PING: Probabilistic Inference for Nucleosome Positioning with MNase-based or Sonicated Short-read Data.

Xuekui Zhang*and Raphael Gottardo†

August 13, 2012

This vignette presents a workflow to use PING on paired-end sequencing data.

Contents

*ubcxzhang@gmail.com
†rgottard@fhcrc.org

1 Licensing and citing

Under the Artistic License 2.0, you are free to use and redistribute this software.

If you use this package for a publication, we would ask you to cite the following:

> Xuekui Zhang, Gordon Robertson, Sangsoon Woo, Brad G. Hoffman, and Raphael Gottardo. (2012). Probabilistic Inference for Nucleosome Positioning with MNase-based or Sonicated Short-read Data. PLoS ONE 7(2): e32095.

2 Introduction

For an introduction to the biological background and PING method, please refer to the PING user guide.

3 PING analysis steps

A typical PING analysis consists of the following steps:

1. Extract reads and chromosomes from bam files.
2. Segment the genome into candidate regions that have sufficient aligned reads via 'segmentPING'
3. Estimate nucleosome positions and other parameters with PING
4. Post-process PING predictions to correct certain predictions

As with any R package, you should first load it with the following command:

```
> library(PING)
```

4 Data Input and Formatting

In order to use the PE version of PING, the input has to be slightly different. Instead of a GRanges object, the new segmentation method use a list of reads and a chromosome.

We provide a dataset for the chromosome I of yeast.

```
> data(yeast_chrI)
> head(reads$P)

           qname pos.+ pos.-
1  6:13194:12920     9   214
2  14:15977:3164     9   214
3 117:4743:11663     9   214
4 24:12054:10535    11   214
5 11:10786:12847    41   179
6  53:15735:7927    41   193

> chrs

[1] "chrI"
```

5 PING analysis

5.1 Genome segmentation

PING is used the same way for paired-end and single-end sequencing data. The function `segmentPING` will decide which segmentation method should be used based on the data type. Paired-end reads should be passed as a list with at least the three elements P, yFm, and yRm. With P being the paired-end reads, yFm and yRm being the reads where one end is missing. When dealing with paired-end data, four new arguments have to be passed to the function: a chromosome chr and three parameters used in candidate region selection: islandDepth, min_cut and max_cut.

In order to improve the computational efficiency of the PING package, if you have access to multiple cores we recommend that you do parallel computations via the `parallel` package. In what follows, we assume that `parallel` is installed on your machine. If it is not, you could omit the first line, and calculations will occur on a single CPU. By default the command is not run. Note that the `segmentPING` and `PING` functions will automatically detect whether you have initialized a cluster and will use it if you have.

```
> library(parallel)

> segPE <- segmentPING(reads, chr = chrs, islandDepth = 3, min_cut = 50,
     max_cut = 1000)
```

It returns a `segReadsListPE` object.

5.2 Parameter estimation

The only difference when using `PING` for paired-end data is the argument PE that has to be set to TRUE.

```
> paraP <- setParaPriorPING(xi = 150, rho = 1.2, alpha = 12, beta = 20000,
     lambda = -6.4e-05, dMu = 200)
> ping <- PING(segPE, paraPrior = paraP, PE = TRUE)
```

The returned object is of class pingList and can be post-processed.

6 Post-processing PING results

Here again, we set the argument PE to TRUE, and use postPING normally.

```
> {
     sigmaB2 = 3600
     rho2 = 15
     alpha2 = 98
     beta2 = 2e+05
 }
> PS = postPING(ping, segPE, paraPrior = paraP, rho2 = rho2, alpha2 = alpha2,
     beta2 = beta2, sigmaB2 = sigmaB2, PE = TRUE)

 The 5 Regions with following IDs are reprocessed for singularity problem:
(208,414]177 (208,414]180 (208,414]192 (208,414]201 (208,414]207
         384          387          399          408          414

 The 62 Regions with following IDs are reprocessed for atypical delta:
[1]  51 390 356 175  21  46

 The 3 Peaks with following IDs are reprocessed for atypical sigma:
[1]   2  93 253

 The 929 regions with following IDs are reprocessed for Boundary problems:
[1]  3  4  6  8  9 10
```

The result output PS is a dataframe that contains estimated parameters of each nucleosome, users can use write.table command to export the selected columns of the result.

```
> head(PS)
```

	ID	chr	w	mu	delta	sigmaSqF	sigmaSqR	se
22201	12	chrI	0.2042955	1270.848	191.3817	2525.083	1500.750	3.440431
22491	38	chrI	0.1645027	13698.663	218.1655	1624.778	2641.707	3.461760
2277	305	chrI	0.1847783	147906.099	205.8027	1979.346	1660.142	3.954364
22021	284	chrI	0.2755595	136799.683	193.7497	1185.758	1085.523	3.364485
22541	38	chrI	0.1162624	14270.464	186.3182	1339.635	1289.626	3.518761
22191	12	chrI	0.1056965	1151.883	190.7205	1154.974	1601.838	4.184562

	score	scoreF	scoreR	minRange	maxRange	seF	seR	rank
22201	253074.6	121509.3	131565.25	672	2307	3.951147	3.997939	1
22491	247208.6	128213.3	118995.32	13112	14788	4.244206	4.578116	2
2277	219554.8	111453.4	108101.39	147090	148499	4.916332	4.785048	3
22021	218716.8	119833.3	98883.44	136235	137314	3.995425	4.163645	4
22541	203632.8	105587.4	98045.44	13112	14788	4.341040	4.223326	5
22191	201956.9	120671.3	81285.54	672	2307	4.821774	5.178364	6