Knowledge-Guided Sample Clustering and Gene Prioritization

KnowEnG Center

PowerPoint by Amin Emad

Summary

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

knoweng
BIG DATA TO KNOWLEDGE CENTER OF EXCELLENCE

- We will focus on the Spreadsheet Visualization, Clustering, and Gene Prioritization pipelines implemented in KnowEnG

- We will try both network-guided and standard modes of operation for the pipelines (if applicable)

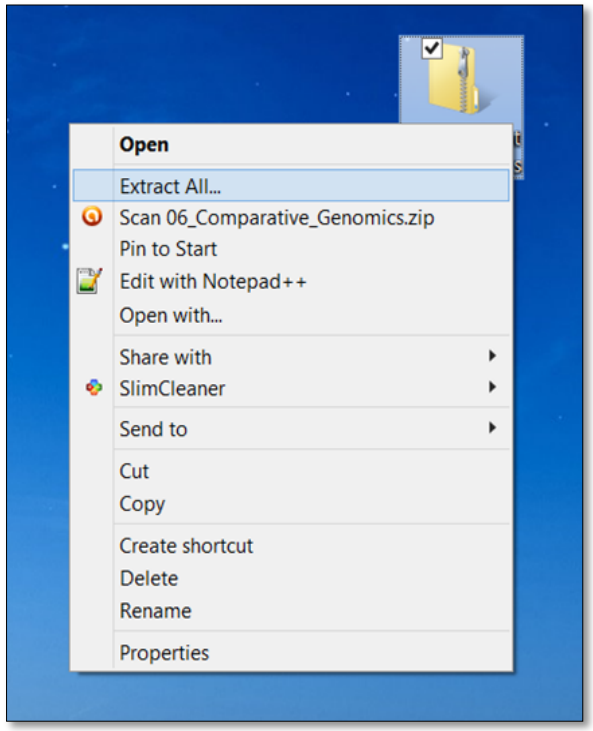
Data

- First download the data which we will use from the link below:

  http://publish.illinois.edu/computational-genomics-course/files/2019/06/08_Clustering_and_Prioritization.zip

- After the download is complete, Right Click and Extract the contents of the archive to your course directory. We will use the files found in:

- [course_directory]/08_Clustering_and_Prioritization/



Step 1: Sign Into KnowEnG Platform

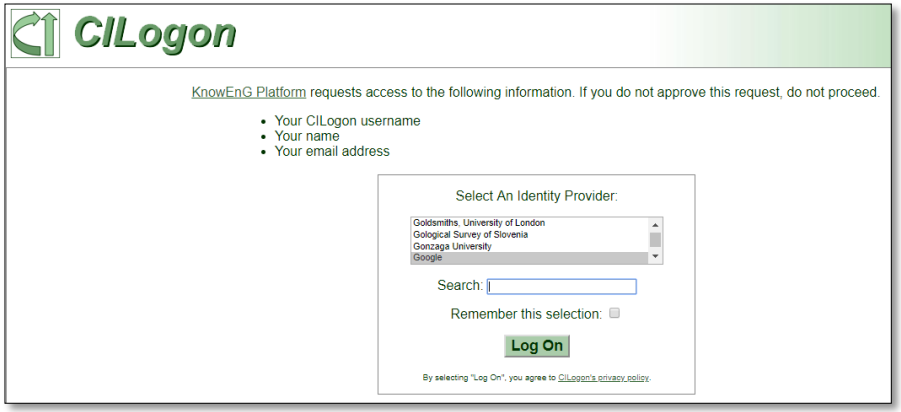
KnowEnG Platform:
https://knoweng.org/analyze/

Go to development version:
https://dev.knoweng.org/

(will be at end of course)

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

Search: **Urbana, Mayo, Google, Github**



Visualization and simple analysis of genomic spreadsheets:

STEP2: Spreadsheet Visualization

- We will use KnowEnG's Spreadsheet Visualization pipeline to explore various properties of a transcriptomic spreadsheet and the relationship between transcriptomic features and different clinical phenotypes

- We will use data corresponding to breast tumor samples from the METABRIC study

Nature. 2012 Apr 18;486(7403):346-52. doi: 10.1038/nature10983.

The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.

Curtis C[1], Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Gräf S, Ha G, Haffari G, Bashashati A, Russell R, McKinney S; METABRIC Group, Langerød A, Green A, Provenzano E, Wishart G, Pinder S, Watson P, Markowetz F, Murphy L, Ellis I, Purushotham A, Børresen-Dale AL, Brenton JD, Tavaré S, Caldas C, Aparicio S.

STEP2: Spreadsheet Visualization

Dataset characteristics:

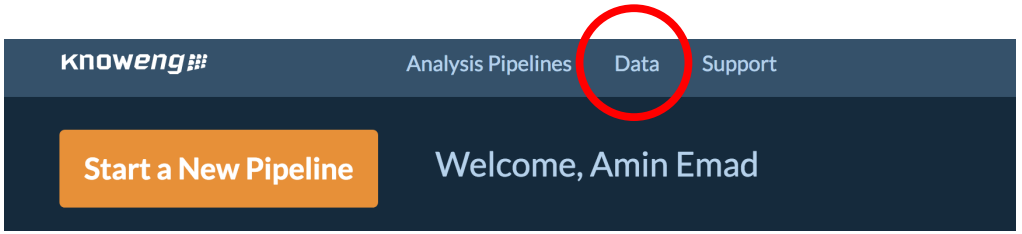
Name	Description
Expression_METABRIC_Demo1	A matrix of (gene x samples) containing the expression (microarray) of 233 genes in 1058 samples. The expression profiles are normalized in advance.
Phenotype_METABRIC_Demo1	A matrix of (samples x clinical phenotypes) including PAM50 subtype, treatment, stage, survival years, etc.

STEP2: Spreadsheet Visualization

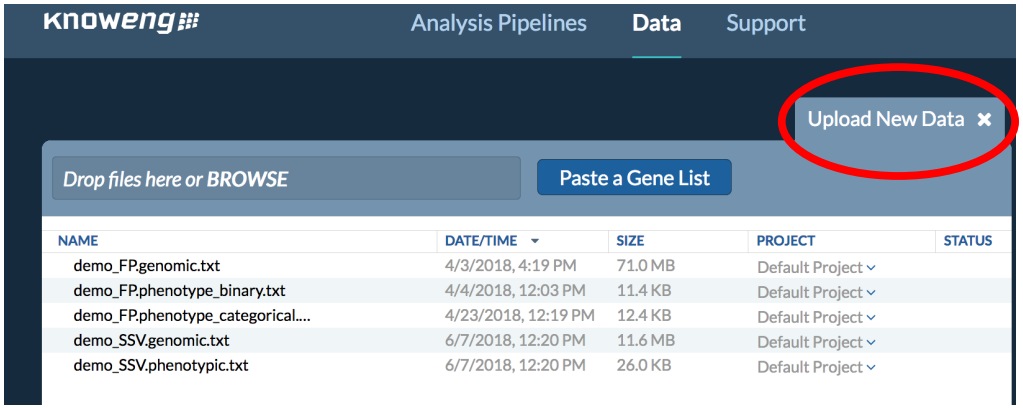
knoweng BIG DATA TO KNOWLEDGE CENTER OF EXCELLENCE

Upload the data:

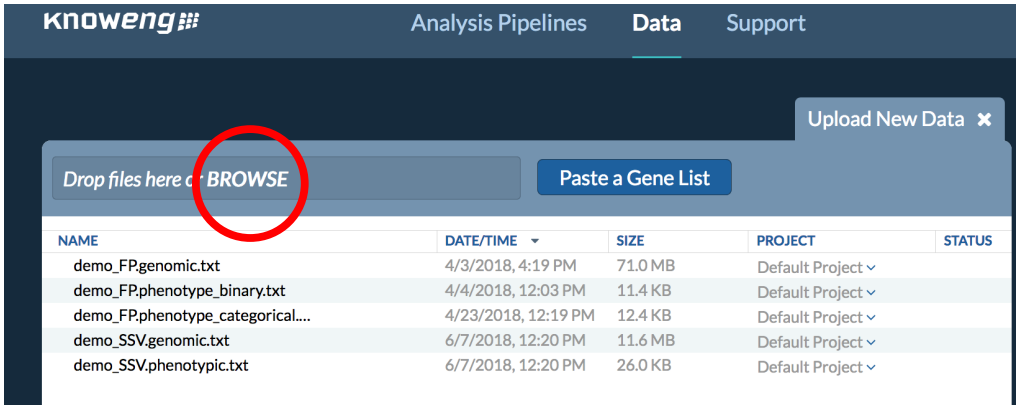
- Select "**Data**" at the top of the page



- Click on "**Upload New Data**"



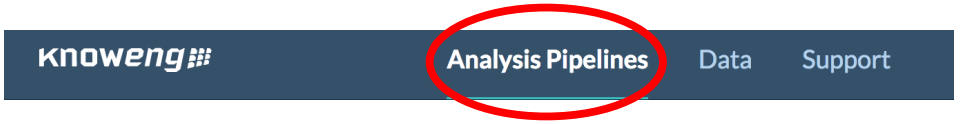
- Click "**BROWSE**" and find the files to upload:
  - Expression_METABRIC_Demo1
  - Phenotype_METABRIC_Demo1



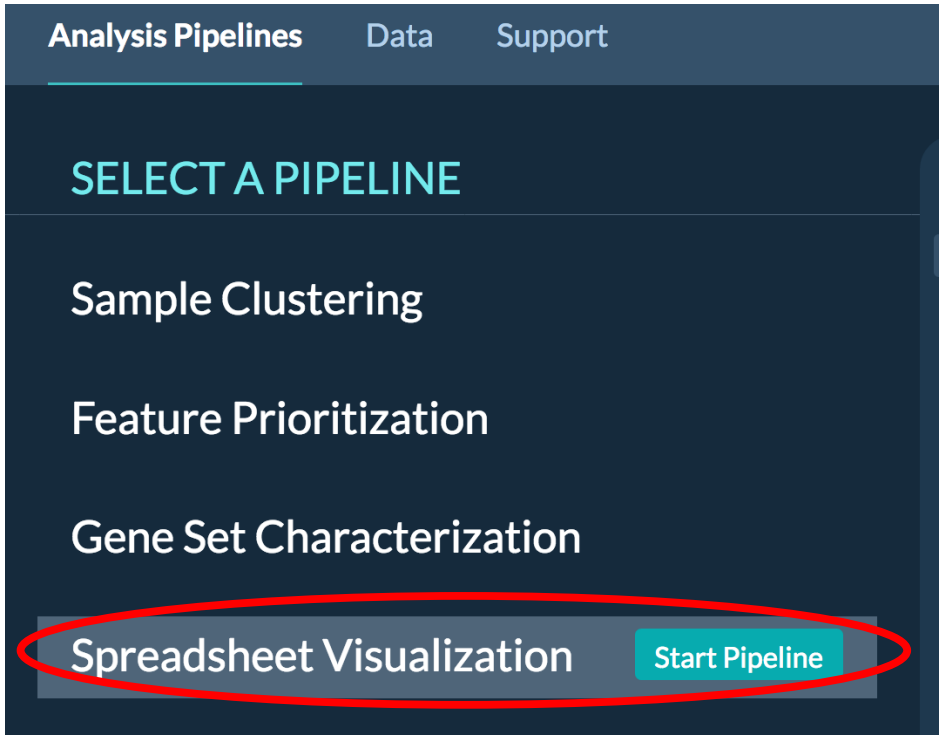
STEP2: Spreadsheet Visualization

Select the pipeline:

- Select "**Analysis Pipelines**" at the top of the page



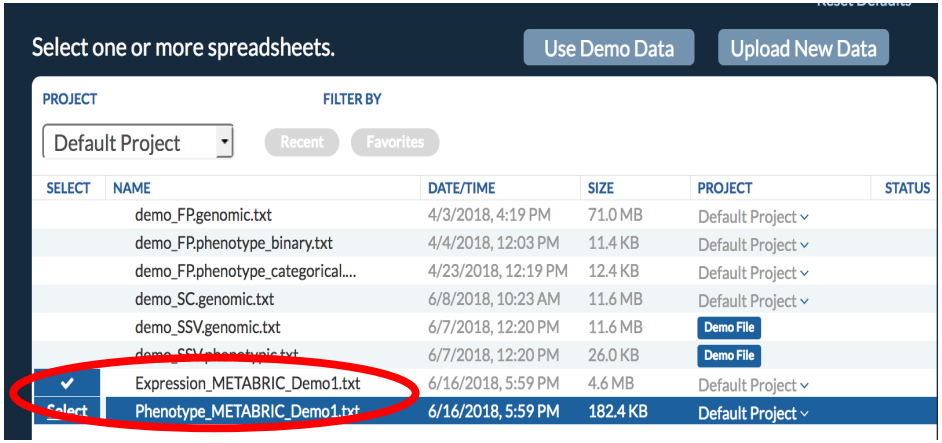
- Select "**Spreadsheet Visualization**" and Click on "**Start Pipeline**"



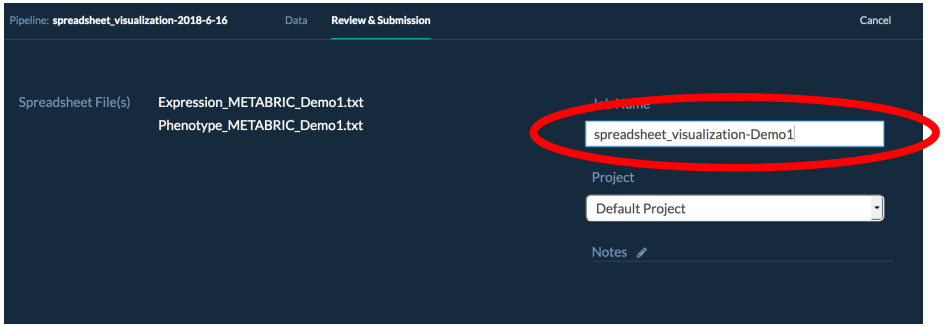
STEP2: Spreadsheet Visualization

Configure the pipeline:

- Select the files:
  - Expression_METABRIC_Demo1.txt
  - Phenotype_METABRIC_Demo1.txt

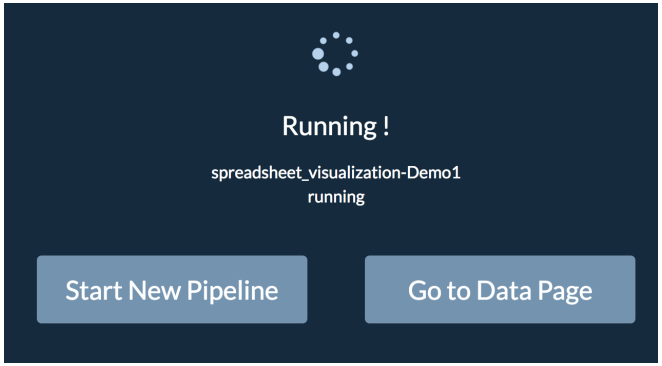


- Select "**Next**" at the right bottom corner of the page



- You can change the name of the results

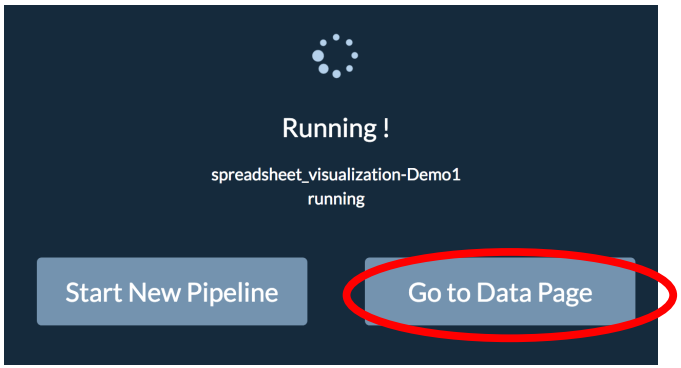
- Then press "**Submit Job**"



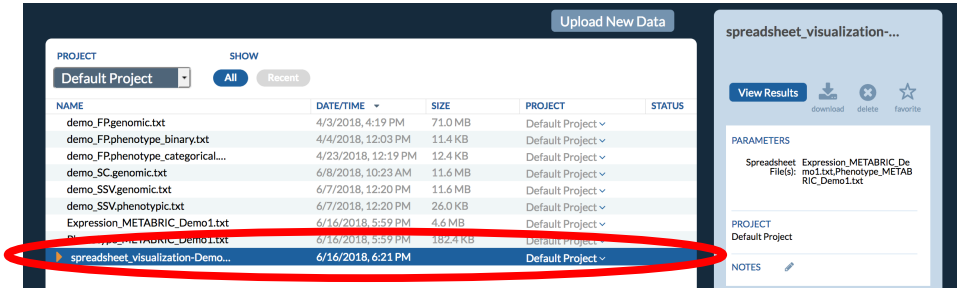
STEP2: Spreadsheet Visualization

The results:

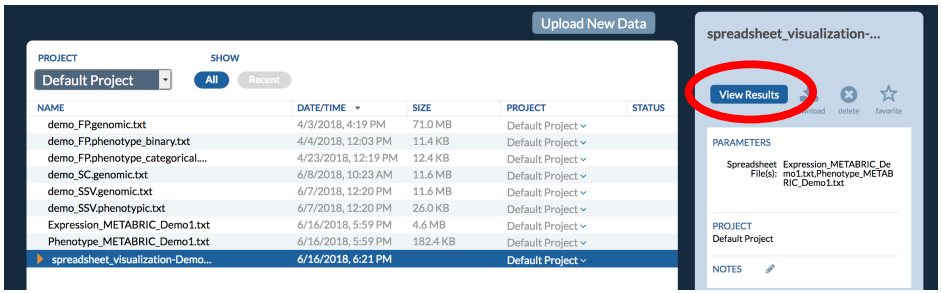
- Select "**Go to Data Page**"



- Select the job you just ran



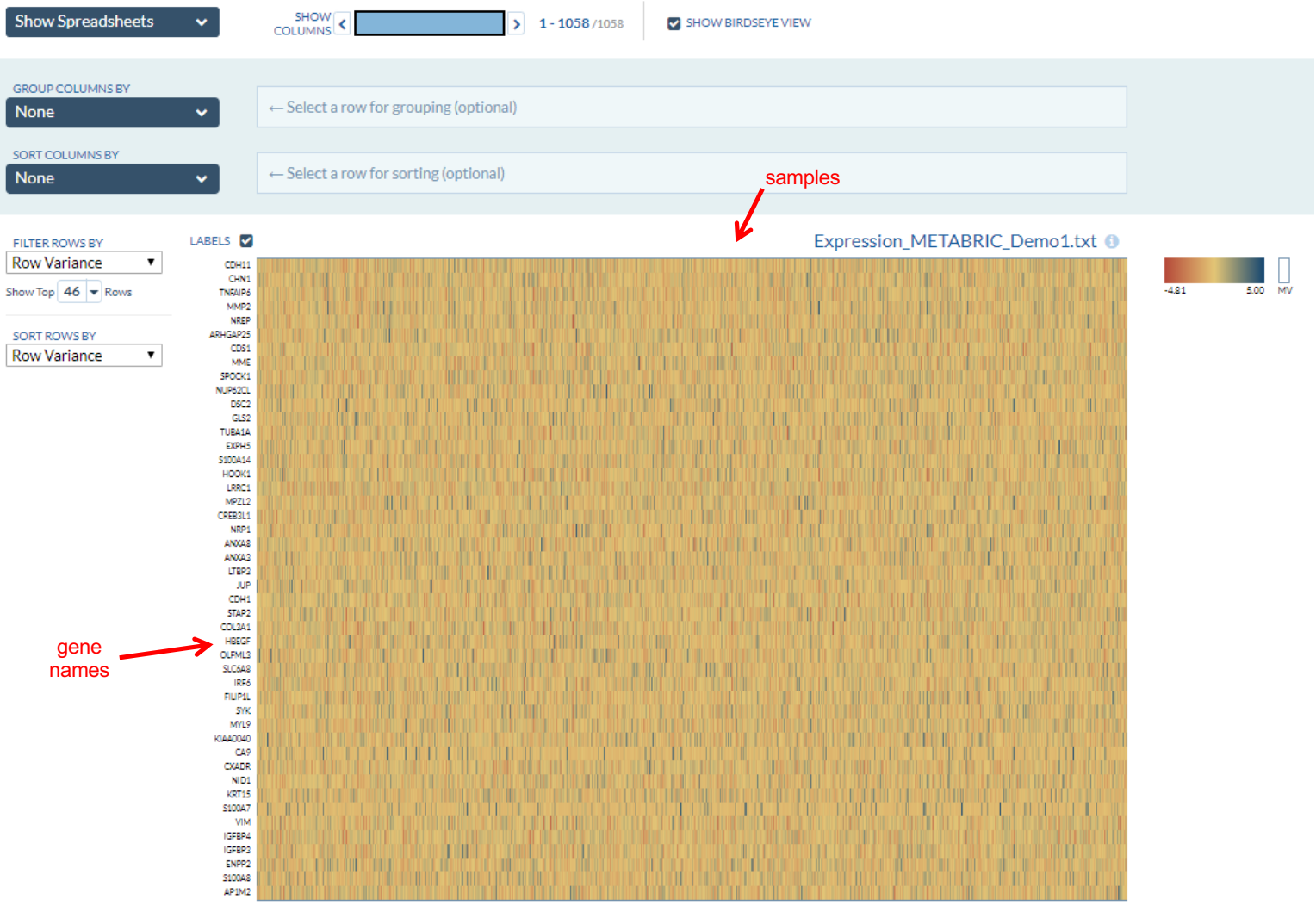
- Then "**View Results**"



