Omics Playground Documentation

Jan 26, 2024

GETTING STARTED

Omics Playground is a comprehensive self-service platform platform for visualization, analytics and exploration of Big Omics Data. It allows users to apply a multitude of state-of-the-art analysis tools to their own data to explore and discover underlying biology in a short time.

The platform offers a unique combination of features that distinguishes it from the other analytics platforms currently available. We believe that data preprocessing (primary analysis) and statistical testing (secondary analysis) are now well established, and the most challenging task is currently data interpretation (tertiary analysis) that often takes the longest time but where actual insights can be gained. Therefore Omics Playground focuses strongly on tertiary analysis while providing good support for secondary analysis.

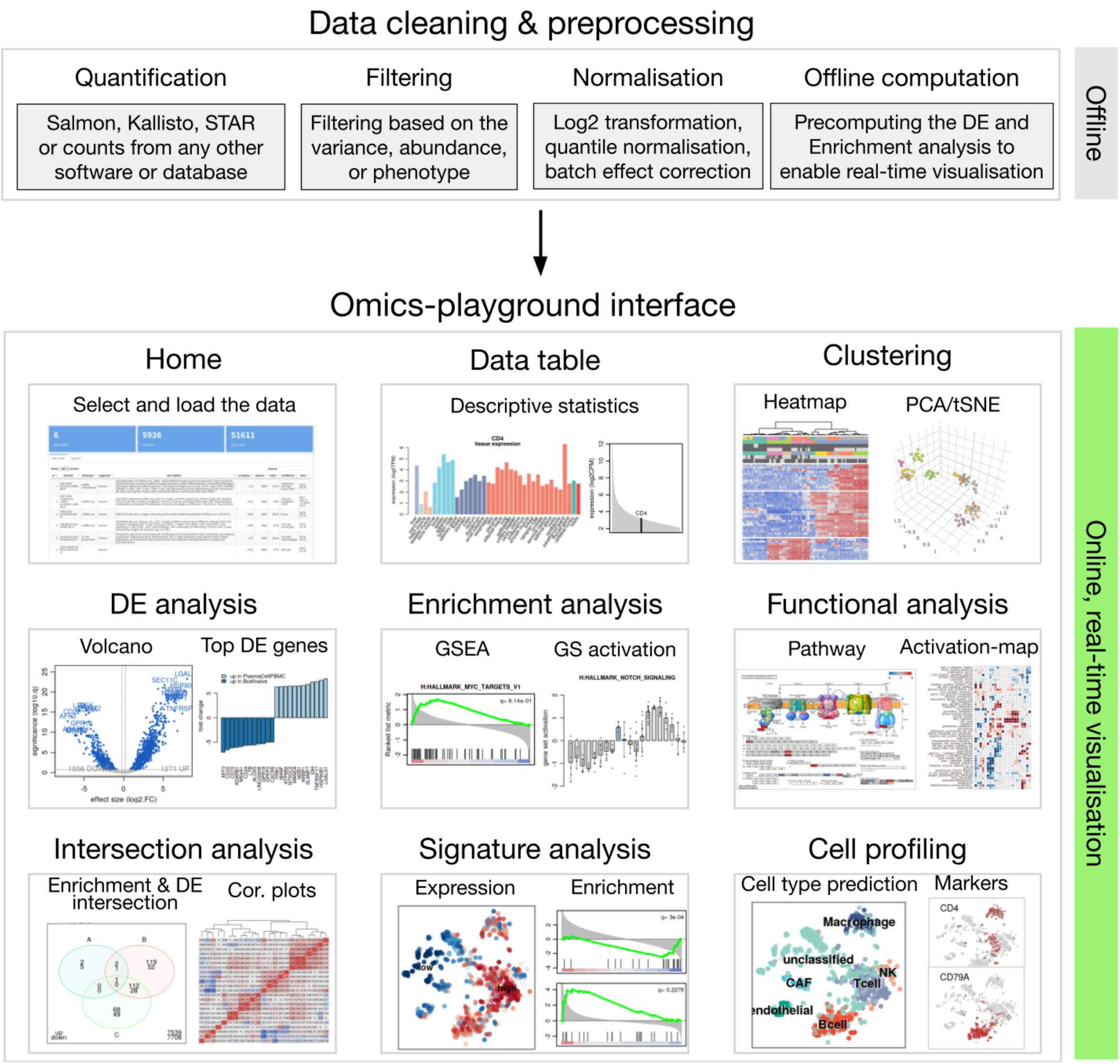
CHAPTER ONE

OVERVIEW

Omics Playground is a user-friendly and interactive web-based platform for the analysis and visualization of transcriptomics and proteomics data. Currently the platform handles any kind of transcriptomics and proteomics data annotated at the gene/protein level, and supports two species, human and mouse. Omics Playground has been in particular devised to also support single cell RNA-seq data, as well as traditional gene expression experiments.

The overview of the platform is shown in the figure below. It consists of two main components. The first component addresses the data importing and preprocessing, which includes preparation of the input data, filtering, normalisation and precomputation of statistics for some analyses. The second part is composed of the online interface, which supports the real-time visualisation and interaction with users. The interface contains several optional settings in order to provide a customisable experience suited to each user's background.

The platform can be accessed through the following link: https://bigomics.ch/omics-playground/


Data cleaning & preprocessing
Quantification
Salmon, Kallisto, STAR or counts from any other software or database
Filtering
Filtering based on the variance, abundance, or phenotype
Normalisation
Log2 transformation, quantile normalisation, batch effect correction
Offline computation
Precomputing the DE and Enrichment analysis to enable real-time visualisation
Offline
Omics-playground interface
Home
Select and load the data
Data table
Descriptive statistics
Clustering
Heatmap
PCA/tSNE
DE analysis
Volcano
Top DE genes
Enrichment analysis
GSEA
GS activation
Functional analysis
Pathway
Activation-map
Intersection analysis
Enrichment & DE intersection
Cor. plots
Signature analysis
Expression
Enrichment
Cell profiling
Cell type prediction
Markers
Macrophage
unclassified
CAF
NK
Tcell
endothelial
Bcell
CD4
CD79A
Online, real-time visualisation

CHAPTER

TWO

INTRODUCTION

The platform requires the transcriptomics and proteomics data to be in a structured format as an input. The easiest way is to prepare two csv files: **counts.csv**, **samples.csv** and an optional **contrast.csv**. When these files are ready, users can upload them direcly in Omics Playground.

2.1 Input file requirements

1. *Counts file* Count or expression .csv file with gene on rows, samples as columns.
2. *Samples file* Samples .csv file with samples on rows, phenotypes as columns.
3. *Contrasts file* Contrast .csv file with samples on rows, conditions as columns. (optional)

See also:

If you have raw file formats, such as FASTQ files or LC-MS proteomics data (mzML, RAW, WIFF…), check our tutorials on how to prepare the counts matrix from these raw formats: *data preparation examples*.

CHAPTER THREE

SAMPLES FILE

The samples file (*samples.csv*) contains the phenotypic information of each sample. The first column contains the sample name, which must be unique, and has to match the name given in the read counts file.

The samples file is a tabular (csv) file with the samples in the rows and the phenotypic data (metadata) in the columns. Note that the platform will not accept purely numerical values as phenotypes.

If we are analyzing a human study (it can be applied to any study) as seen in the `samples.csv` table below, the rows should be anonymized patients, identifyied uniquely by the first column (sample1, sample2…), and the other columns would be sample metadata or phenotypes (hair color, country, weight, age, etc.).

	hair_color	country	age
sample1	blond	Japan	old
sample2	dark	Switzerland	young
sample3	blond	USA	young
sample4	dark	Switzerland	old
sample5	dark	USA	old

As mentioned above, the age was converted from numeric (12, 52, 87) to young and old, since the platform currently does not support continuous values.

Note: All phenotypes must contain at least one alphabet letter. This is done to avoid continuous values (as in the case of weight), since the platform expects discrete ranges. Having excessive numbers of phenotypic groups may also result in errors.

See also:

If you are familiar with R, you can think of the samples file as a data.frame object. We provide an example samples file that can be accessed by installing playbase `devtools::install_github("bigomics/playbase")` and running `playbase::SAMPLES`.

CHAPTER

FOUR

COUNTS FILE

The file 'counts' contains the measurements (genes, proteins, etc..) for each sample listed in the samples file. Just like the samples, the `counts.csv` file is tabular (.csv), where each row describes the features (genes, proteins, etc..) and each column describes the samples.

The rows contains gene IDs, which can be in most common formats (such as HGCN or Ensembl), but not in the Entrez number format. If you are using Entrez numbers, please convert them to Ensembl IDs using tools such as Syngo.

The values should always be numerical, with the exception of "NA" in case of a lack of data. Failure to do so will result in an error.

Below is a simple example of how a `counts.csv` file should look like.

	sample1	sample2	sample3	sample4	sample5
gene1	543.6	1556.1	413.0	887.9	123.4
gene2	6.5	14.7	2.3	42.4	56.7
gene3	10.4	763.5	NA	0	89.0
gene4	3217.4	0	4983.2	7493.8	210.2
gene5	98770.5	113498.0	498351.6	88134.1	345.6
gene6	0	NA	14.9	0	789.0
gene7	47648.8	0	32682.0	93873.2	123.4

Note: The formats accepted as features (genes, proteins are ENSEMBL, ENSEMBLTRAN, UNIGENE, REFSEQ, ACCNUM and UNIPROT and gene SYMBOL). Also note that the platform will not accept transcript IDs. You will need to convert them to Gene IDs. This will result in multiple gene entries that the platform will merge.

See also:

If you are familiar with R, you can think of the counts file as a data.frame object. We provide an example samples file that can be accessed by installing playbase `devtools::install_github("bigomics/playbase")` and running `playbase::COUNTS`.

CHAPTER

FIVE

CONTRASTS FILE (OPTIONAL)

The contrasts are an optional input in Omics Playground. It contains the groups (treatment versus controls, mutant versus wild-type, etc..) that will be tested against each other. The contrasts are a list of comparisons between groups.

There are two ways to define the contrasts: the long contrast form and the short contrast form. The contrast name rows need to be unique, both in short and long forms.

5.1 Sample-wise contrast (long form)

The long contrast form has the same number of rows of the samples file. That means the first column is the sample names (matching samples file), and the second column (onwards) contains the phenotypes.

The first row contains the name of the pairwise comparisons. All pairwise contrasts names must follow the format shown below with the groups joined together by "_vs_" (e.g. piperaquine_vs_control).

..note::

1. All comparison names in the header **must** contain the separator "_vs_"
2. Please use the exact condition names in the title as well as in their columns

For example, if you have 3 groups (A, B, C), you can test the following contrasts: A_vs_B, A_vs_C, B_vs_C.

Here is a minimal example of how the `contrasts.csv` should look like. In this case, the groups will be hair color (blond vs. dark) and country (Japan vs. Switzerland), as derived from the samples file. Use NA (or empty field) to skip samples that do not belong to a contrast.

	dark_vs_blond	japan_vs_switzerland
sample1	blond	japan
sample2	dark	switzerland
sample3	blond	japan
sample4	dark	switzerland
sample5	dark	NA

The column names (dark_vs_blond) will be used to name the comparisons. The first name (dark) will be used as the numerator of the contrast, and the second name (blond) will be used as the denominator of the contrast. Alternatively, you can code the groups with -1 (reference group), 1 (main group) and 0 (not in comparison). This is shorter than writing the entire word but you have to be careful which group to assign +1 and -1 to. The -1 refers to the 'control' group that is the denominator or 'on-the-right' of the '_vs_' in the comparison name.

	dark_vs_blond	japan_vs_switzerland
sample1	-1	1
sample2	1	-1
sample3	-1	1
sample4	1	-1
sample5	1	0

Any of the tables above can be provided to the platform. The zero means the sample5 will not be include in the comparison of japan_vs_switzerland.

5.2 Group-wise contrasts (short form)

Sample-wise contrasts, as introduced in the previous paragraph, become tedious in datasets with large number of samples, as assigning the contrasts to each sample can be cumbersome. Instead, you can specify group-wise contrasts (or short form). However, this approach only works best if users are focusing on a single phenotype in their dataset.

Following the example above, if we would like to create a contrast between the two countries Japan and Switzerland, we will need to create a column called **group** in the sample file, which will contain the phenotypes (in this case country). To simplify, we could simply change the name of the column **country** to **group**.

Once we have the group column in the sample file, we can assign the contrasts as in the table below.

	japan_vs_switzerland	USA_vs_switzerland	all_vs_switzerland
japan	1	0	1
USA	0	1	1
switzerland	-1	-1	-1

We will search for the group column in the samples file, and we will create the contrasts based on the groups. For the first comparison, samples with 'japan' in the group column will be the numerator, and samples with 'switzerland' in the group column will be the denominator. For the second comparison sample with 'USA' (as main group) will be compared to the samples with 'switzerland' (as reference group). The third comparison specifies both Japan and USA (together as one group) to be compared to Switzerland.

..note::

1. Group-wise contrasts require a **group** column in the samples file.
2. The group column name must contain the word *group* (e.g. group, groups, group_name, etc..).
3. If multiple group columns are found, only the first one will be matched.

..seealso::

If you are familiar with R, you can think of the contrasts file as a data.frame object. We provide an example samples file that can be accessed by installing playbase `devtools::install_github("bigomics/playbase")` and running `playbase::CONTRASTS`.

CHAPTER

SIX

CHECKS BEFORE UPLOADING

Check #1. Never use the special characters in your samples and contrasts, as certain characters are reserved for programming purposes. If you need to connect multiple elements, use underscore, "_" instead.

We test regularly which characters are allowed or not, we will update this list accordingly. The characters below are strictly forbidden.

Input file	Part	Characters not allowed
sample.csv	column names	@ , : empty spaces +
sample.csv	row names	@ , : empty spaces
sample.csv	phenotype names	@ , :
contrasts.csv	column names	@ , : empty spaces +
contrasts.csv	contrast names	@ , :

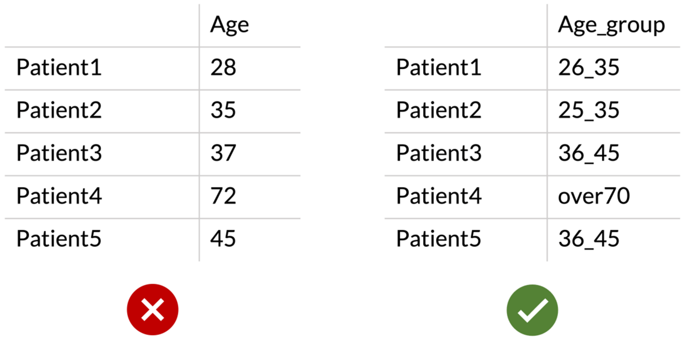
Note: We cannot guarantee that special characters (+, -, *, /, %, etc..) and empty space will work in all Omics Playground modules, so we recommend substituting them with underscore '_'.

Check #2 Avoid starting your sample, phenotype or contrast names with special characters like (+, -, *, /, %, etc..). While the platform will accept these characters, they may be converted into a standard symbol like X. For example, if we label the sample.csv country column as *%country*, we will see X.country in some analysis in Omics Playground.

Check #3 Make sure you have the words 'samples', 'counts' and 'contrasts' in the corresponding filenames. For example, we accept experiment23_samples.csv, samples_experiment23.csv, but not experiment23.csv as sample input.

Check #4. Define intervals instead of using numeric phenotypes. The platform does not cope with continuous numeric variables for phenotypes yet. To avoid that, our coders added a filter that flags phenotypes names "Time" or "Age" as unacceptable. The same applies for other continuous variables, such as height, weight, length, etc.

Instead, you should cluster the various numeric values into definite intervals and then name them accordingly (e.g. "Age_groups", "Time_intervals", etc…)



	Age
Patient1	28
Patient2	35
Patient3	37
Patient4	72
Patient5	45

✗

	Age_group
Patient1	26_35
Patient2	25_35
Patient3	36_45
Patient4	over70
Patient5	36_45

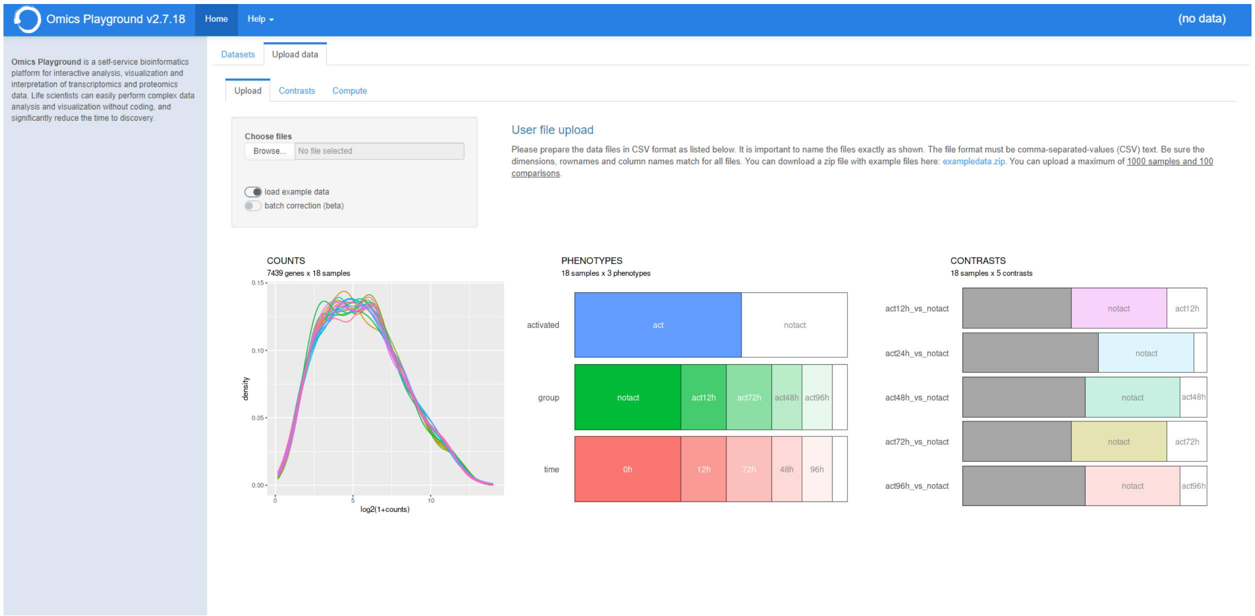
✓

CHAPTER

SEVEN

UPLOADING YOUR DATA IN OMICS PLAYGROUND

Users can import their data from the **Upload data** panel located under the *Load Panel* module. The platform requires a file with the read counts and one with a description of the samples at the minimum. An optional file with the desired contrasts can also be provided. The format of files must be comma-separated-values (CSV) text. It is important to name the files exactly as explained in the previous sections.



Omics Playground is a self-service bioinformatics platform for interactive analysis, visualization and interpretation of transcriptomics and proteomics data. Life scientists can easily perform complex data analysis and visualization without coding, and significantly reduce the time to discovery.

User file upload

Please prepare the data files in CSV format as listed below. It is important to name the files exactly as shown. The file format must be comma-separated-values (CSV) text. Be sure the dimensions, rownames and column names match for all files. You can download a zip file with example files here: exampledata.zip. You can upload a maximum of 1000 samples and 100 comparisons.

Choose files

Browse... No file selected

● load example data

○ batch correction (beta)

CHAPTER

EIGHT

WHAT IF I HAVE MANY DATASETS?

The first step in any OmicsPlayground analysis is to upload data which can be used to create a pgx object. The pgx object is basically the core data structure in the OmicsPlayground upon which most analysis and plotting functions operate.

If you have many datasets (tens to hundreds), it might make sense to create the pgx object with an R script.

We made it very easy to create a pgx object from your own data. With a few R functions, you can convert several datasets into pgx objects that can be uploaded in Omics Playground.

Here we check that your input files do not have problems

```r
# install necessary packages

install.packages("devtools")
devtools::install_github("bigomics/playbase")

library(playbase)

# These are the possible errors you can encounter

playbase::PGX_CHECKS

# individual file checks

SAMPLES = playbase::pgx.checkINPUT(playbase::SAMPLES, type = "SAMPLES")
COUNTS = playbase::pgx.checkINPUT(playbase::COUNTS, type = "COUNTS")
CONTRASTS = playbase::pgx.checkINPUT(playbase::SAMPLES, type = "CONTRASTS")

# Checks across input files

INPUTS_CHECKED <- pgx.crosscheckINPUT(SAMPLES, COUNTS, CONTRASTS)

SAMPLES = INPUTS_CHECKED$SAMPLES
COUNTS = INPUTS_CHECKED$COUNTS
CONTRASTS = INPUTS_CHECKED$CONTRASTS
```

If no errors are reported (and *PASS* is *TRUE*), these new checked files SAMPLES, COUNTS and CONTRASTS can be used safely in the next step.

Here we create a pgx object that can be used in Omics Playground:

```r
# step 1: create a pgx object with your samples, counts and contrasts
pgx <- playbase::pgx.createPGX(
```

(continues on next page)

(continued from previous page)

```
  counts = playbase::COUNTS,
  samples = playbase::SAMPLES,
  contrasts = playbase::CONTRASTS
)

# step 2: compute the pgx object

pgx <- playbase::pgx.computePGX(
  pgx = pgx
)

# save the pgx object to your computer
save(pgx, file = "choose_a_name.pgx.")
```

All you have to do is substitute `playbase::COUNTS`, `playbase::SAMPLES`, and `playbase::CONTRASTS` with your own data. You can then import the pgx object into Omics Playground.

See also:

In reality, there is a lot more happening behing the *pgx.createPGX* and *pgx.computePGX*. If you are interested in learning more, please see our Github Wiki with more details on the statistics, normalization and filtering steps that are performed. Wiki. You can find it here.

CHAPTER

NINE

PREPARING INPUT FILES FROM RAW DATA FORMATS

If you have a raw data format (FASTQ, mzML, mzML, RAW, WIFF, etc..) or want to perform additional pre-processing steps (normalization or filtering), in this section we show tutorials on how to perform such steps. We also show how to prepare data files from public databases such as GEO.

We provide four types of example cases to guide users for preparing their input objects and injecting it into the platform. Basically, the example cases illustrate how to prepare an input data:

1. *from FASTQ files*,
2. *from gene counts table or from the GEO repository*,
3. *from single-cell data*,
4. *from LC-MS/MS proteomics data*.

All the necessary scripts for data cleaning and preprocessing examples can be found under the `scripts/` folder.

9.1 From FASTQ files

Starting from FASTQ files, we recommend using the GREP2 package to obtain gene counts through quality control, trimming, quantification of gene abundance, and so on. Afterwards, the user can refer to the examples in the next section for preparing an input data from the gene counts.

9.2 From gene counts table or GEO repository

Users can prepare an input data from their own gene counts or download a relevant dataset from repositories such as GEO. Some examples are provided in the following scripts:

- TCGA-BRCA: `pgx-tcga-brca.R`
- TCGA-PRAD: `pgx-tcga-prad.R`
- GSE10846: `pgx-GSE10846-dlbcl.R`
- GSE114716: `pgx-GSE114716-ipilimumab.R`
- GSE22886: `pgx-GSE22886-immune.R`
- GSE28492: `pgx-GSE28492-roche.R`
- GSE32591: `pgx-GSE32591-lupusnephritis.R`
- GSE53784: `pgx-GSE53784-wnvjev.R`
- GSE88808: `pgx-GSE88808-prostate.R`

9.3 From single-cell data

Single-cell RNA sequencing experiments have been valuable to provide insights into complex biological systems, reveal complex and rare cell populations, uncover relationships between genes, and track the trajectories of cell lineages. Below we provide some data preparation examples from single-cell experiments:

- GSE72056: `pgx-GSE72056-scmelanoma.R`
- GSE92332: `pgx-GSE92332-scintestine.R`
- GSE98638: `pgx-GSE98638-scliver.R`

9.4 From LC-MS/MS proteomics data

Two examples are provided below for LC-MS/MS proteomics data preprocessing:

- Geiger et al. 2016: `pgx-geiger2016-arginine.R`
- Rieckmann et al. 2017: `pgx-rieckmann2017-immprot.R`

CHAPTER

TEN

LOADING CUSTOM GENESET FILES

GMT (Gene Matrix Transposed) files are a format commonly used in bioinformatics and gene set enrichment analysis (GSEA).

Gene sets are collections of genes that share certain biological characteristics, such as being involved in the same biological pathway or having similar functions. GSEA is a computational method used to determine whether a particular gene set is significantly enriched in a given dataset of gene expression.

In a GMT file, the first column is the geneset name, followed by a description on the second column. Each subsequent column line represents a gene set and consists of a gene set name, and the list of genes belonging to that set. The gene names are typically represented using gene symbols or other identifiers. Each line in a GMT file will correspont to a different geneset.

Download here an example GMT file from genes activated by the transcription factor EGFR: `EGFR_TARGET_GENES.v2023.1.Hs.gmt`

Tip: If you want to create your own genesets, the easiest way is to download our example GMT file and open it in excel as tab-separated values. The first column is the geneset name (EGFR_TARGET_GENES), followed by any description (can be left empty) or the geneset source website in the second column. Then finally the genes should be listed on the third and following columns (ABCA7, ACTB, etc.). After editing, save the file as a tab-separated values file and change the extension to .gmt.

In Omics playground, GMT files are the backbone of several modules, where we aim to identify enriched gene sets associated with specific biological processes, diseases, or experimental conditions.

You can find the geneset availables in omicsplayground with the R package playbase. First, install playbase with `devtools::install_github("omicsplayground/playdata")` and then run the command `playdata::GSET_INFO`.

If you have your own geneset, or you want to include a GMT file from a database, you can upload directly in the Upload module under Options.

See also:

If you are working in R, you can add your gmt file to the `playbase::pgx.computePGX` by providing the `custom.geneset` argument. custom.geneset argument. That should be a list where the first argument `gmt` is the gmt file (see `playbase::EXAMPLE_GMT`), and the second argument is the name of the gmt file.

Building on what we learned in the section *Computing PGX* , we can add our own geneset to the analysis with the followign call:

```
filePath <- getwd() # path to gmt file

custom.geneset <- list()
```

(continues on next page)

(continued from previous page)

```
custom.geneset$gmt <- playbase::read.gmt(filePath)

# perform some basic checks
gmt.length <- length(custom.geneset$gmt)
gmt.is.list <- is.list(custom.geneset$gmt)

# clean genesets

custom.geneset$gmt <- list(CUSTOM = custom.geneset$gmt)

# convert gmt to OPG standard
custom.geneset$gmt <- playbase::clean_gmt(custom.geneset$gmt,"CUSTOM")

# compute custom geneset stats
custom.geneset$gmt <- custom.geneset$gmt[!duplicated(names(custom.geneset$gmt))]
custom.geneset$info$GSET_SIZE <- sapply(custom.geneset$gmt,length)

# pass the custom.geneset as argument to pgx.computePGX
pgx <- playbase::pgx.computePGX(
    pgx = pgx,
    custom.geneset = custom.geneset
)
```

CHAPTER

ELEVEN

DATA PROCESSING INSIDE OMICS PLAYGROUND

11.1 Filtering of features (genes) and samples

The data preprocessing includes some filtering criteria, such as filtering of genes based on variance, the expression across the samples, and the number of missing values. Similarly, samples can also be filtered based on the read quality, total abundance, unrelated phenotype, or an outlier criterion.

11.2 Normalisation

The raw counts are converted into counts per million (CPM) and log2. Depending on the data set, a quantile normalization can be applied. Known batches in the data can be corrected with limma or ComBat. Other unknown batch effects and unwanted variation can be further removed using surrogate variable analysis in the sva package.

11.3 Offline computation

Statistics for the differentially expressed genes analysis and gene set enrichment analysis are precomputed to accelerate the visualisation on the interface.

CHAPTER

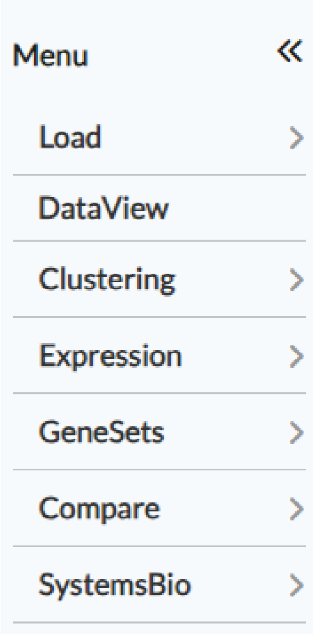
TWELVE

OUTLINE

12.1 Main menu

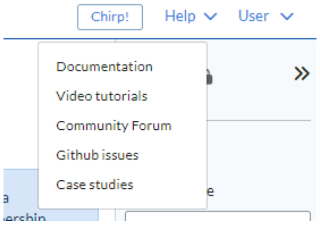
Using the main menu on top, you can navigate through the different analysis modules. Generally you want to start from top to bottom, from specific gene-wise to the more higher-level functional analysis modules. Some users prefer the other way around.

By default, the following menu items are present, namely Load, *Data View*, *Clustering*, *Expression*, *GeneSets*, *Compare* and *SystemsBio*.



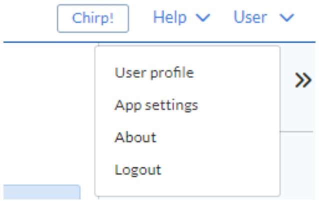
12.2 Help menu

The **Help** menu is located on the top right hand corner. The menu links to the online platform documentation (the page you are currently viewing), a collection of video tutorials on youtube, the community forum of the platform, the Github webpage, where busg can be reported, and a collection of case studies based on Omics Playground.



12.3 User menu

The **User** menu is located on the top right hand corner, right next to the **Help** menu. It links to the user profile, where information on the usage of the platform can be found, the app settings, the "About" pop-up that shows information about the version of the platform currently deployed and finally the "Logout" option to disconnect from the platform.



Clicking "App settings" will take the user to a new page with two tabs: **App settings & News** and **Resource info**.

There are two panels under the **App settings & News** tab: **Application options** and **New features**.

Application options
Through this panel, users can enable beta features, disable alerts and enable captions for the plots.

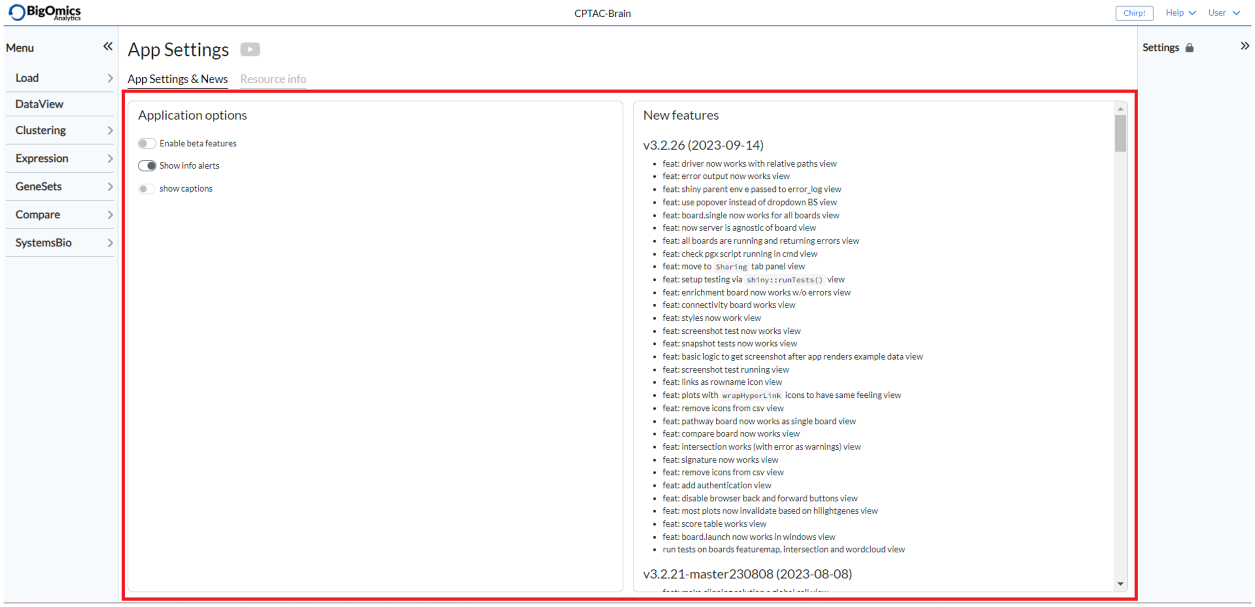
New features
This panel provides a list of new features implemented with each new update of the platform.

Three panels are found under **Resource info**: **Timings**, **PGX slot sizes** and **R object sizes**.

Timings
The timings table reports more detailed information about the object dimensions, object sizes and execution times of the methods.

PGX slot sizes
This table provides details about the pgx object.

R object sizes
This table provides size details about R objects.



Application options

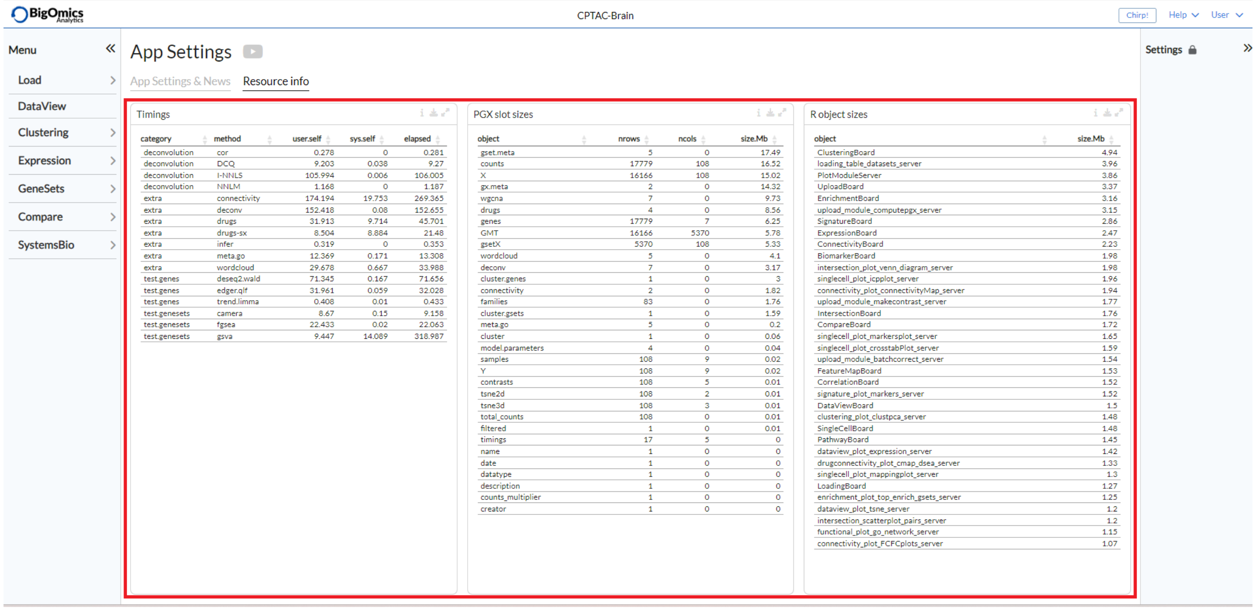
- Enable beta features
- Show info alerts
- show captions

New features

v3.2.26 (2023-09-14)

- feat: driver now works with relative paths view
- feat: error output now works view
- feat: shiny parent env e passed to error_log view
- feat: use popover instead of dropdown BS view
- feat: board.single now works for all boards view
- feat: now server is agnostic of board view
- feat: all boards are running and returning errors view
- feat: check pgx script running in cmd view
- feat: move to `Sharing` tab panel view
- feat: setup testing via `shiny::runTests()` view
- feat: enrichment board now works w/o errors view
- feat: connectivity board works view
- feat: styles now work view
- feat: screenshot test now works view
- feat: snapshot tests now works view
- feat: basic logic to get screenshot after app renders example data view
- feat: screenshot test running view
- feat: links as rowname icon view
- feat: plots with `wrapHyperLink` icons to have same feeling view
- feat: remove icons from csv view
- feat: pathway board now works as single board view
- feat: compare board now works view
- feat: intersection works (with error as warnings) view
- feat: signature now works view
- feat: remove icons from csv view
- feat: add authentication view
- feat: disable browser back and forward buttons view
- feat: most plots now invalidate based on hilightgenes view
- feat: score table works view
- feat: board.launch now works in windows view
- run tests on boards featuremap, intersection and wordcloud view

v3.2.21-master230808 (2023-08-08)



Timings

category	method	user.self	sys.self	elapsed
deconvolution	cor	0.278	0	0.281
deconvolution	DCQ	9.203	0.038	9.27
deconvolution	I-NNLS	105.994	0.006	106.005
deconvolution	NNLM	1.168	0	1.187
extra	connectivity	174.194	19.753	269.365
extra	deconv	152.418	0.08	152.655
extra	drugs	31.913	9.714	45.701
extra	drugs-sx	8.504	8.884	21.48
extra	infer	0.319	0	0.353
extra	meta.go	12.369	0.171	13.308
extra	wordcloud	29.678	0.667	33.988
test.genes	deseq2.wald	71.345	0.167	71.656
test.genes	edger.qlf	31.961	0.059	32.028
test.genes	trend.limma	0.408	0.01	0.433
test.genesets	camera	8.67	0.15	9.158
test.genesets	fgsea	22.433	0.02	22.063
test.genesets	gsva	9.447	14.089	318.987

PGX slot sizes

object	nrows	ncols	size.Mb
gset.meta	5	0	17.49
counts	17779	108	16.52
X	16166	108	15.02
gx.meta	2	0	14.32
wgcna	7	0	9.73
drugs	4	0	8.56
genes	17779	7	6.25
GMT	16166	5370	5.78
gsetX	5370	108	5.33
wordcloud	5	0	4.1
deconv	7	0	3.17
cluster.genes	1	0	3
connectivity	2	0	1.82
families	83	0	1.76
cluster.gsets	1	0	1.59
meta.go	5	0	0.2
cluster	1	0	0.06
model.parameters	4	0	0.04
samples	108	9	0.02
Y	108	9	0.02
contrasts	108	5	0.01
tsne2d	108	2	0.01
tsne3d	108	3	0.01
total_counts	108	0	0.01
filtered	1	0	0.01
timings	17	5	0
name	1	0	0
date	1	0	0
datatype	1	0	0
description	1	0	0
counts_multiplier	1	0	0
creator	1	0	0

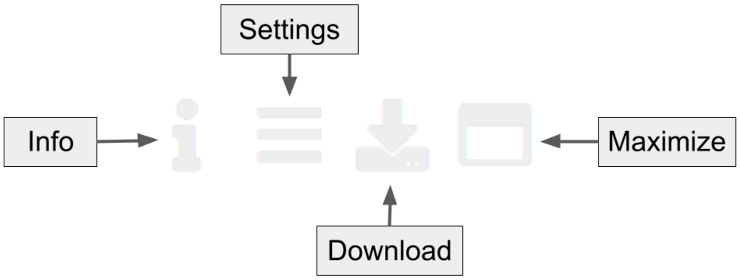
R object sizes

object	size.Mb
ClusteringBoard	4.94
loading_table_datasets_server	3.96
PlotModuleServer	3.86
UploadBoard	3.37
EnrichmentBoard	3.16
upload_module_computepgx_server	3.15
SignatureBoard	2.86
ExpressionBoard	2.47
ConnectivityBoard	2.23
BiomarkerBoard	1.98
intersection_plot_venn_diagram_server	1.98
singlecell_plot_icpplot_server	1.96
connectivity_plot_connectivityMap_server	1.94
upload_module_makecontrast_server	1.77
IntersectionBoard	1.76
CompareBoard	1.72
singlecell_plot_markersplot_server	1.65
singlecell_plot_crosstabPlot_server	1.59
upload_module_batchcorrect_server	1.54
FeatureMapBoard	1.53
CorrelationBoard	1.52
signature_plot_markers_server	1.52
DataViewBoard	1.5
clustering_plot_clustpca_server	1.48
SingleCellBoard	1.48
PathwayBoard	1.45
dataview_plot_expression_server	1.42
drugconnectivity_plot_cmap_dsea_server	1.33
singlecell_plot_mappingplot_server	1.3
LoadingBoard	1.27
enrichment_plot_top_enrich_gsets_server	1.25
dataview_plot_tsne_server	1.2
intersection_scatterplot_pairs_server	1.2
functional_plot_go_network_server	1.15
connectivity_plot_FCFCplots_server	1.07

12.4 Figure & table tags

Each figure or table on the platform is assigned the following interactive buttons, where:

- Info: provides detailed information about the figure or table.
- Settings: users can specify additional settings if applicable.
- Download: downloads a figure as a PNG or PDF file or a table in CSV format.
- Maximize: shows a larger version of a figure in a separate window.



12.5 Glossary

- Signature: a list of selected genes (e.g. by significance or fold change),
- Condition: a specific phenotype group (e.g. tumor or control),
- Contrast: a comparison between two conditions (e.g. tumor vs control),
- Profile: a vector of fold changes corresponding to a certain comparison,
- Hierarchical clustering: a method that groups similar samples into groups,
- Q value: an FDR-adjusted p value,
- Biomarker: a biological feature (gene, mutation or gene set) that characterises a specific physiological or pathological process.

CHAPTER

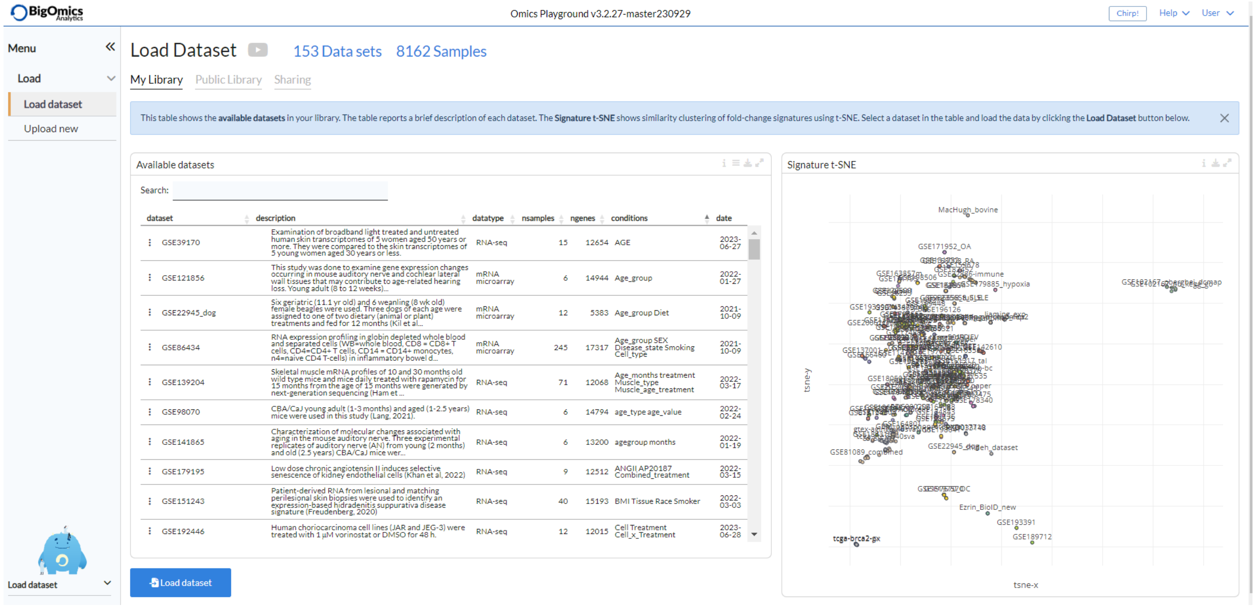
THIRTEEN

LOAD PANEL

13.1 Selecting a dataset

The **Load** panel panel shows the available datasets within the platform. Each dataset contains a brief description as well as the total number of samples, genes, corresponding phenotypes and the collection date.

Users can select a dataset in the table and then click on the "Load dataset" button at the bottom left to load a dataset into the platform. On the right hand side a signature t-SNE plot is displayed . Each dot corresponds to a specific comparison. Signatures/datasets that are clustered closer together, are more similar.

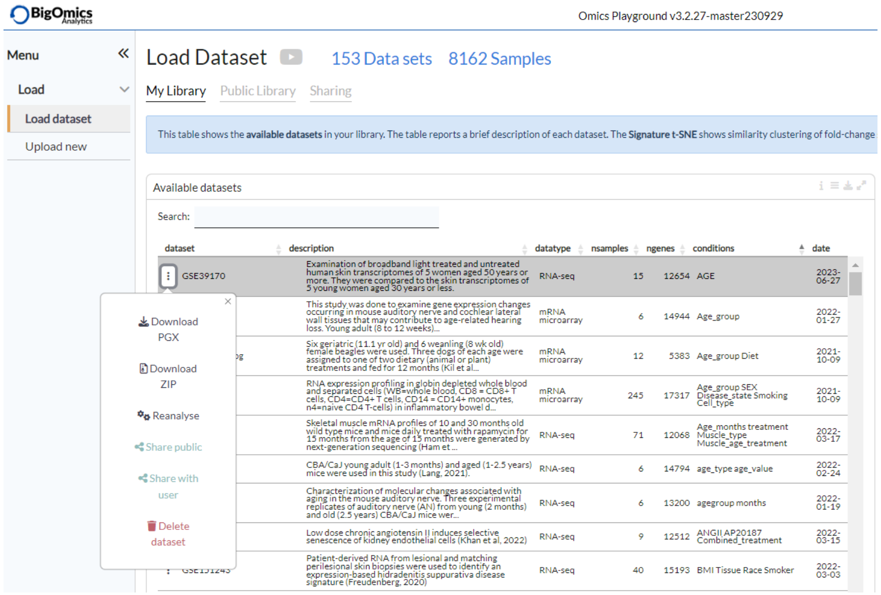


This table shows the **available datasets** in your library. The table reports a brief description of each dataset. The **Signature t-SNE** shows similarity clustering of fold-change signatures using t-SNE. Select a dataset in the table and load the data by clicking the **Load Dataset** button below.

dataset	description	datatype	nsamples	ngenes	conditions	date
⋮ GSE39170	Examination of broadband light treated and untreated human skin transcriptomes of 5 women aged 50 years or more. They were compared to the skin transcriptomes of 5 young women aged 30 years or less.	RNA-seq	15	12654	AGE	2023-06-27
⋮ GSE121856	This study was done to examine gene expression changes occurring in mouse auditory nerve and cochlear lateral wall tissues that may contribute to age-related hearing loss. Young adult (8 to 12 weeks)...	mRNA microarray	6	14944	Age_group	2022-01-27
⋮ GSE22945_dog	Six geriatric (11.1 yr old) and 6 weanling (8 wk old) female beagles were used. Three dogs of each age were assigned to one of two dietary (animal or plant) treatments and fed for 12 months (Kil et al...	mRNA microarray	12	5383	Age_group Diet	2021-10-09
⋮ GSE86434	RNA expression profiling in globin depleted whole blood and separated cells (WB=whole blood, CD8 = CD8+ T cells, CD4=CD4+ T cells, CD14 = CD14+ monocytes, n4=naive CD4 T-cells) in inflammatory bowel d...	mRNA microarray	245	17317	Age_group SEX Disease_state Smoking Cell_type	2021-10-09
⋮ GSE139204	Skeletal muscle mRNA profiles of 10 and 30 months old wild type mice and mice daily treated with rapamycin for 15 months from the age of 15 months were generated by next-generation sequencing (Ham et ...	RNA-seq	71	12068	Age_months treatment Muscle_type Muscle_age_treatment	2022-03-17
⋮ GSE98070	CBA/CaJ young adult (1-3 months) and aged (1-2.5 years) mice were used in this study (Lang, 2021).	RNA-seq	6	14794	age_type age_value	2022-02-24
⋮ GSE141865	Characterization of molecular changes associated with aging in the mouse auditory nerve. Three experimental replicates of auditory nerve (AN) from young (2 months) and old (2.5 years) CBA/CaJ mice wer...	RNA-seq	6	13200	agegroup months	2022-01-19
⋮ GSE179195	Low dose chronic angiotensin II induces selective senescence of kidney endothelial cells (Khan et al, 2022)	RNA-seq	9	12512	ANGII AP20187 Combined_treatment	2022-03-15
⋮ GSE151243	Patient-derived RNA from lesional and matching perilesional skin biopsies were used to identify an expression-based hidradenitis suppurativa disease signature (Freudenberg, 2020)	RNA-seq	40	15193	BMI Tissue Race Smoker	2022-03-03
⋮ GSE192446	Human choriocarcinoma cell lines (JAR and JEG-3) were treated with 1 µM vorinostat or DMSO for 48 h.	RNA-seq	12	12015	Cell Treatment Cell_x_Treatment	2023-06-28

A pop-up menu is displayed when clicking on the three dots next toa dataset name. From this menu, users can download either a pgx object, which contains all the calculations and plots, or the raw input data that was used to generate the dataset. The "Reanalyse" option is used to perform a new analysis based on the same input files as the selected dataset. Users can also share a dataset with another user (via the "Share with user" option) or make a dataset public (via the "Share public" option). Finally, "Delete dataset" is used to remove a dataset permanently from the platform.

N.B. Some of these features are (such as the delete option) are not available on trial accounts.



This table shows the **available datasets** in your library. The table reports a brief description of each dataset. The **Signature t-SNE** shows similarity clustering of fold-change

dataset	description	datatype	nsamples	ngenes	conditions	date
GSE39170	Examination of broadband light treated and untreated human skin transcriptomes of 5 women aged 50 years or more. They were compared to the skin transcriptomes of 5 young women aged 30 years or less.	RNA-seq	15	12654	AGE	2023-06-27
	This study was done to examine gene expression changes occurring in mouse auditory nerve and cochlear lateral wall tissues that may contribute to age-related hearing loss. Young adult (8 to 12 weeks)...	mRNA microarray	6	14944	Age_group	2022-01-27
g	Six geriatric (11.1 yr old) and 6 weanling (8 wk old) female beagles were used. Three dogs of each age were assigned to one of two dietary (animal or plant) treatments and fed for 12 months (Kil et al...	mRNA microarray	12	5383	Age_group Diet	2021-10-09
	RNA expression profiling in globin depleted whole blood and separated cells (WB=whole blood, CD8 = CD8+ T cells, CD4=CD4+ T cells, CD14 = CD14+ monocytes, n4=naive CD4 T-cells) in inflammatory bowel d...	mRNA microarray	245	17317	Age_group SEX Disease_state Smoking Cell_type	2021-10-09
	Skeletal muscle mRNA profiles of 10 and 30 months old wild type mice and mice daily treated with rapamycin for 15 months from the age of 15 months were generated by next-generation sequencing (Ham et ...	RNA-seq	71	12068	Age_months treatment Muscle_type Muscle_age_treatment	2022-03-17
	CBA/CaJ young adult (1-3 months) and aged (1-2.5 years) mice were used in this study (Lang, 2021).	RNA-seq	6	14794	age_type age_value	2022-02-24
	Characterization of molecular changes associated with aging in the mouse auditory nerve. Three experimental replicates of auditory nerve (AN) from young (2 months) and old (2.5 years) CBA/CaJ mice wer...	RNA-seq	6	13200	agegroup months	2022-01-19
	Low dose chronic angiotensin II induces selective senescence of kidney endothelial cells (Khan et al, 2022)	RNA-seq	9	12512	ANGII AP20187 Combined_treatment	2022-03-15
GSE151243	Patient-derived RNA from lesional and matching perilesional skin biopsies were used to identify an expression-based hidradenitis suppurativa disease signature (Freudenberg, 2020)	RNA-seq	40	15193	BMI Tissue Race Smoker	2022-03-03

13.2 Load (Upload new)

Under the **Upload new** option of the **Load** panel users can upload their transcriptomics and proteomics data to the platform. The platform requires three data files as listed below: a data file containing counts/expression (counts.csv), a sample information file (samples.csv) and an optional file specifying the statistical comparisons as contrasts (contrasts.csv). The file format must be comma-separated-values (CSV) text. Be sure the gene names match for all files. On the left side of the panel, users need to provide a unique name and brief description for the dataset while uploading.

counts.csv
Count/expression file with gene on rows, samples as columns.

samples.csv
Samples file with samples on rows, phenotypes as columns.

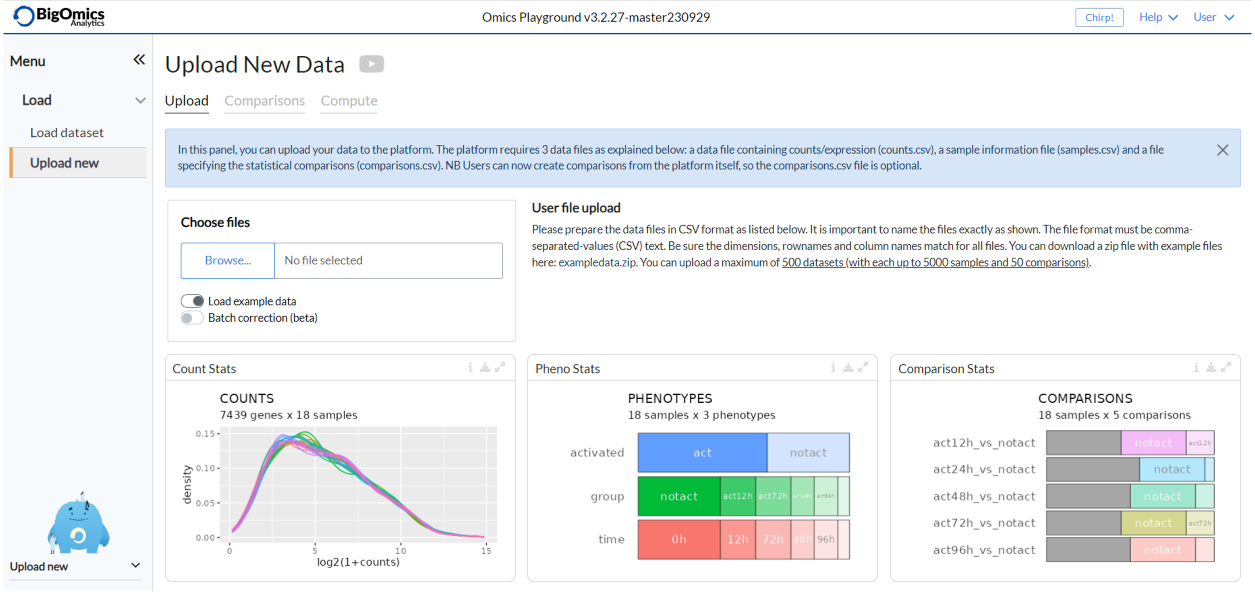
contrasts.csv
Contrast file with conditions on rows, contrasts as columns.

Once uploaded, the platform generates three separate plots showing some stastics and the structure of your input files.

The page also contains a *batch correction* option that allows users to perform batch correction on the data. Selecting it will open a new panel from which users can select the level of batch correction (low, medium or strong). Under *Advanced* users can fine-tune the process. This feature is still in beta and only recommended to users familiar with the batch correction precedure and parameters. More information can be found under *Methods*.

13.3 Upload new (Comparisons)

Users can eschew adding a contrast file and instead generate contrasts via the **Comparisons** subpanel of the **Upload new** panel. Available phenotypes will appear in the “Phenotypes” box as a scrolldown menu. Selecting a phenotype will show the available conditions in the “Conditions” box. Users can then drag individual groups in the “Main group” or “Control Group” boxes. It is also possible to select individual samples from the “Conditions” box and phenotypes can be furthermore combined in the “Phenotypes” box. The platform automatically generates a contrast name in the “Comparison name” box that users can manually edit. Clicking the “add comparison” button will add the selected contrast in the contrast table below. Users can also task the platform to generate comparisons automatically by selecting a phenotype form the “Strata” box and clicking on “add auto-contrasts”. However, we recommend caution when using



Omics Playground v3.2.27-master230929

Upload New Data

In this panel, you can upload your data to the platform. The platform requires 3 data files as explained below: a data file containing counts/expression (counts.csv), a sample information file (samples.csv) and a file specifying the statistical comparisons (comparisons.csv). NB Users can now create comparisons from the platform itself, so the comparisons.csv file is optional.

