A Quick-Start Guide for rSeqDiff

Yang Shi (email: shyboy@umich.edu)
and
Hui Jiang (email: jianghui@umich.edu)

09/05/2013

Introduction

rSeqDiff is an R package that can detect differential gene and isoform expressions from RNA-seq data of multiple biological conditions. The package is in the beta version and the authors are happy to answer any question from the users and will appreciate any feedback and suggestions. The package source file ("rSeqDiff.beta.0.1.tar.gz"), the R script for the package ("rSeqDiff.beta.0.1.R"), the testing dataset containing two files "ASD.sampling_rates" and "control.sampling_rates" and the three csv format output files generated from the testing dataset can be downloaded from the authors' website [1].

This is a quick-start guide for rSeqDiff. A more detailed user manual and R documentations of this package will come soon. Before start to use the package, users are encouraged to read the two references [2] and [3] to get an idea of what the package can do. rSeqDiff considers three cases (models) for each gene: 1) no differential expression, 2) differential expression without differential splicing and 3) differential splicing. For each gene, rSeqDiff will test which of the three models the gene should belong to using a hierarchical likelihood ratio test, and give an estimation of the abundances of the gene and its isoforms (a manuscript presenting the approach is submitted).

The usage of rSeqDiff is dependent on rSeq (another software tool for RNA-seq data analysis developed by Hui Jiang) [4]. Figure 1 illustrates the pipeline of the analysis by rSeq/rSeqDiff, which includes the following three steps (starting from raw RNA-seq read data in the fastq or fasta format): (1) Mapping the reads to the transcript sequences using read aligner software such as BWA, Bowtie/Bowtie 2 or SeqMap, etc. After this step, the sequence alignment files in either SAM or Eland-multiple format should be generated; (2) Using rSeq to process sequence alignment files to generate the ".sampling_rates" files that are required as input files for rSeqDiff. (Note that rSeq can also generate BED format file for visualizing the reads in USCS genome browser or CisGenome browser. See [4] for details) (3) Processing the ".sampling_rates" files using rSeqDiff and obtaining the list of genes belonging to each model and the estimation of gene and isoform abundances.

