

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

Xi Wang^{1,2} and Murray Cairns^{1,2,3}

May 1, 2013

¹School of Biomedical Sciences and Pharmacy, The University of Newcastle, Callaghan, New South Wales, Australia

²Hunter Medical Research Institute, New Lambton, New South Wales, Australia

³Schizophrenia Research Institute, Sydney, New South Wales, Australia

`xi.wang@newcastle.edu.au`

Contents

| | | |
|----------|---|-----------|
| 1 | Introduction | 2 |
| 1.1 | Background | 2 |
| 1.2 | Getting started | 2 |
| 1.3 | Package citation | 2 |
| 2 | Differential splicing analysis and DS scores | 2 |
| 2.1 | The <i>ReadCountSet</i> class | 2 |
| 2.2 | DS analysis and DS scores | 3 |
| 2.3 | DS permutation p-values | 5 |
| 3 | Differential expression analysis and DE scores | 6 |
| 3.1 | Gene read count data from <i>ReadCountSet</i> class | 6 |
| 3.2 | DE analysis and DE scores | 6 |
| 3.3 | DE permutation p-values | 7 |
| 4 | Integrative GSEA runs | 8 |
| 4.1 | DE/DS score integration | 8 |
| 4.2 | Initialization of <i>SeqGeneSet</i> objects | 10 |
| 4.3 | running GSEA with integrated gene scores | 11 |
| 4.4 | <i>SeqGSEA</i> result displays | 11 |
| 5 | Running <i>SeqGSEA</i> with multiple cores | 13 |
| 5.1 | R-parallel packages | 13 |
| 5.2 | Parallelizing analysis on permutation data sets | 14 |
| 6 | An analysis example | 14 |
| 6.1 | Starting from your own RNA-Seq data | 14 |
| 6.2 | An exemplified pipeline | 15 |
| 7 | Session information | 16 |

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 `exonTestability`

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 | 0.5711706 | 0.5279423 | 0.3274380 | 0.6182864 | 1.2228015 | 1.7415362 |
| ENSG00000000005 | 0.3054307 | 0.5937646 | 0.6489390 | 0.1887808 | 0.4136961 | 0.3589379 |
| ENSG00000000419 | 0.5301508 | 0.3378401 | 0.5199378 | 1.0887625 | 0.7012219 | 1.1806827 |
| ENSG00000000457 | 0.9801163 | 0.3986746 | 0.9951899 | 0.4982123 | 0.7348845 | 1.6316861 |
| ENSG00000000460 | 0.8316046 | 0.3786938 | 0.7716115 | 0.5870078 | 0.6806871 | 0.3944324 |
| ENSG00000000938 | 1.1328115 | 1.1063075 | 0.9733312 | 1.2418018 | 0.7524729 | 1.1023085 |
| | result.7 | result.8 | result.9 | result.10 | result.11 | result.12 |
| ENSG00000000003 | 1.0308681 | 1.3795216 | 0.3735805 | 1.3310151 | 0.1670525 | 0.6535881 |
| ENSG00000000005 | 0.5253541 | 0.7741051 | 0.6105869 | 0.3057523 | 0.3029255 | 0.7479774 |
| ENSG00000000419 | 0.4384900 | 2.0405999 | 0.6863701 | 0.6819732 | 0.5370282 | 0.8324662 |
| ENSG00000000457 | 0.5826354 | 1.5606899 | 1.4607706 | 0.3749742 | 1.4206475 | 0.4606007 |
| ENSG00000000460 | 0.6076585 | 0.6473325 | 1.0544343 | 0.5901518 | 1.1056865 | 0.8615408 |
| ENSG00000000938 | 0.8597862 | 1.6251920 | 1.3009659 | 1.1855873 | 0.9638506 | 1.1603940 |
| | result.13 | result.14 | result.15 | result.16 | result.17 | result.18 |
| ENSG00000000003 | 0.6158714 | 0.8343576 | 0.3284674 | 0.8471763 | 0.5437696 | 0.4184979 |
| ENSG00000000005 | 0.3022207 | 0.4260545 | 0.2233850 | 0.5214336 | 0.6576167 | 0.5383698 |
| ENSG00000000419 | 0.6892615 | 0.6392961 | 0.9746763 | 0.2548610 | 0.6651640 | 0.7629032 |
| ENSG00000000457 | 0.4740113 | 0.6605279 | 0.5264745 | 0.4052821 | 1.0482295 | 0.7366859 |
| ENSG00000000460 | 0.5631291 | 1.0797887 | 1.0272799 | 0.3451112 | 0.9128325 | 0.6688654 |
| ENSG00000000938 | 0.9896480 | 1.1954556 | 1.7281802 | 1.6616518 | 1.1949533 | 1.1638789 |
| | result.19 | result.20 | | | | |
| ENSG00000000003 | 0.6078591 | 1.1165454 | | | | |

```

ENSG000000000005 0.2853109 0.6441431
ENSG000000000419 0.4072167 1.0658283
ENSG000000000457 0.6493479 0.5919240
ENSG000000000460 1.1976580 0.6826181
ENSG000000000938 1.1465015 0.9668223

