Example workflow for regsplice package

Lukas M. Weber, Mark D. Robinson

11 October 2016

Package version: regsplice 0.99.0

Contents

1 Introduction

The `regsplice` package implements statistical methods for the detection of differential exon usage (differential splicing) in RNA sequencing (RNA-seq) and exon microarray data sets.

The `regsplice` methods are based on the use of the lasso (L1-regularization) to improve the power of standard generalized linear models. A key advantage of `regsplice` is that runtimes are fast compared to other leading approaches. We anticipate that similar regularization-based methods may also have applications in other settings.

The detailed statistical methodology and performance comparisons with other methods will be described in an upcoming paper.

1.1 Example workflow

This vignette demonstrates an example workflow for the `regsplice` package, using a small simulated RNA-seq data set.

There are two options for running `regsplice`: you can run a complete workflow in one step using the wrapper function `regsplice()`; or you can run the individual functions for each step in sequence, which provides additional flexibility and insight into the methodology. Both options are demonstrated below.

1.2 Data set

The data set used for the example workflow consists of exon-level read counts for a subset of 100 genes from a simulated human RNA-seq data set, consisting of 6 biological samples, with 3 samples in each of 2 conditions.

The original data set is from the paper:

> Soneson, Matthes et al. (2016), *Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage*, Genome Biology, available here

Original data files from this paper, containing the simulated RNA-seq reads (FASTQ and BAM files), are available from ArrayExpress at accession code E-MTAB-3766.

Exon bin counts were generated with the Python counting scripts provided with the DEXSeq package, using the option to exclude exons from overlapping genes instead of aggregating them into multi-gene complexes (see Soneson et al. 2016, Supplementary Material).

For this example workflow, we have selected a subset consisting of the first 100 genes from this simulated data set. The exon-level read counts and the true differential splicing status labels for these 100 genes are saved in the text files `vignette_counts.txt` and `vignette_truth.txt` in the `extdata/` directory in the `regsplice` package source code.

1.3 Exon microarray data

The `regsplice` methods are designed to work with both RNA-seq read counts and exon microarray intensities.

If you are using exon microarray data, the main steps in the workflow are the same as shown below for RNA-seq data. However, the following adjustments to the workflow are required:

- Instead of RNA-seq read counts, a matrix or data frame of exon microarray intensities is provided to the `counts` input argument. The name of the argument is still `counts`, regardless of the input data type.
- Exon microarray intensities should be log2-transformed externally, before they are provided to `regsplice`. This is usually done during pre-processing of microarray data, and may be done automatically depending on your software.
- Filtering of zero-count and low-count exon bins should be disabled, by setting the arguments `filter_zeros = FALSE` and `filter_low_counts = FALSE`.
- Calculation of normalization factors should be disabled, by setting `normalize = FALSE`.
- Calculation of `limma-voom` transformation and weights should be disabled, by setting `voom = FALSE`.

2 Workflow

2.1 Load data and create condition vector

Load the vignette example data file, which contains simulated RNA-seq read counts for 100 genes across 6 biological samples. From the raw data, extract the table of counts (`counts`), gene IDs (`gene_IDs`), and number of exon bins per gene (`n_exons`).

Then create the `condition` vector, which specifies the experimental conditions or treatment groups for each biological sample.

```
# load data
file_counts <- system.file("extdata/vignette_counts.txt", package = "regsplice")
data <- read.table(file_counts, header = TRUE, sep = "\t", stringsAsFactors = FALSE)

head(data)
##                   exon sample1 sample2 sample3 sample4 sample5 sample6
## 1 ENSG00000000003:001     576     506     526     643     482     826
## 2 ENSG00000000003:002     141     122     126     157     121     191
## 3 ENSG00000000003:003     123     102     106     133      99     156
## 4 ENSG00000000003:004      86      76      77      98      72     112
## 5 ENSG00000000003:005      97      83      87     113      76     126
## 6 ENSG00000000003:006     133     107     116     155      97     170

# extract counts, gene_IDs, and n_exons
counts <- data[, 2:7]
tbl_exons <- table(sapply(strsplit(data$exon, ":"), function(s) s[[1]]))
gene_IDs <- names(tbl_exons)
n_exons <- unname(tbl_exons)

dim(counts)
## [1] 3191    6

length(gene_IDs)
## [1] 100

head(gene_IDs)
## [1] "ENSG00000000003" "ENSG00000000005" "ENSG00000000419" "ENSG00000000457"
## [5] "ENSG00000000460" "ENSG00000000938"

length(n_exons)
## [1] 100

sum(n_exons)
## [1] 3191

# create condition vector
condition <- rep(c("untreated", "treated"), each = 3)

condition
## [1] "untreated" "untreated" "untreated" "treated"   "treated"   "treated"
```

2.2 Run workflow with wrapper function

The `regsplice()` wrapper function runs the analysis pipeline with a single command. The required input format for the wrapper function is a `RegspliceData` object, which is created with the `RegspliceData()` constructor function.

