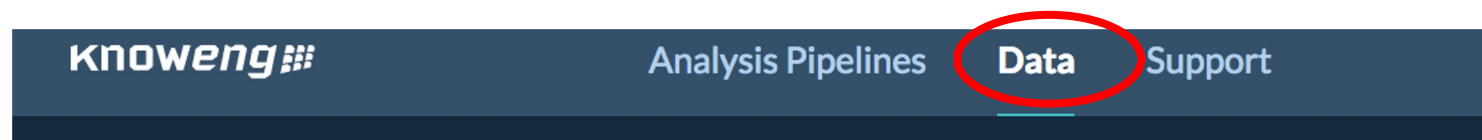
- Review the summary of the job and change the default “**Job Name**” to easily recognize later

Features File	Demo2_Mutation_pancan12_30.tsv	Job Name	SC_HumanNet_clust8
Response File	Demo2_Clinical_pancan12_30.txt	Project	Default Project
Use Network	Yes	Notes	
Species	Human (Hsap)		
Interaction Network	HumanNet Integrated Network		
Network Smoothing	50%		
# Clusters	8		
Use Bootstrapping	Yes		
# Bootstraps	5		
Sample %	80%		

- Press **Submit Job**



# STEP 2C: Standard Clustering Results

- Go to the “Data” page:



- Select “SC\_nonet\_clust8” (or other name you chose for the first run)

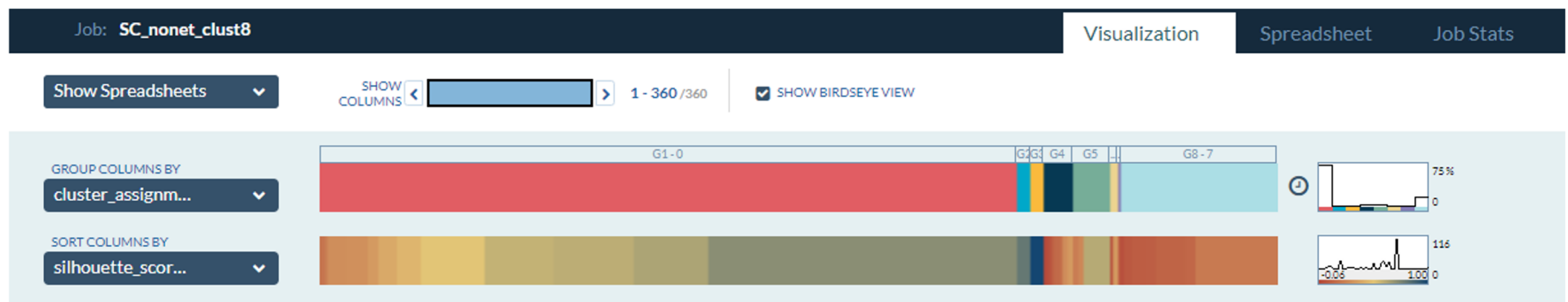
SC_HumanNet_clust8	6/17/2018, 1:27 PM	Default Project ▾
SC_nonet_clust8	6/17/2018, 11:56 AM	Default Project ▾

- Select “View Results” at the top right corner
  - *The option to view results will not be available if*
    - *The job is still running:* 
    - *There was an error:* 
  - *If there’s an error, try repeating the launch steps*

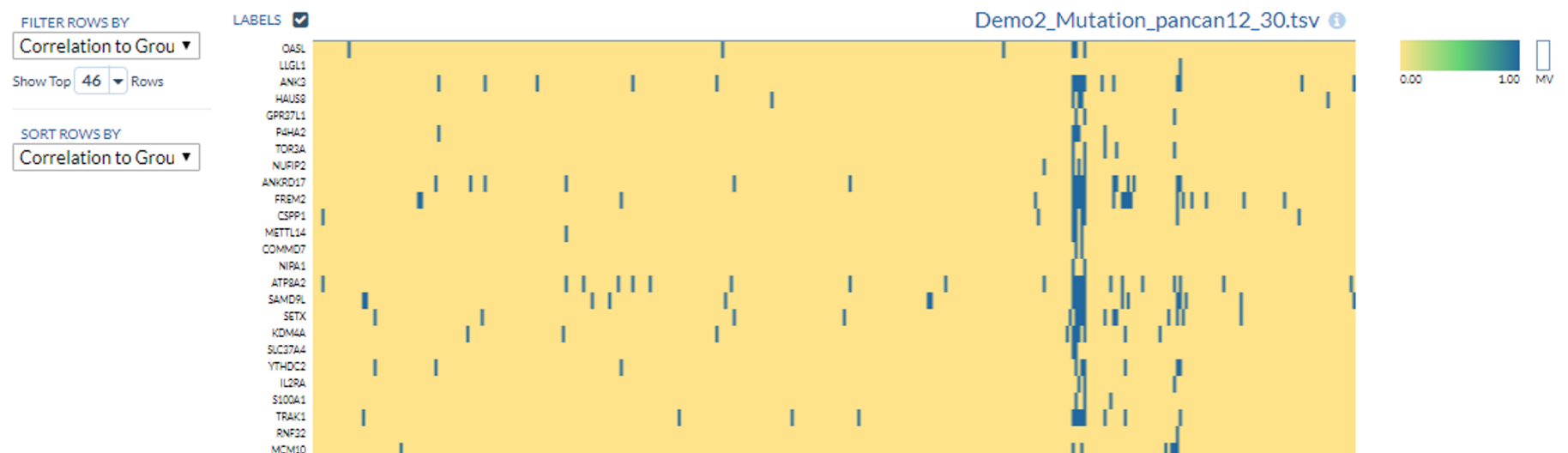
A light blue panel containing a 'View Results' button on the left. To its right are three icons: a download icon, a delete icon (a circle with an X), and a favorite icon (a star). Below each icon is its corresponding label: 'download', 'delete', and 'favorite'. Above the icons, the text 'FEATURES RESPONSE' is on the left, and 'Demo2\_Mutation\_pancan12...' and 'Demo2\_Clinical\_pancan12...' are on the right.

# STEP 2C: Standard Clustering Results

- Visualization shows the cluster sizes and the match of the samples to the cluster (silhouette\_score)



- Heatmap shows genes x samples – significantly correlated mutations

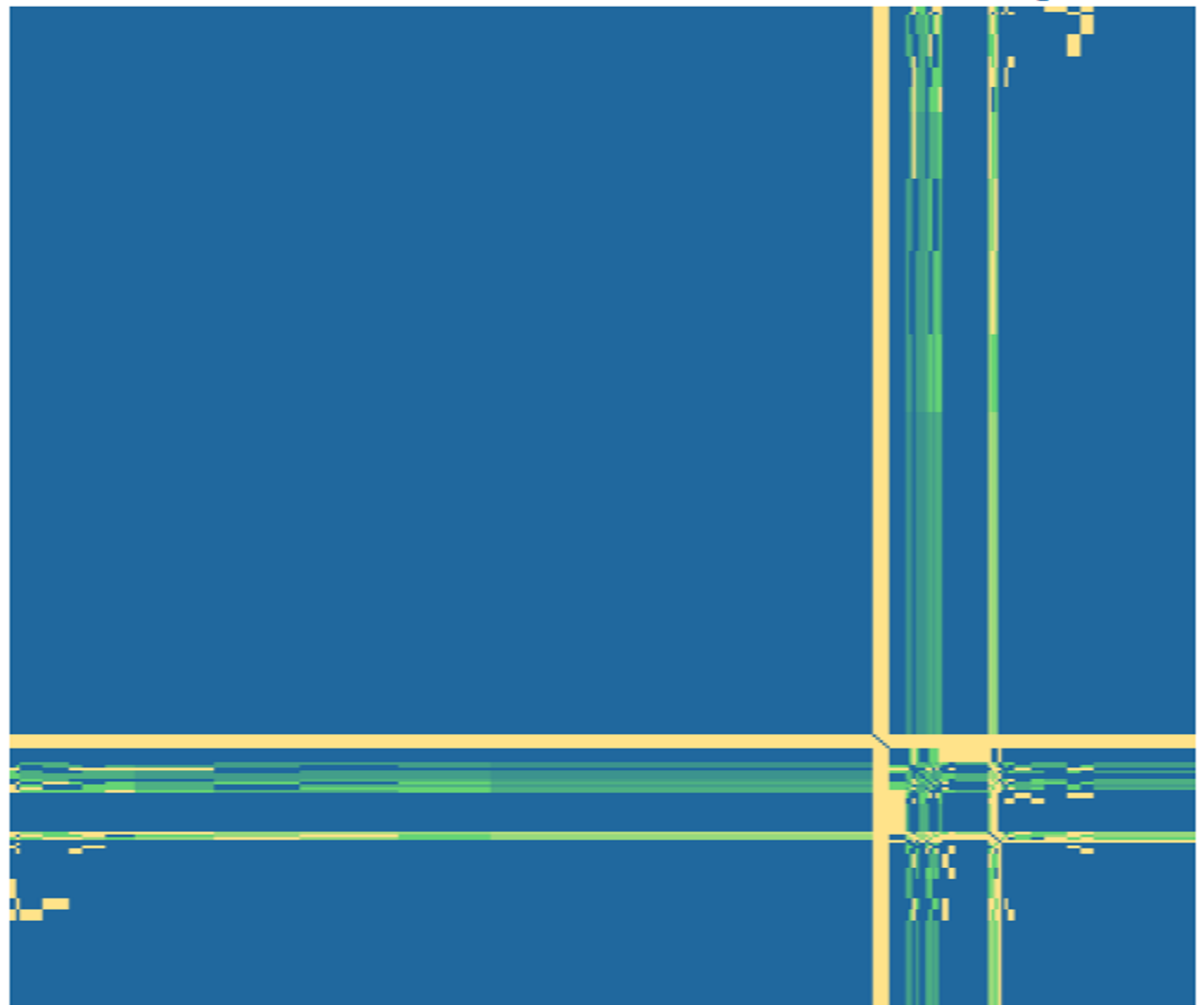




# STEP 2C: Standard Clustering Results

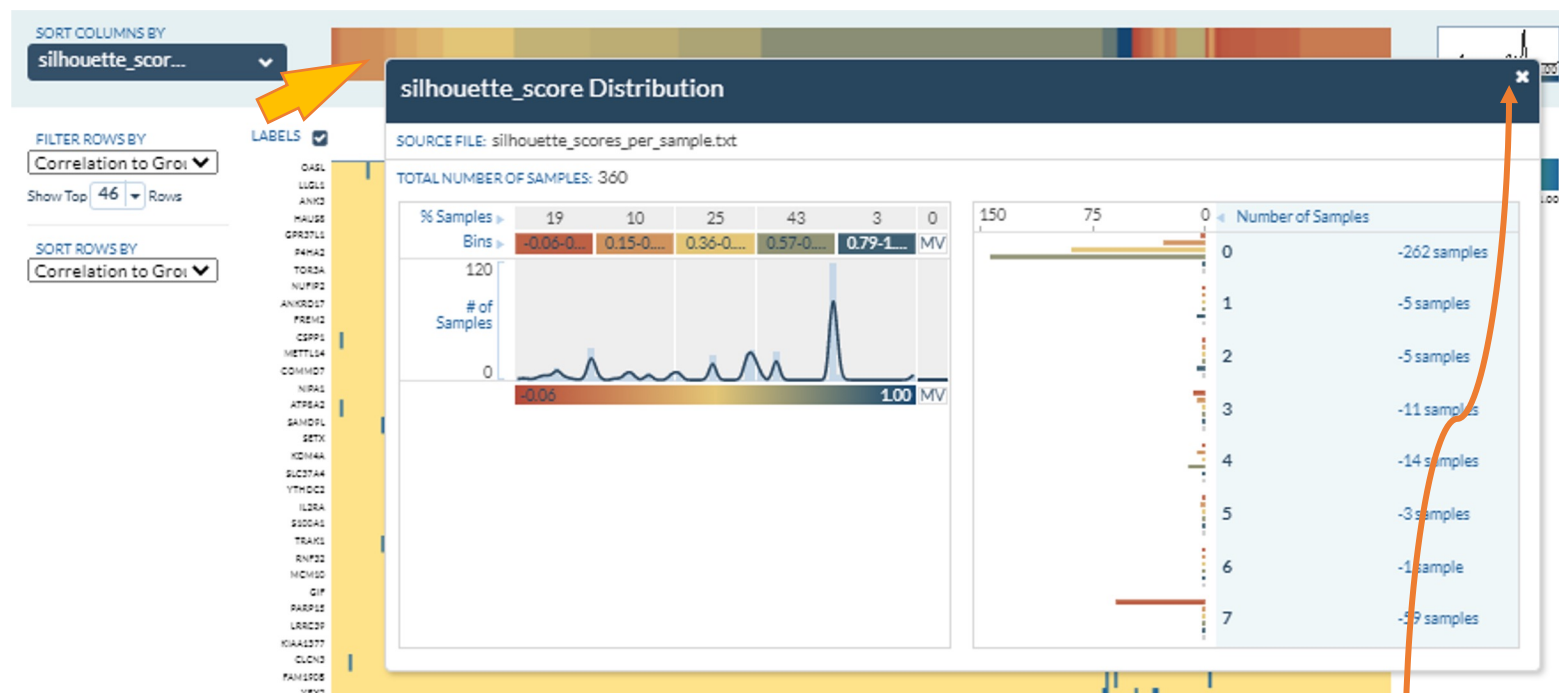
- Heatmap also shows samples x samples co-occurrence

The color of each cell indicates how frequently a pair of patients fell within the same cluster across all samplings



# STEP 2C: Standard Clustering Results

- Click on the **silhouette\_score** Distribution colorbar.
- The **Number of Samples** per cluster show high degree of clustering bias. *262 of the 360 samples are in Cluster 0*



- Close the Distribution panel with the 'X' in the top corner

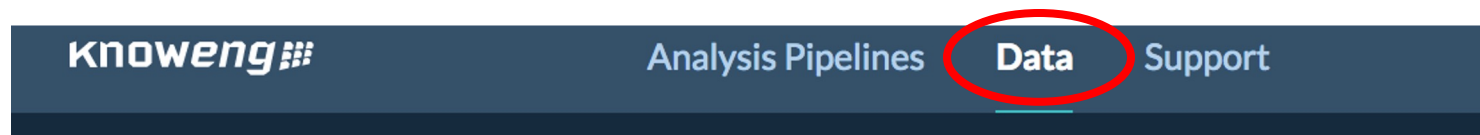
# STEP 2C: Standard Clustering Results

- You can add a phenotype to compare to the clustering at the very bottom of the page
  - click “**Show Rows**”,
  - the name of the clinical file, “**Demo2\_Clinical\_pancan12\_30.txt**”,
  - and select an interesting phenotype, like the “**\_primary\_disease**” type,
  - and click “**Done**”
- This color bar shows the original primary tumor (**\_primary\_disease**) types. Click on the **colorbar** to show which cancer types are present in which clusters



# STEP 2D: Network Clustering Results

- Go to the “Data” page:






- Select “SC\_HumanNet\_clust8” (or any other name you chose)

