Gene set enrichment analysis of RNA-Seq data with the *SeqGSEA* package

Xi Wang[1,2] and Murray Cairns[1,2,3]

February 10, 2014

[1]School of Biomedical Sciences and Pharmacy, The University of Newcastle, Callaghan, New South Wales, Australia
[2]Hunter Medical Research Institute, New Lambton, New South Wales, Australia
[3]Schizophrenia Research Institute, Sydney, New South Wales, Australia

xi.wang@newcastle.edu.au

Contents

1 Introduction

1.1 Background

Transcriptome sequencing (RNA-Seq) has become a key technology in transcriptome studies because it can quantify overall expression levels and the degree of alternative splicing for each gene simultaneously. Many methods and tools, including quite a few R/Bioconductor packages, have been developed to deal with RNA-Seq data for differential expression analysis and thereafter functional analysis aiming at novel biological and biomedical discoveries. However, those tools mainly focus on each gene's overall expression and may miss the opportunities for discoveries regarding alternative splicing or the combination of the two.

SeqGSEA is novel R/Bioconductor package to derive biological insight by integrating differential expression (DE) and differential splicing (DS) from RNA-Seq data with functional gene set analysis. Due to the digital feature of RNA-Seq count data, the package utilizes negative binomial distributions for statistical modeling to first score differential expression and splicing in each gene, respectively. Then, integration strategies are applied to combine the two scores for integrated gene set enrichment analysis. See the publication Wang and Cairns (2013) for more details. The *SeqGSEA* package can also give detection results of differentially expressed genes and differentially spliced genes based on sample label permutation.

1.2 Getting started

The *SeqGSEA* depends on *Biobase* for definitions of class *ReadCountSet* and class *SeqGeneSet*, *DESeq* for differential expression analysis, *biomaRt* for gene IDs/names conversion, and *doParallel* for parallelizing jobs to reduce running time. Make sure you have these dependent packages installed before you install *SeqGSEA*.

To load the *SeqGSEA* package, type `library(SeqGSEA)`. To get an overview of this package, type `?SeqGSEA`.

```
> library(SeqGSEA)
```

```
> ? SeqGSEA
```

In this Users' Guide of the *SeqGSEA* package, an analysis example is given in Section 6, and detailed guides for DE, DS, and integrative GSEA analysis are given in Sections 3, 2, and 4, respectively. A guide to parallelize those analyses is given in Section 5.

1.3 Package citation

To cite this package, please cite the article below:
Wang X and Cairns MJ (2013). Gene set enrichment analysis of RNA-Seq data: integrating differential expression and splicing. BMC Bioinformatics, 14(Suppl 5):S16.

2 Differential splicing analysis and DS scores

2.1 The *ReadCountSet* class

To facilitate differential splicing (DS) analysis, *SeqGSEA* saves exon read count data using *ReadCountSet* class, which is derived from *eSet*. While below is an example showing the steps to create a new `ReadCountSet` object, creating a `ReadCountSet` object from your own data should refer to Section 6.

```
> rcounts <- cbind(t(sapply(1:10, function(x) {rnbinom(5, size=10, prob=runif(1))} )),
+                  t(sapply(1:10, function(x) {rnbinom(5, size=10, prob=runif(1))} )))
> colnames(rcounts) <- c(paste("S", 1:5, sep=""), paste("C", 1:5, sep=""))
> geneIDs <- c(rep("G1", 4), rep("G2", 6))
> exonIDs <- c(paste("E", 1:4, sep=""), paste("E", 1:6, sep=""))
> RCS <- newReadCountSet(rcounts, exonIDs, geneIDs)
> RCS

ReadCountSet (storageMode: environment)
assayData: 10 features, 10 samples
  element names: counts
protocolData: none
phenoData
  sampleNames: S1 S2 ... C5 (10 total)
  varLabels: label
  varMetadata: labelDescription
featureData
  featureNames: 1 2 ... 10 (10 total)
  fvarLabels: exonIDs geneIDs ... padjust (10 total)
  fvarMetadata: labelDescription
experimentData: use 'experimentData(object)'
Annotation:
```

2.2 DS analysis and DS scores

To better illustrate DS analysis functions, we load an example `ReadCountSet` object from a real RNA-Seq data set as follows.

```
> data(RCS_example, package="SeqGSEA")
> RCS_example

ReadCountSet (storageMode: environment)
assayData: 5000 features, 20 samples
  element names: counts
protocolData: none
phenoData
  sampleNames: S1 S2 ... C10 (20 total)
  varLabels: label
  varMetadata: labelDescription
featureData
  featureNames: ENSG00000000003:001 ENSG00000000003:002 ...
    ENSG00000007402:038 (5000 total)
  fvarLabels: exonIDs geneIDs ... padjust (10 total)
  fvarMetadata: labelDescription
experimentData: use 'experimentData(object)'
Annotation:
```

This example `ReadCountSet` object is comprised of 20 samples and 5,000 exons, part of the prostate cancer RNA-Seq data set (Kannan et al., 2011). With the function `geneID` and the script below, we can easily check the number of genes involved in this data set.

```
> length(unique(geneID(RCS_example)))

[1] 182
```

Noticed that some exons are too short or not expressed, we should first filter out these exons from following analysis to secure the robustness of our analysis. By default, function `exonTestablity` marks exons with the sum of read counts across all samples less than `cutoff` (default: 5) to be excluded in downstream analysis. Users can also exclude genes with no or low expression from downstream analysis by checking `geneTestability`.

```
> RCS_example <- exonTestability(RCS_example, cutoff = 5)
```

Then, the main DS analysis is executed using function `estiExonNBstat` for exon DS NB-statistics and function `estiGeneNBstat` for gene DS NB-statistics by averaging exon NB-statistics. Please refer to Wang et al. (2013) for detailed statistic analysis regarding differential splicing from exon count data.

```
> RCS_example <- estiExonNBstat(RCS_example)
> RCS_example <- estiGeneNBstat(RCS_example)
> head(fData(RCS_example)[, c("exonIDs", "geneIDs", "testable", "NBstat")])
```

```
                    exonIDs         geneIDs testable    NBstat
ENSG00000000003:001    E001 ENSG00000000003     TRUE 2.0219857
ENSG00000000003:002    E002 ENSG00000000003     TRUE 0.2486443
ENSG00000000003:003    E003 ENSG00000000003     TRUE 0.1238136
ENSG00000000003:004    E004 ENSG00000000003     TRUE 1.2058520
ENSG00000000003:005    E005 ENSG00000000003     TRUE 2.0668287
ENSG00000000003:006    E006 ENSG00000000003     TRUE 0.2678247
```