Choose files

No file selected

Load example data

Batch correction (beta)

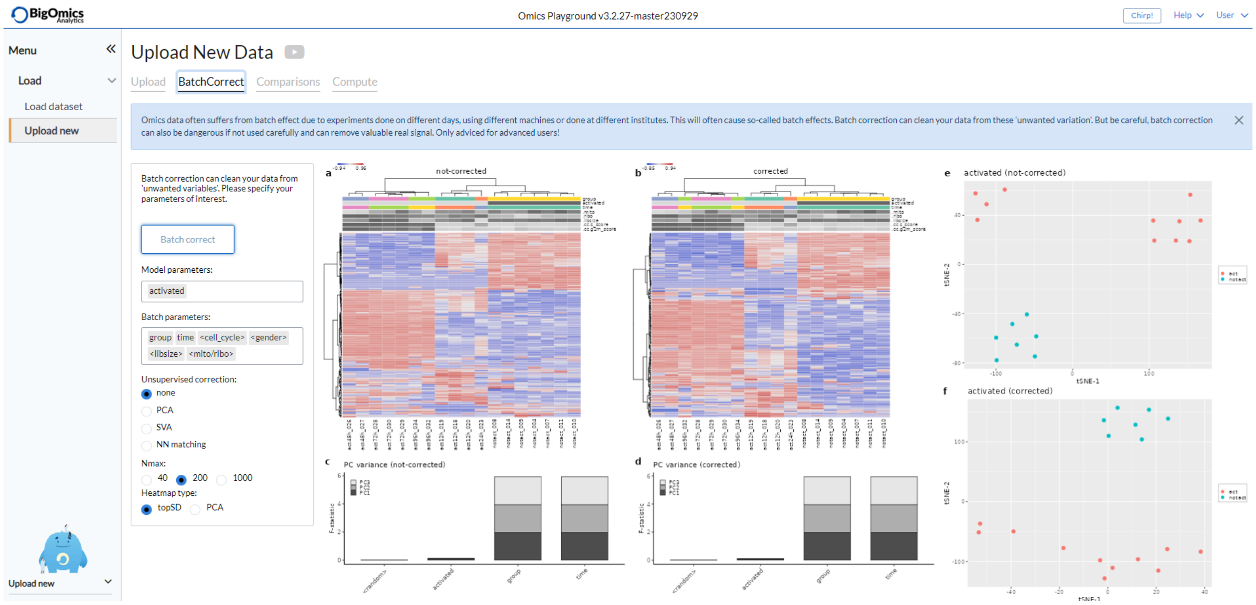
User file upload

Please prepare the data files in CSV format as listed below. It is important to name the files exactly as shown. The file format must be comma-separated-values (CSV) text. Be sure the dimensions, rownames and column names match for all files. You can download a zip file with example files here: exampledata.zip. You can upload a maximum of 500 datasets (with each up to 5000 samples and 50 comparisons).

COUNTS
7439 genes x 18 samples

PHENOTYPES
18 samples x 3 phenotypes

COMPARISONS
18 samples x 5 comparisons



Omics Playground v3.2.27-master230929

Upload New Data

Omics data often suffers from batch effect due to experiments done on different days, using different machines or done at different institutes. This will often cause so-called batch effects. Batch correction can clean your data from these 'unwanted variation'. But be careful, batch correction can also be dangerous if not used carefully and can remove valuable real signal. Only adviced for advanced users!

Batch correction can clean your data from 'unwanted variables'. Please specify your parameters of interest.

Model parameters: activated

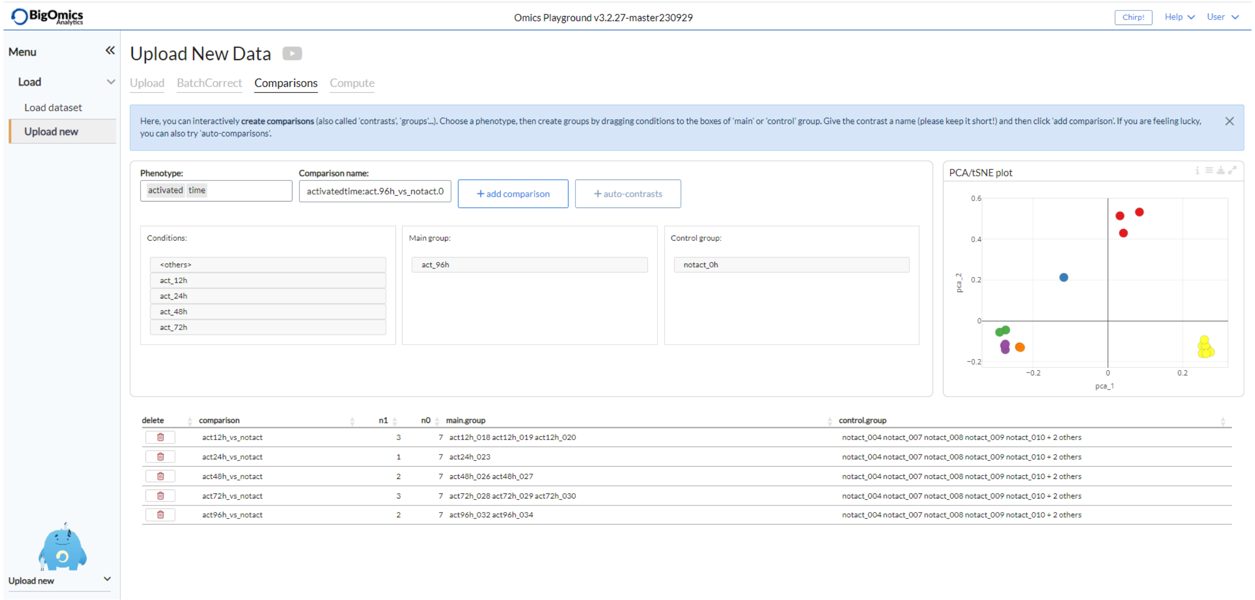
Batch parameters: group time <cell_cycle> <gender> <libsize> <mito/ribo>

Unsupervised correction:
● none
○ PCA
○ SVA
○ NN matching

Nmax: ○ 40 ● 200 ○ 1000

Heatmap type: ● topSD ○ PCA

this function as it is error-prone. The panel also produces a PCA/tSNE plot that users can consult as a guide for generating comparisons.



Here, you can interactively **create comparisons** (also called 'contrasts', 'groups'...). Choose a phenotype, then create groups by dragging conditions to the boxes of 'main' or 'control' group. Give the contrast a name (please keep it short!) and then click 'add comparison'. If you are feeling lucky, you can also try 'auto-comparisons'.

Phenotype: activated time

Comparison name: activatedtime:act.96h_vs_notact.0

Conditions:
<others>
act_12h
act_24h
act_48h
act_72h

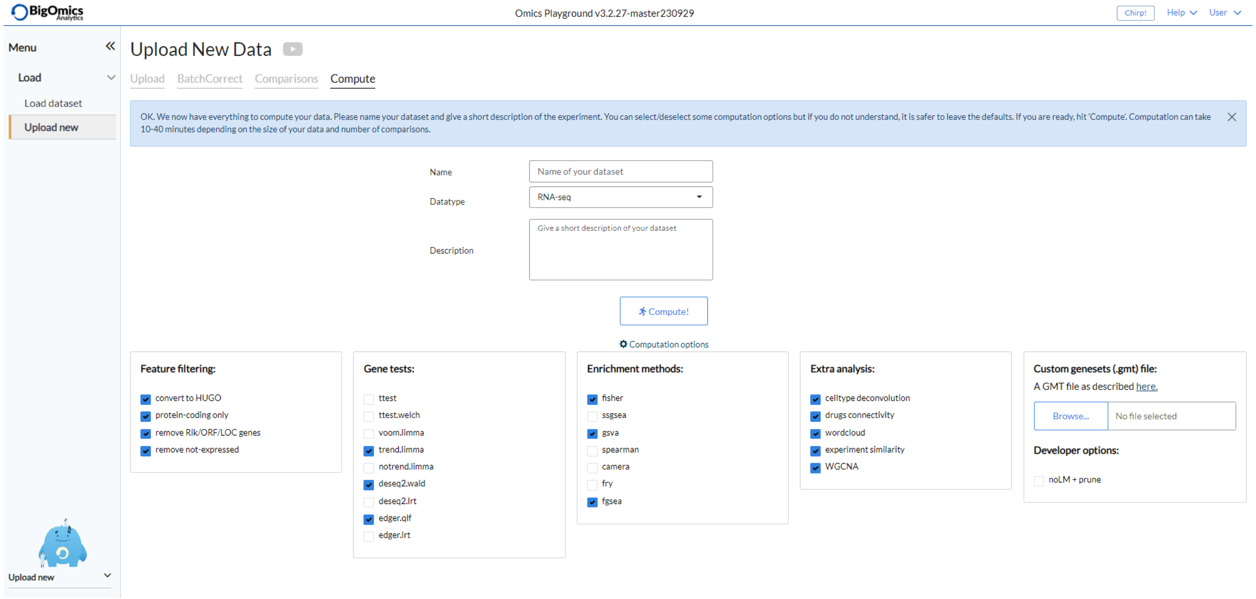
Main group: act_96h

Control group: notact_0h

delete	comparison	n1	n0	main.group	control.group
🗑	act12h_vs_notact	3	7	act12h_018 act12h_019 act12h_020	notact_004 notact_007 notact_008 notact_009 notact_010 + 2 others
🗑	act24h_vs_notact	1	7	act24h_023	notact_004 notact_007 notact_008 notact_009 notact_010 + 2 others
🗑	act48h_vs_notact	2	7	act48h_026 act48h_027	notact_004 notact_007 notact_008 notact_009 notact_010 + 2 others
🗑	act72h_vs_notact	3	7	act72h_028 act72h_029 act72h_030	notact_004 notact_007 notact_008 notact_009 notact_010 + 2 others
🗑	act96h_vs_notact	2	7	act96h_032 act96h_034	notact_004 notact_007 notact_008 notact_009 notact_010 + 2 others

13.4 Upload new (Compute)

The **Compute** subpanel is where users can start the computation of their data. They need to provide a unique name for the dataset, indicate the type of data and provide a short description. Beginners can then click on the “Compute!” button and let the platform run the calculations. Advanced users can click on “Advanced” to access several customisation options. In particular they can de-select feature filtering options, select different gene tests and enrichment methods combinations, include or exclude analysis types and select developer-specific options. A new advanced feature with V3 of the platform is the ability to upload a custom GMT files with a dataset.



OK. We now have everything to compute your data. Please name your dataset and give a short description of the experiment. You can select/deselect some computation options but if you do not understand, it is safer to leave the defaults. If you are ready, hit 'Compute'. Computation can take 10-40 minutes depending on the size of your data and number of comparisons.

Name: Name of your dataset

Datatype: RNA-seq

Description: Give a short description of your dataset

Feature filtering:
- ☑ convert to HUGO
- ☑ protein-coding only
- ☑ remove Rik/ORF/LOC genes
- ☑ remove not-expressed

Gene tests:
- □ ttest
- □ ttest.welch
- □ voom.limma
- ☑ trend.limma
- □ notrend.limma
- ☑ deseq2.wald
- □ deseq2.lrt
- ☑ edger.qlf
- □ edger.lrt

Enrichment methods:
- ☑ fisher
- □ ssgsea
- ☑ gsva
- □ spearman
- □ camera
- □ fry
- ☑ fgsea

Extra analysis:
- ☑ celltype deconvolution
- ☑ drugs connectivity
- ☑ wordcloud
- ☑ experiment similarity
- ☑ WGCNA

Custom genesets (.gmt) file:
A GMT file as described here.
No file selected

Developer options:
- □ noLM + prune

CHAPTER

FOURTEEN

DATA VIEW

The **DataView** module provides information and visualisations of the dataset to quickly lookup a gene, check the counts, or view the data tables. It has five panels, which are briefly explained below, followed by more detailed information for each panel.

The **Gene Overview** panel displays figures related to the expression level of the selected gene, correlation, and average expression ranking within the dataset. In the **Sample QC** panel, the total number of counts (abundance) per sample and their distribution among the samples are displayed. This is most useful to check the technical quality of the dataset, such as total read counts or abundance of ribosomal genes. In **Counts table** panel, the exact expression values across the samples can be looked up, where genes are ordered by the correlation with respect to the first gene. Gene-wise average expression of a phenotype sample grouping is also presented in this table. In the **Sample information** panel, more complete information about samples can be found. Finally, the **Contrasts** panel, shows information about the phenotype comparisons.

14.1 Settings panel

Users can find the settings panel on the right hand side. The panel contains the main settings for the analysis. The analysis can be started by selecting a gene of interest from the `Gene` settings. Users can filter and select samples in the `Filter samples` settings, or collapse the samples by predetermined groups in the `Group by` settings. Under *Options*, it is possible to visualize the information on a raw count level or logarithmic expression level (logCPM).

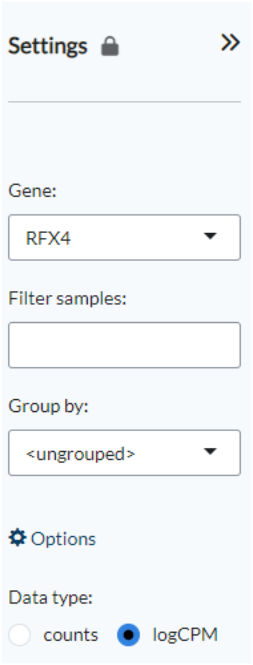
14.2 Gene Overview

The **Gene Overview** panel displays figures related to the expression level of the selected gene, correlation to other genes, and average expression ranking within the dataset. To find out more information from the literature, hyperlinks are provide to connect the selected gene to OMIM, KEGG, and GO databases. It also shows the correlation of the gene in other datasets such as ImmProt and HPA, and plots the cumulative correlation. Furthermore, it displays the tissue expression for a selected gene using the genotype-tissue expression GTEx dataset. For each chart of the panel (left to right, top to bottom), a detailed explanation is provided below.

Gene info
To find out more information from the literature, hyperlinks are provide to connect the selected gene to public databases, including OMIM, KEGG, and GO.

Gene Expression
Expression barplot of grouped samples (or cells) for the selected gene. Samples (or cells) in the barplot can be ungrouped by setting the `Group by` under the main *Options*.

Average Rank
Ranking of the average expression of the selected gene.



Settings

Gene: RFX4

Filter samples:

Group by: <ungrouped>

Options

Data type: ○ counts ● logCPM

t-SNE clustering

T-SNE clustering of samples (or cells) colored by an expression of the gene selected in the `Search gene` dropdown menu. The red color represents an over-expression of the selected gene across samples (or cells).

Top correlated genes

Barplot of the top positively and negatively correlated genes with the selected gene. Absolute expression levels of genes are colored in the barplot, where the low and high expressions range between the light blue and dark blue colors, respectively.

Tissue expression (GTEX)

Tissue expression for the selected gene in the tissue expression GTEx database. Colors corresponds to "tissue clusters" as computed by unsupervised clustering.

14.3 Sample QC

In the **Sample QC** panel, the total number of counts (abundance) per sample and their distribution among the samples are displayed. For each sample, users can also see the percentage of counts in terms of major gene types such as ribosomal genes, heatshock proteins, or kinases. Abnormal abundance of certain genes may indicate technical problems. A detailed explanation is provided below for every chart of the panel (left to right, top to bottom).

Total counts

A barplot of the total number of counts (abundance) for each group. The samples (or cells) can be grouped/ungrouped in the `Group by` setting uder the main *Options*.

Median counts distribution

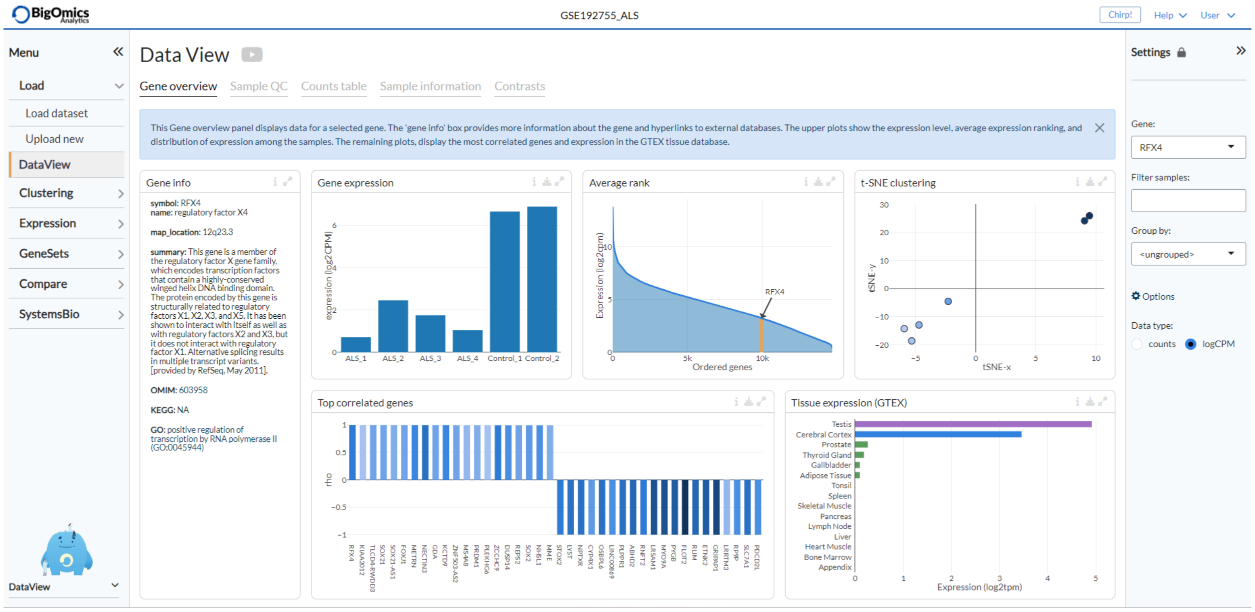
A boxplot of the total number of counts (abundance) for each group.

Density distribution of counts

A plot showing the density distribution of counts for each group.

Dataset abundance of major gene types

A barplot showing the percentage of counts in terms of major gene types such as CD molecules,



GSE192755_ALS

Data View

This Gene overview panel displays data for a selected gene. The 'gene info' box provides more information about the gene and hyperlinks to external databases. The upper plots show the expression level, average expression ranking, and distribution of expression among the samples. The remaining plots, display the most correlated genes and expression in the GTEX tissue database.

Gene info

symbol: RFX4
name: regulatory factor X4

map_location: 12q23.3

summary: This gene is a member of the regulatory factor X gene family, which encodes transcription factors that contain a highly-conserved winged helix DNA binding domain. The protein encoded by this gene is structurally related to regulatory factors X1, X2, X3, and X5. It has been shown to interact with itself as well as with regulatory factors X2 and X3, but it does not interact with regulatory factor X1. Alternative splicing results in multiple transcript variants. [provided by RefSeq, May 2011].

OMIM: 603958

KEGG: NA

GO: positive regulation of transcription by RNA polymerase II (GO:0045944)

Settings

Gene: RFX4

Filter samples:

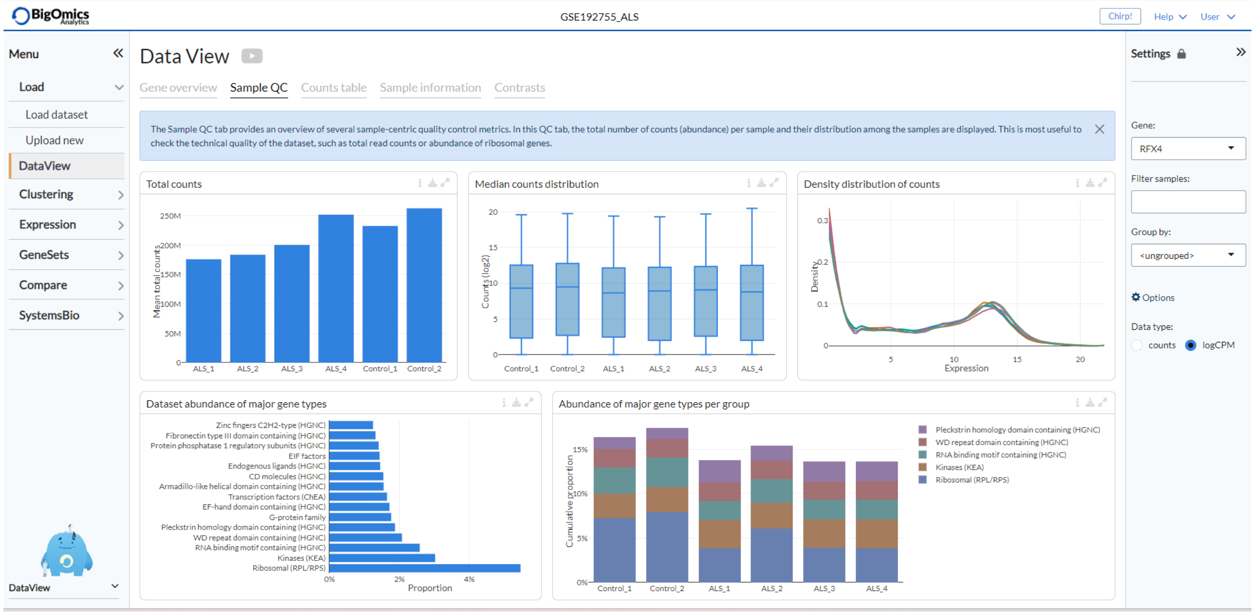
Group by: <ungrouped>

Data type: ○ counts ● logCPM

kinanses or RNA binding motifs for each group.

Abundance of major gene types per group

A barplot showing the average count levels of major gene types such as CD molecules, kinanses or RNA binding motifs for each group.



GSE192755_ALS

Data View

The Sample QC tab provides an overview of several sample-centric quality control metrics. In this QC tab, the total number of counts (abundance) per sample and their distribution among the samples are displayed. This is most useful to check the technical quality of the dataset, such as total read counts or abundance of ribosomal genes.

Settings

Gene: RFX4

Filter samples:

Group by: <ungrouped>

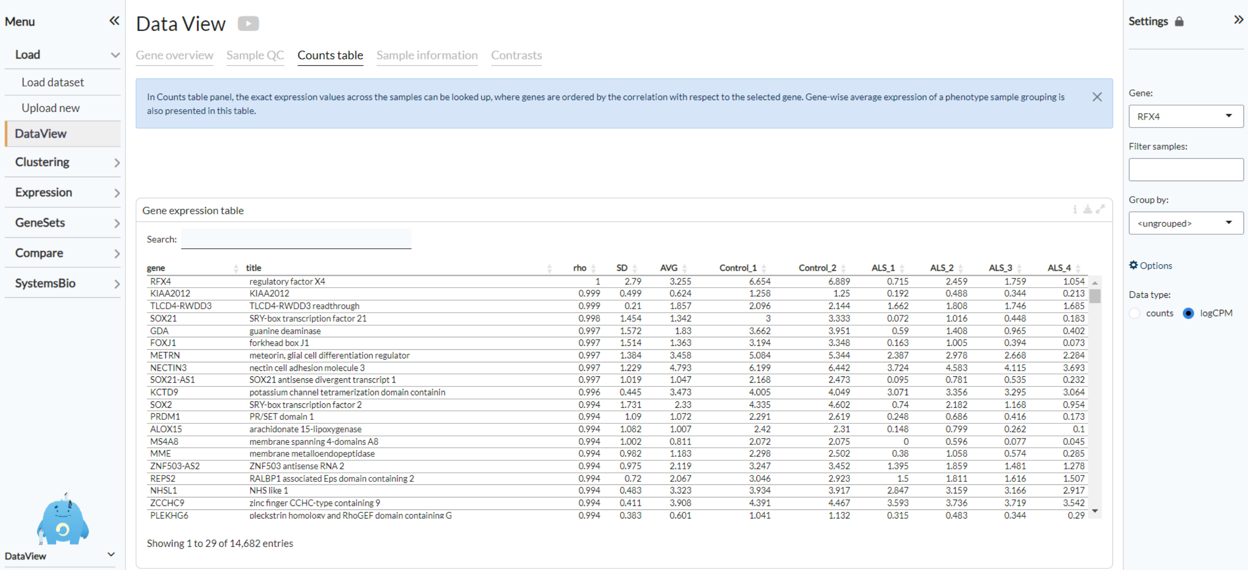
Data type: ○ counts ● logCPM

BigOmics Analytics
Chirp!
Help
User
Menu
Load
Load dataset
Upload new
DataView
Clustering
Expression
GeneSets
Compare
SystemsBio
DataView
Gene overview
Sample QC
Counts table
Sample information
Contrasts
Total counts
Mean total counts
0
50M
100M
150M
200M
250M
ALS_1
ALS_2
ALS_3
ALS_4
Control_1
Control_2
Median counts distribution
Counts (log2)
0
5
10
15
20
Control_1
Control_2
ALS_1
ALS_2
ALS_3
ALS_4
Density distribution of counts
Density
0
0.1
0.2
0.3
5
10
15
20
Expression
Dataset abundance of major gene types
Zinc fingers C2H2-type (HGNC)
Fibronectin type III domain containing (HGNC)
Protein phosphatase 1 regulatory subunits (HGNC)
EIF factors
Endogenous ligands (HGNC)
CD molecules (HGNC)
Armadillo-like helical domain containing (HGNC)
Transcription factors (ChEA)
EF-hand domain containing (HGNC)
G-protein family
Pleckstrin homology domain containing (HGNC)
WD repeat domain containing (HGNC)
RNA binding motif containing (HGNC)
Kinases (KEA)
Ribosomal (RPL/RPS)
0%
2%
4%
Proportion
Abundance of major gene types per group
Cumulative proportion
0%
5%
10%
15%
Control_1
Control_2
ALS_1
ALS_2
ALS_3
ALS_4
Pleckstrin homology domain containing (HGNC)
WD repeat domain containing (HGNC)
RNA binding motif containing (HGNC)
Kinases (KEA)
Ribosomal (RPL/RPS)
Options

14.4 Counts Table

Under the **Counts table** panel, the exact expression values across the samples can be read, where genes are ordered by the correlation with respect to the first gene. Gene-wise average expression of a phenotype sample grouping is also presented in this table.

The samples (or cells) in the table can be ungrouped by setting the `Group by` under the main *Options* to see the exact expression values per sample (or cell). The genes in the table are ordered by the correlation (**rho** column) with respect to the selected gene. **SD** column reports the standard deviation of expression across *all* samples (or cells).



In Counts table panel, the exact expression values across the samples can be looked up, where genes are ordered by the correlation with respect to the selected gene. Gene-wise average expression of a phenotype sample grouping is also presented in this table.

Gene: RFX4

Filter samples:

Group by: <ungrouped>

Options

Data type: ○ counts ● logCPM

Gene expression table

Search:

gene	title	rho	SD	AVG	Control_1	Control_2	ALS_1	ALS_2	ALS_3	ALS_4
RFX4	regulatory factor X4	1	2.79	3.255	6.654	6.889	0.715	2.459	1.759	1.054
KIAA2012	KIAA2012	0.999	0.499	0.624	1.258	1.25	0.192	0.488	0.344	0.213
TLCD4-RWDD3	TLCD4-RWDD3 readthrough	0.999	0.21	1.857	2.096	2.144	1.662	1.808	1.746	1.685
SOX21	SRY-box transcription factor 21	0.998	1.454	1.342	3	3.333	0.072	1.016	0.448	0.183
GDA	guanine deaminase	0.997	1.572	1.83	3.662	3.951	0.59	1.408	0.965	0.402
FOXJ1	forkhead box J1	0.997	1.514	1.363	3.194	3.348	0.163	1.005	0.394	0.073
METRN	meteorin, glial cell differentiation regulator	0.997	1.384	3.458	5.084	5.344	2.387	2.978	2.668	2.284
NECTIN3	nectin cell adhesion molecule 3	0.997	1.229	4.793	6.199	6.442	3.724	4.583	4.115	3.693
SOX21-AS1	SOX21 antisense divergent transcript 1	0.997	1.019	1.047	2.168	2.473	0.095	0.781	0.535	0.232
KCTD9	potassium channel tetramerization domain containin	0.996	0.445	3.473	4.005	4.049	3.071	3.356	3.295	3.064
SOX2	SRY-box transcription factor 2	0.994	1.731	2.33	4.335	4.602	0.74	2.182	1.168	0.954
PRDM1	PR/SET domain 1	0.994	1.09	1.072	2.291	2.619	0.248	0.686	0.416	0.173
ALOX15	arachidonate 15-lipoxygenase	0.994	1.082	1.007	2.42	2.31	0.148	0.799	0.262	0.1
MS4A8	membrane spanning 4-domains A8	0.994	1.002	0.811	2.072	2.075	0	0.596	0.077	0.045
MME	membrane metalloendopeptidase	0.994	0.982	1.183	2.298	2.502	0.38	1.058	0.574	0.285
ZNF503-AS2	ZNF503 antisense RNA 2	0.994	0.975	2.119	3.247	3.452	1.395	1.859	1.481	1.278
REPS2	RALBP1 associated Eps domain containing 2	0.994	0.72	2.067	3.046	2.923	1.5	1.811	1.616	1.507
NHSL1	NHS like 1	0.994	0.483	3.323	3.934	3.917	2.847	3.159	3.166	2.917
ZCCHC9	zinc finger CCHC-type containing 9	0.994	0.411	3.908	4.391	4.467	3.593	3.736	3.719	3.542
PLEKHG6	pleckstrin homology and RhoGEF domain containing G	0.994	0.383	0.601	1.041	1.132	0.315	0.483	0.344	0.29

Showing 1 to 29 of 14,682 entries

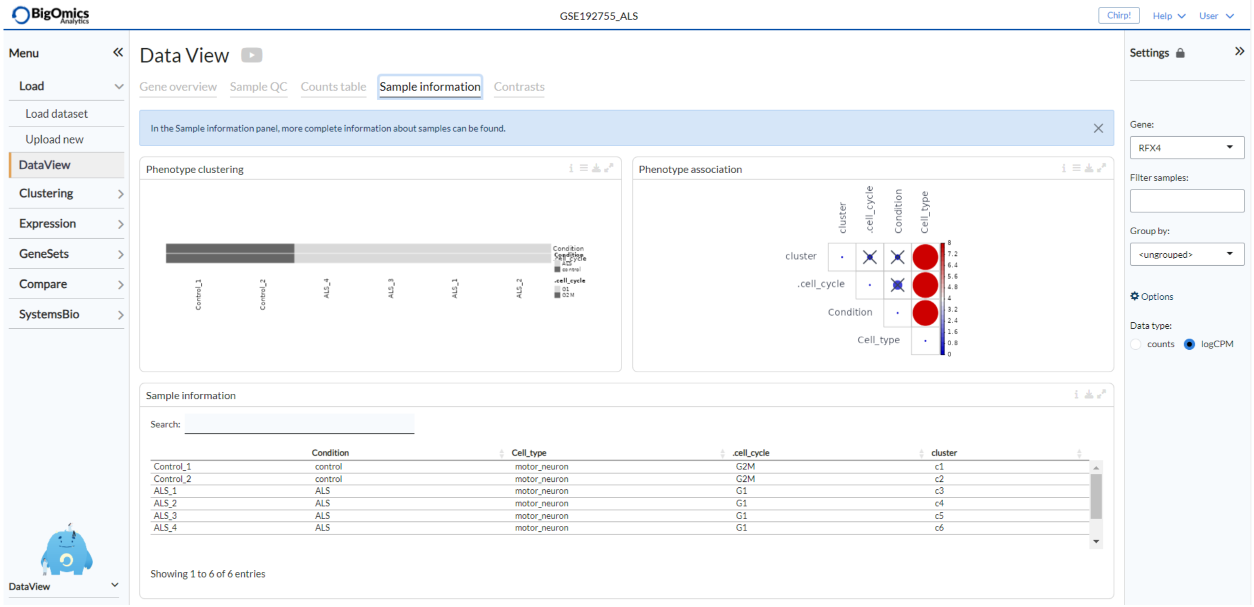
14.5 Sample Information

In the **Sample information** panel, users can check information about samples and their phenotype grouping through three outputs (left to right, top to bottom):

Phenotype clustering
A plot showing phenotype clustering. Phenotypes can be unclustered via the *Settings* icon

Phenotype association
A phenotype association matrix showing correlation between phenotypes.

Sample information
A table with sample information.



In the Sample information panel, more complete information about samples can be found.

Sample information

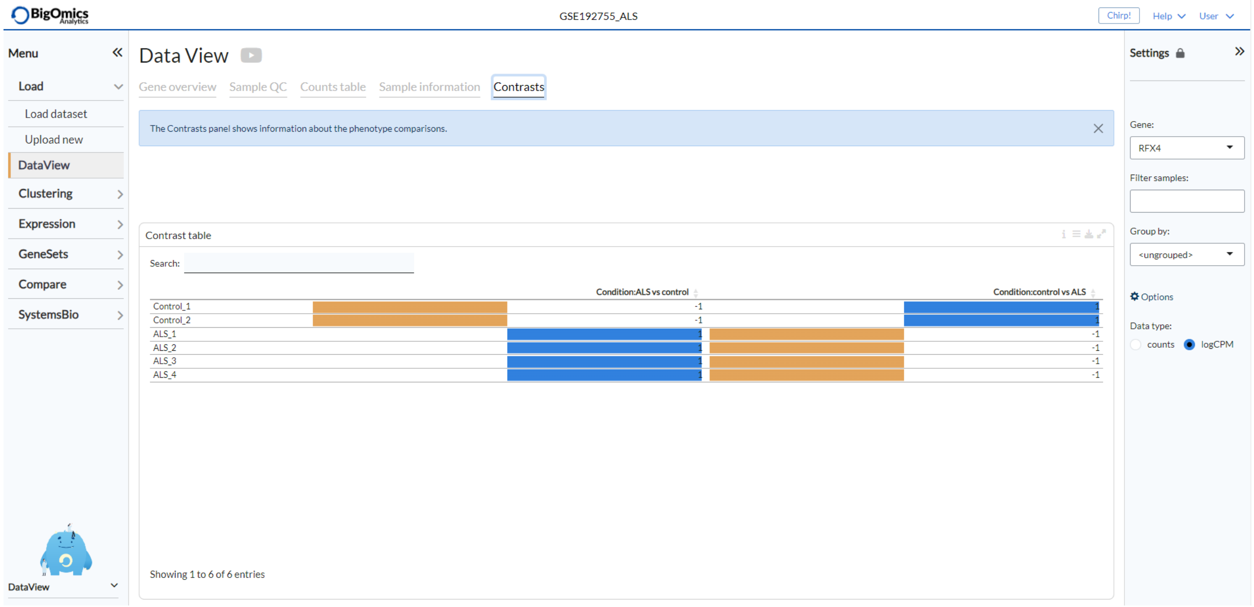
Search:

	Condition	Cell_type	.cell_cycle	cluster
Control_1	control	motor_neuron	G2M	c1
Control_2	control	motor_neuron	G2M	c2
ALS_1	ALS	motor_neuron	G1	c3
ALS_2	ALS	motor_neuron	G1	c4
ALS_3	ALS	motor_neuron	G1	c5
ALS_4	ALS	motor_neuron	G1	c6

Showing 1 to 6 of 6 entries

14.6 Contrasts

In the **Contrasts** panel, more complete information about contrasts can be found. It summarizes the contrasts of all comparisons. Here, users can check which samples belong to which groups for the different comparisons. Non-zero entries '+1' and '-1' correspond to the group of interest and control group, respectively. Zero or empty entries denote samples not use for that comparison.



The Contrasts panel shows information about the phenotype comparisons.

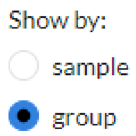
Contrast table

Search:

	Condition:ALS vs control	Condition:control vs ALS
Control_1	-1	1
Control_2	-1	1
ALS_1	1	-1
ALS_2	1	-1
ALS_3	1	-1
ALS_4	1	-1

Showing 1 to 6 of 6 entries

Through the settings icon on top of the plot, users can display samples individually or in groups.



Show by:

○ sample

● group

CHAPTER

FIFTEEN

CLUSTERING

The **Clustering** module performs unsupervised clustering analysis of the data. After having done the QC, it is probably the first way to explore your data. The main purpose is to discover patterns and subgroups in the data, show correlation with known phenotypes, detect outliers, or investigate batch effects.

The module is divided into two submodules: **Samples** and **Features**.

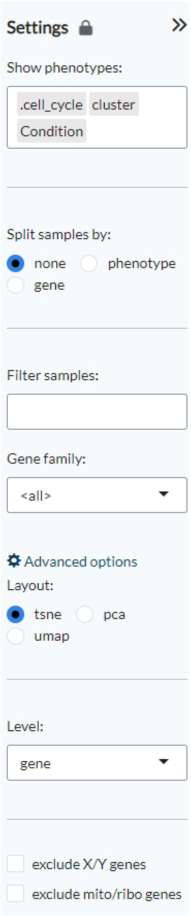
Under the **Samples** submodule you can find classical clustering functions. In the **Heatmap** panel hierarchical clustering can be performed on gene level or gene set level. The **PCA/tSNE** panel shows unsupervised clustering of the samples in 2D/3D as obtained by PCA or tSNE algorithms. The **Parallel** panel displays the expression levels of selected genes across all conditions. On the right, the **Annotate cluster** panel provides a functional annotation for each feature cluster in the heatmap. Users can select from a variety of annotation databases from the literature. The **Phenotypes** panel shows the phenotype distribution as colors on the t-SNE plot. Finally, the **Feature ranking** panel shows a plot that ranks the discriminative power of feature sets (or gene sets) as the cumulative discriminant score for all phenotype variables.

The **Features** submodule performs clustering at either the gene level (**Gene** panel) or at the geneset level (**Geneset** panel). For both a gene or geneset UMAP plot is displayed, next to a "Gene Signatures" UMAP plot, where users can visualise specific phenotypes. Finally, below the plots, a table contains eithe the genes or genesets in a given selected area.

15.1 Samples

15.1.1 Settings panel

The settings panel on the right displays various options to customise the plots. Under `Show phenotypes` users can choose which phenotypes will be displayed in the Phenotype distribution plot under the **PCA/tSNE** panel. The `Split samples by` option applies only to heatmaps. "None" is the default view, "phenotypes" will redraw the heatmap based on the selected phenotype and "gene" allows users to split the heatmap based on the expression level of a gene that can be selected from a scrolldown menu. `Filter samples` can be used to represent only a specific subset of samples in the heatmaps, PCA/tSNE/UMAP plots and the parallel coordinates plot. Under `Gene family` a user can select whether the heatmap and parallel coordinates plot will be built based on all the genes, a specific contrast, a specific gene family or alternatively a custom list of genes provided by the user that can be pasted in the appropriate space. Furthermore, under *Advanced options*, users can choose the layout of the clustering plots (PCA, tSNE or UMAP), the level of analysis (gene or geneset) and exclude mitochondrial and ribosomal genes and/or genes in the X and Y chromosomes.



Settings

Show phenotypes:

.cell_cycle cluster Condition

Split samples by:

● none ○ phenotype ○ gene

Filter samples:

Gene family:

<all>

Advanced options

Layout:

● tsne ○ pca ○ umap

Level:

gene

□ exclude X/Y genes

□ exclude mito/ribo genes

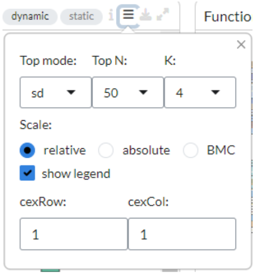
15.1.2 Heatmap

In the **Heatmap** panel hierarchical clustering can be performed on gene level or gene set level expression. For the latter, for each gene set (or pathway), a single-sample enrichment value is computed from the gene expression data using summary methods such as GSVA and ssGSEA.

Next to the plot configuration settings, users can select between a "dynamic" or "static" heatmap. From the plot configuration settings on top of the plot, users can choose various options to customise their heatmaps. It is possible to order the top features under `top mode` as follows:

- sd - features with the highest standard deviation across all the samples,
- PCA - by principal components.
- marker - features that are overexpressed in each phenotype class compared to the rest

In addition, users can specify the `Top N` (50, 150, 500) genes to be used in the heatmap and the number of gene clusters to be displayed under `K`. Users can also choose between 'relative', 'absolute' or 'BMC' (batch-mean centered) expression scale. Under the `CexCol` and `CexRow` settings, it is also possible to adjust the font sizes for the column and row labels. The legend in the heatmap can be disabled by unticking the `show legend` option.



Top mode: sd

Top N: 50

K: 4

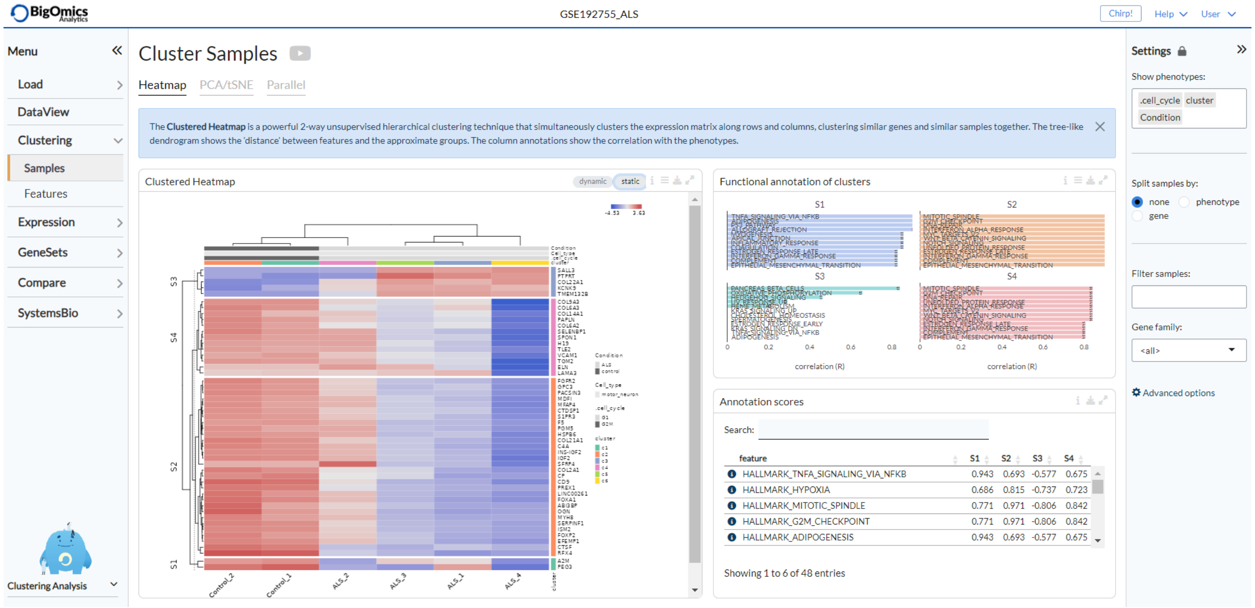
Scale: ● relative ○ absolute ○ BMC

☑ show legend

cexRow: 1

cexCol: 1

The complex heatmap below is generated with the "static" option active. It is a clustered heatmap showing gene expression sorted by 2-way hierarchical clustering. Red corresponds to overexpression, blue to underexpression of the gene. At the same time, gene clusters are functionally annotated in the **Annotate clusters** panel on the right.



The **Clustered Heatmap** is a powerful 2-way unsupervised hierarchical clustering technique that simultaneously clusters the expression matrix along rows and columns, clustering similar genes and similar samples together. The tree-like dendrogram shows the 'distance' between features and the approximate groups. The column annotations show the correlation with the phenotypes.

Split samples by: ● none ○ phenotype ○ gene

Gene family: <all>

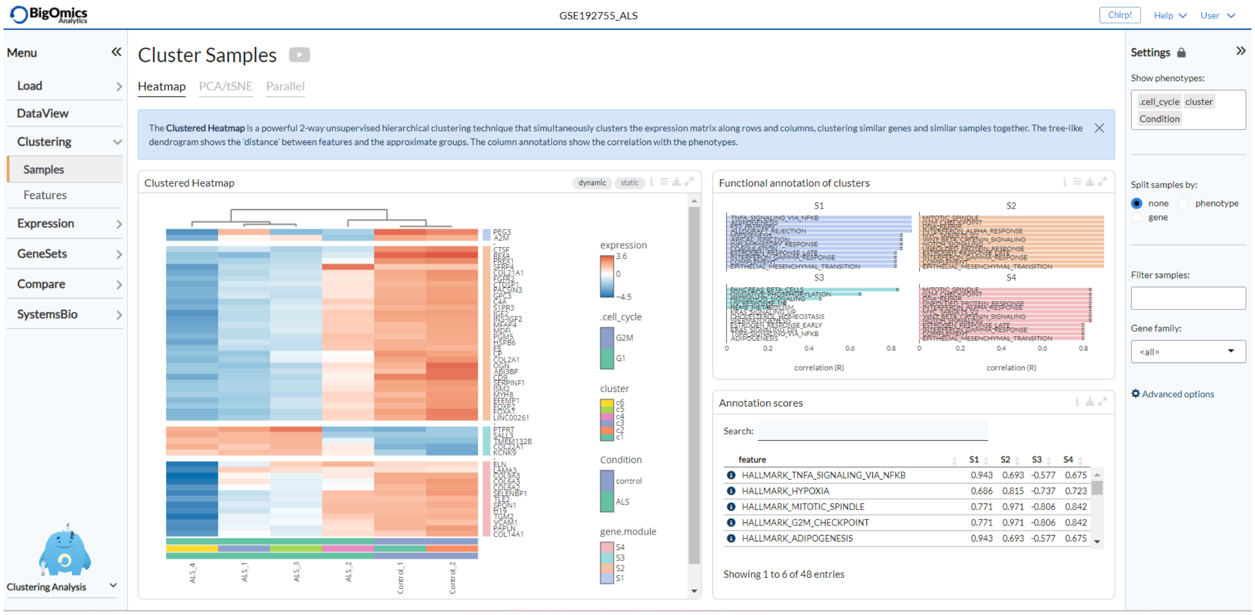
Annotation scores

Search:

feature	S1	S2	S3	S4
HALLMARK_TNFA_SIGNALING_VIA_NFKB	0.943	0.693	-0.577	0.675
HALLMARK_HYPOXIA	0.686	0.815	-0.737	0.723
HALLMARK_MITOTIC_SPINDLE	0.771	0.971	-0.806	0.842
HALLMARK_G2M_CHECKPOINT	0.771	0.971	-0.806	0.842
HALLMARK_ADIPOGENESIS	0.943	0.693	-0.577	0.675

Showing 1 to 6 of 48 entries

Activating the "dynamic" option generates an interactive version of the clustered heatmap. Users should be aware that for large datasets (such as single-cell RNA-seq data) this plot can become rather slow.



The **Clustered Heatmap** is a powerful 2-way unsupervised hierarchical clustering technique that simultaneously clusters the expression matrix along rows and columns, clustering similar genes and similar samples together. The tree-like dendrogram shows the 'distance' between features and the approximate groups. The column annotations show the correlation with the phenotypes.

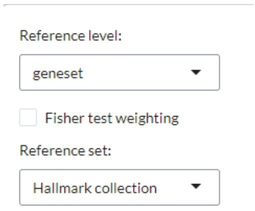
Search:

feature	S1	S2	S3	S4
HALLMARK_TNFA_SIGNALING_VIA_NFKB	0.943	0.693	-0.577	0.675
HALLMARK_HYPOXIA	0.686	0.815	-0.737	0.723
HALLMARK_MITOTIC_SPINDLE	0.771	0.971	-0.806	0.842
HALLMARK_G2M_CHECKPOINT	0.771	0.971	-0.806	0.842
HALLMARK_ADIPOGENESIS	0.943	0.693	-0.577	0.675

Showing 1 to 6 of 48 entries

15.1.3 Annotate clusters

The features in the heatmap are divided into clusters depending on the selected `top mode` in the heatmap panel settings. For each cluster, the **Annotate cluster** section provides a functional annotation using more than 42 published reference databases, including but not limited to well-known databases such as MSigDB, Wikipathways, and GO. In the plot settings, users can specify the level and reference set to be used under the `Reference level` and `Reference set` settings, respectively. Users can also select a Fisher test weighting for gene sets.



Reference level:

geneset

□ Fisher test weighting

Reference set:

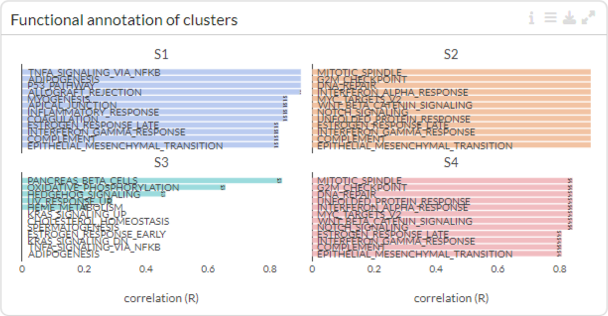
Hallmark collection

The functional annotation for the clusters are displayed below, with the highest ranking annotation features (by correlation) displayed for each gene cluster. Length of the bars corresponds to the average correlation of the cluster with the annotation term. In the table below the barplots, users can view the correlation values of annotation features for each cluster.

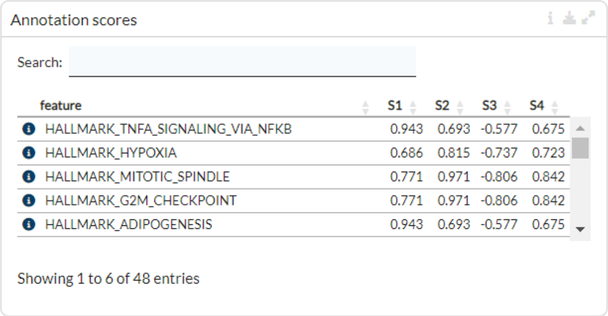
15.1.4 PCA/tSNE

The **PCA/tSNE** panel visualizes unsupervised clustering obtained by the principal components analysis (PCA), t-distributed stochastic embedding (tSNE) or Uniform Manifold Approximation and Projection (`UMAP <https://arxiv.org/abs/1802.03426>__) algorithms. This plot shows the relationship (or similarity) between the samples for visual analytics, where similarity is visualized as proximity of the points. Samples that are 'similar' will be placed close to each other.

Users can customise the PCA/tSNE/UMAP plot in the plot settings, including the `color/label` and `shape` of points using a phenotype class, the placement of the plot legend at the bottom or as a group label, the inclusion of sample labels, the choice between a 2D/3D plot display and the normalisation of the plot matrix.



Functional annotation of clusters

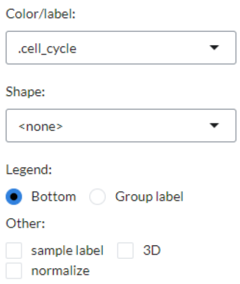


Annotation scores

Search:

feature	S1	S2	S3	S4
HALLMARK_TNFA_SIGNALING_VIA_NFKB	0.943	0.693	-0.577	0.675
HALLMARK_HYPOXIA	0.686	0.815	-0.737	0.723
HALLMARK_MITOTIC_SPINDLE	0.771	0.971	-0.806	0.842
HALLMARK_G2M_CHECKPOINT	0.771	0.971	-0.806	0.842
HALLMARK_ADIPOGENESIS	0.943	0.693	-0.577	0.675

Showing 1 to 6 of 48 entries



Color/label:

.cell_cycle

Shape:

<none>

Legend:

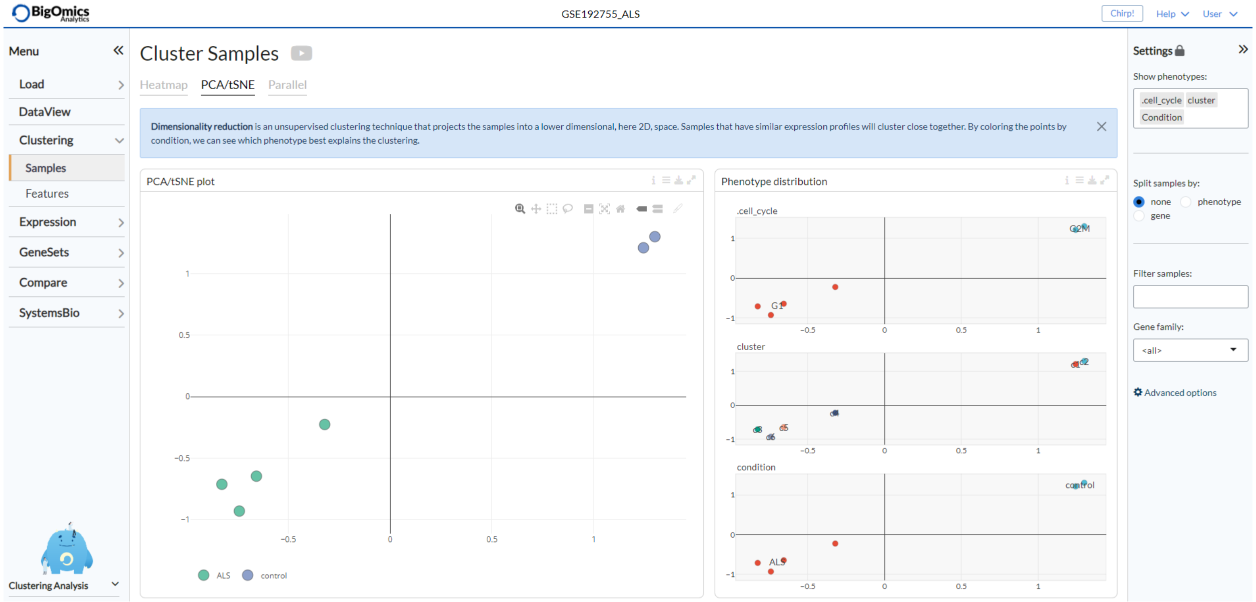
● Bottom ○ Group label

Other:

□ sample label □ 3D

□ normalize

To the right of the PCA/tSNE/UMAP plot (labelled as *PCA/tSNE plot*) the platform also generates a group of plots, entitled *Phenotype distribution*, that visualise the distribution of the available phenotype data. The plots show the distribution of the phenotypes superposed on the t-SNE clustering.



Dimensionality reduction is an unsupervised clustering technique that projects the samples into a lower dimensional, here 2D, space. Samples that have similar expression profiles will cluster close together. By coloring the points by condition, we can see which phenotype best explains the clustering.

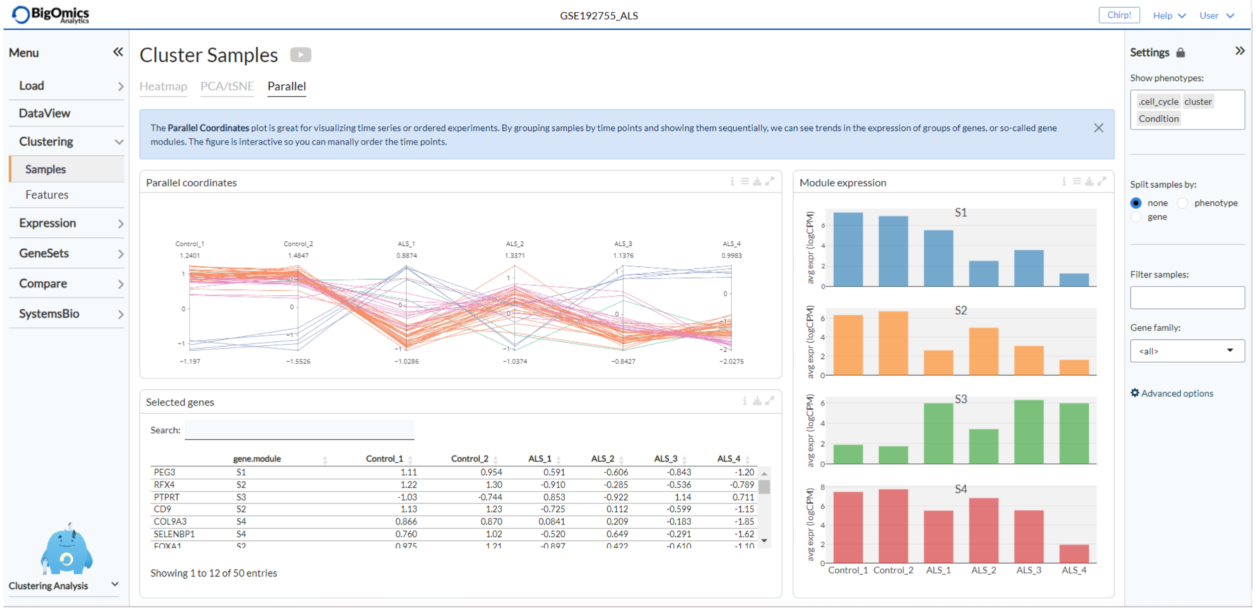
15.1.5 Parallel

The **Parallel** panel visualizes the expression levels of selected genes across all conditions in the plot labelled *Parallel coordinates*. The expression values are scaled but scaling can be removed via the plot settings, where gene expression levels can also be averaged by gene module. This interactive plot is particularly useful to users working with time series experiments, as samples can be grouped by condition (i.e. time) and ordered manually. A table (named *Selected genes*) containing average expression levels of selected genes across conditions is generated below the plot. Finally, to the right of the *Parallel coordinates* plot, a series of histograms are group together in the *Module expression* panel and display the overal expression of each module (the number of which is defined by the `K` value selected in the heatmap settings) by individual sample.