```

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]

ENSG000000000003 1.889718  ENSG000000000005 0.973799  ENSG000000000419 2.240065  ENSG000000000457 3.593726  ENSG000000000460 1.002215

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

      result.1 result.2 result.3 result.4 result.5 result.6
ENSG000000000003 0.7487155 0.6920500 0.4292202 0.8104770 1.6029020 2.2828823
ENSG000000000005 0.6516005 1.2667268 1.3844351 0.4027415 0.8825720 0.7657517
ENSG000000000419 0.7052349 0.4494130 0.6916490 1.4483298 0.9328028 1.5706070
ENSG000000000457 1.2106505 0.4924473 1.2292696 0.6153973 0.9077375 2.0154766
ENSG000000000460 1.1096846 0.5053252 1.0296305 0.7832972 0.9083019 0.5263266
      result.7 result.8 result.9 result.10
ENSG000000000003 1.3513073 1.8083376 0.4897058 1.7447532
ENSG000000000005 1.1207811 1.6514622 1.3026152 0.6522866
ENSG000000000419 0.5833027 2.7145148 0.9130461 0.9071971
ENSG000000000457 0.7196777 1.9277814 1.8043599 0.4631723
ENSG000000000460 0.8108533 0.8637938 1.4070264 0.7874925

```

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 ENSG000000000003 1.4416043  0.05 0.08979592
2 ENSG000000000005 0.4564578  0.50 0.56050955
3 ENSG000000000419 1.6839390  0.05 0.08979592
4 ENSG000000000457 2.9094026  0.00 0.00000000
5 ENSG000000000460 0.7510661  0.45 0.51764706
6 ENSG000000000938 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
ENSG000000000003 495 235 386 272 255 815 1065 803 839 885 278 270 238 175 292
ENSG000000000005 19 1 0 2 2 12 7 3 1 4 4 4 2 0 1
ENSG000000000419 196 134 165 184 132 344 343 307 342 280 179 156 100 120 126
ENSG000000000457 97 78 141 72 102 219 344 277 337 249 62 48 40 43 52
ENSG000000000460 52 35 48 25 47 105 124 80 156 145 48 36 21 34 19
ENSG000000000938 27 44 57 43 14 71 74 146 148 165 48 59 32 79 20
      C6 C7 C8 C9 C10
ENSG000000000003 432 519 621 475 560
ENSG000000000005 9 3 1 14 46
ENSG000000000419 169 255 171 164 201
ENSG000000000457 170 165 131 183 185
ENSG000000000460 68 90 48 72 38
ENSG000000000938 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,] 0.0027994 0.6604707 0.06936950 0.03195676 0.4347785046 1.81777933
[2,] 1.7642470 1.1059357 2.96164754 0.01053080 4.0885858444 0.75292178
[3,] 0.3847206 0.2142478 0.03657615 1.11469611 0.9860910770 1.33137068
[4,] 0.2662528 0.6321351 0.03212019 0.43094434 0.1862124090 0.01073592
[5,] 0.4056088 1.4513642 0.97693327 0.06475785 0.0007172577 0.02512053
      result.7 result.8 result.9 result.10
[1,] 0.03638154 0.1324338 0.5755783007 1.5529527
[2,] 0.09920720 2.1438641 0.0593705105 0.6371794
[3,] 0.20790348 0.3212303 0.3016275106 0.5006260
[4,] 0.05731346 0.7701329 2.1134570572 1.4200775
[5,] 0.04654679 0.2436032 0.0002331531 0.0250130
```

