User Guide for nagnag Package

Xiaoyan Yan[†], Xiaoyong Sun[‡*]

May 4, 2015

[†]Affiliated Hospital of Shandong University of Traditional Chinese Medicine
No. 42 WenhuaWest Road,
Jinan, Shandong 250011, China
[‡]Agricultural Big-Data Center,
College of Information Science and Engineering
Shandong Agricultural University
Taian, Shandong 271018, China

Contents

[*]johnsunx1@gmail.com; johnsunx1@126.com

1 Introduction

The NAGNAG alternative splicing is a regulatory process that controls the mRNA splicing from one gene [1]. NAGNAG alternative splicing has attracted intensive attentions for last decades because of its unique property: inclusion or exclusion of three nucleotides results in difference of one or two amino acids in the final proteins. NAGNAG splicing regulators was first identified as a key regulators [2] accounting for around 20proteome plasticity [3]. Subsequent discoveries investigating the splicesomal machinery and proteome plasticity regulation using the NAGNAG, as a new alternative splice site regulators, has revealed its wide spread occurrences and conservation across several species such as 6annotated genes in Arabidopsis thaliana harbor a genomic NAGNAG acceptor motif [4] as compared to human, where 5motif [2]. Our previous work also proved that this NAGNAG splicing mechanism participates in the splicing process of long intergenic non-coding RNA, which leads to the conclusion that this difference of three nucleotides may have some other biological functions [5].

However, lacking of proper software for detecting NAGNAG splicing significantly hinders the further study of this intriguing system. nagnag is an R package to address this challenge. It can not only detect the NAGNAG alternative splicing through GTF/GFF file, but also calculate the related expression from RNA-Seq data sets, and identify the DNA/RNA/protein difference produced from two isoforms processed through NAGNAG alternative splicing. nagnag is a handy tool for biological community to study the NAGNAG alternative splicing in RNA-Seq data, which is available from https://sourceforge.net/projects/nagnag/files/; or http://genome.sdau.edu.cn/research/software/nagnag.html.

2 Input and output

2.1 Annotation file: gtf/gff file

For requirement of general format, see details at UCSC (http://genome.ucsc.edu/FAQ/FAQformat).

The annotation file should NOT have header, and start as the first line. The file should be tab-delimited text file.

The gtf/gff files from UCSC contains a field to annotate the transcript name. It usually uses ”gene_id” to label gene name, and ”transcript_id” to label transcript name. Please translate your gtf/gff file to this format though we do support one special case: gtf/gff file for Arabidopsis downloaded from TAIR (http://www.arabidopsis.org/).

2.2 Genome sequence file: fasta file

The genome sequence files:fasta file are required to detect nagnag alternative splicing.

2.3 Alignment file: bam file

Each bam file represents one sample, generated from alignment software, such as bowtie or tophat. Paired-end data are treated as single-end data.

2.4 Output

The output columns are,

- **chr**: the chromosome number.
- **source**: the source of the GTF file. It is the column from GTF/GFF file.
- **feature**: the feature of the GTF file, including "exon", "protein", "CDS", etc. It is the column from GTF/GFF file.
- **strand**: the strand of the chromosome.
- **transcriptName**: the unique name for transcripts from the GTF file. Generally it is the last column from GTF/GFF file.
- **gene_id**: gene ID.
- **transcript_id**: the transcript ID.
- **ssite.5**: the alternative splicing site at the upstream.
- **ssite.3**: the alternative splicing site at the downstream.
- **splicingLink**: a character string connecting the alternative splicing site at the upstream to the alternative splicing site at the downstream.
- **asType**: the type of alternative splicing. It can be "acceptor" for alternative acceptor site, or "donor" for alternative donor site.
- **ID**: the unique ID of the alternative splicing event.
- **site1**: the first NAGNAG splicing site (Figure 1).
- **site2**: the second NAGNAG splicing site.
- **seq**: the detailed nucleotide sequences for NAGNAG.
- **site1.splicingLink**: the splicing link for the splicing site 1.
- **site2.splicingLink**: the splicing link for the splicing site 2.
- **site1.transcriptName**: the unique ID for transcripts that contains splicing site 1. It should be the value from the last column from GTF/GFF file.
- **site2.transcriptName**: the unique ID for transcripts that contains splicing site 2. It should be the value from the last column from GTF/GFF file.
- **site1.count**: the number of junction reads support the splicing site 1.
- **site2.count**: the number of junction reads support the splicing site 2.