STEP2: Spreadsheet Visualization

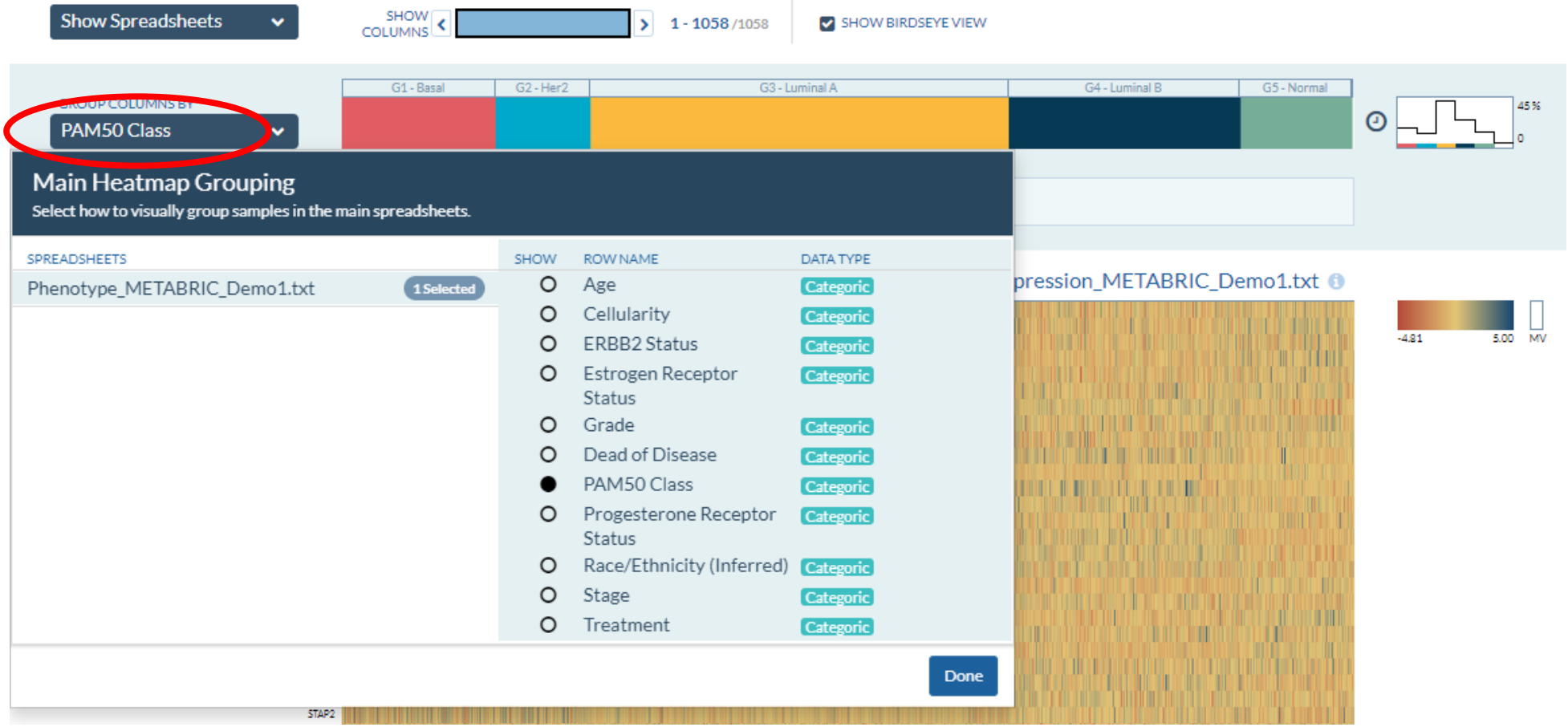
knowENG BIG DATA TO KNOWLEDGE CENTER OF EXCELLENCE

Allows grouping/sorting of columns using another spreadsheet



STEP2: Spreadsheet Visualization

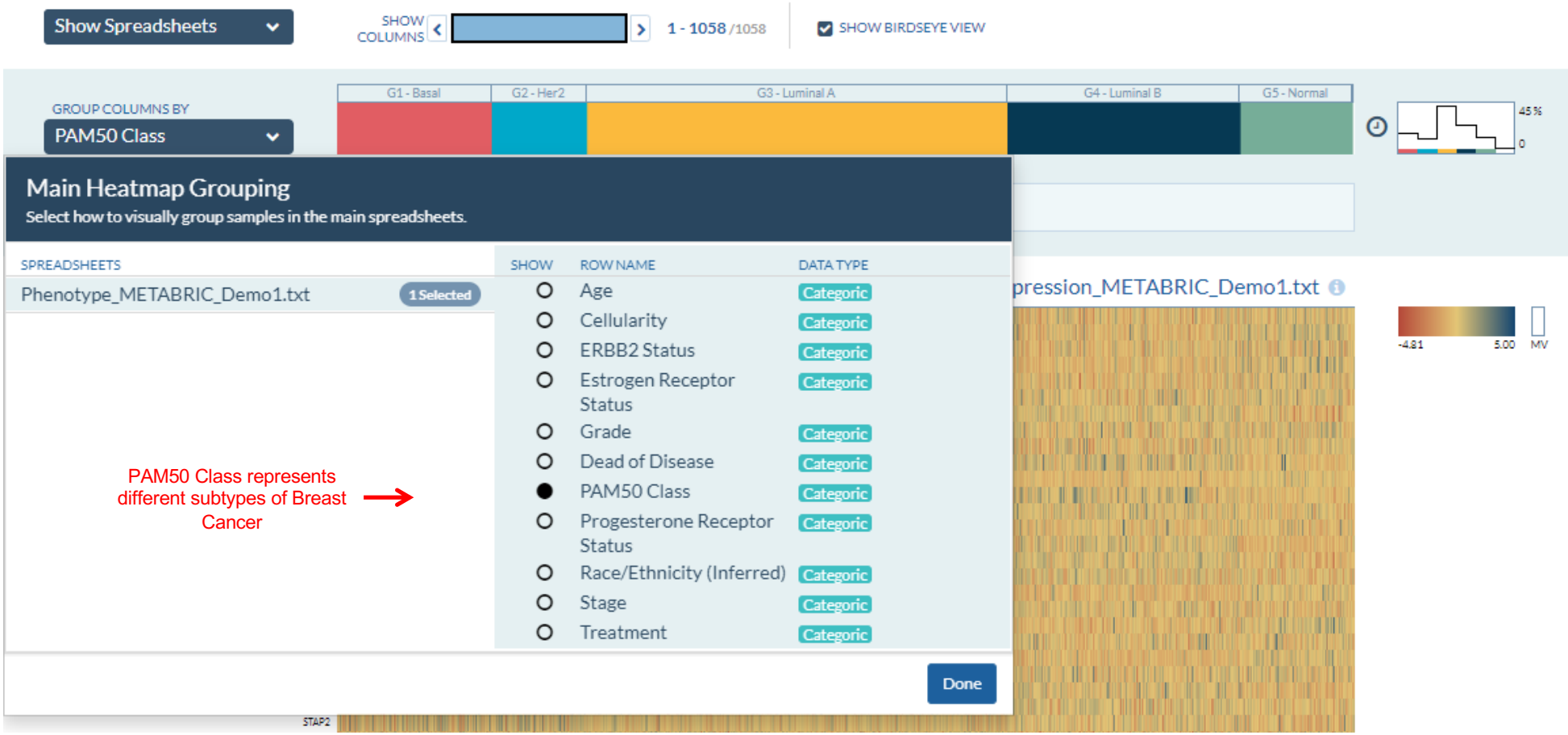
- Click the dropdown "**Group Columns By**" menu and select the phenotype spreadsheet (Phenotype_METABRIC_Demo1.txt)



STEP2: Spreadsheet Visualization

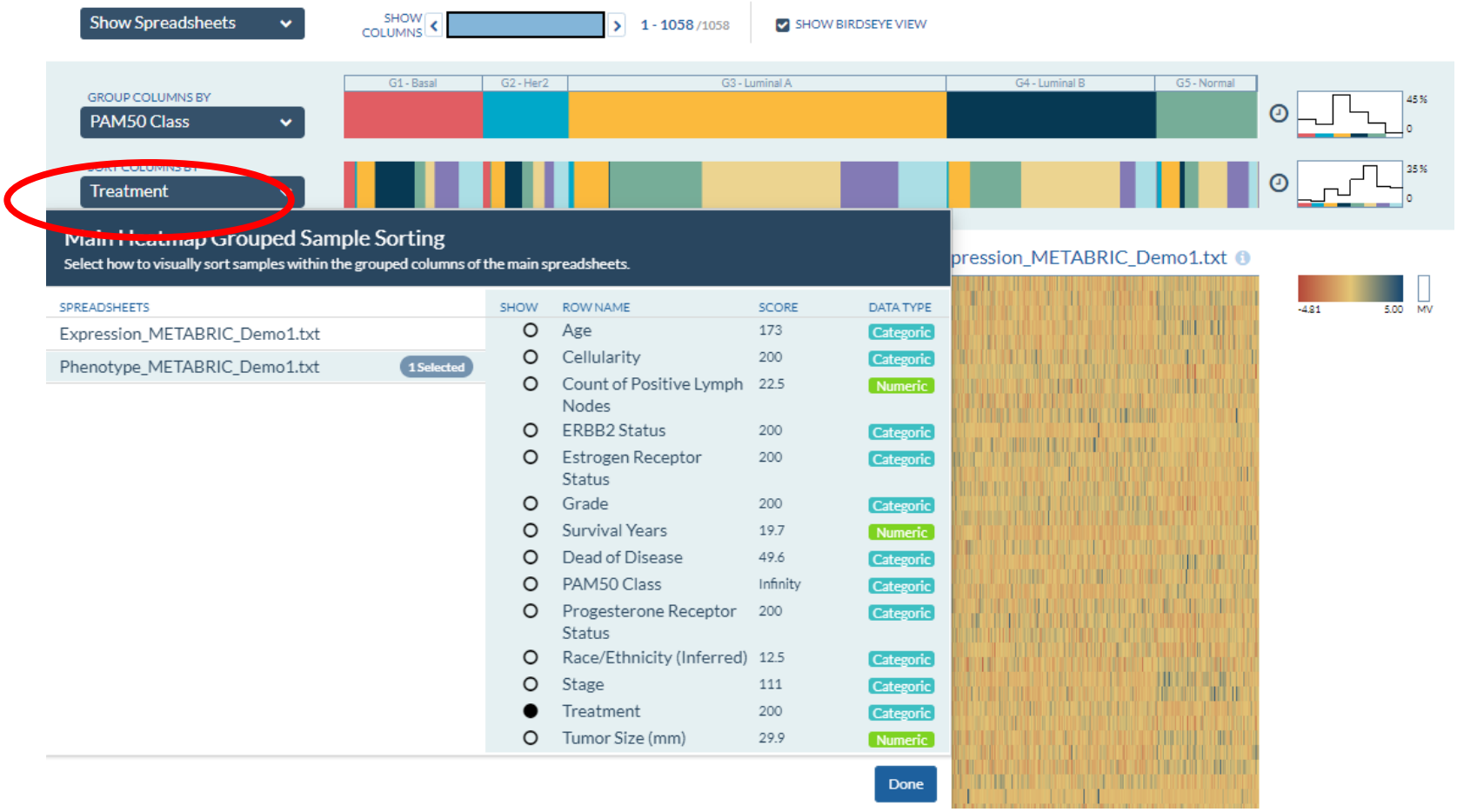
knowengBIG DATA TO KNOWLEDGE CENTER OF EXCELLENCE

- Click the dropdown "**Group Columns By**" menu and select the phenotype spreadsheet (Phenotype_METABRIC_Demo1.txt)
- Select "**PAM50 Class**": the columns of the heatmap will automatically reorganize accordingly. Then press **Done**.



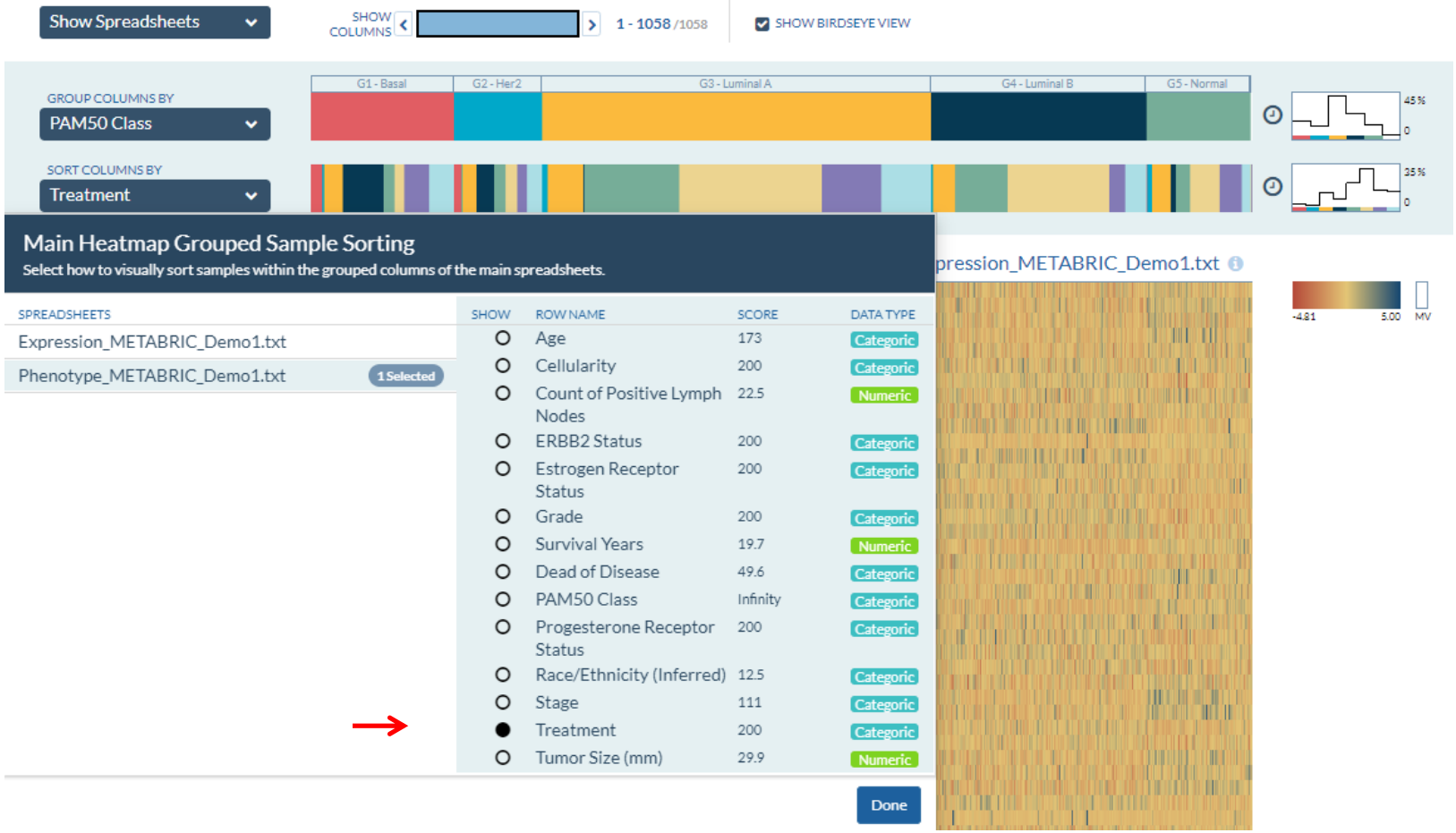
STEP2: Spreadsheet Visualization

- Click the dropdown "**Sort Columns By**" menu and select the phenotype spreadsheet (Phenotype_METABRIC_Demo1.txt) again



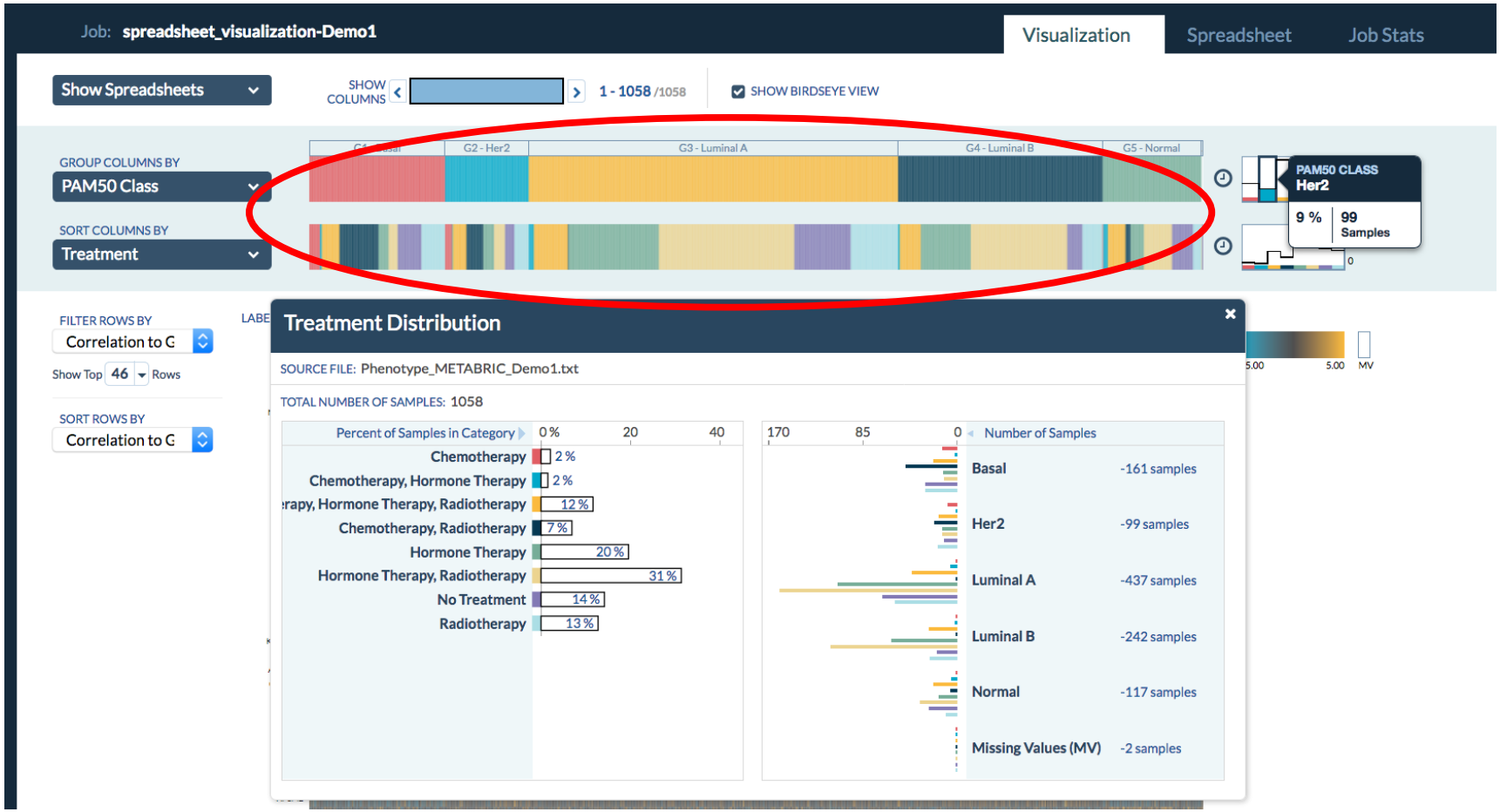
STEP2: Spreadsheet Visualization

- Click the dropdown "**Sort Columns By**" menu and select the phenotype spreadsheet (Phenotype_METABRIC_Demo1.txt) again
- Select "**Treatment**": the columns of the heatmap will automatically reorganize accordingly. Then press **Done**.



