

PROMO 2017a - Tutorial

| | |
|--|----|
| Introduction | 2 |
| Installing PROMO..... | 2 |
| Step 1 - Importing data | 2 |
| Step 2 - Preprocessing | 6 |
| Step 3 – Data Exploration..... | 9 |
| Step 4 – Clustering..... | 13 |
| Step 5 – Analysis of sample clusters | 14 |
| Step 6 – Export..... | 24 |

Introduction

Modern genomic datasets may include hundreds of samples, each measured by several high-throughput technologies and described by extensive clinical information. PROMO (Profiler of Multi-Omics data) is an interactive Matlab-based tool, designed to analyze large versatile datasets. It enables importing multi-label datasets from various file formats, data exploration and visualization, applying unsupervised analysis on both samples and features and utilizing various popular statistical tests including survival analysis. Special features that are specific to multi-omic datasets include dataset integration and joint multi-omic clustering.

Installing PROMO

Installation requirements:

Windows 7 Service Pack 1 and above

At least 4 GB RAM

Any Intel or AMD x86-64 processor

Installation instructions:

Download and Install Matlab Runtime 9.0.1 for Windows using the following link:

<https://www.mathworks.com/products/compiler/mcr.html>

Download and run PROMO_2017a_Installer.exe using the following link:

http://acgt.cs.tau.ac.il/promo/releases/PROMO_2017a_Installer.exe

- Installation path requires read and write permissions.
- PROMO may require a few minutes to load during the first run.

Step 1 - Importing data

Data can be imported using the File menu. You can import data in several different formats:

1. Genomic matrix and Phenome files from UCSC Xena.
2. Tabular file containing a tab delimited table in Expander format.
3. UCSC Cancer Browser zip files.

In this example we will use **Genomic matrix and Phenome files from UCSC Xena**.

1. Enter UCSC Xena using the following link:
<https://xenabrowser.net/datapages/>
2. Click on the dataset you wish to query. For example, TCGA breast cancer:

- ① TCGA Bile Duct Cancer (CHOL) [Visualize](#)
- ① TCGA Bladder Cancer (BLCA) [Visualize](#)
- ① TCGA Breast Cancer (BRCA) [Visualize](#)
- ① TCGA Cervical Cancer (CESC) [Visualize](#)
- ① TCGA Colon and Rectal Cancer (COADREAD) [Visualize](#)

3. In the next page you can see a list of datasets. Click on the dataset that contains the relevant genomic data file:

exon expression RNAseq

- ① exon expression RNAseq (polyA+ IlluminaHiSeq) (n=1,218) TCGA hub

The exon expression profile was measured experimentally using the Illumina HiSeq 2000 RNA Sequencing platform by the University of North Carolina TCGA genome characterization center. Level 3 data was downloaded from TCGA data coordination center. This dataset shows the exon-level transcription estimates, as in RPKM values (Reads Per Kilobase of exon model per Million mapped reads). Exons are mapped onto the human genome coordinates using UCSC Xena

gene expression array

- ① gene expression (AgilentG4502A_07_3) (n=597) TCGA hub

TCGA breast invasive carcinoma (BRCA) gene expression data by agilent array. The gene expression profile was measured experimentally using Agilent 244K custom gene expression G4502A_07_3 microarrays by the University of North Carolina TCGA genomic characterization center. Level 3 interpreted level data was downloaded from TCGA data coordination center. This dataset shows the gene level transcription estimates (level 3 data), as in log2 lowess normalized ratio...

gene expression RNAseq

- ① gene expression RNAseq (polyA+ IlluminaHiSeq pancan normalized) (n=1,218) TCGA hub

TCGA breast invasive carcinoma (BRCA) gene expression by RNAseq, mean-normalized (per gene) across all TCGA cohorts. Values in this dataset are generated at UCSC by combining "gene expression RNAseq" values of all TCGA cohorts, values are then mean-centered per gene, then extracting the converted data only belongs to the this cohort. For comparing data within this cohort, we recommend to use the "gene expression RNAseq" dataset. For questions regarding...

- ① gene expression RNAseq (polyA+ IlluminaHiSeq percentile) (n=1,218) TCGA hub

For each sample, we rank genes RSEM values between 0% to 100%. This dataset is gene expression estimation in percentile rank, which higher value representing higher expression. The dataset can be used to compare this RNAseq data with other cohorts when the other data is processed in the same way (i.e. percentile ranking). Genes are mapped onto the human genome coordinates using UCSC Xena HUGO probeMap (see ID/Gene mapping link below for details).

- ① gene expression RNAseq (polyA+ IlluminaHiSeq) (n=1,218) TCGA hub

The gene expression profile was measured experimentally using the Illumina HiSeq 2000 RNA Sequencing platform by the University of North Carolina TCGA genome characterization center. Level 3 data was downloaded from TCGA data coordination center. This dataset shows the gene-level transcription estimates, as in log2(x+1) transformed RSEM normalized count. Genes are mapped onto the human genome coordinates using UCSC Xena HUGO probeMap (see ID/Gen...

miRNA gene expression RNAseq

- ① miRNA gene expression (IlluminaGA) (n=333) TCGA hub

- ① miRNA gene expression (IlluminaHiSeq) (n=832) TCGA hub

4. Download the dataset by clicking the "Download" button at the right top of the page:

dataset: gene expression RNAseq (polyA+ IlluminaHiSeq pancan normalized)

[Visualize](#)

[Download](#)

TCGA breast invasive carcinoma (BRCA) gene expression by RNAseq (polyA+ IlluminaHiSeq), pancan normalized

TCGA breast invasive carcinoma (BRCA) gene expression by RNAseq, mean-normalized (per gene) across all TCGA cohorts. Values in this dataset are generated at UCSC by combining "gene expression RNAseq" values of all TCGA cohorts, values are then mean-centered per gene, then extracting the converted data only belongs to the this cohort.

For comparing data within this cohort, we recommend to use the "gene expression RNAseq" dataset. For questions regarding the gene expression of this particular cohort in relation to other types tumors, you can use the pancan normalized version of the "gene expression RNAseq" data. For comparing with data outside TCGA, we recommend using the percentile version if the non-TCGA data is normalized by percentile ranking. For more information, please see our Data FAQ: [here](#).

5. Go back to the original page you opened on step 3. This time, click on the **Phenotypes** link:

miRNA gene expression RNAseq

- ① miRNA gene expression (IlluminaGA) (n=333) TCGA hub

- ① miRNA gene expression (IlluminaHiSeq) (n=832) TCGA hub

miRNA isoform expression RNAseq

- ① miRNA isoform expression (IlluminaGA) (n=333) TCGA hub

TCGA breast invasive carcinoma (BRCA) miRNA expression by RNAseq. The miRNA expression profile was measured experimentally using the Illumina Genome Analyzer RNA Sequencing platform by the British Columbia Cancer Agency TCGA genome characterization center. Level 3 interpreted level data was downloaded from TCGA data coordination center. Download data is in the unit of reads per million mapped reads (RPM). This dataset shows the miRNA transcription...

- ① miRNA isoform expression (IlluminaHiSeq) (n=832) TCGA hub

TCGA breast invasive carcinoma (BRCA) miRNA expression by RNAseq. The miRNA expression profile was measured experimentally using the Illumina HiSeq 2000 RNA Sequencing platform by the British Columbia Cancer Agency TCGA genome characterization center. Level 3 interpreted level data was downloaded from TCGA data coordination center. Download data is in the unit of reads per million mapped reads (RPM). This dataset shows the miRNA transcription estimat...

phenotype

- ① Phenotypes (n=1,247) TCGA hub

protein expression RPPA

- ① protein expression RPPA (n=937) TCGA hub

TCGA breast invasive carcinoma (BRCA) protein expression by reverse phase protein array (RPPA). The data was generated and processed at the MD Anderson Cancer Center TCGA proteome characterization center RPPA core. Level 3 interpreted level data was downloaded from TCGA data coordination center. Data normalization from the MDACC RPPA core: under section How are the RPPA data processed.

- ① protein expression RPPA (RBN) (n=747) TCGA hub

TCGA breast invasive carcinoma (BRCA) protein expression data for 131 proteins, measured by RPPA (reverse phase protein array) technology. These data have been normalized by RBN (replicate-base normalization) method developed by MDACC. Details: <https://www.synapse.org/ISynapse:syn1750330> and <http://bioinformatics.mdanderson.org/main/TCGA/Overview>.

somatic mutation (SNPs and small INDELS)

- ① somatic mutation SNPs and small INDELS (wustli curated) (n=982) TCGA hub

TCGA breast invasive carcinoma (BRCA) somatic mutation data. Sequencing data are generated on a Illumina system. The calls are generated at Genome Institute at Washington University Sequencing Center using the WashU pipeline method.

- ① somatic mutation SNPs and small INDELS (wustli) (n=776) TCGA hub

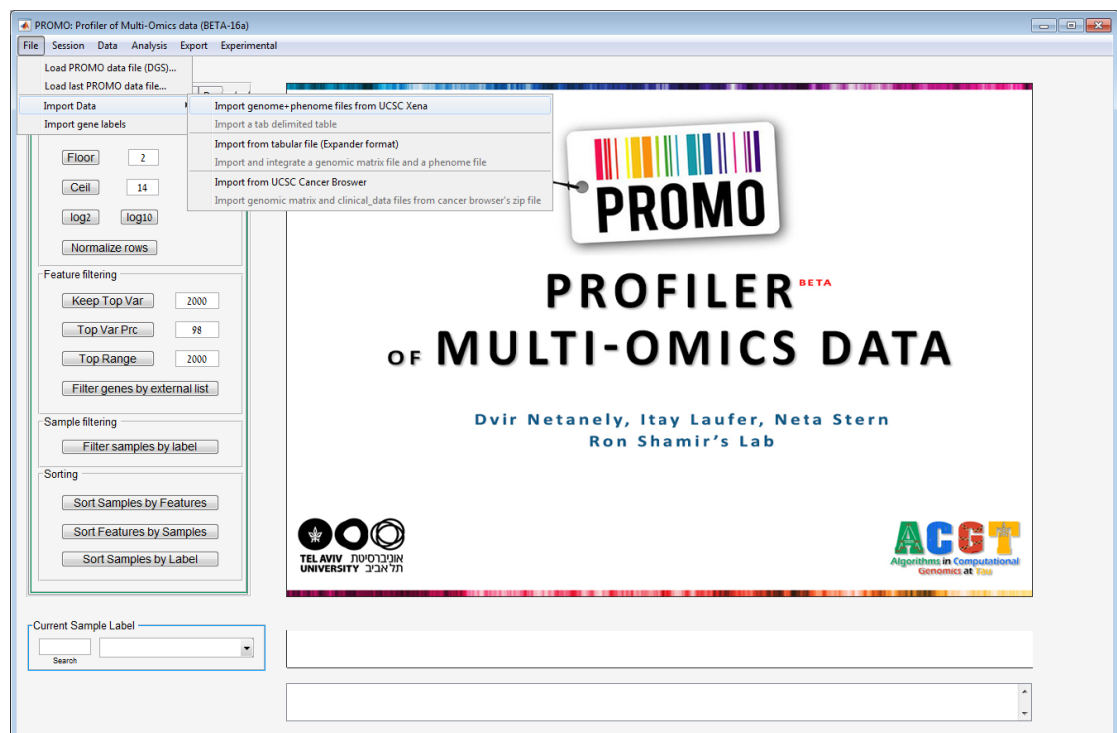
TCGA breast invasive carcinoma (BRCA) somatic mutation data. Sequencing data are generated on a Illumina system. The calls are generated at Genome Institute at Washington University Sequencing Center using the WashU pipeline method.

