Challenges for Single Cell Epigenomics Pipeline

Leandro Lopez Leon (lslopez)

December 12, 2021

1 Introduction

Large scale sequencing techniques measuring genomic features have introduced many computational and statistical challenges. For this study, we focus on ChIP-seq and ATAC-seq assays and their corresponding pipelines. Both pipelines follow similar steps, which include [3, 4]

1. Raw data preprocessing, alignment, and filtering
2. Identifying accessible regions (peak calling)
3. Inference on resulting peak signal (Differential analysis, peak annotation, motif analysis, cell clustering)

At each stage of the pipeline, there are many implementation decisions which irrevocably affect upstream analysis. While there are some heuristics in place which serve as quality control, we believe there are many hyperparameters which require further validation at each stage of this pipeline. In particular, we want to minimize potential biases in the earlier phases which could cascade down to downstream analysis. We address the above steps in more detail in the next sections.

2 Preprocessing

Raw reads data is first trimmed and aligned to a reference genome using dynamic programming. After some filtering and sorting steps, the duplication frequency and library complexity are estimated. At this stage, there are several sanity checks that have to be made on the data depending on the assay. For ATAC-seq, the size distribution over fragment sizes should be periodic and contain 4 modes representing Nucleosome-free regions, and regions with 1,2 or 3 nucleosomes [4]. Additionally, we expect small fragments to be enriched around transcription start sites (TSS) of genes, while larger fragments should be depleted at TSS sites. For large experiments, the above steps can be time consuming, but ensuring data is consistent with our biological understanding is essential. Other than alignment and sorting, the other steps can be carried out in parallel.

If there are multiple libraries being analyzed, normalizing library complexity across them should theoretically account for differences in sequencing depth, but different normalization techniques can lead to biologically unsound results [3]. We don't have any theoretical concerns with with this part of the pipeline.

3 Peak Calling

Next up in the pipeline is the identification of accessible regions. This problem is made harder by the sparsity of valid peaks for any given cell. Several unsupervised and semi-supervised approaches have been developed, but the most popular tools are count-based methods similar to those used for RNA-seq (i.e. MACS3, HOMER)[4]. Theses count-based approaches dominate most pipelines yet they have been found to output inflated P-values in ChIP-seq data [1]. Exacerbating the issue, ChIP enrichments are often marginal and variable across experiments [2] resulting in signals which violate the assumptions of these models. Evaluating these broad enrichment remains an open problem at the crux of the pipeline. After identification of the peaks, one can normalize across peaks and/or across cells using TD-IDF.

4 Statistical Analysis and Inference on Annotated Signal

After we are satisfied with the resulting signal, we can try peaks to genes as well as try to get a glimpse at the underlying regulatory mechanism via motif analysis [4]. These tools are still not robust and have a high FDR's. Similarly we can try to perform feature selection and dimensionality reduction reduction on the outputted peaks in order to cluster cell based on function and gene activity profiles.

In conclusion, there are still gaps in the ChIP-seq/ATAC-seq pipeline that need to be resolved before tackling multi-modal single-cell datasets. This will come at the cost of more compute, but it is the only way to ensure robustness to technical and biological challenges.

References

[1] J. G. Chitpin, A. Awdeh, and T. J. Perkins. RECAP reveals the true statistical significance of ChIP-seq peak calls. *Bioinformatics*, 35(19):3592–3598, Mar. 2019. doi: 10.1093/bioinformatics/btz150. URL https://doi.org/10.1093/bioinformatics/btz150.

[2] C. A. Meyer and X. S. Liu. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews Genetics*, 15(11):709–721, Sept. 2014. doi: 10.1038/nrg3788. URL https://doi.org/10.1038/nrg3788.

[3] J. J. Reske, M. R. Wilson, and R. L. Chandler. ATAC-seq normalization method can significantly affect differential accessibility analysis and interpretation. *Epigenetics & Chromatin*, 13(1), Apr. 2020. doi: 10.1186/s13072-020-00342-y. URL https://doi.org/10.1186/s13072-020-00342-y.

[4] F. Yan, D. R. Powell, D. J. Curtis, and N. C. Wong. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biology*, 21(1), Feb. 2020. doi: 10.1186/s13059-020-1929-3. URL https://doi.org/10.1186/s13059-020-1929-3.