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] 0.79855593 0.21169955 0.04436084 29.05029655 2.05869585
> DEScore.perm[1:5, 1:10]

      result.1 result.2 result.3 result.4 result.5 result.6
[1,] 0.004119553 0.9719385 0.10208308 0.047027077 0.639813334 2.67501599
[2,] 1.491866673 0.9351914 2.50440177 0.008904959 3.457353219 0.63667895
[3,] 0.738644377 0.4113451 0.07022439 2.140161079 1.893245815 2.55616547
[4,] 0.539441003 1.2807360 0.06507705 0.873113896 0.377275268 0.02175148
[5,] 0.592163113 2.1188996 1.42626056 0.094542350 0.001047151 0.03667438
      result.7 result.8 result.9 result.10
[1,] 0.05353852 0.1948875 0.8470121399 2.2853011
[2,] 0.08389070 1.8128751 0.0502043575 0.5388059
[3,] 0.39916434 0.6167462 0.5791098158 0.9611770
[4,] 0.11611982 1.5603261 4.2819653163 2.8771451
[5,] 0.06795536 0.3556452 0.0003403887 0.0365174
```

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
```

```

ENSG000000000005  0.2503510      0.65      1
ENSG000000000419  0.0231052      0.95      1
ENSG000000000457 14.3384053      0.00      0
ENSG000000000460  1.4101270      0.15      1
ENSG000000000938  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 |
|------------------|------------|--------------|--------------|-----------|-----------|
| ENSG000000000003 | 0.5426504 | 3.956985e-01 | 5.408276e-01 | 0.45 | 1 |
| ENSG000000000005 | 0.2503510 | 3.300042e-01 | 4.943803e-01 | 0.65 | 1 |
| ENSG000000000419 | 0.0231052 | 9.244775e-01 | 9.839468e-01 | 0.95 | 1 |
| ENSG000000000457 | 14.3384053 | 9.960426e-05 | 2.589711e-03 | 0.00 | 0 |
| ENSG000000000460 | 1.4101270 | 1.370959e-01 | 2.970412e-01 | 0.15 | 1 |
| ENSG000000000938 | 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 ranges 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.

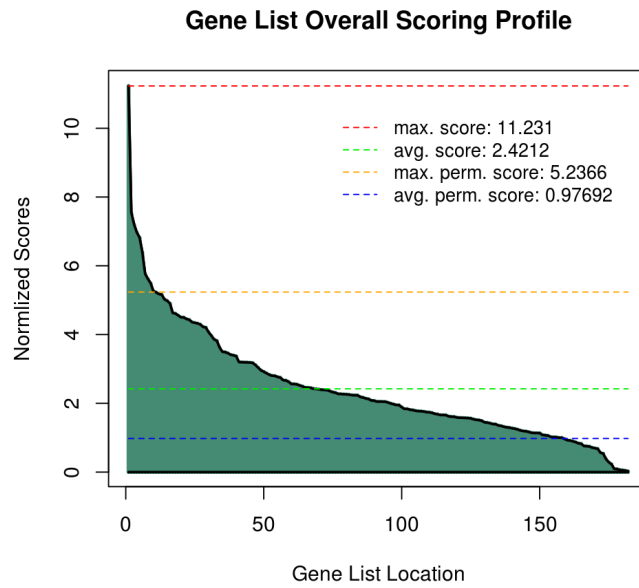