15.2 Features

15.2.1 Settings panel

Using the `Show phenotype` option the phenotypes that will be shown in the *Gene signatures* plots can be selected. Users can also select whether to use all genes for the *Gene UMAP* plot or instead select a combination of gene families under the `Annotate genes` option. The `Annotate genesets` option provides the same functionality for the *Geneset UMAP* plot based on the available geneset collections The `Show full table` option shows the full list of unfiltered genes or genesets. Under the advanced options users can select which sample group to use as a `Reference` to calculate the standard-deviation of log-expression (sd.X), or standard-deviation of the fold-change (sd.FC). If none is selected, the average values for all samples will be used instead. The final option, `UMAP datatype`, is used to select how the UMAP plot will be computed: either using the normalised log-expression (logCPM) or log-fold change matrix (logFC). logCPM is the default choice, while logFC can be used if batch or tissue effects are present in the dataset,



The **Parallel Coordinates** plot is great for visualizing time series or ordered experiments. By grouping samples by time points and showing them sequentially, we can see trends in the expression of groups of genes, or so-called gene modules. The figure is interactive so you can manally order the time points.

Search:

	gene.module	Control_1	Control_2	ALS_1	ALS_2	ALS_3	ALS_4
PEG3	S1	1.11	0.954	0.591	-0.606	-0.843	-1.20
RFX4	S2	1.22	1.30	-0.910	-0.285	-0.536	-0.789
PTPRT	S3	-1.03	-0.744	0.853	-0.922	1.14	0.711
CD9	S2	1.13	1.23	-0.725	0.112	-0.599	-1.15
COL9A3	S4	0.866	0.870	0.0841	0.209	-0.183	-1.85
SELENBP1	S4	0.760	1.02	-0.520	0.649	-0.291	-1.62
FOXA1	S2	0.975	1.21	-0.897	0.422	-0.610	-1.10

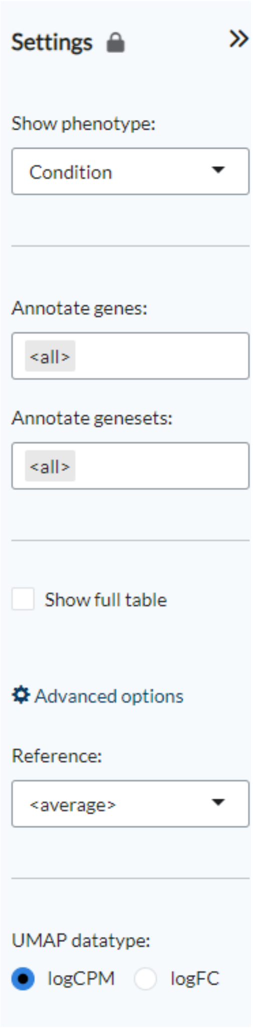
Showing 1 to 12 of 50 entries

Show phenotypes: .cell_cycle cluster Condition

Split samples by: ● none ○ phenotype ○ gene

Filter samples:

Gene family: <all>



Show phenotype: Condition

Annotate genes: <all>

Annotate genesets: <all>

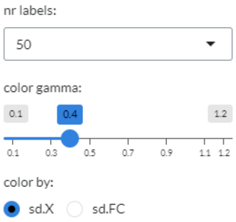
□ Show full table

Reference: <average>

UMAP datatype: ● logCPM ○ logFC

15.2.2 Gene and Geneset UMAP settings

Both the **Gene UMAP** and **Geneset UMAP** panels display the same setting options, as shown below. In it, users can use the `color by` option to select between standard-deviation of log-expression(sd.X) or standard-deviation of the fold-change (sd.FC) to colour individual genes or genesets. Users can also select the range of the colour intensity threshold using the `color gamma` option. Finally, users can choose the number of gene or genesets to be labelled in the UMAP plot under `nr labels`.



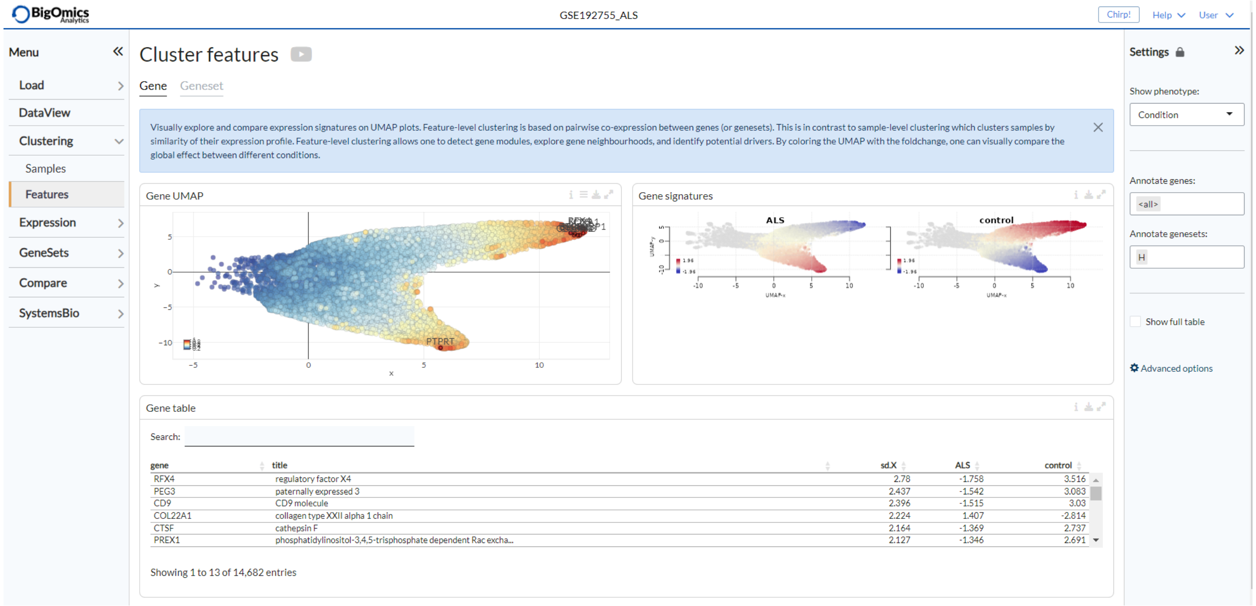
nr labels: 50

color gamma: 0.4

color by: ● sd.X ○ sd.FC

15.2.3 Gene

The **Gene** submodule contains three panels. The **Gene UMAP** panel displays the UMAP clustering of genes coloured by standard-deviation of log-expression(sd.X), or standard-deviation of the fold-change (sd.FC) and using the covariance of gene expression as a distance metric. The **Gene signatures** panel is to the right of the Gene UMAP, which shows the gene UMAP plots for each phenotypic group coloured by relative log-expression. Red indicates a gene is overexpressed in a specific group, blue that it is downregulated compared to the average values of all samples. The **Gene table** is located below the Gene UMAP and shows all the gene contained in the Gene UMAP plot and their relative change in expression. The contents of this table can be subsetted by selecting (by click&drag) on the Gene UMAP plot.



Visually explore and compare expression signatures on UMAP plots. Feature-level clustering is based on pairwise co-expression between genes (or genesets). This is in contrast to sample-level clustering which clusters samples by similarity of their expression profile. Feature-level clustering allows one to detect gene modules, explore gene neighbourhoods, and identify potential drivers. By coloring the UMAP with the foldchange, one can visually compare the global effect between different conditions.

Show phenotype: Condition

Annotate genes: <all>

Annotate genesets: H

□ Show full table

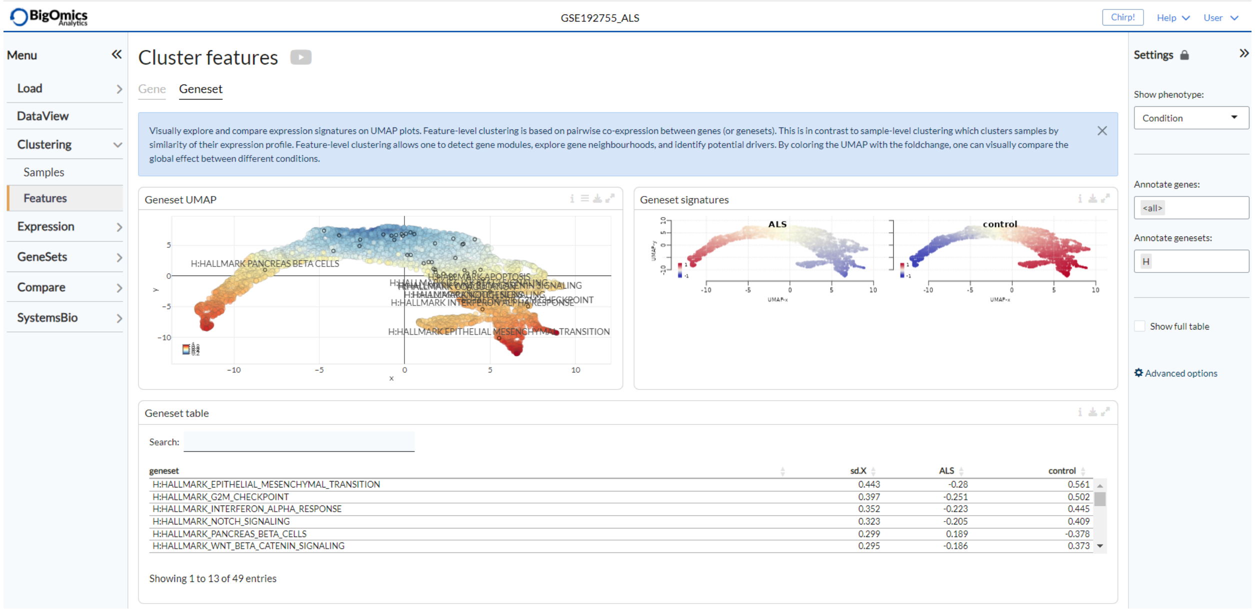
Search:

gene	title	sd.X	ALS	control
RFX4	regulatory factor X4	2.78	-1.758	3.516
PEG3	paternally expressed 3	2.437	-1.542	3.083
CD9	CD9 molecule	2.396	-1.515	3.03
COL22A1	collagen type XXII alpha 1 chain	2.224	1.407	-2.814
CTSF	cathepsin F	2.164	-1.369	2.737
PREX1	phosphatidylinositol-3,4,5-trisphosphate dependent Rac excha...	2.127	-1.346	2.691

Showing 1 to 13 of 14,682 entries

15.2.4 Geneset

The **Geneset** submodule contains three panels. The **Geneset UMAP** panel displays the UMAP clustering of genesets coloured by standard-deviation of log-expression(sd.X), or standard-deviation of the fold-change (sd.FC) and using the covariance of gene expression as a distance metric. The **Geneset signatures** panel is to the right of the Geneset UMAP, which shows the geneset UMAP plots for each phenotypic group coloured by relative log-expression. Red indicates a geneset is overexpressed in a specific group, blue that it is downregulated compared to the average values of all samples. The **Geneset table** is located below the Geneset UMAP and shows all the gene contained in the Geneset UMAP plot and their relative change in expression. The contents of this table can be subsetted by selecting (by click&drag) on the Geneset UMAP plot.



Visually explore and compare expression signatures on UMAP plots. Feature-level clustering is based on pairwise co-expression between genes (or genesets). This is in contrast to sample-level clustering which clusters samples by similarity of their expression profile. Feature-level clustering allows one to detect gene modules, explore gene neighbourhoods, and identify potential drivers. By coloring the UMAP with the foldchange, one can visually compare the global effect between different conditions.

Show phenotype: Condition

Annotate genes: <all>

Annotate genesets: H

□ Show full table

Geneset table

Search:

geneset	sd.X	ALS	control
H:HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	0.443	-0.28	0.561
H:HALLMARK_G2M_CHECKPOINT	0.397	-0.251	0.502
H:HALLMARK_INTERFERON_ALPHA_RESPONSE	0.352	-0.223	0.445
H:HALLMARK_NOTCH_SIGNALING	0.323	-0.205	0.409
H:HALLMARK_PANCREAS_BETA_CELLS	0.299	0.189	-0.378
H:HALLMARK_WNT_BETA_CATENIN_SIGNALING	0.295	-0.186	0.373

Showing 1 to 13 of 49 entries

CHAPTER

SIXTEEN

EXPRESSION

The **Expression** module is divided into three submodules. The **Differential Expression** panel compares expression between two conditions (i.e. tumor versus control), which is one of the fundamental analysis in the transcriptomics data analytics workflow. For each comparison of two conditions (also called a ‘contrast’), the analysis identifies which genes are significantly downregulated or overexpressed in one of the groups.

The **Correlation Analysis** submodule computes the correlation between genes and finds coregulated modules.

Finally, the **Find biomarkers** submodule is used to identify potential biomarkers for a set of phenotypic groups based on gene or protein expression values.

16.1 Differential expression

16.1.1 Settings panel

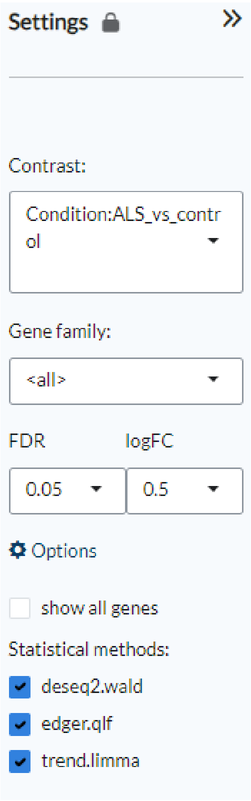
The settings panel on the right contains some settings for the analysis. Users can start the differntial expression (DE) analysis by selecting a contrast of their interest in the `Contrast` setting and specifying a relevent gene family in the `Gene family` setting. It is possible to set the false discovery rate (FDR) and the logarithmic fold change (logFC) thresholds under the `FDR` and `logFC threshold` settings, respectively. Under “Options”, users can select to display all genes in the table (not only significant genes) and also select different combinations of statistical methods for the analysis.

16.1.2 Table

The `Differential expression analysis` table shows the results of the statistical tests slected in the `Statistical methods`. By default, this table reports the meta (combined) results of DESeq2 (Wald), edgeR (QLF), and limma (trend) only. Users can filter top N = {10} differently expressed genes in the table by clicking the `top 10 genes` and also show the q-values from individual statistical methods by ticking `show individual q-values` from the table *Settings*.

For a selected comparison under the `Contrast` setting, the results of the selected methods are combined and reported in the `Differential expression analysis` table, where `meta.q` for a gene represents the highest `q` value among the methods and the number of stars for a gene indicate how many methods identified significant `q` values (`q < 0.05`). The table is interactive (scrollable, clickable); users can sort genes by `logFC`, `meta.q`, or average expression in either conditions.

By clicking on a gene in the the `Differential expression analysis` table (highlighted in grey in the figure), it is possible to see the correlation and enrichment value of gene sets that contain the gene in the the `Gene sets with gene` table.



Settings

Contrast: Condition:ALS_vs_control

Gene family: <all>

FDR 0.05 logFC 0.5

Options

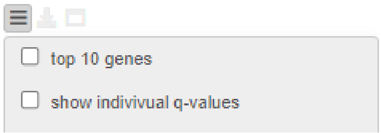
□ show all genes

Statistical methods:

☑ deseq2.wald

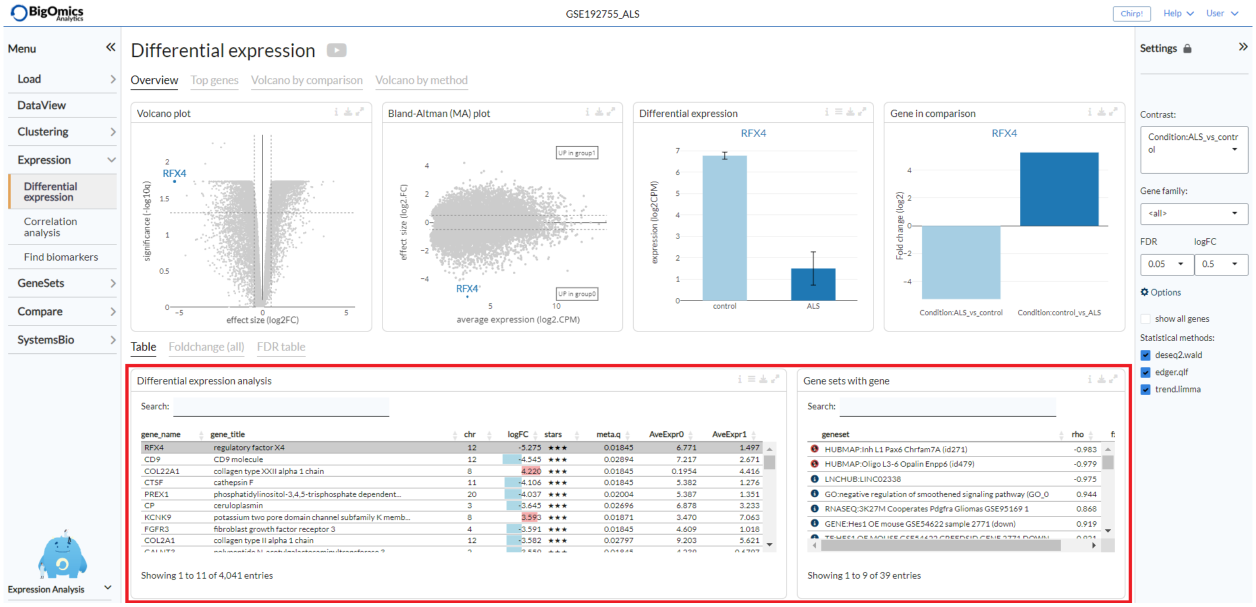
☑ edger.qlf

☑ trend.limma



□ top 10 genes

□ show indivivual q-values



Differential expression analysis

Search:

gene_name	gene_title	chr	logFC	stars	meta.q	AveExpr0	AveExpr1
RFX4	regulatory factor X4	12	-5.275	★★★	0.01845	6.771	1.497
CD9	CD9 molecule	12	-4.545	★★★	0.02894	7.217	2.671
COL22A1	collagen type XXII alpha 1 chain	8	4.220	★★★	0.01845	0.1954	4.416
CTSF	cathepsin F	11	-4.106	★★★	0.01845	5.382	1.276
PREX1	phosphatidylinositol-3,4,5-trisphosphate dependent...	20	-4.037	★★★	0.02004	5.387	1.351
CP	ceruloplasmin	3	-3.645	★★★	0.02696	6.878	3.233
KCNK9	potassium two pore domain channel subfamily K memb...	8	3.593	★★★	0.01871	3.470	7.063
FGFR3	fibroblast growth factor receptor 3	4	-3.591	★★★	0.01845	4.609	1.018
COL2A1	collagen type II alpha 1 chain	12	-3.582	★★★	0.02797	9.203	5.621

Showing 1 to 11 of 4,041 entries

Gene sets with gene

Search:

geneset	rho
HUBMAP:Inh L1 Pax6 Chrfam7A (id271)	-0.983
HUBMAP:Oligo L3-6 Opalin Enpp6 (id479)	-0.979
LNCHUB:LINC02338	-0.975
GO:negative regulation of smoothened signaling pathway (GO_0	0.944
RNASEQ:3K27M Cooperates Pdgfra Gliomas GSE95169 1	0.868
GENE:Hes1 OE mouse GSE54622 sample 2771 (down)	0.919

Showing 1 to 9 of 39 entries

Settings

Contrast: Condition:ALS_vs_control

Gene family: <all>

FDR 0.05 logFC 0.5

Options

□ show all genes

Statistical methods:

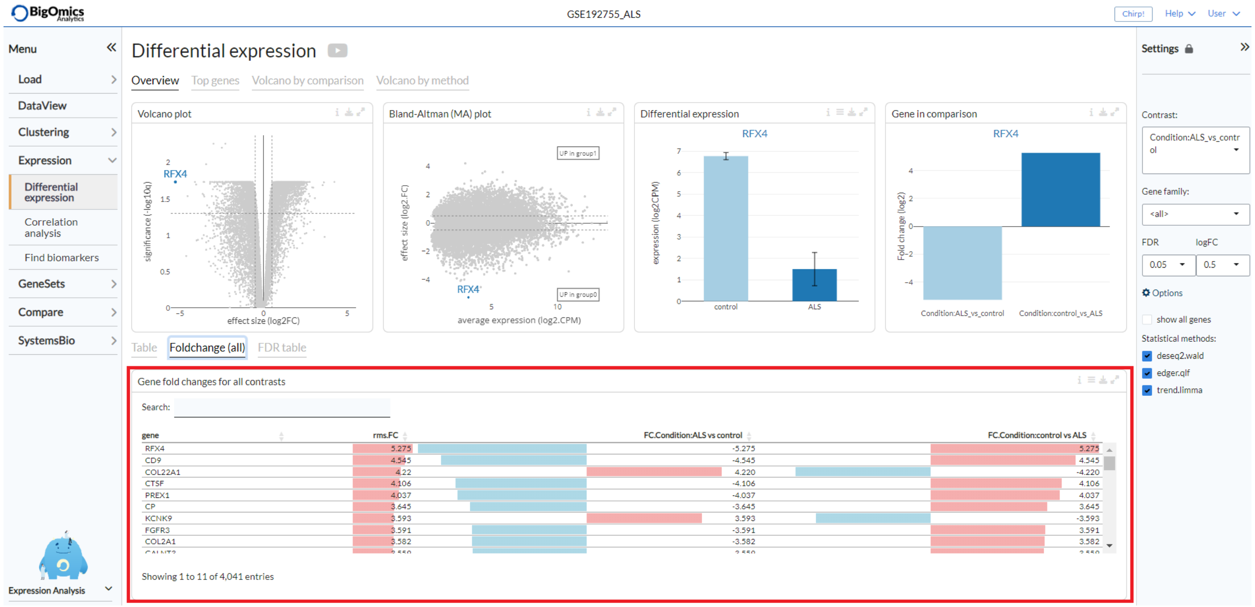
☑ deseq2.wald

☑ edger.qlf

☑ trend.limma

16.1.3 Foldchange (all)

The **Foldchange (all)** tab reports the gene fold changes for all contrasts in the selected dataset. The column `fc.var` corresponds to the variance of the fold-change across all contrasts.



Contrast: Condition:ALS_vs_control

Gene family: <all>

FDR 0.05 logFC 0.5

☐ show all genes

Statistical methods:
☑ deseq2.wald
☑ edger.qlf
☑ trend.limma

Gene fold changes for all contrasts

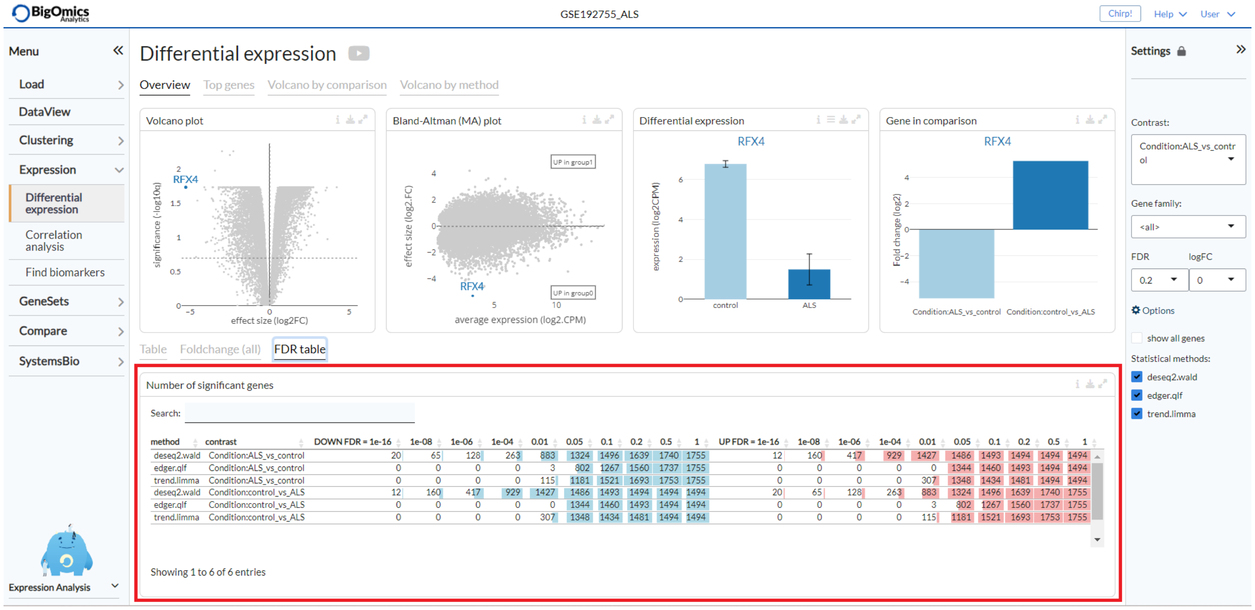
Search:

gene	rms.FC	FC.Condition:ALS vs control	FC.Condition:control vs ALS
RFX4	5.275	-5.275	5.275
CD9	4.545	-4.545	4.545
COL22A1	4.22	4.220	-4.220
CTSF	4.106	-4.106	4.106
PREX1	4.037	-4.037	4.037
CP	3.645	-3.645	3.645
KCNK9	3.593	3.593	-3.593
FGFR3	3.591	-3.591	3.591
COL2A1	3.582	-3.582	3.582

Showing 1 to 11 of 4,041 entries

16.1.4 FDR table

The **FDR table** tab reports the number of significant genes at different FDR thresholds for all contrasts and methods within the dataset. This enables to quickly see which methods are more sensitive. The left part of the table (in blue) correspond to the number of significant down-regulated genes, the right part (in red) correspond to the number of significant overexpressed genes.



Contrast: Condition:ALS_vs_control

Gene family: <all>

FDR 0.2 logFC 0

☐ show all genes

Statistical methods:
☑ deseq2.wald
☑ edger.qlf
☑ trend.limma

Number of significant genes

Search:

method	contrast	DOWN FDR = 1e-16	1e-08	1e-06	1e-04	0.01	0.05	0.1	0.2	0.5	1	UP FDR = 1e-16	1e-08	1e-06	1e-04	0.01	0.05	0.1	0.2	0.5	1
deseq2.wald	Condition:ALS_vs_control	20	65	128	263	883	1324	1496	1639	1740	1755	12	160	417	929	1427	1486	1493	1494	1494	1494
edger.qlf	Condition:ALS_vs_control	0	0	0	0	3	802	1267	1560	1737	1755	0	0	0	0	0	1344	1460	1493	1494	1494
trend.limma	Condition:ALS_vs_control	0	0	0	0	115	1181	1521	1693	1753	1755	0	0	0	0	307	1348	1434	1481	1494	1494
deseq2.wald	Condition:control_vs_ALS	12	160	417	929	1427	1486	1493	1494	1494	1494	20	65	128	263	883	1324	1496	1639	1740	1755
edger.qlf	Condition:control_vs_ALS	0	0	0	0	0	1344	1460	1493	1494	1494	0	0	0	0	3	802	1267	1560	1737	1755
trend.limma	Condition:control_vs_ALS	0	0	0	0	307	1348	1434	1481	1494	1494	0	0	0	0	115	1181	1521	1693	1753	1755

Showing 1 to 6 of 6 entries

16.1.5 Overview Plots

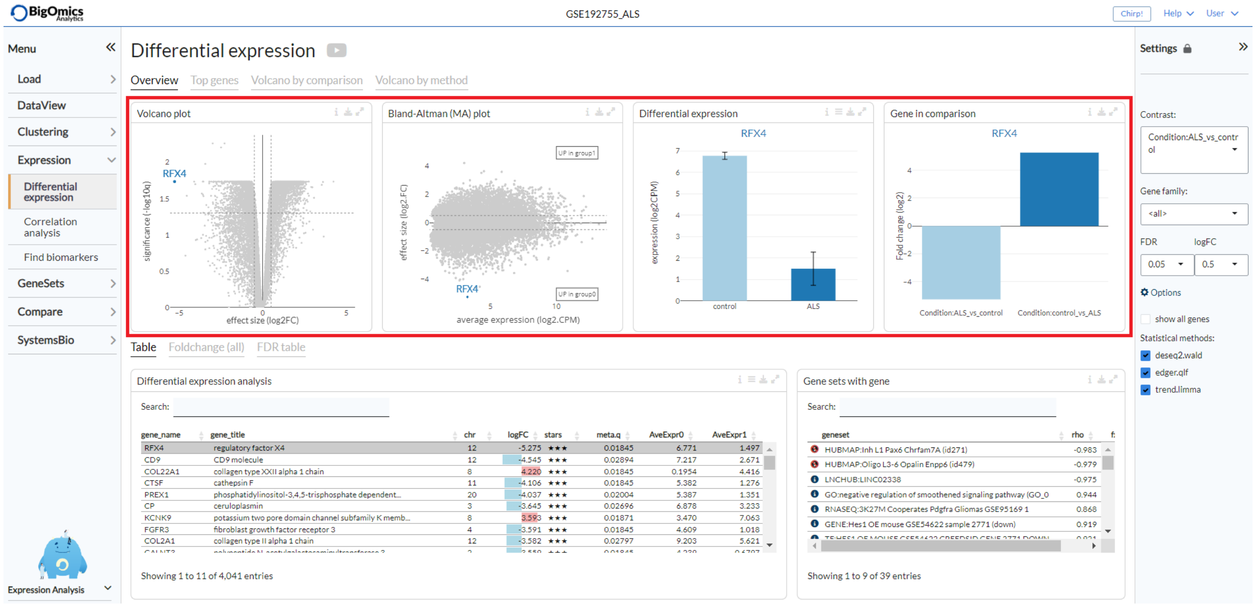
The **Overview** tab shows on top the following plots (from left to right):

Volcano Plot
An interactive volcano plot for the chosen contrast. Unless a specific gene is selected from the differential expression analysis table, all significant genes are highlighted in blue.

Bland-Altman (MA) plot
An interactive MA plot for the chosen contrast. Unless a specific gene is selected from the differential expression analysis table, all significant genes are highlighted in blue.

Differential expression
Differential expression boxplot for a gene that is selected from the differential expression analysis table. Users can customise the plot via the settings icon on top to ungroup samples and change the scale to counts per million (CPM).

Gene in comparison
Fold change summary barplot across all contrasts for a gene that is selected from the differential expression analysis table.



Differential expression analysis

Search:

gene_name	gene_title	chr	logFC	stars	meta.q	AveExpr0	AveExpr1
RFX4	regulatory factor X4	12	-5.275	★★★	0.01845	6.771	1.497
CD9	CD9 molecule	12	-4.545	★★★	0.02894	7.217	2.671
COL22A1	collagen type XXII alpha 1 chain	8	4.220	★★★	0.01845	0.1954	4.416
CTSF	cathepsin F	11	-4.106	★★★	0.01845	5.382	1.276
PREX1	phosphatidylinositol-3,4,5-trisphosphate dependent...	20	-4.037	★★★	0.02004	5.387	1.351
CP	ceruloplasmin	3	-3.645	★★★	0.02696	6.878	3.233
KCNK9	potassium two pore domain channel subfamily K memb...	8	3.593	★★★	0.01871	3.470	7.063
FGFR3	fibroblast growth factor receptor 3	4	-3.591	★★★	0.01845	4.609	1.018
COL2A1	collagen type II alpha 1 chain	12	-3.582	★★★	0.02797	9.203	5.621

Showing 1 to 11 of 4,041 entries

Gene sets with gene

Search:

geneset	rho
HUBMAP:Inh L1 Pax6 Chrfam7A (id271)	-0.983
HUBMAP:Oligo L3-6 Opalin Enpp6 (id479)	-0.979
LNCHUB:LINC02338	-0.975
GO:negative regulation of smoothened signaling pathway (GO_0	0.944
RNASEQ:3K27M Cooperates Pdgfra Gliomas GSE95169 1	0.868
GENE:Hes1 OE mouse GSE54622 sample 2771 (down)	0.919

Showing 1 to 9 of 39 entries

Contrast: Condition:ALS_vs_control

Gene family: <all>

FDR 0.05 logFC 0.5

□ show all genes

Statistical methods:

☑ deseq2.wald

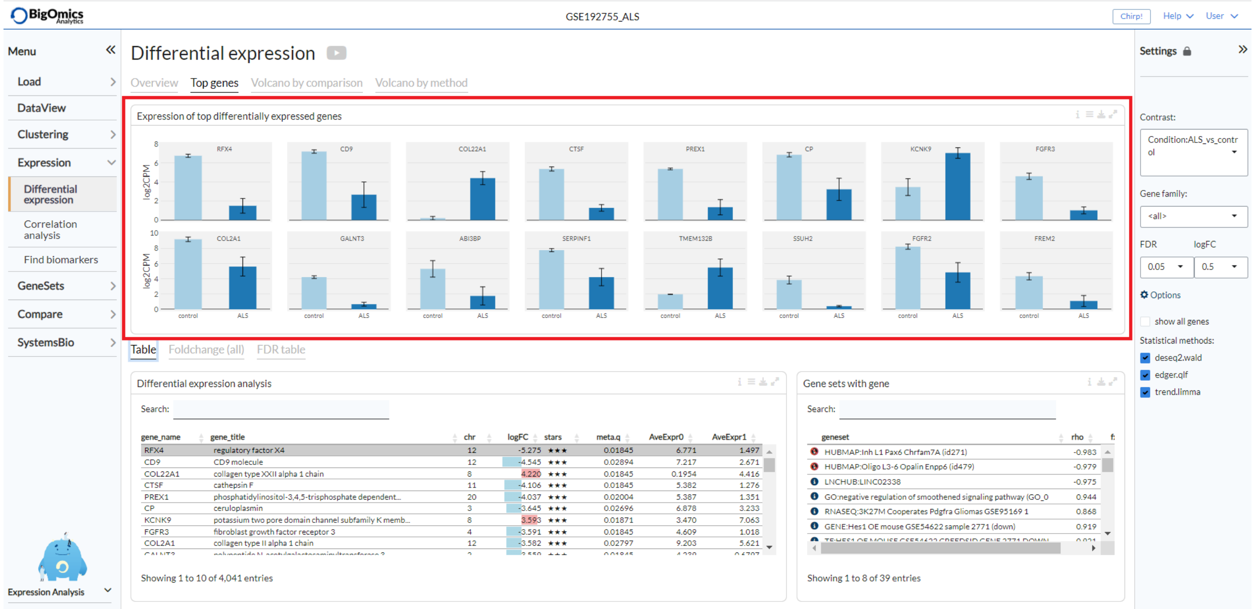
☑ edger.qlf

☑ trend.limma

16.1.6 Top genes

The **Top genes** tab shows the average expression plots across the samples for the top differentially (both positively and negatively) expressed genes for the selected comparison from the `Contrast` setting.

The plot can be customised via the settings to remove the log scale, group samples and show only samples included in the selected contrast.



Differential expression analysis

Search:

gene_name	gene_title	chr	logFC	stars	meta.q	AveExpr0	AveExpr1
RFX4	regulatory factor X4	12	-5.275	★★★	0.01845	6.771	1.497
CD9	CD9 molecule	12	-4.545	★★★	0.02894	7.217	2.671
COL22A1	collagen type XXII alpha 1 chain	8	4.220	★★★	0.01845	0.1954	4.416
CTSF	cathepsin F	11	-4.106	★★★	0.01845	5.382	1.276
PREX1	phosphatidylinositol-3,4,5-trisphosphate dependent...	20	-4.037	★★★	0.02004	5.387	1.351
CP	ceruloplasmin	3	-3.645	★★★	0.02696	6.878	3.233
KCNK9	potassium two pore domain channel subfamily K memb...	8	3.593	★★★	0.01871	3.470	7.063
FGFR3	fibroblast growth factor receptor 3	4	-3.591	★★★	0.01845	4.609	1.018
COL2A1	collagen type II alpha 1 chain	12	-3.582	★★★	0.02797	9.203	5.621

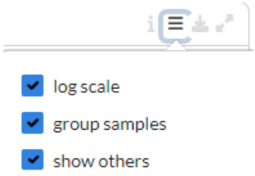
Showing 1 to 10 of 4,041 entries

Gene sets with gene

Search:

geneset	rho
HUBMAP:Inh L1 Pax6 Chrfam7A (id271)	-0.983
HUBMAP:Oligo L3-6 Opalin Enpp6 (id479)	-0.979
LNCHUB:LINC02338	-0.975
GO:negative regulation of smoothened signaling pathway (GO_0	0.944
RNASEQ:3K27M Cooperates Pdgfra Gliomas GSE95169 1	0.868
GENE:Hes1 OE mouse GSE54622 sample 2771 (down)	0.919

Showing 1 to 8 of 39 entries



☑ log scale

☑ group samples

☑ show others

16.1.7 Volcano by comparison

Under the **Volcano by comparison** tab, the platform simultaneously displays multiple volcano plots for genes across all contrasts. By comparing multiple volcano plots, the user can immediately see which comparison is statistically weak or strong. Experimental contrasts with better statistical significance will show volcano plots with 'higher' wings.

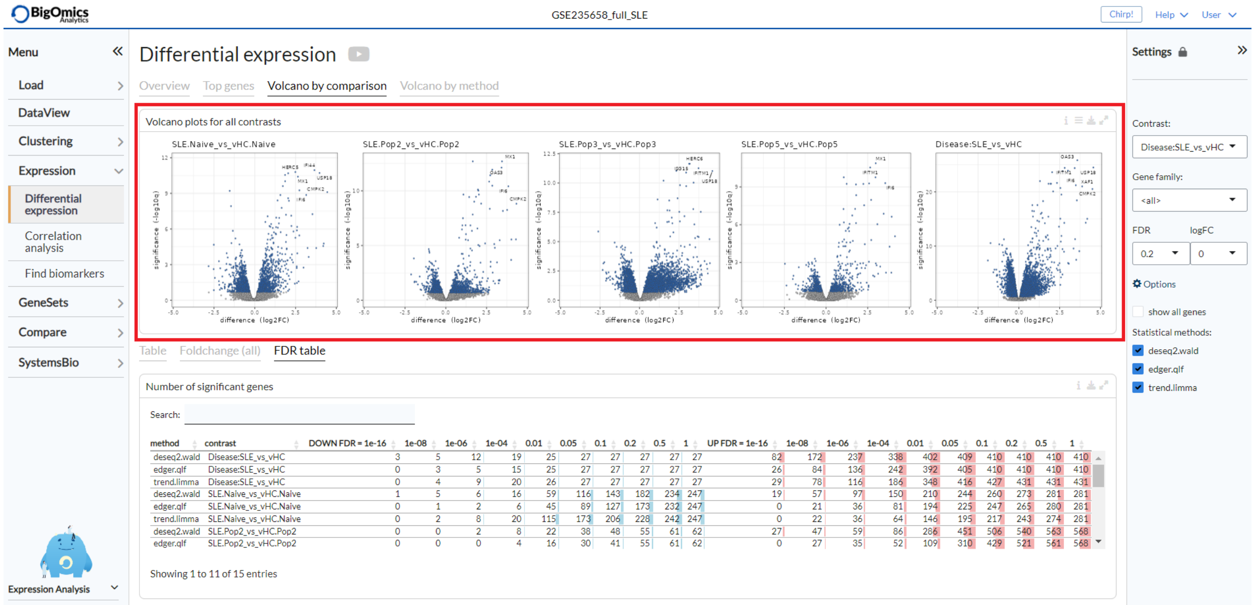
16.1.8 Volcano by method

Under the **Volcano by method** tab, the platform displays the volcano plots provided by multiple differential expression calculation methods for the selected contrast. Methods showing better statistical significance will show volcano plots with 'higher' wings.

16.2 Correlation analysis

16.2.1 Settings panel

The panel contains the main settings for the analysis. The analysis can be started by selecting a gene of interest from the `Gene` settings. Users can filter for a specific gene family by using the `Filter genes` setting. Finally, users can select the number of top genes used to compute partial correlation under the homonymous setting.



Contrast: Disease:SLE_vs_vHC

Gene family: <all>

FDR 0.2 logFC 0

□ show all genes

Statistical methods:

☑ deseq2.wald

☑ edger.qlf

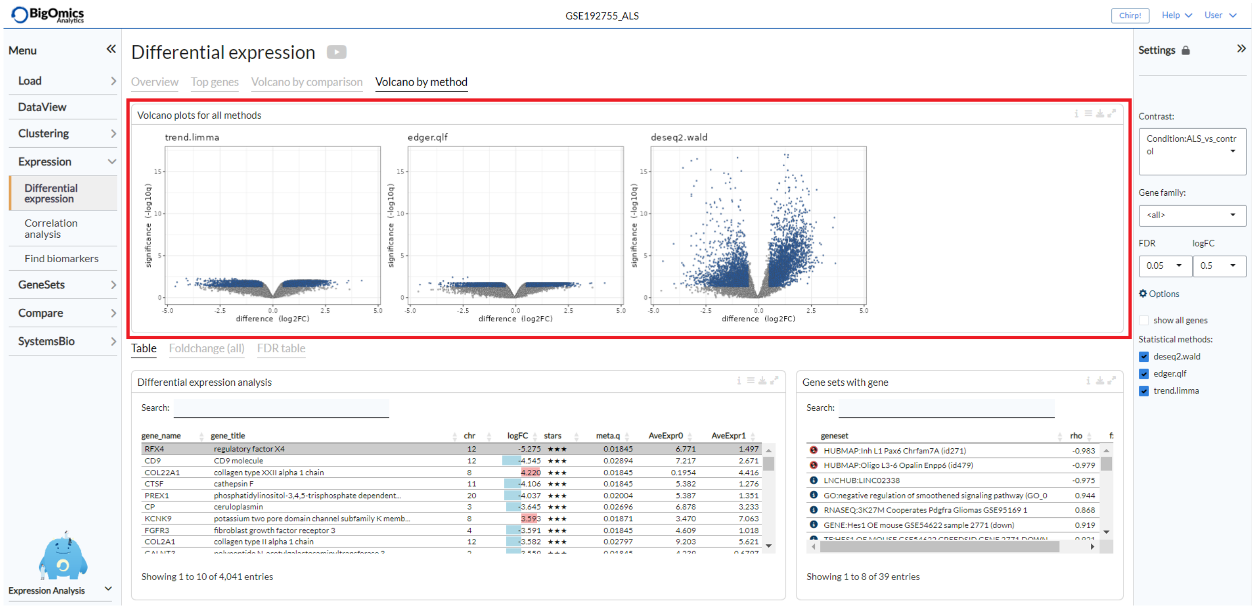
☑ trend.limma

Number of significant genes

Search:

method	contrast	DOWN FDR = 1e-16	1e-08	1e-06	1e-04	0.01	0.05	0.1	0.2	0.5	1	UP FDR = 1e-16	1e-08	1e-06	1e-04	0.01	0.05	0.1	0.2	0.5	1
deseq2.wald	Disease:SLE_vs_vHC	3	5	12	19	25	27	27	27	27	27	82	172	237	338	402	409	410	410	410	410
edger.qlf	Disease:SLE_vs_vHC	0	3	5	15	25	27	27	27	27	27	26	84	136	242	392	405	410	410	410	410
trend.limma	Disease:SLE_vs_vHC	0	4	9	20	26	27	27	27	27	27	29	78	116	186	348	416	427	431	431	431
deseq2.wald	SLE.Naive_vs_vHC.Naive	1	5	6	16	59	116	143	182	234	247	19	57	97	150	210	244	260	273	281	281
edger.qlf	SLE.Naive_vs_vHC.Naive	0	1	2	6	45	89	127	173	232	247	0	21	36	81	194	225	247	265	280	281
trend.limma	SLE.Naive_vs_vHC.Naive	0	2	8	20	115	173	206	228	242	247	0	22	36	64	146	195	217	243	274	281
deseq2.wald	SLE.Pop2_vs_vHC.Pop2	0	0	2	8	22	38	48	55	61	62	27	47	59	86	286	451	506	540	563	568
edger.qlf	SLE.Pop2_vs_vHC.Pop2	0	0	0	4	16	30	41	55	61	62	0	27	35	52	109	310	429	521	561	568

Showing 1 to 11 of 15 entries



Contrast: Condition:ALS_vs_control

Gene family: <all>

FDR 0.05 logFC 0.5

□ show all genes

Statistical methods:

☑ deseq2.wald

☑ edger.qlf

☑ trend.limma

Differential expression analysis

Search:

gene_name	gene_title	chr	logFC	stars	meta.q	AveExpr0	AveExpr1
RFX4	regulatory factor X4	12	-5.275	★★★	0.01845	6.771	1.497
CD9	CD9 molecule	12	-4.545	★★★	0.02894	7.217	2.671
COL22A1	collagen type XXII alpha 1 chain	8	4.220	★★★	0.01845	0.1954	4.416
CTSF	cathepsin F	11	-4.106	★★★	0.01845	5.382	1.276
PREX1	phosphatidylinositol-3,4,5-trisphosphate dependent...	20	-4.037	★★★	0.02004	5.387	1.351
CP	ceruloplasmin	3	-3.645	★★★	0.02696	6.878	3.233
KCNK9	potassium two pore domain channel subfamily K memb...	8	3.593	★★★	0.01871	3.470	7.063
FGFR3	fibroblast growth factor receptor 3	4	-3.591	★★★	0.01845	4.609	1.018
COL2A1	collagen type II alpha 1 chain	12	-3.582	★★★	0.02797	9.203	5.621

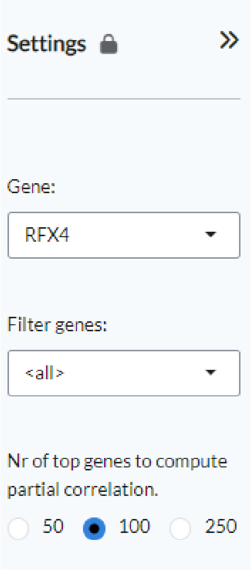
Showing 1 to 10 of 4,041 entries

Gene sets with gene

Search:

geneset	rho
HUBMAP:Inh L1 Pax6 Chrfam7A (id271)	-0.983
HUBMAP:Oligo L3-6 Opalin Enpp6 (id479)	-0.979
LNCHUB:LINC02338	-0.975
GO:negative regulation of smoothened signaling pathway (GO_0	0.944
RNASEQ:3K27M Cooperates Pdgfra Gliomas GSE95169 1	0.868
GENE:Hes1 OE mouse GSE54622 sample 2771 (down)	0.919

Showing 1 to 8 of 39 entries



Settings

Gene:

RFX4

Filter genes:

<all>

Nr of top genes to compute partial correlation.

○ 50 ● 100 ○ 250

16.2.2 Correllation

Under the **Correlation** tab, the platform displays three different outputs:

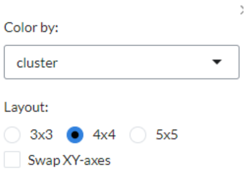
Top correlated genes

A plot displaying the highest correlated genes in respect to the selected gene.

Correlation scatter plots

Pairwise scatter plots for the co-expression of correlated gene pairs across the samples.

The straight line correspond to the (linear) regression fit. Using the settings on top, the plot can be customised by changing the colour of the gene pairs by phenotype using the `Colour by` option. Users can also change the layout of the plots under ``Layout`.



Color by:

cluster

Layout:

○ 3x3 ● 4x4 ○ 5x5

□ Swap XY-axes

Correlation table

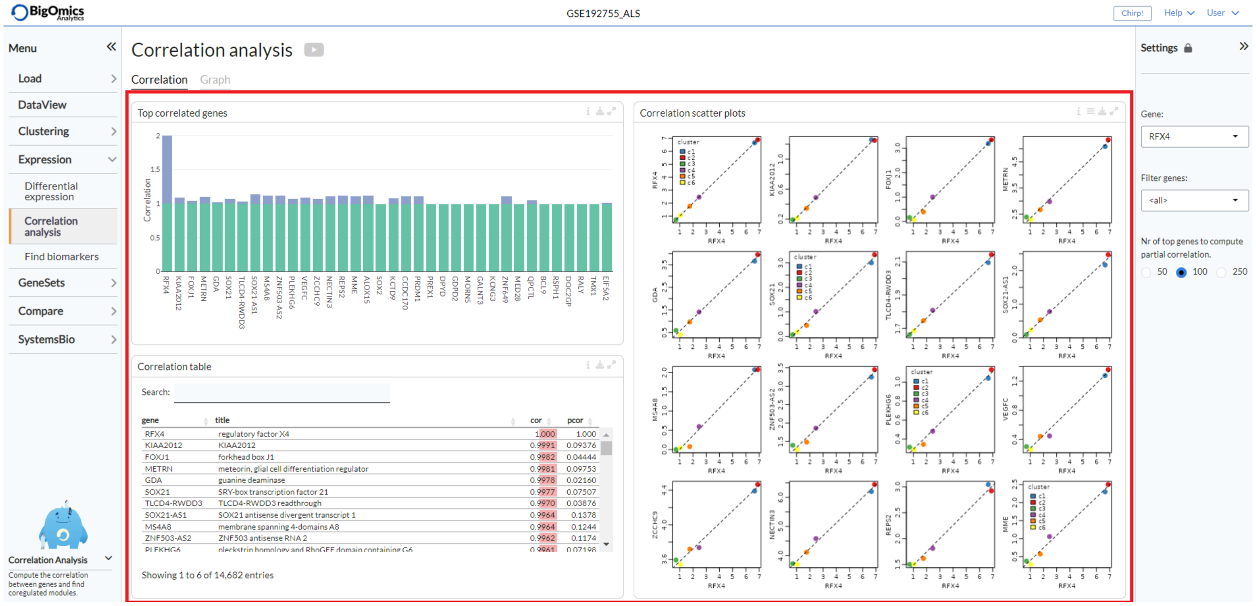
it shows the statistical results from correlated gene pairs.

16.2.3 Graph

Two plots are generated under the **Graph** tab:

Partial correlation network

This plot shows the correlations between gene pairs. Grey edges correspond to positive correlation, red edges correspond to negative correlation. The width of the edge is proportional to the absolute partial correlation value of the gene pair. It can be customised from the plot settings, where the radius of the plot and the partial correlation threshold (`pcor threshold`) can be altered. Users can also the choose from four different layouts (Fruchterman-Reingold, Kamada-Kawai, graphopt and tree layout) under the `Layout` option.



Correlation analysis

Compute the correlation between genes and find coregulated modules.

Gene: RFX4

Filter genes: <all>

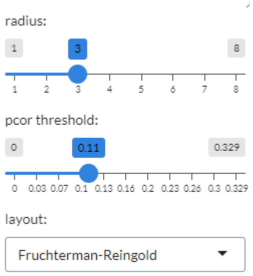
Nr of top genes to compute partial correlation.
○ 50 ● 100 ○ 250

Correlation table

Search:

gene	title	cor	pcor
RFX4	regulatory factor X4	1.000	1.000
KIAA2012	KIAA2012	0.9991	0.09376
FOXJ1	forkhead box J1	0.9982	0.04444
METRN	meteorin, glial cell differentiation regulator	0.9981	0.09753
GDA	guanine deaminase	0.9978	0.02160
SOX21	SRY-box transcription factor 21	0.9977	0.07507
TLCD4-RWDD3	TLCD4-RWDD3 readthrough	0.9970	0.03876
SOX21-AS1	SOX21 antisense divergent transcript 1	0.9964	0.1378
MS4A8	membrane spanning 4-domains A8	0.9964	0.1244
ZNF503-AS2	ZNF503 antisense RNA 2	0.9962	0.1174
PLEKHG6	pleckstrin homology and RhoGEF domain containing G6	0.9961	0.07198

Showing 1 to 6 of 14,682 entries



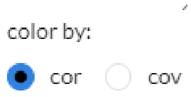
radius: 3

pcor threshold: 0.11

layout: Fruchterman-Reingold

Correlation UMAP

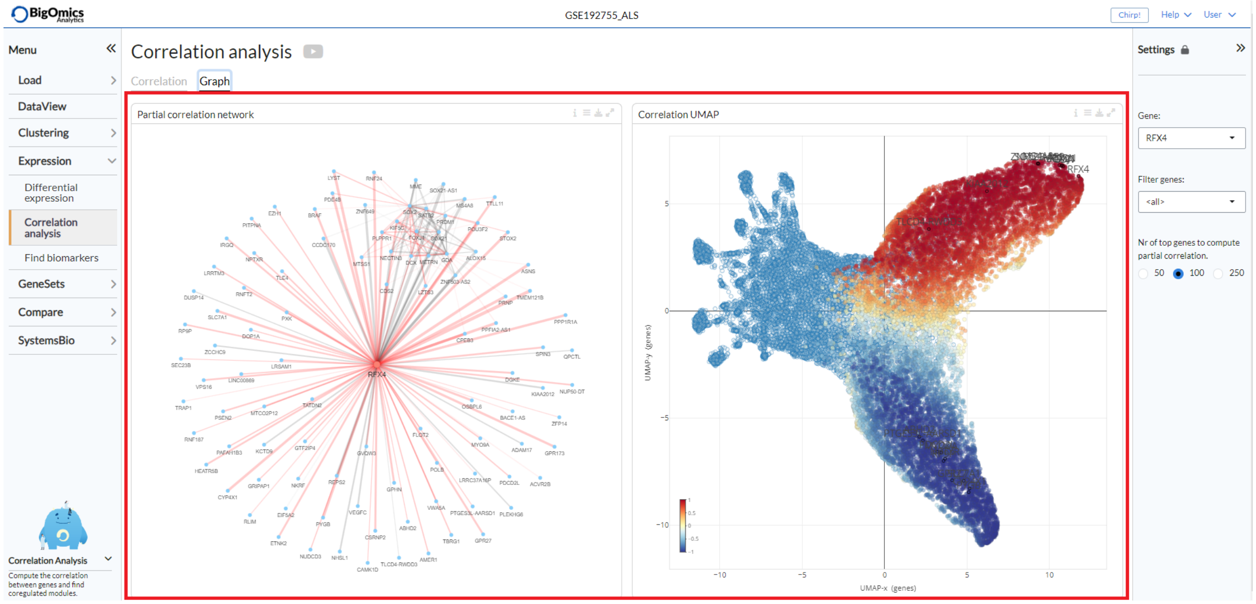
UMAP plot showing the correlation between genes. Genes that are correlated are generally positioned close to each other. Red corresponds to positive correlation/covariance, blue to negative. User can select whether to colour the genes by correlation or covariance in plot settings.



color by:

● cor ○ cov

The **Partial correlation network** is on the left, while the **Correlation UMAP** is on the right of the page.



Gene: RFX4

Filter genes: <all>

Nr of top genes to compute partial correlation.

○ 50 ● 100 ○ 250

16.3 Find Biomarkers

16.3.1 Settings panel

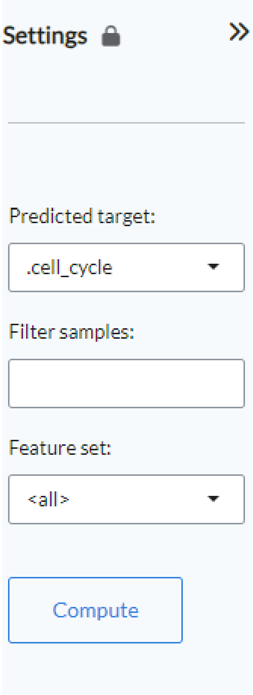
The panel contains the main settings for the analysis. Users select the phenotype on which to perform the analysis with the `Predicted target` setting. In the `Filter samples` setting users can restrict the analysis to a subset of samples rather than the default choice of all samples. The `Feature set` setting will by default be set to "all", but users can restrict the calculations only to a specific gene family or, alteratively, they can paste their own selection of genes. Once all the settings have been assigned, clicking *Compute* will start the calculation.

16.3.2 Feature Selection

Four panels (from left to right and top to bottom) are present under the **Feature Selection** tab:

Variable Importance

Omics Playground calculates a variable importance score for each feature using multiple state-of-the-art machine learning algorithms, including LASSO, elastic nets, random forests, and extreme gradient boosting. Note that we do not use the machine learning algorithms for prediction but we use them just to compute the variable importances according to the different methods. An aggregated score is then calculated as the cumulative rank of the variable importances of the different algorithms.



Predicted target: .cell_cycle

Filter samples:

Feature set: <all>

By combining several methods, the platform aims to select the best possible predictive features. The top features are determined as the features with the highest cumulative ranks. .

Biomarker Expression

These boxplots shows the expression of putative biomarkers across the samples of the identified features.

Heatmap

Expression heatmap of top gene features according to their variable importance. By default the plot will only show the selected samples, but all samples can be displayed via the settings button on top.

