



Tutorial

Julien Moehlin, Bastien Mollet, Bruno Colombo, and Marco Antonio Mendoza-Parra
UMR 8030 Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, University of Evry-val-d'Essonne, University Paris-Saclay,
91057 Évry, France.

I - Installation & run

Multilayer is runnable on every os (windows, linux mac osx) with python 3.
Python version 3.8 is recommended.

1 - Dependencies

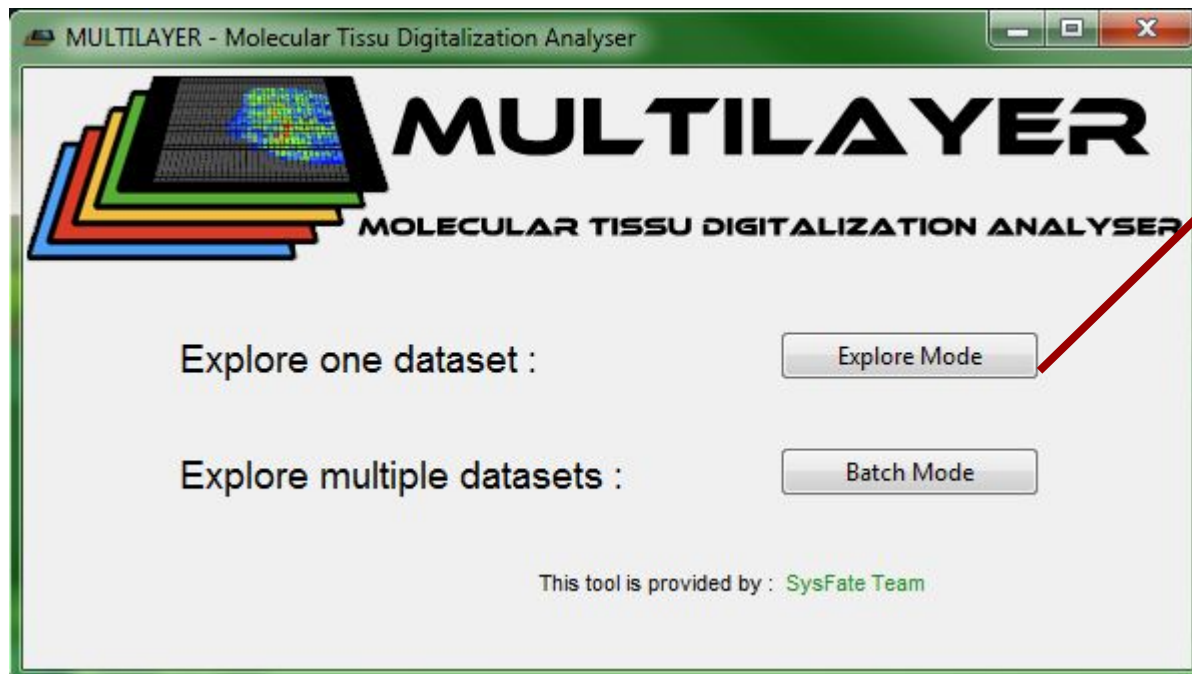
- | | |
|----------------|----------------------------|
| - Numpy | pip install numpy |
| - Matplotlib | pip install matplotlib |
| - Pandas | pip install pandas |
| - Scipy | pip install scipy |
| - Scikit-learn | pip install scikit-learn |
| - Seaborn | pip install seaborn |
| - Networkx | pip install networkx |
| - Louvain | pip install python-louvain |

2 - Run

```
python3 Multilayer.py
```

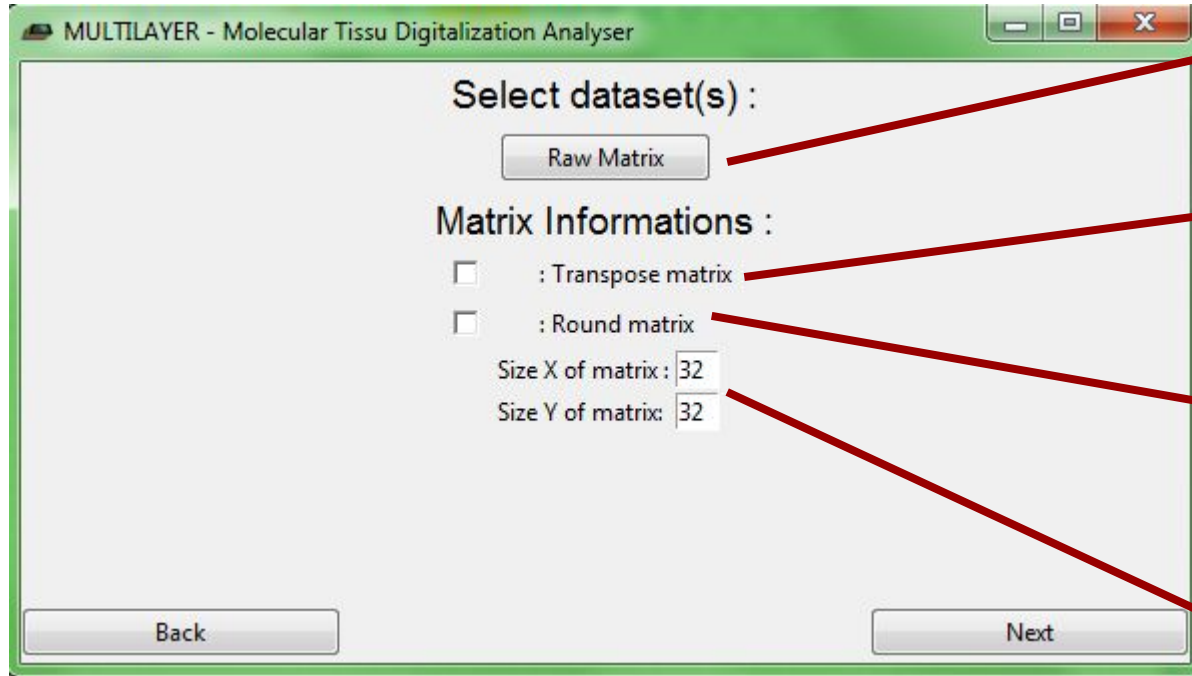
II - Explore molecular tissu with MULTILAYER

II - 1 - Mode



Explore mode : perform analysis over one sample (.tsv file).

II - 2 - Arguments



The screenshot shows a software window titled "MULTILAYER - Molecular Tissu Digitalization Analyser". Inside, there are two main sections: "Select dataset(s) :" and "Matrix Informations :". Under "Select dataset(s) :", there is a button labeled "Raw Matrix". Under "Matrix Informations :", there are two checkboxes: "Transpose matrix" and "Round matrix", both of which are currently unchecked. Below these checkboxes are two input fields: "Size X of matrix :" with the value "32" and "Size Y of matrix :" with the value "32". At the bottom of the window, there are two buttons: "Back" on the left and "Next" on the right. Four red arrows point from the explanatory text on the right to specific elements in the window: the first arrow points to the "Raw Matrix" button, the second to the "Transpose matrix" checkbox, the third to the "Round matrix" checkbox, and the fourth to the "Size X of matrix" and "Size Y of matrix" input fields.

Select dataset(s) :

Raw Matrix

Matrix Informations :

☐ : Transpose matrix

☐ : Round matrix

Size X of matrix : 32

Size Y of matrix: 32

Back

Next

Provide your dataset (.tsv file). The correct format is a matrix with genes name as rows, coordinates (coordinates should be in this format : 'XxY') as columns.

Transpose the dataset is an option. You can use this option if you have a matrix with genes name as columns and coordinate as rows.

Round matrix is an option. The tool needs integer as X & Y (coordinates). This option will round coordinate. For example : a given coordinate '3.36x13.94' will be converted to like this : '3x13'.

Indicate the **size** of dataset. The maximum of X and the maximum of Y. Default value is 32.

MULTILAYER - Molecular Tissu Digitalization Analyser

Arguments for differential gene expression :

Threshold up regulated : 1 ▼
Threshold down regulated : -1 ▼

Patterns detection :

Minimum n° of contiguous gexels : 10

Similarity Methods :

☒ Tanimoto
☐ Dice

Multiprocessing :

Number of thread : 1 max (4).

Back Next

The tool will perform a differential gene expression analysis. Those thresholds are for define up-regulated genes and down-regulated genes.

Pattern detection : The tool will use agglomerative clustering on gexels (<https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html>) for determine a contiguous pattern. The tool will consider a pattern if the number of gexels is upper or equal than this value. Default value is 10.

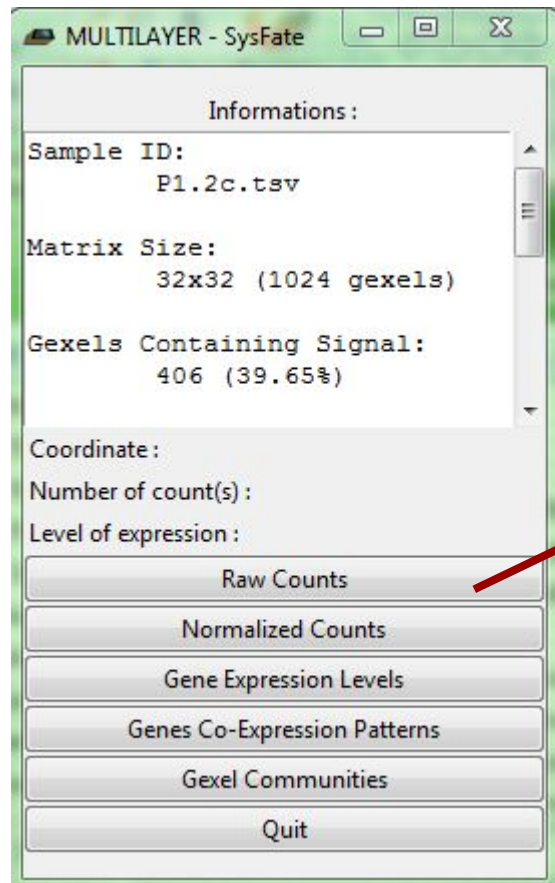
Similarity methods will calculate similarity between patterns.

With Tanimoto, the similarity will be calculated with $A \cup B / A \cap B$.

