HALO: Hierarchical Causal Modeling for Single Cell Multi-Omics Data

Authors: Haiyi Mao†[1,2], Minxue Jia†[1,2], Marissa Di[1,2], Eleanor Valenzi[3], Xiaoyu Tracy Cai[4], Harinder Singh[5], Robert Lafyatis[6], Kun Zhang[7], *Panagiotis Benos[1,2,8]

Affiliations:

[1]Department of Computational and Systems Biology, University of Pittsburgh.
[2]Joint Carnegie Mellon University-University of Pittsburgh Ph.D. Program in Computational Biology.
[3]Division of Pulmonary, Allergy, Critical Care and Sleep Medicine, University of Pittsburgh.
[4]Immunology Discovery and Regenerative Medicine, Genentech, Inc.
[5]Department of Immunology, University of Pittsburgh.
[6]Division of Rheumatology and Clinical Immunology, University of Pittsburgh.
[7]Department of Philosophy, Carnegie Mellon University.
[8]Department of Epidemiology, University of Florida.
†Equal Contributions
*Corresponding author. Email: pbenos@ufl.edu

Abstract: Though open chromatin may promote active transcription, gene expression responses may not be directly coordinated with changes in chromatin accessibility. Most existing methods for single-cell multi-omics data focus only on learning stationary and shared information among these modalities, overlooking modality-specific information delineating cellular states and dynamics resulting from causal relations among modalities. To account for this, the epigenome and transcriptome relationship can be characterized in relation to time as "coupled" (changing dependently) or "decoupled" (changing independently). We propose the framework HALO, which adopts a causal approach to model these temporal causal relations on two levels. On the representation level, HALO factorizes these two modalities into both coupled and decoupled latent representations, identifying the dynamic interplay between chromatin accessibility and transcription through temporal modulations in the latent space. On the individual gene level, HALO matches gene-peak pairs and characterizes changing dynamics between gene expression and local peaks with time. We demonstrate that using coupled and decoupled information, HALO discovers analogous biological functions between modalities, distinguishes epigenetic factors for lineage specification, and identifies temporal *cis*-regulation interactions relevant to cellular differentiation.

Introduction

Single-cell multi-omics technologies have revolutionized our understanding of cellular heterogeneity and complexity, enabling the simultaneous measurement of diverse molecular layers such as transcriptomics, epigenomics, and proteomics within individual cells. These technologies provide a unique opportunity to dissect the regulatory relationships underlying cellular functions and phenotypes. However, the integration and analysis of such heterogeneous data types remain challenging, for example distinguishing the effects of chromatin accessibility on expression across varying cellular states and over time. To address these challenges, we developed HALO (**H**ierarchical C**A**usal Modeling for Single Cell Mu**L**ti-**O**mics Data), a computational framework designed to model the causal relationships within single-cell co-profiled multi-omic data. The key hypothesis of HALO is that changes in chromatin accessibility causally influence gene expression. By leveraging measurements of transcriptomics and chromatin accessibility data from single-cell RNA sequencing (scRNA-seq) and single-cell ATAC sequencing (scATAC-seq), HALO enables a comprehensive analysis of the causal interactions that dictate cell state, function, and regulatory mechanisms.

Despite the established influence of chromatin accessibility on gene expression, the nuances of their respective dynamics require further characterization. Changes in chromatin accessibility and gene expression can exhibit (1) time lag phenomena, where chromatin opening precedes transcription initiation, or (2) chromatin closing and transcription repression occur independently[1]. Previous representation learning methods for single-cell multi-omics data implicitly rely on the strict biological assumption that chromatin accessibility and active transcription are synchronized[2], though some recent methods consider case (1)[1,3]. In contrast to these works, HALO decomposes the relations between chromatin accessibility and gene expression into coupled and decoupled cases. In the coupled case, chromatin accessibility and gene expression exhibit correlated changes over time. Conversely, in the decoupled case, they change independently over time. Due to the high dimensionality and sparsity of single cell genomics data, HALO operates on two levels of analysis: the low-dimensional latent representation level and the individual gene level, incorporating estimations of latent time to account for temporal dynamics in cellular development. The framework is equipped with an interpretable neural network to provide biological meaning to the latent representations. Moreover, HALO employs Granger causality to assess context-specific distal cis-regulation in cases where, despite associated chromatin regions becoming more accessible, gene transcription does not increase correspondingly[3]. We observe these situations frequently occur when the chromatin regions overlap with super enhancer regions.

Results

HALO: a causal machine learning framework to model the interactions between chromatin accessibility and gene expression. HALO models the co-assayed gene expression and chromatin accessibility in a low-dimensional latent space as well as individual gene and its linked peaks (open chromatin regions) level through a causal lens. In our study, we establish a framework to analyze the causal relationships between jointly profiled single-cell RNA sequencing (scRNA-seq) and single-cell ATAC sequencing (scATAC-seq) data, incorporating temporal information. We hypothesize that scATAC-seq data, which is indicative of chromatin accessibility, causally precedes scRNA-seq data due to open chromatin regions influencing gene transcription. Specifically, we distinguish between two types of causal interactions: 1) In the **coupled** case, gene expression and chromatin accessibility exhibit *dependent* changes over time, indicating they are influenced by shared unknown confounders. 2) In the **decoupled** case, certain gene expressions and their local peak patterns change *independently* over time, suggesting the presence of distinct, unknown causal factors affecting gene expressions. This approach extends the concept of time-lagging observed in previous studies[1,4] to a broader framework, encompassing mechanisms that change independently or dependently over time for example, chromatin open regions may become more accessible while gene expression level remains stable. This approach aims to elucidate causal relationships between gene expression and chromatin accessibility at both the levels of representations and individual genes, as depicted in **Figure 1A**.

