EBSEA: Exon Based Strategy for Expression Analysis of genes

Arfa Mehmood

arfa.mehmood (at) utu.fi

March 4, 2016

Contents

1 Introduction

In the conventional RNA-seq pipeline, gene counts are used to find differentially expressed genes in different conditions. EBSEA follows a different approach and it determines differential expression of genes based on the exon counts of the genes. EBSEA calculates the statistical significance of each exon in a gene seperately. The results of the exons in a gene are then aggregated to find the differentially expressed (upregulated/downregulated) genes. The user provides the exon count data, which can be generated for instance, using the python scripts in the DEXSeq R/Bioconductor package.

The statistical significance of each exon in a gene is obtained after normalizing the count data using the trimmed Mean of M values (TMM) method from Bioconductor edgeR package. The normalized counts are used to calculate the statistical significance of exons using the linear modelling approach in the Bioconductor Limma package. The exon results are then aggregated to find gene level estimates. The p-values are determined by comparing the median score to the null distribution. The p-values are further corrected using the Benjamini-Hochberg method.

2 Data

The input data to the EBSEA should consist of a dataframe consisting of counts from each sample. The colnames should represent the sample names. The rownames should consist of a gene followed by an exon number and should be seperated by a colon as shown in example data GeneName:Exonnumber.

The origCounts data is a subset of the first 1000 rows from the exon count data from the Pasilla package in Bioconductor. It consist of seven samples which are treated or untreated. The exon count data example is shown as follows:

```
> library(EBSEA)
> data("origCounts")
> head(origCounts)
```

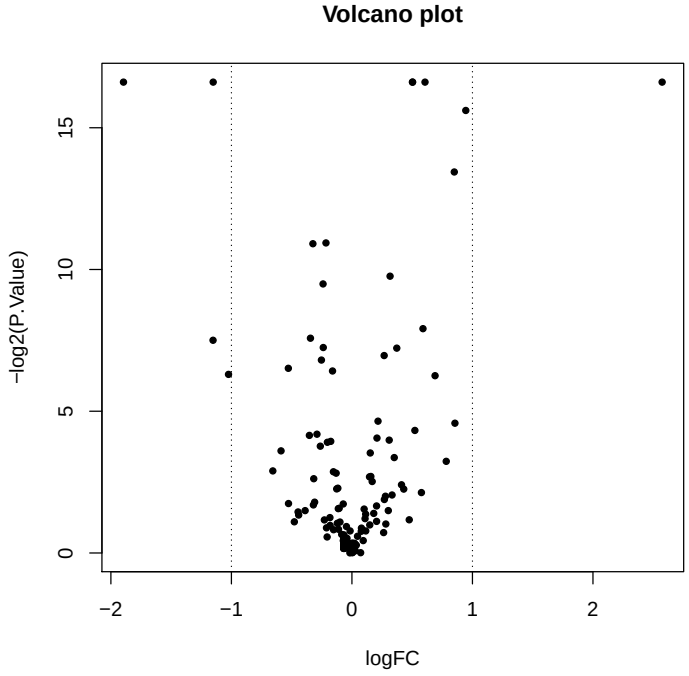
```
                treated1fb treated2fb treated3fb untreated1fb untreated2fb
FBgn0000003:001          0          0          1            0            0
FBgn0000008:001          0          0          0            0            0
FBgn0000008:002          0          0          0            0            0
FBgn0000008:003          0          1          0            1            1
FBgn0000008:004          1          0          1            0            1
FBgn0000008:005          4          1          1            2            2
                untreated3fb untreated4fb
FBgn0000003:001            0            0
```

```
FBgn0000008:001          0          0
FBgn0000008:002          1          0
FBgn0000008:003          1          0
FBgn0000008:004          0          1
FBgn0000008:005          0          1
```

3 Analysis

EBSEA can be run by loading the package and setting the sample groups to be compared. The sample groups should correspond to the colnames of the exon count files. The length of the sample group vector and the number of columns should be the same. If the samples are paired then the user should provide also the information about the pairs. EBSEA can then be called by giving the following parameters:

```
> group <- c('Group1', 'Group1', 'Group1', 'Group2', 'Group2', 'Group2', 'Group2')
> result <- EBSEA(origCounts, group, plot = TRUE)
```