Phenotype_METABRIC_Demo1.txt	6/16/2018, 5:59 PM	182.4 KB	Default Project ▾
▶ SC_HumanNet_clust8	6/17/2018, 1:27 PM		Default Project ▾
▶ SC_nonet_clust8	6/17/2018, 11:56 AM		Default Project ▾

- Select “View Results” at the top right corner

### SC\_HumanNet\_clust8

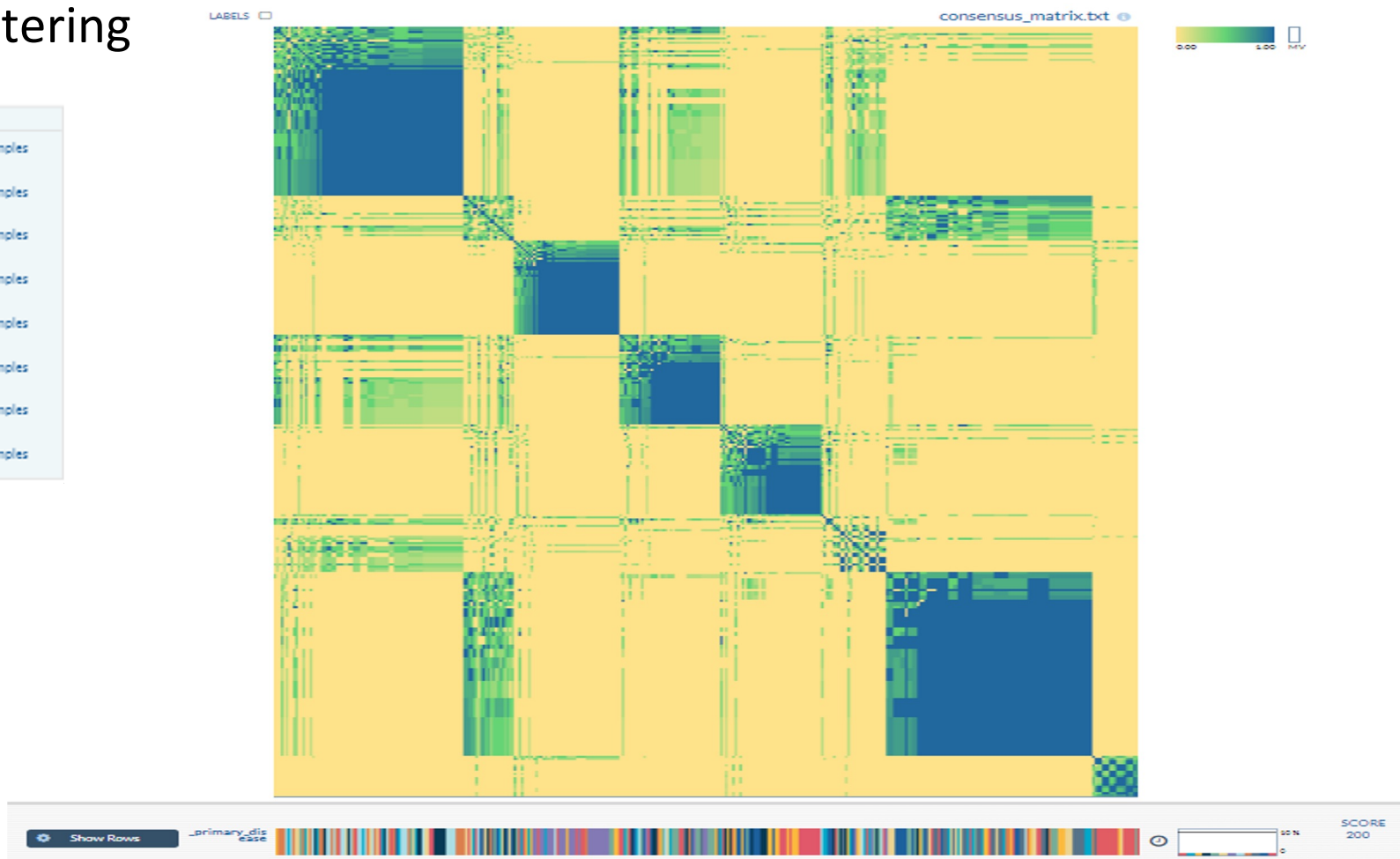
FEATURES      Demo2\_Mutation\_pancan12\_...  
RESPONSE      Demo2\_Clinical\_pancan12\_...

[View Results](#)       download       delete       favorite

# STEP 2D: Network Clustering Results

- As you can see from the sample distribution and the co-occurrence matrix, the network-guided approach provided a more balanced clustering

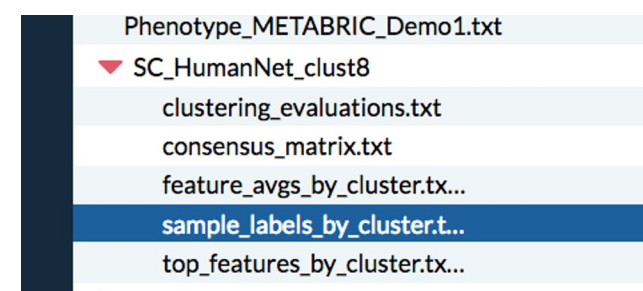
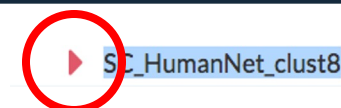
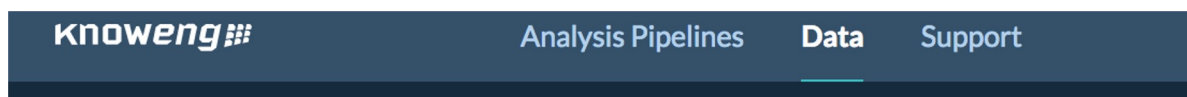
Number of Samples	
0	-79 samples
1	-21 samples
2	-44 samples
3	-42 samples
4	-42 samples
5	-27 samples
6	-86 samples
7	-19 samples



# STEP 2E: Compare in Spreadsheet Visualizer

## To Prepare the Files:

- Go to the “**Data**” page
- Click on triangle by “**SC\_HumanNet\_clust8**” (or any other name you chose)
- Select “**sample\_labels\_by\_cluster.txt**” results file of the network run
- Click on the name at the right top corner to edit and add “**\_HumanNet**” to the end
- Repeat the same for the file in “**SC\_nonet\_clust8**” and add “**\_nonet**” to the end



# STEP 2E: Compare in Spreadsheet Visualizer

- Let's compare the two runs in KnowEnG's Spreadsheet Visualization Tool
- Select **"Analysis Pipelines"**
- Select **"Spreadsheet Visualization"** and Click on **"Start Pipeline"**

The screenshot shows the KnowEnG web interface. At the top, there is a dark blue navigation bar with the KnowEnG logo on the left and three menu items: 'Analysis Pipelines', 'Data', and 'Support'. The 'Analysis Pipelines' menu item is circled in red. Below this bar, the 'Analysis Pipelines' page is displayed. It has a sub-navigation bar with 'Analysis Pipelines', 'Data', and 'Support'. The main content area is titled 'SELECT A PIPELINE' and lists four options: 'Sample Clustering', 'Feature Prioritization', 'Gene Set Characterization', and 'Spreadsheet Visualization'. The 'Spreadsheet Visualization' option is highlighted with a red oval, and a teal 'Start Pipeline' button is visible to its right.

# STEP 2E: Compare in Spreadsheet Visualizer

- Select these four files to evaluate simultaneously and press **Next**:
- Check the summary and change the job name if you like. Press **Submit Job**.


SELECT	NAME
<input checked="" type="checkbox"/>	Demo2_Clinical_pancan12_30.txt
<input checked="" type="checkbox"/>	Demo2_Mutation_pancan12_30.tsv
<input type="checkbox"/>	SC_HumanNet_clust8
<input type="checkbox"/>	clustering_evaluations.txt
<input type="checkbox"/>	consensus_SC_HumanNet_cl...
<input type="checkbox"/>	feature_avgs_by_cluster.tx...
<input checked="" type="checkbox"/>	sample_labels_by_cluster_H...
<input type="checkbox"/>	top_features_by_cluster.tx...
<input type="checkbox"/>	SC_nonet_clust8
<input type="checkbox"/>	clustering_evaluations.txt
<input type="checkbox"/>	consensus_matrix.txt
<input type="checkbox"/>	feature_avgs_by_cluster.tx...
<input checked="" type="checkbox"/>	sample_labels_by_cluster_n...
<input type="checkbox"/>	top_features_by_cluster.tx...
<input type="checkbox"/>	spreadsheet_visualization-Demo...

Spreadsheet File(s)

- Demo2\_Clinical\_pancan12\_30.txt
- Demo2\_Mutation\_pancan12\_30.tsv
- SC\_HumanNet\_clust8/sample\_labels\_by\_cluster\_HumanNet.txt
- SC\_nonet\_clust8/sample\_labels\_by\_cluster\_nonet.txt

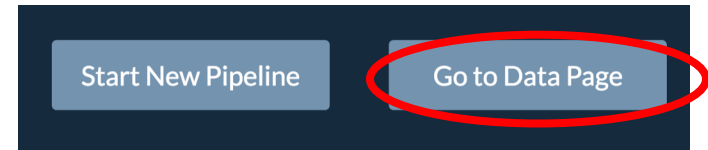
Job Name

Project

Notes 



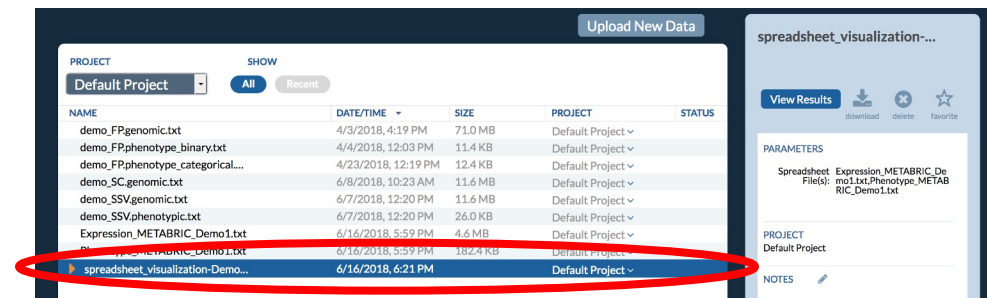
# STEP 2E: Compare in Spreadsheet Visualizer



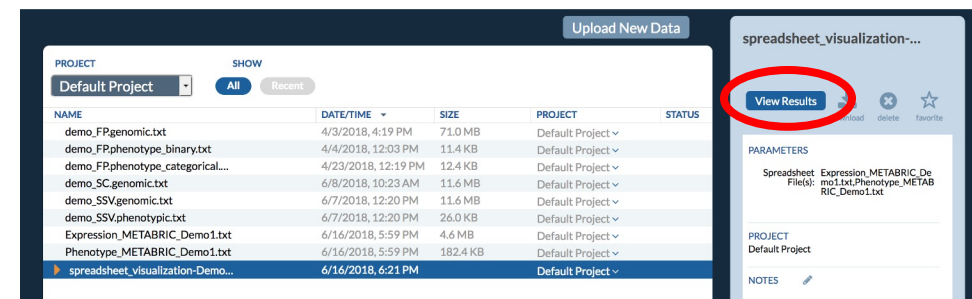
The results:

- Select “Go to Data Page”

- Select the job you just ran

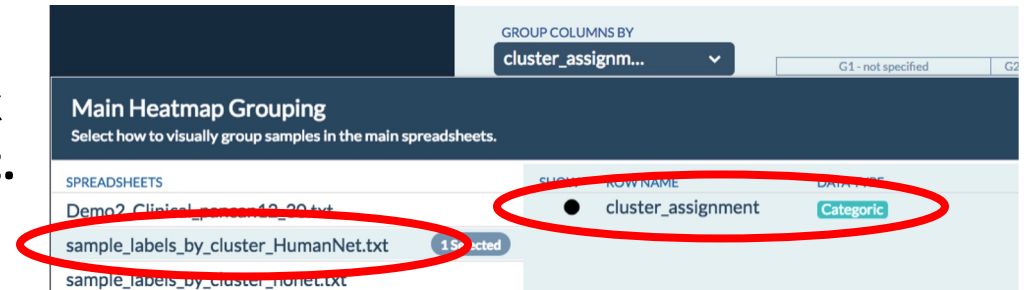


- Then “View Results”



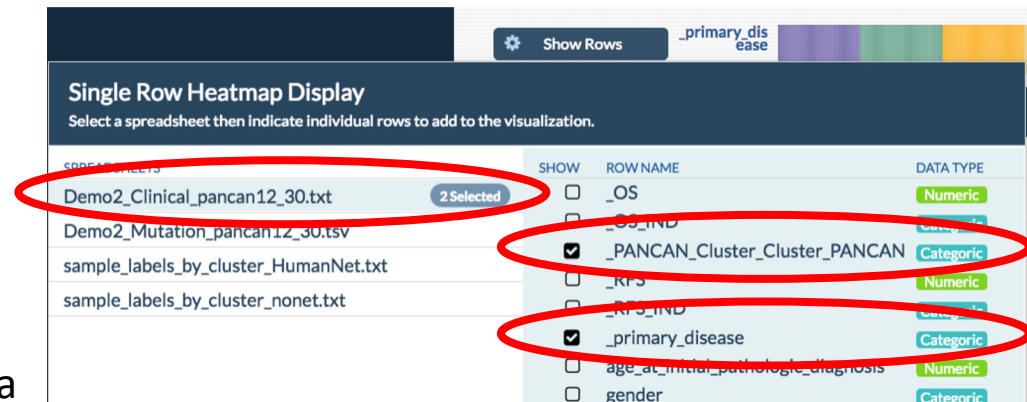
# STEP 2E: Compare in Spreadsheet Visualizer

- In the “**Group Columns By**” drop down click the “**sample\_labels\_by\_cluster\_HumanNet.txt**” network-guided clustering results file; then select “**cluster\_assignment**”

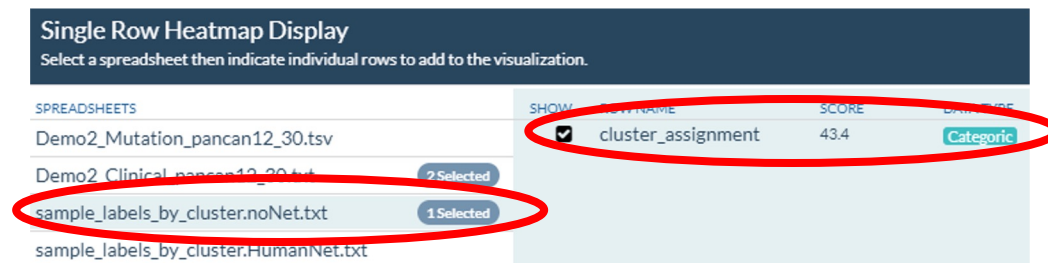


- Click "Done"

- By clicking on “**Show Rows**” at the very bottom of the page add the colorbars
  - “**\_primary\_disease**” and “**\_PANCAN\_Cluster\_Cluster\_PANCAN**” from “**Demo2\_Clinical\_pancan12\_30.txt**” clinical data
  - and “**cluster\_assignment**” from the “**sample\_labels\_by\_cluster\_noNet.txt**” standard clustering results file



- Click "Done"



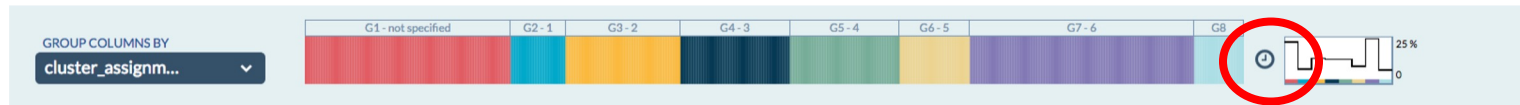
# STEP 2E: Compare in Spreadsheet Visualizer

- You can use this tool to explore top genes, draw Kaplan Meier curves, etc.



