

IVAS : Identification of genetic Variants affecting Alternative Splicing

Seonggyun Han and Sangsoo Kim

April 24, 2017

Contents

1	Introduction	1
2	The input data set	2
2.1	The genotype data	2
2.2	The expression data	2
2.3	The SNP marker position data	2
2.4	The transcripts model data	2
3	The example dataset : data from Geuvadis RNA sequencing project of 1000 Genome samples	3
4	Loading data	3
5	The ASdb object	3
5.1	Searching alternatively spliced exons based on a reference transcript model.	3
5.2	Estimating expression ratio of AS exons with a data set including FPKM values of transcripts	4
5.3	Finding SQTLs	6
6	Identification of SQTLs using multiple cores	6
7	Visualizing the result	6
8	Session Information	7

1 Introduction

Alternative splicing controls relative expression ratios of mature mRNA isoforms from a single gene. Mapping studies of Splicing Quantitative Trait Loci (SQTL), a genetic variant affecting the alternative splicing, are important steps to understand gene regulations and protein activity [1]. We present an effective and user-friendly computational tool to detect SQTLs using transcript expression data from RNA-seq and genotype data, both measured on the same sample. As RNA sequencing (RNA-seq) provides insight into relatively precise measurements of expression level of transcript isoforms from a gene, it is a useful tool to analyze complicated biological phenomenon of RNA transcripts including the alternative splicing [2]. The mapping analysis uses two statistical models : Linear regression model [3] and/or Generalized linear mixed model [5].

2 The input data set

The next subsection introduces the input data. To run this tool, two experimental data sets (an expression data frame from RNA-seq and a genotype data frame) are required. Moreover, we also need a data frame for positions of SNP markers and GTF file for transcript models. As any other genome-wide analyses, it is recommended to use as many samples as possible, usually of population scale, in order to guarantee a statistically significant result.

2.1 The genotype data

The genotype data should be prepared as a simple matrix data. Each column represents an individual and its name should match that of the expression matrix described below (2.2)

	ind1	ind2	ind3	ind4
SNP1	AA	AA	AT	TT
SNP2	CG	CC	GG	CG
SNP3	TT	TT	AT	TT

2.2 The expression data

The expression matrix must comprise expression values of transcripts from RNA-seq. We may obtain them by using alignment tools such as cufflinks. Each column represents an individual and its name should match that of the genotype matrix described above (2.1)

	ind1	ind2	ind3	ind4
transcript1	10.5	15.4	6.7	12.4
transcript2	6.4	7.2	4.5	9.2
transcript3	15.4	14.5	13.2	17.8

2.3 The SNP marker position data

To search SNPs affecting alternative splicing, a data frame comprising genomic location of each SNP is required. It consists of following columns: SNP (SNP marker name), CHR(chromosome number), and locus(SNP position).

SNP	CHR	locus
SNP1	1	4964005
SNP2	1	23513047

2.4 The transcripts model data

We need a reference GTF (General Feature Format) file including information about gene structures such as the positions of exons, introns, and transcripts of genes. The GTF file must be TxDb object from the *GenomicFeatures* package [4].

3 The example dataset : data from Geuvadis RNA sequencing project of 1000 Genome samples

This example uses filtered data from an origin data generated by Geuvadis RNA sequencing project, available at <http://www.geuvadis.org/web/geuvadis/RNAseq-project> [6]. The example expression data includes transcripts of 11 randomly selected genes. The genotype data comprises SNPs in those genes.

4 Loading data

For this analysis, you need to load the *IVAS* package, SNP data, expression data, SNP position data, and TxDb object from GTF.

Loading *IVAS* package :

```
> library(IVAS)
```

Loading expression data :

```
> data(sampleexp)
```

Loading SNP data :

```
> data(samplesnp)
```

Loading SNP position data :

```
> data(samplesnplocus)
```

Loading TxDb object :

```
> sampleDB <- system.file("extdata", "sampleDB", package="IVAS")
> sample.Txdb <- loadDb(sampleDB)
```

If you want to create the TxDb object from a GTF file, you need to use the `makeTxDbFromGFF` function in the *GenomicFeatures* package.

