`ASpediaFI`: Functional Interaction Analysis of AS Events

Doyeong Yu[1], Kyubin Lee[1], Daejin Hyung[1], Soo Young Cho[1], and Charny Park[1]

[1]Bioinformatics Branch, Research Institute, National Cancer Center, Gyeonggi-do, Republic of Korea

October 30, 2019

Contents

1 Introduction

Alternative splicing (AS) is a key contributor to transcriptome and phenotypic diversity. There are hundreds of splicing factors regulating AS events which have a significant impact on diverse biological functions. However, it is a challenge to identify functional AS events related to spliceosome and to explore interplaying genes corresponding to specific pathway. We developed an R package `ASpediaFI` for a systematic and integrative analysis of alternative splicing events and their functional interactions.

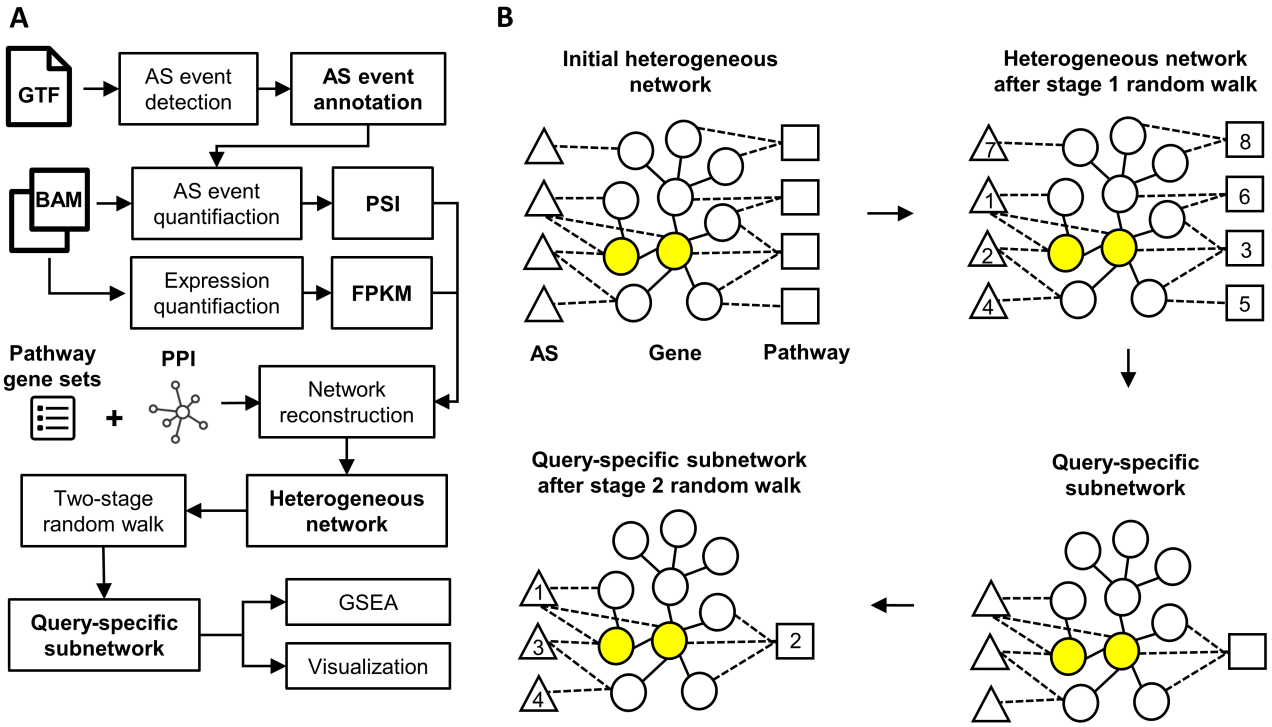

Figure 1: (a) Analytic workflow of `ASpediaFI`. (b) Illustration of the two-stage random walk with restart

Figure 1A shows the analytic workflow of `ASpediaFI`. RNA-Seq BAM files and reference datasets including a GTF file, a gene-gene interaction network, and pathway gene sets are required for this workflow. The workflow begins with obtaining AS event annotations and quantifications. AS event and gene expression quantifications, a gene-gene interaction network, and pathway gene sets are then used to construct a heterogeneous network which contains multiple types of nodes and edges. The initial heterogeneous network (shown in Figure 1B) consists of gene nodes and two types of feature nodes: AS event and pathway.