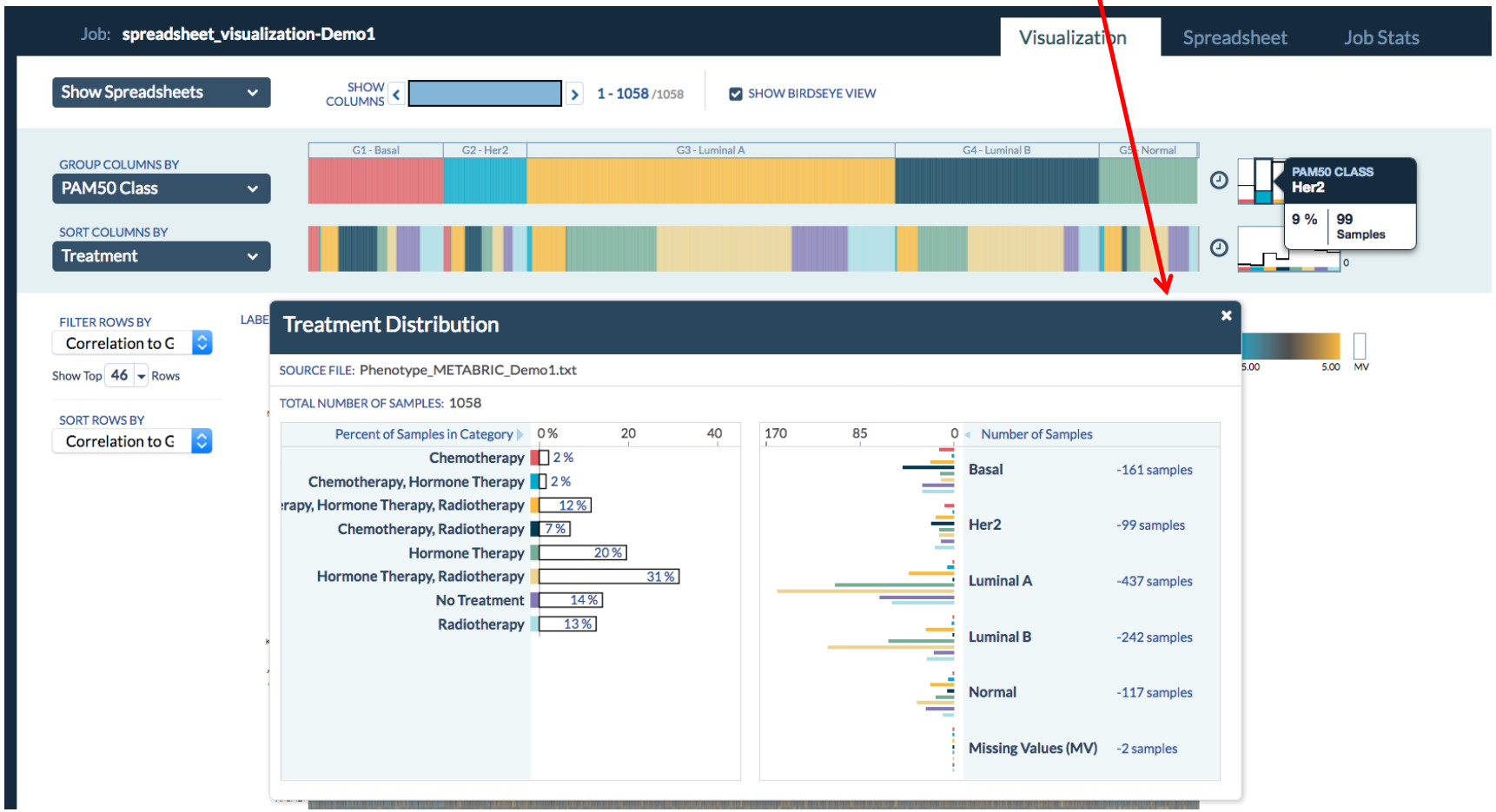
STEP2: Spreadsheet Visualization

- Bars show the status of each sample



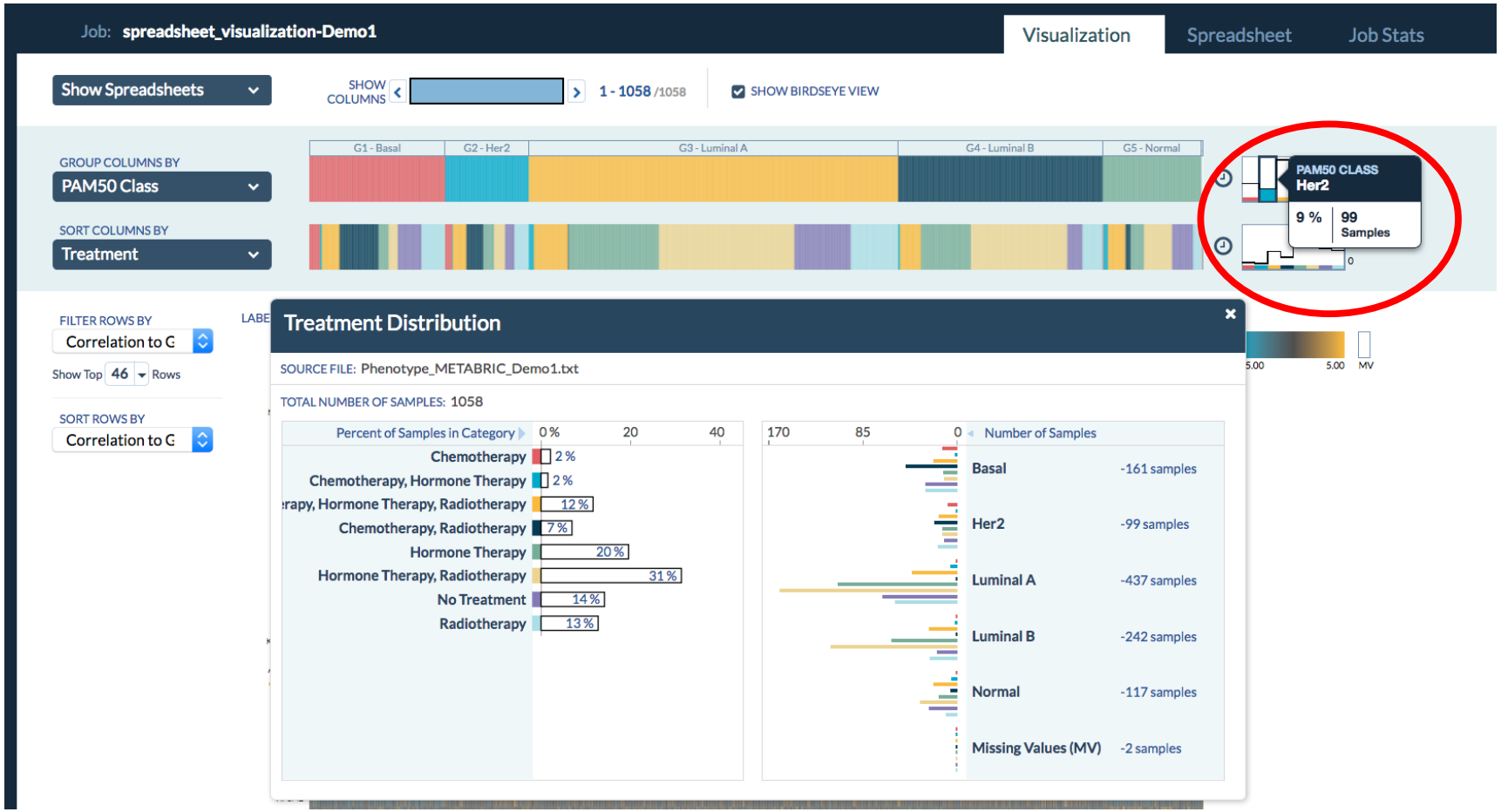
STEP2: Spreadsheet Visualization

- Bars show the status of each sample
- More details can be seen by clicking on the bars



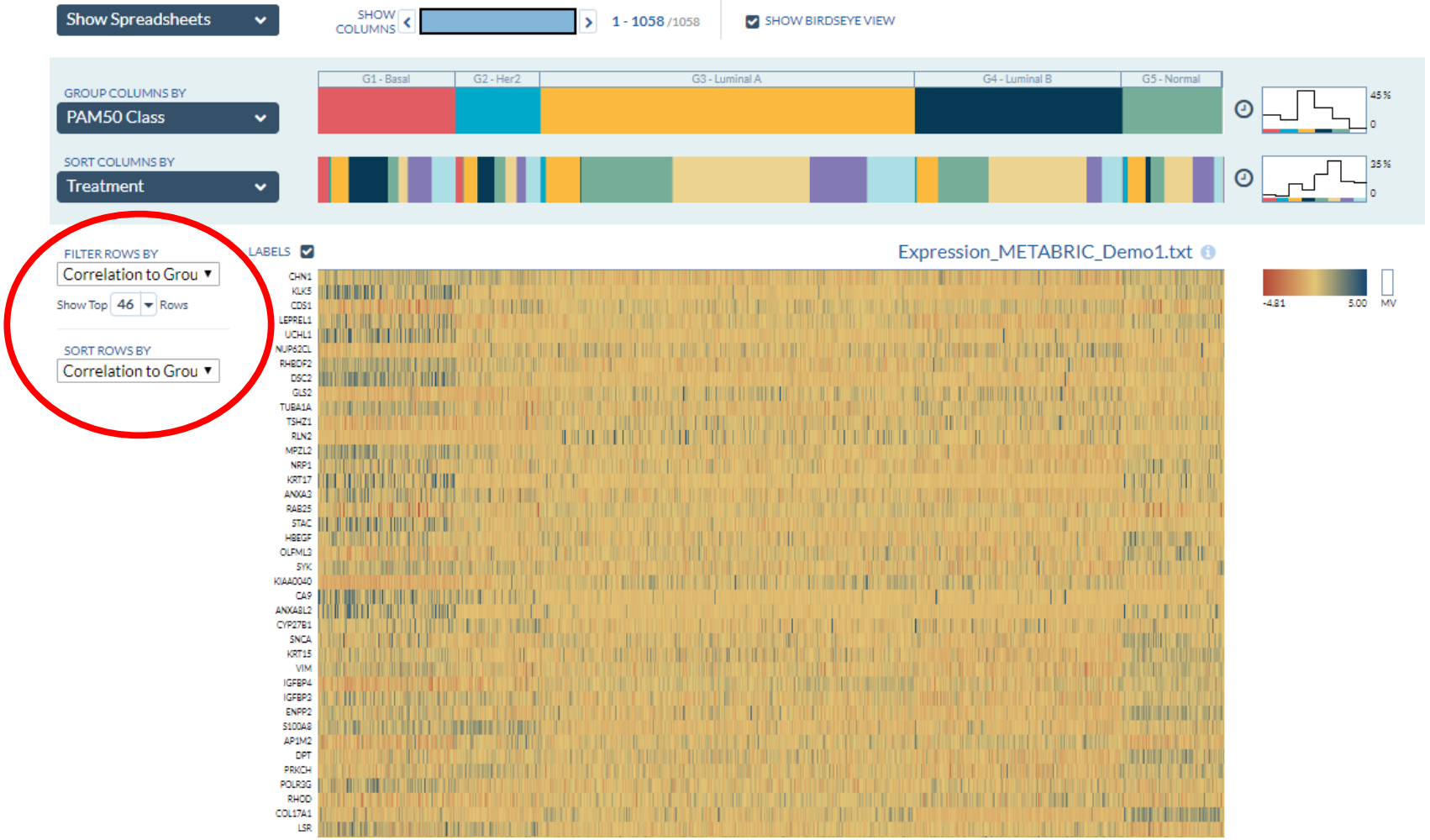
STEP2: Spreadsheet Visualization

- Bars show the status of each sample
- More details can be seen by clicking on the bars
- Bar charts show the histogram of each category



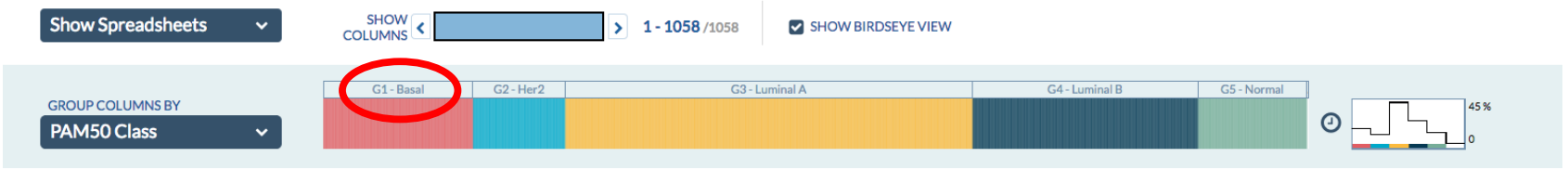
STEP2: Spreadsheet Visualization

- Click the dropdown "**Filter Rows By**" menu and select "**Correlation to Group**". Click the dropdown "**Sort Rows By**" menu and select "**Correlation to Group**".



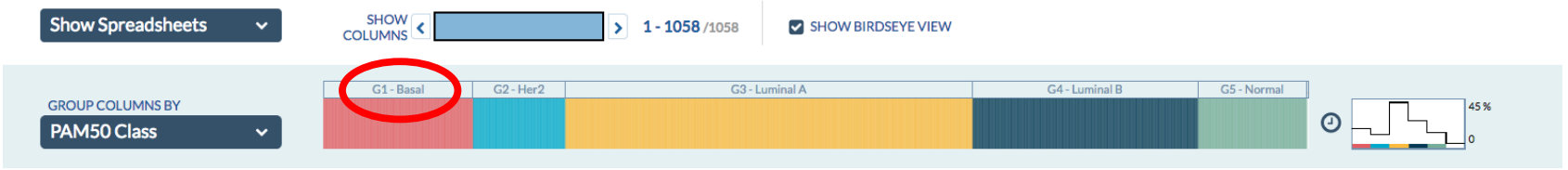
STEP2: Spreadsheet Visualization

- Hover over "**G1-Basal**" and click on it

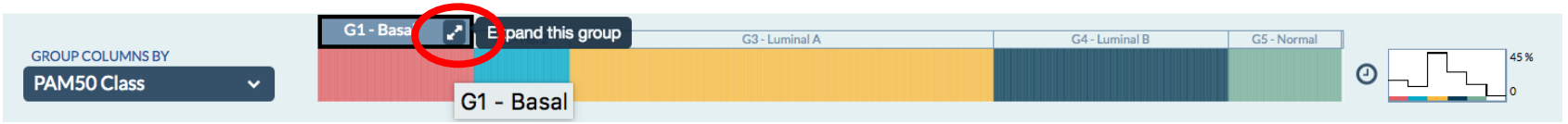


STEP2: Spreadsheet Visualization

- Hover over "**G1-Basal**" and click on it

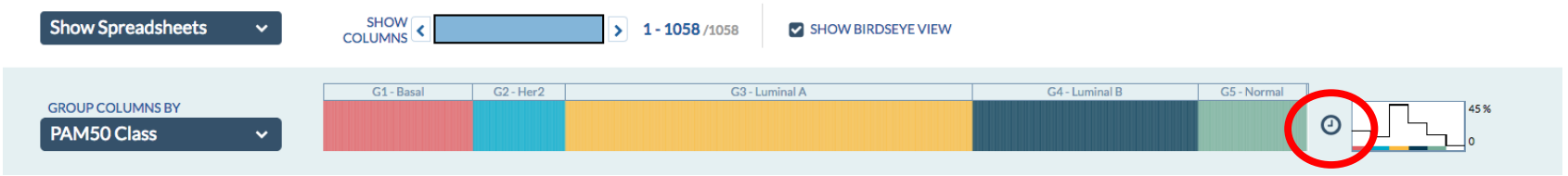


- Click on the arrows to expand the group and observe the expressions

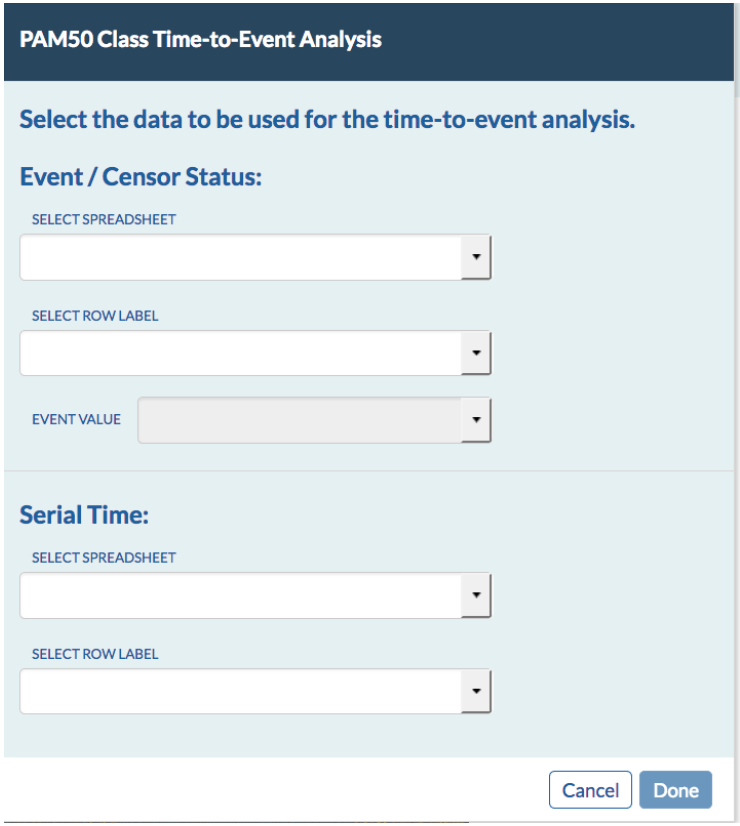


STEP2: Spreadsheet Visualization

- Click on the clock sign to perform Kaplan Meier survival analysis using a set of categories

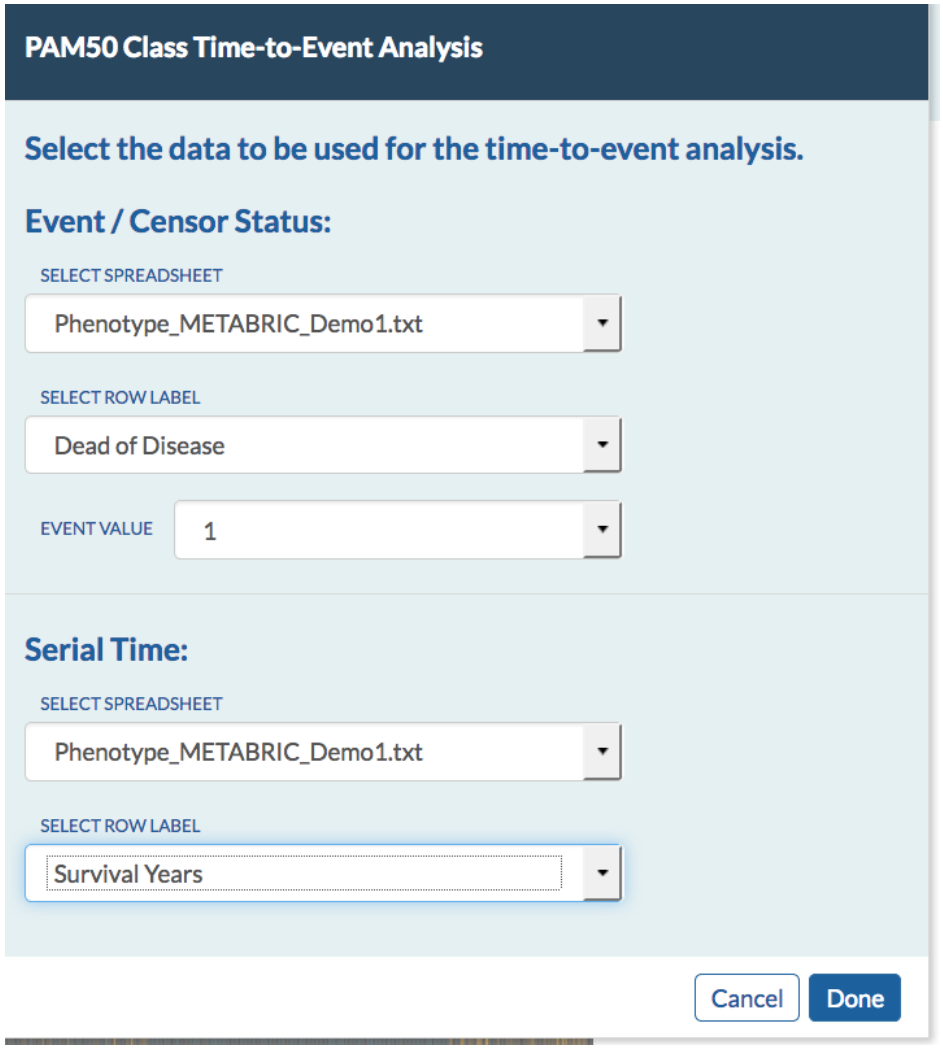


- Use this table to configure Kaplan Meier analysis by selecting the events and time to events



STEP2: Spreadsheet Visualization

- Select the options below for Kaplan Meier analysis and press **Done**.



STEP2: Spreadsheet Visualization

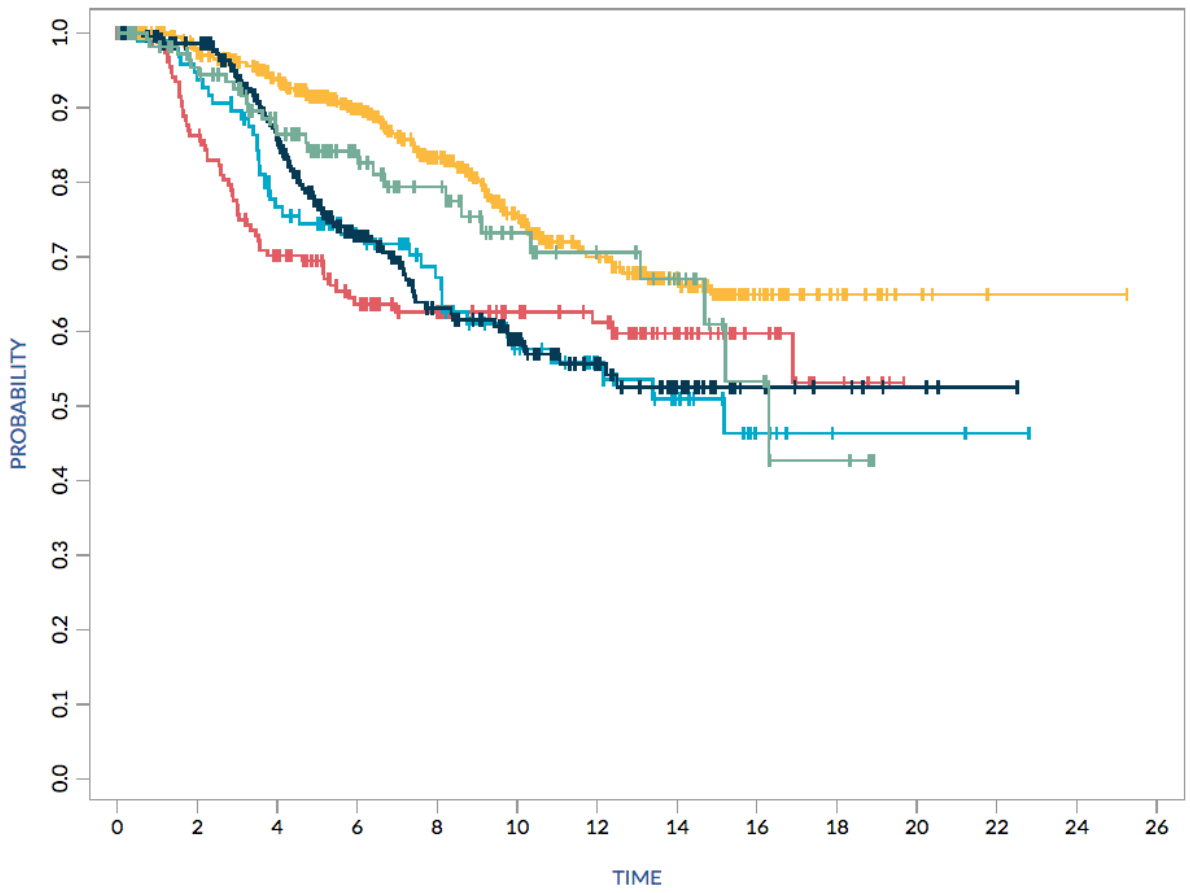
knoweng BIG DATA TO KNOWLEDGE CENTER OF EXCELLENCE

PAM50 Class Time-to-Event Analysis

Edit Setup

Basal
Her2
Luminal A
Luminal B
Normal

Log-Rank p-Value = 9.58e-6



Network-guided clustering of somatic mutations in different cancer types

STEP3: 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[1], Shen JP, Carter H, Gross A, Ideker T.