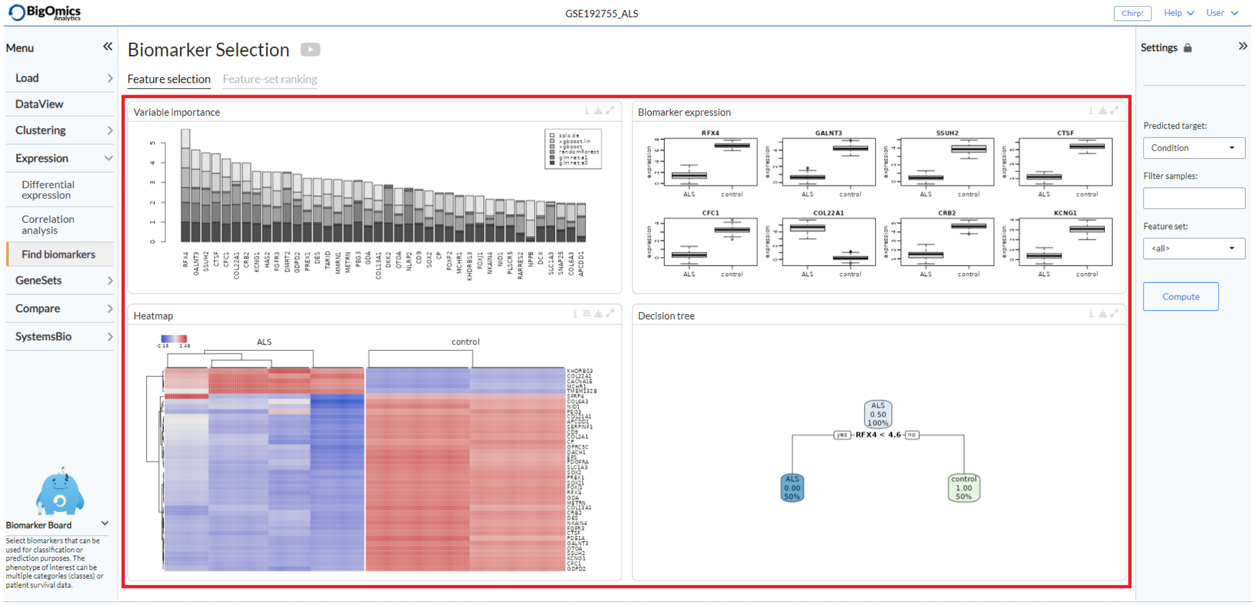
Decision tree

The decision tree shows a tree solution for classification based on the top most important features. The plot provides a proportion of the samples that are defined by each biomarker in the boxes.

16.3.3 Feature-set Ranking

This tab consists of a single plot, the aptly named **Feature-set Ranking**, which displays the ranked discriminant score for top feature sets. The plot ranks the discriminative power of the feature set (or gene family) as a cumulative discriminant score for all phenotype variables. In this way, we can find which feature set (or gene family) can explain the variance in the data the best.

Users can choose between three different methods for the calculation of the plot. P-value based scoring ('p-value') is computed as the average negative log p-value from the ANOVA test. Correlation-based discriminative power ('correlation') is calculated as the average 1-cor between the groups. Thus, a feature set is highly discriminative if the between-group correlation is low. The 'meta' method combines the scores of the two aforementioned methods in a multiplicative manner.

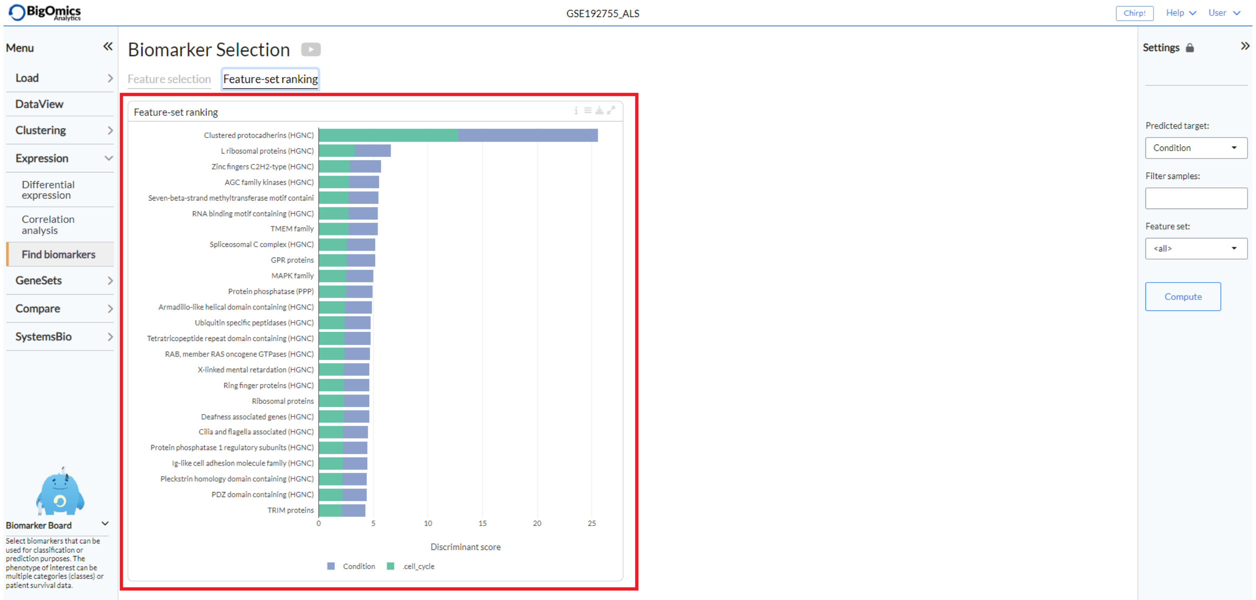


Select biomarkers that can be used for classification or prediction purposes. The phenotype of interest can be multiple categories (classes) or patient survival data.

Predicted target: Condition

Filter samples:

Feature set: <all>

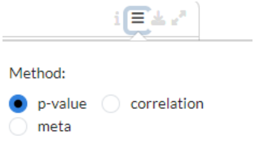

BigOmics Analytics
GSE192755_ALS
Chirp!
Help
User
Menu
Load
DataView
Clustering
Expression
Differential expression
Correlation analysis
Find biomarkers
GeneSets
Compare
SystemsBio
Biomarker Board
Biomarker Selection
Feature selection
Feature-set ranking
Feature-set ranking
Clustered protocadherins (HGNC)
L ribosomal proteins (HGNC)
Zinc fingers C2H2-type (HGNC)
AGC family kinases (HGNC)
Seven-beta-strand methyltransferase motif containi
RNA binding motif containing (HGNC)
TMEM family
Spliceosomal C complex (HGNC)
GPR proteins
MAPK family
Protein phosphatase (PPP)
Armadillo-like helical domain containing (HGNC)
Ubiquitin specific peptidases (HGNC)
Tetratricopeptide repeat domain containing (HGNC)
RAB, member RAS oncogene GTPases (HGNC)
X-linked mental retardation (HGNC)
Ring finger proteins (HGNC)
Ribosomal proteins
Deafness associated genes (HGNC)
Cilia and flagella associated (HGNC)
Protein phosphatase 1 regulatory subunits (HGNC)
Ig-like cell adhesion molecule family (HGNC)
Pleckstrin homology domain containing (HGNC)
PDZ domain containing (HGNC)
TRIM proteins
0
5
10
15
20
25
Discriminant score
Condition
.cell_cycle
Settings
Compute

Select biomarkers that can be used for classification or prediction purposes. The phenotype of interest can be multiple categories (classes) or patient survival data.

Predicted target: Condition

Filter samples:

Feature set: <all>



Method:

● p-value ○ correlation
○ meta

CHAPTER

SEVENTEEN

GENESETS

This module is subdivided into four submodules: **Geneset enrichment**, **Test Geneset**, **Pathway Analysis** and **Word cloud**.

17.1 Geneset enrichment

Similar to the differential gene expression analysis, users can perform differential expression analysis on a geneset level in this page, which is also referred as gene set enrichment (GSE) analysis. The platform has more than 50.000 genesets (or pathways) in total that are divided more than 30 geneset collections such Hallmark, Reactome, Wikipathways, and Gene Ontology (GO). Users have to specify which comparison they want to visually analyze employing a certain geneset collection.

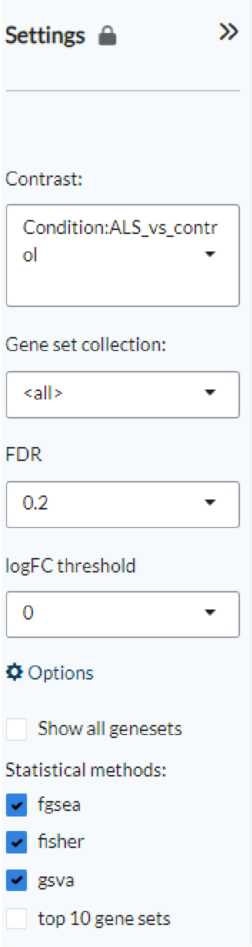
The tab consists of the settings panels as well as four results panels: **Top enriched gene sets**, **Frequency in top gene sets**, a **Table** panel where users can swicth between three tables and **Genes in gene sets**.

17.1.1 Settings panel

The enrichment analysis can be started by selecting a contrats of interest in the `Contrast` and specifying a relevent gene set family in the `Gene set collection`. It is possible to set the false discovery rate (FDR) and the logarithmic fold change (logFC) thresholds under the `FDR` and `logFC threshold` settings, respectively. We allow users to show all gene sets in the``Enrichment analysis`` table , rather than just the significant ones, under the *Options* menu. Users can also select statistical methods for the enrichment analysis from the same menu. To ensure the statistical reliability, the platform performs enrichment analyses using multiple methods, including Spearman rank correlation, GSVA, ssGSEA, Fisher's exact test, GSEA, camera and fry. Finally users can limit the results to only the top 10 up- and down-regulated gene sets.

17.1.2 Top enriched gene sets

This panel displays the enrichment plots of the top differentially enriched gene sets (up and down) for the selected contrast. Black vertical bars indicate the rank of genes in the gene set in the sorted list metric. The green curve corresponds to the enrichment score (ES). The more the green ES curve is shifted to the upper left of the graph, the more the gene set is enriched in the first group. Conversely, a shift of the ES curve to the lower right, corresponds to more enrichment in the second group. This panel will by default display the 12 most up-regulated genesets in the selected contrast. Selecting a specific gene set from the 'Enrichment analysis' table below it will display the selected gene set alone.



Settings

Contrast:

Condition:ALS_vs_control

Gene set collection:

<all>

FDR

0.2

logFC threshold

0

Options

□ Show all genesets

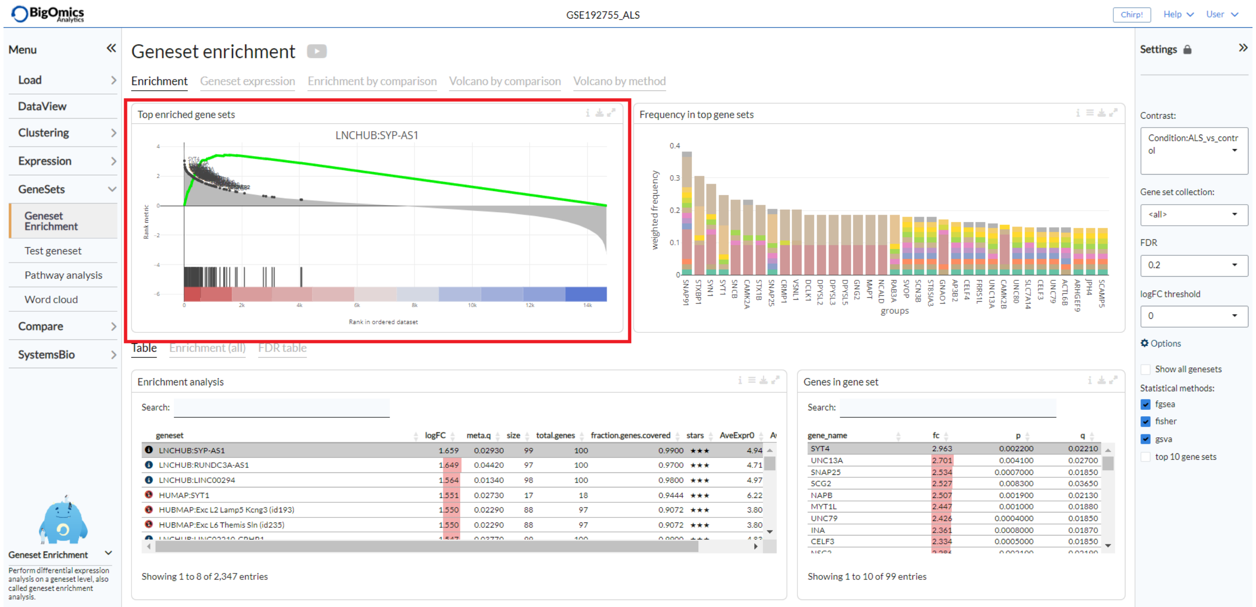
Statistical methods:

☑ fgsea

☑ fisher

☑ gsva

□ top 10 gene sets



Geneset enrichment

Enrichment analysis

Search:

geneset	logFC	meta.q	size	total.genes	fraction.genes.covered	stars	AveExpr0
LNCHUB:SYP-AS1	1.659	0.02930	99	100	0.9900	★★★	4.94
LNCHUB:RUNDC3A-AS1	1.649	0.04420	97	100	0.9700	★★★	4.71
LNCHUB:LINC00294	1.564	0.01340	98	100	0.9800	★★★	4.97
HUMAP:SYT1	1.551	0.02730	17	18	0.9444	★★★	6.22
HUBMAP:Exc L2 Lamp5 Kcng3 (id193)	1.550	0.02290	88	97	0.9072	★★★	3.80
HUBMAP:Exc L6 Themis Sln (id235)	1.550	0.02290	88	97	0.9072	★★★	3.80

Showing 1 to 8 of 2,347 entries

Genes in gene set

Search:

gene_name	fc	p	q
SYT4	2.963	0.002200	0.02210
UNC13A	2.701	0.004100	0.02700
SNAP25	2.534	0.0007000	0.01850
SCG2	2.527	0.008300	0.03650
NAPB	2.507	0.001900	0.02130
MYT1L	2.447	0.001000	0.01880
UNC79	2.426	0.0004000	0.01850
INA	2.361	0.0008000	0.01870
CELF3	2.334	0.0005000	0.01850

Showing 1 to 10 of 99 entries

Settings

Contrast: Condition:ALS_vs_control

Gene set collection: <all>

FDR 0.2

logFC threshold 0

Options

□ Show all genesets

Statistical methods:

☑ fgsea

☑ fisher

☑ gsva

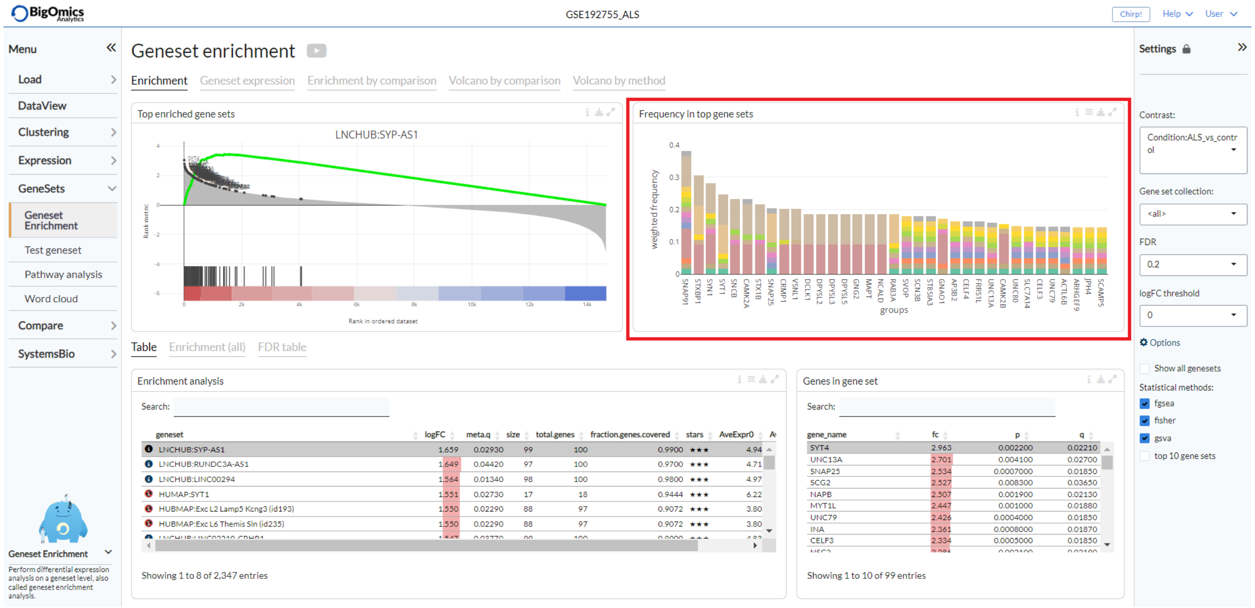
□ top 10 gene sets

Geneset Enrichment

Perform differential expression analysis on a geneset level, also called geneset enrichment analysis.

17.1.3 Frequency in top gene sets

This panel shows the number of times a gene is present in the top-N genesets sorted by frequency. Genes that are frequently shared among the top enriched gene sets may suggest driver genes.



Geneset Enrichment

Perform differential expression analysis on a geneset level, also called geneset enrichment analysis.

Contrast: Condition:ALS_vs_control

Gene set collection: <all>

FDR 0.2

logFC threshold 0

□ Show all genesets

Statistical methods:

☑ fgsea

☑ fisher

☑ gsva

□ top 10 gene sets

Enrichment analysis

Search:

geneset	logFC	meta.q	size	total.genes	fraction.genes.covered	stars	AveExpr0
LNCHUB:SYP-AS1	1.659	0.02930	99	100	0.9900	★★★	4.94
LNCHUB:RUNDC3A-AS1	1.649	0.04420	97	100	0.9700	★★★	4.71
LNCHUB:LINC00294	1.564	0.01340	98	100	0.9800	★★★	4.97
HUMAP:SYT1	1.551	0.02730	17	18	0.9444	★★★	6.22
HUBMAP:Exc L2 Lamp5 Kcng3 (id193)	1.550	0.02290	88	97	0.9072	★★★	3.80
HUBMAP:Exc L6 Themis Sln (id235)	1.550	0.02290	88	97	0.9072	★★★	3.80

Showing 1 to 8 of 2,347 entries

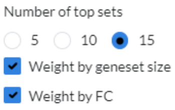
Genes in gene set

Search:

gene_name	fc	p	q
SYT4	2.963	0.002200	0.02210
UNC13A	2.701	0.004100	0.02700
SNAP25	2.534	0.0007000	0.01850
SCG2	2.527	0.008300	0.03650
NAPB	2.507	0.001900	0.02130
MYT1L	2.447	0.001000	0.01880
UNC79	2.426	0.0004000	0.01850
INA	2.361	0.0008000	0.01870
CELF3	2.334	0.0005000	0.01850

Showing 1 to 10 of 99 entries

The settings icon open the settings menu from which users can select the number of top gene sets to be used (`Number of top sets`), whether to weight by geneset size (`Weight by geneset size`) and whether to weight by fold-change (`Weight by FC`).



Number of top sets

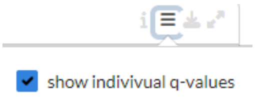
○ 5 ○ 10 ● 15

☑ Weight by geneset size

☑ Weight by FC

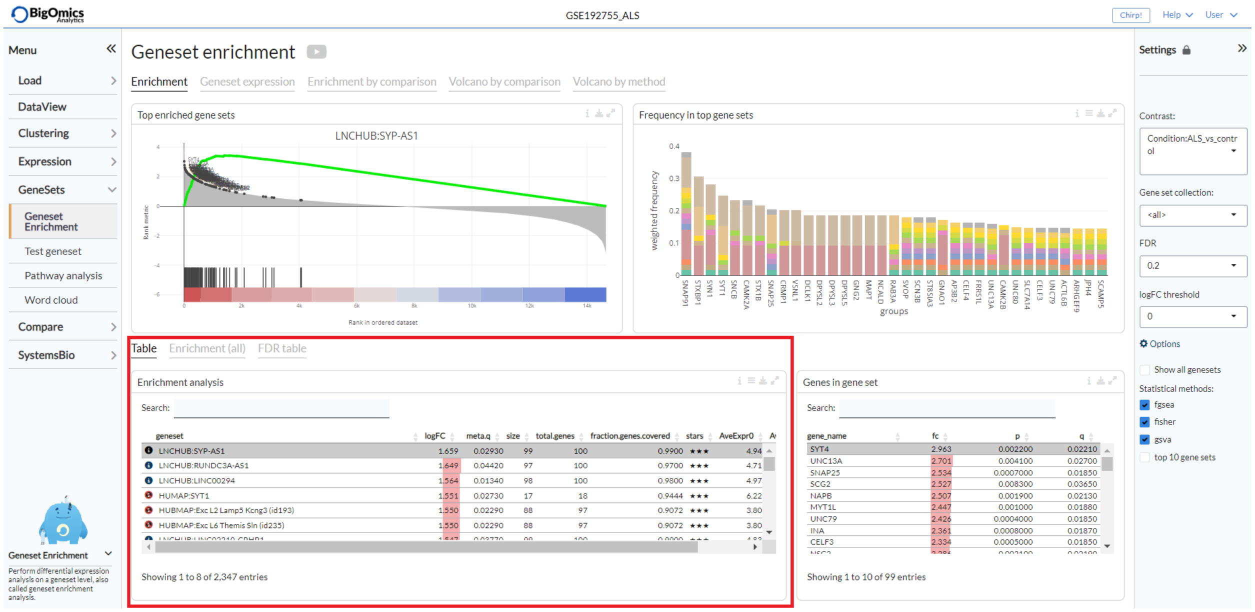
17.1.4 Table

This table shows the enrichment analysis results from the selected statistical methods. With default settings, this table reports the meta (combined) results of camera, fgsea, and Spearman rank correlation only. Users can also display individual q-values for each of the selected analysis methods using the settings on top of the table.



☑ show indivivual q-values

The combined enrichment analysis results from the methods are displayed in this table, where for each geneset the `meta.q` corresponds to the highest `q` value provided by the methods and the number of `stars` indicate how many methods identified the geneset as significant (`q < 0.05`). `AveExpr0` and `AveEprx1` refer to the average gene set expression in group 0 and group 1 of the selected pairwise comparison respectively. The table is interactive; users can sort it by `logFC`, `meta.q`, `AveExpr0`, `AveEprx1` and `stars`.



Contrast: Condition:ALS_vs_control

Gene set collection: <all>

FDR: 0.2

logFC threshold: 0

Options

□ Show all genesets

Statistical methods:

☑ fgsea

☑ fisher

☑ gsva

□ top 10 gene sets

Enrichment analysis

Search:

geneset	logFC	meta.q	size	total.genes	fraction.genes.covered	stars	AveExpr0
LNCHUB:SYP-AS1	1.659	0.02930	99	100	0.9900	★★★	4.94
LNCHUB:RUNDC3A-AS1	1.649	0.04420	97	100	0.9700	★★★	4.71
LNCHUB:LINC00294	1.564	0.01340	98	100	0.9800	★★★	4.97
HUMAP:SYT1	1.551	0.02730	17	18	0.9444	★★★	6.22
HUBMAP:Exc L2 Lamp5 Kcng3 (id193)	1.550	0.02290	88	97	0.9072	★★★	3.80
HUBMAP:Exc L6 Themis Sln (id235)	1.550	0.02290	88	97	0.9072	★★★	3.80

Showing 1 to 8 of 2,347 entries

Genes in gene set

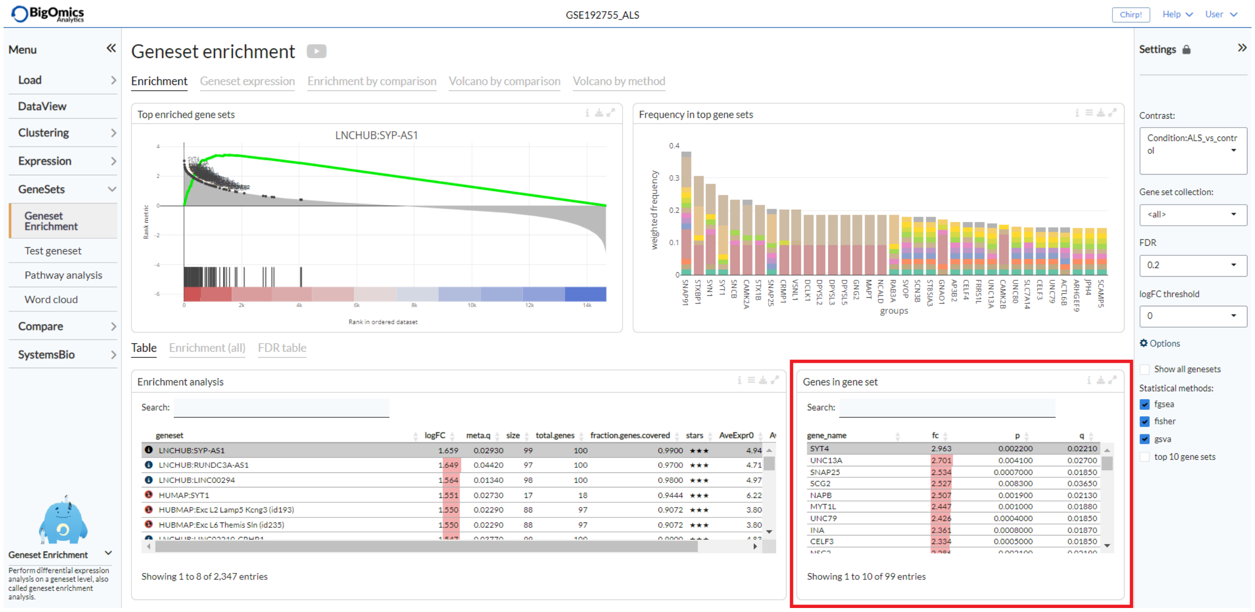
Search:

gene_name	fc	p	q
SYT4	2.963	0.002200	0.02210
UNC13A	2.701	0.004100	0.02700
SNAP25	2.534	0.0007000	0.01850
SCG2	2.527	0.008300	0.03650
NAPB	2.507	0.001900	0.02130
MYT1L	2.447	0.001000	0.01880
UNC79	2.426	0.0004000	0.01850
INA	2.361	0.0008000	0.01870
CELF3	2.334	0.0005000	0.01850

Showing 1 to 10 of 99 entries

17.1.5 Genes in gene set

By clicking on a gene set in the **Table** above, it is possible to see the gene list of that gene set under **Genes in gene set**. this table also reports the fold-change, statistics and correlation of the genes in the selected gene set. By clicking on a gene in this table, users can check the expression status of the gene for the selected contrast in the *Expression barplot* and its correlation to the gene set in the *Enrichment vs. expression* scatter plot under the **Geneset expression** tab, discussed in the next section.


Contrast: Condition:ALS_vs_control

Gene set collection: <all>

FDR: 0.2

logFC threshold: 0

Enrichment analysis

Showing 1 to 8 of 2,347 entries

Genes in gene set

Showing 1 to 10 of 99 entries

17.1.6 Geneset expression

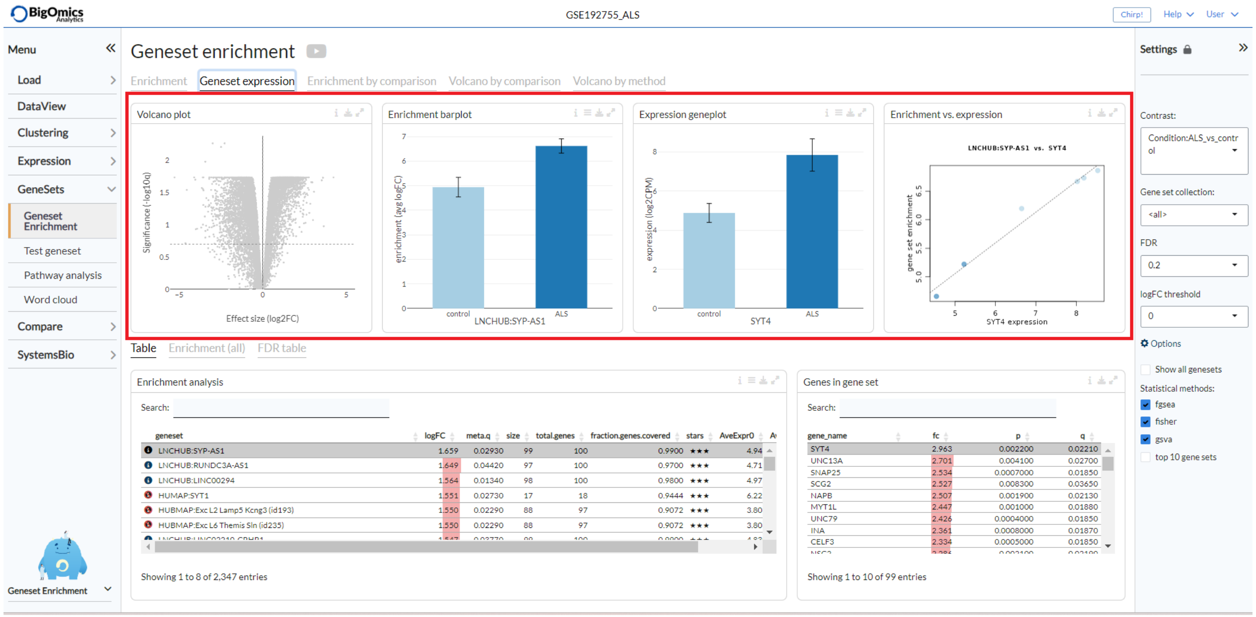
The **Geneset expression** panel provide plots associated with the gene set selected in **Table** and gene selected in **Genes in gene set**, as explained in the previous section.

Volcano Plot
: Volcano-plot of genes showing the significance versus the fold-change on the y and x axes, respectively. Genes in the selected gene set are highlighted in blue.

Enrichment barplot
: Enrichment plot associated with the gene set selected from **Table**. Samples in the barplot can be ungrouped via the plot settings icon.

Expression barplot
: Barplot of the gene expression of the gene. Samples in the barplot can be ungrouped via the plot settings icon.

Enrichment vs. expression
: Scatter plot of the enrichment versus the expression of the selected geneset and gene, on the y and x axes, respectively.



Settings

Contrast: Condition:ALS_vs_control

Gene set collection: <all>

FDR: 0.2

logFC threshold: 0

Options

□ Show all genesets

Statistical methods:

☑ fgsea
☑ fisher
☑ gsva
□ top 10 gene sets

Enrichment analysis

Search:

geneset	logFC	meta.q	size	total.genes	fraction.genes.covered	stars	AveExpr0
LNCHUB:SYP-AS1	1.659	0.02930	99	100	0.9900	★★★	4.94
LNCHUB:RUNDC3A-AS1	1.649	0.04420	97	100	0.9700	★★★	4.71
LNCHUB:LINC00294	1.564	0.01340	98	100	0.9800	★★★	4.97
HUMAP:SYT1	1.551	0.02730	17	18	0.9444	★★★	6.22
HUBMAP:Exc L2 Lamp5 Kcng3 (id193)	1.550	0.02290	88	97	0.9072	★★★	3.80
HUBMAP:Exc L6 Themis Sln (id235)	1.550	0.02290	88	97	0.9072	★★★	3.80

Showing 1 to 8 of 2,347 entries

Genes in gene set

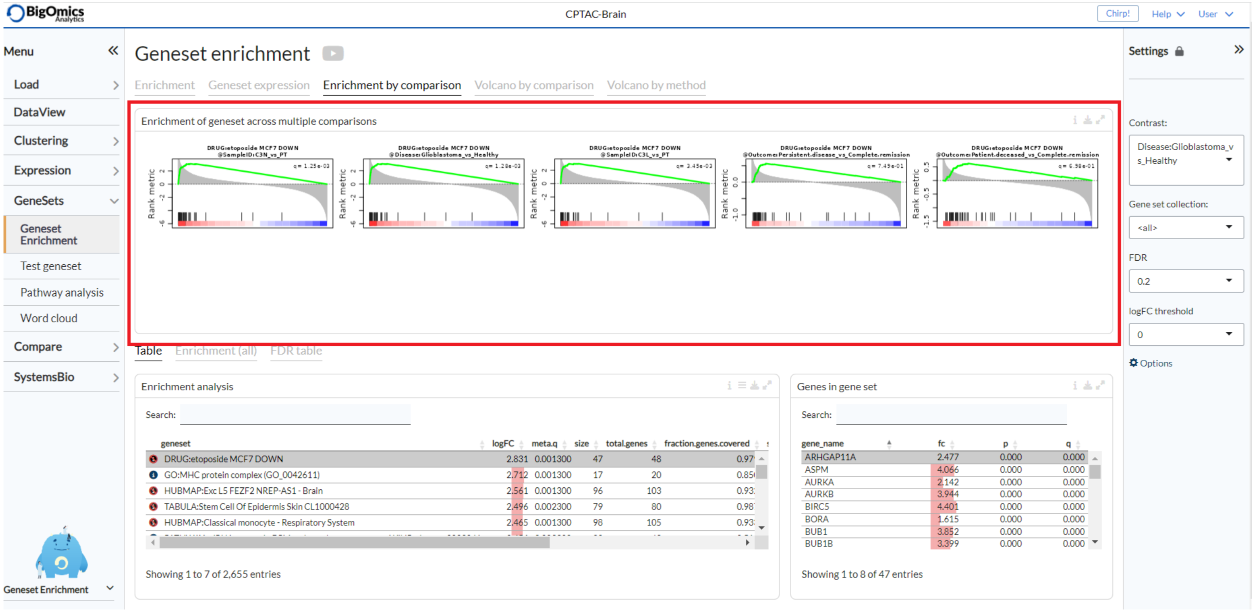
Search:

gene_name	fc	p	q
SYT4	2.963	0.002200	0.02210
UNC13A	2.701	0.004100	0.02700
SNAP25	2.534	0.0007000	0.01850
SCG2	2.527	0.008300	0.03650
NAPB	2.507	0.001900	0.02130
MYT1L	2.447	0.001000	0.01880
UNC79	2.426	0.0004000	0.01850
INA	2.361	0.0008000	0.01870
CELF3	2.334	0.0005000	0.01850

Showing 1 to 10 of 99 entries

17.1.7 Enrichment by comparison

Under the **Enrichment by comparison** panel, users can fin enrichment plots for the selected gene set (in **Table**) across multiple contrasts. The figure allows to quickly compare the enrichment of a certain gene set across all other comparisons.



Contrast: Disease:Glioblastoma_vs_Healthy

Gene set collection: <all>

FDR 0.2

logFC threshold 0

Enrichment analysis

Search:

geneset	logFC	meta.q	size	total.genes	fraction.genes.covered
DRUG:etoposide MCF7 DOWN	2.831	0.001300	47	48	0.97
GO:MHC protein complex (GO_0042611)	2.712	0.001300	17	20	0.85
HUBMAP:Exc L5 FEZF2 NREP-AS1 - Brain	2.561	0.001300	96	103	0.93
TABULA:Stem Cell Of Epidermis Skin CL1000428	2.496	0.002300	79	80	0.98
HUBMAP:Classical monocyte - Respiratory System	2.465	0.001300	98	105	0.93

Showing 1 to 7 of 2,655 entries

Genes in gene set

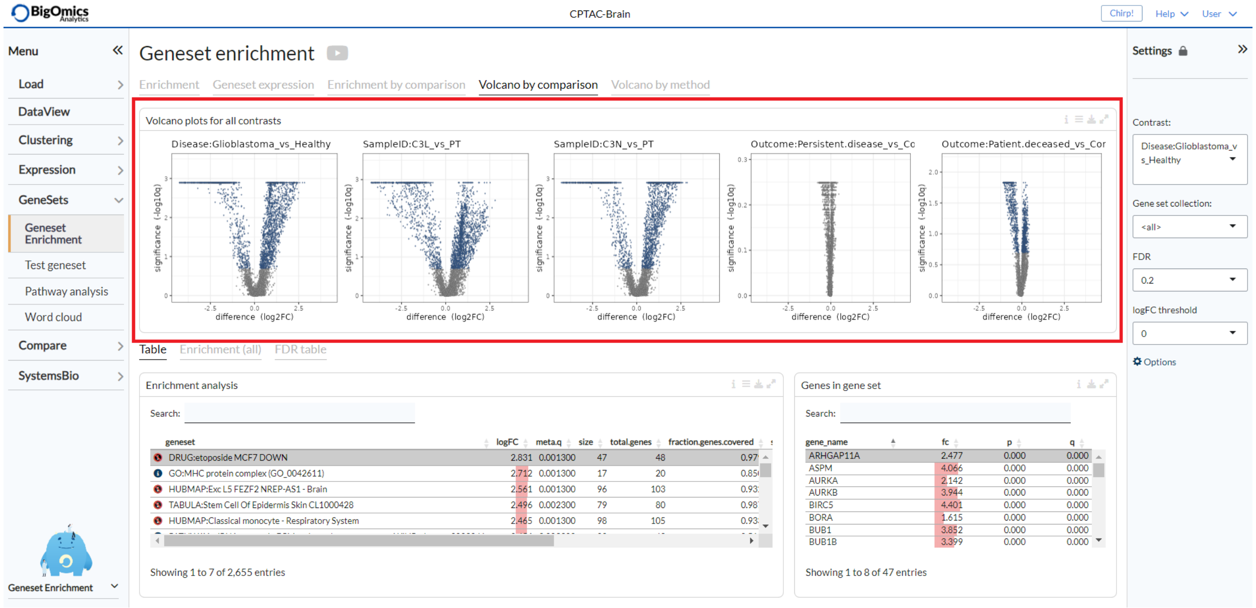
Search:

gene_name	fc	p	q
ARHGAP11A	2.477	0.000	0.000
ASPM	4.066	0.000	0.000
AURKA	2.142	0.000	0.000
AURKB	3.944	0.000	0.000
BIRC5	4.401	0.000	0.000
BORA	1.615	0.000	0.000
BUB1	3.852	0.000	0.000
BUB1B	3.399	0.000	0.000

Showing 1 to 8 of 47 entries

17.1.8 Volcano by comparison

The **Volcano by comparison** panel simultaneously displays volcano plots of gene sets enrichment across all contrasts, showing the enrichment score versus significance on the x and y axes, respectively. This provides users an overview of the statistics across all comparisons. By comparing multiple volcano plots, the user can immediately see which comparison is statistically weak or strong. Experimental contrasts showing better statistical significance will show volcano plots with 'higher' wings.



Contrast: Disease:Glioblastoma_vs_Healthy

Gene set collection: <all>

FDR 0.2

logFC threshold 0

Enrichment analysis

Search:

geneset	logFC	meta.q	size	total.genes	fraction.genes.covered
DRUG:etoposide MCF7 DOWN	2.831	0.001300	47	48	0.97
GO:MHC protein complex (GO_0042611)	2.712	0.001300	17	20	0.85
HUBMAP:Exc L5 FEZF2 NREP-AS1 - Brain	2.561	0.001300	96	103	0.93
TABULA:Stem Cell Of Epidermis Skin CL1000428	2.496	0.002300	79	80	0.98
HUBMAP:Classical monocyte - Respiratory System	2.465	0.001300	98	105	0.93

Showing 1 to 7 of 2,655 entries

Genes in gene set

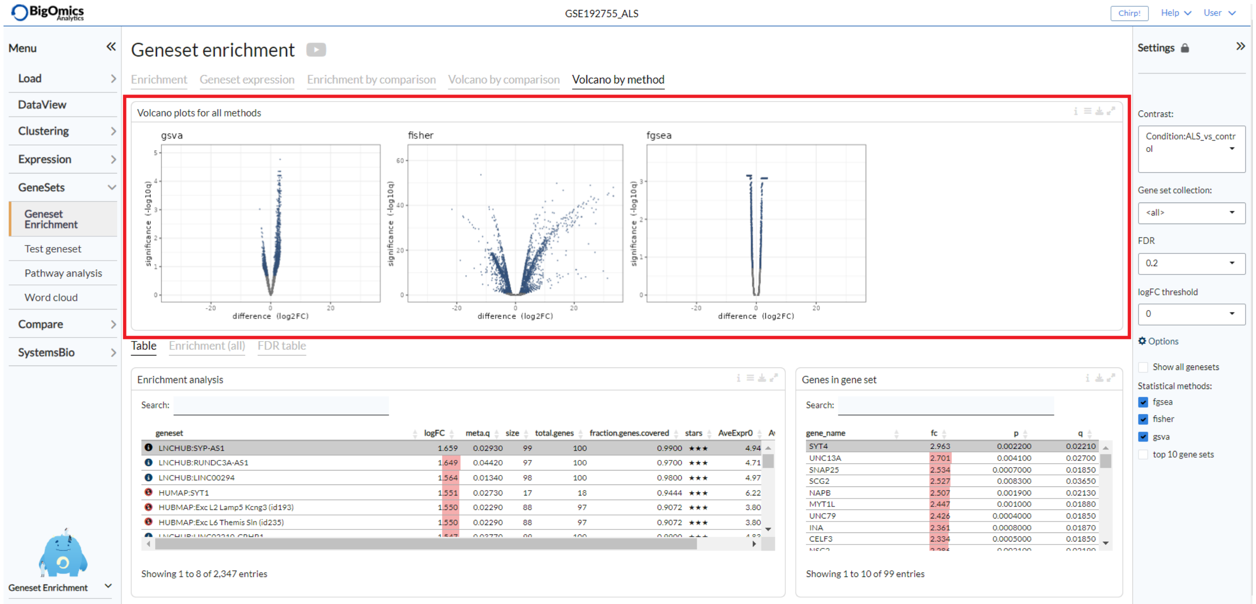
Search:

gene_name	fc	p	q
ARHGAP11A	2.477	0.000	0.000
ASPM	4.066	0.000	0.000
AURKA	2.142	0.000	0.000
AURKB	3.944	0.000	0.000
BIRC5	4.401	0.000	0.000
BORA	1.615	0.000	0.000
BUB1	3.852	0.000	0.000
BUB1B	3.399	0.000	0.000

Showing 1 to 8 of 47 entries

17.1.9 Volcano by method

Under the **Volcano by method** panel, users can see the simultaneous visualisation of volcano plots of gene sets for different enrichment methods. This provides users an quick overview of the sensitivity of the statistical methods at once. Methods showing better statistical significance will show volcano plots with 'higher' wings.



Contrast: Condition:ALS_vs_control

Gene set collection: <all>

FDR: 0.2

logFC threshold: 0

□ Show all genesets

Statistical methods:

☑ fgsea

☑ fisher

☑ gsva

□ top 10 gene sets

Enrichment analysis

Search:

	geneset	logFC	meta.q	size	total.genes	fraction.genes.covered	stars	AveExpr0
●	LNCHUB:SYP-AS1	1.659	0.02930	99	100	0.9900	★★★	4.94
●	LNCHUB:RUNDC3A-AS1	1.649	0.04420	97	100	0.9700	★★★	4.71
●	LNCHUB:LINC00294	1.564	0.01340	98	100	0.9800	★★★	4.97
●	HUMAP:SYT1	1.551	0.02730	17	18	0.9444	★★★	6.22
●	HUBMAP:Exc L2 Lamp5 Kcng3 (id193)	1.550	0.02290	88	97	0.9072	★★★	3.80
●	HUBMAP:Exc L6 Themis Sln (id235)	1.550	0.02290	88	97	0.9072	★★★	3.80

Showing 1 to 8 of 2,347 entries

Genes in gene set

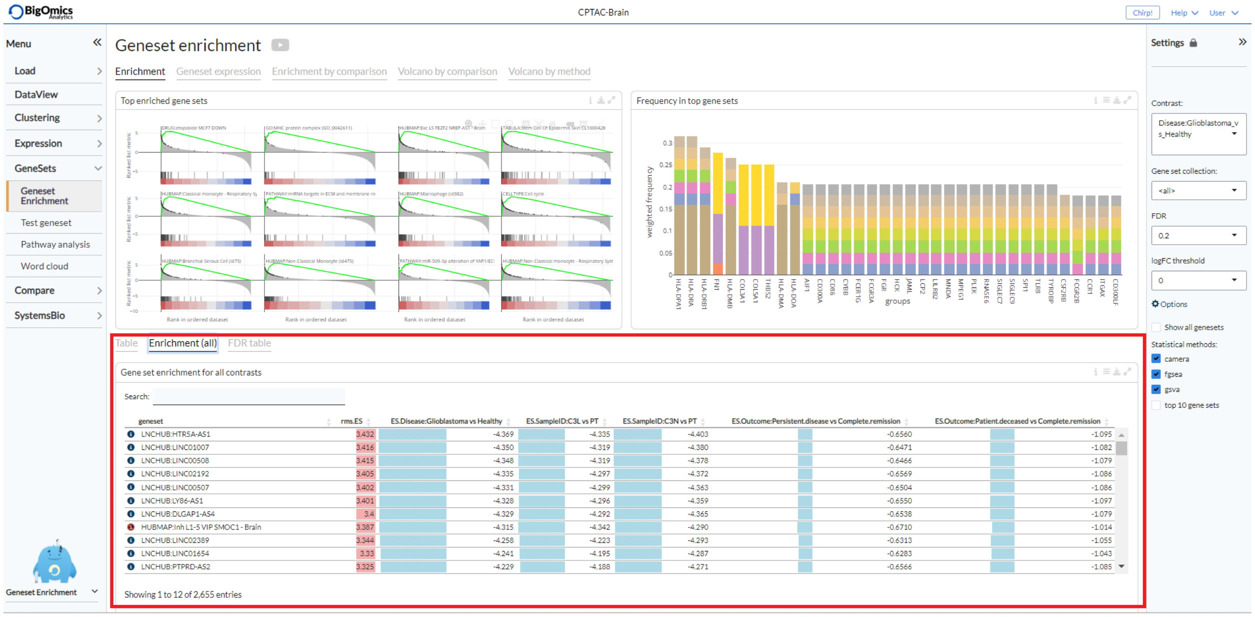
Search:

gene_name	fc	p	q
SYT4	2.963	0.002200	0.02210
UNC13A	2.701	0.004100	0.02700
SNAP25	2.534	0.0007000	0.01850
SCG2	2.527	0.008300	0.03650
NAPB	2.507	0.001900	0.02130
MYT1L	2.447	0.001000	0.01880
UNC79	2.426	0.0004000	0.01850
INA	2.361	0.0008000	0.01870
CELF3	2.334	0.0005000	0.01850

Showing 1 to 10 of 99 entries

17.1.10 Enrichment (all)

The **Enrichment (all)** table provides the enrichment scores of gene sets across all contrasts. It is also possible to visualise all q-values from the setting icon on top of the plot.



Contrast: Disease:Glioblastoma_vs_Healthy

Gene set collection: <all>

FDR: 0.2

logFC threshold: 0

□ Show all genesets

Statistical methods:

☑ camera

☑ fgsea

☑ gsva

□ top 10 gene sets

Gene set enrichment for all contrasts

Search:

	geneset	rms.ES	ES.Disease:Glioblastoma vs Healthy	ES.SampleID:C3L vs PT	ES.SampleID:C3N vs PT	ES.Outcome:Persistent.disease vs Complete.remission	ES.Outcome:Patient.deceased vs Complete.remission
●	LNCHUB:HTR5A-AS1	3.432	-4.369	-4.335	-4.403	-0.6560	-1.095
●	LNCHUB:LINC01007	3.416	-4.350	-4.319	-4.380	-0.6471	-1.082
●	LNCHUB:LINC00508	3.415	-4.348	-4.319	-4.378	-0.6466	-1.079
●	LNCHUB:LINC02192	3.405	-4.335	-4.297	-4.372	-0.6569	-1.086
●	LNCHUB:LINC00507	3.402	-4.331	-4.299	-4.363	-0.6504	-1.086
●	LNCHUB:LY86-AS1	3.401	-4.328	-4.296	-4.359	-0.6550	-1.097
●	LNCHUB:DLGAP1-AS4	3.4	-4.329	-4.292	-4.365	-0.6538	-1.079
●	HUBMAP:Inh L1-5 VIP SMOC1 - Brain	3.387	-4.315	-4.342	-4.290	-0.6710	-1.014
●	LNCHUB:LINC02389	3.344	-4.258	-4.223	-4.293	-0.6313	-1.055
●	LNCHUB:LINC01654	3.33	-4.241	-4.195	-4.287	-0.6283	-1.043
●	LNCHUB:PTPRD-AS2	3.325	-4.229	-4.188	-4.271	-0.6566	-1.085

Showing 1 to 12 of 2,655 entries

17.1.11 FDR table

The **FDR table** panel reports the number of significant gene sets at different FDR thresholds for all contrasts and methods. Using the table the user can determine which statistical methods perform better for a particular contrast. The left part of the table (in blue) correspond to the number of significant down-regulated gene sets, the right part (in red) correspond to the number of significant overexpressed gene sets.

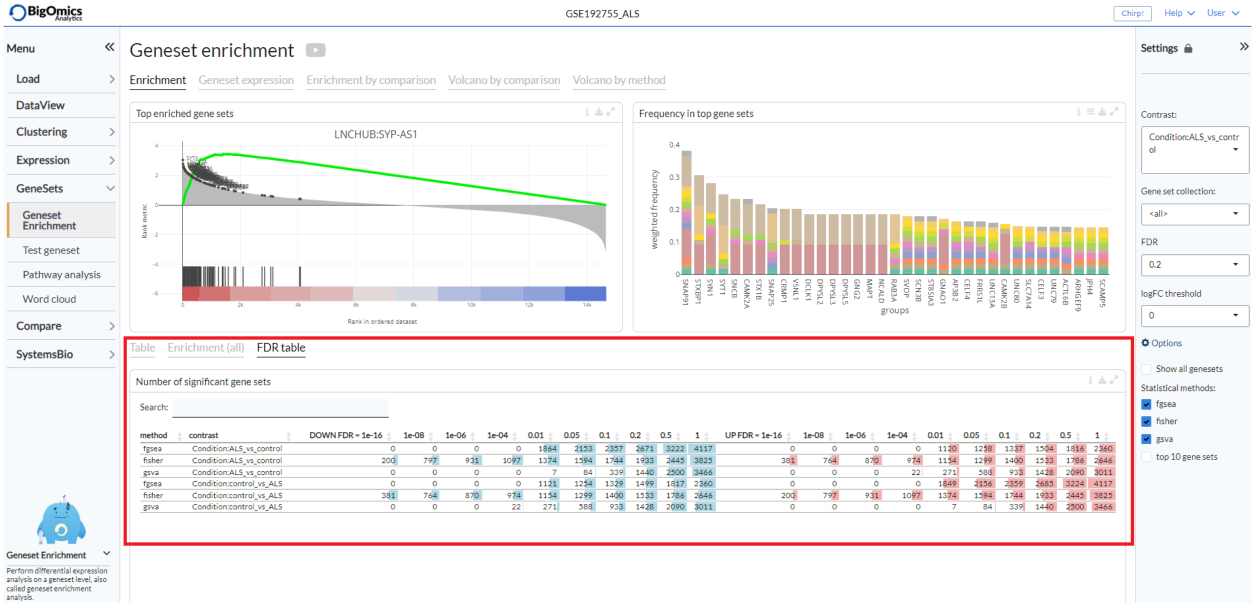


Table | Enrichment (all) | FDR table

Number of significant gene sets

Search:

method	contrast	DOWN FDR = 1e-16	1e-08	1e-06	1e-04	0.01	0.05	0.1	0.2	0.5	1	UP FDR = 1e-16	1e-08	1e-06	1e-04	0.01	0.05	0.1	0.2	0.5	1
fgsea	Condition:ALS_vs_control	0	0	0	0	1864	2153	2357	2671	3222	4117	0	0	0	0	1120	1258	1337	1504	1816	2360
fisher	Condition:ALS_vs_control	200	797	931	1097	1374	1594	1744	1933	2445	3825	381	764	870	974	1154	1299	1400	1533	1786	2646
gsva	Condition:ALS_vs_control	0	0	0	0	7	84	339	1440	2500	3466	0	0	0	22	271	588	933	1428	2090	3011
fgsea	Condition:control_vs_ALS	0	0	0	0	1121	1254	1329	1499	1817	2360	0	0	0	0	1849	2156	2359	2685	3224	4117
fisher	Condition:control_vs_ALS	381	764	870	974	1154	1299	1400	1533	1786	2646	200	797	931	1097	1374	1594	1744	1933	2445	3825
gsva	Condition:control_vs_ALS	0	0	0	22	271	588	933	1428	2090	3011	0	0	0	0	7	84	339	1440	2500	3466

Contrast: Condition:ALS_vs_control

Gene set collection: <all>

FDR: 0.2

logFC threshold: 0

Options

□ Show all genesets

Statistical methods:

☑ fgsea

☑ fisher

☑ gsva

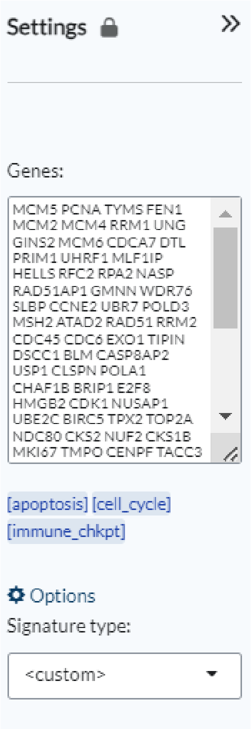
□ top 10 gene sets

17.2 Test geneset

The **Test geneset** submodule is used to test a specific user-defined geneset, or, alternatively, a geneset from the KEGG or Hallmark collection or a list of genes extracted from the available pairwise comparisons, for enrichment in the experimental dataset. The plot contain five tabs: **Enrichment table**, **Volcano plots**, **Enrichment**, **Overlap/similarity** and **Markers**.

17.2.1 Settings panel

The settings panel of the **Test geneset** module consists of a box where users can paste gene lists to be queried. Three examples of gene lists are provided below it, namely the *apoptosis*, *cell_cycle* and *immune_chkpt* lists. Users can also select specific KEGG and hallmark gene sets from the *Options* in the panel, as well as using the differentially expressed genes or proteins from one of the contrasts.



Settings

Genes:

MCM5 PCNA TYMS FEN1
MCM2 MCM4 RRM1 UNG
GINS2 MCM6 CDCA7 DTL
PRIM1 UHRF1 MLF1IP
HELLS RFC2 RPA2 NASP
RAD51AP1 GMNN WDR76
SLBP CCNE2 UBR7 POLD3
MSH2 ATAD2 RAD51 RRM2
CDC45 CDC6 EXO1 TIPIN
DSCC1 BLM CASP8AP2
USP1 CLSPN POLA1
CHAF1B BRIP1 E2F8
HMGB2 CDK1 NUSAP1
UBE2C BIRC5 TPX2 TOP2A
NDC80 CKS2 NUF2 CKS1B
MKI67 TMPO CENPF TACC3

[apoptosis] [cell_cycle]
[immune_chkpt]

Options

Signature type:

<custom>

17.2.2 Enrichment table

This tab is shown indepently of the other four. It display two tables. The first table, *Enrichment by contrasts*, shows the enrichment scores of query signature across all contrasts. The table summarizes the enrichment statistics of the gene list in all contrasts using the GSEA algorithm. The NES corresponds to the normalized enrichment score of the GSEA analysis.

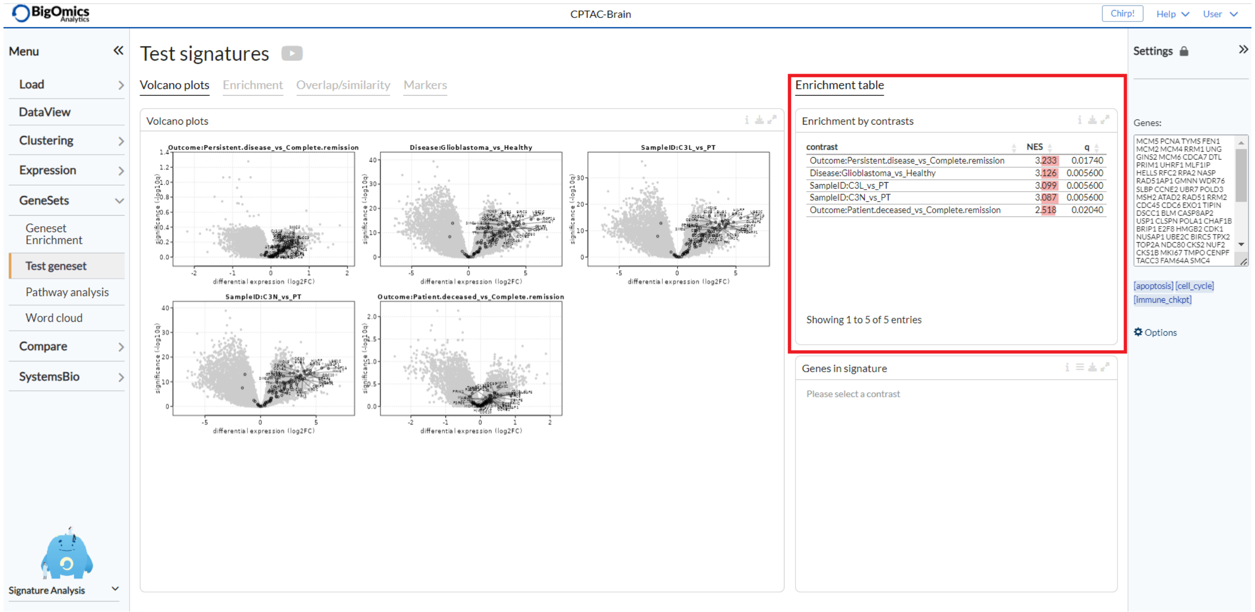
The second table, *Genes in signature*, is based on which pairiwse comparison is selected in *Enrichment by contrasts* and shows the genes of the current signature corresponding to the selected contrast. Genes are sorted by decreasing (absolute) fold-change.