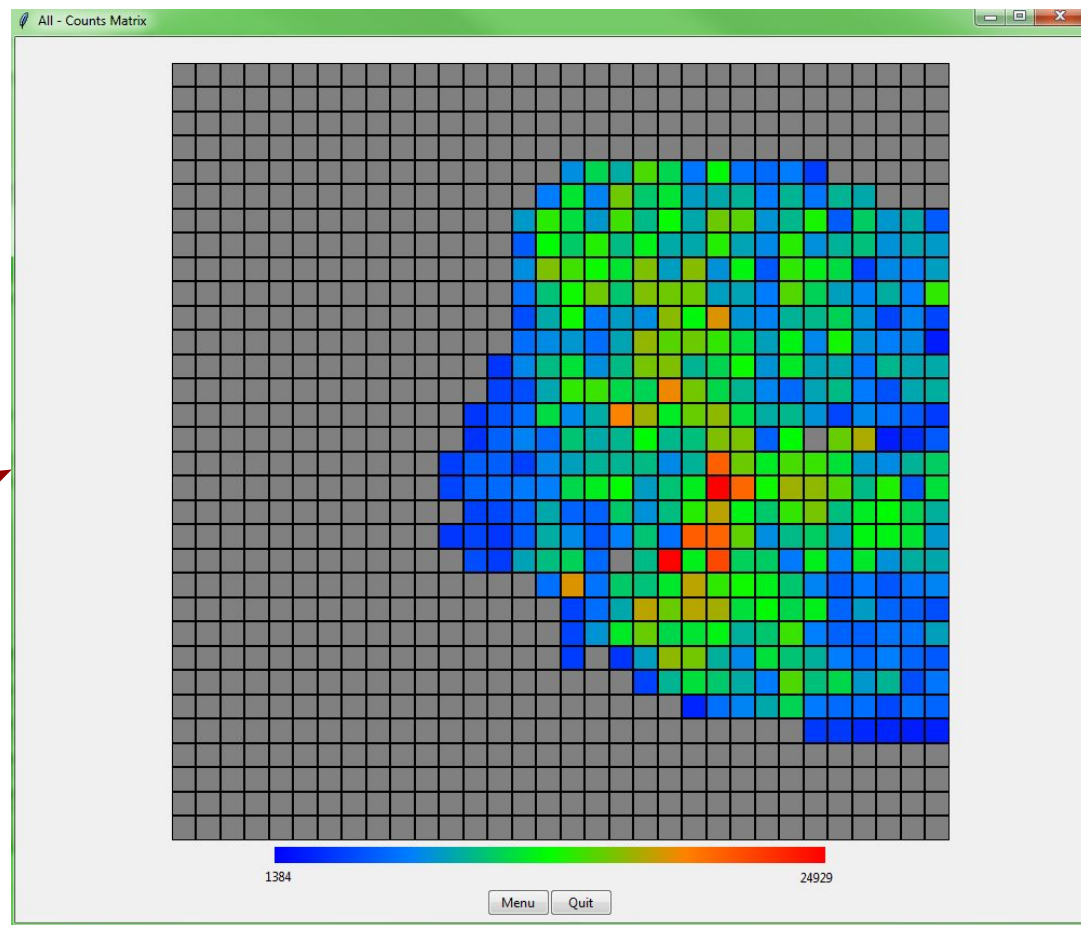
With Dice the similarity will be calculated with $2 * (A \cap B) / (A \cup B)$.

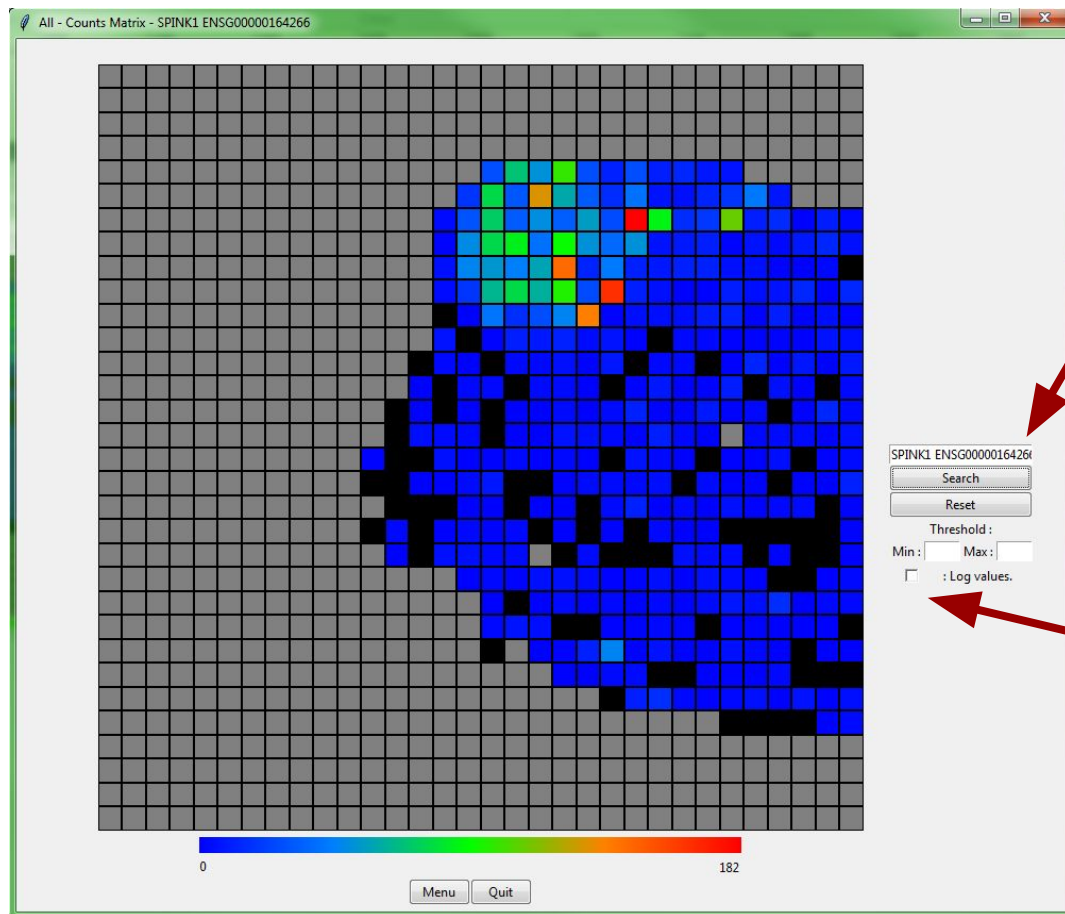
Multiprocessing : provide the number of threads for the analysis (always available for the Batch mode)

II - 3 - Raw dataset



'Raw counts'
shows the input
dataset.



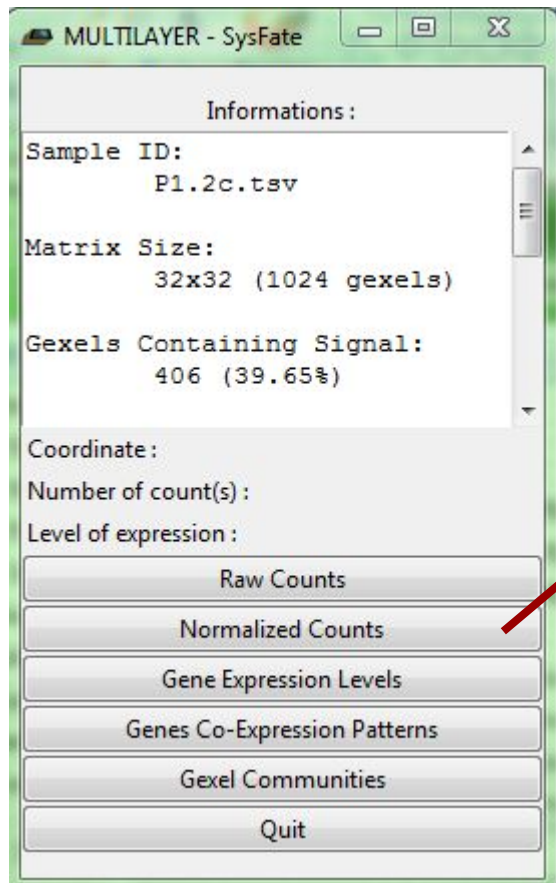


The right panel appears when you hit the menu button.

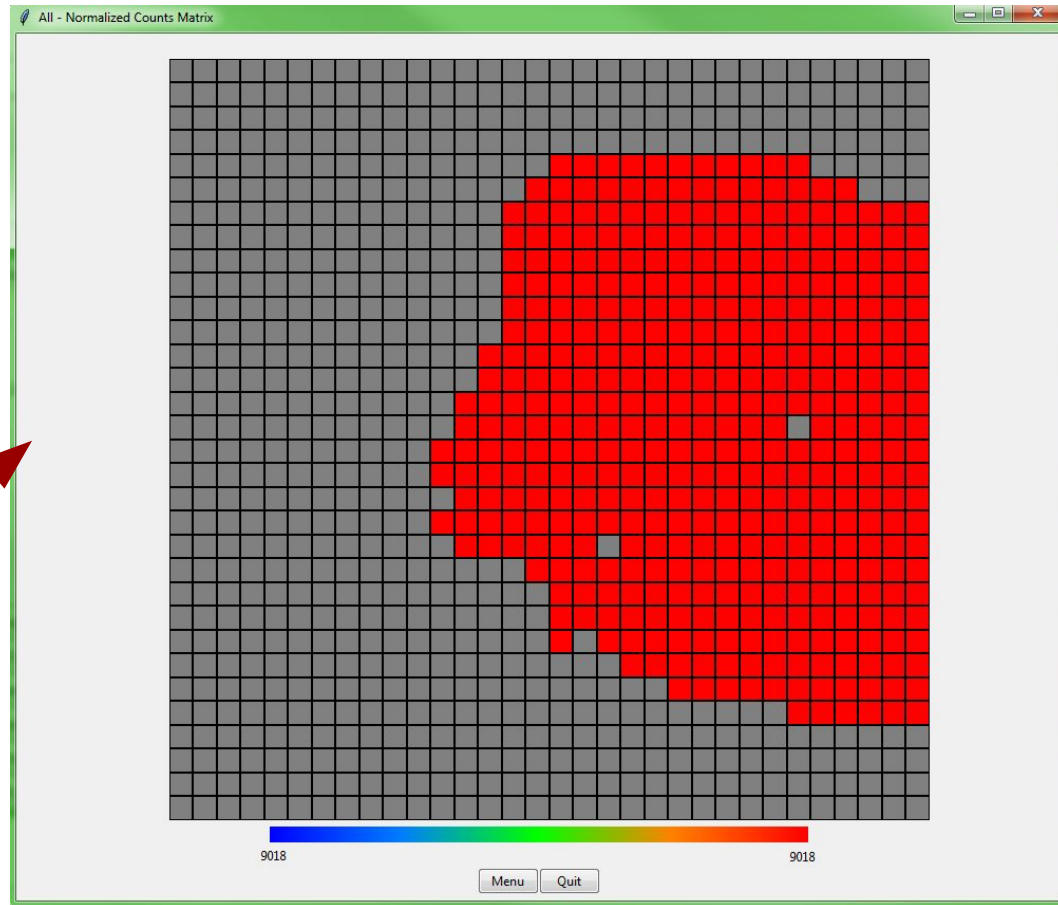
You can query a gene.

