BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 16 2013, pages 2049–2050
doi:10.1093/bioinformatics/btt348

Sequence analysis

Advance Access publication June 20, 2013

PING 2.0: an R/Bioconductor package for nucleosome positioning using next-generation sequencing data

Sangsoon Woo[1,†], Xuekui Zhang[2,†], Renan Sauteraud[1], François Robert[3] and Raphael Gottardo[1,*]

[1]Vaccine and Infectious Diseases and Public Health Sciences Divisions, Fred Hutchinson Cancer Research Center, Seattle, WA 98109-1024, USA, [2]Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205, USA and [3]Laboratory of Chromatin and Genomic Expression, Institut de Recherches Cliniques de Montreal, Montreal, QC H2W 1R7, Canada

Associate Editor: Michael Brudno

ABSTRACT

Summary: MNase-Seq and ChIP-Seq have evolved as popular techniques to study chromatin and histone modification. Although many tools have been developed to identify enriched regions, software tools for nucleosome positioning are still limited. We introduce a flexible and powerful open-source R package, PING 2.0, for nucleosome positioning using MNase-Seq data or MNase– or sonicated– ChIP-Seq data combined with either single-end or paired-end sequencing. PING uses a model-based approach, which enables nucleosome predictions even in the presence of low read counts. We illustrate PING using two paired-end datasets from *Saccharomyces cerevisiae* and compare its performance with nucleR and ChIPseqR.

Availability: PING 2.0 is available from the Bioconductor website at http://bioconductor.org. It can run on Linux, Mac and Windows.

Contact: rgottard@fhcrc.org

Supplementary Information: Supplementary material is available at *Bioinformatics* online.

Received on November 15, 2012; revised on May 20, 2013; accepted on June 13, 2013

1 INTRODUCTION

The nucleosome is the basic structural unit of chromatin, which is composed of a nucleosomal core including 147 bp of DNA wrapped around a central histone octamer (H2A, H2B, H3 and H4) and 'linker' DNA connecting nucleosomal core to the next. Nucleosomes control cellular processes by affecting the accessibility of proteins to the DNA, which can act on gene expression and other DNA-dependent processes (Radman-Livaja *et al.*, 2010). Projects like ENCODE (ENCODE Project Consortium, 2011) have generated large amounts of ChIP-Seq and MNase-Seq data that can be used to improve our understanding of nucleosome positioning and its impact on gene regulation. However, these data require efficient tools that can be used regardless of the experimental protocols employed.

Several statistical approaches were recently developed for analyzing nucleosome-based data. Weiner *et al.* (2010) introduced the *Template Filter* method that models forward and reverse reads based on predefined peak shapes. Zhang *et al.* (2008) developed the *Nucleosome Positioning from Sequencing (NPS)* method that uses read pile-ups followed by wavelet de-noising for identifying positioned nucleosomes. More recent pile-up–based approaches include the *ChIPseqR* (Humburg *et al.*, 2011) and *nucleR* (Flores *et al.*, 2011) R packages, available from Bioconductor. In a Bayesian context, Polishko *et al.* (2012) have developed NOrMAL, which uses mixture models to infer nucleosome positions.

All the tools described above were originally developed for single-end (SE) sequencing data, where data provide only one side of DNA fragments, and as such the length of the DNA fragments must be provided as input or estimated from the data. To resolve this ambiguity and give more information, paired-end (PE) protocols have been developed and are gradually being highlighted in experiments (Kent *et al.*, 2010). Because PE data provide important information about DNA fragment length, they can potentially lead to more precise estimates of nucleosome positions (Fullwood *et al.*, 2009), though formal comparisons of nucleome positioning obtained from SE and PE data are still lacking.

