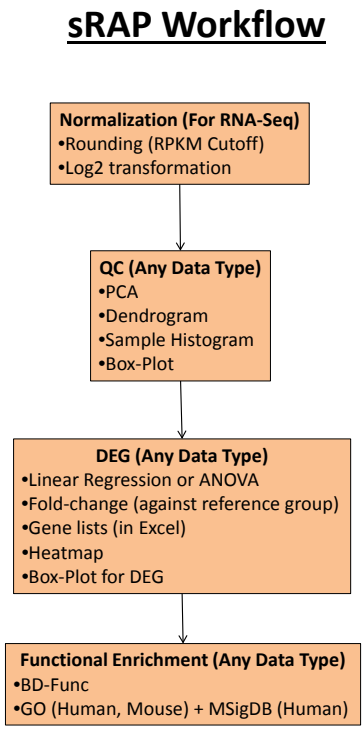
sRAP: Simplified RNA-Seq Analysis Pipeline

Charles Warden

July 27, 2014

1 Introduction

This package provides a pipeline for gene expression analysis. The normalization function is specific for RNA-Seq analysis, but all other functions will work with any type of gene expression data. The output from each function is created in a separate subfolder. Please see the workflow below:



Normalization will take RPKM expression values, round the RPKM values below a given cutoff, and perform a $\log_2$ data transformation. Quality control

metrics (Principal Component Analysis, Hierarchical Clustering, Sample Histograms and Box-Plots, and Descriptive Statistics) can be provided for these normalized expression values. These approximately normal expression distributions are then subject to differential expression. Differentially expressed gene lists are provided in Excel files and data can be visualized using a heatmap for all differentially expressed genes or a box-plot for a specific gene. BD-Func can then be used to calculate functional enrichment (either for fold-change values between two groups or using the normalized expression values).

2 Data

RPKM expression value for MiSeq samples were calculated using TopHat and Cufflinks from raw .fastq files from GSE37703. A template script for this sample preparation (run_RNA_Seq_v2.pl) is available here. This example dataset contains two groups, each with 3 replicates. The dataset is truncated for testing purposes.

```
> library("sRAP")
> dir <- system.file("extdata", package="sRAP")
> expression.table <- file.path(dir,"MiSeq_cufflinks_genes_truncate.txt")
> sample.table <- file.path(dir,"MiSeq_Sample_Description.txt")
> project.folder <- getwd()
> project.name <- "MiSeq"
```

The code for this example assumes all files are in the current working directory. However, you can specify the input and output files in any location (using the complete file path).

3 Data Normalization

To normalize RNA-Seq RPKM values, run the following function

```
> expression.mat <- RNA.norm(expression.table, project.name, project.folder)
```

```
NULL
```

RNA-Seq data is normalized by rounding RPKM (Read Per Kilobase per Million reads, [1]) values by a specified value (default=0.1), followed by a $\log_2$ transformation. This matches the gene expression strategy described in [2].

The output is standard data frame with samples in rows and genes in columns.

An Excel file containing the normalized expression values is created in the "Raw_Data" folder.

If you do not already have a table of RPKM/FPKM expression values, you can use the RNA.prepare.input() function to create such a file. Please see help(RNA.prepare.input) for more details.

4 Quality Control Figures

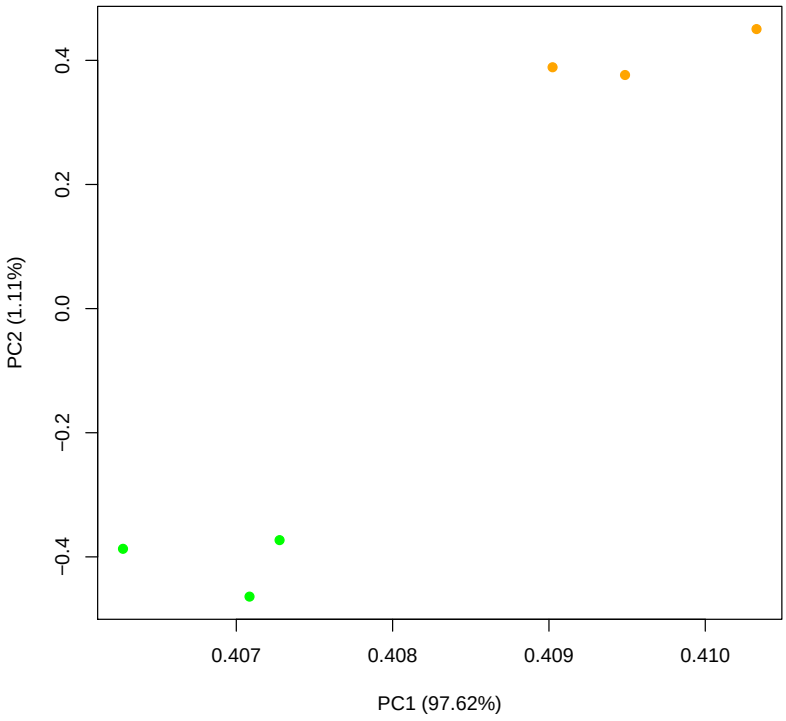
To create quality control figures, run the following function

```
> RNA.qc(sample.table, expression.mat, project.name,
+                  project.folder, plot.legend=F,
+                  color.palette=c("green","orange"))

[1] "SRR493372" "SRR493373" "SRR493374" "SRR493375" "SRR493376" "SRR493377"
[1] "integer"
[1] "Group: HOXA1KD"  "Group: scramble"
[1] "Color: green"  "Color: orange"
NULL
```

