Bioinformatics, 37(22), 2021, 4238–4239
doi: 10.1093/bioinformatics/btab379
Advance Access Publication Date: 17 May 2021
Applications Note

Genome analysis

MotifGenie: a Python application for searching transcription factor binding sequences using ChIP-Seq datasets

Cerag Oguztuzun[1], Pelin Yasar[2], Kerim Yavuz[2], Mesut Muyan[2] and Tolga Can [3,*]

[1]Department of Computer Engineering, Bilkent University, Ankara 06800, Turkey, [2]Department of Biological Sciences, Middle East Technical University, Ankara 06800, Turkey and [3]Department of Computer Engineering, Middle East Technical University, Ankara 06800, Turkey

*To whom correspondence should be addressed.

Associate Editor: Peter Robinson

Received on March 1, 2021; revised on April 15, 2021; editorial decision on May 11, 2021; accepted on May 13, 2021

Abstract

Motivation: Next generation sequencing enabled the fast accumulation of genomic data at public repositories. This technology also made it possible to better understand the regulation of gene expression by transcription factors (TFs) and various chromatin-associated proteins through the integration of chromatin immunoprecipitation (ChIP-Seq). The Cistrome Project has become one of the indispensable research portals for biologists to access and analyze data generated with thousands of ChIP-Seq experiments. Integrative motif analysis on shared binding regions among a set of experiments is not yet achievable despite a set of search and analysis tools provided by Cistrome via its web interface and the Galaxy framework.

Results: We implemented a python command-line tool for searching binding sequences of a TF common to multiple ChIP-Seq experiments. We use the peaks in the Cistrome database as identified by MACS 2.0 for each experiment and identify shared peak regions in a genomic locus of interest. We then scan these regions for binding sequences using a binding motif of a TF obtained from the JASPAR database. MotifGenie is developed in collaboration with molecular biologists and its findings are corroborated by laboratory experiments.

Availability and implementation: MotifGenie is freely available at https://github.com/ceragoguztuzun/MotifGenie.

Contact: tcan@metu.edu.tr

1 Introduction

Transcriptional regulation is a complex process involving a multitude of actors in a cell. In recent years, genome-wide assays, powered by novel sequencing techniques, helped researchers gain new insights toward understanding the intricacies of transcriptional events. One of the most widely applied techniques, chromatin immunoprecipitation (ChIP), allows identification of DNA target sequences of DNA-binding proteins. For a specific DNA-binding protein, binding sites over the whole genome can be determined with a sequencing protocol. ChIP-Seq has provided unprecedented details of gene-specific transcriptional control in physiology and pathophysiology.

The Cistrome Database (Liu *et al.*, 2011), along with the associated search and analysis tools, is an invaluable collection of ChIP-Seq data. It allows users to search for ChIP-Seq data generated with the use of various biological resources for a specific transcription factor (TF). The Cistrome Database also contains data for other chromatin activities, such as histone modification and chromatin accessibility, but in this short applications note, we focus on TFs. One useful set of tools in Cistrome is the Toolkit for Cistrome Data Browser which can be used to find TFs that (i) potentially regulate a gene of interest, (ii) bind to a specific genomic interval or (iii) have significant binding overlap with a set of intervals. The Cistrome Database contains about eleven thousand ChIP-Seq experiments. Although the entire database is searched by the Toolkit for Cistrome Data Browser, the analysis results are given for individual experiments. To the best of our knowledge, searching for binding intervals that are common to a set of ChIP-Seq experiments is not yet achievable.

In this communication, we present MotifGenie, a Python application which allows users to identify common binding intervals of a TF of interest using all available ChIP-Seq data of a selected biological source at the Cistrome database. MotifGenie also searches for binding sequences in these common regions using a binding motif of that TF. If found, these sequences provide another degree of validation and help biologists pinpoint the target regions of TFs in different experimental paradigms.

2 Methods and implementation

MotifGenie is composed of two main modules: (i) finding common binding regions in multiple ChIP-Seq samples and (ii) searching for binding sequences using a binding profile of a TF. The first module is implemented in Python as a web app via Flask on the Google Cloud App Engine. MotifGenie mainly performs the tasks in the second module while programmatically initiates queries for the first one. Peaks identified by MACS 2.0 (Zhang *et al.*, 2008) for 11 286 samples in the Cistrome database are stored privately on the Google Cloud Storage as bed files. Given a cell line, a TF, a genomic locus and an occurrence percentage threshold t, the first module initially identifies the subset of samples for the given cell line and the TF. Then, it scans each bed file in that subset for the given genomic locus. Peak regions that occur in at least t% of the subset of samples are identified at single-base resolution and are post-processed to identify contiguous common binding regions. By increasing t, one can obtain high-confidence binding regions with the cost of reduced sensitivity. The first module can also be accessed as a standalone web application at https://motifgenie-merge-peaks.ue.r.appspot.com/.

The second module uses a JASPAR (Fornes *et al.*, 2020) TF binding profile and searches the genomic sequences in common binding regions by linear scan to find ungapped profile-sequence alignments. The sequences are obtained from the human reference genome, hg38, using the REST API data interface of the UCSC Genome Browser (Kent *et al.*, 2002). The base frequencies in the binding profile are converted into log-odds scores (Altschul *et al.*, 2010) and a match score of +3 and a mismatch penalty of -2 are used to score the ungapped alignment of a subsequence of the binding region with the binding profile. By default, the top 15 highest scoring sequences in binding regions are presented to the user. A sequence logo of the top-scoring binding sequences is also generated if the corresponding command line option is used.

