

Introduction

N6-methyladenosine (m6A) is the most prevalent RNA modification in eukaryotic cell. Similar to the well-studied DNA methylation, m6A is also a reversible modification that orchestrates important functionalities of transcriptome. Over the past 10 years, m6A can only be identified by some conventional low-throughput methods, such as 2D-thin layer chromatography (2D-TLC), mass-spectrography and SCARLET. Until 2012, the development of methylated RNA immunoprecipitation sequencing (MeRIP-seq, also known as “m6A-seq”) technology has done away with this restriction, allowing the profile of transcriptome-wide m6A landscape in a more high-throughput manner. The interpretation of m6A signal from MeRIP-seq data highly relies on computational methods. Recently, several algorithms were designed to meet the computational need for MeRIP-seq-based m6A peak detection. Although this set of newly developed tools have achieved a fairly robust performance, limitations and deficiencies still existed. Therefore, the development of methodologies that enabling the identification of single-nucleotide-resolution m6A sites are of particularly important and necessary.

Here, we present a novel software, DeepRIP, for the identification of single-nucleotide-resolution m6A sites from MeRIP-seq experiment. DeepRIP deploys the Restricted Boltzmann Machine (RBM), a generative stochastic artificial neural network used in Deep Belief Network (DBN), to model the genome-wide distribution of reads enrichment. The RBM model can learn a probability distribution over its set of input using the unsupervised contrastive divergence algorithm, which avoiding the assumption of a predefined distribution and achieving high accuracy in peak detection. Based on the identified peaks, a Multilayer Perceptron (MLP) network is applied to derive single-nucleotide-resolution m6A sites from the peak regions. Combining these two deep learning-based frameworks, DeepRIP can perform peak detection over the MeRIP-seq data at a single nucleotide level with high accuracy.

Usage of DeepRIP

This is the main function (also default function) of DeepRIP for the analysis of MeRIP-seq data.

-predict/-h

If user choose -predict, DeepRIP will call peak. And if user choose -h, DeepRIP will print help message of DeepRIP.

-IP

The IP sample file in BAM format. Please make sure your .bam file can be accessed by this file path. **If mutiple replicates are used, the order of IP samples must correspond to the order of Input samples.**

-Input

The Input sample file in BAM format. Please make sure your .bam file can be accessed by this file path. **If mutiple replicates are used, the order of Input samples must**

correspond to the order of IP samples.

-outdir

DeepRIP will use this specified folder to store temp files and will delete these temp files when finished. DeepRIP will save all output files into specified folder for this option.

-n

Specified for the name of the output file. DeepRIP will use this value to create output file like 'Name.tsv'. So please avoid any confliction between these filenames and your existing files.

-prefix

The prefix of temporary files. DeepRIP will use this value to create temp file like "temp.txt". Default Value: "temp".

-fpr

The false positive rate for peak calling process. DeepRIP will compute the cutoff using this false positive rate using Target-Decoy method. The cutoff value will then be used to filter positive hits.

-outType

The format of result file, can be "tsv" or "bed". Default: "tsv".

-RepType

The replicates type, can be "T" or "B". "T" means technical replicates. "B" means biological replicates. DeepRIP will use different algorithm to combine the replicates according to this option.

-preciseSite

When this flag is true, DeepRIP will predict single-nucleotide-resolution m6A sites from peak regions. DEFAULT: false.

-2Bit

The genome file in 2bit format. DeepRIP will use this .2bit file to get potential sequences in peak regions and predict the precise m6A sites in those sequences.

-T

The prediction threshold of single-nucleotide-resolution m6A sites, can be "High", "Medium" or "Low". Default: "Medium".

Example

Call peaks from MeRIP-seq data without replicates

```
java -Xms8g -jar DeepRIPV1.0.jar -predict -IP SRR2120859.bam -Input  
SRR2120862.bam -outdir ~/DeepRIP/ -n result -fpr 0.01
```

Call peaks from MeRIP-seq data with replicates

```
java -Xms8g -jar DeepRIPV1.0.jar -predict -IP SRR2120859.bam, SRR2120857.bam -  
Input SRR2120862.bam, SRR2120860.bam -outdir ~/DeepRIP/ -n result -fpr 0.01 -RepType  
B
```

Predict single-nucleotide-resolution m6A sites from MeRIP-seq data

```
java -Xms8g -jar DeepRIPV1.0.jar -predict -IP SRR2120859.bam,SRR2120857.bam -  
Input SRR2120862.bam,SRR2120860.bam -outdir ~/DeepRIP/ -n result -fpr 0.01 -RepType  
B -preciseSite True -2Bit hg19.2bit -T Low.
```

Note

1.The physical memory should larger than 8g. JVM memory should set larger than 8g.

Example: java -Xms8g

2. By default, DeepRIP will occupy all threads when running. You can restrict your job to specified number of threads by setting the OMP_NUM_THREADS environment variable before starting your job. **Example: export OMP_NUM_THREADS=8**