- Download the phenotype file by clicking the "Download" button at the right top of the page:

dataset: Phenotypes

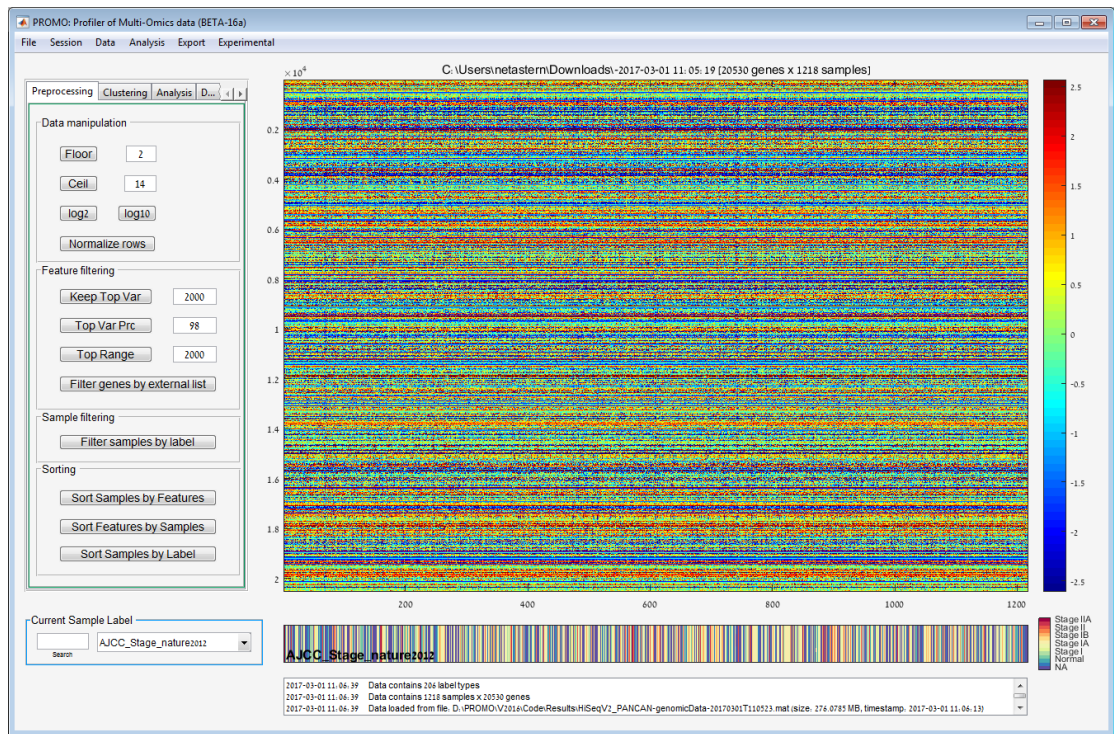
| | | | |
|-------------------|---|-----------|----------|
| | | Visualize | Download |
| cohort | TCGA Breast Cancer (BRCA) | Visualize | |
| dataset ID | TCGA.BRCA.sampleMap/BRCA_clinicalMatrix | | |
| download | https://tcga.xenahubs.net/download/TCGA.BRCA.sampleMap/BRCA_clinicalMatrix; Full metadata | | |
| samples | 1,247 | | |
| version | 2016-04-27 | | |
| hub | TCGA hub | | |
| type of data | phenotype | | |
| raw data | https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/brca/bcr/ | | |
| input data format | ROWS (samples) x COLUMNS (identifiers) (i.e. clinicalMatrix) | | |

- Now import the two files you have downloaded. Open PROMO, and in the top menu, select **File >> Import Data >> Import genome + phenome files from UCSC Xena**:



- Select the Genomic Data file and the Phenome file you have downloaded.
- Select a location for the DGS (PROMO) file.

After import is done, you should see the genomic data presented on the main window:



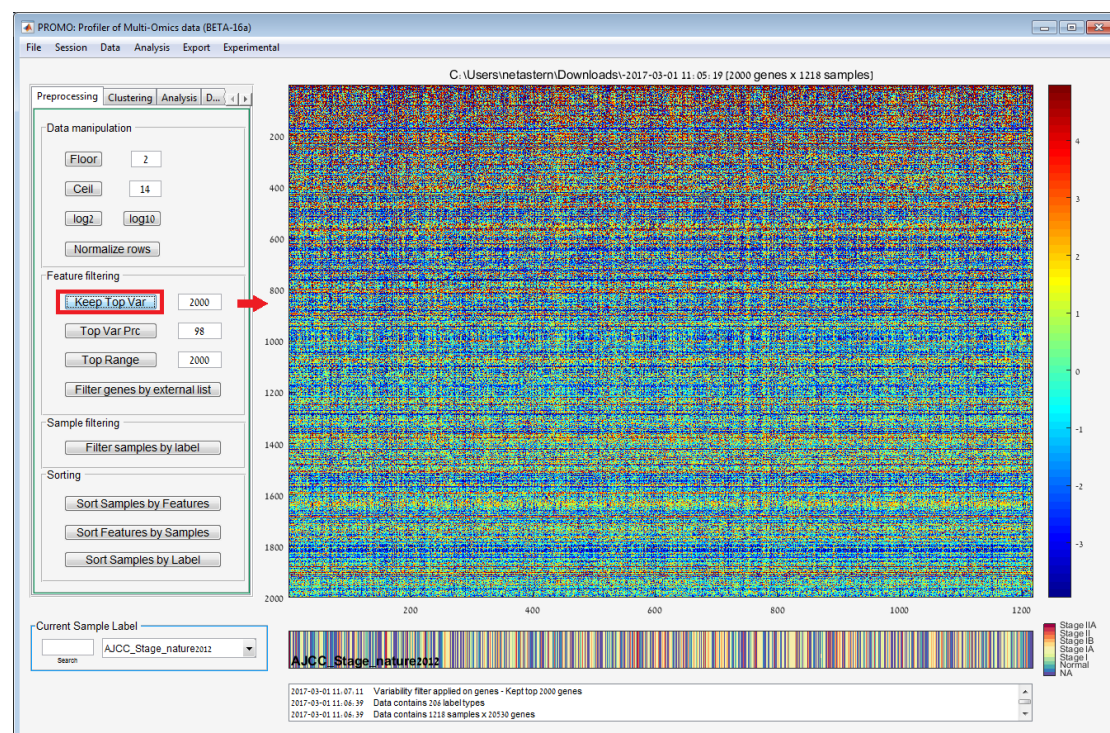
Step 2 - Preprocessing

PROMO offers several preprocessing operations, among them are flooring, which means setting all matrix values that are below a certain threshold (set by the user) to that threshold. The threshold can be set in the adjacent textbox. Other preprocessing options are also available (ceil, log2, log10, and row normalization).

We will now use 2 of PROMO's preprocessing abilities to reduce the data to the most relevant information:

1. **Keep Top Var** (*Preprocessing Tab >> Feature Filtering >> Keep Top Var*): Applies row filtering by keeping the features having the top variance. The number of the features to be kept can be set at the adjacent textbox. Other filtering options are also available (Keep Top Var Percentile, Keep Top Range).

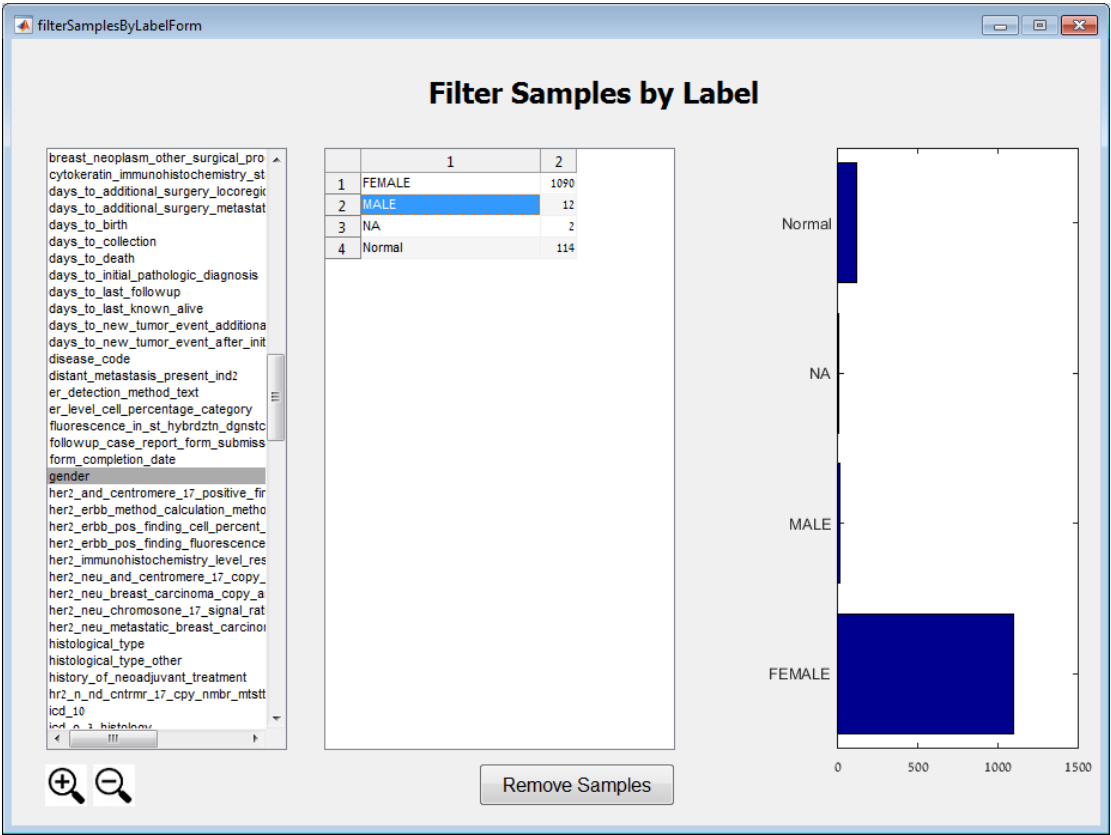
Click on the "Keep Top Var" button to filter the features. Only the 2000 rows with the highest variance will be kept.



