iSeq Tutorial

-- bioinformatics core @ Peking University

About iSeq:

iSeq is an online RNA-seq data analysis tool developed by the bioinformatics core facility at Peking University. It is dedicated to functions including but not limited to differentially expressed genes detection, functional enrichment and data visualization. The interactive and graphical user interface makes everything as easy as possible.

Address: http://iseq.cbi.pku.edu.cn/

1. Upload

The very first step is to initiate gene expression data by uploading two files. The expression file is requisite while the condition file is optional. They should be in the comma separated values (CSV) format and thus terminate their file names with ".csv". Microsoft Excel automatically generates files in such format.

1.1 Expression File of Genes

Each row represents a certain gene and each column represents a certain sample. The value of each entity represents the expression level measured in FPKM or TPM. The first column and the first row should list the names of genes and samples, respectively.

Sample file: http://202.205.131.32:3838/NAT.rawReads.csv

	A	B	C	D	E	F	G	H
1		AD1	AD2	AD3	EM1	EM2	EM3	EM4
2	0610005C13Rik	9	14	21	17	30	19	23
3	0610007P14Rik	428	508	404	374	527	539	501
4	0610009B22Rik	255	257	253	169	280	230	224
5	0610009L18Rik	49	45	60	48	57	60	61
6	0610009O20Rik	676	786	691	617	755	709	712
7	0610010B08Rik	14	14	7	7	5	10	5
8	0610010F05Rik	1204	1099	950	2514	2752	2394	3376
9	0610010K14Rik	390	485	432	727	1167	1001	804
10	0610011F06Rik	347	444	426	185	296	261	232
11	0610012G03Rik	631	794	745	488	628	672	581

1.2 Condition File of Samples

This file allows classifying samples into biological conditions. It should have two rows. The first row lists sample names and the second row lists condition names. Please note to match the sample names with those given in the expression file.

Sample file: http://202.205.131.32:3838/NAT.Condition.csv

	A	B	C	D	E	F	G	H
1	Sample	AD1	AD2	AD3	EM1	EM2	EM3	EM4
2	Condition	AD	AD	AD	EM	EM	EM	EM

Sample files are from http://www.nature.com/neuro/journal/v16/n4/full/nn.3332.html

You may also compare your results with those published in this article.

After successful uploading you will see something like this:

	AD_1	AD_0	AD_2	EM_1	EM_0	EM_2	EM_3
Condition	AD	AD	AD	EM	EM	EM	EM

Show 10 entries

Search:

	AD1	AD2	AD3	EM1	EM2	EM3	EM4
0610005C13Rik	9	14	21	17	30	19	23
0610007P14Rik	428	508	404	374	527	539	501
0610009B22Rik	255	257	253	169	280	230	224
0610009L18Rik	49	45	60	48	57	60	61
0610009O20Rik	676	786	691	617	755	709	712
0610010B08Rik	14	14	7	7	5	10	5
0610010F05Rik	1204	1099	950	2514	2752	2394	3376
0610010K14Rik	390	485	432	727	1167	1001	804
0610011F06Rik	347	444	426	185	296	261	232
0610012G03Rik	631	794	745	488	628	672	581

Showing 1 to 10 of 24,015 entries

Previous 1 2 3 4 5 … 2402 Next

Download the data.

