BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 15 2011, pages 2149–2150
doi:10.1093/bioinformatics/btr345

Genome analysis

Advance Access publication on June 7, 2011

nucleR: a package for non-parametric nucleosome positioning

Oscar Flores[1] and Modesto Orozco[1,2,3,*]

[1]IRB-BSC Joint Research Program on Computational Biology, Institute of Research in Biomedicine, Baldiri i Reixac 10, [2]Department of Biochemistry and Molecular Biology, University of Barcelona, Avinguda Diagonal 645 and [3]Instituto Nacional de Bioinformática. Parc Científic de Barcelona. Baldiri i Reixac 10, Barcelona 08028, Spain

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: nucleR is an R/Bioconductor package for a flexible and fast recognition of nucleosome positioning from next generation sequencing and tiling arrays experiments. The software is integrated with standard high-throughput genomics R packages and allows for *in situ* visualization as well as to export results to common genome browser formats.

Availability: Additional information and methodological details can be found at http://mmb.pcb.ub.es/nucleR

Contact: modesto.orozco@irbbarcelona.org

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on March 15, 2011; revised on April 26, 2011; accepted on June 1, 2011

1 INTRODUCTION

Eukaryotic chromatin is organized in nucleosomes, a structure with approximately 147 basepairs (bp) of DNA wrapped around an octamer of histones (Jiang and Pugh, 2009). Nucleosomes affect the packaging and accessibility of DNA, thus playing a crucial role in defining its functionality (Jiang and Pugh, 2009). The development of high-throughput techniques, such as tiling arrays (TA) and next generation sequencing (NGS), coupled to Micrococal Nuclease (MNase) digestions has enabled the study of nucleosome positioning at the entire genome level for several organisms including humans (Jiang and Pugh, 2009). Some clear features emerged from these studies, such as the presence of well-positioned nucleosomes and their depletion in regions surrounding transcription start sites (TSS; Jiang and Pugh, 2009). However, well-positioned (phased across different cells) nucleosomes coexist with fuzzy (not-phased) ones outside the TSS. This variability makes nucleosome-positioning complex, requiring therefore the development of algorithms to find the 'most-probable' nucleosomal configuration. These approaches include, among others, Hidden Markov Models (HMM) (Lee *et al*., 2007; Yassour *et al*., 2008; Yuan *et al*., 2005), higher order Bayesian Networks (Chen *et al*., 2010) or mixed methods (Di Gesù *et al*., 2009).These methods are very powerful, but the intrinsic assumptions and the level of expertise of the modeler can significantly affect the results.

Here we present a new tool, nucleR, integrated in the open source, multiplatform R/Bioconductor framework. The approach is based on a fast, nonparametric detection of all nucleosome dyads and scoring of the calls. A good performance is achieved by filtering the noise using Fast Fourier Transform (FFT). The user has full freedom to export, select, merge or process suggested nucleosome calls in any desired way, making the method completely flexible. Algorithms presented here are suitable for most TA and single or paired ended NGS platforms.

*To whom correspondence should be addressed.

2 METHODS

nucleR's workflow is presented in Figure 1a. It relies the low-level processing of the genomic data to specialized R/Bioconductor packages, allowing for a wide variety of input formats.

The first step is to convert the input data to obtain 1bp resolution hybridization fluorescence ratios (TA) or short reads coverage (NGS). In the latter case, some extra manipulations are applied to the reads, like correcting the strand bias if working with single-ended sequencing or trimming the reads for remarking the position of the dyad in paired-end cases. Additionally, to reduce any potential bias due to the sequence preferences of MNase (Deniz *et al*., in press), coverage maps obtained from nucleosomal DNA can be easily corrected with those obtained in a parallel experiment for naked DNA. For TA, the main problem is the existence of DNA segments not covered by a probe, which in our procedure are inferred from neighboring probes.

