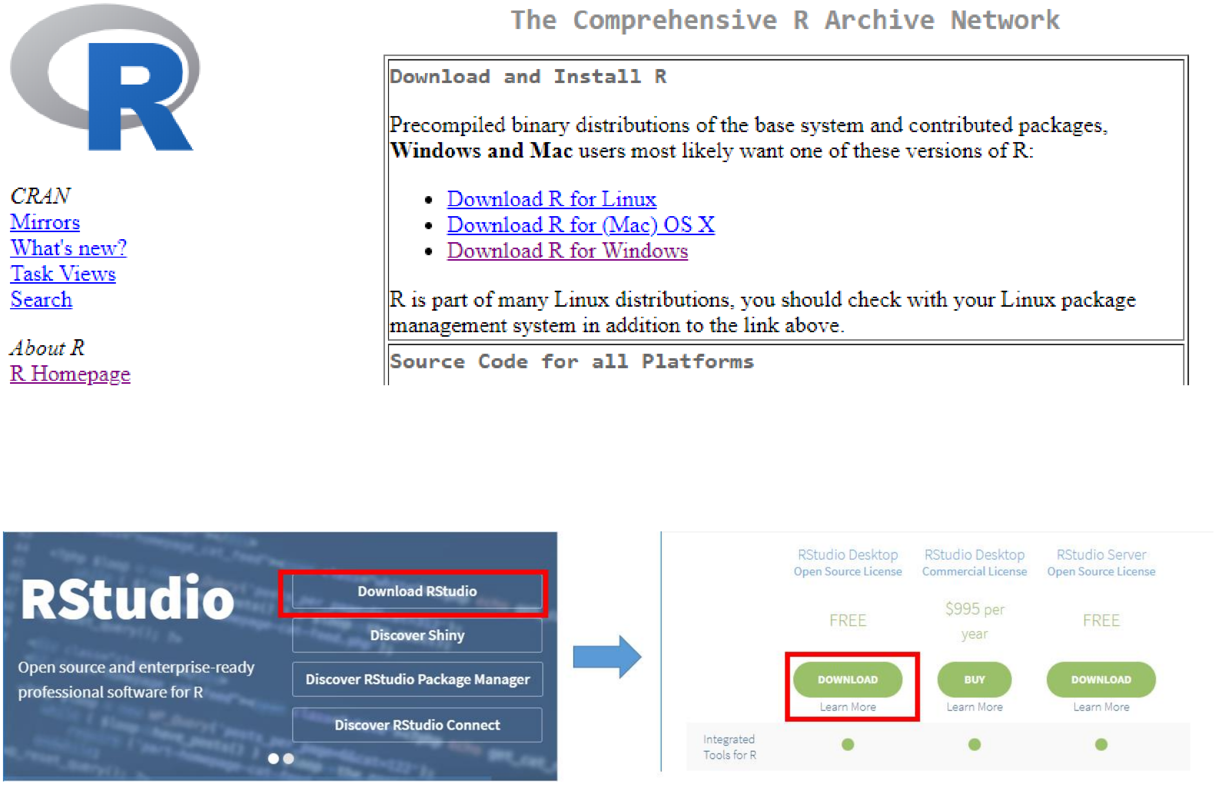
Supplementary Information (ABioTrans User Manual)

ABioTrans - A Biostatistical tool for Transcriptomics Analysis

Updated 25 July 2019

Setup

1. Install R from https://www.r-project.org/ and Rstudio from https://www.rstudio.com/



The Comprehensive R Archive Network

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

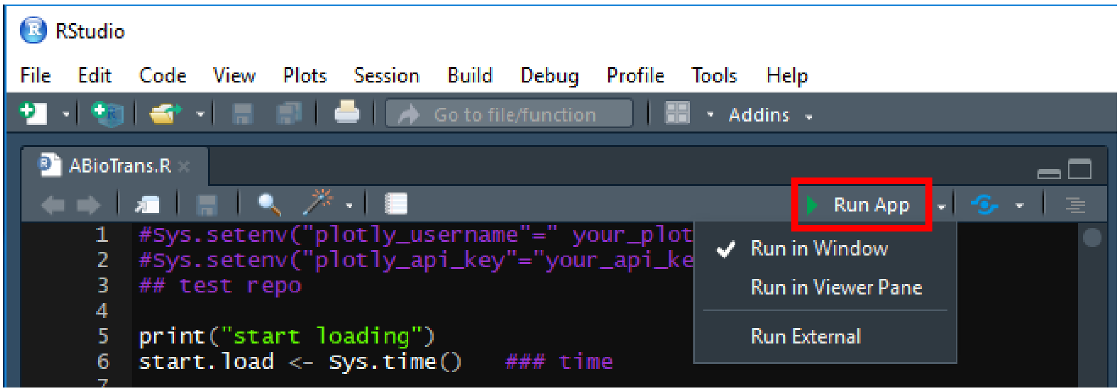
R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Figure 1a: Downloading R from the R project and RStudio from RStudio webpage

2. Download ABioTrans-master.zip on GitHub and unzip it. Please do not modify www inside ABioTrans folder.
3. Open the ABioTrans.R file using RStudio and click RunApp button on the top-right corner. `Run External` opens ABioTrans GUI in your default browser (recommended), whereas `Run in Window` opens the GUI as a RStudio window. When running ABioTrans the first time, it may take up to 30 minutes to install the required R packages to the local computer. For subsequent runs, it only takes 30s to load these packages and start ABioTrans.

Please note that this user manual works for both Windows and Mac operating system.



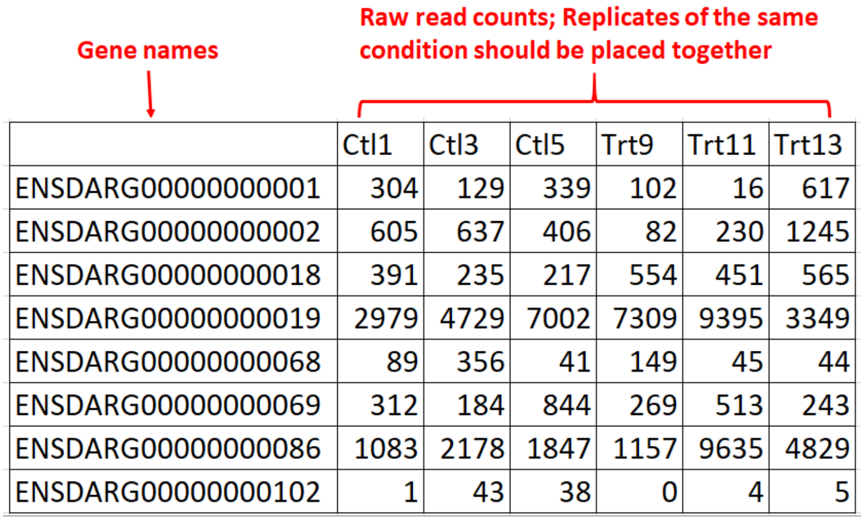
```
#Sys.setenv("plotly_username"=" your_plot
#Sys.setenv("plotly_api_key"="your_api_ke
## test repo

print("start loading")
start.load <- Sys.time()    ### time
```

Figure 1b: Starting ABioTrans from RStudio

You can start your analysis now!

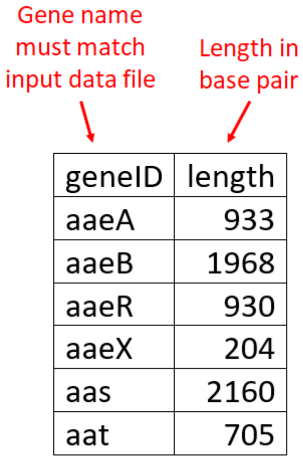
Uploading files

ABiotrans requires all input files in comma-separated value (.csv) format. The data file in .csv should contain the gene names in rows and genotypes (e.g. wild type – mutants or control - treatments, …) in columns, which is similar to the standard transcriptome data file format in GEO database. Supporting files (if applicable) include gene length, list of negative control genes, and metadata file. A number of normalization options are provided in the pre-processing tab depending on the availability of supporting files: RPKM, FPKM, TPM (requiring gene length), RUV (requiring negative control genes), and Upper Quartile (no supporting file needed). The metadata file is required for differential expression analysis, and should specify experimental conditions (e.g. Control/Treated, time 1/time 2/ time3,.) for each genotype listed in the data file. Otherwise, the user can move to the next option to perform analyses (e.g. scatter plot, distribution fitting, Pearson correlation, …) once the data file is loaded (whether normalized or in raw count).