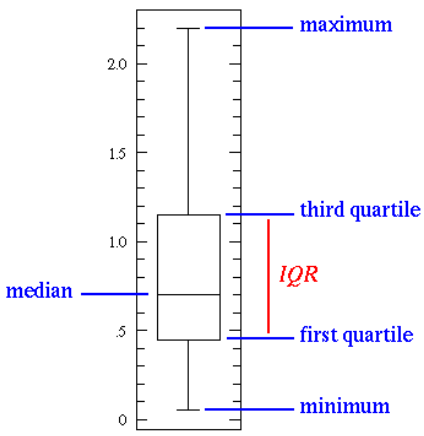
2. Normalization

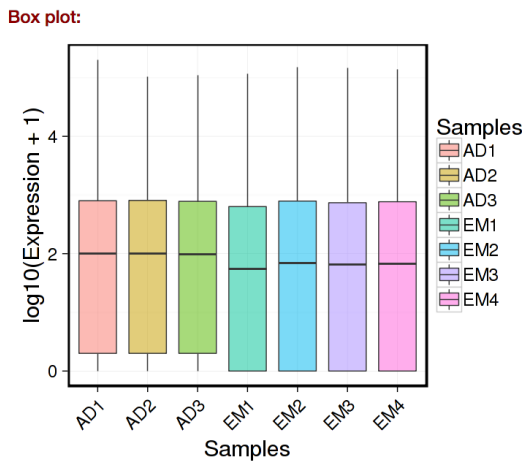
You may normalize your data set using either Quantile Normalization or Size Factor Normalization. You can also perform downstream analysis on the original dataset. After each variation of the normalization method, the quality check reports will be shown on the right, which consists of several plots described below. You may resort to them to find the best normalization strategy for downstream analysis.

2.1 Box Plot

The box plot is a standardized way of displaying the distribution of a set of data points. The central rectangle spans from the first quartile to the third quartile. A segment inside the rectangle shows the median and whiskers stretching outside the box shows the locations of the minimum and maximum.

In this module, each box represents the distribution of gene expression level of a sample. Well normalized expression profiles have expression patterns with great consistency among samples, as is the case with our dataset.

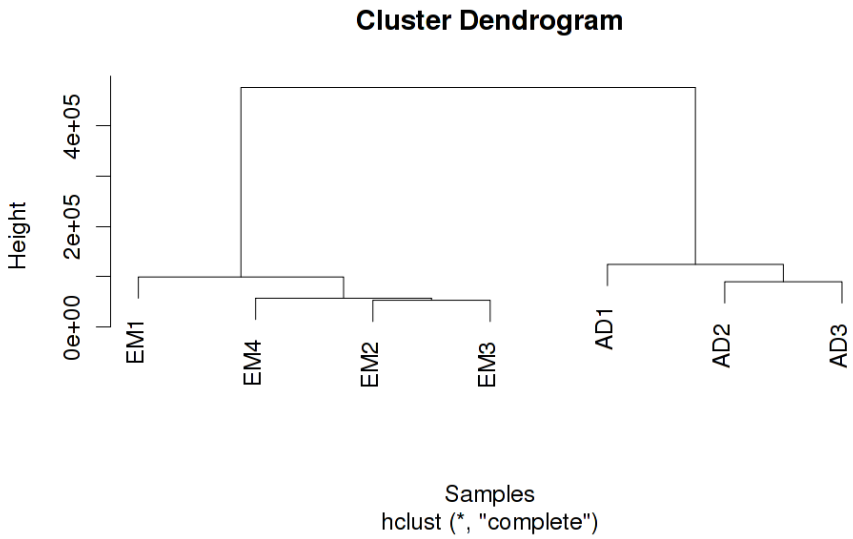




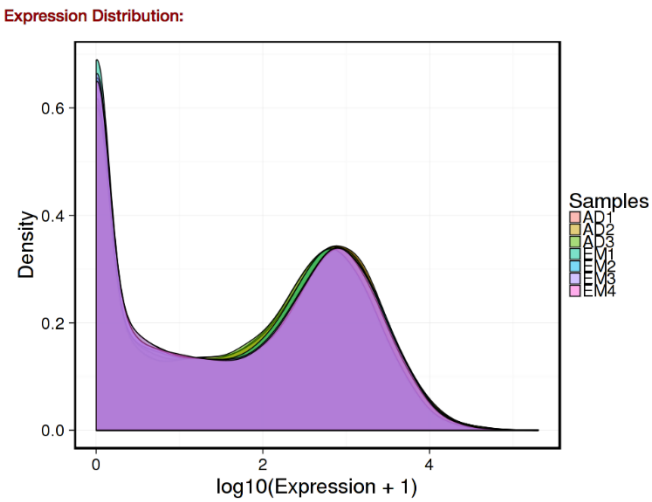
2.2 hierarchical clustering

Hierarchical clustering output a tree structure to visualize similarity relationships among samples. The height of a branching point stands for the similarity among samples in the subtree below it, with more similar samples having lower branching points connecting them.

As is often the case, the clustering result based on our sample files corresponds well with biological condition partitioning.