17.2.3 Volcano plots

This tab contains volcano plots showing where the genes of a test signatures fall in the experimental pairiwse comparisons. For positive enrichment, genes of the query signature would fall on the upper right of the volcano plot, for negative enrichment, on the upper left.

17.2.4 Enrichment

In this tab, multiple plots show the enrichment of the query signature in all constrasts. Positive enrichment means that a particular contrast shows similar expression changes as the query signature. Users can either view the plots of all teh contrasts, while selecting a specific contrast in the *Enrichment by contrasts* table will only display the ernichment plot for it.

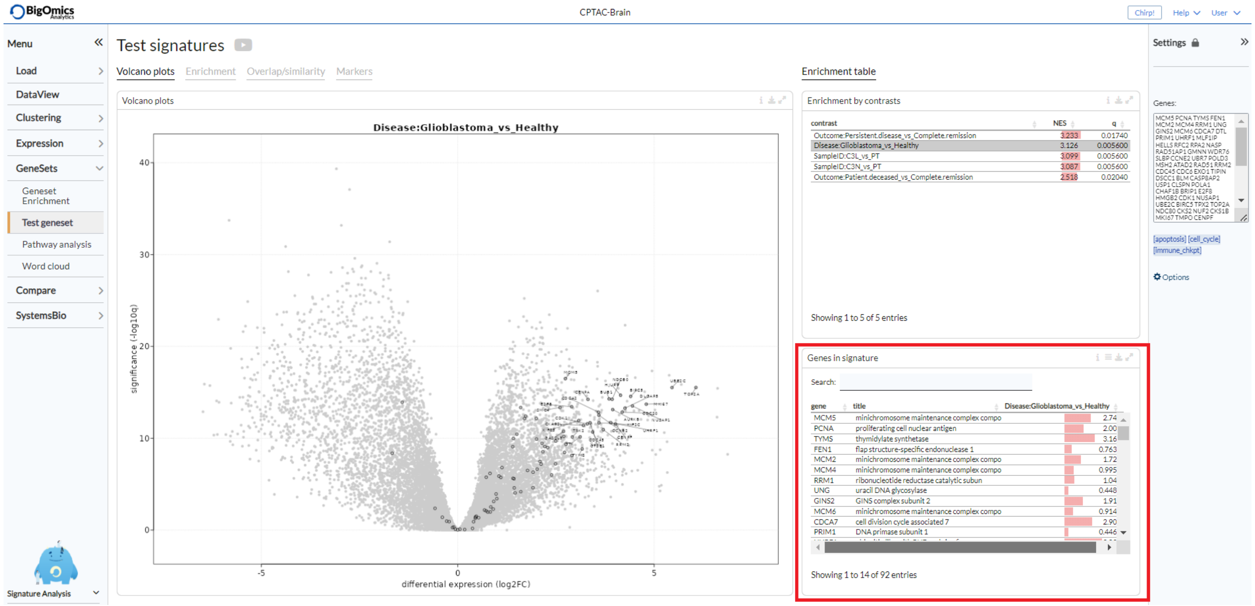


contrast	NES	q
Outcome:Persistent.disease_vs_Complete.remission	3.233	0.01740
Disease:Glioblastoma_vs_Healthy	3.126	0.005600
SampleID:C3L_vs_PT	3.099	0.005600
SampleID:C3N_vs_PT	3.087	0.005600
Outcome:Patient.deceased_vs_Complete.remission	2.518	0.02040

Showing 1 to 5 of 5 entries

Please select a contrast

Genes:
MCM5 PCNA TYMS FEN1
MCM2 MCM4 RRM1 UNG
GINS2 MCM6 CDCA7 DTL
PRIM1 UHRF1 MLF1IP
HELLS RFC2 RPA2 NASP
RAD51AP1 GMNN WDR76
SLBP CCNE2 UBR7 POLD3
MSH2 ATAD2 RAD51 RRM2
CDC45 CDC6 EXO1 TIPIN
DSCC1 BLM CASP8AP2
USP1 CLSPN POLA1 CHAF1B
BRIP1 E2F8 HMGB2 CDK1
NUSAP1 UBE2C BIRC5 TPX2
TOP2A NDC80 CKS2 NUF2
CKS1B MKI67 TMPO CENPF
TACC3 FAM64A SMC4



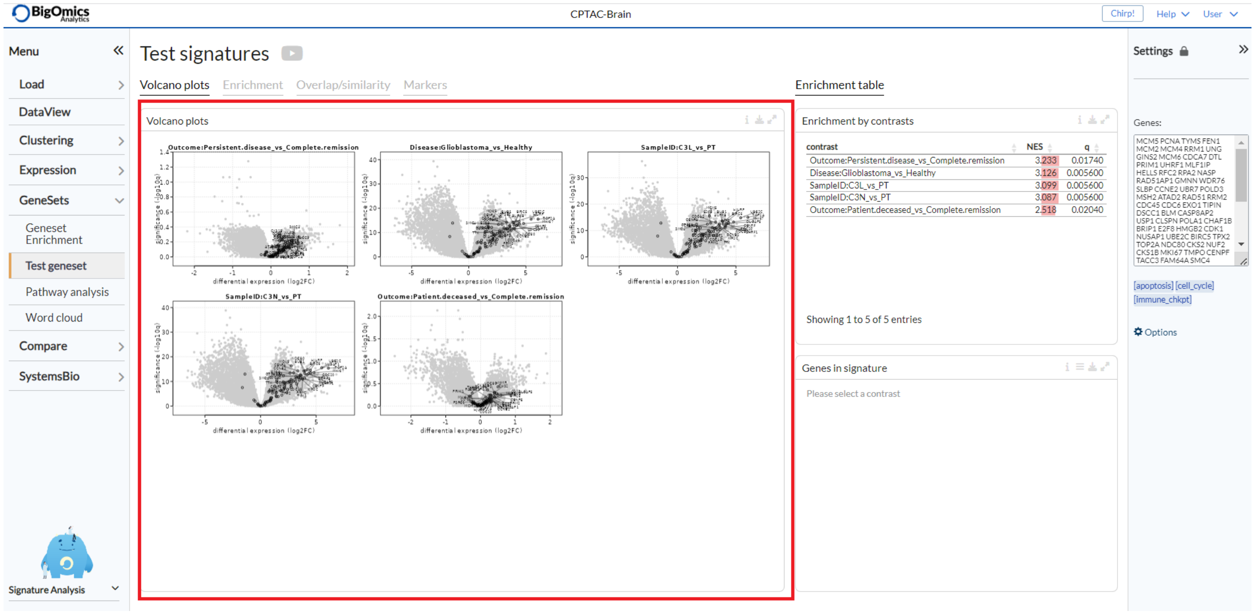
contrast	NES	q
Outcome:Persistent.disease_vs_Complete.remission	3.233	0.01740
Disease:Glioblastoma_vs_Healthy	3.126	0.005600
SampleID:C3L_vs_PT	3.099	0.005600
SampleID:C3N_vs_PT	3.087	0.005600
Outcome:Patient.deceased_vs_Complete.remission	2.518	0.02040

Showing 1 to 5 of 5 entries

Search:

gene	title	Disease:Glioblastoma_vs_Healthy
MCM5	minichromosome maintenance complex compo	2.74
PCNA	proliferating cell nuclear antigen	2.00
TYMS	thymidylate synthetase	3.16
FEN1	flap structure-specific endonuclease 1	0.763
MCM2	minichromosome maintenance complex compo	1.72
MCM4	minichromosome maintenance complex compo	0.995
RRM1	ribonucleotide reductase catalytic subun	1.04
UNG	uracil DNA glycosylase	0.448
GINS2	GINS complex subunit 2	1.91
MCM6	minichromosome maintenance complex compo	0.914
CDCA7	cell division cycle associated 7	2.90
PRIM1	DNA primase subunit 1	0.446

Showing 1 to 14 of 92 entries



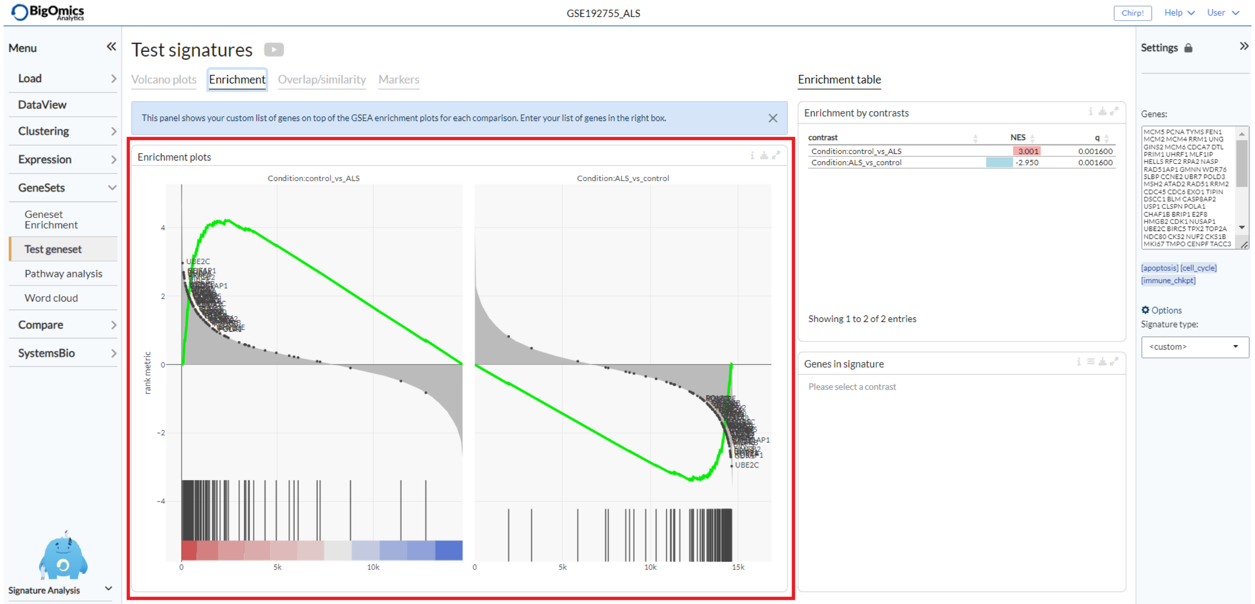
contrast	NES	q
Outcome:Persistent.disease_vs_Complete.remission	3.233	0.01740
Disease:Glioblastoma_vs_Healthy	3.126	0.005600
SampleID:C3L_vs_PT	3.099	0.005600
SampleID:C3N_vs_PT	3.087	0.005600
Outcome:Patient.deceased_vs_Complete.remission	2.518	0.02040

Showing 1 to 5 of 5 entries

Please select a contrast

Genes:

MCM5 PCNA TYMS FEN1 MCM2 MCM4 RRM1 UNG GINS2 MCM6 CDCA7 DTL PRIM1 UHRF1 MLF1IP HELLS RFC2 RPA2 NASP RAD51AP1 GMNN WDR76 SLBP CCNE2 UBR7 POLD3 MSH2 ATAD2 RAD51 RRM2 CDC45 CDC6 EXO1 TIPIN DSCC1 BLM CASP8AP2 USP1 CLSPN POLA1 CHAF1B BRIP1 E2F8 HMGB2 CDK1 NUSAP1 UBE2C BIRC5 TPX2 TOP2A NDC80 CKS2 NUF2 CKS1B MKI67 TMPO CENPF TACC3 FAM64A SMC4



This panel shows your custom list of genes on top of the GSEA enrichment plots for each comparison. Enter your list of genes in the right box.

contrast	NES	q
Condition:control_vs_ALS	3.001	0.001600
Condition:ALS_vs_control	-2.950	0.001600

Showing 1 to 2 of 2 entries

Please select a contrast

Genes:

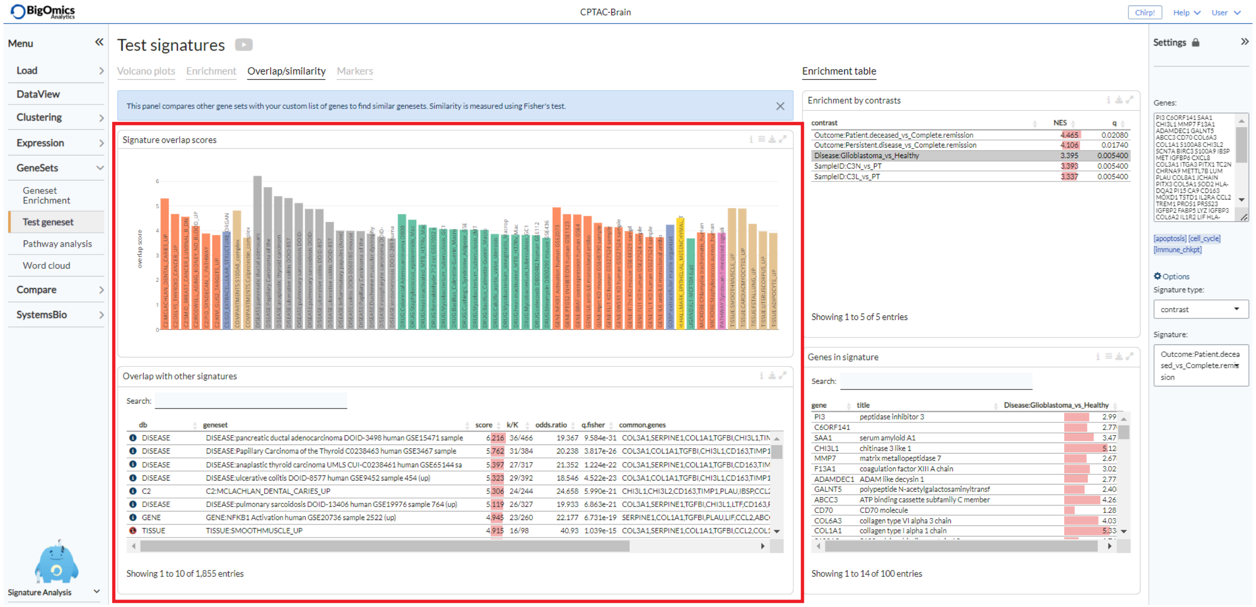
MCM5 PCNA TYMS FEN1 MCM2 MCM4 RRM1 UNG GINS2 MCM6 CDCA7 DTL PRIM1 UHRF1 MLF1IP HELLS RFC2 RPA2 NASP RAD51AP1 GMNN WDR76 SLBP CCNE2 UBR7 POLD3 MSH2 ATAD2 RAD51 RRM2 CDC45 CDC6 EXO1 TIPIN DSCC1 BLM CASP8AP2 USP1 CLSPN POLA1 CHAF1B BRIP1 E2F8 HMGB2 CDK1 NUSAP1 UBE2C BIRC5 TPX2 TOP2A NDC80 CKS2 NUF2 CKS1B MKI67 TMPO CENPF TACC3

Signature type: <custom>

17.2.5 Overlap/similarity

Under the Overlap/similarity tab, users can find the similarity of their gene list with all the gene sets and pathways in the platform, including statistics such as the total number of genes in the gene set (K), the number of intersecting genes between the list and the gene set (k), the overlapping ratio of k/K, logarithm of the odds ratio (log.OR), as well as the p and q values by the Fisher's test for the overlap test.

The *Signature overlap scores* plot is located on top. Its vertical axis shows the overlap score of the gene set which combines the odds ratio and significance (q-value) of the Fisher's test. The*Overlap with other signatures* table is located below it and contain various statistical values.



This panel compares other gene sets with your custom list of genes to find similar genesets. Similarity is measured using Fisher's test.

Overlap with other signatures

Search:

db	geneset	score	k/K	odds.ratio	q.fisher	common.genes
DISEASE	DISEASE:pancreatic ductal adenocarcinoma DOID-3498 human GSE15471 sample	6.216	36/466	19.367	9.584e-31	COL3A1,SERPINE1,COL1A1,TGFBI,CHI3L1,TIN
DISEASE	DISEASE:Papillary Carcinoma of the Thyroid C0238463 human GSE3467 sample	5.762	31/384	20.238	3.817e-26	COL3A1,COL1A1,TGFBI,CHI3L1,CD163,TIMP1
DISEASE	DISEASE:anaplastic thyroid carcinoma UMLS CUI-C0238461 human GSE65144 sa	5.397	27/317	21.352	1.224e-22	COL3A1,SERPINE1,COL1A1,TGFBI,CD163,TIM
DISEASE	DISEASE:ulcerative colitis DOID-8577 human GSE9452 sample 454 (up)	5.323	29/392	18.546	4.522e-23	COL3A1,COL1A1,TGFBI,CHI3L1,CD163,TIMP1
C2	C2:MCLACHLAN_DENTAL_CARIES_UP	5.306	24/244	24.658	5.990e-21	CHI3L1,CHI3L2,CD163,TIMP1,PLAU,IBSP,CCL2
DISEASE	DISEASE:pulmonary sarcoidosis DOID-13406 human GSE19976 sample 764 (up)	5.119	26/327	19.933	6.863e-21	COL3A1,SERPINE1,TGFBI,CHI3L1,LTF,CD163,F
GENE	GENE:NFKB1 Activation human GSE20736 sample 2522 (up)	4.945	23/260	22.177	6.731e-19	SERPINE1,COL1A1,TGFBI,PLAU,LIF,CCL2,ABC
TISSUE	TISSUE:SMOOTHMUSCLE_UP	4.915	16/98	40.93	1.039e-15	COL3A1,SERPINE1,COL1A1,TGFBI,CCL2,COL1

Showing 1 to 10 of 1,855 entries

Enrichment by contrasts

contrast	NES	q
Outcome:Patient.deceased_vs_Complete.remission	4.465	0.02080
Outcome:Persistent.disease_vs_Complete.remission	4.106	0.01740
Disease:Glioblastoma_vs_Healthy	3.395	0.005400
SampleID:C3N_vs_PT	3.393	0.005400
SampleID:C3L_vs_PT	3.337	0.005400

Showing 1 to 5 of 5 entries

Genes in signature

Search:

gene	title	Disease:Glioblastoma_vs_Healthy
PI3	peptidase inhibitor 3	2.99
C6ORF141		2.77
SAA1	serum amyloid A1	3.47
CHI3L1	chitinase 3 like 1	5.12
MMP7	matrix metallopeptidase 7	2.67
F13A1	coagulation factor XIII A chain	3.02
ADAMDEC1	ADAM like decysin 1	2.77
GALNT5	polypeptide N-acetylgalactosaminyltransf	2.40
ABCC3	ATP binding cassette subfamily C member	4.26
CD70	CD70 molecule	1.28
COL6A3	collagen type VI alpha 3 chain	4.03
COL1A1	collagen type I alpha 1 chain	5.33

Showing 1 to 14 of 100 entries

Genes:

PI3 C6ORF141 SAA1 CHI3L1 MMP7 F13A1 ADAMDEC1 GALNT5 ABCC3 CD70 COL6A3 COL1A1 S100A8 CHI3L2 SCN7A BIRC3 S100A9 IBSP MET IGFBP6 CXCL8 COL3A1 ITGA3 PITX1 TC2N CHRNA9 METTL7B LUM PLAU COL8A1 JCHAIN PITX3 COL5A1 SOD2 HLA-DQA2 PI15 CA9 CD163 MOXD1 TSTD1 IL2RA CCL2 TREM1 PROS1 PRSS23 IGFBP2 FABP5 LYZ IGFBP3 COL6A2 IL1R2 LIF HLA-

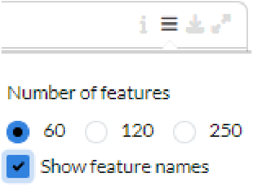
[apoptosis] [cell_cycle] [immune_chkpt]

Options

Signature type: contrast

Signature: Outcome:Patient.deceased_vs_Complete.remission

The *Signature overlap scores* plot can also be customised via the settings icon, where users can select the number of signatures to display (`Number of features`) and can toggle the signature names on and off (`Show features names`).



Number of features

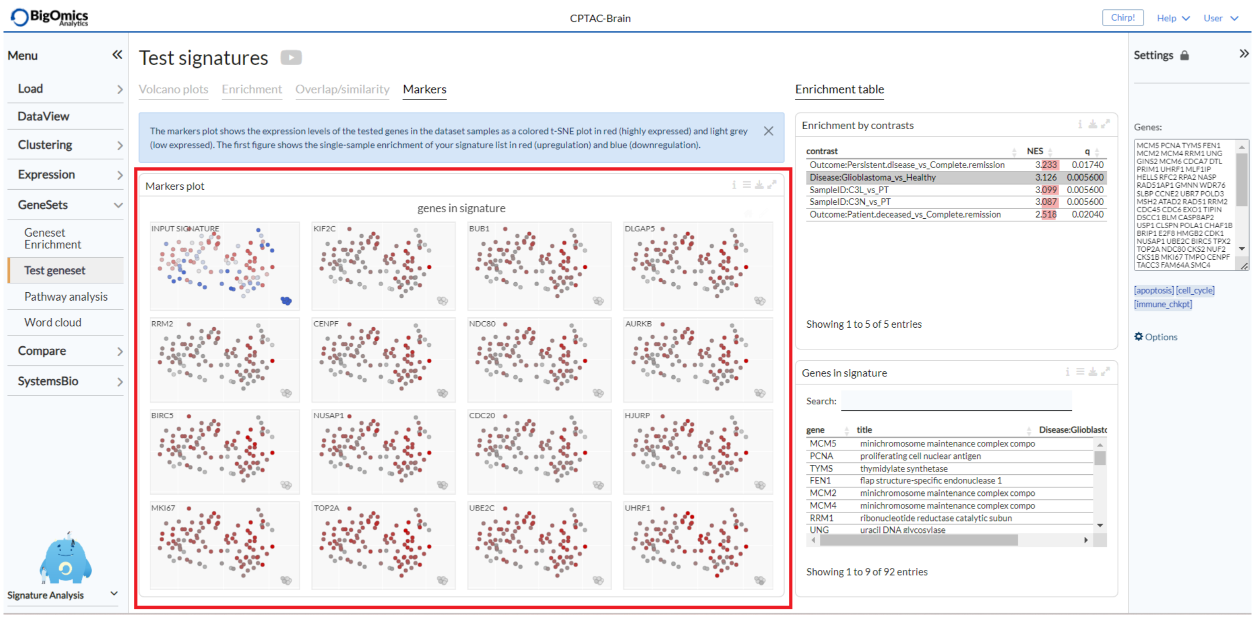
● 60 ○ 120 ○ 250

☑ Show feature names

17.2.6 Markers

The **Markers** tab contains a series of plots (under *Markers plot*) showing the expression levels of the tested genes in the dataset samples as a colored t-SNE plot in red (highly expressed) and light grey (low expressed). The first figure shows the single-sample enrichment of your signature list in red (upregulation) and blue (downregulation).

The plots can be sorted via the `Sort by` option in the settings icon by correlation (default), probability or name. `Layout` swaps the layout of the plots between a 4x4 (default) and a 6x6 grid.



BigOmics Analytics

CPTAC-Brain

Chirp! Help User

Menu

- Load
- DataView
- Clustering
- Expression
- GeneSets
  - Geneset Enrichment
  - Test geneset
  - Pathway analysis
  - Word cloud
- Compare
- SystemsBio

Signature Analysis

Test signatures

Volcano plots | Enrichment | Overlap/similarity | Markers

The markers plot shows the expression levels of the tested genes in the dataset samples as a colored t-SNE plot in red (highly expressed) and light grey (low expressed). The first figure shows the single-sample enrichment of your signature list in red (upregulation) and blue (downregulation).

Markers plot

Enrichment table

Enrichment by contrasts

contrast	NES	q
Outcome:Persistent.disease_vs_Complete.remission	3.233	0.01740
Disease:Glioblastoma_vs_Healthy	3.126	0.005600
SampleID:C3L_vs_PT	3.099	0.005600
SampleID:C3N_vs_PT	3.087	0.005600
Outcome:Patient.deceased_vs_Complete.remission	2.518	0.02040

Showing 1 to 5 of 5 entries

Genes in signature

Search:

gene	title	Disease:Glioblastc
MCM5	minichromosome maintenance complex compo	
PCNA	proliferating cell nuclear antigen	
TYMS	thymidylate synthetase	
FEN1	flap structure-specific endonuclease 1	
MCM2	minichromosome maintenance complex compo	
MCM4	minichromosome maintenance complex compo	
RRM1	ribonucleotide reductase catalytic subun	
UNG	uracil DNA glycosylase	

Showing 1 to 9 of 92 entries

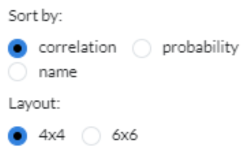
Settings

Genes:

MCM5 PCNA TYMS FEN1 MCM2 MCM4 RRM1 UNG GINS2 MCM6 CDCA7 DTL PRIM1 UHRF1 MLF1IP HELLS RFC2 RPA2 NASP RAD51AP1 GMNN WDR76 SLBP CCNE2 UBR7 POLD3 MSH2 ATAD2 RAD51 RRM2 CDC45 CDC6 EXO1 TIPIN DSCC1 BLM CASP8AP2 USP1 CLSPN POLA1 CHAF1B BRIP1 E2F8 HMGB2 CDK1 NUSAP1 UBE2C BIRC5 TPX2 TOP2A NDC80 CKS2 NUF2 CKS1B MKI67 TMPO CENPF TACC3 FAM64A SMC4

[apoptosis] [cell_cycle] [immune_chkpt]

Options



Sort by:

● correlation ○ probability
○ name

Layout:

● 4x4 ○ 6x6

17.3 Pathway analysis

This module performs specialized enrichment analysis providing higher level functional and visual interpretation.

The **WikiPathways** panel maps the differential fold-changes onto the WikiPathways pathway maps. The **Reactome** panel does the same for the Reactome pathway maps. Under the **GO** panel, a graph-based enrichment analysis is done using the Gene Ontology (GO) graph structure.

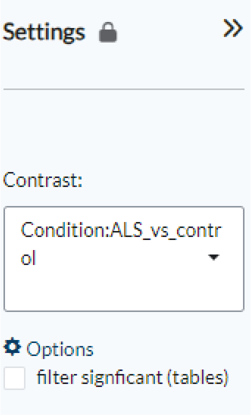
17.3.1 Settings panel

Users can specify the contrast of their interest in the `Contrast` settings. Under the main *Options*, users can select `filter significant (tables)` to keep only significant entries in the table.

17.3.2 WikiPathways

WikiPathways is a public collection of manually curated pathways representing the current knowledge of molecular interactions, reactions and relation networks as pathway maps. In the **WikiPathways** panel, each pathway is scored for the selected contrast profile and reported in the table. An activation-heatmap comparing the activation levels of pathways across multiple contrast profiles is generated. This facilitates the quick detect the similarities between profiles in certain pathways. More detailed explaination of each output is provided below.

WikiPathway
: In the pathway map, genes are colored according to their upregulation (red) or downregulation (blue) in the contrast profile. Each pathway is scored for the selected contrast profile and reported in the table below the plot.



Settings

Contrast:

Condition:ALS_vs_control

Options

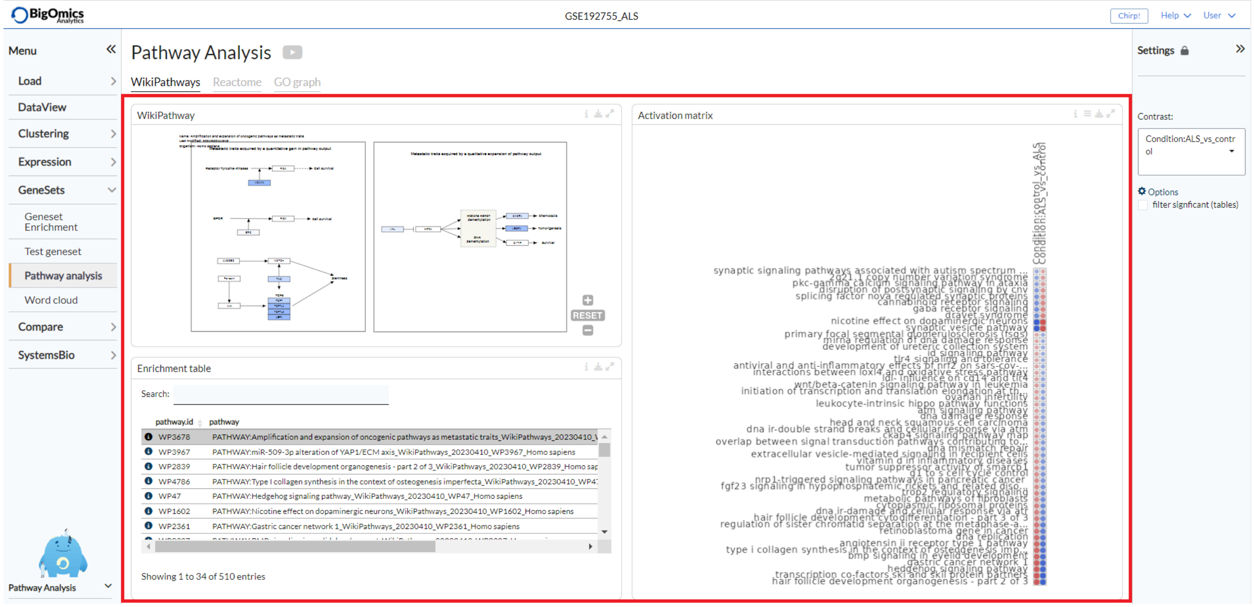
□ filter signficant (tables)

Enrichment table

The table is interactive; enabling user to sort on different variables (WikiPathways id, pathway, logFC and meta q-values) and select a pathway by clicking on the row in the table.

Activation matrix

The activation matrix visualizes the activation levels of pathways (or pathway keywords) across multiple contrast profiles. This facilitates to quickly see and detect the similarities of certain pathways between contrasts. The size of the circles correspond to their relative activation, and are colored according to their upregulation (red) or downregulation (blue) in the contrast profile. The matrix can be normalised from the *settings* icon.



BigOmics Analytics
GSE192755_ALS
Chirp!
Help
User
Menu
Load
DataView
Clustering
Expression
GeneSets
Geneset Enrichment
Test geneset
Pathway analysis
Word cloud
Compare
SystemsBio
Pathway Analysis
Pathway Analysis
WikiPathways
Reactome
GO graph
WikiPathway
RESET
Activation matrix
Condition:control_vs_ALS
Condition:ALS_vs_control
synaptic signaling pathways associated with autism spectrum ...
2q21.1 copy number variation syndrome
pkc-gamma calcium signaling pathway in ataxia
disruption of postsynaptic signaling by cnv
splicing factor nova regulated synaptic proteins
cannabinoid receptor signaling
gaba receptor signaling
dravet syndrome
nicotine effect on dopaminergic neurons
synaptic vesicle pathway
primary focal segmental glomerulosclerosis (fsgs)
mirna regulation of dna damage response
development of ureteric collection system
id signaling pathway
tlr4 signaling and tolerance
antiviral and anti-inflammatory effects of nrf2 on sars-cov-...
interactions between loxl4 and oxidative stress pathway
ldl- influence on cd14 and tlr4
wnt/beta-catenin signaling pathway in leukemia
initiation of transcription and translation elongation at th...
ovarian infertility
leukocyte-intrinsic hippo pathway functions
atm signaling pathway
dna damage response
head and neck squamous cell carcinoma
dna ir-double strand breaks and cellular response via atm
ckap4 signaling pathway map
overlap between signal transduction pathways contributing to...
dna mismatch repair
extracellular vesicle-mediated signaling in recipient cells
vitamin d in inflammatory diseases
tumor suppressor activity of smarcb1
g1 to s cell cycle control
nrp1-triggered signaling pathways in pancreatic cancer
fgf23 signaling in hypophosphatemic rickets and related diso...
trop2 regulatory signaling
metabolic pathways of fibroblasts
cytoplasmic ribosomal proteins
dna ir-damage and cellular response via atr
hair follicle development cytodifferentiation - part 3 of 3
regulation of sister chromatid separation at the metaphase-a...
retinoblastoma gene in cancer
dna replication
angiotensin ii receptor type 1 pathway
type i collagen synthesis in the context of osteogenesis imp...
bmp signaling in eyelid development
gastric cancer network 1
hedgehog signaling pathway
transcription co-factors ski and skil protein partners
hair follicle development organogenesis - part 2 of 3
Settings
Options

Enrichment table

Search:

	pathway.id	pathway
ℹ	WP3678	PATHWAY:Amplification and expansion of oncogenic pathways as metastatic traits_WikiPathways_20230410_\
ℹ	WP3967	PATHWAY:miR-509-3p alteration of YAP1/ECM axis_WikiPathways_20230410_WP3967_Homo sapiens
ℹ	WP2839	PATHWAY:Hair follicle development organogenesis - part 2 of 3_WikiPathways_20230410_WP2839_Homo sap
ℹ	WP4786	PATHWAY:Type I collagen synthesis in the context of osteogenesis imperfecta_WikiPathways_20230410_WP4
ℹ	WP47	PATHWAY:Hedgehog signaling pathway_WikiPathways_20230410_WP47_Homo sapiens
ℹ	WP1602	PATHWAY:Nicotine effect on dopaminergic neurons_WikiPathways_20230410_WP1602_Homo sapiens
ℹ	WP2361	PATHWAY:Gastric cancer network 1_WikiPathways_20230410_WP2361_Homo sapiens

Showing 1 to 34 of 510 entries

Contrast:

Condition:ALS_vs_control

□ filter signficant (tables)

17.3.3 Reactome

Reactome is a collection of manually curated pathways representing the current knowledge of molecular interactions, reactions and relation networks as pathway maps. In the **Reactome** panel, each pathway is scored for the selected contrast profile and reported in the table. A unique feature of the platform is that it provides an activation-heatmap comparing the activation levels of pathways across multiple contrast profiles. An activation-heatmap comparing the activation levels of pathways across multiple contrast profiles is generated. More detailed explaination of each output is provided below.

Reactome

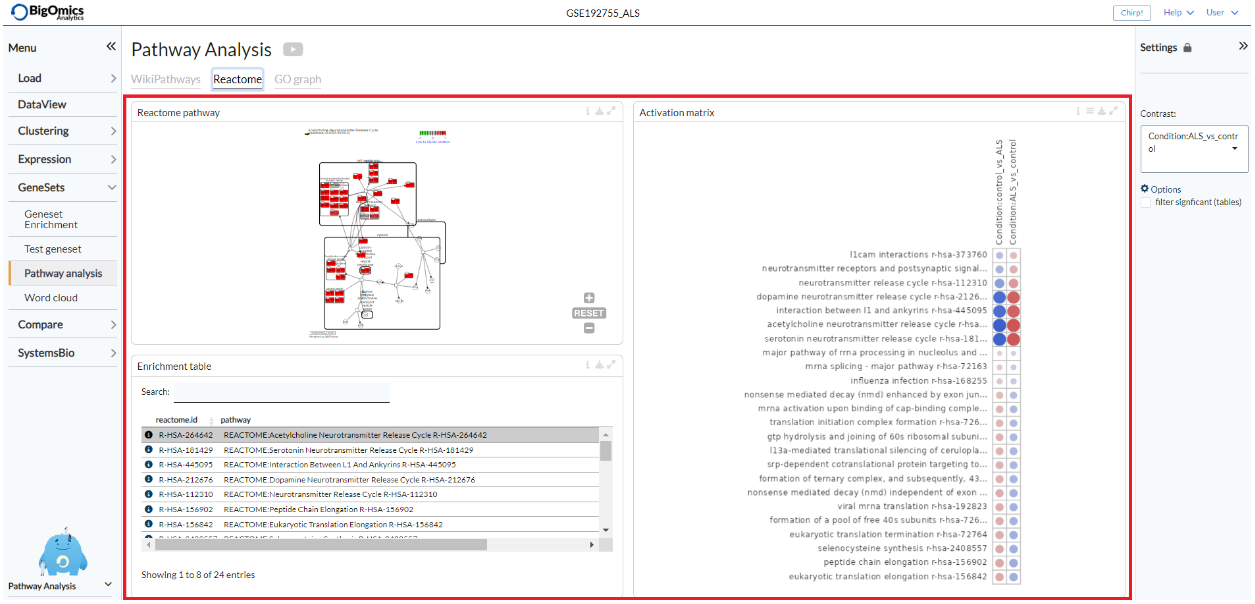
In the pathway map, genes are colored according to their upregulation (red) or downregulation (green) in the contrast profile. Each pathway is scored for the selected contrast profile and reported in the table below the plot.

Enrichment table

The table is interactive; enabling user to sort on different variables (Reactome id, pathway, logFC and meta q-values) and select a pathway by clicking on the row in the table.

Activation matrix

The activation matrix visualizes the activation levels of pathways (or pathway keywords) across multiple contrast profiles. This facilitates to quickly see and detect the similarities of certain pathways between contrasts. The size of the circles correspond to their relative activation, and are colored according to their upregulation (red) or downregulation (blue) in the contrast profile. The matrix can be normalised from the *settings* icon.


BigOmics Analytics
GSE192755_ALS
Chirp!
Help
User
Menu
Load
DataView
Clustering
Expression
GeneSets
Geneset Enrichment
Test geneset
Pathway analysis
Word cloud
Compare
SystemsBio
Pathway Analysis
Pathway Analysis
WikiPathways
Reactome
GO graph
Reactome pathway
RESET
Activation matrix
Condition:control_vs_ALS
Condition:ALS_vs_control
l1cam interactions r-hsa-373760
neurotransmitter receptors and postsynaptic signal...
neurotransmitter release cycle r-hsa-112310
dopamine neurotransmitter release cycle r-hsa-2126...
interaction between l1 and ankyrins r-hsa-445095
acetylcholine neurotransmitter release cycle r-hsa...
serotonin neurotransmitter release cycle r-hsa-181...
major pathway of rrna processing in nucleolus and ...
mrna splicing - major pathway r-hsa-72163
influenza infection r-hsa-168255
nonsense mediated decay (nmd) enhanced by exon jun...
mrna activation upon binding of cap-binding comple...
translation initiation complex formation r-hsa-726...
gtp hydrolysis and joining of 60s ribosomal subuni...
l13a-mediated translational silencing of ceruloplа...
srp-dependent cotranslational protein targeting to...
formation of ternary complex, and subsequently, 43...
nonsense mediated decay (nmd) independent of exon ...
viral mrna translation r-hsa-192823
formation of a pool of free 40s subunits r-hsa-726...
eukaryotic translation termination r-hsa-72764
selenocysteine synthesis r-hsa-2408557
peptide chain elongation r-hsa-156902
eukaryotic translation elongation r-hsa-156842
Settings
Options

Contrast: Condition:ALS_vs_control

□ filter signficant (tables)

Enrichment table

Search:

reactome.id	pathway
R-HSA-264642	REACTOME:Acetylcholine Neurotransmitter Release Cycle R-HSA-264642
R-HSA-181429	REACTOME:Serotonin Neurotransmitter Release Cycle R-HSA-181429
R-HSA-445095	REACTOME:Interaction Between L1 And Ankyrins R-HSA-445095
R-HSA-212676	REACTOME:Dopamine Neurotransmitter Release Cycle R-HSA-212676
R-HSA-112310	REACTOME:Neurotransmitter Release Cycle R-HSA-112310
R-HSA-156902	REACTOME:Peptide Chain Elongation R-HSA-156902
R-HSA-156842	REACTOME:Eukaryotic Translation Elongation R-HSA-156842

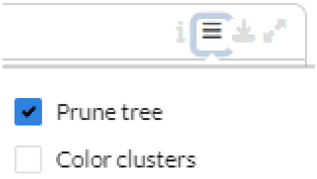
Showing 1 to 8 of 24 entries

17.3.4 GO graph

In the **GO** panel, users can perform GO analysis. GO defines functional concepts/classes and their relationships as a hierarchical graph. The GO database provides a computational representation of the current knowledge about roles of genes for many organisms in terms of molecular functions, cellular components and biological processes. All the features described under the **KEGG pathway** panel, such as scoring the gene sets and drawing an activation-heatmap, can be performed for the GO database under the GO graph tab. Instead of pathway maps, an annotated graph structure provided by the GO database is potted for every selected gene set. Each output chart/table of the panel is describer below in detail.

Gene Ontology graph

The structure of GO can be described in terms of a graph, where each GO term is a node, and the relationships between the terms are edges between the nodes. GO is loosely hierarchical, with 'child' terms being more specialized than their 'parent' terms. The graph is interactive. You can move the graph and zoom in using the mouse. Under the graph *settings*, users can select `Prune tree` to prune the tree only with significant branches and `color custers` to highlight clusters with different colors

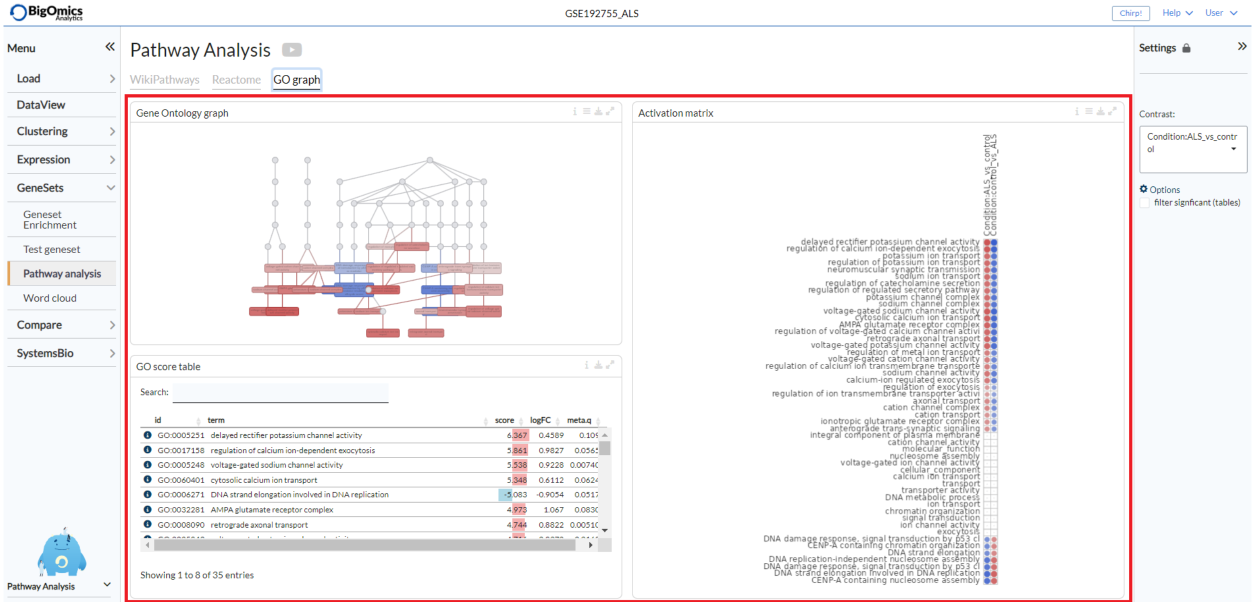

☑ Prune tree
□ Color clusters

GO score table

The scoring of a GO term is performed by considering the cumulative score of all terms from that term to the root node. That means that GO terms that are supported by higher level terms levels are preferentially scored.

Activation matrix

The GO activation matrix visualizes the activation of GO terms across conditions. From this figure, you can easily detect GO terms that are consistently up/down across conditions. The size of the circles correspond to their relative activation, and are colored according to their upregulation (red) or downregulation (blue) in the contrast profile. The matrix can be normalised from the *settings* icon.


BigOmics Analytics
GSE192755_ALS
Chirp!
Help
User
Menu
Load
DataView
Clustering
Expression
GeneSets
Geneset Enrichment
Test geneset
Pathway analysis
Word cloud
Compare
SystemsBio
Pathway Analysis
Pathway Analysis
WikiPathways
Reactome
GO graph
Gene Ontology graph
Activation matrix
Settings
Contrast:
Condition:ALS_vs_control
Options
filter signficant (tables)
Condition:ALS_vs_control
Condition:control_vs_ALS
delayed rectifier potassium channel activity
regulation of calcium ion-dependent exocytosis
potassium ion transport
regulation of potassium ion transport
neuromuscular synaptic transmission
sodium ion transport
regulation of catecholamine secretion
regulation of regulated secretory pathway
potassium channel complex
sodium channel complex
voltage-gated sodium channel activity
cytosolic calcium ion transport
AMPA glutamate receptor complex
regulation of voltage-gated calcium channel activi
retrograde axonal transport
voltage-gated potassium channel activity
regulation of metal ion transport
voltage-gated cation channel activity
regulation of calcium ion transmembrane transporte
sodium channel activity
calcium-ion regulated exocytosis
regulation of exocytosis
regulation of ion transmembrane transporter activi
axonal transport
cation channel complex
cation transport
ionotropic glutamate receptor complex
anterograde trans-synaptic signaling
integral component of plasma membrane
cation channel activity
molecular_function
nucleosome assembly
voltage-gated ion channel activity
cellular_component
calcium ion transport
transport
transporter activity
DNA metabolic process
ion transport
chromatin organization
signal transduction
ion channel activity
exocytosis
DNA damage response, signal transduction by p53 cl
CENP-A containing chromatin organization
DNA strand elongation
DNA replication-independent nucleosome assembly
DNA damage response, signal transduction by p53 cl
DNA strand elongation involved in DNA replication
CENP-A containing nucleosome assembly

GO score table

Search:

id	term	score	logFC	meta.q
GO:0005251	delayed rectifier potassium channel activity	6.367	0.4589	0.109
GO:0017158	regulation of calcium ion-dependent exocytosis	5.861	0.9827	0.0565
GO:0005248	voltage-gated sodium channel activity	5.538	0.9228	0.00740
GO:0060401	cytosolic calcium ion transport	5.348	0.6112	0.0624
GO:0006271	DNA strand elongation involved in DNA replication	-5.083	-0.9054	0.0517
GO:0032281	AMPA glutamate receptor complex	4.973	1.067	0.0830
GO:0008090	retrograde axonal transport	4.744	0.8822	0.00510

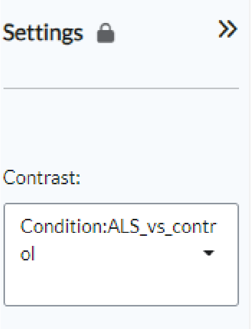
Showing 1 to 8 of 35 entries

17.4 Word Cloud

The **WordCloud** panel performs “keyword enrichment analysis”. It computes enrichment of a selected keyword across all contrasts. Select a keyword by clicking a word in the ‘Enrichment table’. Keyword enrichment is computed by running GSEA on the enrichment score profile for all contrasts. We defined the test set as the collection of genesets that contain the keyword in the title/description.

17.4.1 Settings panel

Users can specify the contrast of their interest in the `Contrast` settings.



Contrast:

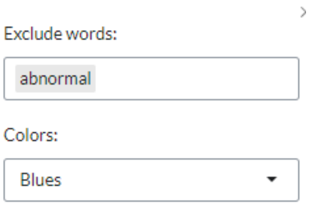
Condition:ALS_vs_control

17.4.2 Main panel

The main panel consists of five different outputs:

Enrichment plots
The plots visualize the enrichment of the selected keyword in the contrasts. Black vertical bars indicate the position of gene sets that contains the *keyword* in the ranked list of enrichment scores. The curve in green corresponds to the ‘running statistic’ of the keyword enrichment score. The more the green ES curve is shifted to the upper left of the graph, the more the keyword is enriched in the first group. Conversely, a shift of the green ES curve to the lower right, corresponds to keyword enrichment in the second group.

Word cloud
The word cloud plot visualizes the frequency/enrichment of keywords for the data set. Select a keyword in the ‘Enrichment table’. In the plot settings, users can exclude certain words from the figure, or choose the color palette. The sizes of the words are relative to the normalized enrichment score (NES) from the GSEA computation. Keyword enrichment is computed by running GSEA on the mean (squared) enrichment profile (averaged over all contrasts). For each keyword, we defined the ‘keyword set’ as the collection of genesets that contain that keyword in the title/description.



Exclude words:

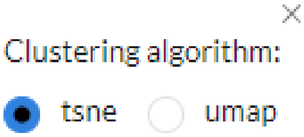
abnormal

Colors:

Blues

Word t-SNE

The Word t-SNE plot visualizes the similarity of the keywords that were found in the title/description of gene sets. Keywords that are often found together in title/descriptions are placed close together in the t-SNE. For each keyword we computed enrichment using GSEA on the mean (absolute) enrichment profiles (averaged over all contrasts). Statistically significant gene sets (q<0.05) are colored in red. The sizes of the nodes are proportional to the normalized enrichment score (NES) of the keyword. In the plot settings, the user can choose between t-SNE and "Uniform Manifold Approximation and Projection" (UMAP).



Clustering algorithm:

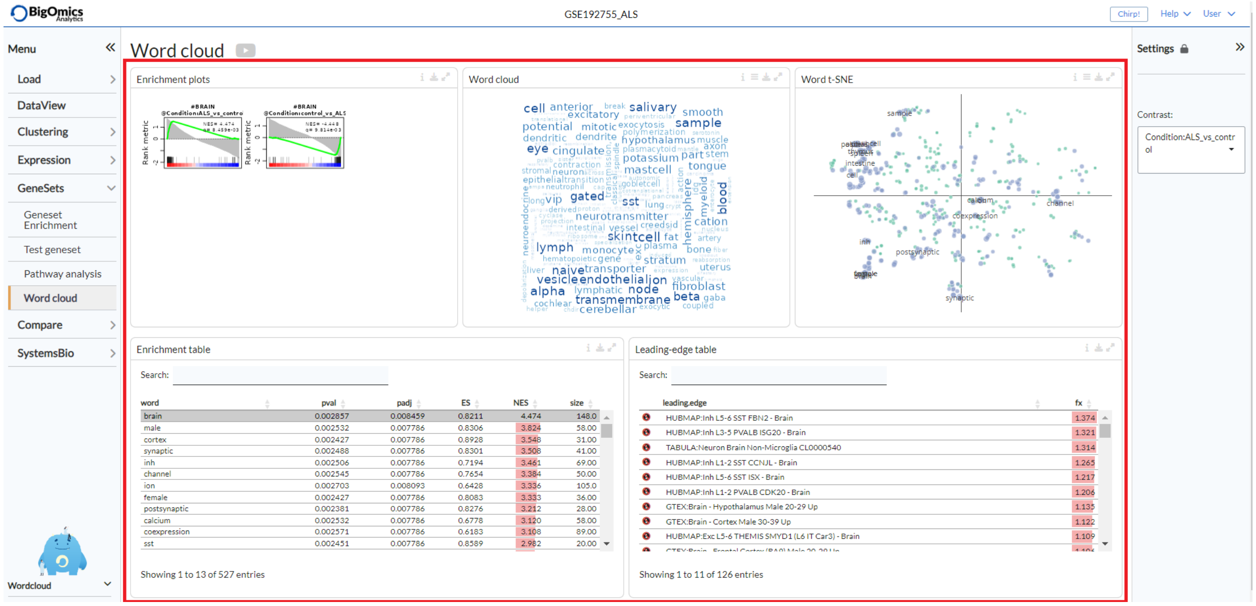
● tsne ○ umap

Enrichment table

The Enrichment table summarizes the results from the enrichment test for the tested keywords. The NES corresponds to the normalized enrichment score from the GSEA analysis.

Leading-edge table

The Leading-edge table shows the geneset titles that have contributed to the enrichment of the selected keyword.



Contrast: Condition:ALS_vs_control

Enrichment table

Search:

word	pval	padj	ES	NES	size
brain	0.002857	0.008459	0.8211	4.474	148.0
male	0.002532	0.007786	0.8306	3.824	58.00
cortex	0.002427	0.007786	0.8928	3.548	31.00
synaptic	0.002488	0.007786	0.8301	3.508	41.00
inh	0.002506	0.007786	0.7194	3.461	69.00
channel	0.002545	0.007786	0.7654	3.384	50.00
ion	0.002703	0.008093	0.6428	3.336	105.0
female	0.002427	0.007786	0.8083	3.333	36.00
postsynaptic	0.002381	0.007786	0.8276	3.212	28.00
calcium	0.002532	0.007786	0.6778	3.120	58.00
coexpression	0.002571	0.007786	0.6183	3.108	89.00
sst	0.002451	0.007786	0.8589	2.982	20.00

Showing 1 to 13 of 527 entries

Leading-edge table

Search:

leading.edge	fx
HUBMAP:Inh L5-6 SST FBN2 - Brain	1.374
HUBMAP:Inh L3-5 PVALB ISG20 - Brain	1.321
TABULA:Neuron Brain Non-Microglia CL0000540	1.314
HUBMAP:Inh L1-2 SST CCNJL - Brain	1.265
HUBMAP:Inh L5-6 SST ISX - Brain	1.217
HUBMAP:Inh L1-2 PVALB CDK20 - Brain	1.206
GTEX:Brain - Hypothalamus Male 20-29 Up	1.135
GTEX:Brain - Cortex Male 30-39 Up	1.122
HUBMAP:Exc L5-6 THEMIS SMYD1 (L6 IT Car3) - Brain	1.109

Showing 1 to 11 of 126 entries

CHAPTER

EIGHTEEN

COMPARE

The **Compare** module consists of three submodules: **Compare signatures**, **Compare datasets** and **Similar experiments**.

Compare signatures allows users to compare experiments by intersecting their signature genes. Under **Compare datasets**, users can compare pairwise comparisons from the dataset in use and another dataset stored in the platform. Finally, with the **Similar experiments** panel, users can perform a large-scale comparison across all uploaded datasets or a pre-loaded collection of more than 6000 public dataset.

18.1 Compare signatures

The main goal of this submodule is to identify contrasts showing similar profiles and find genes that are commonly up/down regulated between two contrasts. The panel is divided into two subpanels: **Pairwise scatter** and **Signature clustering**.

18.1.1 Settings panel

Users can select contrasts to compare from the `Contrast` settings in the input panel on the left. Under *Options*, the `Level` setting allows users to toggle between gene or gene set level analysis, while users can set the `Filter` for selecting specific features (e.g. a specific gene family or gene set).

18.1.2 Pairwise scatter

The **Pairwise scatter** panel shows three outputs:

Scatterplot pairs

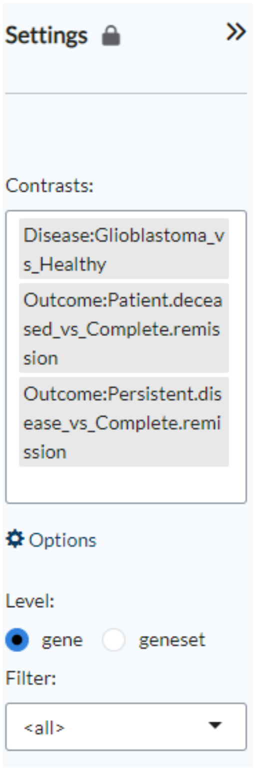
The Pairs plot provides interactive pairwise scatterplots for the differential expression profiles corresponding to multiple contrasts. The main purpose of this panel is to identify similarity or dissimilarity between selected contrasts. When K>=3 contrasts are selected, the figure shows a KxK scatterplot matrix. Via the *Settings*, users can disable the highlighting of genes on the plot.

Venn diagram

The Venn diagram visualizes the number of intersecting genes between the selected contrast profiles. The diagram can be customised via the settings icon by altering `logFC` and `FDR` thresholds and by selecting whether to view disregulated genes jointly or separated by up- and down-regulation under `Counting`.

Leading-edge table

Venn diagram areas can be selected via the settings icon (`Filter intersection`) and are represented by corresponding letters (e.g. ‘ABC’ represents the intersection of contrasts A, B and C). Contrast letter identifiers are provided in the Venn Diagram.



