Bioinformatics, 35(21), 2019, 4392–4393
doi: 10.1093/bioinformatics/btz232
Advance Access Publication Date: 28 March 2019
Applications Note

Genome analysis

epic2 efficiently finds diffuse domains in ChIP-seq data

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

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

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on December 4, 2018; revised on March 11, 2019; editorial decision on March 22, 2019; accepted on March 26, 2019

Abstract

Summary: Data from chromatin immunoprecipitation (ChIP) followed by high-throughput sequencing (ChIP-seq) generally contain either narrow peaks or broad and diffusely enriched domains. 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 and implementation: The MIT-licensed code is available at https://github.com/biocore-ntnu/epic2.

Contact: endrebak85@gmail.com

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

1 Introduction

Chromatin immunoprecipitation (ChIP) followed by high-throughput sequencing (ChIP-seq; Park, 2009) is an experimental method for the genome-wide identification of genomic sites that interact with a given target protein. In brief, DNA is cross-linked to interacting proteins and fragmented, the target protein and its cross-linked DNA fragments are enriched by using an antibody specific to the target protein, the DNA fragments are sequenced, and the resulting sequence reads are aligned to the target genome. Depending on the characteristics of the target protein, different algorithms are then needed to identify the protein's interaction sites in the genome. Transcription factors have distinct binding sites resulting in narrow peaks of reads when the ChIP-seq data are aligned to the genome. In contrast, histone modifications such as histone 3 lysine 27 trimethylation (H3K27me3) occur over longer regions resulting in diffuse signals in the aligned data.

MACS2 (Zhang *et al.*, 2008) is a popular ChIP-seq caller used to identify short- to medium-size peaks. The MACS2 software has an option for finding broad peaks, but this option merely links together narrow peaks, which are not necessarily found in ChIP-seq data with diffuse signals. To identify such diffuse ChIP-seq signals, SICER (Zang *et al.*, 2009) collects all genomic bins that pass a score threshold for ChIP-seq read enrichment, and then merges these bins into regions. If a region's score is higher than a threshold computed to control the statistical significance of regions, the region becomes a candidate region. Whether this region is truly enriched is assessed by comparing the number of ChIP reads in the region to the region's number of background (input) reads, which are determined by sequencing the DNA fragments used as input to the antibody-based enrichment. Finally, the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995) is used to adjust *P*-values 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.

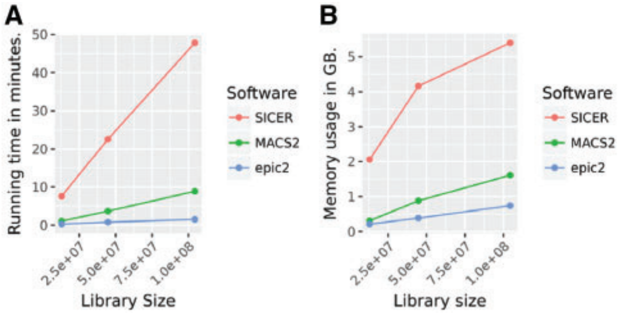

Fig. 1. (**A**) Running time and (**B**) memory usage for SICER, MACS2 and epic2 on increasing dataset sizes. The results are from a Linux server running on a 48 core virtual machine with 192 gigabytes of RAM

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 numbers 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 eight unaligned bytes per element on 64-bit CPUs, strong and distinct peaks in ChIP-seq experiments typically have read depths of less than 10 000 (Rye *et al.*, 2011). Our Cython implementation therefore uses 16 bit integers for storing bin counts, which means that each bin at most can represent 65 536 reads. 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's memory requirements were less than 1/7 and 1/2 of SICER's and MACS2's requirements, respectively (Fig. 1B). In addition to these performance improvements, we made our implementation easy to install and use from the command line, and we added support for both single and paired-end data in BAM, SAM, and regular and compressed (GZIP) BED and BEDPE file formats. Furthermore, we have added new features that make epic2 easy to use both with existing genomes and custom genomes and assemblies. For example, epic2 detects the library's read length and automatically chooses the appropriate precomputed effective genome fraction for the user's specified genome. To ensure the correctness of epic2, we created a Snakemake pipeline (Koster and Rahmann, 2012) to compare the results of SICER and epic2. When tested on three publicly available datasets (Pena-Diaz *et al.*, 2013; Roadmap Epigenomics Consortium *et al.*, 2015) the SICER and epic2 results were essentially identical (Supplementary analysis); the differences were due to a bug where the original SICER implementation includes reads mapping beyond the chromosome borders. This bug potentially increases the total number of reads recognized by SICER and thereby affects all computed *P*-values, as these are scaled by the number of ChIP reads divided by the number of input reads. If run with the same bug (option "–original-algorithm"), epic2 produces the same results as SICER.

The epic2 software and detailed documentation of its features are available at https://github.com/biocore-ntnu/epic2. epic2 is also available in the Bioconda channel (Gruning *et al.*, 2018) of the Conda package manager.

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 investigating epigenetic marks we expect epic2 to be of great use for researchers.

Funding

This work was supported by the Research Council of Norway [grant number 230338]; and Stiftelsen K.G. Jebsen.

Conflict of Interest: none declared.

References

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.

Gruning,B. *et al.* (2018) Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods*, **15**, 475–476.

Koster,J. and Rahmann,S. (2012) Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522.

Park,P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–680.

Pena-Diaz,J. *et al.* (2013) Transcription profiling during the cell cycle shows that a subset of Polycomb-targeted genes is upregulated during DNA replication. *Nucleic Acids Res.*, **41**, 2846–2856.

Roadmap Epigenomics Consortium *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.

Rye,M.B. *et al.* (2011) A manually curated ChIP-seq benchmark demonstrates room for improvement in current peak-finder programs. *Nucleic Acids Res.*, **39**, e25.

Steinhauser,S. *et al.* (2016) A comprehensive comparison of tools for differential ChIP-seq analysis. *Brief. Bioinform.*, **17**, 953–966.

Zang,C. *et al.* (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*, **25**, 1952–1958.

Zhang,Y. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.