doi 10.18699/bgrs2024-1.2-13

Exploring the interplay between reproducibility of open chromatin regions and transcription factor functional activity

Kolmykov S.*, Kulyashov M., Sokolova T., Prasolov D., Kolpakov F.
Department of Computational Biology, Sirius University of Science and Technology, Sirius, Russia
** kolmykovsk@gmail.com*

Key words: transcription factors binding regions; open chromatin regions; ChIP-seq; DNase-seq; scATAC-seq

Motivation and Aim: Exploring the relationship between transcription factor binding regions (TFBRs) and open chromatin regions (OCRs) is crucial for comprehending gene expression regulation. Currently, there are thousands of publicly available next-generation sequencing (NGS) experiments on transcription regulation, such as ChIP-seq, DNase-seq, ATAC-seq, etc. Despite the enormous variety of experimental conditions, it is possible to identify groups of experiments conducted under the similar conditions. Hence, exploring the reproducibility of regulatory regions across experiments under the similar conditions alongside their functional analyses holds promise in advancing our understanding of transcriptional regulation mechanisms. The primary goal of this study is to assess the positional specificity of TFBRs relative to OCRs and their reproducibility, and to explore how this characteristic relates with the functional activity of TFBRs.

Methods and Algorithms: In this study, we utilized 1239 ChIP-seq and 500 DNase-seq experiments from the GTRD database [1]. Open chromatin regions (OCRs) and TF binding regions (TFBRs) were overlapped with each other within cell lines/tissues and under similar treatment conditions. OCRs were pooled and merged across cell lines, and for each region, its reproducibility within the cell line was evaluated. Sets of reliable TFBRs were selected using the METARA algorithm [2].

Functional annotation of TFs was carried out based on the Gene Ontology (GO) database. The TFs were divided into 3 groups: positive regulators (GO:0051254), negative regulators (GO:0051253), TFs with mixed activity. Additionally, TF knockdown/knockout RNA-seq datasets from the KnockTF database were utilized to assess the functional activity of TFBRs [3].

Position weight matrices (PWM) from the HOCOMOCO database [4] were used to identify TF binding motifs within TFBRs. The evolutionary conservativity of TFBRs was evaluated using the phastCons100way and phyloP100way tracks from the UCSC database.

To investigate the reproducibility of open chromatin regions at the single-cell level, we analyzed scATAC-seq data (GSE162690) using the ArchR R-library [5] and a homebrew R-script.

Results: To determine whether TFBRs tend to localize within OCRs, we overlapped TFBRs and OCRs from the GTRD database (Fig. 1). Notably, TFs such as ATF2, MAFK, MYB, SPI1, FOXA2, and CEBPB exhibited reduced ratios of TFBRs within OCRs. To gain deeper insight into this observation, we investigated whether these proteins have the capacity to influence chromatin accessibility. Pioneer and flanking accessibility-associated TFs were identified based on the study by Lemma et al. [6]. The

majority of TFs linked with chromatin remodeling exhibit a lower percentage of TFBRs within OCRs. Moreover, these TFs are characterized by lower reproducibility of OCRs.

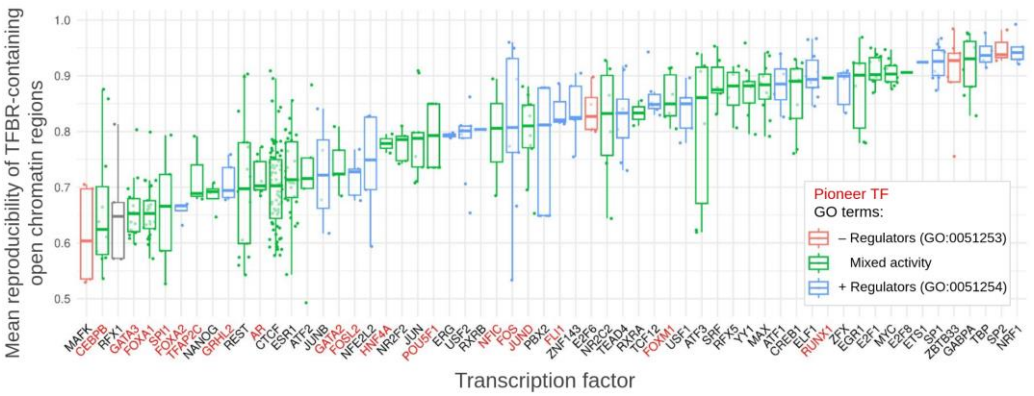

Fig. 1. The average reproducibility of OCRs overlapped with TFBRs in the cell type with the number of DNase-seq experiments greater than 4