We run DS analysis on the permutation data sets as well. Here we set to run permutation 20 times for demonstration; however, in practice at least 1,000 permutations are recommended. To do so, we first generate a permutation matrix, each column corresponding to each permutation; then run DS analysis on the permutation data sets, and updated `permute_NBstat_gene` slot for results.

```
> permuteMat <- genpermuteMat(RCS_example, times=20)
> RCS_example <- DSpermute4GSEA(RCS_example, permuteMat)
> head(RCS_example@permute_NBstat_gene)
```

```
                 result.1  result.2  result.3  result.4  result.5  result.6
ENSG00000000003 0.4646897 0.1260575 0.6149483 0.4489870 0.4948214 0.5315174
ENSG00000000005 0.5356080 0.3363536 0.1704470 0.6456308 0.4412707 0.9279137
ENSG00000000419 0.6575135 0.5714363 0.8102716 0.9453603 0.7468450 0.6225239
ENSG00000000457 1.4381939 0.9180847 1.0670226 0.7770455 1.1478453 0.8626420
ENSG00000000460 0.9771965 0.9525316 1.2331664 0.9587737 0.6856721 0.7314033
ENSG00000000938 0.9918206 1.1312817 1.0750347 1.0949351 0.9143670 1.1026667
                 result.7  result.8  result.9 result.10 result.11 result.12
ENSG00000000003 0.4112458 0.3973255 0.9889178 0.8250040 0.9024498 1.2880614
ENSG00000000005 0.8038640 0.3243202 0.6378915 0.3885627 0.7616259 0.3456666
ENSG00000000419 0.4421745 0.8413454 1.8852414 0.5105660 0.3612803 1.2144647
ENSG00000000457 1.0674979 0.8369454 1.1451959 0.3344314 0.5265102 1.4213672
ENSG00000000460 1.1822440 0.9425534 0.7456354 0.4995767 1.3568056 0.8025370
ENSG00000000938 0.9213397 1.0290492 0.9272678 1.3589191 1.0878784 1.2607846
                result.13 result.14 result.15 result.16 result.17 result.18
ENSG00000000003 0.3912895 0.8716686 1.4492936 0.5710443 0.6222306 1.4141205
ENSG00000000005 0.9162519 0.3669275 0.3813821 0.5413718 0.5402945 0.5200415
ENSG00000000419 2.0124274 0.8895559 0.7519798 0.4361571 1.3088472 1.4203867
ENSG00000000457 1.4821813 1.0312590 0.6750650 0.7053969 0.7102569 0.4312357
ENSG00000000460 0.9527295 1.0291442 0.9256182 1.1110357 0.6949418 0.7951901
ENSG00000000938 1.5854014 1.1030971 1.3725386 1.0577913 1.2204159 1.4646221
```

```
                result.19 result.20
ENSG00000000003 0.2732064 0.8638212
ENSG00000000005 0.5447938 0.5826617
ENSG00000000419 0.8266270 1.0044120
ENSG00000000457 1.0330846 0.4223636
ENSG00000000460 0.8647431 0.4745880
ENSG00000000938 0.7891187 1.7110794
```

The DS NB-statistics from the permutation data sets offer an empirical background of NB-statistics on the real data set. By normalizing NB-statistics against this background, we get the DS scores, which will be used in integrated GSEA runs (Section 4).

```
> DSscore.normFac <- normFactor(RCS_example@permute_NBstat_gene)
> DSscore <- scoreNormalization(RCS_example@featureData_gene$NBstat,
+                               DSscore.normFac)
> DSscore.perm <- scoreNormalization(RCS_example@permute_NBstat_gene,
+                                    DSscore.normFac)
> DSscore[1:5]

ENSG00000000003 ENSG00000000005 ENSG00000000419 ENSG00000000457 ENSG00000000460
      2.0667125       0.8521664       1.8444609       3.2266420       0.8384265

> DSscore.perm[1:5,1:10]

                 result.1  result.2  result.3  result.4  result.5  result.6
ENSG00000000003 0.6661883 0.1807186 0.8816021 0.6436766 0.7093857 0.7619938
ENSG00000000005 0.9999329 0.6279424 0.3182095 1.2053358 0.8238133 1.7323329
ENSG00000000419 0.7201911 0.6259087 0.8875109 1.0354770 0.8180382 0.6818662
ENSG00000000457 1.5950136 1.0181921 1.1833701 0.8617740 1.2730056 0.9567039
ENSG00000000460 1.0908594 1.0633256 1.3766024 1.0702937 0.7654262 0.8164766
                 result.7  result.8  result.9 result.10
ENSG00000000003 0.5895702 0.5696137 1.4177321 1.1827420
ENSG00000000005 1.5007432 0.6054772 1.1908870 0.7254122
ENSG00000000419 0.4843250 0.9215469 2.0649526 0.5592359
ENSG00000000457 1.1838972 0.9282054 1.2700674 0.3708976
ENSG00000000460 1.3197570 1.0521867 0.8323642 0.5576851
```

2.3 DS permutation p-values

Besides calculating DS scores, based on the NB statistics on the real data set and the permutation data sets, we can also calculate a permutation p-value for each gene's DS significance in the studied data set.

```
> RCS_example <- DSpermutePval(RCS_example, permuteMat)
> head(DSresultGeneTable(RCS_example))

           geneID    NBstat pvalue    padjust
1 ENSG00000000003 1.4416043   0.05 0.08888889
2 ENSG00000000005 0.4564578   0.60 0.64785276
3 ENSG00000000419 1.6839390   0.10 0.15575221
4 ENSG00000000457 2.9094026   0.00 0.00000000
5 ENSG00000000460 0.7510661   0.70 0.72470588
6 ENSG00000000938 1.7949049   0.00 0.00000000
```

The adjusted p-values accounting for multiple testings were given by the BH method (Benjamini and Hochberg, 1995). Users can also apply function `topDSGenes` and function `topDSExons` to quickly get the most significant DS genes and exons, respectively.