# STEP 2E: Compare in Spreadsheet Visualizer

- Click on the clock sign to perform Kaplan Meier survival analysis using the selected network-guided cancer subtypes



- Use this table to configure Kaplan Meier analysis by selecting the events and time to events

Select the data to be used for the time-to-event analysis.

**Event / Censor Status:**

SELECT SPREADSHEET

SELECT ROW LABEL

EVENT VALUE

**Serial Time:**

SELECT SPREADSHEET

SELECT ROW LABEL

# STEP 2E: Compare in Spreadsheet Visualizer

- Select the parameters below and press **Done** to see Kaplan Meier curves of clusters identified using HumanNet network

cluster\_assignment Time-to-Event Analysis

Select the data to be used for the time-to-event analysis.

**Event / Censor Status:**

SELECT SPREADSHEET  
Demo2\_Clinical\_pancan12\_30.txt

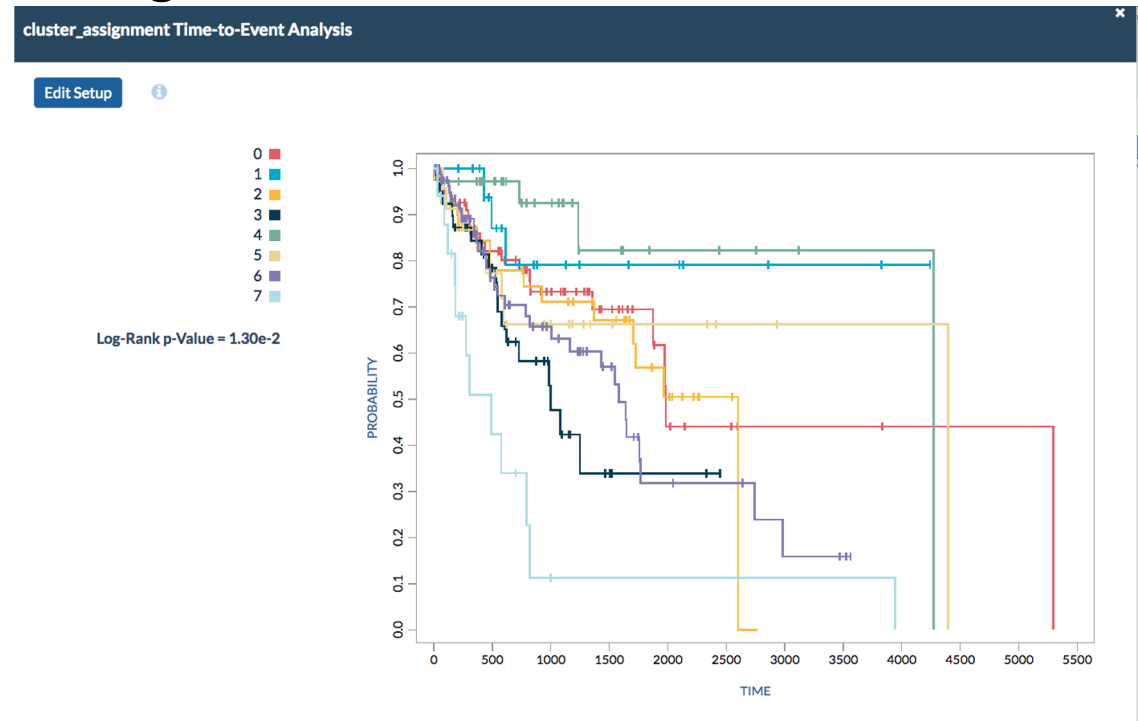
SELECT ROW LABEL  
\_OS\_IND

EVENT VALUE  
1

**Serial Time:**

SELECT SPREADSHEET  
Demo2\_Clinical\_pancan12\_30.txt

SELECT ROW LABEL  
\_OS



# Finding Genes Correlated with Drug Response

In this exercise, we will use cell line gene expression data and cytotoxicity experiments with knowledge-guided methods to find genes that may predict drug response.

# STEP 3: Feature (Gene) Prioritization

- We will use KnowEnG's gene prioritization pipeline to perform network-guided feature (gene) prioritization
- The network-guided gene prioritization implemented in KnowEnG is a method called **ProGENI**:

Genome Biol. 2017 Aug 11;18(1):153. doi: 10.1186/s13059-017-1282-3.

**Knowledge-guided gene prioritization reveals new insights into the mechanisms of chemoresistance.**

Emad A<sup>1</sup>, Cairns J<sup>2</sup>, Kalari KR<sup>3</sup>, Wang L<sup>4</sup>, Sinha S<sup>5</sup>.

- We will use samples from the Cancer Cell Line Encyclopedia (CCLE)

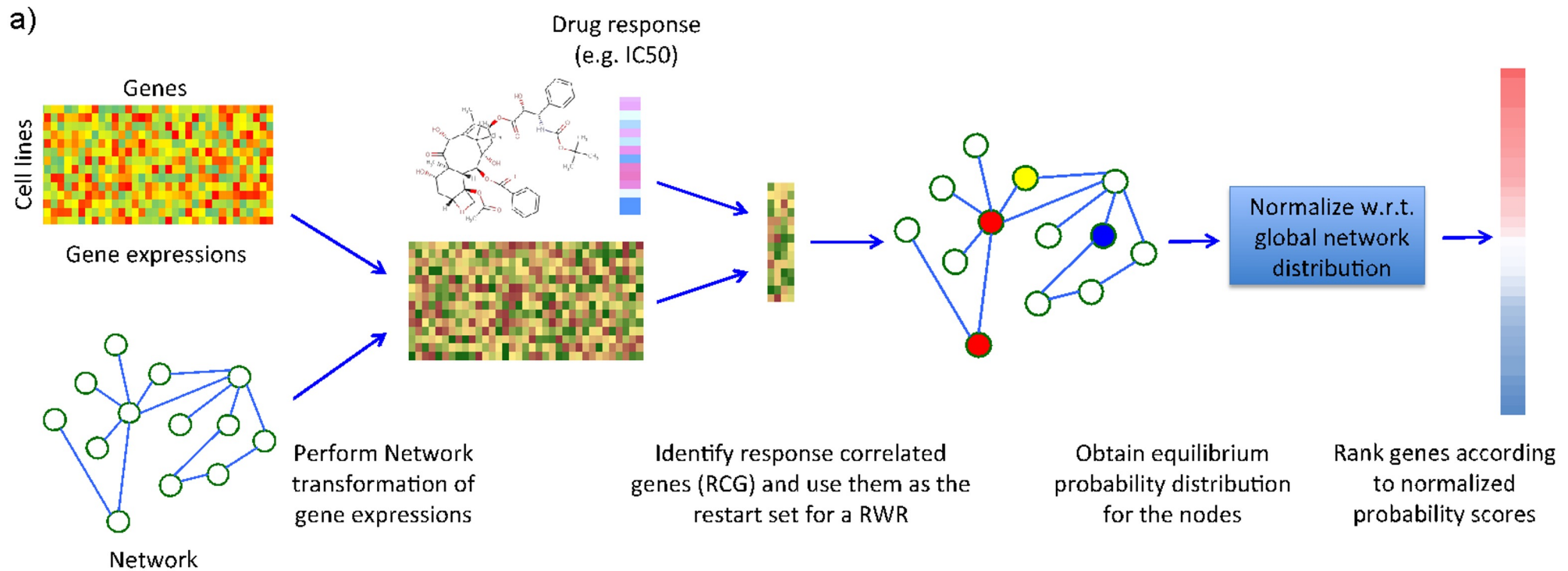
Nature. 2012 Mar 28;483(7391):603-7. doi: 10.1038/nature11003.

**The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.**

Barretina J<sup>1</sup>, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, Reddy A, Liu M, Murray L, Berger MF, Monahan JE, Morais P, Meltzer J, Korejwa A, Jané-Valbuena J, Mapa FA, Thibault J, Bric-Furlong E, Raman P, Shipway A, Engels IH, Cheng J, Yu GK, Yu J, Aspesi P Jr, de Silva M, Jagtap K, Jones MD, Wang L, Hatton C, Paescandolo E, Gupta S, Mahan S, Sougnéz C, Onofrio RC, Liefeld T, MacConaill L, Winckler W, Reich M, Li N, Mesirov JP, Gabriel SB, Getz G, Ardlie K, Chan V, Myer VE, Weber BL, Porter J, Warmuth M, Finan P, Harris JL, Meyerson M, Golub TR, Morrissey MP, Sellers WR, Schlegel R, Garraway LA.

# STEP 3: Gene Prioritization

## Outline of ProGENI:



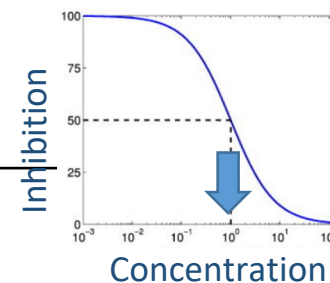


# STEP 3: Gene Prioritization

Find the files in this slide under [\[course\\_directory\]/08\\_Clustering\\_and\\_Prioritization](#)

Dataset characteristics:

Name	Description
demo_FP.genomic	A matrix of (gene x samples) containing the expression of ~17k genes in ~500 cell lines. The expression profiles are normalized in advance.
demo_FP.phenotypic	A matrix of (samples x drugs) containing IC50 values for 24 cytotoxic treatments.



# STEP 3A: Gene Prioritization (network-guided)

Select the pipeline:

- Select “**Analysis Pipelines**” at the top of the page
- Select “**Feature Prioritization**” and Click on “**Start Pipeline**”

The screenshot shows the Knoweng website navigation bar with the 'Analysis Pipelines' link circled in red. Below the navigation bar, a 'SELECT A PIPELINE' menu is displayed with the following options: Sample Clustering, Feature Prioritization (circled in red), Gene Set Characterization, Signature Analysis, and Spreadsheet Visualization. A 'Start Pipeline' button is visible next to the 'Feature Prioritization' option.

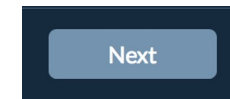
# STEP 3A: Gene Prioritization (network-guided)

## Configure the pipeline:

- For the “omics” file select “Use Demo Data”



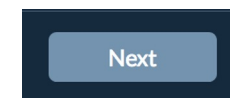
- Click “Next” at the bottom right corner



- For the “response” file select “Use Demo Data”



- Click “Next” at the bottom right corner



# STEP 3A: Gene Prioritization (network-guided)

- Select “**Yes**” in response to using the knowledge network:
  - *This allows us to perform network-guided prioritization (ProGENI)*
- Keep the species as “**Human**”
- Select “**STRING Experimental PPI**” as the network
  - *This network connects genes by the physical protein-protein interactions between their corresponding proteins*
- Keep network smoothing at **50%**:
  - *This controls how much importance is put on network connections instead of the correlation with drug response*
- Click “**Next**”

Do you want to use the Knowledge Network?

Yes  No

1 Select species

Human (Hsap) default

2 Select Interaction Network for analysis

STRING Experimental PPI default

3 Choose the amount of network smoothing

50 % default

# STEP 3A: Gene Prioritization (network-guided)

- Keep the default parameters on this page

Used for continuous-valued response

The screenshot shows a configuration panel with four numbered steps:

- 1 Select the primary prioritization method**  
Dropdown menu: Absolute Pearson Correlation (default)
- 2 Select the method to handle missing "omics" values**  
Dropdown menu: Average (default)
- 3 Number of response-correlated features**  
Input field: 100 (default)
- 4 Number of exported features per phenotype**  
Input field: 100 (default)

Red arrows point from the text 'Used for continuous-valued response' to the 'Absolute Pearson Correlation' dropdown, and from 'Size of RCG set' to the '100' input field in step 3.

- Choose “No” for bootstrapping


Do you want to use bootstrapping?

Yes

No

# STEP 3A: Gene Prioritization (network-guided)

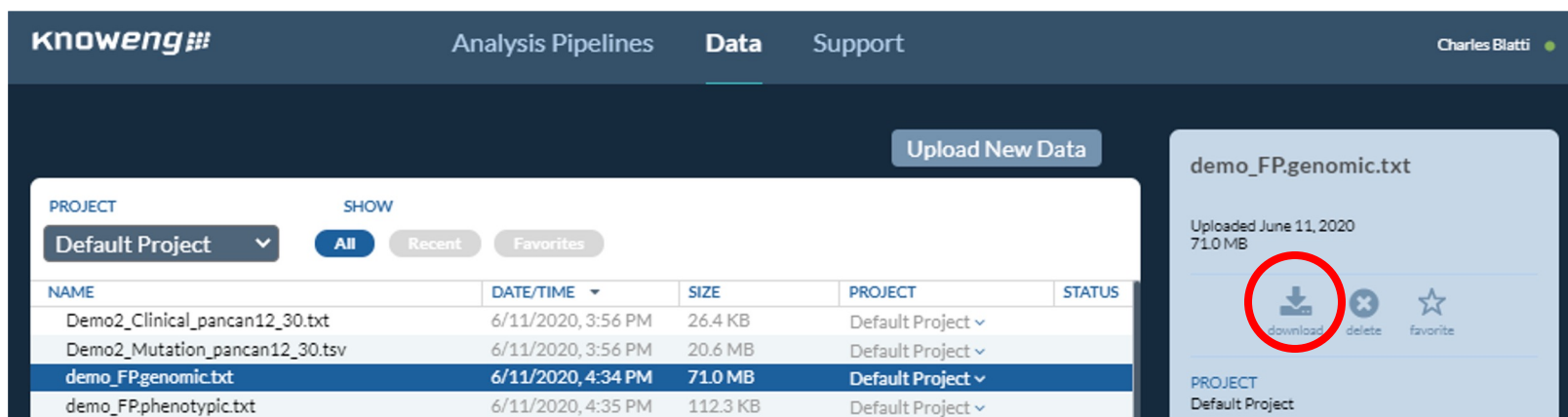
- Review the summary of the job and change its name if you like

Features File	demo_FP.genomic.txt	Job Name	<input type="text" value="feature_prioritization-PPI"/>
Response File	demo_FP.phenotypic.txt	Project	<input type="text" value="Default Project"/>
Use Network	Yes	Notes	
Species	Human (Hsap)		
Interaction Network	STRING Experimental PPI		
Network Influence	50%		
Method	Absolute Pearson Correlation		
Missing Values	Average		
# Response-Correlated Features	100		
# Exported Features	100		
Use Bootstrapping	No		

- **Submit** the job



# STEP 3A: Gene Prioritization (network-guided)

- **Note:** If you ever want to view your data or results outside of the KnowEnG system, just go to the Data page, click on the file or run, and select “download” on the far right panel.