5 The ASdb object

The ASdb object is a `s4` type class object, and the object is used by the *IVAS* package to store the results from functions in this *IVAS* package. The functions of *IVAS* will save their results by adding a slot. Each slot contains a list object consisting of three elements named as "ES", "ASS", and "IR" for each alternatively splicing pattern type (i.e. ES, ASS, and IR means exon skipping, alternative splice site, and intron retention, respectively).

5.1 Searching alternatively spliced exons based on a reference transcript model.

The `Splicingfinder` function tabulates patterns of alternatively spliced exons. This needs the TxDb object from `makeTxDbFromGFF` by reading a reference GTF file for reference transcript models. The `Splicingfinder` function categorizes alternatively spliced exons into four types of AS patterns (i.e. exon skipping, alternative 3-prime splice site, alternative 5-prime splice site, and intron retention). The result will be saved in the "SplicingModel" slot of ASdb.

To use this function :

```
> ASdb <- Splicingfinder(GTFdb=sample.Txdb,calGene=NULL,Ncor=1,out.dir=NULL)
```

```
[1] "-----Processing : chr 2 -----"
[1] "-----Processing : chr 3 -----"
[1] "-----Processing : chr 6 -----"
[1] "-----Processing : chr 8 -----"
[1] "-----Processing : chr 9 -----"
[1] "-----Processing : chr 11 -----"
[1] "-----Processing : chr 17 -----"
[1] "-----Processing : chr 19 -----"
```

```
> ASdb
```

```
Splicing Models : ES = 182 Rows & ASS = 11 Rows & IR = 2 Rows
```

```
#ASdb object with SplicingModel
```

```
> head(slot(ASdb,"SplicingModel")$"ASS")
```

	Index	EnsID	Nchr	Strand	ShortEX	LongEX
1	"ASS1"	"ENSG00000186001"	"3"	"+"	"197562545-197562609"	"197562545-197562693"
2	"ASS2"	"ENSG00000183826"	"6"	"-"	"38565686-38565833"	"38565686-38565897"
3	"ASS3"	"ENSG00000183826"	"6"	"-"	"38565686-38565833"	"38565686-38565897"
4	"ASS4"	"ENSG00000172728"	"8"	"-"	"33319006-33319245"	"33318890-33319243"
5	"ASS5"	"ENSG00000172728"	"8"	"-"	"33318930-33319243"	"33318890-33319243"
6	"ASS6"	"ENSG00000166263"	"17"	"+"	"53076993-53077203"	"53076987-53077203"
	ShortNeighborEX		LongNeighborEX		Short_des	Long_des
1	"197566192-197566268"		"197566192-197566268"		"197562545-197562609"	"197562545-197562693"
2	"38607576-38607924"		"38607576-38607700"		"38565686-38565833"	"38565686-38565897"
3	"38607576-38607924"		"38580610-38580809"		"38565686-38565833"	"38565686-38565897"
4	"33310734-33311028"		"33310734-33311028"		"33319006-33319245"	"33318890-33319243"
5	"33310734-33311028"		"33310734-33311028"		"33318930-33319243"	"33318890-33319243"
6	"53076706-53076812"		"53076706-53076812"		"53076993-53077203"	"53076987-53077203"
	ShortNeighbor_des		LongNeighbor_des		Types	
1	"197566192-197566268"		"197566192-197566268"		"A5SS"	
2	"38607576-38607924"		"38580610-38580809,38607576-38607700"		"A5SS"	
3	"38607576-38607924"		"38580610-38580809,38607576-38607700"		"A5SS"	
4	"33310734-33311028"		"33310734-33311028"		"A3SS"	
5	"33310734-33311028"		"33310734-33311028"		"A3SS"	
6	"53076706-53076812"		"53076706-53076812"		"A3SS"	

You are able to define only a single gene if the single gene is inputted. The first column, named by "Index", is a generally used as an identifier and commonly used in other functions of IVAS.