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

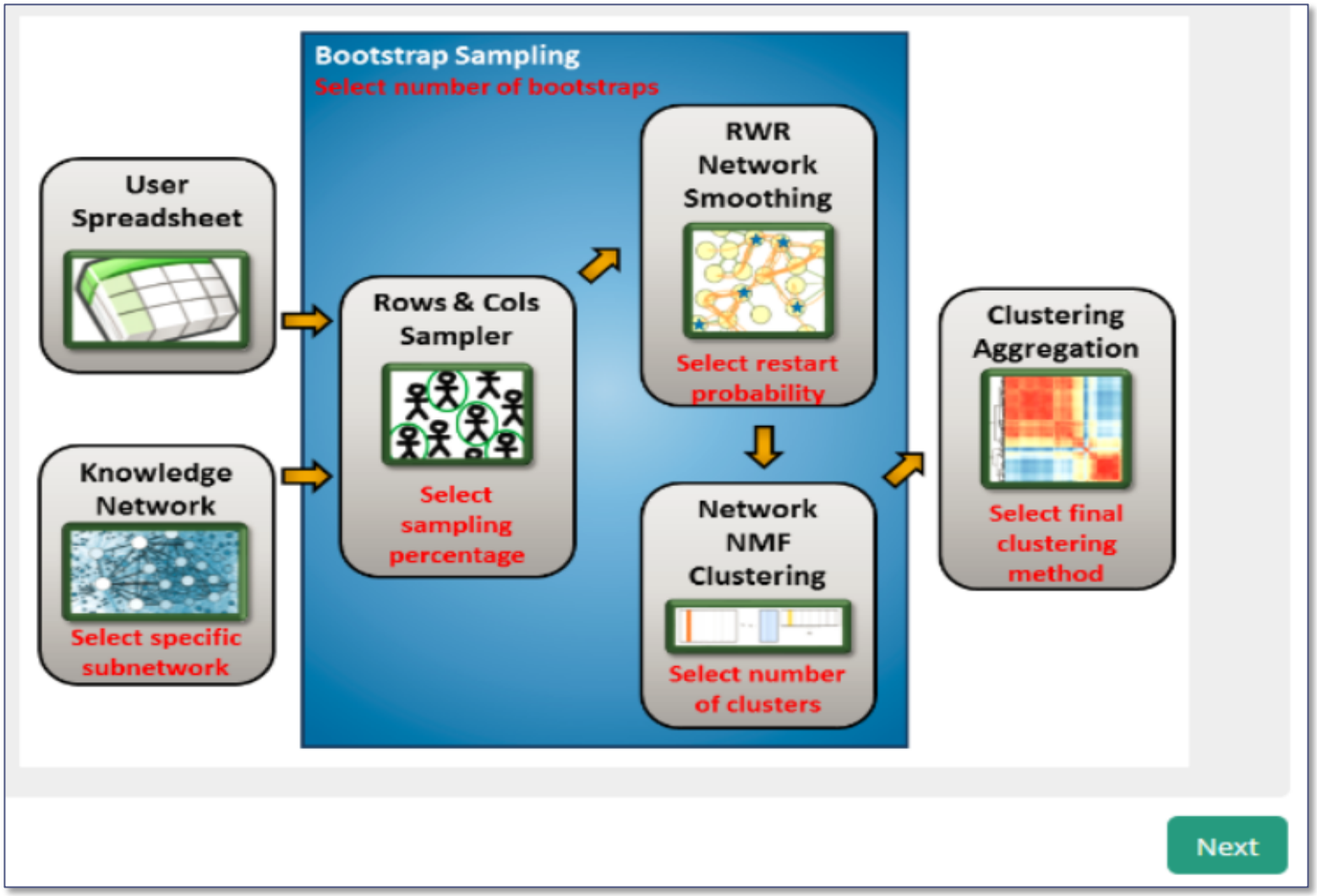
Cell. 2014 Aug 14;158(4):929-944. doi: 10.1016/j.cell.2014.06.049. Epub 2014 Aug 7.

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

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

STEP3: Sample Clustering

Outline of Network-based Stratification:



STEP3: Sample Clustering

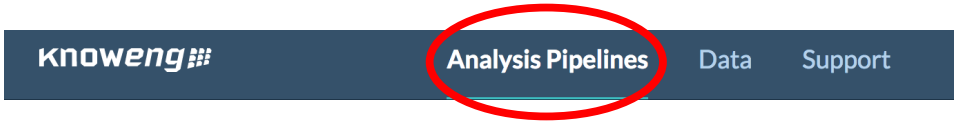
Dataset characteristics:

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.
Demo2_Clinical_pancan12_30	A matrix of (samples x clinical phenotypes) including primary disease, PANCAN consensus cluster, survival years, etc.

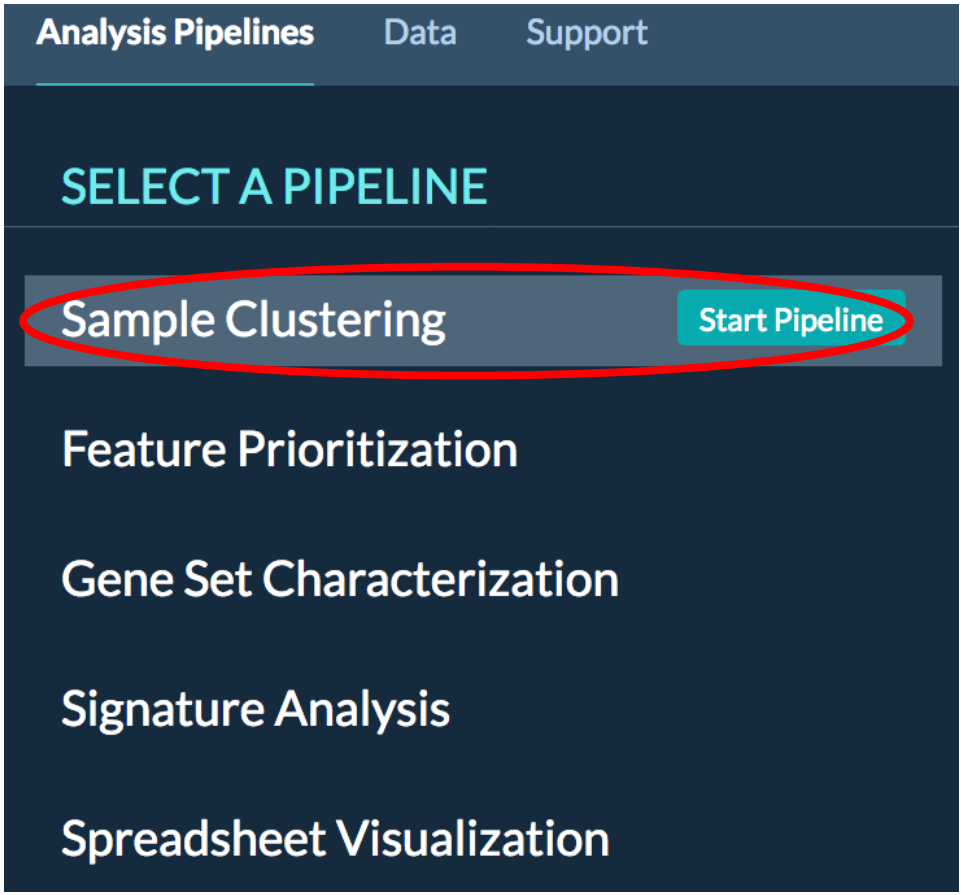
STEP3: Sample Clustering (standard)

Select the pipeline:

- Select "**Analysis Pipelines**" at the top of the page



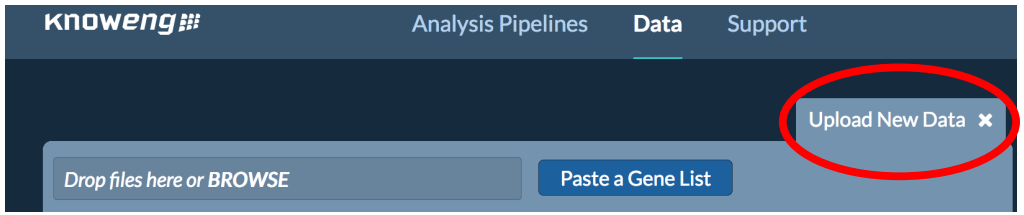
- Select "**Sample Clustering**" and Click on "**Start Pipeline**"



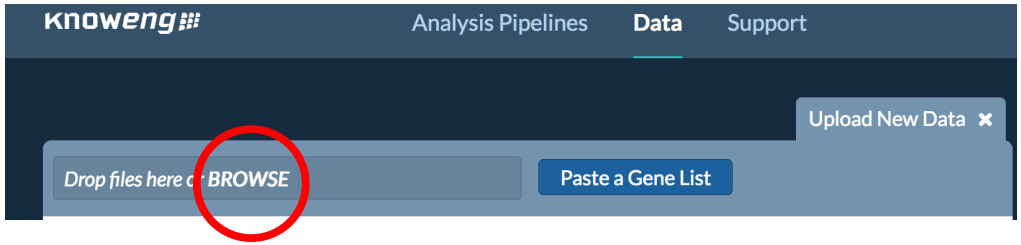
STEP3: Sample Clustering (standard)

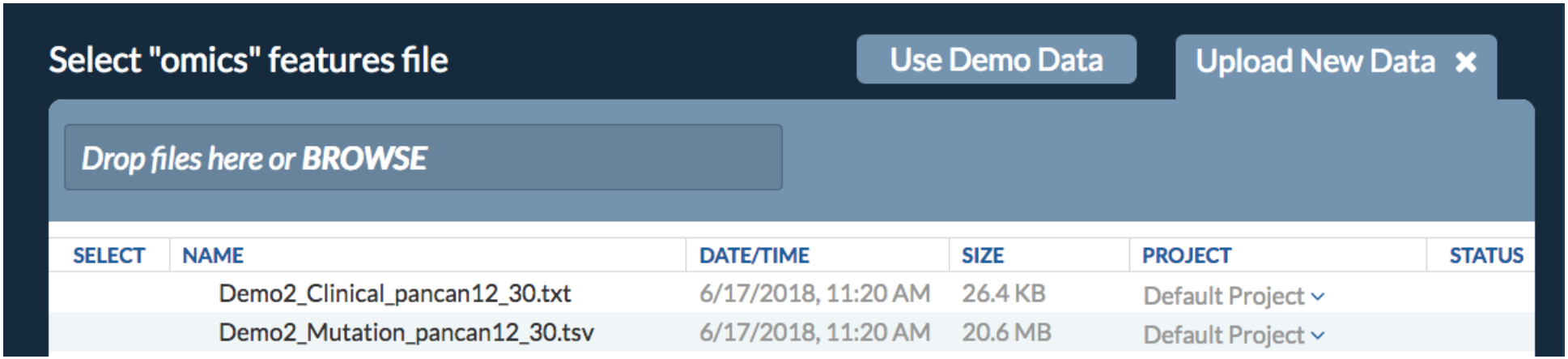
Upload the data:

- Click on "**Upload New Data**"



- Click "**BROWSE**" and find the files to upload:
  - Demo2_Clinical_pancan12_30
  - Demo2_Mutation_pancan12_30

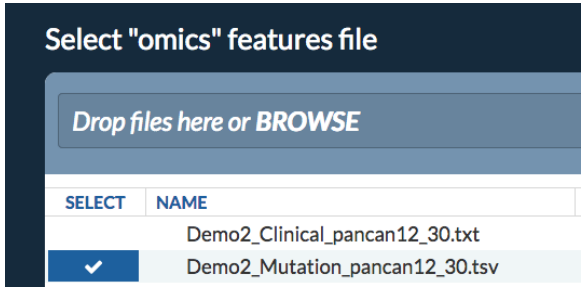




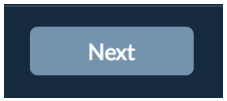
STEP3: Sample Clustering (standard)

Configure the pipeline:

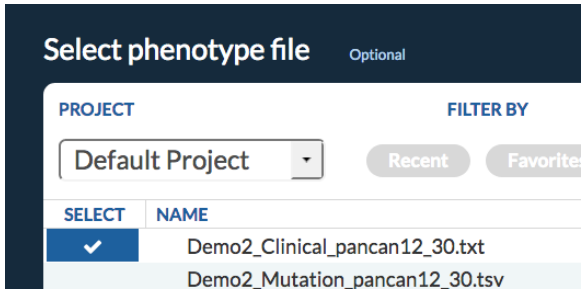
- For the "**omics**" file select:
  - Demo2_Mutation_pancan12_30



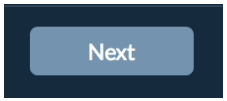
- Click "**Next**" at the bottom right corner



- For the "**phenotype**" file select:
  - Demo2_Clinical_pancan12_30

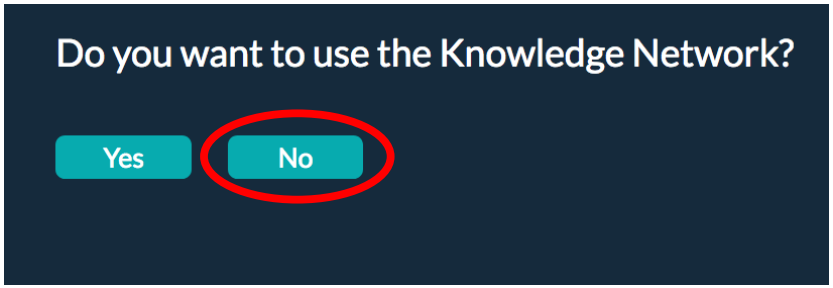


- Click "**Next**" at the bottom right corner

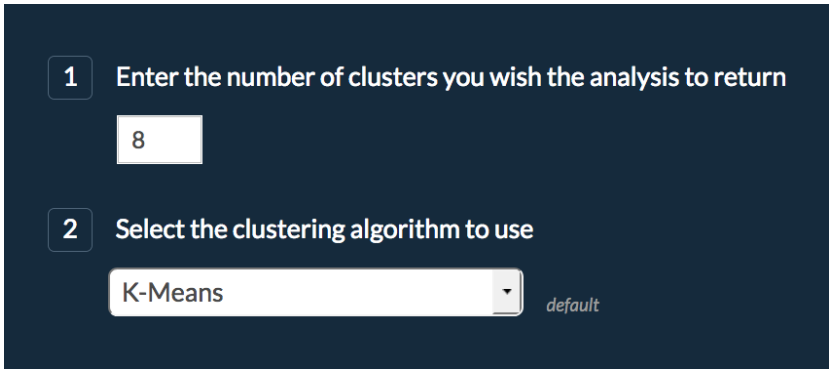


STEP3: Sample Clustering (standard)

- Select "**No**" in response to using the knowledge network:
  - This allows us to perform standard clustering on the data

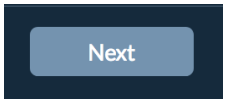


- Choose **8** as number of clusters



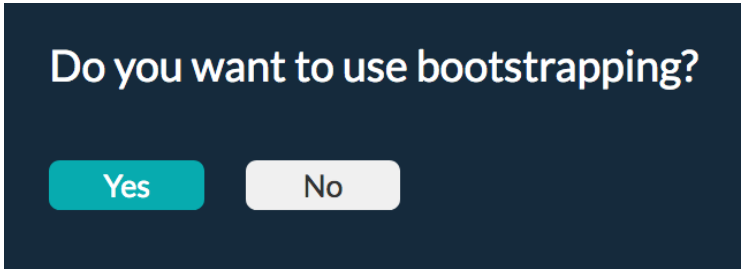
- We will use the default "**K-Means**" clustering algorithm

- Click on "Next" at the bottom right corner

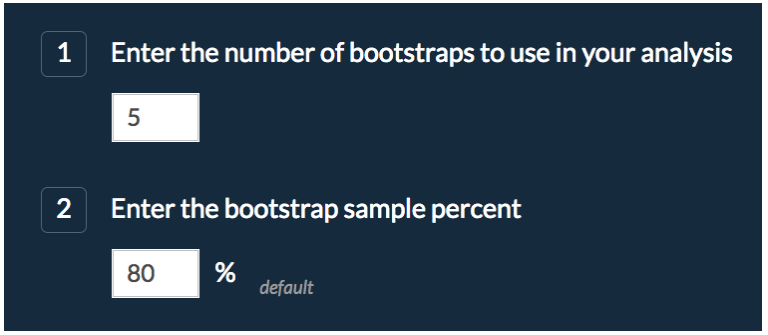


STEP3: Sample Clustering (standard)

- Select "**Yes**" in response to using bootstrap sampling:
  - This allows us to obtain a more robust final clustering

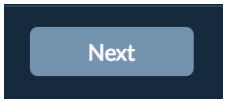


- Choose **5** as number of bootstraps



- We will use the default **80%** rate to sample the data in each bootstrap

- Click on "Next" at the bottom right corner



STEP3: Sample Clustering (standard)

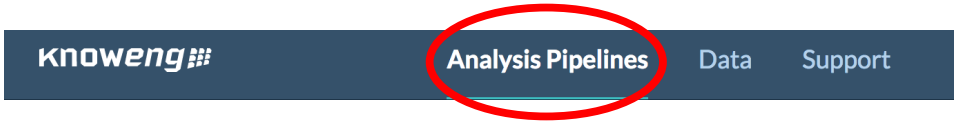
- Review the summary of the job and change the default "**Job Name**" to easily recognize later

- Submit the job

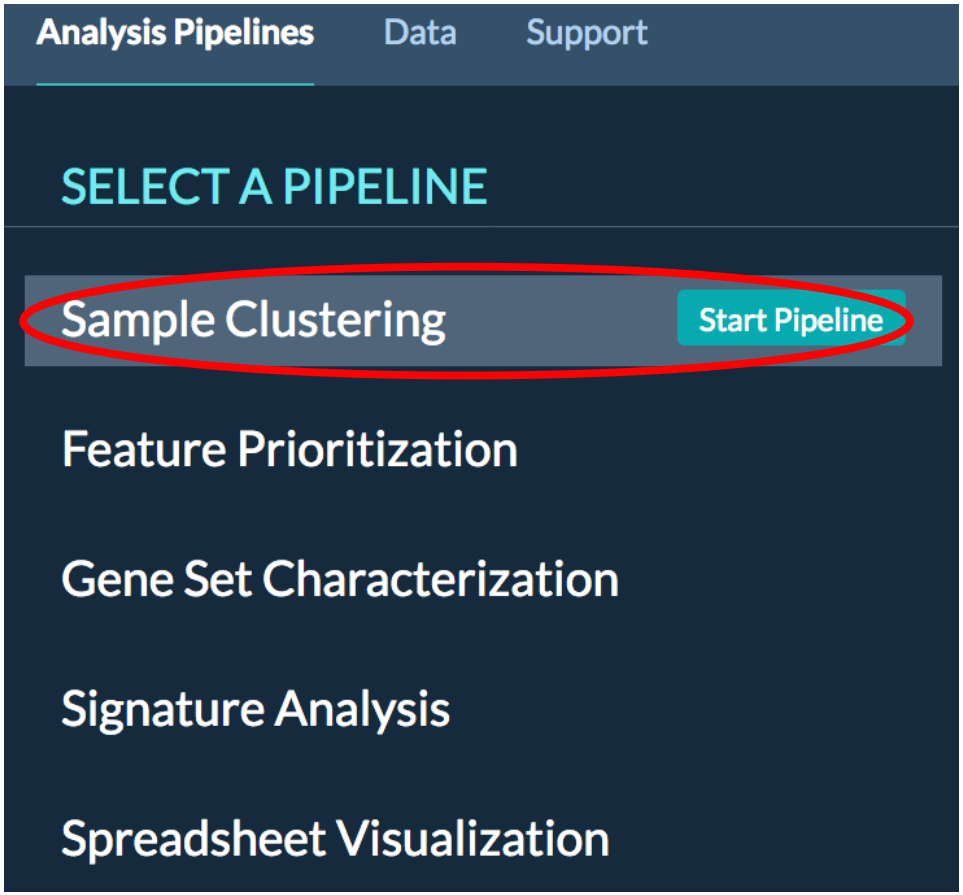
STEP3: Sample Clustering (network-guided)

Select the pipeline:

- Select "**Analysis Pipelines**" at the top of the page



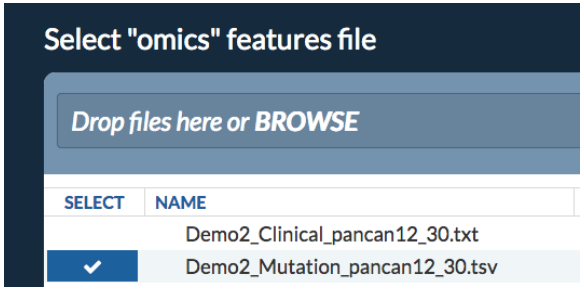
- Select "**Sample Clustering**" and Click on "**Start Pipeline**"



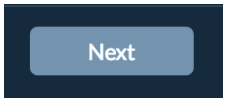
STEP3: Sample Clustering (network-guided)

Configure the pipeline:

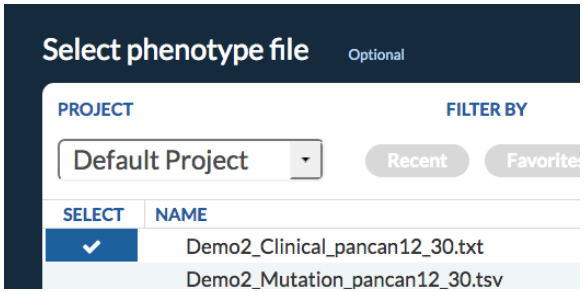
- For the "**omics**" file select:
  - Demo2_Mutation_pancan12_30



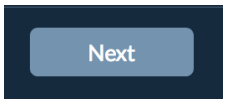
- Click "**Next**" at the bottom right corner



- For the "**phenotype**" file select:
  - Demo2_Clinical_pancan12_30

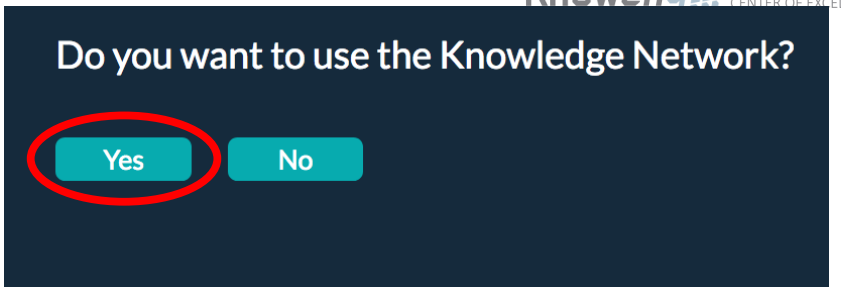


- Click "**Next**" at the bottom right corner

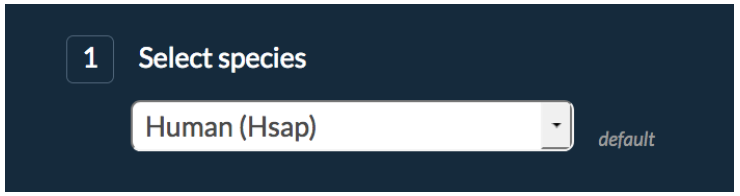


STEP3: Sample Clustering (network-guided)