The next step is to run DRaWR (Discriminative Random Walk with Restarts) [1] on the heterogeneous network to rank AS events and pathways for their relevance to a gene set of interest which is called 'query'. The DRaWR algorithm is demonstrated in Figure 1B. Given the heterogeneous network and a query gene set (colored in yellow), two stages of random walk with restart (RWR) are performed. In the first stage, RWR is run twice on the initial network, one with the query gene set and another with all genes in the network as the restart set. Feature nodes are ranked by the difference between the converged probability distributions in two times of RWR. All feature nodes except top k ranked nodes are removed to reconstruct a query-specific subnetwork. In the second stage, RWR is run on the query-specific subnetwork to obtain final rankings of genes and features. `ASpediaFI` provides the user with the final subnetwork and ranked lists of genes and features for further analysis including gene set enrichment analysis and visualization.

2 Installation

To install ASpediaFI, enter the following commands:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("ASpediaFI")
```

3 Package contents and overview

3.1 Overview of ASpediaFI

ASpediaFI provides the following functionalities:

- AS event detection and annotation
- AS event quantification
- Functional interaction analysis of AS events
- Visualization of AS events and pathways

This package uses a S4 class ASpediaFI as a wrapper of its functionalities (methods) and a container of inputs and outputs (slots).

```
#Load the ASpediaFI package
library(ASpediaFI)
names(getSlots("ASpediaFI"))
```

```
[1] "samples"       "events"        "psi"           "gtf"
[5] "network"       "gene.table"    "as.table"      "pathway.table"
```

The ASpediaFI class contains the following slots:

- samples: a data frame containing information about samples. The first three columns should be names, BAM file paths, and conditions.
- events: a list of AS events extracted from a GTF file.
- psi: a SummarizedExperiment object containing AS event quantification
- gtf: a GRanges object containing genomic features extracted from a GTF file.
- network: an igraph object containing a query-specific subnetwork as a result of DRaWR.
- gene.table, as.table, pathway.table: data frames containing gene nodes, AS event nodes, and pathway nodes.

ASpediaFI performs the analysis by stepwise manner using the following methods:

- annotateASevents: detects AS events from a GTF file and save it in the events field. Also extract features from a GTF file and save in the gtf field.
- quantifyPSI: quantifies AS events using BAM files specified in the samples field.
- analyzeFI: constructs a heterogeneous network of genes, AS events, and pathways and performs DRaWR.
- visualize: visualizes AS event or pathway nodes.
- exportNetwork: exports a subnetwork related to the given pathway to GML format which can be directly used in Cytoscape.

3.2 Case study: SF3B1 mutation in myelodysplastic syndrome

We provide a case study dataset of myelodysplastic syndrome (MDS) patients, GEO accession GSE114922 [2]. This dataset contains 82 MDS patient samples, 28 of which harbored SF3B1 mutations. We prepared gene expression and PSI profiles for the case study. The package provides the profiles of 40 MDS patients, due to file size limitations.

In MDS, hotspot mutation of splicing factor SF3B1 is known to induce distinct subtype, and the study investigated biological functions regulated by the splicing factor mutation. In the following sections, we will walk through the `ASpediaFI` workflow shown in Figure 1 to identify AS events associated with SF3B1 mutation and explore their functional interactions.

4 Workflow

4.1 Input data preparation

Before starting the workflow, `ASpediaFI` requires the following data files to be prepared:

- a GTF file for gene model
- an `igraph` object containing a gene-gene interaction network
- a GMT file containing pathways gene setse
- RNA-Seq BAM files

To begin, we create the `ASpediaFI` object using the constructor function which requires sample names, paths to RNA-Seq BAM files, and sample conditions. We then obtain AS event annotations from a GRCh38 GTF file using the `annotateASevents` method. Due to file size limitations, we extract AS event annotations from a subset of GRCh38 GTF file provided in the `extdata` directory of the package.

```
#Create ASpediaFI object
bamWT <- system.file("extdata/GSM3167290.subset.bam", package = "ASpediaFI")
GSE114922.ASpediaFI <- ASpediaFI(sample.names = "GSM3167290",
                                 bam.files = bamWT, conditions = "WT")

#Detect and annotate AS events from a subset of the hg38 GTF file
gtf <- system.file("extdata/GRCh38.subset.gtf", package = "ASpediaFI")
GSE114922.ASpediaFI <- annotateASevents(GSE114922.ASpediaFI,
                                        gtf.file = gtf, num.cores = 1)
```