3 Differential expression analysis and DE scores

3.1 Gene read count data from *ReadCountSet* class

For gene DE analysis, read counts on each gene should be first calculated. With *SeqGSEA*, users usually analyze DE and DS simultaneously, so the package includes the function `getGeneCount` to facilitate gene read count calculation from a `ReadCountSet` object.

```
> geneCounts <- getGeneCount(RCS_example)
> dim(geneCounts) # 182  20

[1] 182  20

> head(geneCounts)

                 S1  S2  S3  S4  S5  S6   S7  S8  S9 S10  C1  C2  C3  C4  C5
ENSG00000000003 495 235 386 272 255 815 1065 803 839 885 278 270 238 175 292
ENSG00000000005  19   1   0   2   2  12    7   3   1   4   4   4   2   0   1
ENSG00000000419 196 134 165 184 132 344  343 307 342 280 179 156 100 120 126
ENSG00000000457  97  78 141  72 102 219  344 277 337 249  62  48  40  43  52
ENSG00000000460  52  35  48  25  47 105  124  80 156 145  48  36  21  34  19
ENSG00000000938  27  44  57  43  14  71   74 146 148 165  48  59  32  79  20
                 C6   C7  C8  C9 C10
ENSG00000000003 432  519 621 475 560
ENSG00000000005   9    3   1  14  46
ENSG00000000419 169  255 171 164 201
ENSG00000000457 170  165 131 183 185
ENSG00000000460  68   90  48  72  38
ENSG00000000938 103 1285 137 156  90
```

This function results in a matrix of 182 rows and 20 columns, corresponding to 182 genes and 20 samples.

3.2 DE analysis and DE scores

DE analysis has been implemented in several R/`Bioconductor` packages, of which *DESeq* (Anders and Huber, 2010) is mainly utilized in *SeqGSEA* for DE analysis. With *DESeq*, we can model count data with negative binomial distributions for accounting biological variations and various biases introduced in RNA-Seq. Given the read count data on individual genes and sample grouping information, basic DE analysis based on *DESeq* including size factor estimation and dispersion estimation, is encapsulated in the function `runDESeq`.

```
> label <- label(RCS_example)
> DEG <- runDESeq(geneCounts, label)
```

The function `runDESeq` returns a `CountDataSet` object, which is defined in the *DESeq* package. The DE analysis in the *DESeq* package continues with the output `CountDataSet` object and conducts negative-binomial-based statistical tests for DE genes (using `nbinomTest` or `nbinomGLMTest`). However, in this *SeqGSEA* package, we define NB statistics to quantify each gene's expression difference between sample groups.

The NB statistics for DE can be achieved by the following scripts.

```
> DEGres <- DENBStat4GSEA(DEG)
> DEGres[1:5, "NBstat"]

[1]  0.5426504  0.2503510  0.0231052 14.3384053  1.4101270
```

Similarly, we run DE analysis on the permutation data sets as well. The `permuteMat` should be the same as used in DS analysis on the permutation data sets.

```
> DEpermNBstat <- DENBStatPermut4GSEA(DEG, permuteMat)
> DEpermNBstat[1:5, 1:10]

      result.1  result.2   result.3   result.4  result.5    result.6
[1,] 1.8692547 0.8213232 0.09132766 0.58758090 0.1075498 0.001307067
[2,] 2.2015174 0.8331144 6.56414436 1.01761840 4.5251370 0.016317593
[3,] 2.9609004 0.3128548 0.45632175 0.32311173 0.5166499 0.027741041
[4,] 0.9390268 2.7765719 0.01513649 0.04289388 0.6860026 0.037567244
[5,] 0.8982830 1.5601273 0.76208668 1.53935215 2.1748038 0.146443776
        result.7   result.8   result.9 result.10
[1,] 0.039538646 1.05962387 0.26546063 0.4273147
[2,] 1.681705844 0.06648999 0.02355817 1.1210531
[3,] 0.338406073 2.14049659 0.44541465 0.4664437
[4,] 0.002014416 0.22603359 0.70698151 0.6077455
[5,] 0.003562648 0.28149831 0.26544262 1.0652773
```

Once again, the DE NB-statistics from the permutation data sets offer an empirical background, so we can normalize NB-statistics against this background. By doing so, we get the DE scores, which will also be used in integrated GSEA runs (Section 4).

```
> DEscore.normFac <- normFactor(DEpermNBstat)
> DEscore <- scoreNormalization(DEGres$NBstat, DEscore.normFac)
> DEscore.perm <- scoreNormalization(DEpermNBstat, DEscore.normFac)
> DEscore[1:5]

[1]  0.79528665  0.22079582  0.02207349 23.23097380  1.86253740

> DEscore.perm[1:5, 1:10]

     result.1  result.2   result.3   result.4  result.5    result.6    result.7
[1,] 2.739505 1.2036984 0.13384616 0.86113500 0.1576207 0.001915585 0.057946254
[2,] 1.941617 0.7347611 5.78921479 0.89748354 3.9909223 0.014391221 1.483172187
[3,] 2.828687 0.2988849 0.43594563 0.30868383 0.4935799 0.026502322 0.323295238
[4,] 1.521404 4.4985804 0.02452402 0.06949633 1.1114561 0.060866159 0.003263741
[5,] 1.186479 2.0606623 1.00658662 2.03322185 2.8725452 0.193427271 0.004705650
       result.8   result.9 result.10
[1,] 1.55294225 0.38904845 0.6262552
[2,] 0.05864052 0.02077701 0.9887073
[3,] 2.04491706 0.42552556 0.4456156
[4,] 0.36621787 1.14544600 0.9846645
[5,] 0.37181130 0.35060446 1.4070497
```

3.3 DE permutation p-values

Similar to DS analysis, comparing NB-statistics on the real data set and those on the permutation data sets, we can get permutation p-values for each gene's DE significance.

```
> DEGres <- DEpermutePval(DEGres, DEpermNBstat)
> DEGres[1:6, c("NBstat", "perm.pval", "perm.padj")]

                    NBstat perm.pval perm.padj
ENSG00000000003  0.5426504      0.45         1
```

```
ENSG00000000005  0.2503510      0.50         1
ENSG00000000419  0.0231052      1.00         1
ENSG00000000457 14.3384053      0.00         0
ENSG00000000460  1.4101270      0.20         1
ENSG00000000938  2.1976989      0.00         0
```

For a comparison to the nominal p-values from exact testing and forming comprehensive results, users can run `DENBTest` first and then `DEpermutePval`, which generates results as follows.

```
> DEGres <- DENBTest(DEG)
> DEGres <- DEpermutePval(DEGres, DEpermNBstat)
> DEGres[1:6, c("NBstat", "pval", "padj", "perm.pval", "perm.padj")]

                    NBstat         pval         padj perm.pval perm.padj
ENSG00000000003  0.5426504 3.956985e-01 5.408276e-01      0.45         1
ENSG00000000005  0.2503510 3.300042e-01 4.943803e-01      0.50         1
ENSG00000000419  0.0231052 9.244775e-01 9.839468e-01      1.00         1
ENSG00000000457 14.3384053 9.960426e-05 2.589711e-03      0.00         0
ENSG00000000460  1.4101270 1.370959e-01 2.970412e-01      0.20         1
ENSG00000000938  2.1976989 7.309013e-07 4.434134e-05      0.00         0
```