- Select "**Yes**" in response to using the knowledge network:
  - This allows us to perform network-guided clustering

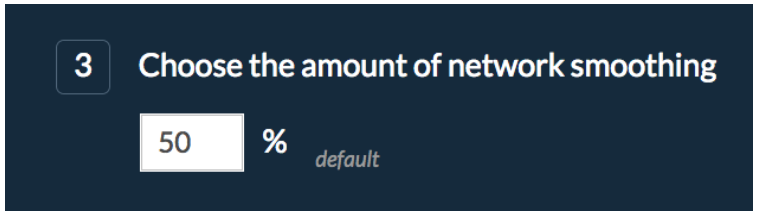


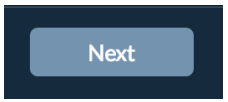
- Keep the species as "**Human**"



- Select "**HumanNet Integrated Network**" as the network

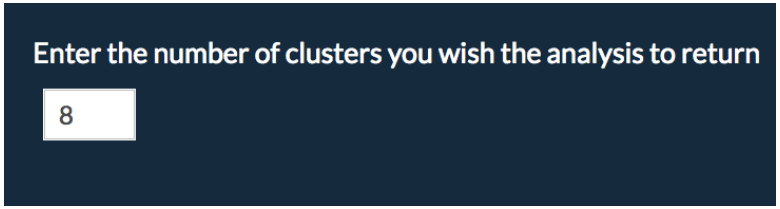
- Keep network smoothing at 50% and click **Next**:
  - This controls how much importance is put on network connections instead of the somatic mutations



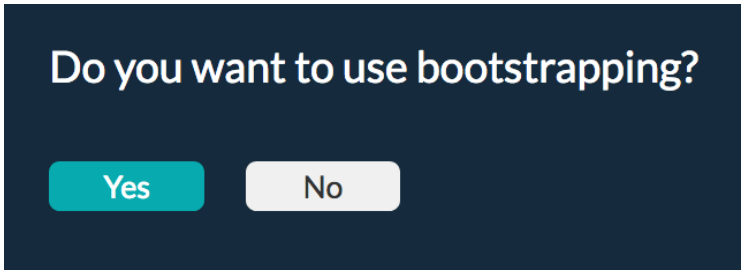


STEP3: Sample Clustering (network-guided)

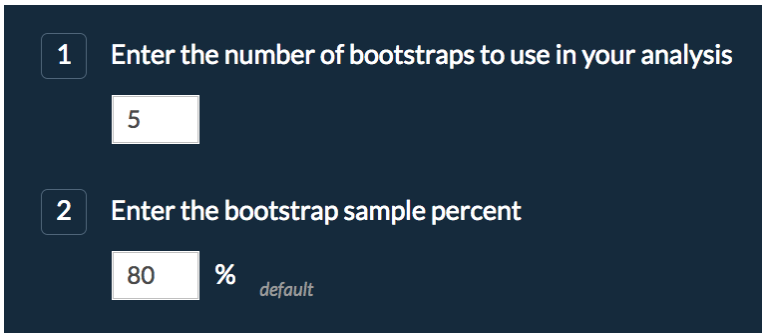
- Choose **8** as number of clusters and click **Next**



- Select "**Yes**" in response to using bootstrap sampling:
  - This allows us to obtain a more robust final clustering



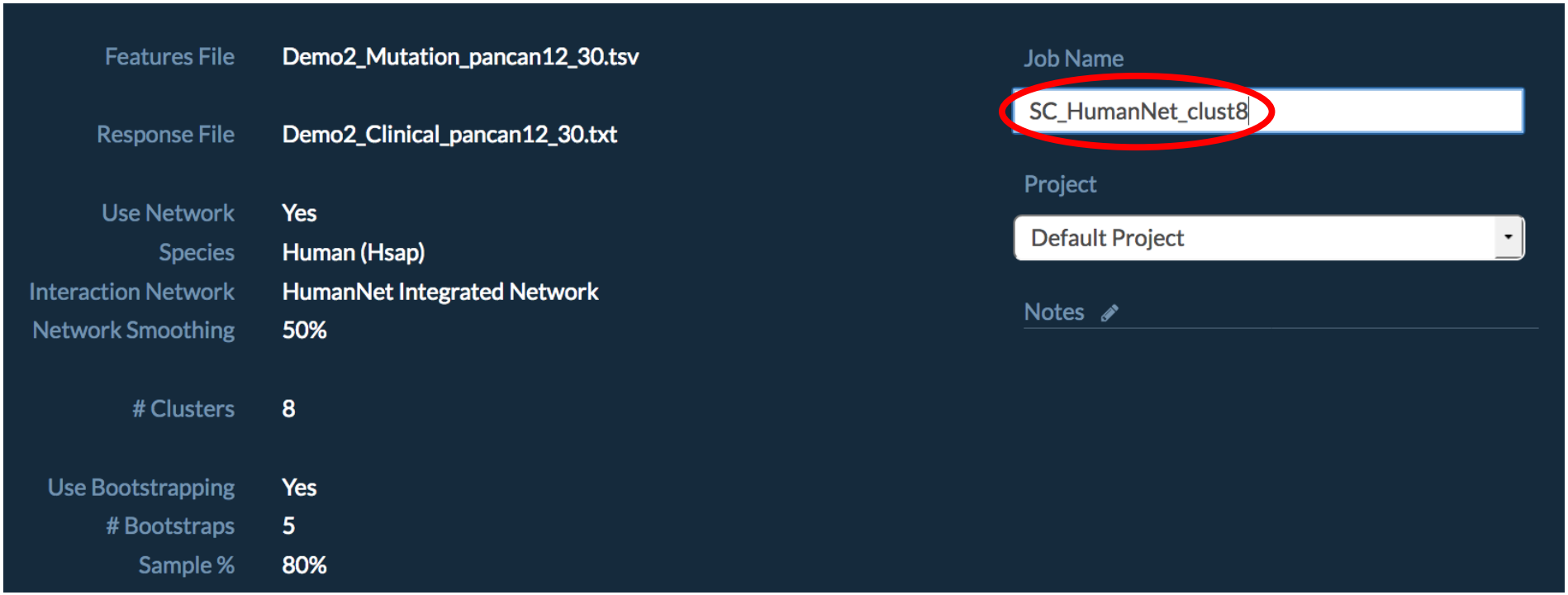
- Choose **5** as number of bootstraps



- We will use the default **80%** rate to sample the data in each bootstrap

STEP3: Sample Clustering (network-guided)

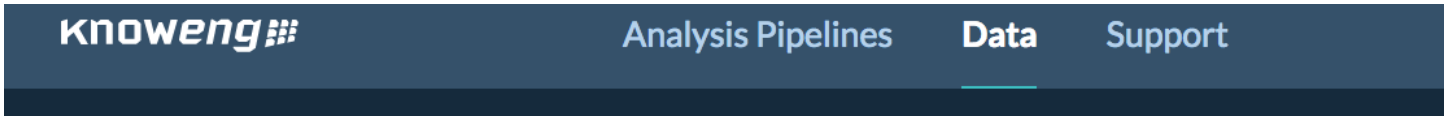
- Review the summary of the job and change the default "**Job Name**" to easily recognize later



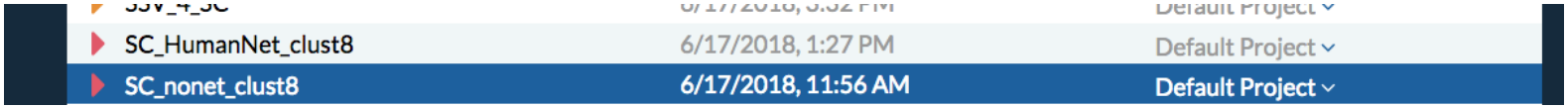
- Press **Submit Job**

STEP3: Sample Clustering (standard vs. network)

- Go to the "**Data**" page:



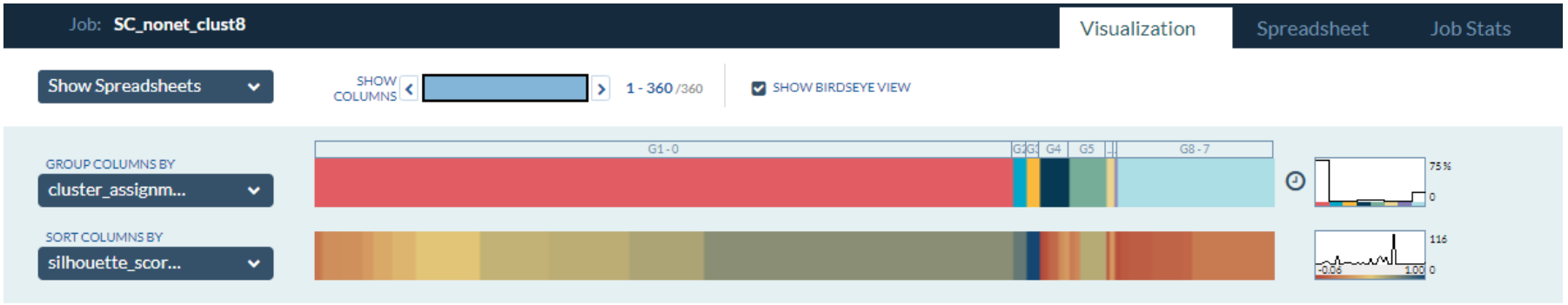
- Select "**SC_nonet_clust8**" (or any other name you chose)



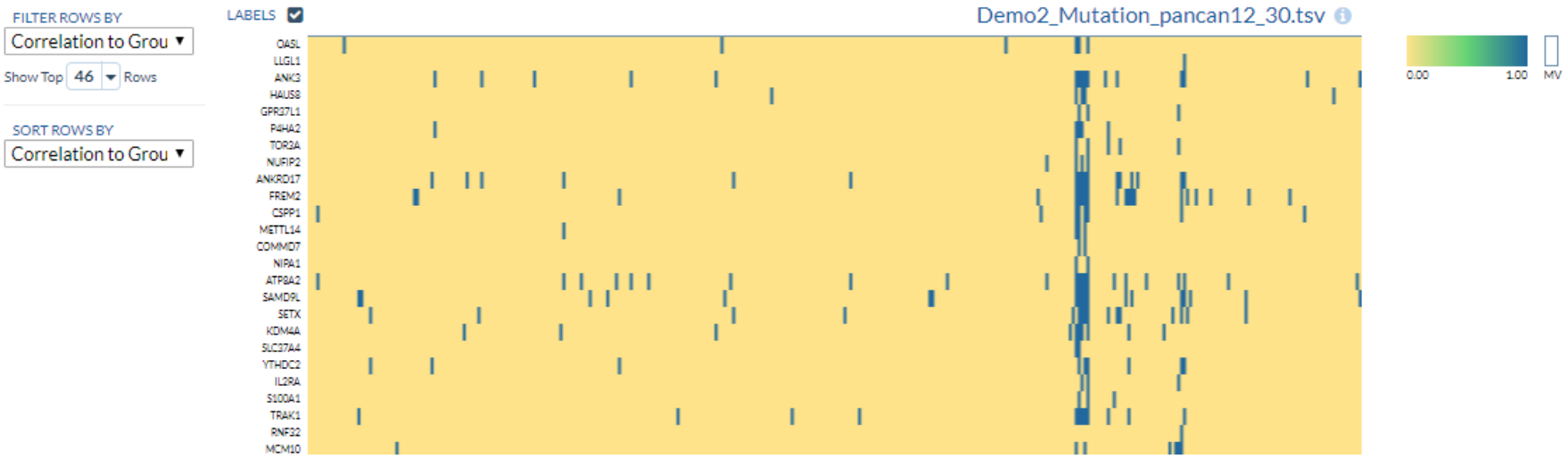
- Select "**View Results**" at the top right corner

STEP3: Sample Clustering (standard vs. network)

- Visualization shows the cluster sizes and the match of the samples to the cluster

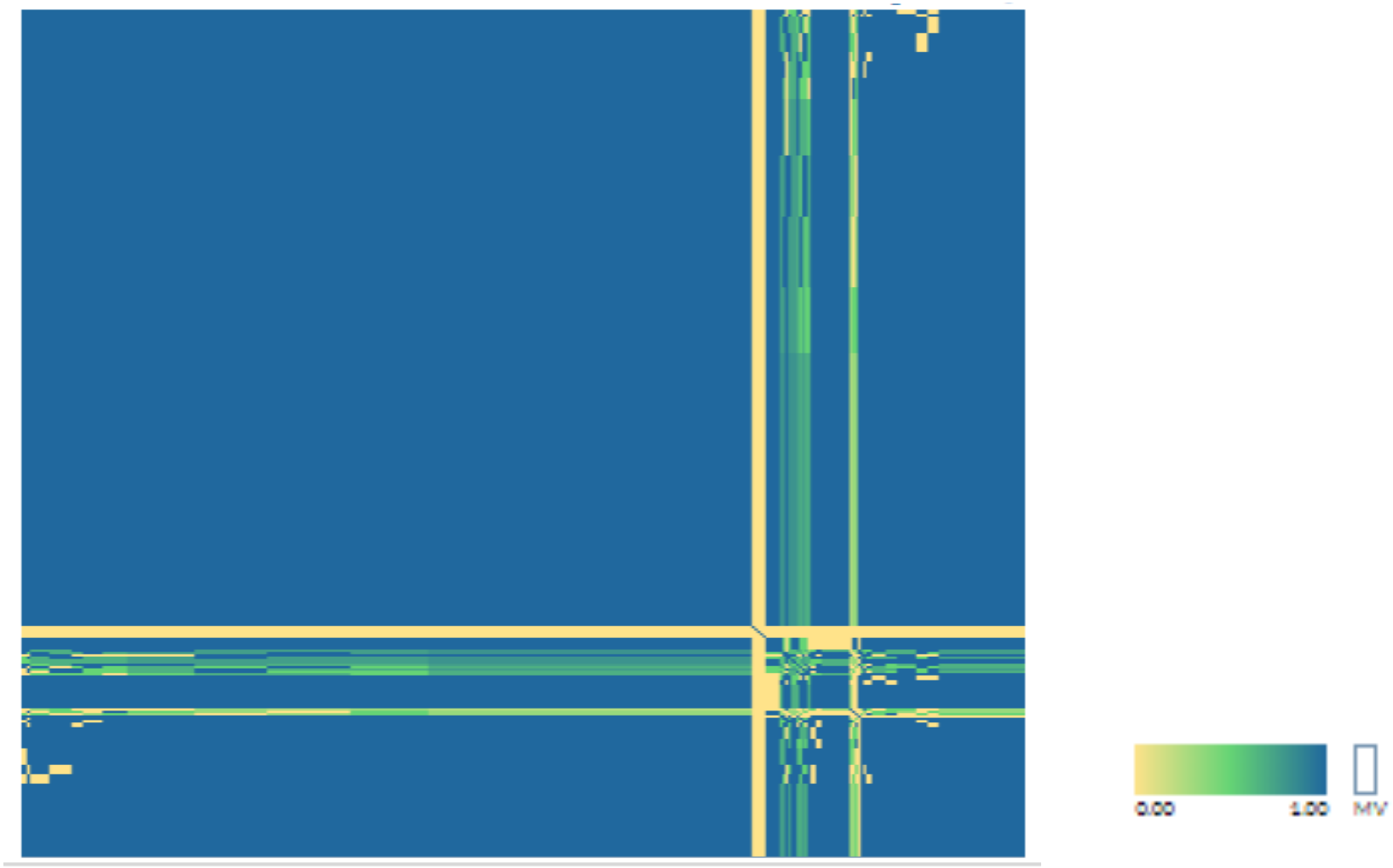


- Heatmap shows the features x samples – significantly correlated mutations



STEP3: Sample Clustering (standard vs. network)

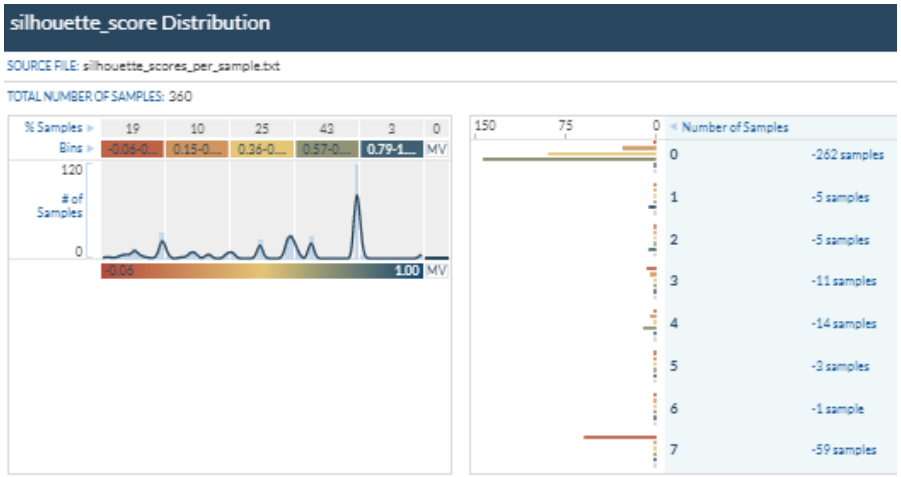
- Heatmap also shows samples x samples co-occurence



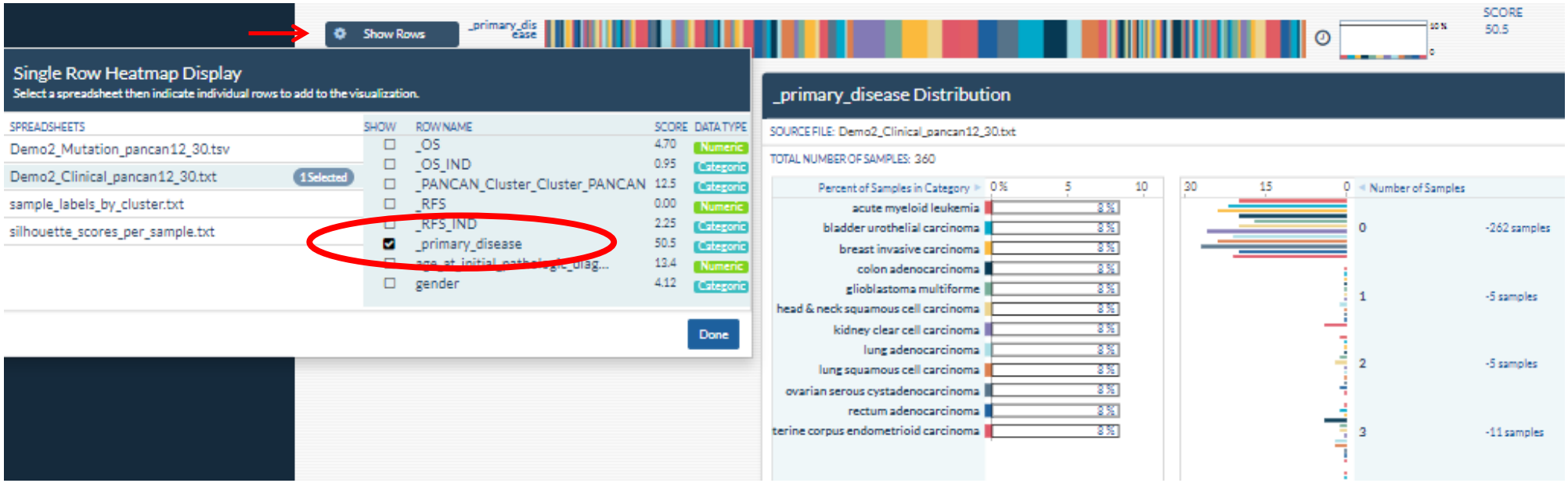
The color of each cell indicates how frequently a pair of patients fell within the same cluster across all samplings

STEP3: Sample Clustering (standard vs. network)

- High degree of clustering bias

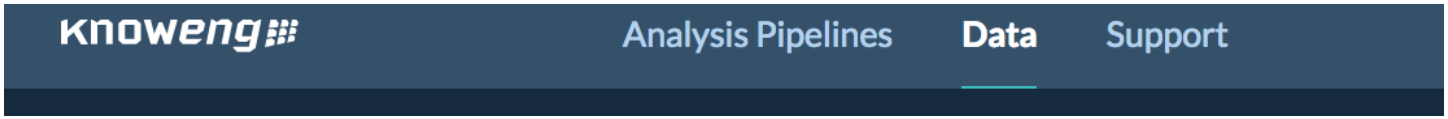


- You can add a phenotype to compare with with the "**Show Rows**"

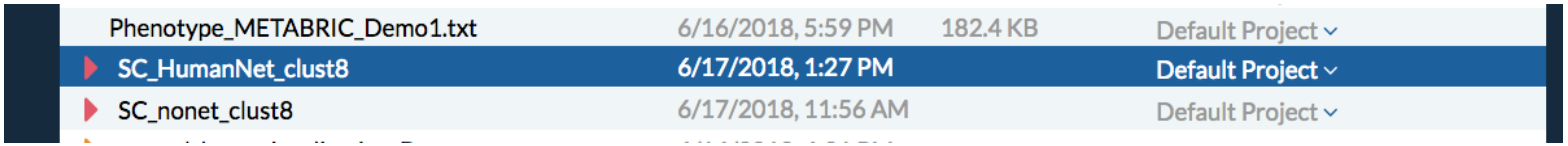


STEP3: Sample Clustering (standard vs. network)

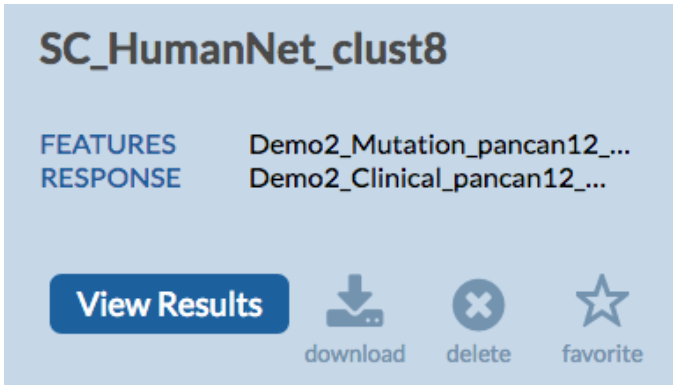
- Go to the "**Data**" page:



- Select "**SC_HumanNet_clust8**" (or any other name you chose)

