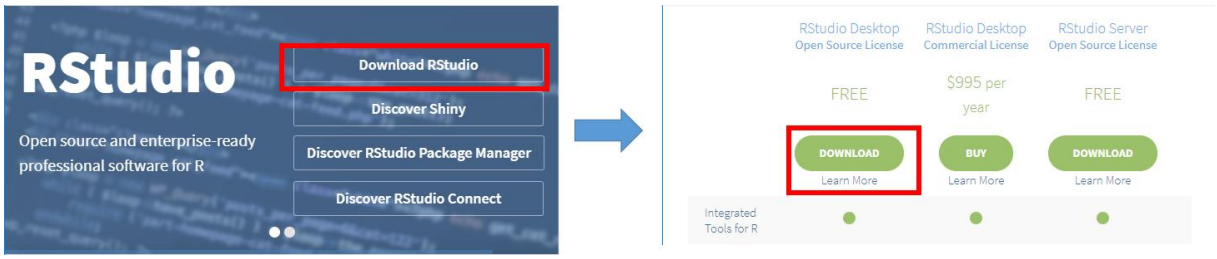
ABioTrans - A Biostatistical tool for Transcriptomics Analysis

User manual

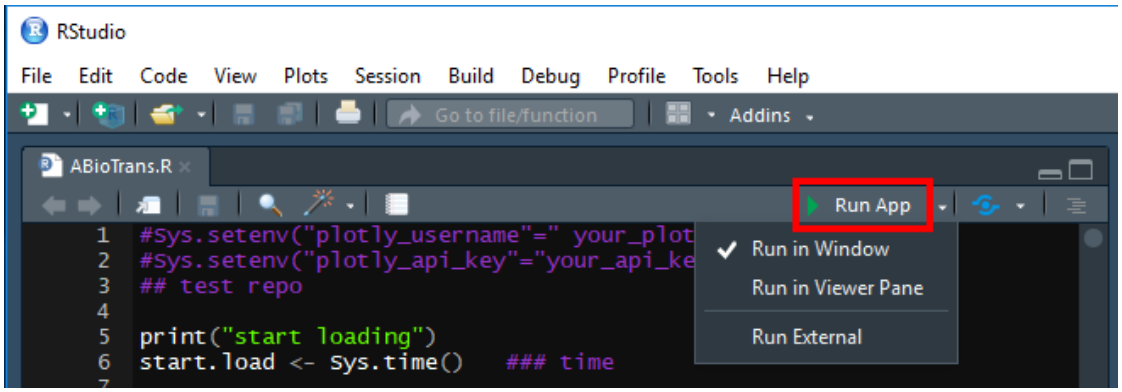
Updated 19 Feb 2019

Setup

1. Install Rstudio from https://www.rstudio.com/



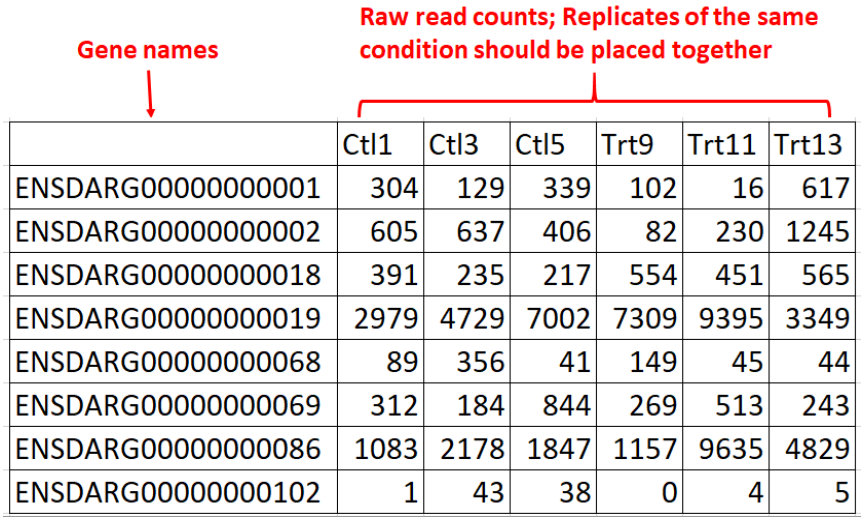
2. Download ABioTrans-master.zip on GitHub and unzip it. Please do not modify www inside ABioTrans folder.
3. Open the ABioTrans.R file using RStudio and click RunApp button on the top-right corner. `Run External` opens ABioTrans GUI in your default browser (recommended), whereas `Run in Window` opens the GUI as a RStudio window.



You can start your analysis now!

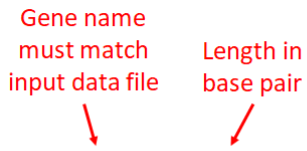
Uploading files

ABiotrans requires all input files in .csv format. The data file, in comma-separated value (.csv) format, should contain the gene names in rows and genotypes (conditions: wildtype, mutants, replicates, etc.) in columns, following the usual format of files deposited in the 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 (eg. 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/click all available analysis buttons (scatter plot, distribution fit, Pearson Correlation, etc.) once a data file is loaded (whether normalized or in raw count).



	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 1a: Data file format in raw counts

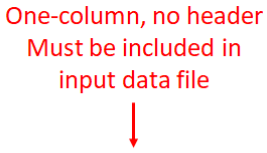


geneID	length
aaeA	933
aaeB	1968
aaeR	930
aaeX	204
aas	2160
aat	705

Figure 1b: Gene length file format

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

Figure 1c: Metadata file format



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

Figure 1d: Negative control gene file format

4. 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 to csv files in test data folder

https://github.com/buithuytien/ABioTrans/tree/master/Test%20data . Once data files and supporting files are uploaded, user can press Submit button and ABioTrans will jump to the next tab `Preprocessing`. In case user accidentally upload the wrong data file, user can overwrite each of them by uploading new files, or press the Reset button to erase all uploaded files.