- **nag**: logical value. It check whether or not the alternative splicing event is NAGNAG alternative splicing.
- **locusType**: the genome annotation of alternative splicing site, including “5UTR”, “exon”, and “3UTR” .
- **frame**: the reading frame at the alternative splicing site.
- **nagSpliceDiff**: the splicing difference of two isoforms at the NAGNAG splicing site.
- **nagType**: the splicing type, including “nagnag+”, or “nagnag++”. “nagnag+” means that the NAGNAG splicing site is the only difference between two isoforms. “nagnag++” means that there are more differences between two isoforms including NAGNAG splicing site.
- **firstSeq**: the upstream sequence before NAGNAG splicing site.
- **lastSeq**: the downstream sequence after NAGNAG splicing site, including NAGNAG splicing site.

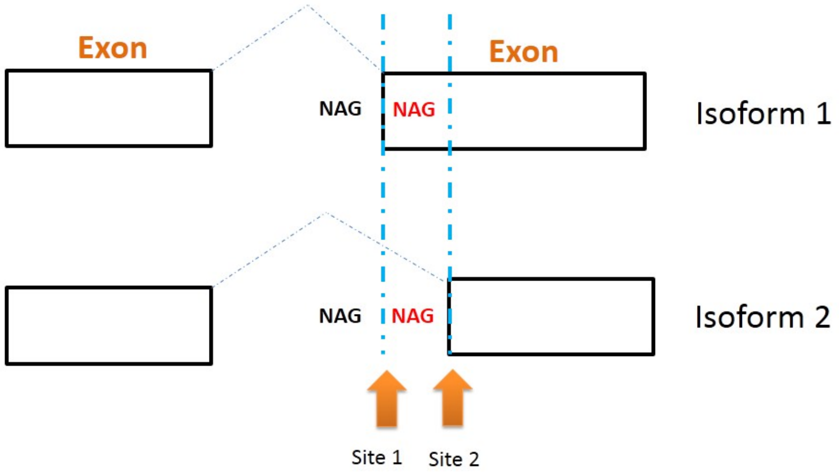

Figure 1: The NAGNAG splicing sites, including splicing site 1 and splicing site 2.

3 Function description

The main features of this package are 1) to detect NAGNAG alternative splicing sites (`altSearch, nagnagDetect`); 2) to count the reads for NAGNAG alternative splicing (`nagnagExpr`); 3) to identify the DNA/RNA/protein sequence difference generated from NAGNAG alternative splicing (nagnagSeq); 4) to summarize the results (nagnagSum).

3.1 Detect NAGNAG alternative splicing sites

3.1.1 altSearch

```
## GTF/GFF are UCSC format
altSearch(gtfFile, genomeFasta, AStype="acceptor", transcriptNameStyle="standard")

## GTF/GFF are TAIR10 format
altSearch(gtfFile, genomeFasta, AStype="acceptor", transcriptNameStyle="arab")
```

It detects all the alternative acceptor sites or alternative donor sites.

Use ? altSearch for details about parameters.

3.1.2 nagnagDetect

```
nagnagDetect(altRef, genomeFasta)
```

It detects all the NAGNAG alternative splicing events.

Use ? nagnagDetect for details about parameters.

3.2 Count the reads for NAGNAG alternative splicing

3.2.1 nagnagExpr

```
## Count the expression of NAGNAG splicing without a reference
nag.expr <- nagnagExpr(bamFile, genomeFasta, nagRef=NULL)

## Count the expression of NAGNAG splicing with a reference
nag.Ref <- nagnagExpr(bamFile, genomeFasta, nagRef)
```

Count the junction reads for novel NAGNAG splicing sites (nagRef=NULL) or for known NAGNAG splicing sites.

Use ? nagnagExpr for details about parameters.

3.3 Identify the DNA/RNA/protein sequence difference generated from NAGNAG alternative splicing

3.3.1 nagnagSeq

```
## produce DNA sequence
n.df <- nagnagSeq(nagEvent, genomeFasta, gtfFile, type="DNA")
```

```
## produce RNA sequence
n.df <- nagnagSeq(nagEvent, genomeFasta, gtfFile, type="RNA")

## produce amino acid sequences
n.df <- nagnagSeq(nagEvent, genomeFasta, gtfFile, type="AA")
```

Use ? `combineCount` for details about parameters.