Settings

Contrasts:

Disease:Glioblastoma_vs_Healthy

Outcome:Patient.deceased_vs_Complete.remission

Outcome:Persistent.disease_vs_Complete.remission

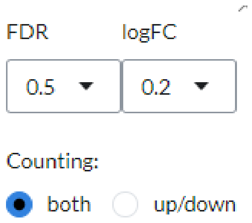
Options

Level:

● gene ○ geneset

Filter:

<all>

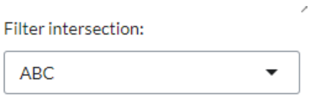


FDR logFC

0.5 0.2

Counting:

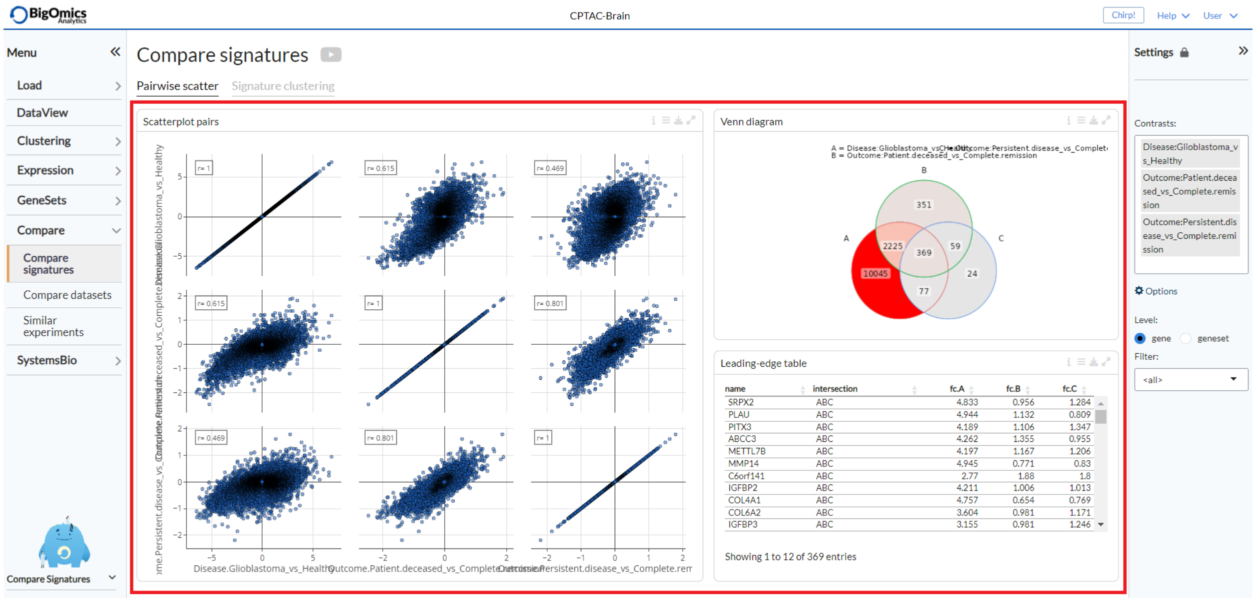
● both ○ up/down



Filter intersection:

ABC

The three output panels are highlighted in the figure below.



Leading-edge table

name	intersection	fc.A	fc.B	fc.C
SRPX2	ABC	4.833	0.956	1.284
PLAU	ABC	4.944	1.132	0.809
PITX3	ABC	4.189	1.106	1.347
ABCC3	ABC	4.262	1.355	0.955
METTL7B	ABC	4.197	1.167	1.206
MMP14	ABC	4.945	0.771	0.83
C6orf141	ABC	2.77	1.88	1.8
IGFBP2	ABC	4.211	1.006	1.013
COL4A1	ABC	4.757	0.654	0.769
COL6A2	ABC	3.604	0.981	1.171
IGFBP3	ABC	3.155	0.981	1.246

Showing 1 to 12 of 369 entries

Contrasts:
Disease:Glioblastoma_vs_Healthy
Outcome:Patient.deceased_vs_Complete.remission
Outcome:Persistent.disease_vs_Complete.remission

Options

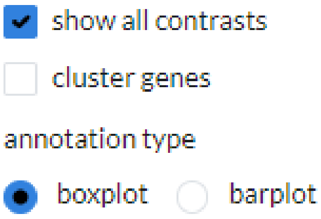
Level: ● gene ○ geneset

Filter: <all>

18.1.3 Signature clustering

Two plots are showed in this panel:

Foldchange heatmap
The foldchange heatmap shows the similarity of the contrasts visualized as a clustered heatmap. Contrasts that are similar will be clustered close together. The plot can be customised via the settings icon. Users can select to show only the selected contrasts (default is to show all), cluster the genes on the heatmap and change the annotation type between logFC (*boxplot*) and cumulative logFC (*barplot*).



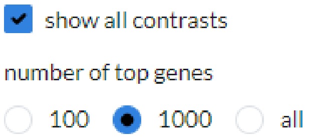
☑ show all contrasts

□ cluster genes

annotation type

● boxplot ○ barplot

Contrast correlation
Contrasts that are similar will be clustered close together. The numeric value in the cells correspond to the Pearson correlation coefficient between contrast signatures. Red corresponds to positive correlation and blue to negative correlation. Under the plot settings, users can use `show all contrasts` (default) to show all contrasts or only the selected ones and change the `number of top genes` to specify the number of top genes for computations (default=1000).

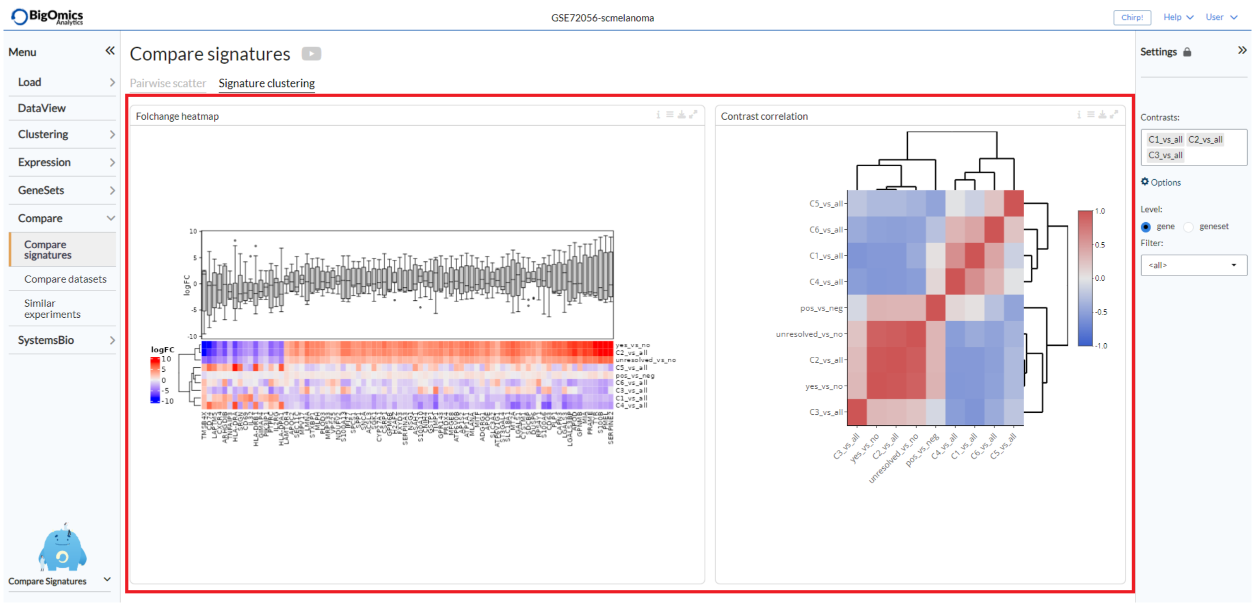


☑ show all contrasts

number of top genes

○ 100 ● 1000 ○ all

A typical output can be seen below.



Contrasts: C1_vs_all C2_vs_all C3_vs_all

Level: ● gene ○ geneset

Filter: <all>

18.2 Compare datasets

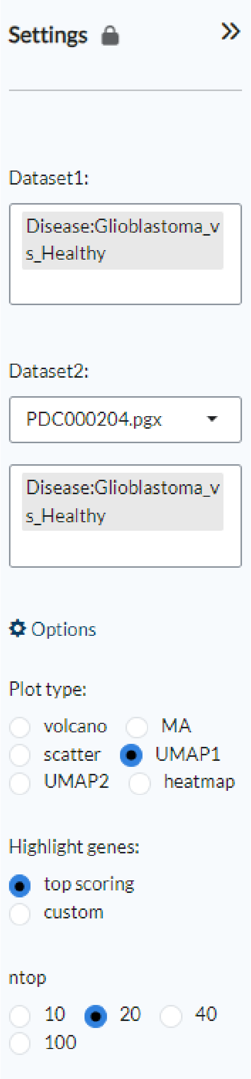
With this submodule, users can compare pairwise comparisons across datasets that have been uploaded into the platform. The submodule is split into thee tabs: **Compare expression**, **Foldchange** and **Gene Correlation**.

18.2.1 Settings panel

Users can select the pairwise comparisons to be selcted from `Dataset1` and `Dataset2`, from which they can also select a dataset from the list of uploaded experiments. Under *Options*, users can set the `Plot type` (default: UMAP1) to be displayed in the **Dataset1** and **Dataset2** panels. `Highlight genes` is used to label genes in the **Dataset1** and **Dataset2** plots. The choice is between highlighting top genes/proteins (default) or a custom list that users can type or paste in the corresponding box. With `ntop` users can define how many genes or proteins to label.

18.2.2 Compare expression

This panel shows plots for selected pairwise comparisons from the current dataset (*Dataset1*) and a second dataset (*Dataset2*) selected from a list of uploaded experiments. The type of plot can be selected via the main submodule **Settings**. Users can select between plotting the genes or proteins as UMAP plots based on either dataset 1 or dataset 2 (UMAP1 and UMAP2), as volcano, MA plots, scatter plots or heatmaps.



Settings

Dataset1:

Disease:Glioblastoma_vs_Healthy

Dataset2:

PDC000204.pgx

Disease:Glioblastoma_vs_Healthy

Options

Plot type:

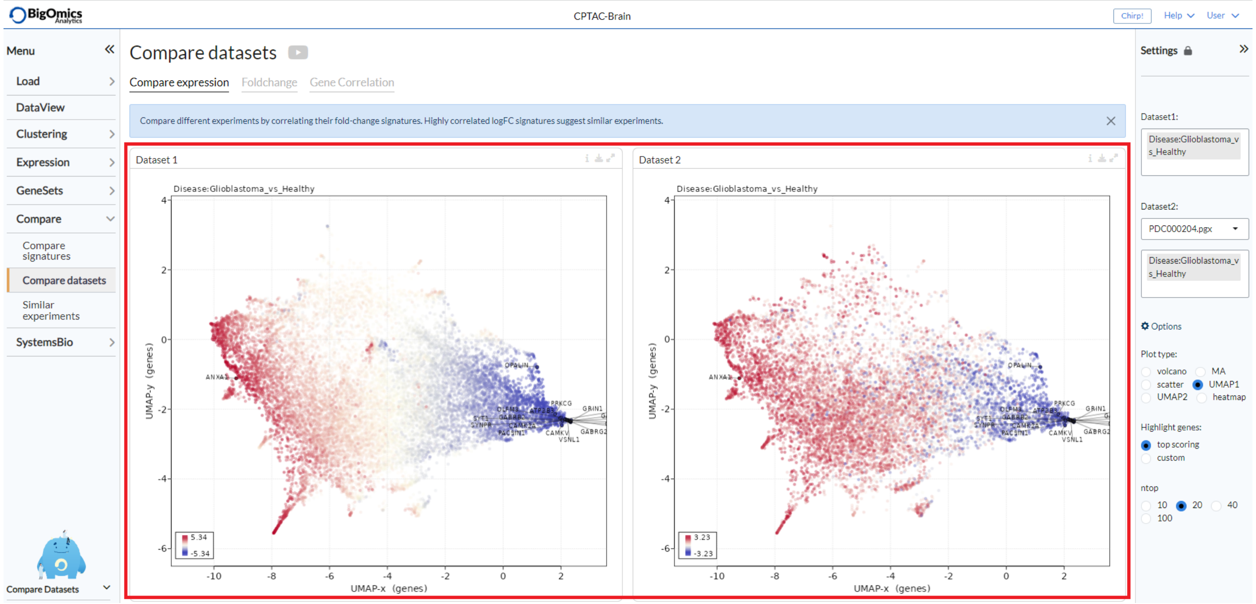
○ volcano ○ MA
○ scatter ● UMAP1
○ UMAP2 ○ heatmap

Highlight genes:

● top scoring
○ custom

ntop

○ 10 ● 20 ○ 40
○ 100



Compare datasets

Compare different experiments by correlating their fold-change signatures. Highly correlated logFC signatures suggest similar experiments.

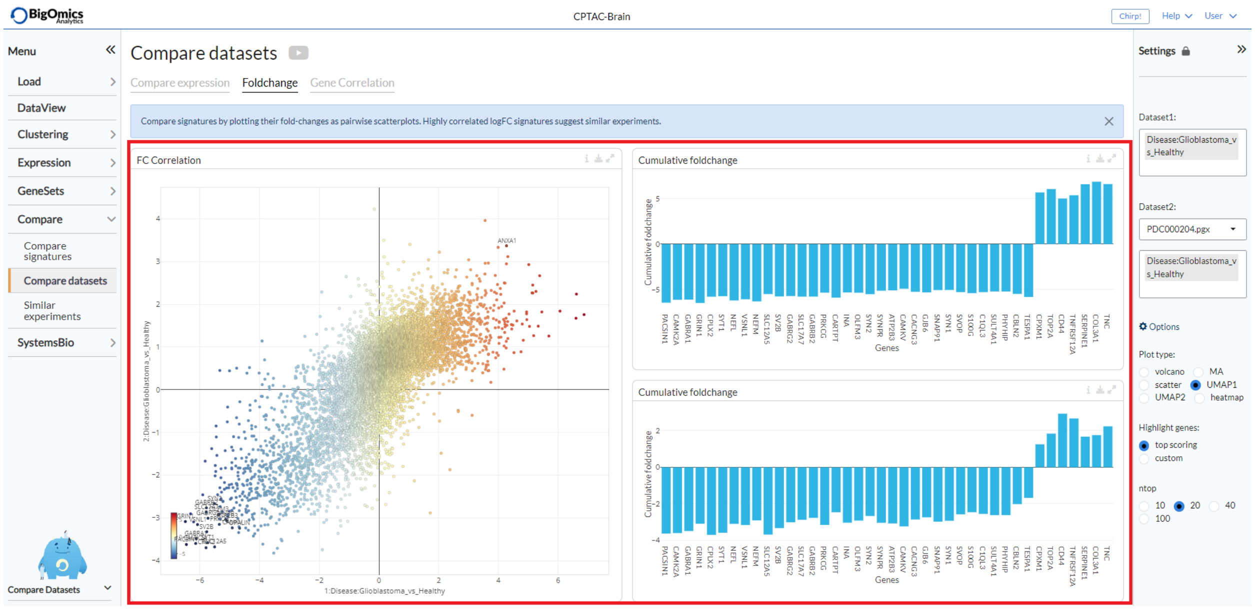
18.2.3 Fold change

This tab contains three panels: **FC correlation**, which contains the plot between the pairwise comparsions selected between two datasets and the **Cumulative foldchange** panels that show barplot highlighting the fold changes in expression for each of the selected pairwise comparisons.

FC correlation
: Scatter plot of gene expression scatter values between two contrasts. Scatters that are similar show high correlation, i.e. are close to the diagonal.

Cumulative foldchange upper
: Barplot showing the cumulative fold changes on dataset 1.

Cumulative foldchange lower
: Barplot showing the cumulative fold changes on dataset 2.



Compare signatures by plotting their fold-changes as pairwise scatterplots. Highly correlated logFC signatures suggest similar experiments.

Dataset1: Disease:Glioblastoma_vs_Healthy

Dataset2: PDC000204.pgx

Disease:Glioblastoma_vs_Healthy

Plot type: ○ volcano ○ MA ○ scatter ● UMAP1 ○ UMAP2 ○ heatmap

Highlight genes: ● top scoring ○ custom

ntop: ○ 10 ● 20 ○ 40 ○ 100

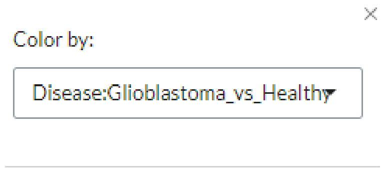
18.2.4 Gene correlation

The **Gene Correlation** tab is used to compare the expression levels of individual genes or proteins between pairwise comparions across datasets. It can also be used for the combined analysis of proteomics and transcriptomics datasets. It contains three panels: **Expression**, **Correlation score** and **Gene correlation**.

Expression
: Barplots of expression values for multiple comparisons in the two datasets (blue and green). Bars are labelled by pairwise comparison groups.

Correlation score
: In this searchable table, users can check mean expression values and correlation scores of genes/proteins across the selected pairwise comparisons.

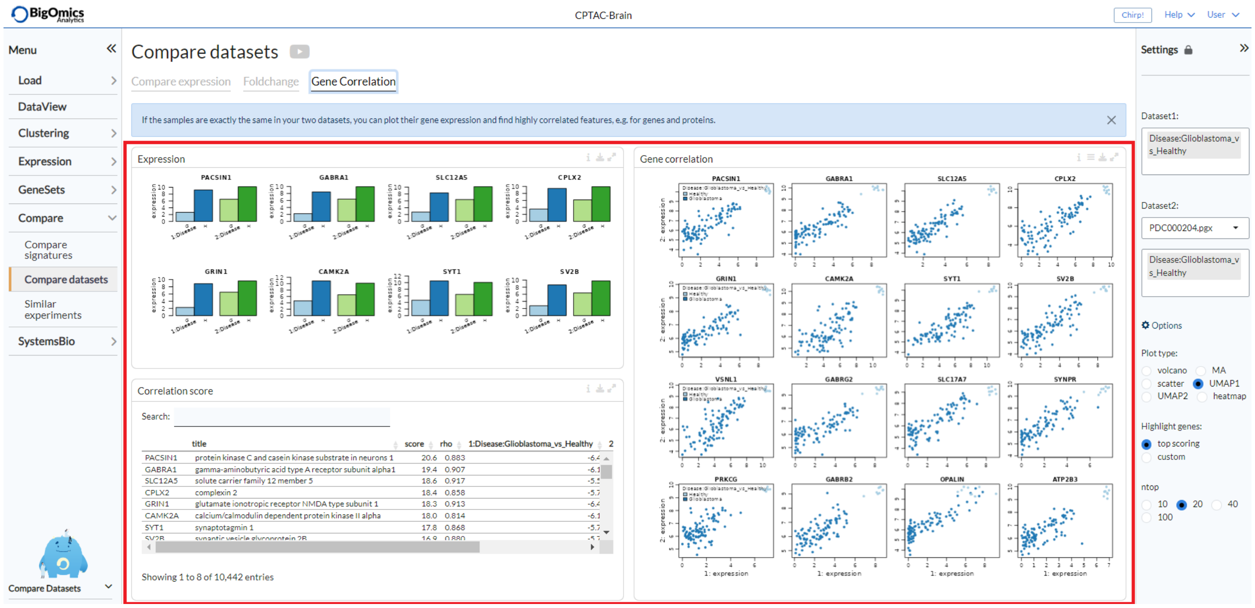
Gene correlation
: Scatter plots of gene expression scatter values between two contrasts. Scatters that are similar show high correlation, i.e. are close to the diagonal. This plot is only available for studies with matched sample Ids and can be used to compare proteomics and transcriptomics datasets from the same samples. Users can select by which pairwise comparison to colour the samples via the settings icon.



Color by:

Disease:Glioblastoma_vs_Healthy

The genes or proteins appearing in the **Expression** barplots and **Gene correlation** scatter plots will be determined in the **Correlation score** table.



If the samples are exactly the same in your two datasets, you can plot their gene expression and find highly correlated features, e.g. for genes and proteins.

Correlation score

Search:

	title	score	rho	1:Disease:Glioblastoma_vs_Healthy	2
PACSIN1	protein kinase C and casein kinase substrate in neurons 1	20.6	0.883	-6.4	
GABRA1	gamma-aminobutyric acid type A receptor subunit alpha1	19.4	0.907	-6.1	
SLC12A5	solute carrier family 12 member 5	18.6	0.917	-5.5	
CPLX2	complexin 2	18.4	0.858	-5.7	
GRIN1	glutamate ionotropic receptor NMDA type subunit 1	18.3	0.913	-6.4	
CAMK2A	calcium/calmodulin dependent protein kinase II alpha	18.0	0.814	-6.1	
SYT1	synaptotagmin 1	17.8	0.868	-5.7	
SV2B	synaptic vesicle glycoprotein 2B	16.9	0.880	-5.7	

Showing 1 to 8 of 10,442 entries

Dataset1: Disease:Glioblastoma_vs_Healthy

Dataset2: PDC000204.pgx

Disease:Glioblastoma_vs_Healthy

Options

Plot type: ○ volcano ○ MA ○ scatter ● UMAP1 ○ UMAP2 ○ heatmap

Highlight genes: ● top scoring ○ custom

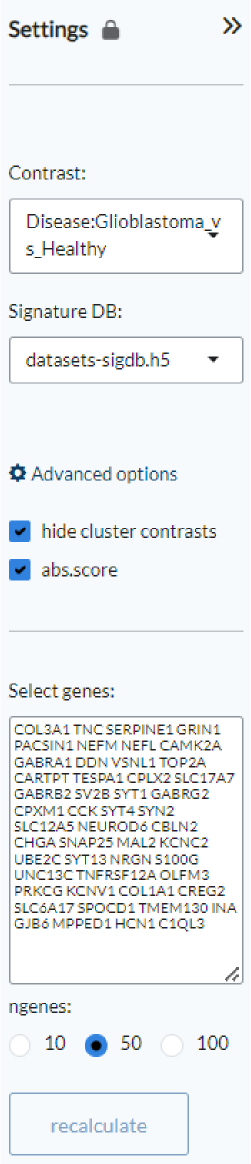
ntop: ○ 10 ● 20 ○ 40 ○ 100

18.3 Similar experiments

With the final submodule, users can compare contrasts from different experiments against a selected pairwise comparisons from the current dataset simultenously. Rather than a pairwise analysis between two datasets, this analysis will take into accounts all uploaded datasets or, alternatively, access two databases of more than 6000 individual experiments collected from the GEO database. The submodule is split into three tabs: **FC correlation**, **FC heatmap** and **Meta-network**.

18.3.1 Settings panel

Through the **Settings** panel, an experimental `Contrast` from the selection of available pairwise comparisons in the dataset can be set. Under `Signature DB`, users can select from three databases: *datasets-sigdb.h5* corresponds to the datasets that have been uploaded by teh user into the platform, *sigdb-archs4.h5* correspond to a collection of more than 6000 datasets from the GEO database, while *sigdb-virome.h5* contains a selection of more than 300 viral infection GEO datasets. Under the *Advanced options*, users can choose whether to `hide cluster contrasts` (default) and show the absolute score (`abs.score`), which is the default choice. Under `Select genes` it is possible to type or paste a user-generated list of genes/proteins and also select the number of genes to be labelled (default: 50). Clicking `Recalculate` will then generate new plots based on the inputed gene list.



Contrast: Disease:Glioblastoma_vs_Healthy

Signature DB: datasets-sigdb.h5

Advanced options

☑ hide cluster contrasts

☑ abs.score

Select genes:

```
COL3A1 TNC SERPINE1 GRIN1
PACSIN1 NEFM NEFL CAMK2A
GABRA1 DDN VSNL1 TOP2A
CARTPT TESPA1 CPLX2 SLC17A7
GABRB2 SV2B SYT1 GABRG2
CPXM1 CCK SYT4 SYN2
SLC12A5 NEUROD6 CBLN2
CHGA SNAP25 MAL2 KCNC2
UBE2C SYT13 NRGN S100G
UNC13C TNFRSF12A OLFM3
PRKCG KCNV1 COL1A1 CREG2
SLC6A17 SPOCD1 TMEM130 INA
GJB6 MPPED1 HCN1 C1QL3
```

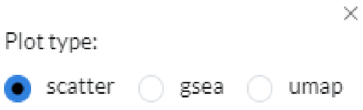
ngenes: ○ 10 ● 50 ○ 100

18.3.2 FC correlation

With this tab it is possible to compare different experiments by correlating their fold-change signatures. The tab consists of three panels: **FC scatter plots**, **Similarity scores** and **FC-FC scatterplot**.

FC scatter plots

Scatter plots of gene expression foldchange values between two contrasts. Foldchanges that are similar show high correlation, i.e. are close to the diagonal. You can switch to gsea or UMAP enrichment plots in the `plot type` option in the settings icon.



Plot type:

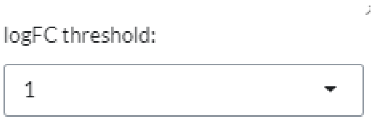
● scatter ○ gsea ○ umap

Similarity score

In this searchable table, Normalized enrichment scores (NES) and Pearson correlation (rho) of reference profiles with respect to the currently selected contrast are displayed. The top 100 up/down genes are considered for the calculation of rho or NES. The score is calculated as rho^2*NES. Highlighting a specific dataset will change the FC-FC scatterplot accordingly.

FC-FC scatterplot

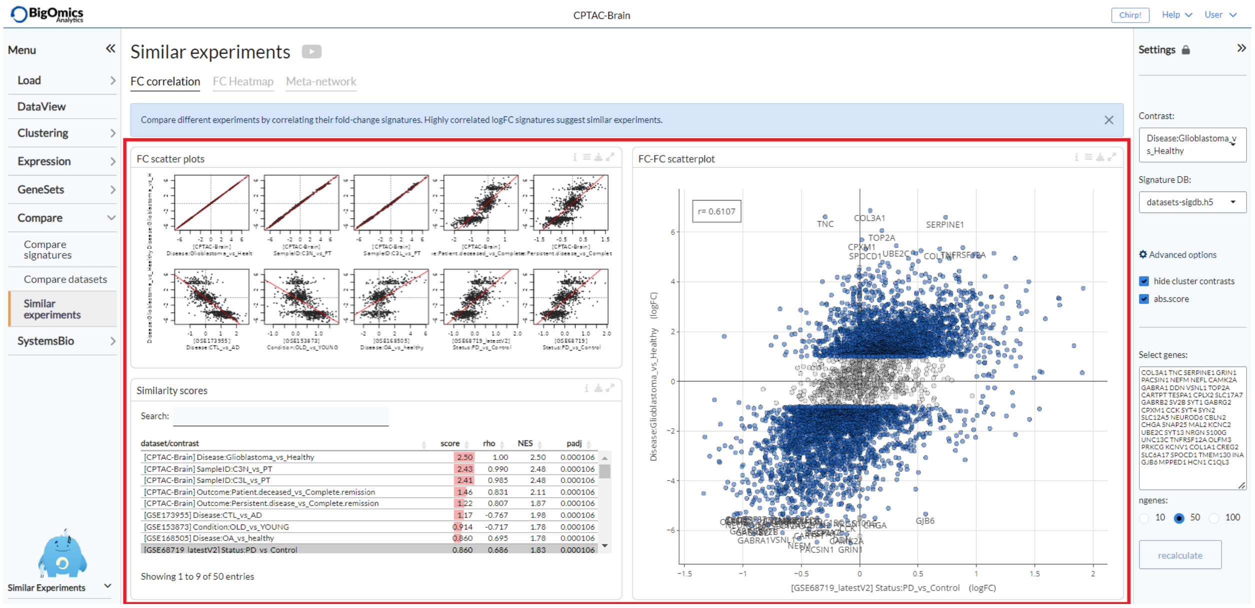
This plot provides a pairwise scatterplot of logFC fold-change profiles for the selected contrasts. The main purpose of this panel is to identify similarity or dissimilarity between selected contrasts. The scatter plot is interactive and shows information of each gene by hovering over it with the mouse. The `logFC threshold` can be set via the settings icon.



logFC threshold:

1

The three panels are highlighted in the image below.



Compare different experiments by correlating their fold-change signatures. Highly correlated logFC signatures suggest similar experiments.

Similarity scores

Search:

dataset/contrast	score	rho	NES	padj
[CPTAC-Brain] Disease:Glioblastoma_vs_Healthy	2.50	1.00	2.50	0.000106
[CPTAC-Brain] SampleID:C3N_vs_PT	2.43	0.990	2.48	0.000106
[CPTAC-Brain] SampleID:C3L_vs_PT	2.41	0.985	2.48	0.000106
[CPTAC-Brain] Outcome:Patient.deceased_vs_Complete.remission	1.46	0.831	2.11	0.000106
[CPTAC-Brain] Outcome:Persistent.disease_vs_Complete.remission	1.22	0.807	1.87	0.000106
[GSE173955] Disease:CTL_vs_AD	1.17	-0.767	1.98	0.000106
[GSE153873] Condition:OLD_vs_YOUNG	0.914	-0.717	1.78	0.000106
[GSE168505] Disease:OA_vs_healthy	0.860	0.695	1.78	0.000106
[GSE68719_latestV2] Status:PD_vs_Control	0.860	0.686	1.83	0.000106

Showing 1 to 9 of 50 entries

Contrast: Disease:Glioblastoma_vs_Healthy

Signature DB: datasets-sigdb.h5

Advanced options

☑ hide cluster contrasts

☑ abs.score

Select genes: COL3A1 TNC SERPINE1 GRIN1 PACSIN1 NEFM NEFL CAMK2A GABRA1 DDN VSNL1 TOP2A CARTPT TESPA1 CPLX2 SLC17A7 GABRB2 SV2B SYT1 GABRG2 CPXM1 CCK SYT4 SYN2 SLC12A5 NEUROD6 CBLN2 CHGA SNAP25 MAL2 KCNC2 UBE2C SYT13 NRGN S100G UNC13C TNFRSF12A OLFM3 PRKCG KCNV1 COL1A1 CREG2 SLC6A17 SPOCD1 TMEM130 INA GJB6 MPPED1 HCN1 C1QL3

ngenes: ○ 10 ● 50 ○ 100

recalculate

18.3.3 FC heatmap

The **FC heatmap** tab provides three panels for more comparative analysis: **Connectivity map**, **Fold-change table** and **Connectivity heatmap**.

Connectvity map

The connectivity map plot shows the similarity of logFC signatures as a t-SNE plot. Signatures that are similar will be clustered close together, signatures that are different are placed farther away. The plot can be extensively customised via the options in the settings icon. `Layout` is usde to alter the layout of the plot, namely pca, tsne (default) and volcano. `Score threshold` thresholds the points by minimum score.``Color by`` is used to colour the samples by either score (default) or dataset. `Color gamma` is used for colour adjustment. Under `Other options`, users can add lables to the plot, turn it into a 3D plot, change the background colour to black and increase the size of the dots.

Plot type:

● scatter ○ gsea ○ umap

Fold-change table

In this searchable table, gene expression fold-changes (log2FC) of similar signatures across different experiments are indicated.

Connectivity heatmap

The heatmap clusters contrasts that are similar close together to provide an overview of the top most correlated contrasts with th equeried pairwise comparison. Through the settings icon, users can alter the `Number of signatures` to be displayed (default=20), they can also `Cluster genes` rather than sorting them by expression (default: off), `Use absolute foldchange` for calculating the cumulative sum and finally can `Reverse negative contrasts` (default: on), so as not to consider the direction of the correlation but only its strength.

These plots are complementary to the plots provided by the **FC correlation** tab, to explore more in detail the nature of the correlation between contrasts.

18.3.4 Meta-network

The final tab of the **Similar experiments** submodule provides some comparative gene network analysis as well as geneset level annotation of the most frequently enriched genesets across the correlated pairwise comparisons. It consists of four panels: **Leading-edge graph**, **Cumulative foldchange**, **Enrichment graph** and **Cumulative enrichment**.

Leading-edge graph

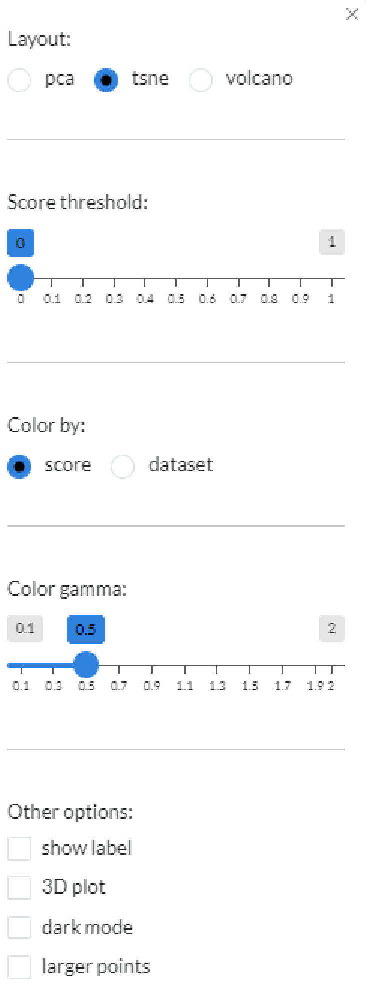
This graph displays the connections between genes shared across the correlated signatures. The edge width corresponds to the number of signatures that share that pair of genes in their top differentially expressed genes. Through the settings icon, it is possible to set the `Edge threshold`, select the number of signatures (`nr of signatures`) to be used (default=10) and select the `Size` parameter for the nodes (default: cumFC).

Cumulative foldchange

The barplot visualizes the cumulative foldchange between the top-10 most similar profiles. Genes that are frequently shared with high foldchange will show a higher cumulative score. You can choose between signed or `Absolute foldchange` in the options and use `order by` to order the plot by FC or cumulative FC (cumFC, default).

Enrichment graph

In this graph, the edge width corresponds to the number of signatures that share that pair of genesets in



Layout:

○ pca ● tsne ○ volcano

Score threshold:

0 … 1

Color by:

● score ○ dataset

Color gamma:

0.1 … 0.5 … 2

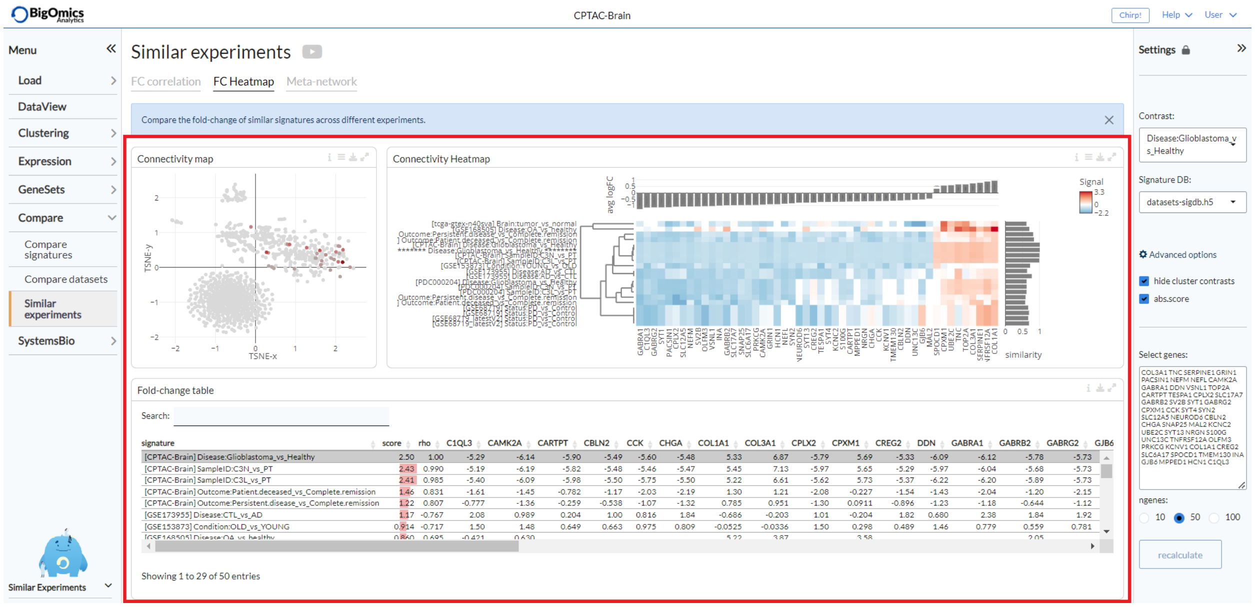
Other options:

□ show label

□ 3D plot

□ dark mode

□ larger points



Similar experiments

Compare the fold-change of similar signatures across different experiments.

Fold-change table

Search:

signature	score	rho	C1QL3	CAMK2A	CARTPT	CBLN2	CCK	CHGA	COL1A1	COL3A1	CPLX2	CPXM1	CREG2	DDN	GABRA1	GABRB2	GABRG2	GJB6
[CPTAC-Brain] Disease:Glioblastoma_vs_Healthy	2.50	1.00	-5.29	-6.14	-5.90	-5.49	-5.60	-5.48	5.33	6.87	-5.79	5.69	-5.33	-6.09	-6.12	-5.78	-5.73	
[CPTAC-Brain] SampleID:C3N_vs_PT	2.43	0.990	-5.19	-6.19	-5.82	-5.48	-5.46	-5.47	5.45	7.13	-5.97	5.65	-5.29	-5.97	-6.04	-5.68	-5.73	
[CPTAC-Brain] SampleID:C3L_vs_PT	2.41	0.985	-5.40	-6.09	-5.98	-5.50	-5.75	-5.50	5.22	6.61	-5.62	5.73	-5.37	-6.22	-6.20	-5.89	-5.73	
[CPTAC-Brain] Outcome:Patient.deceased_vs_Complete.remission	1.46	0.831	-1.61	-1.45	-0.782	-1.17	-2.03	-2.19	1.30	1.21	-2.08	-0.227	-1.54	-1.43	-2.04	-1.20	-2.15	
[CPTAC-Brain] Outcome:Persistent.disease_vs_Complete.remission	1.22	0.807	-0.777	-1.36	-0.259	-0.538	-1.07	-1.32	0.785	0.951	-1.30	0.0911	-0.896	-1.23	-1.18	-0.644	-1.12	
[GSE173955] Disease:CTL_vs_AD	1.17	-0.767	2.08	0.989	0.204	1.00	0.816	1.84	-0.686	-0.203	1.01	-0.204	1.82	0.680	2.38	1.84	1.92	
[GSE153873] Condition:OLD_vs_YOUNG	0.914	-0.717	1.50	1.48	0.649	0.663	0.975	0.809	-0.0525	-0.0336	1.50	0.298	0.489	1.46	0.779	0.559	0.781	
[GSE168505] Disease:OA_vs_healthy	0.860	0.695	-0.421	0.630					5.22	3.87		3.58				2.05		

Showing 1 to 29 of 50 entries

Settings

Contrast: Disease:Glioblastoma_vs_Healthy

Signature DB: datasets-sigdb.h5

Advanced options

☑ hide cluster contrasts

☑ abs.score

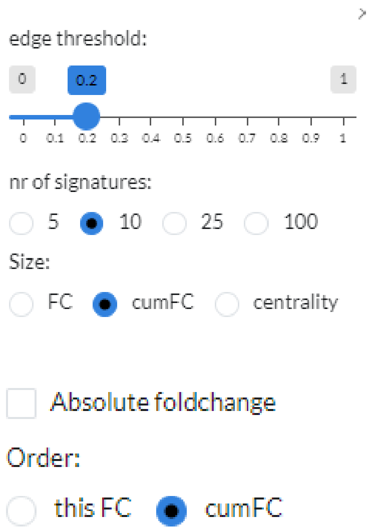
Select genes:

COL3A1 TNC SERPINE1 GRIN1 PACSIN1 NEFM NEFL CAMK2A GABRA1 DDN VSNL1 TOP2A CARTPT TESPA1 CPLX2 SLC17A7 GABRB2 SV2B SYT1 GABRG2 CPXM1 CCK SYT4 SYN2 SLC12A5 NEUROD6 CBLN2 CHGA SNAP25 MAL2 KCNC2 UBE2C SYT13 NRGN S100G UNC13C TNFRSF12A OLFM3 PRKCG KCNV1 COL1A1 CREG2 SLC6A17 SPOCD1 TMEM130 INA GJB6 MPPED1 HCN1 C1QL3

ngenes:

○ 10 ● 50 ○ 100

recalculate



edge threshold:

0 0.2 1

nr of signatures:

○ 5 ● 10 ○ 25 ○ 100

Size:

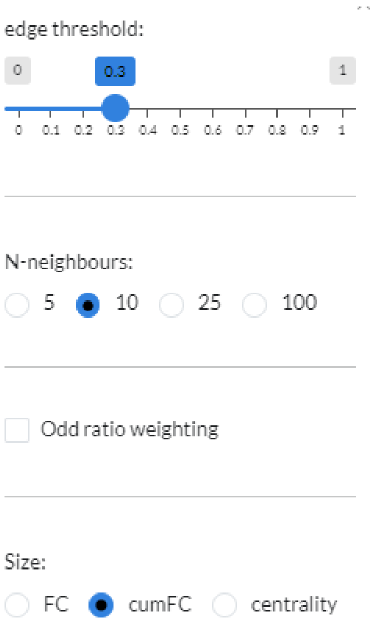
○ FC ● cumFC ○ centrality

☐ Absolute foldchange

Order:

○ this FC ● cumFC

their top enriched genesets. In the plot options you can set the `Edge threshold`, the number of similar experiments to consider (`N-neighbours`, default=10), whether to add a `Odd ratio weighting` (default: off) and the parameter to be used for the node `Size` (default: cumFC).



edge threshold:

0 0.3 1

N-neighbours:

○ 5 ● 10 ○ 25 ○ 100

☐ Odd ratio weighting

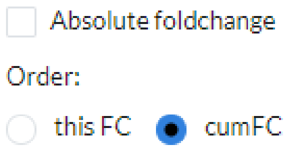
Size:

○ FC ● cumFC ○ centrality

Cumulative enrichment

In this plot, gene sets that are frequently shared with high enrichment will show a higher cumulative scores. You can choose between signed or `Absolute foldchange` in the options and use `order by` to order the plot by FC or cumulative FC (cumFC, default).

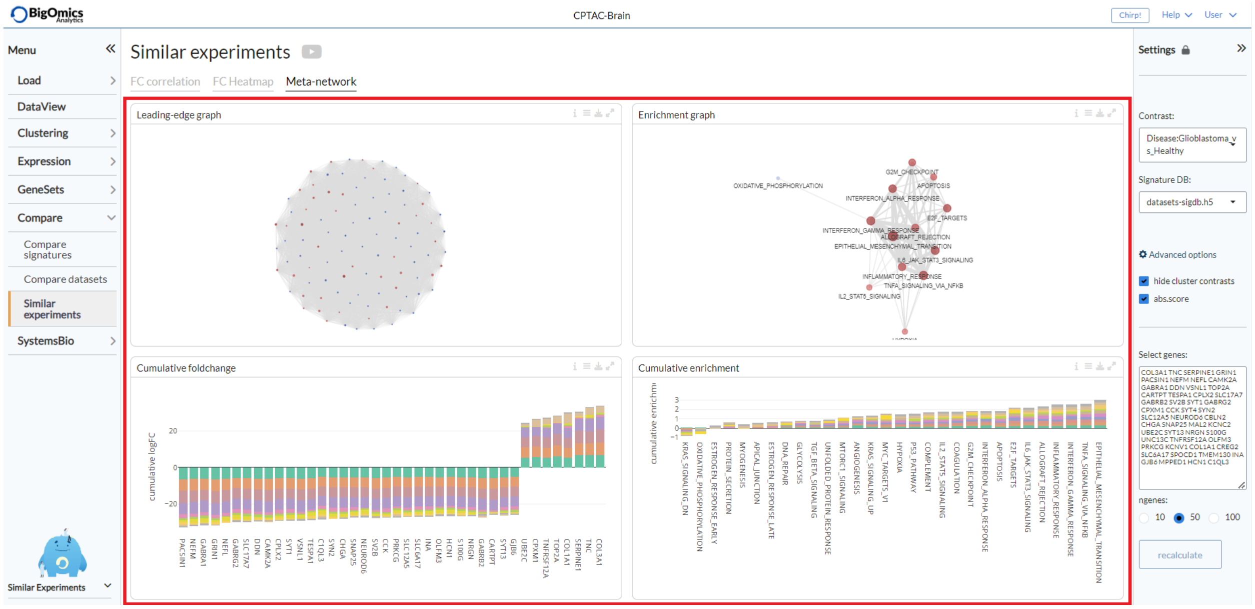
These plots are complementary to the plots provided by the **FC correlation** tab, to explore more in detail the nature of the correlation between contrasts.



□ Absolute foldchange

Order:

○ this FC ● cumFC



Contrast: Disease:Glioblastoma_vs_Healthy

Signature DB: datasets-sigdb.h5

☑ hide cluster contrasts

☑ abs.score

Select genes:

COL3A1 TNC SERPINE1 GRIN1 PACSIN1 NEFM NEFL CAMK2A GABRA1 DDN VSNL1 TOP2A CARTPT TESPA1 CPLX2 SLC17A7 GABRB2 SV2B SYT1 GABRG2 CPXM1 CCK SYT4 SYN2 SLC12A5 NEUROD6 CBLN2 CHGA SNAP25 MAL2 KCNC2 UBE2C SYT13 NRGN S100G UNC13C TNFRSF12A OLFM3 PRKCG KCNV1 COL1A1 CREG2 SLC6A17 SPOCD1 TMEM130 INA GJB6 MPPED1 HCN1 C1QL3

ngenes: ○ 10 ● 50 ○ 100

CHAPTER

NINETEEN

SYSTEMSBIO

The final module of the platform is divided into three submodules: **Drug connectivity**, **Cell profiling** and **WGCNA**.

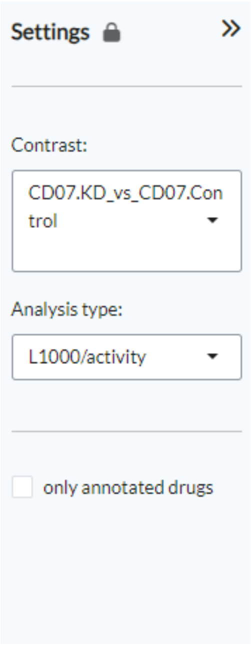
19.1 Drug connectivity

In the **Drug connectivity** submodule, users can correlate their signature with more than 5000 known drug profiles from the L1000 database, as well as with drug sensitivity profiles from the CTRP v2 and GDSC databases. Additionally, a separate list of shRNA- and cDNA-perturebed datasets from the L1000 database is also available (gene/L1000).

An activation-heatmap compares drug activation profiles across multiple contrasts. This facilitates to quickly see and detect the similarities between contrasts for certain drugs.

19.1.1 Settings panel

In the **Settings** panel, users can specify the contrast of their interest with the `Contrast` setting. Under `Analysis type` users can select from four databases, including the L1000 drug connectivity map (L1000/activity), the L1000 gene perturbation (L1000/gene) database, the CTRP v2 drug sensitivity (CTRP_v2/sensitivity) database and the GDSC drug sensitivity (GDSC/sensitivity) database (default: L1000/activity). The `only annotated drugs` option is used to exclude drugs without a known mechanism of action.



Settings

Contrast: CD07.KD_vs_CD07.Control

Analysis type: L1000/activity

□ only annotated drugs

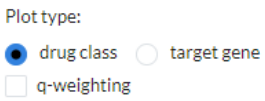
19.1.2 Drug enrichment

There are four main panels in the **Drug enrichment** tab:

Drug connectivity
The Drug Connectivity panel correlates your signature with profiles from the L1000 (activity/L1000 and gene/L1000), CTRP and GDSC databases. It shows the top N=12 similar and opposite profiles as GSEA plots by running the GSEA algorithm on the contrast-drug profile correlation space.

Enrichment table
Enrichment is calculated by correlating your signature with the profiles from the chosen database. Because of multiple perturbation experiments for a single small molecule, they are scored by running the GSEA algorithm on the contrast-small molecule profile correlation space. In this way, we obtain a single score for multiple profiles of a single small molecule. The table can be customised via the table *Settings* to only show annotated drugs.

Mechanism of action
This plot visualizes the mechanism of action (MOA) across the enriched drug profiles. On the vertical axis, the number of drugs with the same MOA are plotted. You can switch to visualize between MOA or target gene. Under the plots *Settings*, users can select the plot type of MOA analysis: by class description (`drug class`) or by target gene (`target gene`). They can also apply q-value weighting for NES scoe values (`q-weighting`).

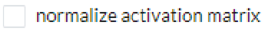


Plot type:

● drug class ○ target gene

□ q-weighting

Activation matrix
The **Activation matrix** visualizes the correlation of small molecule profiles with all available pairwise comparisons. The size of the circles correspond to the strength of their correlation, and are colored according to their positive (red) or negative (blue) correlation to the contrast profile. The matrix can be normalised via the settings icon by ticking the `normalize activation matrix` option.



□ normalize activation matrix

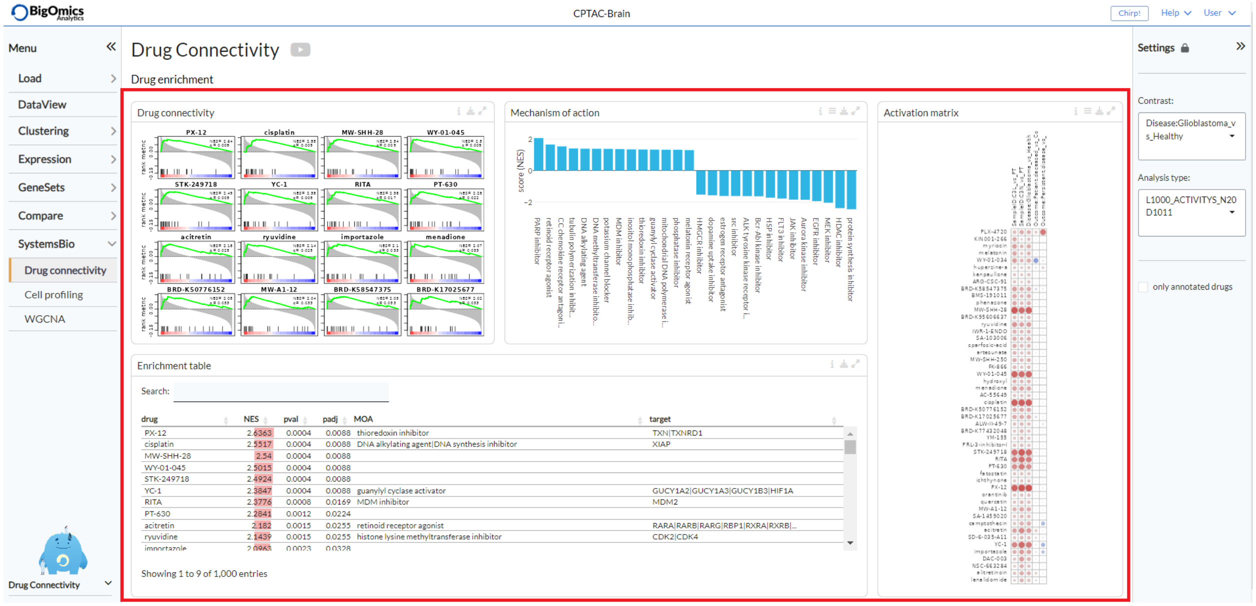
This tab can have many applications, which include understanding the MOA of a novel compund, identifying drugs that can be repurposed for treating a disease, identifying suitable partner drugs for the tested compound or target genes for intervention.

19.2 Cell Profiling

The **Cell Profiling** tab is specifically developed for the analysis and visualization of single-cell datasets. The main applications are identification of immune cell types and visualisations of markers, phenotypes, and proportions across the cells.

The **Cell type** tab infers the type of cells using computational deconvolution methods and reference datasets from the literature.

The **Mapping** tab provides a visualization of the inferred cell types matched to the phenotype variable of the data set, as well as a proportion plot visualizing the interrelationships between two categorical variables (so-called cross tabulation). This can be used to study the composition of a sample by cell type, for example.



Contrast: Disease:Glioblastoma_vs_Healthy

Analysis type: L1000_ACTIVITYS_N20D1011

Enrichment table

Search:

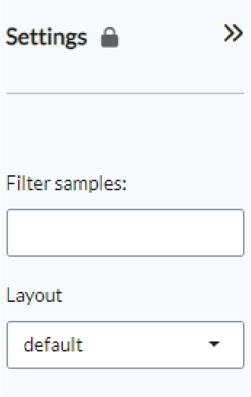
drug	NES	pval	padj	MOA	target
PX-12	2.6363	0.0004	0.0088	thioredoxin inhibitor	TXN|TXNRD1
cisplatin	2.5517	0.0004	0.0088	DNA alkylating agent|DNA synthesis inhibitor	XIAP
MW-SHH-28	2.54	0.0004	0.0088		
WY-01-045	2.5015	0.0004	0.0088		
STK-249718	2.4924	0.0004	0.0088		
YC-1	2.3847	0.0004	0.0088	guanylyl cyclase activator	GUCY1A2|GUCY1A3|GUCY1B3|HIF1A
RITA	2.3776	0.0008	0.0169	MDM inhibitor	MDM2
PT-630	2.2841	0.0012	0.0224		
acitretin	2.182	0.0015	0.0255	retinoid receptor agonist	RARA|RARB|RARG|RBP1|RXRA|RXRB|...
ryuvidine	2.1439	0.0015	0.0255	histone lysine methyltransferase inhibitor	CDK2|CDK4
importazole	2.0963	0.0023	0.0328		

Showing 1 to 9 of 1,000 entries

The **Markers** tab provides potential marker genes, which are the top genes with the highest standard deviation within the expression data across the samples. It also generates a plot mimicking the scatter plots used for gating in flow cytometry analysis.

19.2.1 Settings panel

Users can filter relevant samples in the `Filter samples` settings under the the main `Options` in the input panel. They can also specify the `layout` for the figures by chooisng between pca, tsne or umap options (default: tsne).



Filter samples:

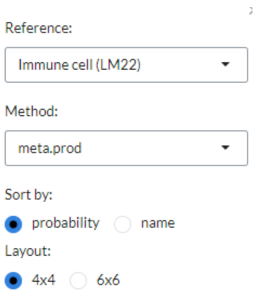
Layout

default

19.2.2 Cell type

The **Cell type** tab contains two panels: **Cell type profiling** and **Phenotypes**.

Cell type profiling infers the type of cells using computational deconvolution methods and reference datasets from the literature. In the plot settings menu, users can select the reference dataset and the method for the cell type prediction in the `reference` and `method` settings, respectively. Currently, we have implemented a total of 7 methods (EPIC, DeconRNAseq, DCQ, I-NNLS, NNLM, correlation-based and a meta-method) and 9 reference datasets to predict immune cell types (4 datasets: LM22, ImmProt, DICE and ImmunoStates), tissue types (2 datasets: HPA and GTEx), cell lines (2 datasets: HPA and CCLE) and cancer types (1 dataset: CCLE). Not all methods or databases may be available for a dataset, the availability depends on the pre-processing done. From the settings, users can also sort plots by either probability or name and change the layout (`sort by`).



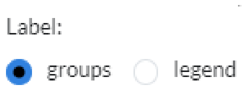
Reference: Immune cell (LM22)

Method: meta.prod

Sort by: ● probability ○ name

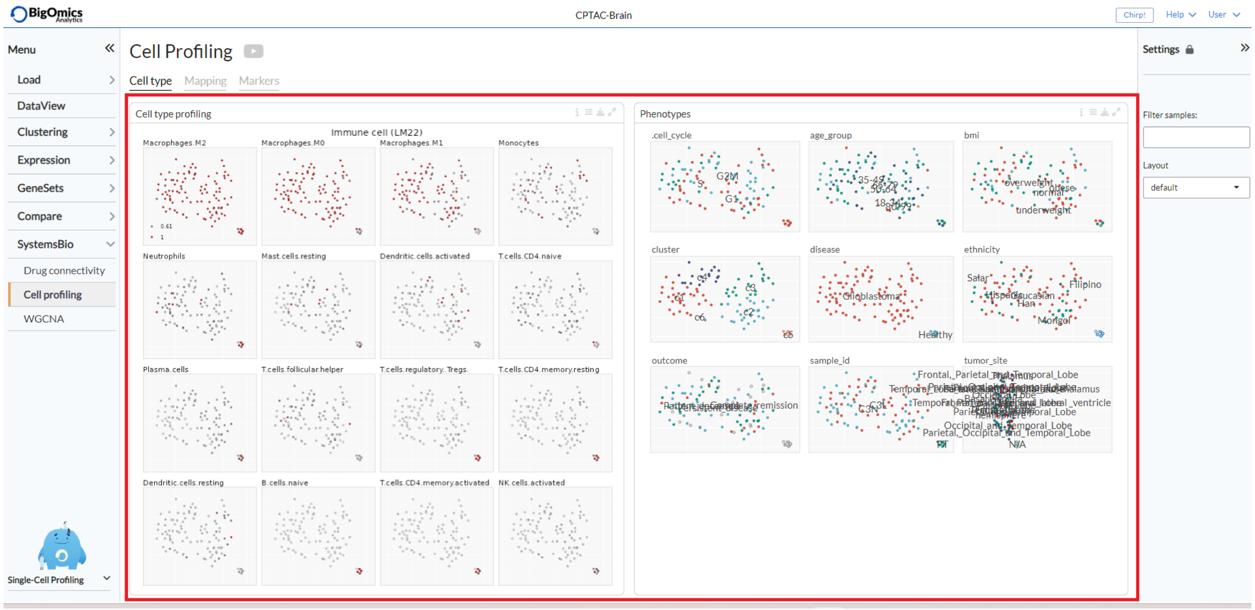
Layout: ● 4x4 ○ 6x6

The **Phenotypes** tab displays plots that show the distribution of the phenotypes superposed on the t-SNE clustering. Often, we can expect the t-SNE distribution to be driven by the particular phenotype that is controlled by the experimental condition or unwanted batch effects. Users can customise the plot via the settings icon, where they can `label` the plot groups or add a legend instead.