On the representation level, we model the lower dimensional latent space interactions between ATAC and RNA modalities from a causal perspective. Specifically, HALO learns the representations that contain the "coupled" information between scATAC-seq $\mathbf{X}^{\mathrm{A}} \in \mathbb{R}^{n_{\mathrm{cells}} \times n_{\mathrm{peaks}}}$ and scRNA-seq $\mathbf{X}^{\mathrm{R}} \in \mathbb{R}^{n_{\mathrm{cells}} \times n_{\mathrm{genes}}}$ data as well as the independent "decoupled" information, such that the causal relations are preserved even across time (**Figure 1B**). Coupled representations encapsulate information where gene expression changes are dependent on chromatin accessibility over time, reflecting shared information across modalities. In contrast, decoupled representations extract information where gene expression changes independently of chromatin accessibility over time, emphasizing modality-specific information. The latent representations of scATAC-seq data can be decomposed to $\mathbf{Z}^{\mathrm{A}} = [\mathbf{Z}_{\mathrm{c}}^{\mathrm{A}}, \mathbf{Z}_{\mathrm{d}}^{\mathrm{A}}]$. The latent representations of scRNA-seq data $\mathbf{Z}^{\mathrm{R}}$may be similarly decomposed: $\mathbf{Z}^{\mathrm{R}} = [\mathbf{Z}_{\mathrm{c}}^{\mathrm{R}}, \mathbf{Z}_{\mathrm{d}}^{\mathrm{R}}]$. $\mathbf{Z}_{\mathrm{c}}^{\mathrm{R}}$ $[\mathbf{Z}_{\mathrm{c}}^{\mathrm{R}}, \mathbf{Z}_{\mathrm{d}}^{\mathrm{R}}]$. $\mathbf{Z}_{\mathrm{c}}^{\mathrm{R}}$ and $\mathbf{Z}_{\mathrm{d}}^{\mathrm{R}}$ represent the decoupled latent representations derived from scRNA-seq and scATAC-seq data while $\mathbf{Z}_{\mathrm{c}}^{\mathrm{R}}$ and $\mathbf{Z}_{\mathrm{c}}^{\mathrm{A}}$denote the coupled latent representations for scRNA-seq and scATAC-seq data, respectively. These representations capture distinct aspects of information from the two modalities, each varying over time or along a latent temporal variable. **Figure 1A** illustrates the underlying causal relationship, wherein $\mathbf{Z}_{\mathrm{d}}^{\mathrm{A}}$ causes $\mathbf{Z}_{\mathrm{d}}^{\mathrm{R}}$, and $\mathbf{Z}_{\mathrm{c}}^{\mathrm{A}}$ causes $\mathbf{Z}_{\mathrm{c}}^{\mathrm{R}}$. However, $\mathbf{Z}_{\mathrm{d}}^{\mathrm{A}}$ and $\mathbf{Z}_{\mathrm{d}}^{\mathrm{R}}$ are influenced by independent causative factors, which are temporally related. This implies that, despite their causal linkage, $\mathbf{Z}_{\mathrm{d}}^{\mathrm{A}}$ and $\mathbf{Z}_{\mathrm{d}}^{\mathrm{R}}$ exhibit independent temporal changes. $\mathbf{Z}_{\mathrm{c}}^{\mathrm{R}}$ and $\mathbf{Z}_{\mathrm{c}}^{\mathrm{A}}$ denote the coupled latent representations for scRNA-seq and scATAC-seq data, respectively. In contrast to the decoupled representations, both $\mathbf{Z}_{\mathrm{c}}^{\mathrm{A}}$ and $\mathbf{Z}_{\mathrm{c}}^{\mathrm{R}}$ are influenced by common latent confounders, which also vary with time. This indicates that the changes in $\mathbf{Z}_{\mathrm{c}}^{\mathrm{A}}$ and $\mathbf{Z}_{\mathrm{c}}^{\mathrm{R}}$ are synchronized over time, reflecting their shared temporal dynamics.

To ensure the delineation of these relationships, HALO utilizes paired scRNA-seq and scATAC-seq data as its input. We employ two distinct encoders to derive the latent representations $\mathbf{Z}^{\mathrm{A}}$ and $\mathbf{Z}^{\mathrm{R}}$. Moreover, we have formulated specific decoupled and coupled causal constraints (See Methods). A Multi-Layer Perceptron (MLP) is used to model the concept that $\mathbf{Z}_{\mathrm{d}}^{\mathrm{A}}$ causes $\mathbf{Z}_{\mathrm{d}}^{\mathrm{R}}$, while a decoupled constraint enforces the independent functional relations between them. Similarly, another MLP is used to align the coupled representations $\mathbf{Z}_{\mathrm{c}}^{\mathrm{A}}$ to $\mathbf{Z}_{\mathrm{c}}^{\mathrm{R}}$, but with a coupled constraint (See Methods). Additionally, we developed a nonlinear interpretable decoder (**Figure 1B**) that allows us to interpret the latent representations by decomposing the reconstruction of genes or peaks into additive contributions from individual representations (see Methods).

At the individual gene level, our approach begins with the use of denoised gene expressions and peaks (**Figure 1C**), applying negative binomial regression to correlate local peaks with gene expression. This method allows us to match local peaks to corresponding gene expression enabling the subsequent calculation of decouple and couple scores at the individual gene level. Specifically, the decouple score quantifies the independence of gene expression level changes in relation to local peaks over time. Conversely, the couple score evaluates the extent to which gene expression changes are dependent on local peaks throughout the temporal course (See Methods). Finally, through the application of Granger causality analysis, we explore the underlying mechanisms of distal peak-gene regulatory interactions. This analytical approach allows us to elucidate instances where local peaks increase, yet corresponding gene expression remains largely unchanged.

In summary, HALO offers several key contributions to the field of multi-omics analysis, which we enumerate as follows: (1) HALO learns latent representations that are causally informed, enhancing our understanding of the interactions across different omics modalities. (2) HALO enables learning of interpretable latent representations. (3) HALO causally characterizes the relations between gene expression and associated chromatin regions over time, unlocking insights about regulatory dynamics. (4) HALO identifies distal *cis*-regulation interactions between chromatin regions and nearby genes, specifically for chromatin regions overlapping with super enhancers.

HALO effectively separates coupled and decoupled representations, enhancing the analysis and interpretation of modality-shared and modality-specific information in mouse skin follicle data. Through the integration of information from both ATAC-seq and RNA-seq modalities, HALO is able to discern cell types and capture latent temporal dynamics. The ATAC and RNA representations from HALO are concatenated and visualized according to cell type (Figure 2A and Figure S1A) and latent time (Figure S1C). **Figure 2B and F** display the UMAP projections constructed by the coupled RNA and ATAC representations, $\mathbf{Z}_{\mathrm{c}}^{\mathrm{A}}$ and $\mathbf{Z}_{\mathrm{c}}^{\mathrm{R}}$, respectively. In accordance with the concept of coupled representations, $\mathbf{Z}_{\mathrm{c}}^{\mathrm{A}}$ and $\mathbf{Z}_{\mathrm{c}}^{\mathrm{R}}$ capture analogous information across the two modalities, which is reflected in the resemblance between their UMAP visualizations. However, the decoupled representations $\mathbf{Z}_{\mathrm{d}}^{\mathrm{A}}$ and $\mathbf{Z}_{\mathrm{d}}^{\mathrm{R}}$ convey different information. In particular, $\mathbf{Z}_{\mathrm{d}}^{\mathrm{R}}$ retains cell type information, while $\mathbf{Z}_{\mathrm{d}}^{\mathrm{A}}$ does not (**Figure 2C and G)**.