Gene names
Raw read counts; Replicates of the same condition should be placed together

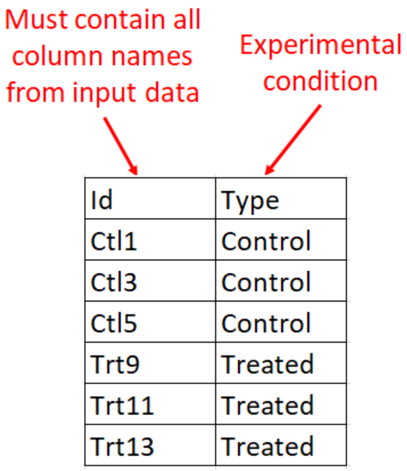
	Ctl1	Ctl3	Ctl5	Trt9	Trt11	Trt13
ENSDARG00000000001	304	129	339	102	16	617
ENSDARG00000000002	605	637	406	82	230	1245
ENSDARG00000000018	391	235	217	554	451	565
ENSDARG00000000019	2979	4729	7002	7309	9395	3349
ENSDARG00000000068	89	356	41	149	45	44
ENSDARG00000000069	312	184	844	269	513	243
ENSDARG00000000086	1083	2178	1847	1157	9635	4829
ENSDARG00000000102	1	43	38	0	4	5

Figure 2a: Required format for raw counts data file


Gene name must match input data file
Length in base pair

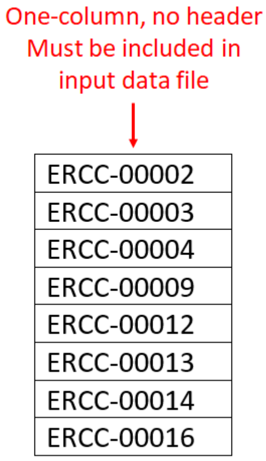
geneID	length
aaeA	933
aaeB	1968
aaeR	930
aaeX	204
aas	2160
aat	705

Figure 2b: Gene length file format


Must contain all column names from input data
Experimental condition

Id	Type
Ctl1	Control
Ctl3	Control
Ctl5	Control
Trt9	Treated
Trt11	Treated
Trt13	Treated

Figure 2c: Metadata file format

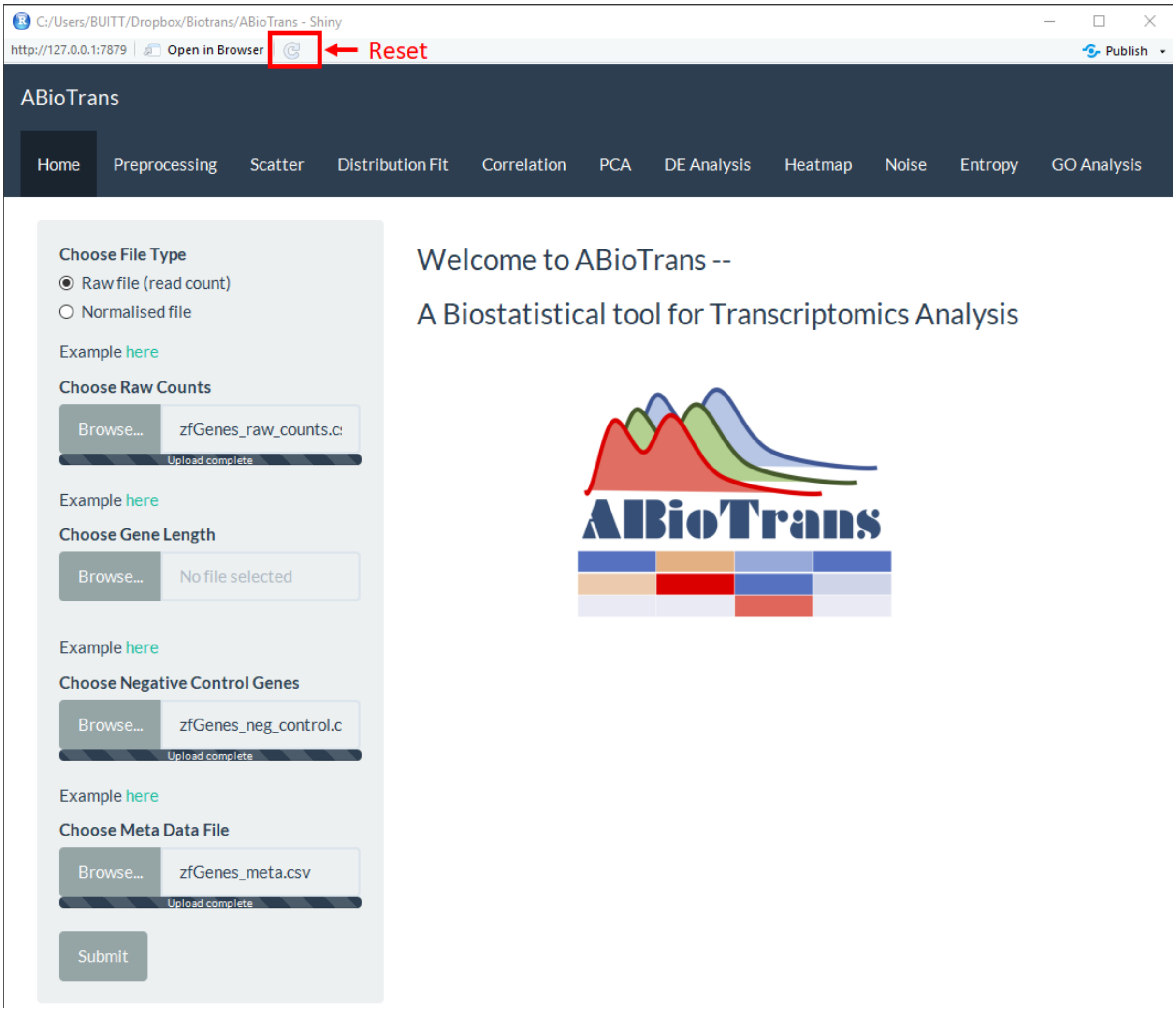

One-column, no header Must be included in input data file

ERCC-00002
ERCC-00003
ERCC-00004
ERCC-00009
ERCC-00012
ERCC-00013
ERCC-00014
ERCC-00016

Figure 2d: Negative control gene file format

In this demo, we will be using the zebra fish data from NCBI GEO database GSE53334: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53334 . The data has been assembled and converted to csv files in the test data folder https://github.com/buithuytien/ABioTrans/tree/master/Test%20data .

Once data files and supporting files are loaded into ABioTrans, user can press Submit button and ABioTrans will automatically proceed to the next tab `Preprocessing`. In case user accidentally upload wrong data file or supporting files, user can overwrite each of them by uploading new files, or press the Reset button to erase all uploaded files and reset the whole program.



Choose File Type

● Raw file (read count)

○ Normalised file

Example here

Choose Raw Counts: zfGenes_raw_counts.c:

Example here

Choose Gene Length: No file selected

Example here

Choose Negative Control Genes: zfGenes_neg_control.c

Example here

Choose Meta Data File: zfGenes_meta.csv

Welcome to ABioTrans --

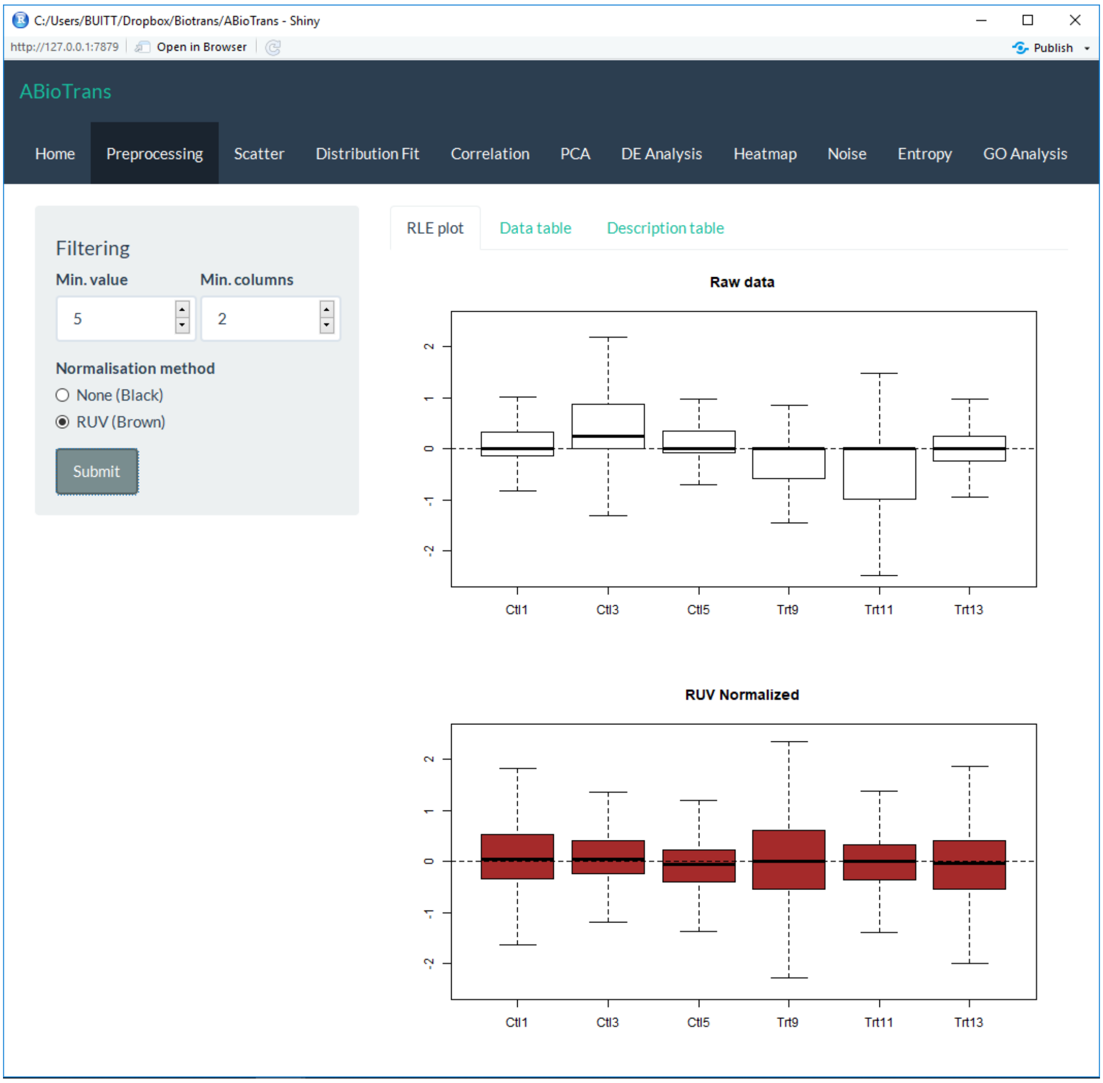
A Biostatistical tool for Transcriptomics Analysis

Figure 3: Home page and file upload – zebra fish data

Preprocessing

Preprocessing involves two steps: removing lowly expressed genes and normalizing the remaining gene expression. First, the user need to specify cut-off expression values (which must be in same unit to the input data file - either raw read counts or normalized expression), and the minimum number of data columns that must exceed the threshold value. Normalization methods are available upon the availability of supporting data files: normalization for sequencing depth, including TPM and RPKM, requires gene length and normalization for sample variation, including RUV, requires negative control genes. Relative Log Expression (RLE) plots of raw and processed data are displayed to visualize the effects of normalization.

In this demo, we carried out RUV normalization to remove the unwanted variation between samples (also known as batch effect correction, or between lane normalization). The rest of the analysis pipeline will be implemented on this filtered, RUV-normalized data.



Filtering

Min. value: 5

Min. columns: 2

Normalisation method

○ None (Black)

● RUV (Brown)

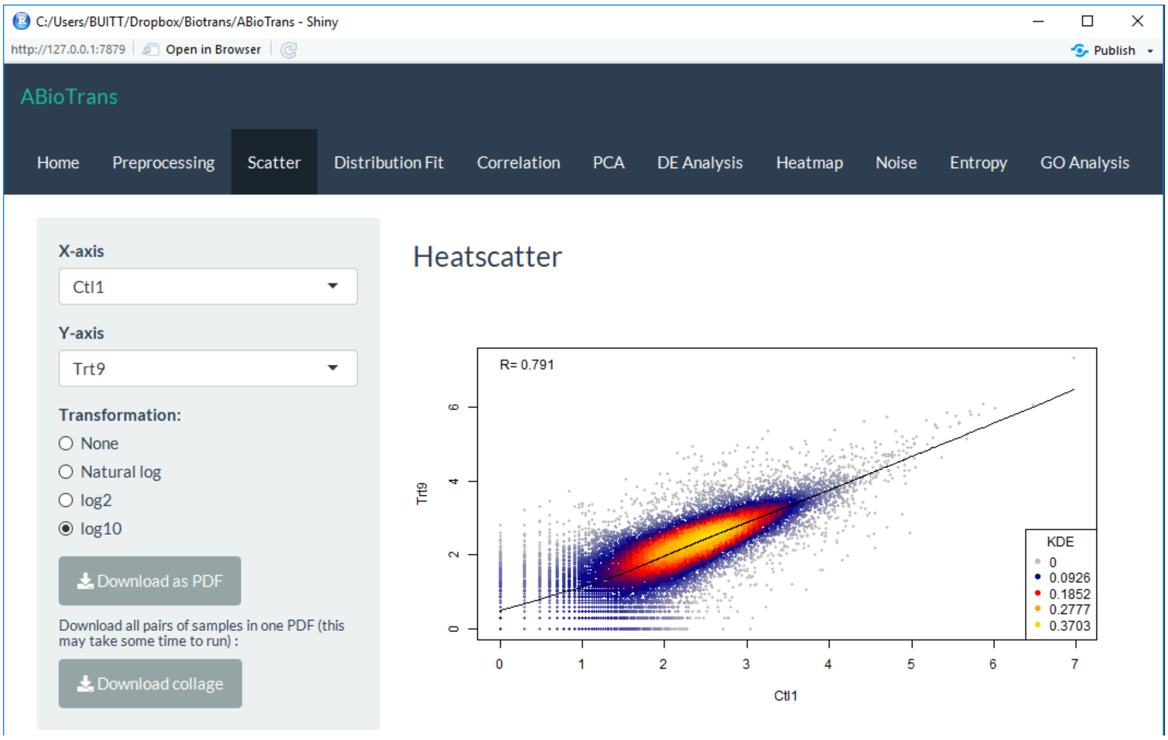
Submit

Figure 4: Preprocessing panel with RLE plots of raw data (upper figure) and filtered, RUV-normalized data (lower figure). Gene expression with minimum five counts in at least two columns are retained.

Scatter plot

Scatter plot compares any 2 samples (or 2 replicates) by displaying the respective expression of all genes in 2D space. Log transformation on the values usually provides better visualization due to the natural skewness of gene expression data. The user can choose to download each single scatter plot, or to download all pairs of samples scatter plot in one PDF file, which may take some time to run.

It is recommended to preform normalization for sequencing depth (TPM, RPKM, FPKM) for this step (and so does distribution fitting, correlation, hierarchical clustering, noise and entropy). However, since gene lengths are not available for this data set, the demo will use RUV- normalized data for all the analysis



X-axis
Ctl1

Y-axis
Trt9

Transformation:

○ None
○ Natural log
○ log2
● log10

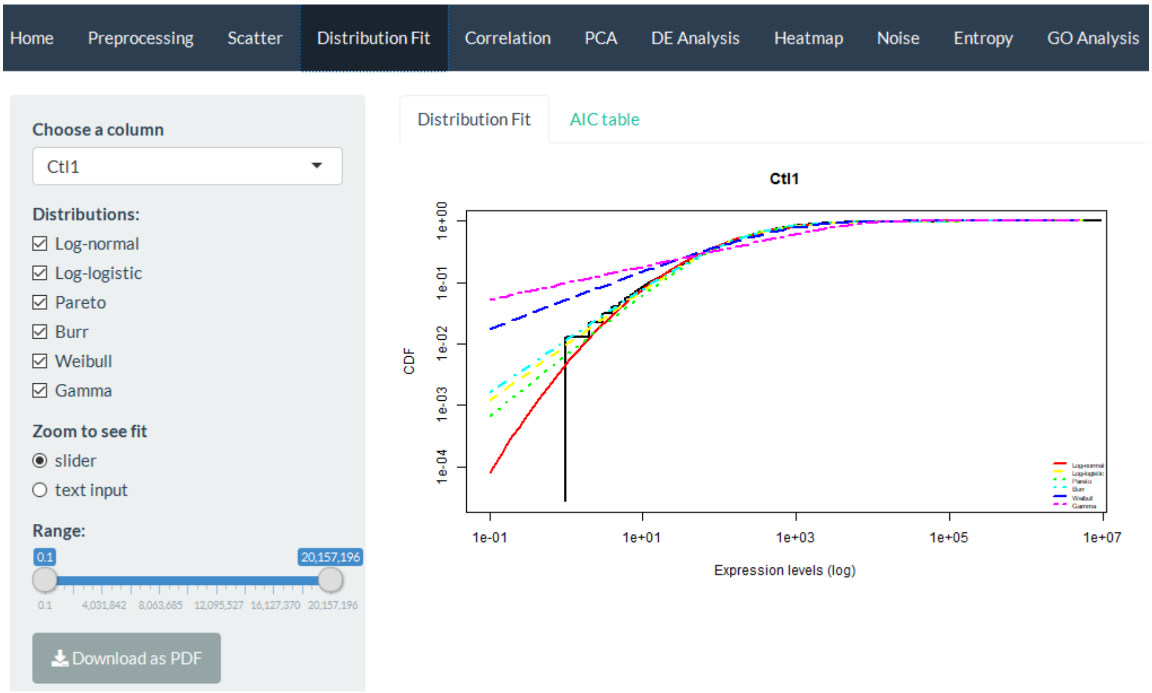
Download all pairs of samples in one PDF (this may take some time to run) :