Here we present a new software tool for nucleosome positioning, PING 2.0, which extends our probabilistic framework for SE ChIP-Seq and MNase-Seq data (Zhang *et al.*, 2011, 2012). PING 2.0 is a major update of an early (unpublished) version of the software and includes a novel statistical framework for PE data, new functions for pre-processing and reading raw data, and better integration with other Bioconductor packages. The result is a complete framework for nucleosome positioning within Bioconductor that can handle sonicated and MNase protocols combined with either SE or PE sequencing. We compare PING 2.0 with two other Bioconductor packages, nucleR and ChIPseqR using two novel PE *Saccharomyces cerevisiae* datasets (available from GEO: GSE47073) generated specifically for this work.

2 METHODS

Hereafter, we use **bold**face to refer to software packages and `teletype` font to refer to object, classes and functions. For simplicity, we also use the term ChIP-seq to refer to both ChIP-Seq and MNase-Seq.

Architecture: In **PING**, we provide a user-friendly R interface to our underlying C code, which is written for optimal utilization of system resources. **PING** facilitates **parallel** processing for large datasets via the parallel package. Also, **PING** uses the S4 system to define object-oriented classes and methods.

Data input: **PING**'s basic data input is a `GRanges` object containing the directional aligned reads (SE data) or reconstructed DNA fragments (PE data). The `GRanges` class serves as the foundation for representing

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

genomic locations within the Bioconductor project. `GRanges` can be derived from classical formats such as BED and BAM using Bioconductor's infrastructure. For convenience, **PING** includes a `bam2gr` function that can be used to convert BAM objects to a `GRanges` object. `bam2gr` can be applied to both SE and PE data, and in the case of PE data, discards mismatched pairs. We hope to support mismatched reads in a future version of **PING**, as these could provide additional information.

Genome segmentation: Because ChIP-seq data are sparse, consisting of large regions with few or no reads, we first preprocess the data by segmenting the genome into regions, each of which has a minimum number of reads that aligned to the genome of interest. This is done with the `segmentPING` function that scans the genome with a sliding-window approach to identify a set of *candidate* regions; `seg<-segmentPING(gr, PE=TRUE)`, where `gr` is the `GRanges` object including aligned reads. For PE data, the *PE* argument should be set to TRUE, and its value is stored in the result.

Statistical inference: After genome segmentation, the PING model is fit to all regions: `ping<-PING(seg)`. For SE data, the forward and reverse read positions are modeled separately using *t*-distributions. However, for PE data, the paired reads are jointly modeled using a bivariate *t*-distribution; see Supplementary Material. The `PING` function returns nucleosome positions (i.e. their centers), as well as nucleosome enrichment scores (reflecting the number of reads within the nucleosomes) and fuzziness (defined as the standard errors of the estimated centers) (Fig. 1).

Post-processing: Because `PING` is based on a parametric model, departures from the model assumptions, e.g. background noise or repetitive regions, could lead to noisy nucleosome estimates. The `postPING` function was devised to detect and correct such problematic estimates. The function takes both a `PING` result and a segmentation result objects: `psPING<-postPING(ping, seg)`, and returns a data frame containing all estimated parameters for each nucleosome. The `makeRangedDataOutput` function can then convert these results into a `GRanges` object for exporting to the `BED` or `WIG` formats. Post-processing only modifies a small fraction of estimated nucleosomes (<5% for the data used here), most of which have small score. More information about the correction procedure can be found in the Supplementary Material.

Graphical representation: **PING 2.0** provides the `plotSummary` function, for rapid visualization of identified nucleosomes. The function is basically an interface to the `plotTracks` function of the `Gviz` package and takes as input a `PING` result and the `GRanges` used in `segmentPING` to show pile-up profiles along with the predicted nucleosome positions and calculated scores. The user can also include a list of several `PING` results corresponding, for example, to different biological conditions.