5.2 Estimating expression ratio of AS exons with a data set including FPKM values of transcripts

The RatioFromFPKM function calculates expression ratio between transcripts with and without alternatively spliced exons. First, RatioFromFPKM divides the isoforms from a single gene into two groups: transcripts with and without an alternatively spliced exon. Then, the ratio of the group totals of transcript FPKM values is calculated. The RatioFromFPKM requires expression data set of transcript FPKM values and ASdb with the "SplicingModel" slot. The result will be saved in the "Ratio" slot of ASdb

```
> ASdb <- RatioFromFPKM(GTFdb=sample.Txdb,ASdb=ASdb,Total.expdata=sampleexp,
+ CalIndex="ASS7",Ncor=1,out.dir=NULL)
```

```
> ASdb
```

```
Splicing Models : ES = 182 Rows & ASS = 11 Rows & IR = 2 Rows
```

```
Ratio : ES = 0 Rows by 0 samples & ASS = 1 Rows by 78 samples & IR = 0 Rows by 0 samples
```

```
#ASdb object with SplicingModel & Ratio
```

```
> head(slot(ASdb,"Ratio")$"ASS")
```

Index	EnsID	Nchr	Strand	ShortEX	LongEX
1	"ASS7" "ENSG00000170889"	"19"	"+"	"54704610-54704756"	"54704740-54704829"
	ShortNeighborEX	LongNeighborEX	Short_TX		
1	"54705028-54705149"	"54705028-54705149"	"ENST00000302907 ENST00000391751"		
	Long_TX	Types	NA06984	NA06986	
1	"ENST00000391752 ENST00000402367"	"A5SS"	"0.0370610175563808"	"0.0754673755080699"	
	NA06989	NA06994	NA07037	NA07048	
1	"0.431995041306961"	"0.352248098956179"	"0.615508066951179"	"0.2535297934717"	
	NA07051	NA07056	NA07347	NA07357	
1	"0.396359920018477"	"0.229019337579839"	"0.147021679772774"	"0.294091318766693"	
	NA10847	NA10851	NA11829	NA11830	
1	"0.0835188716083212"	"0.030954840680335"	"0.0174246902189581"	"0.030532429762246"	
	NA11831	NA11843	NA11892	NA11893	
1	"0.15728432880497"	"0.215269984759597"	"0.136324142619792"	"0.38436403201718"	
	NA11894	NA11920	NA11930	NA11931	
1	"0.212453507751045"	"0.0333217262293684"	"0.0681235360867984"	"0.0248643687301508"	
	NA11992	NA11993	NA11994	NA11995	
1	"0.245183066925427"	"0.276368360584032"	"0.114340505887804"	"0.0688073456257966"	
	NA12004	NA12006	NA12043	NA12044	
1	"0.0218694795539099"	"0.719903020532971"	"0.0253480818915533"	"0.0691011133998759"	
	NA12045	NA12058	NA12144	NA12154	
1	"0.269579049697507"	"0.35466877311412"	"0.495392792194793"	"0.0353058516847381"	
	NA12155	NA12249	NA12272	NA12273	
1	"0.0356549500182518"	"0.332527122764556"	"0.547392066663861"	"0.0461822327489977"	
	NA12275	NA12282	NA12283	NA12286	
1	"0.134086715517285"	"0.584161781407799"	"0.0376982756006002"	"0.242375101101388"	
	NA12287	NA12340	NA12341	NA12342	
1	"0.298063167714008"	"0.344188773640231"	"0.0398630023057284"	"0.189423547621436"	
	NA12347	NA12348	NA12383	NA12399	
1	"0.304069583466665"	"0.0175489426208136"	"0.0285158376928488"	"0.0208852172856115"	
	NA12400	NA12413	NA12489	NA12546	
1	"0.162129766403219"	"0.272617060489197"	"0.102893668902972"	"0.0480414433339426"	
	NA12716	NA12717	NA12718	NA12749	
1	"0.38349497995995"	"0.180658216895047"	"0.269490808164129"	"0.0967290881103072"	
	NA12750	NA12751	NA12761	NA12763	
1	"0.226219189907337"	"0.024698616842959"	"0.167469444660537"	"0.0288808382631836"	
	NA12775	NA12777	NA12778	NA12812	
1	"0.475715991162855"	"0.281683152810493"	"0.00965356629010742"	"0.0351200909069428"	
	NA12814	NA12815	NA12827	NA12829	
1	"0.583521251701744"	"0.500254757772165"	"0.313010002548281"	"0.035631624411982"	
	NA12830	NA12842	NA12843	NA12872	
1	"0.0102342527143945"	"0.684792301681123"	"0.251585485678078"	"0.220474614626964"	
	NA12873	NA12874	NA12889	NA12890	
1	"0.353762604584194"	"0.0362672467194004"	"0.222764959132477"	"0.27775528737905"	

In this example, we will estimate ratio in the "ASS7" index among splicing models in ASdb.