Heatscatter

Figure 5: Scatter plot between Ctrl1 and Trt9 samples

Distribution fitting

Distribution fitting compares the gene expression to a number of statistical continuous distributions, which can be used to validate the data. To visualize the comparison, ABioTrans displays the Cumulative Distribution Function of the preprocessed gene expression data with the user-selected theoretical distributions. Once it is confirmed that the gene set follow a distribution, it would be safe conclude the validity of the gene expression data. AIC table is also provided in `AIC table` tab to show the best fitted distribution in each sample



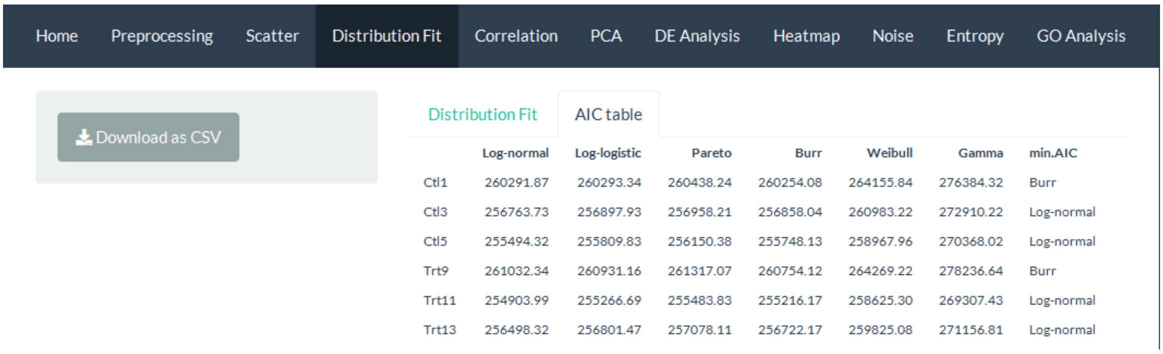
Choose a column: Ctl1

Distributions:
☑ Log-normal
☑ Log-logistic
☑ Pareto
☑ Burr
☑ Weibull
☑ Gamma

Zoom to see fit
● slider
○ text input

Range: 0.1 – 20,157,196
0.1 4,031,842 8,063,685 12,095,527 16,127,370 20,157,196

Figure 6a: Comparing Cumulative Distribution Functions of raw count data (black) with lognormal (red colour), log-logistic (yellow colour), Pareto (green colour), Burr (cyan colour), Weibull (blue colour) and gamma (purple colour) distributions in Ctrl 1 replicate.

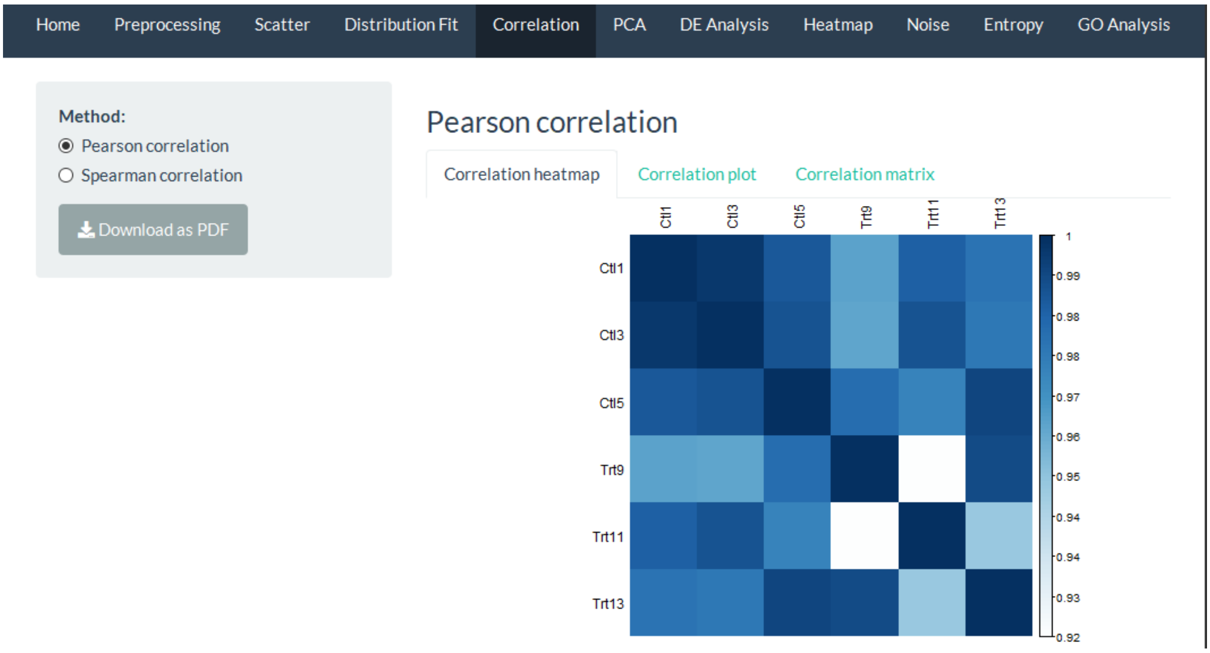


	Log-normal	Log-logistic	Pareto	Burr	Weibull	Gamma	min.AIC
Ctl1	260291.87	260293.34	260438.24	260254.08	264155.84	276384.32	Burr
Ctl3	256763.73	256897.93	256958.21	256858.04	260983.22	272910.22	Log-normal
Ctl5	255494.32	255809.83	256150.38	255748.13	258967.96	270368.02	Log-normal
Trt9	261032.34	260931.16	261317.07	260754.12	264269.22	278236.64	Burr
Trt11	254903.99	255266.69	255483.83	255216.17	258625.30	269307.43	Log-normal
Trt13	256498.32	256801.47	257078.11	256722.17	259825.08	271156.81	Log-normal

Figure 6b: Table of AIC values for 6 distributions (Log-normal, Log-logistic, Pareto, Burr, Weibull, and Gamma) in all 6 replicates

Correlation analysis

In this panel, Pearson and Spearman correlations between any two samples are computed. The correlation values are presented in numeric table (`Correlation matrix` tab) and heatmap (`Correlation heatmap` and `Correlation plot` tab).

