iNPS: An improved algorithm for accurate nucleosome positioning from sequencing data

User Manual for Version 1.2.2

1 Overview

iNPS is improved from X. S. Liu's NPS algorithm, for high quality nucleosome positioning from MNase-seq data. Our procedure contains the following eight steps. (1) Generate a wave-form nucleosome profile, with the resolution of 10 bp, by extending each tag from the 5' end by 150 bp, and taking the middle 75 bp as the enrichment of nucleosome signal. For paired-end sequencing data, the middle 50% part of each tag is taken as the enrichment of nucleosome signal. (2) Perform Gaussian convolution and first/second/third derivative of Gaussian convolution to smooth the nucleosome profile and find extremum/infection/most-winding points. (3) Distinguish each pair of inflection points as a candidate of "main" nucleosome peak or "shoulder". (4) Determine whether a "shoulder" candidate should be an independent nucleosome, or the dynamic part of the adjacent "main" nucleosome peak. (5) Adjust the inflection borders of the preliminary nucleosome detection. (6) Merge the closely located nucleosome peaks as "doublets". (7) Filter some nucleosome peaks with bad shapes. (8) Perform statistical tests to quantify the confidence level of each nucleosome.

2 Environment

iNPS was developed with python 3.2, so the python 3 environment must be installed under a Linux system.

3 Usage

3.1 Command line:

```
$ python3   iNPS_V1.2.2.py   -i   -o   -c   -l   --s_p
```

3.2 For help, please try:

```
$ python3   iNPS_V1.2.2.py   -h
```

3.3 Arguments for command line:

--version	show program's version number and exit
-h, --help	show help message and exit
-i, --input	"/path/filename" INPUT_FILE file of sequencing tags in a standard BED format (chromosome <tab> start <tab> end <tab> name <tab> score <tab> strand).
-o, --output	"/path/filename" ● Here, the name extension is unnecessary. ● Software will output two result files, "filename_[ChromosomeName].like_b ed" and "filename_[ChromosomeName].like_wig", to record coordinates and profiles of detected nucleosomes respectively. ● The chromosome name will be added as suffix in the file names. ● If your detect nucleosomes on multiple chromosomes, for each chromosome, software will

	output two result files "filename_[ChromosomeName].like_bed" and "filename_[ChromosomeName].like_wig" respectively. And finally, a file "filename_Gathering.like_bed" will gather the detected nucleosomes on every chromosome. ● Note that a path "/path/filename/" or "/path/filename_[ChromosomeName]/" will be built to record the preliminary and intermediate data.
-c, --chrname	● Specify the name (or abbreviation) of the chromosome, if you would like to do nucleosome detection ONLY on ONE single chromosome. ● For nucleosome detection on multiple chromosomes, please do NOT use this parameter. That is, if your do NOT use this parameter, software will detect nucleosome on each chromosome ONE-BY-ONE in the input data as default.
-l, --chrlength	● The length of the chromosome specified by parameter "-c" or "--chrname". ● ONLY used for nucleosome detection on ONE single chromosome (parameter "-c" or "--chrname" is setted). ● If you do NOT use this parameter, software will find the maximum coordinate in the input data to represent the chromosome length as default. ● For nucleosome detection on multiple chromosomes, please do NOT use this parameter. The length of each chromosome will be determined by the tag with maximum coordinate of the corresponding chromosome respectively.
--s_p	● "s" or "p". ● Default = s ● Set to "p" if the input data is paired-end tags. ● Otherwise, set to "s" or use the default setting if the input data is single-end tags.
--pe_max	● The superior limit of the length of paired-end tags. ● Default = 200 ● The tags longer than the cutoff will be ignored. ● This parameter is ONLY available for paired-end sequencing data. ● Please avoid using too large value.
--pe_min	● The inferior limit of the length of paired-end tags. ● Default = 100 ● The tags shorter than the cutoff will be ignored. ● This parameter is ONLY available for paired-end sequencing data. ● Please avoid using too small value.

3.4 Examples.

Taking an example, a file "InputFile.bed" which includes the MNase-seq tags on the whole human genome (coordinate system hg18) is used as input data. Here are the detailed explanations for the following commands and parameter setting.

- **Example 1:**

```
$ python3   iNPS_V1.2.2.py   -i /PathA/InputFile.bed   -o /PathB/Output   -c chr1   -l 247249719
```

Do nucleosome detection ONLY on chromosome 1, as the parameter "-c" has been set to "chr1". And since the "-l" has been set to 247249719, the maximum coordinate of resulted nucleosome profiles will be 247249719. The output files are listed in the following table:

/PathB/Output_chr1.like_bed	Results	Coordinates of detected nucleosomes in chr1
/PathB/Output_chr1.like_wig	Results	Detected nucleosome profiles in chr1
/PathB/Output_chr1/chr1.bed	Intermediate records	MNase-seq tags of chr1, extracted from the input file "InputFile.bed"
/PathB/Output_chr1/InputData_Summary.txt	Intermediate records	Recording the number of tags of chr1, the maximum coordinate among the tags of chr1, and the chromosome length of chr1.

- **Example 2:**

```
$ python3   iNPS_V1.2.2.py   -i /PathA/InputFile.bed   -o /PathB/Output   -c chr1
```

Do nucleosome detection ONLY on chromosome 1, as the parameter "-c" has been set to "chr1". Without "-l" setting, software will use the maximum coordinate of MNase-seq tag of chromosome 1 as the length of chromosome 1. The output files are listed in the following table:

/PathB/Output_chr1.like_bed	Results	Coordinates of detected nucleosomes in chr1
/PathB/Output_chr1.like_wig	Results	Detected nucleosome profiles in chr1
/PathB/Output_chr1/chr1.bed	Intermediate records	MNase-seq tags of chr1, extracted from the input file "InputFile.bed"
/PathB/Output_chr1/InputData_Summary.txt	Intermediate records	Recording the number of tags of chr1, the maximum coordinate among the tags of chr1, and the chromosome length of chr1.