3.4 Summarize the results

3.4.1 nagnagSum

```
altRef <- altSearch(gtfFile, AStype="acceptor", transcriptNameStyle="arab")
nagnagSum(altRef)

nagRef <- nagnagDetect(altRef, genomeFasta)
nagnagSum(nagRef)

nag.expr <- nagnagExpr(bamFile, genomeFasta, nagRef=NULL)
nagnagSum(nag.expr)
```

This function provides summary for the results from `altSearch`, `nagnagDetect` and `nagnagExpr` function.

3.5 Visualize the results

3.5.1 nagnagVis

```
altRef <- altSearch(gtfFile, AStype="acceptor", transcriptNameStyle="arab")
jpeg(file="test.jpg")
nagnagVis(altRef)
dev.off()

nagRef <- nagnagDetect(altRef, genomeFasta)
jpeg(file="test.jpg")
nagnagVis(nagRef)
dev.off()

nag.expr <- nagnagExpr(bamFile, genomeFasta, nagRef=NULL)
jpeg(file="test.jpg")
nagnagVis(nag.expr)
dev.off()
```

This function provides visualization for the results from `altSearch`, `nagnagDetect` and `nagnagExpr` function. It generates figure files as PDF, jpeg, tiff, etc.

4 Demonstration

1. RNA-Seq data for Arabidopsis.
Example data is available at http://genome.sdau.edu.cn/research/software/nagnag.html.

```
> library(nagnag)
> gtfFile <- "TAIR10_gene.gff"
> genomeFasta <- "TAIR10_all.fa"

> altRef <- altSearch(gtfFile, genomeFasta, AStype="acceptor",
                      transcriptNameStyle="arab")
```

The altRef includes the following columns:

```
> head(altRef)
       chr source feature strand     transcriptName   gene_id transcript_id
63495 chr1 TAIR10    exon      + Parent=AT1G28490.1 AT1G28490   AT1G28490.1
63514 chr1 TAIR10    exon      + Parent=AT1G28490.2 AT1G28490   AT1G28490.2
7416  chr1 TAIR10    exon      + Parent=AT1G03960.1 AT1G03960   AT1G03960.1
7443  chr1 TAIR10    exon      + Parent=AT1G03960.2 AT1G03960   AT1G03960.2
64509 chr1 TAIR10    exon      + Parent=AT1G29120.1 AT1G29120   AT1G29120.1
64535 chr1 TAIR10    exon      + Parent=AT1G29120.2 AT1G29120   AT1G29120.2
       ssite.5  ssite.3      splicingLink   asType ID
63495 10016680 10016869 10016680-10016869 acceptor  1
63514 10016680 10016865 10016680-10016865 acceptor  2
7416   1014543  1014767   1014543-1014767 acceptor  3
7443   1014543  1014621   1014543-1014621 acceptor  4
64509 10174670 10175216 10174670-10175216 acceptor  5
64535 10174670 10175216 10174670-10175216 acceptor  6

> nagRef <- nagnagDetect(altRef, genomeFasta)
[1] "start.."
[1] "Site search..Done.."
[1] "NAG search..Done.."
[1] "Done.."
> head(nagRef)
           ID  chr  site1  site2 strand    seq site1.splicingLink
538 4468-4465 chr1  35471  35474      - TAGCAG        35471-35567
539 4468-4466 chr1  35471  35474      - TAGCAG        35471-35567
540 4468-4467 chr1  35471  35474      - TAGCAG        35471-35567
541 4468-4469 chr1  35471  35474      - TAGCAG        35471-35567
87    692-693 chr1 375122 375125      + AAGCAG      374922-375122
544 4543-4544 chr1 513246 513249      - AAGCAG      513246-513332
    site1.transcriptName site2.splicingLink site2.transcriptName
538   Parent=AT1G01060.4        35474-35567   Parent=AT1G01060.1
539   Parent=AT1G01060.4        35474-35567   Parent=AT1G01060.2
540   Parent=AT1G01060.4        35474-35567   Parent=AT1G01060.3
541   Parent=AT1G01060.4        35474-35567   Parent=AT1G01060.5
87    Parent=AT1G02080.1      374922-375125   Parent=AT1G02080.2
544   Parent=AT1G02470.1      513249-513332   Parent=AT1G02470.2
```