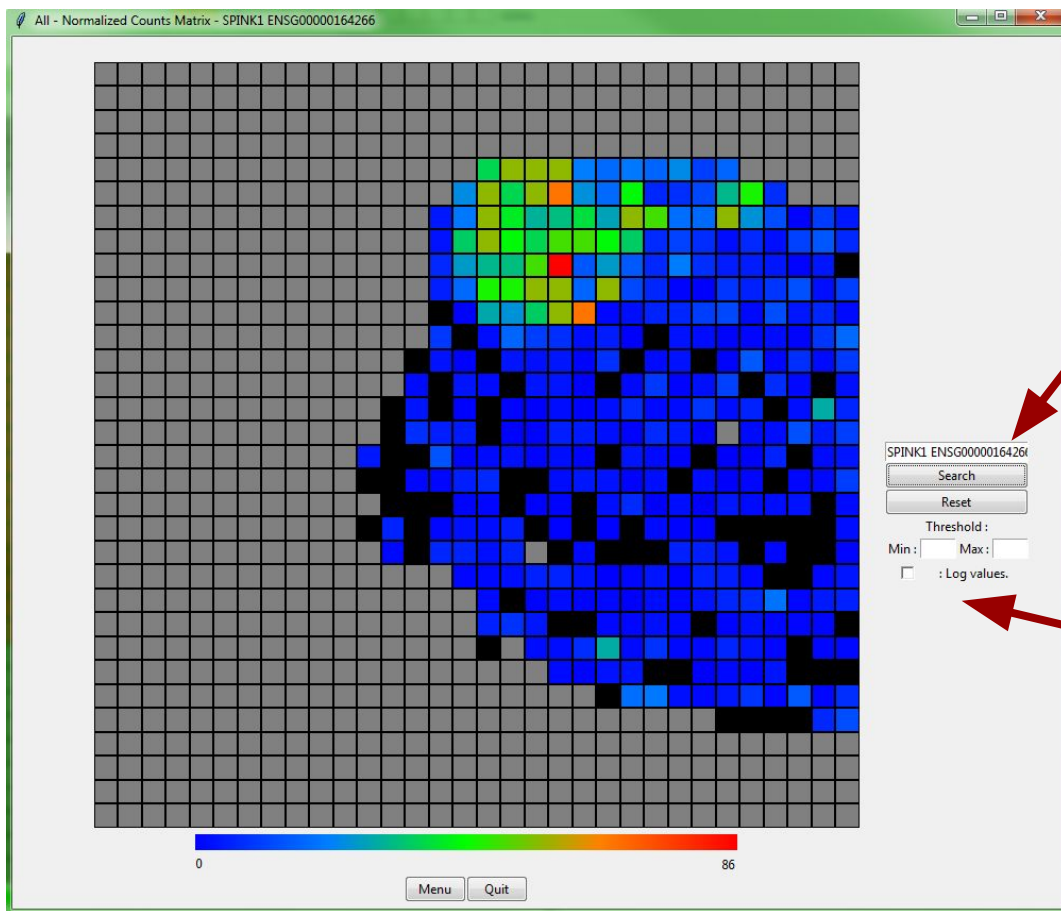
Display gene with log values

II - 3 - Normalized dataset



'Normalized counts' shows the normalized dataset.



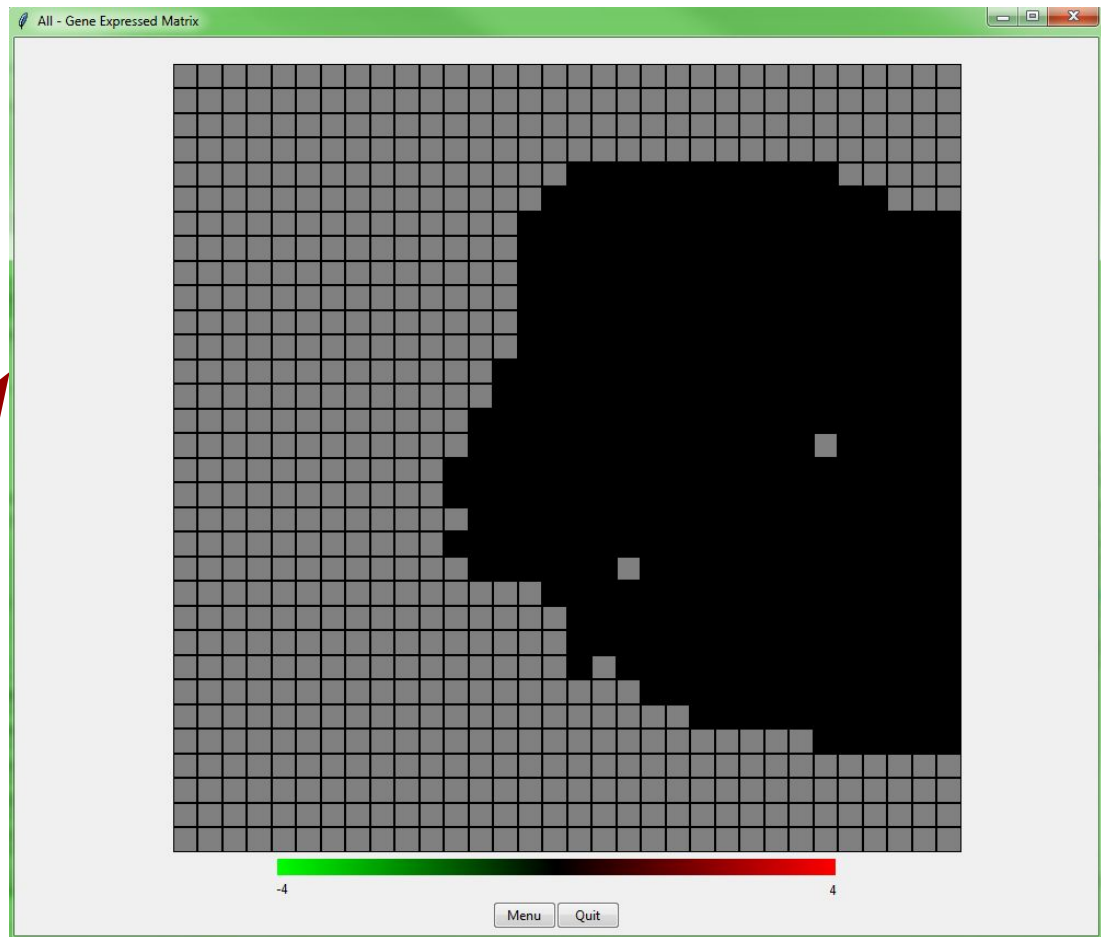
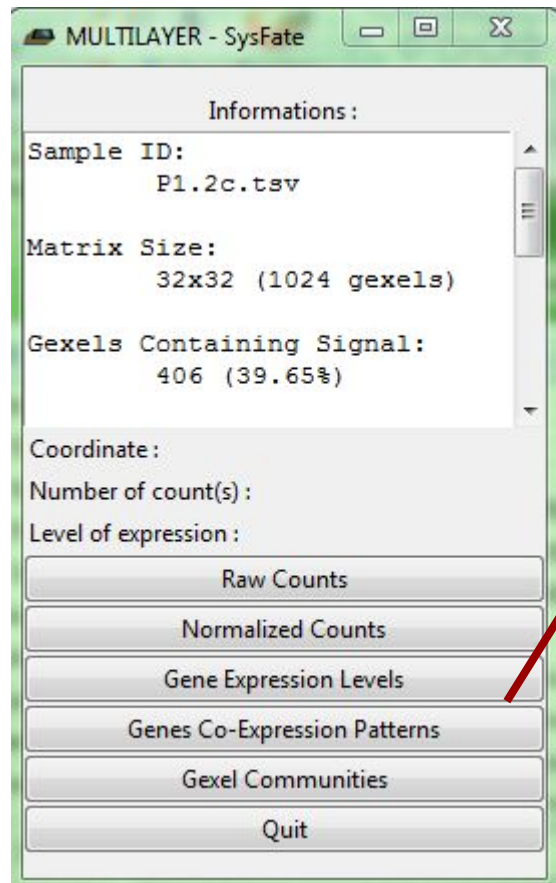


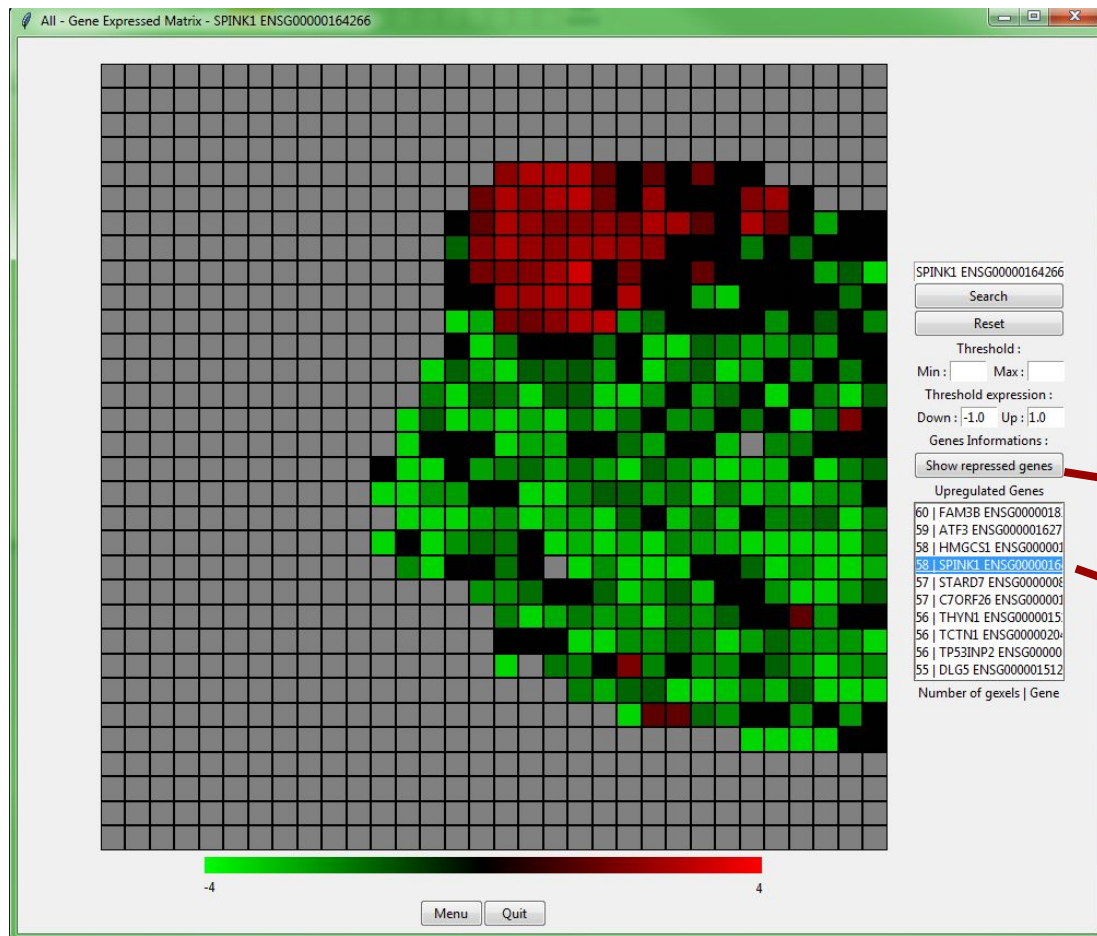
The right panel appears when you hit the menu button.

You can query a gene.

Display gene with log values

II - 3 - Gene expressed levels dataset



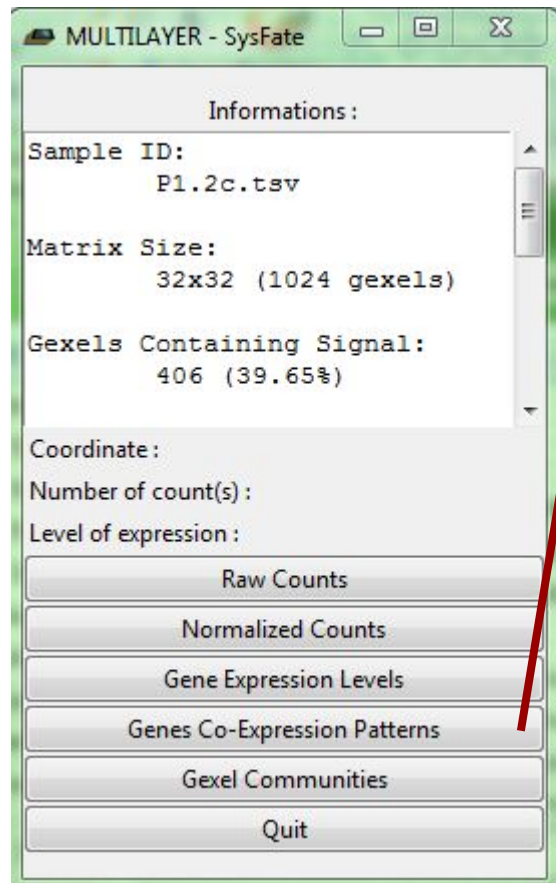


You can query a gene.

