TRAPR : Total RNA-seq Analysis Package for R User's Guide

JH Lim, SY Lee,

First edition 3 September 2012

TRAPR topics documents

1. Introduce TRAPR

RNA-Seq, is a standard technology for measuring gene expression at an unprecedented accuracy. Numerous Bioconductor packages have been developed for statistical analysis of RNA-Seq data. However, those tools focus on specific aspects of the data analysis pipeline and are hard to integrate appropriately with each other because of their ununified data structure and processing methods. They also lack visualization methods to confirm data integrity and process. Here, we present an R based RNA-Seq analysis pipeline TRAPR, an integrated tool including statistical analysis and visualization of RNA-Seq expression data. TRAPR provides various functions for data management, filtering of low quality data, normalization, transformation, statistical analysis, data visualization, and result visualization that allow researchers to build customized analysis pipelines

TRAPR is written in R (as of version 2.15) and is available at http://www.snubi.org/software/trapr

2. Start TRAPR

* Our manual was made based on window 7

2.1 Set up TRAPR package in your R

2.2 Previous set up packages

TRAPR must be needs some R packages(vsn, preprocessCore, edgeR, gridExtra, ggplot2, reshape2). We will show description of package setup method. For example, we explain vsn.

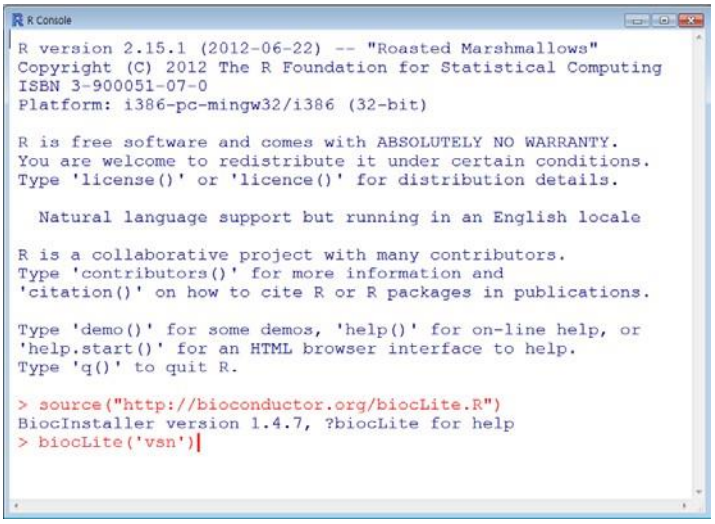


Fig1 Dependent packages installation

```
> biocLite('vsn')
> biocLite('preprocessCore')
> biocLite('gridExtra') >
biocLite('ggplot2')
> biocLite('reshape2')
```

2.3 Loading TRAPR package

```
> library(TRAPR)
```

2.4 Format of input file

TRAPR input file consist of Tab-deliminated text file. Column is sample list. Row is gene list. Each cell is filled gene expression values(Fig2).

Gene	1017-NOR	1079-NOR	110-NOR	1199-NOR	1207-NOR	1264-NOR
1/2-SBSRNA4	1.05316	1.12635	1.16598	0.903904	1.388	1.7264
A1BG	1.58885	1.53534	1.42561	1.45638	2.60126	1.46374
A1BG-AS1	0.813991	1.18167	0.549843	0.893526	1.82539	0.835946
A1CF	0	0	0	0.0512491	0	0.02684
A2LD1	1.47371	1.40145	1.66409	1.50761	0.89793	3.16842
A2M	677.94	788.356	555.677	536.844	299.616	789.546
A2ML1	0.0461678	0.0332678	0.0441994	0	0.0242496	0
A2MP1	0.161584	0	0.151889	0	0.204479	0.219432
A4GALT	6.09765	6.41299	2.60222	4.72684	4.70813	3.9633
A4GNT	0.306902	0	0.501595	1.12183	0.395527	0.266301
AA06	0	0	0	0	0	0
AAA1	0.130131	0	0	0	0	0
AAAS	8.66273	11.1746	12.4727	17.3145	20.0061	14.5962
AACS	4.82528	4.76136	5.15812	9.67915	6.27222	5.8261
AACSP1	0.163496	0.142302	0.516818	1.56209	0.420916	0.0664654
AADAC	14.6518	4.90271	9.24786	17.3143	3.16036	20.1152
AADACL2	1.91869	0	0.113368	0.964742	0	0.165812
AADACL3	0	0	0.047547	0	0	0
AADACL4	0.112633	0.131725	0	0.0943328	0.0909002	0
AADAT	4.43236	3.92109	5.01229	3.6641	3.4742	5.99568
AAGAB	12.5497	11.6327	11.1433	13.3532	9.16678	12.4696
AAK1	2.6064	2.49532	3.17003	2.4327	2.15502	2.50611

