PROMO – Profiler of Multi-Omics Data

Integrative Multi-omic Analysis Tutorial

Introduction	2
Multi-Omic Dataset Collection Management	3
Inter-Omic Feature Correlation Analysis	11
Multi-Omic Clustering	17

Introduction

PROMO (Profiler of Multi-Omics data) is an interactive tool, designed to analyze large versatile datasets together with their clinical labels. This tutorial demonstrates PROMO's more advanced features for integrative analysis of multi-omic datasets (multiple datasets describing the same set of patients but generated using different omic technologies).

Use this tutorial if you wish to:

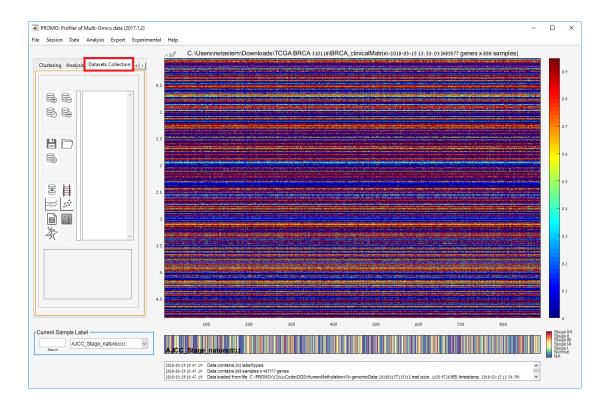
- Assemble a multi-omic collection by importing several different omic matrices using PROMO.
- Edit your multi-omic dataset collection by adding/removing datasets or by editing a specific dataset within the dataset collection.
- Intersect the collection datasets in order to keep only samples appearing in all collection datasets.
- Merge the collection datasets to yield a single featureconcatenated matrix.
- Inter-omic feature correlations Identify correlations of features in two different omics.
- **Multi-omic sample clustering** Cluster the patients using any subset of the datasets in your collection.

Multi-Omic Dataset Collection Management

Multi-omic dataset collection is a set of datasets for the same group of paitents, each possibly generated by a different omic technology. Datasets in a collection can be used in PROMO for various integrative multi-omic analyses.

Follow these instructions to assemble a dataset collection:

- 1. Import the first dataset as described in section 'Step 1 Importing data' in PROMO's basic tutorial: <u>http://acgt.cs.tau.ac.il/promo/tutorial/PROMO Example Tutorial.pdf</u> The dataset you imported is now your active dataset.
- 2. On the left panel of the main form, switch to the tab 'Datasets Collection':



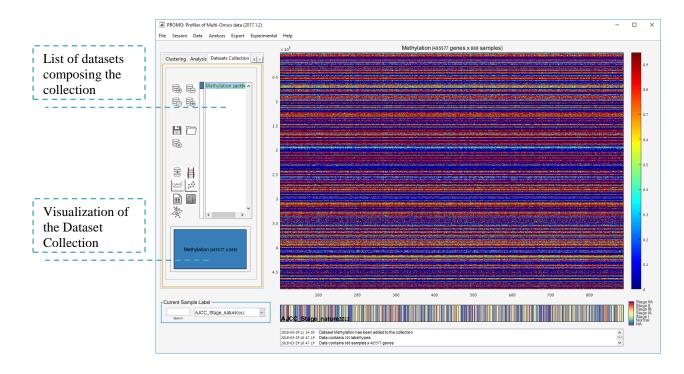
 Add the active dataset to your dataset collection by pressing the 'Add Dataset' button:

PROMO: Profiler of Multi-Omics data (2017.1.2) e Session Data Analysis Export Experimen	al Help	-		×
	×10 ⁵ C: \Users\netastern\Downloads\TCGA BRCA 310118\BRCA_clinicalMatrix\-2018-03-15 13: 53: 03 (485577 genes x 888 samples)			
Clustering Analysis Datasets Collection				0.9
				0.8
8 🗁	13 13 14			0.7
50	$_2$			0.6
8 4				0.5
 ■ 	³ Instanting the enders of the enders o		_	0.3
				0.2
	15			0.1
	2415254014141545400000000000000000000000			0
Current Sample Label AJCC_Stage_nature2012	AUCC_Stage_neture2012		Sta Sta Sta Sta Nor	ige IIA ige II ige IB ige IA ige I ige I
	2018-05-1918 47.19 Data contains 2018bellypes 2018-05-1918 47.19 Data contains 88 samples 45577 genes 2018-05-1918 47.19 Data backed from E. C. PROLOX2018 CodeDOS/HumanMethylation/30-genomicData-201803157135313 mat size 1450 9728 MB, timestamp. 2018-05-1513 54.59	^ ~	_ 18	

4. Enter a name for the dataset:

承 Enter Da	ataset Name	_		×
	Please enter data	aset nam	e:	
	Methylation	ı		
	ОК			

5. Press 'OK'. The active dataset will now be added to the dataset collection. You can see the names of the datasets composing the dataset collection on the list box under the 'Dataset Collection' tab on the left side of the screen.



6. Repeat steps 1-5 for every dataset you wish to add to your collection.

The following screenshot shows a dataset collection containing three different omics for TCGA lung cancer cohort: RNA-Seq, Protein, and Methylation. Notice that the name of the currently active dataset (RNA-Seq) is colored in green.



We will now edit and update one of the datasets within our dataset collection.

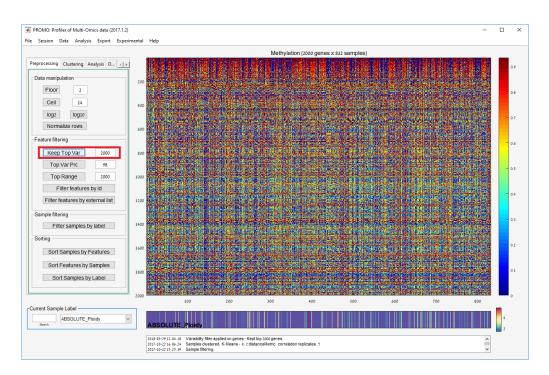
7. On the left panel, double-click the dataset that you wish to edit in order to make it the active dataset. In this example, we will edit the Methylation dataset. Notice that the dataset name on the left is now colored in green. Methylation is now the **active** dataset and we can make changes to this dataset using the various single-omic methods provided by PROMO.



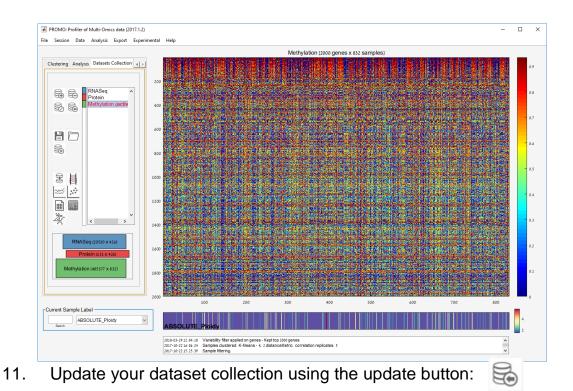
8. On the left panel of the main form, switch to tab 'Preprocessing':

ROMO: Profiler of Multi-Omics data (2017.1.2) File Session Data Analysis Export Experim	sedal Hele	-	□ ×
ne sesson bata mayas export experim	×10 ⁵ Methylation (485577 genes x 832 samples)		
Preprocessing Clustering Analysis D			0.9
Ceil 14 log2 log10			- 0.7
Normalize rows Feature filtering Keep Top Var 2000	15 A space parameter of the end of the end of the space of the space of the space of the end of the end of the space parameter of the space of th		- 0.6
Top Var Prc 98 Top Range 2000 Filter features by Id	23		- 0.5
Filter features by external list Sample filtering Filter samples by label			- 0.3
Sorting Sort Samples by Features			- 0.2
Sort Features by Samples Sort Samples by Label	45		0.1
Current Sample Label	100 200 300 400 500 600 700 800	,	0 4 2
	1017-10-2116-0, 34 Samples clustered K-Neans- k. 1 distanceMetric correlation replicates 5 3017-10-2113 12 34 Sample Maring 37 femm of yes, Linning "dessassi vel value filormali removed, keeping 313 samples	^ ~	