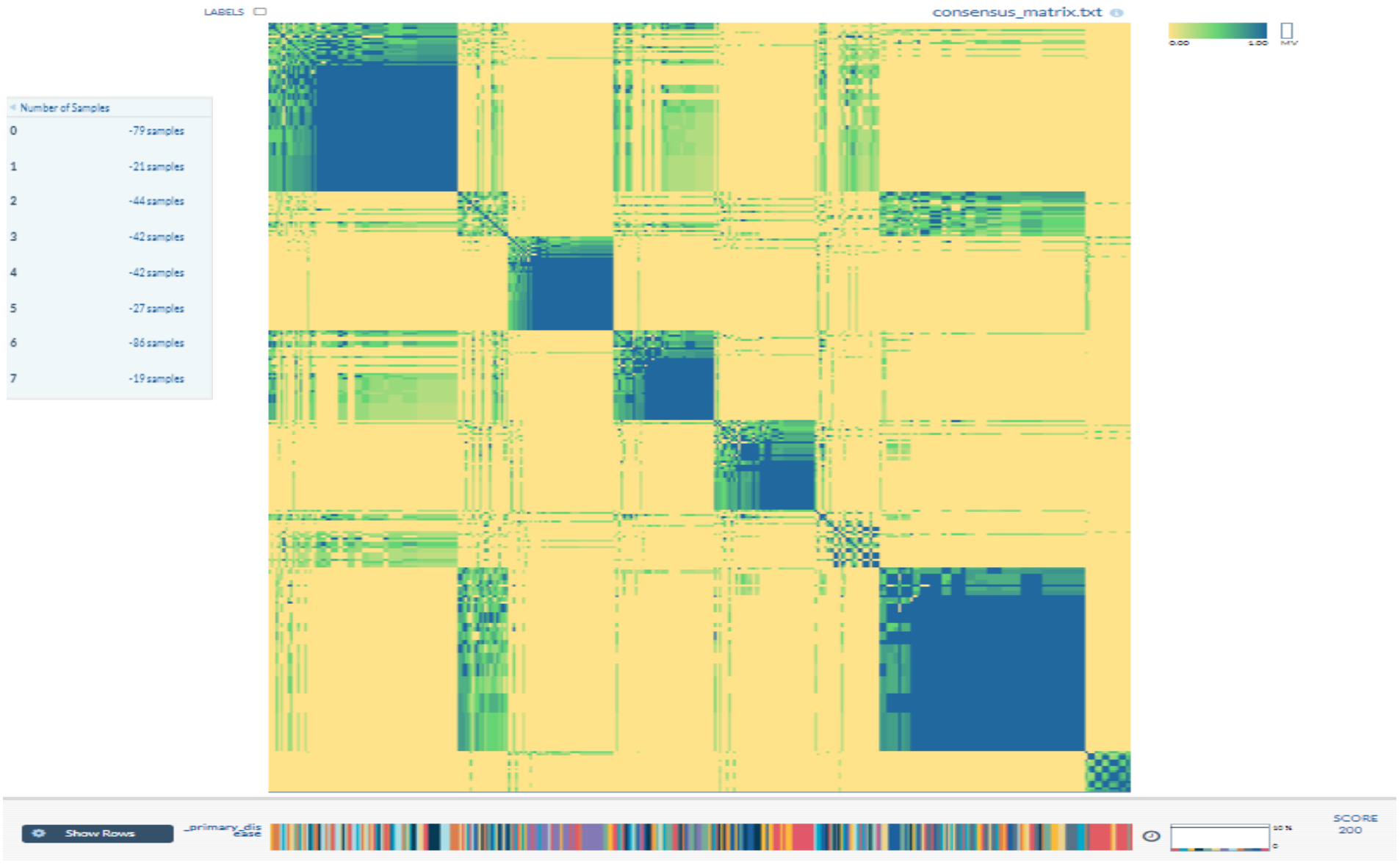


- Select "**View Results**" at the top right corner



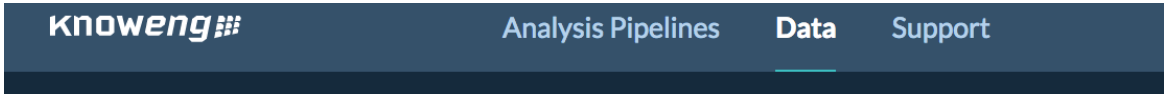
STEP3: Sample Clustering (standard vs. network)

- A more balanced clustering

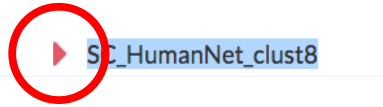


STEP3: Sample Clustering (standard vs. network)

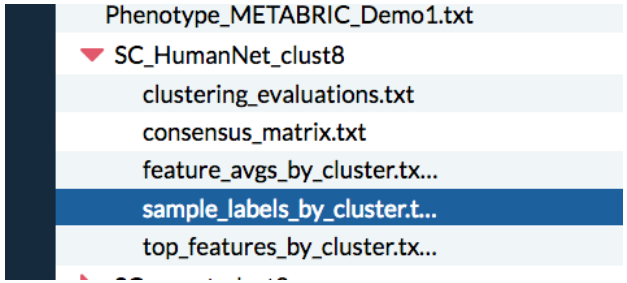
- Go to the “Data” page



- Click on triangle by “**SC_HumanNet_clust8**”



- Select “**sample_labels_by_cluster**”



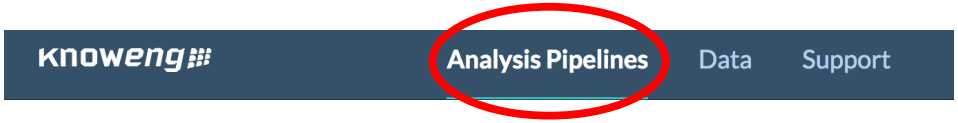
- Click on the name at the right top corner to edit and add “**_HumanNet**” to the end

- Repeat the same for “**SC_nonet_clust8**” and add “**_nonet**” to the end

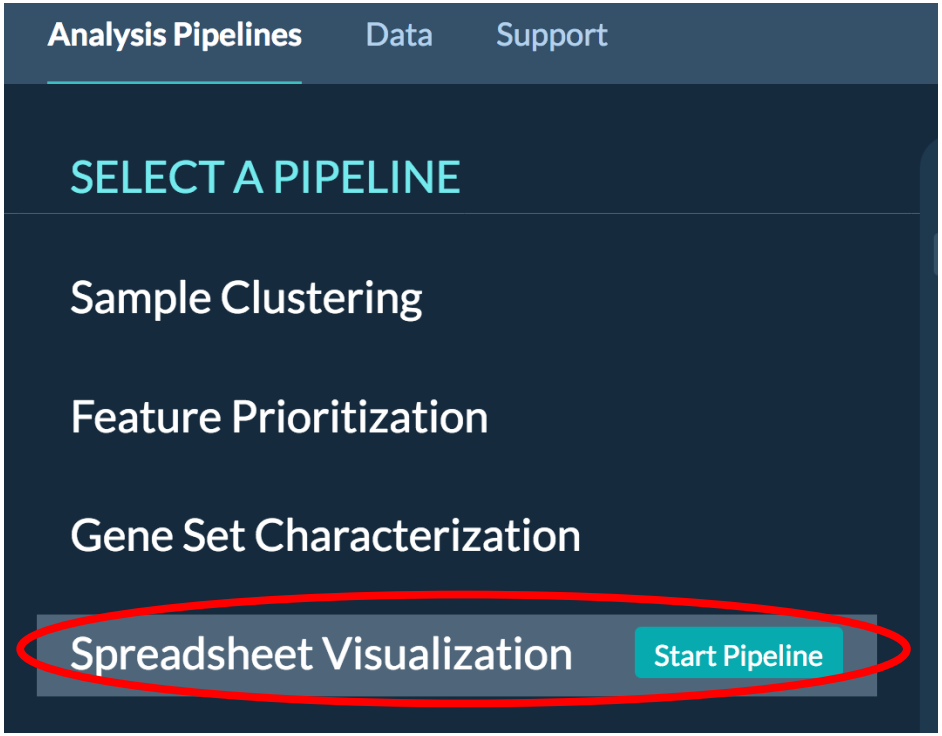
STEP3: Sample Clustering (standard vs. network)

Let's evaluate the results in SSV

- Select "**Analysis Pipelines**"

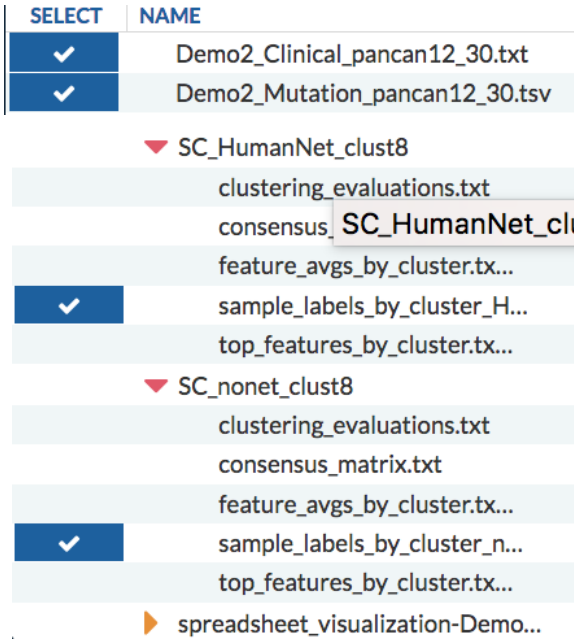


- Select "**Spreadsheet Visualization**" and Click on "**Start Pipeline**"

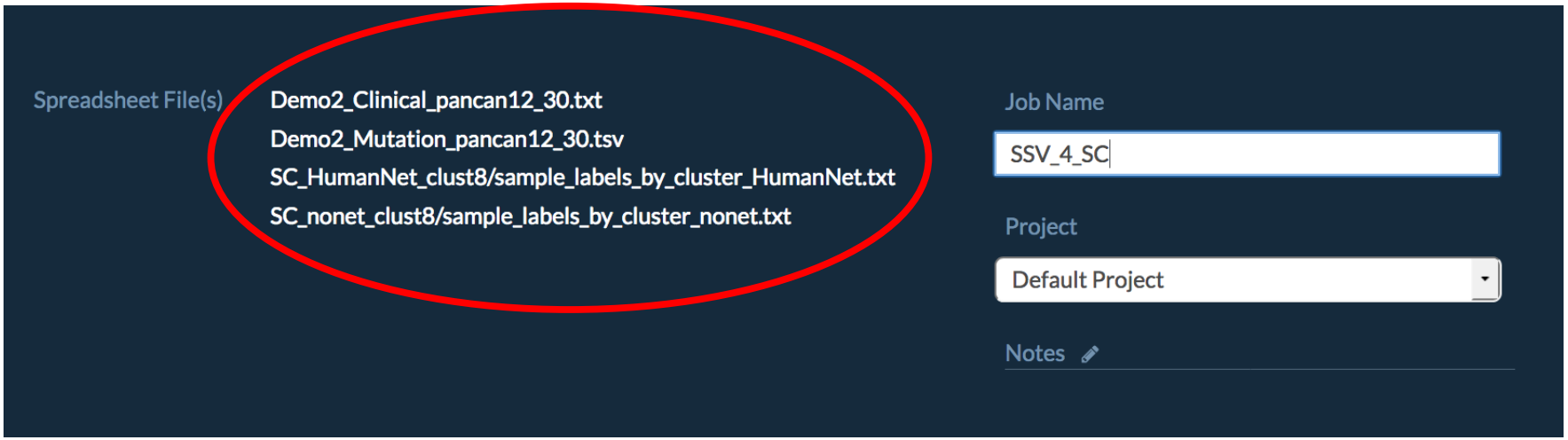


STEP3: Sample Clustering (standard vs. network)

- Select these four files to evaluate simultaneously and press **Next**:



- Check the summary and change the job name if you like. Press **Submit Job**.

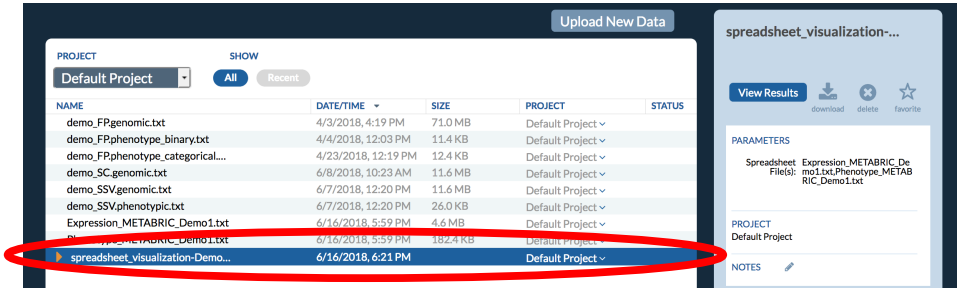


STEP3: Sample Clustering (standard vs. network)

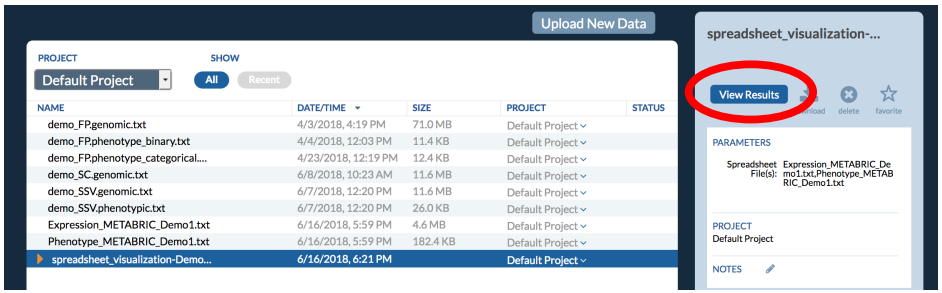
The results:

- Select "**Go to Data Page**"

- Select the job you just ran

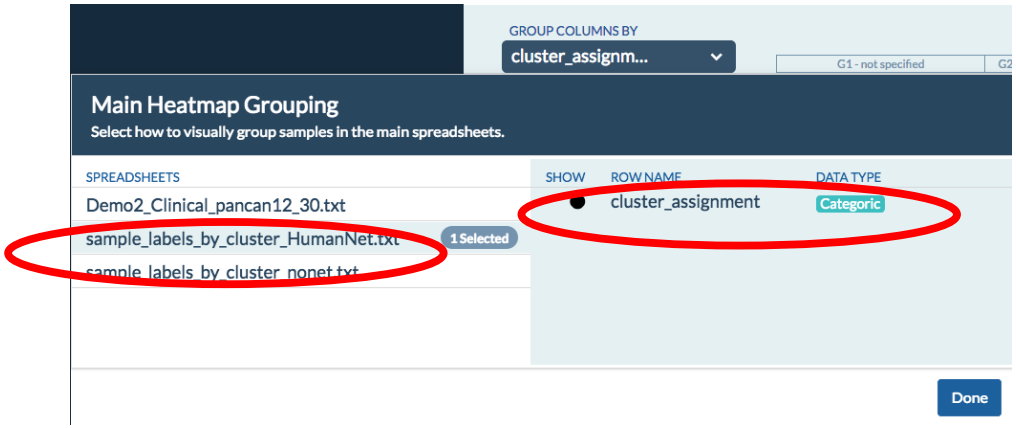


- Then "**View Results**"

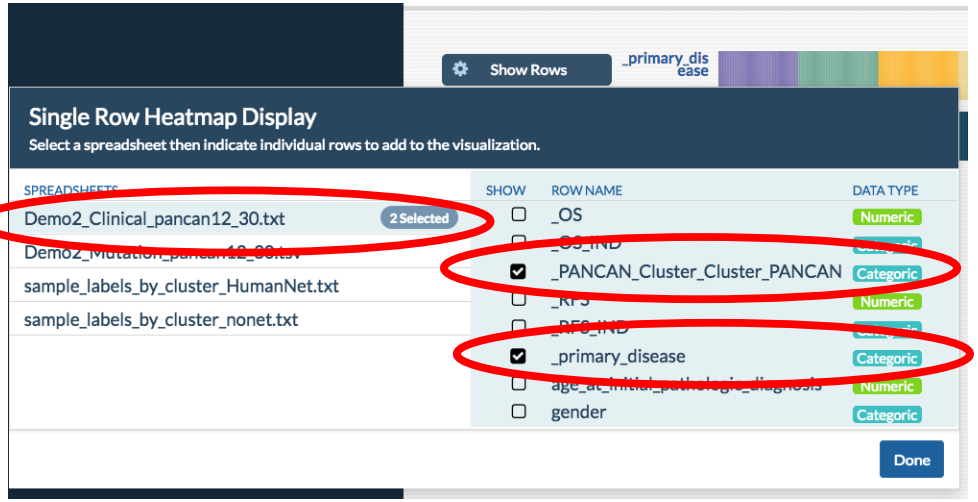


STEP3: Sample Clustering (standard vs. network)

- In "**Group Columns By**" select "**cluster_assignment**" from the "sample_labels_by_cluster_HumanNet.txt"

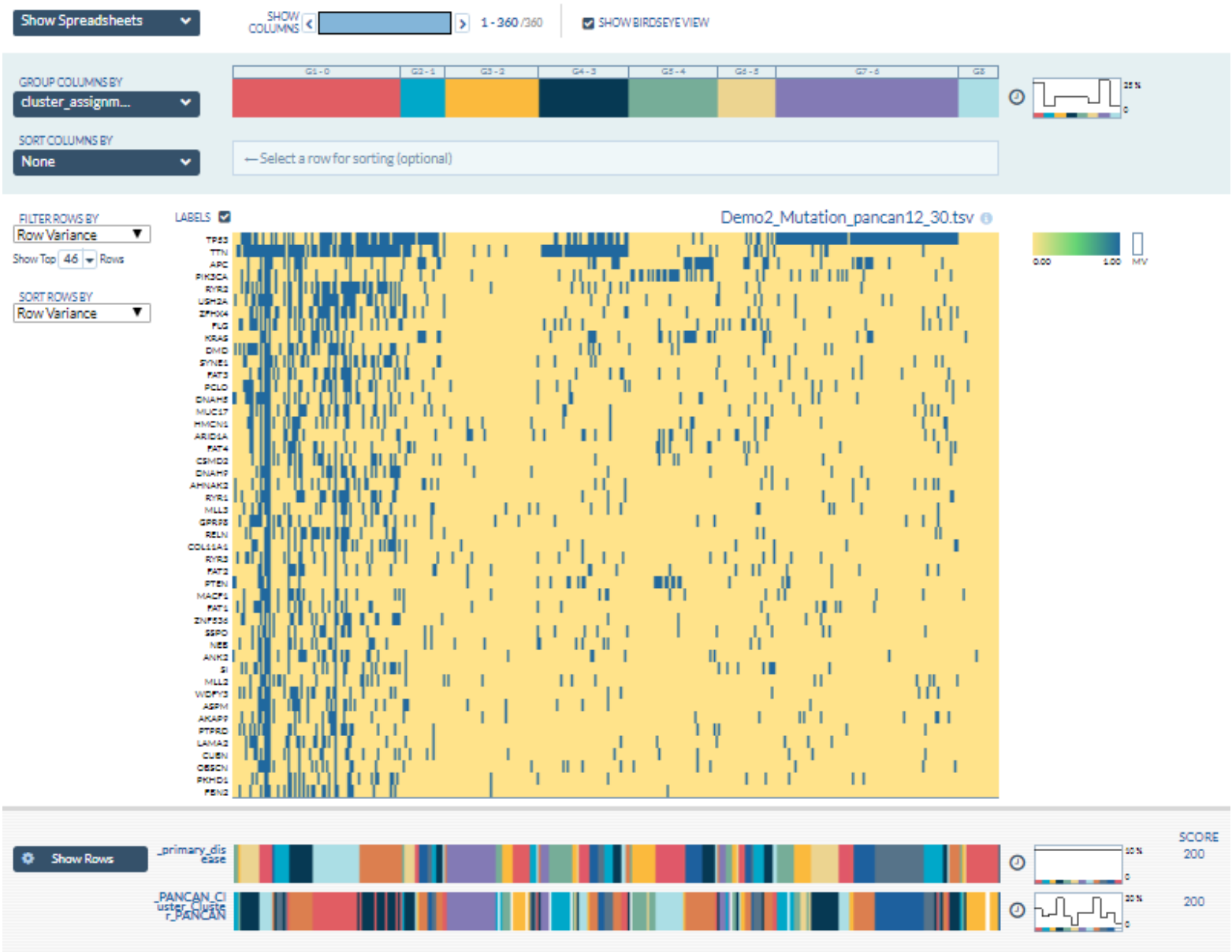


- By clicking on "**Show Rows**" add "**_primary_disease**" and "_PANCAN_Cluster_Cluster_PANCAN" from "Demo2_Clinical_pancan12_30.txt"



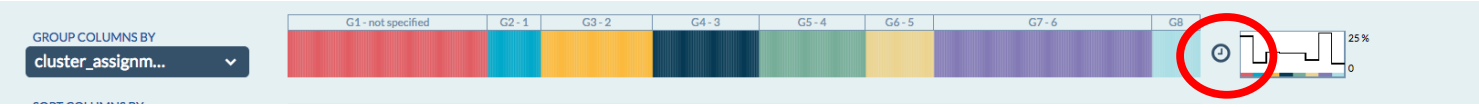
STEP3: Sample Clustering (standard vs. network)

- You can explore top genes, draw Kaplan Meier curves, etc.