```
[1] "--------------------Processing : chr11 --------------------"
```

```
sapply(events(GSE114922.ASpediaFI), length)
```

```
A5SS A3SS   SE  MXE   RI
  35   21   49   40   56
```

```
head(events(GSE114922.ASpediaFI)$SE)
```

```
  EnsID                Nchr    Strand 1stEX                 DownEX
1 "ENSG00000256269.10" "chr11" "+"    "119089683-119089760" "119089217-119089272"
2 "ENSG00000256269.10" "chr11" "+"    "119089082-119089131" "119088635-119088848"
3 "ENSG00000256269.10" "chr11" "+"    "119089082-119089131" "119088635-119088707"
4 "ENSG00000256269.10" "chr11" "+"    "119089100-119089131" "119088635-119088707"
5 "ENSG00000256269.10" "chr11" "+"    "119092125-119092163" "119091413-119091526"
6 "ENSG00000256269.10" "chr11" "+"    "119092125-119092163" "119091861-119091874"
  UpEX
```

```
1 "119089990-119090067"
2 "119089217-119089272"
3 "119089217-119089272"
4 "119089217-119089272"
5 "119092404-119092523"
6 "119092404-119092523"
  EventID
1 "HMBS:SE:chr11:119089217:119089272:119089683:119089760:119089990:119090067"
2 "HMBS:SE:chr11:119088635:119088848:119089082:119089131:119089217:119089272"
3 "HMBS:SE:chr11:119088635:119088707:119089082:119089131:119089217:119089272"
4 "HMBS:SE:chr11:119088635:119088707:119089100:119089131:119089217:119089272"
5 "HMBS:SE:chr11:119091413:119091526:119092125:119092163:119092404:119092523"
6 "HMBS:SE:chr11:119091861:119091874:119092125:119092163:119092404:119092523"
```

The `annotateASevents` method identifies five types of AS events:

- A5SS (alternative 5' splice site)
- A3SS (alternative 3' splice site)
- SE (skipped exon)
- MXE (mutually exclusive exon)
- RI (retained intron)

A list of AS event annotations contains Ensembl ID, chromosome, strand, genomic coordinates of exons, and AS event ID. AS event ID is written in the format of [gene symbol]:[event type]:[chromosome]:[genomic coordinates of exon boundaries] , as defined by ASpedia (https://combio.snu.ac.kr/aspedia/help.html) [3]. Note that the `events` function is available for accessing to AS event annotations. The `annotateASevents` method also extracts genomic features from a GTF file and save in a `gtf` slot as a `GRanges` object for the visualization of AS events.

Next, we quantify AS events from BAM files using the `quantifyPSI` method. The `quantifyPSI` method computes PSI (Percentage Spliced In), the fraction of mRNAs containing the alternatively spliced exon [4]. The user needs to specify a type of RNA-Seq reads (single or paired), read length, insert size, and a minimum number of reads mapped to a given exon. At this point, we compute PSI values from a subset of two BAM files for demonstration. The `quantifyPSI` method saves PSI values in the `psi` slot as a `SummarizedExperiment` object with sample information. Similarly, PSI values can be accessed using the `psi` function. Note that row names of PSI values are AS event IDs.

```
#Compute PSI values of AS events
GSE114922.ASpediaFI <- quantifyPSI(GSE114922.ASpediaFI, read.type = "paired",
                                   read.length = 100, insert.size = 300,
                                   min.reads = 3, num.cores = 1)
```

```
[1] "Calculating PSI of SE events"
[1] "Calculating PSI of MXE events"
[1] "Calculating PSI of RI events"
[1] "Calculating PSI of ALSS events"
```

```
tail(assays(psi(GSE114922.ASpediaFI))[[1]])
```

```
                                                    GSM3167290
HMBS:RI:chr11:119088635:119088707:119089100:119089131       1.00
HMBS:RI:chr11:119089683:119089760:119089990:119090067       0.49
HMBS:RI:chr11:119092125:119092163:119092404:119092523       0.68
HMBS:RI:chr11:119092125:119092163:119092758:119092811       1.00
HMBS:RI:chr11:119092125:119092523:119092758:119092811       0.55
HMBS:RI:chr11:119087987:119088078:119088255:119088308       0.96
```

Since we need PSI and gene expression profiles for all samples to construct a heterogeneous network, we load example datasets in the package. We update the `psi` and `samples` slots with PSI values and sample information stored in the example dataset.

