GRAIGH: Gene Regulation accessibility integrating GeneHancer database

GRAIGH: Gene Regulation accessibility integrating GeneHancer database

Lorenzo Martini
Email: lorenzo.martini@polito.it
0000-0002-7794-7791

Roberta Bardini
Email: roberta.bardini@polito.it
0000-0002-1809-3212

Alessandro Savino
Email: alessandro.savino@polito.it
0000-0003-0529-7950

Stefano Di Carlo
Politecnico di Torino Control and Computer Engineering Department
Torino,10129, Italy
Email: stefano.dicarlo@polito.it
0000-0002-7512-5356

***Abstract*— Single-cell assays for transposase-accessible chromatin sequencing data represent a potent tool for exploring the epigenetic heterogeneity within cell populations. Despite their power, understanding the chromatin accessibility landscape poses challenges. This study introduces Gene Regulation Accessibility Integrating GeneHancer (GRAIGH), a novel approach to interpreting genome accessibility by integrating information from the GeneHancer database, detailing genome-wide enhancer-to-gene associations. Initially, we outline the methods for integrating GeneHancer with scATAC-seq data. This involves creating a new matrix where GeneHancer element IDs replace traditional accessibility peaks as features. Subsequently, the paper assesses the method's ability to analyze data and detect cellular heterogeneity. Notably, our findings demonstrate the selective accessibility of GeneHancer elements for distinct cell types, with connected genes serving as precise marker genes. Furthermore, we explore the specificity of GeneHancer element accessibility, highlighting their high selectivity against gene activity. This investigation underscores the potential of Gene Regulation Accessibility Integrating GeneHancer in unraveling the complexities of chromatin accessibility, offering insights into the nuanced relationship between accessibility and cellular heterogeneity.**

***Index Terms*—Bioinformatics, scATAC-seq, Enhancers, GeneHancer**

I. Introduction

Unraveling the complexity of cellular biology poses a fundamental challenge in the field [1]. Despite continuous discoveries, many aspects, particularly those related to DNA and transcription mechanisms, remain elusive but are pivotal for comprehending cell identities and functions [2]. Gene expression plays a crucial role in cellular homeostasis, particularly in multicellular organisms where it influences distinct functionalities [3]. For studying cell type heterogeneity and its relevance to various pathologies, transcriptomic analysis proves invaluable [4].

Recent advances in experimental techniques, such as Next-Generation Sequencing (NGS) and, specifically, single-cell RNA sequencing (scRNA-seq), enable high-resolution profiling of gene expression at the single-cell level [5] [6]. However, while powerful, these data do not provide a comprehensive view of the intricate regulatory mechanisms underlying gene transcription [7]. On the other hand, single-cell assays for transposase-accessible chromatin sequencing (scATAC-seq) data presents unique challenges. Unlike scRNA-seq, scATAC-seq features are experiment-dependent genomic coordinates, complicating dataset comparisons and the interpretation of accessible genomic regions [8].

This work introduces Gene Regulation Accessibility Integrating GeneHancer (GRAIGH), a novel computational approach designed to interpret scATAC-seq features and extract meaningful information. GRAIGH aims to integrate scATAC-seq datasets with the GeneHancer database, detailing genome-wide enhancer-to-gene and promoter-to-gene associations. This integration helps overcome scATAC-seq data limitations. Our study demonstrates the strength of using GeneHancer associations to investigate cellular heterogeneity with higher specificity compared to other methods like Gene Activity (GA).

II. Background

scATAC-seq is a powerful technique for probing chromatin accessibility at the single-cell level [9]. In a typical scATAC-seq dataset, each column represents a single cell, and each row corresponds to a specific genomic locus. Binary values in the matrix indicate the chromatin accessibility status of each genomic locus in individual cells. However, defining peaks as regulatory elements is a computational process heavily reliant on experimental parameters. Consequently, these experiment-specific peaks lack the well-defined characteristics of genes [10] [11].

