`ASpediaFI`: Functional Interaction Analysis of AS Events

Doyeong Yu[1]

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

Contents

1 Introduction **2**

2 Installation **2**

3 Package contents and overview **3**
3.1 Overview of `ASpediaFI` 3
3.2 Case study: SF3B1 mutation in myelodysplastic syndrome 4

4 Workflow **4**
4.1 Input data preparation 4
4.2 Functional interaction analysis of AS events 6
4.3 Reporting 7
4.4 Visualization 8

5 References **9**

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 challenging to identify functional events related to a specific splicing factor among thousands of them and explore their associations with genes and pathways. We developed an R package `ASpediaFI` for a systematic and integrative analysis of alternative splicing events and their functional interactions.

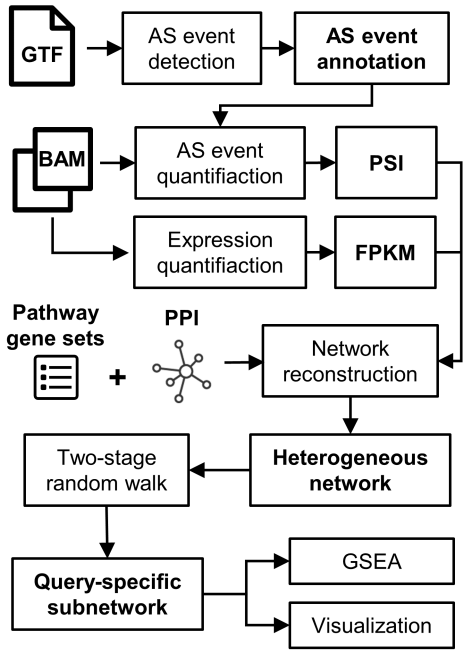

Figure 1: Analytic workflow of `ASpediaFI`

Figure 1 shows the analytic workflow of `ASpediaFI`. The workflow consists of the following steps:

1. Annotate and quantify AS events.
2. Prepare gene expression quantification, a gene-gene interaction network, and pathway gene sets.
3. Construct a heterogeneous network comprising gene, AS event, and pathway nodes.
4. Run DRaWR on the heterogeneous network to rank AS events and pathways for their relevance to a splicing factor or a gene set of interest.

At the end of the workflow, a relevant subnetwork and a ranked list of AS events and pathways will be generated for further 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 reference class ASpediaFI as a wrapper of its functionalities (methods) and a container of inputs and outputs (fields).

```
#Load the ASpediaFI package
library(ASpediaFI)

#Fields of the ASpediaFI reference class
ASpediaFI$fields()
```

```
               samples                 events                    psi
         "data.frame"                 "list" "SummarizedExperiment"
                 gtf                network             gene.table
            "GRanges"              "igraph"           "data.frame"
            as.table          pathway.table
         "data.frame"           "data.frame"
```

The ASpediaFI reference class contains the following fields:

- 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 PSI values of AS events.
- 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.

```
#Methods of the ASpediaFI reference class
setdiff(ASpediaFI$methods(), setRefClass("default")$methods())
```

```
[1] "analyze"   "detect"    "quantify"  "visualize"
```

Other than default methods provided for all reference classes, the ASpediaFI reference class includes the following methods:

- detect: 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.
- quantify: computes PSI values of AS events from BAM files specified in the samples field.
- analyze: constructs a heterogeneous network of genes, AS events, and pathways and performs DRaWR.
- visualize: visualizes AS event or pathway nodes.

3.2 Case study: SF3B1 mutation in myelodysplastic syndrome

In this vignette, we will explore all these functionalities using a dataset of myelodysplastic syndrome (MDS) patients from the study GSE114922 [1]. This dataset contains 82 MDS patient samples, 28 of which harbored SF3B1 mutations. The RNA-Seq reads from the GEO database were aligned to the GRCh38 genome using STAR and quantified using RSEM.

SF3B1 is one of the most frequently mutated splicing factor in MDS, and the study have shown its association with AS events of genes involved in heme metabolism. 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

To begin, we instantiate the `ASpediaFI` class and obtain a list of AS event annotations from a GRCh38 GTF file using the `detect` 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.

```
#Instantiate the ASpediaFI reference class
GSE114922.ASpediaFI <- ASpediaFI()

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

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

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

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

```
head(GSE114922.ASpediaFI$events$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 `detect` 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. The `detect` method also extracts genomic features from a GTF file and save in a `gtf` field as a `GRanges` object for the visualization of AS events.