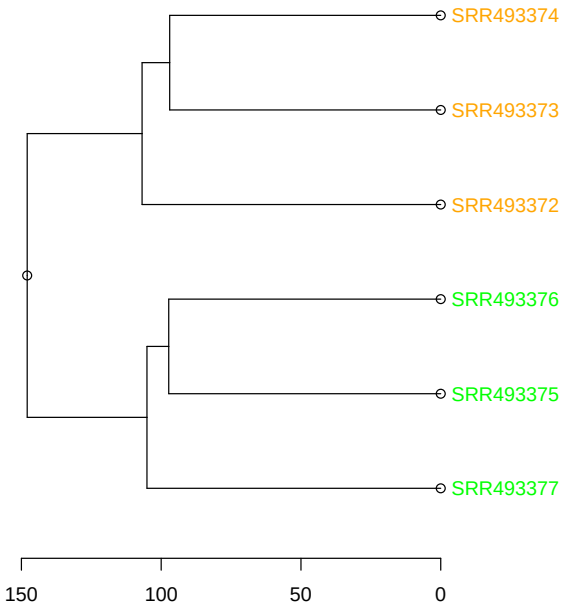
The input is a matrix of normalized expression values, possibly created from the RNA.norm function.

This function creates quality control figures within the "QC" subfolder. Quality control figures / tables include: Principal Components Analysis (figure for 1st two principal components, table for all principal components), Sample Dendrogram, Sample Histogram, Box-Plot for Sample Distribution, as well as a table of descriptive statistics for each sample (median, top/bottom quartile, maximum, and minimum). Please see the example figures displayed below:

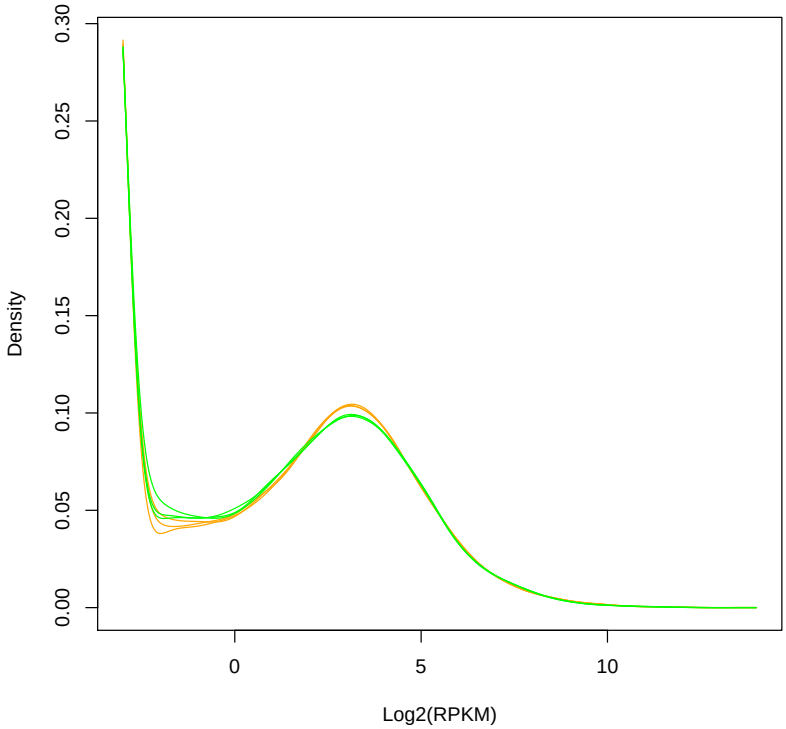
RNA.qc PCA Plot

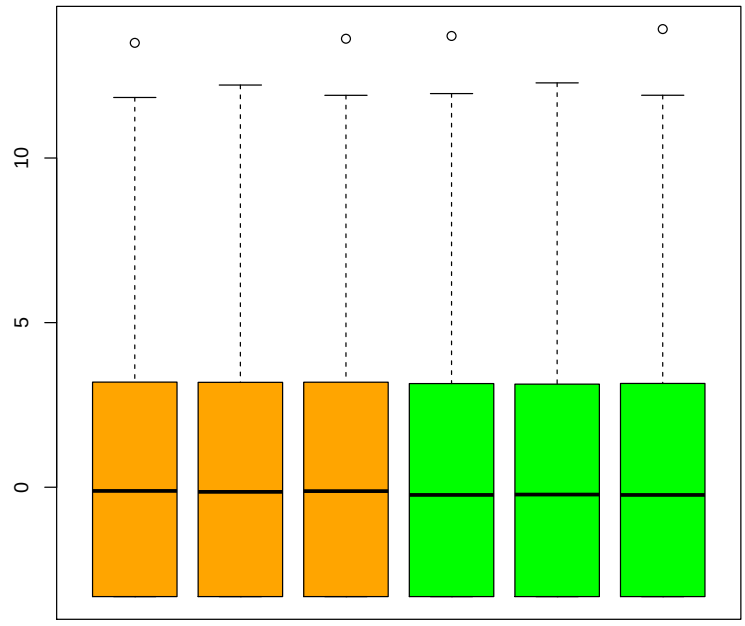


RNA.qc Dendrogram



RNA.qc Sample Histogram





This step is optional - this function is not needed for downstream analysis. However, this function is likely to be useful to identifying outliers, overall quality of the data, etc.

5 Differential Expression

To define differentially expressed genes, run the following function

```
> stat.table <- RNA.deg(sample.table, expression.mat,
+                         project.name, project.folder, box.plot=FALSE,
+                         ref.group=T, ref="scramble",
+                         method="aov", color.palette=c("green","orange"))

[1] 388   4
[1] 388   4
[1] 388   6
[1] "Group: HOXA1KD"  "Group: scramble"
```

```
[1] "Color: green"  "Color: orange"
NULL
```

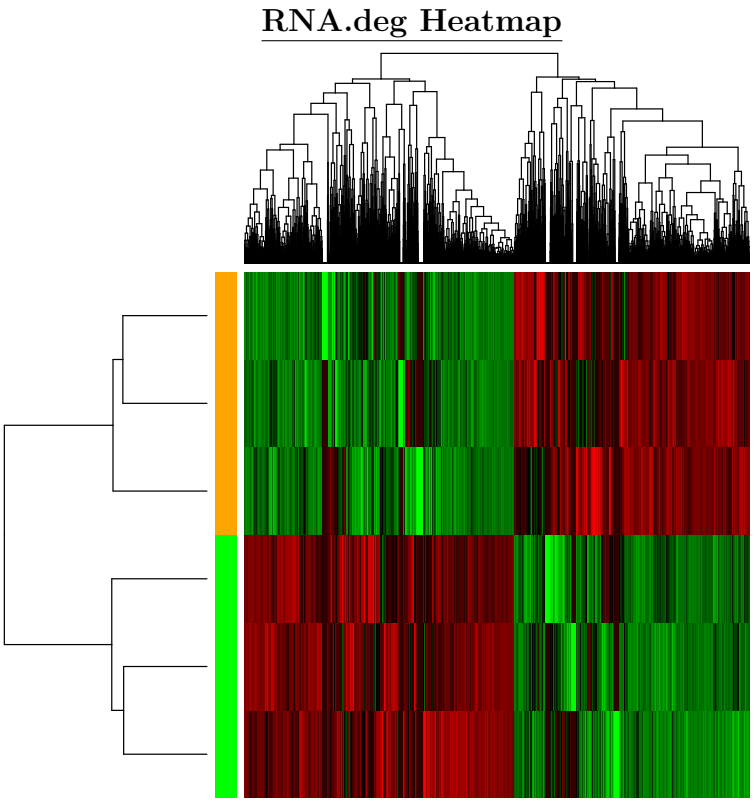
The input is a matrix of normalized expression values, possibly created from the RNA.norm function.

The function returns a table of differential expression statistics for all genes.