```
#Load PSI and gene expression data
data("GSE114922.fpkm")
data("GSE114922.psi")

#Update the "samples" and "psi" fields
psi(GSE114922.ASpediaFI) <- GSE114922.psi
samples(GSE114922.ASpediaFI) <- as.data.frame(colData(GSE114922.psi))

head(samples(GSE114922.ASpediaFI))
```

```
                 name path condition
GSM3167358 GSM3167358            MUT
GSM3167313 GSM3167313            MUT
GSM3167375 GSM3167375            MUT
GSM3167359 GSM3167359            MUT
GSM3167335 GSM3167335            MUT
GSM3167370 GSM3167370            MUT
```

4.2 Functional interaction analysis of AS events

The `analyzeFI` method performs data preprocessing, network construction and DRaWR (Discriminative Random Walk with Restart). As the DRaWR algorithm requires a query gene set as input, we first detect genes differentially expressed in SF3B1-mutated samples using the `limma` package (other DEG analysis tools such as edgeR and DESeq2 can also be used). Assuming DEGs to represent a functional gene set, we use them as a query to identify AS events and pathways closely related to SF3B1 mutation. If a query is given as a character vector, all genes in the query have equal weights. The user can attribute distinct weights by providing a data frame containing the weights in the second column as a query.

```
#Choose query genes based on differential expression
library(limma)

design <- cbind(WT = 1, MvsW = samples(GSE114922.ASpediaFI)$condition == "MUT")
fit <- lmFit(log2(GSE114922.fpkm + 1), design = design)
fit <- eBayes(fit, trend = TRUE)
tt <- topTable(fit, number = Inf, coef = "MvsW")
query <- rownames(tt[tt$logFC > 1 &tt$P.Value < 0.1,])
head(query)
```

```
[1] "HBB"      "FTL"      "HBA2"     "HBA1"     "ATP6V0D2" "RPL37A"
```

The `analyzeFI` method allows the user to change options for data preprocessing, network construction, and DRaWR. `restart` and `num.feats` define a restart probability and the number of features to be retained in the final subnetwork. The restart probability is the probability of jumping back to the restart set (query). If the restart probability is small, the walk tends to move around the neighbors of query nodes [5]. `num.folds` specifies the number of folds in cross-validation for DRaWR. `low.expr`, `low.var`, `prop.na`, and `prop.extreme` are options for filtering AS events. `cor.threshold` defines a threshold of Spearman's correlation for connecting AS event nodes and gene nodes in a heterogeneous network. Please see help(analyzeFI) for details.

```
#Perform functional interaction analysis of AS events
GSE114922.ASpediaFI <- analyzeFI(GSE114922.ASpediaFI, query = query,
                                 expr = GSE114922.fpkm, restart = 0.9,
                                 num.folds = 5, num.feats = 200,
```

```
                                    low.expr = 1, low.var = 0, prop.na = 0.05,
                                    prop.extreme = 1, cor.threshold = 0.3)
```

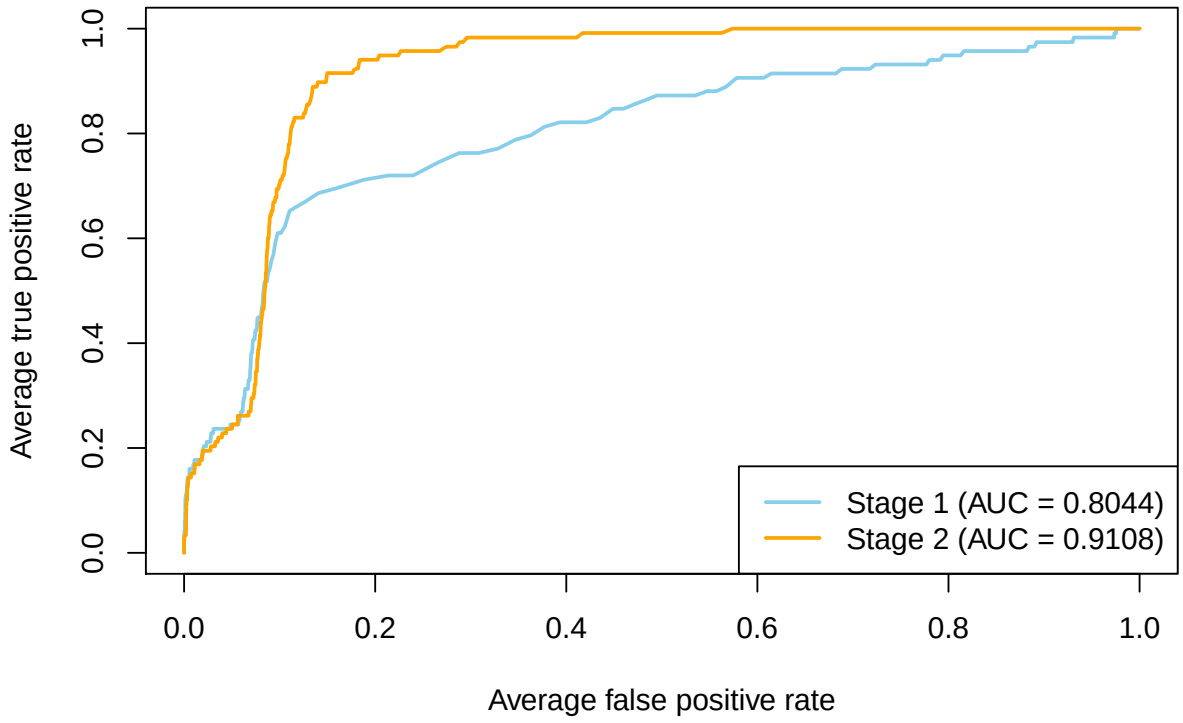

Figure 2: Cross-validation performance of DRaWR

Figure 2 shows an ROC plot from the cross-validation produced by the `analyze` method. 10% of a query gene set is held out as a test set, and the remaining gene set is used as query. Using the stationary probabilities of gene nodes after the first stage and second stage RWR, ROC curves for two stages are computed.

4.3 Reporting

The `analyzeFI` method saves top-ranked AS events and pathways in the `as.table` and `pathway.table` fields, respectively. `as.table` contains the information about AS events including AS event ID, gene symbol, AS event type, final ranking, stationary probability, and PSI values of samples in each condition.