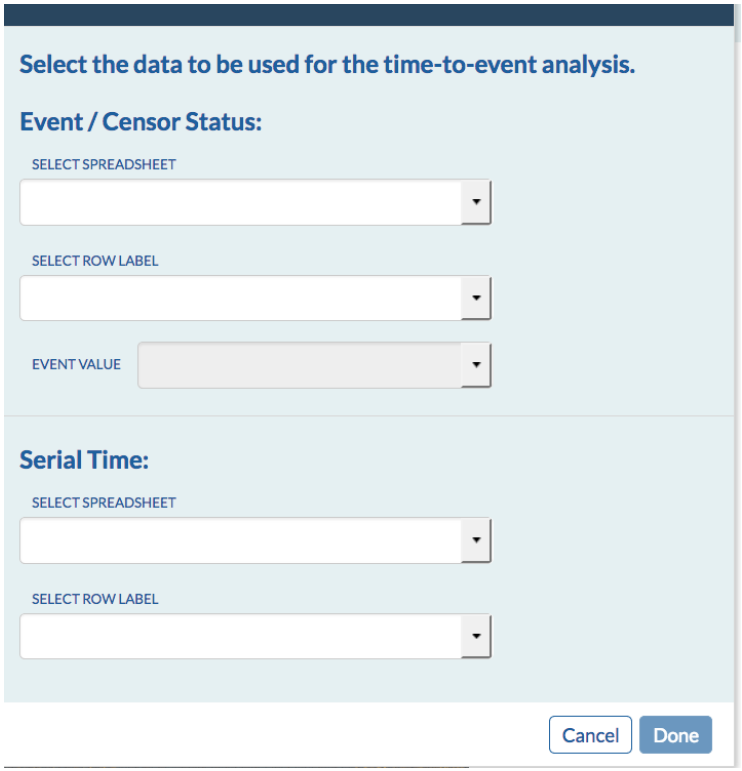


STEP3: Sample Clustering (standard vs. network)

- Click on the clock sign to perform Kaplan Meier survival analysis using any of the categories

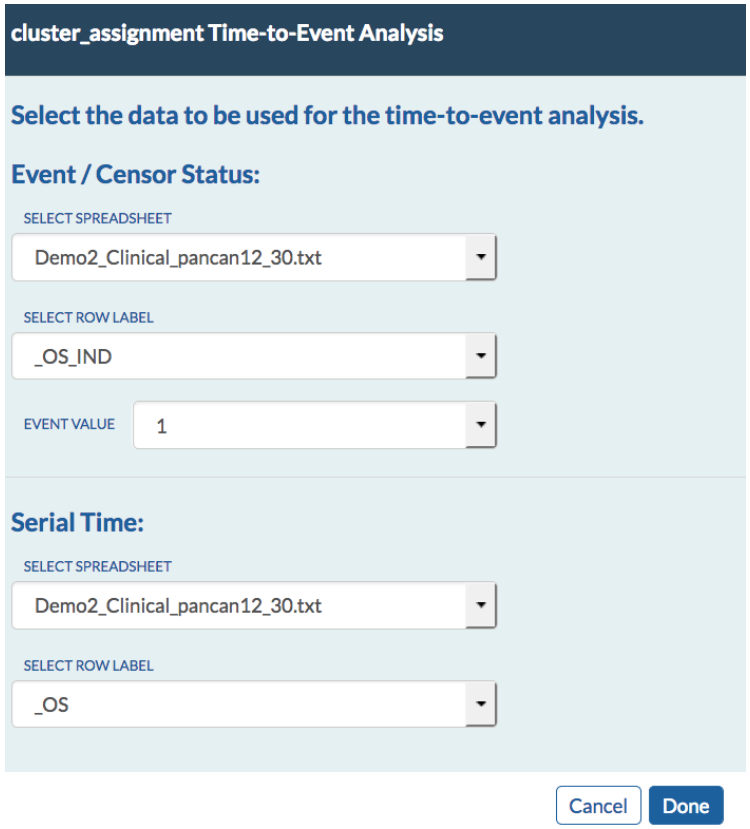


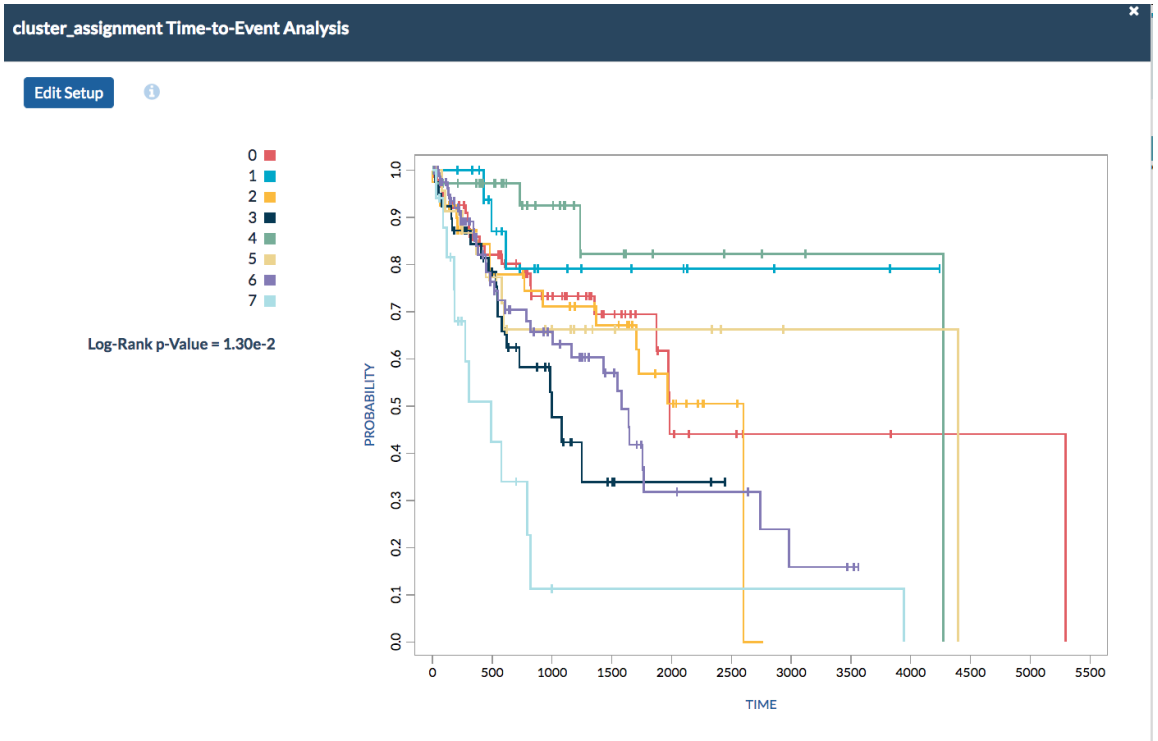
- Use this table to configure Kaplan Meier analysis by selecting the events and time to events



STEP3: Sample Clustering (standard vs. network)

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





Network-guided gene prioritization

STEP4: Gene Prioritization

- We will use KnowEnG's gene prioritization pipeline to perform network-guided 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[1], Cairns J[2], Kalari KR[3], Wang L[4], Sinha S[5].

- We will use samples from the CCLE dataset

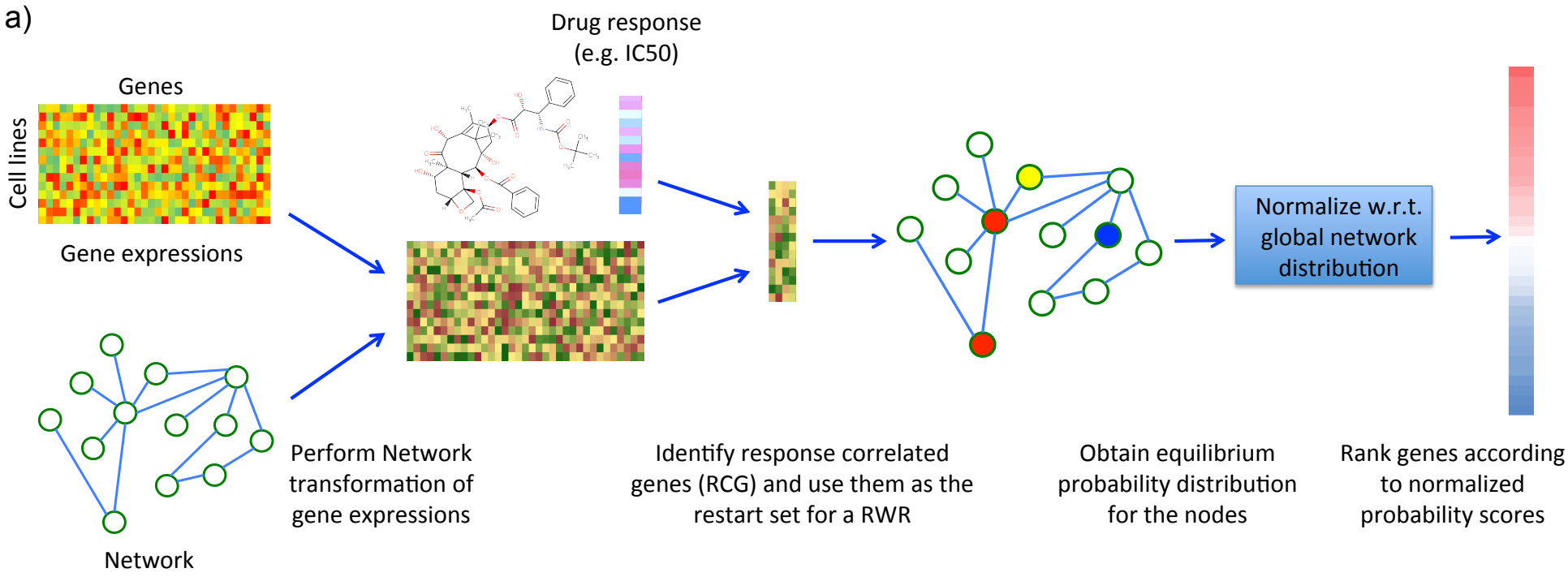
Nature. 2012 Mar 28;483(7391):603-7. doi: 10.1038/nature11003.

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

Barretina J[1], 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, Palescandolo E, Gupta S, Mahan S, Sougnez 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.

STEP4: Gene Prioritization

Outline of ProGENI:



STEP4: Gene Prioritization

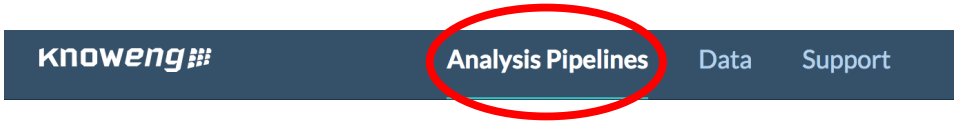
Dataset characteristics:

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.

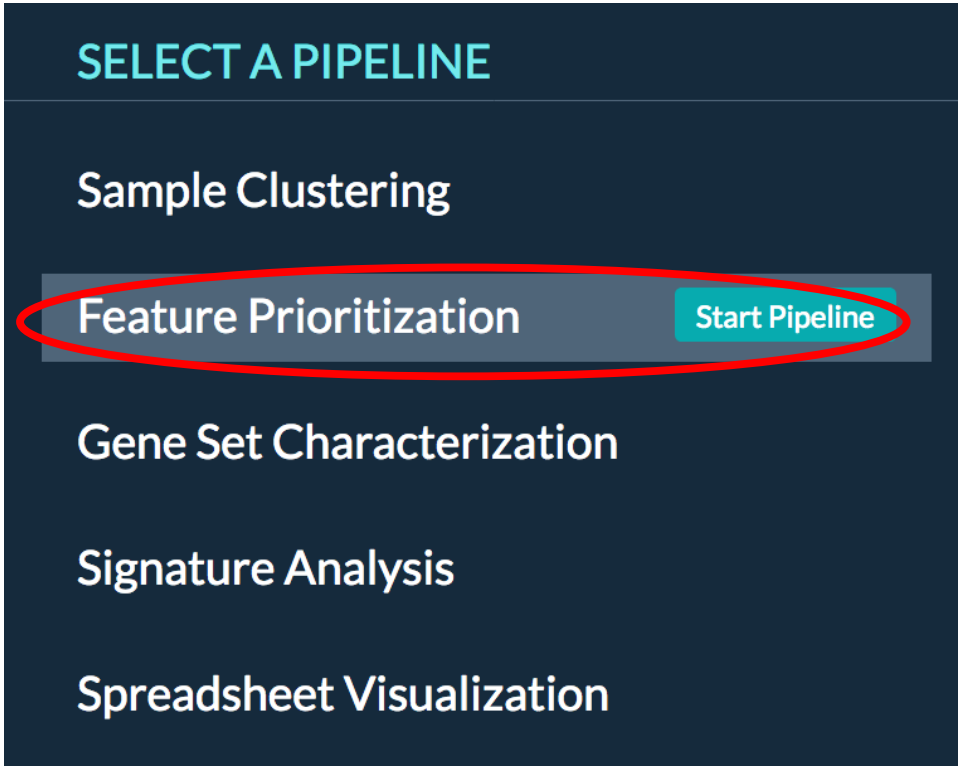
STEP4: Gene Prioritization (network-guided)

Select the pipeline:

- Select "**Analysis Pipelines**" at the top of the page



- Select "**Feature Prioritization**" and Click on "**Start Pipeline**"

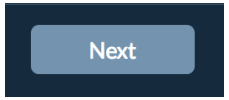


STEP4: Gene Prioritization (network-guided)

Configure the pipeline:

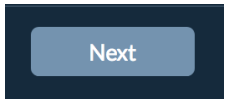
- For the "**omics**" file select "**Use Demo Data**"

- Click "**Next**" at the bottom right corner



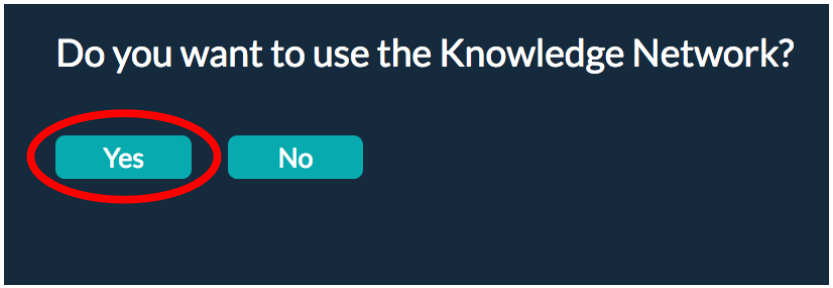
- For the "**response**" file select "**Use Demo Data**"

- Click "**Next**" at the bottom right corner

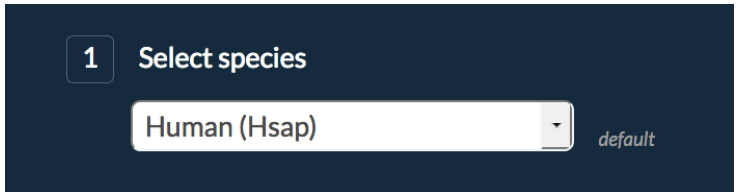


STEP4: Gene Prioritization (network-guided)

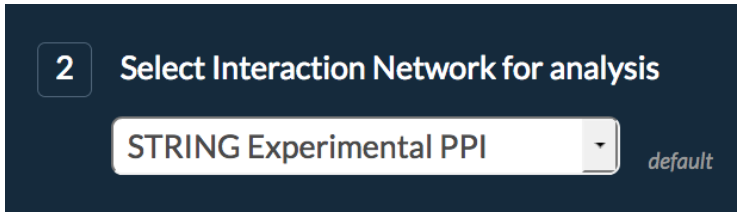
- Select "**Yes**" in response to using the knowledge network:
  - This allows us to perform network-guided prioritization (ProGENI)



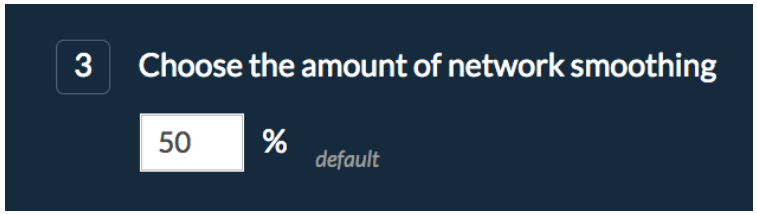
- Keep the species as "**Human**"



- Select "**STRING Experimental PPI**" as the network

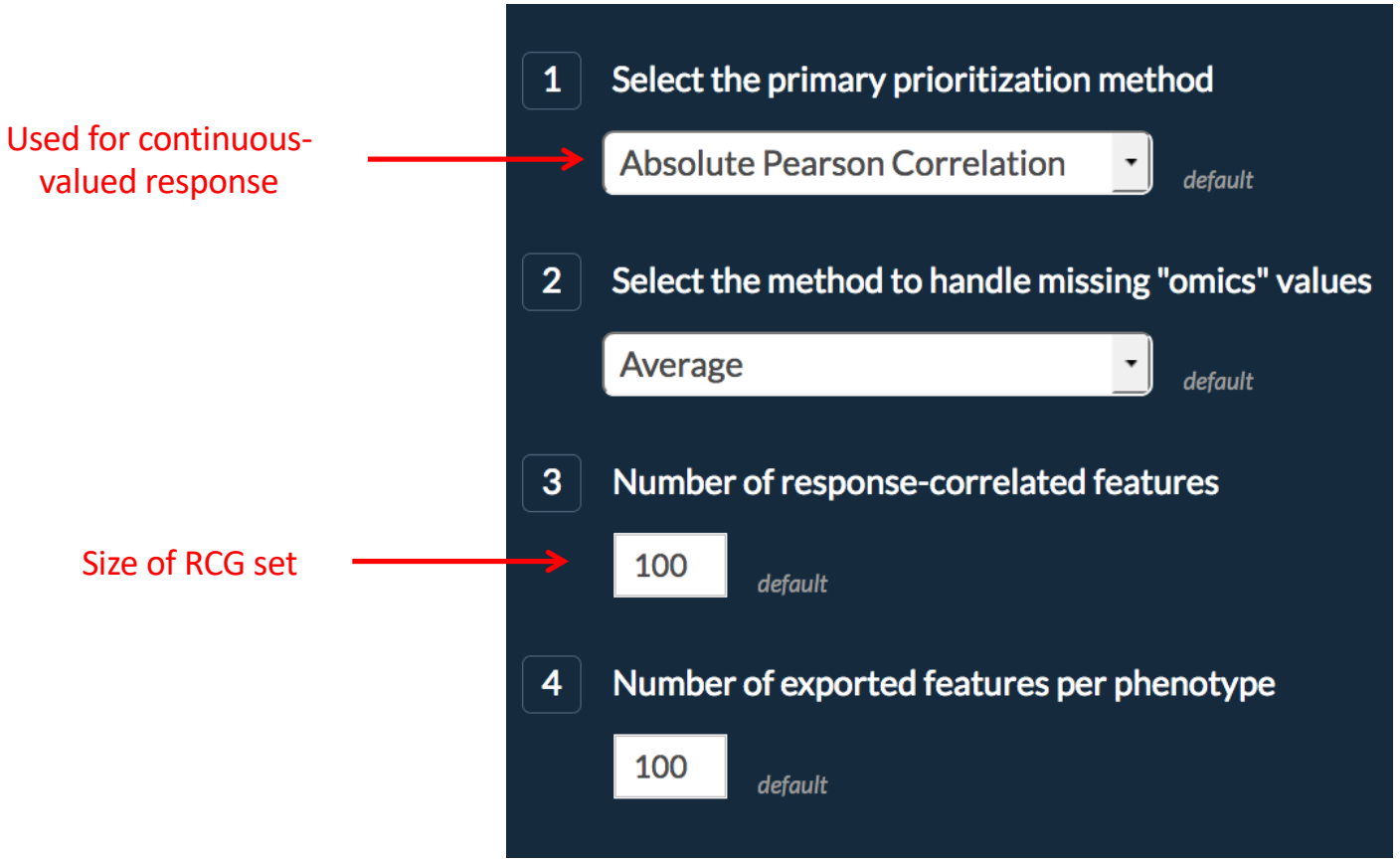


- Keep network smoothing at 50%:
  - This controls how much importance is put on network connections instead of the somatic mutations

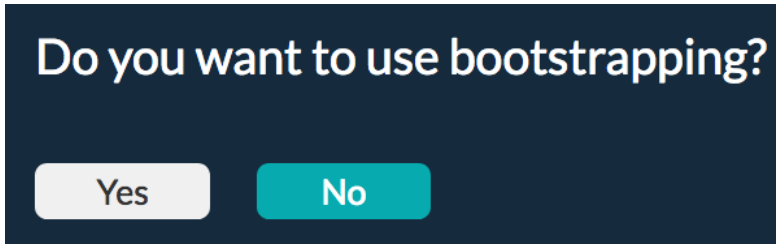


STEP4: Gene Prioritization (network-guided)

- Keep the default parameters on this page

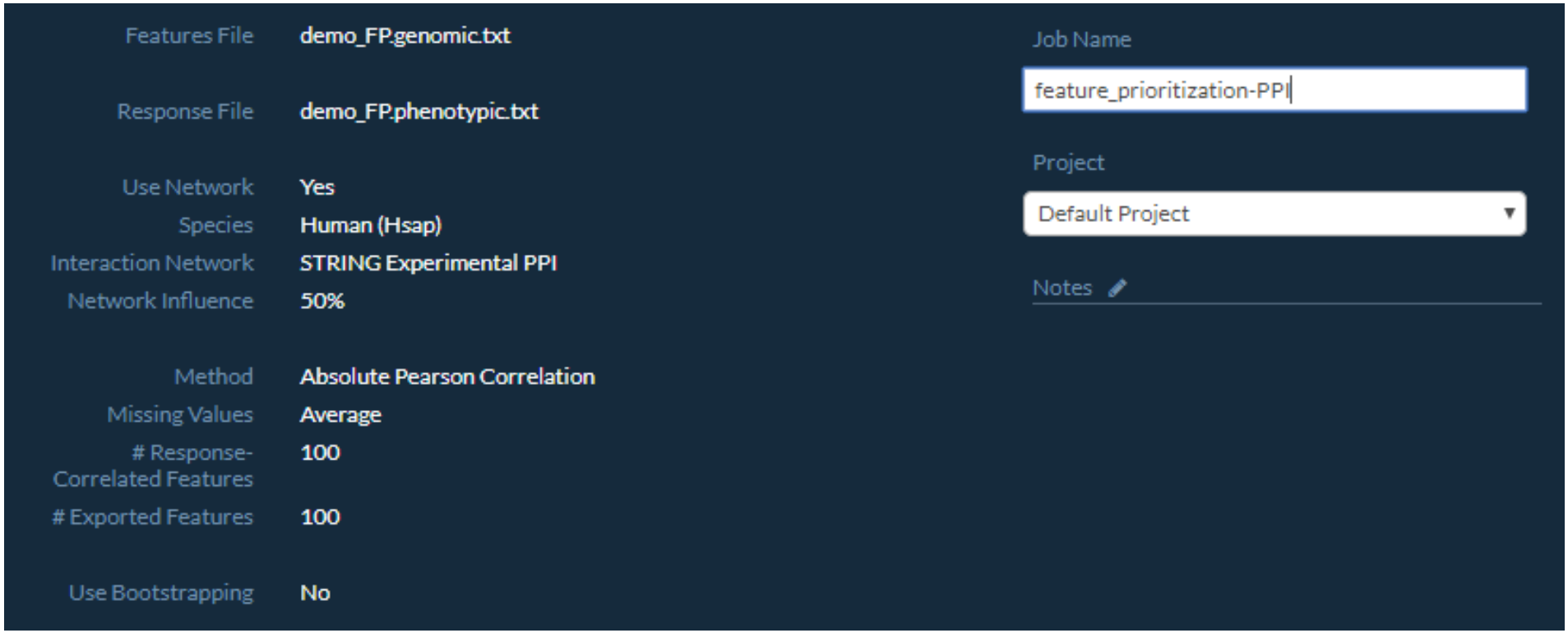


- Choose "**No**" for bootstrapping



STEP4: Gene Prioritization (network-guided)