4 Integrative GSEA runs

4.1 DE/DS score integration

We have proposed two strategies for integrating normalized DE and DS scores (Wang and Cairns, 2013), one of which is the weighted summation of the two scores and the other is a rank-based strategy. The functions `geneScore` and `genePermuteScore` implement two methods for the weighted summation strategy: weighted linear combination and weighted quadratic combination. Scripts below show a linear combination of DE and DS scores with weight for DE equal to 0.3. Users should keep the weight for DE in `geneScore` and `genePermuteScore` the same, and the weight rangs from 0 (i.e., DS only) to 1 (i.e., DE only). Visualization of gene scores can be made by applying the `plotGeneScore` function.

```
> gene.score <- geneScore(DEscore, DSscore, method="linear", DEweight = 0.3)
> gene.score.perm <- genePermuteScore(DEscore.perm, DSscore.perm,
+                                     method="linear",  DEweight=0.3)
> plotGeneScore(gene.score, gene.score.perm)
```

The plot generated by the `plotGeneScore` function (Fig. 1) can also be saved as a PDF file easily with the `pdf` argument of `plotGeneScore`.

The functions `geneScore` and `genePermuteScore` also implement one method for the rank-based integration strategy: using data-set-specific ranks. The plot for integrated gene scores is shown in Fig. 2.

```
> gene.score <- geneScore(DEscore, DSscore, method="rank", DEweight = 0.3)
> gene.score.perm <- genePermuteScore(DEscore.perm, DSscore.perm,
+                                     method="rank",  DEweight=0.3)
> plotGeneScore(gene.score, gene.score.perm)
```

Rather than the above method to integrate scores with data-set-specific ranks, an alternative method is implemented with the `rankCombine` function, which takes only the ranks from the real data set for integrating DE and DS scores on both real and permutation data sets. This provides a method in a global manner. The plot of gene scores is shown in Fig. 3.

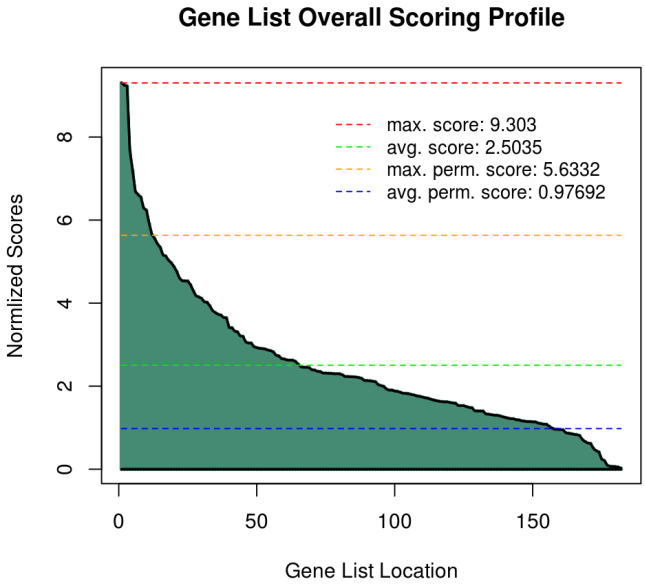

Figure 1: Gene scores resulted from linear combination. Scores are sorted from the largest to the smallest. Red, green, orange, blue dotted horizontal lines represent the maximum score, average score on the real data set, and the maximum score, average score on the permutation data sets.

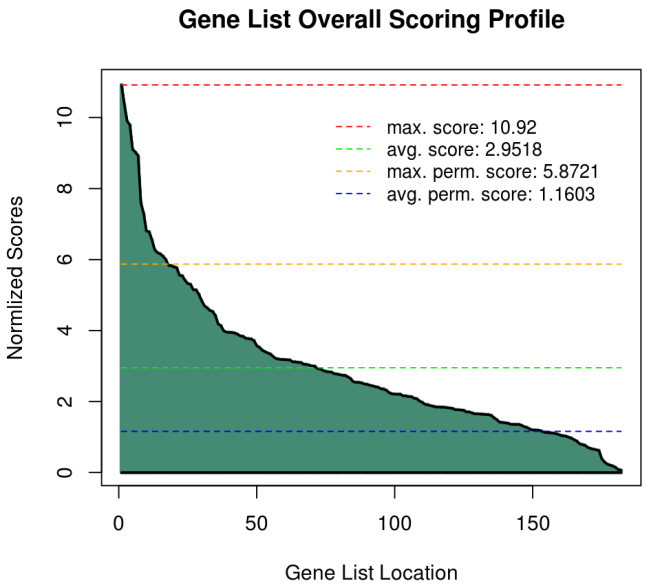

Figure 2: Gene scores resulted from rank-based combination with data-set-specific ranks. Scores are sorted from the largest to the smallest. Red, green, orange, blue dotted horizontal lines represent the maximum score, average score on the real data set, and the maximum score, average score on the permutation data sets.

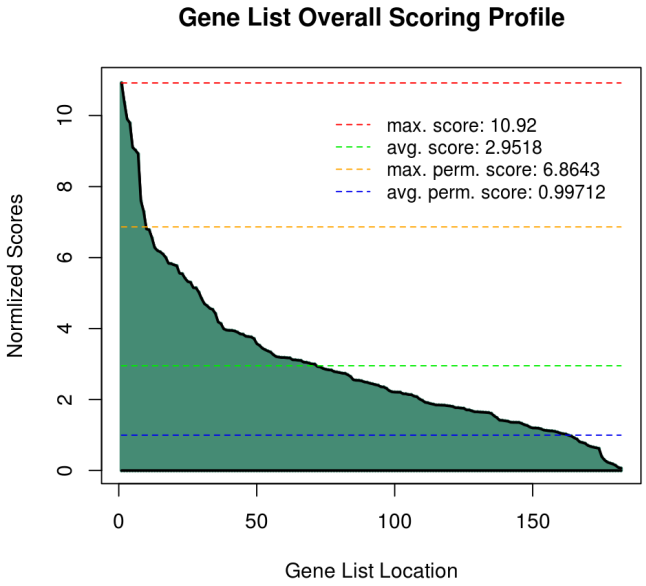

Figure 3: Gene scores resulted from rank-based combination with the same rank got from the real data set. Scores are sorted from the largest to the smallest. Red, green, orange, blue dotted horizontal lines represent the maximum score, average score on the real data set, and the maximum score, average score on the permutation data sets.

```
> combine <- rankCombine(DEscore, DSscore, DEscore.perm, DSscore.perm, DEweight=0.3)
> gene.score <- combine$geneScore
> gene.score.perm <- combine$genePermuteScore
> plotGeneScore(gene.score, gene.score.perm)
```

Basically the integrated gene scores are distributed similarly with the three integration methods at DE weight 0.3 (Figs. 1, 2, and 3); however, according to the analysis in Wang and Cairns (2013), SeqGSEA can detect slightly more significant gene sets with rank-based integration strategy than with linear combination.

