The development of high-throughput techniques, such as tiling using Fast Fourier Transform (FFT). The user has full freedom to configure. These approaches include, among others, Hidden Markov Models (HMM) (Lee et al., 2007; Yassour et al., 2009; Yuan et al., 2010), higher order Bayesian Networks (Chen et al., 2007; Yassour et al., 2009) 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, multipurpose 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.

2 METHODS

nucleR’s workflow is presented in Figure 1a. It relies on 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 removing 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 knocked out before performing the inverse FFT. In the case of raw and filtered profiles. The following step is the detection of nucleosome dye, 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...
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 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


