Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Application Note

OXFORD

Genome Analysis

epic2 efficiently finds diffuse domains in ChIP-Seq data

Endre Bakken Stovner [1,2,3,4], Pål Sætrom [1,2,3,4]*

[1]Department of Computer Science, Norwegian University of Science and Technology, Trondheim, 7013, Norway,
[2]Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, 7013, Norway
[3]Bioinformatics Core Facility, Norwegian University of Science and Technology, Trondheim, 7013, Norway
[4]K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, Norwegian University of Science and Technology, Trondheim, Norway

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: As the size and number of ChIP-Seq experiments quickly grow, faster methods to find peaks in ChIP-Seq data are required. The SICER ChIP-Seq caller has proven adept at finding diffuse domains in ChIP-Seq data, but it is slow, requires much memory, needs manual installation steps and is hard to use. epic2 is a complete rewrite of SICER that is focused on speed, low memory overhead and ease-of-use.
Availability: epic2 is freely available from https://github.com/biocore-ntnu/epic2
Contact: paalsat@gmail.com

1 Introduction

ChIP-Seq (Park (2009)) is an experimental method used to analyze how a protein of interest interacts with DNA. The end result of a ChIP-Seq experiment is a pool of DNA fragments that bind to the protein. These fragments must be aligned and the alignments further analyzed to find so-called enriched regions - the genomic regions where the protein binds.

Different algorithms are needed depending on the characteristics of the protein. Transcription factors (TFs) have distinct binding sites resulting in short and distinct peaks in the ChIP-seq data, whereas histone modifications such as histone 3 lysine 27 tri-methylation (H3K27me3) occur over longer regions resulting in diffuse and rolling hills-like signals in the data.

MACS2 (Zhang *et al*. (2008)) is a popular ChIP-Seq caller that uses the expected shift between peaks aligning on the Watson and Crick strand to identify short and punctuate to medium-size peaks. The MACS2 software also has an option for finding broad peaks, but this option merely links together punctuate peaks, which are not necessarily found in ChIP-Seq types with diffuse signals

To identify such diffuse ChIP-seq signals, SICER (Zang *et al*. (2009)) collects all ChIP bins that pass a score threshold for enrichment, and then merges these bins into regions. If the region score is higher than a threshold computed to control the statistical significance of regions, it becomes a candidate region. Whether this region is truly enriched is assesed by comparing the number of ChIP reads in the region to the number of background (input) reads. The resulting p-values are finally false discovery rate (FDR) controlled to adjust for multiple testing.

Benchmarks have shown that SICER is one of the best tools for finding diffuse ChIP-seq signals (Steinhauser *et al*. (2016)); however, the SICER software is cumbersome, slow, and has high memory requirements, making SICER impractical for large-scale data analyses. To address these shortcomings, we have created epic2, which is a complete reimplementation of SICER. The epic2 software is about 30 times faster and uses less than 1/7 of the memory of SICER on relevant genome-scale ChIP-seq data.

2 Improvements and new features

The SICER algorithm's memory requirements are due to binning the genome and counting the number of reads per bin, whereas its running time depends on the number of input reads and genomic bins. To improve on the original Python implementation of SICER, we decided to use Cython, as this language is compatible with Python 2 and 3, gives the running time performance of compiled languages, and provides additional data type control. Specifically, whereas Python lists of integers use about 8 unaligned bytes per element on 64 bit CPUs, strong and distinct peaks in ChIP-seq experiments typically have read depths of less than 10000 Rye *et al*. (2011). Our Cython implementation therefore uses 16 bit integers for storing bin counts. As the majority of the runtime is used to read data into

memory, the parsers for supported input file formats are written in Cython and C++. In addition, we have arranged the data contiguously to ensure memory-locality and fast iteration.

We benchmarked our epic2 implementation on an in-house dataset of H3K27me3 ChIP-seq data. Varying the amount of input data revealed that epic2 was up to 32 and 6 times faster than SICER and MACS2, respectively (Fig. 1A). Moreover, epic2 used less than 1/7 and 1/2 as much memory as SICER and MACS2 (Fig. 1B).

In addition to these performance improvements, we made our implementation easy to install and use from the command line on single and paired-end data in bam, sam and both regular and gzipped bed and bedpe file formats. Furthermore, we have added new features that makes epic2 easy to use both with existing genomes and custom genomes and assemblies. For example, epic2 detects the readlength, by which it can automatically choose the appropriate precomputed effective genome fraction for the genome chosen. A complete list of new features in epic2 is available at https://github.com/biocore-ntnu/epic2

3 Conclusion

epic2 is a fast, low-memory, easy to use and install reimplementation of the extremely popular SICER ChIP-Seq caller. As ChIP-Seq is a fundamental technology for investingating epigenetic marks we expect epic2 to be of great use for researchers.

References

Steinhauser, S. and Kurzawa, N. and Eils, R. and Herrmann, C. (2016) A comprehensive comparison of tools for differential ChIP-seq analysis, *Briefings in Bioinformatics*, **6**, 953-966.

Zang, C., Schones D. E., Zeng C., Cui K. Zhao, K. and Peng W. (2009), A clustering approach for identification of enriched domains from histone modification ChIP-Seq data, *Bioinformatics*, **25**, 1952-1958

Zhang Y., Liu T., Meyer C. A., Eeckhoute J., Johnson, D. S., Bernstein B. E., Nusbaum C., Myers Richard M., Brown M., Li W., Liu X. S. (2008), Model-based Analysis of ChIP-Seq (MACS), *Genome Biology*, **9**, R137

Park, Peter J. (2009), ChIP-Seq: advantages and challenges of a maturing technology, *Nature Reviews Genetics*, **10**, 669EP

Rye M. B. and Sætrom P., and Drabløs F. (2011), A manually curated ChIP-seq benchmark demonstrates room for improvement in current peak-finder programs, *Nucleic Acids Research*, **339**, e25