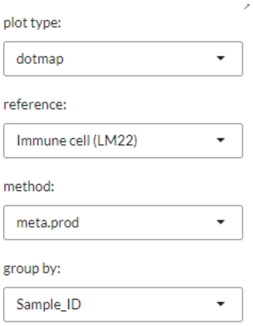
Label: ● groups ○ legend

The cell type profiling tab displays the two panels side by side.



19.2.3 Mapping

The **Mapping** panel contains two panels. The **Cell type mapping** panel contains a plot representing the cell type mapping across all samples. This plot can be customised via the *Settings* menu. Through it, users can change the `plot type` between a dotmap and a heatmap and select the `reference` dataset, select the analysis `method`. The reference datasets and the methods available are the same as indicated in the **Cell type profiling** panel under the **Cell type** tab. Users can also use `group by` to group samples by input phenotypes.



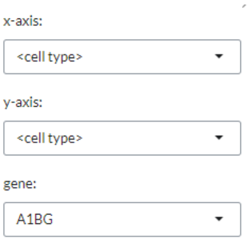
plot type: dotmap

reference: Immune cell (LM22)

method: meta.prod

group by: Sample_ID

The **Proportions** panel contains a proportion plot visualizes the overlap between two categorical variables. This may be useful for bulk RNA datasets, as it can provide information about the proportion of different cell types in the samples. From the settings icon, users can select whwther to display the <cell type> (based on the chosen reference dataset) or select one of the available phenotypes on the x- and y-axes of the plot. By selecting a gene with `gene` they can also add an expression barplot that indicates the expression level (high or low is based on average sample expression for the gene) of the selected gene for each of the sample groups as well as adding the total number of read counts of the selected gene per sample group.



x-axis: <cell type>

y-axis: <cell type>

gene: A1BG

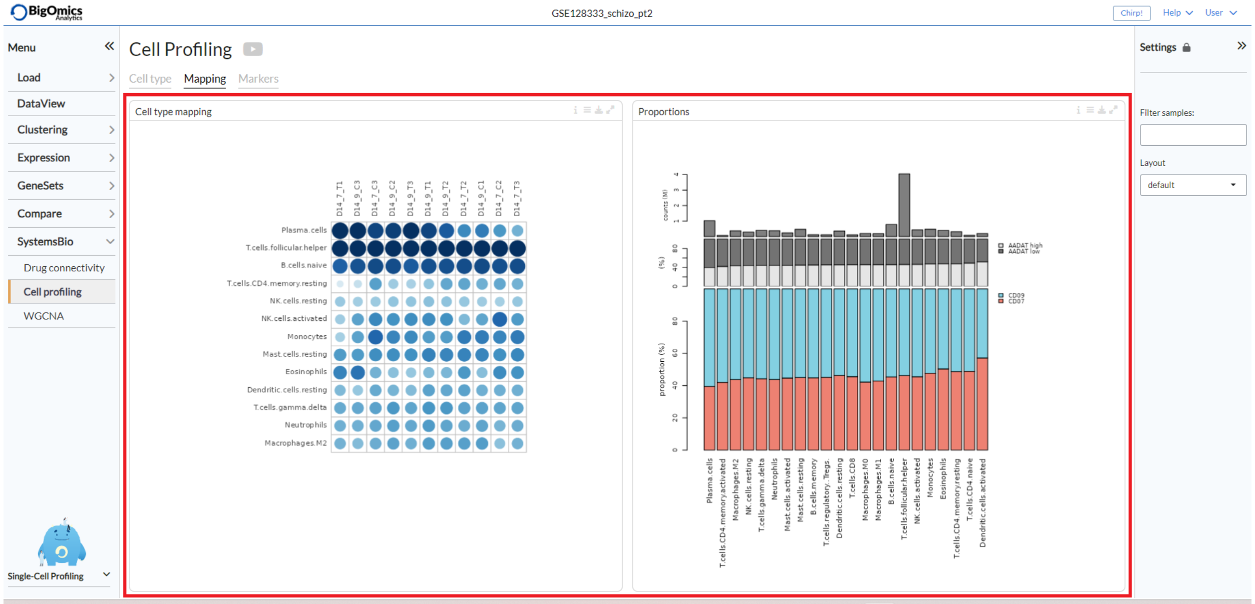
The two panels are displayed side by side in the tab.

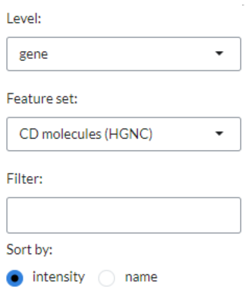
19.2.4 Markers

The **Markers** tab consists of two panels: **Expression of marker genes** and **Cytometry plot**.

Expression of marker genes consists of 25 t-SNE plots of the genes with the highest standard deviation that could represent potential biomarkers. The red color shading is proportional to the (absolute) expression of the gene in corresponding samples. In the settings icon, users can specify the `Level` of the marker analysis: gene or gene set level. They can also restrict the analysis by selecting a particular functional group in the `Feature set`, where genes are divided into 89 groups, such as chemokines, transcription factors, genes involved in immune checkpoint inhibition, and so on (default: CD molecules (HGNC)). In addition, it is possible to filter markers by a specific keywords in the `Filter` setting and sort them by intensity (default) or name (`sort by`).

For each gene pairs combination, the panel also generates a cytometry-like plot (**Cyto plot**) of samples. The aim of this feature is to observe the distribution of samples in relation to the selected gene pairs. For instance, when applied to single-cell sequencing data from immunological cells, it can mimic flow cytometry analysis and distinguish T helper cells from other T cells by selecting the CD4 and CD8 gene combination. Under the plot settings icon, users can select





Level:

gene

Feature set:

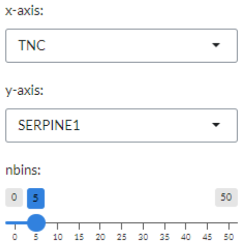
CD molecules (HGNC)

Filter:

Sort by:

● intensity ○ name

their prefered genes on the x- and y-axes in the `x-axis` and `y-axis`, respectively. They can also set the maximum number of bins for histgram distribution (`nbins`). This will be used to calculate the density distribution of the gene pairs selected in the `x-axis` and `y-axis`.

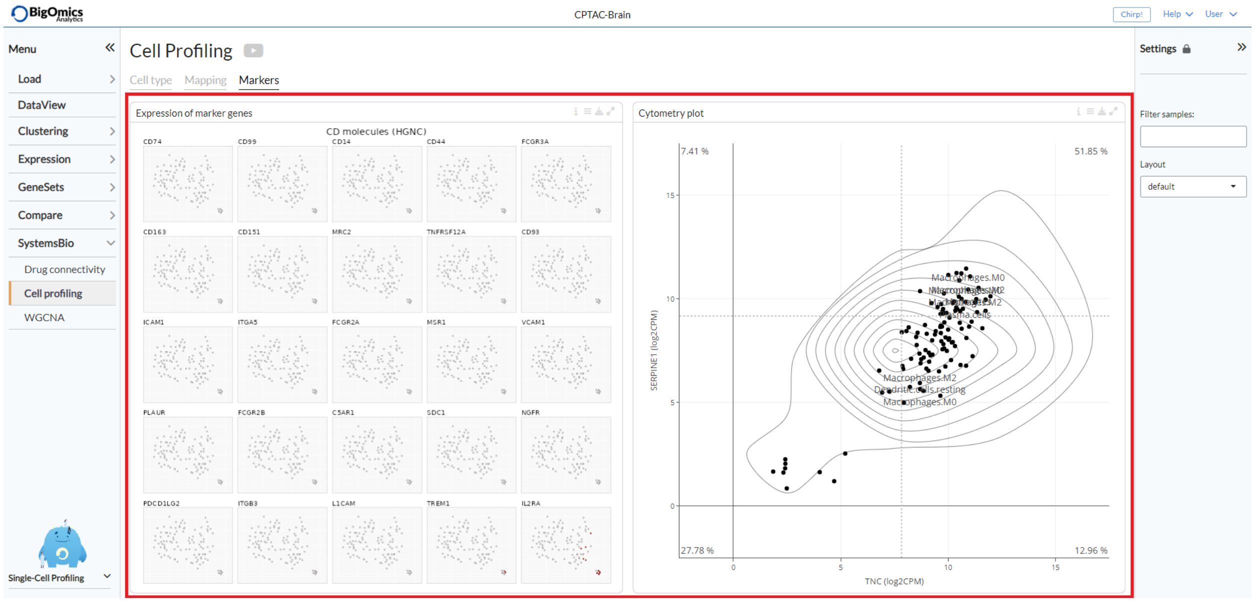


x-axis: TNC

y-axis: SERPINE1

nbins: 5

The two panels are displayed side by side in the tab.

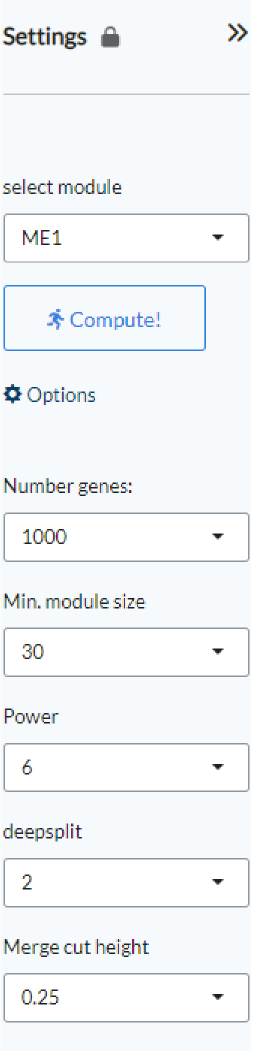


19.3 WGCNA

The final submodule under **SystemsBio** is dedicated to weighted correlation network analysis (**WGCNA**), which serves the purpose of identifying clusters (modules) comprising highly correlated genes. These clusters can be summarized using either the module eigengene or an intramodular hub gene. WGCNA also facilitates the association of modules with each other and external sample traits through eigengene network methodology. Furthermore, it allows for the computation of module membership measures.

19.3.1 Settings panel

WGCNA modules are selected from the **Settings** panel (`select module`) and plots can be recalcluated based on the selected module. Under *Options*, the number of genes (`Number genes`, default=1000), the miinum module size (`Min. module size`, default=30), the `Power` (default=6), the `deepsplit` (default=2) and the `Merge cut height`, default=0.25) can be set.



select module: ME1

Number genes: 1000

Min. module size: 30

Power: 6

deepsplit: 2

Merge cut height: 0.25

19.3.2 WGCNA

The **WGCNA** tab consists of fiive panels (from left to right and top to bottom): **Gene dendrogram and gene modules**, **Scale independence and mean connectivity**, **TOM heatmap**, **Gene clustering** and **Module graph**.

Gene dendrogram and gene modules

In this panel, gene modules are detected as branches of the resulting cluster tree using the dynamic branch cutting approach. Genes inside a given module are summarized with the module eigengene. The module eigengene of a given module is defined as the first principal component of the standardized expression profiles.

Scale independence and mean connectivity

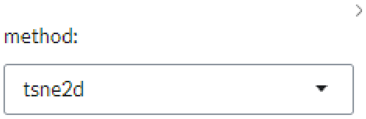
This panel is used for the the analysis of network topology for various soft-thresholding powers. The left plot shows the scale-free fit index (y-axis) as a function of the soft-thresholding power (x-axis). The right plot displays the mean connectivity (degree, y-axis) as a function of the soft-thresholding power (x-axis).

TOM heatmap

The panel displays the Topological Overlap Matrix (TOM) as a heatmap, which shows the correlation among gene module memberships.

Gene clustering

Dimensionality reduction maps colored by WGCNA module. Via the settings icon, the layout can be changed between tsne (default), pca and umap.

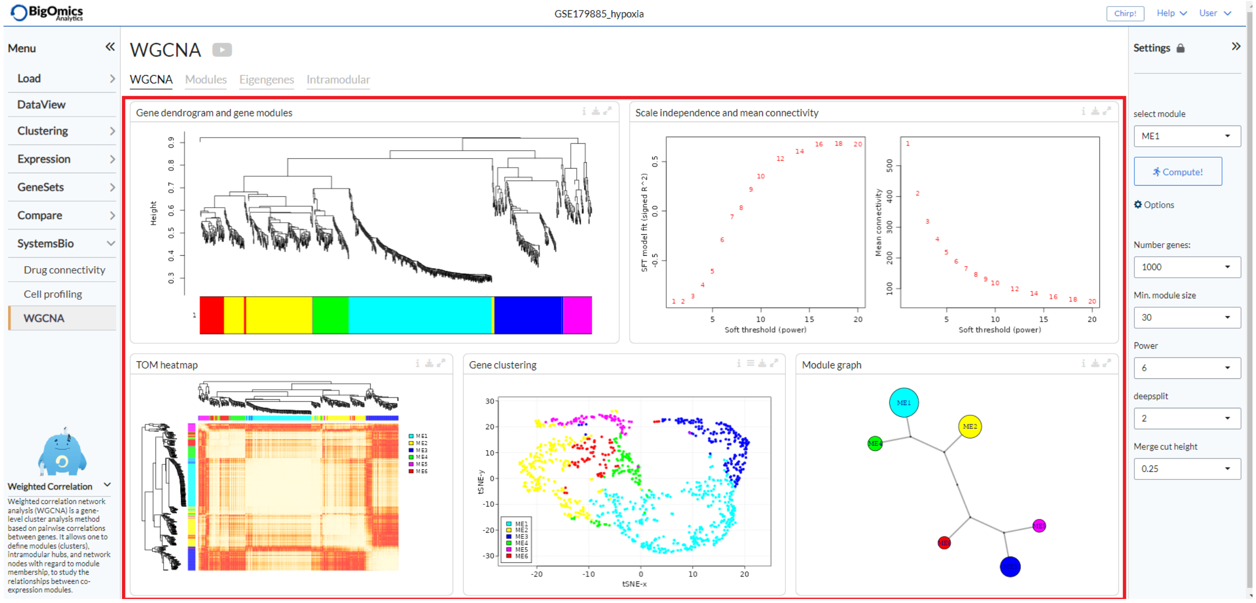


method:

tsne2d

Module graph

The final panel contains the raph network of WGCNA modules, which represents the relationship betweem of the gene modules.



Weighted Correlation

Weighted correlation network analysis (WGCNA) is a gene-level cluster analysis method based on pairwise correlations between genes. It allows one to define modules (clusters), intramodular hubs, and network nodes with regard to module membership, to study the relationships between co-expression modules.

select module: ME1

Number genes: 1000

Min. module size: 30

Power: 6

deepsplit: 2

Merge cut height: 0.25

19.3.3 Modules

The **Modules** tab contains five panels (left to right, top to bottom): **Module-Trait relationships**, **Correlation network**, **Module Enrichment (plot)**, **Module genes** and **Module enrichment (table)**.

Module-Trait relationships

Module-trait analysis identifies modules that are significantly associated with the measured clinical traits by quantifying the association as the correlation of the eigengenes with external traits. The relationships between the various WGCNA modules and the phenotypic groups in the dataset are displayed as a heatmap, with shades of red indicating a negative correlation and shades of green indicating a positive correlation. The continuous variables can be binarised via the settings icon (`binarize continuous vars`).

□ binarize continuous vars

Correlation network

A partial correlation graph centered on module eigengene with top most correlated features. Green edges correspond to positive (partial) correlation, red edges to negative (partial) correlation. Width of the edges is proportional to the correlation strength of the gene pair. The regularized partial correlation matrix is computed using the 'graphical lasso' (Glasso) with BIC model selection.

Module Enrichment Plot

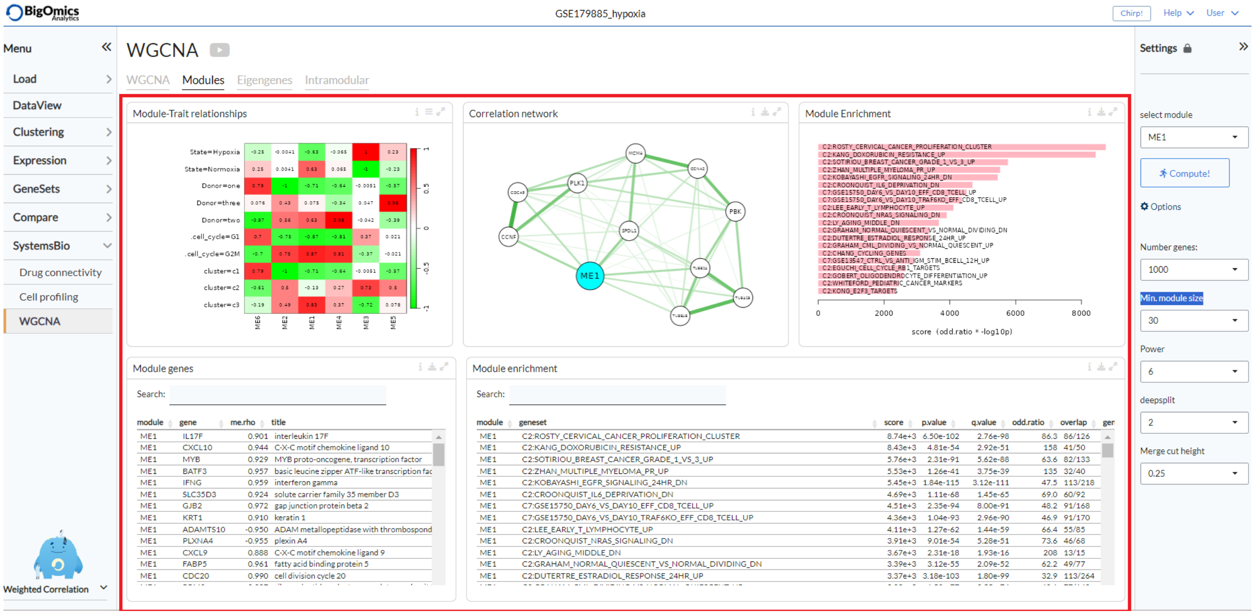
A plot that displays the functional enrichment of top most enriched genesets from a selection of GO and MSigDB genesets.

Module genes

A table showing the genes in the WGCNA module selected via the **Settings** panel. The value me.rho represents the correlation between the gene expression and the module.

Module Enrichment Table

Functional enrichment of the module calculated using Fisher's exact test. In this table, users can check mean expression values of features across the conditions for the selected module.



BigOmics Analytics
GSE179885_hypoxia
Chirp!
Help
User
Menu
Load
DataView
Clustering
Expression
GeneSets
Compare
SystemsBio
Drug connectivity
Cell profiling
WGCNA
Weighted Correlation
WGCNA
WGCNA
Modules
Eigengenes
Intramodular
Module-Trait relationships
Correlation network
CDCA8
PLK1
MCM4
CCNA2
PBK
CCNF
SPDL1
ME1
TUBB2A
TUBA1B
TUBB2B
Module Enrichment
C2:ROSTY_CERVICAL_CANCER_PROLIFERATION_CLUSTER
C2:KANG_DOXORUBICIN_RESISTANCE_UP
C2:SOTIRIOU_BREAST_CANCER_GRADE_1_VS_3_UP
C2:ZHAN_MULTIPLE_MYELOMA_PR_UP
C2:KOBAYASHI_EGFR_SIGNALING_24HR_DN
C2:CROONQUIST_IL6_DEPRIVATION_DN
C7:GSE15750_DAY6_VS_DAY10_EFF_CD8_TCELL_UP
C7:GSE15750_DAY6_VS_DAY10_TRAF6KO_EFF_CD8_TCELL_UP
C2:LEE_EARLY_T_LYMPHOCYTE_UP
C2:CROONQUIST_NRAS_SIGNALING_DN
C2:LY_AGING_MIDDLE_DN
C2:GRAHAM_NORMAL_QUIESCENT_VS_NORMAL_DIVIDING_DN
C2:DUTERTRE_ESTRADIOL_RESPONSE_24HR_UP
C2:GRAHAM_CML_DIVIDING_VS_NORMAL_QUIESCENT_UP
C2:CHANG_CYCLING_GENES
C7:GSE13547_CTRL_VS_ANTI_IGM_STIM_BCELL_12H_UP
C2:EGUCHI_CELL_CYCLE_RB1_TARGETS
C2:GOBERT_OLIGODENDROCYTE_DIFFERENTIATION_UP
C2:WHITEFORD_PEDIATRIC_CANCER_MARKERS
C2:KONG_E2F3_TARGETS
0
2000
4000
6000
8000
score (odd.ratio * -log10p)
1
0.5
0
-0.5
-1
Settings
Compute!
Options

Module-Trait relationships

	ME6	ME2	ME1	ME4	ME3	ME5
State=Hypoxia	-0.25	-0.0041	-0.63	-0.065	1	0.23
State=Normoxia	0.25	0.0041	0.63	0.065	-1	-0.23
Donor=one	0.79	-1	-0.71	-0.64	-0.0051	-0.57
Donor=three	0.076	0.43	0.075	-0.34	0.047	0.96
Donor=two	-0.87	0.56	0.63	0.98	-0.042	-0.39
.cell_cycle=G1	0.7	-0.75	-0.87	-0.81	0.37	0.021
.cell_cycle=G2M	-0.7	0.75	0.87	0.81	-0.37	-0.021
cluster=c1	0.79	-1	-0.71	-0.64	-0.0051	-0.57
cluster=c2	-0.61	0.5	-0.13	0.27	0.73	0.5
cluster=c3	-0.19	0.49	0.83	0.37	-0.72	0.078

select module: ME1

Number genes: 1000

Min. module size: 30

Power: 6

deepsplit: 2

Merge cut height: 0.25

Module genes

Search:

module	gene	me.rho	title
ME1	IL17F	0.901	interleukin 17F
ME1	CXCL10	0.944	C-X-C motif chemokine ligand 10
ME1	MYB	0.929	MYB proto-oncogene, transcription factor
ME1	BATF3	0.957	basic leucine zipper ATF-like transcription fac
ME1	IFNG	0.959	interferon gamma
ME1	SLC35D3	0.924	solute carrier family 35 member D3
ME1	GJB2	0.972	gap junction protein beta 2
ME1	KRT1	0.910	keratin 1
ME1	ADAMTS10	-0.950	ADAM metallopeptidase with thrombospond
ME1	PLXNA4	-0.955	plexin A4
ME1	CXCL9	0.888	C-X-C motif chemokine ligand 9
ME1	FABP5	0.961	fatty acid binding protein 5
ME1	CDC20	0.990	cell division cycle 20

Module enrichment

Search:

module	geneset	score	p.value	q.value	odd.ratio	overlap	gen
ME1	C2:ROSTY_CERVICAL_CANCER_PROLIFERATION_CLUSTER	8.74e+3	6.50e-102	2.76e-98	86.3	86/126	
ME1	C2:KANG_DOXORUBICIN_RESISTANCE_UP	8.43e+3	4.81e-54	2.92e-51	158	41/50	
ME1	C2:SOTIRIOU_BREAST_CANCER_GRADE_1_VS_3_UP	5.76e+3	2.31e-91	5.62e-88	63.6	82/133	
ME1	C2:ZHAN_MULTIPLE_MYELOMA_PR_UP	5.53e+3	1.26e-41	3.75e-39	135	32/40	
ME1	C2:KOBAYASHI_EGFR_SIGNALING_24HR_DN	5.45e+3	1.84e-115	3.12e-111	47.5	113/218	
ME1	C2:CROONQUIST_IL6_DEPRIVATION_DN	4.69e+3	1.11e-68	1.45e-65	69.0	60/92	
ME1	C7:GSE15750_DAY6_VS_DAY10_EFF_CD8_TCELL_UP	4.51e+3	2.35e-94	8.00e-91	48.2	91/168	
ME1	C7:GSE15750_DAY6_VS_DAY10_TRAF6KO_EFF_CD8_TCELL_UP	4.36e+3	1.04e-93	2.96e-90	46.9	91/170	
ME1	C2:LEE_EARLY_T_LYMPHOCYTE_UP	4.11e+3	1.27e-62	1.44e-59	66.4	55/85	
ME1	C2:CROONQUIST_NRAS_SIGNALING_DN	3.91e+3	9.01e-54	5.28e-51	73.6	46/68	
ME1	C2:LY_AGING_MIDDLE_DN	3.67e+3	2.31e-18	1.93e-16	208	13/15	
ME1	C2:GRAHAM_NORMAL_QUIESCENT_VS_NORMAL_DIVIDING_DN	3.39e+3	3.12e-55	2.09e-52	62.2	49/77	
ME1	C2:DUTERTRE_ESTRADIOL_RESPONSE_24HR_UP	3.37e+3	3.18e-103	1.80e-99	32.9	113/264	

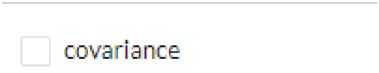
19.3.4 Eigengenes

The **Eigengenes** tab is used to visualise the network of eigengenes and study the relationships among the found modules. One can use the eigengenes as represetative profiles and quantify module similarity by eigengene correlation. For each module, we also define a quantitative measure of 'module membership' (MM) as the correlation of the module eigengene and the gene expression profile. This allows us to quantify the similarity of all genes to every module.

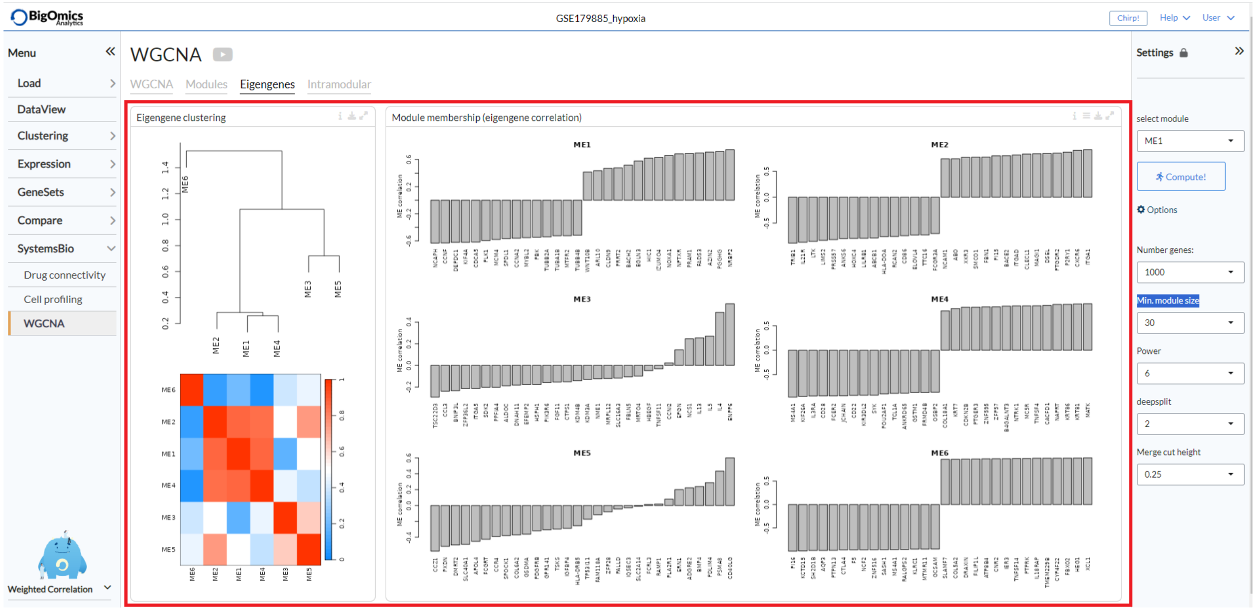
The tab contains two panels: **Eigengene clustering** and **Module membership (eigengene correlation)**.

Eigengene clustering
: A cluster heatmap that shows the relationship between the different modules produced by the platform.

Module membership (eigengene correlation)
: This panels contains a series of plots for each one of the generated modules. For each module, we define a quantitative measure of module membership (MM) as the correlation of the module eigengene and the gene expression profile. This allows us to quantify the similarity of all genes on the array to every module. Users can also select to include the `covariance` for each gene (default:off).



□ covariance



select module: ME1

Number genes: 1000

Min. module size: 30

Power: 6

deepsplit: 2

Merge cut height: 0.25

19.3.5 Intramodular

The **Intramodular** tab is used to quantify associations of individual genes with the trait of interest (weight) by defining Gene Significance (GS) as (the absolute value of) the correlation between the gene and the trait. For each module, a quantitative measure of module membership (MM) as the correlation of the module eigengene and the gene expression profile is also defined. Using the GS and MM measures, users can identify genes that have a high significance for weight as well as high module membership in interesting modules.

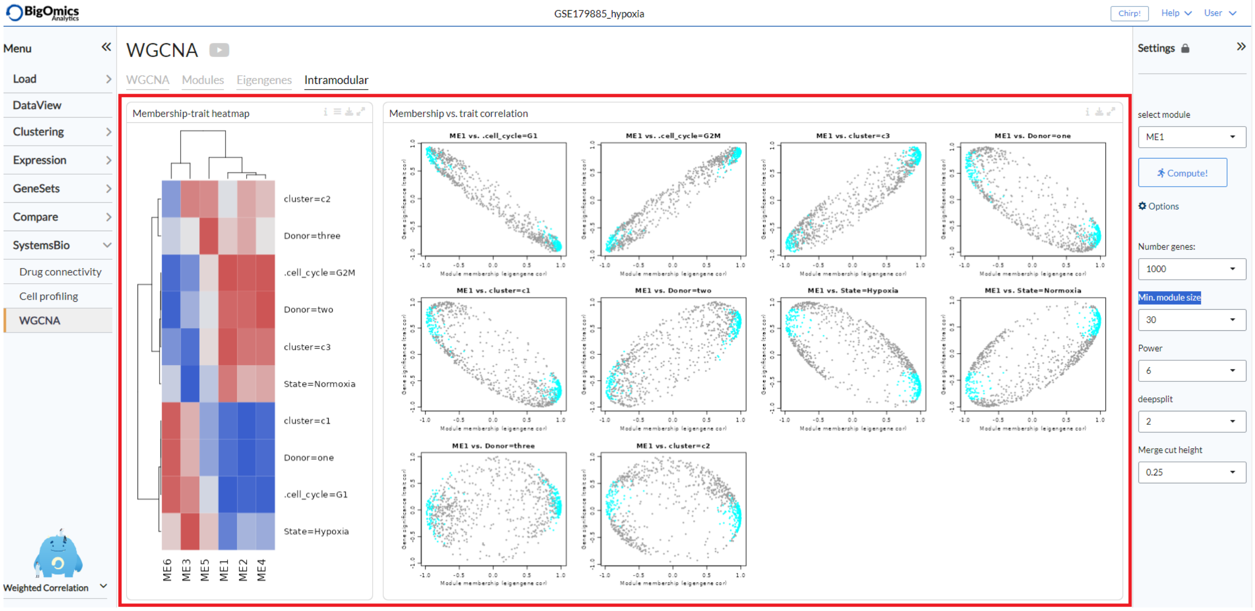
The tab contains two panels: **Membership-trait heatmap** and **Membership vs. trait correlation**.

Membership-trait heatmap

For each module, a quantitative measure of module membership (MM) as the correlation of the module eigengene and the gene expression profile is defined. This allows us to quantify the similarity of all genes on the array to every module and represent them as a heatmap.

Membership vs. trait correlation

In this panel, the MM and trai correlations are represented as a series of scatterplots.



select module: ME1

Number genes: 1000

Min. module size: 30

Power: 6

deepsplit: 2

Merge cut height: 0.25

CHAPTER

TWENTY

REANALYZING PUBLIC DATASETS

To illustrate the use case of the Omics Playground, we reanalyzed different types of publics datasets, including microarray, bulk RNA-seq, single-cell RNA-seq and proteomic datasets to recapitulate the results.

20.1 Single-cell RNA-seq data

For single-cell RNA-seq data, we downloaded the melanoma data set GSE72056 of Tirosh et al.. Our platform recapitulates well the original findings of the paper. The t-SNE clustering (*Figure 1*) separates the different cell types. *Figure 2* and *Figure 3* show the volcano plot, MA plot and most differentially expressed genes between malignant and non-malignant cells. The CNV map (*Figure 4*) confirms the major chromosomal copy number variations found in the malignant cells. *Figure 5* shows high enrichment of a immune checkpoint signature, particularly concentrated in the T cells. The biomarker heatmap (*Figure 6*) highlights the marker genes for each cell type. Each gene cluster is furthermore automatically annotated with the most correlated gene sets (*Figure 7*).

20.1.1 tSNE plot

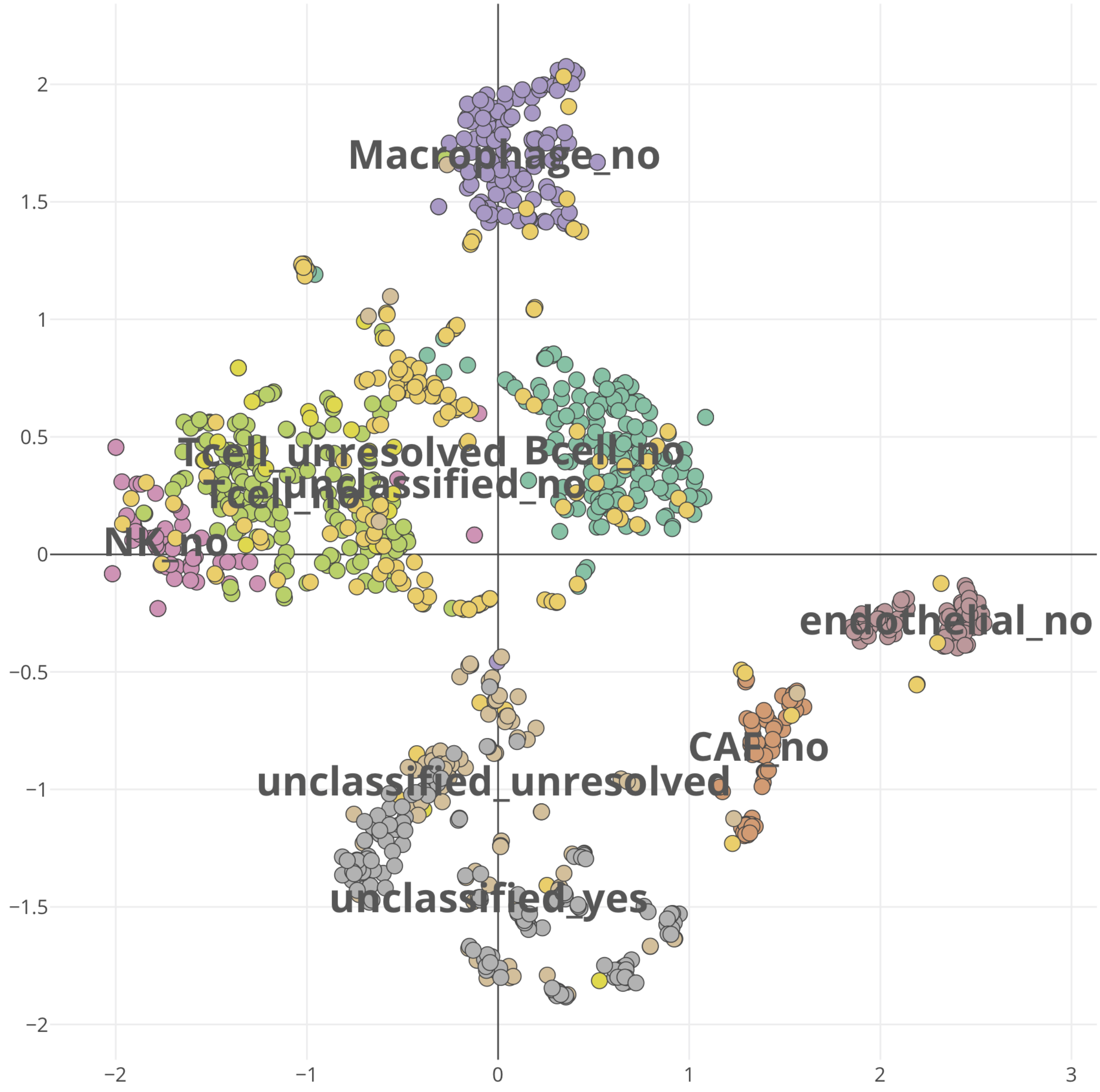
Figure 1. The t-SNE clustering with cell type annotation for the `GSE72056-scmelanoma` dataset. To reproduce the figure on the platform, select and load `GSE72056-scmelanoma` dataset, and go to the **PCA/tSNE** panel of the **Clustering** module. From the plot *Settings*, set the `color: group`, `layout: tsne`, and leave other settings as default.

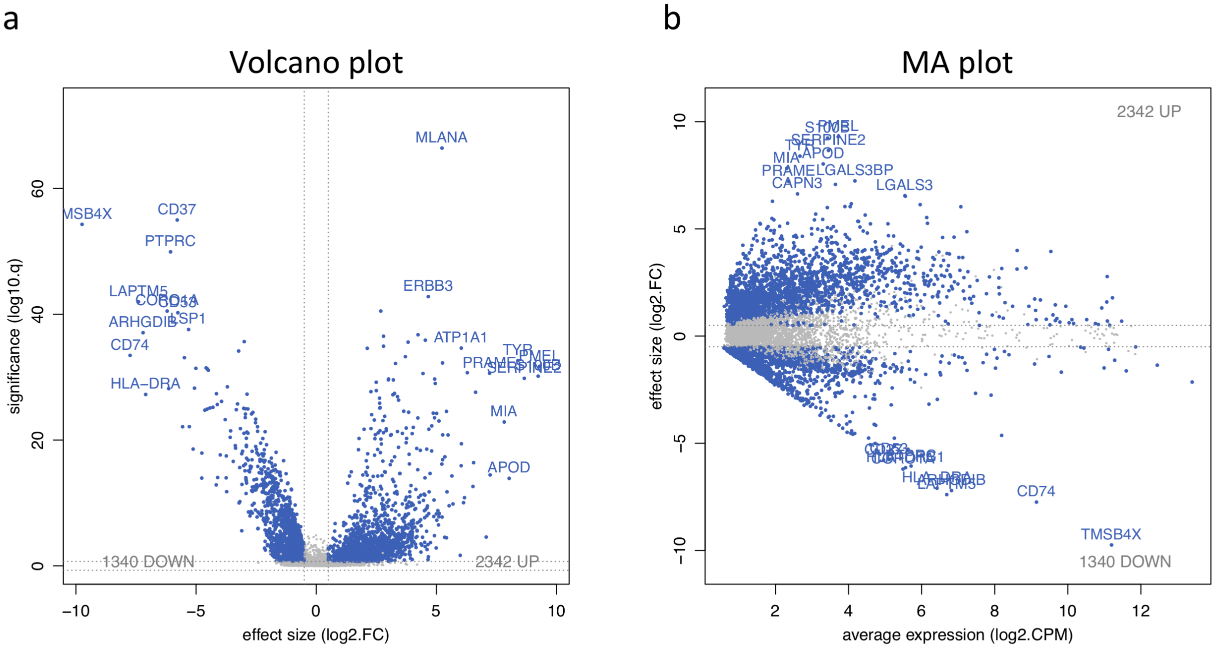
20.1.2 Volcano and MA plot

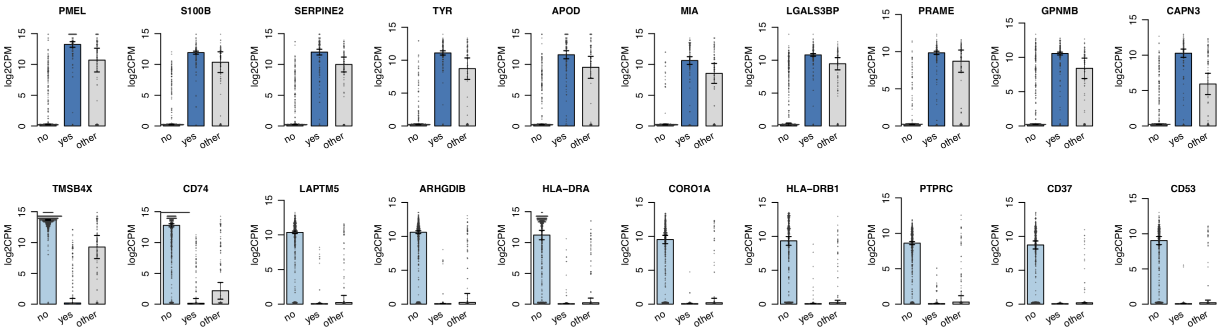
Figure 2. Volcano and MA plot for the malignant versus non-malignant contrast. To replicate the figure on the platform, go to the **Plots** panel of the **Expression** module. From the input slider, set the `Contrast: yes_vs_no` and `Gene family: all`.

20.1.3 Differentially expressed genes

Figure 3. Barplot of corresponding differentially expressed genes. To obtain the figure on the platform, go to the **Top genes** panel of the **Expression** module. From the input slider, set the `Contrast: yes_vs_no` and `Gene family: all`.


Macrophage_no
Tcell_unresolved
Bcell_no
unclassified_no
Tcell_no
NK_no
endothelial_no
CAF_no
unclassified_unresolved
unclassified_yes
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2
−1
0
1
2
3


a
Volcano plot
MLANA
MSB4X
CD37
PTPRC
ERBB3
LAPTM5
CORO1A
CD53
ARHGDIB
LSP1
ATP1A1
CD74
TYR
PMEL
PRAME
S100B
SERPINE2
HLA-DRA
MIA
APOD
1340 DOWN
2342 UP
significance (log10.q)
effect size (log2.FC)
0
20
40
60
−10
−5
0
5
10
b
MA plot
2342 UP
S100B
PMEL
SERPINE2
TYR
MIA
APOD
PRAME
LGALS3BP
CAPN3
LGALS3
CD53
CORO1A
PTPRC
HLA-DRB1
HLA-DRA
ARHGDIB
LAPTM5
CD74
TMSB4X
1340 DOWN
effect size (log2.FC)
average expression (log2.CPM)
10
5
0
−5
−10
2
4
6
8
10
12


PMEL
S100B
SERPINE2
TYR
APOD
MIA
LGALS3BP
PRAME
GPNMB
CAPN3
TMSB4X
CD74
LAPTM5
ARHGDIB
HLA-DRA
CORO1A
HLA-DRB1
PTPRC
CD37
CD53
log2CPM
0
5
10
15
no
yes
other

20.1.4 Inferred copy number

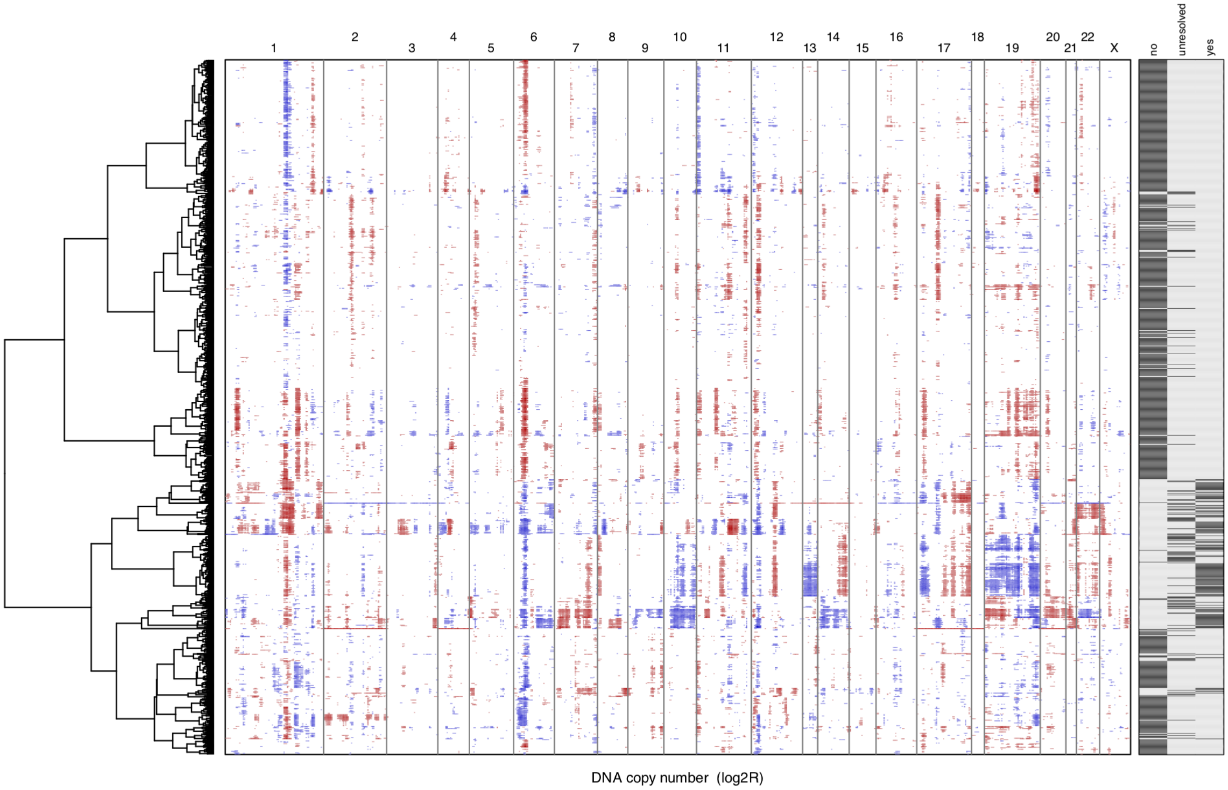


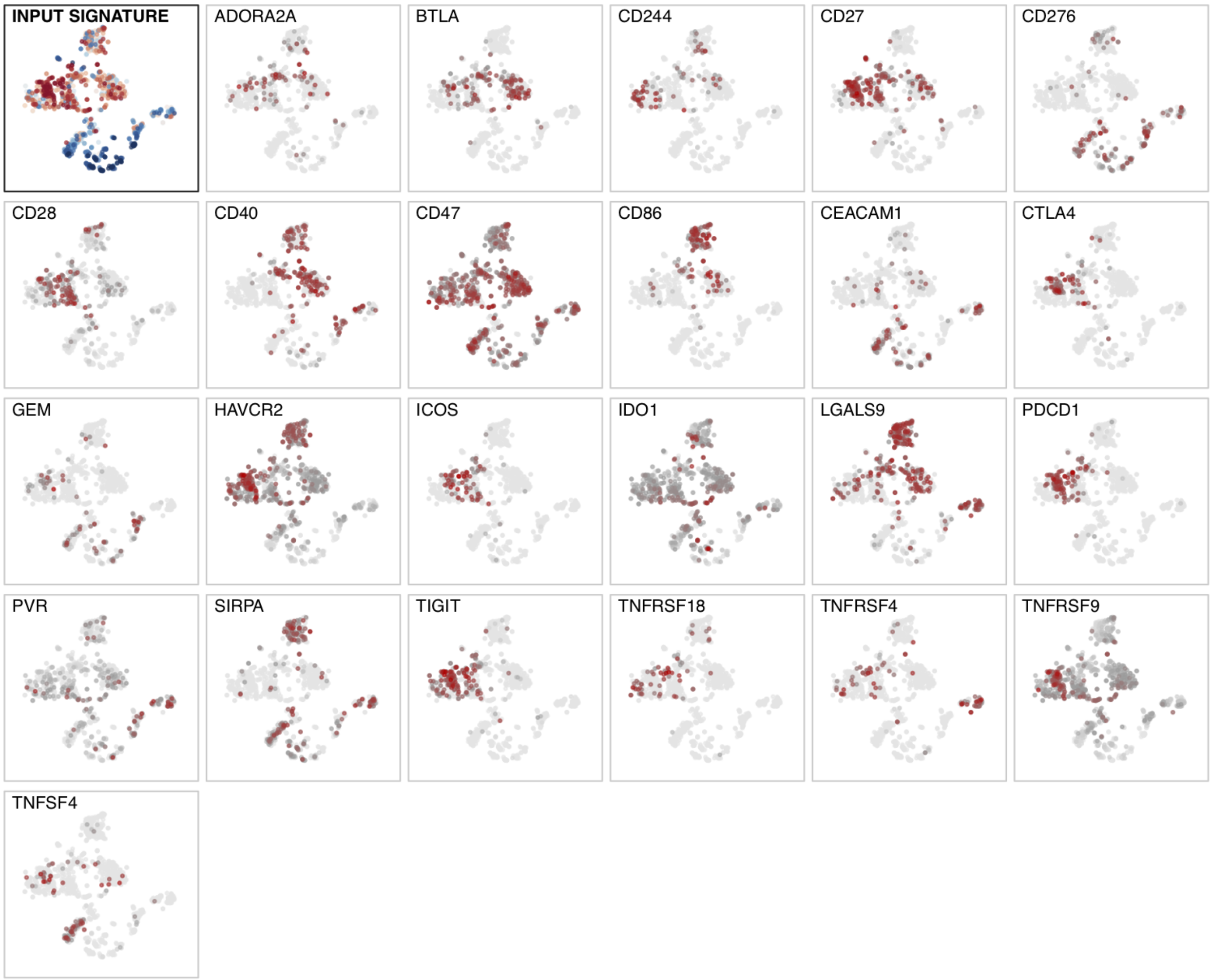
Figure 4. Inferred copy number for sample Cy80. To reproduce the figure on the platform, go to the **CNV** panel of the **scProfiling** module. From the plot *Settings*, set the `Annotate with: malignant` and `Order samples by: clust`.

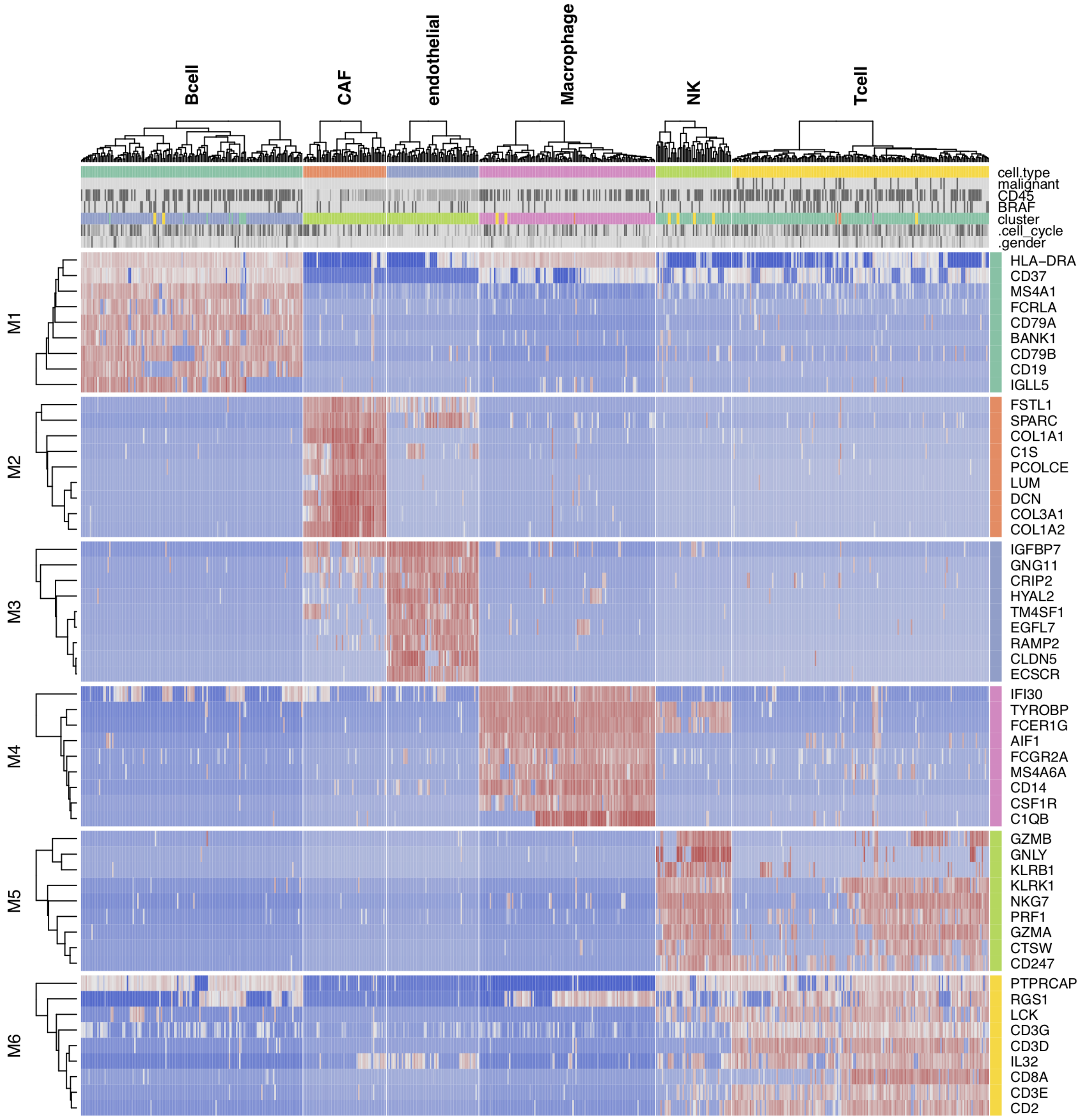
20.1.5 Immune checkpoint signature

Figure 5. Enrichment distribution for an immune checkpoint signature showing high enrichment in T and B cells . To regenerate the figure on the platform, go to the **Marker** panel in the **Signature** module. From the input slider, select `Contrast: custom` and `Signature: immune_chkpt` as it is provided in the sample list.

20.1.6 Biomarker heatmap

Figure 6. Biomarker heatmap for non-malignant cells. To reproduce the figure on the platform, go to the **Heatmap** panel in the **Clustering** module. From the input slider, set the `Filter samples: cell.type={Bcell, CAF, endothelial, Macrophage, NK, Tcell}`. In the plot *Settings*, set `Plot type: ComplexHeatmap`, `split by: cell.type`, and `top mode: specific`.


