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

Xuekui Zhang*and Raphael Gottardo†

August 30, 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 M of yeast.

```
> data(yeast_chrM)
> head(yeast_chrM$P)
```

```
                qname pos.- pos.+
4059237  120:6253:2074   338   187
4059238  42:9052:11042   313   194
4059239  17:6495:10151   341   209
4059240  81:14542:7245   341   209
4059241 87:14926:13898   341   209
4059242 101:5324:18045   341   209
```

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(yeast_chrM, chr = "chrM", 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.

```
> ping <- PING(segPE, 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, rho2 = rho2, alpha2 = alpha2, beta2 = beta2,
     sigmaB2 = sigmaB2, PE = TRUE)

 The 6 Regions with following IDs are reprocessed for singularity problem:
(0.773,114]80   (114,228]39   (114,228]51   (114,228]79   (114,228]82
          80           153           165           193           196
 (114,228]106
          220

 The 17 Regions with following IDs are reprocessed for atypical delta:
[1] 155 190 129 142  41  37
[1] "No predictions with atypical sigma"

 The 172 regions with following IDs are reprocessed for Boundary problems:
[1]  4  6  7 12 18 20
```

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    score
6831 97 chrM 0.3309686 35700.34 148.9233 1344.1709 1613.7732 9.224526 820464.3
6861 38 chrM 0.2665151 13610.62 149.7794 1294.9684 1173.7339 8.996622 721442.7
6821 97 chrM 0.2721901 35550.59 151.4008 1453.3622 1497.1270 8.296995 707296.8
6851 38 chrM 0.2554825 13457.08 157.8576  957.1654  795.3726 5.717227 693150.9
```

```
694  95 chrM 0.4507928 34780.25 149.3691 1188.2719 1181.5768 9.706183 636567.1
6811 97 chrM 0.2641553 35397.71 154.8389 1352.6040 1506.7203 7.511593 622421.2
       scoreF   scoreR minRange maxRange      seF      seR rank
6831 424378.1 396086.2    34788    36009 9.765269 9.082267    1
6861 339502.5 381940.3    12907    14105 8.988645 9.428187    2
6821 353648.4 353648.4    34788    36009 8.955503 8.068337    3
6851 311210.6 381940.3    12907    14105 6.008960 6.446341    4
694  353648.4 282918.7    34351    35423 9.980666 9.838740    5
6811 311210.6 311210.6    34788    36009 7.965341 7.626689    6
```