9. Apply row filtering for keeping the features with the highest variance: On the adjacent text box to the "Keep Top Var" Button, Enter the number of features to keep. In our example, we kept the top 2000 most variable features:



10. Switch back to the datasets collection tab. Notice that the Methylation dataset name is now colored in red, indicating the active dataset contains pending changes that were not saved to the collection. You can perform similar data filtering steps on the other datasets in the collection.



PROMO: Profiler of Multi-Omics data (2017.1.2) \times -File Session Data Analysis Export Experimental Help Methylation [2000 genes x 832 samples] Clustering Analysis Datasets Collection 200 RNASeq Protein Methylation (activ 50 50 400 60 800 3 100 *** 120 凌 RNASeq (20530 x 426) 0.2 16 Methylation (485577 x 832) Current Sample Label ~ ABSOLUTE_Ploidy 2018-03-2912.04-18 Variability fitter applied on genes - Kept top 2000 genes 2017-10-22.16.06.24 Samples clustered. K-Means - k. 2 distanceMetric: correlation replicates. 5 2017-10-22.15.25.39 Sample filtering. **^**

Notice the active dataset name (Methylation) is now colored in green. The changes made to the methylation dataset are now saved and our dataset collection has been updated.

12. Save the dataset collection to your local disk using the save button:



13. Choose a name for your dataset collection file and press "save":

Output file name						
→ ✓ ↑ - This PC → Local Disk (C:) → PRC	DMO > V2016 > Code > DSC > SAV	ED_DSC	√ Ō	Search SAVED_DSC		۶
Organize 🔻 New folder					•== •	
🕂 Downloads 🖈 ^ Name ^	Date modified	Туре	Size			
🔮 Documents 🖈	No. Server and					
📰 Pictures 🖈	No items ma	tch your search.				
🔄 DGS 🛛 🖈						
DSC						
PROMO Tutorial						
Results						
🐉 Dropbox						
· ·						
ConeDrive						
💻 This PC						
A Network						
File name: LUNG_DATASET_COLLECTION.mat						_
Save as type: MAT-files (*.mat)						
Hide Folders				Save	Cancel	

Congratulations! You are now the official owner of a dataset collection. Other buttons you might find useful while working on your dataset collection are:

90	Remove the selected dataset from the dataset collection.	
ŝ	Remove the whole dataset collection	

90	Rename the selected dataset.
\square	Load dataset collection from file.
	Intersect common samples (Leave only the samples that appear in all selected datasets).
070	Merge all selected datasets by intersecting the samples and concatenating the data of all of the selected datasets. The result is a unified feature-concatenated matrix.

More advanced features will be discussed further in the tutorial.

Inter-Omic Feature Correlation Analysis

PROMO enables the identification of features that are correlated on two different datasets contained in a dataset collection.

Using the dataset collection described on the previous section (TCGA's lung cancer datasets), we will demonstrate how to identify the most correlated features between the RNA-Seq and Protein datasets.

1. On the 'Dataset Collection' tab, press the inter-omic correlation button on the left panel:



2. In the 'Dataset Selection' window, select the two datasets you wish to use (this type of analysis requires exactly two datasets) and add them to the list using the add button.

Here, we selected and added the RNASeq and Protein datasets.

承 Dat	asets	Selection	_		×
	Plea	Datasets Selection	-		
	Meth	ylation ~	4	dd	
		Dataset Name	D	elete	
	1	RNASeq		elete (
	2	Protein	0)elete	
				ОК	

3. Label based sample filtering - At this point, you can select which samples you wish to remove from the data before continuing the analysis, based on their clinical label values.