2.3 expressional distribution plot



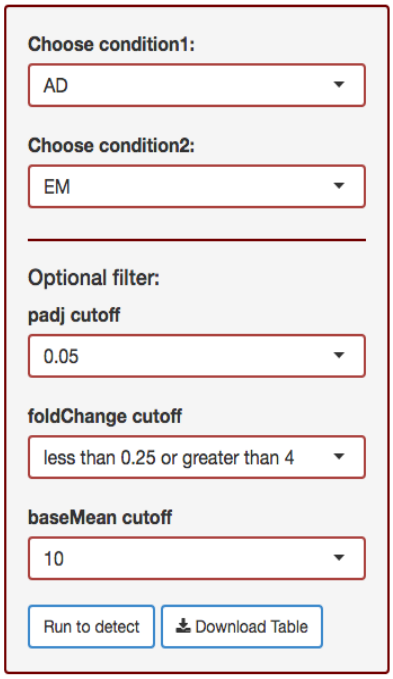
The distribution density of the expression profile of each sample. The abscissa stands for the base-10 logarithm of the expression level plus 1. Samples are color-coded.

Distribution patterns are commonly assumed to be similar among samples. If one curve largely differs from others, it often implies low quality of the corresponding sample.

3. DEG Calling

3.1 DESeq

DESeq is a tool for calling differentially expressed genes (DEGs) between two biological conditions of interest. There are several parameters to run this test, as described below:



1) **padj cutoff:** Adjusted p-value cutoff. The adjusted p-value, also called the false discovery rate (FDR), is an indicator to estimate the expected proportion of discoveries (identified as differentially expressed) that are false. Genes with smaller FDRs are thought to be differentially expressed with higher statistical significance. Setting a smaller cutoff value will result in a more stringent test and less reported genes, and vice versa.

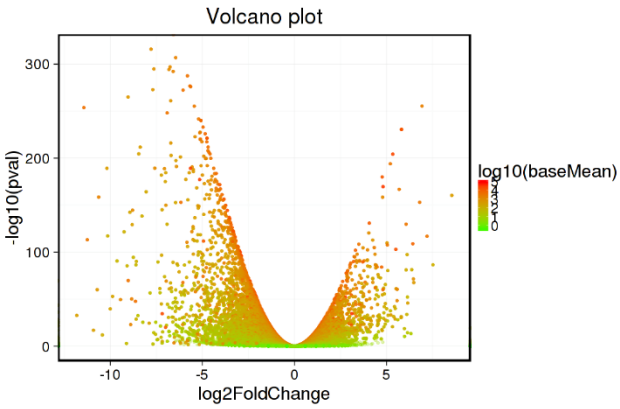
2) **foldChange cutoff:** The fold change is defined as the mean expression level under condition 2 divided by that under condition 1. The greater the relative difference, the further the fold change departs from 1. Setting a greater cutoff value will result in a more stringent test and less reported genes, and vice versa.

3) Mean expression cutoff. The mean expression level is with regards to all samples under both conditions. This filter is intended to remove genes with inappreciable expression levels, which open leads to unduly large fold change values.

Click the “Run to detect” button to start running. When it’s finished, you will get the differentially expressed gene list as below. You may click the “Download” button to download the list in .csv format.

	baseMean	baseMeanA	baseMeanB	foldChange	log2FoldChange	pval	padj
Cntnap1	3176.143	7310.333	75.50	0.010327846	-6.597317	0.000000e+00	0.000000e+00
Ddn	13368.286	30755.667	327.75	0.010656573	-6.552113	0.000000e+00	0.000000e+00
S100b	1264.429	2932.667	13.25	0.004518072	-7.790077	1.242254e-316	8.344221e-313
Plekhb1	2703.571	6213.667	71.00	0.011426426	-6.451482	1.261489e-307	6.355069e-304
Pcp4l1	1618.429	3730.333	34.50	0.009248503	-6.756564	1.044693e-297	4.210322e-294
Ankrd33b	1028.571	2384.000	12.00	0.005033557	-7.634206	1.405367e-295	4.719926e-292
Itpka	1480.286	3413.667	30.25	0.008861439	-6.818243	5.468982e-295	1.574364e-291
Gfap	2561.429	5895.000	61.25	0.010390161	-6.588638	6.106919e-293	1.538256e-289
Zfp365	5700.429	12992.667	231.25	0.017798502	-5.812100	2.797389e-288	6.263354e-285
Rasgrp1	5168.143	11752.667	229.75	0.019548755	-5.676779	1.719610e-277	3.465187e-274

Showing 1 to 10 of 2,829 entries Previous 1 2 3 4 5 ... 283 Next



The p value is an indicator to estimate the expected probability of making a false discovery given the gene is actually not differentially expressed. Plotting the p value against the logarithm of fold change results in a volcano plot as below, which gives us clues about the distribution of differentially expressed genes.