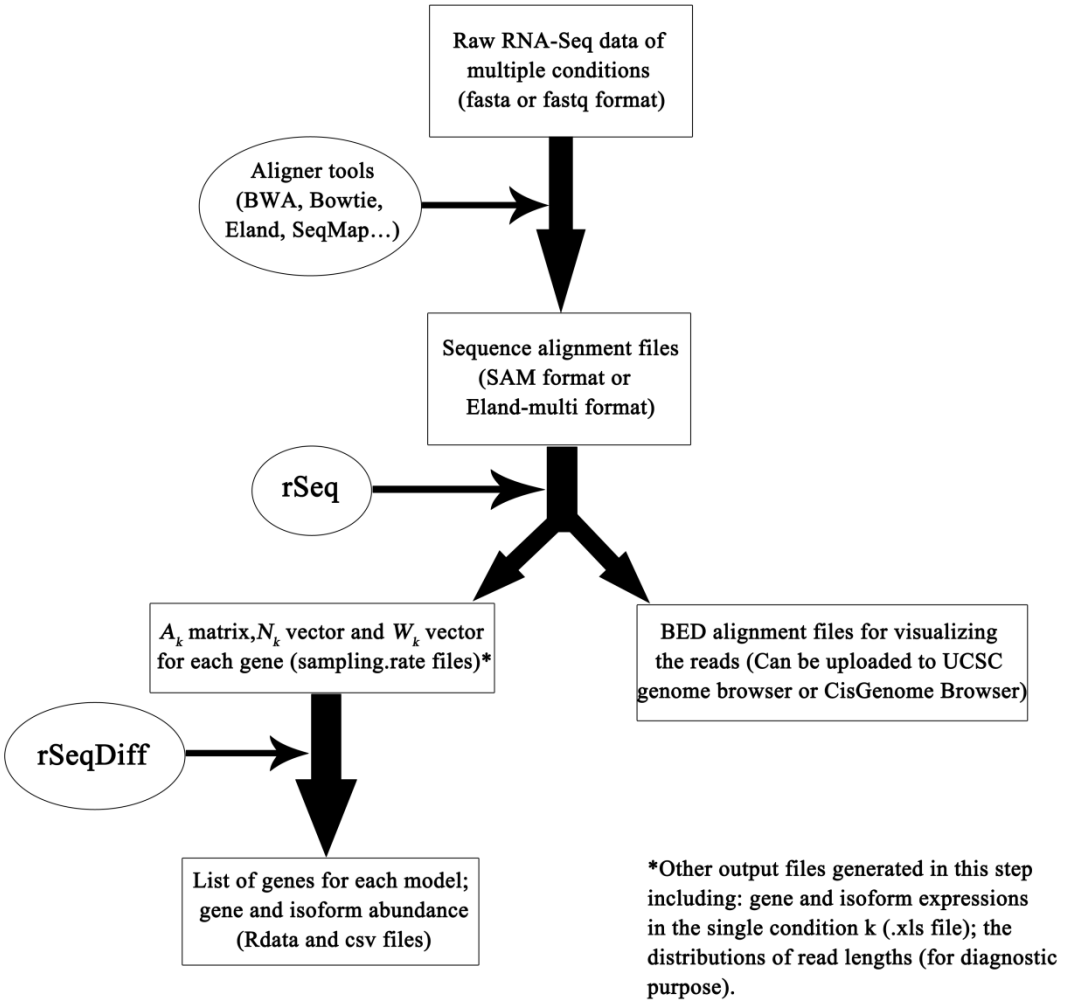

Figure 1. The pipeline of the analysis by rSeq/rSeqDiff.

Detailed description of Step 1 and 2 is given in the documentations of rSeq [4]. Below we demonstrate Step 3 using a testing dataset, which contains two files: "ASD.sampling_rates" and "control.sampling_rates" generated by rSeq. These files contain the sampling rate matrices for 200 genes, which is a subset of totally 25999 genes in the analysis that is presented in our submitted manuscript. Basically, we are testing differentially spliced genes between autism spectrum diseased (ASD) and normal brain samples. The original data are from [5].

Methods

1. *Download and install the package.*

Currently, rSeqDiff can be used in UNIX or Windows machine (R≥2.14.0 should be installed in the machine. For UNIX machine, the user also need to install Rtools).

To use it in a UNIX machine, download the package source file ("rSeqDiff.beta.0.1.tar.gz") from the authors' website [1], then cd to the directory containing this file and use the following Rtools command to install the package:

```
$R CMD INSTALL -l   /path to R library directory/   rSeqDiff.beta.0.1.tar.gz
```

Please ignore the warning message, since the documentations of the package are not completed yet at this moment.

Then, start R and load the package:

```
>library(rSeqDiff)
```

If the package cannot be installed successfully in the UNIX machine (in that case, the users are very welcomed and encouraged to report the errors to the authors), or the user want to use the package in a Windows machine, the user can download the R script ("rSeqDiff_beta.R") from the authors' website [1] and run the script in R, and then continue to the next step.

2. Run the analysis.

First read the two ".sampling_rates" files into list objects using the function "read_file":

```
> ASD=read_file('path to ASD.sampling_rates ')
> control=read_file(' path to control.sampling_rates ')
```

Then make a collective list of all the above lists, using the following code:

```
> list_2_files=list(ASD, control)
```

At this step, the user can give the names of the two biological conditions (ASD and control) that are under comparison using the following code, so that the condition names will appear in the final result. If the user doesn't give the names of the conditions, rSeqDiff will automatically name the conditions as: condition_1, condition_2…

```
>ASDvsControl=c('ASD', 'control')
```

Then the user can run the analysis by calling the "isoform_estimation" function, the main function of the package:

```
>expression=isoform_estimation(list_2_files, ASDvsControl)
```

The input parameters of this function is:

```
isoform_estimation(list_K_files, condition_names, CI=FALSE, alpha=0.05, read_lim=5)
```

where the input parameters are:

-list_K_files: this is a collective list of the ".sampling.rates" files from all conditions. In this example, it is "list_2_files".

