Advances in Supervised Interpretation and Preprocessing of ATAC-seq

Single-cell assay for transposase-accessible chromatin with sequencing (scATAC-seq) has been widely used to profile chromatin accessibility and explore the epigenomic properties in heterogeneous cell populations. Furthermore, transcription factors (TFs) bind to accessible sites, and scATAC-seq analysis could associate TF binding specificities, typically available as sequence motifs, with variance in chromatin accessibility.

Here we present a new supervised TF motif regression method, MotifReg, for computational analysis and interpretation of scATAC-seq data. MotifReg can efficiently analyze large-scale scATAC-seq datasets and learn interpretable associations between TF motifs and accessible peaks. The resultant low-dimensional cell representations in the TF activity space facilitate downstream cell type-specific analyses of scATAC-seq data. The supervised formulation in MotifReg enables ranking TF motifs by significance of their association with chromatin accessibility, implicating cell type-specific TFs.

Furthermore, by comparing MotifReg performance across data analysis settings, we develop guidelines and best practices for ATAC-seq peak calling, data preprocessing and normalization. We leverage these best practices to show that aggregation of biologically similar ATAC data allows for the discovery of chromosomal regions enriched for chromatin interaction, known motif binding, or ATAC co-accessibility. We provide an automated methodology to quickly and efficiently scrape, compile, and preprocess this data from online sources and illustrate how construction of these high-quality atlases can complement existing analyses.

Reading List

Book

1. Bioinformatics Algorithms, Phillip Compeau and Pavel Pevzner, Chapters 1-10

Single Cell Modality Papers

1. Heumos, L., Schaar, A.C., Lance, C. et al. Best practices for single-cell analysis across modalities. Nat Rev Genet 24, 550–572 (2023). https://doi.org/10.1038/s41576-023-00586-w
2. Lause, J., Berens, P. & Kobak, D. Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. Genome Biol 22, 258 (2021). https://doi.org/10.1186/s13059-021-02451-7
3. Martens, L.D., Fischer, D.S., Yépez, V.A. *et al.* Modeling fragment counts improves single-cell ATAC-seq analysis. *Nat Methods* **21**, 28–31 (2024). https://doi.org/10.1038/s41592-023-02112-6

ATAC Analysis Papers

1. Yuan, H. & Kelley, D. scbasset: sequence-based modeling of single-cell atac-seq using convolutional neural networks. Nature Methods 19 (2022). https://www.nature.com/articles/s41592-022-01562-8
2. Zhang, Y. et al. Model-based analysis of chip-seq (macs). Genome Biology 9, R137 – R137 (2008). https://genomebiology.biomedcentral.com/articles/10.1186/gb-2008-9-9-r137
3. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromvar: Inferring transcriptionfactor-associated accessibility from single-cell epigenomic data. Nature methods 14, 975 – 978 (2017). https://www.nature.com/articles/nmeth.4401
4. Bentsen, M., Goymann, P., Schultheis, H. *et al.* ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nat Commun* **11**, 4267 (2020). https://doi.org/10.1038/s41467-020-18035-1
5. ChromBPNet: bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants Anusri Pampari, Anna Shcherbina, Evgeny Z. Kvon, Michael Kosicki, Surag Nair, SoumyaKundu, Arwa S. Kathiria, Viviana I. Risca, Kristiina Kuningas, Kaur Alasoo, William JamesGreenleaf, Len A. Pennacchio, Anshul Kundaje bioRxiv 2024.12.25.630221; doi: https://doi.org/10.1101/2024.12.25.630221

Motif Library Paper

1. Vierstra2020 Vierstra, J., Lazar, J., Sandstrom, R. *et al.* Global reference mapping of human transcription factor footprints. *Nature* **583**, 729–736 (2020)

Sequencing Paper

1. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015 Jul 23;523(7561):486-90. doi: 10.1038/nature14590. Epub 2015 Jun 17. PMID: 26083756; PMCID: PMC4685948.