This filtering will not affect the original dataset collection. Therefore, you can repeat this analysis using different filters on the original dataset collection.

In this example, we chose to remove all the male samples from the data used in the following analysis.

Dataset Collection Label Filtering	_	
Dataset Collection Label Filtering - Add Filter	×	
Label Based Sample Removal		
	el value	s
Select dataset: RNASeq		~
Select label. gender		\sim
Select value: MALE (249)		\sim
	Add Filt	er
Dataset Name Label Name Label Value	Delet	e
1 RNASeq gender MALE	Delet	e

4. Feature Preprocessing – at this point you can define the number of the most variable features in each dataset to be included in the analysis, and indicate whether to normalize the features before starting the analysis. This preprocessing will not affect the original dataset collection, only the data used in the current analysis.

In this example, we chose to keep the top 2000 features from the RNA-Seq dataset and the top 100 features from the Protein dataset.

Datasets Preprocessing		- 🗆	\times
F	eatures Preprocessing		
1. RNA Seq (20530 features)	Keep top 2000 features with the highest variance	Normalize rows	
2. Protein (131 features)	Keep top 100 features with the highest variance	Normalize rows	
		ОК	

- 5. We will now select the criteria for the analysis.
 - a. Select the type of correlation you want to use. Available options are Spearman or Pearson. In this example, we will use Spearman correlation.
 - b. Define filters to be applied on the results. You can choose to filter by R values, FDR corrected p-values, or keep a predefined number of top results. You can also combine several types of

filtering techniques to make sure all the correlations you will get meet all the conditions specified.

In our example we chose to keep the top 200 correlations, both negative and positive.

r Dataset Correlations	_	
Inter-Dataset Correla	tion Crite	ria
Correlation type:	Spearman	~
Filtering		
R threshold:	0.6	
O Keep values above the threshold		
O Keep values below the threshold		
Keep values whose absolute value is	above the threshold	
FDR Corrected p-Value threshold	1.0e-10	
Keep top 200 Correlations.		
O Keep top positive correlations		
○ Keep top negative correlations		
Keep both negative and positive top co	orrelations	

6. The results of the analysis appear on the 'Dataset Correlation Results' window, which contains the list of top inter-omic correlations as well as various visualizations of the top correlations:

承 Dataset Correlation Results

APCDD1L

25

APCDD1L

Inter Correlation Summary Results Table Correlation Results - Graph Visualization (force) Correlation Results - Graph Visualization (layered) 🖁 📢 🕨