4.2 Initialization of *SeqGeneSet* objects

To facilitate running gene set enrichment analysis, *SeqGSEA* implements a *SeqGeneSet* class. The *SeqGeneSet* class has several slots for accommodating a category of gene sets derived from any biological knowledge-based databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG). However, we recommend to start with the formatted gene-set files from the well-maintained resource Molecular Signatures Database (MSigDB, `http://www.broadinstitute.org/gsea/msigdb/index.jsp`) (Subramanian et al., 2005). After downloading a `gmt` file from the above URL, users can use `loadGenesets` to initialize a `SeqGeneSet` object easily. Please note that with the current version of *SeqGSEA*, only gene sets with gene symbols are supported, though read count data's gene IDs can be either gene symbols or Ensembl Gene IDs.

Below is shown an example of the `SeqGeneSet` object, which contains information such as how many gene sets in this object and the names/sizes/descriptions of each gene set.

```
> data(GS_example, package="SeqGSEA")
> GS_example
```

```
SeqGeneSet object: gs_symb.txt
GeneSetSourceFile: /Library/Frameworks/R.framework/Versions/2.15/Resources/library/SeqGSEA/extdat
GeneSets: ERB2_UP.V1_DN
          AKT_UP_MTOR_DN.V1_UP
          ...
          KRAS.600.LUNG.BREAST_UP.V1_DN
  with the number of genes in respective sets: 6, 6, ..., 5
  brief descriptions:
          http://www.broadinstitute.org/gsea/msigdb/cards/ERB2_UP.V1_DN
          http://www.broadinstitute.org/gsea/msigdb/cards/AKT_UP_MTOR_DN.V1_UP
          ...
          http://www.broadinstitute.org/gsea/msigdb/cards/KRAS.600.LUNG.BREAST_UP.V1_DN
  # gene sets passed filter: 11  (#genes >= 5 AND <= 1000)
  # gene sets excluded: 178  (#genes < 5 OR > 1000)
ES scores: not computed
ES postions: not computed
Permutated ES scores: not performed
ES scores normalized: No
ES p-value: not computed
ES FWER: not computed
ES FDR: not computed
```

4.3 running GSEA with integrated gene scores

With the initialized `SeqGeneSet` object and integrated gene scores as well as gene scores on the permutation data sets, the main `GSEnrichAnlyze` can be executed; and the `topGeneSets` allows users promptly access to the top significant gene sets.

```
> GS_example <- GSEnrichAnalyze(GS_example, gene.score, gene.score.perm)
> topGeneSets(GS_example, 5)
```

```
                          GSName GSSize       ES ES.pos pval     FDR FWER
8                 HOXA9_DN.V1_UP      5 1.682211     38 0.05 0.00000 0.30
9                     TBK1.DF_UP      5 1.900735     27 0.00 0.00000 0.20
5                   PKCA_DN.V1_DN      5 1.482182     71 0.20 0.33333 0.65
11 KRAS.600.LUNG.BREAST_UP.V1_DN      5 1.344223     29 0.30 0.50000 0.85
10                      NFE2L2.V2      9 1.301831     66 0.25 0.60000 0.85
```

The main GSEA includes several steps detailed in Wang and Cairns (2013) and its original paper Subramanian et al. (2005). In *SeqGSEA*, functions `calES`, `calES.perm`, `normES` and `signifES` are implemented to complete the analysis. Advanced users may set up customized pipelines with the functions above themselves.

4.4 *SeqGSEA* result displays

Several functions in *SeqGSEA* can be employed for visualization of gene set enrichment analysis running results. The `plotES` function is to plot the distribution of normalized enrichment scores (NES) of all gene sets in a `SeqGeneSet` object on the observed data set versus its empirical background provided by the NES on the permutation data sets (Fig. 4).

```
> plotES(GS_example)
```

The `plotSig` function plots the distributions of permutation p-value, false discovery rate (FDR) and family-wise error rate (FWER) versus NES. The example plot is not shown in this vignette as the distributions can be far from the real ones due to the limited permutation times.

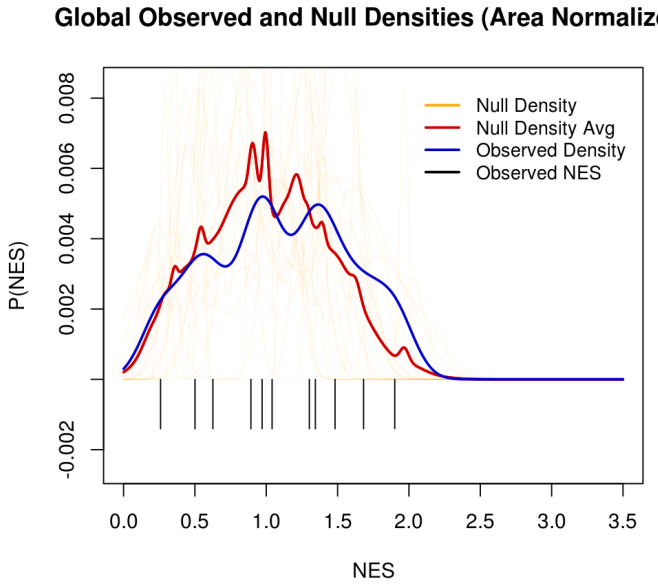

Figure 4: Distribution of normalized enrichment scores (NES) on the observed and permutation (null) data sets. Blue: observed NES density; Orange: each for NES density on one permutation data set; Red: the average density on all permutation data sets; Black: observed NES values.

```
> plotSig(GS_example)
```

The `plotSigGS` function is to plot detailed results of a particular gene set that has been analyzed. Information in the plot includes running enrichment scores, null NES on the permutation data sets. See Fig. 5 for an example.

```
> plotSigGeneSet(GS_example, 9, gene.score) # 9th gene set is the most significant one.
```

