SEGtool:

an R package for specifically expressed gene detection

Author list: Zhang Qiong; Liu Chun Jie; Lin Sheng Yan

Feb 29, 2016

College of Life Science and Technology, HUST, Wuhan,China

Contents

1 Introduction

SEGtool is an R package used for expression datasets analysis in purpose to detect specifically expressed genes (SEGs, also known as tissue specific genes). SEGs are essential outliers in a given condition (or different treatment, tissue). In order to detect such outliers, we combines one-step Tukey-biweight, modified fuzzy C-mean (FCM) clustering algorithm, Jaccard and modified Simulated Annealing (SA) methods to analyze special expression patterns with the special focus on gene-centered expression data. It is designed for a distinct purpose of mining SEG. In addition to the main purpose of SEG detection, it includes other methods such as get_visualisation_html, get_heatmap and get_PCA which produce text files and html webpage report. This package can be freely obtained from the website: (http://bioinfo.life.hust.edu.cn/software/SEGtool). SEGtool is suitable for one-factor experimental design with multiple treatment levels, different tissues and SEG finding.

The procedures included in SEGtool analysis are:

- Specific Expression Patterns Detection, using a Tukey-biweight modified fuzzy C-mean (FCM) clustering algorithm method
- Principle Component Analysis (PCA) for the samples with SEGs
- Cluster analysis for genes and samples
- Represent SEGs in different samples
- Plotting all the analysis results
- Generate html report

2 System requirements

This package can be run on LINUX/UNIX and WINDOWS OS, it was developed under R 3.2.1 in the ubuntu12.04 operating system.

The memory occupation is dependent on the sample size of the input datasets. 170M, 220M, 336M memory and 4min, 5min, 8min were taken in an E7- 4820 computer using 4 Cores while handling 39, 60 and 100 samples which all of them have 60533 genes.

In order to implement our package, the R software and R packages including

hwriterPlus, parallel, pheatmap, ggplot2 , svglite , stringr are required and all of them can be downloaded from CRAN or http://www.bioconductor.org/

3 Installation

3.1 Installation from Rgui for Windows(07/08)

Select Menu Packages in the gui interface of R, click Install package from local file or from the internet. Choose the file SEGtool_*.tar.gz and install it.

3.2 Installation on Linux/Unix

Copy SEGtool_*.tar.gz file to the R directory and change to the R directory, then type the follow command: R CMD INSTALL SEGtool. If there is no error message, it means that the package is installed in the default directory such as R/library. The complete command is as follow:

cp SEGtool_*.tar.gz R_DICTIONARY(This is the location where R been installed)

cd R_DICTIONARY

R CMD INSTALL SEGtool

4 Data and function list

Here are the example and functions which were used in SEGtool:

4.1 Input data format

The input data should be prepared by users before this package loaded. Two expression formats are accepted by this package:

1) All genes expression data in numeric format.

2) Log2-format expression data (If user input this kind of data, please make sure any of the number is less than 0, SEGtool detect the negative value and then translate all input data to normal format by f(x)=2^x, x is input data).

The input data file must be a tab delimited text file containing expression data of all genes. The first row contains sample name with no replicates; the first column contains the GeneID (Gene name, or Gene symbol or any flag represents gene) of each gene.

Here is an example format:

GeneID	Adipose	Muscle	Heart	Ovary	Uterus
ENSG00000223972	0	0	0	0	0
ENSG00000227232	5.8	6.7	6.0	2.9	10.4
ENSG00000243485	0	0	0	0	0
ENSG00000237613	0	0	0	0	0

The input data are composed by normal expression data obtained from other corresponding pre-treatments, SEGtool is suitable for analyzing datasets derived from variant kinds of expression format (sequencing raw tag, RPM, TPM, RPKM, FPKM, RMS, MAS, etc.). If samples/tissues/conditions/treatments in input datasets have replicates, integrating expression values of replicates for each treatment by each gene is recommended. User could use mean/median/quantile methods to integrating values of replicates or use SEGtool build-in function(replicates_value_integration) to gain a M-ESTIMATOR value (based on one-step Tukey-biweight value) as follow.

R> library(SEGtool);

Data_integrate<-replicates_value_integration(x,y)

replicates_value_integration : A function to integrate expression of replicates. The x is a data.frame object and y is a factor vector. The same index in y represents the same group/treatment/tissue/condition at the corresponding location in column name of x .

4.2 A guide for main functions

To help users understand the usage of SEGtool, this section will show a few demo codes with dataset distributed within the package.

Use the build-in dataset:

R> library(SEGtool);

R>data (EbiHumanExpression)

Or

R>source("SEGtool.R")

R>load("SEGtool_dictionary/data/ EbiHumanExpression.rdata")

EbiHumanExpression is an expression data. rame object (which is defined by this package) derived from the 39-tissues 60533 genes' expression data of EBI expression datasets.