Figure 1: Gene scores resulted from linear combination. Scores are sorted from the largest to the smallest. Red, green, orange, blue dotted horizontal lines represent the maximum score, average score on the real data set, and the maximum score, average score on the permutation data sets.

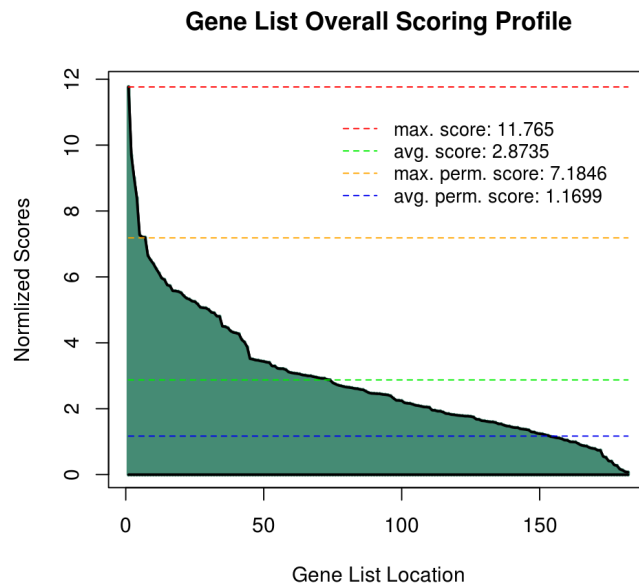


Figure 2: Gene scores resulted from rank-based combination with data-set-specific ranks. Scores are sorted from the largest to the smallest. Red, green, orange, blue dotted horizontal lines represent the maximum score, average score on the real data set, and the maximum score, average score on the permutation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

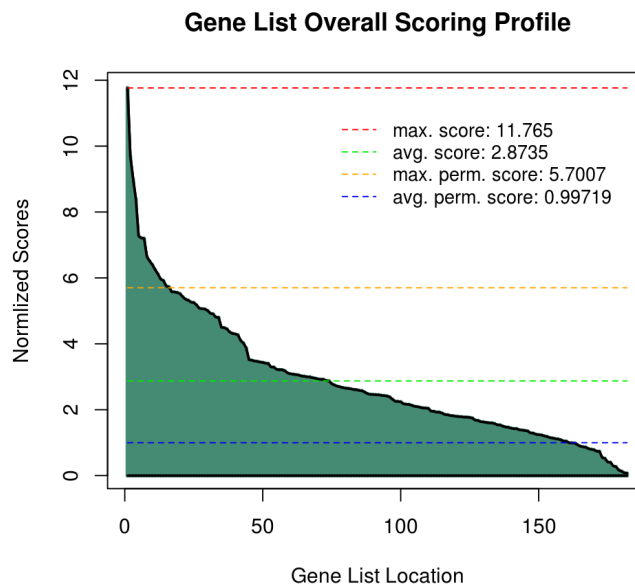


Figure 3: Gene scores resulted from rank-based combination with the same rank got from the real data set. Scores are sorted from the largest to the smallest. Red, green, orange, blue dotted horizontal lines represent the maximum score, average score on the real data set, and the maximum score, average 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 `GSEnrichAnalyze` 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 |
|---|-----------------|--------|-----------|--------|------|---------|------|
| 9 | TBK1.DF_UP | 5 | 1.9938784 | 18 | 0.0 | 0.00000 | 0.05 |
| 8 | HOXA9_DN.V1_UP | 5 | 1.4492422 | 51 | 0.3 | 0.33333 | 0.80 |
| 1 | ERB2_UP.V1_DN | 6 | 1.5125698 | 19 | 0.1 | 0.50000 | 0.65 |
| 5 | PKCA_DN.V1_DN | 5 | 1.3129966 | 43 | 0.2 | 0.50000 | 0.90 |
| 4 | ATF2_S_UP.V1_UP | 5 | 0.3958253 | 152 | 0.9 | 0.90909 | 1.00 |

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.

Global Observed and Null Densities (Area Normalized)

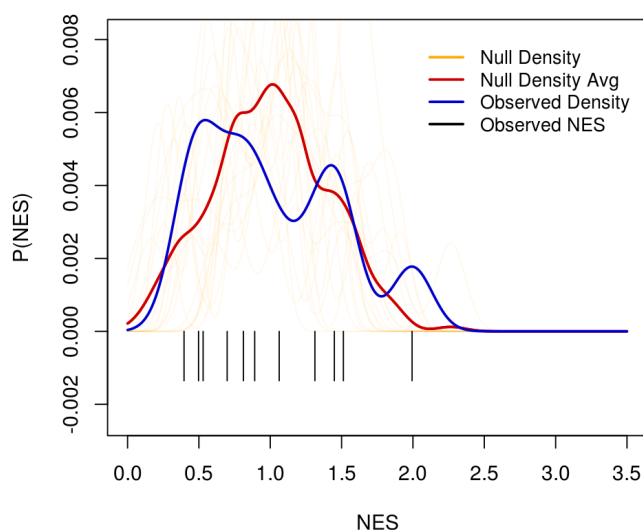


Figure 4: Distribution of normalized enrichment scores (NES) on the observed and permutation (null) data sets. Blue: observed NES density; Orange: each for NES density on one permutation data set; Red: the average density on all permutation data sets; Black: observed NES values.