5.3 Finding SQTls

Using "SplicingModel" and "Ratio" slots in ASdb from Splicingfinder and RatioFromFPKM, respectively, the sQTlsFinder function can identify significant SNPs associated with alternative splicing rate (ratio). The result will be saved in the "sQTls" slot of ASdb

```
> ASdb <- sQTlsFinder(ASdb=ASdb, Total.snpdata=samplesnp,
+   Total.snplocus=samplesnplocus, method="lm", Ncor=1)
> ASdb
```

Splicing Models : ES = 182 Rows & ASS = 11 Rows & IR = 2 Rows

Ratio : ES = 0 Rows by 0 samples & ASS = 1 Rows by 78 samples & IR = 0 Rows by 0 samples

sQTls : ES = 0 Rows & ASS = 1 Rows & IR = 0 Rows

#ASdb object with SplicingModel & Ratio & sQTls

```
> head(slot(ASdb, "sQTls")$"ASS")

      SNP      Index  EnsID      Strand Nchr Types  ShortEX
[1,] "rs3810232" "ASS7" "ENSG00000170889" "+"    "19" "A5SS" "54704610-54704756"
      LongEX      ShortNeighborEX      LongNeighborEX      pByGeno
[1,] "54704740-54704829" "54705028-54705149" "54705028-54705149" "3.98508717225347e-13"
      FdrByGeno      diff      met
[1,] "3.98508717225347e-13" "diff" "lm"
```

In this example, we will run the function with the linear regression model. sQTlsFinder shows chromosome numbers during mapping analysis.

6 Identification of SQTls using multiple cores

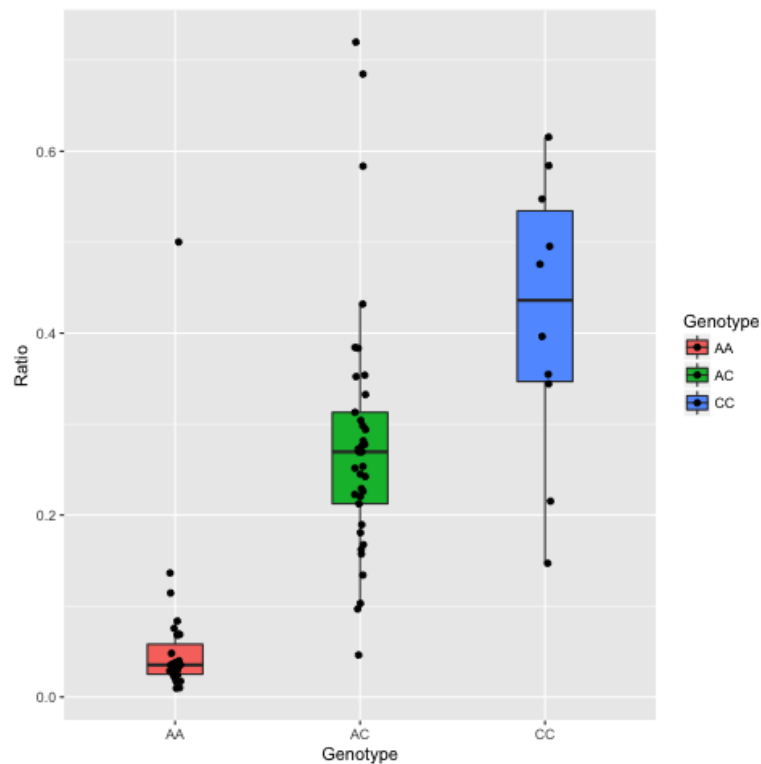
Splicingfinder, RatioFromFPKM, and sQTlsFinder functions provide to use multi-thread through foreach function. The last argument "Ncor" of the functions denotes the number of threads.

```
> ASdb <- Splicingfinder(GTFdb=sample.Txdb, calGene=NULL, Ncor=4)
> ASdb <- RatioFromFPKM(GTFdb=sample.Txdb, ASdb=ASdb, Total.expdata=sampleexp, Ncor=4)
> ASdb <- sQTlsFinder(ASdb=ASdb, Total.snpdata=samplesnp,
+   Total.snplocus=samplesnplocus, method="lm", Ncor = 4)
> ASdb
```