The results of a `regsplice` analysis consist of a set of multiple testing adjusted p-values (Benjamini-Hochberg false discovery rates, FDR) quantifying the statistical evidence for differential exon usage (DEU) for each gene. The adjusted p-values are used to rank the genes in the data set according to their evidence for DEU, and a significance threshold can be specified to generate a list of genes with statistically significant evidence for DEU.

The wrapper function returns gene names, fitted model objects and results, raw p-values, multiple testing adjusted p-values, likelihood ratio (LR) test statistics, and degrees of freedom of the LR tests.

The required inputs for the `RegspliceData()` constructor function are `counts` (matrix or data frame of RNA-seq read counts or exon microarray intensities), `gene_IDs` (vector of gene IDs), `n_exons` (vector of exon lengths, i.e. number of exon bins per gene), and `condition` (vector of experimental conditions for each biological sample).

Alternatively, the inputs can be provided as a `SummarizedExperiment` object, which will be parsed to extract these components. This may be useful when running `regsplice` as part of a pipeline with other packages.

See `?RegspliceData` and `?regsplice` for additional details, including other available inputs and options. The `seed` argument is used to generate reproducible results. Note that the progress bar does not display well in the vignette; it should display on a single line on your screen.

```
library(regsplice)

rs_data <- RegspliceData(counts, gene_IDs, n_exons, condition)

res <- regsplice(rs_data, seed = 123)
## removed 936 exon(s) with zero counts
## removed 1 remaining single-exon gene(s)
## removed 240 low-count exon(s)
## removed 4 remaining single-exon gene(s)
## Fitting regularized (lasso) models...
##
  |
  |                                                                      |   0%
  |
  |================================                                      |  50%
  |
  |======================================================================| 100%
## Fitting null models...
```

2.2.1 Summary table of results

The function `summary_table()` is used to generate a summary table of the results.

The results are displayed as a data frame of the top `n` most highly significant genes, ranked according to either the false discovery rate (FDR) or raw p-values, up to a specified significance threshold (e.g. FDR < 0.05).

The argument `rank_by` chooses whether to rank by FDR or raw p-values.

To display results for all genes up to the significance threshold, set the argument `n = Inf`. To display results for all genes in the data set, set both `n = Inf` and `threshold = 1`.

For more details, see `?summary_table`.

```
summary_table(res)
##          gene_IDs       p_vals        p_adj   LR_stats df_tests
## 1 ENSG00000004766 1.963457e-17 5.890372e-16 137.057172       25
## 2 ENSG00000001461 1.044422e-05 1.566633e-04  25.811578        3
## 3 ENSG00000003436 1.316879e-04 1.316879e-03  17.870151        2
## 4 ENSG00000003249 2.207283e-03 1.655462e-02   9.368648        1
```

2.3 Run workflow using functions for individual steps

Alternatively, the `regsplice` analysis pipeline can be run using the individual functions for each step, which provides additional flexibility and insight into the statistical methodology. The steps are described below.

2.3.1 Create RegspliceData object

The first step is to create a `RegspliceData` object, which contains data in the format required by the functions in the `regsplice` analysis pipeline.

The `RegspliceData` format is based on the `SummarizedExperiment` container from Bioconductor. The main advantage of this format is that subsetting operations automatically keep data and meta-data for rows and columns in sync, which helps avoid errors caused by selecting incorrect row or column indices.

Initially, the `RegspliceData` objects contain raw data along with meta-data for rows (genes and exon bins) and columns (biological samples). During subsequent steps in the `regsplice` analysis pipeline, the data values are modified, and additional data and meta-data are added. Final results are stored in a `RegspliceResults` object.

The required inputs are `counts` (matrix or data frame of RNA-seq read counts or exon microarray intensities), `gene_IDs` (vector of gene IDs), `n_exons` (vector of exon lengths, i.e. number of exon bins per gene), and `condition` (vector of experimental conditions for each biological sample).