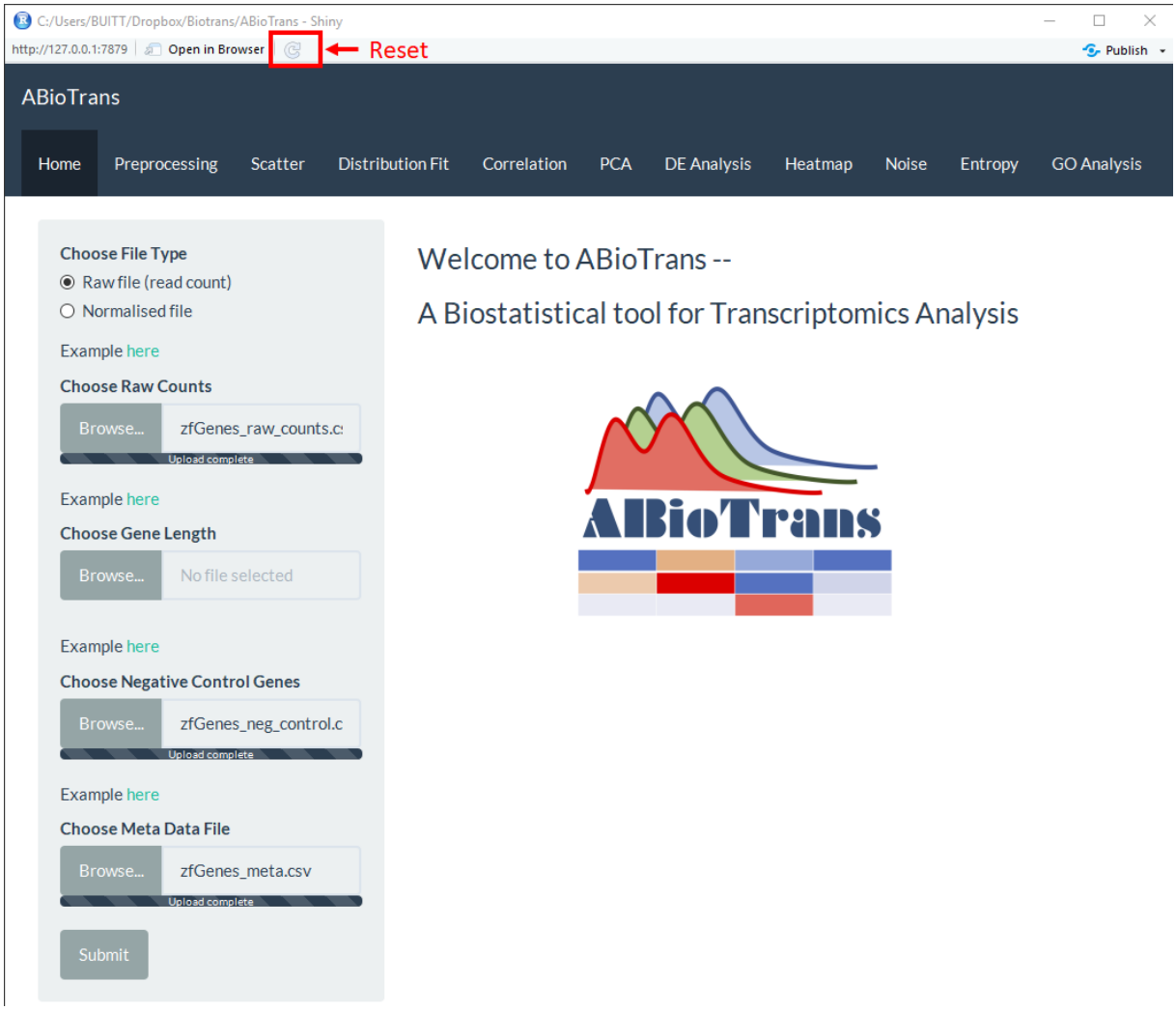


Figure 2: Home page and file upload – zebra fish data

Preprocessing

Preprocessing involves 2 step: removing lowly expressed genes and normalization. First, you need to specify the cut-off expression values (same unit to your input data file - either raw read counts or normalized expression), and the minimum number of columns (samples / replicate) whose expression is above threshold value. Normalization methods are available upon your input of supporting data files (gene length and negative control genes). Relative Log Expression (RLE) plot of raw and processed data are displayed to visualize the effects of normalization.

In this demo, we carried out RUV method 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.

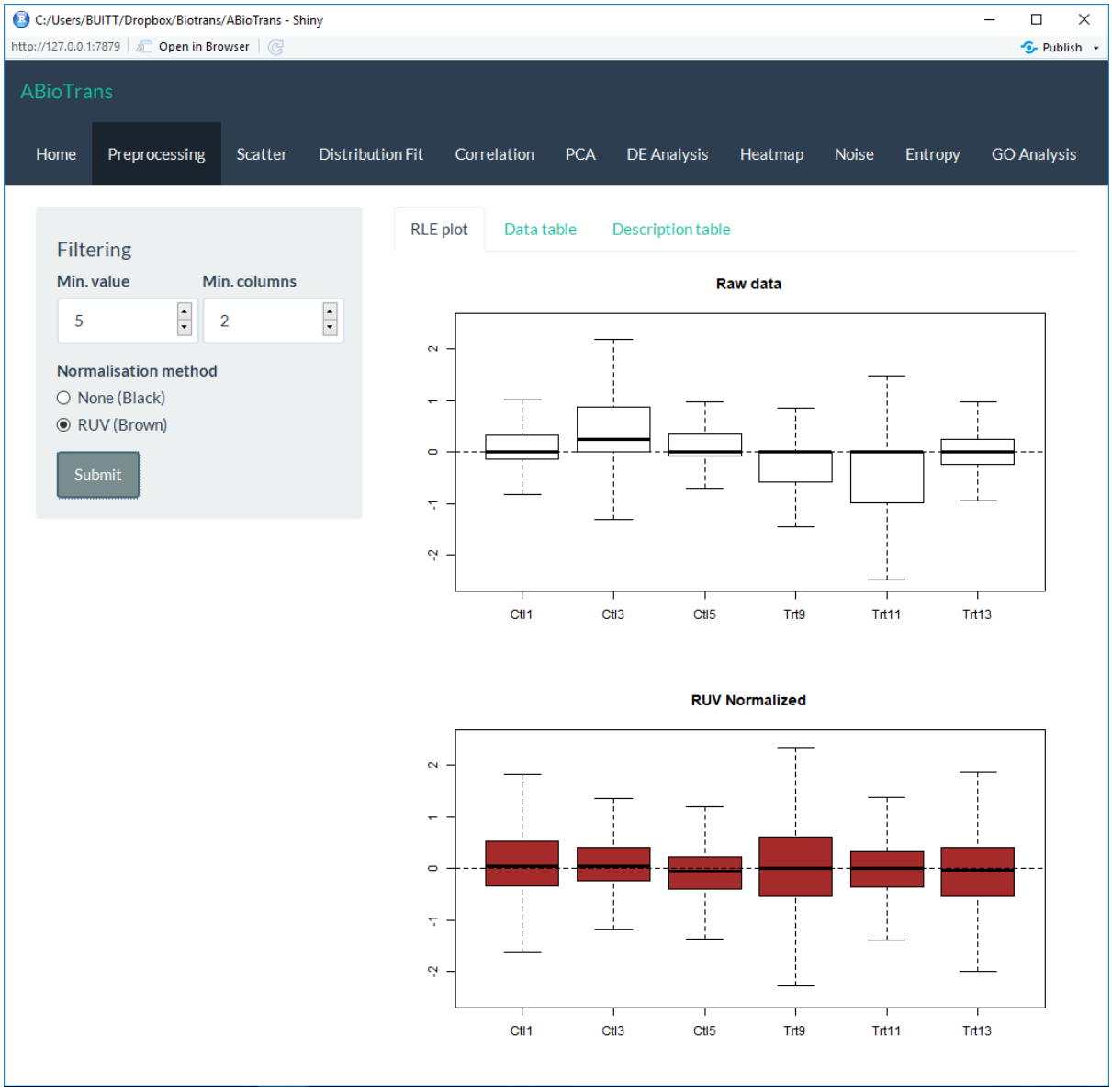


Figure 3: Preprocessing – including low count filtering and RUV normalization

Scatter plot

Scatter plot compares any 2 samples (or 2 replicates) by displaying expression of all genes in 2D space. Log transformation on the values will provide better visualization due to the skewed distribution of gene expression in nature. The colours on scatter plot refer to kernel density estimation (KDE). You can download a single plot as PDF. You can also 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