This variability in peak definitions poses challenges when comparing and interpreting scATAC-seq data from different sources. Moreover, linking genes and their transcription to specific peaks is not straightforward. To address this, a common approach is to use a Gene Activity Matrix (GAM), encoding genomic accessibility information to transform features into genes. However, existing methods often narrowly focus on well-defined gene body and promoter regions, accounting for less than 20% of the total epigenetic information [12]. This

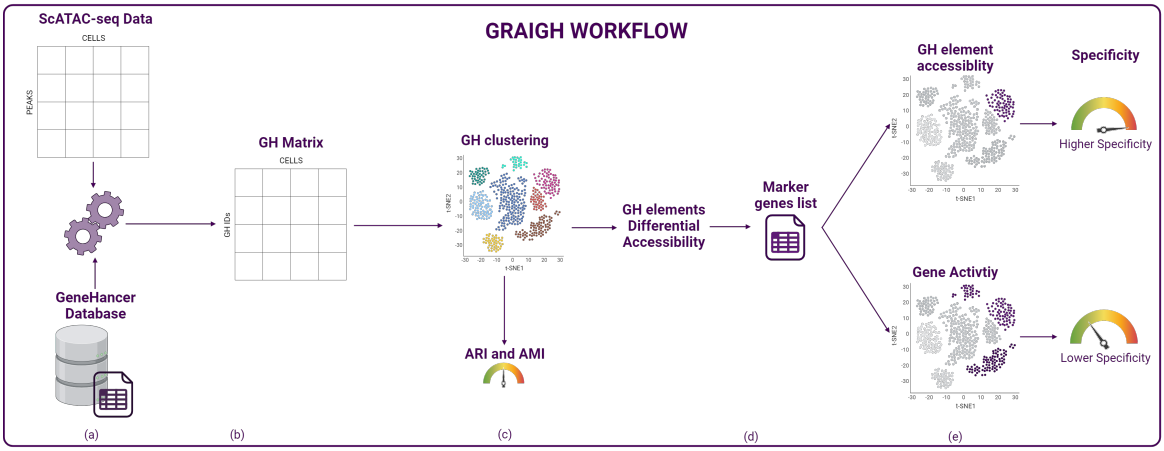


Fig. 1: Workflow of Gene Regulation Accessibility Integrating GeneHancer (GRAIGH). (a) Integration of the scATAC-seq matrix, using peaks as features, with the GeneHancer database. (b) Generation of the GH Matrix with GH_{el} IDs as features. (c) Processing of the GH Matrix to obtain unsupervised clustering. A comparison with clustering from the original scATAC-seq data demonstrates that the GH matrix introduces no critical biases. (d) Conducting Differential Accessibility Analysis of GH_{el}, revealing enhancer elements associated with known marker genes. (e) Comparative analysis of gene activity and GH_{el} accessibility for specific marker genes. Results indicate that GH_{el} exhibits higher specificity.

focus results in the loss of valuable epigenetic data, particularly related to enhancer regions, which remain unlinked to their respective target genes.

The GenHancer database [13] emerges as a potential solution. GenHancer provides associations between enhancers, promoter elements, and their corresponding genes. In the following sections, this paper introduces an innovative method (as illustrated in Fig. 1) for integrating GenHancer with scATAC-seq data. The study aims to demonstrate the efficacy of this integration in interpreting and analyzing scATAC-seq data, providing a promising solution to current challenges in the field.

III. Materials and Methods

A. Materials

This study utilizes the most recent version of GeneHancer and applies the proposed methodology to a publicly accessible 10xGenomic scATAC-seq dataset, comprising 10,246 human peripheral blood mononuclear cells (PBMC) cells and encompassing 165,434 peaks [14].

B. GeneHancer

GeneHancer [13], a component of the GeneCard database [15], provides comprehensive insights into human genes. It offers genome-wide enhancer-to-gene and promoter-to-gene associations, covering up to 18% of the genome. These regulatory elements result from a thorough cross-source investigation involving nine data sources, ensuring reliable, non-redundant information with functional annotations. Each GeneHancer element (GH_{el}) uniquely corresponds to a regulatory element, identified by genomic coordinates, connected genes, and a confidence score. 'Elite' connections undergo multiple source verification. A limitation is that the database is specific to the hg38 genome version, restricting its use to matching scATAC-seq data. The current version comprises 393,464 GH_{el} and 2,408,198 GH-gene connections, encompassing 18% of the genome.

C. scATAC-seq data processing

The workflow for processing a scATAC-seq dataset, represented as matrix $\mathbf{A}_{|P|\times|C|}$ (with P denoting peaks and C cells), follows established procedures [16]. This study employs the widely-used R package Seurat for data processing.