Besides the functions to generate plots, the `writeSigGeneSet` function can write the detailed information of any analyzed gene sets, including NES, p-values, FDR, and the leading set (see the definition in Wang and Cairns (2013)). An example is shown below.

```
> writeSigGeneSet(GS_example, 9, gene.score) # 9th gene set is the most significant one.
```

```
GSEA result for gene set No. 9:
genesetName         gs_symb.txt:TBK1.DF_UP
genesetSize         5
genesetDesc         http://www.broadinstitute.org/gsea/msigdb/cards/TBK1.DF_UP
NES         1.90073475156584
Pos         27
pvalue         0
FDR         0
FWER         0.2

Leading set:
ENSG00000002919         9.10366103395513
ENSG00000001167         7.2994306528263
ENSG00000005194         5.14755492019088
```

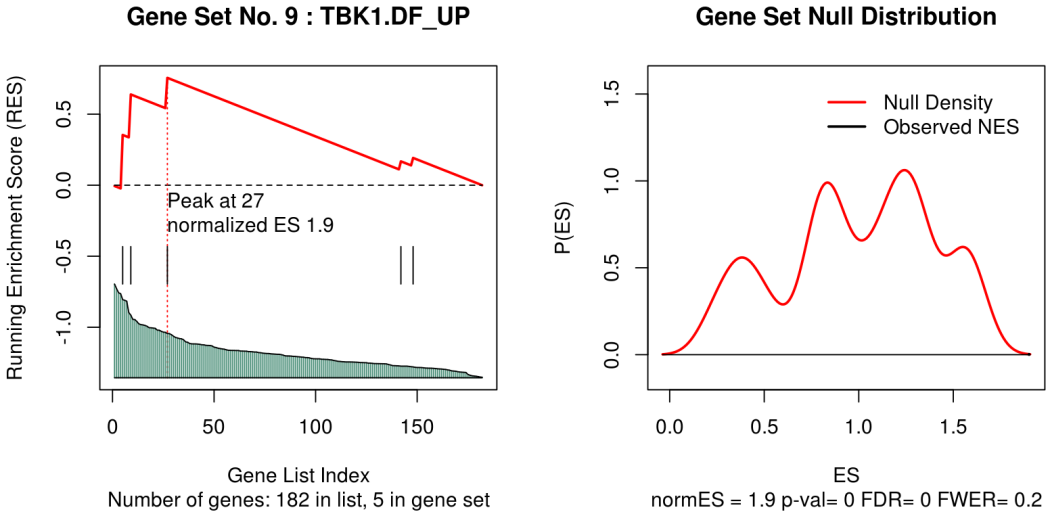

Figure 5: Left: gene locations of a particular gene set according to the gene score rank and running enrichment scores; Right: null NES distribution and the relative position of the observed NES.

```
Whole gene set:
ENSG00000002919        9.10366103395513
ENSG00000001167        7.2994306528263
ENSG00000005194        5.14755492019088
ENSG00000006576        1.35863953641559
ENSG00000005059        1.26725834863045
```

The `GSEAresultTable` generates a summary table of the GSEA analysis, which can also be output with customized scripts. An example can be found in Section 6.

5 Running *SeqGSEA* with multiple cores

5.1 R-parallel packages

There are many R packages for facilitating users in running R scripts in parallel, including *parallel*, *snowfall*, *multicore*, and many others. While experienced users may parallelize *SeqGSEA* runnings with the above packages themselves to reduce the running time, we provide with in the *SeqGSEA* package vignette an general way for users to parallelize their runnings utilizing the *doParallel* package (which depends on *parallel*).

First, we show a toy example for a basic idea how *doParallel* works. Basically, *doParallel* is a *parallel backend* for the *foreach* package using *parallel*, which provides a mechanism to execute *foreach* loops in parallel. With the `foreach` function in the *foreach* package, we can specify which `foreach` loops need to be parallelized using the `%dopar%` operator. However, without a registered parallel backend, the `foreach` loops will be executed sequentially even if the `%dopar%` operator is used. In those cases, the *foreach* package will issue a warning that it is running sequentially. Below are two running examples showing how the task is running sequentially and in parallel, respectively.

Run sequentially without parallel backend registered

```
> library(doParallel)
> a <- matrix(1:16, 4, 4)
```

```
> b <- t(a)
> foreach(b=iter(b, by='col'), .combine=cbind) %dopar%
+   (a %*% b)
```

```
     [,1] [,2] [,3] [,4]
[1,]  276  304  332  360
[2,]  304  336  368  400
[3,]  332  368  404  440
[4,]  360  400  440  480
```

Although the warning message didn't appear here, you would definitely see a warning message when you run the scripts above, like:
Warning message:
executing %dopar% sequentially: no parallel backend registered

Run in parallel with two cores

```
> library(doParallel)
> cl <- makeCluster(2) # specify 2 cores to be used in this computing
> registerDoParallel(cl)
> getDoParWorkers() # 2
```

```
[1] 2
```

```
> a <- matrix(1:16, 4, 4)
> b <- t(a)
> foreach(b=iter(b, by='col'), .combine=cbind) %dopar%
+   (a %*% b)
```

```
     [,1] [,2] [,3] [,4]
[1,]  276  304  332  360
[2,]  304  336  368  400
[3,]  332  368  404  440
[4,]  360  400  440  480
```

The parallel backend registration was done with `registerdoParallel`. For more details please refer to *doParallel*'s vignette (`http://cran.r-project.org/web/packages/doParallel/index.html`).

5.2 Parallelizing analysis on permutation data sets

In *SeqGSEA*, the loops for analyzing permutation data sets are implemented by `foreach` with `%dopar%` operator used. Those loops include DS, DE, and GSEA analyses, which are the most time consuming parts. Although there are three parts can take the advantage of parallel running, users only need to register parallel backend once at the beginning of all analyses. See an analysis example in the next section (Section 6).

6 Analysis examples

6.1 Starting from your own RNA-Seq data

With this *SeqGSEA* package, we provide complementary Python scripts for counting reads on exons of each genes from SAM/BAM files: two scripts `prepare_exon_annotation_refseq.py` and `prepare_exon_annotation_ensembl.py` for preparing (sub-)exon annotation, and `count_in_exons.py` for counting reads. The scripts are based on the `HTSeq` Python package (`http://www-huber.embl.de/users/anders/HTSeq/`). Please install it before using the Python scripts provided. The scripts can be found in the directory given by the following command.

```
> system.file("extscripts", package="SeqGSEA", mustWork=TRUE)
```

```
[1] "/tmp/RtmpoA7tNr/Rinst1d5e370236fc/SeqGSEA/extscripts"
```

Simply by typing "python" + the file name of echo script in your shell console, the help documentation will be on your screen.