Fig2. input format of TRAPR

2.5 Sample data

Our sample data file name is 'sample.txt'. This file contained in TRAPR. The origin of sample.txt is breast cancer data from TCGA. It consist of 9 normal tissue cancer sample and 10 cancer sample using RNA-seq technology and It's value is FPKM(Fig3).

Gene	Cancer1	Cancer2	Cancer3	Cancer4	Cancer5	Cancer6	Cancer7	Cancer8	Cancer9	Normal1	Normal2	Normal3	Normal4	Normal5	Normal6	Normal7	Normal8	Normal9	Normal10
ARHGEF1(	7.301864	10.42624	6.250692	4.971898	7.597803	5.756496	2.814361	7.28613	7.000573	7.638267	10.8043	12.41658	2.908491	6.339586	9.085635	9.176228	10.65157	12.23706	11.25679
HIF3A	0.337063	0.892587	0.57944	0.765272	0.721414	0.531417	0.575267	0.740609	0.738899	10.35939	6.684232	6.964437	15.2138	6.890075	5.656143	2.436059	4.585604	3.6078	1.503526

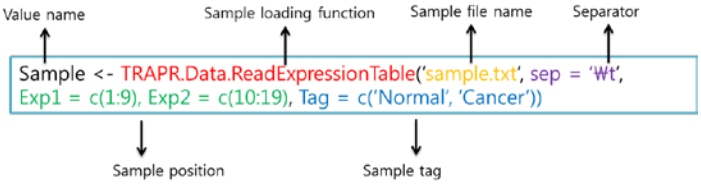
Fig3 sample.txt in TRAPR

2.6 Start TRAPR analysis

2.6.1 Data Manipulation

```
Sample <- TRAPR.Data.ReadExpressionTable('sample.txt', sep = '\t', Exp1 = c(1:9), Exp2 = c(10:19), Tag = c('Normal', 'Cancer'))
```

Function loading sample file is TRAPR.Data.ReadExpressionTable. The code component explain showing Fig3. More Information in reference manual.



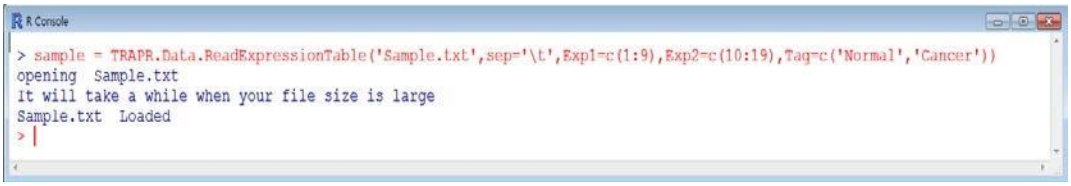


Fig4 sample loading code & example

Sample position (green color) is description for sample column position. Tag (blue color) is description for sample labels. Loaded sample data is saving 'Sample' value and after using function of filtering and normalization. More Information in reference manual. 2.6.2 Filter

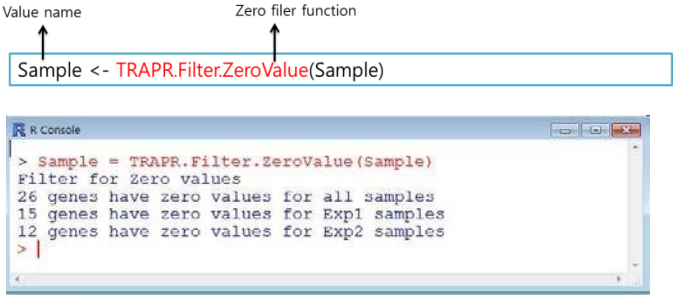


Fig5 Zero value filtering function code & example

```
> Sample <- TRAPR.Filter.ZeroValue(Sample)
```

