Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

PREVIEW

From Reads to Insights: Integrative Pipelines for Biological Interpretation of ATAC-seq Data

Ya Cui*, **Jason Sheng Li**, **Wei Li**

Division of Computational Biomedicine, Department of Biological Chemistry, School of Medicine, University of California, Irvine, Irvine, CA 92697, USA

Received 15 June 2021; revised 30 June 2021; accepted 7 July 2021
Available online 15 July 2021

Handled by Zhihua Zhang

Underlying the regulation of mammalian gene expression at the level of transcription is the structure and modifications of chromatin. Understanding the twisting structures of DNA wrapped around histones and their higher-level ordering allows us to peek into a vast regulatory landscape. Looking closer, the physical accessibility of the genome provides information on the positions of nucleosomes and biologically-active regions — promoters, enhancers, insulators, and other regulatory elements. Currently, Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq) is a widely used technology for detecting genome-wide chromatin accessibility [1,2]. Compared to other methods, such as micrococcal nuclease sequencing (MNase-seq) [3], formaldehyde-assisted isolation of regulatory elements sequencing (FAIRE-seq) [4], and DNase I hypersensitive sites sequencing (DNase-seq) [5], ATAC-seq offers a simpler and quicker protocol by incorporating hyperactive Tn5 transposase to simultaneously cut open chromatin and ligate high-throughput sequencing adapters at chromatin-accessible regions [1,2]. Importantly, ATAC-seq can also be applied to samples with limited starting cell material (500–50,000 cells) [1,2].

These advantages of ATAC-seq — simplicity, speed, and low input material requirements — have made it highly popular in recent years. Publications and datasets using ATAC-seq have increased exponentially [6]. Several international consortia, including The Cancer Genome Atlas (TCGA) [7], CommonMind [8], FOUNDIN-PD [9], AD Knowledge Portal [10], and iPSCORE [11], use ATAC-seq data for population-scale studies. Despite the growing need to process ATAC-seq data, only a few analytical tools have been developed specifically for such data. Most of the current tools used for ATAC-seq data analysis are adopted from DNase-seq or chromatin immunoprecipitation sequencing (ChIP-seq) data analysis suites assuming similar data characteristics [12]. Typically, there are five steps for ATAC-seq data analysis: 1) quality control (QC); 2) read alignment; 3) peak calling; 4) downstream analysis (such as peak differential analysis, peak annotation, motif enrichment, and nucleosome position analysis); and 5) integration with multi-omics data [6]. However, there is no comprehensive and standalone analytical pipeline defined to guide ATAC-seq users.

In this issue, Liu et al. [13] and Qiu et al. [14] make great strides toward a standalone analytical ATAC-seq pipeline. Liu et al. present the ATAC-seq Integrative Analysis Package (AIAP), which defines a series of ATAC-seq-specific QC metrics to optimize data and further uses a pseudo single-end strategy (PE-asSE) specifically developed for ATAC-seq to improve data analysis [13]. Using these ATAC-seq-specific metrics (*e.g.*, reads under peak

*Corresponding author.
E-mail: yac7@uci.edu (Cui Y).

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.
https://doi.org/10.1016/j.gpb.2021.06.002

ratio, background, promoter enrichment, and subsampling enrichment) in conjunction with other traditional QC metrics, Liu et al. conducts QC checks at different steps of data processing and provide a user-friendly visualization of quality reports. Even further, Liu et al. incorporate a PE-asSE strategy to better analyze the Tn5 transposase insertion event within ATAC-seq data. Briefly, Liu et al. shift each end of the non-redundant uniquely mapped read pair +4 bp/−5 bp to define the Tn5 insertion position and then further extend 75 bp in both directions around the Tn5 insertion position. By applying this strategy, one non-redundant uniquely mapped read pair is divided into two single-end fragments. Applying the PE-asSE approach to ATAC-seq data of GM12878 cells leads to the identification of ~ 99.9% peaks that can be identified by traditional strategies and an additional ~ 23% (20,918) peaks that cannot be discovered by traditional methods. Further analysis shows that most of these PE-asSE-specific peaks overlap with known GM12878 DNase I hypersensitive sites (DHSs) and are highly enriched in all active histone modifications. Taken together, the additional open chromatin regions (OCRs) identified by PE-asSE are likely to be true functional regulatory elements rather than false positives, highlighting dramatic improvements in ATAC-seq OCR discovery. Given these improvements, Liu et al. further extend the usage of AIAP to differentially accessible region (DAR) analysis, applying AIAP to ATAC-seq data of mouse liver at embryonic day 11.5 (E11.5) and postnatal day 0 (P0). Peak differential analysis of data collected at these two time points results in an ~ 35% increase in the number of DARs identified. Overall, Liu et al. present an ATAC-seq QC and analysis pipeline that dramatically improves the sensitivity of both peak calling and differential analysis.