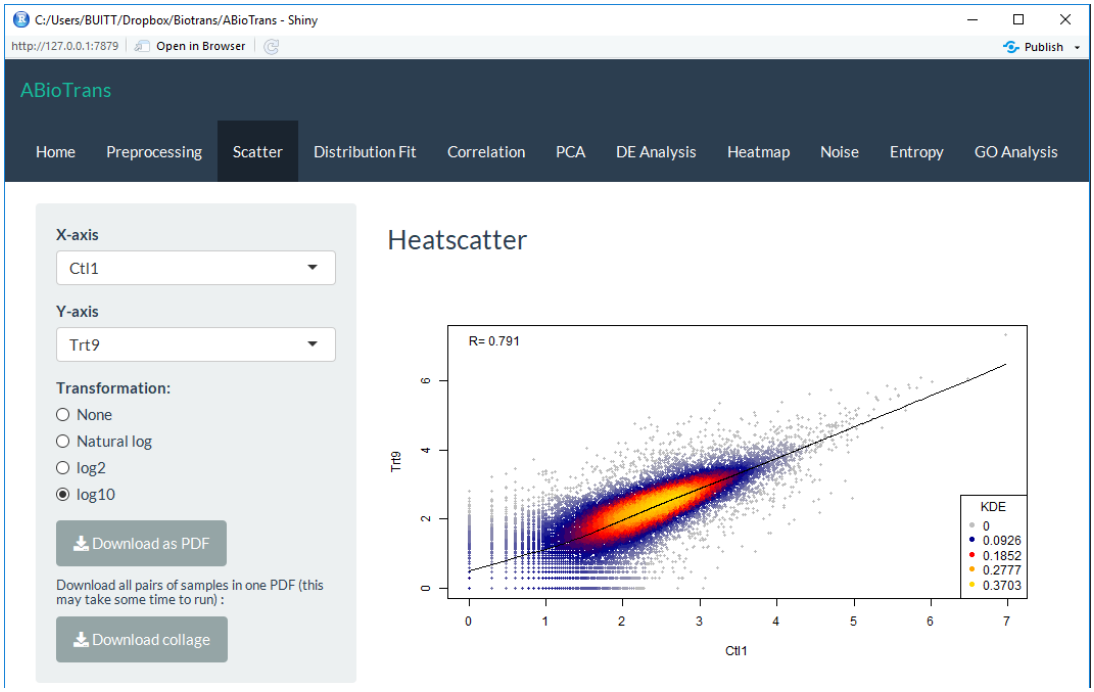


Figure 4: Scatter plot between Ctrl1 and Trt9 samples

Distribution fitting

Distribution fitting can be used a Quality Control (QC) step, in which it compares the gene expression in 1 sample to various statistical continuous distributions. Once we confirmed that the gene set follow a distribution, we can safely conclude the validity of our data. AIC table is also provided to show the best fitted distribution in each sample

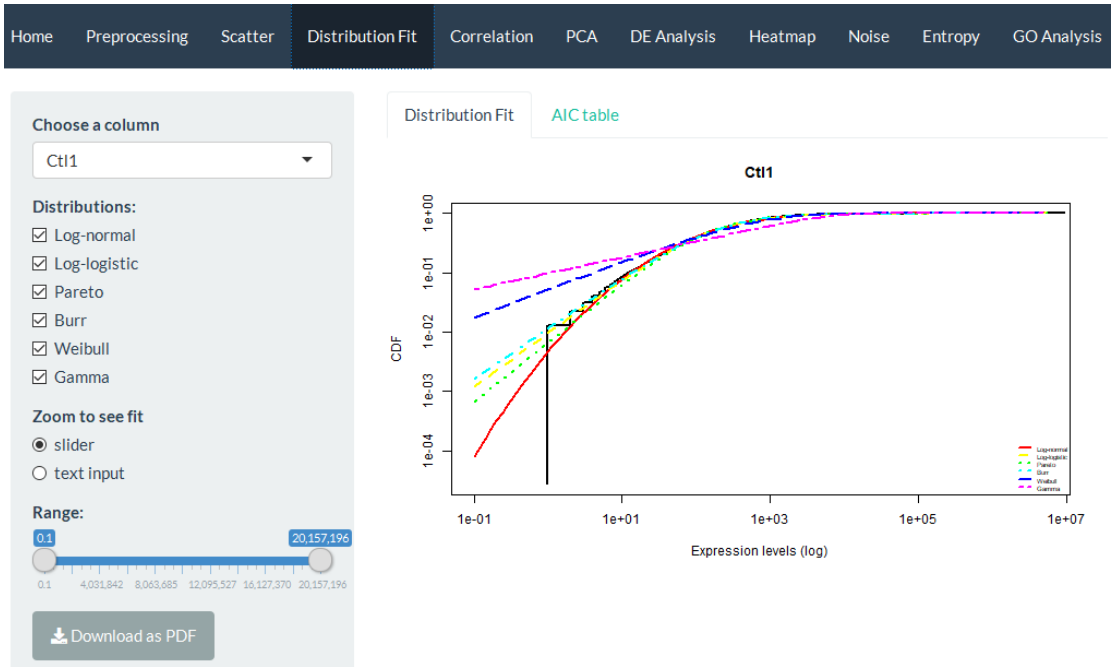
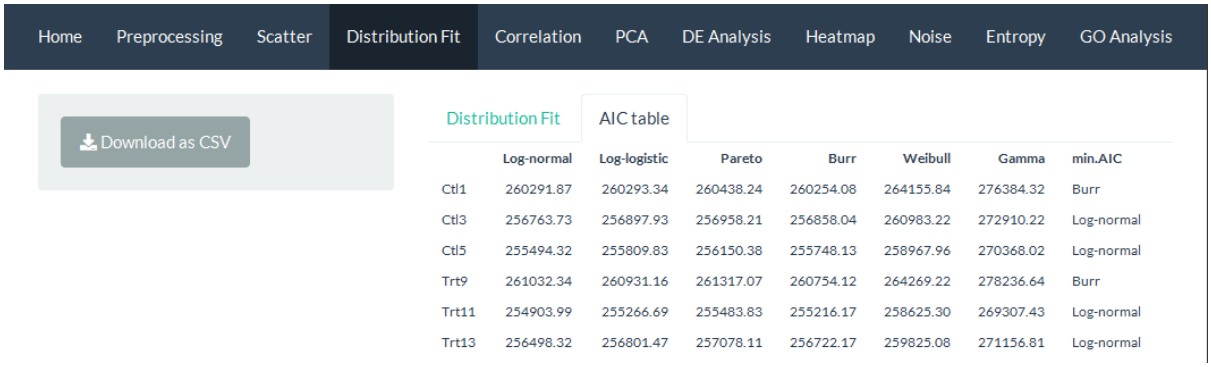


Figure 5a: Comparing Cumulative Distribution Functions of raw count data (black) with lognormal (red colour), Pareto (green colour), and Weibull (blue colour).