Remove zero values from sample file. TRAPR provieds 6 kinds of filter condition ('Filter for Gene List', 'TRAPR.Filter.LowVariance', 'TRAPR.Filter.LowExpression', 'TRAPR.Filter.SampleDeletion', 'TRAPR.Filter.GeneDeletion'). If user want other function just typing replace 'TRAPR.Filter.ZeroValue'(read color). More Information in reference manual. 2.6.3

Normalization

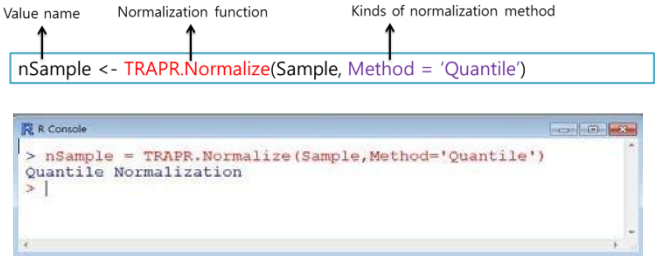


Fig6 Normalization function code & example

```
> nSample = TRAPR.Normalize(Sample,Method='Quantile')
```

TRAPR provide 4 method for normalization ('UpperQuartile', 'Quantile', 'Median', 'Mean'). If you needs another function, just write method name replace 'Quantile'(purple color).

2.6.4 Data Visualization

TRAPR provide pre-processing (filtering + normalization) result visualization. Data visualization function include 5 kinds of plots(Fig7).