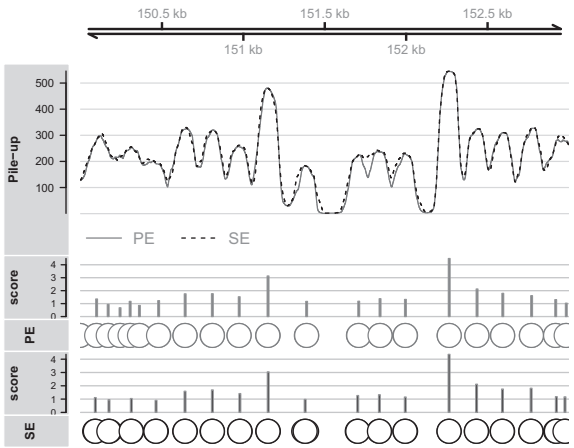

Fig. 1. Nucleosome positions identified by **PING** in a subset of ChrI from our Total MNase data using both SE and PE data. The SE data were generated by simply discarding one PE read at random, and the corresponding pile-up profile was obtained by extended reads to an average length of the PE DNA fragments. The number of nucleosomes estimated from PE is greater, as it provides more detailed information as shown by the PE pile-up profile

3 RESULTS

We applied **PING 2.0** to two PE *S.cerevisiae* datasets generated for this article (Total and H2A.Z MNase) and compared its performance with **nucleR** and **ChIPseqR**. As can be seen in Figure 1, the PE version of `PING` leads to the detection of more nucleosomes that are supported by additional variation in the pile-up profile. However, the SE version also leads to accurate nucleosome positioning. This is confirmed in our comparison study where we show that **PING 2.0** SE-based predictions match their PE-based ones well and better than **nucleR** and **ChIPseqR**. In particular, **nucleR**'s PE predictions appear to be biased toward high-density regions where the algorithms tend to predict many overlapping nucleosomes (Supplementary Figs S3 and S4). In our comparison, we also looked at the stability of the predictions when the datasets are subsampled. Globally, **nucleR** and **PING** lead to stable predictions for both SE and PE, while **ChIPseqR**'s predictions are more noisy and poorly match the results from the full data. See Supplementary Material for more details.

ACKNOWLEDGEMENT

The authors thank Gordon Robertson for helpful discussion.

Funding: NSERC Postdoctoral Fellowship (to X.Z.); National Institutes of Health [R01 HG005692 (to S.W., R.S. and R.G.)].

Conflict of Interest: none declared.

REFERENCES

ENCODE Project Consortium. (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.*, **9**, e1001046.

Flores *et al.* (2011) nucleR: a package for non-parametric nucleosome positioning. *Bioinformatics*, **27**, 2149–2150.

Fullwood *et al.* (2009) Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analysis. *Genome Res.*, **19**, 521–532.

Humburg *et al.* (2011) ChIPseqR: analysis of ChIP-seq experiments. *BMC Bioinformatics*, **12**, 39.

Kent *et al.* (2010) Chromatin particle spectrum analysis: a method for comparative chromatin structure analysis using paired-end mode next-generation DNA sequencing. *Nucleic Acids Res.*, **39**, e26.

Polishko *et al.* (2012) NORMAL: accurate nucleosome positioning using a modified Gaussian mixture model. *Bioinformatics*, **28**, i242–i249.

Radman-Livaja *et al.* (2010) Nucleosome positioning: how is it established, and why does it matter. *Dev. Biol.*, **339**, 258–266.

Weiner *et al.* (2010) High-resolution nucleosome mapping reveals transcription-dependent promotor packaging. *Genome Res.*, **20**, 90–100.

Zhang *et al.* (2008) Identifying positioned nucleosomes with epigenetic marks in human from ChIP-Seq. *BMC Genomics*, **9**, 537.

Zhang *et al.* (2011) PICS: probabilistic inference for ChIP-seq. *Biometrics*, **67**, 151–163.

Zhang *et al.* (2012) Probabilistic inference for nucleosome positioning with MNase-based or sonicated short-read data. *PLoS One*, **7**, e32095.