	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 5b: Table of AIC values for 6 distributions (Log-normal, Log-logistic, Pareto, Burr, Weibull, and Gamma) in all 6 replicates

Correlation analysis

Linear (Pearson) and monotonic non-linear (Spearman) correlations between any two samples are computed, in actual values in a table or as a density gradient plot between the samples.

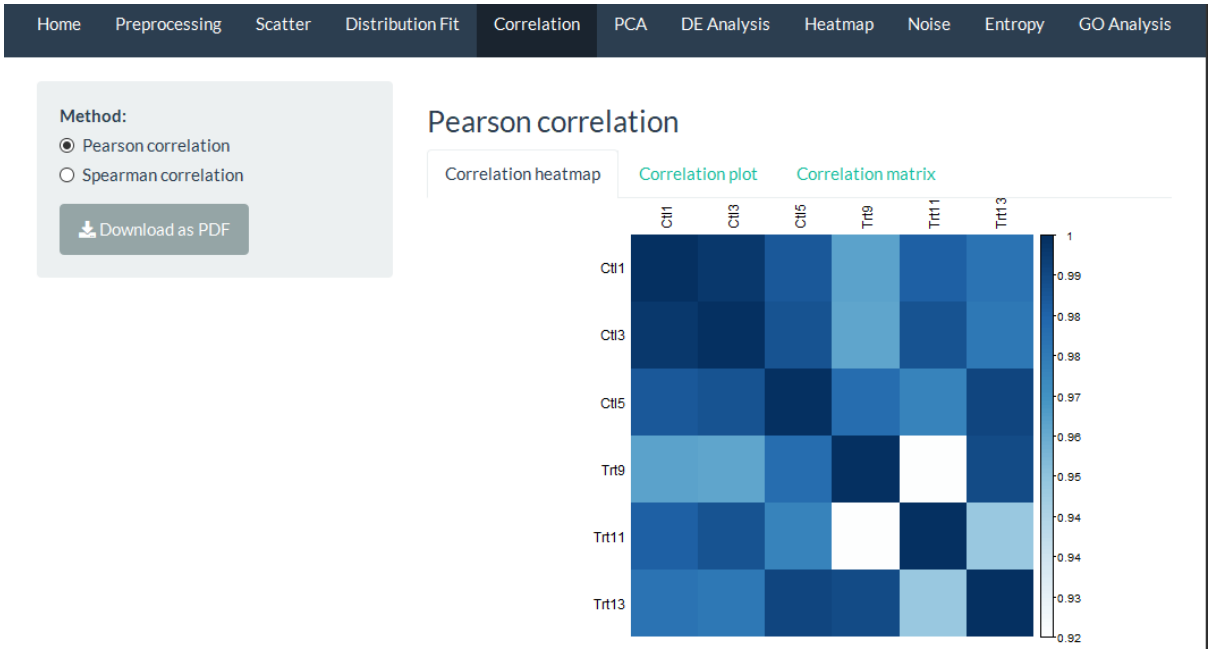


Figure 6: Heat map of Pearson correlation

PCA and K-means clustering

PCA is a linear dimensional reduction method, which is utilized to visualize our datasets on 2D space (the original data consisting of thousands of genes are in thousand-scale space). By default, data from all genes are taken to compute PC values; however, you can visualize a subset of genes on this PC space by choosing a gene sample size and a gene sample order. Then the variance percentage of all principal components (scree plot), 2-D plot of any PC-axis combination and 3-D plot will be shown. K-means clustering is available for PCA 2-D and 3-D plots.

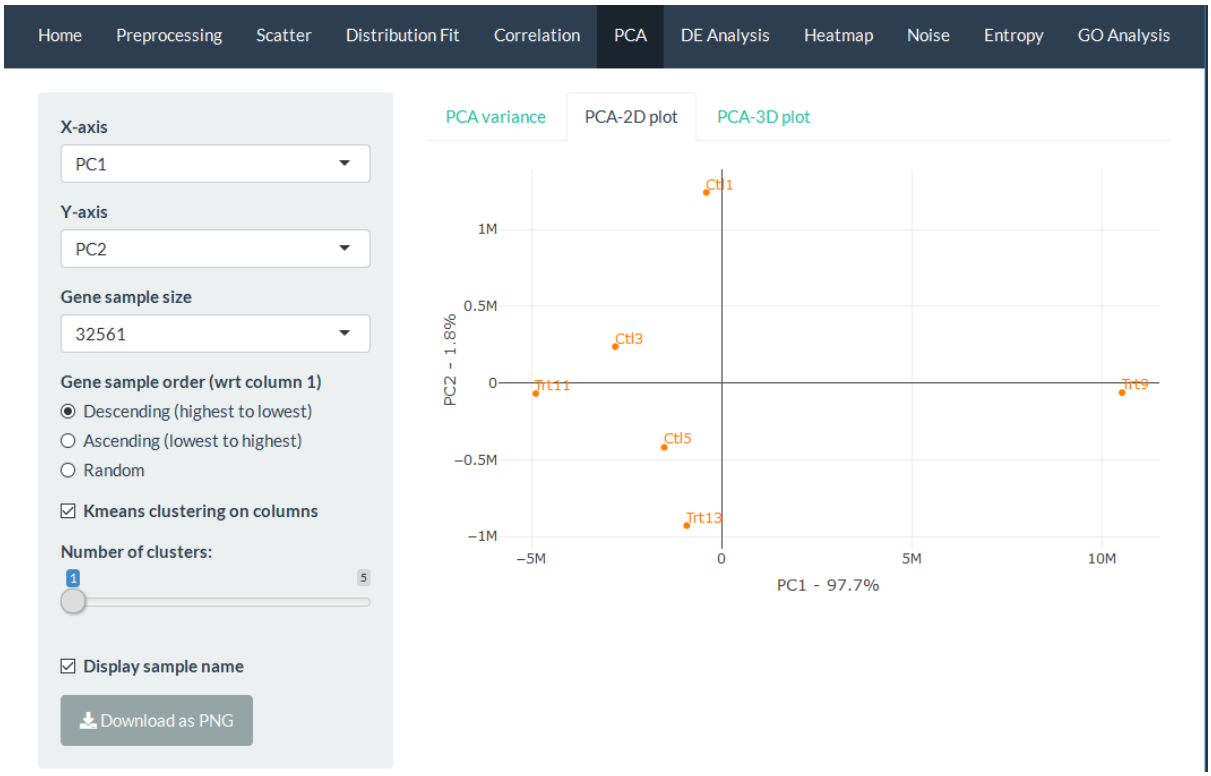


Figure 7: RUV-normalized count data of 6 replicates on PC1-PC2 space

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 you need to specify DE methods, two conditions to compare (condition 2 is compared against condition 1), and fold change and False Discovery Rate (FDR or adjusted p-value) threshold. By convention, DE genes are thresholded at 0.05 (FDR) and 2-fold change. As a result, table of DE genes, volcano plot of DE result and dispersion plot of input data are displayed. Please note that volcano plot and dispersion plot are only available for edgeR and DESeq2 methods.