Or reading files from user's own experiment data:

R> EbiHumanExpression <- read.table("EBI_human_merged_tissue_expression.matrix", header=T,row.names=1,sep="\t")

Key step for specific expressed gene detection is as follow:

SEGtool_result <- SEGtool (EbiHumanExpression, exp_cutoff = 3,multi_cpu = 4, detect_mod=2, result_outdir='SEGtool_result', draw_heatmap=TRUE, draw_pca=TRUE, draw_plot=TRUE, html_report=TRUE)

EbiHumanExpression: A data.frame or matrix-like object with column name and row name. This dataset is used for SEG detection.

exp_cutoff: Expression cutoff for each gene. If all expressiond of a gene in samples ares less than the threshold, the genes will be discarded.

multi_cpu: The cpu number called for SEG detection. This option only works on the Linux-like platform which support mclapply function in R.

detect_mod: SEGtool build-in 3 (1, 2, 3) mods to detect SEG. Mod 3 is more strict than mod 2, perform the most accuracy at the cost of sensitivity. Mod 2 is designed as a moderate mode for the balance of accuracy and sensitivity. Mod 1 is the most sensitive mod which find SEG as many as possible, this may lost accuracy.

result_outdir: Dictionary where stores the result. Default is "SEGtool_result".

draw_heatmap: if this option is set true, package will draw heatmap for samples and SEGs after SEG detection. Default is TRUE (details see section 4.3).

draw_pca: if this option is set true, package will do PCA analysis and draw figure for PCA result. Default is TRUE (details see section 4.4).

draw_plot: if this option is set true, package will draw plot figure (x axis is sample, y is expression) for each SEG. Default is FALSE (details see section 4.5).

html_report: if this option is set true, package will generate a html webpage report for the result. Complete html report needs draw_heatmap, draw_pca and draw_plot options to be set TRUE. Report will be generated in the result_outdir direcotry which is set before. Default is FALSE (details see section 4.6).

SEGtool_result is a list object containing 3 elements: Allsummary, SEGinSample and p_value.

Allsummary: Summary for the result of SEGtool package. A demo is as follow:

total_genes	total_samples	total_SEGs	samples_have_SEG	high_SEGs	low_SEGs	overlap_high_low_SEGs
60433	38	2742	38	2700	39	3

SEGinSample: A data.frame stores the SEG pattern in samples. Each row represents a gene's expression situation in samples. -1,0,1 in this data.frame means different SEG pattern, -1 means the gene in this sample is detected by SEGtool as a candidate low SEG, 0 is NON-SEG pattern and 1 represents high SEG. Part of a demo is like this:

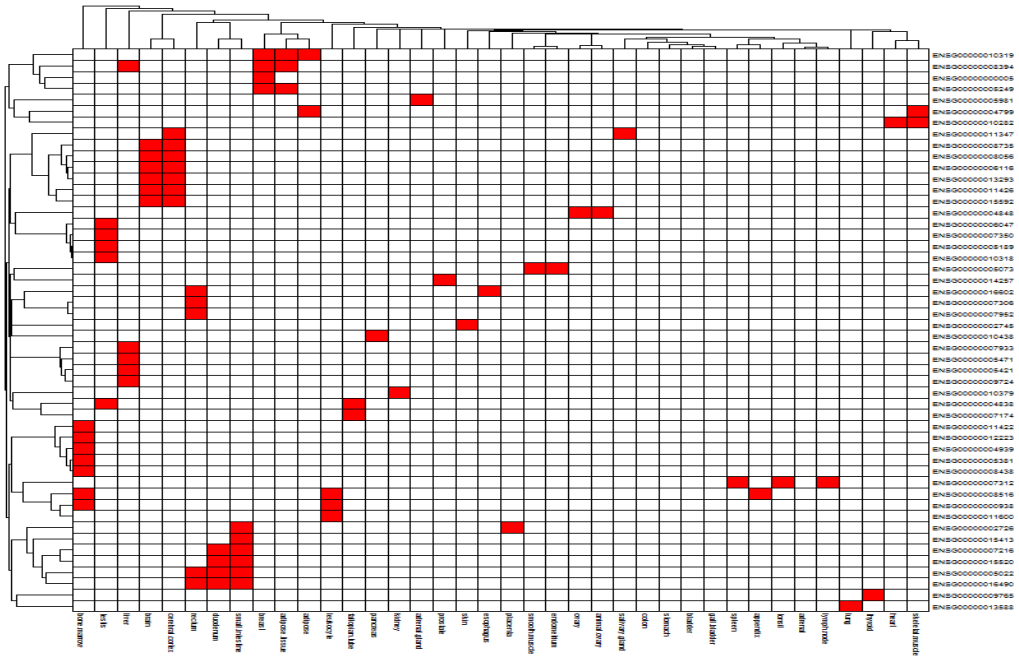
	Adipose	Muscle	Heart	Ovary	Uterus
ENSG00000223972	1	0	0	0	0
ENSG00000227232	0	0	0	-1	1
ENSG00000243485	0	0	0	0	0
ENSG00000237613	0	-1	0	0	0

p_value: a data.frame contains 4 columns to illustrate the SEGresult: max_exp, SEG_p_value, max_SEG_p_value and p_len. MAX_exp is maximal value of the SEG's expression. SEG_p_value is binomial test's p value for SEG in corresponding sample, the p value has been treated by -log10 function. Max_SEG_p_value is the maximal value of SEG_p_value. The p_len means how many samples in this gene are SEGs. Part of this part is shown as follow:

	max_exp	SEG_p_value	max_SEG_p_value	p_len
ENSG00000204983	61264.99	Inf	Inf	1
ENSG00000115386	36461.64	Inf	Inf	1
ENSG00000091704	33834.74	Inf	Inf	1