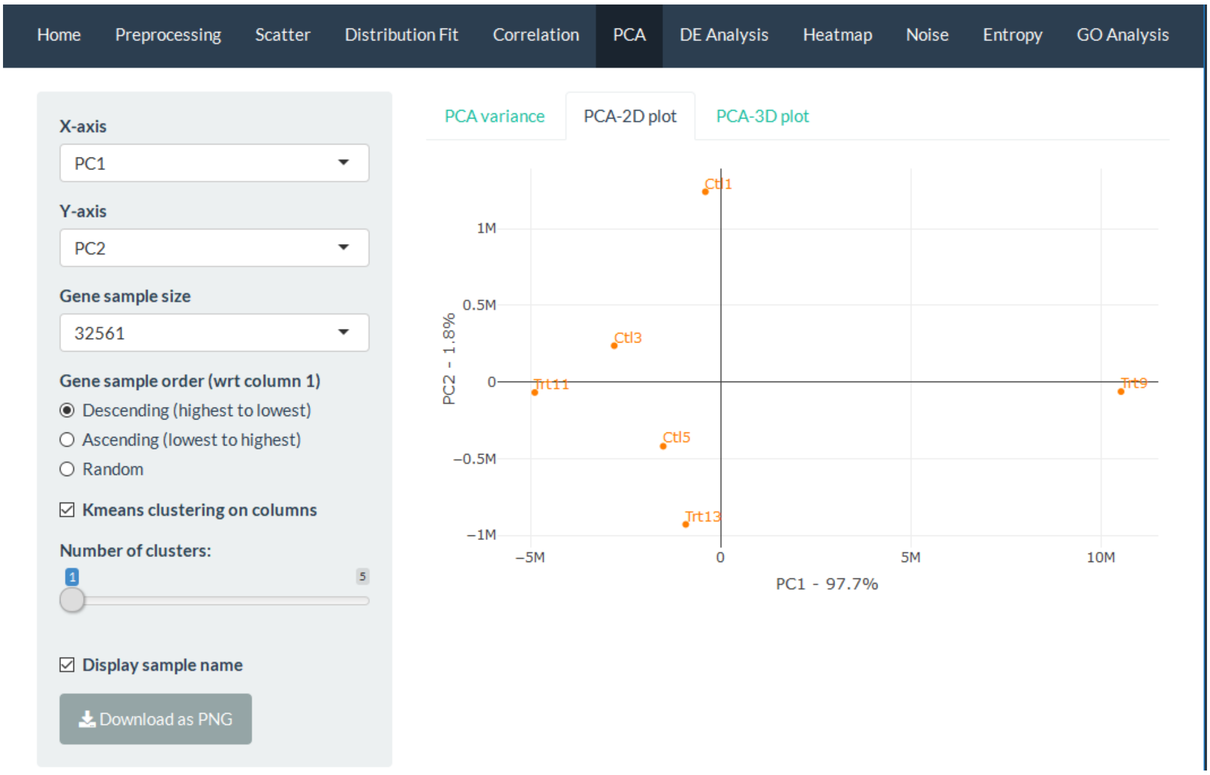


Method:
● Pearson correlation
○ Spearman correlation

Figure 7: Heatmap of Pearson correlation between all pairs of replicates

PCA and K-means clustering

PCA is a linear dimensional reduction method, which is utilized to visualize gene expression data on 2D space (the gene expression data consists of thousands of genes, corresponding to thousand-dimensional space). By default, expression data from all genes are taken to compute PC values; however, the user can visualize a subset of genes on this PC space by choosing a gene sample size and a gene sample order. For example, to visualize the top 1000 expression genes, the gene sample size to choose is 1000, and the gene sample order is descending. ABioTrans displays the variance percentage of all principal components (scree plot) in `PCA variance` tab, 2D plot of any PC-axis combination in `PCA-2D plot` tab and 3D plot in `PCA-3D plot` tab. K-means clustering is available for PCA 2-D and 3-D plots.



X-axis
PC1

Y-axis
PC2

Gene sample size
32561

Gene sample order (wrt column 1)
● Descending (highest to lowest)
○ Ascending (lowest to highest)
○ Random

☑ Kmeans clustering on columns

Number of clusters:
1

☑ Display sample name

Figure 8: RUV-normalized count data of 6 replicates on PC1-PC2 space

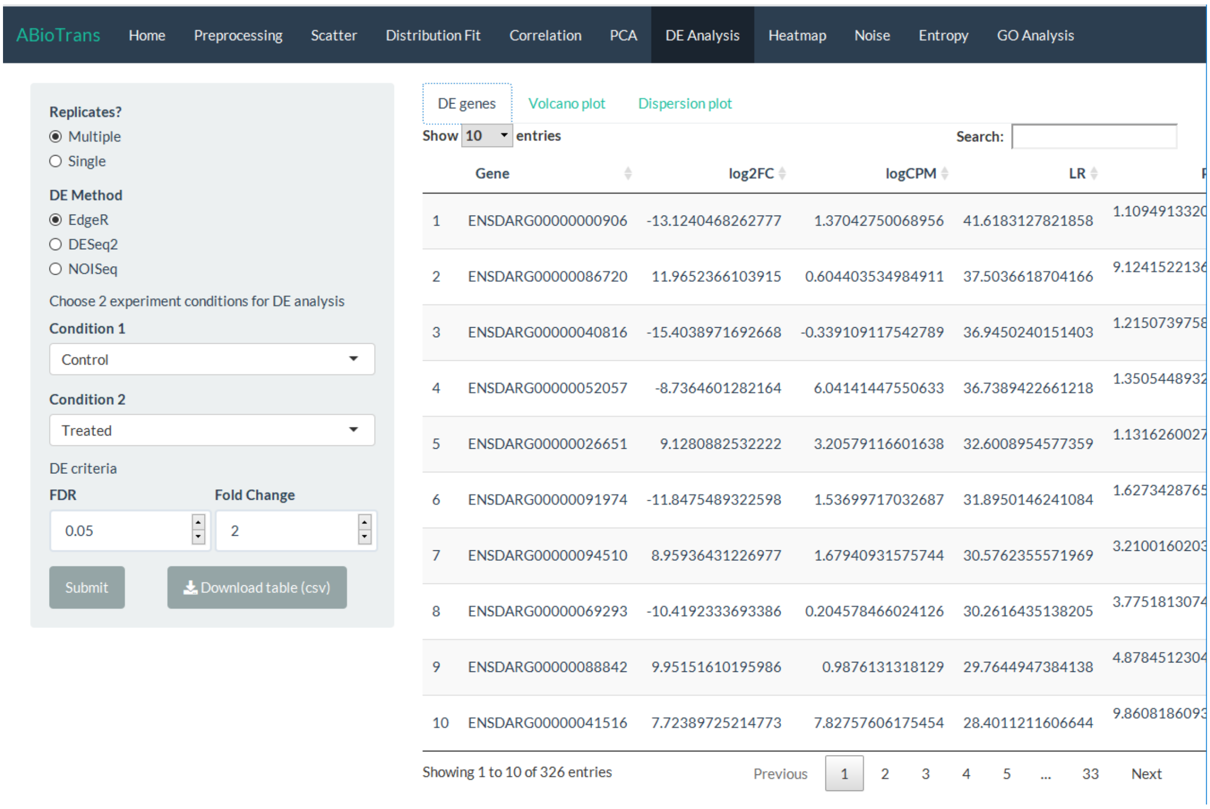
Differential Expression Analysis

ABioTrans provides 3 Differential Expression (DE) Analysis methods for multiple replicate dataset: edgeR, DESeq2 and NOISeq. For data with single replicate in all experiment condition, NOISeq method can simulate technical replicates to carry out DE test. Metadata file is required for DE Analysis. Please make sure metadata contains all column names from input data file and match them with experimental condition

For edgeR and DESeq2, raw read counts data file must be provided. For NOISeq, gene expression should be normalized for sequencing depths (by select normalization method in Preprocessing tab if raw counts file is inputted, or by directly providing normalized gene expression)

To carry out the analysis, first the user needs to specify DE methods, two conditions to compare (condition 2 is compared against condition 1), fold change cut-off value and False Discovery Rate (FDR or adjusted p-value) threshold, then hit the “Submit” button. By convention, DE genes are thresholded at FDR = 0.05 and 2-fold change. When the computation finishes, table of DE genes, volcano plot of DE result and dispersion plot of input data are displayed in their respective tabs. Please note that volcano plot and dispersion plot are only available for edgeR and DESeq2 methods.

In this demo, we perform DE analysis using edgeR; the result will also be used in Heat map and hierarchical clustering analysis.