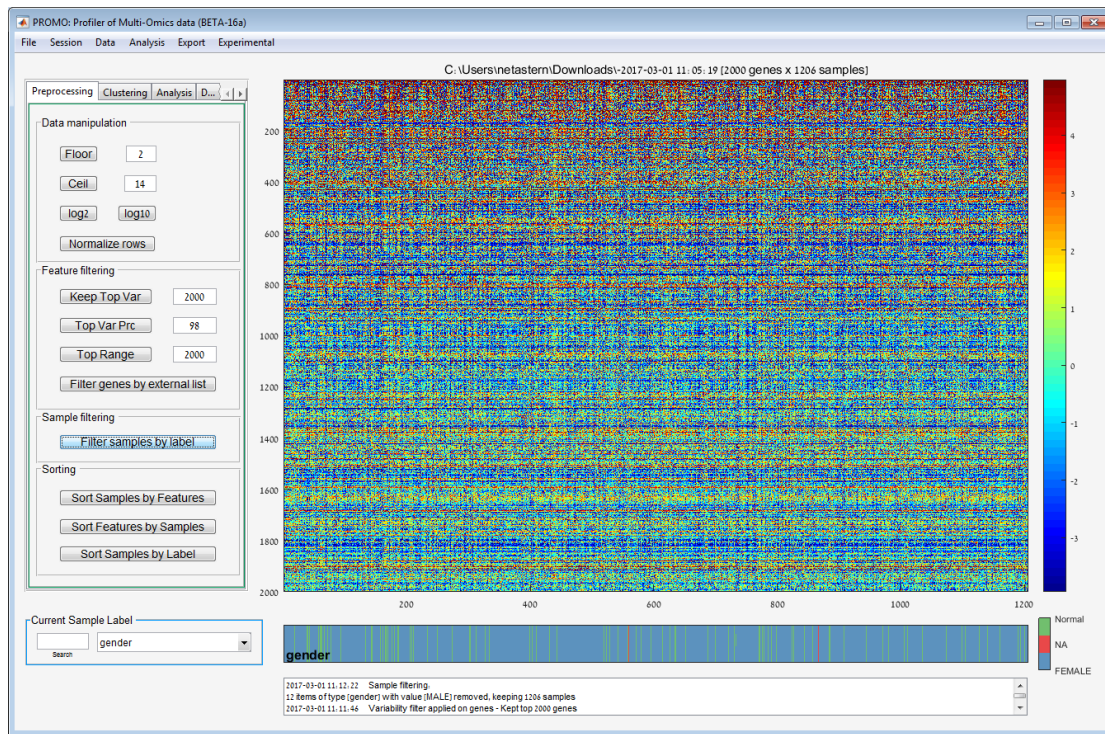
2. **Sample Filtering** (*Preprocessing Tab >> Sample Filtering >> filter samples by label*):

This option allows you to remove samples with a given label value.

In this example, after clicking on the "Filter samples by label" button, we should see a gender label on the left list.



We wish to remove all the male samples out of the data. Select the gender label and the MALE value and click on "Remove Samples". All samples containing the MALE label value will be deleted from the data.



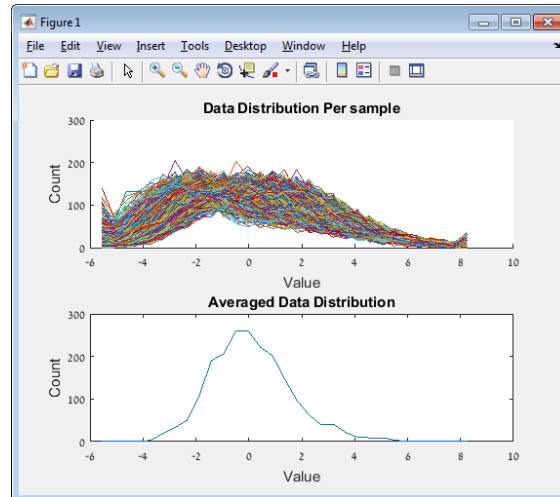
Step 3 – Data Exploration

Data exploration can be performed using several methods:

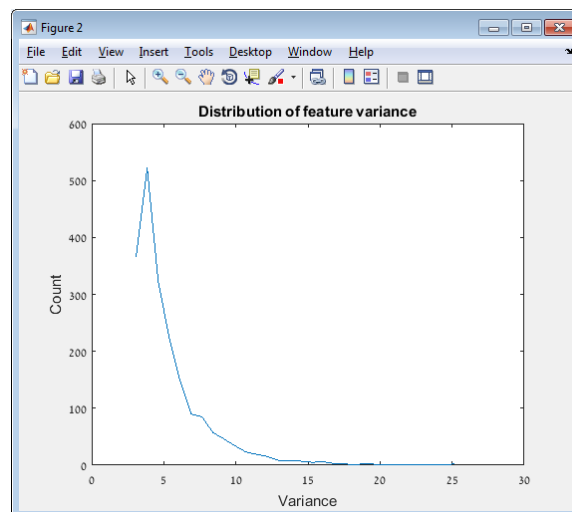
1. Data Distribution (Analysis menu >> Data Distribution)

This option allows you to view several plots describing data distribution, including:

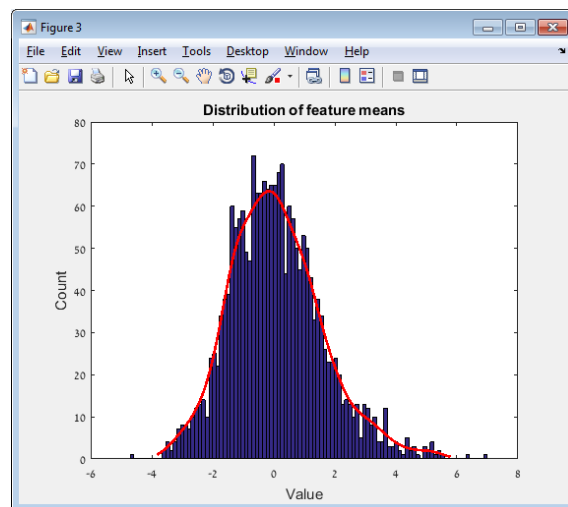
a. Data Distribution Per Sample



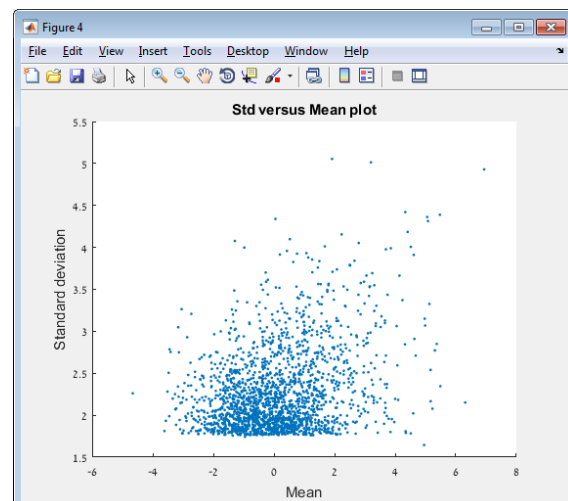
b. Distribution of feature variance



c. Distribution of feature means



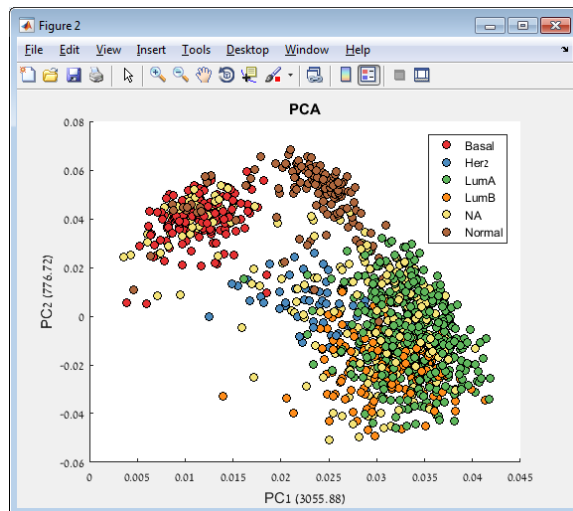
d. Std versus mean plot



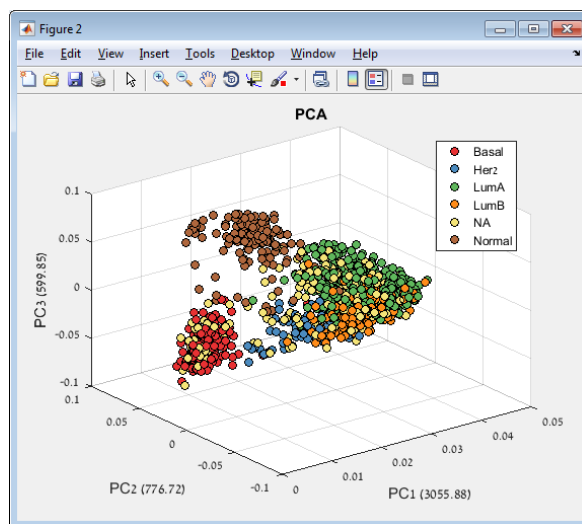
2. **PCA** (Analysis menu >> PCA)

You can view several different PCA plots using PROMO. The plots are generated using the selected label in the main window.