The processing initiates with peak calling, data normalization, scaling, and dimensionality reduction using Latent semantic indexing (LSI), chosen over PCA for this data type [17]. Further dimension reduction is achieved using Uniform Manifold Approximation and Projection (UMAP) (or Stochastic Neighbor Embedding (tSNE)) to provide a two-dimensional dataset representation. Unsupervised clustering reveals cellular heterogeneity but not cell types. Seurat integrates cross-modality data and classifies cells using labels from an external pre-labeled single-cell RNA sequencing (scRNA-seq) dataset of the same biological sample. The reference dataset used in this study is a PBMC scRNA-seq dataset recommended by Seurat guidelines, available at [18]. These classification results serve as a ground truth for comparisons.

As discussed in Section II, GAM is a standard tool for processing epigenetic data. A GAM is represented as matrix $\mathbf{A}^{act}_{|G|\times|C|}$ (with G for genes and C cells), measuring gene accessibility. Signac, a companion package of Seurat [19], computes GAM. However, it is important to note that Signac, like other GAM methods, calculates gene activity solely from gene bodies and imputed promoters, disregarding enhancer regions [20] [21].

D. GH matrix creation and processing

The proposed idea in this paper for integrating GeneHancer with a scATAC-seq dataset involves creating a new matrix where features (rows) represent GH_{el} and columns represent cells. The algorithm for creating this matrix receives the list of GH_{el} ($\mathbf{GH_{el}} = GH_{el1}, GH_{el2}, ..., GH_{elN}$) and the set of peaks ($P = p_1, p_2, ..., p_N$) genomic coordinates and generates an association matrix $\mathbf{GP}_{|\mathbf{GHel}|\times|P|}$ such that each element is one if a peak overlaps a GH_{el}. This process translates the epigenetic features (i.e., peaks) into something univocally defined and comparable with other experiments (i.e., GH_{el}). Multiplying the $\mathbf{GP}_{|\mathbf{GHel}|\times|P|}$ matrix by the $\mathbf{A}_{|P|\times|C|}$ matrix produces the $\mathbf{GH}_{|\mathbf{GH_{el}}|\times|C|}$ matrix, providing a new representation of the original scATAC-seq data that can be processed with the same workflow exposed in Section III-C. This procedure leads to a new 2D visualization and clustering of the cells. Comparing the clustering obtained with the two matrices is essential to demonstrate that the analysis of scATAC-seq datasets can be reliably performed on the GH data without losing information. There is a significant benefit of employing uniquely identified and interoperable features (the GH_{el} elements), which carry meaningful biological insights.

E. Differentially accessible features and their specificity

In addition to proposing a novel representation for scATAC-seq data, this paper investigates whether integrating GeneHancer can improve support for exploring cellular heterogeneity. Given that GH_{el} has well-defined connections to genes, it could potentially assist in identifying cell types within the dataset, akin to the use of gene markers in scRNA-seq data analysis.

Central to this approach is differential analysis, a well-established method in single-cell analysis pipelines for identifying distinguishing features among cell groups or clusters. In scRNA-seq data, Differential Expression (DE) analysis targets genes, while in scATAC-seq data, it focuses on peaks in Differential Accessibility (DA) analysis. Utilizing the Seurat suite, this study conducts DA analysis on GH_{el} between cell types, yielding a list of top GH_{el} with the highest average log-fold change for each cell type.

Subsequently, the study examines these elements by retrieving their associated genes with elite connections, including known marker genes. This assessment highlights the coherence of DA GH_{el} with specific cell types, providing insights into dataset cell heterogeneity. Additionally, the research introduces a quantitative analysis to explore the relationship between GH_{el} and cell types, enhancing the characterization of both components.

In the context of single-cell RNA sequencing (scRNA-seq) data, examining the expression patterns of well-established marker genes is a common method for deciphering the diversity of cell populations within a dataset. These markers serve to distinguish distinct clusters as specific cell types. Given this, conducting a similar investigation using the GeneHancer matrix becomes intriguing.

Using an approach similar to the one described in the preceding section, this research compiles a list of established marker genes and identifies the associated GH_{el} with strong connections. These GH_{el} entities are expected to be uniquely accessible to the same cell type, holding potential as epigenetic markers across various experimental setups.