Next, we quantify AS events from BAM files using the `quantify` method. The `quantify` method requires the `samples` field to have three columns in the following order: name (sample ID), path (BAM file path), and condition (sample condition). Also, a type of RNA-Seq reads (single or paired), read length, insert size, and a minimum number of reads mapped to a given exon need to be specified. At this point, we compute PSI values from a subset of one BAM file for demonstration. The `quantify` method saves PSI values in the `psi` field as a `SummarizedExperiment` object with sample information. Note that row names of PSI values are AS event IDs.

```
#Compute PSI values of AS events
bam <- system.file("extdata/GSM3167287.subset.bam", package = "ASpediaFI")
GSE114922.ASpediaFI$samples <- data.frame(name = "GSM3167287", path = bam,
                                          condition = "")
GSE114922.ASpediaFI$quantify(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 ASS events"
```

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

```
                                                    GSM3167287
HMBS:RI:chr11:119088635:119088707:119089100:119089131       1.00
HMBS:RI:chr11:119089683:119089760:119089990:119090067       0.45
HMBS:RI:chr11:119092125:119092163:119092404:119092523       0.61
HMBS:RI:chr11:119092125:119092163:119092758:119092811       1.00
HMBS:RI:chr11:119092125:119092523:119092758:119092811       0.52
HMBS:RI:chr11:119087987:119088078:119088255:119088308       0.96
```

In addition to AS event annotations and quantifications, the following are required for network reconstruction:

- a named list of pathway gene sets
- an `igraph` object containing a gene-gene interaction network
- a matrix or `SummarizedExperiment` object containing gene expression profiles (FPKM)

If the first two inputs are not given, `ASpediaFI` uses a combined list of HALLMARK, KEGG, and REACTOME pathway gene sets and a network with gene interactions collected from BIND, DIP, HPRD, and REACTOME. Gene expression quantifications must be prepared by the user using quantification tools such as RSEM and Cufflinks. Here, we use gene expression profiles of MDS patients stored in the package as an example dataset. Since we need sample information and AS event quantifications for all samples, we also load the example dataset containing PSI values and sample information to update the `psi` and `samples` fields.

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

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

head(GSE114922.ASpediaFI$samples)
```

```
                 name path condition
GSM3167287 GSM3167287            MUT
GSM3167290 GSM3167290             WT
GSM3167294 GSM3167294            MUT
GSM3167295 GSM3167295             WT
GSM3167297 GSM3167297            MUT
GSM3167298 GSM3167298             WT
```

4.2 Functional interaction analysis of AS events

The `analyze` method performs data preprocessing, network construction and DRaWR (Discriminative Random Walk with Restart) [2]. AS event and gene expression quantifications, a gene-gene interaction network, and pathway gene sets are used to construct a heterogeneous network composed of gene, AS event, and pathway nodes. DRaWR is then applied to identify AS events and pathways associated with a gene set of interest.

The DRaWR algorithm consists of two stages of random walk with restart (RWR). RWR is run on the heterogeneous network twice in the first stage, one with a query gene set and another with all genes as the restart set. AS events and pathways are ranked by the difference between the converged probability distributions in two times of RWR. In the second stage, RWR is run on a subnetwork composed of all gene nodes and top k ranked feature (AS event or pathway) nodes to obtain final rankings of genes and features.

As the DRaWR algorithm requires a query gene set as input, we first perform t-test on the gene expression dataset to detect genes differentially expressed in SF3B1-mutated samples. These DEGs are used 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
pvalues <- apply(log2(GSE114922.fpkm + 1), 1, function(x)
    t.test(x ~ GSE114922.ASpediaFI$samples$condition)$p.value)
query <- names(pvalues)[pvalues < 0.01]
head(query)
```

```
[1] "ALAS2"    "ATP6V0D2" "BLVRB"    "C1QC"     "CD163"    "CD74"
```