			Top li	nter-Da	ataset (Correla	ations			
	PS1	GS1	MEAN1	PS2	GS2	MEAN2	RHO	P-VAL	Show Plot	_
1	IGFBP2	IGFBP2	11.42	IGFBP2	IGFBP2	0.28	0.9016	2.7071e-60	Show Plots	
2	кп	КП	9.20	СКГГ	СКП	0.75	0.7936	1.3220e-34	Show Plots	i r
3	ADH1B	ADH1B	8.22	CAVEOLIN1	CAVEOLIN1	1.28	0.6409	4.9274e-17	Show Plots	
4	IFNG	IFNG	3.17	CASPASE7C	CASPASE7CLE	0.31	0.6174	2.7904e-15	Show Plots	
5	SCN7A	SCN7A	6.35	CYCLINB1	CYCLINB1	-0.49	-0.6055	1.7528e-14	Show Plots	
6	SFTPC	SFTPC	10.66	CAVEOLIN1	CAVEOLIN1	1.28	0.6034	2.0485e-14	Show Plots	
7	C16orf89	C160rf89	9.77	CYCLINB1	CYCLINB1	-0.49	-0.5910	1.3329e-13	Show Plots	
8	ADAMTS8	ADAMTS8	5.23	CYCLINB1	CYCLINB1	-0.49	-0.5903	1.3329e-13	Show Plots	
9	SOX2	SOX2	7.85	MIG6	MIG6	0.05	-0.5843	2.5783e-13	Show Plots	
10	PCDP1	PCDP1	6.29	CYCLINB1	CYCLINB1	-0.49	-0.5839	2.5783e-13	Show Plots	
11	SUSD2	SUSD2	9.98	CYCLINB1	CYCLINB1	-0.49	-0.5838	2.5783e-13	Show Plots	
12	ABCA8	ABCA8	5.69	CYCLINB1	CYCLINB1	-0.49	-0.5835	2.5783e-13	Show Plots	
13	SFTA1P	SFTA1P	6.38	CYCLINB1	CYCLINB1	-0.49	-0.5747	9.2503e-13	Show Plots	
14	ATP1A2	ATP1A2	4.12	CYCLINB1	CYCLINB1	-0.49	-0.5742	9.3783e-13	Show Plots	
15	OXGR1	OXGR1	2.19	MIG6	MIG6	0.05	-0.5708	1.4542e-12	Show Plots	
16	ADH1B	ADH1B	8.22	CYCLINB1	CYCLINB1	-0.49	-0.5701	1.5151e-12	Show Plots	
17	C7	<u>C7</u>	9.56	CAVEOLIN1	CAVEOLIN1	1.28	0.5683	1.8636e-12	Show Plots	
18	GRIA1	GRIA1	3.29	CAVEOLIN1	CAVEOLIN1	1.28	0.5655	2.6534e-12	Show Plots	
19	C1orf116	C10rf116	11.01	PKCALPHAP	PKCALPHAPS6	0.02	0.5650	2.6964e-12	Show Plots	
20	IL1F7	IL1F7	3.10	PKCALPHAP	PKCALPHAPS6	0.02	0.5630	3.2113e-12	Show Plots	
21	ABCC6	ABCC6	7.30	CYCLINB1	CYCLINB1	-0.49	-0.5627	3.2113e-12	Show Plots	
22	C7	<u>C7</u>	9.56	CYCLINB1	CYCLINB1	-0.49	-0.5626	3.2113e-12	Show Plots	
23	MARK1	MARK1	7.34	MIG6	MIG6	0.05	-0.5623	3.2113e-12	Show Plots	
24	GNLY	GNLY	7.06	CASPASE7C	CASPASE7CLE	0.31	0.5622	3.2113e-12	Show Plots	
	1									