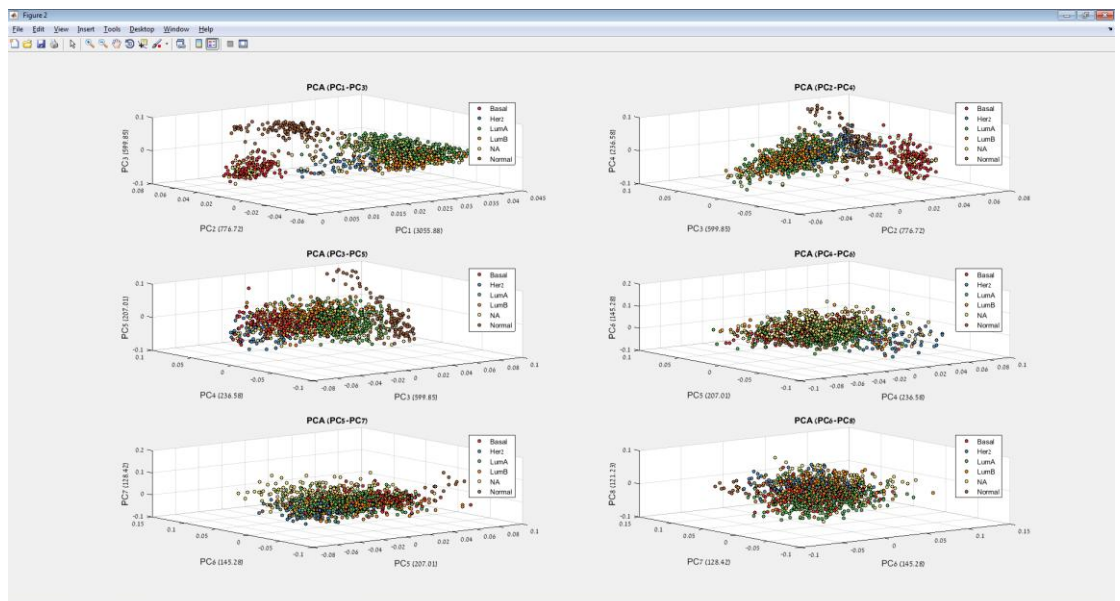
a. PCA 2 Dimensional Plots



b. PCA 3 Dimensional Plots

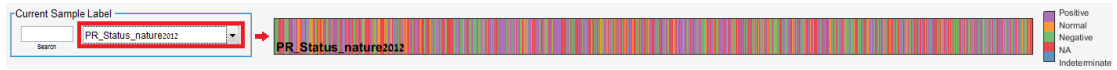


c. PCA Multi Dimensional Plots



3. Exploring Labels

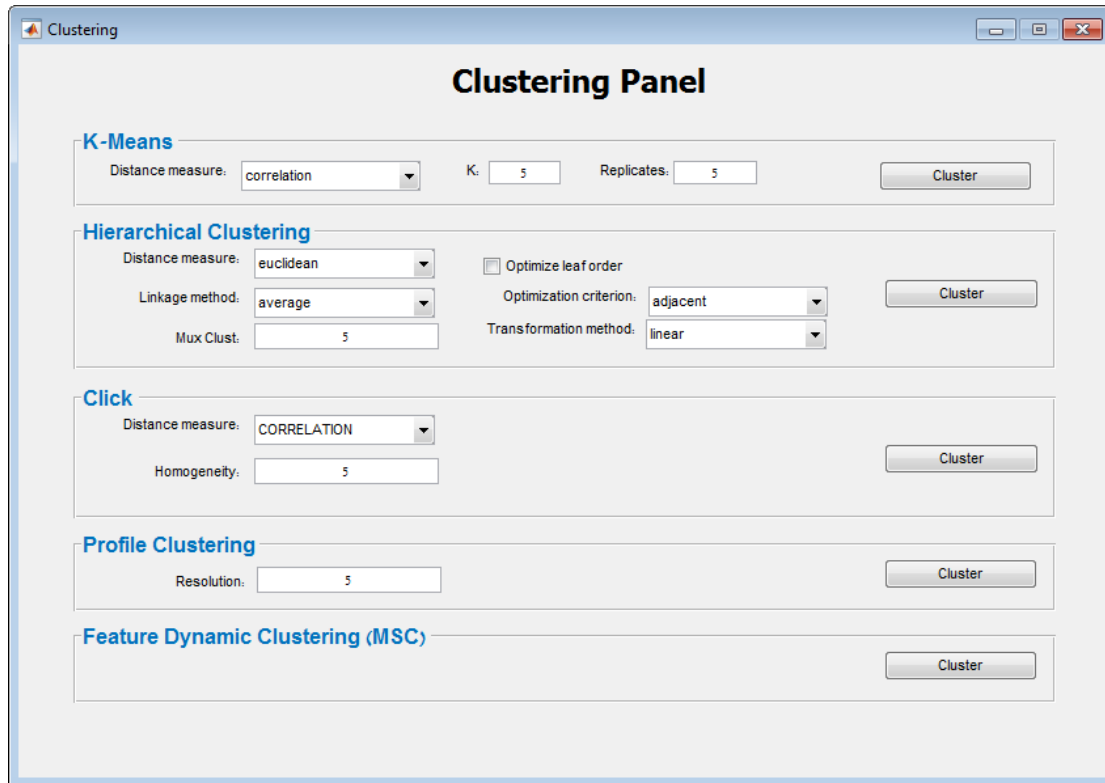
You can select different labels using the label list in the current sample label panel at the left bottom of the main window, and view the label values belonging to each sample in the corresponding plot on the right.



Step 4 – Clustering

PROMO allows clustering both on the samples and the features separately. Clustering can be done using the Clustering tab in the main window or using the Analysis menu (Analysis >> Clustering >> Cluster samples/Cluster genes).

Clustering samples and features

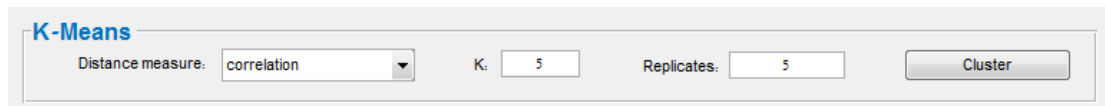


The screenshot shows the 'Clustering Panel' window with the following sections:

- K-Means**: Distance measure: correlation, K: 5, Replicates: 5, Cluster button.
- Hierarchical Clustering**: Distance measure: euclidean, Linkage method: average, Mux Clust: 5, Optimize leaf order: ☐, Optimization criterion: adjacent, Transformation method: linear, Cluster button.
- Click**: Distance measure: CORRELATION, Homogeneity: 5, Cluster button.
- Profile Clustering**: Resolution: 5, Cluster button.
- Feature Dynamic Clustering (MSC)**: Cluster button.

Both types of clustering are done using the Clustering Panel. The clustering panel allows you to choose the algorithm used for clustering and set its parameters.

In this example we will use the **K-Means** algorithm. Other clustering solutions include hierarchical clustering and click.



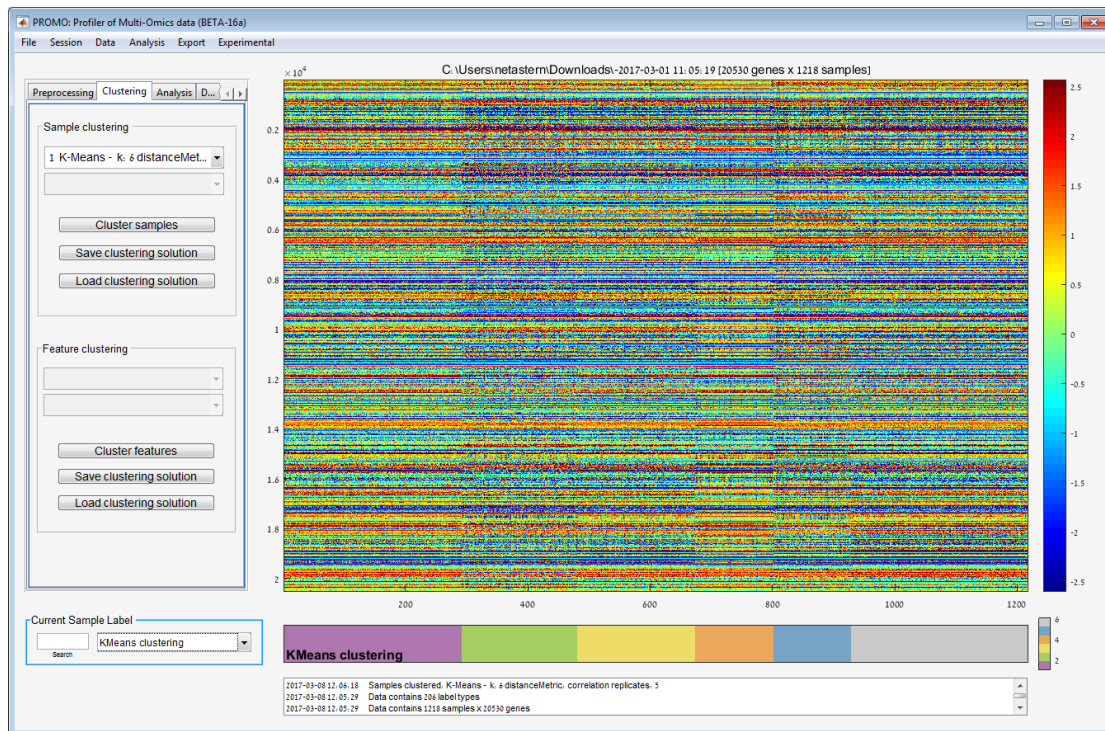
This close-up shows the 'K-Means' section with the following settings:

- Distance measure: correlation
- K: 5
- Replicates: 5
- Cluster button

Change the K value to 6 and click on the Cluster button.

The expected results should be viewed in the main window. Notice the current sample label has changed to the cluster you have just created.

The clustering solution is added to the label list. You can walk through cluster solutions and save and load clustering solutions.



Step 5 – Analysis of sample clusters

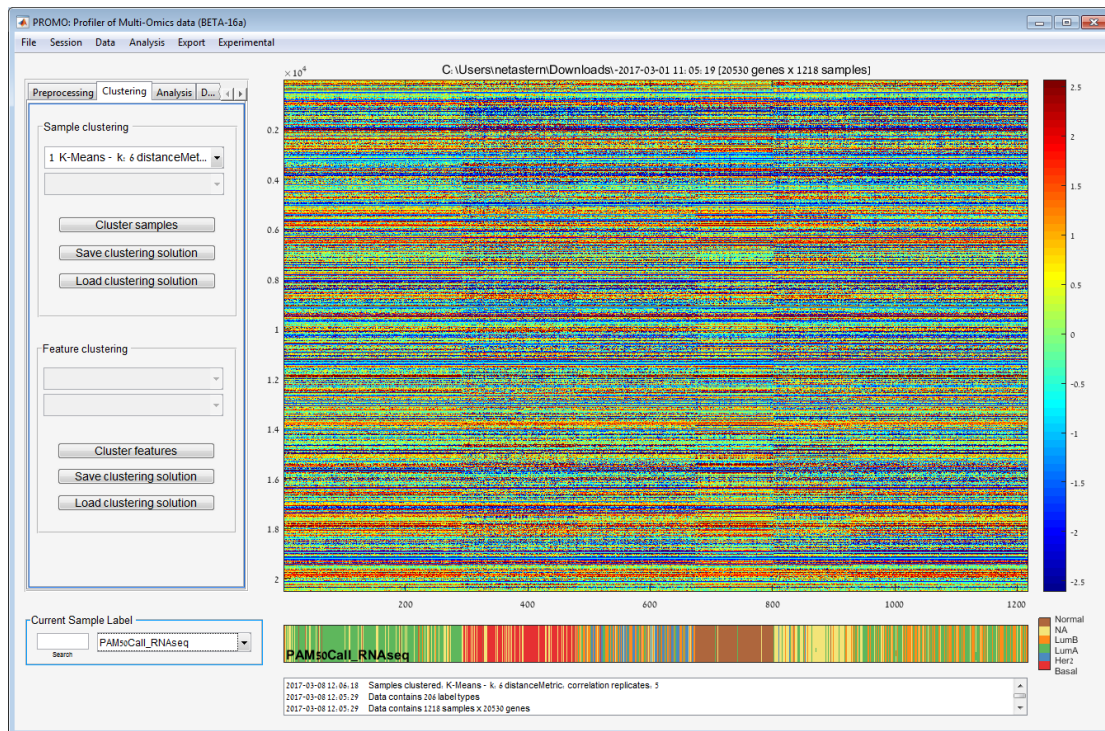
Single label analysis (Analysis >> Analysis of sample clusters >> Selected label analysis)

Use single label analysis to analyze label enrichment on clusters for a specific label. The analyzed label is set in the Current Sample Label list at the left bottom of the main window.