```
#Table of AS nodes in the final subnetwork
as.table(GSE114922.ASpediaFI)[1:5, 2:5]
```

```
  GeneSymbol EventType Rank Probability
1    TIMM17B        SE    1     0.00103
2       NKTR        RI    2     0.00099
3       MCM7        RI    3     0.00099
4     COMMD3        RI    4     0.00094
5      MOV10        RI    5     0.00089
```

```
#Table of GS nodes in the final subnetwork
pathway.table(GSE114922.ASpediaFI)[1:5, 1:11]
```

```
                      Pathway Rank Probability Pvalue Adj.Pvalue
1        REACTOME_IMMUNE_SYSTEM    1     0.00103 0.0001     0.0013
2      HALLMARK_HEME_METABOLISM    2     0.00089 0.0001     0.0013
3           HALLMARK_COMPLEMENT    3     0.00066 0.0001     0.0013
4 REACTOME_INNATE_IMMUNE_SYSTEM    4     0.00057 0.0067     0.0410
5          HALLMARK_P53_PATHWAY    5     0.00054 0.0016     0.0122
  EnrichmentScore NormalizedEnrichmentScore Size Count AvgRank NumEvents
1            0.35                       1.4  933   618    1663        83
2            0.55                       2.0  200   133    1500        83
3            0.53                       1.9  200   119    1178        83
4            0.37                       1.4  279   153    1153        83
5            0.41                       1.5  200   121    1886        83
```

`gs.table` includes the following information about pathway nodes:

- `Pathway`: name of pathway
- `Rank`: final ranking
- `Probability`: stationary probability
- `Pvalue`: GSEA P-value
- `Adj.Pvalue`: GSEA adjusted P-value
- `EnrichmentScore`: enrichment score
- `NormalizedEnrichmentScore`: enrichment score normalized to the average of random samples
- `Size`: the number of total genes in the pathway gene set
- `Count`: the number of genes in the pathway gene set that are also present in the network
- `AvgRank`: the average final ranking of genes in the pathway gene set
- `NumEvents`: the number of AS events in the final subnetwork connected to genes in the pathway gene set
- `Genes`: genes in the pathway gene set that are also present in the network

The results from GSEA are included in `pathway.table` if the `samples` field contains information about sample conditions (e.g. SF3B1 mutation).

The `analyzeFI` method saves the final query-specific subnetwork in the `network` field as an `igraph` object. The user can explore the interactions between AS events and genes as follows:

```
#Extract AS-gene interactions from the final subnetwork
library(igraph)
edges <- as_data_frame(network(GSE114922.ASpediaFI))
AS.gene.interactions <- edges[edges$type == "AS", c("from", "to")]

head(AS.gene.interactions)
```

```
      from
27032  A2M
27033  A2M
27034  A2M
27035  A2M
27036  A2M
27037  A2M
                                                                                to
27032   CLK4:SE:chr5:178623416:178623256:178620673:178619838:178618778:178618556
27033           MRRF:A5SS:chr9:122285169:122285287:122285946:122286039:122286166
27034 VPS29:SE:chr12:110499546:110499457:110497002:110496823:110496203:110496046
27035        ERCC8:SE:chr5:60898400:60898276:60893643:60892037:60891086:60890889
27036                       CDK5RAP3:RI:chr17:47974400:47974448:47975159:47975337
27037       RBM4B:SE:chr11:66668685:66668615:66666493:66666272:66665578:66665548
```

ASpediaFI also allows the user to export the entire subnetwork or a subnetwork related to specific pathway using the **exportNetwork** method. Given a pathway node, the **exportNetwork** method extracts a pathway-specific network and exports it to GML format which can be directly used in Cytoscape. If a pathway node is not given, the entire final subnetwork is exported.