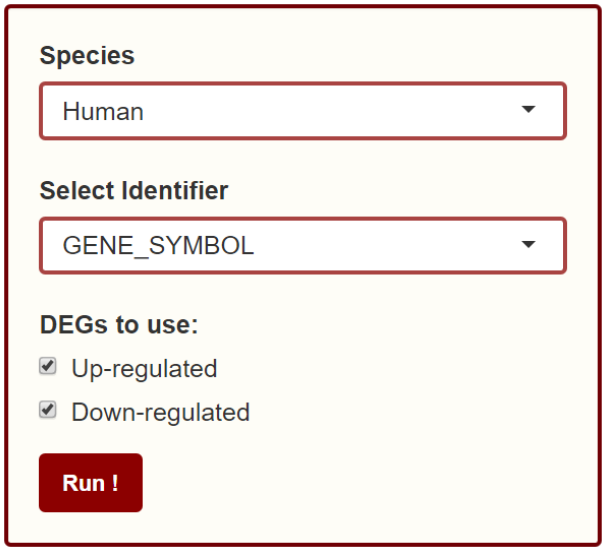
3.2 Fold Change

Calling differentially expressed genes based solely on the fold change value. This mode is typically used when lacking biological replicates.

Please note that the Y axis of the volcano plot in this mode represents the mean expression level of gene, rather than p value in a canonical volcano plot.

4. Function

This module provides several tools to reveal the biological meaning behind a select set of genes. Generally, it tests which gene categories have more genes (i.e. be enriched) than they would have if the gene set is chosen randomly. Pathway functionality tests the enrichment of KEGG pathways, while other tools test the enrichment of Gene Ontoloty (GO) terms.



Before running the functional enrichment test, you should specify the species, because genes with the same name may be present in different species with different functions.

Be sure to select the right gene identifier for correct recognition of the genes. For gene lists with mixed name types, david.ncifcrf.gov/conversion.jsp may help convert their names.

The gene set is by default DEGs output from the DEG Calling module. You can choose only up-regulated or down-regulated genes, or you can choose both. “Up-regulated” means having a higher expression level under condition 2 than condition 1.

When it finishes running, you will get the list of enriched GO categories (or pathways), along with their corresponding count number (number of genes within), enrichment p value and other related information.

	Category	Term	Count	Fold.Enrichment	PValue
1	GOTERM_MF_FAT	GO:0022803~passive transmembrane transporter activity	125	3.10211943214992	1.83746459124047e-31
2	GOTERM_MF_FAT	GO:0015267~channel activity	125	3.10211943214992	1.83746459124047e-31
3	GOTERM_MF_FAT	GO:0022836~gated channel activity	100	3.59805697243538	7.30051949707309e-31
4	GOTERM_MF_FAT	GO:0005216~ion channel activity	114	3.13725630967843	3.53638889404961e-29
5	GOTERM_MF_FAT	GO:0022838~substrate-specific channel activity	116	3.09277587817683	4.42973799528529e-29
6	GOTERM_BP_FAT	GO:0043269~regulation of ion transport	139	2.6685670982029	1.38386351188716e-27
7	GOTERM_CC_FAT	GO:0005576~extracellular region	639	1.40399632304742	5.56006759498538e-27
8	GOTERM_BP_FAT	GO:0006811~ion transport	240	1.98051843340266	2.00753461400238e-26
9	GOTERM_BP_FAT	GO:0034220~ion transmembrane transport	151	2.45230690275501	7.55467203764771e-26
10	GOTERM_BP_FAT	GO:0006952~defense response	251	1.89367294526679	9.36713954017327e-25