The result consist of a list of two dataframes:

- Exon statistic table
- Gene statistic table

The exon statistics are as follows:

```
> head(result$ExonTable)
```

```
           GeneExon  AveExpr       logFC   P.Value       FDR
433 FBgn0000003:001 1.458686  0.47543587 0.4439905 0.9784763
736 FBgn0000008:001 1.232262 -0.06643235 0.8954555 0.9784763
466 FBgn0000008:002 1.458686 -0.47797968 0.4670941 0.9784763
299 FBgn0000008:003 2.137955 -0.69723494 0.3097988 0.9784763
659 FBgn0000008:004 2.137955  0.19204208 0.7566457 0.9784763
403 FBgn0000008:005 3.027786  0.49434677 0.4111072 0.9784763
```

The column names represent the following:

- **Gene and Exon:** Gene with its respective exon
- **AveExpr:** Average Expression
- **FC:** Fold change
- **logFC:** Log fold change
- **FDR:** False discovery rate
- **P.Value:** p-value

The gene statistics are as follows:

```
> head(result$GeneTable)
```

```
         Gene Median(signed_P.Value) ExonCount       logFC      P.Value
1 FBgn0000003              0.4439905         1  0.47543587 0.4442355576
2 FBgn0000008              0.9126710        14 -0.07169295 0.7392126079
3 FBgn0000014              0.8954555        10 -0.06643235 0.7401025990
4 FBgn0000015              0.8954555        12 -0.06643235 0.7127428726
5 FBgn0000017              0.1527009         9 -0.32396370 0.0005199948
6 FBgn0000018              0.5503789         2 -0.12156711 0.4839951600
          FDR
1 0.922812200
2 0.968428002
3 0.968428002
4 0.968428002
5 0.007539925
6 0.948368895
```

The column names represent the following:

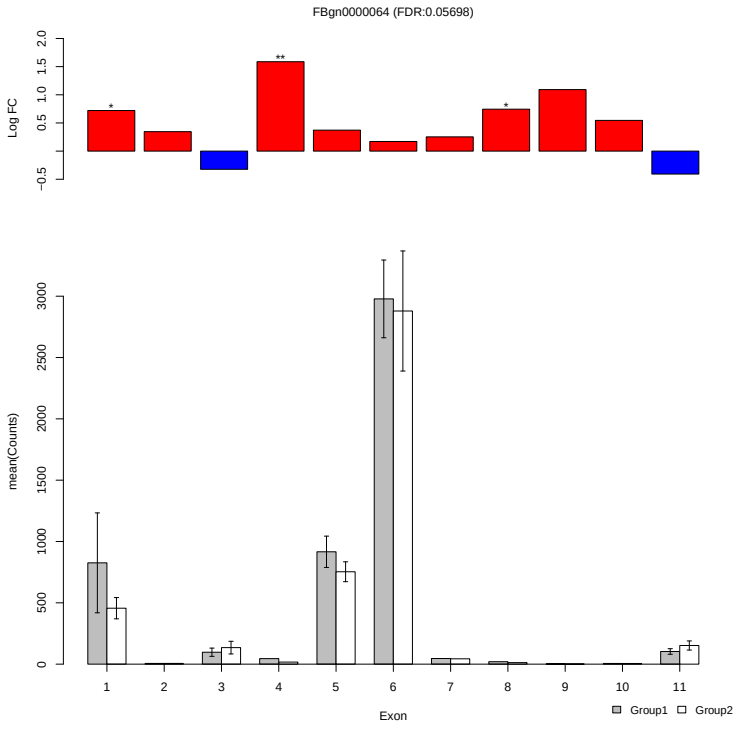
- **Gene**: Gene Name
- **Median:** Median value of the exon p-values

- **ExonCount:** The number of exons in a gene
- **FC:** Fold change
- **logFC:** Log fold change
- **FDR:** False discovery rate
- **P.Value:** p-value

The results can be viewed, stored or processed further.

The user can visualize gene information by providing an identifier of the gene of interest:

```
> visualizeGenes('FBgn0000064', result)
```



4 References

Laiho, A. et al.,*A note on an exon-based strategy to identify differentially expressed genes in RNA-seq experiments.* PloS One, 2014.