```
#Export a pathway-specific subnetwork to GML format
exportNetwork(GSE114922.ASpediaFI, node = "HALLMARK_HEME_METABOLISM",
              file = "heme_metabolism.gml")
```

4.4 Visualization

The **visualize** method enables visualization of AS events or pathways. If the user provides an AS event nodes as input, it produces a plot describing the AS event and a boxplot of PSI values. Note that the **gtf** field must contain a **GRanges** object with genomic features extracted from the GTF file. The genomic region around the AS event can be zoomed by setting **zoom** to **TRUE**. Figure 3 illustrates the mutually exclusive exons of HMBS, which has been shown to be associated with SF3B1 mutation in MDS.

```
#Check if any event on the HMBS gene is included in the final subnetwork
as.nodes <- as.table(GSE114922.ASpediaFI)$EventID
HMBS.event <- as.nodes[grep("HMBS", as.nodes)]

#Visualize event
visualize(GSE114922.ASpediaFI, node = HMBS.event, zoom = FALSE)
```

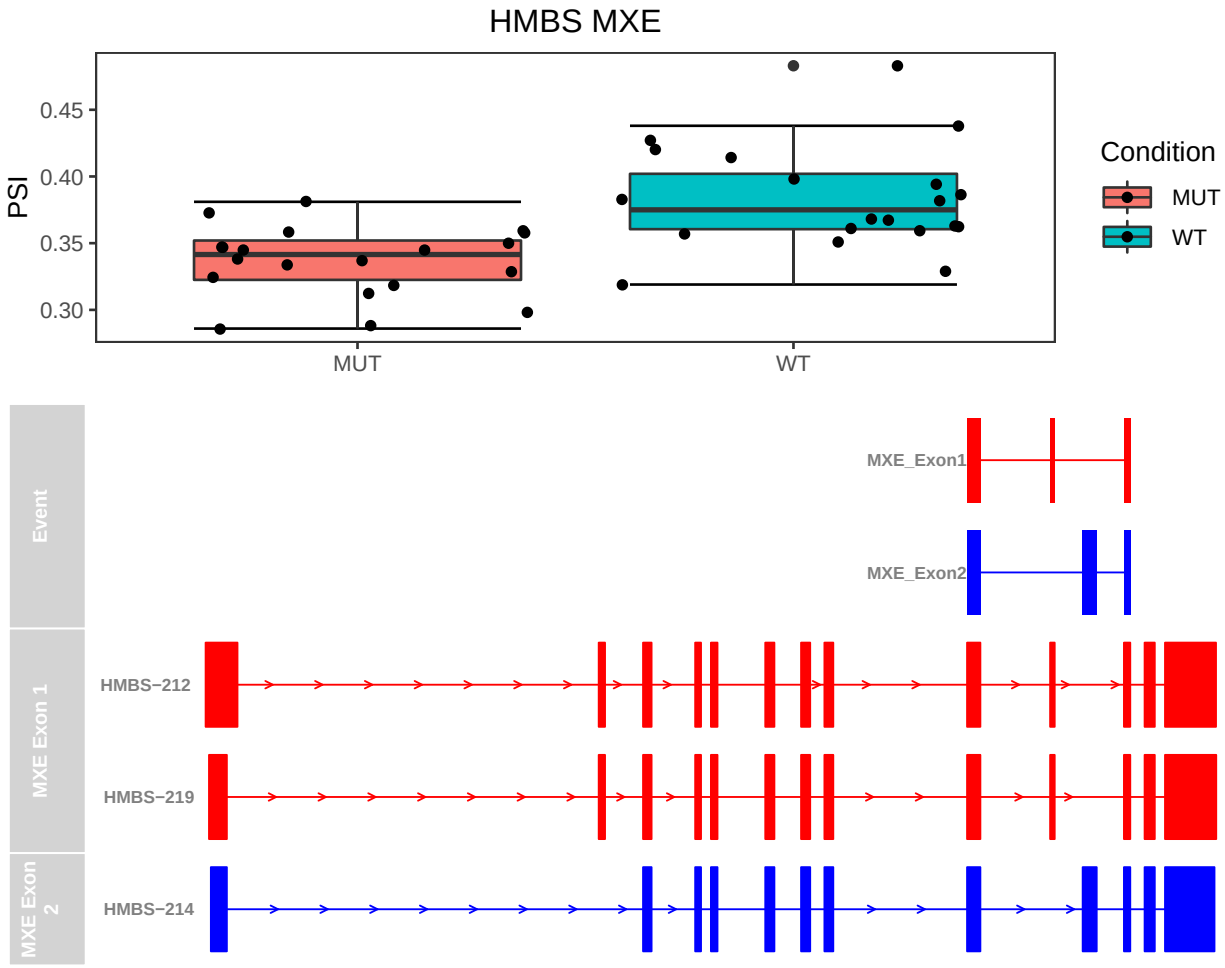

Figure 3: AS event visualization

If a pathway node is given, the `visualize` method shows a subnetwork consisting of highly ranked gene nodes and AS event nodes connected to the given pathway. The user can change the number of gene and AS event nodes to be shown in the subnetwork by setting `n`. Figure 4 demonstrates the subnetwork related to the hallmark pathway of heme metabolism which has also been shown to be associated with SF3B1 mutation in MDS.