We will be carrying out DE analysis using edgeR in this demo, the result will also be used in Heat map and hierarchical clustering analysis.

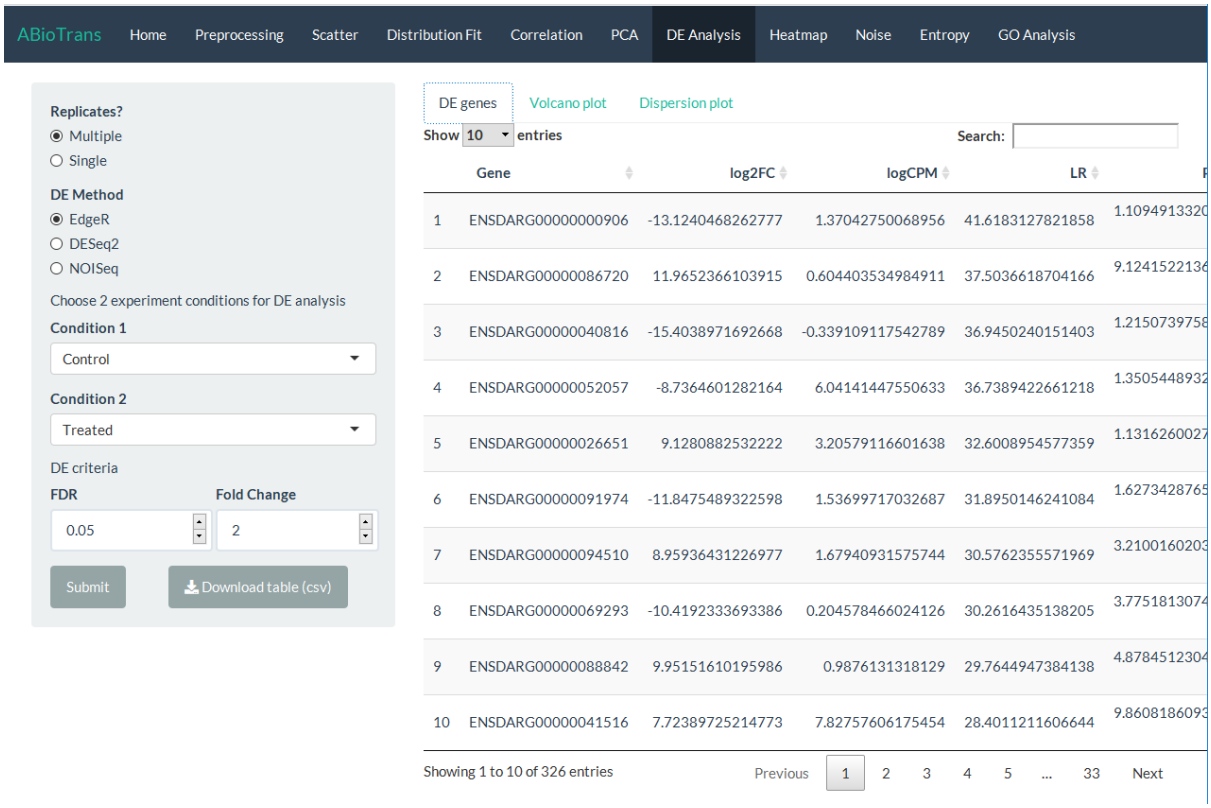


Figure 8a: DE analysis using edgeR method between Treated and Control conditions.