After we get all the potential NAGNAG splicing sites, let's input the RNA-Seq data.

```
> bamFile <- c("accepted_hits.bam")
> nag.expr <- nagnagExpr(bamFile, genomeFasta, nagRef=NULL)
```

```
[1] "Start analysis.."
[1] "Read bam file..Done.."
[1] "Check junction reads..Done.."
[1] "Find NAGNAG candidate..Done.."
[1] "Count NAGNAG reads..Done.."
>
> nagEvent <- nagRef[1,]
> n.df <- nagnagSeq(nagEvent, genomeFasta, gtfFile, type="AA")
```

```
> nagnagSum(altRef)
$donor
  AlterSiteNo LocusNo
1           2    1889
2           3     747
3           4     186
4           5      77
5           6      21
6           7       8
7           8       3
8           9       1
9          10       0

$acceptor
  AlterSiteNo LocusNo
1           2     811
2           3     166
3           4      55
4           5      13
5           6       4
6           7       1
7           8       1
8           9       0

> nagnagSum(nagRef)
$nagnagNo
[1] 526

$nag
GAG AAG TAG CAG
 64 246 247 495

$nagnag

GAGGAG GAGTAG AAGGAG GAGAAG CAGGAG TAGGAG AAGTAG AAGAAG GAGCAG CAGTAG TAGAAG
     1      1      2      3     14     15     17     27     27     29     30
TAGTAG AAGCAG CAGAAG TAGCAG CAGCAG
    33     54     86     89     98

> nagnagSum(nag.expr)
```

```
$nagnagNo
[1] 8

$nag
GAG AAG CAG TAG
  1   3   4   8

$nagnag

CAGAAG CAGGAG TAGAAG TAGCAG TAGTAG
     1      1      2      2      2
#jpeg("test1.jpg")
#nagnagVis(altRef)
#dev.off()

#jpeg("test2.jpg")
#nagnagVis(nagRef)
#dev.off()

#jpeg("test3.jpg")
#nagnagVis(nag.expr)
#dev.off()

#jpeg("test4.jpg", width=5, height=5, units="in", res=500)
#vis.df <- nagnagVis(nag.expr, showPvalue=TRUE)
#dev.off()
%@
```

2. RNA-Seq data for human

```
library(nagnag)
gtfFile <- "hg19.gtf"
genomeFasta <- "hg19.fa"
altRef <- altSearch(gtfFile, AStype="acceptor")
nagRef <- nagnagDetect(altRef, genomeFasta)

bamFile <- c("accepted_hits.bam")
nag.expr <- nagnagExpr(bamFile, genomeFasta, nagRef)

nagEvent <- nagRef[1,]
n.df <- nagnagSeq(nagEvent, genomeFasta, gtfFile, type="AA")
n.df

nagnagSum(altRef)
nagnagSum(nagRef)
nagnagSum(nag.expr)

jpeg("test5.jpg")
nagnagVis(altRef)
dev.off()
```

```
jpeg("test6.jpg")
nagnagVis(nagRef)
dev.off()

jpeg("test7.jpg")
nagnagVis(nag.expr)
dev.off()

jpeg("test8.jpg", width=5, height=5, units="in", res=500)
vis.df <- nagnagVis(nag.expr, showPvalue=TRUE)
dev.off()
```

References

[1] Shi Y, Sha G, Sun X. (2014). **Genome-wide study of NAGNAG alternative splicing in Arabidopsis.** *Planta.* **239(1)**:127-138.

[2] Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, Schreiber S, Backofen R, Platzer M. (2004). **Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity.** *Nat Genet.* **36**:1255-1257.

[3] Bradley R, Merkin J, Lambert N, Burge C. (2012). **Alternative splicing of RNA triplets is often regulated and accelerates proteome evolution**. *PLoS Biol.* **10(1)**:e1001229.

[4] Schindler S, Szafranski K, Hiller M, Ali GS, Palusa SG, Backofen R, Platzer M, Reddy AS. (2008). **Alternative splicing at NAGNAG acceptors in Arabidopsis thaliana SR and SR-related protein-coding genes.** *BMC Genomics.* **9**:159.

[5] Sun X, Lin SM, Yan X. (2014). **Computational evidence of NAGNAG alternative splicing in human large intergenic noncoding RNA.** *Biomed Res Int.* **2014**:736798.