In another paper of this issue, Qiu et al. present the Containerized Bioinformatics workflow for Reproducible ChIP/ATAC-seq Analysis (CoBRA), which provides a comprehensive and customizable ChIP and ATAC-seq analysis pipeline [14]. The strength of CoBRA is to integrate multiple commonly used functions, including normalization, copy number variation adjustment, sample clustering, differential peak calling, motif enrichment, Cistrome DB Toolkit [15] analysis, and Gene Set Enrichment Analysis (GSEA) [16] pathway analysis into the same package for scientists with limited computational experience. Based on the snakemake system [17], CoBRA can also easily integrate additional analytical tools in the future. CoBRA is well-documented and provides step-by-step tutorials for 3 case studies to guide users with limited computational experience. As an example, Qiu et al. have applied CoBRA to the ATAC-seq data of HL-60 promyelocytes differentiating into macrophages at five time points (0 h, 3 h, 24 h, 96 h, and 120 h). Unsupervised analysis results in three clusters showing clear differences in open chromatin between the early (0 h and 3 h), intermediate (24 h), and late stage (96 h and 120 h) time points. Further motif enrichment analysis in each cluster also identifies potential functional transcriptional regulators, such as early growth response protein (EGR) and Maf, in macrophage differentiation. Finally, Qiu et al. have integrated ATAC-seq data with sample-matched RNA-seq data, highlighting that genes differentially expressed during macrophage differentiation are flanked by changes in open chromatin structure.

In short, AIAP and CoBRA, the two useful ATAC-seq analysis pipelines developed by Liu et al. and Qiu et al., will provide a convenient entry and general guideline for scientists with limited computational experience to explore ATAC-seq data. Both pipelines use docker to allow compatibility on different operating systems and are able to generate high-quality figures automatically. These two studies represent significant progress in ATAC-seq analysis. Despite their benefits, more efforts are needed to extend the power of these ATAC-seq pipelines in the future. For example, more alternative tools may be included for each analytical step. Currently, both AIAP and CoBRA pipelines only include a popular software MACS2 [18] for peak calling, but not HMMRATAC [19], an alternative peak caller that is specifically developed for ATAC-seq and outperforms MACS2 [19]. Furthermore, to take advantage of multi-omics data from the same individual, more advanced analytical functions are required to integrate ATAC-seq data with other -omics data such as those from genotyping, RNA-seq, and ChIP-seq. For example, the integration of genotype and ATAC-seq modalities to identify quantitative trait loci for chromatin accessibility (caQTLs) has been recently used to evaluate genetic effects on chromatin accessibility [20]. caQTLs have been applied to the fine mapping of causal variants in noncoding chromatin accessible regions in multiple cell lines and tissues, such as blood [21], liver [22], and brain [23], furthering our understanding of regulatory mechanisms in human diseases at the tissue or cell type-specific levels. We expect more ATAC-seq-specific analytic tools to be developed in the future. Benchmarking studies for ATAC-seq analytical tools will help guide developers of comprehensive ATAC-seq pipelines in selecting the appropriate tools to be included. As ATAC-seq data continue to be generated for more individuals, cell types, and molecular processes, comprehensive and user-friendly ATAC-seq analytic pipelines will aid in the interpretation of biological mechanisms underlying ATAC-seq data.

CRediT author statement

Ya Cui: Conceptualization, Writing - original draft, Writing -

reviewing & editing, Supervision. **Jason Sheng Li:** Writing - original draft, Writing - reviewing & editing. **Wei Li:** Conceptualization, Writing - reviewing & editing, Funding acquisition, Supervision. All authors have read and approved the final manuscript.