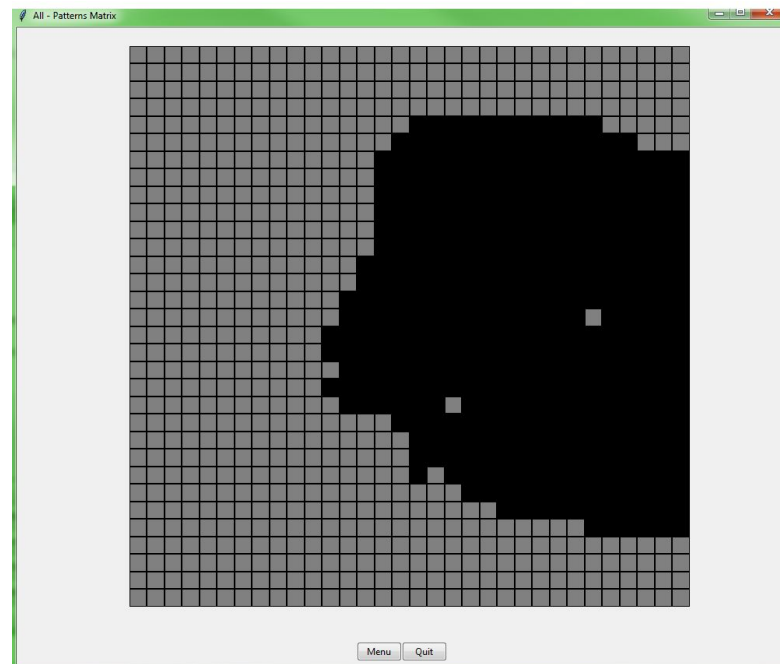
The button 'show repressed genes' show the top ranking down-regulated genes based on the number of gexels.

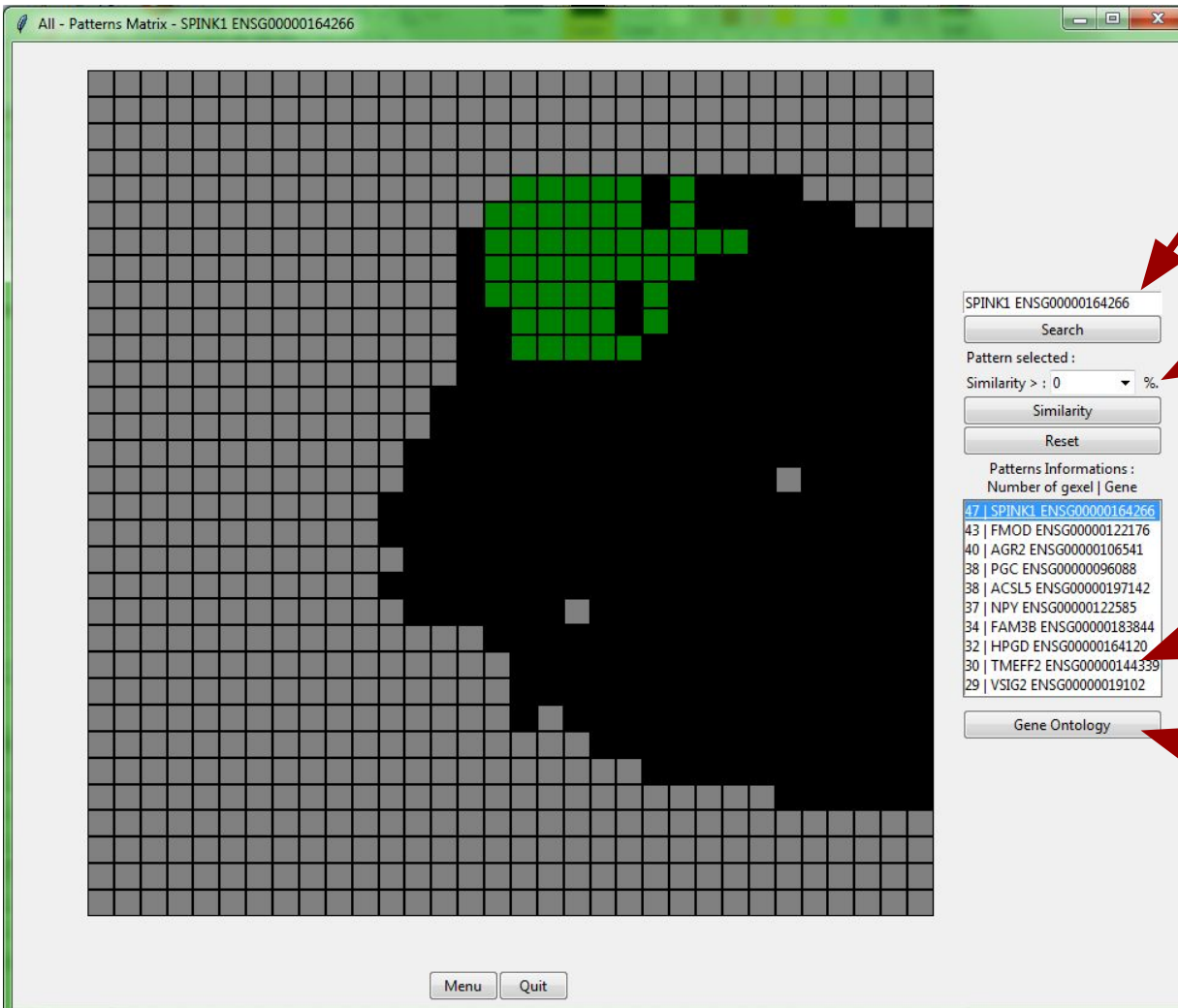
Top ranking of up-regulated genes based on the number of gexels.

II - 4 - Co-expression Patterns



The value is displayed here is the default value with whom the analysis was made. If you change this value, the tool will perform another agglomerative clustering for determine new clusters (this step can take a little time).





You can query a gene.

Pattern of co-expressed genes analysis based on their similarity.

Top ranking genes based on the number of gexels per pattern.

Gene ontology analysis.

II - 4 - A - Pattern of co-expressed genes analysis

SPINK1 ENSG00000164266

Search

Pattern selected :

Similarity > : 0

Similarity

Reset

Patterns Informations :
Number of gexel | Gene

47	SPINK1 ENSG00000164266
43	FMOD ENSG00000122176
40	AGR2 ENSG00000106541
38	PGC ENSG00000096088
38	ACSL5 ENSG00000197142
37	NPY ENSG00000122585
34	FAM3B ENSG00000183844
32	HPGD ENSG00000164120
30	TMEFF2 ENSG00000144339
29	VSIG2 ENSG00000019102

Gene Ontology

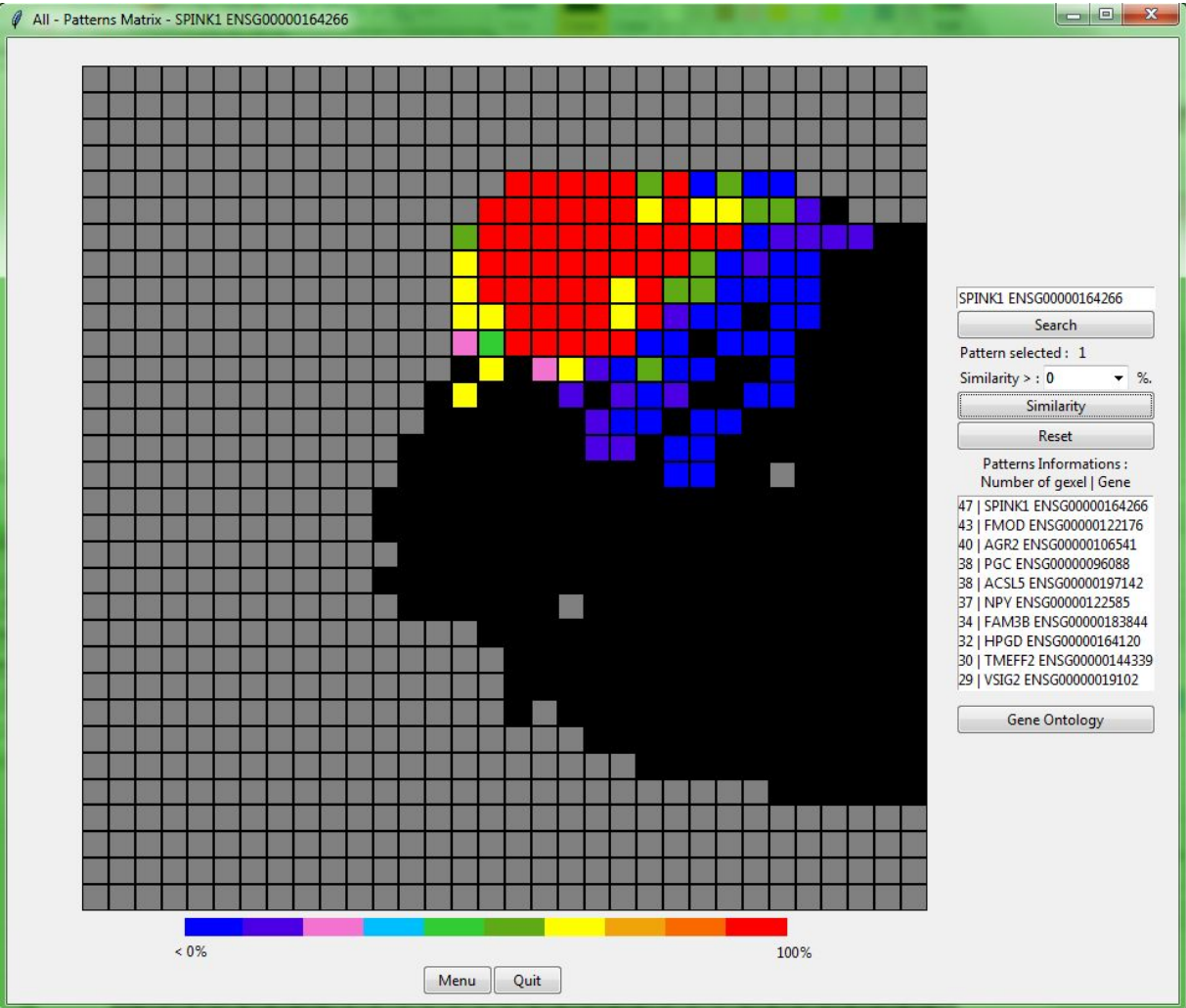
First step:

Select a pattern of genes by clicking on one gexel.

The selected pattern will appear here.

Second step:

Select a threshold of similarity and press 'Similarity button'.



Result of Co-Expression Pattern analysis for SPINK1 ENSG00000164266

40	30	20	10	0
GD ENSG000000	KCNN4 ENSG000000	ORM2 ENSG000000	THYN1 ENSG000000	CFD ENSG0000001
F3 ENSG0000001	LIPF ENSG0000001	CAB39L ENSG000000	TNS1 ENSG000000	NOMO2 ENSG000000
LIM5 ENSG000000	ADAMTSL1 ENSG	APOD ENSG000000	AHL1 ENSG0000001	THBS4 ENSG000000
BS4 ENSG000000	ORM1 ENSG000000	PROC ENSG000000	CST1 ENSG0000001	DHR57 ENSG000000
CAPD3 ENSG000000	RP11-314O13.1 E	MFAP4 ENSG000000	LCN2 ENSG000000	EFNA1 ENSG000000
CA ENSG0000000	HMGCS2 ENSG000000	STARD10 ENSG000000	PIGR ENSG0000001	IGF2 ENSG0000001
DN3 ENSG000000		HGS ENSG0000001	CLN3 ENSG000000	COL1A1 ENSG000000
ACR2B ENSG000000		TRPM4 ENSG000000	COTL1 ENSG000000	C3 ENSG000000125
		KLK4 ENSG0000001		MGP ENSG0000001
		WNK2 ENSG000000		GSTP1 ENSG000000
		CYP27A1 ENSG000000		C15 ENSG00000018
		RAP1GAP ENSG000000		
		TMSB10 ENSG000000		
		REG4 ENSG0000001		
		ATF5 ENSG0000001		

Save Quit

The result of the co-expression pattern analysis

List of genes in each intervals of similarity.