Examining the analysis of the decoupled ATAC representation $\mathbf{Z}_{\mathbf{d}}^{\mathbf{A}}$, we performed clustering based on it. Cluster 4 (**Figure 2D)** is notably characterized by the decoupled ATAC latent representation 14 (**Figure 2H**). Additionally, we have mapped cluster 4 onto the original SHARE-seq UMAP projection (**Figure 2E** and **Figure S1B**), revealing that cluster 4 corresponds to the novel root cells[4]. These novel roots cells were previously identified by chromatin potential, which is a quantitative measure of chromatin lineage-priming and used for cell fate prediction. Moreover, it is also demonstrated that novel roots (cluster 4) express distinct marker genes in comparison to expected root cells (**Figure 2K).** To validate this novel cell type in HALO's framework, we evaluated the decoupled ATAC and RNA representations and found that, consistent with the original study[4], the cosine distance between the decoupled representations is maximal within the novel roots (**Figure 2I - J)**. By utilizing the interpretable decoder, HALO can identify the enriched transcription factors (TFs) within specific latent ATAC representations. **Figure 2N** shows the significantly enriched TFs in coupled representation 3 and decoupled representation 14, including Wt1 and Pou5fl. Transcription factors Wt1 and Pou5f1 are important in the Wnt/β-Catenin signaling pathway, exerting regulatory control over Lef1 and Dach1[5,6], which serve as marker genes for novel root cells Click or tap here to enter text.. In coupled ATAC representation 2, the transcription factor Runx1 is enriched, which directly promotes the proliferation of hair follicle stem cells and affects hair morphogenesis and differentiation. Coupled RNA representation 3, which is highly correlated with coupled ATAC representation 3, captures Eda, Wnt, and Shh signaling pathways, which are important for hair follicle morphogenesis[8].

To validate the relations between the coupled and decoupled representations of the two modalities, we show the Pearson correlation matrix for the latent representations (**Figure 2L**), which reveals a strong correlation between the ATAC and RNA coupled representations. In contrast, the decoupled representations exhibit a weaker correlation. Additionally, we assessed the correlations of latent representations between modalities using various representation learning methods, including MIRA[9], GLUE[10], MultiVI[11], and scMVP[12]. The comparative analysis (**Figure 2M, S1D)** indicates that HALO's decoupled representations maintain the lowest correlations, whereas its coupled representations are highly correlated across modalities with other four distinct datasets: Mouse Brain from 10X genomics, NEAT-Seq[13], NeurIPS[14] and Systemic sclerosis (SSc) pulmonary epithelium. We further interrogate whether the decoupled representations identified by HALO represent modality-specific batch bias or truly modality-specific biological information. To address this, we compared HALO's performance in batch correction against that of other frameworks on the NeurIPS datasets[14], which are known to contain substantial batch effects (**Figure S1E**), by silhouette score and Hilbert-Schmidt Independence Criterion (HSIC) (See methods). HALO is particularly effective at removing batch information, thereby confirming that its decoupled representations predominantly capture modality-specific biological information rather than mere batch effects.

To further examine our approach, we have developed a genomic matching score (See Methods) for coupled RNA and ATAC representation pairs that assesses the distance between their important genes and peaks by calculating the ratio of peaks that are located within the *cis*-regulation regions of the genes. **Figure 2O** presents the genomic matching scores for the coupled representations in mouse skin dataset, which demonstrates that HALO's coupled representations could capture regulatory interactions between peaks and gene expressions from the extent to which they are aligned with each other.

HALO characterizes gene-peak interactions in a temporal causal perspective. To further characterize the causal temporal relations between individual genes and peaks, we categorize gene-peak pairs into coupled genes and decoupled genes (**Figure 1A, 3B**). Employing negative binomial regression from denoised nearby peaks to denoised gene expression penalized by genomic distance, we align gene-peak pairs (See Methods). Upon aggregating these matched peaks, HALO calculates decouple and couple scores to quantitatively assess the extent of "decoupledness" and "coupledness" in gene-peak relationships, with positive scores suggesting decoupled and coupled, respectively**. Figure S3A-C** display the simulation outcomes for the decouple scores of both simulated decoupled and coupled gene-peaks pairs (**see Methods**). The SHARE-seq hair follicle data represent several lineages, differentiated from transit-amplifying cells (TACs) to inner root sheath (IRS), medulla, and cuticle/cortex cells (Figure 3A). **Figure 3C & D** and **S5 A & B** highlight the decoupled/coupled gene-peaks across all branches**,** with the respective couple and decouple scores (**Figure 3H, S4A & C**). Additionally, HALO takes into account distinct developmental lineages: Cortex lineage (**Figure 3E, S5C & D**), IRS (**Figure 3F, S5E & F**), and Medulla (**Figure 3G, S5G & H**). By examining the decoupled genes (Abl2, Dnajb1, Dlx3 and Btg2) across all branches or within specific lineages, it becomes apparent that gene expression levels and corresponding peaks change independently over time. Specifically, Dlx3 exhibits decoupled behaviors on all three branches (**Figure 3H**); the gene expression remains relatively stable, but the aggregated peaks keep going up with time (**Figure 3E-G**).

We address the question of why these peaks continue to increase by examining the downstream effect of the accessibility of a local chromatin region on the gene expression of potential regulation targets other than the corresponding gene. Granger causality is well suited for the task of inferring these regulatory relationships due to the *time lag* between a local peak and target gene expression change. By utilizing the Granger causality test (see Methods), we are able to identify the distal regulation relations among Dlx3's corresponding peaks and nearby genes other than Dlx3 (**see Methods**)**.** In **Figure 3I**, the pink loops are peak-gene interactions inferred by DORC[4], while the blue loops are uniquely detected by negative binomial regression (See Methods). By predicting gene expression using the peaks of preceding cells with respect to ascending latent time, we find that nine local peaks within mouse skin hair follicle super enhancer region exhibit Granger causal relations with the expression of Itgs (**Figure 3J-K** and **Figure S4B**)[15]. In summary, by analyzing gene-peak pairs at the individual gene level, HALO effectively distinguishes between coupled and decoupled interactions, as well as illustrates that local peaks of decoupled genes may contribute to expression of nearby genes.

HALO uncovers regulatory factors in primary human CD4 memory T cells assayed by NEAT-seq. NEAT-seq profiles the intra-nuclear protein epitope abundance of a panel T cell master transcription factors (TFs), chromatin accessibility, and transcriptome in single cells[13]. Due to posttranscriptional regulatory mechanisms affecting GATA3 protein expression[16], we can leverage the protein quantification of the Th2 master TF GATA3 for more cell state information than can be gleaned from using RNA expression levels alone. HALO infers latent representations utilizing nuclear protein level of GATA3 as a proxy variable for latent time (**See Methods**) and constructs UMAP embedding to visualize distinct T cell subsets (**Figure 4A, S6 A-C, S7 & S8**). Notably, two coupled latent representations, RNA coupled 9 and ATAC coupled 9, show negative correlations with the nuclear protein level of GATA3 (**Figure 4 16B-C**). RNA coupled

9 captures mTORC1 signaling (**Figure 4D**), which negatively regulates Th2 differentiation[17]. Consistently, ATAC coupled 9 is enriched in ZFX/NR4A motifs (**Figure 4E**), where NR4A is known to suppresses Th2 genes[18,19]. Conversely, Th17-specific ATAC coupled representation 12 exhibits enrichment in NR2F6/ARID5A motifs (**Figure S8 & 4E**), which are essential for the differentiation and function of Th17 cells[20,21].