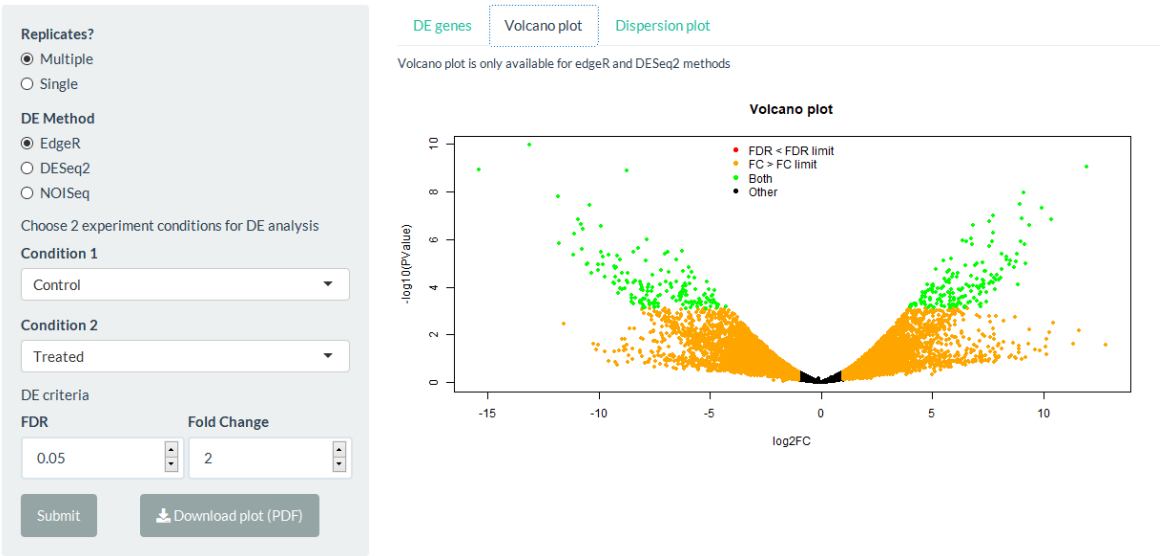


Figure 8b: Volcano plot from edgeR DE analysis result

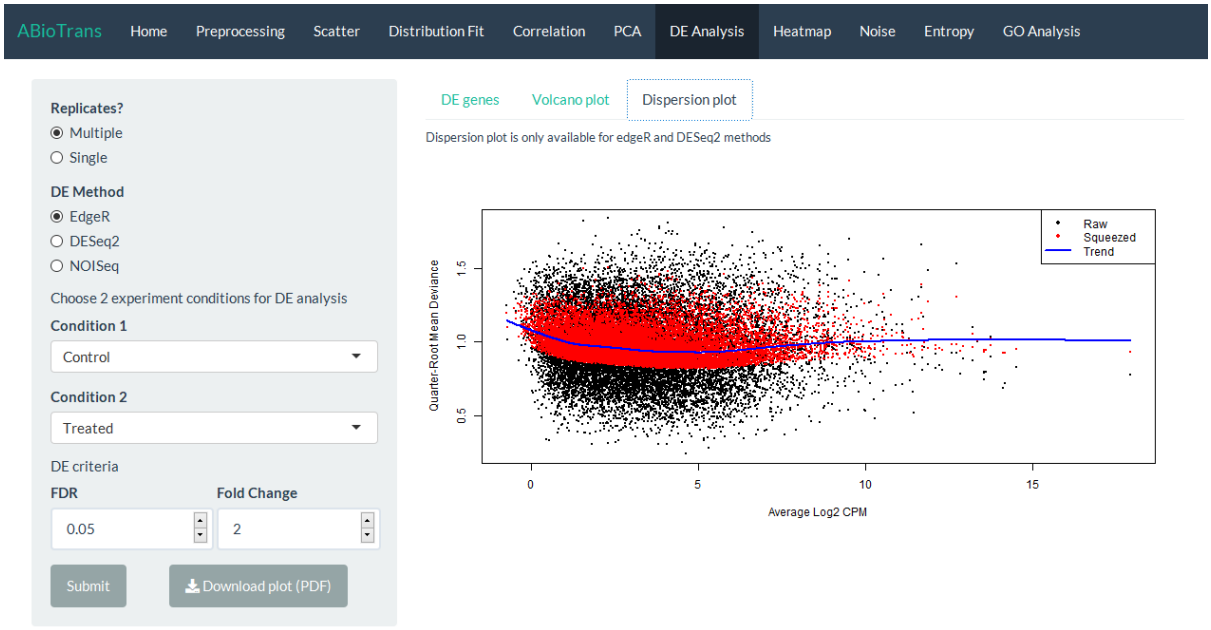


Figure 8c: Plot of dispersion generated by edgeR method

Heat map and clustering

In this demo, we apply hierarchical clustering on the output of DE analysis using edgeR (326 genes). Alternatively, 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 list the name of genes for each cluster in the `Gene clusters` tab

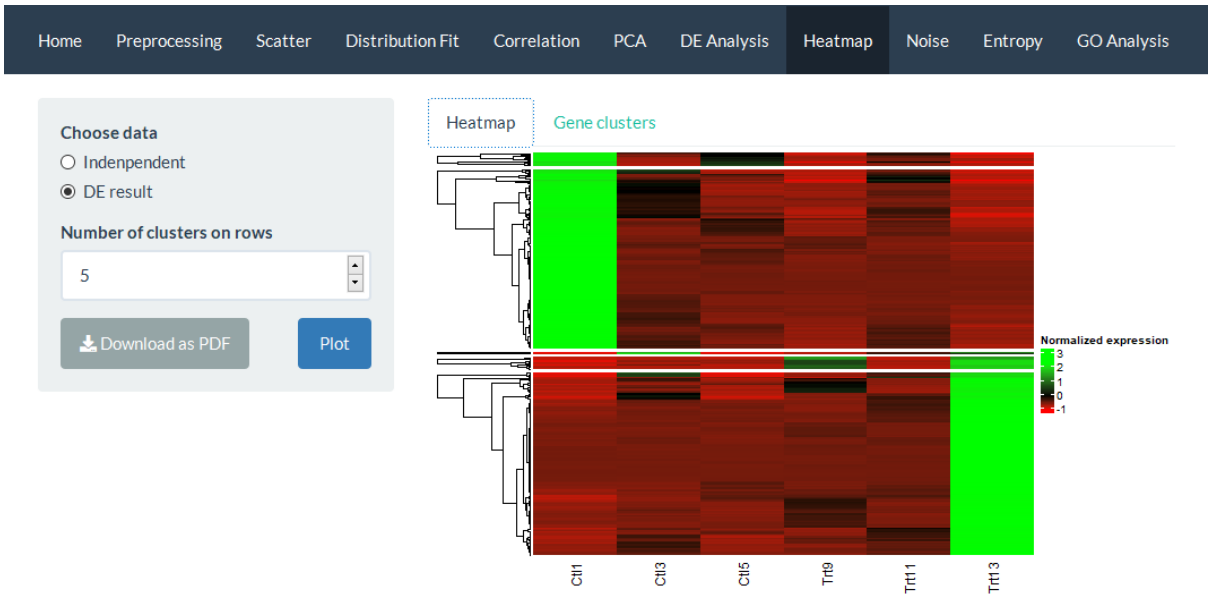


Figure 9a: Heat map and hierarchical clustering (5 clusters) on DE genes resulted from edgeR DE analysis

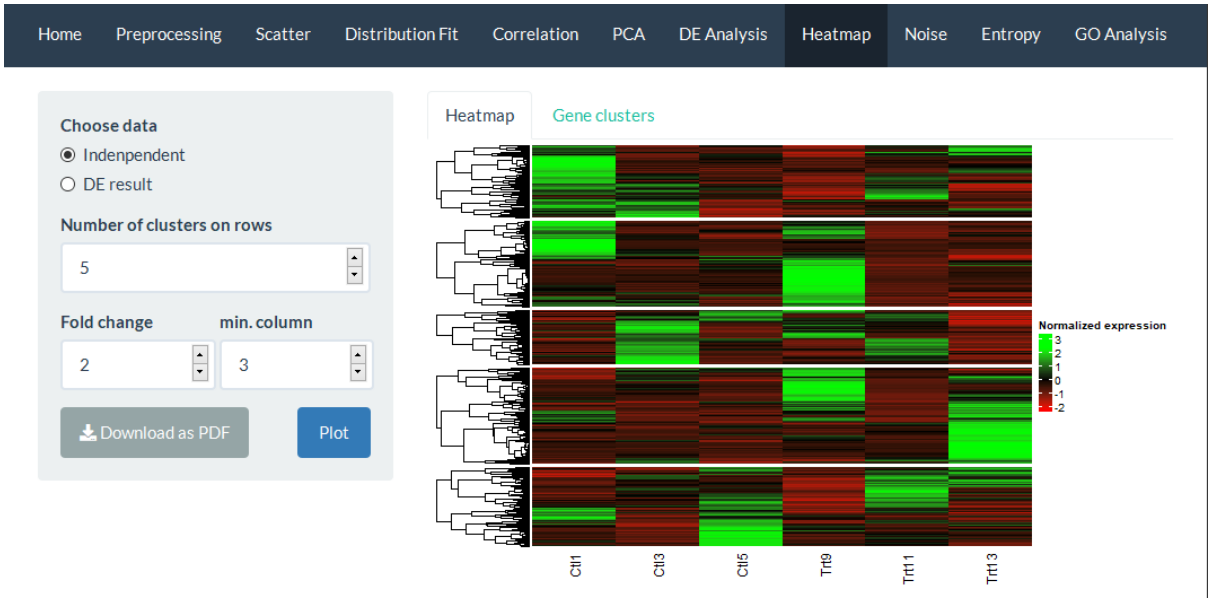


Figure 9b: Heat map and hierarchical clustering (5 clusters) on genes with at least 2-fold change in minimum 3 columns (carried out independently from DE analysis)

Noise analysis

Noise indicates the variability among samples of the same experimental condition. The noise is computed as the squared coefficient of variation [13], defined as the variance ($\sigma 2$) of expression divided by the square mean expression ($\mu 2$), 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