7 Visualizing the result

To visualize the results into boxplot, the IVAS package provides the saveBplot function. Using the data frame from the output of sQTlsFinder function, saveBplot can make the boxplot.

```
> saveBplot(ASdb=ASdb, Total.snpdata=samplesnp, Total.snplocus=samplesnplocus,
+   CalIndex="ASS7", out.dir="./result")
```



The output png files are saved in "result" folder.

8 Session Information

R version 3.4.0 (2017-04-21)

Platform: x86_64-apple-darwin15.6.0 (64-bit)

Running under: OS X El Capitan 10.11.6

Matrix products: default

BLAS: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRblas.0.dylib

LAPACK: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRlapack.dylib

locale:

[1] C/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:

[1] stats4 parallel stats graphics grDevices utils datasets methods

[9] base

other attached packages:

[1] IVAS_1.96.0	ggplot2_2.2.1	GenomicFeatures_1.28.0
[4] AnnotationDbi_1.38.0	Biobase_2.36.0	GenomicRanges_1.28.0
[7] GenomeInfoDb_1.12.0	IRanges_2.10.0	S4Vectors_0.14.0
[10] BiocGenerics_0.22.0		

loaded via a namespace (and not attached):

[1] SummarizedExperiment_1.6.0	splines_3.4.0	lattice_0.20-35
[4] colorspace_1.3-2	htmltools_0.3.5	rtracklayer_1.36.0

[7] yam1_2.1.14	XML_3.98-1.6	nloptr_1.0.4
[10] DBI_0.6-1	BiocParallel_1.10.0	ggfortify_0.4.1
[13] matrixStats_0.52.2	GenomeInfoDbData_0.99.0	foreach_1.4.3
[16] plyr_1.8.4	stringr_1.2.0	zlibbioc_1.22.0
[19] Biostrings_2.44.0	munsell_0.4.3	gtable_0.2.0
[22] codetools_0.2-15	memoise_1.1.0	evaluate_0.10
[25] labeling_0.3	knitr_1.15.1	biomaRt_2.32.0
[28] doParallel_1.0.10	Rcpp_0.12.10	backports_1.0.5
[31] scales_0.4.1	DelayedArray_0.2.0	XVector_0.16.0
[34] lme4_1.1-13	Rsamtools_1.28.0	gridExtra_2.2.1
[37] BiocStyle_2.4.0	digest_0.6.12	stringi_1.1.5
[40] dplyr_0.5.0	rprojroot_1.2	grid_3.4.0
[43] tools_3.4.0	bitops_1.0-6	magrittr_1.5
[46] RCurl_1.95-4.8	lazyeval_0.2.0	RSQLite_1.1-2
[49] tibble_1.3.0	tidyr_0.6.1	MASS_7.3-47
[52] Matrix_1.2-9	minqa_1.2.4	iterators_1.0.8
[55] assertthat_0.2.0	rmarkdown_1.4	R6_2.2.0
[58] nlme_3.1-131	GenomicAlignments_1.12.0	compiler_3.4.0

References

- [1] Keyan Zhao, Zhi-xiang Lu, Juw Won Park, Qing Zhou, Yi Xing. 2013. GLiMMPS: robust statistical model for regulatory variation of alternative splicing using RNA-seq data. *Genome Biol* 14, R74.
- [2] Joseph K. Pickrell, John C. Marioni, Athma A. Pai, Jacob F. Degner, Barbara E. Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, Jonathan K. Pritchard. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464, 768-722.
- [3] N.E. Breslow and D.G. Clayton. 1993. Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association* 88 421: 9-25.
- [4] Michael Lawrence, et al. 2013. Software for Computing and Annotating Genomic Ranges. *PLoS Comput Biol*. 9(8): e1003118.
- [5] Chambers, J. M. 1992. Linear models. Chapter 4 of *Statistical Models in S* eds J. M. Chambers and T. J. Hastie, Wadsworth, and Brooks Cole.
- [6] Tuuli Lappalainen, et al. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506-511.