GATA3 regulates the Th2 cell fate decision in CD4+ T cells, while Tbet orchestrates Th1 differentiation. Both T-bet and GATA3 are co-expressed in Th1 cells, rather than being exclusively present in their respective lineages (**Figure 4F**). Within Th1 cells, T-bet suppresses Th2 genes by redistributing GATA3 from Th2 gene loci to Tbet binding sites within Th1 gene regions[22]. Granzyme genes (GZMA, GZMB, GZMK) possess binding sites specifically targeted by both T-bet and GATA3 in Th1 cells[23]. To dissect the regulation mechanisms of different TFs, we calculate decouple scores for Th1 cells using protein levels of GATA3 and T-bet as temporal information. Intriguingly, expression of the GZMA gene and its local peaks exhibit decoupled dynamics in Th1 cells (**Figure 4 G-H**). As GATA3 protein levels rise, local peaks of GZMA become less accessible, even though GZMA gene expression is increased at high GATA3 protein levels (**Figure 4H**). Consistently, CD4+ T cells co-expressing T-bet and GATA3 exhibit upregulation of GZMA gene expression, compared to cells expressing only T-bet (**Figure S6D**). Conversely, increases in nuclear T-bet levels enhance the accessibility of local GZMA peaks, while GZMA gene expression level remains constant (**Figure 4G**). Using nuclear protein levels as temporal information, Granger causality tests reveal that local peaks of GZMA mediate the gene expression of GZMK (**Figure 4 I-L**). Strikingly, local peaks of GZMA within the Th1-specific T-bet super enhancer region are also crucial for regulating gene expression of GZMK[24]. Leveraging TF protein levels as proxies for temporal information, HALO coupled representations not only exhibit strong correlation between RNA and ATAC modalities, but also capture similar biological contexts (**Figure S6E**). HALO allows us to further investigate the regulation of a TF using Granger causality to determine downstream enhancer and its gene targets.

Discussion

The relationship between chromatin accessibility and gene expression is complex and often asynchronous. Previous methods assume that the two are synced with respect to time, missing nuances like time-lagging effects and independent regulatory mechanisms. HALO is the first to address these gaps by differentiating between coupled (dependent) and decoupled (independent) changes from a causal perspective. The framework operates at both representations and individual gene levels to learn interpretable modality-specific and shared information, as well as characterizes gene and associated peak interactions.

We conduct extensive benchmarks to evaluate the performance of HALO from different perspectives. Across several datasets, we demonstrate that HALO effectively learns highly coupled information across modalities while preserving decoupled and modality-specific information rather than batch bias. The highly coupled RNA and ATAC representation pairs, characterized by high Pearson correlation, capture functionally analogous biological contexts. In contrast, decoupled ATAC representations can inform our knowledge of chromatin accessibility-mediated cell fate potentials during the differentiation of mouse skin hair follicles.

At the individual gene level, HALO distinguishes between coupled and decoupled gene-peak pairs. For decoupled cases where associated chromatin regions become more accessible but gene expression remains stable or decreases, HALO employs Granger causality to assess distal regulatory interactions between associated peaks and nearby genes. We find that peaks overlapping with super enhancer regions often exhibit such decoupled behavior and are involved in distal regulation during cellular transition or developmental trajectories.

To enhance HALO's applicability, future work can extend the framework to include profiling of additional modalities, such as methylation and protein levels, to provide a more comprehensive understanding of gene regulation dynamics. Additionally, leveraging spatial epigenome-transcriptome co-profiling technologies could allow HALO to uncover spatiotemporal dynamics and genome-wide gene regulation mechanisms within tissue contexts[25]. These advancements would extend HALO’s utility as a valuable tool for understanding the functions and regulatory mechanisms of cell populations.

Methods

See the Supplementary Methods Section.

Bibliography

1. Li, C., Virgilio, M. C., Collins, K. L. & Welch, J. D. Multi-omic single-cell velocity models epigenome–transcriptome interactions and improves cell fate prediction. *Nat Biotechnol* **41**, 387–398 (2023).
2. Liu, J. *et al.* Jointly defining cell types from multiple single-cell datasets using LIGER. *Nat Protoc* **15**, 3632–3662 (2020).
3. Wu, A. P., Singh, R. & Berger, B. Granger causal inference on DAGs identifies genomic loci regulating transcription. *ArXiv* **abs/2210.10168**, (2022).
4. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**, 1103-1116.e20 (2020).
5. Huggins, I. J. *et al.* The WNT target SP5 negatively regulates WNT transcriptional programs in human pluripotent stem cells. *Nat Commun* **8**, 1034 (2017).
6. Paris, N. D., Coles, G. L. & Ackerman, K. G. Wt1 and β-catenin cooperatively regulate diaphragm development in the mouse. *Dev Biol* **407**, 40–56 (2015).
7. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**, 1103-1116.e20 (2020).
8. Sulic, A.-M. *et al.* Transcriptomic landscape of early hair follicle and epidermal development. *Cell Rep* **42**, 112643 (2023).
9. Lynch, A. W. *et al.* MIRA: joint regulatory modeling of multimodal expression and chromatin accessibility in single cells. *Nat Methods* **19**, 1097–1108 (2022).
10. Cao, Z.-J. & Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat Biotechnol* **40**, 1458–1466 (2022).
11. Ashuach, T. *et al.* MultiVI: deep generative model for the integration of multimodal data. *Nat Methods* **20**, 1222–1231 (2023).
12. Li, G. *et al.* A deep generative model for multi-view profiling of single-cell RNA-seq and ATAC-seq data. *Genome Biol* **23**, 20 (2022).
13. Chen, A. F. *et al.* NEAT-seq: simultaneous profiling of intra-nuclear proteins, chromatin accessibility and gene expression in single cells. *Nat Methods* **19**, 547–553 (2022).
14. Lance, C. *et al.* Multimodal single cell data integration challenge: Results and lessons learned. in *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track* (eds. Kiela, D., Ciccone, M. & Caputo, B.) vol. 176 162–176 (PMLR, 2022).
15. Adam, R. C. *et al.* Pioneer factors govern super-enhancer dynamics in stem cell plasticity and lineage choice. *Nature* **521**, 366–370 (2015).
16. Cook, K. D. & Miller, J. TCR-Dependent Translational Control of GATA-3 Enhances Th2 Differentiation. *The Journal of Immunology* **185**, 3209–3216 (2010).
17. Chi, H. Regulation and function of mTOR signalling in T cell fate decisions. *Nat Rev Immunol* **12**, 325–338 (2012).
18. Sekiya, T. *et al.* Suppression of Th2 and Tfh immune reactions by Nr4a receptors in mature T reg cells. *Journal of Experimental Medicine* **212**, 1623–1640 (2015).
19. Sekiya, T. *et al.* Regulation of peripheral Th/Treg differentiation and suppression of airway inflammation by Nr4a transcription factors. *iScience* **24**, 102166 (2021).
20. Masuda, K. *et al.* Arid5a regulates naive CD4+ T cell fate through selective stabilization of Stat3 mRNA. *Journal of Experimental Medicine* **213**, 605–619 (2016).
21. Hermann-Kleiter, N. & Baier, G. Orphan nuclear receptor NR2F6 acts as an essential gatekeeper of Th17 CD4+ T cell effector functions. *Cell Communication and Signaling* **12**, 38 (2014).
22. Hertweck, A. *et al.* The TH1 cell lineage-determining transcription factor T-bet suppresses TH2 gene expression by redistributing GATA3 away from TH2 genes. *Nucleic Acids Res* **50**, 4557–4573 (2022).
23. Kanhere, A. *et al.* T-bet and GATA3 orchestrate Th1 and Th2 differentiation through lineage-specific targeting of distal regulatory elements. *Nat Commun* **3**, 1268 (2012).
24. Hertweck, A. *et al.* T-bet Activates Th1 Genes through Mediator and the Super Elongation Complex. *Cell Rep* **15**, 2756–2770 (2016).
25. Zhang, D. *et al.* Spatial epigenome–transcriptome co-profiling of mammalian tissues. *Nature* **616**, 113–122 (2023).
26. Mitra, S. *et al.* Single-cell multi-ome regression models identify functional and disease-associated enhancers and enable chromatin potential analysis. *Nat Genet* **56**, 627–636 (2024).

