Pipeline for Identifying Interacting DNA Elements

NEW YORK UNIVERSITY

Anni Zheng, Catherine Hua, Lyric Li

Center for Data Science

Introduction & Background

Enhancers play a critical role in regulating gene expression, often working with other enhancers to influence transcription. Despite their importance, the mechanisms by which enhancers interact remain poorly understood, with limited evidence supporting the existence of significant interactions. Recent studies have demonstrated that enhancer effects generally combine multiplicatively and exhibit rare interactions that are challenging to detect. To advance this understanding, a fully functional processing pipeline and analytical framework was developed under the NYU High Performance Computing (HPC) environment. This framework integrates the generalized linear model (GLM) for imputing gene expression with the Poisson regression model for analyzing associations between enhancer accessibility and gene expression at the single-cell level. Lastly, the project studied enhancer pair activities specific to one cell type, CD14-Mono, which were identified based on the model's results and showed that these activities may be saturated and non-linear.

Dataset

The dataset used for this study comprises three key files: atac_matrix, rna_matrix, and meta_data, each serving a distinct purpose in the analysis pipeline. These files provide the foundation for mapping enhancer activity, gene expression, and cell metadata, enabling a comprehensive exploration of enhancer-gene interactions.

1. The **atac_matri**x file contains chromatin accessibility data for each enhancer region across all cells in the dataset. This matrix provides a quantitative measure of enhancer activity, with values ranging from 0(no activity) to 1(full activity).
2. The **rna_matrix** file contains single-cell RNA sequencing (scRNA-seq) data, capturing gene expression levels for each gene across all cells. Expression values range from 0 to 5585, representing the raw count of transcripts per gene.
3. The **meta_data** file includes additional information about each cell, such as: nUMI: Total number of unique molecular identifiers, reflecting sequencing depth celltype: The biological classification of each cell. For this project, we used celltype CD14-mono for analysis, and percent.mito: The percentage of mitochondrial gene expression, often used to filter out low-quality cells.

Environment Set Up

NYU-IT HPC Greene cluster was utilized to execute the project in our system, replacing the original LSF session scheduler. The Greene cluster operates on a Linux operating system. Resource management and job scheduling were handled by the Slurm software system. The Snakemake pipeline, a hybrid of Python and shell scripting, was employed for workflow management. 'SBATCH' commands were used for handling job and error files, as well as the 'module load' command to activate the integrated Snakemake environment within the HPC system.

References

[1] Tian, J., Lou, J., Cai, Y., Rao, M., Lu, Z., Zhu, Y., Zou, D., Peng, X., Wang, H., Zhang, M., Niu, S., Li, Y., Zhong, R., Chang, J., & Miao, X. (2020) Risk SNP-Mediated Enhancer-Promoter Interaction Drives Colorectal Cancer through Both FADS2 and AP002754.2. Cancer Research 80(9):1804-1818. doi: 10.1158/0008-5472.CAN-19-2389. Epub 2020 Mar 3. PMID: 32127356.

[2] Mitra, S., Malik, R., Wong, W., Rahman, A., Hartemink, A.J., Pritykin, Y., Dey, K.K., & Leslie, C.S. (2024) Single-cell multi-ome regression models identify functional and disease-associated enhancers and enable chromatin potential analysis. Nature Genetics 56(4):627-636. doi: 10.1038/s41588-024-01689-8. Epub 2024 Mar 21. Erratum in: Nature Genetics 56(6):1319. doi: 10.1038/s41588-024-01805-8. PMID: 38514783; PMCID: PMC11018525.

[3] Sakaue, S., Weinand, K., Isaac, S., Dey, K.K., Jagadeesh, K., Kanai, M., Watts, G.F.M., Zhu, Z.; Accelerating Medicines Partnership® RA/SLE Program and Network; Brenner, M.B., McDavid, A., Donlin, L.T., Wei, K., Price, A.L., & Raychaudhuri, S. (2024) Tissue-specific enhancer-gene maps from multimodal single-cell data identify causal disease alleles. Nature Genetics 56(4):615-626. doi: 10.1038/s41588-024-01682-1. Epub 2024 Apr 9. PMID: 38594305; PMCID: PMC11456345.

[4] Zhou, J., Guruvayurappan, K., Toneyan, S., Chen, H.V., Chen, A.R., Koo, P., & McVicker, G. (2023) Analysis of single-cell CRISPR perturbations indicates that enhancers act multiplicatively and provides limited evidence for epistatic-like interactions. bioRxiv 2023.04.26.538501; doi: https://doi.org/10.1101/2023.04.26.538501.

Methodology

SCENT (Single-Cell Enhancer-Target Gene Mapping): Models the causal relationship between enhancers and genes in single cells for both common and rare diseases. This algorithm uses Poisson regression, due to the sparsity of RNA and ATAC data, as well as bootstrap-based significance testing to obtain empirical p-values, thereby mapping enhancers to target genes. The algorithm also ensures that the output gene-peak-peak pairs pass a threshold with a non-zero proportion $> 5\%$ before applying the Poisson regression. Given the selected significant gene-peak-peak pairs, we performed multiple-testing correction on the p-values and filtered for all enhancer-gene pairs with an FDR (False Discovery Rate) < 0.1.

Epistasis Model: With the obtained enhancer-gene pairs as one of the inputs to the epistasis model, we first filtered the RNA, ATAC, and metadata files based on the cell type (e.g., CD14-Mono). To test the individual activity of each enhancer-gene pair in cells, we created three models: *cells10*, where enhancer2 is inactive regardless of enhancer1; *cells01*, where enhancer1 is inactive regardless of enhancer2; and *cells11*, where both enhancers are active.