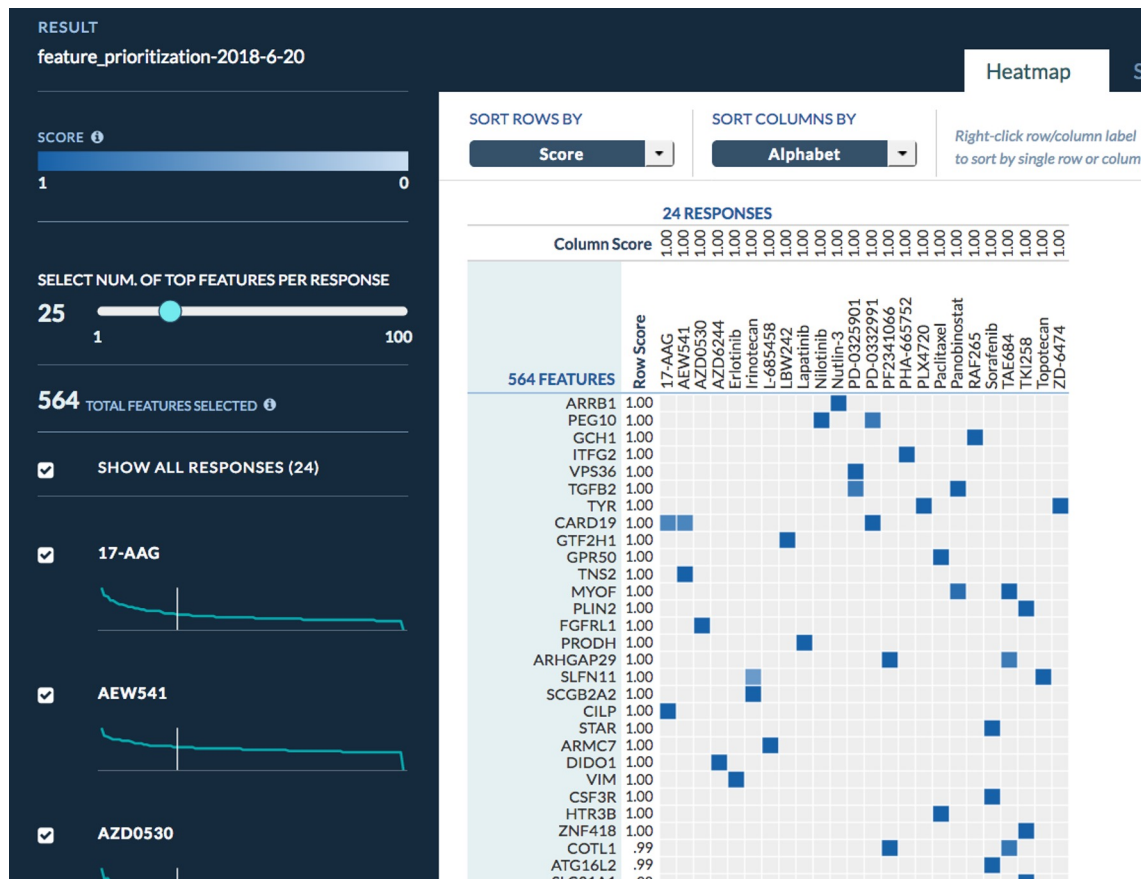
The screenshot shows the KnowEnG interface. At the top, there are navigation tabs for 'Analysis Pipelines', 'Data', and 'Support', with 'Data' selected. The user's name 'Charles Blatti' is in the top right. Below the navigation is an 'Upload New Data' button. The main content area is divided into two panels. The left panel shows a table of files under the 'Default Project'. The right panel shows a detailed view of the selected file 'demo\_FPgenomic.txt', including its upload date and size, and a set of action icons: 'download' (circled in red), 'delete', and 'favorite'.

NAME	DATE/TIME	SIZE	PROJECT	STATUS
Demo2_Clinical_pancan12_30.txt	6/11/2020, 3:56 PM	26.4 KB	Default Project	
Demo2_Mutation_pancan12_30.tsv	6/11/2020, 3:56 PM	20.6 MB	Default Project	
demo_FPgenomic.txt	6/11/2020, 4:34 PM	71.0 MB	Default Project	
demo_FPphenotypic.txt	6/11/2020, 4:35 PM	112.3 KB	Default Project	

- **Reminder:** The option to view results will not be available if
  - The job is still running: 
  - There was an error: 
    - If there's an error, repeat the launch steps and check your job summary matches exactly
- *This job takes about five minutes, so you are welcome to skip ahead and start Step 4 on slide 55 and come back later to finish Step 3*

# STEP 3A: Gene Prioritization (network-guided)

- Go to the **Data** page
- Select **“View Results”** when the job is done

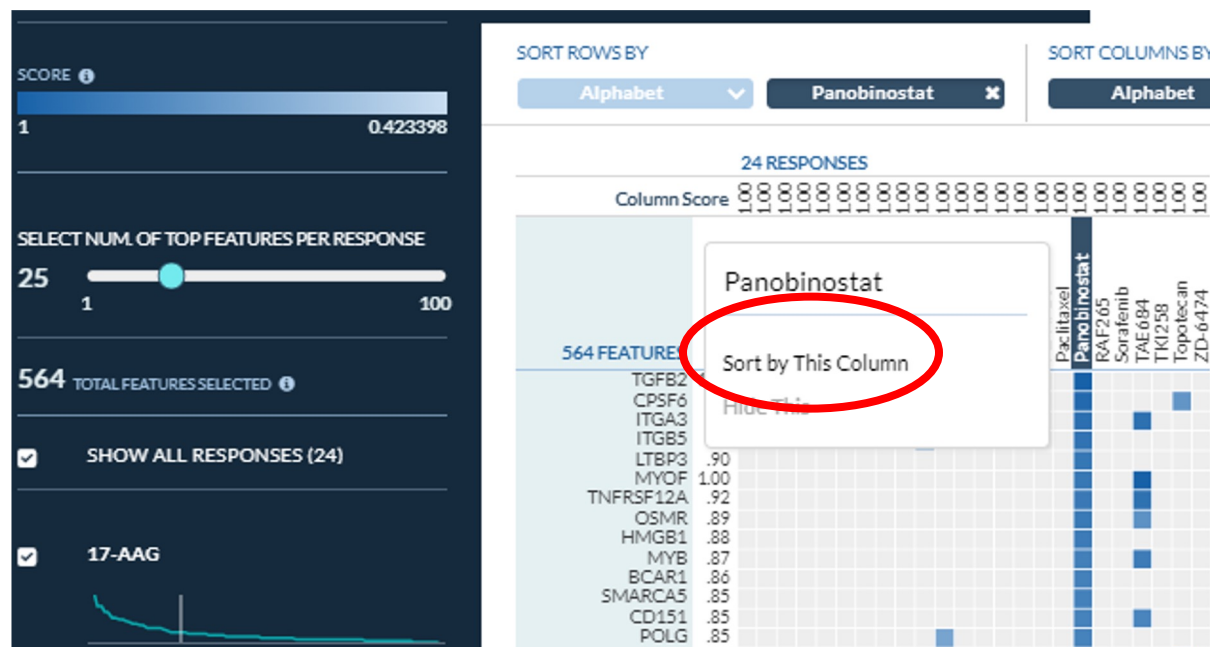


Heatmap shows the top gene rows identified for each drug column



# STEP 3A: Gene Prioritization (network-guided)

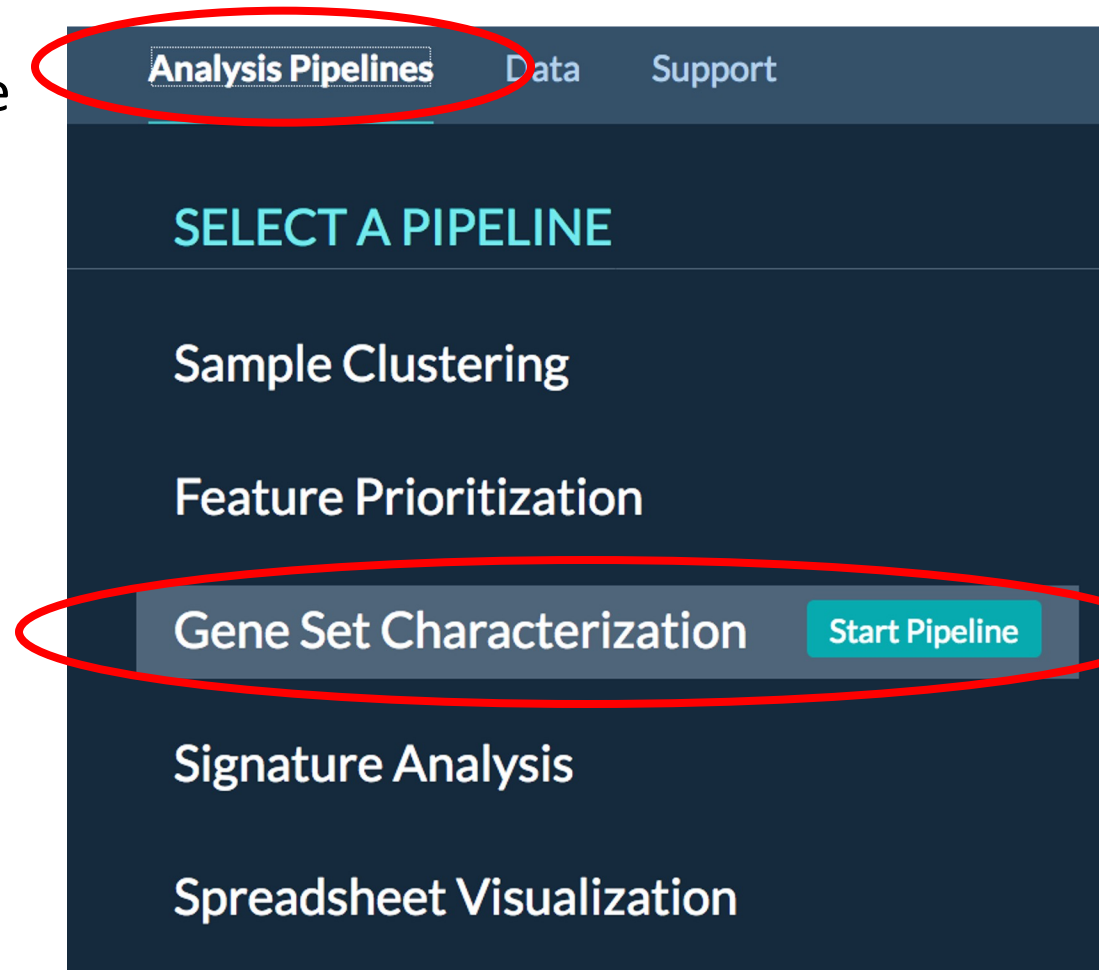
- You can “right-click” on a **drug column name** to sort rows and see its top genes



- You can also sort columns by a gene to see drugs for which the gene was among the top list
- Panobinostat (HDAC inhibitor) prevents chromatin formation which is tied to the transforming growth factor beta signaling pathway (TGFβ2 is top result).*

# STEP 3B: Gene Prioritization (network-guided)

- Let's see the enrichment of the top genes in different Gene Ontology (GO) terms
- Go to “**Analysis Pipelines**” page
- Select “**Gene Set Characterization**” pipeline



# STEP 3B: Gene Prioritization (network-guided)

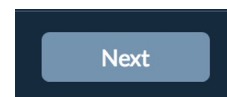
- Select the green triangle by the gene prioritization job you ran

Expression_METABRIC_Demo1.txt	6/16/2018, 5:59 PM	4.6 MB	Default Project ▾
Phenotype_METABRIC_Demo1.txt	6/16/2018, 5:59 PM	182.4 KB	Default Project ▾
▶ feature_prioritization-2018-6-...	6/20/2018, 9:48 AM		Default Project ▾
SSV_4_SC	6/17/2018, 3:32 PM		Default Project ▾
SC_HumanNet_clustR	6/17/2018, 1:27 PM		Default Project ▾

- Select **“top\_features\_per\_phenotype\_matrix”** which contains the ProGENI top gene lists for each of the 24 drugs

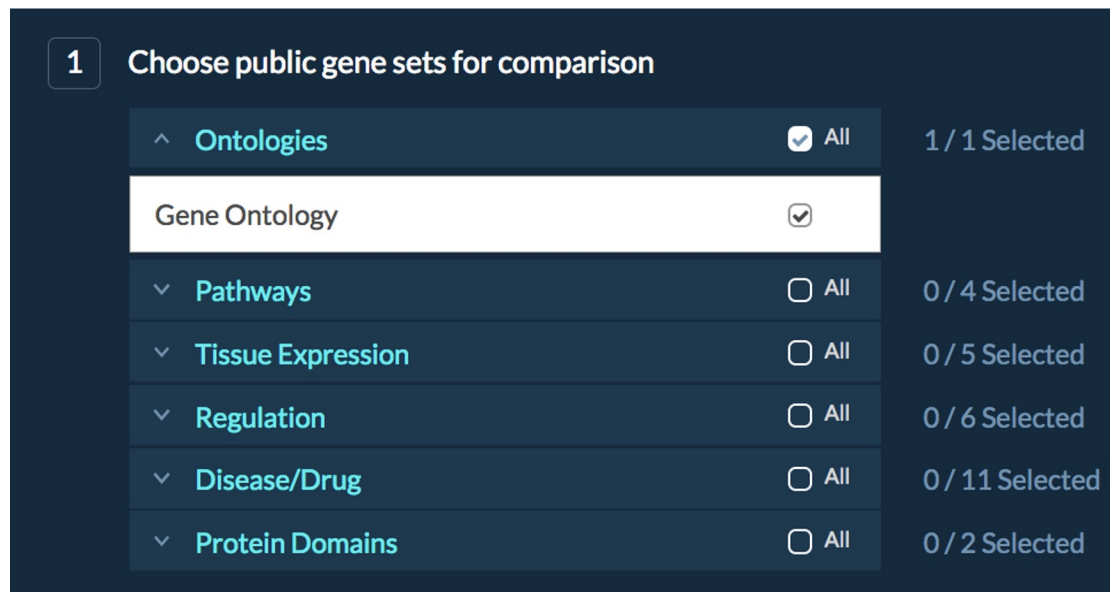
Phenotype_METABRIC_Demo1.txt	6/16/2018, 5:59 PM	182.4 KB	Default Project ▾
▼ feature_prioritization-2018-6-...	6/20/2018, 9:48 AM		Default Project ▾
features_ranked_per_phenot...	6/20/2018, 9:50 AM	16.5 MB	Default Project ▾
✓ top_features_per_phenotype...	6/20/2018, 9:50 AM	899.9 KB	Default Project ▾

- Press **Next**



# STEP 3B: Gene Prioritization (network-guided)

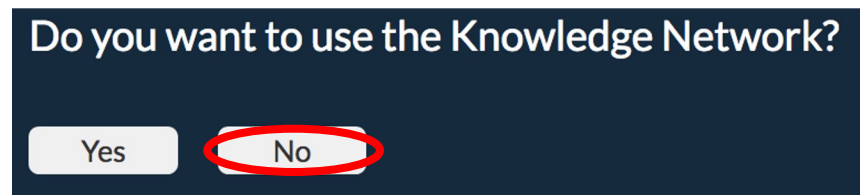
- For gene sets, select your gene sets of interest (e.g. GO) and press **Next**



1 Choose public gene sets for comparison

^ Ontologies	<input checked="" type="checkbox"/> All	1 / 1 Selected
Gene Ontology	<input checked="" type="checkbox"/>	
∨ Pathways	<input type="checkbox"/> All	0 / 4 Selected
∨ Tissue Expression	<input type="checkbox"/> All	0 / 5 Selected
∨ Regulation	<input type="checkbox"/> All	0 / 6 Selected
∨ Disease/Drug	<input type="checkbox"/> All	0 / 11 Selected
∨ Protein Domains	<input type="checkbox"/> All	0 / 2 Selected

- Say “**No**” to using the knowledge network (we will do that later) and press **Next**. Then press **Submit Job**.
  - *This job should take about one minute.*



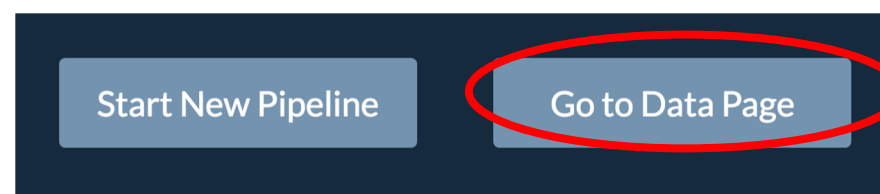
Do you want to use the Knowledge Network?

