DOI 10.18699/SBB-2022-338

Integration of ChIP-seq and RNA-seq data in structure-function analysis of *cis*-regulatory elements

Dolgikh V.[1, 2]*, Wiebe D.[1, 2], Levitsky V.[1, 2], Zemlyanskaya E.[1, 2]

[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*

[2] *Novosibirsk State University, Novosibirsk, Russia*

* *volleydollmc1@gmail.com*

Key words: regulation of transcription, *cis*-elements, *cis*-regulatory code, transcription factors

Motivation and Aim: The precision of gene regulation by transcription factors (TFs) is controlled at multiple levels, including the physical properties of TF–DNA interactions and TF oligomerization, among others. This is reflected in a complex structure of *cis*-regulatory elements. They may represent *cis*-regulatory modules consisting of several TF binding sites; besides, divergent binding sites may correspond to one TF. Due to a high combinatorics, the systemic experimental studies of *cis*-regulatory complexity are tricky. Bioinformatics integration of ChIP-seq data (which, along with the target TF binding site, contain the footprints of its partners) with RNA-seq data (which reflect transcriptional response of potential TF targets) can facilitate the systemic studies of *cis*-regulatory elements. However, to our knowledge, no software for such integrated analysis is currently available. Therefore, we developed a pipeline for structure-function analysis of *cis*-regulatory elements based on integration of ChIP-seq and RNA-seq data.
Methods and Algorithms: The pipeline is implemented in R programming language. It consists of three modules. Module (I) performs a repetitive *de novo* motif search in ChIP-seq peaks with HOMER, followed by motif clustering, cluster analysis and annotation of the inferred motifs using universalMotif and igraph R packages. Module (II) analyzes pairwise co-occurrence of the inferred motifs in ChIP-seq peaks with MCOT. Module (III) estimates an enrichment of the putative regulatory elements defined at the previous stages in certain promoter regions of differentially expressed genes using Fisher's exact test.
Results: The pipeline uses ChIP-seq peaks for TF of interest, the set of differentially expressed genes, which reflect transcriptional activity of potential TF targets, and the background set of genes as input data. Noteworthy, the algorithm beneath the pipeline is capable of automatically (with no expert involvement) reducing technical artifacts of *de novo* motif search, and segregating related motifs, which could potentially represent divergent binding sites for the same TF. The pipeline output contains the list of motifs and composite elements (the simplest *cis*-regulatory modules consisting of two TF binding sites), which potentially regulate the target TF-driven transcription, as well as the functional characteristics of each regulatory element, including association with transcriptional response. The output is supplemented with clear visualizations. The pipeline application for analysis of *cis*-elements involved in regulation of hormone responses in *Arabidopsis thaliana* identified different modes of transcriptional regulation and allowed classifying hormone-related TFs.
Conclusion: Our results demonstrate that the developed pipeline can foster the screening studies of structure-function organization of *cis*-regulatory elements
Acknowledgements: This work was supported by the RSF grant No. 20-14-00140.