Bioinformatics, 32(16), 2016, 2528–2530
doi: 10.1093/bioinformatics/btw206
Advance Access Publication Date: 22 April 2016
Applications Note

OXFORD

Sequence analysis

Pasha: a versatile R package for piling chromatin HTS data

Romain Fenouil[1,*,†], Nicolas Descostes[2,†], Lionel Spinelli[3,†], Frederic Koch[4], Muhammad Ahmad Maqbool[5], Touati Benoukraf[6], Pierre Cauchy[7], Charlène Innocenti[8], Pierre Ferrier[3] and Jean-Christophe Andrau[5,*]

[1]Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA, [2]Department of Biochemistry and Molecular Pharmacology, Howard Hughes Medical Institute, New York University Langone School of Medicine, New York, NY, USA, [3]Centre D'Immunologie De Marseille-Luminy, Aix Marseille Université UM2, Inserm, U1104, CNRS UMR7280, 13288 Marseille, France, [4]Department of Developmental Genetics, Max Planck Institute for Molecular Genetics, Berlin 14195, Germany, [5]Institute of Molecular Genetics of Montpellier (IGMM), UMR5535 CNRS, 34293 Montpellier, France, [6]Cancer Science Institute of Singapore, National University of Singapore, Singapore, Singapore, [7]Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham B15 2TT, UK and [8]CHU Montpellier, INSERM U1203, Institute of Regenerative Medicine and Biotherapy, Montpellier, France

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.
Associate Editor: John Hancock

Received on December 18, 2015; revised on April 7, 2016; accepted on April 8, 2016

Abstract

Summary: We describe an R package designed for processing aligned reads from chromatin-oriented high-throughput sequencing experiments. Pasha (preprocessing of aligned sequences from HTS analyses) allows easy manipulation of aligned reads from short-read sequencing technologies (ChIP-seq, FAIRE-seq, MNase-Seq, ...) and offers innovative approaches such as ChIP-seq reads elongation, nucleosome midpoint piling strategy for positioning analyses, or the ability to subset paired-end reads by groups of insert size that can contain biologically relevant information.
Availability and implementation: Pasha is a multi-platform R package, available on CRAN repositories under GPL-3 license (https://cran.r-project.org/web/packages/Pasha/).
Contacts: rfenouil@gmail.com or jean-christophe.andrau@igmm.cnrs.fr
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Transcription regulation analyses benefited from the transition to High Throughput Sequencing (HTS) technology, which offers an unprecedented coverage by sequencing a large number of small sequences in parallel. However, while library preparation protocols recommend the selection of sheared DNA fragments ranging from less than 100 bp to more than 500 bp, the read size generated by the sequencing process is often shorter (25–100 bp). As a consequence, typical single-end sequencing reads will display a gap between enrichments on both strands after alignment to the reference genome. In most analyses scenarios, peak detection algorithms are making use of the lag between the strands to infer the theoretical center of enrichment (Liu, 2014; Valouev *et al.*, 2008). However, while the localization of binding events is a crucial step for analyzing stable binding factors in ChIP-seq experiments, an additional layer of qualitative and quantitative analyses are often required for an exhaustive description of transient or dynamic chromatin association (e.g. RNA Polymerase). To characterize such profiles, it is critical to

restore the missing coverage information from read sequences in order to preserve the qualitative and quantitative aspect of the data. Here, we describe a piling tool that addresses this question by transforming aligned reads to piled enrichment scores. The modular structure of this pipeline makes it adaptable to most experimental scenarios for chromatin-oriented analyses, as well as short RNA sequencing. Pasha pipeline is for instance able to process seamlessly paired-ends or single-ends dataset by loading appropriate module, and allows the user to focus on biologically relevant parameters. Importantly, a versatile set of options decorates the core pipeline and enables innovative approaches for downstream analyses, such as an alternative midpoint piling strategy often used for nucleosome positioning analyses, or the ability to subset paired-end reads by groups of insert size for a finer analysis of fragments population. Finally, in order to maximize the compatibility with public tools and facilitate the integration to existing pipelines, a broad range of input and output file formats were implemented (Supplementary Fig. 1).

2 Description

2.1 Elongating from the edges

We developed an original method to estimate the original DNA fragments size from the observed lag between both strands. Because it relies on a simple multiplicative operation on piled scores, most represented fragment sizes are given more importance during elongation estimation, which confers a significant advantage for samples displaying a vast heterogeneity in fragments size as compared to concurrent techniques often based on cross-correlation.

By applying this method, and adjusting the coordinates of the region covered by reads, we were able to restore the crucial information required for an accurate representation of reads coverage (Supplementary Fig. 2).