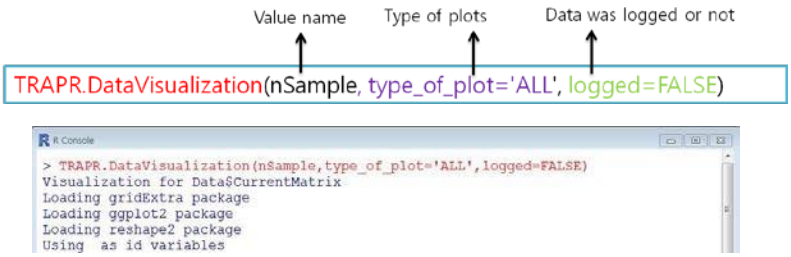


Fig7 Data visualization function code & example

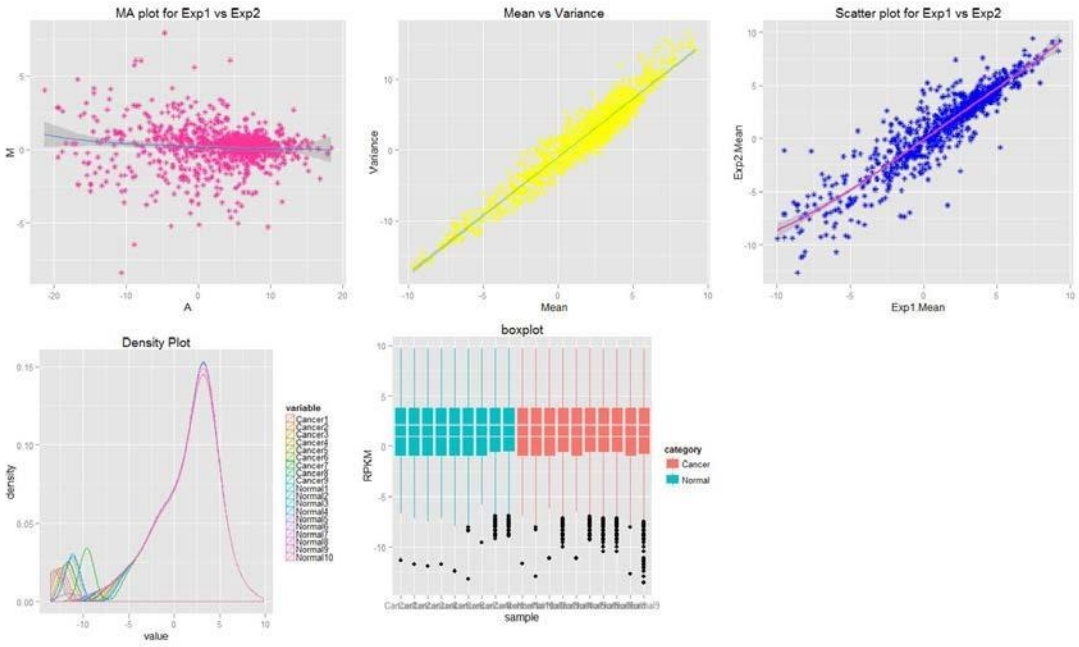


Fig8 Result example plot of data visualization

```
> TRAPR.DataVisualization(nSample, type_of_plot = 'ALL', logged = FALSE)
```

2.6.5 Statistical Test

TRAPR provide 4 method for statistical test ('ttest', 'wilcoxon', 'edgeR', 'FC'). If you needs another function, just write method name replace ' ttest '(purple color). TRAPR also provide adjust test method('holm', 'hotchberg', 'hommel', 'bonferroni', 'BH', 'BY', 'fdr', 'none').

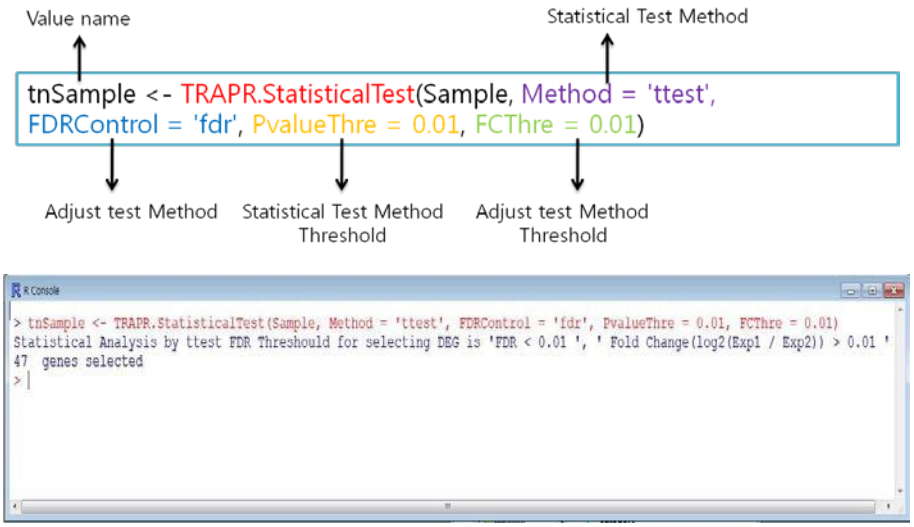


Fig9 Statistical Test function code & example

2.6.6 Result Visualization

TRAPR provide Result plot of after statistical test. We provide 2 kinds of plots (Heatmap, volcano plot)(Fig11).

