POLITECNICO DI TORINO
Repository ISTITUZIONALE

GAGAM v1.2: An Improvement on Peak Labeling and Genomic Annotated Gene Activity Matrix Construction

Original
GAGAM v1.2: An Improvement on Peak Labeling and Genomic Annotated Gene Activity Matrix Construction / Martini, L., Bardini, R., Savino, A., DI CARLO, S.. - In: GENES. - ISSN 2073-4425. - ELETTRONICO. - 14:1(2023). [10.3390/genes14010115]

Availability:
This version is available at: 11583/2974377 since: 2023-01-05T16:34:05Z

Publisher:
MDPI

Published
DOI:10.3390/genes14010115

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

25 November 2026

genes

MDPI

Article

GAGAM v1.2: An Improvement on Peak Labeling and Genomic Annotated Gene Activity Matrix Construction

Lorenzo Martini *, Roberta Bardini, Alessandro Savino and Stefano Di Carlo

Politecnico di Torino, Control and Computer Engineering Department, 10129 Torino, Italy
* Correspondence: lorenzo.martini@polito.it

Abstract: Single-cell Assay for Transposase-Accessible Chromatin using sequencing (scATAC-seq) is rapidly becoming a powerful technology for assessing the epigenetic landscape of thousands of cells. However, the sparsity of the resulting data poses significant challenges to their interpretability and informativeness. Different computational methods are available, proposing ways to generate significant features from accessibility data and process them to obtain meaningful results. Foremost among them is the peak calling, which interprets the raw scATAC-seq data generating the peaks as features. However, scATAC-seq data are not trivially comparable with single-cell RNA sequencing (scRNA-seq) data, an increasingly pressing challenge since the necessity of multimodal experiments integration. For this reason, this study wants to improve the concept of the Gene Activity Matrix (GAM), which links the accessibility data to the genes, by proposing an improved version of the Genomic-Annotated Gene Activity Matrix (GAGAM) concept. Specifically, this paper presents GAGAM v1.2, a new and better version of GAGAM v1.0. GAGAM aims to label the peaks and link them to the genes through functional annotation of the whole genome. Using genes as features in scATAC-seq datasets makes different datasets comparable and allows linking gene accessibility and expression. This link is crucial for gene regulation understanding and fundamental for the increasing impact of multi-omics data. Results confirm that our method performs better than the previous GAMs and shows a preliminary comparison with scRNA-seq data.

Keywords: epigenomic single-cell data; gene activity matrix; bioinformatics

check for updates

Citation: Martini, L.; Bardini, R.; Savino, A.; Di Carlo, S. GAGAM v1.2: An Improvement on Peak Labeling and Genomic Annotated Gene Activity Matrix Construction. *Genes* **2023**, *14*, 115. https://doi.org/10.3390/genes14010115

Academic Editors: Alexandre G. de Brevern, Ignacio Rojas, Francisco Ortuño, Alfredo Benso, Jean-Marc Schwartz and Olga Valenzuela

Received: 21 October 2022
Revised: 20 December 2022
Accepted: 24 December 2022
Published: 30 December 2022

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Recent advances in Next Generation Sequencing (NGS) technologies paved the way for single-cell multi-omics data analysis, which captures different facets of cells' regulative state, including the epigenome, the genome, the transcriptome, and the proteome [1]. Multi-omics approaches increase resolution and sensitivity in the characterization of cellular states, the identification of known or new cellular phenotypes, and the understanding of cell dynamics [2], supporting a quantitative and a comprehensive approach to studying cellular heterogeneity [3].

In particular, the combination of transcriptomic and epigenomic data provides integrated information on the functional activation of genes and the structural organization of chromatin. There are different experimental approaches to generating epigenomic data. These include accessibility measurements, which indicate whether chromatin is open or closed at genomic locations, exposing other genomic regions for transcriptional and regulatory processes [4]. These data have a very different organization than transcriptomic data indicating the expression level of genes.

In general, analyzing data simultaneously from multiple omics does not guarantee a gain in the informativeness of the analysis of the cellular system. Different mechanisms underlie each omic layer, and complex relations link different layers. Purely data-driven multi-omic combinations overlook the connections between omic layers. These methods fail to provide a systemic analysis and can also lead to inaccurate assumptions on such

relations [5]. Furthermore, it is well-known that expression levels show almost no correlation to protein quantity in the cell. This is due to complex and non-obvious regulatory dynamics between transcriptional and translational regulations. In these dynamics, it is not sure that a large pool of mRNA for a gene will yield a large quantity of protein and vice-versa. Due to the uncertainty and complexity of biological dynamics, it is not trivial to treat transcriptomic and proteomic data under the assumption that the way they measure functional activation is comparable. Epigenomic data allow us to extrapolate a measure of transcriptional activity at the relevant regions for specific genes. This is often considered a predictor of the expression levels of the same gene. Yet, drawing a linear relation between activity and expression levels is far from trivial. Accessibility is a necessary but not sufficient condition for a gene for its transcription: a completely accessible gene can show a full range of expression levels, including null expression. This does not support the assumption that the relation between accessibility and expression is linear in multi-omics analyses [6]. Thus, building better models of the relations among the different omic layers is pivotal to truly leverage multi-omics analysis and gain a systemic understanding of the cellular state. Multi-omics analysis strategies must combine data-driven and model-driven approaches to account for the complex relations among different omic layers explicitly and formally [5].

Focusing on the relations between the epigenomic and transcriptomic layers, the concept of gene activity, i.e., the overall accessibility of a gene allowing its transcription inside the cell [7], is an example of a data- and model-driven approach which facilitates the joint analysis of epigenomic and transcriptomic data. Chromatin accessibility refers to the structural organization of the DNA strands in eukaryotic genomes. Chromatin organization changes across cell types and cell differentiation phases. On a genome-wide scale, epigenetic regulations, by affecting chromatin openness, determine the accessibility of genomic regions to transcriptional processes. Complex patterns of chromatin occupancy by such epigenetic factors determine inverse patterns of chromatin accessibility, yielding structural information about binding of the transcriptional machinery and transcriptional regulators at the DNA [8]. Accessible regions can occur on coding regions (i.e., exons or introns), promoters, Transcription Starting Sites (TSS), and cis- and trans-regulatory elements, such as enhancers. Each of these types of regions has a different implication on the activity of a gene, linked to the functional effect determined by the effective binding of their specific binding proteins. After the assumption that regulatory elements that exhibit higher effective activation also show higher accessibility, epigenomic data can support the genome-wide identification of genomic regions that can discriminate different cell types or states at a functional level [9]. While epigenetic regulation can completely subtract genomic portions from transcriptional activity, an open genomic region undergoes more nuanced epigenetic regulations with both positive and negative implications for transcriptional activity. Accessibility patterns can make the same region more or less accessible and, depending on the functionality of accessible elements, have different functional consequences on the transcriptional process. Considering the fundamental limitation that accessibility is a necessary but not sufficient condition for a gene to be transcribed and acknowledging the different contributions of coding and non-coding accessible regions to the activity of a gene provide suitable premises to build a model of gene activity to support the multimodal analysis of epigenomic and transcriptomic data.

Current approaches build a bridge between transcriptomics and epigenomics, supporting a direct comparison of gene activity and gene expression levels. In particular, GAMs [7] provide an effective way to summarize accessibility information deriving from single-cell experiments in a score for gene activity. A GAM, similarly to an expression matrix, has a subset of genes from the reference genome as features and cells as data points. Each element of the matrix ($GAM_{g,c}$) represents the Gene Activity Score (GAS) of a gene g in a cell c [7], based on the gene activity model of choice. Using the same set of genes in expression and activity matrices makes transcriptomic and epigenomic data easy to analyze with the same pipelines. Yet, as better explained in Section 2, even if current approaches to

compute gene activity in GAMs aim to derive a measure of activity from accessibility data, they show limitations in capturing the contextual meaning and the regulative implications of epigenomic data. This hampers the consistent integration of activity and expression and, thus, a robust approach to epigenomic and transcriptomic multimodal analysis.

To overcome this limitation, this work introduces Genomic-Annotated Gene Activity Matrix (GAGAM) v1.2, a data- and model-driven analysis strategy for epigenomic data. It combines accessibility data with genomic annotations. This combination allows us to weigh the contributions of different accessible regions to the overall gene activity and functional significance. This model is an extension of a preliminary model published by Martini et al. (GAGAM v1.0) that was already an improvement to other GAM-based approaches. GAGAM v1.0 included an explicit, more comprehensive, and modular model of the different contributions to gene activity from the different annotated region types [10]. GAGAM v1.2 improves differential activity analyses' resolution, explainability, and interpretability, supporting the study of cellular heterogeneity based on accessibility data alone. The GAS model underlying GAGAM v1.2 improves the precision in weighting contributions of accessible regions and the specificity of their attribution to each gene in the matrix. The capability of GAGAM v1.2 was extensively validated, showing improved performance compared to GAGAM v1.0 on the same datasets and comparable performance on newly acquired datasets. Moreover, this work provides preliminary results on the analytical comparison between epigenomic and transcriptomic data based on GAGAM v1.2 and expression matrices from multimodal datasets.

2. Background

The availability of data from multiple omic pools is the enabler of multimodal single-cell analysis and catalyzed a revolution in medicine and biology. These data paved the way for integrated, systemic, and data-driven approaches to biological complexity [5,11]. The multimodal analysis includes the integration of data from different experiments considering different omics, as well as experiments using multi-omic sequencing technologies. Multi-omic sequencing technologies sample multiple omic pools from the same cells, characterizing their state more comprehensively than the sequencing of single omics. Integrating information from more than one omic pool in the same cell ensures higher consistency, increasing reliability and accuracy of cellular heterogeneity studies, cell type identification, and inference of differentiation trajectories.

While gaining information richness, multi-omic approaches can lose throughput. As a result, their overall throughput is bound to the lowest throughput among the sequencing technologies employed.

Considering the integration of transcriptomic and epigenomic data, this issue emerges clearly. Single-cell transcriptomics is the most mature among single-cell methods. As illustrated in [12], sequencing technologies for single-cell transcriptomics relate to methods from the 1990s, such as single-cell quantitative polymerase chain reaction (qPCR) [13], based on the consideration of a small number of genes. However, the first example of single-cell transcriptomics considering large sets of genes dates back to 2009 [14]. During the last decade, technological advances in microfluidics and automation allowed extending the sequencing process to thousands of cells per experiment. While scRNA-seq can scale to hundreds of thousands of cells, most single-cell epigenomic sequencing technologies tend to have lower throughput [4]. In general, available single-cell epigenomics considers either the methylation pattern at the DNA or the chromatin state [15]. Single-cell bisulfite sequencing (scBS-seq) addresses methylation profiles in single cells [16]. Methylation of CpG sites often indicates repression of gene transcription, but methylation of the gene body links to transcriptional activation. Single-cell methylation profiling requires high numbers of cells in input, yet low throughput (tens to thousands of cells) [17]. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) measures the actual presence of transcriptional regulators at the genome, capturing not only genomic accessibility but the presence of the actors of the transcriptional process as well [18]. Yet, data generated by

single-cell ChIP-seq of individual cells are very sparse: downstream analyses require an experiment to interrogate thousands of cells to ensure robust results [19]. To reduce the number of input cells required, alternative methods to ChIP-seq, not depending on immunoprecipitation, but on a reaction occurring within the nucleus, such as CUT&Tag [20], have been developed [21]. scATAC-seq [22] measures chromatin accessibility profiles. As previously discussed, accessibility indicates an open state of a genomic region, which often implies that the region - either coding or non-coding - is available for the transcriptional machinery or its regulators to bind and exert their function. scATAC-seq is rapidly becoming the primary way to assess the accessibility of the whole genome at the single-cell resolution, supporting the interrogation of numbers of cells ranging from hundreds (plate-based technologies) [23] to tens of thousands (droplet-based technologies) [24]. Analysis techniques for scATAC-seq datasets employ different ways to define meaningful features for analyzing accessibility profiles, as shown in [25]. One of the most popular is the "peak calling," which defines peaks, i.e., intervals on the genome that have a local enrichment of transposase cut-sites, meaning they are highly accessible [26]. This organizes accessibility data referring them to specific genomic regions. Yet, peaks come from an experiment-dependent set of chromosomal regions, which impedes direct comparison between epigenomic and transcriptomic datasets.

As described before, a GAM effectively defines robust accessibility features. The GAM considers the overall accessibility of the genomic regions linked to specific genes as defined in single-cell transcriptomics. Using scATAC-seq data, the GASs composing the elements of a GAM can be computed after the accessibility of the peaks related to the corresponding genomic region to such genes. However, the way to link peaks to the correct genetic region on the genome is not unique. The literature reports three main GAM-based strategies:

1. GeneScoring sums the peaks in a broad region before and after a gene's TSS, weighted by their distance from it [27]. This is the easiest way to define the activity of a gene. Nevertheless, it is simplistic since it overlooks the precise location of the genomic region relative to the gene and the distinction between coding and non-coding regulatory elements.
2. Cicero defines the activity of a gene as the accessibility of the peaks overlapping the TSS and the accessibility of all the co-accessible peaks, i.e., the peaks whose accessibility has a high correlation among cells. This definition supports the claim that these peaks participate in transcriptional regulation of the genomic region [7]. This method, compared with GeneScoring, implements a strategy to factor in regulatory elements at the DNA as well. However, the way it identifies the genes is very simplistic since it relies on a single DNA base, i.e., the TSS. This limits the approach's effectiveness in clearly linking genomic regions to genes. Moreover, co-accessibility evaluation is a very long and computationally heavy process. The GAS estimation does not consider the meaningfulness of the peaks, meaning if the co-accessible peaks are regulatory regions.
3. Signac builds a GAM by counting all the raw fragment reads (coming from the experimental Tn5 Transposase insertions) overlapping the gene body [28]. The main limitation of this method is that it overlooks the functional significance of specific reads completely. It requires a fragment file related to the dataset containing all the fragments in each cell. It is a large file and rarely available, thus making the computation often impossible.

In general, all these methods oversimplify the relationship between a gene and the accessibility of its genomic region. The accessibility profile of a genomic region provides information about the regulation and resulting expression of the corresponding gene. However, this does not imply a direct nor linear relation between region accessibility and gene expression levels. If a gene is accessible, it is not necessarily expressed: accessibility only gives an insight into chromatin structure, determining whether transcription is possible.

The rise of multi-omic approaches - comparing single-omic information and measuring multiple omic pools in the same experiment - makes this point even more relevant. In

particular, the emergence of new multi-omic NGS techniques allows performing both scATAC-seq and scRNA-seq simultaneously, posing the challenge of effectively analyzing and comparing information from transcriptomics and epigenomics. In this case, one way to achieve multi-omic consistent integration is to employ a model-driven approach for GAS computation.

To work towards this goal, we previously introduced in [10] the concept of Genomic-Annotated Gene Activity Matrix (GAGAM): a new, model-driven way to construct a GAM based on the functional genomic annotation of the peaks. A GAGAM differs from purely data-driven state-of-the-art scATAC-seq annotation methodologies since it considers that the contributions of different sets of accessible regulatory elements have a different meaning and impact on the GAM computation. This provides a methodological innovation. Beyond the bare superposition of accessibility data with genomic annotations, a GAGAM captures the functional impact of selected sets of accessible regions. This model builds an additional information layer on the genomic accessible regions not provided by routine pipelines for scATAC-seq annotation. The choices made to build the genomic model at the base of the GAGAM reflect the objective of multimodal epigenomic and transcriptomic single-cell data analysis. A pivotal point to support multimodal analysis is to allow comparison and joint analysis of different omic pools provided their structural differences.

This work introduces GAGAM v1.2, an improved version of GAGAM v1.0, to enhance the integration of accessibility and transcriptomic data. In GAGAM v1.2, the underlying gene activity model relies on an enhanced peak labeling process that focuses on the protein-coding gene regions that are more important from a transcriptomic standpoint. GAGAM v1.2 considers the contribution of the peaks inside the gene body more strictly than GAGAM v.1.0. Instead of accounting for the whole gene body (meaning both exon and intron parts), GAGAM v1.2 focuses on exon regions since they are the coding regions that are important from a transcriptomic standpoint. Moreover, during annotation, the identification of promoter peaks has been improved in GAGAM v1.2 using two approaches: the first based on the promoter signature of the candidate cis-Regulatory Elements (cCREs) tracks, the second on peaks overlapping the genes' TSS. Eventually, GAGAM v1.2 introduces a normalization of the different contributions composing the final matrix scores not available in GAGAM v1.0 and improves the quality of the computed scores.

Compared to GAGAM v1.0, this work advances the GAGAM pipeline in methodology and performance. In particular, it implements specific strategies to capture cellular heterogeneity better, yielding improved performances. In addition, it implements an innovative version of the genomic model to allow direct qualitative comparison between GAGAM-based activity and gene expression for multimodal datasets.

3. Materials and Methods

Figure 1 introduces the workflow for computation and evaluation of GAGAM starting from a scATAC-seq dataset. Throughout the paper, the acronym GAGAM refers to the generic concept of the Genomic-Annotated Gene Activity Matrix. A specific version identifier allows us to refer to features implemented in one variant of GAGAM.

A scATAC-seq dataset contains a set P of peaks observed in a group of C cells. Each peak corresponds to a region of the target genome and is defined by its chromosome and a genomic coordinate pair $p = (ch, start, stop)$. The dataset is a binary matrix $\mathbf{D}_{|P|\times|C|}$ where rows are associated with peaks and columns with cells. An element of $\mathbf{D}$ equal to 1 denotes a peak (row) accessible in a cell (column).

The main contribution of GAGAM is to exploit information regarding overlaps of peaks, gene bodies, and genetic regulatory regions (i.e., promoters and enhancers) to build a GAM with higher information content.

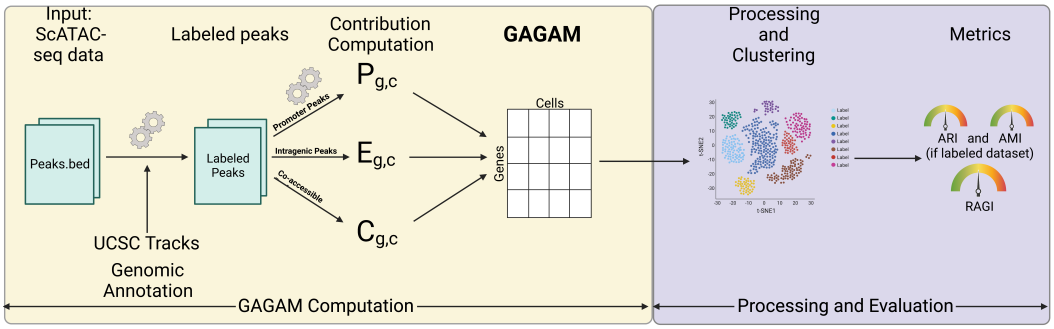


Figure 1. Workflow for computation and evaluation of GAGAM. The workflow starts with scATAC-seq data, and labels the peaks with the help of genomic annotations and UCSC tracks. Then it computes the three contributions forming GAGAM. GAGAM is evaluated and compared to other GAMs through clustering experiments with three well-established metrics: ARI [29], and AMI [30] (if the dataset is labeled) or RAGI [31] (if the dataset is not labeled).

3.1. Genomic Annotation

Understanding the biological meaning of the genomic region underlined by the peaks is fundamental for their proper model-driven usage. Therefore, the genomic annotation of peaks is the first step to constructing GAGAM. This process aims to enrich information regarding peaks with data from different genomic annotations useful for a model-driven construction of a GAM. Figure 2 shows the genomic model considered in this work. It represents a genomic unit, which includes three parts: (i) the gene body region, divided into Exons (the coding parts) and Introns (the non-coding parts), starting at the TSS, (ii) the gene Promoter, preceding the coding region, and (iii) a set of Enhancers that are distal to the gene.

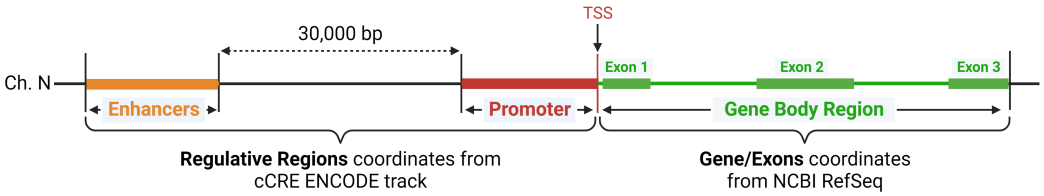


Figure 2. Genomic Model. The genomic model consists of the coordinates of all the genomic regions related to the gene. The gene body region (green) comes from the NCBI RefSeq Genes annotations. The regulative regions, i.e., Promoter (in red) and Enhancers (in orange) come from the cCREs ENCODE tracks.

The gene coding region is defined using NCBI RefSeq Genes [32] annotations, consisting of genes' genomic coordinates. The NCBI RefSeq annotations are accessible using the NCBI Eukaryotic Genome Annotation Pipeline [33]. It consists of an annotated and curated information list of protein-coding and non-protein-coding genes.

Since GAGAM aims to link the epigenomic information to the transcriptomic one, it only considers the protein-coding gene regions. In GAGAM v1.2, pseudogenes, miRNA, and lncRNA regions, also present in the annotation, are discarded. This differs from GAGAM v1.0, which includes the lncRNA genes. LncRNA genes are coding regions but not for mRNA products measured in transcriptomic experiments and are not helpful for comparison with transcriptomic experiments. Every gene g in the target genome G is a tuple defining the gene's chromosome and its genomic coordinates pair (i.e., $g = (ch, start, stop)$). From the same coordinates, one defines the TSS of a gene as a small region corresponding to the two bases of the first exon, i.e., $g_{TSS} = (ch, start, start + 1)$.

Instead of accounting for the whole gene body (meaning both exon and intron parts) as in GAGAM v.1.0 [10], GAGAM v1.2 focuses on exon regions since they are the coding regions that are important from a transcriptomic standpoint. So for each gene g, there is a list of n exon regions defined by their coordinates, i.e., $g_{e_1} = (ch, start, stop_{e_1}) ... g_{e_n} = (ch, start_{e_n}, stop)$.

The regulative genomic regions are elements on the DNA footprints for the trans-acting proteins involved in transcription, either for the positioning of the basic transcriptional machinery or for the regulation. The annotation tracks are associations between a genomic region and a label indicating the function of the region. Given a target genome G, it is possible to define a set R of regulative regions with each region defined by the corresponding chromosome, the genomic coordinates pair, and a label (e.g., promoter or enhancer) indicating the function of the region ($r = (ch, start, stop, l)$).

Information regarding regulative gene regions is available from the Encyclopedia of DNA Elements (ENCODE) project, which provides an extensive collection of cell- and tissue-based genomic annotations tracks, including, for example, transcription, chromatin organization, epigenetic landscape dynamics, and protein binding sites from the mouse and human genomes [34]. ENCODE data are available through the ENCODE data portal [35].

This work only considers genomic annotations relative to promoter and enhancer functions. These regulatory elements are derived from the ENCODE cCREs. cCREs provides an extensive collection of annotated regions for the human and mouse genomes. Predicted classification of cCREs relies on biochemical signatures deriving from considering DNase hypersensitivity, histone methylation, acetylation, and CTCF binding data [34]. It is important to highlight that, given the current nature of the ENCODE annotation, it is impossible to get information for more complex regulatory elements. It is the case, for example, of the alternative promoters, which in the development and differentiation stages of some cell types allow for specific expression of alternate transcripts. Currently, GAGAM v1.2 does not handle alternative promoter usage since it considers a single promoter for each gene, thus excluding the contribution of alternative promoters either upstream or downstream of the TSS. To acknowledge different isoforms for each gene, it will be necessary to consider them as additional features, extending the structure of the GAGAM accordingly, which is not part of the current version.

In addition, since this work aims to label peaks from both human and mouse datasets, cCREs tracks (in BigBed format [36]) were collected from ENCODE for both the human [37] and mouse [38] genomes. To be precise, the cCREs tracks are only aligned to the most recent genome assemblies (hg38 for humans and mm10 for mice). Therefore, when dealing with data aligned to older assemblies, there is a need to convert peak coordinates to the track assembly. To achieve this, this work leverages the UCSC `LiftOver` tool [39] to map genomic coordinates of the peaks before performing peak labeling. This step can insert some uncertainty concerning the original data like some peaks cannot be aligned. However, this error is negligible since it regards just some tens of peaks, a small portion compared to the hundreds of thousands in the data.

The goal of the genomic annotation process is to associate each peak $p \in P$ obtained from a scATAC-seq dataset $\mathbf{D}$ to a set of genomic annotation labels by analyzing how the peak overlaps with the different genomic regions.

GAGAM labels each peak $p \in P$ with four possible tags: (1) `prom` for peaks overlapping a promoter region or a TSS, (2) `enhD` for peaks overlapping a distal enhancer region and not a promoter region, (3) `exon` for the peaks contained into a gene's exon region, and (4) `empty` in all other cases. The rule to assign the label is summarized in the following equation:

$$\mathcal{PL} : p \in P \mapsto \begin{cases} \texttt{prom} & \text{if } (\exists r \in R | r \subseteq p \wedge r_l = \text{prom}) \vee \\ & (\exists g_{tss} \in G | p \subseteq g_{tss}) \\ \texttt{enhD} & \text{if } (\exists r \in R | r \subseteq p \wedge r_l = \text{enh}) \wedge \\ & (\nexists r \in R | r \subseteq p \wedge r_l = \text{prom}) \\ \texttt{exon} & \text{if } \exists g_e \in G | p \subseteq g_e \\ \texttt{empty} & \text{otherwise} \end{cases} \tag{1}$$

The operator $a \subseteq b$ is used here to denote that the two regions a and b belong to the same chromosome with b overlapping a (i.e., $a_{start} \geq b_{start} \wedge a_{end} \leq b_{end}$). The computation

of the intersection between peaks and annotation regions leverages the `bigBedToBed` tool from ENCODE [40].

Given the list of annotated peaks, GAGAM builds a gene activity matrix as a weighted sum of three separated matrices: (i) the promoter peaks matrix (**P**) indicating accessibility of genes associated with promoter peaks, (ii) the co-accessibility matrix (**C**) indicating the accessibility of genes associated with distal enhancer peaks, and (iii) the exon peaks matrix (**E**) indicating the accessibility of genes containing exon peaks.

The weights w_p, w_c, and w_e, as done in [10], are binary numbers that provide a formal and modular way to select which contributions to include in the final matrix computation. Therefore, these weights are not estimated from any training step but instead set at the beginning of the analysis. This work defines two combinations of weight values for GAGAM construction, one with all three contributions (i.e., $w_p = 1$, $w_c = 1$, and $w_e = 1$), and one that excludes the contribution of exon peaks (i.e., $w_p = 1$, $w_c = 1$, and $w_e = 0$)). In general, this allows us to study the contribution of each sub-matrix composing GAGAM to the gene activity and, consequentially, if some influence the gene expression in a specific way. This is not part of this specific paper but will be part of future analyses.

This yields a final curated and model-driven evaluation of the activity of the genes:

$$\mathbf{GAGAM} = w_p \cdot \mathbf{P} + w_c \cdot \mathbf{C} + w_e \cdot \mathbf{E} \tag{2}$$

It is relevant to notice that, in GAGAM v1.0 [10], some peaks simultaneously overlapping enhancers signatures and intronic regions were contributing only to the Intragenic Matrix (the old version of the Exon Matrix). Thus their regulative information was lost. GAGAM v1.2 labels these peaks more precisely as `enhD` peaks, contributing to the co-accessibility Matrix.

The following paragraphs illustrate the modeling assumptions and construction process for each contribution to GAGAM.

3.2. Promoter Peaks Matrix

The promoter peaks matrix exploits annotation-based information about promoter peaks to identify relevant genes in the GAM. The identification of promoter peaks in GAGAM v1.2 has changed compared to GAGAM v1.0 [10].

The golden rule applied in GAGAM is that *a gene in a cell is active if, and only if, its promoter peak is accessible.* This rule reduces the set of interesting genes to consider when constructing a GAM.

Let us denote with $P^p \subseteq P$ the subset of peaks in the dataset **D** annotated as promoters (i.e., $\mathcal{PL}(p) = \texttt{prom}\ \forall p \in P^p$) and with $\mathbf{D}^p_{|P^p|\times|C|}$ the submatrix of **D** including only rows associated to promoter peaks.

GAGAM v1.2 constructs a binary matrix $\mathbf{GP}_{|G^p|\times|P^p|}$ associating the set of genes with active promoter peaks (G^p) to their related peaks. To associate a promoter peak to a gene, GAGAM v1.2 considers overlapping an enlarged gene body region, including 500 bp before the TSS with the peak region. This value was selected as an estimation of the mean peak length, but other values can be used depending on the available data.

The identification of promoter peaks has changed compared to GAGAM v1.0 [10]. GAGAM v1.2 employs two rules to define the promoter peaks. It relies on the promoter signature of the cCREs tracks, but it also considers the peaks overlapping the genes' TSS.

The motivation for this choice is twofold.

First, even if the cCREs track information is curated and well-organized, it is still a prediction, and therefore, it has intrinsic limitations linked to partial prediction accuracy. Indeed, some genes in the NCBI RefSeq Genes annotation appear to have no promoter signature near the TSS (Figure 3). So, to not lose possible information about the activity of this category of genes, it is convenient to label the peaks overlapping their TSS as `prom`.

Second, the genes are directional (meaning they organize along an intended direction for the transcriptional process to occur, where, for example, the promoter regions precede the coding region). Still, the cCREstracks signatures are not. This can bring problems when

assigning the promoter peaks to the right genes. For example, in genomic regions where two divergent genes overlap, a `prom` peak could be wrongly assigned to the most overlapping gene (Figure 4). Therefore, GAGAM v1.2 employs the innate TSS-gene association to correctly set the `prom` peaks to the genes.

Based on this, the promoter peaks matrix is a binary matrix computed as:

$$\mathbf{P}_{|G^p|\times|C|} = \mathbf{GP}_{|G^p|\times|P^p|} \times \mathbf{D}^p_{|P^p|\times|C|} \tag{3}$$

This matrix is a GAM based on accessibility data for the subset of genes in the dataset exhibiting promoter peaks. In this way, GAGAM leverages available knowledge on transcriptional regulatory regions to select, among the entire group of genes considered in the scATAC-seq dataset, only the active ones, based on the modeling assumption that promoter accessibility is a necessary condition for gene regulation and transcription.

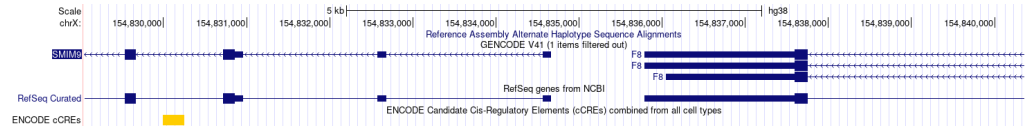


Figure 3. Genomic region of SMIM9 gene. The figure shows there are no prom signature regions, in the proximity of the TSS of SMIM9, in the cCREs track, meaning it would be left out from the gene activity computation if looking only at Promoter signatures.

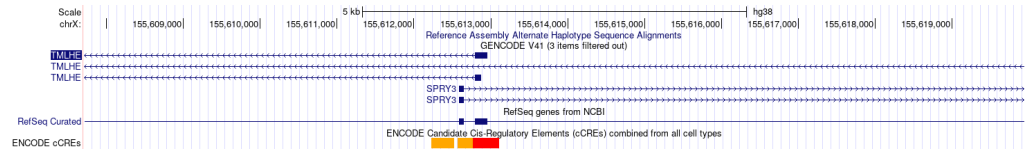


Figure 4. Genomic region of TMLHE and SPRY3 genes. In this case, the promoter signature (red) is an ambiguous position between the two divergent genes, and due to some overlapping, the prom peaks could be assigned to the wrong genes.

3.3. Promoter-Enhancer Co-Accessibility Matrix

The co-accessibility matrix accounts for the connections between promoters and enhancers. It is the second most important contribution and employs the `enhD` peaks.

GAGAM v1.2 leverages Cicero [7] to calculate the co-accessibility of the peaks. The co-accessibility represents how couples of peaks are simultaneously accessible in the cells, expressing it in a range between 0 and 1. This calculation *"connects regulatory elements to their putative target genes"* [7], meaning it may find connections between the promoters and the distal regulatory regions where different elements like Transcriptional Factors (TF) bind and enable transcription. The first step to compute this matrix is to calculate the co-accessibility from the scATAC-seq data with the Cicero function `run_cicero` (for the explanation of the calculation, refer to [7]). The result is a list of peaks couples with their co-accessibility value (ca) and distance (d) in the form $conn = (p^1, p^2, ca, d)$. However, the co-accessibility alone does not carry out the information on the gene regulatory regions. Connections found by the algorithm do not describe gene regulation links, so many are unnecessary for this model.

GAGAM v1.2 selects only couples of promoter-enhancer peaks, i.e., couples with $p^1 \in P^p$ and $p^2 \in P^e$ (or vice versa), with $P^e \subseteq P$ representing the subset of peaks in the dataset $\mathbf{D}$ annotated as enhancers (i.e., $\mathcal{PL}(p) = \texttt{enhD}\ \forall p \in P^e$). Compared to the GAGAM v1.0 [10], GAGAM v1.2 is more precise in selecting couples of peaks, making sure to select all the combinations without repetition.

GAGAM v1.2 considers only strong and relevant peaks connections, which are more likely to represent actual regulatory links. This is done thanks to two criteria: the distance between connected peaks and the strength of the connections, represented by the two parameters d_{th} and c_{am}. These parameters derive from the Cicero pipeline co-accessibility calculation function [7], a well-established co-accessibility analysis pipeline.

Co-accessibility scores tend to have low values. GAGAM v1.2 uses the ca_m parameter to determine the strongest connections (couples with $ca \geq ca_m$) and filter out connections that are not relevant to the gene to reduce noise. c_{am} is the mean value of all the non-zero co-accessibility scores. This threshold is automatically computed based on co-accessibility data from each specific dataset, adapting the selection of the considered connections. This provides a pretty strict selection (10% at most) of connections and prevents inconsistencies linked to inter-dataset differences since the co-accessibility results can vary a lot between experiments.

The distance d_{th} is a threshold to define the span of the genomic region where connections are most probably actual direct regulation links. It allows selecting only connections within a relevant distance range in the genome (i.e., $d \leq d_{th}$). Unfortunately, the literature lacks a clear indication for choosing the value of this parameter. Therefore, GAGAM v1.2 employs the exact value the Cicero guidelines suggest (i.e., 30,000 bp) [7].

At this point, the analysis identifies co-accessible couples of genomic regions, selects couples composed of a promoter and an enhancer, and among them, keeps only promoter-enhancer couples lying within the $d \leq d_{th}$ range of 30,000 bp and whose accessibility score was above the mean value of all the co-accessibility scores above zero (ca_m). This is the starting point to build the co-accessibility contribution for GAGAM computation.

To calculate the co-accessibility matrix $\mathbf{C}$, GAGAM uses three matrices. First, the binary matrix $\mathbf{GP}_{|G^p|\times|P^p|}$ previously defined in Section 3.2 and associating genes with promoter peaks. Second, the matrix $\mathbf{PE}_{|P^p|\times|P^e|}$ associating promoter peaks and enhancer peaks. The elements of this matrix are the co-accessibility values ca of couples of peaks available in the list produced by Cicero and 0 otherwise. Third, the matrix $\mathbf{D}^e_{|P^e|\times|C|}$ is a submatrix of $\mathbf{D}$ including only rows associated with enhancer peaks.

Based on this, the co-accessibility matrix is as follows:

$$\mathbf{C}_{|G^p|\times|C|} = \mathbf{GP}_{|G^p|\times|P^p|} \times \mathbf{PE}_{|P^p|\times|P^e|} \times \mathbf{D}^e_{|P^e|\times|C|} \tag{4}$$

3.4. Exon Peaks Matrix

The third contribution to the final GAGAM relies on exon accessibility within the gene coding region.

GAGAM v1.2 considers the contribution of the peaks inside the gene body more strictly than GAGAM v.1.0 [10]. As aforementioned, intronic regions of the gene bodies can contain regulatory regions. These do not necessarily relate to the gene containing them. Thus, instead of labeling all the peaks inside the gene body as `intra`, GAGAM v1.2 distinguishes between exons and introns. Therefore, as discussed in Section 3.1, only the peaks overlapping exons take the `exon` label and constitute the third contribution to the GAS.

As detailed in Section 5, the current GAGAM computation uses a single promoter for each coding region. For this reason, the calculation of accessible exons' contribution to gene activity is referred to as the selected promoter.

Let us denote with $P^i \subseteq P$ the subset of peaks in the dataset $\mathbf{D}$ annotated as intragenic (i.e., $\mathcal{PL}(p) = \texttt{intra}\ \forall p \in P^i$) and with $\mathbf{D}^i_{|P^i|\times|C|}$ the submatirx of $\mathbf{D}$ including only rows associated to exon peaks. GAGAM v1.2 constructs a matrix $\mathbf{GI}_{|G^p|\times|P^i|}$ associating genes with active promoter peaks (G^p) to their related exon peaks. This matrix only considers genes with active promoter peaks to follow the GAGAM golden rule (Section 3.2). The reason is that some of the identified exon peaks could be part of genes that do not have a promoter peak. Moreover, it could happen that given a gene region inside a cell, exon peaks could be accessible even if the promoter peak is not.

The number and the length of exons vary a lot between the genes of a genome. Thus, statistically, there will be more peaks overlapping exons of a long gene, meaning it might have a higher score after its length. To prevent this bias, GAGAM employs a strategy from the GeneScoring [27] method to compute the elements of $\mathbf{GI}$. It weighs the contribution of the exon peaks with an exponentially decaying function of their distance from the TSS (i.e., $\mathbf{GI}_{g,p} = a \cdot e^{-\frac{d}{5000}}$ where $a = 1$ if $p \subseteq g$, 0 otherwise and d is the distance of the peak from

TSS). In this way, very long genes are not over-represented because the most crucial part of the gene's activity is near the promoter. Therefore, the peaks near it are weighted more. GAGAM v1.2 improves the peak distance calculation by considering the distance of the center of the peak to the TSS, not just the nearest base.

Based on this, the exon peaks matrix is a matrix computed as:

$$\mathbf{I}_{|G^p|\times|C|} = \mathbf{GI}_{|G^p|\times|P^i|} \times \mathbf{D}^i_{|P^i|\times|C|} \quad (5)$$

3.5. Final Matrix Construction

After the computation of the three contributions above, the final step is constructing the GAGAM v1.2. As done in [10], this work defines two combinations of contributions for GAGAM construction, one with all three contributions (i.e., $w_p = 1$, $w_c = 1$, and $w_e = 1$), and one without the exon peaks part (i.e., $w_p = 1$, $w_c = 1$, and $w_e = 0$)). This allows us to assess the exon peaks matrix contribution and, thus, the role of exon accessibility to the analysis performance.

Moreover, GAGAM v1.2 adds another element to the final construction. GAGAM v1.0 [10] lacked normalizing the final matrix scores. Since there is not a reliable normalization method suited for these contributions, GAGAM v1.2 employs the normalization approach proposed by Cicero (for the explanation of the method, refer to [7] and its documentation). This ensures at least a normalization between the different impacts of the three contributions. To establish the actual effects of the normalization, Section 4 shows the results on the normalized GAGAM v1.2.

4. Results and Discussion

This section explains the strategy proposed in this work to validate GAGAM v1.2, the metrics chosen to measure performance, the analysis pipelines, and the datasets employed. Finally, it shows and discusses the results obtained.

4.1. Validation Strategy

GAGAM models gene activity based on scATAC-seq data, providing an enhanced tool to study cellular heterogeneity based on epigenomic data and to build a robust comparison with transcriptomics. Based on this consideration, following the same approach proposed in [10], the proposed validation strategy aims to assess the capabilities of GAGAM v1.2, considering performance under two lenses: (1) the capability of correctly identifying cellular heterogeneity and (2) the enhancement of biological significance for multi-omics studies.

The first validation step refers to (1), which evaluates GAGAM v1.2 capabilities to identify cellular heterogeneity. This is measured by quantifying the consistency of the results obtained with GAGAM-based clustering analysis versus those obtained with established approaches on the same datasets. Classification performance measures and comparisons rely on Adjust Rand Index (ARI) [29] and Adjust Mutual Information (AMI) [30], whose usage is thoroughly discussed in Section 4.2.

The second validation step refers to (2). It is to prove that GAGAM v1.2, and, in particular, its model-based selection of genes to consider for the analysis, have biological meaning, intended in two aspects: the intrinsic informativeness of the GAGAM and its capability to support a comparison with transcriptomics. The first aspect is measured by quantifying the capability of the analysis to detect significant differences in gene activity among identified clusters. This measurement relies on the Residual Average Gini Index (RAGI) metric (see Section 4.2). The second aspect is investigated with a preliminary comparative analysis between Differentially Active (DA) and Differentially Expressed (DE) genes (see Section 4).

4.2. Metrics Definition

The validation strategy proposed in Section 4.1 relies on the unsupervised clustering of cells based on the genes selected for GAGAM construction. Clustering analysis outputs

clusters of cells modeling the heterogeneous types of cells in a cell population. The first step includes evaluating whether GAGAM-based clustering analysis yields clusters that correctly represent cell heterogeneity. Comparing results from clustering classification based on GAGAM v1.2 with ground truth is necessary. Intending to do so, there are two scenarios: (i) the dataset includes cell labels from pre-existing classifications. Thus, each cell has a label identifying its cell type, or (ii) the dataset does not include cell labels, so there is no ground truth to compare. In the first scenario, it is possible to directly compare GAGAM-based clustering classification with the cell-type labels leveraging ARI and AMI metrics to assess similarity. Since there is no reference classification in the second scenario, ARI and AMI cannot be used as in the previous case. An alternative method is to employ clustering-based labels obtained from the raw scATAC-data analysis as ground truth and then use ARI and AMI to measure the results' similarity. Otherwise, it is possible to use the RAGI metric to assess and compare the intrinsic capability of different classification methods to distinguish cells based on activity (or accessibility) profiles.

4.2.1. ARI and AMI

This work relies on the ARI [29] and AMI [30] metrics from information theory to measure similarity in classification results. These two metrics are often employed for this type of evaluation. In particular, ref [31] uses them for benchmarking, and they are the metrics employed to evaluate GAGAM v1.0 [10], which supports the direct comparison of the new results with those previously produced. ARI and AMI similarity measures range between 0 and 1, where 1 is a perfect match, and 0 is complete dissimilarity. The main difference between the two is that ARI performs better when the ground truth clustering has large clusters with equal cell numerosity. In contrast, AMI performs better when the ground truth clustering yields unbalanced clusters in size, including small clusters [41]. Considering this, even though AMI is preferable for most cases considered in this work, the results rely on both metrics.

4.2.2. RAGI

The RAGI [31] is based on the hypothesis that good clustering results should exhibit the high activity of marker genes in specific clusters and activity of housekeeping genes in all the cells from every cluster. A good GAM should convey meaningful biological information that should translate into a difference in the activity patterns across clusters between the two sets of genes. Computing the difference between the Gini index of markers and housekeeping genes respectively, the RAGI can measure the intrinsic capability of GAGAM-based clustering analysis to distinguish groups of cells with the same functional profile (i.e., the activity of specific marker genes), relying on internal control (i.e., activity the housekeeping genes). On the other hand, RAGI is highly dependent on the genes chosen to calculate it. Indeed, the RAGI algorithm requires two lists of genes, housekeeping genes, and marker genes, which are not trivial to select because there is no universal definition of housekeeping genes or marker genes. Therefore, the selection of housekeeping genes is partly arbitrary and could bring errors to the calculation. This is even worse in the case of marker genes, which are highly debatable [42] and should be decided after the cell types present in the dataset, which is only partially defined. Therefore, it is essential to carefully choose the right set of marker genes strictly related to the dataset sample. In this work, the list of housekeeping genes is based on the literature, particularly on [43] for human cells and on [44] for murine cells. On the other hand, the marker genes for the human Peripheral Blood Mononuclear Cell (PBMC) datasets are based on a curated list of markers for PBMC classification from [45], while the marker genes for the mouse brain dataset were defined based on [46,47].

4.3. Datasets

This work analyzes GAGAM v1.2 performance on six datasets (see Table 1 and 2). Three of these datasets supported the validation of GAGAM v1.0 [10]; the others are new.

Five out of six datasets are from the 10XGenomic platform [24] since it provides high-quality datasets generated by various technologies. The datasets shared with [10] consist of a collection of respectively 5335 (*10X v1.0.1 PBMC* [48]) and 4623 (*10X V2.0.0 PBMC* [49]) cells from human Peripheral Blood Mononuclear Cells (PBMC) samples. Additionally, this study also employed the dataset of bone marrow (*Buenrostro2018*) from [50], consisting of 2034 cells with provided cell-type classification. The three new datasets consist of one scATAC-seq dataset of mouse brain with 7729 cells (*10X V2.1.0 Brain* [51]), and two from the single cell Multiome technology, consisting respectively of 10,691 cells (*10X Multiome Controller PBMC* [52]) and 10,970 cells (*10X Multiome Chromium X PBMC* [52]) from human PBMC samples. These two datasets provide scATAC-seq and scRNA-seq data, supporting the preliminary comparison between the activity and expression of genes. None of the three new datasets have cell labels. Therefore, their ARI and AMI evaluations leverage raw scATAC-seq clustering-based labels.

Table 1. Old datasets employed.

Dataset	Species	Tissue	Cells Labels	Reference
10X v1.0.1 PBMC	Human	PBMC	No	[48]
10X V2.0.0 PBMC	Human	PBMC	No	[49]
Buenrostro2018	Human	Bone Marrow	Yes	[50]

Table 2. New datasets employed.

Dataset	Species	Tissue	Cells Labels	Reference
10X Multiome Controller PBMC	Human	PBMC	No	[52]
10X Multiome Chromium X PBMC	Human	PBMC	No	[52]
10X V2.1.0 Brain	Mouse	Brain Cortex	No	[51]

4.4. Pipelines

For clustering analysis and preliminary comparison with transcriptomics, the pipeline of choice is Monocle3 [53], given its simplicity, the straightforward workflow, and the fact that Cicero GAM (see Section 2) is dependent on it (Figure 1). The standard Monocle workflow starts with a GAM, performs Principal Component Analysis (PCA), visualizes the cells in 2D using UMAP [54], and, most importantly, conducts cell clustering. However, Latent Semantic Indexing (LSI) is an alternative approach to the usual PCA, especially for scATAC-seq data. Thus, since GAGAM v1.2 is a direct manipulation of scATAC-seq data, it makes sense to employ LSI.

Concerning performance evaluation, ARI and AMI computation relies on the R package ARICODE [55]. In contrast, RAGI computation is based on a custom Python script (see **Data Availability Statement**).

4.5. Results

This Section provides the results of the validation strategy proposed in Section 4.1. Results refer to the following comparisons:

- GAGAM v1.2 *versus* GAGAM v1.0 [10], over the old datasets;
- GAGAM v1.2 *versus* Cicero [7] and GeneScoring [27], over the new datasets.

Besides considering both GAGAM v1.0 and GAGAM v1.2, the experimental setup considers two versions of GAGAM v1.2:

- GAGAM1 v1.2 created using the complete GAGAM workflow (i.e., $w_p = 1$, $w_i = 1$, and $w_c = 1$)
- GAGAM2 v1.2 constructed considering only the promoter peaks and the co-accessibility (i.e., $w_p = 1$, $w_i = 0$, and $w_c = 1$).

4.5.1. Peak Labeling

As a first insight into the GAGAM v1.2 construction process, Table 3 shows the final number of labeled peaks per dataset, focusing on their labeling and the final number of genes employed for GAGAM construction.

Table 3. Numbers of peak labels.

Dataset	Total Peaks	*Prom* Peaks	*enhD* Peaks	*Exon* Peaks	Final Genes
10X v1.0.1 PBMC	97,124	11,651	1940	7763	11,488
10X V2.0.0 PBMC	135,373	15,720	79,481	3650	13,298
Buenrostro2018	491,232	22,847	190,298	127,667	14,905
10X Multiome Controller PBMC	115,179	15,528	67,790	3207	13,230
10X Multiome Chromium X PBMC	111,743	15,438	66,211	3064	13,153
10X V2.1.0 Brain	176,651	14,662	50,784	9768	13,023

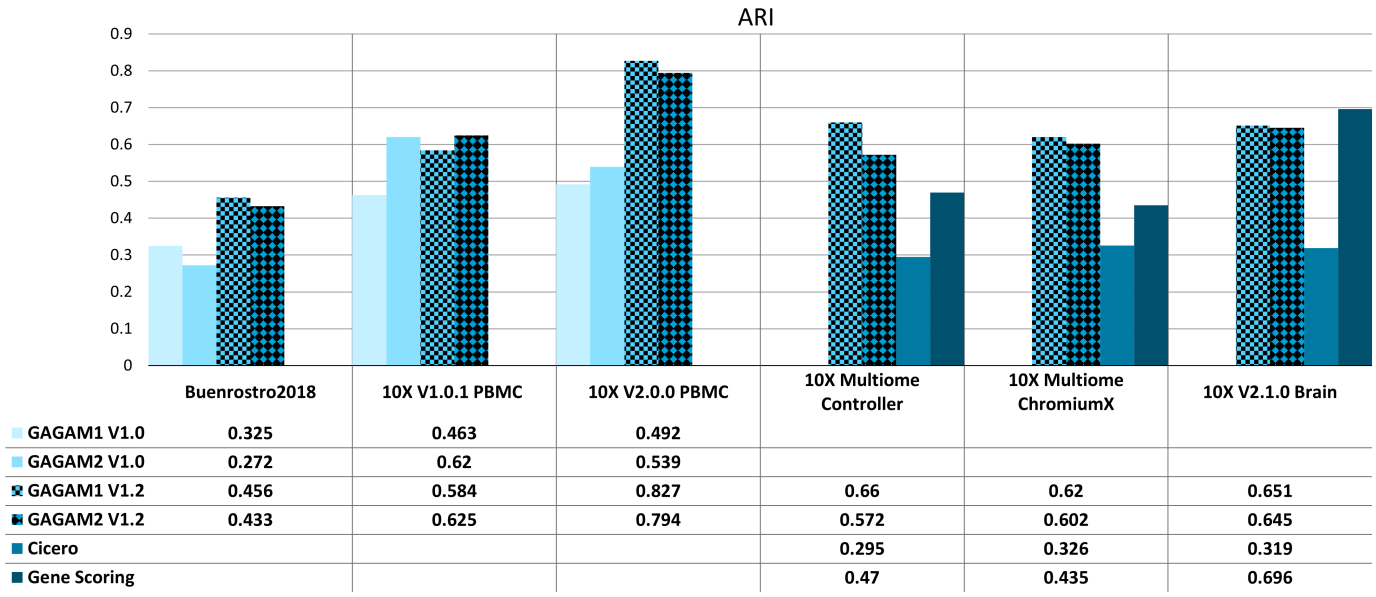
4.5.2. Clustering Similarity

The first step of the validation strategy described in Section 4.1 aims to evaluate GAGAM v1.2 capability of correctly identifying cellular heterogeneity. At this aim, this section compares the similarity of results obtained with GAGAM-based clustering analysis versus results obtained with other approaches. Measures and comparisons rely on the ARI and AMI metrics (see Section 4.2).

Figures 5 and 6 report AMI and ARI scores for all methods under analysis over the considered datasets. Comparisons rely on ground truth labels. For example, clustering labels are compared with pre-existing cell type labels for the *Buenrostro2018* dataset [50]. In contrast, clustering labels are compared with the raw scATAC-seq clustering results for the other datasets. Figures display the ARI and AMI scores for the two GAGAM v1.2 versions (GAGAM1 v1.2 and GAGAM2 v1.2, respectively, patterned bars) against GAGAM v1.0 from [10] (lighter blue tones) for the three Old datasets (left), and Cicero and GeneScoring (darker blue tones) for the three New datasets (right). Results show that both versions of GAGAM v1.2 perform better than GAGAM v1.0, Cicero, and GeneScoring. However, GAGAM v1.2 offers higher performances than GAGAM v1.0, with some significant improvements. In particular for *10X V2.0.0 PBMC* where the AMI of GAGAM1 v1.2 has a gain of about 23%. Only on *10X V2.1.0 Brain*, GeneScoring has a higher metrics value than GAGAM. However, as discussed in [10], even if GeneScoring has, sometimes, higher performances of AMI and ARI, it has abysmal results of the RAGI, lowering the impact of these higher metrics. Generally, both versions of GAGAM v1.2 score higher in ARI and AMI metrics than clustering results from the scATAC-seq data. This suggests that GAGAM v1.2 can identify cellular heterogeneity at least at the same level as the direct scATAC analysis.

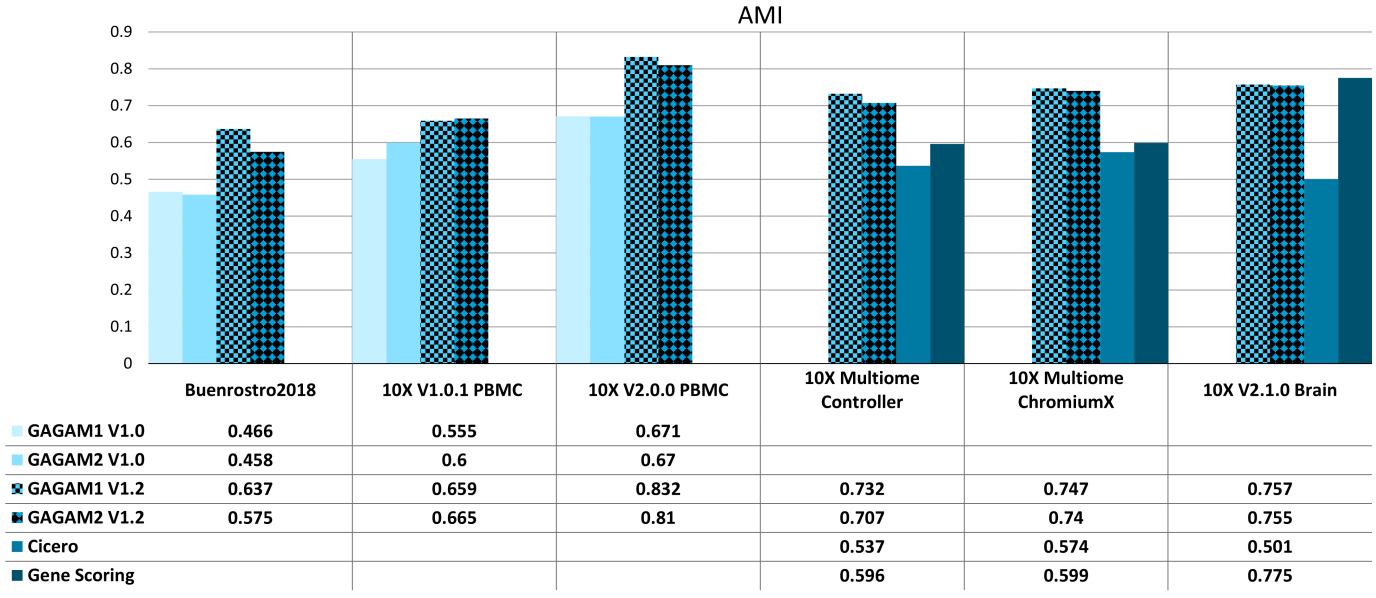
4.5.3. Biological Significance

The second step of the validation strategy described in Section 4.1 aims to assess GAGAM v1.2 biological significance, proving the model-based selection of genes to consider and their elaboration enhances informativeness and biological meaning. This is measured by quantifying the capability of the analysis to detect significant differences in gene activity among identified clusters. This measurement relies on the RAGI metric (see Section 4.2).



	Buenrostro2018	10X V1.0.1 PBMC	10X V2.0.0 PBMC	10X Multiome Controller	10X Multiome ChromiumX	10X V2.1.0 Brain
GAGAM1 V1.0	0.325	0.463	0.492			
GAGAM2 V1.0	0.272	0.62	0.539			
GAGAM1 V1.2	0.456	0.584	0.827	0.66	0.62	0.651
GAGAM2 V1.2	0.433	0.625	0.794	0.572	0.602	0.645
Cicero				0.295	0.326	0.319
Gene Scoring				0.47	0.435	0.696

Figure 5. ARI results for both Old (**left**) and New (**right**) datasets. The Old datasets are compared with GAGAM v1.0 from [10], while the New datasets are with Cicero and GeneScoring.

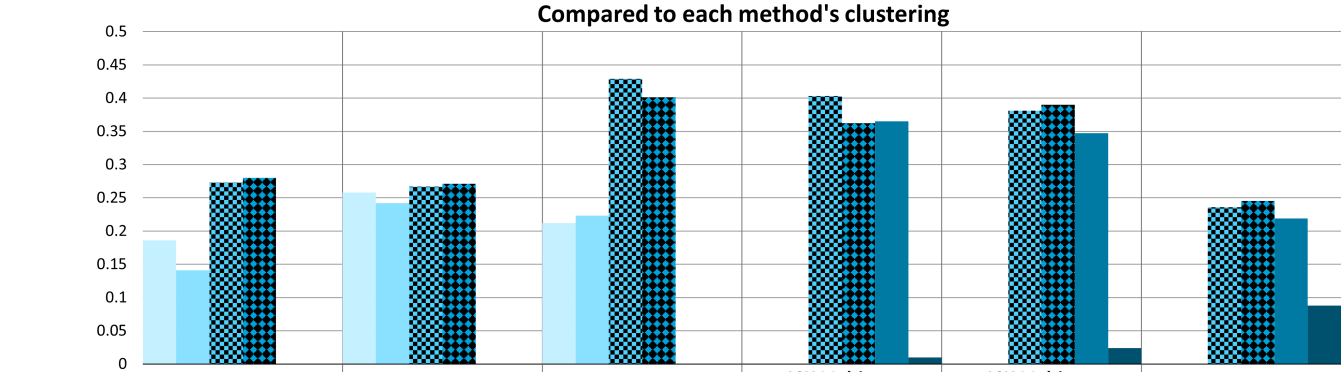


	Buenrostro2018	10X V1.0.1 PBMC	10X V2.0.0 PBMC	10X Multiome Controller	10X Multiome ChromiumX	10X V2.1.0 Brain
GAGAM1 V1.0	0.466	0.555	0.671			
GAGAM2 V1.0	0.458	0.6	0.67			
GAGAM1 V1.2	0.637	0.659	0.832	0.732	0.747	0.757
GAGAM2 V1.2	0.575	0.665	0.81	0.707	0.74	0.755
Cicero				0.537	0.574	0.501
Gene Scoring				0.596	0.599	0.775

Figure 6. AMI results for both Old (**left**) and New (**right**) datasets. The Old datasets are compared with GAGAM v1.0 from [10], while the New datasets are with Cicero and GeneScoring.

Figures 7 and 8 show the results of RAGI computation, assessing the information content of the GAMs (GAGAM1 v1.2, GAGAM2 v1.2, GAGAM1 v1.0, GAGAM2 v1.0, Cicero and GeneScoring GAMs, respectively) compared to other clustering classifications. Following the same logic from ARI and AMI results presentation (see Section 4.5.2), the two versions of GAGAM v1.2 are shown with patterned bars, the two versions of GAGAM v1.0 from [10] in lighter blue tones for the three Old datasets (left), and Cicero and GeneScoring are shown in darker blue tones for the three New datasets (right). To perform RAGI computation (as described in Section 4.2.2), this work employed three different sets of curated marker genes, one for each type of tissue in the datasets (human PBMC, human bone marrow, mouse brain) and two different sets of housekeeping genes, one for each species in the datasets (human and mouse, respectively).

Results for biological significance organize into two groups. The first group (Figure 7) leverages RAGI scores to compare informativeness between the clustering results obtained with the different GAMs. The second group (Figure 8) leverages RAGI scores to compare the clustering results of each GAM with classification labels from pre-existing cell type labels (in the case of the *Buenrostro2018* dataset [50]) or from clustering-based labels obtained from clustering analysis on raw scATAC-seq data. This way, all methods are evaluated against the same partition to understand which GAM is the most biologically consistent.



	Buenrostro2018	10X V1.0.1 PBMC	10X V2.0.0 PBMC	10X Multiome Controller	10X Multiome ChromiumX	10X V2.1.0 Brain
GAGAM1 V1.0	0.186	0.258	0.212			
GAGAM2 V1.0	0.141	0.242	0.223			
GAGAM1 V1.2	0.273	0.267	0.429	0.403	0.381	0.236
GAGAM2 V1.2	0.28	0.271	0.401	0.362	0.39	0.245
Cicero				0.365	0.347	0.219
Gene Scoring				0.01	0.024	0.088

Figure 7. RAGI results of both Old (**left**) and New (**right**) datasets when compared to each method's clustering. The Old datasets are compared with the previous GAGAM v1.0 from [10], while the New datasets are with Cicero and GeneScoring.

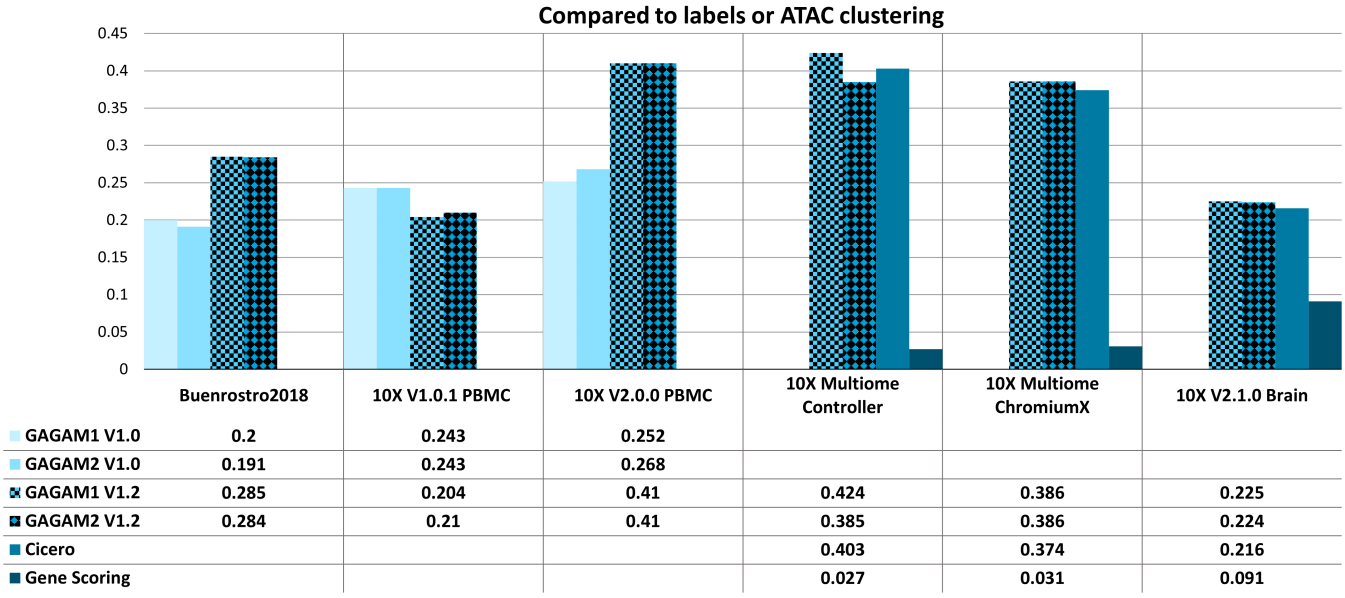
In general, results show how GAGAM v1.2 has good RAGI scores, which are either better or comparable to scores of other methods. Improvements here are less significant than those on ARI and AMI scores. Only the *10X v1.0.1 PBMC* dataset shows lower performance than the scATAC-seq clustering results. Consistently with what is shown in [10], although GeneScoring's clustering results are consistent with the ground truth classification (as indicated by ARI and AMI), RAGI results highlight how GeneScoring GAM performs poorly as for both RAGI scores (Figures 7 and 8) and the respective p-values over the tolerable threshold of 0.05 (data not shown, see **Data Availability Statement**), meaning they are not statistically significant. This highlights the importance of evaluating the GAMs performances on both aspects of clustering similarity to identify cellular heterogeneity and biological significance.

Overall, the results show that GAGAM1 v1.2 (the version with all three contributions) generally performs better, or at least equally, than GAGAM2 v1.2, highlighting that considering all three contributions is relevant to maximize performance. This emerges more clearly in this work than in [10].

4.5.4. Preliminary Comparison with Transcriptomics

Finally, the proposed validation strategy aims to support the comparison between epigenomic and transcriptomic data for multi-omics studies. As discussed, multi-omics datasets allow for comparing the GAGAM results with the scRNA-seq data. This work proposes a preliminary qualitative comparison between the activity and expression of DA genes identified by GAGAM v1.2 and scRNA-seq clustering analysis. At first, this process performs differential analysis on the GAGAM v1.2 clustering. This yields the most DA genes for each cluster. Then, to investigate if the activity and the expression of specific

genes are consistent, it provides two types of visualizations: violin plots and a heatmap, showing the activity and the expression levels of the most DA genes. It is possible to investigate if the activity and the expression of specific genes are consistent.



	Buenrostro2018	10X V1.0.1 PBMC	10X V2.0.0 PBMC	10X Multiome Controller	10X Multiome ChromiumX	10X V2.1.0 Brain
GAGAM1 V1.0	0.2	0.243	0.252			
GAGAM2 V1.0	0.191	0.243	0.268			
GAGAM1 V1.2	0.285	0.204	0.41	0.424	0.386	0.225
GAGAM2 V1.2	0.284	0.21	0.41	0.385	0.386	0.224
Cicero				0.403	0.374	0.216
Gene Scoring				0.027	0.031	0.091

Figure 8. RAGI results of both Old (**left**) and New (**right**) datasets when compared to labels or ATAC clustering. The Old datasets are compared with the previous GAGAM v1.0 from [10], while the New datasets are with Cicero and GeneScoring.

As an example, Figures 9 and 10 shows the results on both 10X multiome datasets, *10X Multiome Controller PBMC* and *10X Multiome Chromium X PBMC*, considering the GAGAM1 v1.2 (the version including all three contributions). This plot visualizes, for each cluster, the expression level distribution of the top DA considered genes across cells. Even if not the same, the activity and expression profiles tend to be consistent. Anyway, these preliminary results are promising. For example, Figure 10 shows how the activity profile of the gene MS4A1 (a well-known marker for PBMC cells) is highly consistent with its expression profile, meaning it is both active and expressed in the same cells.

Figures 11 and 12 provides heatmaps, again for both 10X multiome datasets, that allow a similar comparison, extending it to a broader panel of DA genes (specifically the top 10 DA genes per cluster). In addition, the heatmaps show the activity (on the top) and the expression (on the bottom) by cluster throughout all the datasets, allowing a more comprehensive investigation and comparison of differential activity and expression profiles, respectively. Again, Figure 11 shows overall consistency between the activity and expression of these genes throughout the cells of the dataset.

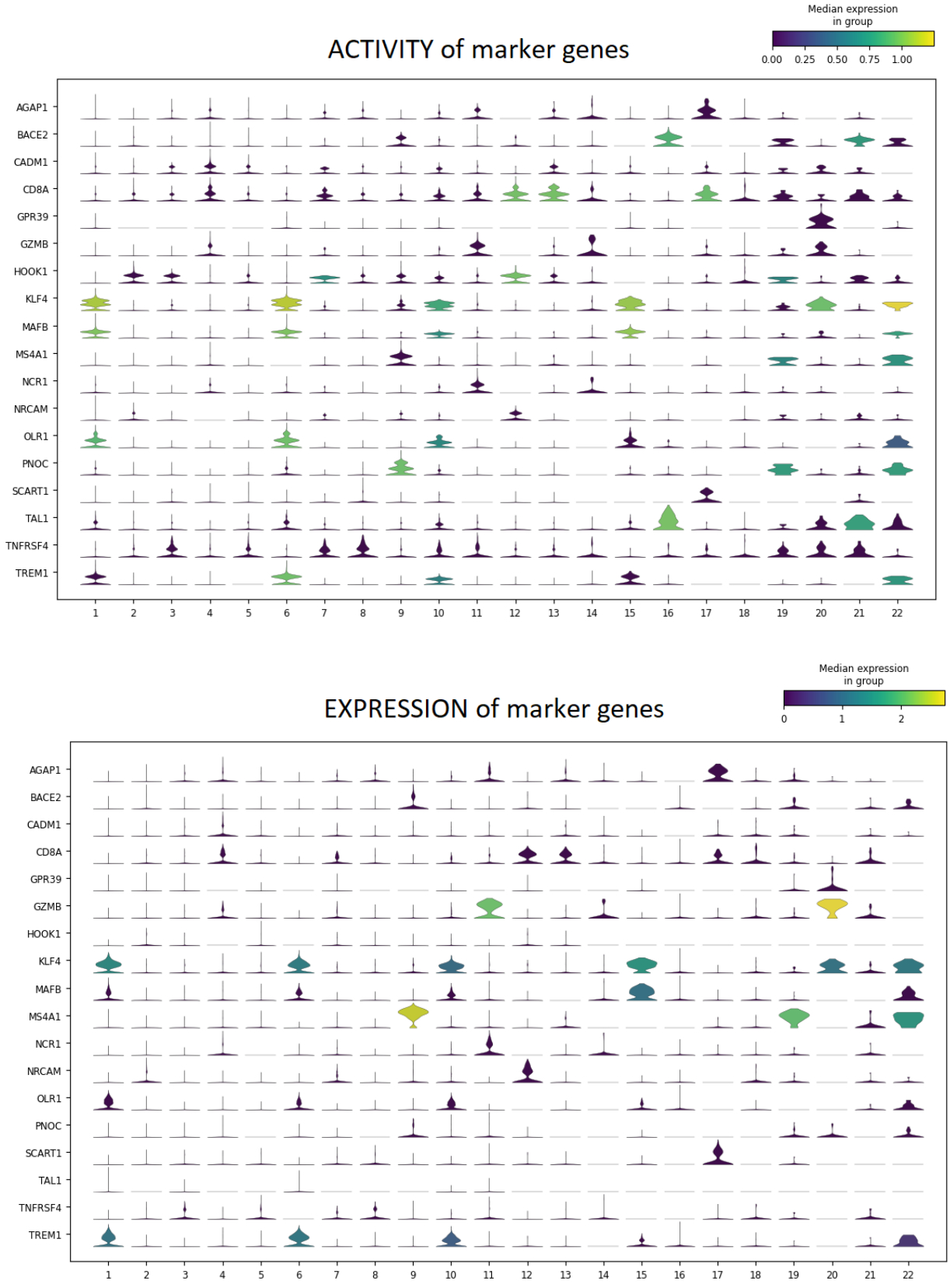


Figure 9. Violin Plots for the top DA genes for each cluster, of their Activity (on the **top**) and Expression (on the **bottom**), for the *10X Multiome Controller PBMC* dataset. The two figures show a general coherence (especially for genes like MS4A1, AGAP1, and KLF4) between the activity and expression profiles on the same cells, showing consistency between the two biological levels.

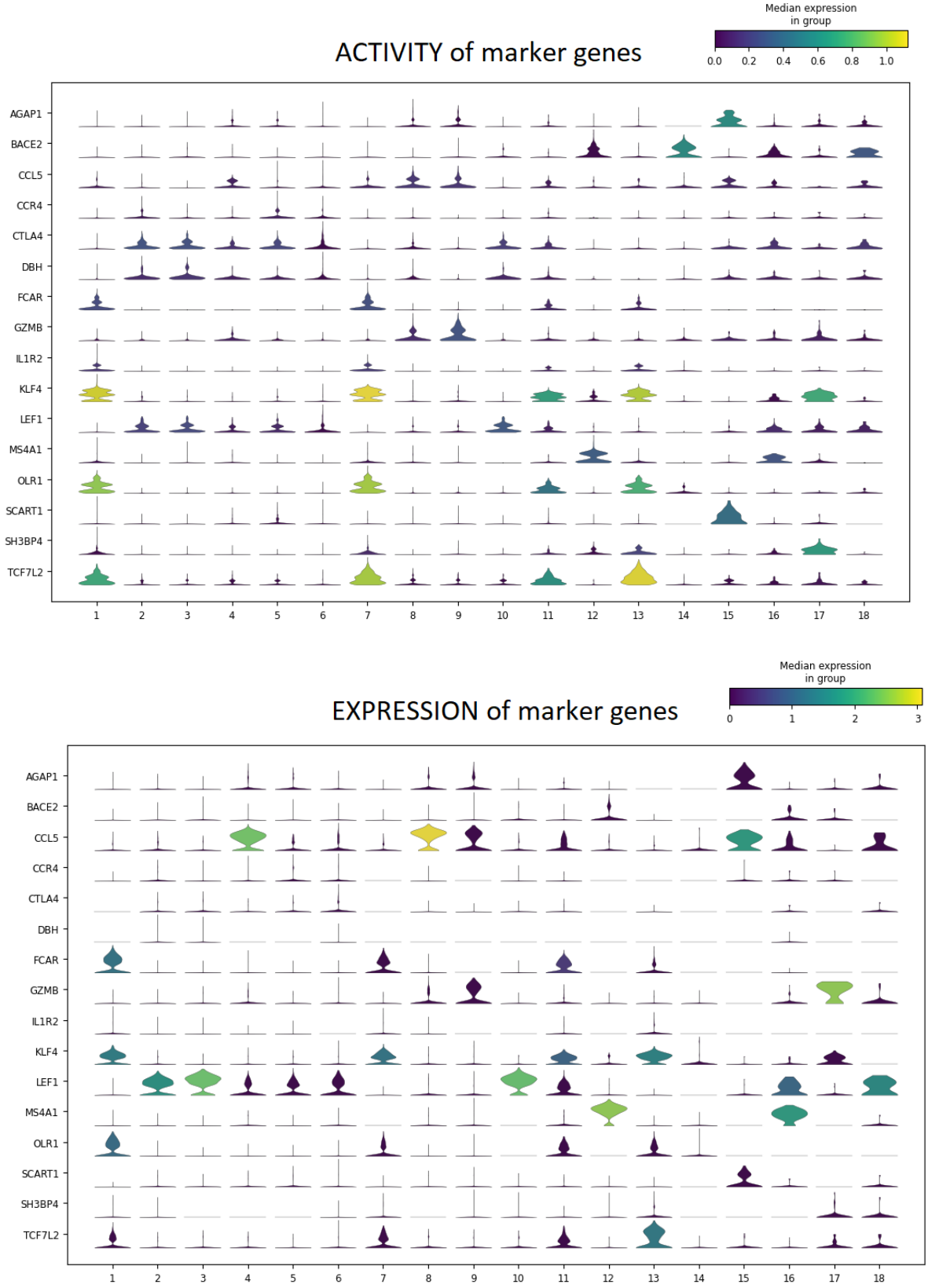


Figure 10. Violin Plots for the top DA genes for each cluster, of their Activity (on the **top**) and Expression (on the **bottom**), for the *10X Multiome Chromium X PBMC* dataset. The two figures show a general coherence (especially for genes like MS4A1, AGAP1, and KLF4) between the activity and expression profiles on the same cells, showing consistency between the two biological levels.

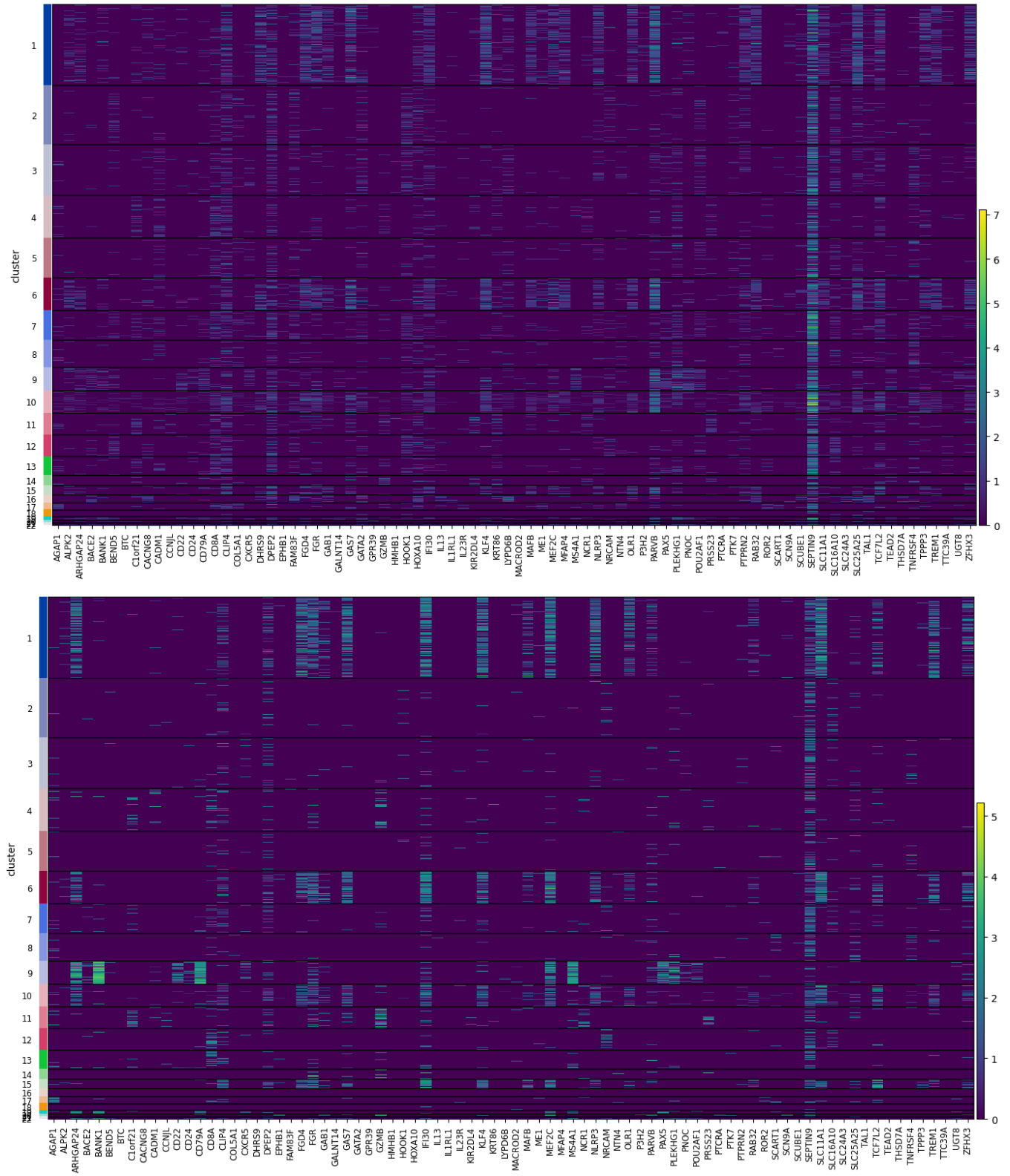


Figure 11. Heatmap for the top 10 DA genes, showing their Activity (on the **top**) and Expression (on the **bottom**) profiles throughout the whole *10X Multiome Controller PBMC* dataset. Again, one can notice repeated patterns between the two.

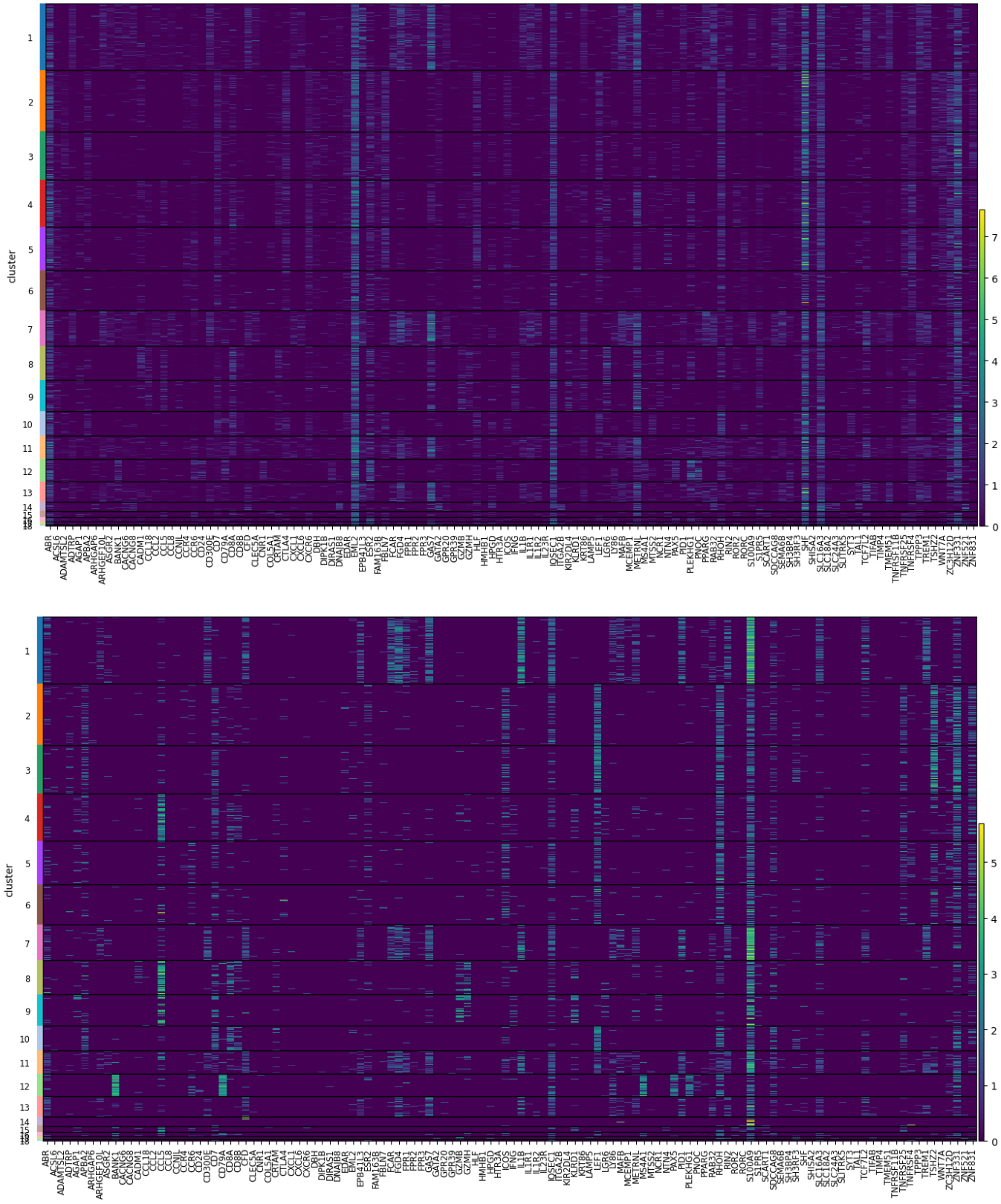


Figure 12. Heatmap for the top 10 DA genes, showing their Activity (on the **top**) and Expression (on the **bottom**) profiles throughout the whole *10X Multiome Chromium X PBMC* dataset. Again, one can notice repeated patterns between the two, even if, in this dataset, the discrepancies are more and the correlation is less clear.

However, as expected, there is no complete concordance, especially in the case of the *10X Multiome Chromium X PBMC* (which shows more discrepancies), but this is a preliminary qualitative investigation. Further investigation of these differences between the activity and expression of genes could be informative and will follow. In fact, of great interest will be to understand what the relationships are between highly correlated and poorly correlated activity-expression genes. One explanation of the discrepancies between activity and expression can derive from the different time scales between the two biological processes, meaning that the variation in gene expression has a faster dynamic than the epigenetic changes. Moreover, the scATAC-seq has a much lower resolution than the scRNA-seq, making it difficult to completely grasp the actual epigenetic state. Even if this is a preliminary and qualitative analysis, it demonstrates how the content of GAGAM v1.2 has a biological meaning that supports comparison with the gene expression. This is not trivial and establishes the GAGAM as a powerful tool to assess the relationship between accessibility and expression. More generally, a tool for multi-omics data analysis. Here the paper shows this first qualitative analysis that allows investigation of the general relationship of activity and expression patterns of some genes. However, as discussed in Section 3.1, the modular nature of GAGAM will allow easy follow-up analyses. Specifically, it will be interesting to inspect if different accessibility patterns of each contribution affect the expression levels, discriminating between highly expressed transcripts vs. moderate or low expressed transcripts. In addition, an interesting point would be to calculate the ratio between the promoter and exon accessibility to investigate its possible correlation with the gene expression subsequently. This will surely be a future improvement, sensibly extending GAGAM data exploration and interpretation capability.

5. Conclusions

The presented work focuses on the capabilities of GAGAM to yield a biologically significant and understandable way to compute gene activity and compare it with gene expression levels. In conclusion, GAGAM v1.2 improves on the previous version of GAGAM v1.0, maintaining its novelty as a method to obtain a GAM from scATAC-seq data. Furthermore, it is based on a model-driven approach leveraging genomic annotations of genes and functional elements.

GAGAM v1.2 improves the correct and thoughtful labeling of the peaks. Specifically, it enhances the precision of the *prom* labeling, the proper association with the right genes (covering some limit cases), the focus on exons instead of on the entire gene body, and the appropriate identification of *enhD* peaks from intronic regions.

Experimental results demonstrate how GAGAM v1.2 improved performances compared to the previous version, confirming its dominance towards Cicero and GeneScoring. Furthermore, GAGAM v1.2 improved its scores for all the metrics employed, showing that the GAGAM v1.2-based clustering is comparable with the one obtained from the raw scATAC-seq data, supporting the claim that GAGAM v1.2 can identify cellular heterogeneity.

In addition, GAGAM is a suitable method to interpret accessibility data in general. Indeed, since it employs genes as features, it allows analyzing scATAC-seq data through well-studied and investigated concepts like marker genes. The same analysis would not be possible with raw accessibility data. The analysis of the RAGI score supports this claim and highlights the gene activity differences between marker and housekeeping genes. The direct comparison with transcriptomics further and more coherently supports this. Indeed, as a further step, the activity of the genes is compared with their expression on the same cells, highlighting consistency between the two biological levels.

The GAGAM approach stems from a radical innovation of genomic annotation of accessibility data toward multimodal single-cell data analysis. The GAGAM genomic model captures non-trivial contributions of different genomic regions to the potential functional regulation of transcriptions at a gene, interpreting and weighing them in a separate and specific way, and rationally organizing them in the final GAS computation.

This fully interprets the final GAS and supports biologically meaningful accessibility data analysis and direct comparison with gene expression levels.

On the other hand, the non-triviality of the comparisons between activity and expression data highlights the importance of interpretability and biological significance of GAS computation. The genomic model at the core of GAGAM computation works and will further develop in this direction, to characterize the complex relationship between genomic accessibility and transcription. The modularity of final matrix computation allows the separate handle of the contribution of well-defined parts, quantifying the impact of each contribution to the final GAGAM computation and subsequent analysis. For example, it will be exciting to investigate whether different gene expression levels correspond to specific patterns in overall gene activity or any specific individual contribution of GAGAM. Lastly, the modular nature of GAGAM will enable the seamless extension of the genomic model in a modular way, progressively including new contributions in the final GAGAM computation and the possibility to select which ones to include every time. For instance, this could ease the implementation of a GAGAM version, which will include different isoforms of genes. Generally, it provides a tool for the bioinformatician to tailor the GAGAM application for specific purposes, explicitly evaluating and transparently choosing which contributions to consider.

Besides the improvement and modular expansion of the genomic model, another direction for improvement for GAGAM is toward usability. In the direction of existing analysis pipelines, accessibility and usability are essential to allow the bioinformatics community to leverage the GAGAM approach's capabilities easily. Thus it would be valuable to develop the GAGAM pipeline to include an improved data exploration and interpretation capability. Furthermore, future works aim to allow easier customization of the analysis parameters to allow users to adapt the GAGAM pipeline to their specific analyses. In this direction, this work already provides the complete analysis pipeline and guidance to reproduce the proposed experiments (see *Data Availability statement*).

In conclusion, GAGAM v1.2 is a promising and reliable way to interpret scATAC-seq data, which focuses on the accessibility of the genes and their regulatory elements, with the potential to act as a direct link between epigenomic and transcriptomic, proving to be a promising tool, especially for multi-omic data analysis.

Author Contributions: Conceptualization, L.M.; methodology, L.M. and R.B.; software, L.M. and R.B.; validation, L.M. and R.B.; formal analysis, L.M. and R.B.; investigation, L.M. and R.B.; resources, L.M. and R.B.; data curation, L.M. and R.B.; writing—original draft preparation, L.M., R.B. and S.D.C.; writing—review and editing, L.M., R.B., A.S. and S.D.C.; visualization, L.M., R.B., A.S. and S.D.C.; supervision, A.S. and S.D.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in NCBI GENE Expression Omnibus (GEO) with accession number GSE96769, and from the freely available 10XGenomic platform [24] datasets. All the code and material created are available at https://github.com/smilies-polito/GAGAM Accessed on 29 December 2022.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, Y.; Ma, L.; Wu, D.; Chen, G. Advances in bulk and single-cell multi-omics approaches for systems biology and precision medicine. *Brief. Bioinform.* **2021**, *22*, bbab024. [CrossRef] [PubMed]
2. Hu, Y.; An, Q.; Sheu, K.; Trejo, B.; Fan, S.; Guo, Y. Single cell multi-omics technology: Methodology and application. *Front. Cell Dev. Biol.* **2018**, *6*, 28. [CrossRef]
3. Carter, B.; Zhao, K. The epigenetic basis of cellular heterogeneity. *Nat. Rev. Genet.* **2021**, *22*, 235–250. [CrossRef]

4. Kelsey, G.; Stegle, O.; Reik, W. Single-cell epigenomics: Recording the past and predicting the future. *Science* **2017**, *358*, 69–75. [CrossRef] [PubMed]
5. Subramanian, I.; Verma, S.; Kumar, S.; Jere, A.; Anamika, K. Multi-omics data integration, interpretation, and its application. *Bioinform. Biol. Insights* **2020**, *14*, 1177932219899051. [CrossRef] [PubMed]
6. Preissl, S.; Gaulton, K.J.; Ren, B. Characterizing cis-regulatory elements using single-cell epigenomics. *Nat. Rev. Genet.* **2022**, *2022*, 1–23. [CrossRef]
7. Pliner, H.A.; Packer, J.S.; McFaline-Figueroa, J.L.; Cusanovich, D.A.; Daza, R.M.; Aghamirzaie, D.; Srivatsan, S.; Qiu, X.; Jackson, D.; Minkina, A.; et al. Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol. Cell* **2018**, *71*, 858–871. [CrossRef]
8. Shashikant, T.; Ettensohn, C.A. Genome-wide analysis of chromatin accessibility using ATAC-seq. *Methods Cell Biol.* **2019**, *151*, 219–235.
9. Minnoye, L.; Marinov, G.K.; Krausgruber, T.; Pan, L.; Marand, A.P.; Secchia, S.; Greenleaf, W.J.; Furlong, E.E.; Zhao, K.; Schmitz, R.J.; et al. Chromatin accessibility profiling methods. *Nat. Rev. Methods Prim.* **2021**, *1*, 1–24. [CrossRef]
10. Martini, L.; Bardini, R.; Savino, A.; Di Carlo, S. GAGAM: A Genomic Annotation-Based Enrichment of scATAC-seq Data for Gene Activity Matrix. In *Proceedings of the Bioinformatics and Biomedical Engineering*; Rojas, I., Valenzuela, O., Rojas, F., Herrera, L.J., Ortuño, F., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2022; pp. 18–32.
11. Santiago-Rodriguez, T.M.; Hollister, E.B. Multi omic data integration: A review of concepts, considerations, and approaches. In *Seminars in Perinatology*; Elsevier: Amsterdam, The Netherlands, 2021; Volume 45, p. 151456.
12. Aldridge, S.; Teichmann, S.A. Single cell transcriptomics comes of age. *Nat. Commun.* **2020**, *11*, 4307 . [CrossRef]
13. Ståhlberg, A.; Kubista, M. Technical aspects and recommendations for single-cell qPCR. *Mol. Asp. Med.* **2018**, *59*, 28–35. [CrossRef] [PubMed]
14. Tang, F.; Barbacioru, C.; Wang, Y.; Nordman, E.; Lee, C.; Xu, N.; Wang, X.; Bodeau, J.; Tuch, B.B.; Siddiqui, A.; et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **2009**, *6*, 377–382. [CrossRef] [PubMed]
15. Stein, C.M.; Weiskirchen, R.; Damm, F.; Strzelecka, P.M. Single-cell omics: Overview, analysis, and application in biomedical science. *J. Cell. Biochem.* **2021**, *122*, 1571–1578. [CrossRef]
16. Ahn, J.; Heo, S.; Lee, J.; Bang, D. Introduction to single-cell DNA methylation profiling methods. *Biomolecules* **2021**, *11*, 1013. [CrossRef] [PubMed]
17. Niemöller, C.; Wehrle, J.; Riba, J.; Claus, R.; Renz, N.; Rhein, J.; Bleul, S.; Stosch, J.M.; Duyster, J.; Plass, C.; et al. Bisulfite-free epigenomics and genomics of single cells through methylation-sensitive restriction. *Commun. Biol.* **2021**, *4*, 153. [CrossRef] [PubMed]
18. Nakato, R.; Sakata, T. Methods for ChIP-seq analysis: A practical workflow and advanced applications. *Methods* **2021**, *187*, 44–53. [CrossRef]
19. Ma, S.; Zhang, Y. Profiling chromatin regulatory landscape: Insights into the development of ChIP-seq and ATAC-seq. *Mol. Biomed.* **2020**, *1*, 9. [CrossRef]
20. Kaya-Okur, H.S.; Wu, S.J.; Codomo, C.A.; Pledger, E.S.; Bryson, T.D.; Henikoff, J.G.; Ahmad, K.; Henikoff, S. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **2019**, *10*, 1930.
21. Harada, A.; Kimura, H.; Ohkawa, Y. Recent advances in single-cell epigenomics. *Curr. Opin. Struct. Biol.* **2021**, *71*, 116–122. [CrossRef]
22. Buenrostro, J.D.; Wu, B.; Litzenburger, U.M.; Ruff, D.; Gonzales, M.L.; Snyder, M.P.; Chang, H.Y.; Greenleaf, W.J. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **2015**, *523*, 486–490. [CrossRef]
23. Xu, W.; Wen, Y.; Liang, Y.; Xu, Q.; Wang, X.; Jin, W.; Chen, X. A plate-based single-cell ATAC-seq workflow for fast and robust profiling of chromatin accessibility. *Nat. Protoc.* **2021**, *16*, 4084–4107. [CrossRef] [PubMed]
24. Satpathy Ansuman T., Granja Jeffrey M., Y.K.E.; et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **2019**, *37*, 925–936. [CrossRef] [PubMed]
25. Danese A., Richter M.L., C.K.; et al. EpiScanpy: Integrated single-cell epigenomic analysis. *Nat. Commun.* **2021**, *12*, 5228. [CrossRef] [PubMed]
26. Yan, F.; Powell, D.R.; Curtis, D.J.; Wong, N.C. From reads to insight: A hitchhiker's guide to ATAC-seq data analysis. *Genome Biol.* **2020**, *21*, 22. [CrossRef]
27. Lareau, C.A.; Duarte, F.M.; Chew, J.G.; Kartha, V.K.; Burkett, Z.D.; Kohlway, A.S.; Buenrostro, J.D. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.* **2019**, *37*, 916–924. [CrossRef]
28. Stuart, T.; Srivastava, A.; Madad, S.; Lareau, C.A.; Satija, R. Single-cell chromatin state analysis with Signac. *Nat. Methods* **2021**, *18*, 1333–1341. [CrossRef]
29. Hubert, L.; Arabie, P. Comparing partitions. *J. Classif.* **1985**, *2*, 193–218. [CrossRef]
30. Vinh, N.X.; Epps, J.; Bailey, J. Information Theoretic Measures for Clusterings Comparison: Is a Correction for Chance Necessary? In Proceedings of the 26th Annual International Conference on Machine Learning; Association for Computing Machinery, Montreal, QC, Canada, 14–18 June 2009; pp. 1073–1080. [CrossRef]

31. Chen, H.; Lareau, C.; Andreani, T.; Vinyard, M.E.; Garcia, S.P.; Clement, K.; Pinello, L. Assessment of Computational Methods for the Analysis of Single-Cell ATAC-Seq Data. *Genome Biol.* **2019**, *20*, 241–241. [CrossRef]
32. O'Leary, N.A.; Wright, M.W.; Brister, J.R.; Ciufo, S.; Haddad, D.; McVeigh, R.; ; Pruitt, K.D. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **2016**, *44*, D733–45. [CrossRef]
33. Thibaud-Nissen, Francoise and Souvorov, Alexander and Murphy, Terence and DiCuccio, Michael and Kitts, Paul Eukaryotic Genome Annotation Pipeline. In *The NCBI Handbook [Internet]*; 2nd ed.; National Center for Biotechnology Information (US): Bethesda, MD, USA, 2013. Available online: https://www.ncbi.nlm.nih.gov/books/NBK169439/ (accessed on 29 December 2022).
34. Moore, J.E.; Purcaro, M.J.; Pratt, H.E.; Epstein, C.B.; Shoresh, N.; Adrian, J.; Kawli, T.; Davis, C.A.; Dobin, A.; Kaul, R.; et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **2020**, *583*, 699–710.
35. ENCODE. ENCODE Data Portal. Available online: https://www.encodeproject.org (accessed on 29 December 2022).
36. Kent, W.J.; Zweig, A.S.; Barber, G.; Hinrichs, A.S.; Karolchik, D. BigWig and BigBed: Enabling browsing of large distributed datasets. *Bioinformatics* **2010**, *26*, 2204–2207. [CrossRef] [PubMed]
37. UCSC. UCSC Human cCRE Track Download. Available online: http://hgdownload.soe.ucsc.edu/gbdb/hg38/encode3/ccre/ (accessed on 29 December 2022).
38. UCSC. UCSC Mouse cCRE Track Download. Available online: http://hgdownload.soe.ucsc.edu/gbdb/mm10/encode3/ccre/ (accessed on 29 December 2022).
39. Luu, P.L.; Ong, P.T.; Dinh, T.P.; Clark, S.J. Benchmark study comparing liftover tools for genome conversion of epigenome sequencing data. *NAR Genom. Bioinform.* **2020**, *2*, lqaa054. [CrossRef] [PubMed]
40. UCSC. bigBedToBed too. Available online: https://genomebrowser.wustl.edu/goldenPath/help/bigBed.html (accessed on 29 December 2022).
41. Romano, S.; Vinh, N.X.; Bailey, J.; Verspoor, K. Adjusting for Chance Clustering Comparison Measures. *J. Mach. Learn. Res.* **2016**, *17*, 4635–4666.
42. Martini, L.; Bardini, R.; Di Carlo, S. Meta-Analysis of cortical inhibitory interneurons markers landscape and their performances in scRNA-seq studies. In Proceedings of the 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Houston, TX, USA, 9–12 December 2021; pp. 253–258. [CrossRef]
43. Eli, E.; Levanon, E.Y. Human housekeeping genes, revisited. *Trends Genet. TIG* **2013**, *29*, 569–74.
44. Hounkpe, B.W.; Chenou, F.; de Lima, F.; De Paula, E.V. HRT Atlas v1.0 database: Redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. *Nucleic Acids Res.* **2021**, *49*, D947–D955. [CrossRef]
45. Pliner, H.A. Supervised classification enables rapid annotation of cell atlases. *Nat. Methods* **2019**, *16*, 983–986. [CrossRef]
46. Institute, A. © 2010 Allen Cell Types Database. Available online: https://portal.brain-map.org/atlases-and-data/rnaseq (accessed on 29 December 2022).
47. Martini, L. Study of Cellular Heterogeneity of Mouse Cerebral Cortex, through Joint scRNA-seq and scATAC-seq Analysis, Derived from SNARE-seq Technique., Supervisors Stefano Di Carlo, Roberta Bardini. Politecnico di Torino, Corso di Laurea Magistrale in Physics of Complex Systems (Fisica Dei Sistemi Complessi), 2020 Available online: http://webthesis.biblio.polito.it/id/eprint/16750 (accessed on 29 December 2022).
48. 10XGenomics. 5k Peripheral Blood Mononuclear Cells (PBMCs) from a HEALTHY donor Single Cell ATAC Dataset by Cell Ranger ATAC 1.0.1, 10x Genomics, (2019, December 17th). Available online: https://www.10xgenomics.com/resources/datasets/5-k-peripheral-blood-mononuclear-cells-pbm-cs-from-a-healthy-donor-1-standard-1-0-1 (accessed on 29 December 2022).
49. 10XGenomics. Peripheral Blood Mononuclear Cells (PBMCs) from a Healthy Donor Single Cell ATAC Dataset by Cell Ranger ATAC 2.0.0, 10x Genomics, (2021, May 3rd). Available online: https://www.10xgenomics.com/resources/datasets/10-k-peripheral-blood-mononuclear-cells-pbm-cs-from-a-healthy-donor-next-gem-v-1-1-1-1-standard-2-0-0 (accessed on 29 December 2022).
50. Buenrostro, J.D.; Corces, M.R.; Lareau, C.A.; Wu, B.; Schep, A.N.; Aryee, M.J.; Greenleaf, W.J. Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **2018**, *173*, 1535–1548.e16. [CrossRef]
51. 10XGenomics. 8k Adult Mouse Cortex Cells from a Healthy Donor Single Cell ATAC Dataset by Cell Ranger ATAC 2.1.0, 10x Genomics, (2022, March 29th). Available online: https://www.10xgenomics.com/resources/datasets/8k-adult-mouse-cortex-cells-atac-v1-1-chromium-x-1-1-standard (accessed on 29 December 2022).
52. 10XGenomics. 10k Peripheral blood mononuclear cells (PBMCs) from a healthy donor Single Cell Multiome ATAC + Gene Expression Dataset by Cell Ranger ARC 2.0.0, 10x Genomics, (2021, August 9th). Available online: https://www.10xgenomics.com/resources/datasets/10-k-human-pbm-cs-multiome-v-1-0-chromium-controller-1-standard-2-0-0.
53. Trapnell, C.; Cacchiarelli, D.; Grimsby, J.; Pokharel, P.; Li, S.; Morse, M.; Rinn, J.L. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **2014**, *32*, 381–386. [CrossRef]

54. McInnes, L.; Healy, J.; Saul, N.; Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **2018**, *3*, 861. [CrossRef]
55. Chiquet, J. Aricode: Efficient Computations of Standard Clustering Comparison Measures. Available online: https://cran.r-project.org/web/packages/aricode/index.html (accessed on 29 December 2022).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.