2.2 Paired-ends and orphan reads

Alternative paired-end sequencing technique waives the requirement for elongation estimation by performing the simultaneous sequencing of both ends of each DNA fragment. Their subsequent alignment to the reference genome provides direct information about the initial fragment length. In some cases—because of insufficient sequence quality, insertions/deletions events, or repeated regions on the genome—one read from a pair fails to align to the reference genome. These broken pairs can represent a considerable amount of information that is lost when the orphan reads are discarded. In an attempt to provide a thorough handling of paired-end experiments to our pipeline, we developed additional options to rehabilitate orphan reads by elongating them to the median fragment size represented by the remainder of the sample.

2.3 Midpoint piling

A common theme in nucleosomal studies (MNase-seq) is to represent data as midpoint to emphasize positioning. Whereas densities (generated by tag elongation) will allow better characterization of nucleosome apparent depletion, midpoints will give a better view of the precise positioning of nucleosome more specifically at promoters (Pugh, 2010; Seila *et al.*, 2008; Weiner *et al.*, 2010). This type of analysis reduces reads coverage to a single base footprint, centered on the theoretical region covered by the initial DNA fragment. While this compromises the quantitative and qualitative aspect of the observations, it allows a more resolutive outlook for factors with a high density of binding events, and is particularly suited for analyzing binding patterns in large scale analyses (e.g. average profiles).

This method is available as an alternative piling method in our package and provides additional options (summarized in Supplementary Fig. 3) such as the possibility to highlight borders of binding events rather than their centered position, as seen in (Schones *et al.*, 2008).

2.4 MultiReads

Some genomic regions (eventually large, e.g. telomeric and centromeric) are composed of repeated sequences that can interfere with the assignment of unambiguous coordinates during alignment. These reads are often discarded although they can sum a precious amount of information. Our pipeline integrates specific strategies designed to rehabilitate signal in these regions: one affects a weight divided by the number of candidate positions for each read, while the other iterative algorithm preferentially affect reads to positions showing surrounding enrichments (Chung *et al.*, 2011). In both cases, the complexity threshold is determined by adjusting the maximal number of candidate regions reported during the alignment step. Although their use prevents a strict quantitative comparison between genomic regions, these methods have proven to highlight some signal on relevant repeated genomic features (i.e. miRNAs, snRNAs; Supplementary Fig. 4) and centromeric and telomeric regions.

2.5 Size segregation

Because chromatin fragmentation process is likely to be affected by biological factors (e.g. MNase treatment is less efficient where nucleosomes are bound to DNA), it is expected that subpopulations of reads in the fragmented population can reflect relevant biological features. This approach has been used in several studies coupled with MNase-Seq approach to probe for subnucleosomal structures linked to chromatin (Henikoff *et al.*, 2011).

Our pipeline allows for isolation of fragments size subpopulations in paired-end samples and automatically generates reports for specified ranges, possibly revealing relevant biological information in ChIP-seq experiments (Supplementary Fig. 5). Other specific protocols targeting small populations of RNAs (Core *et al.*, 2008; Fenouil *et al.*, 2012; Seila *et al.*, 2008) can also benefit from this option as RNA populations can be are characterized by their size (ie. MicroRNAs, tRNAS).

3 Conclusion

The described pipeline is a versatile tool that summarizes chromatin-oriented HTS data to enrichment scores using various strategies. It integrates several options allowing a seamless adaptation to various experimental setups. Additionally, the R package provides several tools for programmers that need to develop or integrate additional features.

Funding

Work in the JCA laboratory is supported by Centre National de la Recherche Scientifique (CNRS), Agence Nationale de la Recherche (ANR), Institut National du Cancer (INCa) and Commission of the European Communities. ND was supported by grant from the Ligue Nationale contre le Cancer and MAM by a grant from the ANR iSPLICE ANR-11-BSV8-0013.

Conflict of Interest: none declared.

References

Chung,D. *et al.* (2011) Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-Seq data. *PLoS Comput. Biol.*, 7, e1002111.

Core,L.J. *et al*. (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, **322**, 1845–1848.

Fenouil,R. *et al*. (2012) CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. *Genome Res*., **22**, 2399–2408.

Henikoff,J.G. *et al*. (2011) Epigenome characterization at single base-pair resolution. *Proc. Natl. Acad. Sci. USA*, **108**, 18318–18323.

Liu,T. (2014) Use model-based Analysis of ChIP-Seq (MACS) to analyze short reads generated by sequencing protein-DNA interactions in embryonic stem cells. *Methods Mol. Biol*., **1150**, 81–95.

Pugh,B.F. (2010) A preoccupied position on nucleosomes. *Nat. Struct. Mol. Biol*., **17**, 923.

Schones,D.E. *et al*. (2008) Dynamic regulation of nucleosome positioning in the human genome. *Cell*, **132**, 887–898.

Seila,A.C. *et al*. (2008) Divergent transcription from active promoters. *Science*, **322**, 1849–1851.

Valouev,A. *et al*. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods*, **5**, 829–834.

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