-condition_names: a vector of the names of the conditions that you are analyzing. In this example, it is "ASDvsControl".

-CI: if TRUE, the program will calculate the 95% confidence intervals for the gene and isoform abundances, but it will take longer time; if FALSE (by default), the program will not calculate the 95% confidence interval for isoform abundance.

-alpha: the significance level of the hierarchical likelihood ratio test. By default, it is 0.05. This parameter can be specified by the user between 0 to 1

-read_lim: the cut-off value to filter out genes with very low read counts. The default value is 5, so genes with less than 5 reads mapped in all the conditions will be filtered out in the analysis. This parameter can be specified by the user between 0 to 50.

The running time should be in several minutes, and then the following messages should appear:

```
[1] "Number of genes that have been successfully processed:"
[1] 200
[1] "Number of genes with low reads mapped to:"
[1] 60
[1] "Number of genes that belongs model 0:"
[1] 44
[1] "Number of genes that belongs model 1:"
[1] 70
[1] "Number of genes that belongs model 2:"
```

[1] 26

3. Examine the result

So expression is a list contains all the results for the 200 genes. To check the result for individual gene, say the "NRCAM" gene, use the following code (the meaning of each part of the list is given below):

```
> expression[['NRCAM']]

$gene_name ##the name of the gene
[1] "NRCAM"

$isoform_names ##the names of the isoforms of the gene
[1] "NM_001037132" "NM_005010" "NM_001193582" "NM_001193583"
"NM_001193584"

$condition_names ##the names of the conditions under comparison
[1] "ASD" "control"

$MLE_under_3_models ##this part contains the MLE for all three models
$MLE_under_3_models$Model0_MLE ##MLE under model 0 (no differential
expression)
$MLE_under_3_models$Model0_MLE$theta_MLE ##The MLE of isoform
abundance under model 0
NM_001037132 NM_005010 NM_001193582 NM_001193583
NM_001193584
  30.5719758 12.6998976 0.5617613 16.9546911 3.4083055

$MLE_under_3_models$Model0_MLE$loglikelihood_MLE ##the maximum
loglikelihood under model 0
          [,1]
[1,] 28091.02

$MLE_under_3_models$Model0_MLE$gene_MLE ##The MLE of gene abundance
under model 0
[1] 64.19663

$MLE_under_3_models$Model1_MLE ##MLE under model 1 (differential expression
without differential splicing)
```

$MLE_under_3_models$Model1_MLE$basic_theta_vector_MLE ##The MLE of basic isoform abundance vector under model 1

```
                       [,1]
NM_001037132 58.363558
NM_005010       24.232936
NM_001193582    1.072823
NM_001193583 32.361085
NM_001193584    6.528064
```

$MLE_under_3_models$Model1_MLE$tau_MLE ##The MLE of isoform ratio vector under model 1

```
        ASD     control
0.3893905 0.6106095
```

$MLE_under_3_models$Model1_MLE$theta_matrix_MLE ##The MLE of isoform abundance under model 1

```
          NM_001037132      NM_005010      NM_001193582      NM_001193583
NM_001193584
ASD              22.72622    9.436076       0.4177469        12.60110        2.541966
control       35.63734 14.796860       0.6550756        19.75998        3.986097
```

$MLE_under_3_models$Model1_MLE$loglikelihood_MLE ##the maximum loglikelihood under model 1

```
[1] 29294.29
```

$MLE_under_3_models$Model1_MLE$gene_MLE ##The MLE of gene abundance under model 1

```
       ASD    control
47.72311 74.83536
```

$MLE_under_3_models$Model2_MLE ##MLE under model 2 (differential splicing)

$MLE_under_3_models$Model2_MLE$theta_matrix_MLE ##The MLE of isoform abundance under model 2

NM_001037132 NM_005010 NM_001193582 NM_001193583 NM_001193584
ASD 16.47850 19.27949 0.5006271 8.492616 2.972129
control 37.97617 10.19038 0.6252672 22.254635 3.784595