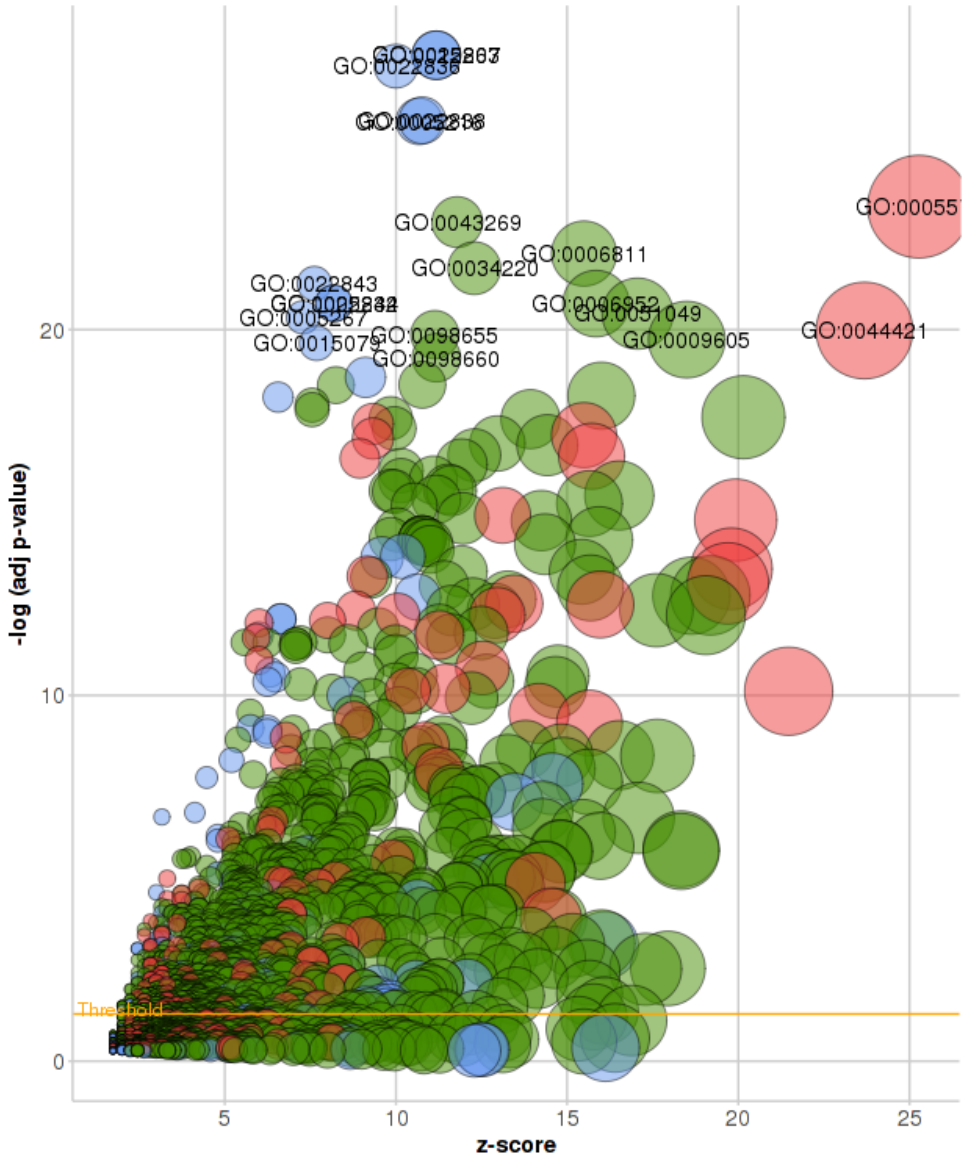
Showing 1 to 10 of 2,418 entries

Previous 1 2 3 4 5 ... 242 Next

There are two available methods to run the GO test, with different visualization outputs. The DAVID method shows enriched GO terms in a scatter plot as following. Each category is represented by a colored circle with the size proportional to the number of genes within. The Y axis measures the statistical significance of enrichment. The X axis is the z-score defined as