Save the liste in tsv file by clicking on 'save' button.

II - 4 - B - Pattern of co-expressed genes - Gene ontology

SPINK1 ENSG00000164266

Search

Pattern selected :

Similarity : [10, 0[%.

Similarity

Reset

Patterns Informations :
Number of gexel | Gene

47		SPINK1	ENSG00000164266
43		FMOD	ENSG00000122176
40		AGR2	ENSG00000106541
38		PGC	ENSG00000096088
38		ACSL5	ENSG00000197142
37		NPY	ENSG00000122585
34		FAM3B	ENSG0000018384
32		HPGD	ENSG00000164120
30		TMEFF2	ENSG000001443
29		VSIG2	ENSG00000019102

Gene Ontology

Gene Ontology Analysis - All - Patterns Matrix - SPINK1 ENSG00000164266

GO terms	P-Values	Genes
----------	----------	-------

Select a GO terms database and press 'Run'.
Genes can be copied to clipboard by double-click on the table.

ARCHS4_Cell-lines

Run Save

Barplot Heatmap Quit

Select a GO database. The user can provide his own database (paste the file with the correct format in the directory called 'GO_DB').

Once a database selected, click on 'run' button (this step can take few seconds).

Gene Ontology Analysis - All - Patterns Matrix - SPINK1 ENSG00000164266

GO terms	P-Values	Genes
Adenoma	4.606729889305634e-10	AGR2 APOE EFNA1 KLK4 C3 HGS API
Adenocarcinoma	3.5534874027347053e-09	AGR2 TMSB10 APOE ANPEP EFNA1
Prostatic Neoplasms	2.0138140432939682e-08	AGR2 STARD10 APOE SPON2 KLK4 K
Malignant neoplasm of prostate	3.3494919909225766e-08	AGR2 STARD10 APOE SPON2 ANPEP
Precancerous Conditions	2.9563078179287126e-07	AGR2 IGF2 PTGDS GSTP1 MGP SPINK
Malignant tumor of colon	3.29519650115297e-07	APOE SPON2 ANPEP CLN3 KLK4 C3
Colon Carcinoma	7.075737610983052e-07	APOE SPON2 ANPEP CLN3 EFNA1 KI
Colorectal Carcinoma	4.55792525570373e-06	AGR2 APOE SPON2 EFNA1 KLK4 KEA
Liver carcinoma	6.1097097777728139e-06	ATF5 AGR2 PROC APOE SPON2 ANP
High-Grade Prostatic Intraepithelial	6.599754624149162e-06	AGR2 GSTP1 APOD SPINK1 PSCA
Vascular inflammations	9.066017026042678e-06	KEAP1 THBS4 APOE ELF3 GJB1 LCN2
Hepatitis	1.214695051421898e-05	KEAP1 C1S GSTP1 APOE CYP27A1 PE
Neoplasm Metastasis	1.321232645743605e-05	AGR2 TMSB10 APOE ANPEP DHRS7
Stomach Neoplasms	2.213155065216265e-05	IGF2 REG4 GSTP1 APOE CLDN3 CST1
Prostate carcinoma	2.296127210839665e-05	AGR2 APOE ANPEP DHRS7 EFNA1 KL

Select a GO terms database and press 'Run'.
Genes can be copied to clipboard by double-click on the table.

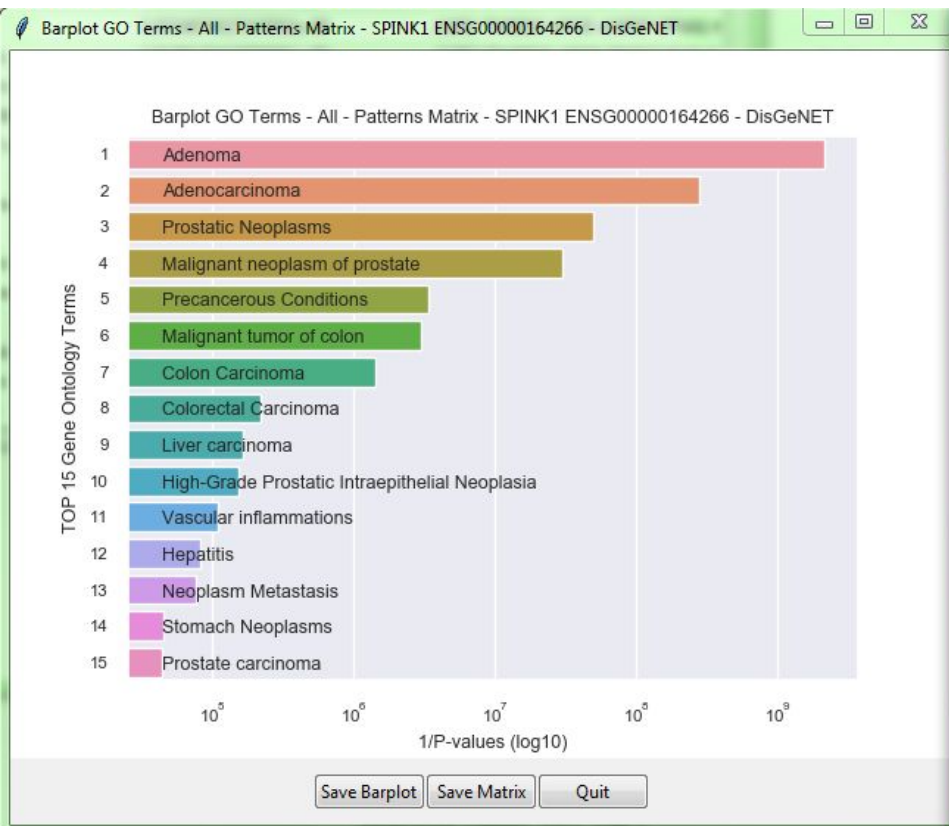
DisGeNET

Run Save

Barplot Heatmap Quit

Save the result (the table in 3 columns) in a tsv file by clicking on 'save' button.

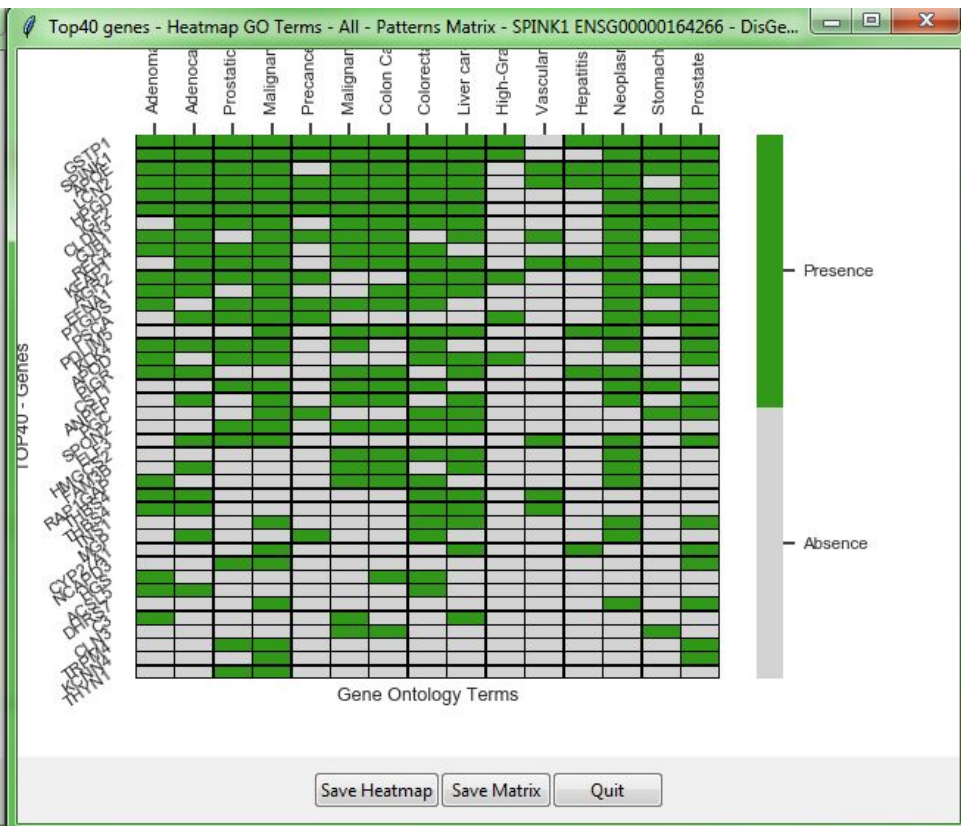
Barplot and heatmap will be introduced on the next page.



Barplot shows TOP 15 best GO terms based on p-values.

Save plot by clicking on 'save barplot'

Save the full matrix by clicking on 'save matrix'.

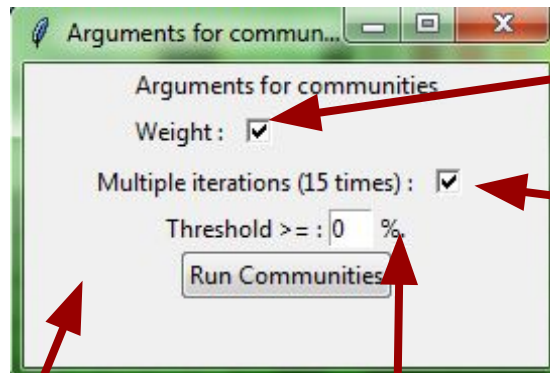
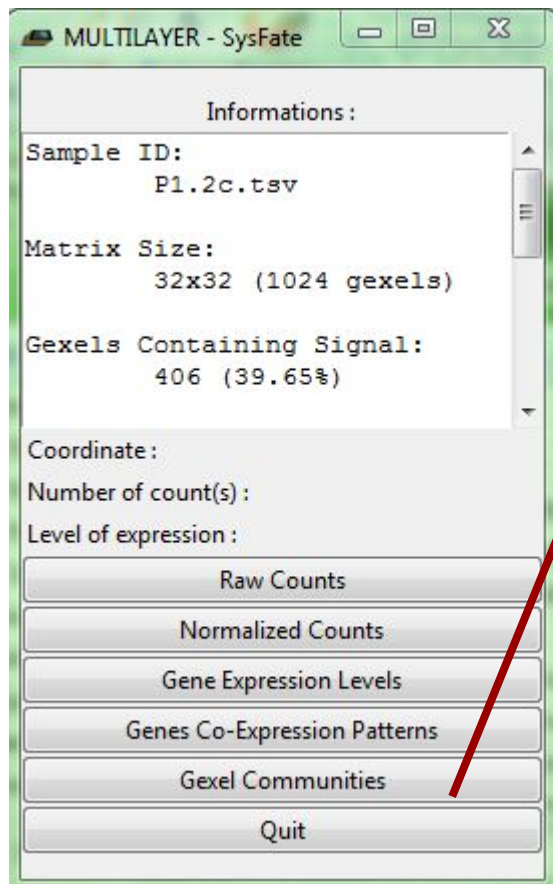


Heatmap shows TOP 15 best GO terms and the TOP 40 genes.