Yes No

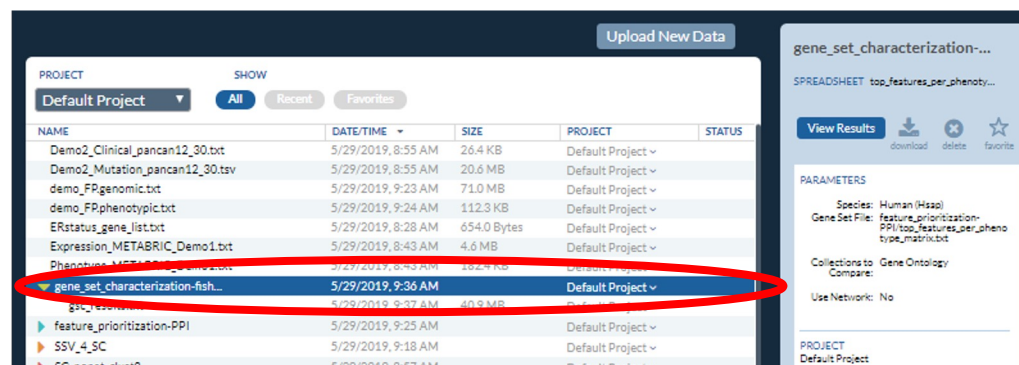
# STEP 3B: Gene Prioritization (network-guided)

The Gene Ontology enrichment results:

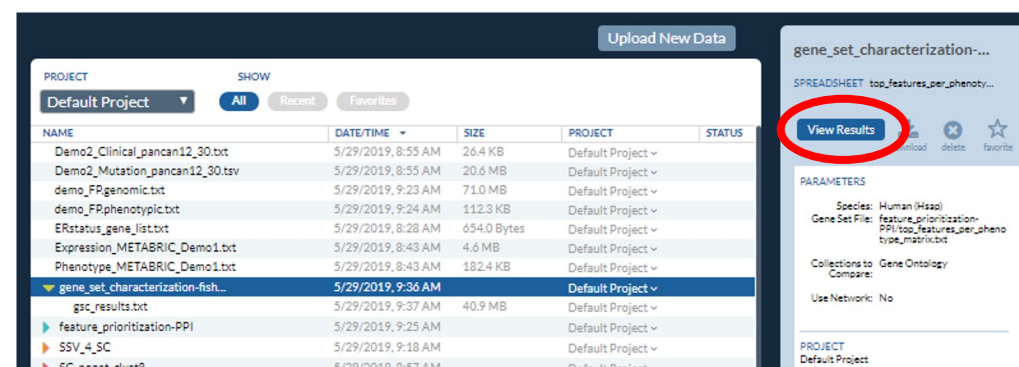
- Select “Go to Data Page”



- Select the job you just ran

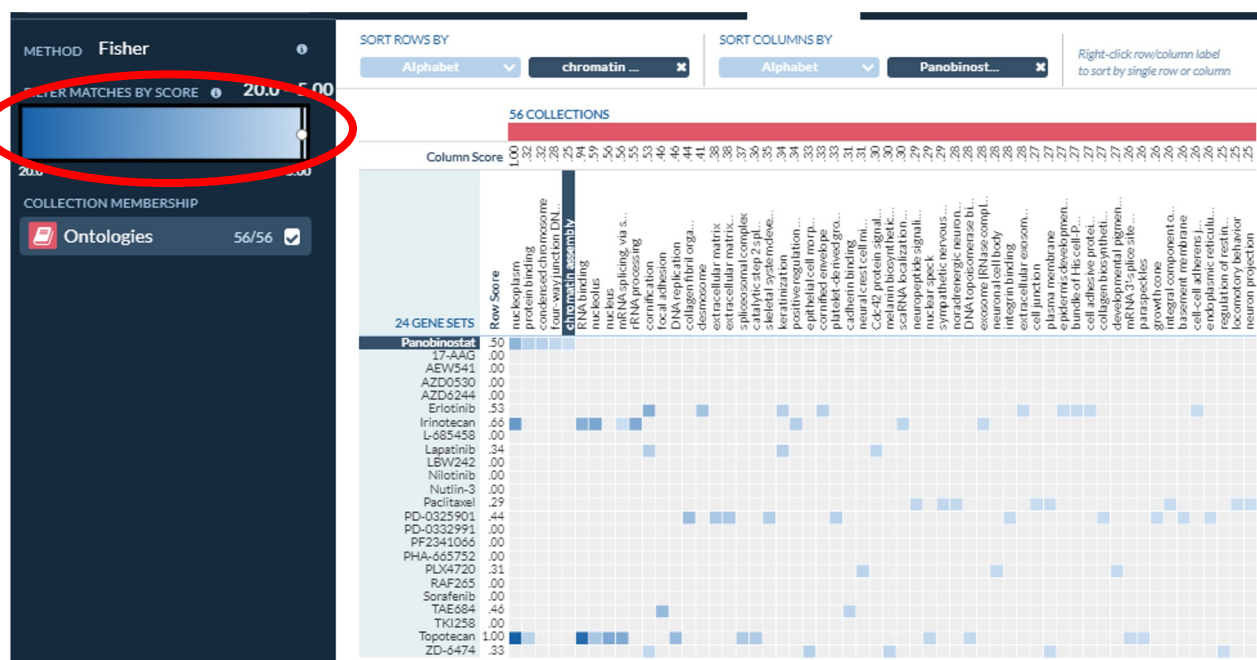


- Then “View Results”



# STEP 3B: Gene Prioritization (network-guided)

- This page shows the enriched GO gene sets for each drug to gene list
- You can change the filter (scores represent  $-\log_{10}$  (p-value) of enrichment) to see fewer or **more** enriched gene sets



- *The network-guided genes whose expression correlated with the response to Panobinostat are enriched with terms related to chromatin assembly*

# Creating a Novel Gene Expression Signature

In this exercise, we will use the integrative iLINCS data portal to extract gene expression data from TCGA Breast Invasive Carcinoma (BRCA) samples and build a gene signature based on the estrogen receptor status.

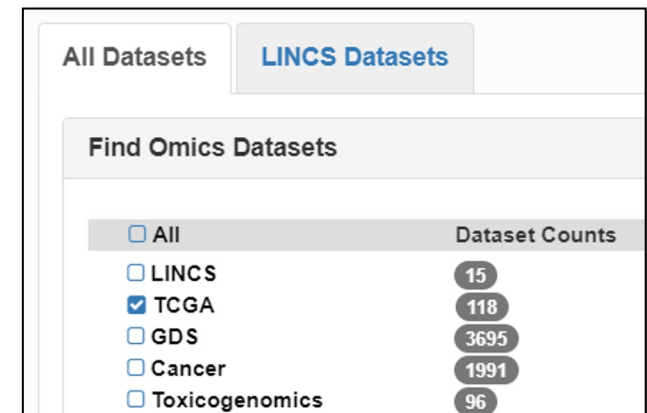
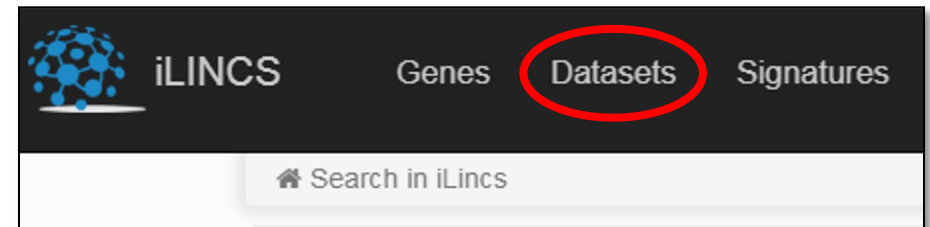
# Step 4: Perturbagen and Disease Datasets

- Open your web browser and go to the iLINCS data portal:  
<http://www.ilincs.org/ilincs/>
- This portal, curated by the LINCS Data Coordination and Integration Center, contains transcriptomic and proteomic datasets from the many LINCS affiliated projects, including the LINCS L1000 assay. It also contains several other large public datasets of perturbations to cell lines and samples of disease.
- We will define a custom gene signature from TCGA data and see how it can be used in various network related analyses.



# Step 4A: Select Breast Cancer Dataset

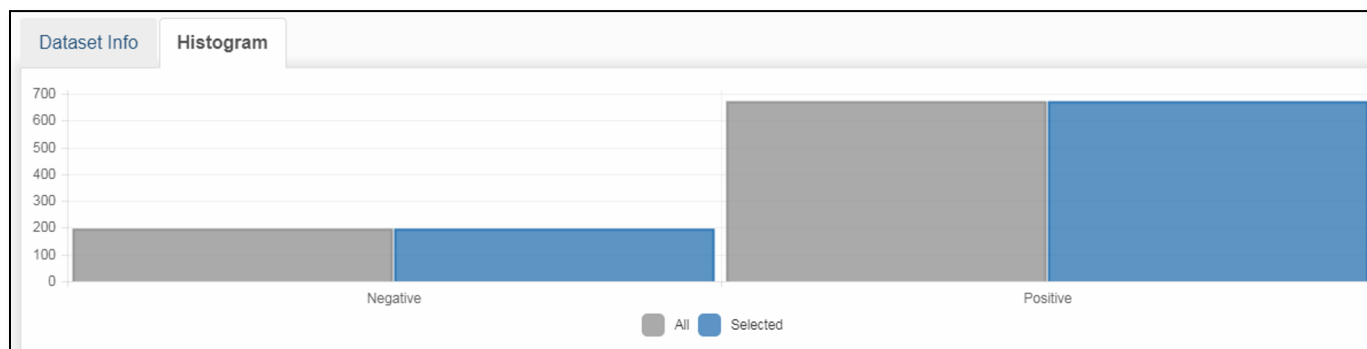
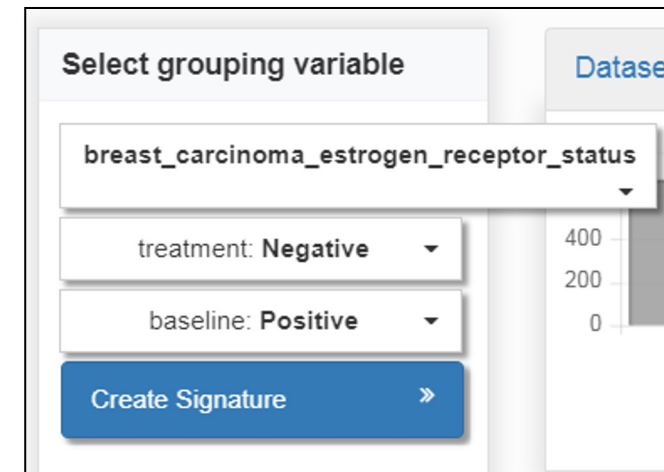
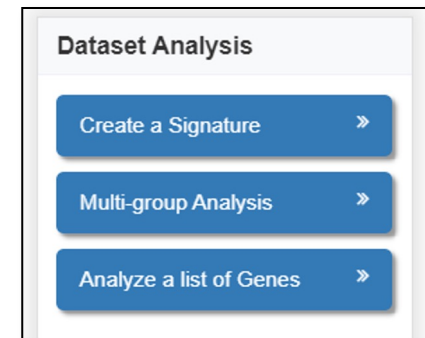
- Click on **“Datasets”** in the header
- In the **“All Datasets”** tab, for Choose Dataset, select only the **TCGA** datasets
- Scroll down and enter **“919”** on the Description box to find **“919 mRNA-seq breast invasive carcinoma (BRCA) samples from TCGA project”** by Collins, et al. Click **“Analyze”**.