Alternatively, the inputs can be provided as a `SummarizedExperiment` object, which will be parsed to extract these components. This may be useful when running `regsplice` as part of a pipeline with other packages.

For more details, see `?RegspliceData`.

```
library(regsplice)

Y <- RegspliceData(counts, gene_IDs, n_exons, condition)
```

2.3.2 Filter zero-count exon bins

Next, use the function `filter_zeros()` to filter exon bins (rows) with zero counts in all biological samples (columns).

Any remaining single-exon genes are also removed (since differential splicing requires multiple exons).

If you are using exon microarray data, this step should be skipped.

For more details, see `?filter_zeros`.

```
Y <- filter_zeros(Y)
## removed 936 exon(s) with zero counts
## removed 1 remaining single-exon gene(s)
```

2.3.3 Filter low-count exons

Filter low-count exon bins with `filter_low_counts()`.

The arguments `filter_min_per_exon` and `filter_min_per_sample` control the amount of filtering. Default values are provided; however, these should be adjusted depending on the total number of samples and the number of samples per condition.

Any remaining single-exon genes are also removed.

If you are using exon microarray data, this step should be skipped.

For more details, see `?filter_low_counts`.

```
Y <- filter_low_counts(Y)
## removed 240 low-count exon(s)
## removed 4 remaining single-exon gene(s)
```

2.3.4 Calculate normalization factors

The function `run_normalization()` calculates normalization factors, which are used to scale library sizes.

By default, `run_normalization()` uses the TMM (trimmed mean of M-values) normalization method (Robinson and Oshlack, 2010), implemented in the `edgeR` package. For more details, see the documentation for `calcNormFactors()` in the `edgeR` package.

This step should be done after filtering. The normalization factors are then used by `limma-voom` in the next step.

If you are using exon microarray data, this step should be skipped.

For more details, see `?run_normalization`.

```
Y <- run_normalization(Y)
```

2.3.5 'voom' transformation and weights

The next step is to use `limma-voom` to transform the counts and calculate exon-level weights. This is done with the `run_voom()` function.

The `limma-voom` methodology transforms counts to log2-counts per million (logCPM), and calculates exon-level weights based on the observed mean-variance relationship. This is required because raw or log-transformed counts do not fulfill the statistical assumptions required for linear modeling (i.e. equal variance). After the `limma-voom` transformation and weights have been calculated, linear modeling methods can be used.

For more details, see the following paper, which introduced `voom`; or the limma User's Guide (section "Differential splicing") available on Bioconductor.

- Law et al. (2014), *voom: precision weights unlock linear model analysis tools for RNA-seq read counts*, Genome Biology, available here

Note that `voom` assumes that exon bins (rows) with zero or low counts have already been removed, so this step should be done after filtering with `filter_zeros()` and `filter_low_counts()`.

If normalization factors are available (from previous step with `run_normalization()`), they will be used by `voom` to calculate normalized library sizes. If they are not available, `voom` will use non-normalized library sizes (columnwise total counts) instead.

If you are using exon microarray data, this step should be skipped.

For more details, see `?run_voom`.

```
Y <- run_voom(Y)

# view column meta-data including normalization factors and normalized library sizes
colData(Y)
## DataFrame with 6 rows and 4 columns
##   sample_names   condition norm_factors lib_sizes
##    <character> <character>    <numeric> <numeric>
## 1      sample1   untreated    0.8720188  348678.4
## 2      sample2   untreated    1.0031023  327179.9
## 3      sample3   untreated    1.0828220  315878.7
## 4      sample4     treated    1.0338759  305822.6
## 5      sample5     treated    0.9833141  359353.1
## 6      sample6     treated    1.0385113  300454.8
```

2.3.6 Initialize RegspliceResults object

The `initialize_results()` function creates a `RegspliceResults` object, which will contain the results of the analysis. This object will be populated in the subsequent steps.

For more details, see `?initialize_results`.

```
res <- initialize_results(Y)
```

2.3.7 Fit models

There are three model fitting functions:

- `fit_reg_multiple()` fits regularized (lasso) models containing an optimal subset of exon:condition interaction terms for each gene. The model fitting procedure penalizes the interaction terms only, so that the main effect terms for exons and samples are always included. This ensures that the null model is nested, allowing likelihood ratio tests to be calculated.
- `fit_null_multiple()` fits the null models, which do not contain any interaction terms.
- `fit_full_multiple()` fits "full" models, which contain all exon:condition interaction terms for each gene.