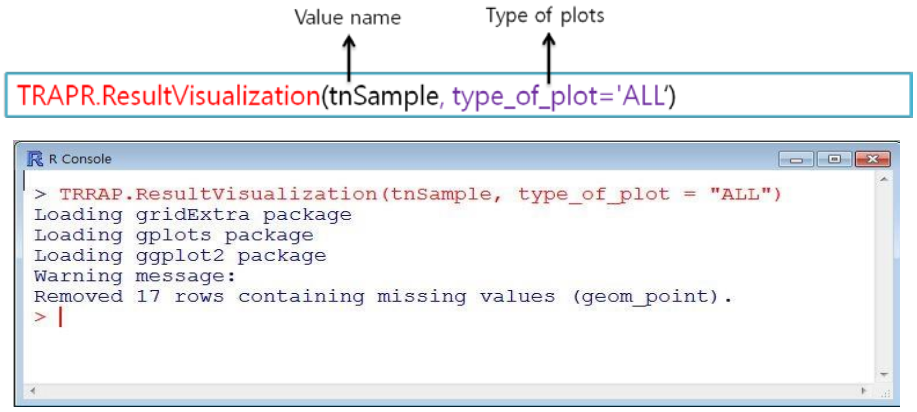


Fig10 Result visualization function code & example

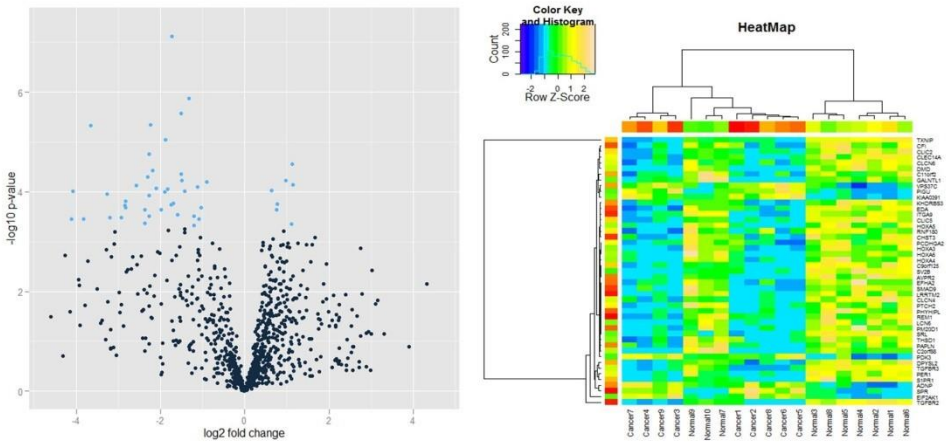


Fig11 Result example plot of result visualization

```
> TRAPR.ResultVisualization(tnSample, type_of_plot='ALL')
```

2.6.7 Out-print of DEG Result file

TRPR provide out-print your DEG result file in your directory. User just typing file name(blue color) in code.

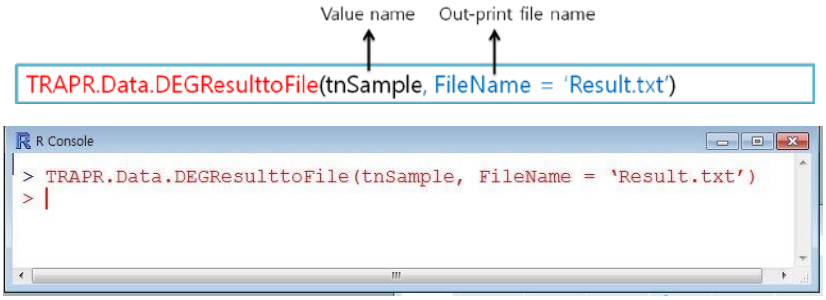


Fig12 Out-print of DEG Result file function code & example

```
TRAPR.Data.DEGResulttoFile(tnSample, FileName = 'Result.txt')
```

3. Reference

Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data, Genome biology, 11, R106.

Auer, P.L., Srivastava, S. and Doerge, R.W. (2012) Differential expression--the next generation and beyond, Briefings in functional genomics, 11, 57-62.

Costa, V., et al. (2012) RNA-Seq and human complex diseases: recent accomplishments and future perspectives, European journal of human genetics : EJHG.

Garber, M., et al. (2011) Computational methods for transcriptome annotation and quantification using RNA-seq, Nature methods, 8, 469-477.

Gentleman, R.C., et al. (2004) Bioconductor: open software development for computational biology and bioinformatics, Genome biology, 5, R80.

Hardcastle, T.J. and Kelly, K.A. (2010) baySeq: empirical Bayesian methods for identifying differential expression in sequence count data, BMC bioinformatics, 11, 422.

Langmead, B., Hansen, K.D. and Leek, J.T. (2010) Cloud-scale RNA-sequencing differential expression analysis with Myrna, Genome biology, 11, R83.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics, 26, 139-140.

Sultan, M., et al. (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome, Science, 321, 956-960.

Tarazona, S., et al. (2011) Differential expression in RNA-seq: a matter of depth, Genome research, 21, 2213-2223.

Wang, L., et al. (2010) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data, Bioinformatics, 26, 136-138.

Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics, Nature reviews. Genetics, 10, 57-63.