Organism	Collection	Data Type	Sample Type	Sample Count	Description	Reference	
All					919		
human	TCGA	Gene Expression	tissue	919	919 RNA-seq breast invasive carcinoma (BRCA) samples from TCGA project. The data was processed using... <a href="#">More</a>	Collins FS, Barker AD. Mapping... <a href="#">More</a> ID=TCGA_BRCA_RNA SeqV2	<a href="#">Analyze</a>

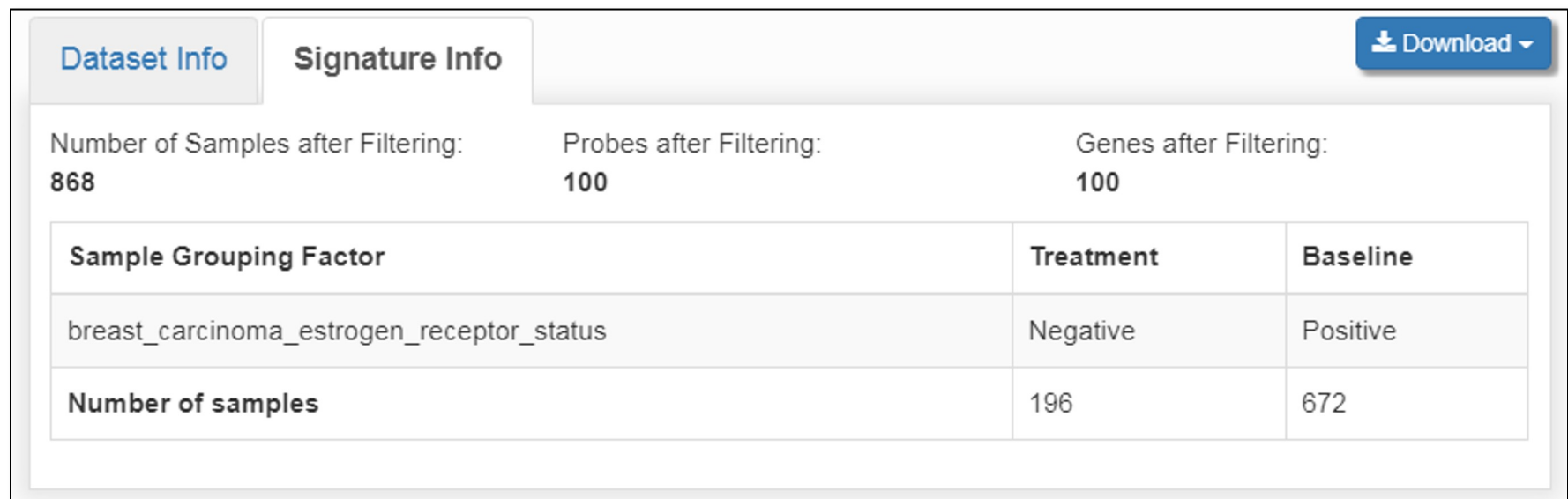
# Step 4B: Creating a Novel Gene Signature

- Click on “**Create a Signature**”
- In “**Select grouping variable**” dropdown select “**breast\_carcinoma\_estrogen\_receptor\_status**”
- In “**Select treatment group**” dropdown select “**Negative**”
- In “**Select baseline group**” dropdown select “**Positive**”
- Finally, click on “**Create Signature**” button



# Step 4B: Our ER Status Gene Signature

- When the signature is calculated, a quick summary of the number of samples from each group is presented in the “**Signature Info**” tab



The screenshot displays the 'Signature Info' tab with a 'Download' button in the top right. Summary statistics are shown as follows:

Number of Samples after Filtering:	Probes after Filtering:	Genes after Filtering:
868	100	100

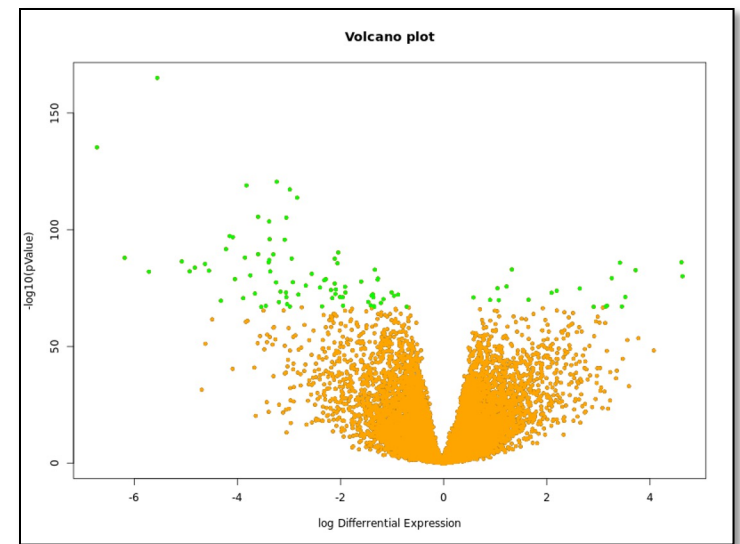
  

Sample Grouping Factor	Treatment	Baseline
breast_carcinoma_estrogen_receptor_status	Negative	Positive
<b>Number of samples</b>	196	672

Next, we will look more closely at the genes involved in our signature.

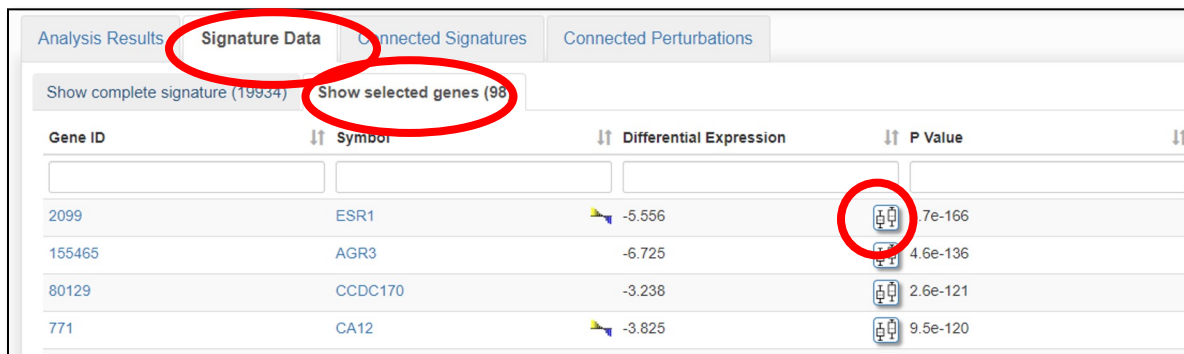
# Step 4C: Examining Gene Expression of our Signature

- To get statistics about how the signature is defined, we will select “**Modify the list of selected genes**” on the left
- We are presented with a volcano plot for the log fold change (x-axis) and differential expression significance (y-axis) of each gene.
- The green genes are the ones included in the gene expression signature.
- They have a log Differential Expression of at least +/- 0.6 and significance p-value less than  $10^{-60}$ .

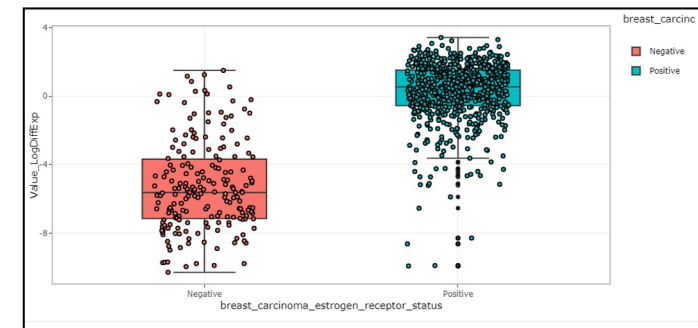


# Step 4C: Examining Gene Expression of our Signature

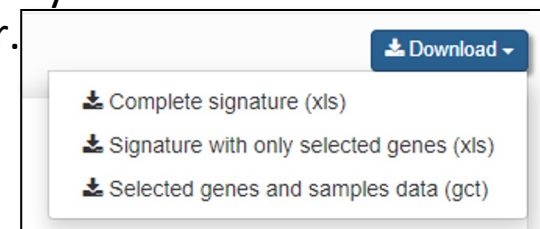
- Click the “**Signature Data**” tab and “**Show selected genes**” to see the list of selected signature genes



Gene ID	Symbol	Differential Expression	P Value
2099	ESR1	-5.556	7e-166
155465	AGR3	-6.725	4.6e-136
80129	CCDC170	-3.238	2.6e-121
771	CA12	-3.825	9.5e-120



- Note that ESR1, estrogen receptor 1, is the most significantly differentially expressed gene, which is consistent with the immunohistochemical staining assay result that defined the positive and negative groups. Click on the **plot icon** on the right to see ESR1’s measured gene expression values.
- Because of the number of samples (868) is high, the differential expression p-values are very significant for these top signature genes
- You can click the “**Download**” button on the top right of the website and save “**Signature with only selected genes**” table as an Excel file if you want to save the details of the 100 selected genes. We will use part of this file later.

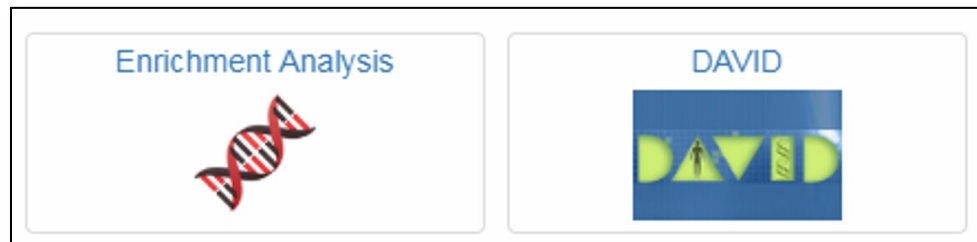


# Discovering Pathways Related to Our Gene Signature

In this section, we will consider some of the characterization resources that are available for gene signatures and gene sets.

# Step 5: Standard Gene Set Enrichment

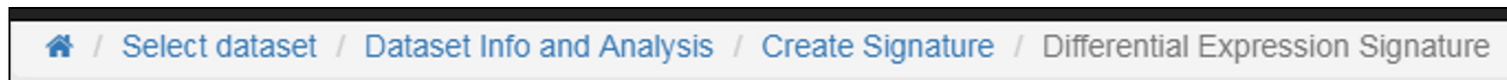
- Back in iLINCS, the “**Analysis Results**” tab which contains many different methods for analyzing our novel ER status gene signature.
- Two of the tools listed are links to **Enrichr** and **DAVID**.



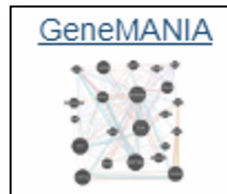
- Both tools use standard statistical enrichment tests to examine the overlap of the 100 genes of our ER status gene signature with Gene Ontology term annotations, pathways, and other gene sets.
- These tools output the results in slightly different ways, so you may want to explore them in your own time.

# Step 5B: GeneMANIA

- Return to the analysis result by clicking on “**Differential Expression Signature**” in the tool bar at the top



- The last linked tool we will explore today from iLINCS is GeneMANIA.

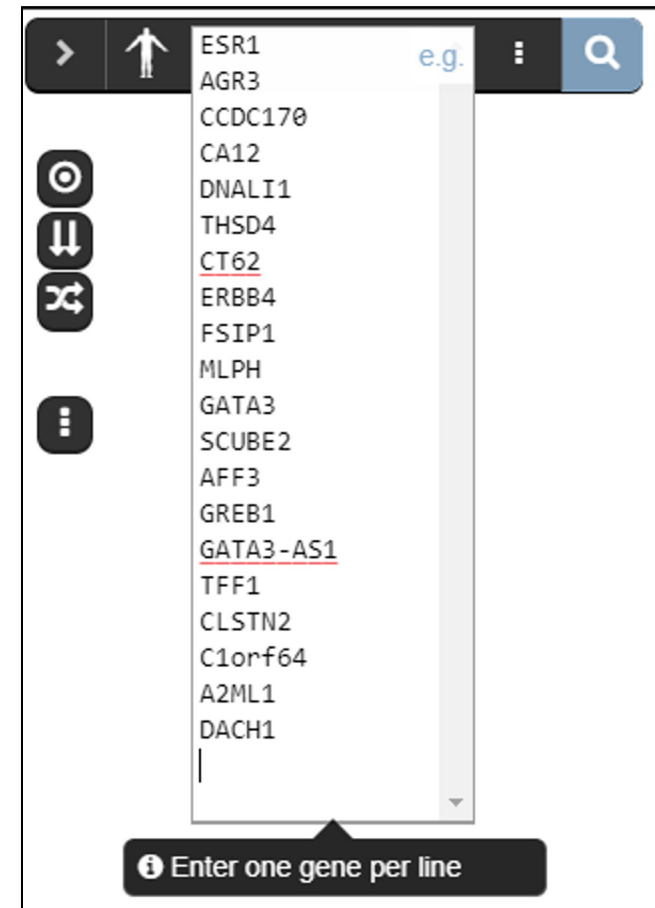


- GeneMANIA is a network-based guilt-by-association algorithm that finds the network neighbors of an input gene set from a heterogeneous collection of interaction networks
- Go to <https://genemania.org/>





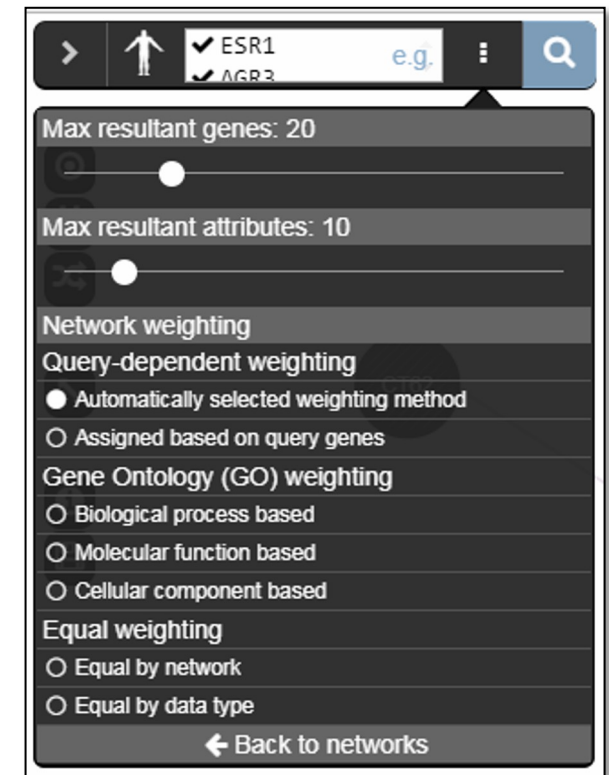
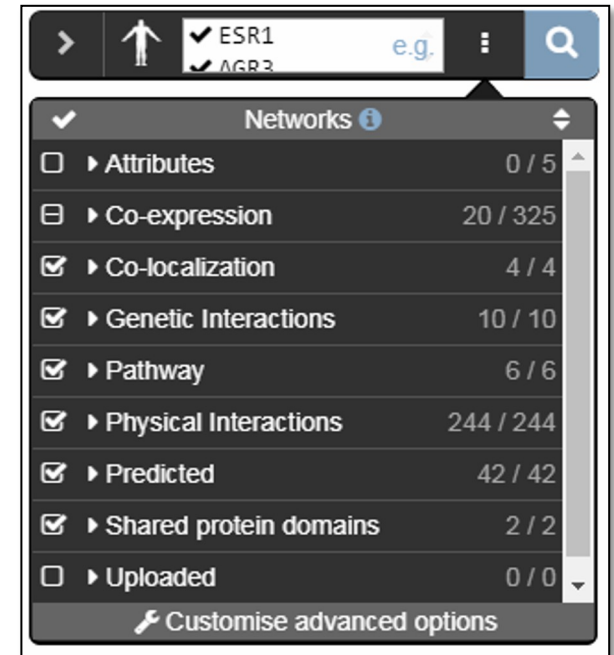
# Step 5B: GeneMANIA

- We are going to enter the top 20 differentially expressed genes from our ER status gene signature. We will use GeneMANIA to return 20 additional network neighbor genes (not necessarily differentially expressed themselves)
- Then we will look at functional enrichment of this combined set of 40 genes.
- Find, open in a editor, and copy the contents of the file: [course\_directory]/07\_Signatures\_and\_Characterization/ERstatus\_top20.txt
- This is the top 20 differentially expressed genes of our ER status signature extracted from the “Name\_GeneSymbol” column in the Excel download
- Paste this list into the text box at the top left corner of the main page. Make sure to delete any previous list in the text box.