The `analyze` 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. `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$analyze(query = query, expr = GSE114922.fpkm,
                            restart = 0.7, num.folds = 5, num.feats = 100,
                            low.expr = 1, low.var = NULL, prop.na = 0.05,
                            prop.extreme = 1, cor.threshold = 0.3)
```

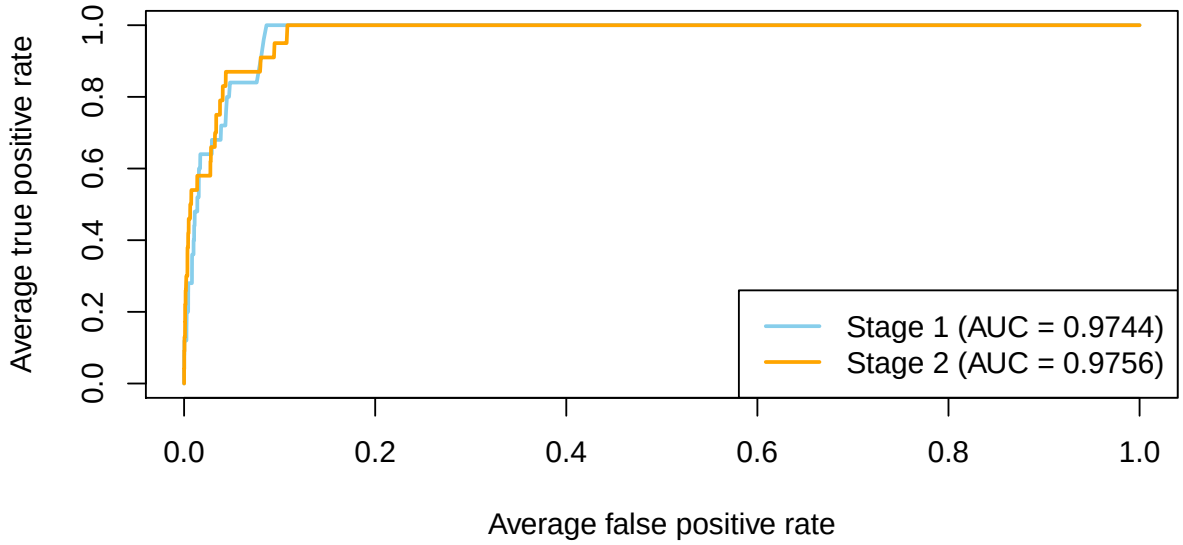

Figure 2: Performance of DRaWR

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 converged probability distributions after the first stage and second stage RWR, ROC curves for two stages are computed.

4.3 Reporting

The `analyze` method saves top-ranked AS events and pathways in the `as.table` and `pathway.table` fields, respectively. If the `samples` field contains information about sample conditions (e.g. SF3B1 mutation), the results from gene set enrichment analysis are included in the `pathway.table` field. The column `avg.rank` is the average rank of gene nodes in the corresponding pathway. The last column `neighborAS` is the number of AS event nodes in the final subnetwork connected to the pathway.

```
#Table of AS nodes in the final subnetwork
head(GSE114922.ASpediaFI$as.table, 5)
```

```
                                                                                node
3                                          VPS52:RI:chr6:33264900:33264782:33264497:33264374
4                        HMBS:SE:chr11:119092125:119092163:119092404:119092523:119092758:119092811
5 TUBGCP4:MXE:chr15:43400044:43400221:43401716:43403799:43404413:43404552:43405202:43405323
6                                    NAP1L4:A3SS:chr11:2976123:2976024:2972279:2972243:2972102
8   QTRT1:MXE:chr19:10707302:10707380:10707500:10707615:10712161:10712628:10712758:10712867
    prob
3 0.0041
4 0.0039
5 0.0038
6 0.0038
8 0.0037
```

```
#Table of GS nodes in the final subnetwork
head(GSE114922.ASpediaFI$pathway.table, 5)
```

```
                      node   prob   pval   padj   ES  NES size avg.rank
1     HALLMARK_P53_PATHWAY 0.0089 0.1313 0.3759 0.24 1.22  121      709
2 HALLMARK_HEME_METABOLISM 0.0077 0.0001 0.0019 0.52 2.66  134      736
3      REACTOME_CELL_CYCLE 0.0037 0.0001 0.0019 0.28 1.51  323     1422
4      REACTOME_HEMOSTASIS 0.0028 0.6385 0.8923 0.17 0.95  257     1317
5            KEGG_LYSOSOME 0.0026 0.0211 0.0989 0.32 1.50   71      960
  neighborAS
1         58
2         59
3         59
4         59
5         59
```

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 skipped exon of HMBS, which has been shown to be associated with SF3B1 mutation in MDS.