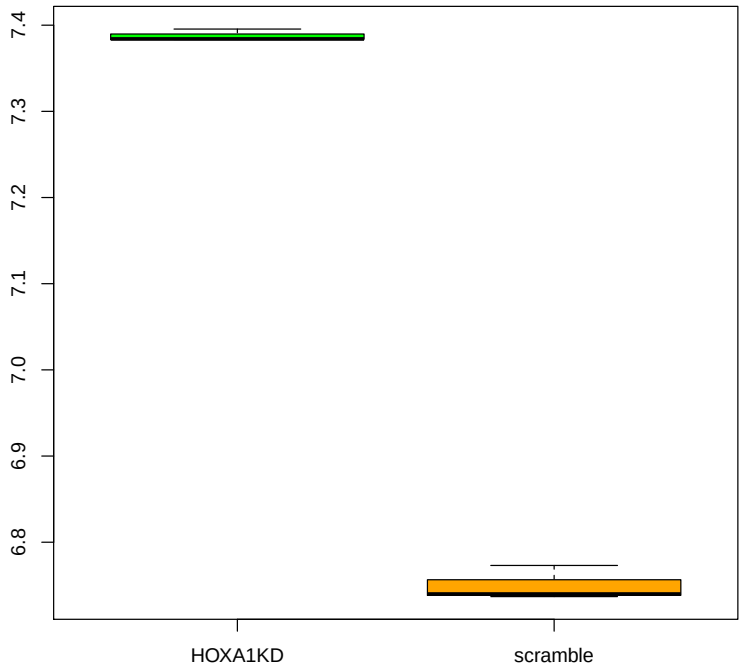
In all cases, p-values can be calculated via linear regression of ANOVA, and false-discovery rates (FDR) are calculated by the method of [3]. It is assumed that expression values are on a $\log_2$ scale, as described in [2].

If the primary variable (defined in the second column of the sample description table) is a factor with two groups and a specified reference, then fold-change values can also used to select differentially expressed genes (along with p-value and FDR values).

The function creates lists of differentially expressed genes (as well as a table of statistics for all genes) in Excel files. A heatmap of differentially expressed genes is also displayed . If desired, the user can also create box-plots for all differentially expressed genes. Please see the example figures below:



RNA.deg Box-Plot



An Excel file containing the table of differential expression statistics for all genes is created in the "Raw_Data" folder. All other outputfiles are created in the "DEG" folder (DEG stands for "Differenitally Expressed Genes").

6 Functional Enrichment

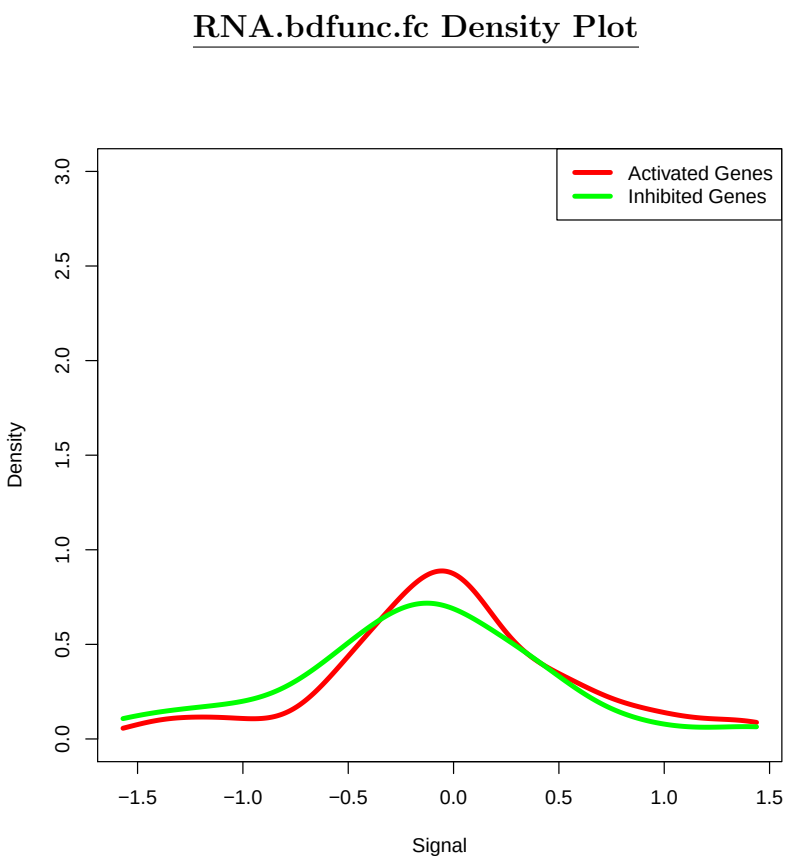
To identify functional categories subject to differential expression, run the following function

```
> #data(bdfunc.enrichment.human)
> #data(bdfunc.enrichment.mouse)
> RNA.bdfunc.fc(stat.table, plot.flag=FALSE,
+                 project.name, project.folder, species="human")

NULL

> RNA.bdfunc.signal(expression.mat, sample.table, plot.flag=FALSE,
+                 project.name, project.folder, species="human")
```