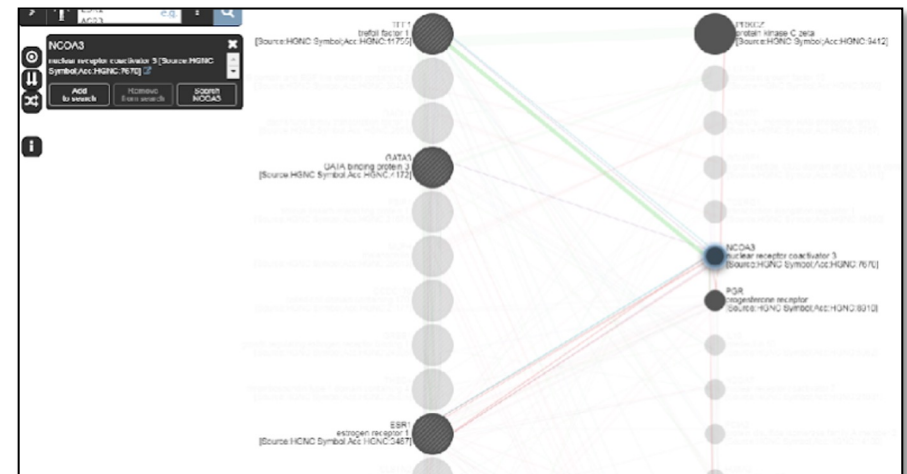
# Step 5B: GeneMANIA

- Click on the **stacked-dots** options button next to where you pasted the list 
- This first list shows all the possible networks that GeneMANIA will consider combining for the analysis of our twenty genes
- Select **“Customise advanced options”**
- This menu shows that we are going to find at most 20 neighbors using the automatic network weighting scheme, which is based on our 20 query genes
- Click the **search** magnifying glass. 



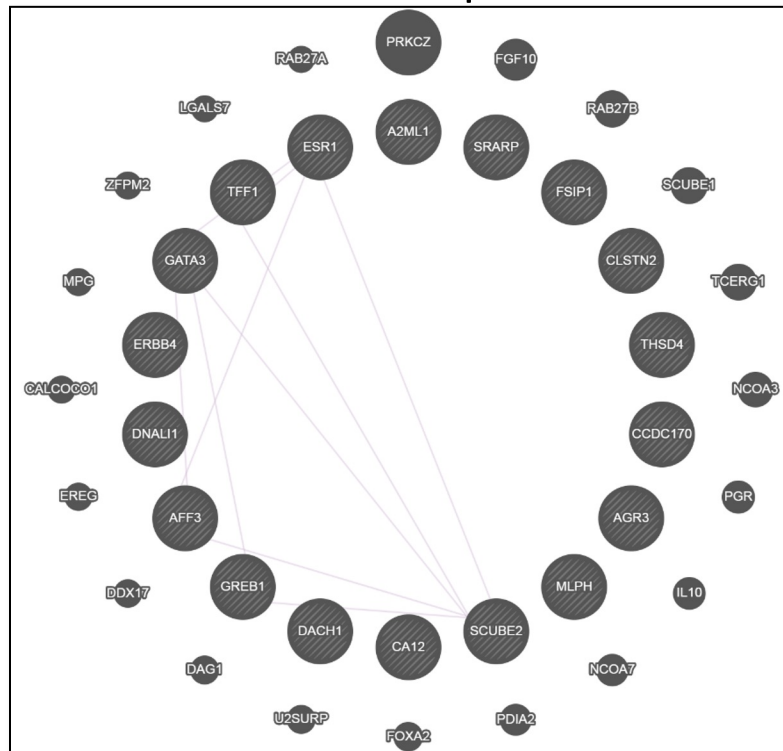
# Step 5B: GeneMANIA

- The resulting network contains our 20 input genes (striped) and our 20 predicted network neighbors (solids). The size of the network neighbors indicates its final guilt-by-association value on the composite affinity network.
- You may choose between three arrangements of the graph. The **stacked** arrangement may be easiest for understanding the nodes. You can hover over any node to highlight its neighbors.
- For example,
  - NCOA7 is also known as Estrogen Nuclear Receptor Coactivator 1
  - NCOA3 is associated with Estrogen-Receptor Positive Breast Cancer
- Both are connected to ESR1 (and other top 20 genes) through pathways edges and neither are in our original 100 differentially expressed gene signature



# Step 5B: GeneMANIA

- On the right side (hidden unless you click on the button with 3 horizontal bars) is the selected interaction networks that were relevant to the 20 input genes, sorted by type and by weight. You can toggle the networks to display any set of edges.
- The highest weighted co-expression network is from breast tumors and relates the top 20 genes to each other fairly well, but does not connect them to the predicted 20.




Expand: all top none ✕

- Genetic Interactions** 37.80%
- BIOGRID-SMALL-SCALE-STUDIES 30.85%
- IREF-SMALL-SCALE-STUDIES 6.40%
- Lin-Smith-2010 0.55%
- Co-expression** 31.17%
- Perou-Botstein-2000** 8.09%
- Molecular portraits of human breast tumours. Perou et al (2000). [Nature](#)
- Co-expression with 189,373 interactions from supplementary material
- Bild-Nevins-2006 B 7.45%
- Chen-Brown-2002 3.30%
- Ross-Perou-2001 3.11%
- Dobbin-Giordano-2005 2.27%
- Wang-Maris-2006 2.21%
- Roth-Zlotnik-2006 1.83%
- Mallon-McKay-2013 1.06%
- Ramaswamy-Golub-2001 1.02%
- Wu-Garvey-2007 0.84%
- Physical Interactions** 18.05%
- Pathway** 8.59%
- Co-localization** 4.38%

# Step 5B: GeneMANIA

- Finally, we can perform the standard enrichment tests incorporating our predicted neighbors into our gene set.
- Click on the pie chart in the bottom left corner
- We see most of the results relating to hormone and steroid signaling pathways and receptors.

Function	FDR	Coverage
<input type="checkbox"/> response to steroid hormone	1.07e-2	6 / 159
<input type="checkbox"/> regulation of protein kinase B signaling	1.07e-2	6 / 191
<input type="checkbox"/> steroid hormone mediated signaling pathway	1.07e-2	5 / 102
<input type="checkbox"/> protein kinase B signaling	1.08e-2	6 / 201
<input type="checkbox"/> hormone-mediated signaling pathway	1.25e-2	5 / 119
<input type="checkbox"/> cellular response to steroid hormone stimulus	1.27e-2	5 / 124
<input type="checkbox"/> establishment of melanosome localization	1.45e-2	3 / 19
<input type="checkbox"/> establishment of pigment granule localization	1.45e-2	3 / 20
<input type="checkbox"/> melanosome localization	1.45e-2	3 / 20
<input type="checkbox"/> pigment granule transport	1.45e-2	3 / 19
<input type="checkbox"/> pigment granule localization	1.78e-2	3 / 22
<input type="checkbox"/> mesenchymal cell differentiation	4.43e-2	5 / 185
<input type="checkbox"/> muscle cell proliferation	6.91e-2	4 / 107
<input type="checkbox"/> growth factor receptor binding	6.91e-2	4 / 106
<input type="checkbox"/> intracellular receptor signaling pathway	6.91e-2	5 / 210
<input type="checkbox"/> mesenchyme development	6.96e-2	5 / 216
<input type="checkbox"/> RNA polymerase II-specific DNA-binding transcription factor binding	8.33e-2	5 / 230
<input type="checkbox"/> mammary gland development	8.33e-2	3 / 42
<input type="checkbox"/> cellular pigmentation	9.84e-2	3 / 46





# Attention!

- For this last section (Slides 71-84), please only continue on if you:
    1. Have **at least 20 min** before we end for the day
  - AND-
  - 2. Feel confident working with KnowEnG on your own
- **If not, please stop here.**

# Gene Set Characterization Using Discriminative Random Walks

In this final exercise, we will find terms related to the 100 top differentially expressed genes of our ER status signature using the DRaWR method that incorporates the functional annotation terms directly in the network-based algorithm.

# Step 6: Login Into KnowEnG Platform

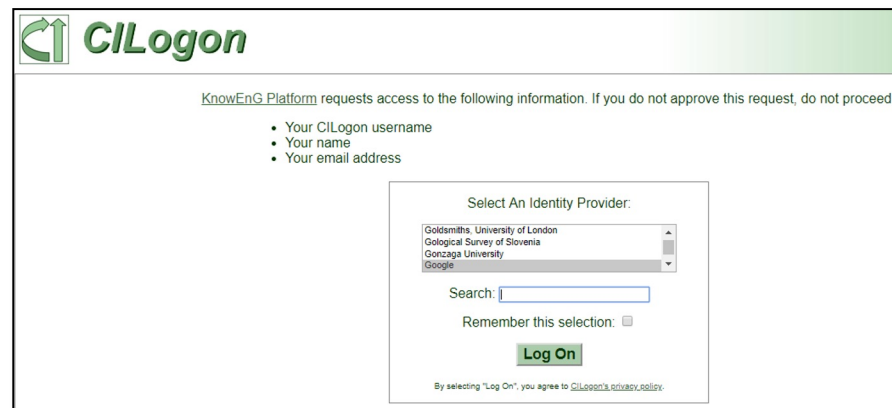
Return to KnowEnG Platform:

<https://platform.knoweng.org/static/#/home>

If necessary,

Login with **CILogon** - Login service through other accounts

Search: **Urbana, Mayo, Google, GitHub**



**CILogon**

KnowEnG Platform requests access to the following information. If you do not approve this request, do not proceed.

- Your CILogon username
- Your name
- Your email address

Select An Identity Provider:

Goldsmiths University of London  
Geological Survey of Slovenia  
Gonzaga University  
Google

Search:

Remember this selection:

**Log On**

By selecting "Log On", you agree to CILogon's privacy policy.



# STEP 6A: Gene Set Characterization

Select the pipeline:

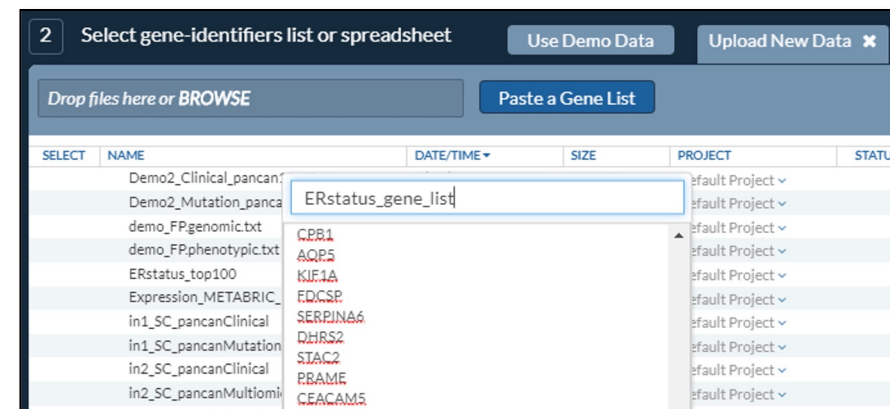
- Select “**Analysis Pipelines**” at the top of the page
- Select “**Gene Set Characterization**” and Click on “**Start Pipeline**”

The screenshot shows the top navigation bar of the Knoweng website. The 'knoweng' logo is on the left, and the navigation menu on the right includes 'Analysis Pipelines', 'Data', and 'Support'. The 'Analysis Pipelines' link is circled in red. Below the navigation bar is a dark blue panel titled 'SELECT A PIPELINE'. It lists several pipeline options: 'Sample Clustering', 'Feature Prioritization', 'Gene Set Characterization', 'Signature Analysis', 'Spreadsheet Visualization', and 'Network Preparation'. The 'Gene Set Characterization' option is highlighted with a light blue background and has a 'Start Pipeline' button next to it. This entire option is circled in red.

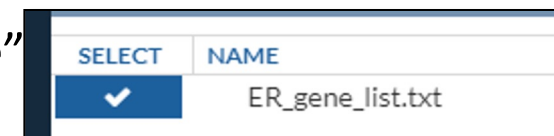
# Step 6A: Upload Data

Find the file in this slide under [course\_directory]/07\_Signatures\_and\_Characterization/

- Leave the default species “**Human**”
- Find, open in a text editor, and copy the contents of the file [course\_directory]/07\_Signatures\_and\_Characterization/**ERstatus\_top100.txt**
- This is the top 100 differentially expressed genes of our ER status gene signature extracted from the Name\_GeneSymbol column of our earlier Excel download



- Click on the “**Upload New Data**” tab
- Select the “**Paste a Gene List**” button.
- Give your gene list a name, e.g. “**ERstatus\_gene\_list**”
- Paste the file contents into the gene list text box. Click “**Done**”



- Click “**Select**” next to the name of your pasted list and you should see a checkmark
- Click “**Next**”

# Step 6A: Configure Algorithm Parameters

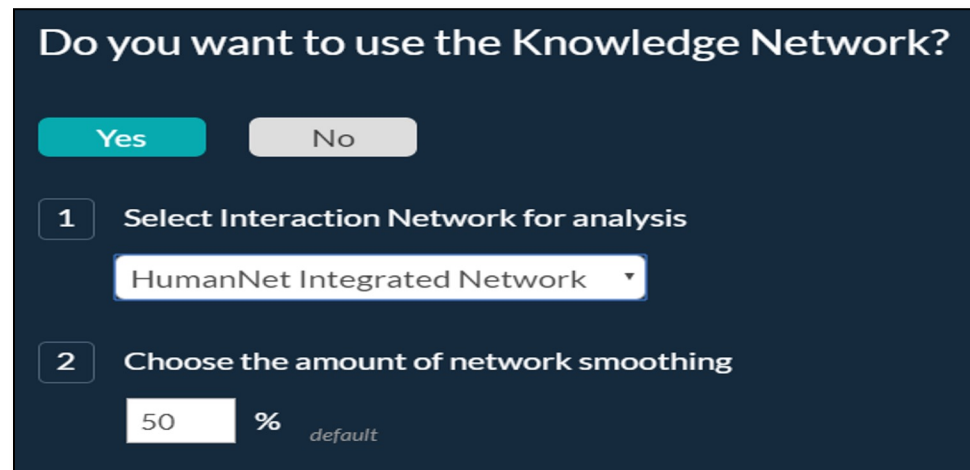
- We will choose to use a subset of 4 gene set collections available in the knowledge network
- **Ontologies:** Gene Ontology (default)
- **Pathways:** Enrichr Pathway Membership (**must add**)
- **Pathways:** Reactome Pathways Curated (**must add**)
- **Tissue Expression:** GEO Expression Set (**must add**)
- (**unclick Protein Domains:** Pfam Protein Domains)