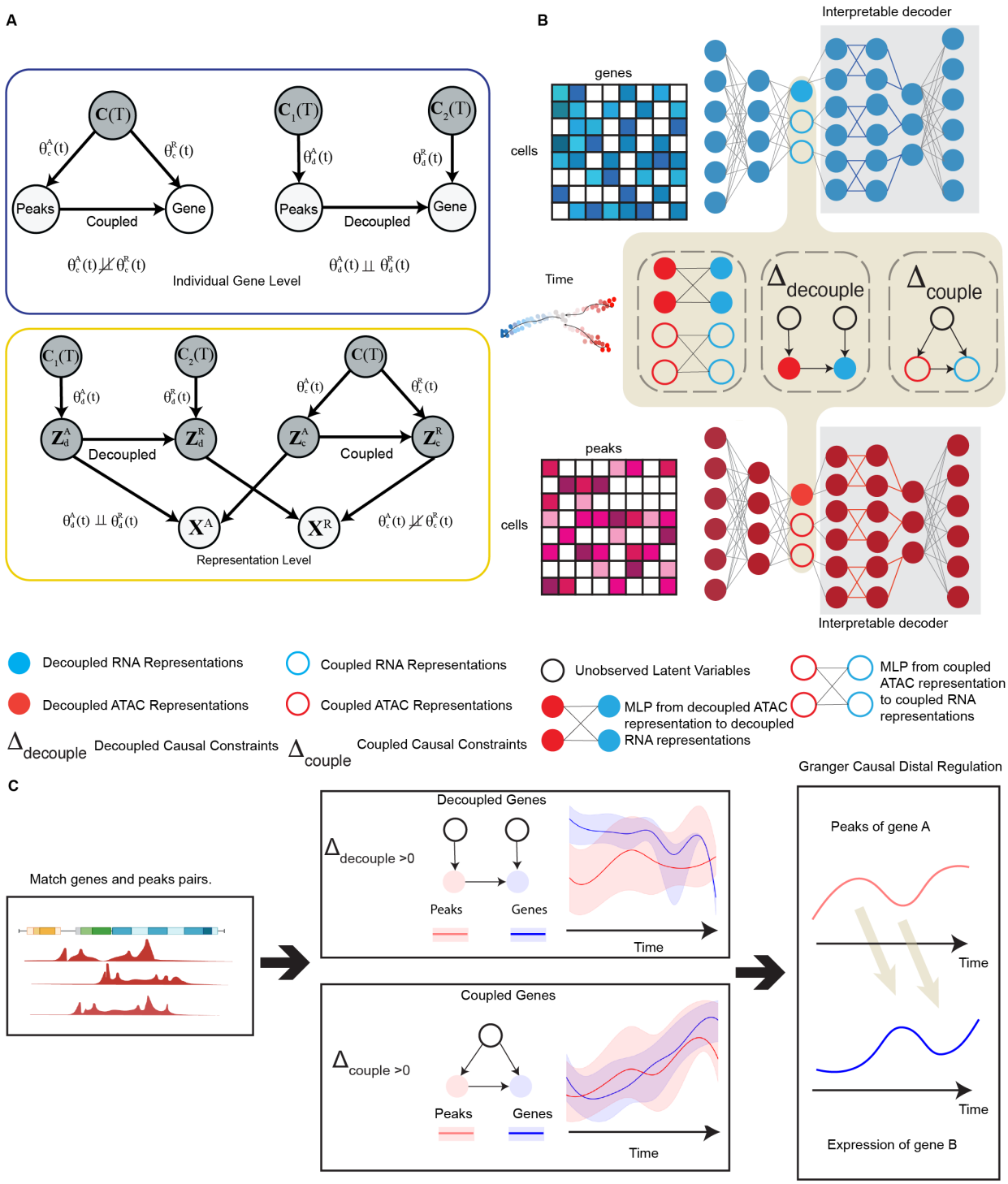
A
C(T)
Peaks
Gene
Coupled
Decoupled
Individual Gene Level
Representation Level
B
Interpretable decoder
genes
cells
Time
peaks
Δdecouple
Δcouple
Decoupled RNA Representations
Coupled RNA Representations
Unobserved Latent Variables
MLP from coupled ATAC representation to coupled RNA representations
Decoupled ATAC Representations
Coupled ATAC Representations
MLP from decoupled ATAC representation to decoupled RNA representations
Decoupled Causal Constraints
Coupled Causal Constraints
C
Match genes and peaks pairs.
Decoupled Genes
Δdecouple >0
Coupled Genes
Δcouple >0
Granger Causal Distal Regulation
Peaks of gene A
Expression of gene B

Figure 1. The Main Framework of HALO.

A. Top: The causal diagram of individual gene expression and its corresponding peaks; Bottom: the causal diagram of representation level of scRNA-seq and scATAC-seq.

B. Depicts the architecture for representation learning within a causal regularized variational autoencoder (VAE) framework, focusing on the derivation of latent representations. For the ATAC modality, the latent space, denoted as $\mathbf{Z_A}$ is the latent representation of ATAC modality, is divided into $\{\mathbf{Z_A^d}, \mathbf{Z_A^c}\}$, representing the decoupled and coupled latent representations, respectively. Similarly, the RNA modality's latent space comprises $\mathbf{Z_R^d}$ for the decoupled and $\mathbf{Z_R^c}$ is the latent decoupled representation. The decoupled representations, $\mathbf{Z_A^d}$ and $\mathbf{Z_R^d}$ are adhere to the decoupled causal constraints $\Delta_{decouple}$, whereas the coupled representations, $\mathbf{Z_A^c}$, $\mathbf{Z_R^d}$ conform to the coupled causal constraints Δ_{couple}. This structure enables a nuanced understanding of the latent dynamics within each modality by distinguishing between decoupled and coupled genetic influences.