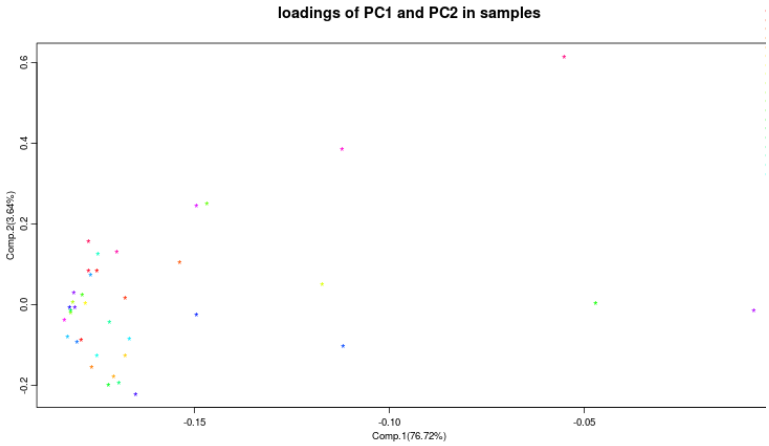
4.3 draw_heatmap function

This option is used to draw heatmap (top 50 SEG arranged by p value, most 50 samples) for SEGs, make it clear to see the gene in which sample is SEG. The red color means high SEG, white is non-SEG and green is low. Row represents gene and column indictes sample.



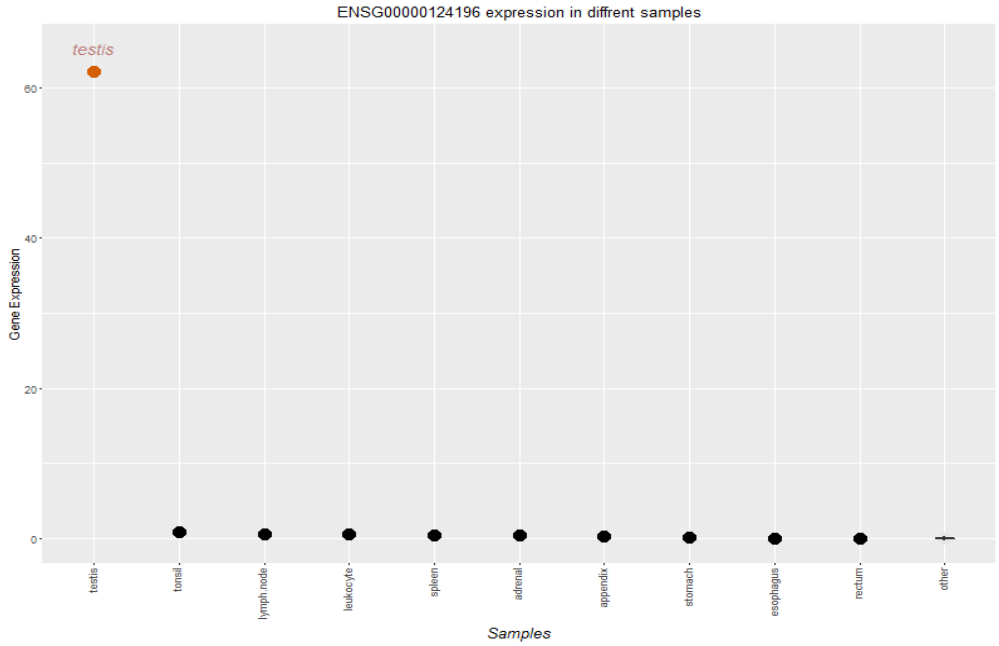
4.4 draw_pca function

This option is used to do PCA analysis performed only after SEGs has been already detected and sorted by build-in function which lists genes in the order of their overall specificity (judged by p value). Then we can perform PCA for all SEGs and use ***the first two components*** to plot pictures.



4.5 draw_plot function

This option is used to draw expression picture for all SEGs (each gene has a figure). All samples will be rearranged by expression, the x axis exhibits top 10 samples, if sample number is more than 10, the rest of the samples will at the end of x axis been named other with the expression displaying as box-plot of those expression. High SEG in the sample is orange color while low is green.



4.6 html_report functions

This function will generate html webpage report for current analysis result, complete html report contains some basic statistic information of the result and requires draw_* function. A demo of html webpage report is contained in the example dictionary of package.