The fitting functions fit models for all genes in the data set. The functions are parallelized using `BiocParallel` for faster runtime. For `fit_reg_multiple()`, the default number of processor cores is 8, or the maximum available if less than 8. For `fit_null_multiple()` and `fit_full_multiple()`, the default is one core, since these functions are already extremely fast.

Note that in this example, we have used a single core for `fit_reg_multiple()`, in order to simplify testing and compilation of this vignette. We have also used `suppressWarnings()` to hide warning messages related to the small number of observations per gene in this data set.

For more details, see `?fit_reg_multiple`, `?fit_null_multiple`, or `?fit_full_multiple`.

```
# set random seed for reproducibility
seed <- 123

# fit regularized models
res <- suppressWarnings(fit_reg_multiple(res, Y, n_cores = 1, seed = seed))
## Fitting regularized (lasso) models...

# fit null models
res <- fit_null_multiple(res, Y, seed = seed)
```

```
## Fitting null models...

# fit "full" models (not required if 'when_null_selected = "ones"' in next step)
res <- fit_full_multiple(res, Y, seed = seed)
## Fitting full models...
```

2.3.8 Calculate likelihood ratio tests

The function `LR_tests()` calculates likelihood ratio (LR) tests between the fitted models and null models.

If the fitted regularized (lasso) model contains at least one exon:condition interaction term, the LR test compares the lasso model against the nested null model. However, if the lasso model contains zero interaction terms, then the lasso and null models are identical, so the LR test cannot be calculated. The `when_null_selected` argument lets the user choose what to do in these cases: either set p-values equal to 1 (`when_null_selected = "ones"`); or calculate a LR test using the "full" model containing all exon:condition interaction terms (`when_null_selected = "full"`), which reduces power due to the larger number of terms, but allows the evidence for differential exon usage among these genes to be distinguished. You can also return NAs for these genes (`when_null_selected = "NA"`).

The default option is `when_null_selected = "ones"`. This simply calls all these genes non-significant, which in most cases is sufficient since we are more interested in genes with strong evidence for differential exon usage. However, if it is important to rank the low-evidence genes in your data set, use the `when_null_selected = "full"` option. If `when_null_selected = "ones"` or `when_null_selected = "NA"`, the "full" fitted models are not required.

The results object contains gene names, fitted model objects and results, raw p-values, multiple testing adjusted p-values (Benjamini-Hochberg false discovery rates, FDR), likelihood ratio (LR) test statistics, and degrees of freedom of the LR tests.

For more details, see `?LR_tests`.

```
res <- LR_tests(res)
```

2.3.9 Summary table of results

The function `summary_table()` is used to generate a summary table of the results.

The results are displayed as a data frame of the top `n` most highly significant genes, ranked according to either the false discovery rate (FDR) or raw p-values, up to a specified significance threshold (e.g. FDR < 0.05).

The argument `rank_by` chooses whether to rank by FDR or raw p-values.

To display results for all genes up to the significance threshold, set the argument `n = Inf`. To display results for all genes in the data set, set both `n = Inf` and `threshold = 1`.

For more details, see `?summary_table`.

```
summary_table(res)
##          gene_IDs       p_vals        p_adj  LR_stats df_tests
## 1 ENSG00000004766 5.573266e-22 1.839178e-20 130.22077       12
## 2 ENSG00000001461 1.311623e-05 2.164178e-04  25.33879        3
## 3 ENSG00000003436 2.418061e-03 2.659867e-02  18.46368        5
## 4 ENSG00000003249 5.907319e-03 4.873538e-02  10.26313        2
```

3 Analyze results

For the simulated data set in this vignette, the true differential splicing status of each gene is known. In this section, we show how to analyze the results and calculate a contingency table showing the number of true positives, true negatives, false positives, and false negatives.

3.1 Summary of all significant genes

As shown in the workflow above, we can use the `summary_table()` function with argument `n = Inf` to display a list of all genes with significant evidence for differential exon usage (DEU).

```
summary_table(res, n = Inf)
##          gene_IDs       p_vals        p_adj  LR_stats df_tests
## 1 ENSG00000004766 5.573266e-22 1.839178e-20 130.22077       12
## 2 ENSG00000001461 1.311623e-05 2.164178e-04  25.33879        3
## 3 ENSG00000003436 2.418061e-03 2.659867e-02  18.46368        5
## 4 ENSG00000003249 5.907319e-03 4.873538e-02  10.26313        2
```

The total number of genes with significant evidence for DEU at a given threshold can also be calculated.