Competing interests

The authors declare no competing financial interests.

Acknowledgments

We thank members of the Li lab for helpful discussions. This work was supported by the National Institutes of Health, United States, to Wei Li (Grant Nos. R01HG007538, R01CA193466, and R01CA228140).

ORCID

0000-0003-1574-0928 (Ya Cui)
0000-0003-0485-7194 (Jason Sheng Li)
0000-0001-9931-5990 (Wei Li)

References

[1] Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: a method for assaying chromatin accessibility genome-wide. Curr Protoc Mol Biol 2015;109:21.29.1–9.

[2] Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 2013;10:1213–8.

[3] Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, et al. Dynamic regulation of nucleosome positioning in the human genome. Cell 2008;132:887–98.

[4] Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res 2007;17:877–85.

[5] Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell 2008;132:311–22.

[6] Yan F, Powell DR, Curtis DJ, Wong NC. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. Genome Biol 2020;21:22.

[7] Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, et al. The chromatin accessibility landscape of primary human cancers. Science 2018;362:eaav1898.

[8] Bryois J, Garrett ME, Song L, Safi A, Giusti-Rodriguez P, Johnson GD, et al. Evaluation of chromatin accessibility in prefrontal cortex of individuals with schizophrenia. Nat Commun 2018;9:3121.

[9] Bressan E, Reed X, Bansal V, Hutchins E, Cobb MM, Webb MG, et al. The Foundational data initiative for Parkinson's disease (FOUNDIN-PD): enabling efficient translation from genetic maps to mechanism. bioRxiv 2021;446785.

[10] De Jager PL, Ma Y, McCabe C, Xu J, Vardarajan BN, Felsky D, et al. A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. Sci Data 2018;5:180142.

[11] Panopoulos AD, D′Antonio M, Benaglio P, Williams R, Hashem SI, Schuldt BM, et al. iPSCORE: a resource of 222 iPSC lines enabling functional characterization of genetic variation across a variety of cell types. Stem Cell Reports 2017;8:1086–100.

[12] Chang P, Gohain M, Yen MR, Chen PY. Computational methods for assessing chromatin hierarchy. Comput Struct Biotechnol J 2018;16:43–53.

[13] Liu S, Li D, Lyu C, Gontarz PM, Miao B, Madden PAF, et al. AIAP: a quality control and integrative analysis package to improve ATAC-seq data analysis. Genomics Proteomics Bioinformatics 2021;19:641–51.

[14] Qiu X, Feit AS, Feiglin A, Xie Y, Kesten N, Taing L, et al. CoBRA: Containerized Bioinformatics workflow for Reproducible ChIP/ATAC-seq Analysis. Genomics Proteomics Bioinformatics 2021;19:652–61.

[15] Zheng R, Wan C, Mei S, Qin Q, Wu Q, Sun H, et al. Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. Nucleic Acids Res 2019;47:D729–35.

[16] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545–50.

[17] Koster J, Rahmann S. Snakemake — a scalable bioinformatics workflow engine. Bioinformatics 2018;34:3600.

[18] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol 2008;9:R137.

[19] Tarbell ED, Liu T. HMMRATAC: a hidden Markov ModeleR for ATAC-seq. Nucleic Acids Res 2019;47:e91.

[20] Kumasaka N, Knights AJ, Gaffney DJ. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nat Genet 2016;48:206–13.

[21] Alasoo K, Rodrigues J, Mukhopadhyay S, Knights AJ, Mann AL, Kundu K, et al. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. Nat Genet 2018;50:424–31.

[22] Currin KW, Erdos MR, Narisu N, Rai V, Vadlamudi S, Perrin HJ, et al. Genetic effects on liver chromatin accessibility identify disease regulatory variants. Am J Hum Genet 2021;108:1169–89.

[23] Liang D, Elwell AL, Aygun N, Krupa O, Wolter JM, Kyere FA, et al. Cell-type-specific effects of genetic variation on chromatin accessibility during human neuronal differentiation. Nat Neurosci 2021;24:941–53.