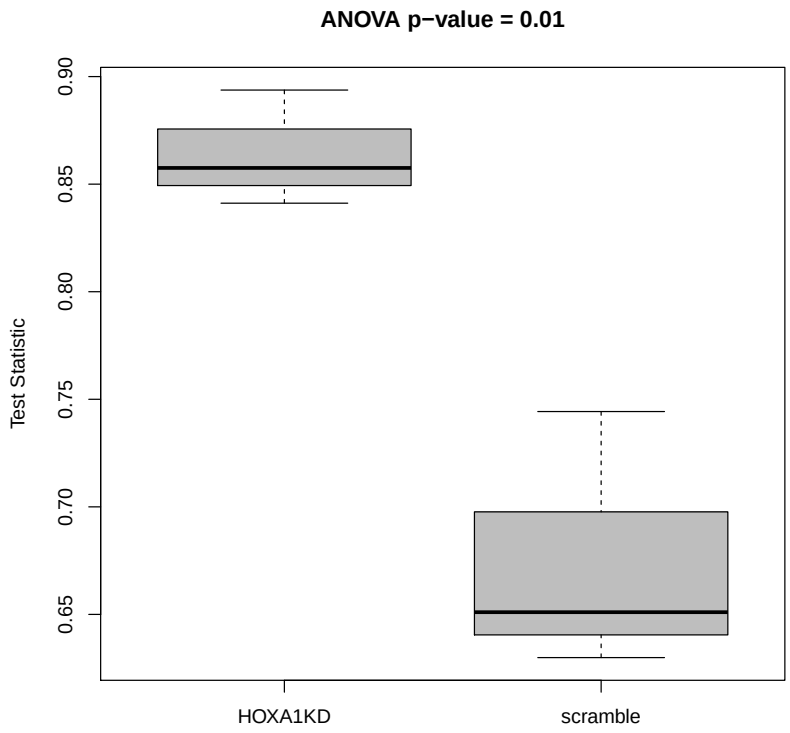
```
NULL
```

This is an implementation of the Bi-Directional FUNCtional enrichment (BD-Func [4]) algorithm. Briefly, p-values quantifying the difference between up- and down-regulated genes can be calculated via t-test, Mann-Whitney U test, or K-S test. If desired, false discovery rates (FDR) can be calculated using either the method of Benjamini and Hochberg [3] or the Storey q-value [5].

The input for the RNA.bdfunc.fc function is a table of differential expression statisics (like that created by the RNA.deg function). If desired, the this function can create density plots for all gene lists (see below).



The input for the RNA.bdfunc.signal function is a table of normalized expression values (like that created by the RNA.norm function). If desired, the this function can create box-plots for enrichment scores across all gene lists (see below).



In both cases, the output files are created with in the "BD-Func" subfolder. The goal of BD-Func is to calculate functional enrichment by comparing lists of activated and inhibited genes for a functional category, pathway, and/or network.

This package includes pre-defined enrichment lists are available for human and mouse gene symbols. The human enrichment list is based upon Gene Ontology [6] and MSigDB [7] gene lists. The mouse enrichment list is based upon Gene Ontology categories. Additional gene lists will need to be imported using the enrichment.file parameter.

This step is optional - there are no other functions that depend on the results of this analysis.

References

[1] Mortazavi A et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Meth*, 5:621–628, 2008.

[2] Warden CD et al. Optimal Calculation of RNA-Seq Fold-Change Values. *Int J Comput Bioinfo In Silico Model*, 2:285-292, 2014

[3] Benjamini Y and Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B*, 57:289–300, 1995.

[4] Warden CD et al. BD-Func: A Streamlined Algorithm for Predicting Activation and Inhibition of Pathways. *peerJ*, 1:e159, 2013.

[5] Storey JD and Tibshirani R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100:9440–9445, 2003.

[6] Ashburner M et al. Gene Ontology: tool for the unification of biology. *Nat Genet*, 25:25–29, 2000.

[7] Liberzon A et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27:1739–1740, 2011.