```
> 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              1.99387835096076
Pos              18
pvalue           0
FDR              0
FWER             0.05
```

Leading set:

```
ENSG00000005194      9.01002344580312
ENSG00000002919      6.11731371980889
ENSG00000001167      5.57073690355341
```

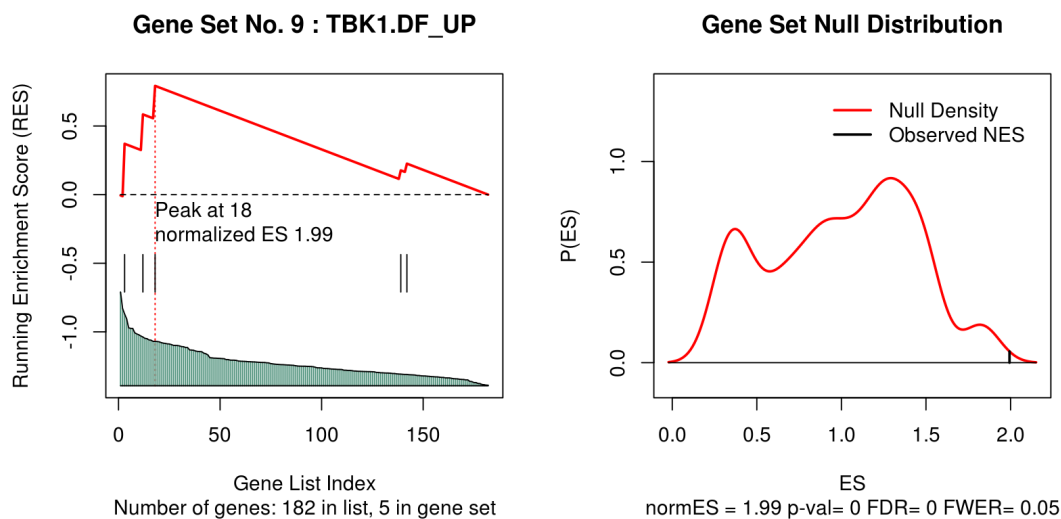


Figure 5: Left: gene locations of a particular gene set according to the gene score rank and running enrichment scores; Right: null NES distribution and the relative position of the observed NES.

Whole gene set:

| | |
|-----------------|------------------|
| ENSG00000005194 | 9.01002344580312 |
| ENSG00000002919 | 6.11731371980889 |
| ENSG00000001167 | 5.57073690355341 |
| ENSG00000006576 | 1.4657092355431 |
| ENSG00000005059 | 1.4229126332451 |

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 An analysis example

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/Rtmpi4a7YZ/Rinst4a7e7b3c6b43/SeqGSEA/extscripts"
```

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

6.2 An exemplified pipeline

Below is shown a real example with the data enclosed in the *SeqGSEA*, 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 exprssion
> 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)
> # visualization 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=""))

```

Fixme: If users would like to run *SeqGSEA* with different weighting parameters or different integration methods, how can they avoid repetitive calculations of DE and DS scores.

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=""))

```

7 Session information

```
> sessionInfo()
```

```
R version 3.0.0 (2013-04-03)
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=C                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_1.0.1 iterators_1.0.6 SeqGSEA_1.0.2 foreach_1.4.0
[5] biomaRt_2.16.0 DESeq_1.12.0 lattice_0.20-15 locfit_1.5-9.1
[9] Biobase_2.20.0 BiocGenerics_0.6.0
```

loaded via a namespace (and not attached):

```
[1] AnnotationDbi_1.22.5 DBI_0.2-6 IRanges_1.18.0
[4] RColorBrewer_1.0-5 RCurl_1.95-4.1 RSQLite_0.11.3
[7] XML_3.96-1.1 annotate_1.38.0 codetools_0.2-8
[10] compiler_3.0.0 genefilter_1.42.0 geneplotter_1.38.0
[13] grid_3.0.0 splines_3.0.0 stats4_3.0.0
[16] survival_2.37-4 tools_3.0.0 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.