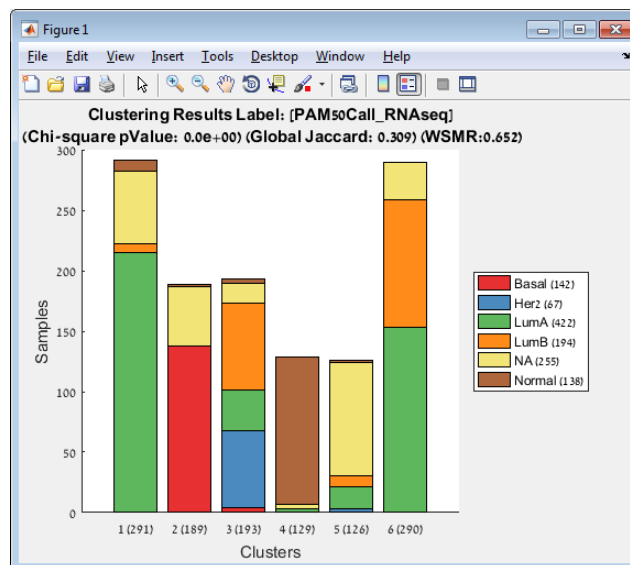
Select the label you wish to analyze:

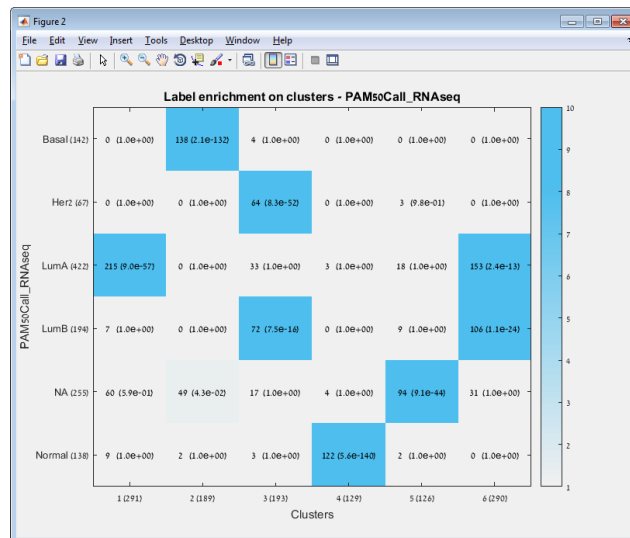
Current Sample Label

Search PAM50Call_RNAseq



Click the Selected label analysis menu item and view the results.

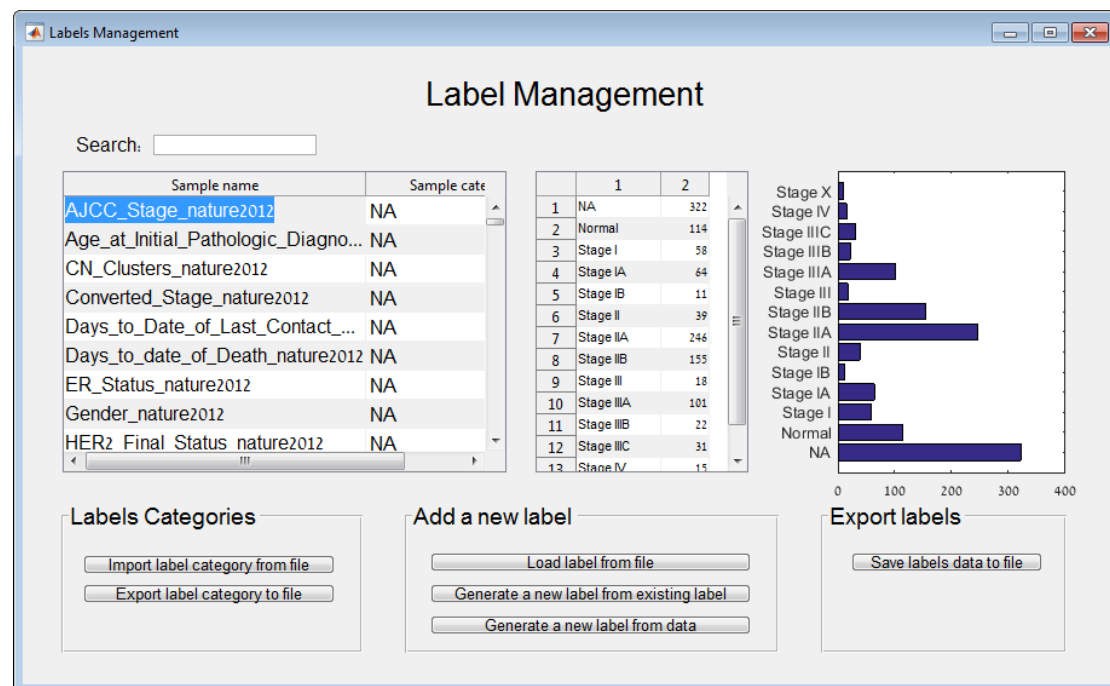




Multi label analysis (Analysis >> Analysis of sample clusters >> Multi label analysis)

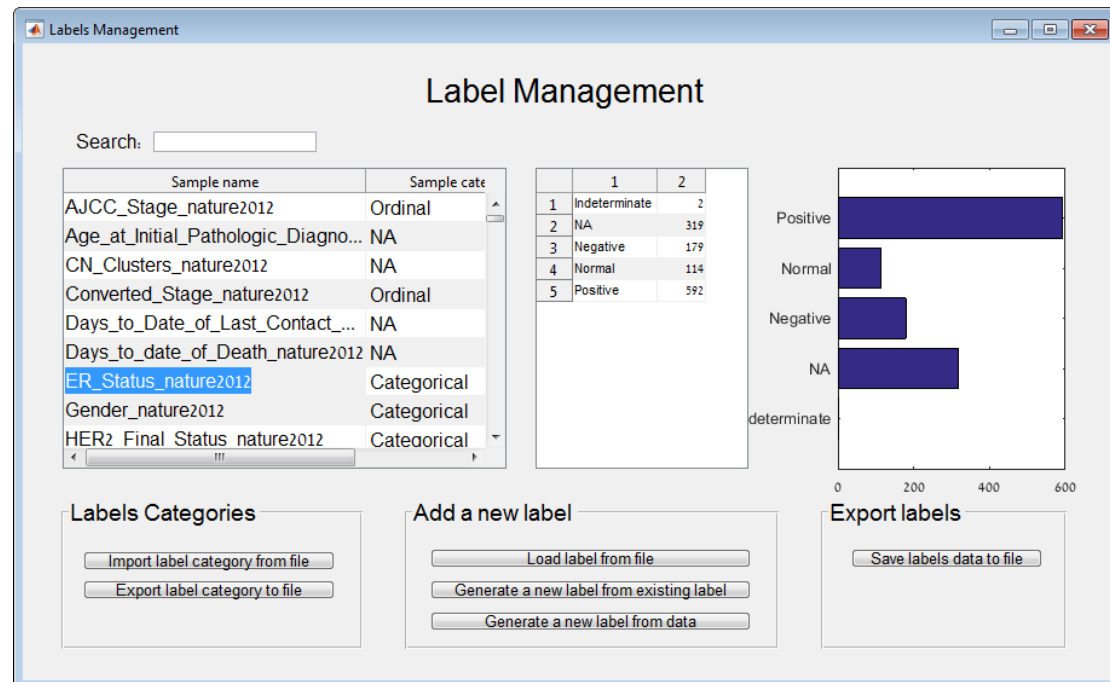
We now wish to execute multi label analysis on our clustering solution. In order to do that, first we need to define the labels categories.

Enter the Label Management Panel using the Data menu (Data >> Label Management).



The Label Management Panel allows us to set the category of each label. This is done using the second column of the left table. The right table and plot describing the distribution of the values of the label can help you decide which category to define for the label you are interested of. Also, you can click on the button "Import label category from file" to load a pre prepared list of label categories.

Click on the button "Import label category from file" and select the file 'TCGA labels category.txt' in the Resources folder (in the installation folder of PROMO).