C. Illustrates the gene-peak level analysis process. Initially, genes and ATAC peaks within gene proximities are matched using non-negative binomial regression, linking gene expression to neighboring ATAC peaks. Subsequently, this analysis computes decouple and couple scores to categorize gene-peaks into either decoupled or coupled genes. Finally, the analysis employs the Granger causality test to identify distal regulatory relationships between peaks and genes, uncovering potential mechanisms of genetic regulation. The Granger causality test to identify distal regulatory relationships between peaks and genes, uncovering potential mechanisms of genetic regulation.

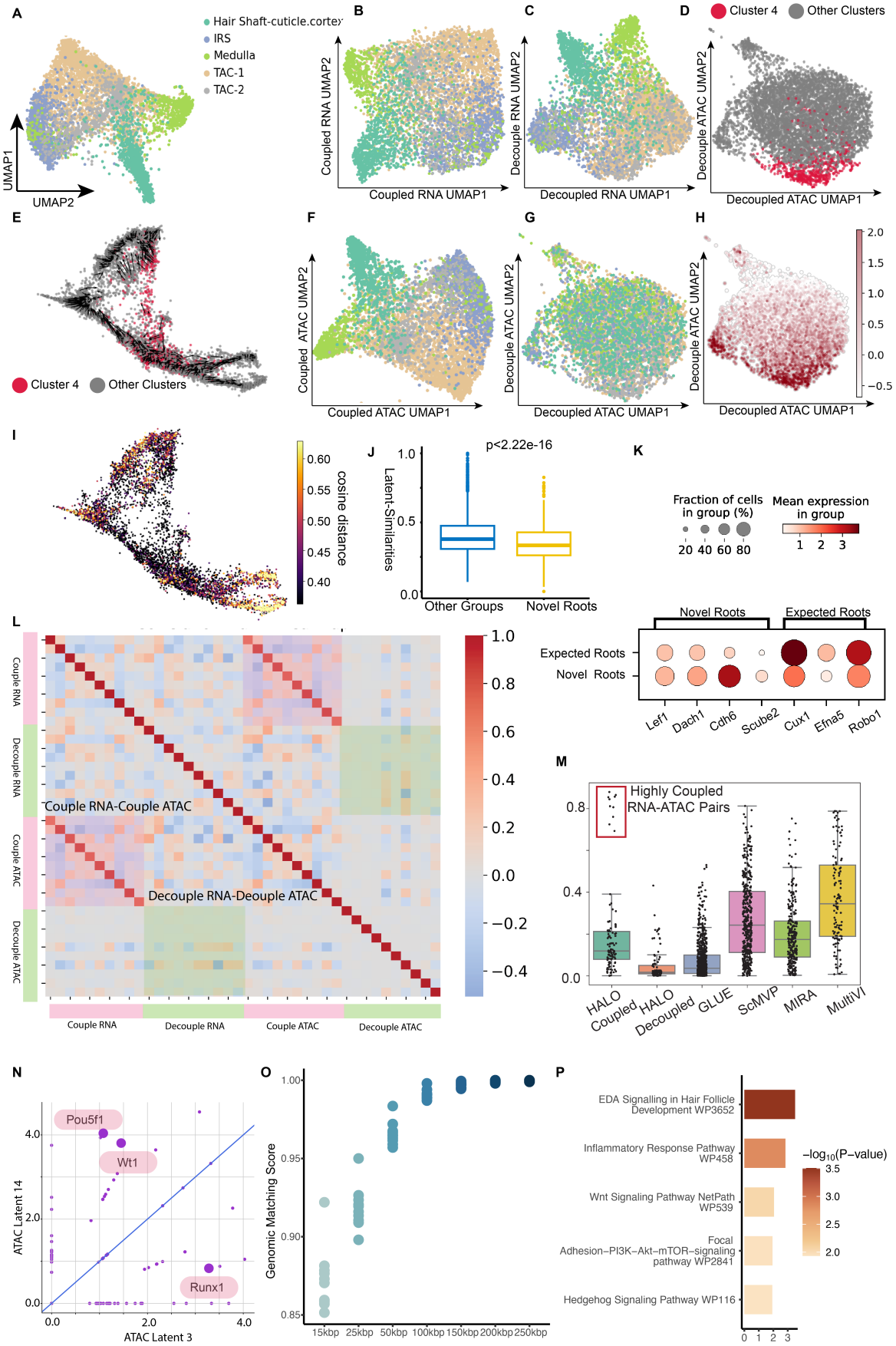
A
Hair Shaft-cuticle.cortex
IRS
Medulla
TAC-1
TAC-2
UMAP1
UMAP2
B
Coupled RNA UMAP2
Coupled RNA UMAP1
C
Decouple RNA UMAP2
Decoupled RNA UMAP1
D
Cluster 4
Other Clusters
Decouple ATAC UMAP2
Decoupled ATAC UMAP1
E
Cluster 4
Other Clusters
F
Coupled ATAC UMAP2
Coupled ATAC UMAP1
G
Decouple ATAC UMAP2
Decoupled ATAC UMAP1
H
Decouple ATAC UMAP2
Decoupled ATAC UMAP1
I
cosine distance
J
p<2.22e-16
Latent-Similarities
Other Groups
Novel Roots
K
Fraction of cells in group (%)
Mean expression in group
Novel Roots
Expected Roots
Expected Roots
Novel Roots
Lef1
Dach1
Cdh6
Scube2
Cux1
Efna5
Robo1
L
Couple RNA
Decouple RNA
Couple ATAC
Decouple ATAC
Couple RNA-Couple ATAC
Decouple RNA-Deouple ATAC
M
Highly Coupled RNA-ATAC Pairs
HALO Coupled
HALO Decoupled
GLUE
ScMVP
MIRA
MultiVI
N
Pou5f1
Wt1
Runx1
ATAC Latent 14
ATAC Latent 3
O
Genomic Matching Score
15kbp
25kbp
50kbp
100kbp
150kbp
200kbp
250kbp
P
EDA Signalling in Hair Follicle Development WP3652
Inflammatory Response Pathway WP458
Wnt Signaling Pathway NetPath WP539
Focal Adhesion−PI3K−Akt−mTOR−signaling pathway WP2841
Hedgehog Signaling Pathway WP116
−log10(P−value)

Figure 2. The representation level results of SHARE-seq mouse hair dataset. The decouple representations contain the chromatin potential information in the latent space.