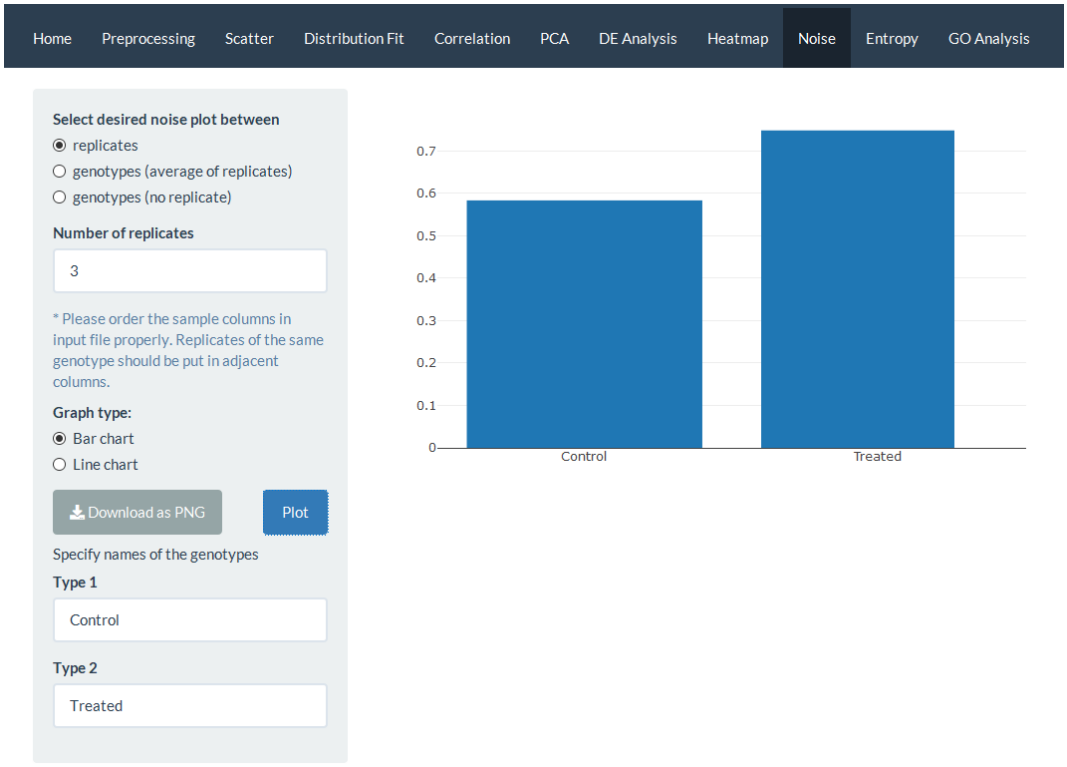


Figure 10a: Noise within Control and Treated groups

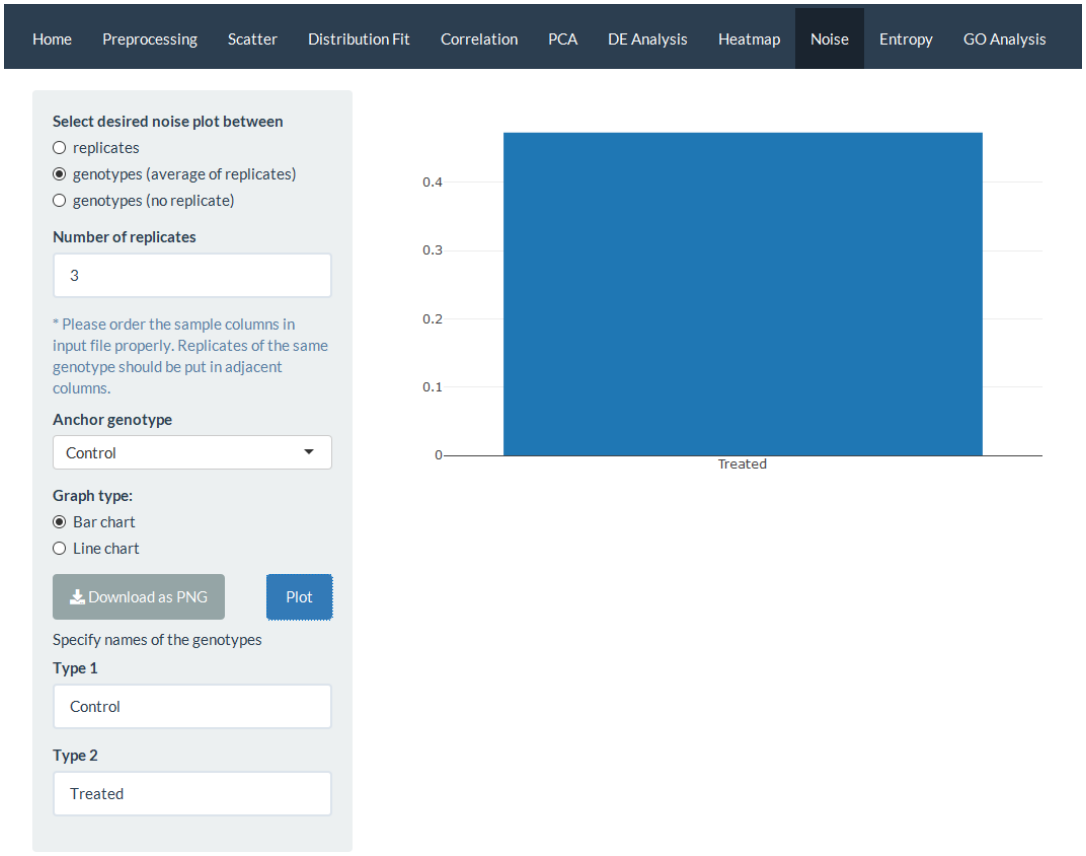


Figure 10b: Noise of Treated group against Control group

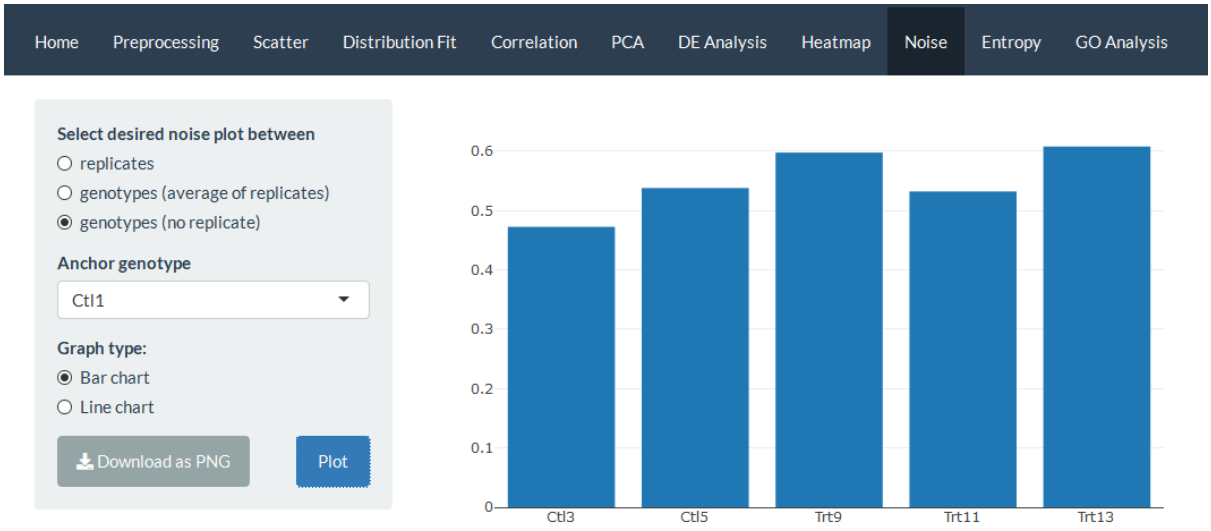


Figure 10c: Noise of all replicates against Ctl1 replicate

Entropy analysis

Shannon entropy is another measure of sample variability, which is calculated for each replicate.