However, it could be argued that the Gene Activity Matrix (GAM) accomplishes a comparable inquiry with fewer steps by directly assessing the activity of marker genes. Consequently, this study demonstrates that employing the GH_{el}-based approach yields superior outcomes in recognizing distinct cell types than relying solely on gene activity evaluations. Indeed, one relevant characteristic of a cell type marker for heterogeneity investigation is its specificity [22], defined as:

$$\textit{Specificity} = \frac{\text{True Negatives (TN)}}{\text{True Negatives (TN)} + \text{False Positives (FP)}} \quad (1)$$

The concept of "high specificity" refers to a characteristic where a feature utilized to distinguish a particular cell type is exclusively present in cells belonging to that specific type. This study assesses the level of specificity in terms of both the accessibility of GH_{el} and the activity of literature marker genes for various cell types derived from the Seurat integration, subsequently drawing comparisons between them.

When dealing with a marker gene, the initial step involves selecting the GH_{el} associated with that particular gene. Since a single gene can be linked to multiple GH_{el}, this research focuses on the elite GH_{el} generated through DE analysis, a collection referred to as $\mathbf{GH^g_{el}}$. Subsequently, for a given marker gene g corresponding to a specific cell type t, the activity vector of g denoted as $\mathbf{A}_{g\times|C|}$, the vector of cell-type labels CT, and the accessibility vector of each associated GH_{el}, denoted as $\mathbf{GH}_{gh\times|C|}$, are binarized. With the cell-type labels vector CT serving as the reference truth and the activity vector as predictive data, the study calculates the specificity pertaining to the activity. Subsequently, for each element gh within the $\mathbf{GH^g_{el}}$ list, the algorithm computes its specificity, thereby establishing the accessibility specificity as the average value across all elements.

In conclusion, the algorithm computes the disparity between the specificity concerning GH_{el} accessibility and activity.

IV. Results

This section presents results from applying the proposed methods to the previously mentioned PBMC dataset. The aim is to demonstrate that a GH matrix is equivalent to the original data and can be used for cell heterogeneity studies with the advantage of easy comparison among multiple datasets.

Fig. 3a illustrates the cell clustering performed on the original data. Cells are grouped into two significant populations, and some smaller groups, aligning with expectations for this sample type [17]. The unsupervised clusterization algorithm divides the cells into 20 clusters. This result aligns with the cell-type classification obtained with the Seurat integration shown in Fig. 3b. The algorithm identifies T cells (subdivided into subtypes) as the major population, followed by Monocytes. The two smaller groups represent Natural Killer (NK)

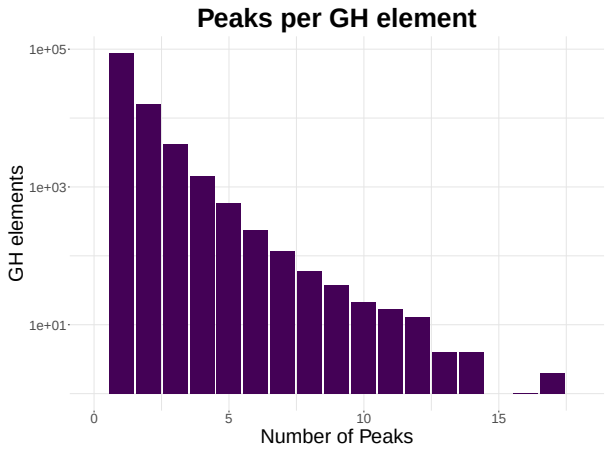


Fig. 2: Barplot of Peaks per GH_{el}. The y-axis is in logarithmic scale. As the plot shows, the great majority of GH_{el} have one or few peaks overlapping them. However, in many cases, the GH_{el} have many peaks overlapping them since they are longer than the peaks.

cells and B cells, concluding with a few cells labeled Dendritic cells. The proposed approach overlaps the GH_{el} with the coordinates of the peaks, creating the connection matrix. Fig. 2 shows the distribution of how many peaks are connected to the GH_{el}. Most GH_{el} have one or few peaks overlapping them, with a reduced number overlapping many peaks (up to 27). The reason for many overlaps stems from the length of the peaks being much smaller than the GH_{el}. After filtering out the GH_{el} with no overlaps, 109,620 GH_{el} remains, generating the final 109,620x10,246 GH matrix.