Save plot by clicking on 'save heatmap'.

Save the full matrix by clicking on 'save matrix'.

II - 5 - Gexel communities

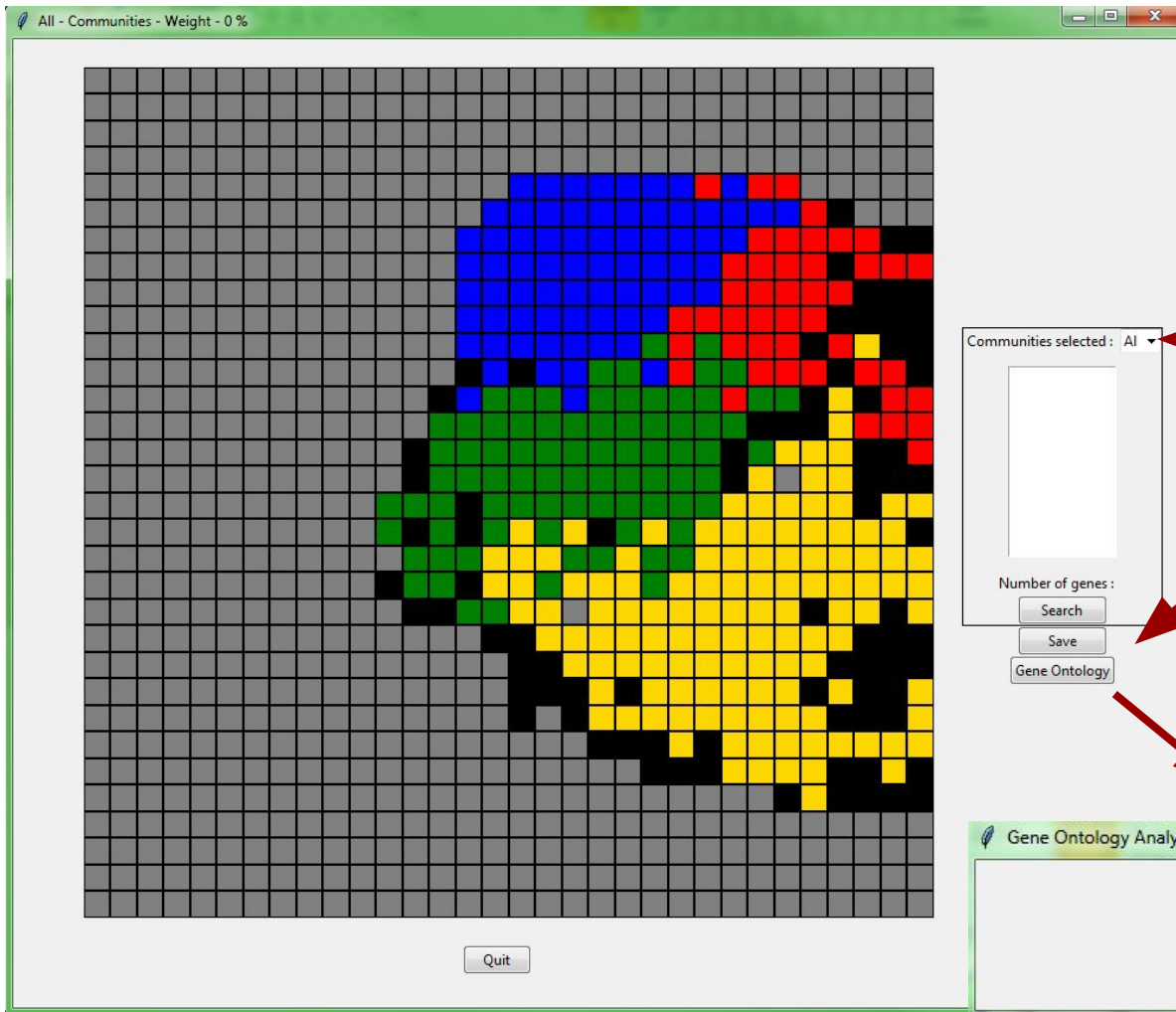


Consider the similarity weight as parameter in louvain algorithm.

Multiple iterations, perform 15 times louvain algorithm and keep the more frequent. Result will be show in the console.

Filter genes by similarity.
Default value is 0, all genes will be consider.

```
#####  
# LOG #  
#####  
1 [*] 80.0%  
2 [ ] 13.333333333333334%  
3 [ ] 6.666666666666667%
```



All communities are displayed. If there are overlaps between communities, the tool will display the gexel with community with the highest similarity.

Communities selected : All

Display all communities or a specific one.

Save a tsv file (result of louvain)

Number of genes :

Search

Save

Gene Ontology

Select database and 'run'.

Gene Ontology Analysis - All - Communities - Weight - 0 %

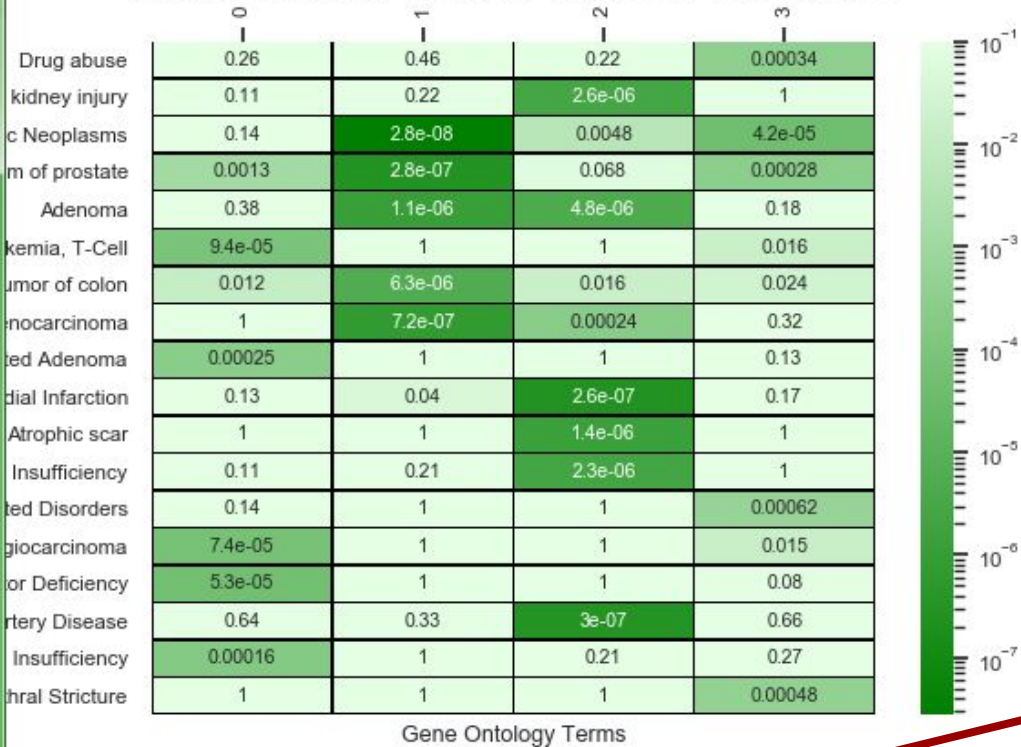
Select a GO terms database and press 'Run'.

DisGeNET

Run

Quit

Heatmap GO Terms - All - Communities - Weight - 0 % - TOP5 - DisGeNET



Gene Ontology Terms

Save Heatmap

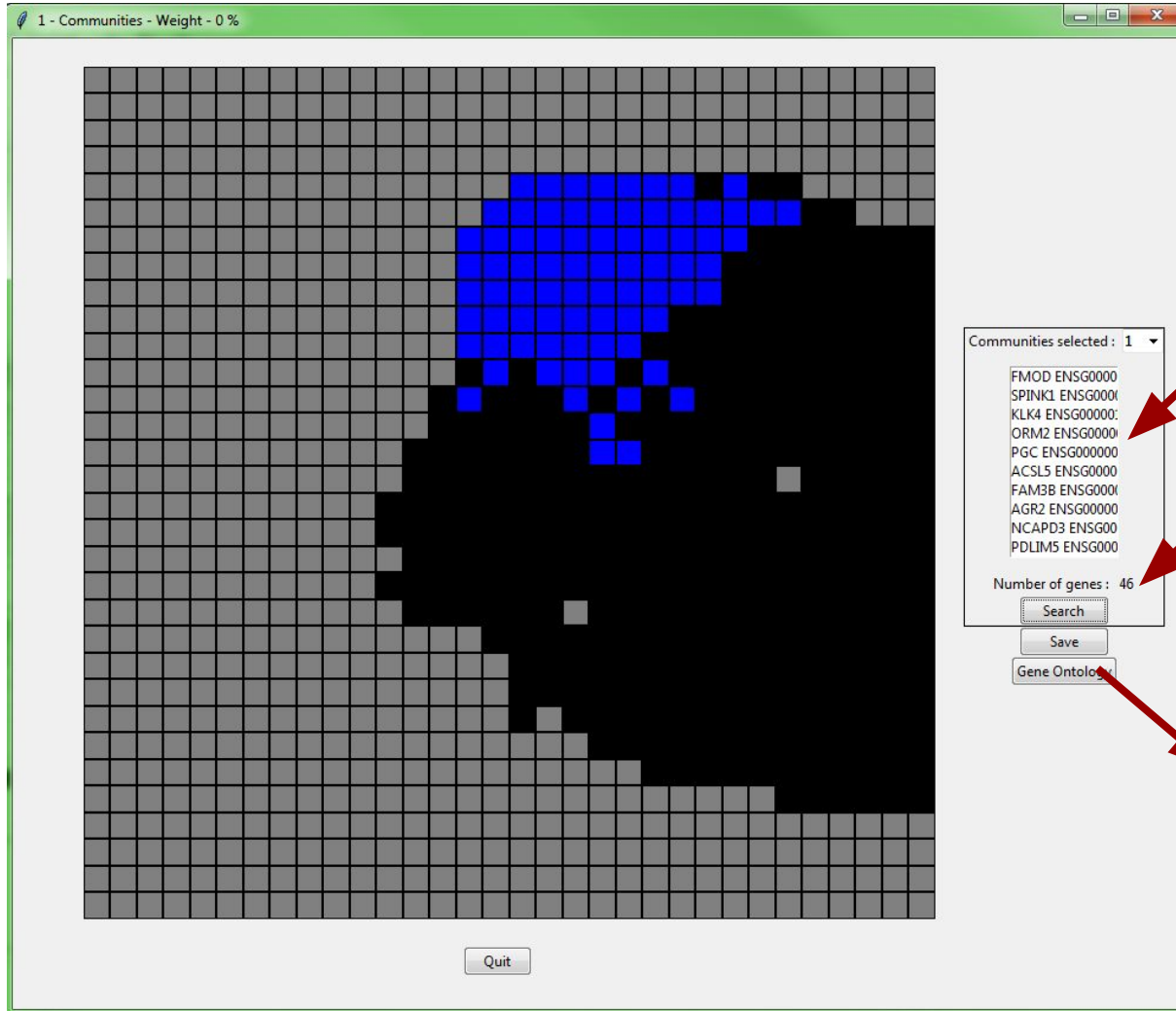
Save Matrix

Quit

For each community, get the top 5 best GO terms.

Save plot by clicking on 'save heatmap'

Save matrix by clicking on 'save matrix'



A REFAIRE

All communities are displayed. If there are overlaps between communities, the tool will display the gexel with community with the highest similarity.