```
#Visualize nework pertaining to specific pathway
visualize(GSE114922.ASpediaFI, node = "HALLMARK_HEME_METABOLISM", n = 10)
```

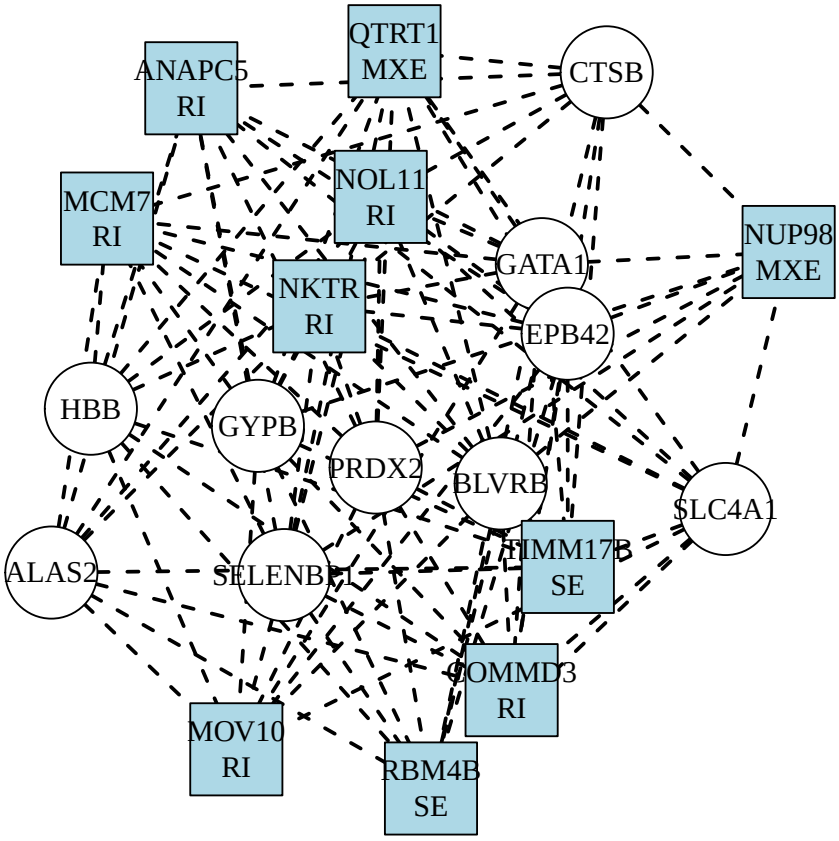

Figure 4: Pathway visulization

5 Session Info

```
sessionInfo()

R version 3.6.1 (2019-07-05)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows Server 2012 R2 x64 (build 9600)

Matrix products: default

locale:
[1] LC_COLLATE=C                           LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252 LC_NUMERIC=C
[5] LC_TIME=English_United States.1252

attached base packages:
[1] parallel  stats4    stats     graphics  grDevices utils     datasets
[8] methods   base

other attached packages:
 [1] igraph_1.2.4.1              limma_3.42.0
```

```
 [3] ASpediaFI_1.0.0             ROCR_1.0-7
 [5] gplots_3.0.1.1              SummarizedExperiment_1.16.0
 [7] DelayedArray_0.12.0         BiocParallel_1.20.0
 [9] matrixStats_0.55.0          Biobase_2.46.0
[11] GenomicRanges_1.38.0        GenomeInfoDb_1.22.0
[13] IRanges_2.20.0              S4Vectors_0.24.0
[15] BiocGenerics_0.32.0         knitr_1.25