General Form: $E_i \sim \text{Poisson}(\lambda_i)$

$$\log(\lambda_i) = \beta_0 + \beta_{\text{enhancer}} X_{\text{enhancer}} + \beta_{\text{mito}} X_{\text{mito}} + \beta_{\text{UMI}} X_{\text{UMI}}$$

Three models:

$$\log(\lambda_{i10}) = \beta_0 + \beta_{\text{enhancer1}} X_{\text{enhancer1}} + \beta_{\text{mito}} X_{\text{mito}} + \beta_{\text{UMI}} X_{\text{UMI}}$$

$$\log(\lambda_{i01}) = \beta_0 + \beta_{\text{enhancer2}} X_{\text{enhancer2}} + \beta_{\text{mito}} X_{\text{mito}} + \beta_{\text{UMI}} X_{\text{UMI}}$$

$$\log(\lambda_{i11}) = \beta_0 + \beta_{\text{enhancer1}} X_{\text{enhancer1}} + \beta_{\text{enhancer2}} X_{\text{enhancer2}} + \beta_{\text{mito}} X_{\text{mito}} + \beta_{\text{UMI}} X_{\text{UMI}}$$

The models return the values of intercepts, beta estimates and corresponding p-values. For the p-values that were below 0.1, an additional bootstrapping was applied to improve confidence.

Workflow

```
snakemake —latency-wait 60 —forceall —use-conda —jobs 32 —cluster
'sbatch —time=48:00:00 —mem=32G —cpus-per-task=8 -o out.%J.txt -e
err.%J.txt'
```

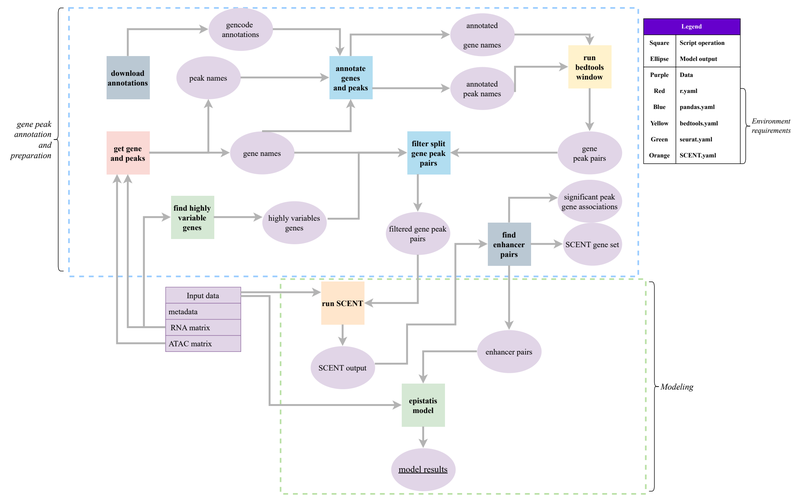
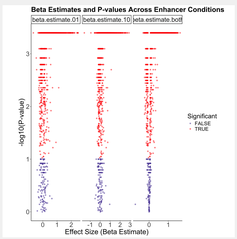


Figure 1. Pipeline workflow and outputs

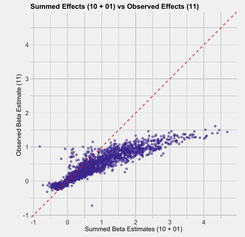
Result and Analysis

1. In the QQ plot(figure(a)), the observed values (black dots) largely align with the expected values (reddiagonal line) in the lower p-value range, indicating that the model outputs generally conform to expectations under the null. However, deviations from the diagonal line are observed at higher -log10(p) values, suggesting the presence of statistically significant enhancer-gene associations that exceed the null hypothesis, supporting the hypothesis of **biologically meaningful enhancer-gene associations.**
2. From the volcano plot(figure(b)), for all three comparisons(01vs00, 10vs00, 11vs00), a subset of enhancer-gene pairs shows positive beta estimates, indicating a positive association with gene expression when enhancers are active. Beta estimates for 11 vs 00 (both enhancers active) tend to be larger compared to 01 vs 00 and 10 vs 00, suggesting **an synergistic effect when both enhancers are active comparing to only one active.**
3. In the scatterplot of the summed effect of 10 + 01 and the observed effect of 11(figure(c)). The majority of points fall below the red dashed line, meaning that the beta estimates for 11 are generally lower than the summed values of 10 + 01. This deviation suggests that **enhancer interactions may not be purely additive and could involve regulatory limitations or saturation effects.**

(a) bootstrapped p_value from SCENT output



(b) Interaction Effects



(c) Beta of enhancer 1 + enhancer 2 vs. enhancer both

Discussion and Future Work

HPC Environment Limitation

1. The standard HPC job limit of 48 hours often constrained pipeline execution, especially for larger datasets. Larger cell types with extensive datasets frequently exceeded the time limit, causing incomplete runs.
2. Large input files demanded significant resources and time, which caused delays on debugging processes.
3. Certain tasks required older Python methods, which were unavailable in the default HPC environment. Although some packages were successfully installed locally, cluster execution occasionally failed to locate them. To address this, specific methods were sourced from open-source Python GitHub repositories without altering the original code.
4. Snakemake, which we had built our pipeline upon was also unstable, either tolerating incomplete rule outputs or halting prematurely.

Future Work

The pipeline and analysis have only been tested on one cell type out of 30 due to limited time and resources. For future work, it would be beneficial to test all cell types to gain a more accurate and comprehensive conclusion. Furthermore, the current analysis focused on identifying pairwise enhancer interactions using generalized linear models. Future work could explore more sophisticated models, such as non-linear models (e.g., neural networks or tree-based methods) to capture complex enhancer synergies.