$MLE_under_3_models$Model2_MLE$loglikelihood_MLE ##the maximum loglikelihood under model 2

[1] 29330.45

$MLE_under_3_models$Model2_MLE$gene_MLE ##The MLE of gene abundance under model 1

ASD control
47.72336 74.83105

$OPT_model_estimation ##this part contains the MLE under the best model selected by the hierarchical likelihood ratio test

$OPT_model_estimation$opt_model ##this is the best model, which is model 2 for this gene

[1] "model_2"

$OPT_model_estimation$p_value ##the p values from the hierarchical likelihood ratio test

model1_vs_model0 model2_vs_model0 model2_vs_model1
0.000000e+00 0.000000e+00 7.327472e-15

$OPT_model_estimation$MLE
$OPT_model_estimation$MLE$theta_matrix_MLE ##The MLE of isoform abundance under the best model (model 2)

NM_001037132 NM_005010 NM_001193582 NM_001193583 NM_001193584
ASD 16.47850 19.27949 0.5006271 8.492616 2.972129
control 37.97617 10.19038 0.6252672 22.254635 3.784595

$OPT_model_estimation$MLE$loglikelihood_MLE ##the maximum loglikelihood

under the best model (model 2)

[1] 29330.45

$OPT_model_estimation$MLE$gene_MLE ##The MLE of gene abundance under the best model (model 2)

```
      ASD   control
47.72336 74.83105
```

$OPT_model_estimation$p1 ##This is the p value from the likelihood ratio test between model 1 and 0, which can be used to examine the differential expression of the gene

```
model1_vs_model0
               0
```

$OPT_model_estimation$log2.fold_change ##This is the log2 fold changes of the estimated gene abundance between ASD and control

```
       ASD
-0.6490321
```

$OPT_model_estimation$p2 ##This is the p value from the likelihood ratio test between model 2 and 0, which can be used to examine the differential splicing of the gene. Note: for genes with a single isoform, this value is NA.

```
model2_vs_model0
               0
```

$OPT_model_estimation$T_value ##This is the T statistic (see below for definition)

[1] 0.2816434

We define the following *T* statistic to rank the levels of differential splicing of genes:

$$T = \frac{1}{2}\left\| \frac{\hat{\theta}_1}{\| \hat{\theta}_1 \|_1} - \frac{\hat{\theta}_2}{\| \hat{\theta}_2 \|_1} \right\|_1$$

where $\| \cdot \|_1$ denotes the vector L_1 norm, and $\hat{\theta}_1$ and $\hat{\theta}_2$ are the estimated isoform abundance vectors of the two conditions. Large *T* values indicate high level of differential

splicing of isoforms [6].

To see the genes of each model, we can write the results into .csv files using the following code. The .csv files can be opened by Microsoft Excel software.

###write the list of genes of each model to .csv files

##model 0
```
> gene_list_model0=output_each_model(expression, model='model0')
[1] "44 genes belong to model 0."
>write.csv(gene_list_model0, file = '/ directory / model0_list.csv',row.names=F)
```

##model 1
```
> gene_list_model1=output_each_model(expression, model='model1')
[1] "70 genes belong to model 1."
> write.csv(gene_list_model1, file = '/ directory / model1_list.csv',row.names=F)
```

Note: Genes belonging to model 1 are ranked by the log2 (fold change) in the .csv file.

##model 2
#In this example, we are testing two conditions, and the genes that belong to model 2 can be ranked by T statistic (see below for details). The following code will write genes of model 2 to csv files and rank them by the T statstics

```
> gene_list_model2=output_each_model(expression, model='model2')
[1] "26 genes belong to model 2."
> write.csv(gene_list_model2, file='/directory/model2_list.csv', row.names=F)
```

Note: Genes belonging to model 2 are ranked by the T values in the .csv file.

After running these codes, three .csv files are generated (these files can be downloaded from the authors' website [1] for the users to check). The model0_list.csv, model1_list.csv and model2_list.csv files are the lists of genes that belong to model 0, model 1 and model 2, respectively.