- **Example 3:**

```
$ python3   iNPS_V1.2.2.py   -i /PathA/InputFile.bed   -o /PathB/Output
```

Do nucleosome detection on each chromosome in "InputFile.bed". Software will use the tag with maximum coordinate of each chromosome as the length of the corresponding chromosome respectively. The output files are listed in the following table:

/PathB/Output_chr1.like_bed /PathB/Output_chr2.like_bed … /PathB/Output_chrX.like_bed /PathB/Output_chrY.like_bed	Results	Coordinates, shape properties, and statistical scores of the detected nucleosomes in each of the 24 chromosomes (1 ~ 22, X, and Y) respectively.
/PathB/Output_Gathering.like_bed	Results	Gather the nucleosome information of the 24 "like_bed" files for each of the 24 chromosomes respectively.
/PathB/Output_chr1.like_wig /PathB/Output_chr2.like_wig … /PathB/Output_chrX.like_wig /PathB/Output_chrY.like_wig	Results	Detected nucleosome profiles in each of the 24 chromosomes (1 ~ 22, X, and Y) respectively.
/PathB/Output/chr1.bed /PathB/Output/chr2.bed	Intermediate records	Splitting the input file "InputFile.bed" by chromosomes

… /PathB/Output/chrX.bed /PathB/Output/chrY.bed		
/PathB/Output/InputData_Summary.txt	Intermediate records	Recording the number of tags, the maximum coordinate among the tags, and the chromosome length of each of the 24 chromosomes (1 ~ 22, X, and Y) respectively.

4 Inputs

4.1 Single-end sequencing data.

4.1.1 Input file of **single-end** sequencing tags should be a standard BED format (https://genome.ucsc.edu/FAQ/FAQformat.html), which contains the following 6 columns segregated by <tab>.

chromosome start end name score strand

4.1.2 To have an intuitive look at the BED format, please see the tag coordinate bed files on the webpage (http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcellnucleosomes.aspx). And here is an example fragment.

chr1	121186537	**121186560**	U0	0	**–**
chr1	**223780047**	223780070	U0	0	**+**
chr1	**77322505**	77322528	U0	0	**+**
chr1	173286280	**173286303**	U0	0	**–**
chr1	**51114393**	51114416	U0	0	**+**

Here, not all the information in the table above is necessary. If the sequencing tag is in the forward strand (column 6 is "+"), the coordinate in column 2 is needed, otherwise, if the sequencing tag is in the reverse strand (column 6 is "–"), the coordinate in column 3 is needed.

4.1.3 If your inputting data is incomplete, please make sure that all the data as highlighted in the table above should be kept in the inputting file, and other places in the table could be filled with "None", as shown in the following table.

chr1	None	**121186560**	None	None	**–**
chr1	**223780047**	None	None	None	**+**
chr1	**77322505**	None	None	None	**+**
chr1	None	**173286303**	None	None	**–**
chr1	**51114393**	None	None	None	**+**

4.1.4 Even if you don't know which chromosome these tags belong to, but if you can make sure that all the sequencing tags should be in **ONE** single chromosome, iNPS still can be used for nucleosome detection by inputting data as following table.

None	None	**121186560**	None	None	**–**
None	**223780047**	None	None	None	**+**
None	**77322505**	None	None	None	**+**
None	None	**173286303**	None	None	**–**
None	**51114393**	None	None	None	**+**

4.2 Paired-end sequencing data.

Input file of **paired-end** sequencing tags should be a 3-column BED format, which contains the following 3 columns segregated by <tab>.

chromosome start end

To have an intuitive look at the BED format, please see the example file downloaded from the GEO repository with accession number GSM849959 (ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM849nnn/GSM849959/suppl/GSM849959_GA2807_CMT1_shH2A.Z-2d_MNase_0.1U_r520l2.bed.gz). And here is an example fragment.

chr4	138987819	138987972
chr11	114706061	114706216
chr11	16157850	16158040
chr15	88796655	88796835
chr8	86556663	86556822

5 Outputs

iNPS outputs two result files: ***.like_wig** and ***.like_bed**.

5.1 *.like_wig A result file records nucleosome profiles. There are 7 columns in this file. Users could extract their interesting part and view the profile easily with some software as Microsoft Excel.

Column 1: Coordinate (10bp resolution)
Column 2: Original nucleosome profile
Column 3: Gaussian convolution smoothed profile
Column 4: Laplacian of Gaussian convolution (LoG)
Column 5: Milder LoG with a smaller deviation
Column 6: Tag accumulation
Column 7: Detected peaks

5.2 *.like_bed A result file records detected nucleosome coordinates and the shape properties. There are 10 columns in this file.

Column 1: Chromosome.
Column 2: Coordinate of the beginning inflection boundary of a detected nucleosome.
Column 3: Coordinate of the ending inflection boundary of a detected nucleosome.
Column 4: Nucleosome index number.
Column 5: Length between two inflection points.
Column 6: The peak height of the detected nucleosome.
Column 7: Area under curve.
Column 8: Shape of the detected nucleosome.

“MainPeak”:	an isolated “main” nucleosome peak
“MainPeak+Shoulder”:	a “main” peak associated with a “shoulder”
“MainPeak:doublet”:	a merged “doublet”
“Shoulder”:	an independent “shoulder”

Column 9: “-log10(Pvalue_of_peak)”, the tag enrichment within the peak region
Column 10: “-log10(Pvalue_of_valley)”, the tag depletion within the flanking valley region