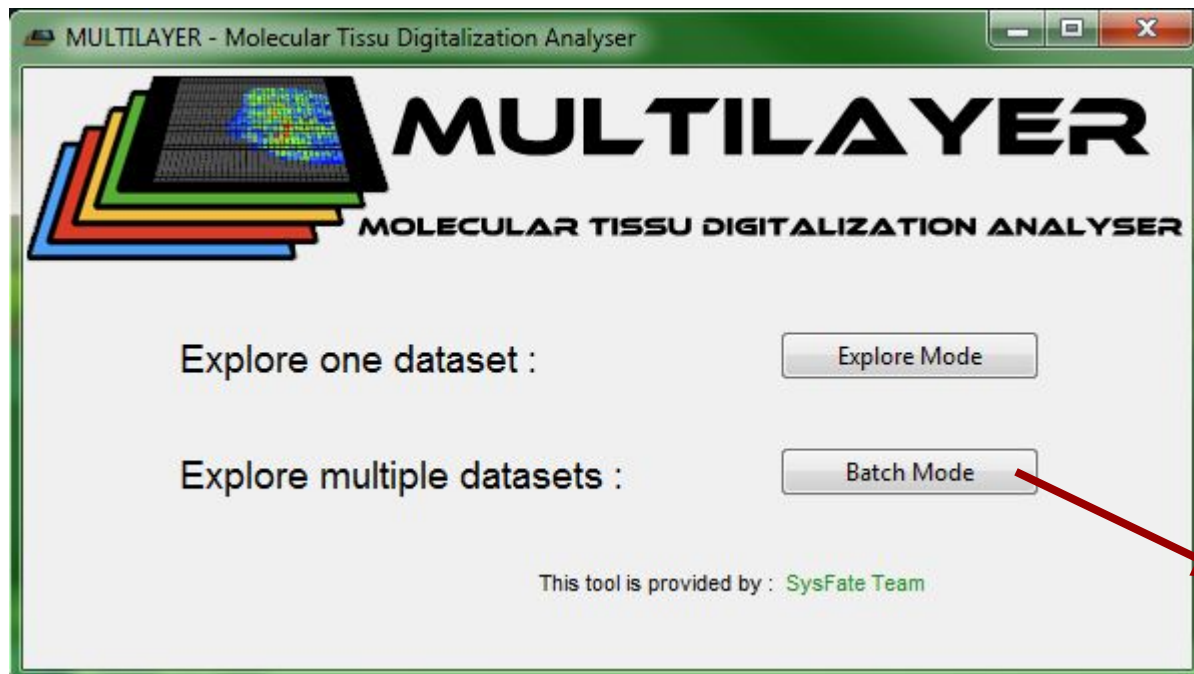
If a specific community is selected, genes which compose the given community will be display.

Toltal number of genes in the community

Same as parttern of co-expressed genes. Go back to pages 16 & 17.

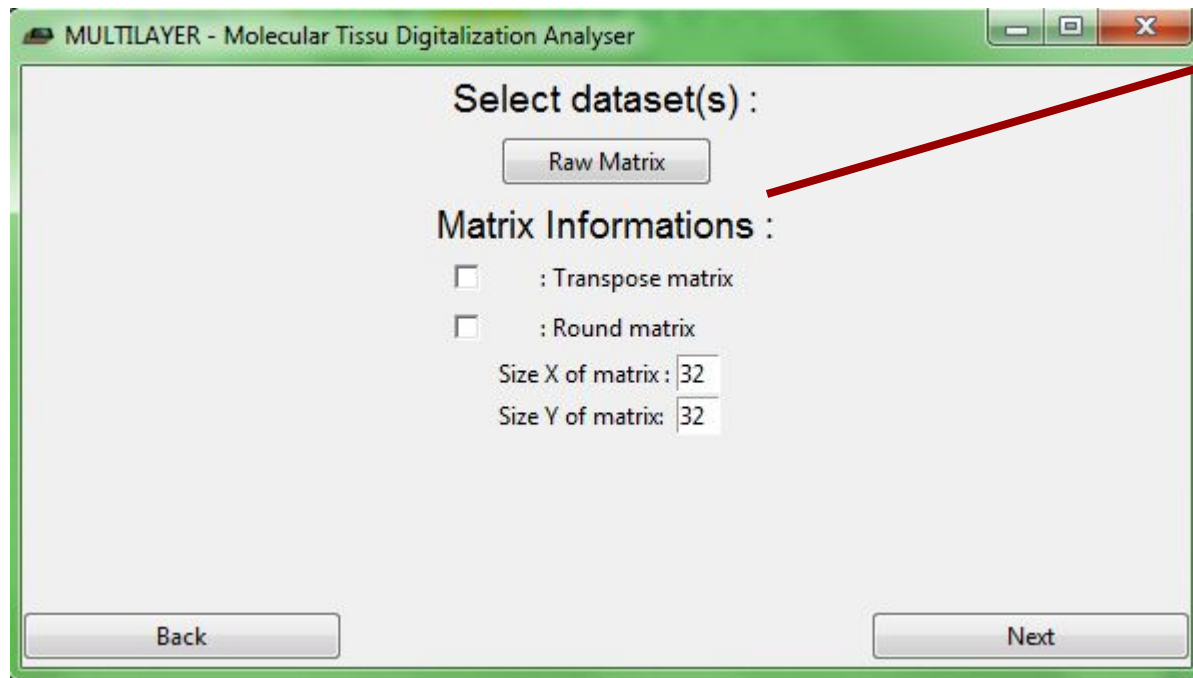
III - Batch mode with MULTILAYER (multiple samples)

III - 1 - Mode



Explore mode : perform analysis over several samples (.tsv files)

III - 2 - Arguments



MULTILAYER - Molecular Tissue Digitalization Analyser

Select dataset(s) :

Raw Matrix

Matrix Informations :

☐ : Transpose matrix

☐ : Round matrix

Size X of matrix : 32

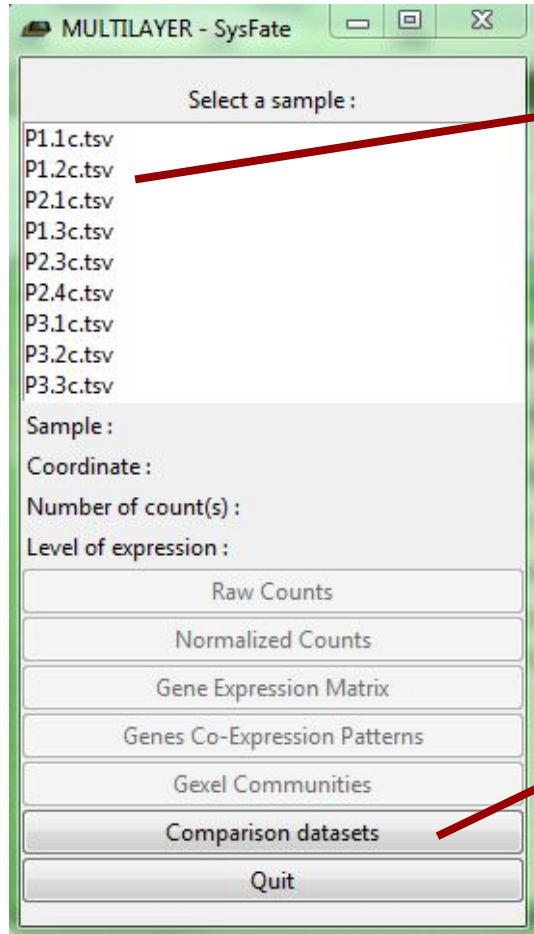
Size Y of matrix: 32

Back

Next

Same as Explore mode (page 4)

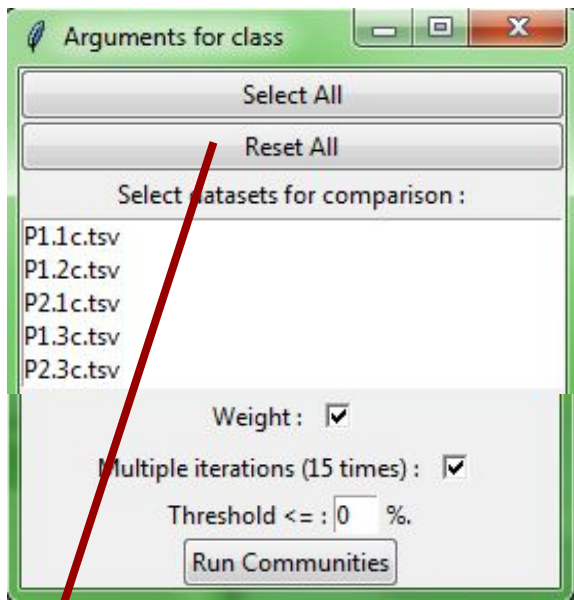
The single difference :
By clicking on raw matrix, you will choose a directory. **All tsv files in the selected directory will be consider as a input Matrix.**



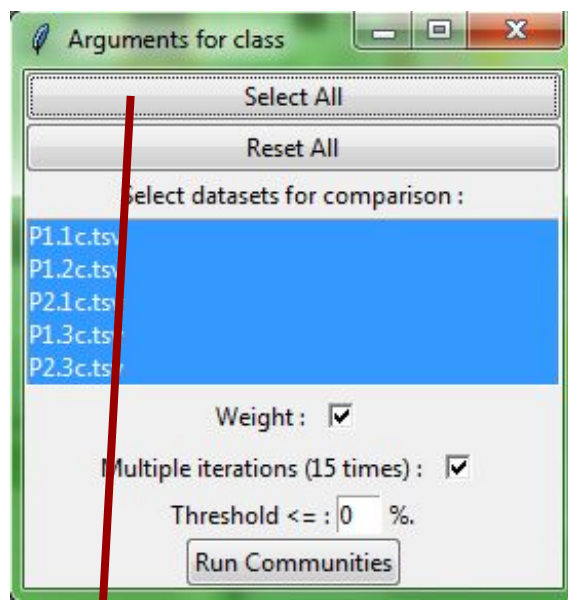
Select a specific sample for perform all previous analysis from pages 6 to 21.

Important Note : Raw counts and normalized counts are not stock in memory, they will be calculate every each click on 'raw counts' and 'normalized counts' buttons.

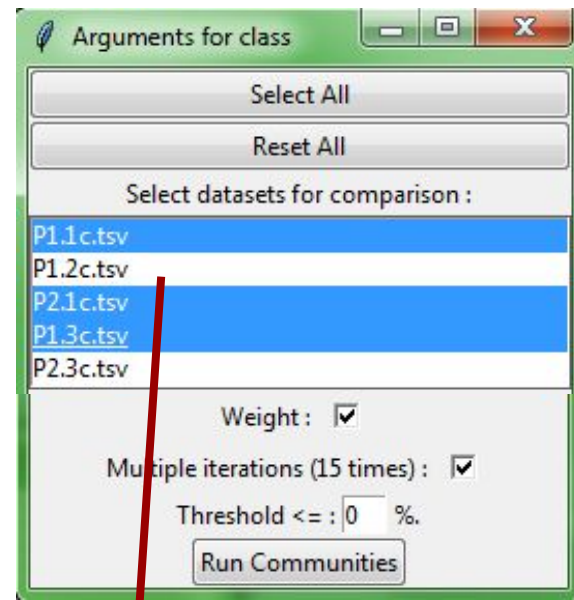
Comparison of all communities from all datasets.