The next step is 'profile cleaning' based on Fourier analysis, which simultaneously smoothes the signal and cleans the distortions in the coverage peaks. Noise removal from coverage profile is performed following signal theory (Smith, 1999). Accordingly, the original complex signal is described as a combination of simple periodic waves. By transforming the original profile into the Fourier Space using FFT, one can analyze the power spectrum of single frequencies, i.e. the contribution of every frequency to the original signal. High frequencies are usually echoes of lower frequencies and are sources of noise. They can therefore be removed without affecting the final profile (Smith, 1999). In our case, a small number of components are chosen depending on the nature of the experiment (typically 1% for TA and 2% for NGS; see Supplementary Material) and the rest are knockedout before performing the inverse FFT; see Figure 1b for an example of raw and filtered profiles. The following step is the detection of nucleosome dyad, which is done using a simple local maxima search and is largely facilitated by the clarity of the filtered profile. Nucleosome calls are determined by selecting the surrounding bases around the dyad position, and are scored based on the height and sharpness of the peak; giving high score to large and sharp peaks and penalizing fuzziness (Fig. 1b).

Once nucleosome calling and scoring is done, the user can manipulate the calls with standard R/Bioconductor tools to select, merge or perform further study on nucleosome positioning in a way that fulfills his/her specific needs. Methods for visualizing the results and exporting the data in BED and WIG formats are also provided. The nucleR package has been created to manipulate large datasets and offers an efficient usage of FFT (see Supplementary Material) and support for parallel processing in multicore machines.

3 RESULTS

In order to illustrate the performance of nucleR, we have analyzed two datasets derived from MNase treatment of yeast chromatin: TA

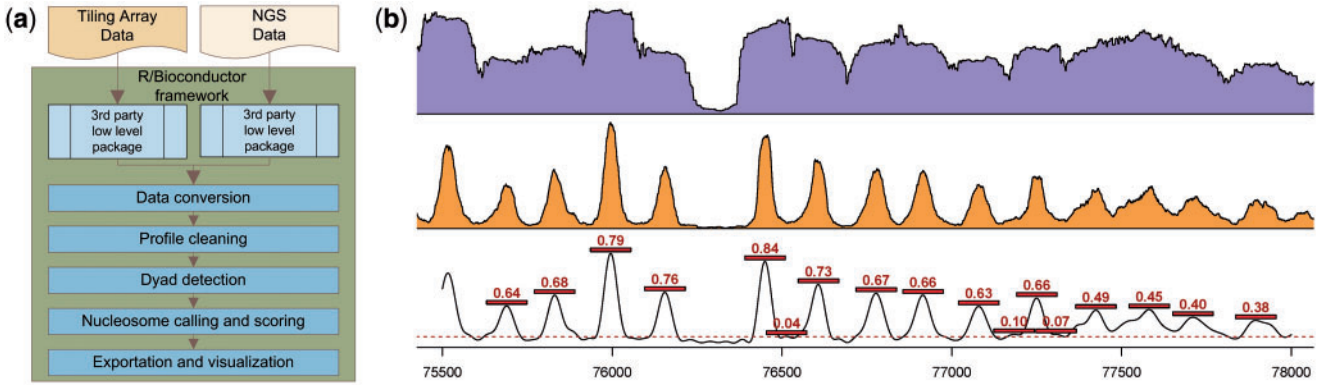

Fig. 1. (**a**) nucleR workflow diagram. (**b**) Top: raw coverage from NGS paired-end reads; middle: coverage using trimmed reads, remarking the dyad location; bottom: filtered trimmed coverage and scored nucleosome positions. Detection threshold is marked as a red dashed line.

from Nislow's group (Lee *et al*., 2007) and an NGS experiment performed in our group (Deniz *et al*., in press). A comparative analysis has been performed using HMM results provided by Lee *et al*. (2007) and the package ChIPSeqR, available from R/Bioconductor repository. We selected the first method as it is widely used in the literature and the second one for being the only package that enables a nucleosome positioning analysis in R/Bioconductor.