Fig. 4 displays a 2D representation of the data obtained from processing the GH matrix, showcasing cluster patterns similar to those illustrated in Fig. 3a. In this instance, the unsupervised clustering algorithm has identified 19 clusters, and the evident resemblance among the identified clusters can be easily discerned. Additionally, it is feasible to assess their similarity using metrics such as Adjust Rand Index (ARI) and Adjust Mutual Information (AMI), commonly employed to gauge classification similarities [17] [22].

When these metrics are computed for the two clustering results, they yield an AMI of 0.867 and an ARI of 0.804, underscoring a significant likeness between the two clusterings. This outcome highlights the credibility of the GH matrix approach in producing comparable results to the original data while avoiding noticeable biases. Consequently, it confirms the effectiveness of this approach, which leverages biologically meaningful features to not only appropriately analyze scATAC-seq data but also directly compare results across different experiments.

In addition to the previous results, further experiments have been conducted to explore the effectiveness of the GeneHancer (GH_{el}) and, more broadly, the utility of this technique in analyzing cell heterogeneity. These results are also compared to those of the Gene Activity Matrix (Gene Activity Matrix). During the processing of the initial dataset, a GAM comprising 19,607 genes is generated. This specific GAM serves as the benchmark against which the final analysis is evaluated.

The DA analysis on the GH matrix identifies 76,081 imputed DA GeneHancers (GH_{el}), of which 73,235 are positive markers, indicating their positive characterization of the corresponding cell type. Investigation into some of these elements reveals intriguing GH_{el}. For instance, elements GH02J086783 and GH02J086805, specific to CD8 subtype cells, serve as elite enhancers for CD8A and CD8B genes, recognized markers of the homonymous cells. A similar analysis applies to GH05J140611 and GH05J140596, differentially accessible for Monocytes and elite enhancers of the CD14 gene.

These results demonstrate that the DA GH_{el} are selectively accessible for distinct cell types. Furthermore, the genes connected to them serve as markers for these distinct cell types. This highlights the coherence of accessibility of regulatory elements and known cell types, providing crucial information for cellular heterogeneity investigation.

This study examines the discriminative potential of GH_{el} linked to cell-type marker genes in analyzing cell heterogeneity. It explicitly compares the specificity of GH_{el} accessibility to the specificity of their target marker gene activity. Table I presents these specificity values for the considered marker genes, along with their differences.

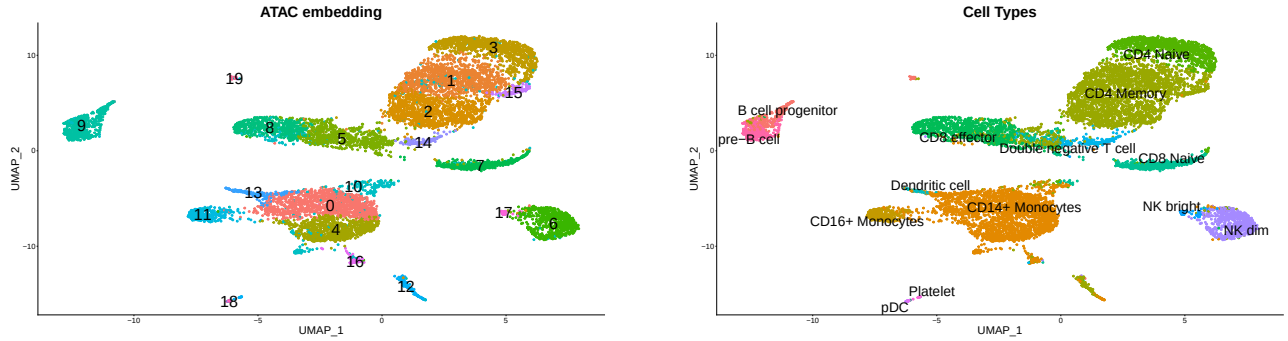
The reported results are intriguing. Firstly, all the differences are positive, indicating that GH_{el} specificity is consistently higher than activity specificity. Moreover, the differences are particularly significant, especially for more populated cell types like T cell subtypes and Monocytes, reaching values of 0.266 and 0.394, respectively. The difference is lower for smaller populations, such as Dendritic and B cells, where well-separated subtypes tend to be well-defined at the activity level. However, their GH_{el} specificity remains consistently higher than 0.9, demonstrating the method's reliability.

This becomes even more evident when visualizing the features on the dataset. Fig. 5 displays both the accessibility of GH05J140611 (a), an enhancer element of the CD14 gene marker for Monocytes, and the gene's activity (b). It is immediately apparent how the gene activity spreads throughout the dataset, while the accessibility of its enhancer element is specific to the Monocytes population. Similarly, Fig. 6 presents the same representation for the CD4 gene and the accessibility of the connected element GH12J006784. In this case as well, the accessibility of the GH_{el} is more specific than the activity.

These results show how the use of GH_{el} as features is a reliable way to interpret the scATAC-seq data, and, more importantly, they have greater specificity in detecting the heterogeneity of the dataset than the gene activity.

V. Conclusions

In conclusion, GRAIGH introduces a novel approach for interpreting scATAC-seq data. This technology, revealing chromatin accessibility at the single-cell level, poses challenges due to the absence of well-defined features akin to genes in scRNA-seq. This study integrates scATAC-seq data with the



(a) (b)

Fig. 3: (a) UMAP visualization obtained from processing the original scATAC-seq data, with its unsupervised clusterization. (b) Same UMAP embedding but colored with the cell-type labels obtained from the Seurat label transfer integration.

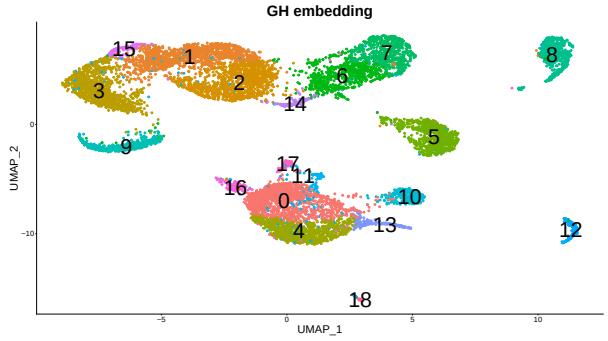


Fig. 4: UMAP embedding and clustering result, obtained from the processing of the GH matrix. Comparing with Fig. 3.a, it is evident that the clustering subdivision of the cells is mostly coherent, demonstrating that the GH matrix does not introduce relevant biases and can be used to process scATAC-seq data.

TABLE I: The table reports for each main cell type and marker gene the specificity of the gene activity and the mean of the GH_{el} accessibility. The last row is the difference which highlights how the GH_{el} accessibility has always higher specificity.

Cell Type	Marker	Accessibility	Activity	Δ
CD4+ T cells	CD4	0.918	0.523	0.394
	CCR7	0.795	0.448	0.346
	IL7R	0.815	0.505	0.309
CD8+ T cells	CD8A	0.854	0.633	0.221
Monocytes	CD14	0.988	0.722	0.266
	MS4A7	0.935	0.749	0.185
NK cells	GNLY	0.962	0.870	0.090
Dendritic	FCER1A	0.987	0.914	0.073
	CST3	0.947	0.861	0.086
B cells	MS4A1	0.957	0.928	0.028

GeneHancer database, delineating genome-wide enhancer-to-gene and promoter-to-gene associations. The unique identifiers of GH_{el} ensure interoperability across diverse datasets.

This research demonstrates that integrating GeneHancer data with scATAC-seq and using GH_{el} as features is a robust method for exploring single-cell epigenomic data. The approach is validated by comparing results with the original scATAC-seq data, revealing no significant biases. Furthermore, the paper underscores that GH_{el} accessibility corresponds to specific cell types. By analyzing GH_{el} accessibility of known marker genes, it surpasses traditional gene activity analysis in identifying cell types.

However, limitations exist. GeneHancer data is currently exclusive to the human genome, and understanding transcriptomic regulation remains a complex, evolving process. Future work may incorporate motif information from the peaks and leverage multi-omic datasets to explore the correlation between GH_{el} accessibility and gene expression.

In summary, GRAIGH offers a valuable means of interpreting scATAC-seq data, shedding light on intricate regulatory mechanisms underlying cellular heterogeneity. This method opens new avenues for understanding gene regulation, cellular dynamics, and their relevance to various medical conditions. In conclusion, integrating scATAC-seq data with the GeneHancer database is a promising step toward unraveling the complexities of cellular biology at the epigenomic level.

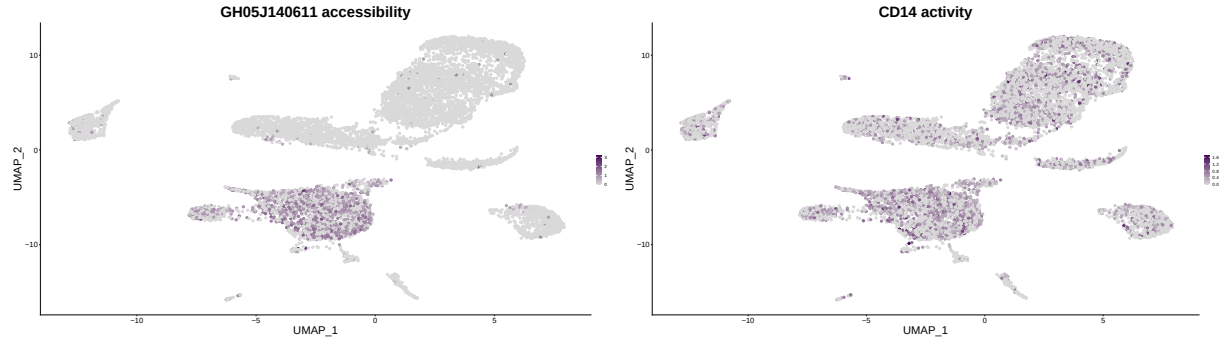
VI. Data and Code Availability

GeneCard allows direct download of the older database 2017 version https://www.genecards.org/GeneHancer_Version_4-4, but it is possible to request the access to the latest versions from the online platform https://www.genecards.org/Guide/DatasetRequest. The 10X genomics dataset is freely available at https://www.10xgenomics.com/resources/datasets/10k-human-pbmcs-atac-v2-chromium-controller-2-standard. All the code employed in this study is publicly available on the GitHub repository at https://github.com/smilies-polito/GRAIGH.

References

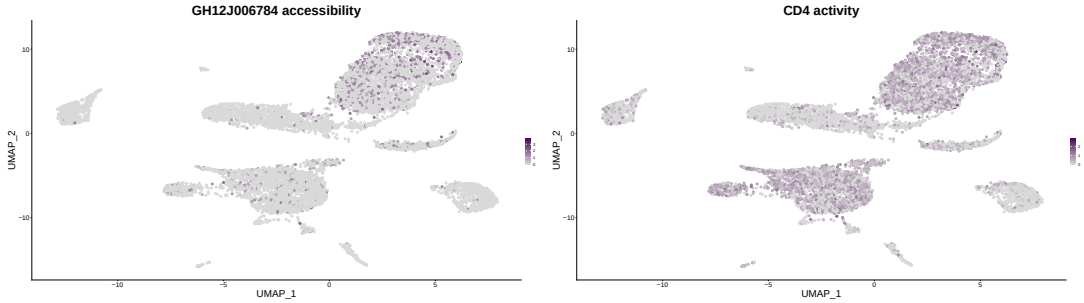
[1] S. J. Altschuler and L. F. Wu, "Cellular heterogeneity: Do differences make a difference?" *Cell*, vol. 141, no. 4, p. 559–563, 2010.

[2] Q. Chen, J. Shi, Y. Tao, and M. Zernicka-Goetz, "Tracing the origin of heterogeneity and symmetry breaking in the early mammalian embryo," *Nat. Commun.*, vol. 9, no. 1, 2018.



(a) (b)

Fig. 5: The CD14 gene is a marker for the Monocytes. The accessibility of its enhancer GH05J140611 is specifically accessible in the monocyte population (a), while its gene activity (b) is more spread out in many other cells.



(a) (b)

Fig. 6: The CD4 gene is a marker for the homonymous T cells. Its enhancer GH12J006784 is specifically accessible in the CD4+ T cells population (a), while its gene activity (b) is more spread out in many other cells.

[3] A. S. N. Seshasayee, "Gene expression homeostasis and chromosome architecture," *Bioarchitecture*, vol. 4, no. 6, pp. 221–225, 2014.

[4] S. L. Goldman, M. MacKay, E. Afshinnekoo, A. M. Melnick, S. Wu, and C. E. Mason, "The impact of heterogeneity on single-cell sequencing," *Front. Genet.*, vol. 10, p. 8, 2019.

[5] X. Li and C.-Y. Wang, "From bulk, single-cell to spatial RNA sequencing," *Int. J. Oral Sci.*, vol. 13, no. 1, p. 36, 2021.

[6] A. Haque, J. Engel, S. A. Teichmann, and T. Lönnberg, "A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications," *Genome Med.*, vol. 9, no. 1, 2017.

[7] A. Moosavi and A. Motevalizadeh Ardekani, "Role of epigenetics in biology and human diseases," *Iran. Biomed. J.*, vol. 20, no. 5, pp. 246–258, 2016.

[8] 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 Biol.*, vol. 21, no. 1, p. 22, 2020.

[9] G. Kelsey, O. Stegle, and W. Reik, "Single-cell epigenomics: Recording the past and predicting the future," *Science*, vol. 358, no. 6359, pp. 69–75, 2017.

[10] H. Jeon, H. Lee, B. Kang, I. Jang, and T.-Y. Roh, "Comparative analysis of commonly used peak calling programs for ChIP-Seq analysis," *Genomics Inform.*, vol. 18, no. 4, p. e42, 2020.

[11] Z. Ji, W. Zhou, W. Hou, and H. Ji, "Single-cell ATAC-seq signal extraction and enhancement with SCATE," *Genome Biol.*, vol. 21, no. 1, p. 161, 2020.

[12] L. Martini, R. Bardini, A. Savino, and S. Di Carlo, "Meta-analysis of gene activity (maga) contributions and correlation with gene expression, through gagam," in *Bioinformatics and Biomedical Engineering*. Springer Nature Switzerland, 2023, pp. 193–207.

[13] S. Fishilevich, R. Nudel, N. Rappaport, R. Hadar, I. Plaschkes, T. Iny Stein, N. Rosen, A. Kohn, M. Twik, M. Safran, D. Lancet, and D. Cohen, "GeneHancer: genome-wide integration of enhancers and target genes in GeneCards," *Database (Oxford)*, vol. 2017, 2017.

[14] 10XGenomics, "10k cryopreserved human peripheral blood mononuclear cells (pbmcs) from a healthy donor single cell atac dataset by cell ranger atac 2.1.0, 10x genomics, (2022, march 29th)."

[15] G. Stelzer, N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, T. I. Stein, R. Nudel, I. Lieder, Y. Mazor, S. Kaplan, D. Dahary, D. Warshawsky, Y. Guan-Golan, A. Kohn, N. Rappaport, M. Safran, and D. Lancet, "The GeneCards suite: From gene data mining to disease genome sequence analyses," *Curr. Protoc. Bioinformatics*, vol. 54, no. 1, pp. 1.30.1–1.30.33, 2016.

[16] B. Hwang, J. H. Lee, and D. Bang, "Single-cell RNA sequencing technologies and bioinformatics pipelines," *Exp. Mol. Med.*, vol. 50, no. 8, pp. 1–14, 2018.

[17] H. Chen, C. Lareau, T. Andreani, M. E. Vinyard, S. P. Garcia, K. Clement, M. A. Andrade-Navarro, J. D. Buenrostro, and L. Pinello, "Assessment of computational methods for the analysis of single-cell ATAC-seq data," *Genome Biol.*, vol. 20, no. 1, p. 241, 2019.

[18] T. Stuart, "Seurat reference dataset." [Online]. Available: https://signac-objects.s3.amazonaws.com/pbmc_10k_v3.rds

[19] S. R. Stuart T. *et al.*, "Single-cell chromatin state analysis with signac," *Nature Methods*, 2021.

[20] L. Martini, R. Bardini, A. Savino, and S. Di Carlo, "GAGAM v1.2: An improvement on peak labeling and genomic annotated gene activity matrix construction," *Genes (Basel)*, vol. 14, no. 1, p. 115, 2022.

[21] ——, "GAGAM: A genomic annotation-based enrichment of scATAC-seq data for gene activity matrix," in *Bioinformatics and Biomedical Engineering*. Cham: Springer International Publishing, 2022, pp. 18–32.

[22] L. Martini, R. Bardini, and S. Di Carlo, "Meta-Analysis of cortical inhibitory interneurons markers landscape and their performances in scRNA-seq studies," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2021.