The headers of the model0_list.csv file are:

- gene_name: the name of the gene.

- isoform_name: the names of the isoforms of the gene.

- model: which model the gene belongs to.

- log_likelihood: the log-likelihood under the MLE of the isoform abundance.

- p_value: the p_value(s) of the hierarchical likelihood ratio test.

- theta_MLE: the MLE of the isoform abundance (the order of the values correspond to the order of isoform_name).

- gene_MLE: the MLE of the gene abundance.

-p1: This is the p value from the likelihood ratio test between model 1 and 0, which can be used to examine the differential expression of the gene.

- log2.fold_change: This is the log2 fold changes of the estimated gene abundance between ASD and control.

-p2: This is the p value from the likelihood ratio test between model 2 and 0, which can be used to examine the differential splicing of the gene

-T_value: This is the T statistic described above. Large T values indicate high level of differential splicing of isoforms.

The headers of the model1_list.csv files are:

- gene_name, isoform_name, model, log_likelihood, p_value, p1, log2.fold_change, p2, T_value have exact the same meaning as in the model0_list.csv file.

- basic_theta_vector_MLE: the MLE of basic isoform abundance vector under model 1

- tau_MLE_ASD: the MLE of isoform abundance ratio in ASD

- tau_MLE_control: the MLE of isoform abundance ratio in control

- theta_MLE_ASD: the MLE of isoform abundance in ASD

- theta_MLE_ASD: the MLE of isoform abundance in control

- gene_MLE_ASD: the MLE of gene abundance in ASD

- gene_MLE_control: the MLE of gene abundance in control

The headers of the model2_list.csv files are:

- gene_name, isoform_name, model, log_likelihood, p_value, theta_MLE_ASD, theta_MLE_ASD, gene_MLE_ASD, gene_MLE_control, p_value, p1, log2.fold_change, p2, T_value have exactly the same as in the model1_list.csv file.

To check the overall pattern of gene differential expression, we can draw the scatter plot of the p values from the likelihood ratio test between model 1 and 0 v.s. the log2 fold changes of the estimated gene abundance ("volcano plot"), using the following function volcano_DF:

```
> volcano_DF(expression, xlim=c(-3,3))
```

The function is defined as following:

```
volcano_DF=function(expression, pch=20, ...)
```

Where the 1st argument of this function is expression, the list that contains all the results for the 200 genes, and pch by default is set to be 20, and other arguments … is the same as the function plot.

The plot looks like following:

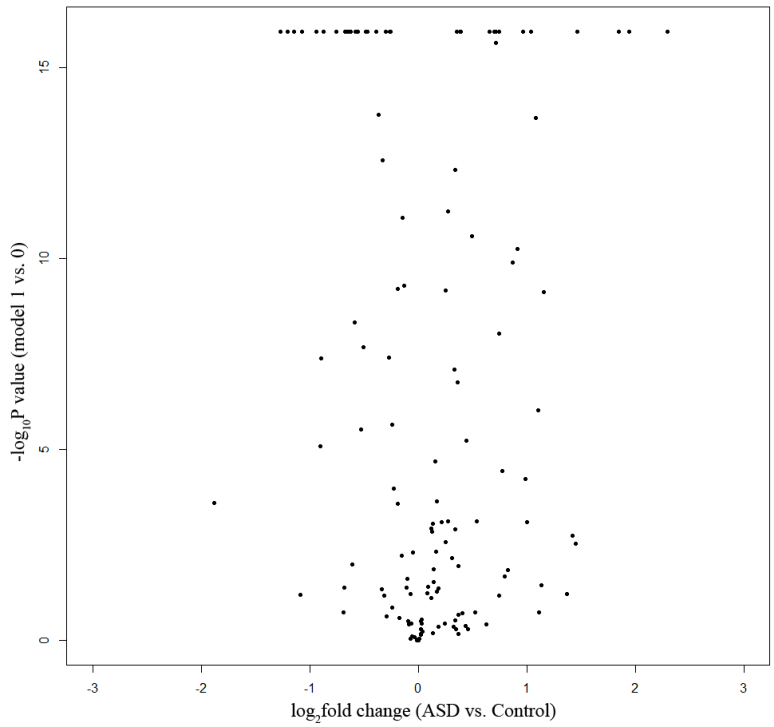

Figure 2. Plot of the -log 10 based p values from the likelihood ratio test between model 1 and 0 v.s. the log2 fold changes of the estimated gene abundance, which can be used for visualizing differential expression of each gene.