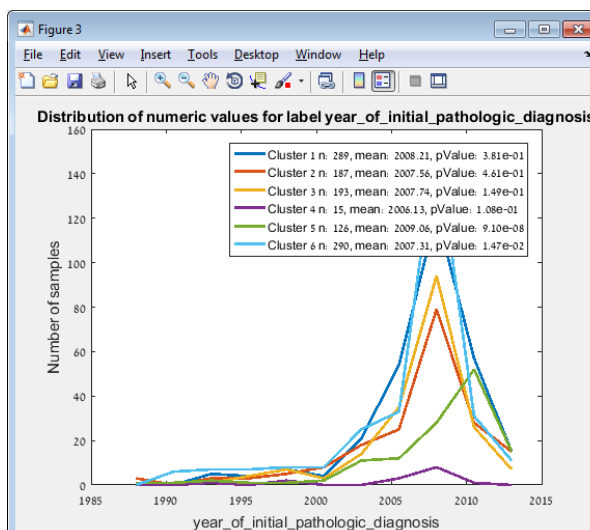
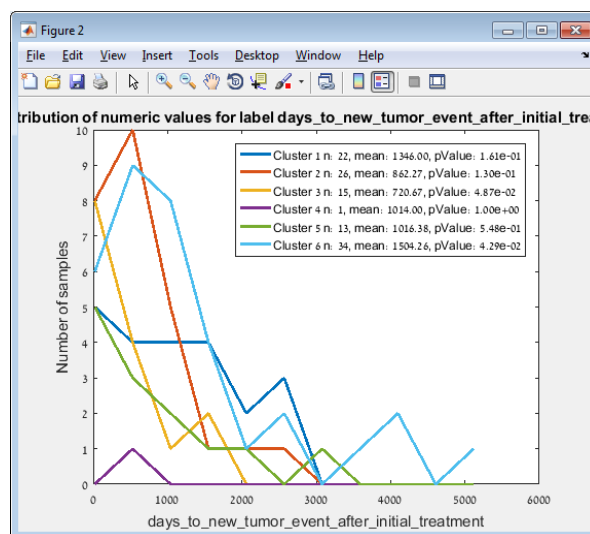
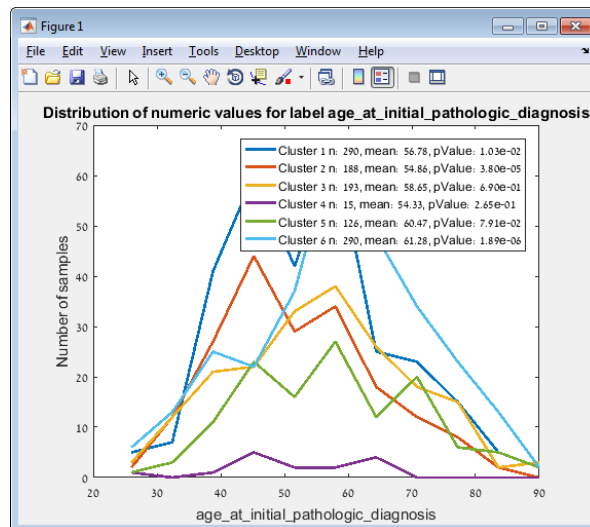
You should now see that some of the label categories have been set. Now you can execute multi label analysis on the clustering solution. Close the Labels Management panel and return to the main window. **Use the Analysis menu to execute multi label analysis (Analysis >> Analysis of sample clusters >> Multi label analysis).** A new panel will open, allowing you to define the parameters for the analysis:

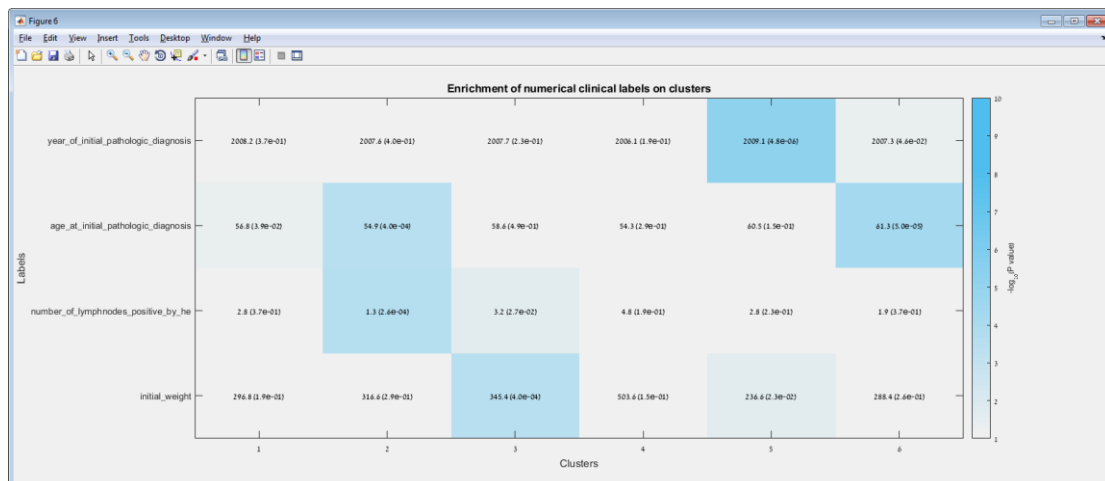
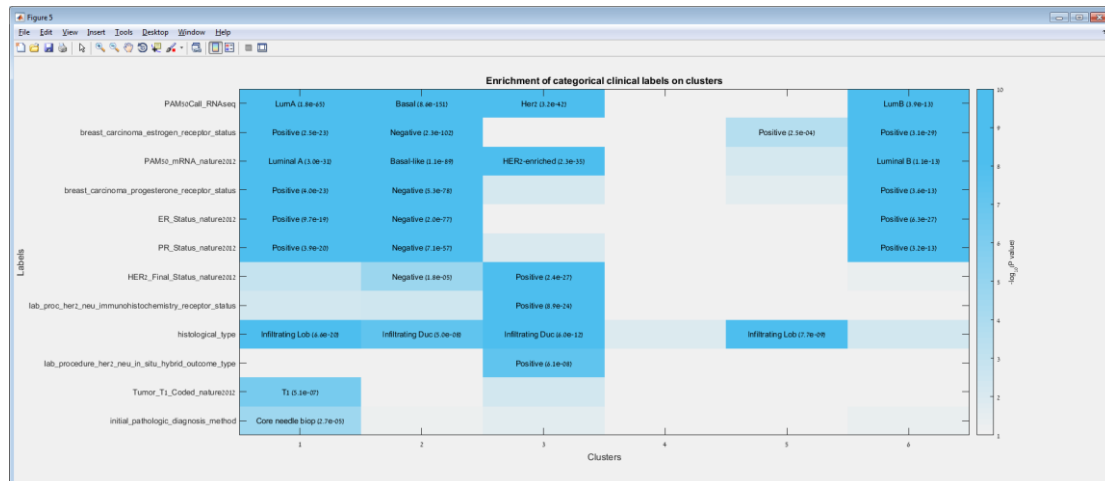
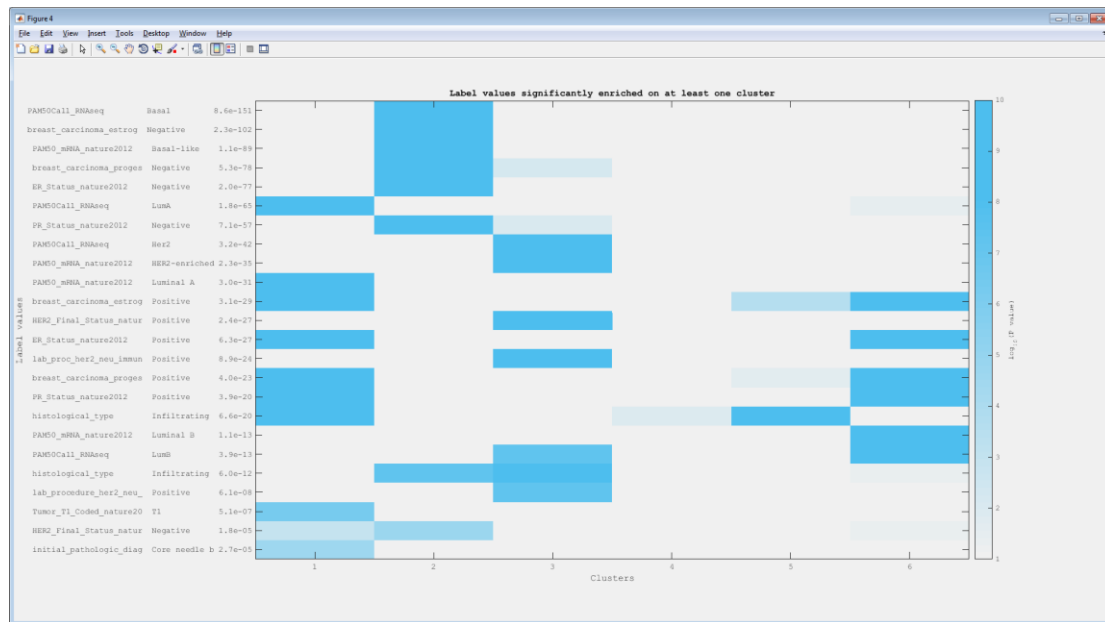
The 'Clinical I...' dialog box contains the following parameters:

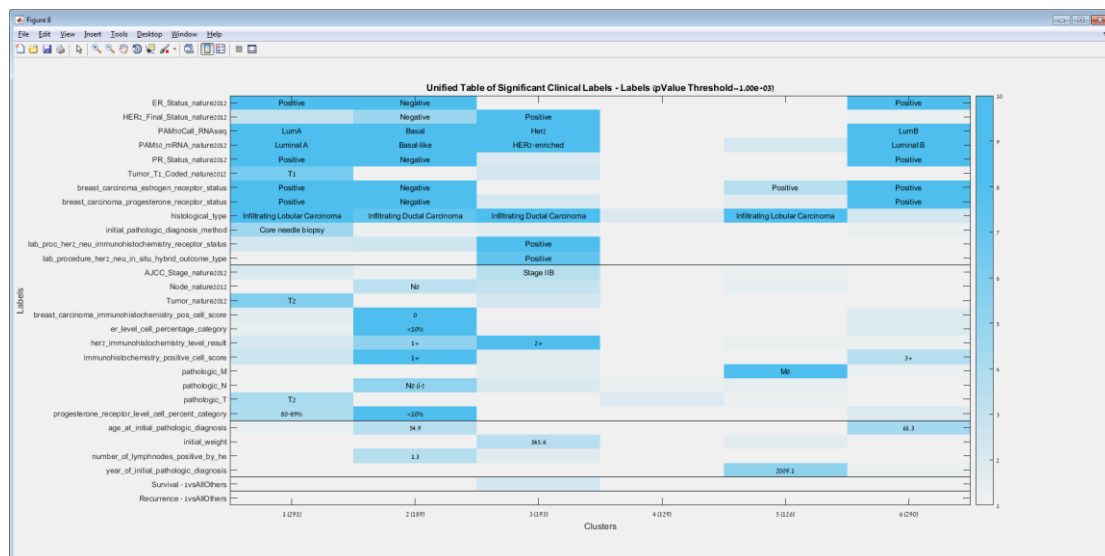
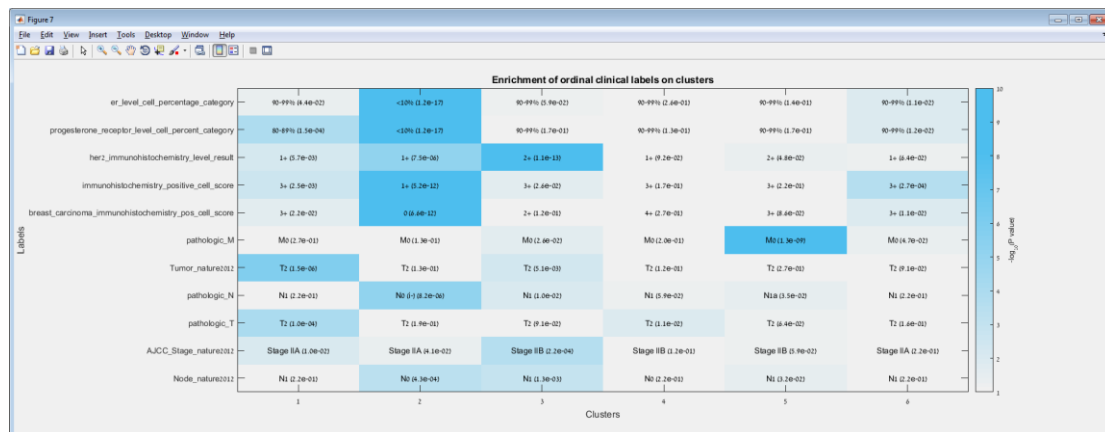
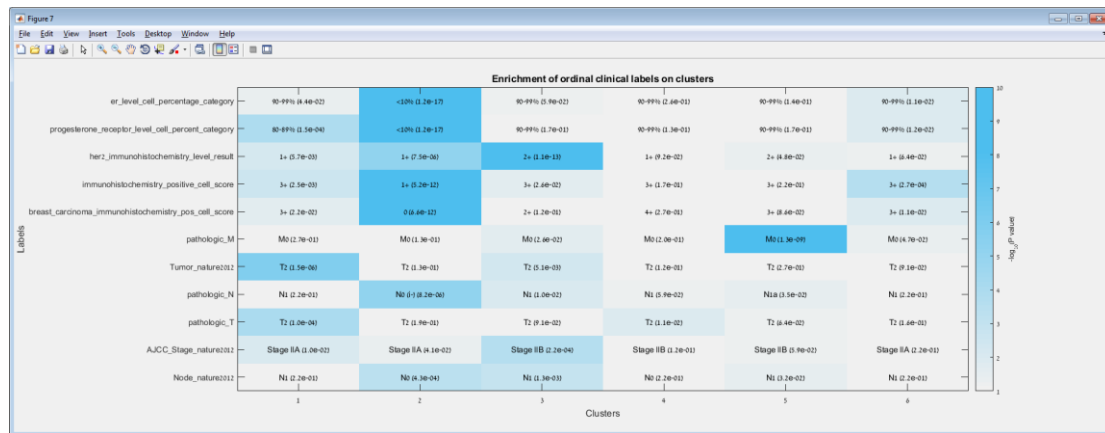
- Alpha: 0.001
- Show Figures: 0=no, 1=yes (0)
- Save Figures: 0=no, 1=yes (0)
- Ignored label values: NA Normal
- Verbose: 0=no, 1=yes (0)
- Minimum cluster size: 3
- Max label values for categorical test: 20
- Group Figures: 0=no, 1=yes (0)

Buttons: OK, Cancel

Leave the default parameters as they are and click OK. You should see the results found by PROMO shown using different plots in separated windows.







Biomarker Identification (Analysis >> Supervised tests)

If you want to identify features that are differentially expressed between two sample groups (defined by any given label), you can use the supervised analysis window.

- Select the test you wish to execute.** Available tests are: Ranksum test, Student t-test (unequal variance), Student t-test (equal variance), ANOVA, Kruskal-Wallis.

Supervised Tests Panel

Supervised tests

Test type: **Ranksum test** ☒ Apply FDR correction ☒ Add gene groups to clustering solutions list

KMeans clustering

| | 1 | 2 |
|---|---|----|
| 1 | 1 | 34 |
| 2 | 2 | 43 |
| 3 | 3 | 24 |
| 4 | 4 | 26 |
| 5 | 5 | 32 |
| 6 | 6 | 14 |

Group 1

Add to Group 1

Clear group

Group 2

Add to Group 2

Clear group

Filtering

☒ Significance filter: 0.01

☐ Fold change filter: 2 >< Median Difference

☐ Top genes filter: 2000

Run

2. Select a label.

Supervised Tests Panel

Supervised tests

Test type: **Ranksum test** ☒ Apply FDR correction ☒ Add gene groups to clustering solutions list

KMeans clustering

| | 1 | 2 |
|---|---|----|
| 1 | 1 | 34 |
| 2 | 2 | 43 |
| 3 | 3 | 24 |
| 4 | 4 | 26 |
| 5 | 5 | 32 |
| 6 | 6 | 14 |

Group 1

Add to Group 1

Clear group

Group 2

Add to Group 2

Clear group

Filtering

☒ Significance filter: 0.01

☐ Fold change filter: 2 >< Median Difference

☐ Top genes filter: 2000

Run

- ## 3. Build two groups for comparison by associating the label values.
- You can associate a group with a label value by selecting the value from the left table and clicking the "Add to Group" button above the relevant group.

Supervised Tests Panel

Supervised tests

Test type: Ranksum test ☒ Apply FDR correction ☒ Add gene groups to clustering solutions list

KMeans clustering

| | 1 | 2 |
|---|---|----|
| 1 | 1 | 24 |
| 2 | 2 | 42 |
| 3 | 3 | 36 |
| 4 | 4 | 19 |
| 5 | 5 | 22 |
| 6 | 6 | 30 |

Group 1

Add to Group 1

Clear group

Group 2

Add to Group 2

Clear group

Filtering

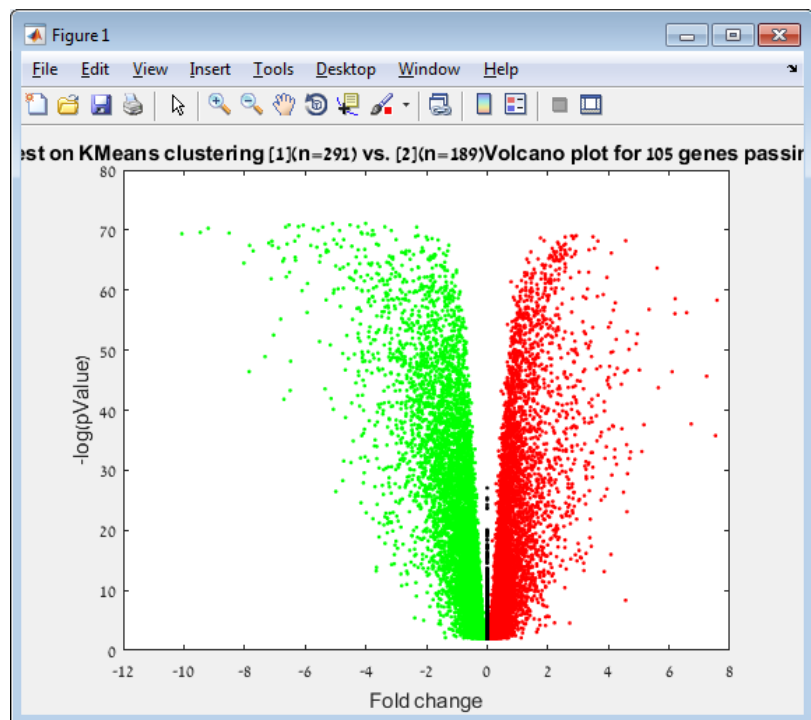
☒ Significance filter: 0.01

☐ Fold change filter: 2 >< Median Difference

☐ Top genes filter: 2000

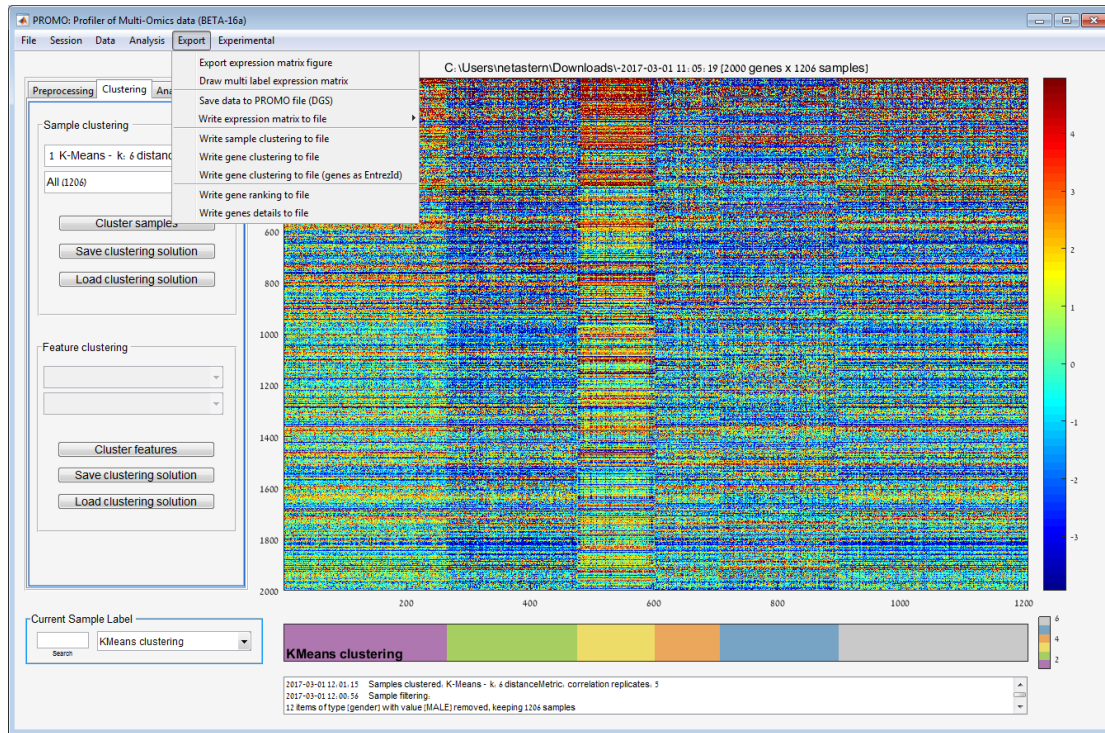
Run

Click the Run button and view the results.