loaded via a namespace (and not attached):
  [1] snow_0.4-3               backports_1.1.5          Hmisc_4.2-0
  [4] fastmatch_1.1-0          BiocFileCache_1.10.0     plyr_1.8.4
  [7] lazyeval_0.2.2           splines_3.6.1            ggplot2_3.2.1
 [10] digest_0.6.22            foreach_1.4.7            ensembldb_2.10.0
 [13] htmltools_0.4.0          gdata_2.18.0             ggfortify_0.4.7
 [16] magrittr_1.5             checkmate_1.9.4          memoise_1.1.0
 [19] BSgenome_1.54.0          cluster_2.1.0            doParallel_1.0.15
 [22] Biostrings_2.54.0        askpass_1.1              prettyunits_1.0.2
 [25] colorspace_1.4-1         blob_1.2.0               rappdirs_0.3.1
 [28] xfun_0.10                dplyr_0.8.3              crayon_1.3.4
 [31] RCurl_1.95-4.12          lme4_1.1-21              zeallot_0.1.0
 [34] survival_2.44-1.1        VariantAnnotation_1.32.0 iterators_1.0.12
 [37] glue_1.3.1               gtable_0.3.0             zlibbioc_1.32.0
 [40] XVector_0.26.0           scales_1.0.0             DBI_1.0.0
 [43] DRaWR_1.0.1              Rcpp_1.0.2               progress_1.2.2
 [46] htmlTable_1.13.2         foreign_0.8-72           bit_1.1-14
 [49] Formula_1.2-3            htmlwidgets_1.5.1        httr_1.4.1
 [52] fgsea_1.12.0             RColorBrewer_1.1-2       acepack_1.4.1
 [55] pkgconfig_2.0.3          XML_3.98-1.20            Gviz_1.30.0
 [58] nnet_7.3-12              dbplyr_1.4.2             labeling_0.3
 [61] tidyselect_0.2.5         rlang_0.4.1              reshape2_1.4.3
 [64] AnnotationDbi_1.48.0     munsell_0.5.0            tools_3.6.1
 [67] RSQLite_2.1.2            evaluate_0.14            stringr_1.4.0
 [70] bit64_0.9-7              caTools_1.17.1.2         purrr_0.3.3
 [73] AnnotationFilter_1.10.0  IVAS_2.6.0               ismev_1.42
 [76] nlme_3.1-141             biomaRt_2.42.0           compiler_3.6.1
 [79] rstudioapi_0.10          curl_4.2                 e1071_1.7-2
 [82] tibble_2.1.3             stringi_1.4.3            highr_0.8
 [85] GenomicFeatures_1.38.0   lattice_0.20-38          ProtGenerics_1.18.0
 [88] Matrix_1.2-17            nloptr_1.2.1             vctrs_0.2.0
 [91] pillar_1.4.2             lifecycle_0.1.0          GSA_1.03.1
 [94] data.table_1.12.6        bitops_1.0-6             rtracklayer_1.46.0
 [97] R6_2.4.0                 latticeExtra_0.6-28      KernSmooth_2.23-16
[100] gridExtra_2.3            codetools_0.2-16         dichromat_2.0-0
[103] boot_1.3-23              MASS_7.3-51.4            gtools_3.8.1
[106] assertthat_0.2.1         openssl_1.4.1            GenomicAlignments_1.22.0
[109] Rsamtools_2.2.0          GenomeInfoDbData_1.2.2   mgcv_1.8-30
[112] hms_0.5.1                grid_3.6.1               rpart_4.1-15
[115] tidyr_1.0.0              mGSZ_1.0                 class_7.3-15
[118] minqa_1.2.4              biovizBase_1.34.0        base64enc_0.1-3
```

6 References

[1] Blatti, C. et al. (2016). Characterizing gene sets using discriminative random walks with restart on heterogeneous biological networks. *Bioinformatics*, **32**, 2167–2175.

[2] Pellagatti, A. et al. (2018). Impact of spliceosome mutations on RNA splicing in myelodysplasia: dysregulated genes/pathways and clinical associations. *Blood*, **132**, 1225–1240.

[3] Hyung, D. et al. (2018). ASpedia: a comprehensive encyclopedia of human alternative splicing. *Nucleic acids research*, **46**, D58–D63.

[4] Katz, Y. et al. (2018). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature methods*, **7**, 1009–1015.

[5] Jin, W. et al. (2018). Supervised and extended restart in random walks for ranking and link prediction in networks. *PloS one*, **14**, e0213857.