Other than the Python scripts provided, users who prefer playing with R/`Bioconductor` packages can also use `easyRNASeq` in *easyRNASeq*, `summarizeOverlaps` in *GenomicRanges*, and `featureCounts` in *Rsubread* to count reads that mapped to each exon. Please refer to respective packages for detailed usage.

For users who are not familiar with RNA-Seq data processing, the upstream steps of counting reads are (1) data preprocessing, including adapter removal, low-quality read filtering, data quality-control analysis, and (2) read mapping. R/`Bioconductor` users can apply *Rsubread* to map reads based on a seed-and-vote approach, as well as a few QC analysis. Users familiar with command-line can choose from a wide range of tools, such as already widely used ones including `TopHat` (`http://tophat.cbcb.umd.edu`), `START` (`http://code.google.com/p/rna-star`), and etc..

6.2 Exemplified pipeline for integrating DE and DS

Below is shown a typical SeqGSEA running example with the data enclosed with the *SeqGSEA* package, which are a part of the prostate cancer data set (Kannan et al., 2011). We divide the process into five steps for a complete SeqGSEA run.

Step 0: Initialization. (Users should change values in this part accordingly.)

```
> rm(list=ls())
> # input count data files
> data.dir <- system.file("extdata", package="SeqGSEA", mustWork=TRUE)
> case.pattern <- "^SC"  # file name starting with "SC"
> ctrl.pattern <- "^SN"  # file name starting with "SN"
> case.files <- dir(data.dir, pattern=case.pattern, full.names = TRUE)
> control.files <- dir(data.dir, pattern=ctrl.pattern, full.names = TRUE)
> # gene set file
> geneset.file <- system.file("extdata", "gs_symb.txt",
+                             package="SeqGSEA", mustWork=TRUE)
> # output file prefix
> output.prefix <- "SeqGSEA.test"
> # setup parallel backend
> library(doParallel)
> cl <- makeCluster(2) # specify 2 cores to be used in computing
> registerDoParallel(cl) # parallel backend registration
> # setup permutation times
> perm.times <- 20 # change the number to >= 1000 in your analysis
```

Step 1: DS analysis

```
> # load exon read count data
> RCS <- loadExonCountData(case.files, control.files)
> # remove genes with low expression
> RCS <- exonTestability(RCS, cutoff=5)
> geneTestable <- geneTestability(RCS)
> RCS <- subsetByGenes(RCS, unique(geneID(RCS))[ geneTestable ])
> # get gene IDs, which will be used in initialization of gene set
> geneIDs <- unique(geneID(RCS))
> # calculate DS NB statistics
```

```
> RCS <- estiExonNBstat(RCS)
> RCS <- estiGeneNBstat(RCS)
> # calculate DS NB statistics on the permutation data sets
> permuteMat <- genpermuteMat(RCS, times=perm.times)
> RCS <- DSpermute4GSEA(RCS, permuteMat)
```

Step 2: DE analysis

```
> # get gene read counts
> geneCounts <- getGeneCount(RCS)
> # calculate DE NB statistics
> label <- label(RCS)
> DEG <-runDESeq(geneCounts, label)
> DEGres <- DENBStat4GSEA(DEG)
> # calculate DE NB statistics on the permutation data sets
> DEpermNBstat <- DENBStatPermut4GSEA(DEG, permuteMat) # permutation
```

Step 3: score integration

```
> # DE score normalization
> DEscore.normFac <- normFactor(DEpermNBstat)
> DEscore <- scoreNormalization(DEGres$NBstat, DEscore.normFac)
> DEscore.perm <- scoreNormalization(DEpermNBstat, DEscore.normFac)
> # DS score normalization
> DSscore.normFac <- normFactor(RCS@permute_NBstat_gene)
> DSscore <- scoreNormalization(RCS@featureData_gene$NBstat, DSscore.normFac)
> DSscore.perm <- scoreNormalization(RCS@permute_NBstat_gene, DSscore.normFac)
> # score integration
> gene.score <- geneScore(DEscore, DSscore, DEweight=0.5)
> gene.score.perm <- genePermuteScore(DEscore.perm, DSscore.perm, DEweight=0.5)
> # visilization of scores
> # NOT run in the example; users to uncomment the following 6 lines to run
> #plotGeneScore(DEscore, DEscore.perm, pdf=paste(output.prefix,".DEScore.pdf",sep=""),
> #              main="Expression")
> #plotGeneScore(DSscore, DSscore.perm, pdf=paste(output.prefix,".DSScore.pdf",sep=""),
> #              main="Splicing")
> #plotGeneScore(gene.score, gene.score.perm,
> #              pdf=paste(output.prefix,".GeneScore.pdf",sep=""))
```

Step 4: main GSEA

```
> # load gene set data
> gene.set <- loadGenesets(geneset.file, geneIDs, geneID.type="ensembl",
+                          genesetsize.min = 5, genesetsize.max = 1000)
> # enrichment analysis
> gene.set <- GSEnrichAnalyze(gene.set, gene.score, gene.score.perm, weighted.type=1)
> # format enrichment analysis results
> GSEAres <- GSEAresultTable(gene.set, TRUE)
> # output results
> # NOT run in the example; users to uncomment the following 4 lines to run
> #write.table(GSEAres, paste(output.prefix,".GSEA.result.txt",sep=""),
> #            quote=FALSE, sep="\t", row.names=FALSE)
> #plotES(gene.set, pdf=paste(output.prefix,".GSEA.ES.pdf",sep=""))
> #plotSig(gene.set, pdf=paste(output.prefix,".GSEA.FDR.pdf",sep=""))
```

For gene sets used in Step 4, while we recommend users directly download and use those already well-formatted gene sets from MSigDB (`http://www.broadinstitute.org/gsea/msigdb/index.jsp`), users can also feed whatever gene sets to SeqGSEA as long as they are in the GMT format. Please refer to the following URL for details: `http://www.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats`.

6.3 Exemplified pipeline for DE-only analysis

For the demanding of DE-only analysis, such as for organisms without much alternative splicing annotated, here we show an exemplified pipeline for such analysis. It includes 4 steps as follows.