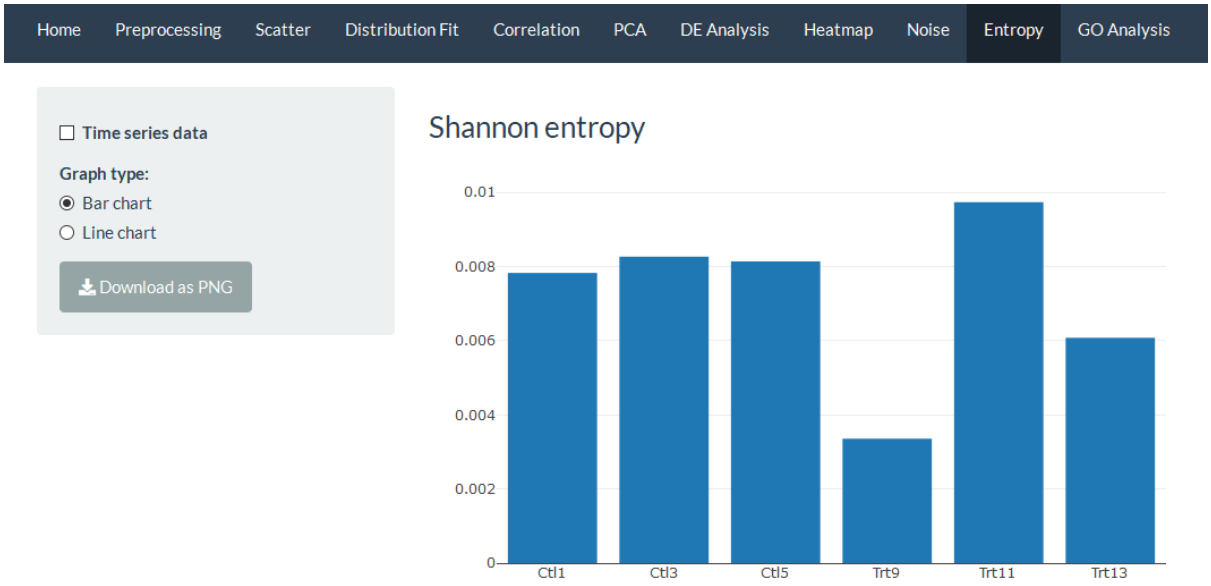


Figure 11: Entropy of all replicates

Gene ontology analysis

This function results in a list significantly enriched Gene Ontology terms. 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. Note that the over-representation test was not performed; the pie chart simply displays all terms associated to every single gene. Also, there will be overlapping genes in each term displayed (since one gene usually take part in multiple biological functions)

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

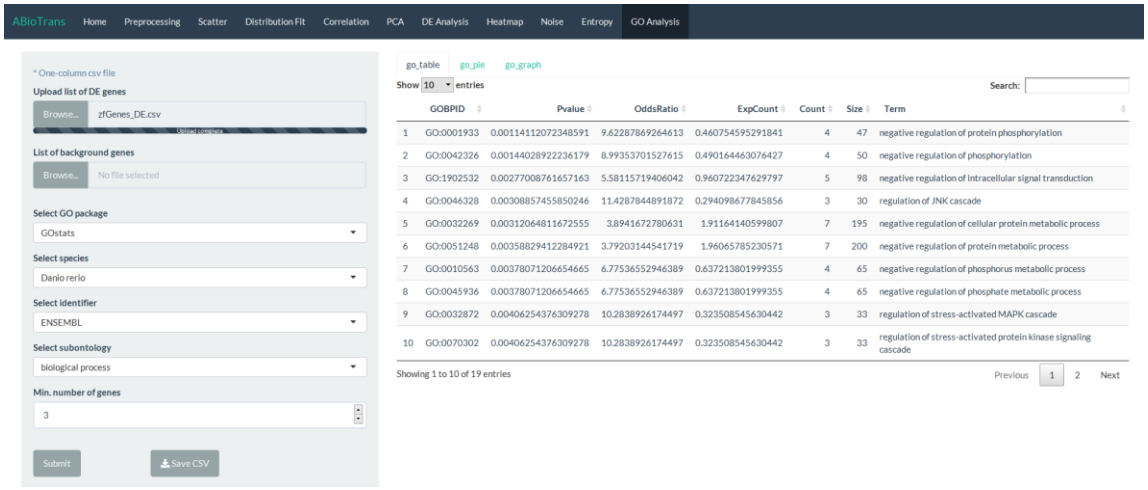


Figure 12a: Enriched biological processes in DE genes (that resulted from edgeR DE analysis)

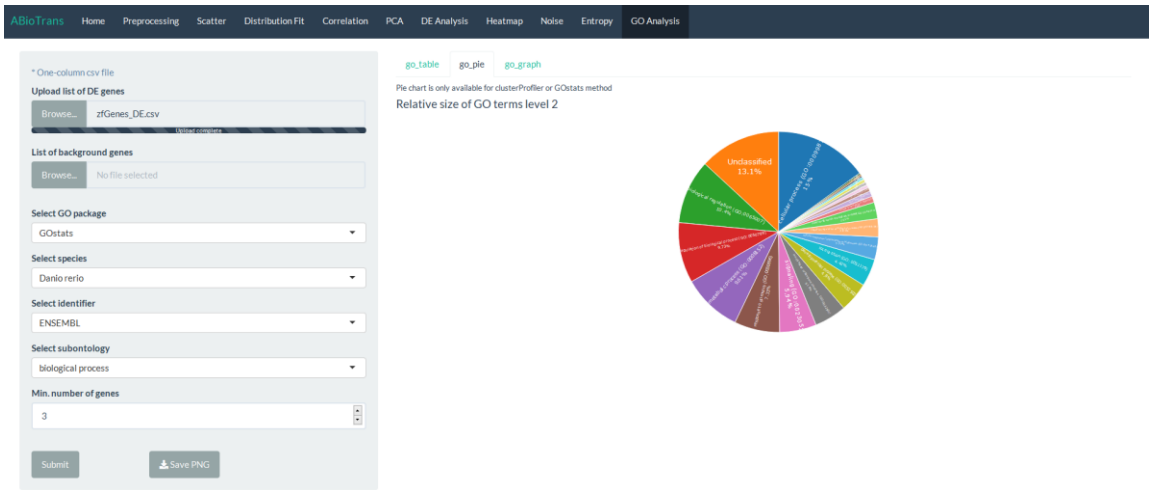


Figure 12b: Pie chart of all associated biological processes to input DE gene set.