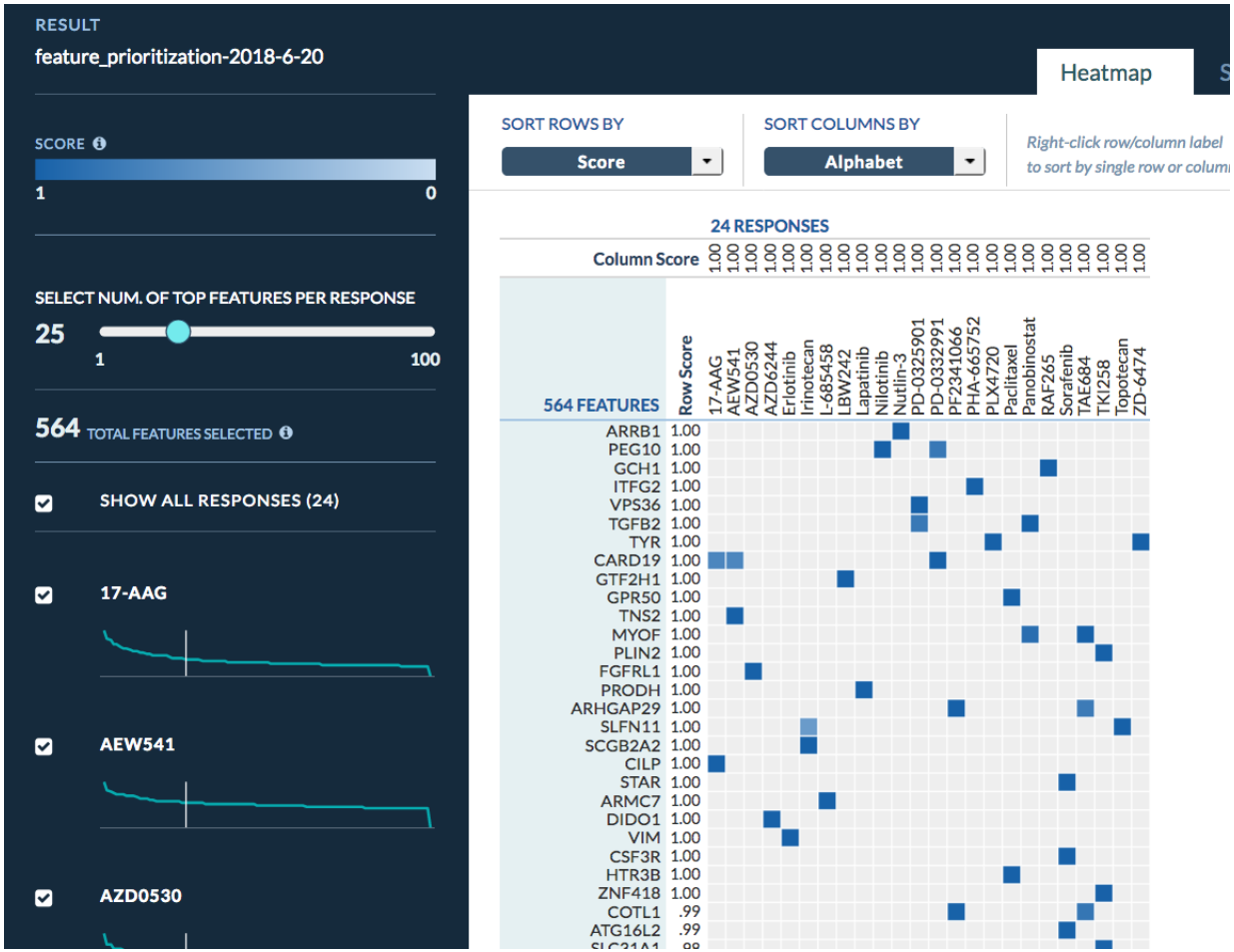
- Review the summary of the job and change its name if you like



- Submit the job

STEP4: Gene Prioritization (network-guided)

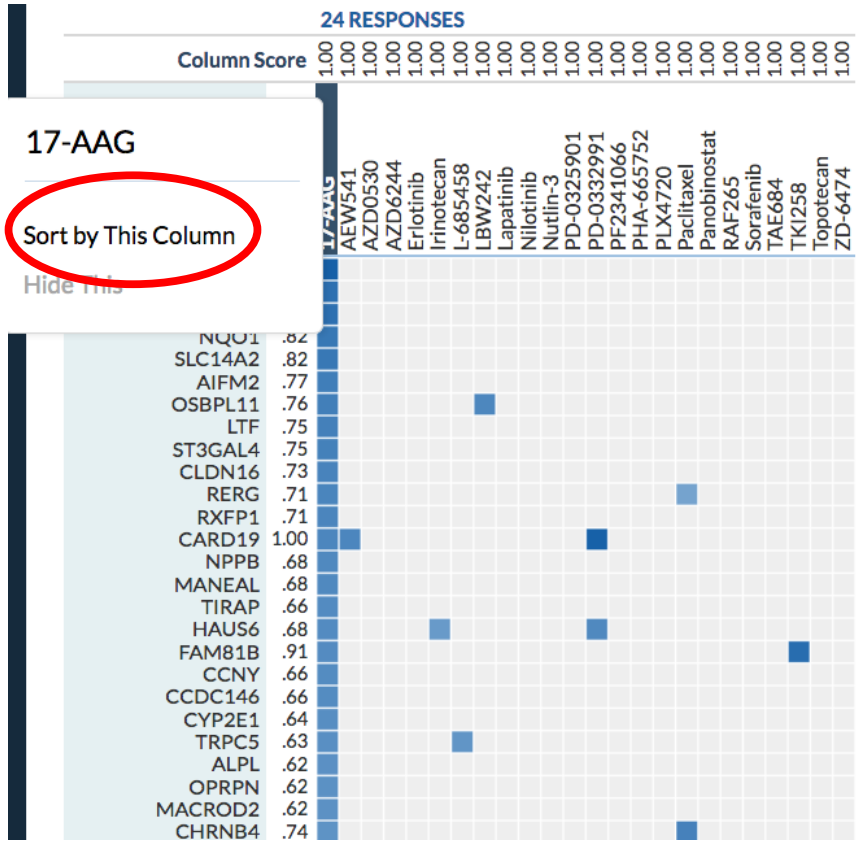
- Go to the Data page
- Select "**View Results**" when the job is done



Heatmap shows the top genes identified for each drug

STEP4: Gene Prioritization (network-guided)

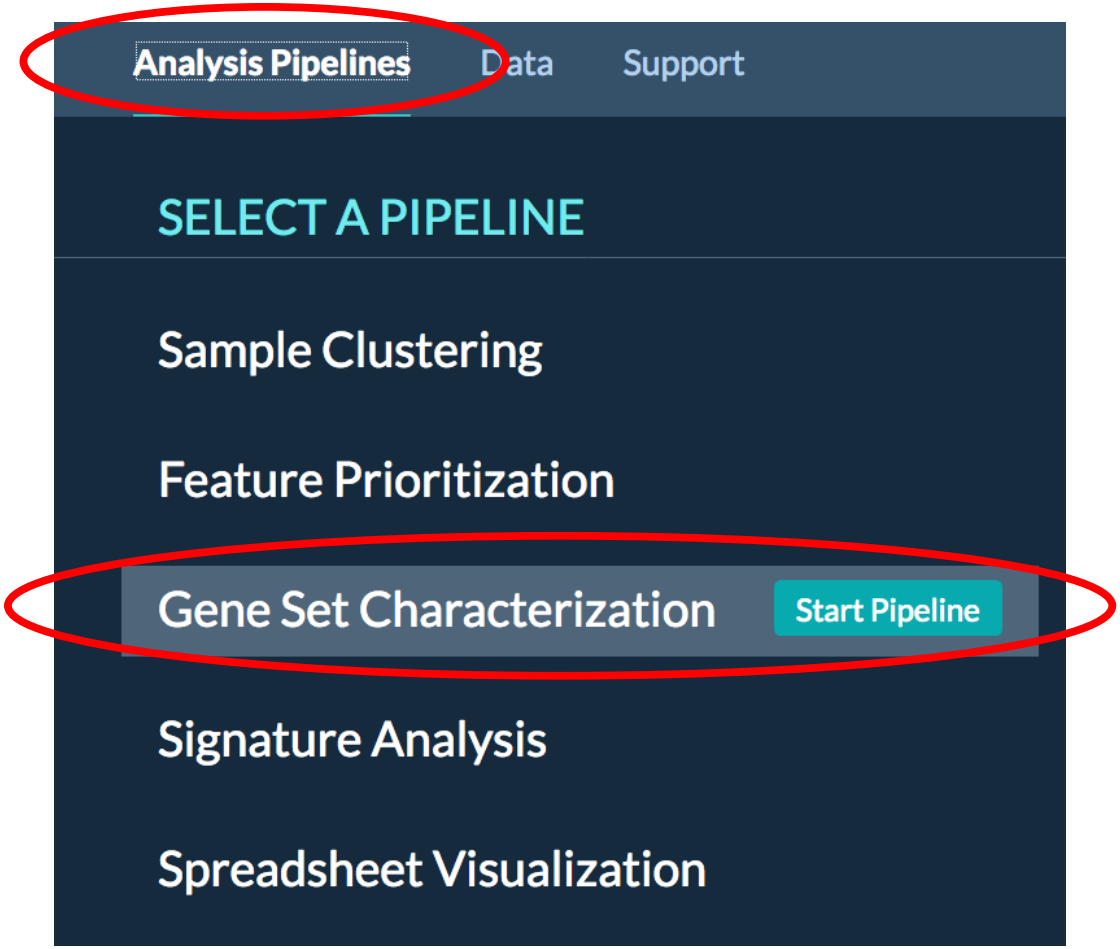
- You can “right-click” on a drug to sort rows it and see its top genes



- You can also sort columns by a gene to see drugs for which the gene was among the top list

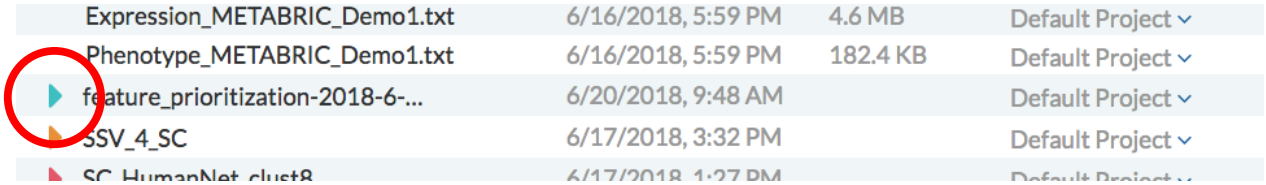
STEP4: Gene Prioritization (network-guided)

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

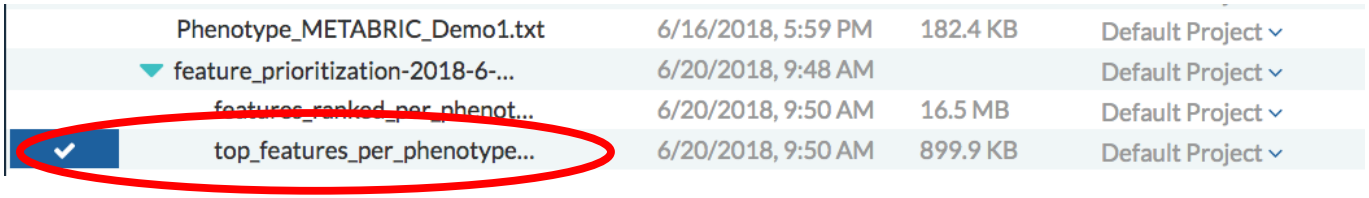


STEP4: Gene Prioritization (network-guided)

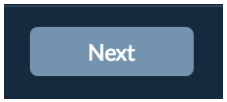
- Select the green triangle by the gene prioritization job you ran



- Select “top_features_per_phenotype_matrix”

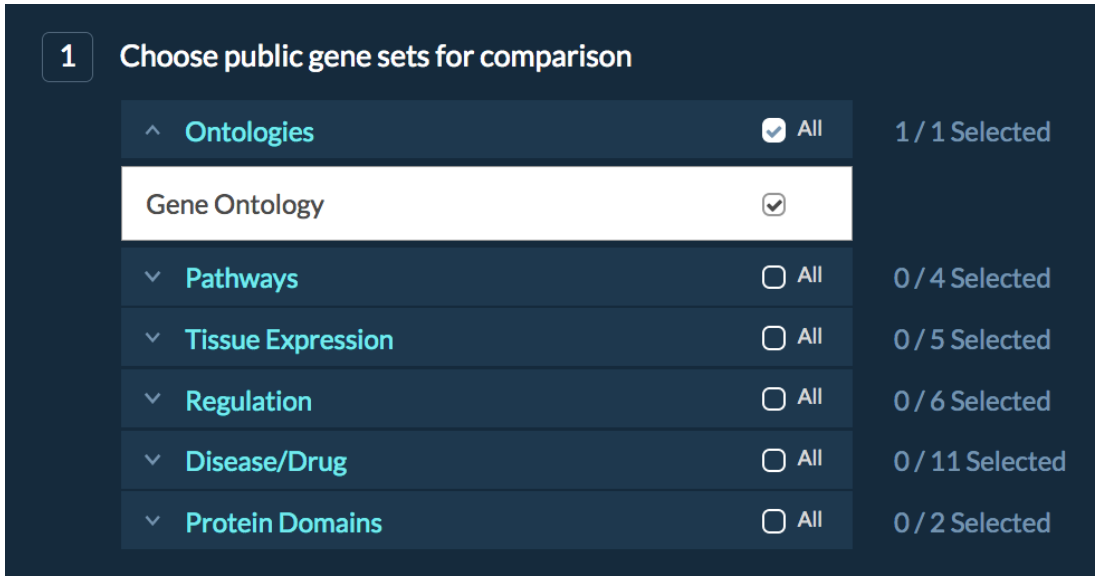


- Press Next

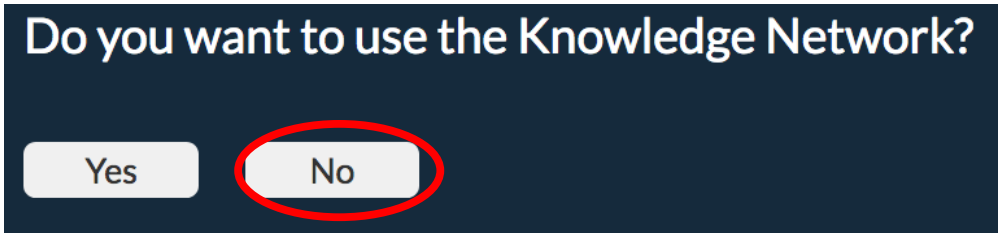


STEP4: Gene Prioritization (network-guided)

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



- Say "**No**" to using the knowledge network and press **Next**. Then press **Submit Job.**

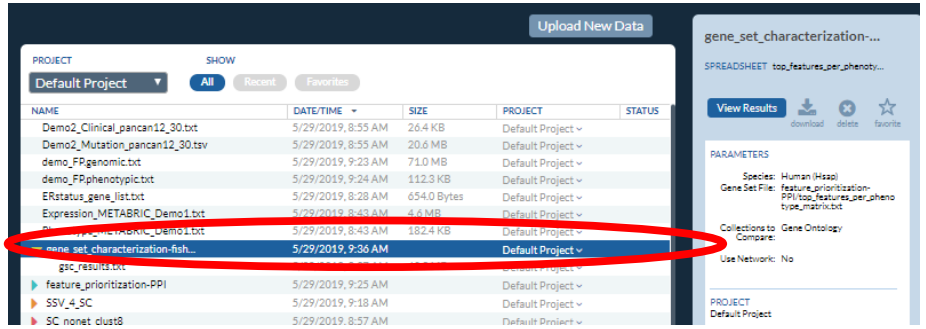


STEP4: Gene Prioritization (network-guided)

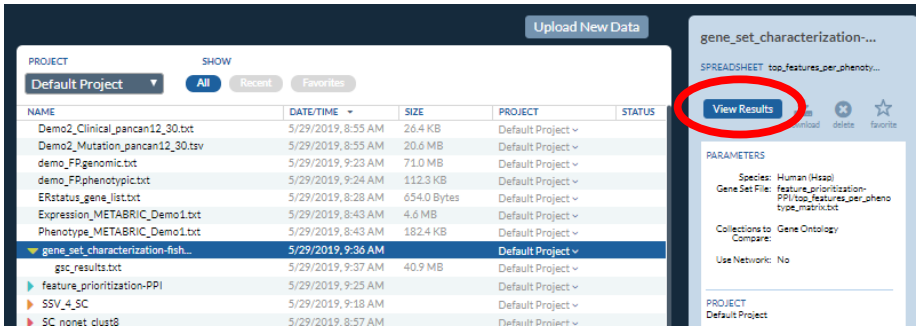
The results:

- Select “**Go to Data Page**”

- Select the job you just ran

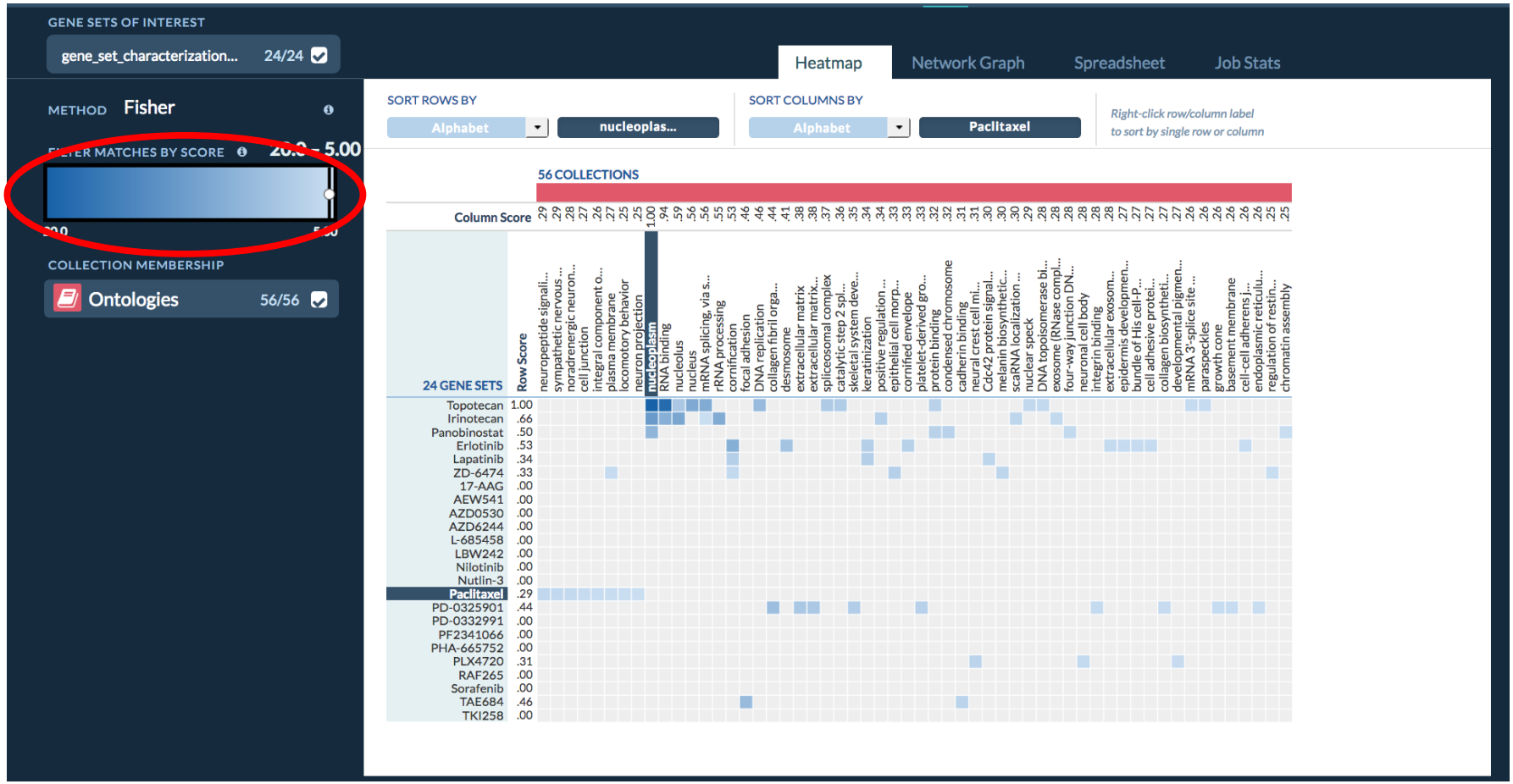


- Then “**View Results**”



STEP4: Gene Prioritization (network-guided)

- This page shows the enriched gene sets for each drug
- You can change the filter (scores represent –log10 (p-value) of enrichment) to see fewer or more enriched gene sets



Resources

- Tutorials:
  - Quickstarts: https://knoweng.org/quick-start/
  - YouTube: https://www.youtube.com/channel/UCjyIIolCaZIGtZC20XLBOyg
- Resources:
  - Data Preparation Guide: 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
- Source Code:
  - Docker Images: https://hub.docker.com/u/knowengdev/
  - Github Repos: https://knoweng.github.io/
- Other Cloud Platforms
  - https://cgc.sbgenomics.com/public/apps#q?search=knoweng
- Research
  - TCGA Analysis Paper: https://www.biorxiv.org/content/10.1101/642124v1
  - TCGA Analysis Walkthrough: https://github.com/KnowEnG/quickstart-demos/tree/master/publication_data/blatti_et_al_2019
- Contact Us with Questions and Feedback: knoweng-support@illinois.edu