INPUT SIGNATURE
ADORA2A
BTLA
CD244
CD27
CD276
CD28
CD40
CD47
CD86
CEACAM1
CTLA4
GEM
HAVCR2
ICOS
IDO1
LGALS9
PDCD1
PVR
SIRPA
TIGIT
TNFRSF18
TNFRSF4
TNFRSF9
TNFSF4


Bcell
CAF
endothelial
Macrophage
NK
Tcell
cell.type
malignant
CD45
BRAF
cluster
.cell_cycle
.gender
M1
M2
M3
M4
M5
M6
HLA−DRA
CD37
MS4A1
FCRLA
CD79A
BANK1
CD79B
CD19
IGLL5
FSTL1
SPARC
COL1A1
C1S
PCOLCE
LUM
DCN
COL3A1
COL1A2
IGFBP7
GNG11
CRIP2
HYAL2
TM4SF1
EGFL7
RAMP2
CLDN5
ECSCR
IFI30
TYROBP
FCER1G
AIF1
FCGR2A
MS4A6A
CD14
CSF1R
C1QB
GZMB
GNLY
KLRB1
KLRK1
NKG7
PRF1
GZMA
CTSW
CD247
PTPRCAP
RGS1
LCK
CD3G
CD3D
IL32
CD8A
CD3E
CD2

20.1.7 Annotate heatmap clusters

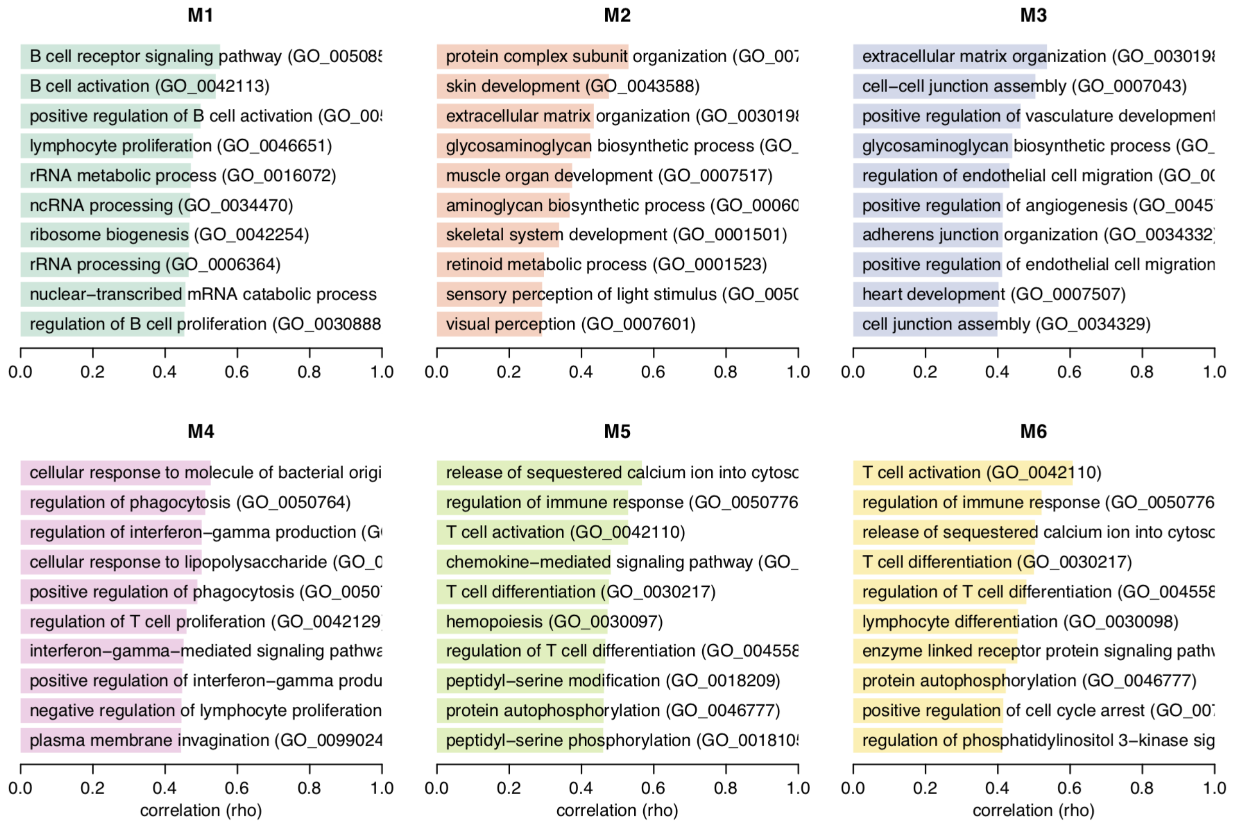


Figure 7. Enrichment annotation of corresponding heatmap clusters from the *Figure 6*. To reproduce the figure on the platform, generate the heatmap in *Figure 6* first, then go to the **Annotate clusters** panel. From the plot *Settings*, set the `Reference set: GOBP`.

20.2 Bulk RNA-seq Data

To elucidate the mechanism of action of a new drug, or for the intention of drug repurposing, it is often useful to find other drugs that have similar or opposing signatures compared to some given fold change profile. This example is illustriated using the RNA-sequencing dataset GSE114716 from Goswami et al., which contains CD4 T cells following ipilimumab therapy. *Figure 8* shows the top ranked drugs with most similar or most opposing signatures to ipilimumab, a novel monoclonal antibody targeting CTLA-4 used in tumour therapy to stimulate the immune system. The complete list contains several compounds that stimulate the immune system, such as alpha-tocopherol (Morel et al.), but also highlights compounds that are not commonly associated with the modulation of immune responses, such as strophanthidin, an intropic drug that has recently been shown to display pro-inflammatory activities (Karas et al.).

20.2.1 Drug enrichment profiles

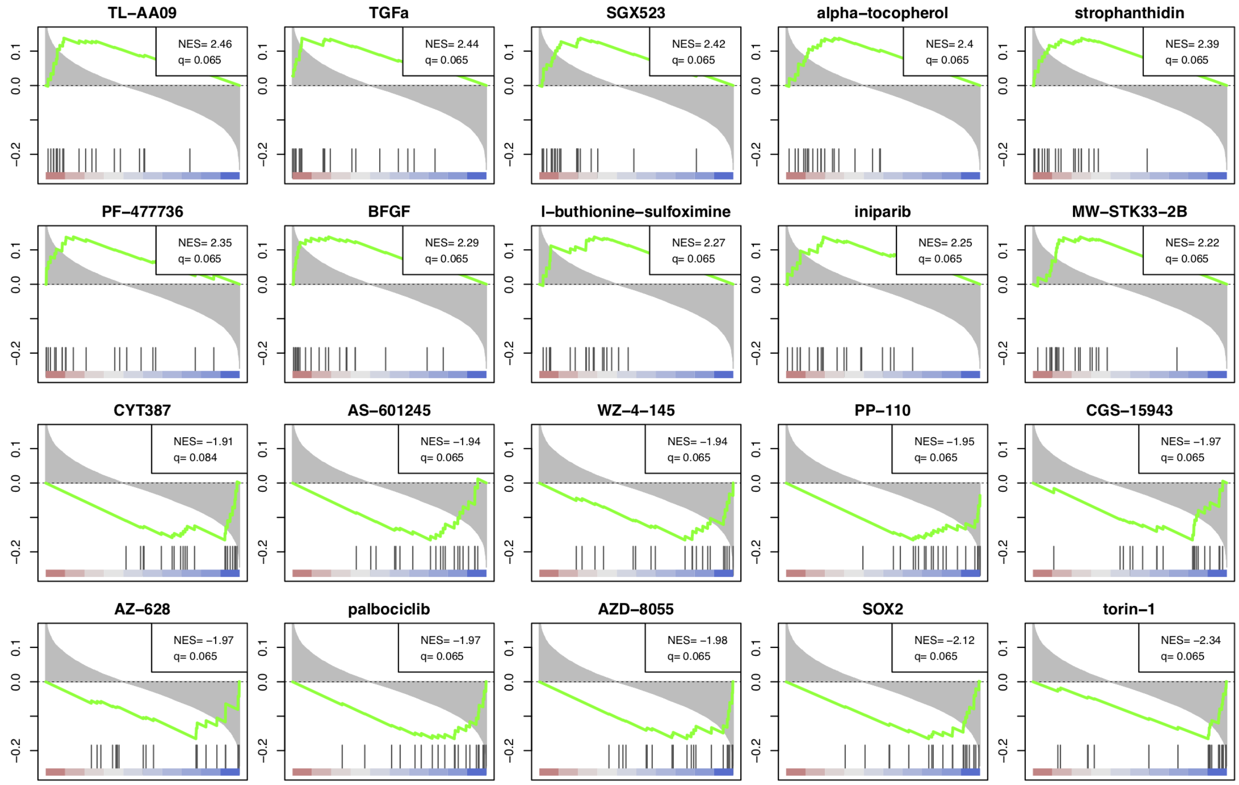


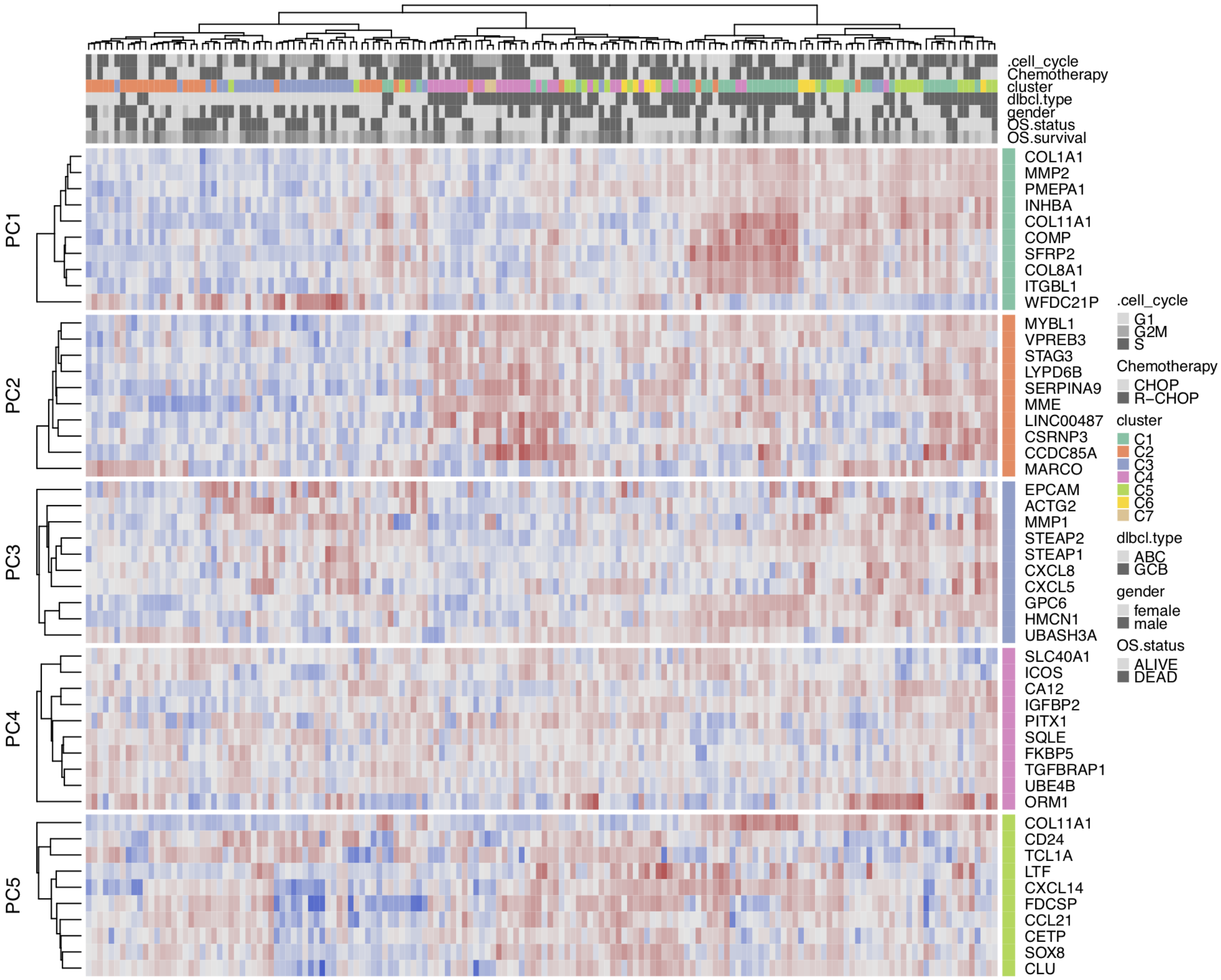
Figure 8. Drug enrichment profiles for most similar and opposing drugs compared to ipilimumab treatment. To reobtain the figure on the platform, select and load `GSE114716-ipilimumab` dataset, go to the **Drug CMap** panel under the **Functional** module, and set the `Contrast:  Ipi_vs_baseline` from the plot *Settings*.

20.3 Microarray Data

In this section, we perform the heatmap clustering, biomarker selection and survival analysis using the GSE10846 from Lenz et al., which is the microarray gene expression dataset of diffuse large B-cell lymphoma (DLBCL) patients. *Figure 9* shows a hierarchical cluster heatmap of microarray gene expression data. *Figure 10* and *Figure 11* show the variable importance plot and a survival tree on the overall survival of the DLBCL patients, respectively.

20.3.1 Hierarchical cluster heatmap

Figure 9. Hierarchical cluster heatmap for `GSE10846-dlbcl` dataset. To replicate the figure, select and load `GSE10846-dlbcl` dataset on the platform. Go to the **Heatmap** panel of the **Clustering** module, and set the `Level: gene` and `Features:  all` from the input panel. In the plot *Settings*, set the `split by:  none` and `Top mode: pca`.


.cell_cycle
Chemotherapy
cluster
dlbcl.type
gender
OS.status
OS.survival
PC1
COL1A1
MMP2
PMEPA1
INHBA
COL11A1
COMP
SFRP2
COL8A1
ITGBL1
WFDC21P
PC2
MYBL1
VPREB3
STAG3
LYPD6B
SERPINA9
MME
LINC00487
CSRNP3
CCDC85A
MARCO
PC3
EPCAM
ACTG2
MMP1
STEAP2
STEAP1
CXCL8
CXCL5
GPC6
HMCN1
UBASH3A
PC4
SLC40A1
ICOS
CA12
IGFBP2
PITX1
SQLE
FKBP5
TGFBRAP1
UBE4B
ORM1
PC5
COL11A1
CD24
TCL1A
LTF
CXCL14
FDCSP
CCL21
CETP
SOX8
CLU
.cell_cycle
G1
G2M
S
Chemotherapy
CHOP
R−CHOP
cluster
C1
C2
C3
C4
C5
C6
C7
dlbcl.type
ABC
GCB
gender
female
male
OS.status
ALIVE
DEAD

20.3.2 Variable importance plot

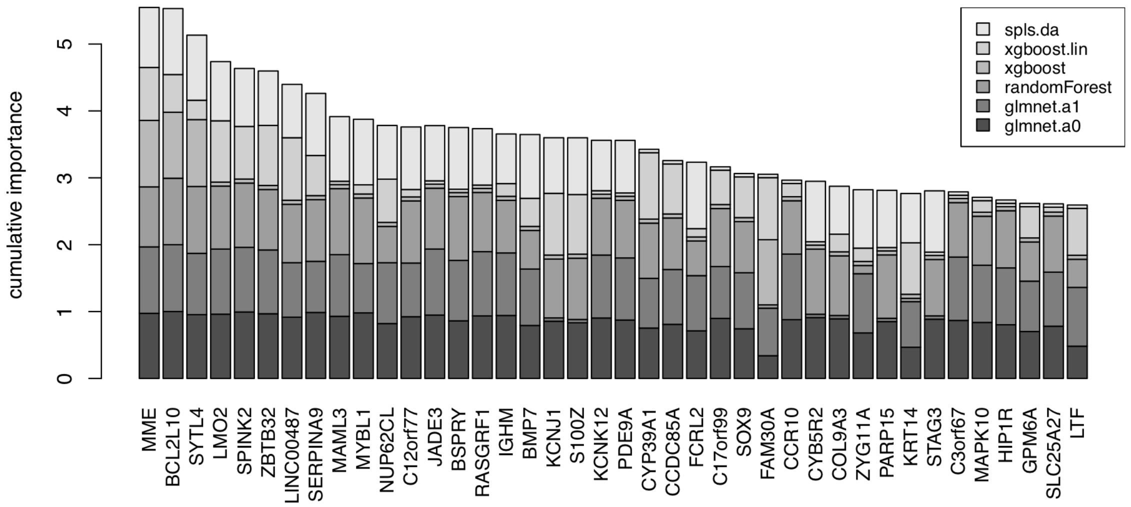


Figure 10. Variable importance plot. To replicate the figure, go to the **Biomarker** module, and set the `Predicted target: dlbcl.type` from the input panel.

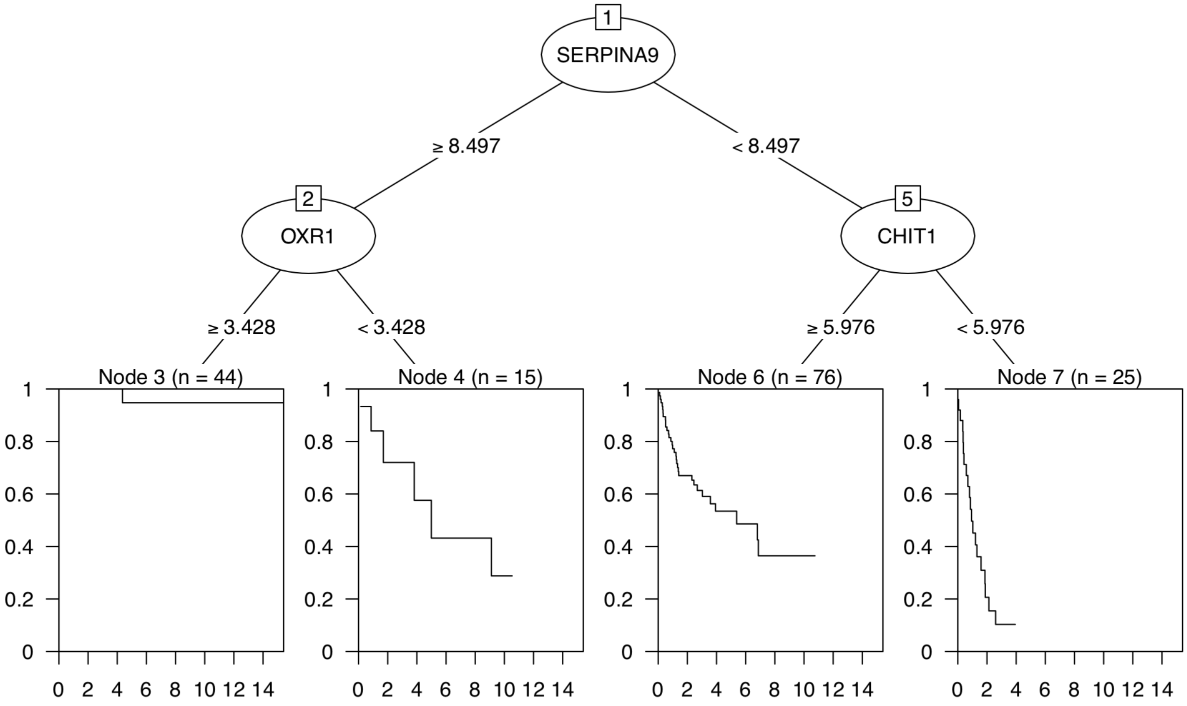
20.3.3 Survival tree

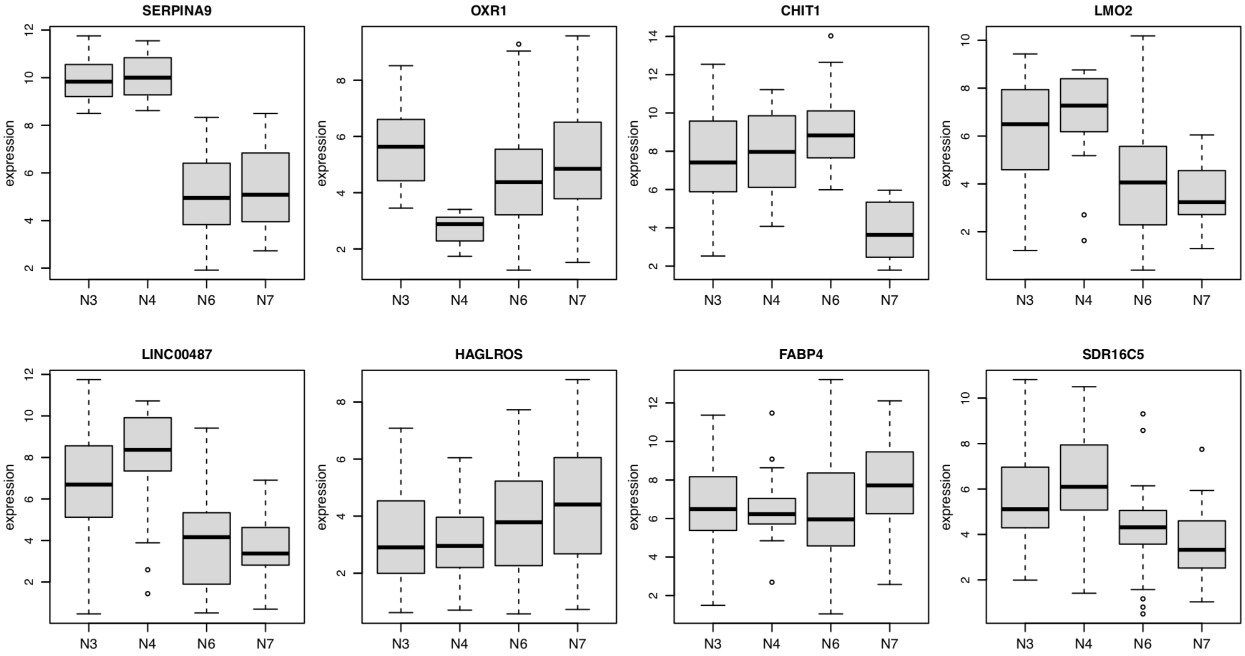
Figure 11. Survival tree analysis for `GSE10846-dlbcl` dataset. To redproduce similar figures, go to the **Biomarker** module, and set the `Predicted target: OS.survival` from the input panel. Note that the survival tree is stochastically built up with some of the top features shown in *Figure 10*; Therefore, users can get a slightly different survival tree every time.

20.4 Proteomic Data

With larger data sets, often the number of contrasts increases and complicates the overall analysis. For example, the proteomics data set of Rieckmann et al. 2017 comprises 26 populations of seven major immune cell types, measured during resting and activated states. There are more than 300 possible comparisons to make. To gain a better overview, gene set activation matrix (*Figure 12*) help visualize the similarities between multiple contrasts on a functional level. Alternatively, similarities can be visualized as a connectivity graph (*Figure 13*). For the same data set, *Figure 14* shows a computed partition tree that classifies the major cell types.

Another example dataset is from Geiger et al., where the proteome profiles of activated vs resting human naive T cells at different times were compared. *Figure 15* shows the volcano plots corresponding to eight different statistical tests comparing time-dependent activation of T cells at 48h vs. 12h. We see that both standard t-test and the Welch t-test show much less power to detect significant genes compared to the other methods. The result from edgeR-QLF is close to those of the two limma based methods, while edgeR-LRT is very similar to the results of DESeq2-Wald.


1
SERPINA9
≥ 8.497
< 8.497
2
OXR1
5
CHIT1
≥ 3.428
< 3.428
≥ 5.976
< 5.976
Node 3 (n = 44)
Node 4 (n = 15)
Node 6 (n = 76)
Node 7 (n = 25)
1
0.8
0.6
0.4
0.2
0
0 2 4 6 8 10 12 14


SERPINA9
OXR1
CHIT1
LMO2
LINC00487
HAGLROS
FABP4
SDR16C5
expression
N3
N4
N6
N7

20.4.1 Activation matrix

Figure 12. Gene Ontology activation matrix. To replicate the figure, select and load the `rieckmann2017-immprot` dataset, and go to the **GO graph** panel of the **Functional** module with default settings.

20.4.2 Contrast heatmap

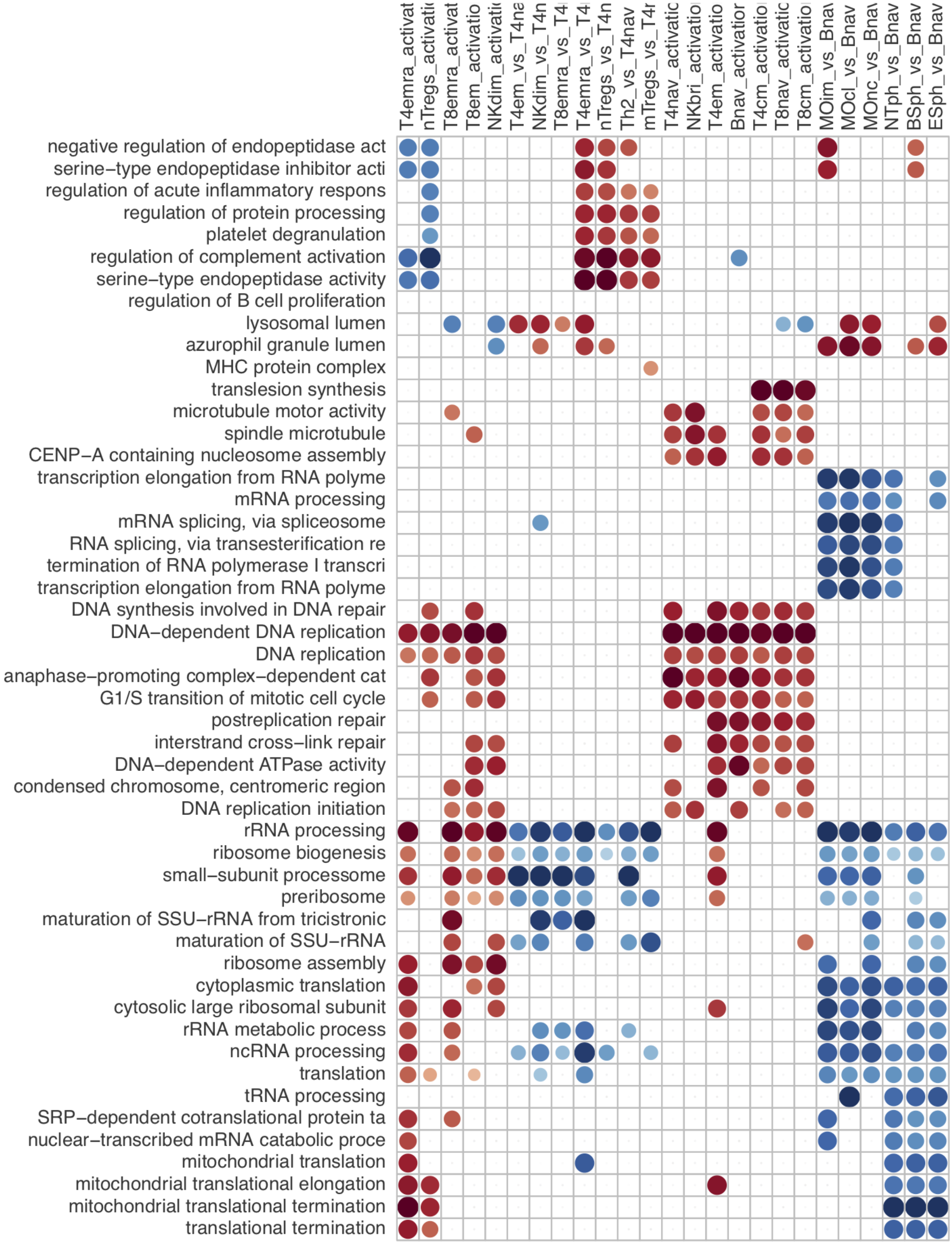
Figure 13. Contrast heatmap for the `rieckmann2017-immprot` dataset. To generate the figure on the platform, go to the **Contrast heatmap** panel of the **Intersection** module with default settings.

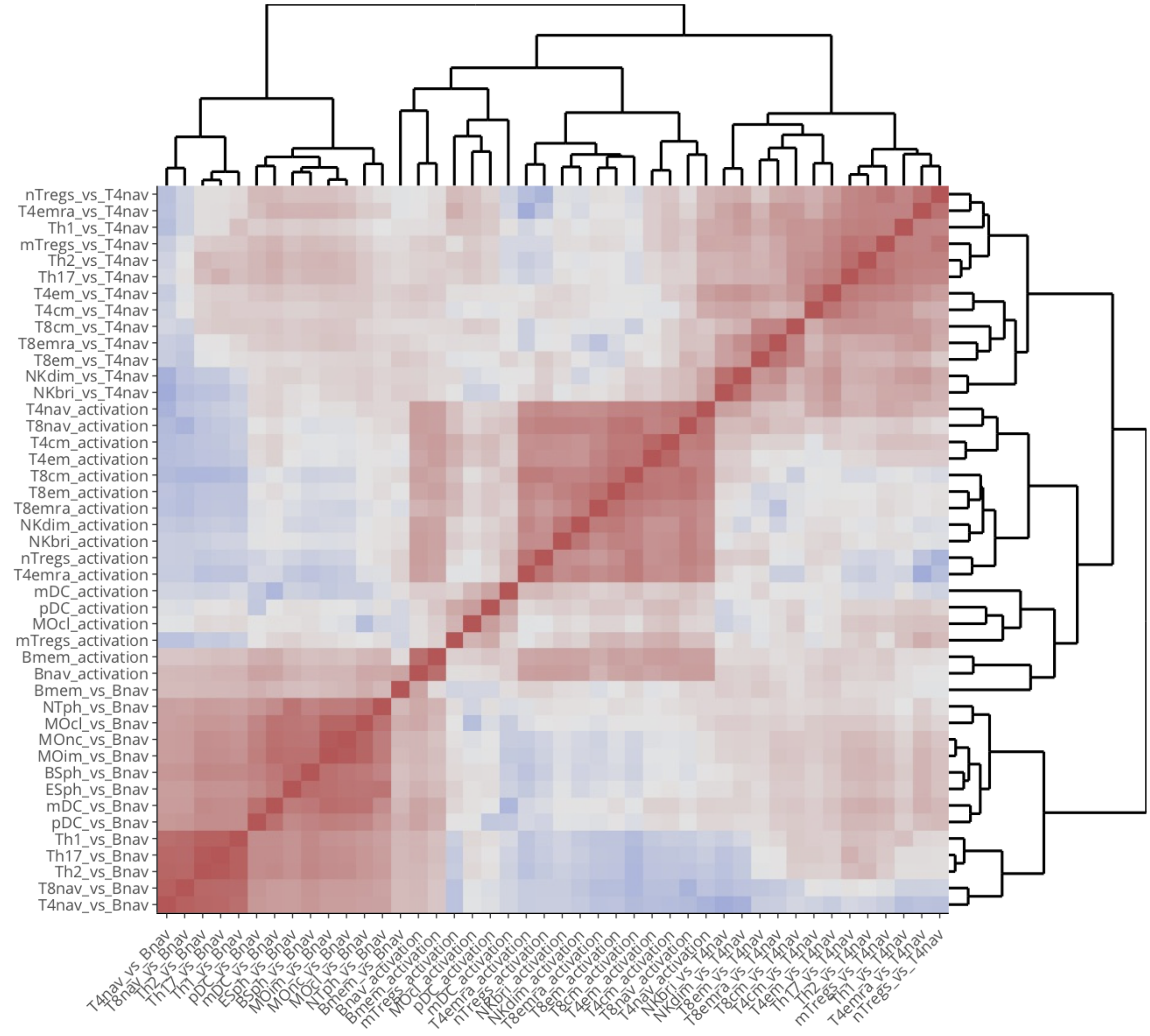
20.4.3 Classification tree

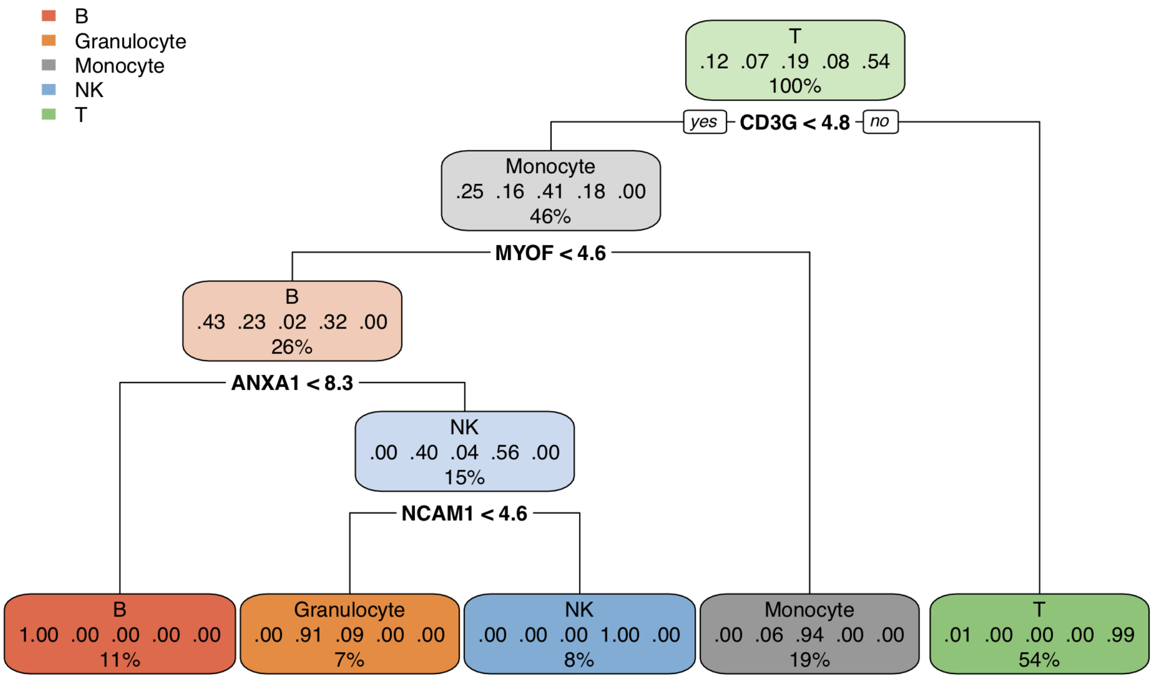
Figure 14. Classification tree for the `rieckmann2017-immprot` dataset. To reproduce similar figures, go to the **Biomarker** module, and set the `Predicted target:  cell.type` from the input panel. Note that the classification tree is stochastically built up with some of the top features shown in *Figure 10*; Therefore, users can get a slightly different survival tree every time.

20.4.4 Volcano plots of methods

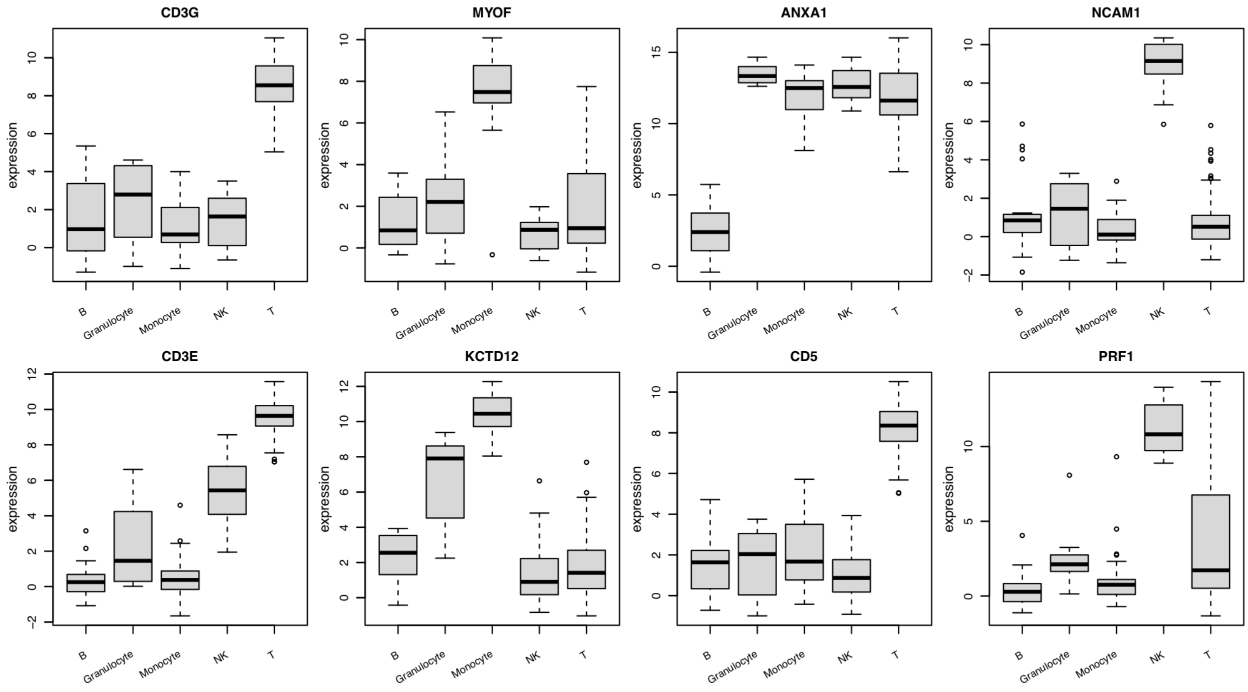
Figure 15. Volcano plots corresponding to eight different statistical methods comparing time-dependent expression of T cell activation at 48h vs. 12h. To regenerate the figure, select and load `geiger2016-arginine` dataset. Go to the **Volcano (methods)** panel under the **Expression** module, and set the `Contrast:  act48h_vs_act12h`.


T4emra_activat
nTregs_activati
T8emra_activat
T8em_activatio
NKdim_activati
T4em_vs_T4na
NKdim_vs_T4n
T8emra_vs_T4
T4emra_vs_T4
nTregs_vs_T4n
Th2_vs_T4nav
mTregs_vs_T4r
T4nav_activatic
NKbri_activatio
T4em_activatio
Bnav_activatior
T4cm_activatio
T8nav_activatic
T8cm_activatio
MOim_vs_Bnav
MOcl_vs_Bnav
MOnc_vs_Bnav
NTph_vs_Bnav
BSph_vs_Bnav
ESph_vs_Bnav
negative regulation of endopeptidase act
serine–type endopeptidase inhibitor acti
regulation of acute inflammatory respons
regulation of protein processing
platelet degranulation
regulation of complement activation
serine–type endopeptidase activity
regulation of B cell proliferation
lysosomal lumen
azurophil granule lumen
MHC protein complex
translesion synthesis
microtubule motor activity
spindle microtubule
CENP–A containing nucleosome assembly
transcription elongation from RNA polyme
mRNA processing
mRNA splicing, via spliceosome
RNA splicing, via transesterification re
termination of RNA polymerase I transcri
transcription elongation from RNA polyme
DNA synthesis involved in DNA repair
DNA–dependent DNA replication
DNA replication
anaphase–promoting complex–dependent cat
G1/S transition of mitotic cell cycle
postreplication repair
interstrand cross–link repair
DNA–dependent ATPase activity
condensed chromosome, centromeric region
DNA replication initiation
rRNA processing
ribosome biogenesis
small–subunit processome
preribosome
maturation of SSU–rRNA from tricistronic
maturation of SSU–rRNA
ribosome assembly
cytoplasmic translation
cytosolic large ribosomal subunit
rRNA metabolic process
ncRNA processing
translation
tRNA processing
SRP–dependent cotranslational protein ta
nuclear–transcribed mRNA catabolic proce
mitochondrial translation
mitochondrial translational elongation
mitochondrial translational termination
translational termination

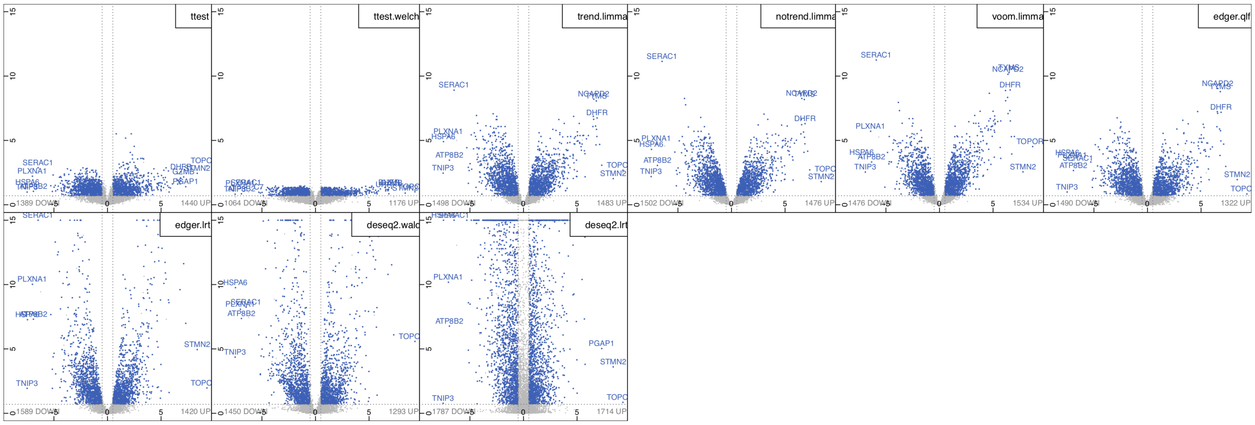

nTregs_vs_T4nav
T4emra_vs_T4nav
Th1_vs_T4nav
mTregs_vs_T4nav
Th2_vs_T4nav
Th17_vs_T4nav
T4em_vs_T4nav
T4cm_vs_T4nav
T8cm_vs_T4nav
T8emra_vs_T4nav
T8em_vs_T4nav
NKdim_vs_T4nav
NKbri_vs_T4nav
T4nav_activation
T8nav_activation
T4cm_activation
T4em_activation
T8cm_activation
T8em_activation
T8emra_activation
NKdim_activation
NKbri_activation
nTregs_activation
T4emra_activation
mDC_activation
pDC_activation
MOcl_activation
mTregs_activation
Bmem_activation
Bnav_activation
Bmem_vs_Bnav
NTph_vs_Bnav
MOcl_vs_Bnav
MOnc_vs_Bnav
MOim_vs_Bnav
BSph_vs_Bnav
ESph_vs_Bnav
mDC_vs_Bnav
pDC_vs_Bnav
Th1_vs_Bnav
Th17_vs_Bnav
Th2_vs_Bnav
T8nav_vs_Bnav
T4nav_vs_Bnav
T4nav_vs_Bnav
T8nav_vs_Bnav
Th2_vs_Bnav
Th17_vs_Bnav
Th1_vs_Bnav
pDC_vs_Bnav
mDC_vs_Bnav
ESph_vs_Bnav
BSph_vs_Bnav
MOim_vs_Bnav
MOnc_vs_Bnav
MOcl_vs_Bnav
NTph_vs_Bnav
Bmem_vs_Bnav
Bnav_activation
Bmem_activation
mTregs_activation
MOcl_activation
pDC_activation
mDC_activation
T4emra_activation
nTregs_activation
NKbri_activation
NKdim_activation
T8emra_activation
T8em_activation
T8cm_activation
T4em_activation
T4cm_activation
T8nav_activation
T4nav_activation
NKbri_vs_T4nav
NKdim_vs_T4nav
T8em_vs_T4nav
T8emra_vs_T4nav
T8cm_vs_T4nav
T4cm_vs_T4nav
T4em_vs_T4nav
Th17_vs_T4nav
Th2_vs_T4nav
mTregs_vs_T4nav
Th1_vs_T4nav
T4emra_vs_T4nav
nTregs_vs_T4nav



B
Granulocyte
Monocyte
NK
T
T
.12 .07 .19 .08 .54
100%
yes
CD3G < 4.8
no
Monocyte
.25 .16 .41 .18 .00
46%
MYOF < 4.6
B
.43 .23 .02 .32 .00
26%
ANXA1 < 8.3
NK
.00 .40 .04 .56 .00
15%
NCAM1 < 4.6
B
1.00 .00 .00 .00 .00
11%
Granulocyte
.00 .91 .09 .00 .00
7%
NK
.00 .00 .00 1.00 .00
8%
Monocyte
.00 .06 .94 .00 .00
19%
T
.01 .00 .00 .00 .99
54%



CD3G
MYOF
ANXA1
NCAM1
CD3E
KCTD12
CD5
PRF1
expression
B
Granulocyte
Monocyte
NK
T


ttest
SERAC1
PLXNA1
HSPA6
DHFR
STMN2
GZMB
PGAP1
TOPC
1389 DOWN
1440 UP
ttest.welch
SERAC1
1064 DOWN
1176 UP
STMN2
TOPC
trend.limma
SERAC1
NCAPD2
TYMS
DHFR
PLXNA1
HSPA6
ATP8B2
TNIP3
TOPC
STMN2
1498 DOWN
1483 UP
notrend.limma
SERAC1
NCAPD2
TYMS
DHFR
PLXNA1
HSPA6
ATP8B2
TNIP3
TOPC
STMN2
1502 DOWN
1476 UP
voom.limma
SERAC1
TYMS
NCAPD2
DHFR
PLXNA1
HSPA6
ATP8B2
TNIP3
TOPOR
STMN2
1476 DOWN
1534 UP
edger.qlf
NCAPD2
TYMS
DHFR
HSPA6
SERAC1
ATP8B2
TNIP3
STMN2
TOPC
1490 DOWN
1322 UP
edger.lrt
SERAC1
PLXNA1
ATP8B2
STMN2
TNIP3
TOPC
1589 DOWN
1420 UP
deseq2.wald
HSPA6
SERAC1
PLXNA1
ATP8B2
TNIP3
TOPC
1450 DOWN
1293 UP
deseq2.lrt
SERAC1
PLXNA1
ATP8B2
PGAP1
STMN2
TNIP3
TOPC
1787 DOWN
1714 UP
-5
0
5
0
5
10
15

CHAPTER

TWENTYONE

METHODS

Below are snippets that you can use to describe the methods when using the Omics Playground. These are just examples and you need to extract and modify the parts you used and need.

21.1 Batch correction

Batch effects, or contamination by unwanted variables, was identified by an F-test for the first three principal components. Continuous variables were dichotomized into high/low before testing. Highly confounding variables would appear as having high relative contribution in the first or second principal component, often higher than the variable of interest. Batch effects were also visually assessed (before and after correction) using annotated heatmaps and t-SNE plots colored by variables.

Batch correction was performed for explicit batch variables or unwanted covariates. Parameters with a correlation r>0.3 with any of variables of interest (i.e. the model parameters) were omitted from the regression. Correction was performed by regressing out the covariate using the 'removeBatchEffect' function in the limma R/Bioconductor package.

Technical correction was performed for intrinsic technical parameters such as library size (i.e. total counts), mitochondrial and ribosomal proportions, cell cycle and gender. These parameters were estimated from the data. The cell cycle was estimated using the Seurat R/Bioconductor package. Gender (if not given) was estimated by checking the absence/presence of expression of gender specific genes on the X/Y chromosome. Parameters with a correlation r>0.3 with any of the model parameters were omitted from the regression. Correction was performed by regressing out the covariate using the 'removeBatchEffect' function in the limma R/Bioconductor package.

Unsupervised batch correction was performed using 'surrogate variable analysis' (SVA) (Leek 2007) by estimating the latent surrogate variables and regressing out using the 'removeBatchEffect' function in the limma R/Bioconductor package.

21.2 Clustering

Heatmaps were generated using the ComplexHeatmap R/Bioconductor package (Gu 2016) on scaled log-expression values (z-score) using euclidean distance and Ward linkage. The standard deviation was used to rank the genes for the reduced heatmaps.

T-distributed stochastic neighbor embedding (t-SNE) was computed using the top 1000 most varying genes, then reduced to 50 PCA dimensions before computing the t-SNE embedding. The perplexity heuristically set to 25% of the sample size or 30 at maximum, and 2 at minimum. Calculation was performed using the *Rtsne* R package.

Uniform Manifold Approximation and Projection (UMAP) was computed using the top 1000 most varying genes, then reduced to 50 PCA dimensions before computing the UMAP embedding. The number of neighbours was heuristically set to 25% of the sample size or 30 at maximum, and 2 at minimum. Calculation was performed using the *uwot* R package.

Principal component analysis (PCA) was computed using the *irlba* R package.

21.3 Statistical testing

Multi-method statistical testing. For gene-level testing, statistical significance was assessed using three independent statistical methods: DESeq2 (Wald test), edgeR (QLF test) and limma-trend (Love 2014; Robinson 2010; Ritchie 2015). The maximum q-value of the three methods was taken as aggregate q-value, which corresponds to taking the intersection of significant genes from all three tests.

Statistical testing of differential enrichment of genesets was performed using an aggregation of multiple statistical methods: Fisher's exact test, fGSEA (Korotkevich 2019), Camera (Wu 2012) and GSVA/limma (Hanzelmann 2013, Ritchie 2015). The maximum q-value of the selected methods was taken as aggregate meta.q value, which corresponds to taking the intersection of significant genes from all tests. As each method uses different estimation parameters (NES for GSEA, odd-ratio for fisher, etc.) for the effect size, for consistency, we took the average log fold-change of the genes in the geneset as sentinel value. We used more than 50000 genesets from various public databases including: MSigDB (Subramanian 2005; Liberzon 2015), Gene Ontology (Ashburner 2000), and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa 2000).

21.4 Functional analysis

Graph-weighted GO analysis. The enrichment score of a GO term was defined as the sum of q-weighted average fold-changes, (1-q)*logFC, of the GO term and all its higher order terms along the shortest path to the root in the GO graph. The fold-change of a gene set was defined as the average of the fold-change values of its member genes. This graph-weighted enrichment score thus reflects the enrichment of a GO term with evidence that is corroborated by its parents in the GO graph and therefore provides a more robust estimate of enrichment. The activation map visualizes the scores of the top-ranked GO terms for multiple contrasts as a heatmap.

KEGG pathway visualization was performed using the Pathview R/Bioconductor package using the foldchange as node color.

21.5 Cell type profiling

Cell type profiling was performed using the LM22 signature matrix as reference data set (Chen 2018). We have evaluated a total of 6 computational deconvolution methods: DeconRNAseq (Gong 2013), DCQ (Altboum 2014), I-NNLS (Abbas 2009), NNLM (Lin 2020), rank-correlation and a meta-method. For NNLM, we repeated NNLM for non-logarithmic (NNLM.lin) and ranked signals (NNLM.rnk). The latter meta-methods, meta and meta.prod, summarize the predictions of all the other methods as the mean and/or geometric mean of the normalized prediction probabilities, respectively.

[1] Gong T, Szustakowski JD. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. Bioinformatics. 2013.

[2] Altboum Z, et al. Digital cell quantification identifies global immune cell dynamics during influenza infection. Mol Syst Biol. 2014 Feb 28;10(2):720.

[3] Abbas A, et al. Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus, PLOS One, 2009.

[4] Lin X, Boutros PC. Optimization and expansion of non-negative matrix factorization. BMC Bioinformatics. 2020.

[5] Chen B, et al. Profiling Tumor Infiltrating Immune Cells with CIBERSORT. Methods Mol Biol. 2018.

21.6 Scripting and visualization

Data preprocessing was performed using bespoke scripts using R (R Core Team 2013) and packages from Bioconductor (Huber 2015). Statistical computation and visualization have been performed using the Omics Playground version vX.X.X (Akhmedov 2020).

21.7 REFERENCES

Akhmedov M, Martinelli A, Geiger R and Kwee I. Omics Playground: A comprehensive self-service platform forvisualization, analytics and exploration of Big Omics Data. NAR Genomics and Bioinformatics, Volume 2, Issue 1, March 2020,

Ashburner et al. Gene ontology: tool for the unification of biology. Nat Genet. May 2000;25(1):25-9.

Huber W, et al. (2015) Orchestrating high-throughput genomic analysis with Bioconductor. Nature Methods 12:115-121; doi:10.1038/nmeth.3252

Kanehisa, M. and Goto, S.; KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30 (2000).

Leek J., Storey J. Capturing heterogeneity in gene expression studies by 'surrogate variable analysis' PLoS Genet. 2007

Love MI, Huber W, Anders S (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome Biology, 15, 550.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies." Nucleic Acids Research, 43(7)

Robinson MD, McCarthy DJ, Smyth GK (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics, 26(1), 139-140.

CHAPTER

TWENTYTWO

FREQUENTLY ASKED QUESTIONS (FAQ)

22.1 Can I use LIMMA, EdgeR and DESeq2 for my proteomics data?

EdgeR and DESeq2 are statistical methods based on the negative binomial model (an overdispersed Poisson model). Poisson and negative binomial models, naturally account for heteroscedasticity and zero values in the data. Also LIMMA is widely used in RNA-seq analysis, and is an emperical Bayesian method based on moderated t-test statistics. While these methods were originally conceived for the for differential expression analysis in RNA-seq data, there is increasingly more acceptance that these methods can also be used for proteomics data.

Langley and Mayer (2015) assessed seven methods for differential expression analysis in proteomics, and showed that DESeq to outperform the more commonly used t-test. Kammers et al. (2015) show that LIMMA shows better results than the t-test. Branson and Fretais (2016) also highlighted the statistical analogy of proteomics label-free spectral count quantification with count data from RNA sequencing and propose to use edgeR, DESEq and baySeq for proteomics data. Gregori et al. (2013) uses EdgeR and Poisson based methods for their LC/MS-MS data. They also created the 'msmsTests' R package. Medo et al (2019) use EdgeR for their study on missing values in the differential analysis of proteomic and phosphoproteomics data. Gatto proposes both edgeR and LIMMA based methods for proteomics statistical analysis. Chen et al. (2020) reviewed bioinformatics methods for proteomics data analysis and reaffirmed that LIMMA can achieve more robust and accurate results than the traditional t-test.

Notice that proper scaling/normalization of quantitative proteomics data is important before using RNA-seq based methods as sometimes the proteomics intensity values may be far greater than those found in RNA-seq. Therefore Omics Playground scales proteomics data automatically to 'counts per million' (CPM). We have also seen that batch correction (e.g. using ComBat) may improve the downstream statistical analysis in proteomics.

References:

1. Kammers et al. "Detecting significant changes in protein abundance". EuPA Open Proteomics Volume 7, June 2015.
2. Langley SR, Mayr M. "Comparative analysis of statistical methods used for detecting differential expression in label-free mass spectrometry proteomics". J Proteomics. 2015.
3. Branson OE, Freitas MA. "A multi-model statistical approach for proteomic spectral count quantitation". J Proteomics. 2016.
4. Gregori et al. "msmsTests-package: LC-MS/MS Differential Expression Tests". R package.
5. Gregori et al. "An Effect Size Filter Improves the Reproducibility in Spectral Counting-based Comparative Proteomics." Journal of Proteomics, 2013.
6. Medo, M, Aebersold, DM and Medová, M "ProtRank: bypassing the imputation of missing values in differential expression analysis of proteomic data." BMC Bioinformatics 20, 563 (2019).
7. Gatto L. "Bioconductor tools for mass spectrometry and proteomics", https://lgatto.github.io/bioc-ms-prot/lab.html#8_statistical_analysis

8. Chen et al. “Bioinformatics Methods for Mass Spectrometry-Based Proteomics Data Analysis.” Int J Mol Sci. 2020;21(8):2873, 2020.

22.2 How are duplicated gene/protein names handled in the counts file?

Duplicated row identifiers (genes/proteins with same name) are handled by summing up their linear intensities/counts. If the data was in logarithm, it will be (hopefully) automatically detected and exponentiated. The rational of summing up the counts (or linear intensities in proteomics) is that we don’t differentiate between possible gene/protein isoforms and sum them up as a group. If you want to retain the isoforms, you may keep the names as GENE.1 and GENE.2 but you must turn off any gene filter. However as currently such gene/protein variants are not recognized in the gene sets, this will result in wrong enrichment test.

22.3 How are “missing values” handled in the counts file?

At the moment missing values are imputed to 0 (zero). The Omics Playground uses the function *pgx.createPGX()* in the file pgx-compute.R. Note that only real NA (or empty) values in the counts file are supposed to be “missing”. Zero values are assumed to be real zeros. If zero-value imputaton is not what you want, and you want to impute your “missing” values differently, (currently) you must do that manually before uploading the CSV file.

22.4 How do I switch between stable, beta & testing docker versions?

You can test the latest features of the Omics Playground by switching to a more experimental docker version. Omics Playground has three different docker versions:

- Stable channel (latest): This channel is fully tested by the Omics Playground team, and is the best choice to avoid crashes and other problems. It’s updated roughly every 2–3 months for minor changes, and every 6 months for major changes.
- Beta channel (beta): To view upcoming improvements and features with low risk, use the Beta channel. It’s updated roughly every month, with updates before the Stable channel gets them.
- Testing channel (testing): The Testing channel gets updated once or twice weekly and is intended only for developers. This build might have serious bugs or highly experimental features. The Testing channel always uses the latest source code and is build irregularly whenever needed.

You can switch between versions by pulling the docker image corresponding to the version latest, beta or testing.

CHAPTER TWENTYTHREE

CONTACT

If you have further questions about how to use Omics Playground, please ask your question in our user forum at Google Groups.

If you've discovered a bug, have a feature request, or want to contribute to the code, please go to our GitHub repository.

If you have other questions you can email us.

CHAPTER TWENTYFOUR

CITATION

To cite Omics Playground in publications please use:

Akhmedov M, Martinelli A, Geiger R and Kwee I. Omics Playground: A comprehensive self-service platform forvisualization, analytics and exploration of Big Omics Data. *NAR Genomics and Bioinformatics, Volume 2, Issue 1, March 2020, lqz019; doi:10.1093/nargab/lqz019*

CHAPTER

TWENTYFIVE

LICENSE

Omics Playground is open source: you are allowed to download the source code and make your own edits. Although Omics Playground is open source, the software is distributed under a dual license: for either non-commerical or commercial use.

25.1 Non-commercial

If you are a non-profit company, academic, or use our products for personal use, you may enjoy our software for free under a Creative Commons (CC) Attribution-NonCommercial licence.

25.2 Commercial/governmental

If you are not entitled to the non-commercial license, or you are a governmental agency, you must purchase a commercial license. Buying such a license is mandatory as soon as you develop commercial activities involving Omics Playground software. These activities include: offering paid services to customers as an ASP, SaaS, professional consulting, or shipping Omics Playground with a closed source product.