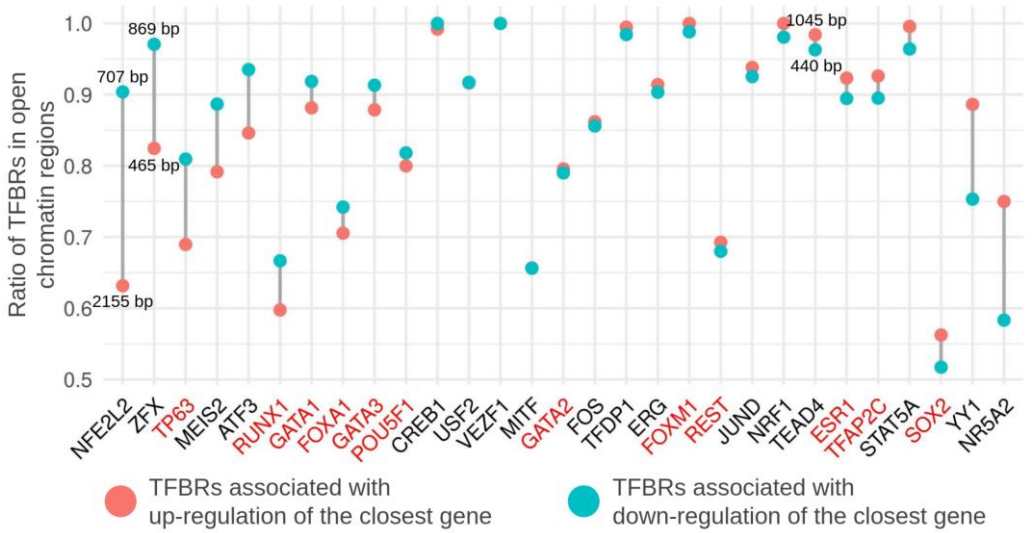

Fig. 2. Relationship between the ratio of TFBRs within OCRs and the pattern of change in the expression level of the closest gene following TF knockout/knockdown. Distances in kb indicate the average distance from the nearest TSS

Since DNase-seq is based on bulk cell analysis, it's likely that lower reproducibility of OCRs may arise due to the OCRs heterogeneity across different cell subpopulations. To explore this further we analyzed reproducibility of OCRs at the single cell level based on the scATAC-seq data.

Additionally, we investigated differences in the ratio of TFBRs within OCRs based on changes in the expression levels of the closest genes (<5kb to TSS) after TF knockout/knockdown. Significant differences in the ratio of TFBRs within OCRs were observed for several TFs depending on whether the expression of the closest gene increased or decreased (Fig. 2). For example, for ZFX and NFE2L2, a higher ratio of TFBRs within OCRs is associated with an increase in gene expression levels after TF knockout, while the reverse pattern is observed for TFAP2C and STAT5A.

Furthermore, a relationship between the distance of TFBRs to the closest TSS and the alterations in gene expression levels was discovered for several TFs: NFE2L2, ZFX and TEAD4. It is worth noting, that for NFE2L2 and ZFX there is also a significant difference between reproducibility of OCRs containing TFBRs associated with up-regulation and down-regulation of expression levels of genes. Moreover, for ZFX, the TFBRs associated with up-regulation of the closest genes are more likely to contain a TF binding motif and are also more evolutionary conserved compared to down-regulation associated TFBRs.

Conclusion: The preferences of TFBRs to be located within OCRs for different TFs were investigated. Most TFs associated with chromatin remodeling have demonstrated a reduced reproducibility of OCRs containing TFBRs. Certain TFs showed a correlation between a higher ratio of TFBRs within OCRs, their reproducibility, and changes in the expression levels of the nearest genes following TF knockout/knockdown. Analysis of OCRs reproducibility at the single-cell level revealed stable regions among OCRs with reduced reproducibility, consistent across cell subpopulations.

Funding: The study was financially supported by the Russian Science Foundation (project No. 24-24-20108, https://rscf.ru/en/project/24-24-20108/).

References

1. Kolmykov S. et al. Meta-analysis of ChIP-seq datasets through the rank aggregation approach. In: Cognitive Sciences, Genomics and Bioinformatics (CSGB-2020). Novosibirsk, 2020;180-184. doi 10.1109/CSGB51356.2020.9214614
2. Kolmykov S. et al. GTRD: an integrated view of transcription regulation. *Nucleic Acids Res*. 2021;D1:D104-D111. doi 10.1093/nar/gkaa1057
3. Feng C. et al. KnockTF 2.0: a comprehensive gene expression profile database with knockdown/knockout of transcription (co-) factors in multiple species. *Nucleic Acids Res*. 2024;52(D1):D183-D193. doi 10.1093/nar/gkad1016
4. Vorontsov I.E. et al. HOCOMOCO in 2024: a rebuild of the curated collection of binding models for human and mouse transcription factors. *Nucleic Acids Res*. 2024;52(D1):D154-D163
5. Granja J.M. et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat Genet*. 2021;53(3):403-411. doi 10.1038/s41588-021-00790-6
6. Lemma R.B. et al. Pioneer transcription factors are associated with the modulation of DNA methylation patterns across cancers. *Epigenet Chromatin*. 2022;15(1):13. doi 10.1186/s13072-022-00444-9