Single-Cell ATAC-seq Data Analysis Pipeline

Emon Karmoker, Syed Nafis, Md. Mansurul Haque, Jony Khan

May 6, 2024

Abstract

Single-cell Assay for Transposase-Accessible Chromatin using sequencing (scATAC-seq) has revolutionized our ability to probe chromatin accessibility at the single-cell level. In this paper, we present a comprehensive computational pipeline leveraging the `Signac` and `Seurat` packages in R for scATAC-seq data analysis. The pipeline encompasses data preprocessing, quality control, normalization, and downstream analysis such as dimensionality reduction and clustering. Applied to a dataset of peripheral blood mononuclear cells (PBMCs), our pipeline effectively identifies distinct cell populations based on chromatin accessibility profiles. Through visualization techniques like UMAP, our pipeline facilitates the exploration of cellular heterogeneity within populations. Overall, our standardized workflow empowers researchers to unravel the intricate landscape of chromatin accessibility regulation, offering new insights into cellular biology at unprecedented resolution.

1 Introduction

Single-cell ATAC-seq (scATAC-seq) is a cutting-edge technology that allows researchers to interrogate chromatin accessibility profiles at single-cell resolution. Unlike bulk ATAC-seq, which provides an average signal across a population of cells, scATAC-seq enables the investigation of cell-to-cell variability in chromatin accessibility within heterogeneous cell populations. However, the analysis of scATAC-seq data presents unique challenges due to the sparsity and noise inherent in the data. Here, we present a computational pipeline for the analysis of scATAC-seq data, leveraging the capabilities of the `Signac` and `Seurat` packages in R.

2 Literature Review

Previous studies have demonstrated the utility of scATAC-seq in various biological contexts, including cell type identification, regulatory element mapping, and cell state characterization [1, 3, 2]. Several computational tools and pipelines have been developed for the analysis of scATAC-seq data, each with its strengths and limitations. Our pipeline builds upon existing methods and provides a streamlined workflow for the analysis of scATAC-seq data, integrating quality control, normalization, and downstream analysis steps.

3 Methodology

The scATAC-seq data analysis pipeline presented in this paper consists of several key steps:

1. Data Preprocessing: Raw scATAC-seq data is preprocessed to generate a count matrix of chromatin accessibility.
2. Quality Control: Quality control metrics such as nucleosome signal score and TSS enrichment score are computed to filter out poor-quality cells.
3. Normalization: Normalization is performed using the TF-IDF method to account for differences in library size and capture efficiency.
4. Dimensionality Reduction: Linear and non-linear dimensional reduction techniques such as SVD and UMAP are applied to reduce the dimensionality of the data.

5. Clustering: Cells are clustered based on their chromatin accessibility profiles to identify distinct cell populations.

We utilize the `Signac` package for data preprocessing, including the creation of a chromatin assay and gene annotation. Quality control metrics such as nucleosome signal score and TSS enrichment score are computed to filter out poor-quality cells. Normalization is performed using the TF-IDF method, followed by linear and non-linear dimensional reduction techniques such as SVD and UMAP. Finally, cells are clustered based on their chromatin accessibility profiles.

4 Results

We applied our scATAC-seq analysis pipeline to a dataset of PBMCs (peripheral blood mononuclear cells) and obtained the following results:

- Identification of distinct cell clusters based on chromatin accessibility profiles
- Visualization of cell clusters using dimensionality reduction techniques such as UMAP

Our analysis revealed cell subpopulations with distinct chromatin accessibility patterns, suggesting heterogeneity within the PBMC population.

5 Discussion

The scATAC-seq analysis pipeline presented in this paper provides a robust framework for the analysis of scATAC-seq data. By integrating quality control, normalization, and downstream analysis steps, our pipeline enables researchers to uncover biological insights from scATAC-seq datasets. Future work may focus on further refining the pipeline and applying it to diverse biological systems to gain a deeper understanding of chromatin accessibility dynamics at the single-cell level.

6 Conclusion

In conclusion, we have presented a comprehensive pipeline for the analysis of scATAC-seq data, leveraging the `Signac` and `Seurat` packages in R. Our pipeline facilitates the identification of cell subpopulations and the exploration of chromatin accessibility dynamics at single-cell resolution. We believe that our pipeline will serve as a valuable resource for researchers interested in studying epigenetic regulation at the single-cell level.

References

[1] Jason D Buenrostro, Beijing Wu, Howard Y Chang, and William J Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490, 2015.

[2] M Ryan Corces, Jason D Buenrostro, Beijing Wu, Peyton G Greenside, Sok Kean Chan, Joshua L Koenig, Michael P Snyder, Jonathan K Pritchard, Anshul Kundaje, William J Greenleaf, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature genetics*, 48(10):1193–1203, 2016.

[3] Darren A Cusanovich, Riza Daza, Andrew Adey, Hannah A Pliner, Lena Christiansen, Kevin L Gunderson, Frank J Steemers, Cole Trapnell, and Jay Shendure. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237):910–914, 2015.