3 Example run

We provide an example usage of MotifGenie using the transcriptional regulation of the MYC oncogene by the TF CTCF in the human colon cancer cell line, HCT-116. CTCF is reported to regulate *MYC* in this cell line (Schuijers *et al.*, 2018). We analyzed a 10-kb region encompassing the *MYC* locus and performed a MotifGenie search with the following parameters: locus = chr8: 127730434–127740951, cell line = HCT-116, JASPAR profile = MA0139.1.pfm and occurrence threshold, $t = 40\%$. There are 15 ChIP-Seq experiments in the Cistrome database using the HCT-116 cell line with CTCF as the TF. Processing the peak regions in these 15 samples, the first module of MotifGenie, identifies three common (i.e. occurring in at least 40% of the samples) contiguous binding regions near and inside the *MYC* locus at chr8: 127733936–127734282, chr8: 127736140–127736349 and chr8: 127737787–127737932. The second module searches the probabilistic binding profile of CTCF on the reference sequences of these common binding regions. The output of MotifGenie and the correspondingsequence logo of the top 15 sequences are shown in Figure 1. The sequence logo agrees with the JASPAR profile at http://jaspar.genereg.net/matrix/MA0139.1/.

4 Discussion

MotifGenie is a tool that extends the available functionality of the Cistrome Database Toolkit. One of the major contributions is to be able to integrate multiple ChIP-Seq experiments and identify high-confidence binding regions. Using MotifGenie, we identified the binding sites of ELF1 and MAZ on the CXXC5 gene and validated these sites using various approaches including electrophoretic mobility shift assay (EMSA) (Yasar *et al.*, 2021). The Cistrome Analysis Pipeline includes a tool for motif analysis, named MISP. However, when we ran MISP via the Galaxy Toolkit (Afgan *et al.*, 2016) on the *CXXC5* locus, MISP was not able to identify the ELF1 and MAZ binding motifs.

SEQUENCE	SCORE	POSITION	PARALLELITY	PEAK
GTGTGCCCTCTACTGGCAG	9.342963	127734105 - 127734123	-	peak1
TTGCGCCACCTGAAGGAGA	8.430384	127737843 - 127737861	-	peak3
CCGCCCCCTCTCTGGGCAG	6.603304	127734156 - 127734174	-	peak1
GCCGCGGGAGGTGGCGGTG	6.159184	127737898 - 127737916	+	peak3
GGGTCAGGTGGGGGCAGGA	6.014313	127734051 - 127734069	+	peak1
CGCCCTTAAGAAGCCGCGG	5.780073	127737910 - 127737928	+	peak3
GCGCCACCTGAAGGAGAAG	5.735097	127737841 - 127737859	+	peak3
GGGGCAGGAGCAGGAGCGT	5.678570	127734041 - 127734059	+	peak1
TCCAGCCACCTCCTTGTTA	5.543466	127733968 - 127733986	-	peak1
TCCACCCTAGCCGGCCGCC	5.353720	127736273 - 127736291	+	peak2
CCACCTGAAGGAGAAGGCG	5.330792	127737838 - 127737856	+	peak3
TTTCCAGCGGGGGAAGGAC	5.273824	127734009 - 127734027	+	peak1
TTGGCTGCAGAAGGTCCGA	5.217327	127733939 - 127733957	+	peak1
CTGCCTCTCGCTGGAATTA	4.891351	127736242 - 127736260	+	peak2
TGGGCGCTAGCGGCTGCGT	4.650896	127737810 - 127737828	+	peak3

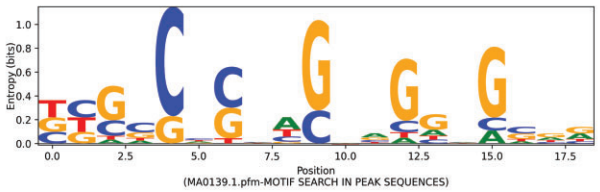

Fig. 1. The output of MotifGenie. The identified top-scoring binding motif sequences in the three common binding regions for CTCF are listed in the top panel with their corresponding similarity scores. The search is performed on both the forward and the reverse strand. The bottom panel shows the sequence logo of the top 15 binding sequences

MotifGenie is not specific for Cistrome; but, it uses the vast amount of preprocessed (i.e. peaks identified with MACS 2.0) ChIP-Seq data provided by the Cistrome Database. MotifGenie is not currently part of the Cistrome Project and is available as a standalone tool. Since MotifGenie is available in source code, analysis of other ChIP-Seq datasets is also feasible.

Although MotifGenie performs motif search on a reference sequence, if aligned sequences of each sample or single nucleotide variations at binding locations are available, MotifGenie may be used to find binding sequences specific to a given condition (or lack of them) toward a better understanding of mechanistic aspects of gene expression associated with a pathological state.

Funding

This work was supported by the Scientific and Technological Research Council of Turkey – Chemistry and Biology Research Support Group (TUBITAK-KBAG) [118Z957] and the Middle East Technical University – Scientific Research Projects (METU-BAP) [108-2021-10640].

Conflict of Interest: none declared.

References

Afgan,E. *et al.* (2016) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.*, **44**, W3–W10.

Altschul,S.F. *et al.* (2010) The construction and use of log-odds substitution scores for multiple sequence alignment. *PLoS Comput. Biol.*, **6**, e1000852.

Fornes,O. *et al.* (2020) JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **48**, D87–D92.

Kent,W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

Liu,T. *et al.* (2011) Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol.*, **12**, R83.

Schuijers,J. *et al.* (2018) Transcriptional dysregulation of MYC reveals common enhancer-docking mechanism. *Cell Rep.*, **23**, 349–360.

Yasar,P. *et al.* (2021) A CpG island promoter drives the CXXC5 gene expression. *submitted*,

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