Replicates?
● Multiple
○ Single

DE Method
● EdgeR
○ DESeq2
○ NOISeq

Choose 2 experiment conditions for DE analysis

Condition 1: Control

Condition 2: Treated

DE criteria

FDR: 0.05

Fold Change: 2

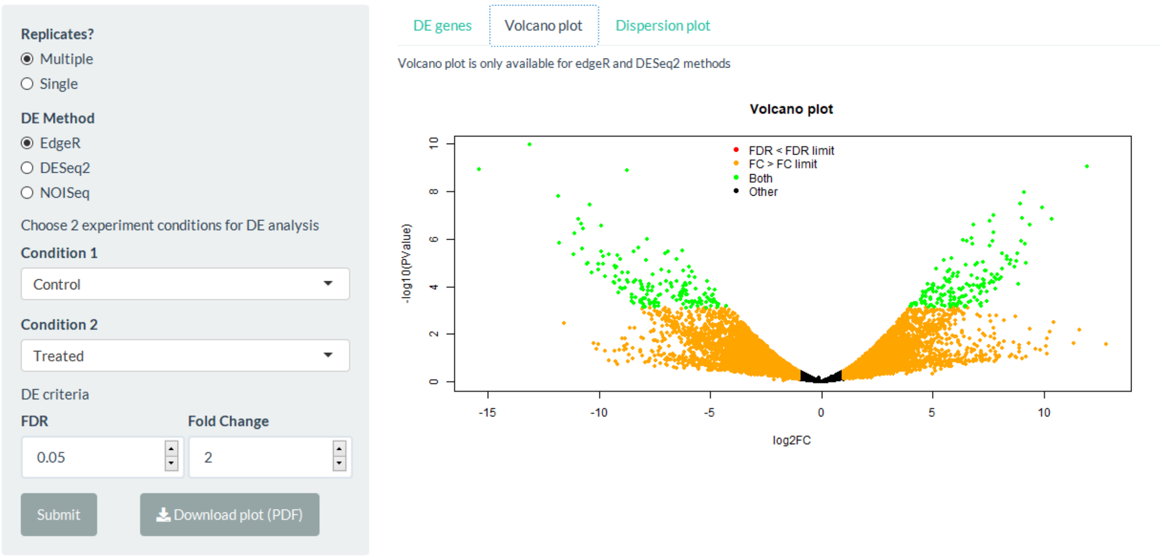
Show 10 entries

Search:

	Gene	log2FC	logCPM	LR	P
1	ENSDARG00000000906	-13.1240468262777	1.37042750068956	41.6183127821858	1.1094913320
2	ENSDARG00000086720	11.9652366103915	0.604403534984911	37.5036618704166	9.1241522136
3	ENSDARG00000040816	-15.4038971692668	-0.339109117542789	36.9450240151403	1.2150739758
4	ENSDARG00000052057	-8.7364601282164	6.04141447550633	36.7389422661218	1.3505448932
5	ENSDARG00000026651	9.1280882532222	3.20579116601638	32.6008954577359	1.1316260027
6	ENSDARG00000091974	-11.8475489322598	1.53699717032687	31.8950146241084	1.6273428765
7	ENSDARG00000094510	8.95936431226977	1.67940931575744	30.5762355571969	3.2100160203
8	ENSDARG00000069293	-10.4192333693386	0.204578466024126	30.2616435138205	3.7751813074
9	ENSDARG00000088842	9.95151610195986	0.9876131318129	29.7644947384138	4.8784512304
10	ENSDARG00000041516	7.72389725214773	7.82757606175454	28.4011211606644	9.8608186093

Showing 1 to 10 of 326 entries

Figure 9a: Table of DE genes computed by edgeR analysis method between Treated and Control conditions.



Replicates?
● Multiple
○ Single

DE Method
● EdgeR
○ DESeq2
○ NOISeq

Choose 2 experiment conditions for DE analysis

Condition 1: Control

Condition 2: Treated

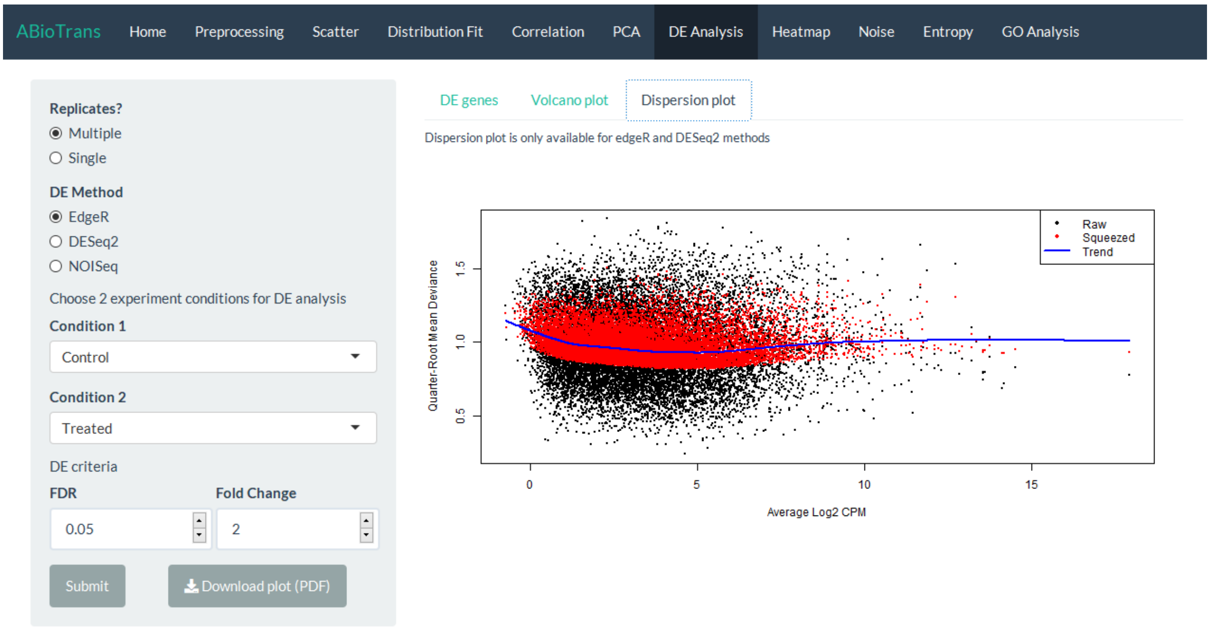
DE criteria

FDR: 0.05

Fold Change: 2

Volcano plot is only available for edgeR and DESeq2 methods

Figure 9b: Volcano plot from edgeR DE analysis between Treated and Control conditions