Step 0: Initialization. (Users should change values in this part accordingly.)

```
> rm(list=ls())
> # input count data files
> data.dir <- system.file("extdata", package="SeqGSEA", mustWork=TRUE)
> count.file <- paste(data.dir,"geneCounts.txt",sep="/")
> # gene set file
> geneset.file <- system.file("extdata", "gs_symb.txt",
+                             package="SeqGSEA", mustWork=TRUE)
> # output file prefix
> output.prefix <- "SeqGSEA.test"
> # setup parallel backend
> library(doParallel)
> cl <- makeCluster(2) # specify 2 cores to be used in computing
> registerDoParallel(cl) # parallel backend registration
> # setup permutation times
> perm.times <- 20 # change the number to >= 1000 in your analysis
```

Step 1: DE analysis

```
> # load gene read count data
> geneCounts <- read.table(count.file)
> # speficify the labels of each sample
> label <- as.factor(c(rep(1,10), rep(0,10)))
> # calculate DE NB statistics
> DEG <-runDESeq(geneCounts, label)
> DEGres <- DENBStat4GSEA(DEG)
> # calculate DE NB statistics on the permutation data sets
> permuteMat <- genpermuteMat(label, times=perm.times)
> DEpermNBstat <- DENBStatPermut4GSEA(DEG, permuteMat) # permutation
```

Step 2: score normalization

```
> # DE score normalization
> DEscore.normFac <- normFactor(DEpermNBstat)
> DEscore <- scoreNormalization(DEGres$NBstat, DEscore.normFac)
> DEscore.perm <- scoreNormalization(DEpermNBstat, DEscore.normFac)
> # score integration - DSscore can be null
> gene.score <- geneScore(DEscore, DEweight=1)
> gene.score.perm <- genePermuteScore(DEscore.perm, DEweight=1)  # visilization of scores
> # NOT run in the example; users to uncomment the following 6 lines to run
> #plotGeneScore(DEscore, DEscore.perm, pdf=paste(output.prefix,".DEScore.pdf",sep=""),
> #              main="Expression")
> #plotGeneScore(gene.score, gene.score.perm,
> #              pdf=paste(output.prefix,".GeneScore.pdf",sep=""))
```

Step 3: main GSEA

```
> # load gene set data
> geneIDs <- rownames(geneCounts)
> gene.set <- loadGenesets(geneset.file, geneIDs, geneID.type="ensembl",
+                          genesetsize.min = 5, genesetsize.max = 1000)
> # enrichment analysis
> gene.set <- GSEnrichAnalyze(gene.set, gene.score, gene.score.perm, weighted.type=1)
> # format enrichment analysis results
> GSEAres <- GSEAresultTable(gene.set, TRUE)
> # output results
> # NOT run in the example; users to uncomment the following 4 lines to run
> #write.table(GSEAres, paste(output.prefix,".GSEA.result.txt",sep=""),
> #            quote=FALSE, sep="\t", row.names=FALSE)
> #plotES(gene.set, pdf=paste(output.prefix,".GSEA.ES.pdf",sep=""))
> #plotSig(gene.set, pdf=paste(output.prefix,".GSEA.FDR.pdf",sep=""))
```

6.4 One-step SeqGSEA analysis

While users can choose to run SeqGSEA step by step in a well-controlled manner (see above), the one-step SeqGSEA analysis with an all-in `runSeqGSEA` function enables users to run SeqGSEA in the easiest way. With the `runSeqGSEA` function, users can also test multiple weights for integrating DE and DS scores. DE-only analysis starting with exon read counts is also supported in the all-in function.

Follow the example below to start your first SeqGSEA analysis now!

```
> ### Initialization ###
> # input file location and pattern
> data.dir <- system.file("extdata", package="SeqGSEA", mustWork=TRUE)
> case.pattern <- "^SC" # file name starting with "SC"
> ctrl.pattern <- "^SN" # file name starting with "SN"
> # gene set file and type
> geneset.file <- system.file("extdata", "gs_symb.txt",
+                             package="SeqGSEA", mustWork=TRUE)
> geneID.type <- "ensembl"
> # output file prefix
> output.prefix <- "SeqGSEA.example"
> # analysis parameters
> nCores <- 8
> perm.times <- 1000 # >= 1000 recommended
> DEonly <- FALSE
> DEweight <- c(0.2, 0.5, 0.8)  # a vector for different weights
> integrationMethod <- "linear"
>
> ### one step SeqGSEA running ###
> # NOT run in the example; uncomment the following 4 lines to run
> # CAUTION: running the following lines will generate lots of files in your working dir
> #runSeqGSEA(data.dir=data.dir, case.pattern=case.pattern, ctrl.pattern=ctrl.pattern,
> #           geneset.file=geneset.file, geneID.type=geneID.type, output.prefix=output.prefix,
> #           nCores=nCores, perm.times=perm.times, integrationMethod=integrationMethod,
> #           DEonly=DEonly, DEweight=DEweight)
```

7 Session information

```
> sessionInfo()

R version 3.0.2 Patched (2013-12-18 r64488)
Platform: x86_64-unknown-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] parallel  stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
 [1] doParallel_1.0.6   iterators_1.0.6    SeqGSEA_1.2.1      foreach_1.4.1
 [5] biomaRt_2.18.0     DESeq_1.14.0       lattice_0.20-24    locfit_1.5-9.1
 [9] Biobase_2.22.0     BiocGenerics_0.8.0

loaded via a namespace (and not attached):
 [1] AnnotationDbi_1.24.0 DBI_0.2-7            IRanges_1.20.6
 [4] RColorBrewer_1.0-5   RCurl_1.95-4.1       RSQLite_0.11.4
 [7] XML_3.98-1.1         annotate_1.40.0      codetools_0.2-8
[10] compiler_3.0.2       genefilter_1.44.0    geneplotter_1.40.0
[13] grid_3.0.2           splines_3.0.2        stats4_3.0.2
[16] survival_2.37-7      tools_3.0.2          xtable_1.7-1
```

Cleanup

This is a cleanup step for the vignette on Windows; typically not needed for users.

```
> allCon <- showConnections()
> socketCon <- as.integer(rownames(allCon)[allCon[, "class"] == "sockconn"])
> sapply(socketCon, function(ii) close.connection(getConnection(ii)) )
```

References

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

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, 57(1):289–300.

Kannan, K., Wang, L., Wang, J., Ittmann, M. M., Li, W., and Yen, L. (2011). Recurrent chimeric rnas enriched in human prostate cancer identified by deep sequencing. *Proc Natl Acad Sci U S A*, 108(22):9172–7.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–50.

Wang, W., Qin, Z., Feng, Z., Wang, X., and Zhang, X. (2013). Identifying differentially spliced genes from two groups of rna-seq samples. *Gene*, 518(1):164–170.

Wang, X. and Cairns, M. (2013). Gene set enrichment analysis of RNA-Seq data: integrating differential expression and splicing. *BMC Bioinformatics*, 14(Suppl 5):S16.