- **A.** The UMAP of RNA and ATAC representations.
- **B.** The UMAP of RNA coupled representations.
- **C.** The UMAP of RNA decoupled representations.
- **D.** The UMAP of cluster 4 and the rest of the clusters based on the ATAC decoupled representations.
- **E.** The cluster 4 and rest of clusters, and chromatin potential vector field shown on previously published UMAP[4,26].
- **F.** The UMAP of ATAC coupled representations.
- **G.** The UMAP of ATAC decoupled representations.
- **H.** The UMAP of ATAC latent representation 14, which mainly characterizes cluster 4.
- **I.** UMAP of SHARE-Seq colored by the cosine distance of decoupled RNA and decoupled ATAC representations.
- **J.** The boxplots of the similarity of decoupled RNA and decoupled ATAC representations between cluster 4 (Novel Roots) and rest of cells, as determined by cosine similarity.
- **K.** The marker gene expression levels of the novel roots and expected roots.
- **L.** The Pearson-correlations matrix of latent representations.
- **M.** The Pearson-correlations for HALO (coupled representations), HALO (decoupled representations), GLUE, scMVP, MIRA, and MultiVI.
- **N.** The enriched transcription factors (TFs) for coupled ATAC latent representation 3 and decoupled ATAC latent representation 14.
- **O.** The Genomic Matching Score of coupled representations with different genomic distances from genes' TSS site.
- **P.** The gene ontology (GO) enrichment of top genes of RNA coupled representation 3.

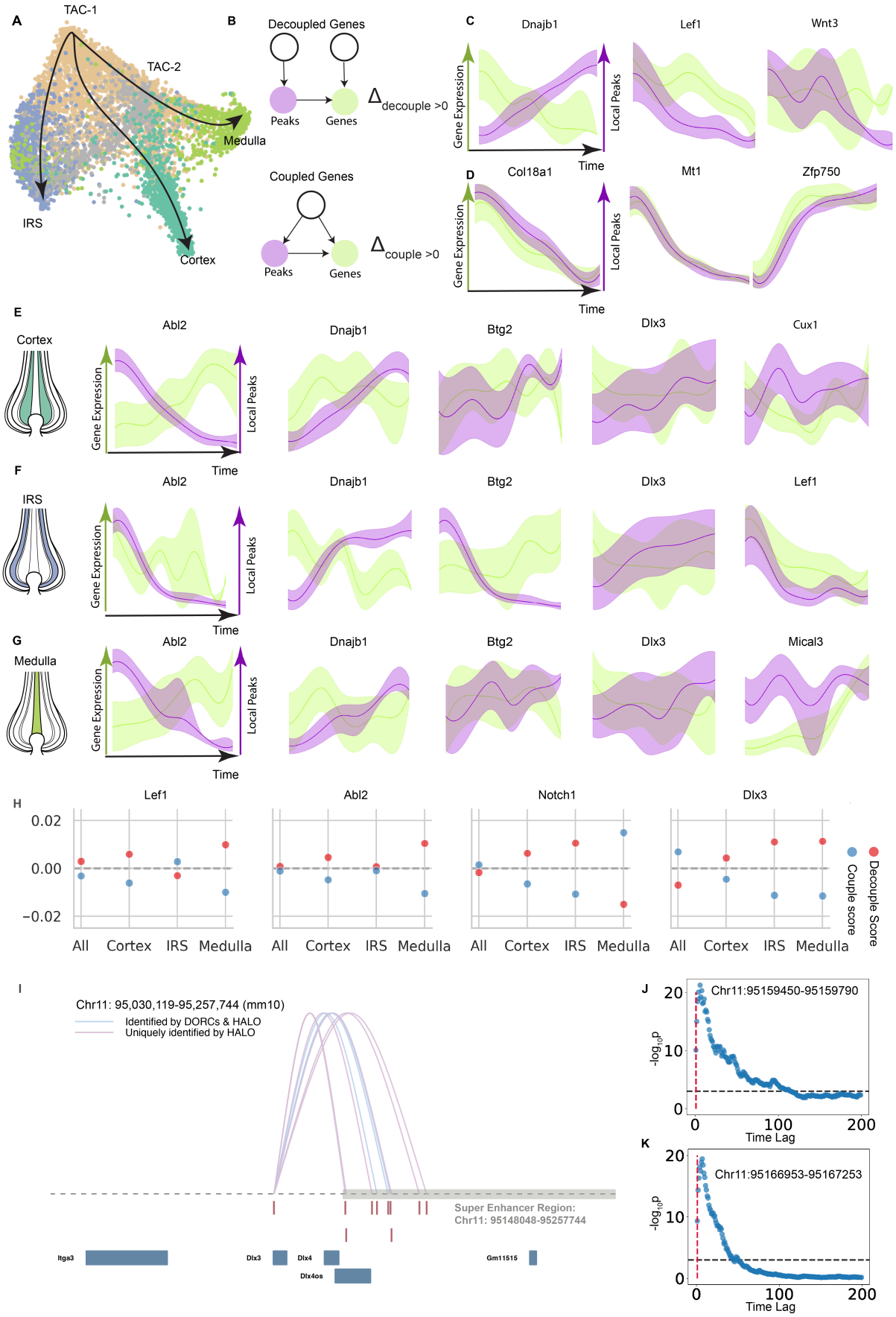
A
TAC-1
TAC-2
Medulla
IRS
Cortex
B
Decoupled Genes
Peaks
Genes
Δdecouple >0
Coupled Genes
Peaks
Genes
Δcouple >0
C
Dnajb1
Lef1
Wnt3
Gene Expression
Local Peaks
Time
D
Col18a1
Mt1
Zfp750
Gene Expression
Local Peaks
Time
E
Cortex
Abl2
Dnajb1
Btg2
Dlx3
Cux1
Gene Expression
Local Peaks
Time
F
IRS
Abl2
Dnajb1
Btg2
Dlx3
Lef1
Gene Expression
Local Peaks
Time
G
Medulla
Abl2
Dnajb1
Btg2
Dlx3
Mical3
Gene Expression
Local Peaks
Time
H
Lef1
Abl2
Notch1
Dlx3
0.02
0.00
−0.02
All
Cortex
IRS
Medulla
Couple score
Decouple Score
I
Chr11: 95,030,119-95,257,744 (mm10)
Identified by DORCs & HALO
Uniquely identified by HALO
Super Enhancer Region:
Chr11: 95148048-95257744
Itga3
Dlx3
Dlx4
Dlx4os
Gm11515
J
Chr11:95159450-95159790
-log10p
Time Lag
K
Chr11:95166953-95167253
-log10p
Time Lag

Figure 3. The individual gene level results of SHARE-seq mouse hair dataset.

A. The development trajectory of mouse hair cells.

B. The causal diagrams of decoupled and coupled individual gene-peak pairs. If the decouple score $\Delta_{decouple} > 0$, we classify the gene-peak pair as decoupled. If the couple score $\Delta_{couple} > 0$, we classify the gene-peak pair as coupled.