- Click **“Next”**

Category	Selection	Count
Ontologies	<input checked="" type="checkbox"/> All	1 / 1 Selected
Pathways	<input checked="" type="checkbox"/> All	2 / 4 Selected
Enrichr Pathway Membership	<input checked="" type="checkbox"/>	
Pathway Commons Pathways	<input type="checkbox"/>	
PPI Complex	<input type="checkbox"/>	
Reactome Pathways Curated	<input checked="" type="checkbox"/>	
Tissue Expression	<input checked="" type="checkbox"/> All	1 / 5 Selected
Regulation	<input type="checkbox"/> All	0 / 6 Selected
Disease/Drug	<input type="checkbox"/> All	0 / 11 Selected
Protein Domains	<input type="checkbox"/> All	0 / 2 Selected

# Step 6A: Configure Network Parameters

- Click “**Yes**” for question about using the Knowledge Network
- The Knowledge Network we will use is an integrated network from the [HumanNet](#) project (“**HumanNet Integrated Network**”)
- Network size information can be found [here](#)
- The amount of network smoothing controls how much importance is put on network connections instead of the original 100 genes. We will use the default of **50%**
- Click “**Next**”

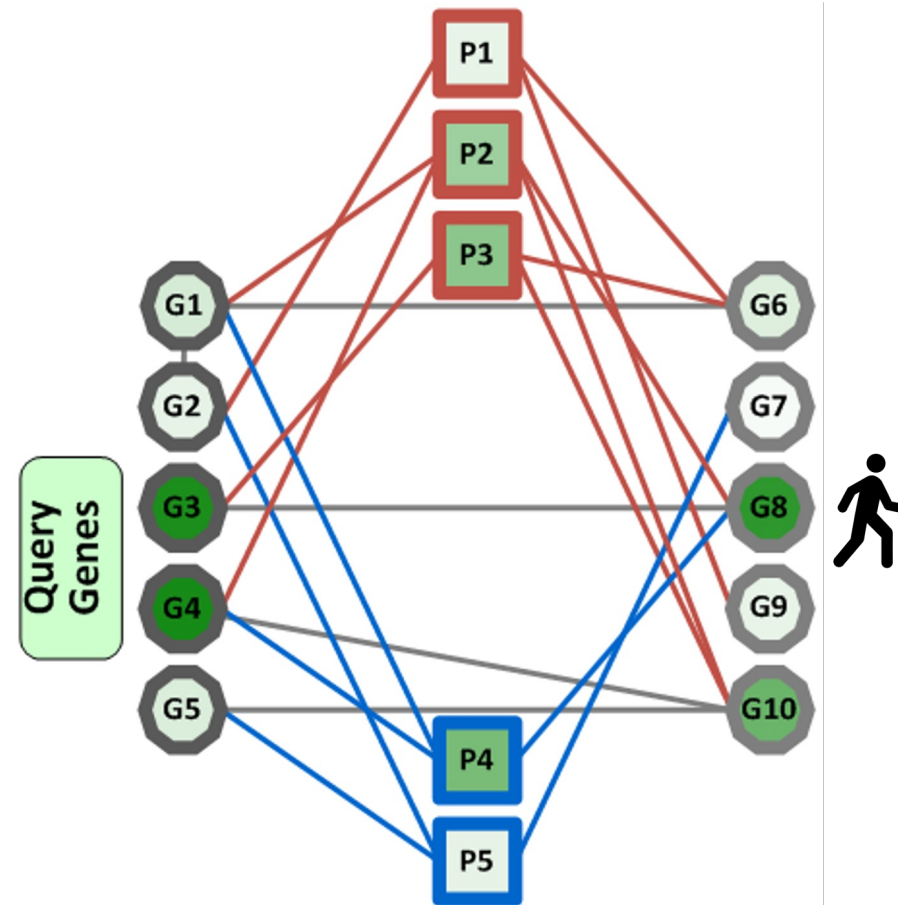


The screenshot shows a dark-themed dialog box with the title "Do you want to use the Knowledge Network?". At the top, there are two buttons: "Yes" (highlighted in teal) and "No" (grey). Below the buttons, there are two numbered steps:

- 1 Select Interaction Network for analysis  
A dropdown menu is open, showing "HumanNet Integrated Network" as the selected option.
- 2 Choose the amount of network smoothing  
A text input field contains the number "50", followed by a percentage sign "%". Below the input field, the word "default" is written in a smaller font.

# Step 6A: Reminder about DRaWR Algorithm

- Squares are the Gene Ontology and pathway terms we selected
- Query Genes are our 100 ER status signature genes
- Gray edges are the HumanNet Integrated Network
- We are asking the algorithm to find property squares that a random walker who is forced to restart often at the query genes will visit unusually frequently



# Step 6A: Launch DRaWR Job

- Change **job name** to “gene\_set\_characterization-DRaWR-HN”
- Verify all the parameters are correct.
- Click “**Submit Job**”
- While this is running (roughly two minutes), we are going to launch the standard fisher exact enrichment tests with the same data sets.
- Click “**Start New Pipeline**”

Species	Human (Hsap)
Gene Set File	ERstatus_gene_list
Collections to Compare	Ontologies: 1 Pathways: 2 Tissue Expression: 1 Regulation: 0 Disease/Drug: 0 Protein Domains: 0
Use Network	Yes
Interaction Network	HumanNet Integrated Network
Network Smoothing	50%

Start New Pipeline

# Step 6B: Launch Standard Enrichment Tests

- Hover over Gene Set Characterization under Analysis Pipelines and click “**Start Pipeline**”
- Click “**Select**” next to the name of your pasted list and you should see a checkmark. Click “**Next**”
- Select same 4 collections:
  - Ontologies: Gene Ontology (default)
  - Pathways: Enrichr Pathway Membership (**must add**)
  - Pathways: Reactome Pathways Curated (**must add**)
  - Tissue Expression: GEO Expression Set (**must add**)
  - (**unclick** Protein Domains: PFam Protein Domains)
- Click “**Next**”
- Click “**No**” for question about using the Knowledge Network. Click “**Next**”
- Change **job name** to “gene\_set\_characterization-fisher”
- Verify all the parameters are correct.
- Click “**Submit Job**”

Species	Human (Hsap)
Gene Set File	ERstatus_gene_list
Collections to Compare	Ontologies: 1 Pathways: 2 Tissue Expression: 1 Regulation: 0 Disease/Drug: 0 Protein Domains: 0
Use Network	No

# Step 6C: View DRaWR Results

- Click the “Go to Data Page” button

PROJECT		SHOW		
Default Project ▾		All	Recent	
NAME	DATE/TIME ▾	SIZE	PROJECT	STATUS
ERstatus_gene_list	6/20/2018, 5:12 AM	654.0 Bytes	Default Project ▾	
▶ gene_set_characterization-fish...	6/20/2018, 5:42 AM		Default Project ▾	🔄
▶ gene_set_characterization-DRaW...	6/20/2018, 5:33 AM		Default Project ▾	

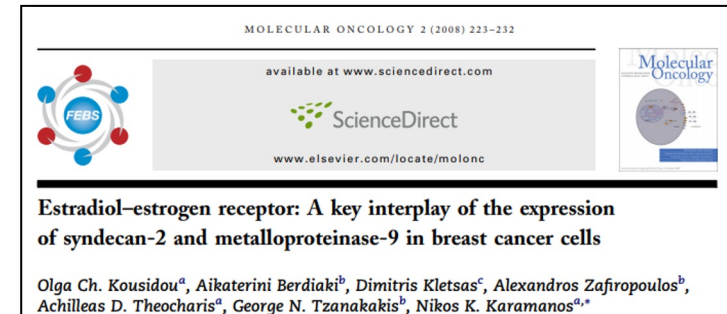
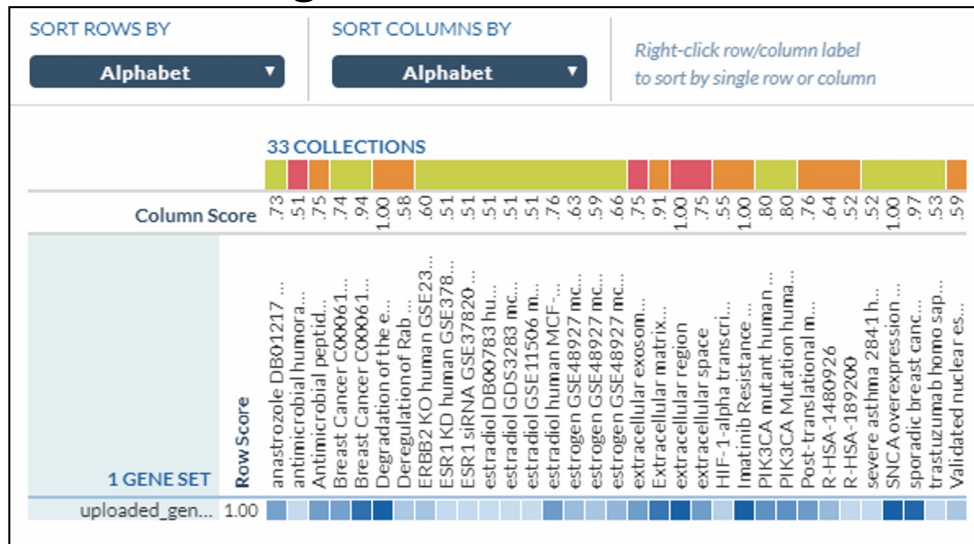
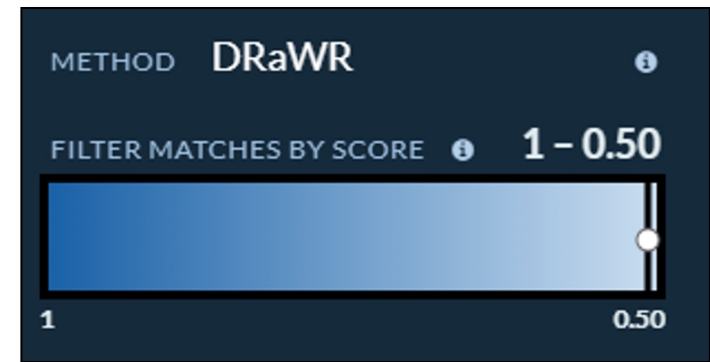
- You can check the status of your jobs here. Gray arrows mean that your job is currently queued or running. A red icon means something went wrong.
- Otherwise, when your job is successfully finished, you should be able to click the green arrow and see the primary result files.
- Click on the DRaWR job “**gene\_set\_characterization-DRaWR-HN**”
- Then click on the “**View Results**” button





# Step 6C: View DRaWR Results

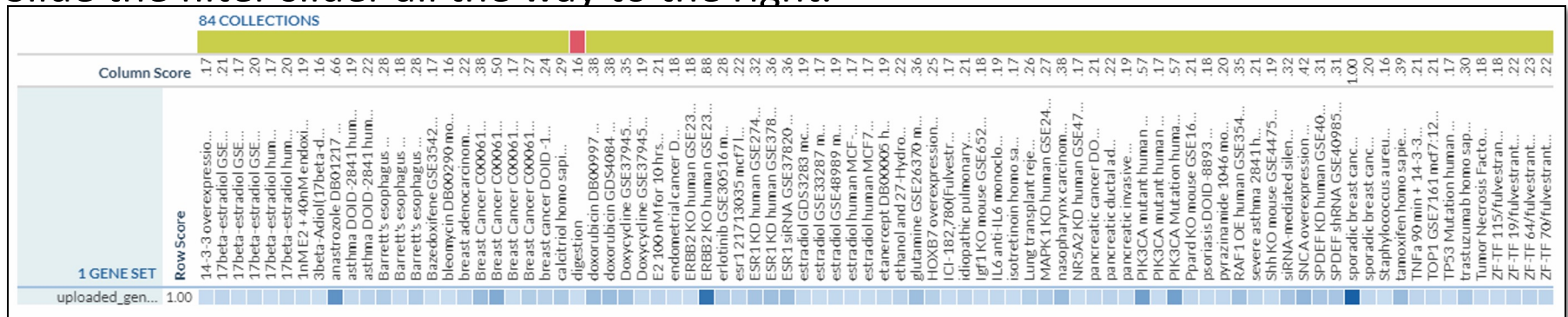
- Slide the filter slider all the way to the right.
- The DRaWR method picks up many GEO Expression gene sets that relate to ESR1 and estrogen and estradiol.



- DRaWR also ranks highly a number of pathway and Gene Ontology terms related to extracellular space, which is known to have many molecules effected by estrogens and related to ER expression

# Step 6C: View Fisher Results

- Click the “**Data**” link at the top of the page
- Click on the DRaWR job “**gene\_set\_characterization-fisher**”
- Then click on the “**View Results**” button
- Slide the filter slider all the way to the right.



- The Fisher method finds the same GEO Expression gene sets that relate to ESR1 and estrogen and estradiol, as well as some additional estradiol ones that DRaWR missed. It also detects many more less obviously related GEO gene sets.
- The standard enrichment method does not detect any highly significant enrichments with pathways or Gene Ontology terms.
- Since it is missing here the extracellular space terms detected by DRaWR are strongly connected to the signature genes, but mostly through their HumanNet network neighbors and not direct connections.

# Main Lab Take Home Message

- Whether it is
  - Sample Clustering
  - Gene Prioritization
  - Gene Set Characterization
- Omics data can be analyzed
  - in the context of a pathway, interaction, or other affinity network
  - to provide complementary insights to standard approaches

# List of Other KnowEnG Resources

- **Other Pipelines:**
  - **Network Preparation** for uploading your custom network to the platform for analysis
  - **Signature Analysis** for mapping samples to signatures by correlation of omics profiles
- **Tutorials:**
  - Quickstarts: <https://knoweng.org/quick-start/>
  - YouTube: <https://www.youtube.com/channel/UCjyIloIcaZIGtZC20XLBOyg>
- **Resources:**
  - Data Preparation Guide: [https://github.com/KnowEnG/quickstart-demos/blob/master/pipeline\\_readmes/README-DataPrep.md](https://github.com/KnowEnG/quickstart-demos/blob/master/pipeline_readmes/README-DataPrep.md)
  - Knowledge Network Contents:
    - Summary: <https://knoweng.org/kn-data-references/>
    - Download: [https://github.com/KnowEnG/KN\\_Fetcher/blob/master/Contents.md](https://github.com/KnowEnG/KN_Fetcher/blob/master/Contents.md)
- **Research:**
  - Knowledge-guided analysis of omics Data (KnowEnG cloud platform paper): <https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000583>
  - TCGA Analysis Walkthrough: [https://github.com/KnowEnG/quickstart-demos/tree/master/publication\\_data/blatti\\_et\\_al\\_2019](https://github.com/KnowEnG/quickstart-demos/tree/master/publication_data/blatti_et_al_2019)
- **Source Code:**
  - Docker Images: <https://hub.docker.com/u/knowengdev/>
  - GitHub Repos: <https://knoweng.github.io/>
- **Other Cloud Platforms:**
  - <https://cgcbioinformatics.com/public/apps?q=search=knoweng>
- Contact Us with Questions and Feedback: [knoweng-support@illinois.edu](mailto:knoweng-support@illinois.edu)