The main difference between Lee's HMM and our approach is that the method presented here is able to detect multiple shifted nucleosomes in a given single position, and not just providing the 'most probable' state for each position. This has a large impact on the final map, where we are able to identify a richer landscape of nucleosome calls in TA experiments. Additionally, the HMM-based approach is difficult to apply to 1-bp resolution experiments, such as NSG, due to the large amount of memory required for backtracking. Apart from this scalability problem, the modeling of background transition probabilities, a key element in HMM, requires a very fine and subjective tuning. This can overconstrain the results, forcing, for example, preferred length linker DNA, strict periodic positioning or inability to detect coverage peaks due to strange chromatin structures, like centromers or tetrasomes. In the previous work of Lee *et al*., 70 873 nucleosomes (40 096 well positioned and 30 777 fuzzy) were detected. nucleR applied to the same dataset is able to detect a total of 151 882 individual scored nucleosome calls. Furthermore, our method displayed a larger ability to detect nucleosome positions in synthetic data (77% hits for nucleR versus 67% for HMM) and also a higher correlation when rebuilding the original read maps obtained from the TA experiment ($P=0.63$ for nucleR versus $P=0.38$ for HMM) (see Supplementary Material).

ChIPSeqR (www.bioconductor.org) provides a similar approach as the one presented here, but lacks support for TA data and a method for noise removal. This generates problems in peak detection since only global maxima above a settled threshold are detected efficiently, missing many relevant sub-peaks (local maximums) and leading to an underestimation of nucleosome density. In our NGS data nucleR detects 100 335 nucleosome calls (repeated genomic regions were not considered for this analysis), comprising all of the local maximums above the default threshold on the smoothed signal. With the same detection threshold, ChIPSeqR detects by default 57 725 nucleosome binding sites and 2 206 180 if asking for sub-peak detection, very far from the expected magnitude of 10^4-10^5 calls, according to yeast genome size and putative nucleosome length. A visual comparison of the mentioned methods is available as Supplementary Material (see Supplementary Fig. S1). Again, a benchmark analysis showed that nucleR was able to recover more nucleosome positions in synthetic data (95% for nucleR versus 81% for ChIPseqR) and to reproduce with higher definition the experimental coverage map ($P=0.29$ for nucleR versus $P=0.03$ for ChIPSeqR) (see Supplementary Material for details).

nucleR has a computational complexity between $O(N)$ and $O(N \log N)$. The package is accessible free of cost under LGPL-3 license scheme from our website (http://mmb.pcb.ub.es/nucleR) and the Instituto Nacional de Bioinformática site (http://www.inab.org) It should also be available on Bioconductor upon publication.

ACKNOWLEDGEMENTS

We thank Carles Fenollosa for testing the software and Özgen Deniz for the experimental support.

Funding: Spanish Ministry of Science and Innovation (BIO2009-10964 and Consolider E-Science); Instituto de Salud Carlos III (INB-Genoma España and COMBIOMED RETICS); Fundación Marcelino Botín.

Conflict of Interest: none declared.

REFERENCES

Chen,X. *et al*. (2010) A dynamic Bayesian network for identifying protein-binding footprints from single molecule-based sequencing data. *Bioinformatics*, **26**, i334–i342.

Deniz,Ö. *et al*. (2011) Physical properties of naked DNA signal gene regulatory regions. (in press).

Di Gesù,V. *et al*. (2009) A multi-layer method to study genome-scale positions of nucleosomes. *Genomics*, **93**, 140–145.

Jiang,C. and Pugh,B.F. (2009) Nucleosome positioning and gene regulation: advances through genomics. *Nature Rev. Genet.*, **10**, 161–172.

Lee,W. *et al*. (2007) A high-resolution atlas of nucleosome occupancy in yeast. *Nature Genet.*, **39**, 1235–1244.

Smith,S.W. (1999) *The Scientist and Engineer's Guide to Digital Signal Processing*, 2nd edn. California Technical Publishing, San Diego, CA, USA.

Yassour,M. *et al*. (2008) Nucleosome positioning from tiling microarray data. *Bioinformatics*, **24**, i139–i146.

Yuan,G.-C. *et al*. (2005) Genome-scale identification of nucleosome positions in S.cerevisiae. *Science*, **309**, 626–630.