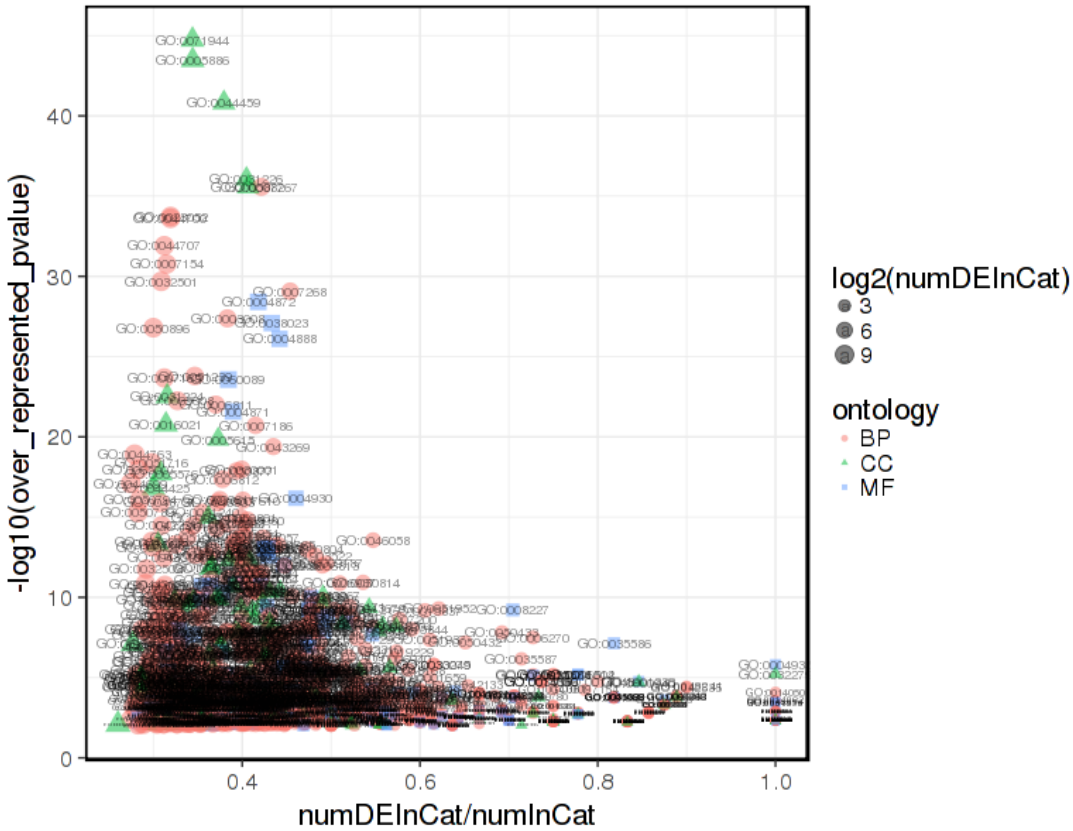
$$\frac{U - D}{\sqrt{U + D}}$$

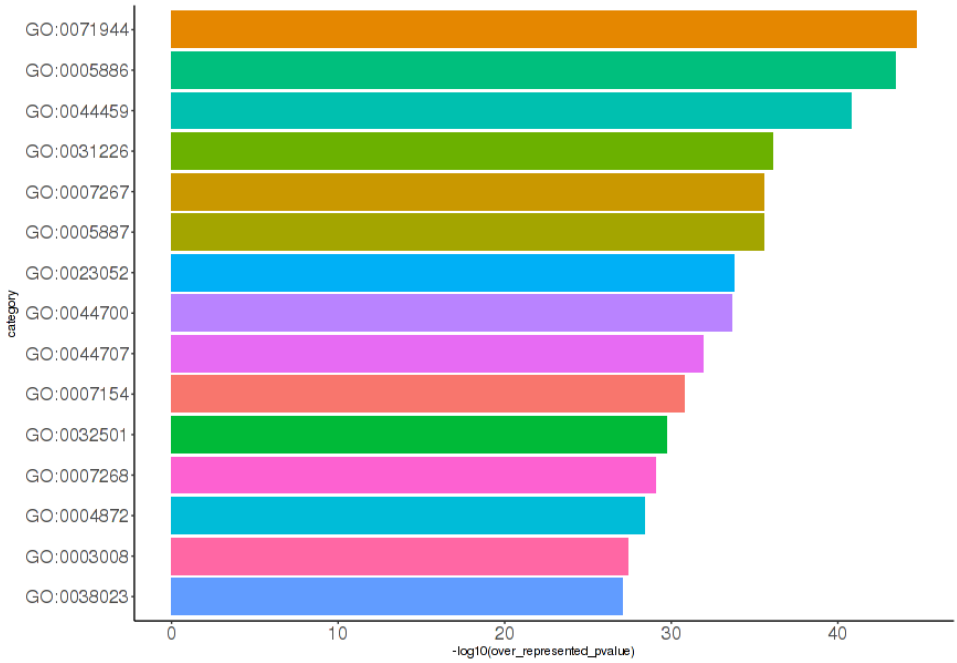
where U (or D) is the number of up- (or down-) regulated genes. Note that if only up-regulated genes are detected in the DEG Calling module, here D will be equal to zero, and vice versa.