Reset all selection

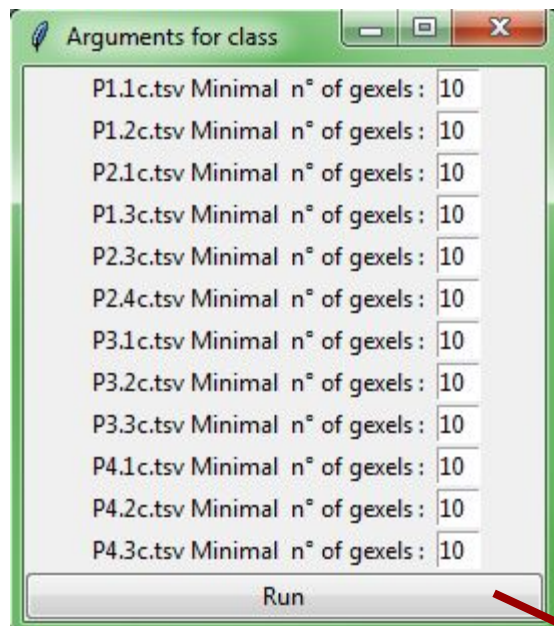


Select all datasets



Select several specifics datasets

Important note : Arguments Weight, Multiple iterations and Filter are commented page 18.

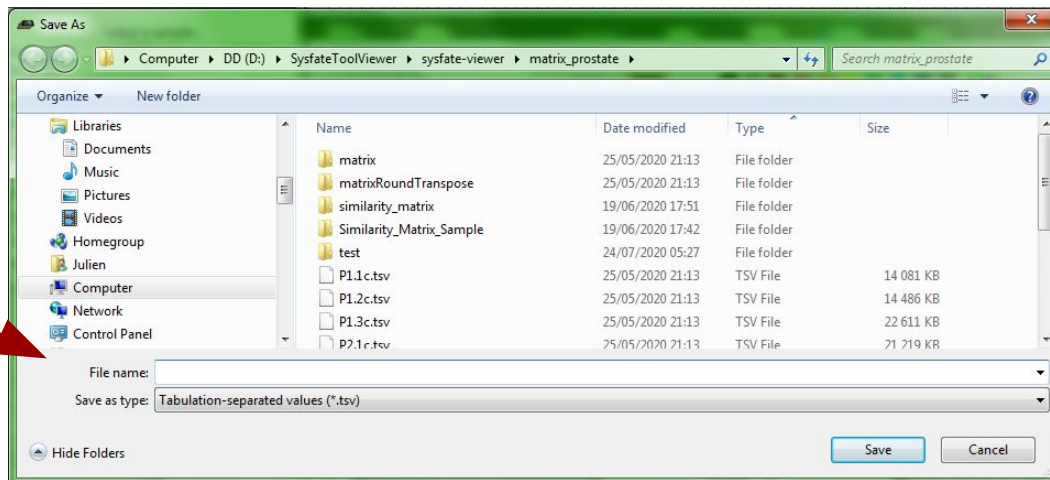


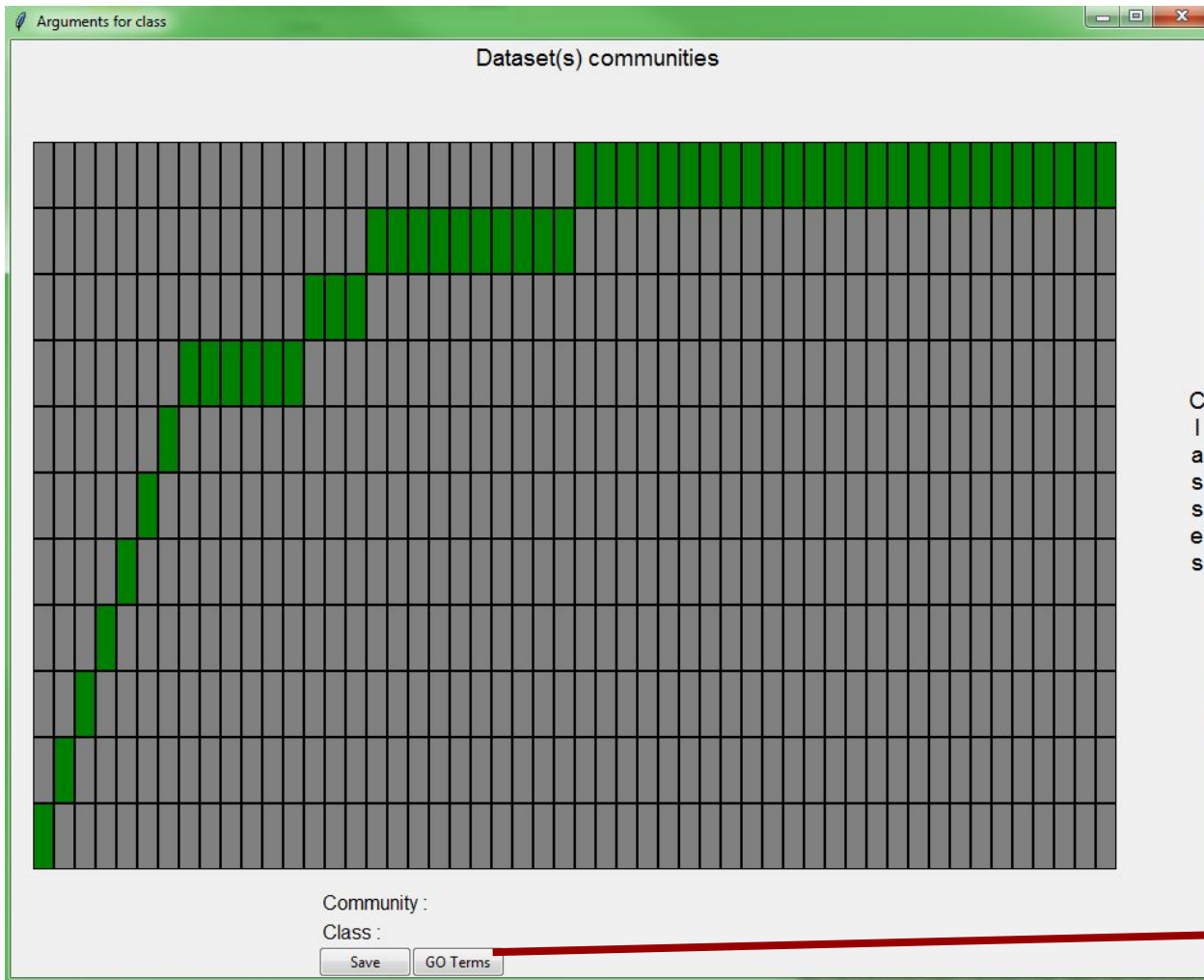
The displayed value is the default value.

For the analysis, you can change minimal number of gexels for define patterns for each datasets separately.

If you change the value, the tool will compute with new value (this step can take a time).

You have to provide a name for saving files. The tool will save 2 files (ex: save.tsv, save_filter.tsv)





Interactive heatmap :

In X axis all datasets communities, in Y axis all classes.

A class is group of 1 or several communities.

GO terms

Gene Ontology Analysis - Class

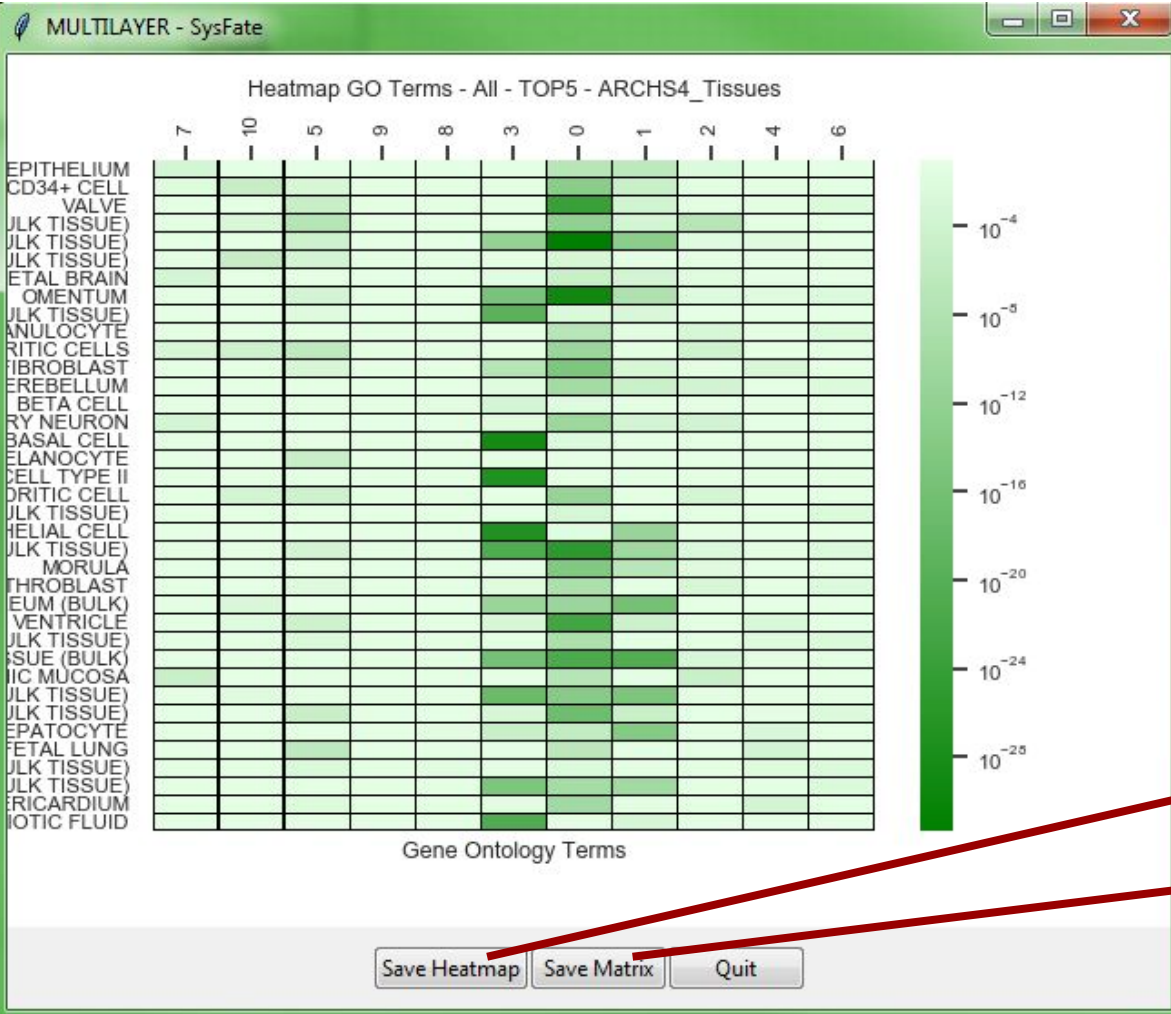
Select a GO terms database and press 'Run'.

Top GO terms : 5 ▼

DisGeNET ▼ Run

Quit

Select the number of GO terms selected.
Select the database.



IV - MULTILAYER compressor

For high resolution dataset, we developed module for compress data. Multilayer compressor is able to merge several gexels in one big gexel. The format has to be in 3 columns like : barcodes (XxY) with header 'bc' - Genes with header 'gene' - Genes counts with header 'count'

Multilayer_Compressor.py has several arguments :

-i : input matrix.

-o : output matrix.

-cx : compressor factor.

-cy : (optional) compressor factor for Y. If -cy is not define -cy will be equal to -cx.

python3 Multilayer_Compressor.py -i input.tsv -o output.tsv -cx 100 -cy 100

In this given example : 10 000 gexels will be merge in 1 gexel.