C. Examples of decoupled gene-peak pairs across all branches.

D. Examples of coupled gene-peak pairs across all branches.

E. Examples of decoupled gene-peak pairs in Cortex.

F. Examples of decoupled gene-peak pairs in inner root sheath (IRS).

G. Examples of decoupled gene-peak pairs in Medulla.

H. The decouple/couple score of genes on different developmental branches.

I. The loops denote significant connections between peaks within super enhancer region and RNA expression of Dlx3. The connections between the expression of Dlx3 and peaks are identified by a gene-peak-linking algorithm and reported by[4].

J. The scatter plot visualizes significance of Granger causal relations between Itga3 gene expression and peaks in Chr11: 95159490-95159790. The X-axis is the time lag (number of cells, sorted by latent time), Y-axis is the Granger causality -log(P-values). Granger causality significance is evaluated by using the peaks of all previous cells $[c_{t-n}, c_t)$ up to lag *n* (number of cells with index preceding time *t*) to determine the gene expression for cell c_t, where *t* is determined by latent time sequential ordering of all cells. We utilize the likelihood ratio test for the Granger-causality based regulation inference.

K. The scatter plot of Granger causal relations of Itga3 gene expression and peaks in Chr11: 95166953-95167253. The X-axis is the time lag (number of cells sorted by latent time), Y-axis is the Granger causality -log(P-values).

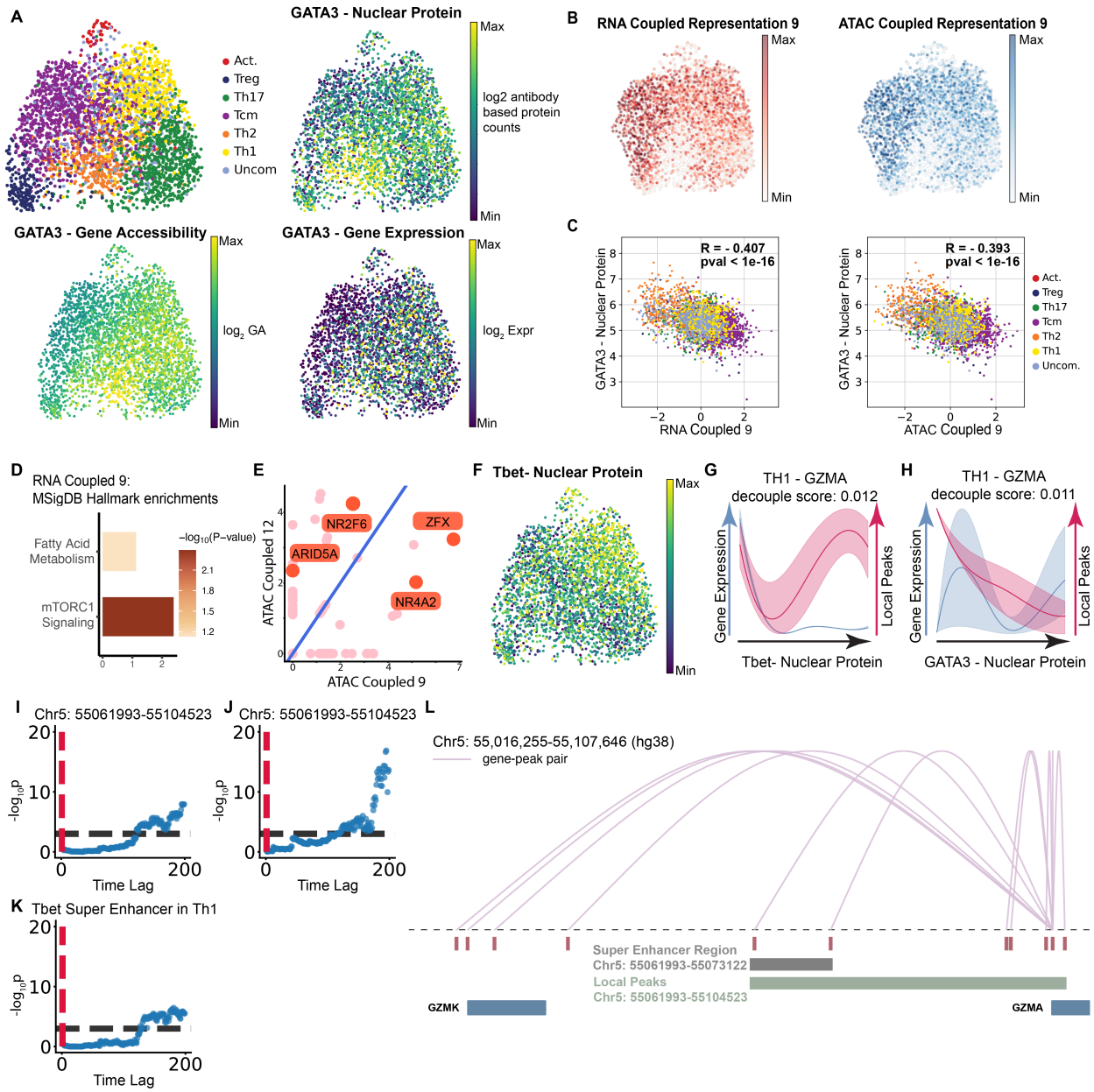

Figure 4. HALO reveals regulatory factors of human CD4 memory T cell dataset

A. UMAP visualization of human CD4 memory T cells colored by original cell type, as well as nuclear protein, gene accessibility, and gene expression of GATA3.
B. RNA coupled 9 and ATAC coupled 9 representations on UMAP embedding of RNA and ATAC representations.
C. Scatter plots visualization GATA3 nuclear protein level and coupled latent representations (RNA coupled 9 and ATAC coupled 9)
D. Gene set enrichment for RNA coupled 9.
E. Motif enrichment for ATAC coupled 9 and ATAC coupled 12.
F. Nuclear protein level of Tbet on UMAP embedding of RNA and ATAC representations.
G. Gene-peak pairs of GZMA in Th1 cells sorted by Tbet nuclear protein level.
H. Gene-peak pairs of GZMA in Th1 cells sorted by GATA3 nuclear protein level.

I. The scatter plot of Granger causal relations of GZMK gene expression and peaks in Chr5: 55061993-55104523. The X-axis is the time lag (number of cells, sorted by Tbet nuclear protein level), Y-axis is the Granger causality -log(P-values). We utilized the F-test for the Granger-causality based regulation inference.
J. The scatter plot of Granger causal relations of GZMK gene expression and peaks in Chr5: 55061993-55104523. The X-axis is the time lag (number of cells, sorted by GATA3 nuclear protein level), Y-axis is the Granger causality -log(P-values).
K. The scatter plot of Granger causal relations of GZMK gene expression and peaks in Th1-specific T-bet super enhancer region (Chr5: 55061993-55073122). The X-axis is the time lag (number of cells, sorted by Tbet nuclear protein level), Y-axis is the Granger causality -log(P-values).
L. The loops denote significant connections between local peaks and RNA expression of GZMA. The connections between the expression of GZMA and peaks are identified by a gene-peak-linking algorithm.