The GOSeq method gives a similar plot except that the X axis is the proportion of genes in each category that are in our gene list. The most significant categories are also

summarized in a horizontal bar plot. The length of each bar represents the statistical significance of enrichment.

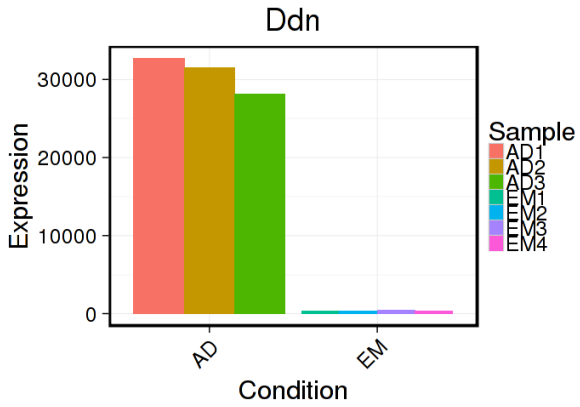




5. Plots

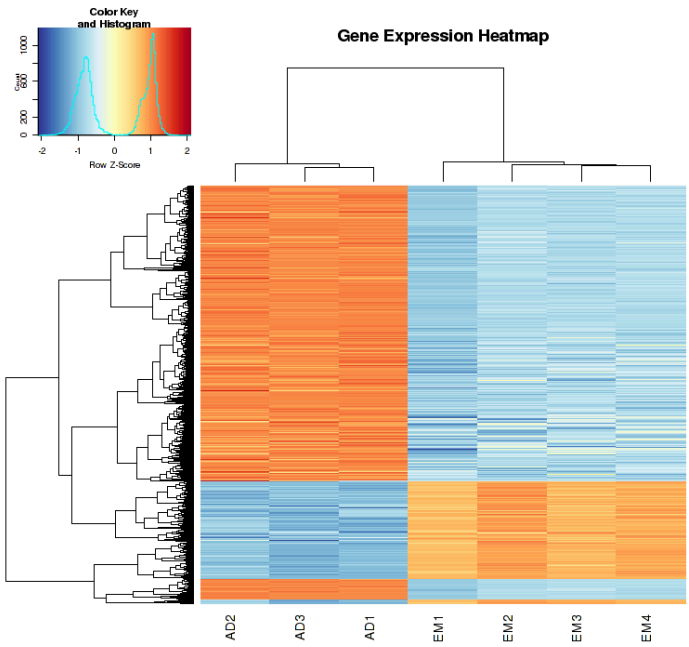
5.1 barplot

The bar plot shows the expression level of a given gene in all samples, grouped by conditions. As showed below, gene Ddn has far higher expression levels in adult samples.

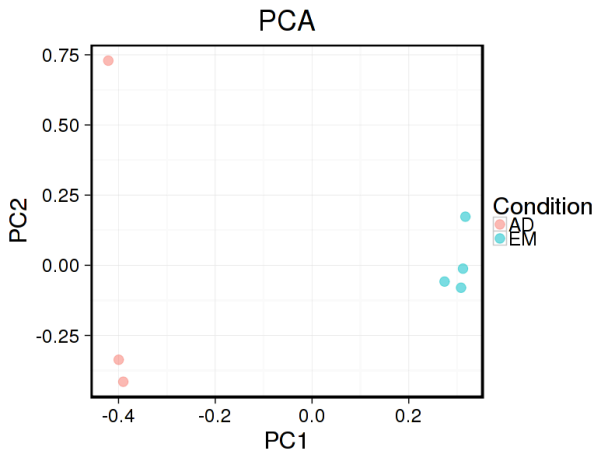


5.2 heatmap

In the heatmap, each row represents a gene, and each column represents a sample. Rows and columns are hierarchically clustered. Gene sets with special expression modes could be identified with heatmaps.