Replicates?
● Multiple
○ Single

DE Method
● EdgeR
○ DESeq2
○ NOISeq

Choose 2 experiment conditions for DE analysis

Condition 1: Control

Condition 2: Treated

DE criteria

FDR: 0.05

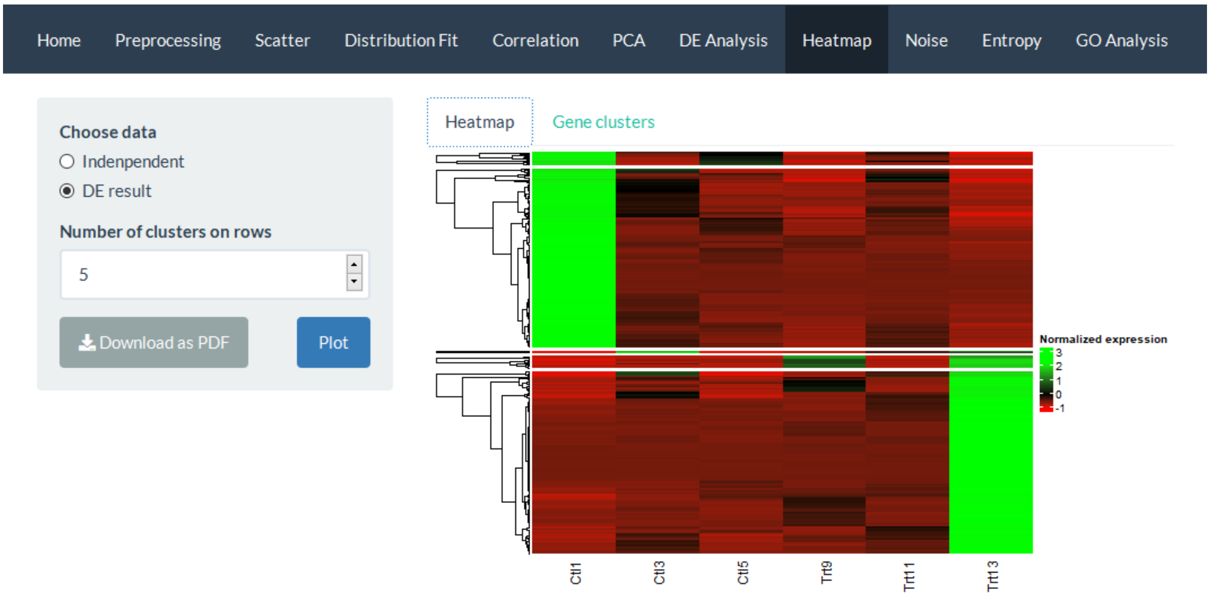
Fold Change: 2

Dispersion plot is only available for edgeR and DESeq2 methods

Figure 9c: Plot of dispersion generated by edgeR method

Heatmap and clustering

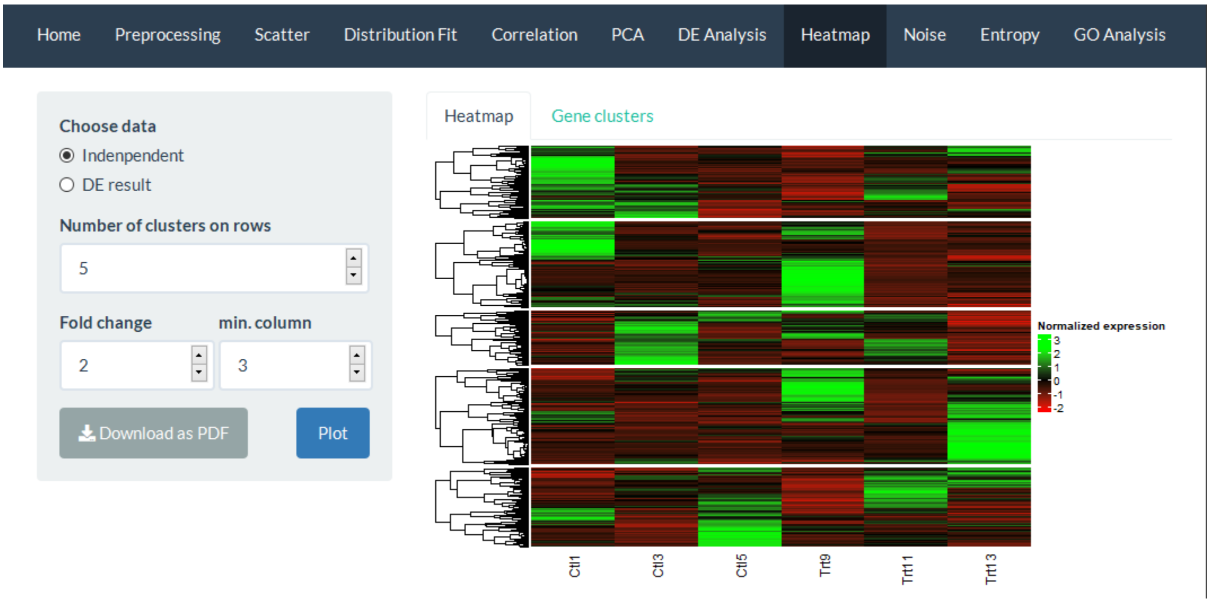
In this demo, we apply hierarchical clustering on the output of DE analysis using edgeR in the previous section (326 genes). Alternatively, the user can carry out clustering independently without going through DE analysis by specifying the minimum fold change of gene expression between 2 samples. ABioTrans also lists the name of genes for each cluster in the `Gene clusters` tab



Choose data
○ Indenpendent
● DE result

Number of clusters on rows: 5

Figure 10a: Heatmap and hierarchical clustering (5 clusters) on DE genes resulted from edgeR DE analysis



Choose data

● Indenpendent

○ DE result

Number of clusters on rows: 5

Fold change: 2

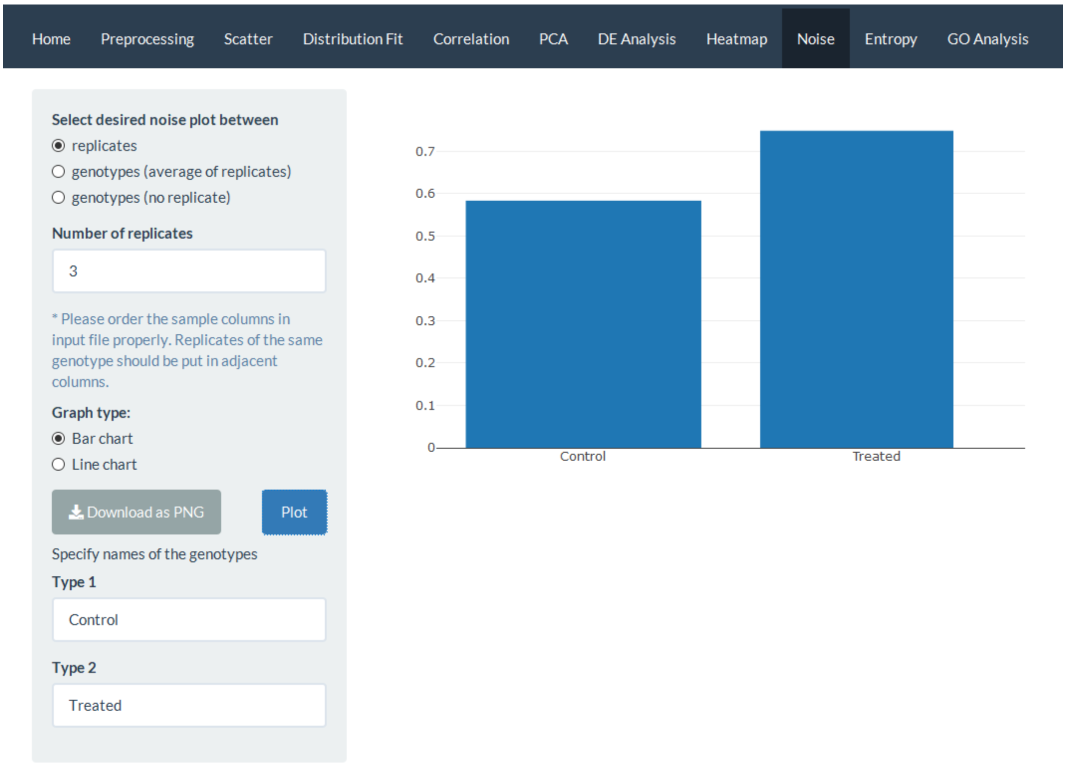
min. column: 3

Figure 10b: Heatmap and hierarchical clustering (5 clusters) on genes with at least 2-fold change in minimum 3 columns (carried out independently from DE analysis)

Noise analysis

Noise indicates the variability among samples of the same experimental condition. The noise is computed as the squared coefficient of variation [5], defined as the variance (σ^2) of expression divided by the square mean expression (μ^2), for all genes between all possible pairs of samples [5]

The dataset used in this demo has 3 replicates each for “Control” and “Treated” condition. For replicates option, noise of each condition are displayed. For genotype (average of replicates) option, the noise of Treated condition is computed based on variance of Treated condition against Control condition (which is called noise of Treated condition against Control condition). Similarly, genotypes (no replicate) is to compute the noise of every replicate against one anchor sample



Select desired noise plot between

● replicates

○ genotypes (average of replicates)

○ genotypes (no replicate)

Number of replicates

3

* Please order the sample columns in input file properly. Replicates of the same genotype should be put in adjacent columns.

Graph type:

● Bar chart

○ Line chart

Specify names of the genotypes

Type 1

Control

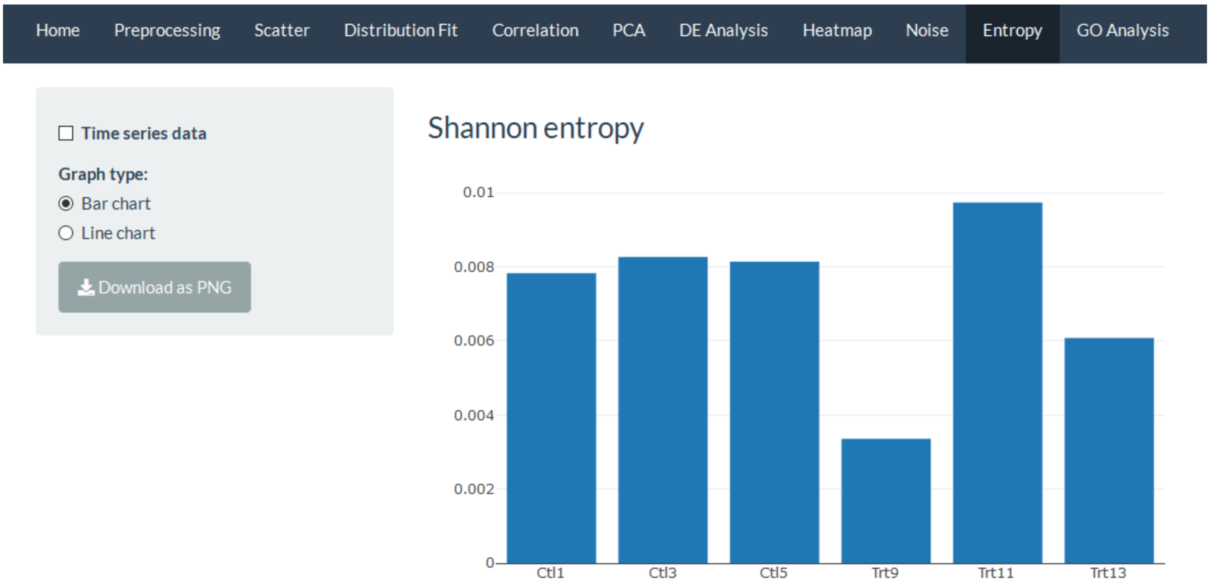
Type 2

Treated

Figure 11: Noise within Control and Treated groups

Entropy analysis

Shannon entropy of gene expression in each replicate (X) is defined as $H(X) = -\sum_{i-1}^{n} p(x_i)\log_2 p(x_i)$, where $p(x_i)$ represents the probability of gene expression value $x = x_i$. Entropy values were obtained through histogram-based partitioning approach and the number of bins is determined using Doane's rule: $b(X) = 1 + \log_2 n + \log_2\left(1 + \frac{|g_X|}{\sigma_g}\right)$, where g_X is the skewness of the expression distribution of each sample, and $\sigma_g = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}}$ [5].