Note that we are using the multiple testing adjusted p-values (Benjamini-Hochberg false discovery rates, FDRs) for this calculation. A standard threshold of FDR < 0.05 implies that 5% of genes in the list are expected to be false discoveries.

```
sum(p_adj(res) < 0.05)
## [1] 4

table(p_adj(res) < 0.05)
##
## FALSE  TRUE
##    77     4
```

3.2 Contingency table

As mentioned above, the true differential splicing (DS) status is known for each gene, since this is a simulated data set. Therefore, we can calculate contingency tables comparing the true and predicted DS status for each gene at a given significance threshold. Increasing the significance threshold returns more genes, at the expense of a larger number of false positives.

```
# load true DS status labels
file_truth <- system.file("extdata/vignette_truth.txt", package = "regsplice")
data_truth <- read.table(file_truth, header = TRUE, sep = "\t", stringsAsFactors = FALSE)

str(data_truth)
## 'data.frame':    100 obs. of  2 variables:
##  $ gene     : chr  "ENSG00000000003" "ENSG00000000005" "ENSG00000000419" "ENSG00000000457" ...
##  $ ds_status: int  0 0 0 0 0 0 0 0 1 0 ...

# remove genes that were filtered during regsplice analysis
data_truth <- data_truth[data_truth$gene %in% gene_IDs(res), ]

dim(data_truth)
## [1] 81  2
```

```
length(gene_IDs(res))
## [1] 81

# number of true DS genes in simulated data set
sum(data_truth$ds_status == 1)
## [1] 6

table(data_truth$ds_status)
##
##  0  1
## 75  6

# contingency table comparing true and predicted DS status for each gene
# (significance threshold: FDR < 0.05)
table(true = data_truth$ds_status, predicted = p_adj(res) < 0.05)
##     predicted
## true FALSE TRUE
##    0    72    3
##    1     5    1

# increasing the threshold detects more genes, at the expense of more false positives
table(true = data_truth$ds_status, predicted = p_adj(res) < 0.99)
##     predicted
## true FALSE TRUE
##    0    59   16
##    1     2    4
```

4 Additional information

4.1 Additional user options

Additional user options not discussed in the workflow above include:

- `alpha`: Elastic net parameter for `glmnet` model fitting functions. The value of `alpha` must be between 0 (ridge regression) and 1 (lasso). The default value is 1, which fits a lasso model. See `glmnet` package documentation for more details.
- `lambda_choice`: Parameter to select which optimal `lambda` value to choose from the `cv.glmnet` cross validation fit. Available choices are `"lambda.min"` (model with minimum cross-validated error) and `"lambda.1se"` (most regularized model with cross-validated error within one standard error of minimum). The default value is `"lambda.min"`. See `glmnet` package documentation for more details.

For further details, including a complete list and description of all available user options, refer to the documentation for the `regsplice()` wrapper function, which can be accessed with `?regsplice` or `help(regsplice)`.

4.2 Design matrices

The function `create_design_matrix()` creates the model design matrix for each gene. This function is called automatically by the model fitting functions, so does not need to be used directly. In this section, we demonstrate how it works for a single gene, and show an example design matrix, in order to provide further insight into the statistical methodology.

The design matrix includes main effect terms for each exon and each sample, and interaction terms between the exons and conditions.

Note that the design matrix does not include main effect terms for the conditions, since these are absorbed into the main effect terms for the samples. In addition, the design matrix does not include an intercept column, since it is simpler to let the model fitting functions add an intercept term later.

For more details, see `?create_design_matrix`.

```
# gene with 3 exons
# 4 biological samples; 2 samples in each of 2 conditions
design_example <- create_design_matrix(condition = rep(c(0, 1), each = 2), n_exons = 3)

design_example
##    Exon2 Exon3 Samp2 Samp3 Samp4 Exon2:Cond1 Exon3:Cond1
## 1      0     0     0     0     0           0           0
## 2      1     0     0     0     0           0           0
## 3      0     1     0     0     0           0           0
## 4      0     0     1     0     0           0           0
## 5      1     0     1     0     0           0           0
## 6      0     1     1     0     0           0           0
## 7      0     0     0     1     0           0           0
## 8      1     0     0     1     0           1           0
## 9      0     1     0     1     0           0           1
## 10     0     0     0     0     1           0           0
## 11     1     0     0     0     1           1           0
## 12     0     1     0     0     1           0           1
```