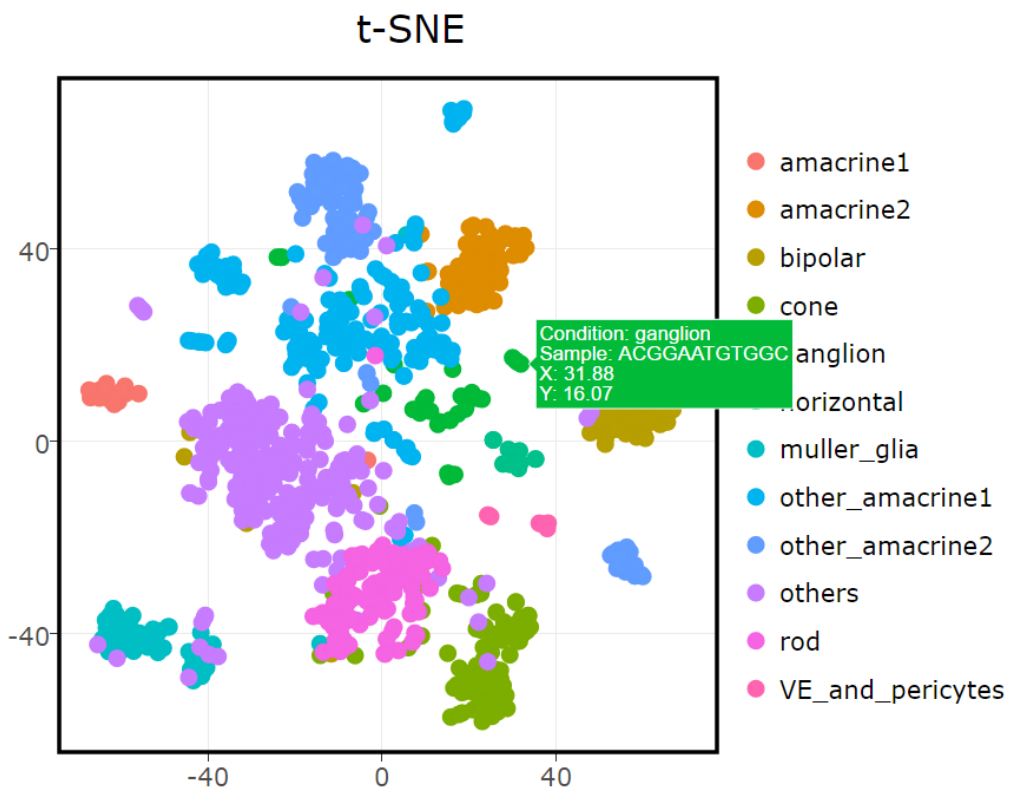
5.3 Principal component analysis (PCA)



PCA projects high-dimensional data points onto a low dimensional space for visualization. The orthogonal axes of the space are named PC1, PC2, and so on. They are chosen in such a way that the original data points have the largest variance in the direction of PC1, the second largest in PC2, and the like. The overall vicinity relationship of data points is supposed to be maintained after the projection.

5.4 T-distributed stochastic neighbor embedding (t-SNE)

T-SNE is a non-linear dimesionality reduction technique that maps points onto a 2D plane, with the goal to keep the neighboring relationship among points. It is particularly suitable for sample clustering and do not have the problem of hiding other dimensions as in PCA. You can hover your mouse over a point to show detailed information about it. It should be noted that the result of t-SNE is only plausible when the number of samples is large enough (at least, say, fifty). You can roughly judge the quality of a t-SNE plot by seeking if distinct clusters can be identified.



Stay in touch

If you find any bugs, have gripes about your experience using iSeq, or would like to offer suggestions to improve it, please contact

zhangchao3@hotmail.com