☐ Time series data

Graph type:
● Bar chart
○ Line chart

Figure 12: Entropy of all replicates

Gene ontology analysis

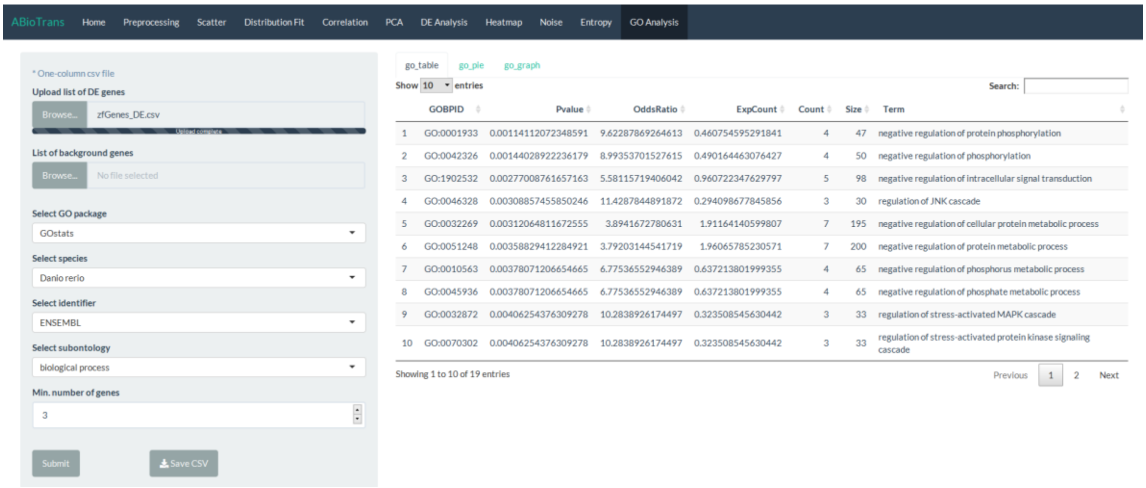
This analysis results in a list significantly enriched Gene Ontology terms. The user can select among 3 gene ontology enrichment test: clusterProfiler and GOstats, and enrichR:

- Both ClusterProfiler and GOstats uses hypergeometric over-representation test. User need to specify the species, gene identifier and sub-ontology before proceeding to the analysis. clusterProfiler method also implements Benjamini & Hochberg method for p-value correction, and apply a threshold of 0.01 for both p-value, q-value (False Discovery rate) and adjusted p-value. On the other hand, GOstats uses a non-conditional test, with p-value = 0.01 being the only cut-off criteria (hence, GOstats is less stringent than clusterProfiler)
- enrichR requires input gene IDs to be in SYMBOL format. Afterwards user needs to specify the database to perform the analysis.

ABioTrans also display a pie chart to visualize the relative size of all level-2 ontology terms associated to the gene set in `go_pie` tab. Note that the pie chart simply displays all GO terms associated to every single gene, which is not the result from the over representation test. Also, there will be overlapping genes in each term displayed (since one gene usually take part in multiple biological functions).

ABioTrans also provides a network visualization to display the mapping between the enriched GO terms and their respective genes in `go_graph` tab. Please note that the network graph is only available when clusterProfiler method is used.

In this demo, edgeR DE analysis result is saved to .csv file to local drive (file named *zfGenes_DE.csv* from test data folder), and then loaded to GO Analysis tab as list of DE genes. The data used in this demo is in ENSEMBL format, from Danio rerio species (Zebra fish). The demo performs a GOstats test with biological process GO terms



* One-column csv file

Upload list of DE genes: zfGenes_DE.csv (Upload complete)

List of background genes: No file selected

Select GO package: GOstats

Select species: Danio rerio

Select identifier: ENSEMBL

Select subontology: biological process

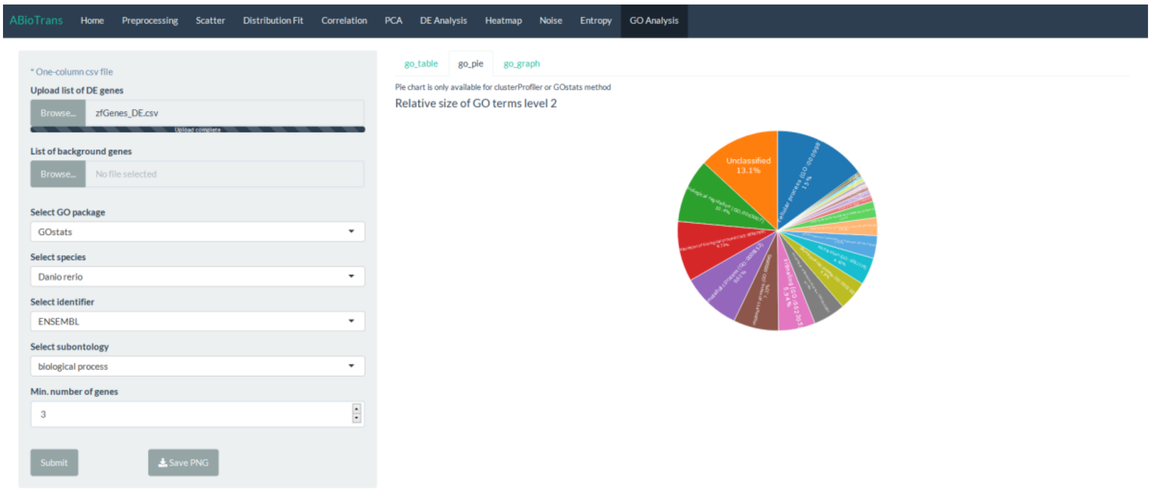
Min. number of genes: 3

Show 10 entries   Search:

	GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
1	GO:0001933	0.00114112072348591	9.62287869264613	0.460754595291841	4	47	negative regulation of protein phosphorylation
2	GO:0042326	0.00144028922236179	8.99353701527615	0.490164463076427	4	50	negative regulation of phosphorylation
3	GO:1902532	0.00277008761657163	5.58115719406042	0.960722347629797	5	98	negative regulation of intracellular signal transduction
4	GO:0046328	0.00308857455850246	11.4287844891872	0.294098677845856	3	30	regulation of JNK cascade
5	GO:0032269	0.00312064811672555	3.8941672780631	1.91164140599807	7	195	negative regulation of cellular protein metabolic process
6	GO:0051248	0.00358829412284921	3.79203144541719	1.96065785230571	7	200	negative regulation of protein metabolic process
7	GO:0010563	0.00378071206654665	6.77536552946389	0.637213801999355	4	65	negative regulation of phosphorus metabolic process
8	GO:0045936	0.00378071206654665	6.77536552946389	0.637213801999355	4	65	negative regulation of phosphate metabolic process
9	GO:0032872	0.00406254376309278	10.2838926174497	0.323508545630442	3	33	regulation of stress-activated MAPK cascade
10	GO:0070302	0.00406254376309278	10.2838926174497	0.323508545630442	3	33	regulation of stress-activated protein kinase signaling cascade

Showing 1 to 10 of 19 entries

Figure 13a: Enriched biological processes in DE genes (that resulted from edgeR DE analysis)



* One-column csv file

Upload list of DE genes: zfGenes_DE.csv (Upload complete)

List of background genes: No file selected

Select GO package: GOstats

Select species: Danio rerio

Select identifier: ENSEMBL

Select subontology: biological process

Min. number of genes: 3

Pie chart is only available for clusterProfiler or GOstats method

Relative size of GO terms level 2

Figure 13b: Pie chart of all associated biological processes to input DE gene set.