The development of high-throughput techniques, such as tiling and 3

1 Instituto Nacional de Bioinformática. Parc Científic de Barcelona. Baldrí i Reixac 10, Barcelona 08028, Spain

...using Fast Fourier Transform (FFT). The user has full freedom to...

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 smooths 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 knocked out 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
for sub-peak detection, very far from the expected magnitude

default 57 725 nucleosome binding sites and 2 206 180 if asking

signal. With the same detection threshold, ChIPSeqR detects by

the local maximums above the default threshold on the smoothed

regions were not considered for this analysis), comprising all of

efficiently, missing many relevant sub-peaks (local maximums) and

since only global maxima above a settled threshold are detected

method for noise removal. This generates problems in peak detection

as the one presented here, but lacks support for TA data and a

work of Lee

chromatin structures, like centromers or tetrasomes. In the previous

positioning or inability to detect coverage peaks due to strange

forcing, for example, preferred length linker DNA, strict periodic

background transition probabilities, a key element in HMM, requires

backtracking. Apart from this scalability problem, the modeling of

such as NSG , due to the large amount of memory required for

based approach is difficult to apply to 1-bp resolution experiments,

of nucleosome calls in TA experiments. Additionally, the HMM-

the final map, where we are able to identify a richer landscape

‘most probable’ state for each position. This has a large impact on

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 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 10^5 – 10^9 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.imab.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


footprints from single-molecule sequencing data. Bioinformatics, 26, i334–i342.

Deniz,O. et al. (2011) Physical properties of naked DNA signal gene regulatory regions.

(in press).

Di Gesu,V. et al. (2009) A multi-layer method to study genome-scale positions of

nucleosomes. Genome, 93, 140–145.


Genet., 39, 1235–1244.


Bioinformatics, 24, i339–i346.

Yuan,G.-C. et al. (2005) Genome-scale identification of nucleosome positions in