To check the overall pattern of gene differential splicing, we can draw the scatter plot of the p values from the likelihood ratio test between model 2 and 0 v.s. the T values, using the following function volcano_DS:

```
> volcano_DS(expression)
```

The function is defined as following:

```
volcano_DS=function(expression, pch=20, …)
```

Where the 1st argument of this function is expression, the list that contains all the results for the 200 genes, and pch by default is set to be 20, and other arguments … is the same as the function plot.

The plot looks like following:

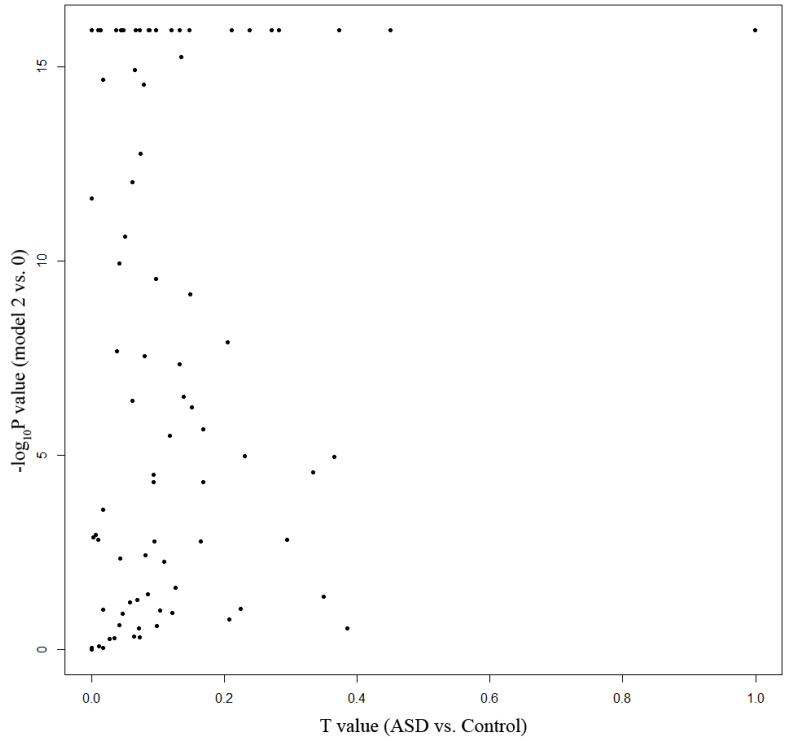

Figure 3. Plot of the -log 10 based p values from the likelihood ratio test between model 2 and 0 v.s the T values, which can be used for visualizing differential splicing of each gene.

Figure 2 and 3 show only the 200 genes in the testing dataset. Please refer to Figure S4 of [6] for the plots of all genes in the dataset.

References

1. rSeqDiff package and its documentations. Available at: http://www-personal.umich.edu/~jianghui/rseqdiff/
2. Jiang H, Wong WH (2009) Statistical inferences for isoform expression in RNA-Seq. Bioinformatics 25: 1026-1032.
3. Salzman J, Jiang H, Wong WH (2011) Statistical modeling of RNA-Seq data. Statistical Science 26: 62-83.
4. rSeq package and its documentations. Available at: http://www-personal.umich.edu/~jianghui/rseq/index.html
5. Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y et al. (2011) Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature 474: 380-384.
6. Shi Y, Jiang H (2013) rSeqDiff: Detecting differential isoform expression from RNA-Seq data using hierarchical likelihood ratio test (submitted).