```
#Visualize AS event
GSE114922.ASpediaFI$visualize(node = GSE114922.ASpediaFI$as.table$node[2],
                              zoom = FALSE)
```

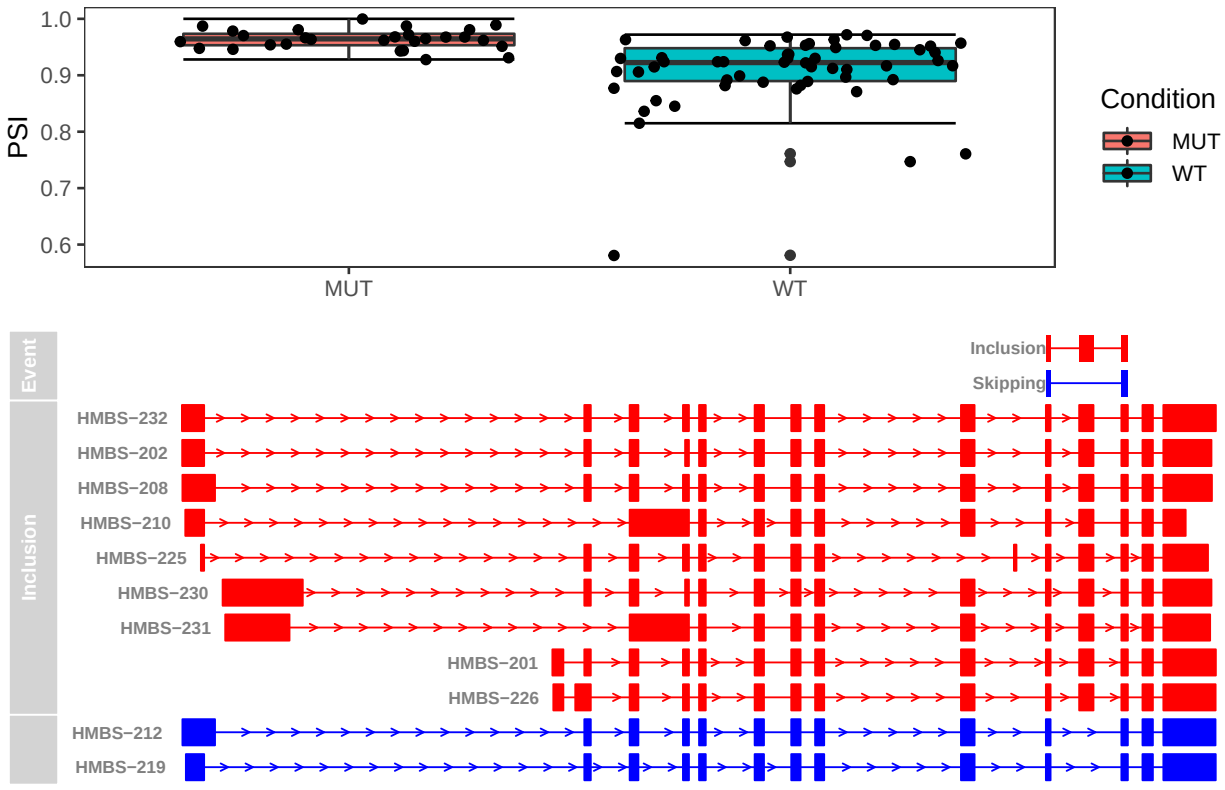

Figure 3: AS event visualization

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
GSE114922.ASpediaFI$visualize(node = GSE114922.ASpediaFI$pathway.table$node[2],
                              n = 10)
```

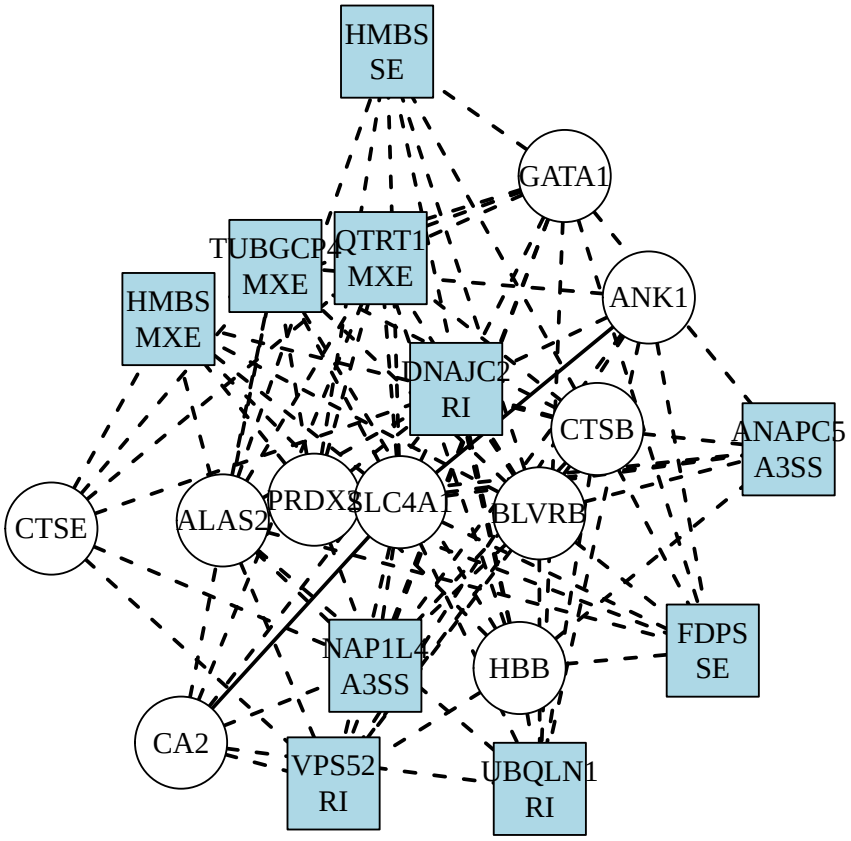

Figure 4: Pathway visulization

5 References

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

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