PAI1

0.55

0.5556

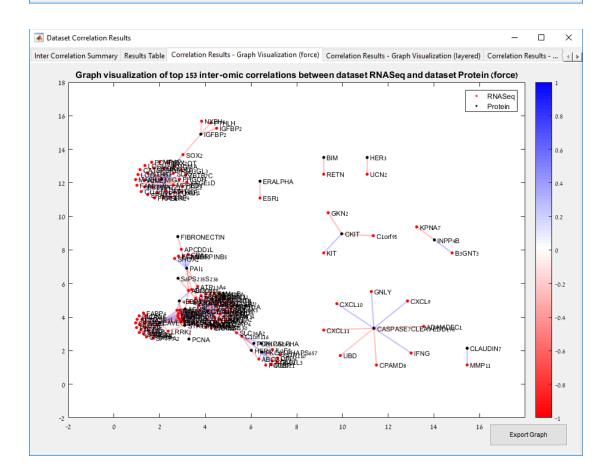
7.6006e-12

Show Plots

v

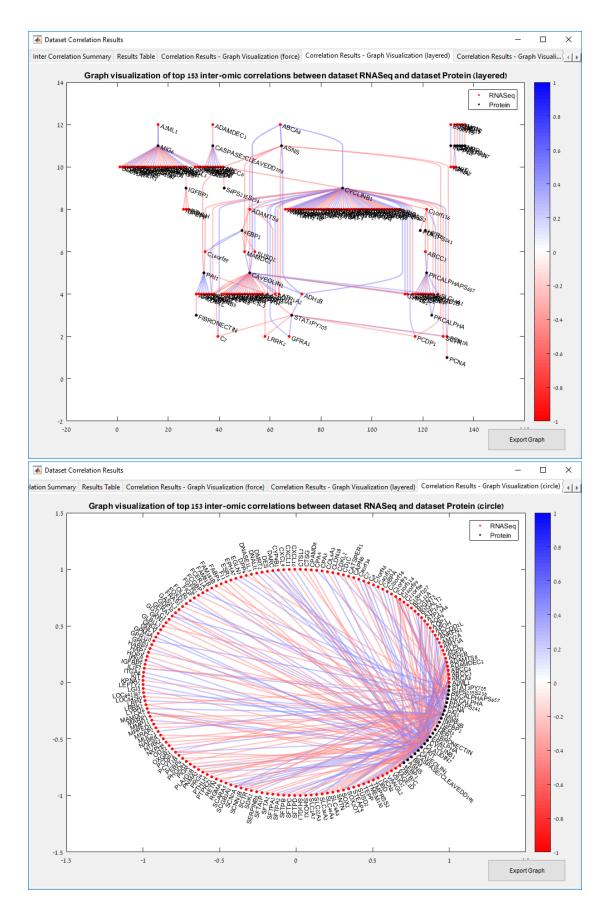
PAI1

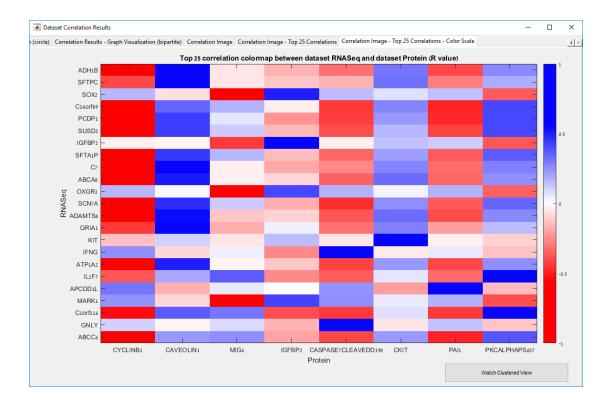
4.46



_

 \times





Multi-Omic Clustering

PROMO offers several methods for multi-omic clustering. Such methods cluster the samples based on several different datasets describing the same set of samples.

Here we show how to apply the multi-omic consensus clustering implemented in PROMO.

Consensus Clustering can be initiated by clicking the Consensus Clustering button on the 'Dataset Collection' tab:

 The next three steps allow selection of the datasets to be included in the analysis, sample filtering and feature preprocessing similarly to the process described in sections 2-4. Notice that this time you can select up to 10 datasets. 2. The next screen enables choosing the parameters for the Consensus Clustering:

Multi-Omic Clustering Parameters Multi-Omic Clusterir	ng Paramete	rs	
Select multi omic clustering algorithm: Consensus	_		~
Internal Clustering Parameters			
Guess number of clusters (k) using:	Silhouette	~]
Maximal K tested:	5]
O Define number of clusters (k):	5]
Internal clustering distance method:	correlation	~	
Number of replicates:	5]
Resampling Parameters			
Number of repeatitions (H):	100		
Sample resampling fraction:	0.8]
Feature resampling fraction:	1]
Similarity Matrix Clustering Parameters			
The clustering method for the similarity matrix:	K-medoids	~]
Number of clusters (k):	3,4,5,6		
			ОК

 Once the clustering completes, a figure displaying the Consensus Clustering results is displayed. The figure includes the original data, iterations matrix, consensus distance matrix and the resulting consensus clustering solution.

Further, a unified feature-concataenated matrix containing all datasets included in the anlsysis will be displayed. In addition, new label representing the assignment of the multi-omic clustering will be added to labels list box.

The following example uses Methylation, Protein and RNA-Seq datasets and K=5 as the final number of clusters:

