genes

MDPI

Article

Cross-Omic Transcription Factor Analysis: An Insight on Transcription Factor Accessibility and Expression Correlation

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

Control and Computer Engineering Department, Politecnico di Torino, 10129 Torino, Italy; lorenzo.martini@polito.it (L.M.); roberta.bardini@polito.it (R.B.); alessandro.savino@polito.it (A.S.)
* Correspondence: stefano.dicarlo@polito.it

Abstract: It is well known how sequencing technologies propelled cellular biology research in recent years, providing incredible insight into the basic mechanisms of cells. Single-cell RNA sequencing is at the front in this field, with single-cell ATAC sequencing supporting it and becoming more popular. In this regard, multi-modal technologies play a crucial role, allowing the possibility to simultaneously perform the mentioned sequencing modalities on the same cells. Yet, there still needs to be a clear and dedicated way to analyze these multi-modal data. One of the current methods is to calculate the Gene Activity Matrix (GAM), which summarizes the accessibility of the genes at the genomic level, to have a more direct link with the transcriptomic data. However, this concept is not well defined, and it is unclear how various accessible regions impact the expression of the genes. Moreover, the transcription process is highly regulated by the transcription factors that bind to the different DNA regions. Therefore, this work presents a continuation of the meta-analysis of Genomic-Annotated Gene Activity Matrix (GAGAM) contributions, aiming to investigate the correlation between the TF expression and motif information in the different functional genomic regions to understand the different Transcription Factors (TFs) dynamics involved in different cell types.

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

Citation: Martini, L.; Bardini, R.; Savino, A.; Di Carlo, S. Cross-Omic Transcription Factor Analysis: An Insight on Transcription Factor Accessibility and Expression Correlation. *Genes* **2024**, *15*, 268. https://doi.org/10.3390/genes15030268

Academic Editors: Ignacio Rojas, Jan Urban, Francisco Ortuño, Olga Valenzuela and Jean-Fred Fontaine

Received: 16 January 2024
Revised: 13 February 2024
Accepted: 17 February 2024
Published: 21 February 2024

1. Introduction

Next Generation Sequencing (NGS) technologies serve as the backbone for cutting-edge cellular biology research, offering a powerful tool to investigate fundamental cell mechanisms. These technologies, especially single-cell RNA sequencing (scRNA-seq) and single-cell assays for transposase-accessible chromatin sequencing (scATAC-seq), significantly contribute to studying cellular states with high resolution, a critical aspect for understanding cellular heterogeneity.

Widely utilized for profiling thousands of single-cell transcriptional profiles, scRNA-seq enables the investigation of cellular heterogeneity based on gene expression [1–3]. Simultaneously, the emerging popularity of scATAC-seq proves invaluable. This technology, by probing the entire genome and assessing accessible chromatin regions, offers complementary insights into gene regulation processes [4] and expression [5]. While the integration of scRNA-seq and scATAC-seq through multi-modal technologies is becoming crucial for understanding cell-related phenomena, the inherent differences in data types between the two technologies pose challenges to joint analysis [6–8]. Correlating the accessibility of a genomic region with gene expression is not straightforward due to the intricate machinery involved in transcriptional regulation. scRNA-seq datasets prioritize genes as prominent features, while scATAC-seq datasets consider genomic regions as features, making their integration challenging.

To bridge this gap, the gene activity (GA) concept is introduced, summarizing genomic accessibility information in a form where features are genes, allowing direct comparison with scRNA-seq matrices [9]. However, defining the relationship between accessible regions and genes remains unclear.

The Genomic-Annotated Gene Activity Matrix (GAGAM) approach [10,11] proposes a promising solution, relying on a genomic model based on annotations to associate genomic regions with accessible genes. This approach constructs a Gene Activity Matrix (GAM) with contributions from different functional genomic regions (promoters, exons, and enhancers). Although GAGAM better models the gene regulatory landscape, it lacks the representation of the complex gene regulation mechanisms [12], especially the involvement of Transcription Factors (TFs).

This work aims to address this gap by the preliminary analyzing the correlation between TF expression and the accessibility of their motifs. Specifically, it explores differences in motif accessibility in promoter and enhancer regions, aiming to tailor TF information with GAGAM contributions in a nuanced manner. This work is an extension of the previously published work documented in [13].

2. Background

To comprehend the proposed analysis, it is essential to introduce the fundamental technologies underpinning this work.

2.1. Single-Cell Sequencing Technologies

A short overview of the scATAC-seq data organization aids in understanding the derived concept of Gene Activity (GA). scATAC-seq is a technology offering insights into the epigenomic state of cells by probing the entire genome. It utilizes the Tn5-transposase to identify regions where chromatin is open, and DNA sequences are accessible [14]. This technology enables the investigation not only of genes, as in scRNA-seq, but also of various functional elements such as enhancers and promoters scattered throughout the genome, crucial for gene regulation [15,16].

While scRNA-seq data use genes as primary features, scATAC-seq data utilize peaks, i.e., short genomic regions described by their coordinates on chromosomes. This difference poses a significant challenge when correlating the two biological levels. One approach to overcome this hurdle is transforming peaks into gene-like data and comparing the two technologies. As the introduction mentions, GA serves as one such method [6].

However, current models for defining GA often oversimplify the relationship between a gene and the accessibility of its genomic region. Certain approaches, such as GeneScoring [17] and Signac [18], indiscriminately consider peak signals overlapping gene body regions without distinguishing between coding and non-coding regulatory elements. In contrast, Cicero [6] adopts a more structured approach, considering various regulatory regions but collapsing the gene region to a single base. These methods retain minimal biological information from raw scATAC-seq data, primarily related to gene-coding regions, despite only representing a small percentage of the entire signal [19].

Beyond these simplistic models, other approaches aim to encompass more accessible genomic regions and their impact on the overall GA. This work specifically employs GAGAM, utilizing curated genomic annotations to functionally label peaks and compute distinct contributions [10].

2.2. GAGAM

GAGAM uses information on various DNA regions, particularly exons and non-coding regions with regulatory roles, to improve the analysis of biological information from scATAC-seq data. This model-driven approach aims to support the study of cellular heterogeneity better. GAGAM lays the groundwork for a detailed investigation into the relationship between accessibility and expression in single-cell data. Its modular structure allows the independent computation of contributions, facilitating specific and separate investigations, which is especially crucial when considering the role of regulatory regions with challenging relationships to gene expression. Table 1 briefly overviews the three contributions constituting GAGAM, to provide the necessary background.

Table 1. Description of GAGAM contribution variables.

Label	Contribution	Description
`prom`	$\mathbf{D}^{prom}_{\lvert P^p \rvert \times \lvert C \rvert}$	Contribution from promoter peaks P^p, i.e., peaks overlapping promoter signatures from ENCODE cCREs annotation. They are linked to the genes by proximity to the nearest TSS of a protein-coding gene.
`exon`	$\mathbf{D}^{exon}_{\lvert P^i \rvert \times \lvert C \rvert}$	Contribution from exon peaks P^i, i.e., peaks overlapping exon regions from the NCBI RefSeq Genes annotation. They are linked to the genes they are in.
`enhD`	$\mathbf{D}^{enhd}_{\lvert P^e \rvert \times \lvert C \rvert}$	Contribution from enhancer peaks P^e, i.e., peaks overlapping enhancer signatures from ENCODE cCREs annotation. They are linked to the genes by co-accessibility scores with promoter peaks.

GAGAM operates solely on preprocessed scATAC-seq data organized in the form of a matrix $\mathbf{D}_{|P|\times|C|}$, where P is the set of peaks in the dataset, and C is the set of available cells. As described in Table 1 GAGAM utilizes the UCSC Genome Browser [20] to obtain genomic annotations, labeling all peaks $p \in P$ overlapping with regions of interest, assigning labels `prom`, `exon`, `enhD` to peaks, linking accessible peaks to their biological functions. The original dataset $\mathbf{D}_{|P|\times|C|}$ is then split into three subsets based on the three sets of labeled peaks: $\mathbf{D}^{prom}_{|P^p|\times|C|}$, $\mathbf{D}^{enhd}_{|P^e|\times|C|}$, $\mathbf{D}^{exon}_{|P^i|\times|C|}$. These three matrices are further processed to obtain the final gene activity matrix. For a further and more in-depth explanation of the GAGAM computation, interested readers may refer to [10,11] for a detailed description.

The current GAGAM model, while implementing cis-regulatory elements to calculate activity scores, solely considers their accessibility. Although it is a good indicator for assessing involvement in transcription, it overlooks the interaction with TFs. scATAC-seq data, however, offer ways to investigate TF interactions. Identifying DNA sequences where TFs can bind, known as Transcription Factor Binding Motifs (TFBMs), is feasible. However, these motifs, i.e., short sequences of a maximum of a dozen base pairs (bp), present limitations compared to the hundreds of bp in peaks [5]. Additionally, a motif does not guarantee TF binding, as the specific TF must be expressed and transcribed by cells to be actively involved in regulation. When a TF is bound to DNA, the region becomes inaccessible to the Tn5 Transposase used in scATAC-seq experiments, leaving a detectable footprint in the signal [14]. However, due to the sparsity of single-cell data, TF footprints are not measurable for each region with a specific TFBM but require studying the average signal from all motif instances.

Various types of information are available to study TFs in single-cell experiments, but looking at only one aspect has limitations. This work proposes a preliminary assessment of the correlation between the TFBM enrichment, TF footprint signal, and actual TF expression. Understanding the intricate dynamics of the TF contribution is crucial for proper modeling in GAGAM. The analysis examines motifs in promoter and enhancer regions independently, as this separation is central to GAGAM, and different TF interactions occur in these functional regions.

3. Materials and Methods

3.1. Dataset

This work requires a multi-omic dataset to allow a direct comparison between the epigenetic information (from scATAC-seq) and the gene expression (from scRNA-seq). The dataset of choice is an open access dataset from the 10X Genomics platform, consisting of 10,691 cells from adult murine peripheral blood mononuclear cell (PBMC) [21]. The scATAC-seq part of the dataset has a total of 115,179 peaks as features, while the scRNA-seq part has 36,601 genes. The tools employed to process and elaborate the data are GAGAM (the focus of this paper, accessible from [10]) and Seurat (v 5.0.1) [1]. The latter is one of the most well-known and highly utilized single-cell pipelines. This allows the processing of the

datasets, which is beneficial since it supports a data structure tailored to contain the results of different epigenetic analyses. Moreover, Seurat provides a dataset integration approach to label the cells with known cell-type labels by employing an external reference dataset. In this way, the cells are divided into cell-type clusters representing the ground truth of the following analyses. The dataset underwent a quality control check and preprocessing following standard thresholds. In detail, for gene expression, the process filters out cells with over 2500 or less than 200 unique gene counts and cells with more than 5% mitochondrial reads. For chromatin accessibility, it filters out cells with over 30,000 and less than 3000 fragments reads and then we also checked for nucleosome signal and TSS enrichment. Then, normalization, scaling, principal component analysis (PCA), uniform manifold approximation and projection (UMAP), and clustering for gene expression have been implemented, while the chromatin accessibility underwent normalization, latent semantic indexing (LSI), UMAP, and clustering.

3.2. Aggregated Cells

Before starting with the actual meta-analysis, it is worth noting that scRNA-seq and scATAC-seq only detect a tiny fraction of the actual signal from each cell (around 10–45% for scRNA-seq and only 1–10% for scATAC-seq [19]). This translates into considerable sparsity for the data. For each cell, the dataset contains several zero entries that could be false negatives [22]. This characteristic introduces noise when trying to correlate accessibility and expression. For this reason, this work explores the idea of performing the analysis based on the concept of *aggregated cell* behavior. Specifically, it aggregates cells from the same cell types obtained from the Seurat integration, representing the average over groups of cells instead of single cells (Figure 1d). In this way, this work computes the correlation not on the single cells $c \in C$ (where C is the set of cells of the dataset) but on the aggregated cells $ct \in CT$ (where CT is the set of cell-types) representing the average behavior over the cell types.

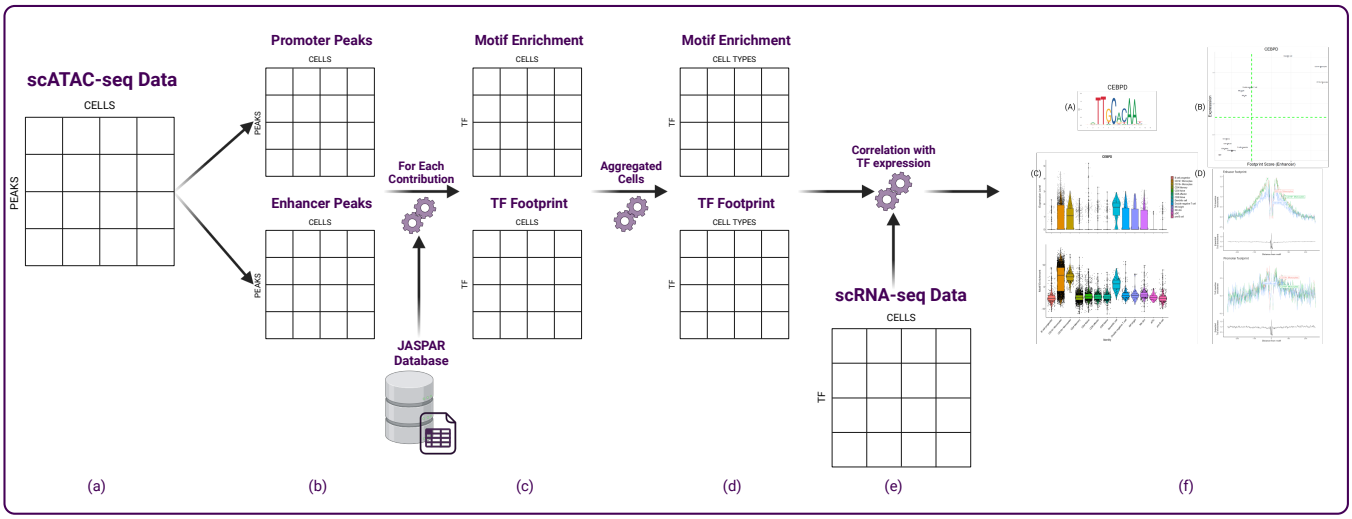

Figure 1. Workflow: (**a**) The scATAC-seq data serve as the input. (**b**) The data are partitioned into GAGAM contributions, resulting in promoter and enhancer matrices. (**c**) For each contribution, utilizing the JASPAR database, both motif enrichment and TF footprint scores are determined for all TFs expressed in the dataset. (**d**) Cells are aggregated based on cell-type annotations. (**e**) The two matrices obtained for each contribution are then compared to the expression matrix to analyze their correlation. (**f**) Results analysis reveals specific correlations and dynamics between TF expression and its motif information, with plots showing the TF motif (A), the motif enrichment (C), and the footprint scores (B,D)

3.3. TF Motifs and Motif Enrichment Analysis

TFBMs are DNA sequences that bind to transcription factors. Known TFBMs are curated in JASPAR [23], an open access database storing manually curated transcription factors binding profiles across multiple species of eukaryotes. These TFBMs are not exact sequences of nucleotides since there is a natural redundancy of sequences recognized by the TFs. Therefore, the motif information is stored in a Position Frequencies Matrix (PFM), representing the probability of finding a specific base for each nucleotide in the binding region. Given that the motifs are short sequences (6–12 bp), it is likely to find redundant and non-relevant matches when searching for DNA sequences matching the motif. Therefore, instead of searching for all possible matches inside accessible regions, it is more common to implement a motif enrichment analysis. In other words, this type of analysis identifies how much a known motif is over- or under-represented in a cell's accessible regions [14]. To do so, this work employs ChromVAR [24], an R tool for analyzing sparse chromatin accessibility data, providing reliable motif enrichment functions. The function requires the motif from JASPAR and calculates the enrichment scores for each of the $m \in M$ motif inside each cell, obtaining a final matrix of $\mathbf{ME}_{|M|\times|C|}$. The set of M motifs in this work does not comprehend the total 632 human motifs provided by JASPAR. It narrows it to the motifs whose corresponding TFs are also expressed in the dataset since the focus is the correlation of the motif information with the TFs expression. However, instead of performing, as commonly done, this analysis on the dataset as a whole, in this work, we are interested in investigating the differences between promoter and enhancer regions. Therefore, given the two scATAC-seq sub-matrices $\mathbf{D}^{prom}_{|P^p|\times|C|}$, $\mathbf{D}^{enhd}_{|P^e|\times|C|}$, the motif enrichment calculation is performed on them separately (Figure 1c), meaning it will capture the over- or under-representation of the motif specifically in promoter and enhancer regions. This calculation results in two matrices $\mathbf{ME}^{prom}_{|M|\times|C|}$ and $\mathbf{ME}^{enhd}_{|M|\times|C|}$, subsequently transformed into their aggregated forms $\mathbf{ME}^{prom}_{|M|\times|CT|}$ and $\mathbf{ME}^{enhd}_{|M|\times|CT|}$, that this work analyzes separately (Figure 1c).

3.4. TF Footprints

TFBMs are good indicators for inferring the interaction between DNA and TFs. However, they only give information on the possible binding locations but do not capture the actual binding events. Indeed, of the millions of motifs detected on the DNA, only a small portion are actual binding regions, and even less will be relevant in a certain cell type. Fortunately, scATAC-seq data can help investigate these binding events through the TF footprint analysis. In scATAC-seq sequencing experiments, when a TF is bound to the DNA, it protects that region from sequencing, while the DNA bases immediately adjacent to TF binding are accessible, leaving, in this way, a sort of footprint in the signal. This footprint appears as a low signal from the center of the TFBM and a stronger signal from its immediate flanking regions. Ideally, the footprint would be detectable for each cell and location, but it would require a much higher sequencing depth than the technology provides. Therefore, the signal from all the motif occurrences is aggregated. Specifically, this work calculates for each motif in each cell the aggregated sequencing signal at the motif and its surrounding ± 250 bp, as shown in Algorithm 1. The result is a list of $|M|$ vectors $\mathbf{FP}_m = [fp_1, ..., fp_n]$ with $n = 500+$ length of the ith motif, where each element represents the average normalized bias-corrected insertion signal for all the base pairs surrounding all the occurrences of the accessible motif inside a cell. This calculation is lengthy and computationally heavy for all the motifs. From it, it is possible to define the footprint score of the motif in the cells as the average signal from the motif flanking regions defined as ± 50 bp from the motif.

Again, this work differentiates the signal from the promoter and enhancer regions. This calculation is performed on the two matrices separately, obtaining $\mathbf{TFP}^{prom}_{|M|\times|C|}$ and $\mathbf{TFP}^{enhd}_{|M|\times|C|}$ where the matrices elements are the footprint scores of motif $m \in M$ in cell $c \in C$. Also, in this case, this work aggregates cells based on their cell-type labels, resulting in the final matrices $\mathbf{TFP}^{prom}_{|M|\times|CT|}$, $\mathbf{TFP}^{enhd}_{|M|\times|CT|}$ where CT is the list of cell-types.

Algorithm 1 Footprint score computation

```
for m in M do                                          ▷ m → is a single motif/TF
    Compute the footprint signal vector FP_m = [fp_1, ..., fp_n]
end for
for c in C do                                          ▷ c → is a single cell
    for m in M do
        Compute the footprint score TFP_{m×c}
    end for
end for
for ct in CT do                                        ▷ ct → is a single cell type
    for m in M do
        Compute TFP_{m,ct} the average footprint score of m ∀c ∈ ct
    end for
end for
```

3.5. Correlation with Expression

As previously highlighted, this study aims not only to explore TFs through their motif accessibility but, more significantly, to comprehend their correlation with expression levels. Before delving into this analysis, a brief overview of the gene expression matrix formalism is necessary. The considered multi-omic dataset provides a gene expression matrix $\mathbf{E}_{|G|\times|C|}$, encompassing 29,372 genes detected in the experiment. The matrix is then narrowed to the set M of genes associated with TFss, resulting in a sub-matrix $\mathbf{E}_{|M|\times|C|}$. Subsequently, as discussed in Section 3.2, the expression values are averaged across cell types, yielding the final $\mathbf{E}_{|G|\times|CT|}$ matrix (Figure 1e). This matrix serves as the foundation for subsequent comparisons. The correlation study follows the approach employed in [13], where the Pearson correlation between the expression and motif information is calculated for each gene in the aggregated cells. These correlations are then presented through scatter plots to illustrate the general correlation within each cell type visually. The investigation covers motif enrichment versus expression and TF footprint scores versus expression, consistently differentiating between promoter and enhancer contributions (Figure 1f).

This methodology enables exploring the intricate relationship between inferable information related to TFs from accessible regions and their expression patterns.

4. Results

4.1. Enhancer Regions Shows More Variability in Motif Information

When examining the motif enrichment matrices $\mathbf{ME}^{prom}_{|M|\times|C|}$ and $\mathbf{ME}^{enhd}_{|M|\times|C|}$, it is noteworthy that the enrichment in enhancer regions exhibits more pronounced variability compared to the enrichment in promoter regions. This contrast is clearly illustrated in Figure 2, where the variability in TFBM enrichment is markedly higher in enhancer regions than in promoters. In studying cellular heterogeneity, this observation underscores the substantial contribution of enhancer regions in conveying relevant differences in motif accessibility. Analyzing data with higher variability is crucial for detecting distinctions between cell populations and emphasizing the significance of enhancer regions in the epigenetic context. However, the information derived from promoters should not be overlooked. Notably, from Figure 2, it is interesting to observe that the most variable TFBMs in promoter regions belong to the FOS and JUN TF families. These families are known to aggregate and form the complex Activator Protein-1 (AP-1), which binds to promoters, regulating the nuclear gene expression in T-cells. Further discussion on this TFs is provided in the subsequent Section 4.2.

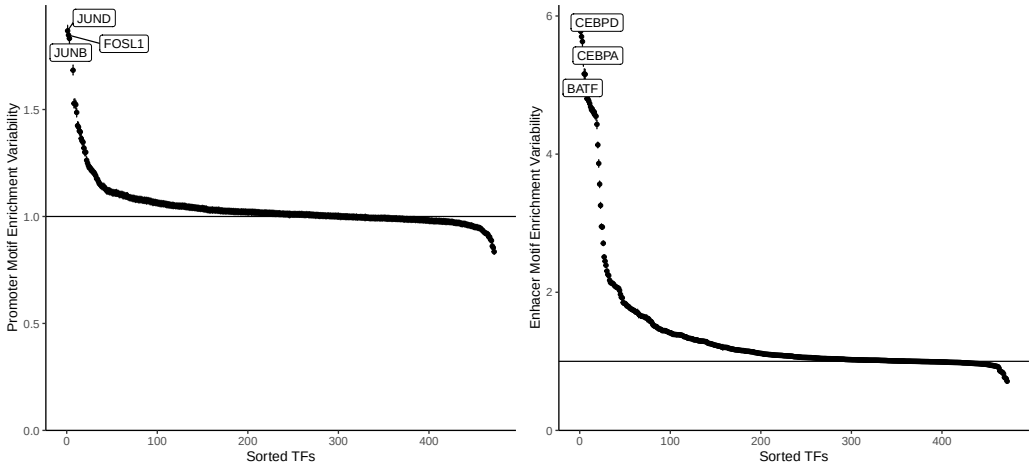

Figure 2. The TFs variability for promoter motif enrichment (**left**) and enhancer motif enrichment (**right**). For the promoter motif enrichment, the top TFs belong to FOS and JUN TF families known to bind to promoter regions.

These distinctions are illustrated in Figure 3, where each black dot in the violin plots represents the motif enrichment of a TF for each cell type. Notably, enhancer regions only exhibit a few TFBMs with the noteworthy enrichment, underscoring their significance in each cell type. It is well established that only a subset of TFs influences gene regulation in a given cell type, especially when regulating cell-type-specific gene pathways [25]. Thus, identifying a limited number of highly enriched TFBMs per cell type aligns with expectations. Consequently, enhancer regions demonstrate a higher sensitivity to cell type-specific motif accessibility than promoter regions. This observation is consistent with the findings in [13], emphasizing the substantial contribution of enhancers to the epigenetic signal in scATAC-seq and their more significant variability in correlation with expression. Once again, these results underscore the importance of modeling promoter and enhancer contributions differently, highlighting the limitations of approaches that solely focus on promoters and overlook valuable information.

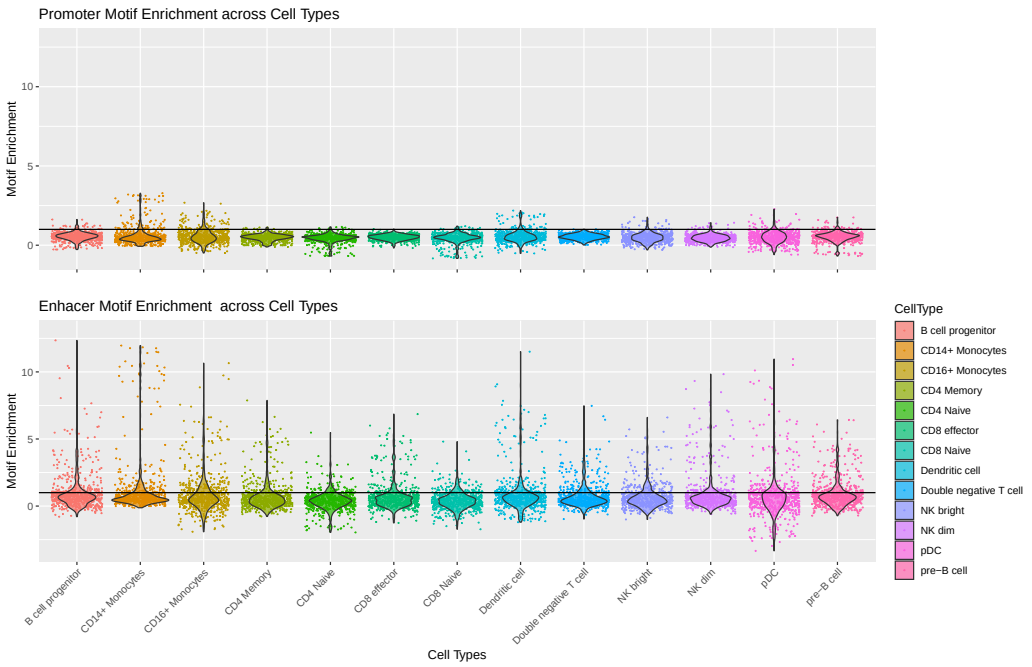

Figure 3. TFs motif enrichment for each cell type. Each black dot represents a TF.

4.2. General Correlation with Expression Is Low

Having explored the distinctions between enhancer and promoter regions, this section delves into their correlation with the actual expression of TFs. In contrast to the findings in [13], interpreting the correlation between TFs expression and their accessibility information is less straightforward. The general Pearson correlation scores are relatively low. Specifically, for the expression-TFBM enrichment correlation in both promoter and enhancer regions, only around 15% of TFs exhibit a higher correlation than 0.5 (with significant p-values < 0.05). Conversely, examining the expression-TF footprint correlation, 19% of TFs show a correlation higher than 0.5 (with significant p-values < 0.05) in enhancer regions, which increases to 42% in promoter regions. This overall low correlation is expected, considering that, as discussed earlier, only a subset of TFs will be relevant in a particular cell type, resulting in coherent and correlated motif information for only some of them.

This observation becomes more evident when referring to Figure 4. Each point on these plots represents a single TF for each cell type. Notably, many TFs are clustered in the bottom left corner of the motif enrichment plots, indicating low expression and motif enrichment. The interesting aspect lies in the TFs characterizing each cell type, identifiable in the top-right section of the plots. Here, TFs that are relevant from both expression and motif perspectives are found. Notably, these TFs tend to be cell type-specific, emphasizing that, for each cell type, a distinct subset of TFs is captured in this quadrant, aligning with the previous comments on the relevance of specific TFs for different cell types.

Furthermore, these cell-type-specific TFs show consistency between the two motif levels. Examples include BACH1, CEBPB, and CEBPD, which are characteristic for CD14+ Monocytes in both motif enrichment and footprint score. This consistent dual information is crucial for identifying and studying cell-type-specific TF mechanisms and regulation, which is not apparent when solely considering expression levels. CTCF, for instance, appears to have a strong signal in all cell types despite its low expression. This characteristic is only discernible when examining enhancers, highlighting a specific correlation between the CTCF and enhancer regions. This aligns with its known role in DNA bending, facilitating interaction between promoters and enhancer regions [26].

Lastly, beyond these general observations, investigating the differences between differentiated cell types, such as naive and memory CD4+ T-cells, could unveil specific TF motif patterns involved in the differentiation or proliferation processes. This aspect is explored in the following section, examining various types of correlation for selected TFs.

4.3. Motif Information Highlights Differences in AP-1 Subunits

The AP-1 TF is a dimeric transcription factor composed of two subunits, typically belonging to the Fos and Jun TF families. It is a key regulator in processes such as cell proliferation, differentiation, and apoptosis [27,28]. Numerous studies highlight its impact on T-cell activation, the initial step driving the differentiation and proliferation of Naive T-cells into various specialized T-cell subsets [29,30]. As evident from the scatter plots (Figure 4), both FOS and JUNB consistently occupy the top region for all cell types. This trend is further elucidated in Figure 5 and Supplementary Figure S1, which present various details for FOS and JUNB, respectively. The violin plots depict the expression (B, top) and motif enrichment (B, bottom) of the TF in each cell type. While the gene expression appears relatively consistent across cell types, there is a distinct separation in motif enrichment.

Crucially, CD4 Memory and CD8 Effector T-cells exhibit significantly higher enrichment than their respective naive T-cells, aligning with the discussed activation mechanism. This pattern is reinforced by the TF footprints, as illustrated in Figure 5C, representing the expression-TF footprint of FOS for each cell type. Once again, naive T-cells exhibit noticeably lower footprint scores than their counterparts. This distinction becomes more apparent in the plots in Figure 5D, showing the average signal around the motifs. For clarity, only CD4 naive and memory cells are depicted, demonstrating the stronger footprint signal for memory T-cells. Importantly, this variation is only evident for enhancers and not promoters, underscoring the notable differences between the signals from functional genomic regions.

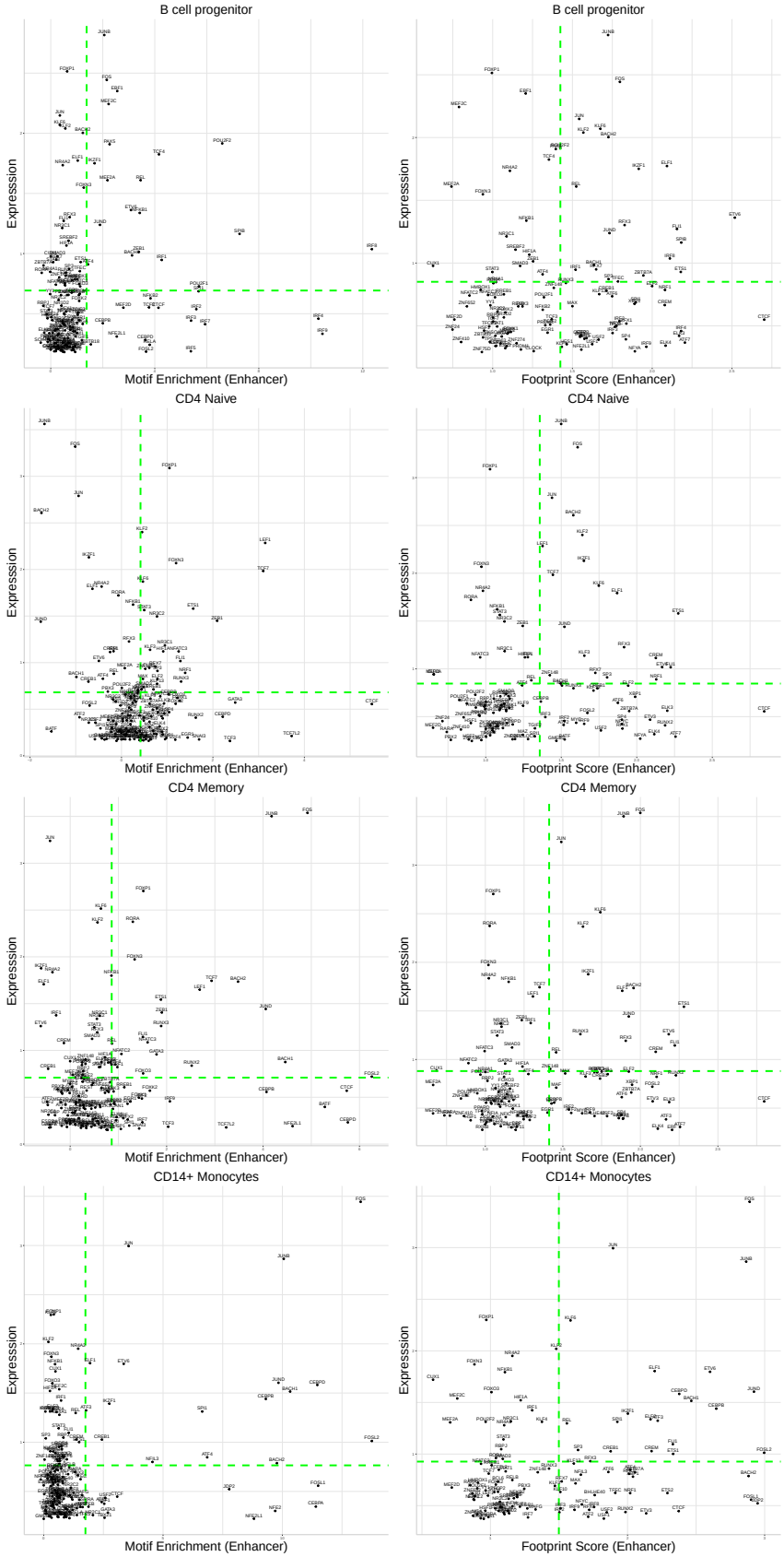

Figure 4. Scatter plots for the expression-motif enrichment and expression-TF footprint correlation.

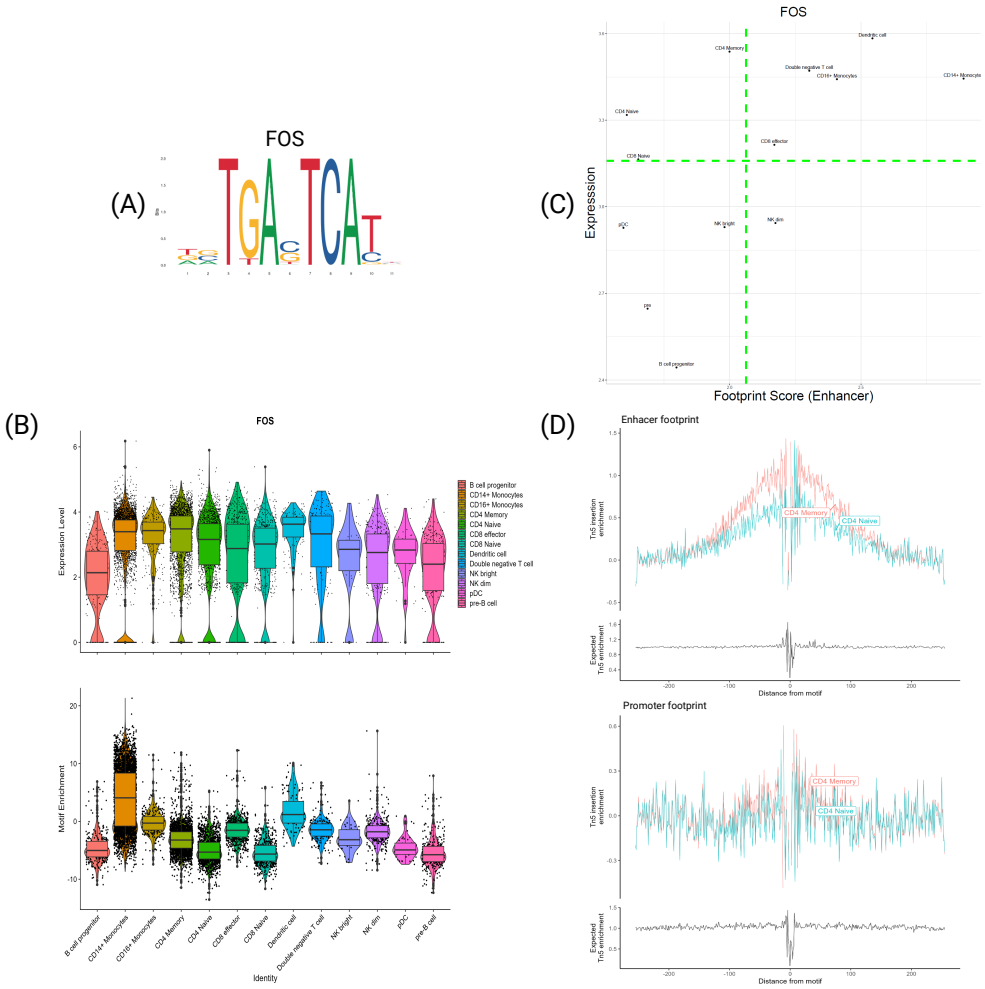

Figure 5. Transcription actor FOS. (**A**) PFM visualization for FOS's motif MA0476.1. (**B**) Violin plot of expression (**top**) and motif enrichment (**bottom**), for all cell types. (**C**) Scatter plot of footprint score-expression of FOS for all cell types. (**D**) Tn5 insertion plots for Memory and Naive CD4 T−cells.

A similar pattern is observed for the gene BATF. As depicted in Supplementary Figure S2, the expression is uniformly low across all cell types. However, both motif enrichment and footprint scores exhibit variations between naive and memory cells, which are especially noticeable between CD8 naive and CD8 effector T-cells. BATF is recognized for its involvement in the functional development of CD8 T-cells [31,32], once again emphasizing how the dynamic of the TFs expression alone may not suffice to discern specific changes between cell types.

In conclusion, the FOS and JUNB TF, subunits of AP-1, exhibit a significant correlation with certain cell types in their motif information despite minimal variability at the transcriptional level. This result, coupled with observations on BATF, underscores that TFs expression alone may not be adequate to identify crucial differences in specific cellular processes; it must be complemented by an examination at the epigenetic level for a more comprehensive understanding.

4.4. Specific TFs Shows Differences Only at the Expression Levels

FOS and JUNB TFs are not the only ones with differences in expression and motif information behavior. Indeed, the TF BACH2 has a similar interesting behavior. BACH2 is a known regulator in the B-cells development [33], by orchestrating the early specification and commitment of B-cell progenitors [34]. However, Figure 6 shows a peculiar behavior in its motif information. Looking at Figure 6B, it is evident that there is a gene expression differentiation between cell types, particularly between B-cell progenitors and pre-B-cells,

that becomes overexpressed in the latter. However, the motif enrichment in these two cell types is pretty much identical, implying that only the TF expression drives this differentiation, an opposite behavior to FOS and JUNB. Furthermore, the footprint plots in Figure 6D show no difference between the two cell types.

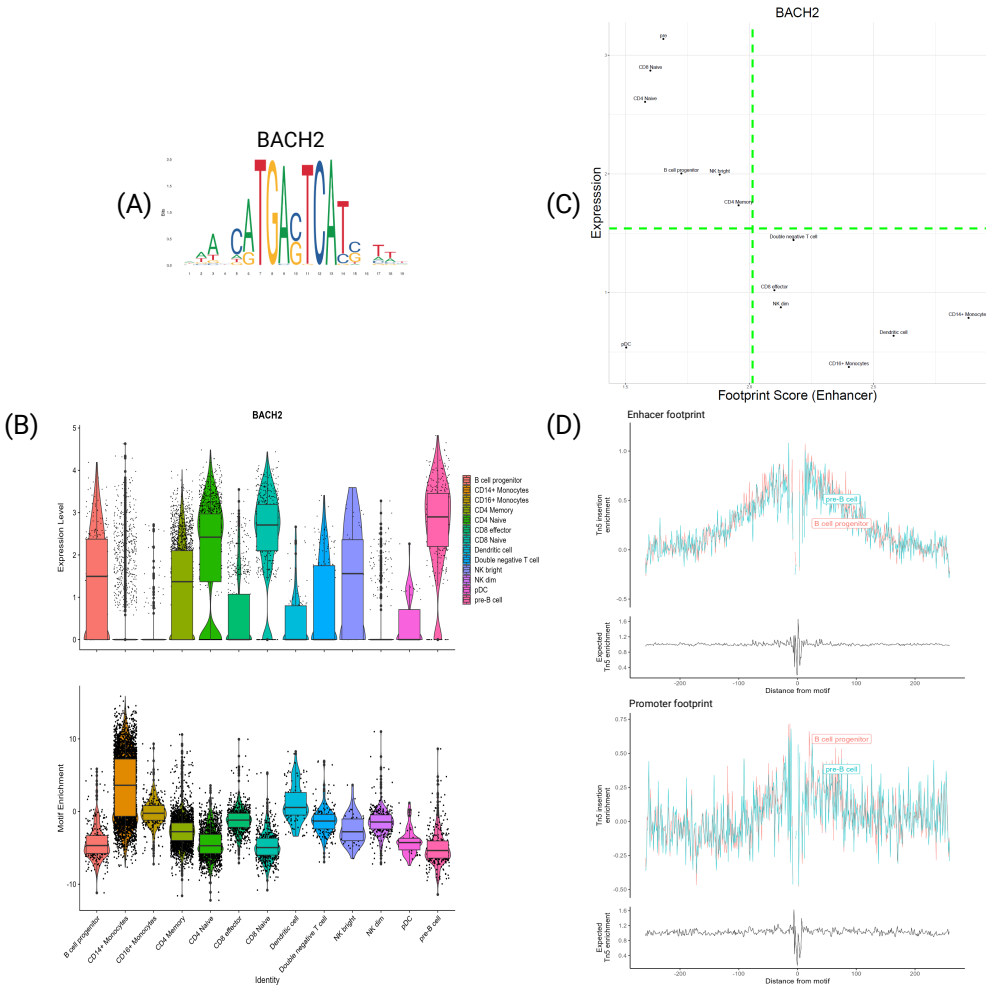

Figure 6. Transcription factor BACH2. (**A**) PFM visualization for BACH2's motif MA1101.2. (**B**) Violin plot of the expression (**top**) and motif enrichment (**bottom**) for all cell types. (**C**) Scatter plot of footprint score−expression of BACH2 for all cell types. (**D**) Tn5 insertion plots for B−cell progenitors and pre−B−cells.

Furthermore, the motif enrichment in the other cell types seems to have an inverse trend to the expression. Indeed, cell types with the highest expression (such as naive T-cells and pre-B-cells) display the lowest enrichment. This inverse correlation is even more evident from Figure 6C, showing how cell types with a lower expression have higher footprint scores and vice versa. This peculiar behavior seems counterintuitive since one would expect not to see a higher expression if the motif is that much accessible. However, from the literature, BACH2 is highly characterized as a repressor TF [35], which regulates B-cells by suppressing specific genes related to the myeloid program. Hence, the observed inverse correlation for this gene may indicate a distinctive repressive dynamic. In cells where BACH2 is active, the motif of this gene becomes generally less accessible but more specific, fine-tuning its repression mechanism and subsequently leading to an increase in its expression. BACH2 exhibits a unique connection between its expression and motif information, serving as an intriguing indicator to discern repressive dynamics from the more common enhancing dynamics. This observation is of the utmost importance, as an effective transcriptional regulation model should accurately represent these two markedly

different dynamics. Currently, there remains a need for consensus or comprehensive information to distinguish the intricacies of silencing processes.

4.5. TFs Characterize Cell Types at Both Expression and Motif Information Levels

In this concluding section, it is pertinent to showcase genes that exhibit coherent and cell-type-specific correlations between their expression and motif information. For instance, the transcription factor CEBPD, depicted in Figure 7, is recognized for its role in the inflammatory response of monocytes [36–38]. This functional role is reflected in its expression and motif information, as illustrated in Figure 7B. CEBPD is selectively expressed in monocytes and dendritic cells while significantly over-enriched in these cell types. This observation is reinforced by the footprint score, where the scatter plot indicates that these subtypes are situated in the top-right corner, signifying that CEBPD plays a regulatory role in these cells at both the transcriptomic and epigenomic levels.

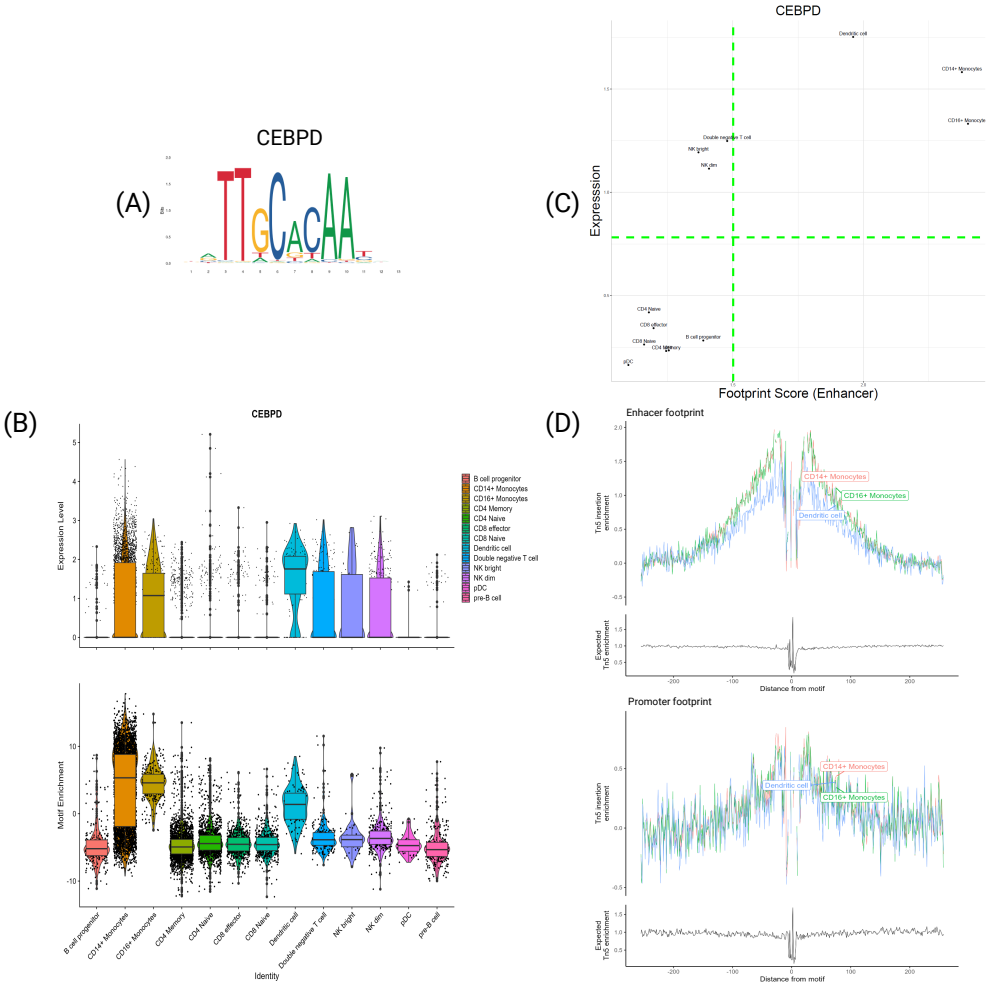

Figure 7. Transcription factor CEBPD. (**A**) PFM visualization for CEBPD's motif MA0836.2. (**B**) Violin plot of expression (**top**) and motif enrichment (**bottom**), for all cell types. (**C**) Scatter plot of footprint score−expression of CEBPD for all cell types. (**D**) Tn5 insertion plots for CD14+ and CD16+ monocytes and dendritic cells.

Similarly, Figure 8 presents the findings for the TF POU2F2. This TF holds considerable significance in B-cells, particularly in motif enrichment, as it exhibits the highest motif enrichment signal among those investigated. This observation aligns with the footprint in Figure 8D, showcasing a prominent signal at the flanking regions for both subtypes. Once again, the scatter plot illustrates that B-cells consistently reside in the top-right corner,

indicating high expression and footprint scores, while other subtypes occupy the opposite bottom-left corner.

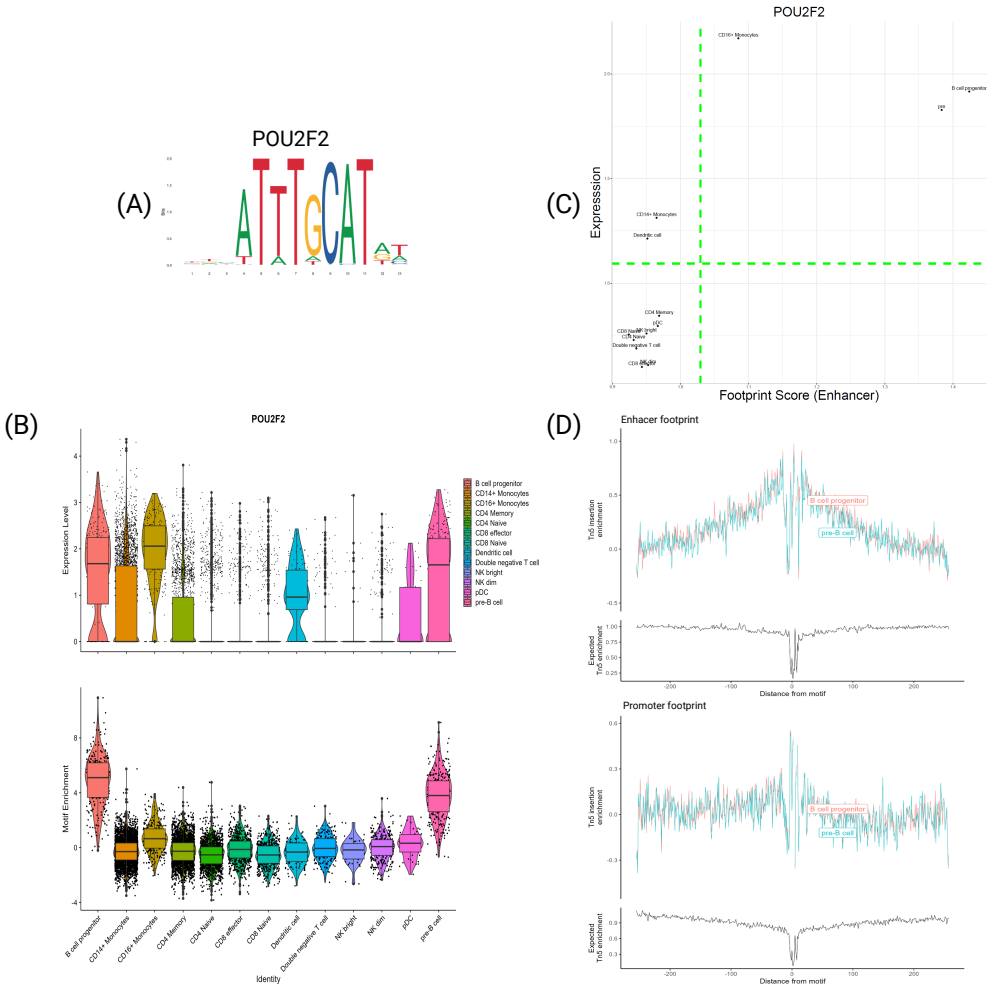

Figure 8. Transcription factor POU2F2. (**A**) PFM visualization for POU2F2's motif MA0507.1. (**B**) Violin plot of expression (**top**) and motif enrichment (**bottom**), for all cell types. (**C**) Scatter plot of footprint score−expression of POU2F2 for all cell types. (**D**) Tn5 insertion plots for B−cell progenitors and pre−B−cells.

Similar observations apply to the TFs TCF7 and LEF1 (see Supplementary Figures S3 and S4). Both are characteristic of T-cells and demonstrate coherence between expression and motif information. Notably, there is a discernible difference in CD8 effector cells, which exhibit a weaker correlation than other T-cell subtypes. This discrepancy may highlight a specific dynamic of these TFs in those subtypes, which could be crucial in distinct biological processes.

In conclusion, what ties these genes together is their specific relevance in distinct subtypes at both the expression and motif information levels. Therefore, modeling their impact on transcriptional regulation is essential, as they likely serve as specialized regulators characterizing various biological processes.

5. Conclusions

This work presents a comprehensive analysis of the correlation between the motif information obtainable from scATAC-seq data and the expression of the TFs themselves. This analysis is crucial for understanding transcriptional regulation, which the TFs are a crucial part of. The increasing power of multi-omic sequencing technologies assists in this, simultaneously allowing the investigation of the expression and DNA accessibility of relevant regions related to transcriptional regulation. Specifically, this work investigates

the motif presence in accessible enhancer and promoter regions distinctively. Two types of information are considered: the motif enrichment, representing how much a motif is over-represented in determined regions, and the TF footprint scores, representing the signal of a TF binding event. This analysis brought some interesting results. First, there is a remarkable difference between the signal from enhancers and promoters, with the first showing a more significant variability between cell types and highlighting different types of TFs. These differences show the importance of a distinct analysis of the two types of functional regions, which need to be studied separately to properly understand the intricacies of transcriptional regulation.

However, the correlation between the motif information and the expression is low, whatever contribution one considers. This result is partially expected since only a small subset of the TFs is cell-type specific and, consequentially, is coherent between the omic levels. However, the reported results highlight the differential behaviors of specific TFs between certain cell types. The exciting part is that the results highlighted different correlation patterns in the motif information, indicating the necessity of modeling their impact on transcription in specific manners.

Like all approaches, this work also has disadvantages. The primary constraint of this method arises from the inherent high sparsity of scATAC-seq data. The limited signal for each cell renders it impractical to explore motif information at a single-cell resolution. Instead, one must rely on the aggregated behavior of a cell type or a group of cells. Furthermore, this study delves into the self-dynamics of transcription factors (TFs) without yet investigating their impact on the genes they regulate.

This work focused on the TFs by themselves, looking at the different information inferred by the multi-omic data. However, this is only the first step in modeling the transcriptional regulation. Future work will aim to understand the correlation between the TFs and their putative target genes, trying to understand how the motif information from scATAC-seq data can influence gene expression. Moreover, it will be relevant to practically model this correlation in an extension of the gene activity concept, specifically GAGAM, which will not only investigate the general DNA accessibility but will consider a higher level of information to model the transcriptional regulation correctly. Recognizing the intricate connection between the expression of TFs and the accessibility of their binding regions will be essential in defining a transcriptional regulation model that can accurately capture their impact on gene expression. Additionally, the diverse behaviors we have highlighted may serve as key indicators of undiscovered dynamics within the TFs themselves. This will be valuable for both understanding cell-type-specific processes and in cellular heterogeneity studies and capture fundamental mechanisms in specific pathologies.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/genes15030268/s1, Figure S1: Transcription Factor JUNB; Figure S2: Transcription Factor BATF; Figure S3: Transcription Factor TCF7; Figure S4: Transcription Factor LEF1.

Author Contributions: Conceptualization, L.M.; methodology, L.M.; software, L.M.; validation, L.M.; formal analysis, L.M. and R.B.; investigation, L.M. and R.B.; resources, L.M.; data curation, L.M.; 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.C. 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 at https://github.com/smilies-polito/GAGAM, accessed on 29 December 2022. All the code employed for this work is available at https://github.com/smilies-polito/MAGA

(accessed on 16 December 2023), including all the Supplementary Materials and figures, accessible in Zenodo with the DOI https://doi.org/10.5281/zenodo.10517230.

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

References

1. Chen, G.; Ning, B.; Shi, T. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Front. Genet.* **2019**, *10*, 317. [CrossRef]
2. 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]
3. Martini, L.; Amprimo, G.; Di Carlo, S.; Olmo, G.; Ferraris, C.; Savino, A.; Bardini, R. Neuronal Spike Shapes (NSS): A straightforward approach to investigate heterogeneity in neuronal excitability states. *Comput. Biol. Med.* **2024**, *168*, 107783. [CrossRef] [PubMed]
4. Buenrostro, J.D.; Corces, M.R.; Lareau, C.A.; Wu, B.; Schep, A.N.; Aryee, M.J.; Majeti, R.; Chang, H.Y.; Greenleaf, W.J. Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **2018**, *173*, 1535–1548.e16. [CrossRef]
5. Baek, S.; Lee, I. Single-cell ATAC sequencing analysis: From data preprocessing to hypothesis generation. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 1429–1439. [CrossRef]
6. Chen S.; Lake, B.B.; Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* **2019**, *37*, 1452–1457. [CrossRef] [PubMed]
7. Hao, Y.; Hao, S.; Andersen-Nissen, E.; Mauck, W.M.; Zheng, S.; Butler, A.; Lee, M.J.; Wilk, A.J.; Darby, C.; Zager, M.; et al. Integrated analysis of multimodal single-cell data. *Cell* **2021**, *184*, 3573–3587. [CrossRef] [PubMed]
8. 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]
9. 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] [PubMed]
10. 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. [CrossRef]
11. 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: Cham, Switzerland, 2022; pp. 18–32.
12. Martini, L.; Savino, A.; Bardini, R.; Carlo, S.D. GRAIGH: Gene Regulation accessibility integrating GeneHancer database. *bioRxiv* **2023**. [CrossRef]
13. Martini, L.; Bardini, R.; Savino, A.; Di Carlo, S. Meta-analysis of Gene Activity (MAGA) Contributions and Correlation with Gene Expression, Through GAGAM. In *Proceedings of the Bioinformatics and Biomedical Engineering*; Springer Nature: Cham, Switzerland, 2023; pp. 193–207.
14. 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] [PubMed]
15. Kelsey, G.; Stegle, O.; Reik, W. Single-cell epigenomics: Recording the past and predicting the future. *Science* **2017**, *358*, 69–75. [CrossRef]
16. Danese A.; Richter M.L.; Chaichoompu, K.; Fischer, D.S.; Theis, F.J.; Colomé-Tatché, M. EpiScanpy: Integrated single-cell epigenomic analysis. *Nat. Commun.* **2021**, *12*, 5228. [CrossRef]
17. Lareau C.A.; Duarte F.M.; Chew, J.G.; Kartha, V.K.; Burkett, Z.D.; Kohlway, A.S.; Pokholok, D.; Aryee, M.J.; Steemers, F.J.; Lebofsky, R.; et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.* **2019**, *37*, 916–924. [CrossRef]
18. 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]
19. Chen, H.; Lareau, C.; Andreani, T.; Vinyard, M.E.; Garcia, S.P.; Clement, K.; Andrade-Navarro, M.A.; Buenrostro, J.D.; Pinello, L. Assessment of Computational Methods for the Analysis of Single-Cell ATAC-Seq Data. *Genome Biol.* **2019**, *20*, 241. [CrossRef]
20. Kent, J.; Sugnet, C.W.; Furey, T.S.; Roskin, K.M.; Pringle, T.H.; Zahler, A.M.; Haussler, D. The Human Genome Browser at UCSC. *Genome Res.* **2002**, *12*, 996–1006. [CrossRef]
21. 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. Available online: https://www.10xgenomics.com/datasets/10-k-human-pbm-cs-multiome-v-1-0-chromium-controller-1-standard-2-0-0 (accessed on 9 August 2021).
22. Hwang B.; Lee, J.; Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **2018**, *50*, 1–14. [CrossRef] [PubMed]

23. Rauluseviciute, I.; Riudavets-Puig, R.; Blanc-Mathieu, R.; Castro-Mondragon, J.A.; Ferenc, K.; Kumar, V.; Lemma, R.B.; Lucas, J.; Chèneby, J.; Baranasic, D.; et al. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **2023**, *52*, D174–D182. [CrossRef] [PubMed]
24. Schep, A.N.; Wu, B.; Buenrostro, J.D.; Greenleaf, W.J. chromVAR: Inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **2017**, *14*, 975–978. [CrossRef]
25. Lee, B.K.; Bhinge, A.A.; Battenhouse, A.; McDaniell, R.M.; Liu, Z.; Song, L.; Ni, Y.; Birney, E.; Lieb, J.D.; Furey, T.S.; et al. Cell-type specific and combinatorial usage of diverse transcription factors revealed by genome-wide binding studies in multiple human cells. *Genome Res.* **2012**, *22*, 9–24. [CrossRef]
26. Holwerda, S.J.B.; de Laat, W. CTCF: The protein, the binding partners, the binding sites and their chromatin loops. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **2013**, *368*, 20120369. [CrossRef]
27. Eferl, R.; Wagner, E.F. AP-1: A double-edged sword in tumorigenesis. *Nat. Rev. Cancer* **2003**, *3*, 859–868. [CrossRef]
28. Hess, J.; Angel, P.; Schorpp-Kistner, M. AP-1 subunits: Quarrel and harmony among siblings. *J. Cell Sci.* **2004**, *117*, 5965–5973. [CrossRef]
29. Yukawa, M.; Jagannathan, S.; Vallabh, S.; Kartashov, A.V.; Chen, X.; Weirauch, M.T.; Barski, A. AP-1 activity induced by co-stimulation is required for chromatin opening during T cell activation. *J. Exp. Med.* **2020**, *217*, jem.20182009. [CrossRef]
30. Atsaves, V.; Leventaki, V.; Rassidakis, G.Z.; Claret, F.X. AP-1 transcription factors as regulators of immune responses in cancer. *Cancers* **2019**, *11*, 1037. [CrossRef]
31. Tsao, H.W.; Kaminski, J.; Kurachi, M.; Barnitz, R.A.; DiIorio, M.A.; LaFleur, M.W.; Ise, W.; Kurosaki, T.; Wherry, E.J.; Haining, W.N.; et al. Batf-mediated epigenetic control of effector CD8+ T cell differentiation. *Sci. Immunol.* **2022**, *7*, eabi4919. [CrossRef]
32. Kurachi, M.; Barnitz, R.A.; Yosef, N.; Odorizzi, P.M.; DiIorio, M.A.; Lemieux, M.E.; Yates, K.; Godec, J.; Klatt, M.G.; Regev, A.; et al. The transcription factor BATF operates as an essential differentiation checkpoint in early effector CD8+ T cells. *Nat. Immunol.* **2014**, *15*, 373–383. [CrossRef]
33. Ochiai, K.; Igarashi, K. Exploring novel functions of BACH2 in the acquisition of antigen-specific antibodies. *Int. Immunol.* **2023**, *35*, 257–265. [CrossRef]
34. Kaiser, F.M.P.; Janowska, I.; Menafra, R.; de Gier, M.; Korzhenevich, J.; Pico-Knijnenburg, I.; Khatri, I.; Schulz, A.; Kuijpers, T.W.; Lankester, A.C.; et al. IL-7 receptor signaling drives human B-cell progenitor differentiation and expansion. *Blood* **2023**, *142*, 1113–1130. [CrossRef] [PubMed]
35. Itoh-Nakadai, A.; Hikota, R.; Muto, A.; Kometani, K.; Watanabe-Matsui, M.; Sato, Y.; Kobayashi, M.; Nakamura, A.; Miura, Y.; Yano, Y.; et al. The transcription repressors Bach2 and Bach1 promote B cell development by repressing the myeloid program. *Nat. Immunol.* **2014**, *15*, 1171–1180. [CrossRef]
36. Spek, C.A.; Aberson, H.L.; Butler, J.M.; de Vos, A.F.; Duitman, J. CEBPD potentiates the macrophage inflammatory response but CEBPD knock-out macrophages fail to identify CEBPD-dependent pro-inflammatory transcriptional programs. *Cells* **2021**, *10*, 2233. [CrossRef]
37. Ko, C.Y.; Chang, W.C.; Wang, J.M. Biological roles of CCAAT/Enhancer-binding protein delta during inflammation. *J. Biomed. Sci.* **2015**, *22*, 6. [CrossRef]
38. Liu, J.; Gao, H.; Li, C.; Zhu, F.; Wang, M.; Xu, Y.; Wu, B. Expression and regulatory characteristics of peripheral blood immune cells in primary Sjögren's syndrome patients using single-cell transcriptomic. *iScience* **2022**, *25*, 105509. [CrossRef]