Natural AntiSense Transcripts Identification using NASTI-seq

Song Li

Duke Unviersity Institute of Genome Science and Policy

sl228@duke.edu

NASTIseq is a software package for the identification of cis-Natural Antisense Transcripts (cis-NAT) using strand specific RNA-seq data. The method is based on the observation that strand specific RNAseq sometimes generate reads from the unexpected strand. We designed a probabilistic model which estimate the strand specificity and account for it using a binomial model.

1 Installation

NASTIseq 1.0 was build under R version 2.15.2. To install the package, first go to website `http://www.genome.duke.edu/labs/ohler/research/NASTIseq/index.html`. Download the source package. Install the package using command line.

```
$R CMD INSTALL NASTIseq_1.0.tar.gz
```

2 Input Data Preparation

2.1 Read Count Preparation

To use NASTI-seq package, load the package into R. Read count data from whole root sample is provided as part of the package, alone with a number of other important data to illustrate the required data formats and how to use NASTI-seq package.

```
> library(NASTIseq)
> data(WholeRoot)
> names(WholeRoot)

[1] "smat"     "genepos"  "pospairs" "asmat"

> head(WholeRoot$smat)

          [,1] [,2] [,3]
AT1G38440    0    2    0
AT1G43171    2    8    1
```

```
AT1G67670    3    7    0
AT1G75280  341  171  142
AT1G03010   56   19   25
AT1G19850  747  357  473
```

```
> head(WholeRoot$asmat)
```

```
          [,1] [,2] [,3]
AT1G38440    0    0    0
AT1G43171    0    0    0
AT1G67670    0    0    2
AT1G75280    0    6    8
AT1G03010    0    0    0
AT1G19850    6   21   15
```

WholeRoot$smat is a N by M matrix of read count for reads that mapped to the sense strand. N is the number of gene loci in Arabidopsis. M is the number of biological replicates in the sample, and in our case, M equals 3. Each rowname must be a unique locus name. For Arabidopsis, The unique IDs from TAIR [2] are used as the row names. WholeRoot$asmat is a N by M matrix of read count to the antisense strand. Read counts can be obtained using popular software such as RSamtools [1]. For a forward strand gene (locus), the number of reads that map to the forward strand will be stored in the "smat" matrix, and the number of reads that map to the reverse strand will be stored in the "asmat" matrix. For a reverse strand gene (locus), the number of reads that map to the forward strand will be stored in the "asmat" matrix, and the number of reads that map to the reverse strand will be stored in the "smat" matrix.

2.2 Known Natural Antisense Annotation

Annotated cis-NATs are used as positive training set to decide the threshold in the NASTI-seq method. The training set for Arabidopsis is provided as part of the package.

```
> dim(WholeRoot$pospairs)
```

```
[1] 874   2
```

```
> head(WholeRoot$pospairs)
```

```
         V1        V2
1 AT2G46910 AT2G46915
2 AT3G12250 AT3G12260
3 AT5G50315 AT5G50320
4 AT1G76780 AT1G76790
5 AT3G10140 AT3G10150
6 AT3G47550 AT3G47560
```

WholeRoot$pospairs is a two column matrix, with each row contains the names of a known pair of cis-natural antisense transcripts. Our positive training set is originally from [3]; only pairs that are still supported by the TAIR10 annotation are used. For other organisms, you can find annotated cis-NATs

from online resources such as http://natsdb.cbi.pku.edu.cn/index.php or http://bis.zju.edu.cn/pnatdb/.

Another important file for the identification of cis-NATs is the genome annotation file.

```
> dim(WholeRoot$genepos)

[1] 33597     9

> head(WholeRoot$genepos)

          seqname         source feature start   end score strand frame
AT1G01010    Chr1 protein_coding    exon  3631  5899     .      +     .
AT1G01020    Chr1 protein_coding    exon  5928  8737     .      -     .
AT1G01030    Chr1 protein_coding    exon 11649 13714     .      -     .
AT1G01040    Chr1 protein_coding    exon 23146 31227     .      +     .
AT1G01046    Chr1          miRNA    exon 28500 28706     .      +     .
AT1G01050    Chr1 protein_coding    exon 31170 33153     .      -     .
          attributes
AT1G01010  AT1G01010
AT1G01020  AT1G01020
AT1G01030  AT1G01030
AT1G01040  AT1G01040
AT1G01046  AT1G01046
AT1G01050  AT1G01050
```

WholeRoot$genepos is similar to the GTF file from TAIR10 annotation. First eight columns are the standard GTF columns. Each row represents a locus, and for each locus, the 'start' and 'end' columns are the genomic coordinates of the locus. 'start' is always smaller than 'end'. In the 9th column, only the unique locus name was used instead of the original annotation field from standard GTF files.

2.3 Running NASTIseq on Whole Root Sample

To predict cis-NATs, first generate NASTI score, which is based on Bayesian information criteria (BIC). Simply type the following commands:

```
> data(WholeRoot)
> WRscore<-getNASTIscore(WholeRoot$smat,WholeRoot$asmat)
>
```

Instead of interpreting the scores at their absolute magnitude, we use training data to find the appropriate threshold in the NASTI scores. To do this, we need to first generate a set of negative training gene pairs that are unlikely to form cis-NAT pairs. The negative pairs can be generated using the provided annotation data. Once the negative pairs are found, the prediction will be made on the positive and negative training pairs.

```
> negpairs<-getnegativepairs(WholeRoot$genepos)
> head(negpairs)
```

```
     locus1    locus2
1 AT1G01070 AT1G01073
2 AT1G01100 AT1G01110
3 AT1G01115 AT1G01120
4 AT1G01180 AT1G01183
5 AT1G01200 AT1G01210
6 AT1G01240 AT1G01250
```

```
> WRpred<-NASTIpredict(WholeRoot$smat,WholeRoot$asmat, WholeRoot$pospairs, negpairs)
>
```

ROCR is an widely used R package for visualizing classifiers. We use the pacakge to create an object that contains the true positive and false positives. We then find the threshold in the NASTI score for a given false discovery rate (FDR) using the training sample and predict cis-NAT pairs.

```
> WRpred_rocr<-prediction(WRpred$predictions,WRpred$labels)
> thr<-defineFDR(WRpred_rocr,0.05)
> WR_names<-FindNATs(WRscore, thr, WholeRoot$pospairs,WholeRoot$genepos)
```

FindNATs is a function that look through the provided annotation file to determine which pair of genes are candidate cis-NATs.

```
> names(WR_names)
```

```
[1] "knownpairs" "newpairs"   "neworphan"
```

```
> head(WR_names$newpairs)
```

```
     [,1]        [,2]
[1,] "AT1G76630" "AT1G76640"
[2,] "AT2G06045" "AT2G06050"
[3,] "AT4G30100" "AT4G30110"
[4,] "AT3G63300" "AT3G63310"
[5,] "AT1G50740" "AT1G50750"
[6,] "AT5G18960" "AT5G18970"
```

```
> head(WR_names$neworphan)
```

```
[1] "ATMG00030" "AT5G49440" "AT2G11240" "AT5G49640" "AT2G33320" "AT2G24840"
```

References

[1] Martin Morgan and Hervé Pagès. *Rsamtools: Binary alignment (BAM), variant call (BCF), or tabix file import*. R package version 1.10.2.

[2] The Arabidopsis Information Resource. https://www.arabidopsis.org/.

[3] X. J. Wang, T. Gaasterland, and N. H. Chua. Genome-wide prediction and identification of cis-natural antisense transcripts in arabidopsis thaliana. *Genome biology*, 6(4):R30, 2005.