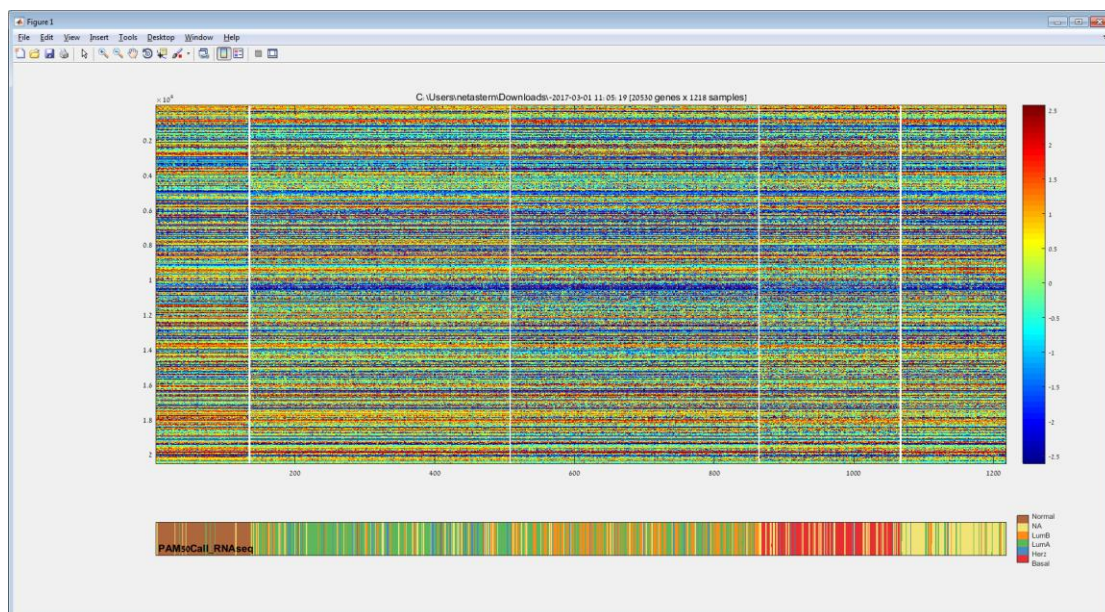
Step 6 – Export

PROMO offers several methods to export your data. One of them allows you to export all data to a tabular file (**Export >> Write expression matrix to file >> to tabular format**). You can also export both sample and feature clustering solutions (**Export >> Write sample clustering to file** or **Export >> Write gene clustering to file**).



Export Figure (Export >> Export expression matrix figure)

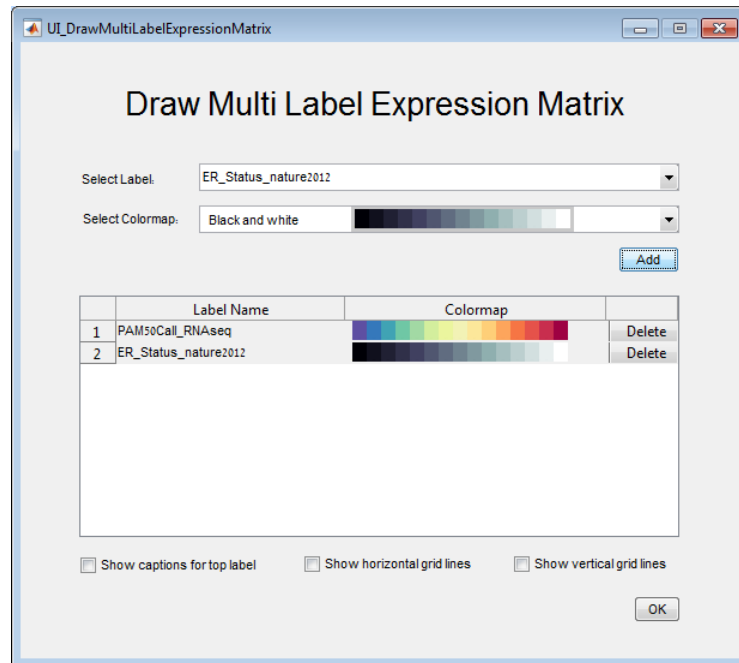
Use the Export expression matrix figure menu to export the dataset visual plot and the selected label plot to an external figure.



Draw Multi label expression matrix (Export >> Draw Multi label expression matrix)

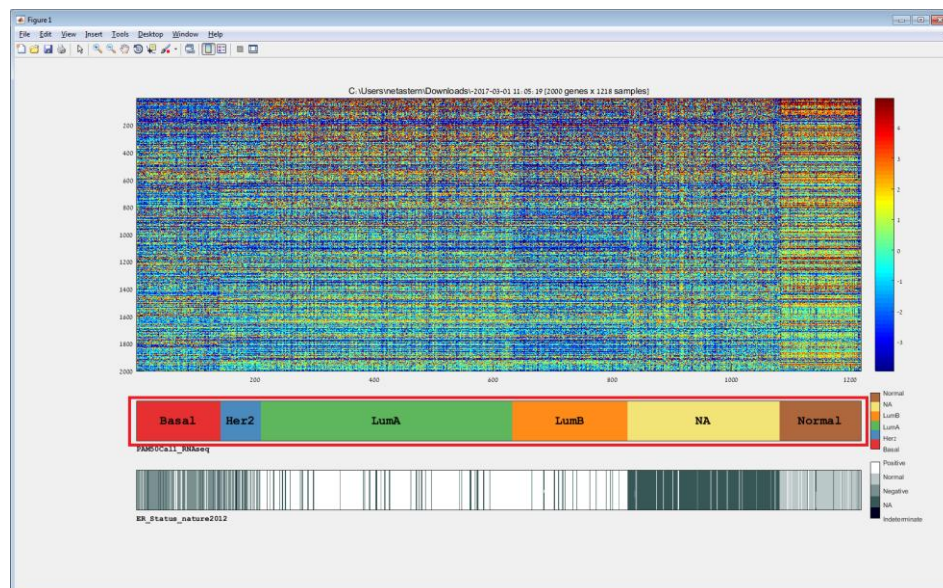
This option allows you to view your main data matrix along with several different labels presented in separated bars.

You can select labels and colormaps in the parameters window:



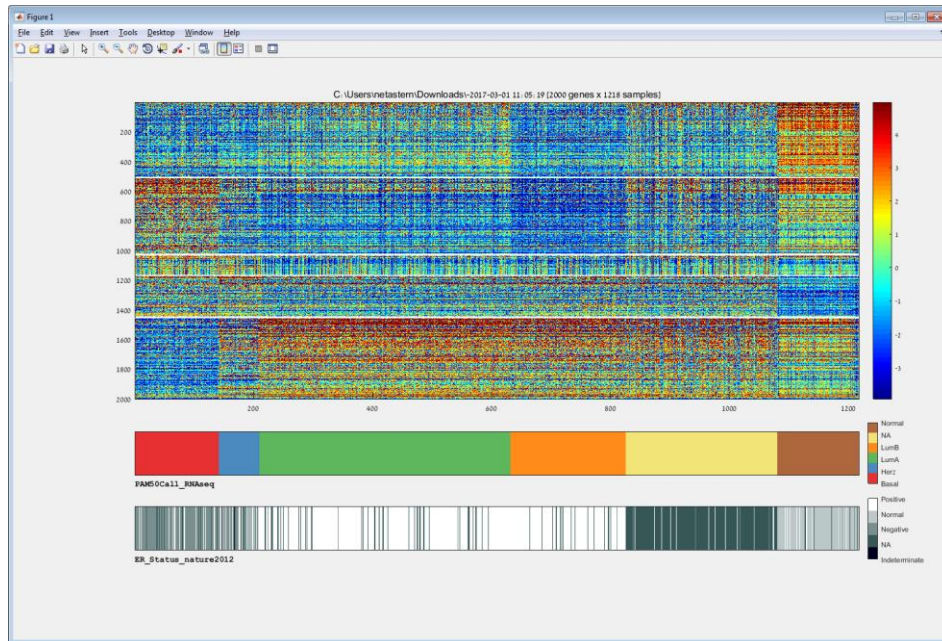
Additional options are:

- **Show captions for top label**
You can add captions representing the first label values to the first label bar. This option is recommended for using only when the samples are sorted by the first label. This can be done using the "Sort Samples by Label" button in the Preprocessing tab.



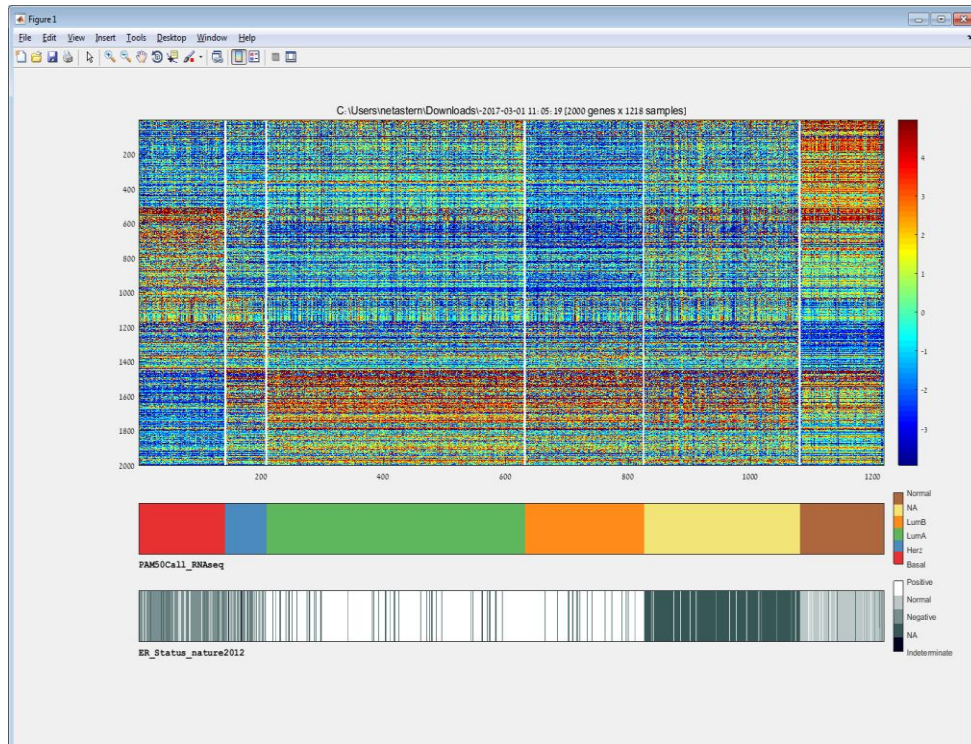
- **Show horizontal grid lines**

When a feature clustering solution exists, you can use this option to present horizontal separators between the different clusters in the main data matrix.



- **Show vertical grid lines**

You can use this option to present vertical grid lines in the main data matrix. The vertical grid lines separate between samples with different label values, referring to the values of the first selected label. That is why this option is recommended for using only when the samples are sorted by the first label.



It is possible to apply all or some of the options by checking the relevant checkboxes in the parameters window:

