Nonparametric Interrogation of Transcriptional Regulation in Single-Cell RNA and Chromatin Accessibility Multiomic Data

Yuriko Harigaya
University of North Carolina
Zhaojun Zhang
University of Pennsylvania
Hongpan Zhang
University of Virginia
Chongzhi Zang
University of Virginia
Nancy Zhang (nzh@wharton.upenn.edu)
University of Pennsylvania https://orcid.org/0000-0002-0880-5749
Yuchao Jiang
University of North Carolina

Article

Keywords: single-cell multiomics, transcriptional regulation, transcription factor, chromatin accessibility

Posted Date: September 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-930184/v1

Nonparametric Interrogation of Transcriptional Regulation in Single-Cell RNA and Chromatin Accessibility Multiomic Data

Yuriko Harigaya[1], Zhaojun Zhang[2], Hongpan Zhang[3,4], Chongzhi Zang[3,4,5], Nancy R Zhang[2,*], Yuchao Jiang[6,7,8,*]

[1] Curriculum in Bioinformatics and Computational Biology, School of Medicine, University of North Carolina, Chapel Hill, NC 27599, USA.

[2] Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA.

[3] Center for Public Health Genomics, University of Virginia, Charlottesville, VA 22908, USA.

[4] Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA 22908, USA.

[5] Department of Public Health Sciences, University of Virginia, Charlottesville, VA 22908, USA.

[6] Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, NC 27599, USA.

[7] Department of Genetics, School of Medicine, University of North Carolina, Chapel Hill, NC 27599, USA.

[8] Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC 27599, USA.

* To whom correspondence should be addressed: nzh@wharton.upenn.edu, yuchaoj@email.unc.edu.

Abstract

Epigenetic control of gene expression is highly cell-type- and context-specific. Yet, despite its complexity, gene regulatory logic can be broken down into modular components consisting of a transcription factor (TF) activating or repressing the expression of a target gene through its binding to a *cis*-regulatory region. Recent advances in joint profiling of transcription and chromatin accessibility with single-cell resolution offer unprecedented opportunities to interrogate such regulatory logic. Here, we propose a nonparametric approach, TRIPOD, to detect and characterize three-way relationships between a TF, its target gene, and the accessibility of the TF's binding site, using single-cell RNA and ATAC multiomic data. We apply TRIPOD to interrogate cell-type-specific regulatory logic in peripheral blood mononuclear cells and contrast our results to detections from enhancer databases, *cis*-eQTL studies, ChIP-seq experiments, and TF knockdown/knockout studies. We then apply TRIPOD to mouse embryonic brain data during neurogenesis and gliogenesis and identified known and novel putative regulatory relationships, validated by ChIP-seq and PLAC-seq. Finally, we demonstrate TRIPOD on SHARE-seq data of differentiating mouse hair follicle cells and identify lineage-specific regulation supported by histone marks for gene activation and super-enhancer annotations.

Keywords: single-cell multiomics, transcriptional regulation, transcription factor, chromatin accessibility.

Context-specific regulation of gene transcription is central to cell identity and function in eukaryotes. Precision of transcriptional control is achieved through multitudes of transcription factors (TFs) that bind to the *cis*-regulatory regions of their target genes, dynamically modulating chromatin accessibility and recruiting transcription complexes in response to developmental and environmental cues[1]. Dissecting this regulatory logic is fundamental to our understanding of biological systems and our study of diseases. Over the past decades, molecular studies have elucidated the structure of TF complexes and provided mechanistic models into their function[2]. Methods based on high-throughput sequencing have enabled the genome-wide profiling of gene expression[3], TF binding[4], chromatin accessibility[5], and 3D genome structure[6]. TF knockdown/knockout studies have also identified, *en masse*, their species-, tissue-, and context-specific target genes[7]. Concurrently, novel statistical approaches have allowed for more precise identification and modeling of TF binding sites[8], and expression quantitative trait loci (eQTLs) databases now include associations that are tissue-specific[9] and will soon be cell-type specific[10]. Yet, despite this tremendous progress, our understanding of gene regulatory logic is still rudimentary.

When a TF j activates or represses the expression of a gene g through binding to a regulatory element t in *cis* to the gene, we call such a relationship a *regulatory trio*. Despite its complexity, gene regulatory logic can be broken down into modular components consisting of such peak-TF-gene trios. In this paper, we focus on the identification of regulatory trios using multiomic experiments that jointly profile gene expression and chromatin accessibility at single-cell resolution.

Single-cell RNA sequencing (scRNA-seq) and single-cell assay of transposase-accessible chromatin sequencing (scATAC-seq), performed separately, have already generated detailed cell-type-specific profiles of gene expression and chromatin accessibility. When the two modalities are not measured in the same cells, the cells can be aligned by computational methods[11], followed by association analyses of gene expression and peak accessibility. While these methods have been shown to align well-differentiated cell types correctly, they often fail for cell populations consisting of transient states. Additionally, the alignment of cells necessarily assumes a peak-gene association model, which is often learned from other datasets. Then, the post-alignment association

analysis is plagued by logical circularity, as it is difficult to disentangle new findings from prior assumptions that underlie the initial cell alignment.

Single-cell multiomic experiments that sequence the RNA and ATAC from the same cells directly enable joint modeling of a cell's RNA expression and chromatin state, yet methods for the analysis of such data are still in their infancy. Almost all existing methods for detecting and characterizing regulatory relationships between TF, regulatory region, and target gene rely on marginal relationships. For example, Signac[12] and Ma *et al.*[13] use marginal associations between peaks and genes to identify putative enhancer regions, while Signac[12] and Seurat V4[14] link differentially expressed TFs to differentially accessible motifs across cell types. Such pairwise marginal associations are sometimes examined manually, in tandem, using low-dimensional embedding. One exception is in PECA[15], which uses a parametric model to characterize the joint distribution of TF expression, regulatory site accessibility, chromatin remodeler expression, and target gene expression. Although PECA was designed to be applied to matched bulk transcriptomic and epigenomic data, such joint modeling concepts could potentially be very powerful for single-cell multiomic data. Yet, PECA relies heavily on parametric assumptions, is computationally intensive to fit to large datasets, and is difficult to diagnose.

Context-specific gene regulation may be masked in marginal associations, as we will show in examples later. We explore in this paper the use of higher-order models that interrogate conditional and three-way interaction relationships for the identification of regulatory trios. First, as proof of principle, we show that a simple model that integrates TF expression with *cis*-peak accessibility significantly improves gene expression prediction, as compared to a comparable model that utilizes peak accessibility alone. We present TRIPOD, a computational framework for *t*ranscription *r*egulation *i*nterrogation through nonparametric *p*artial association analysis of single-cell multi*o*mic sequencing *d*ata. TRIPOD detects two types of trio relationships, which we call conditional level 1 and conditional level 2, through robust nonparametric tests that are easy to diagnose. A novel influence measure allows the detection and visualization of cell states driving these regulatory relationships, applicable both to discrete cell types, as well as continuous cell trajectories.

We first apply TRIPOD to single-cell multiomic data of human peripheral blood mononuclear cells (PBMCs) and compare the regulatory trios detected to relationships detected through marginal associations. We show that the detections are coherent with the vast amounts of existing knowledge from enhancer databases, bulk cell-type-specific chromatin immunoprecipitation followed by sequencing (ChIP-seq) experiments, tissue-specific TF knockdown/knockout studies, and *cis*-eQTL studies, but that conditional and marginal models identify different sets of relationships. We next apply TRIPOD to the interrogation of lineage-specific regulation in the developing mouse brain, where relationships detected by TRIPOD are compared against those derived from existing ChIP-seq and proximity ligation-assisted ChIP-seq (PLAC-seq) data. Here, TRIPOD identifies known trio relationships, as well as putative novel regulatory crosstalk between neuronal TFs and glial-lineage genes. We also apply TRIPOD to SHARE-seq data on mouse hair follicle cell differentiation to illustrate trio detection and influence analysis in data collected from different protocols. Through these analyses, we demonstrate how to harness single-cell multiomic technologies in the study of gene regulation, and how the data from these technologies corroborate and complement existing data.

Results

A simple interaction model between TF expression and peak accessibility improves RNA prediction. We started our analyses by making gene expression predictions, a standard procedure carried out by existing methods[11]. We benchmarked against: (i) Signac[12] and Cicero[16], which compute the gene activity matrix by summing the ATAC reads in gene bodies and promoter regions; (ii) MAESTRO[17], which adopts a regulatory potential model by taking the weighted sum of the ATAC reads based on existing gene annotations; and (iii) sci-CAR[18], which performs a regularized regression using gene expression as the outcome and peak accessibilities as the predictors. We also use a regularized regression model (referred to as peak-TF LASSO model), where instead of peak accessibility, interactions between TF expressions and peak accessibilities serve as predictors. The model considers only peaks within a certain range of the gene's transcription start site (TSS) and only interactions between TFs and peaks containing

high-scoring binding motifs for the TFs. To avoid overfitting, we perform leave-one-out prediction and adopt independent training and testing sets. See Methods for details.

We analyzed single-cell multiomic datasets from different human and mouse tissues generated by different platforms – PBMC by 10X Genomics, embryonic mouse brain by 10X Genomics, mouse skin by SHARE-seq[13], and adult mouse brain by SNARE-seq[19]. Data summaries are included in Supplementary Table 1; reduced dimensions via uniform manifold approximation and projection (UMAP)[20] are shown in Fig. 1a and Supplementary Fig. 1,2a. To mitigate the undesirable consequences of sparsity and stochasticity in the single-cell data, we clustered cells to form metacells[14], and pooled gene expression and chromatin accessibility measurements within each metacell.

Our results show that, across window sizes, the peak-TF LASSO model significantly improves prediction accuracy across the transcriptome (Fig.1b), with examples shown in (Fig. 1c). This improvement in prediction accuracy holds true when an independent dataset is used for validation (Supplementary Fig. 3). For the SNARE-seq data[19], sequencing depth is substantially shallower (Supplementary Fig. 4), thus the improvement of the peak-TF LASSO model is diminished but still evident (Supplementary Fig. 2b). This demonstrates that the product of TF expression and peak accessibility significantly improves RNA prediction accuracy beyond simply using peak accessibility, offering strong empirical evidence of three-way interaction relationships between TF expression, peak accessibility, and target gene expression that can be extracted from such multiomic experiments. However, we will not rely on coefficients from the LASSO model to screen for such trios, as their significance is difficult to compute due to the hazards of post-selection inference[21]. Additionally, accessibility of peaks and expression of TF affecting the same gene are often highly correlated, in which case LASSO tends to select the few with the highest associations and ignore the rest. In such cases, we believe it is more desirable to report all trios.

TRIPOD for the detection of peak-TF-gene trio regulatory relationships by single-cell multiomic data. We propose TRIPOD, a nonparametric method that screens single-cell RNA and ATAC multiomic data for conditional associations and three-way interactions between the expression of a TF j, the accessibility of a peak region

t containing the TF's motif, and the expression of a putative target gene g within a pre-fixed distance of peak t (Fig. 2a). Existing methods[12-14] screen for marginal associations either between the TF and the peak or between the peak and the target gene. However, three-way relationships may be complex, and true associations may be masked by population heterogeneity, as we demonstrate later. When marginal associations are masked, evidence for cooperation between the TF and the peak in the regulation of a gene can be inferred from partial associations: (i) with the peak open at a fixed accessibility, whether cells with higher TF expression have higher gene expression; and (ii) with the TF expression fixed at a value above a threshold, whether cells with higher peak accessibility have higher gene expression. To identify such conditional associations without making linearity assumptions on the marginal relationships, TRIPOD matches metacells by either their TF expressions or peak accessibilities (Fig. 2b): For each matched metacell pair, the variable being matched is controlled for, and differences between the pair in the other two variables are computed. Then, across pairs, the nonparametric Spearman's test is used to assess the association between difference in target gene expression ΔY_g and difference in the unmatched variable (i.e., ΔY_j if the cells were matched by X_t, or ΔX_t if the cells were matched by Y_j). We call this the "conditional level 1 test."

For illustration, consider the metacell denoted by the black point in Fig. 2b: If we were to match by peak accessibility, this metacell would be matched to the metacell colored in red. We would then compute ΔY_j, the difference between TF j expressions of the matched pair. If we were to match by TF expression, the black dot would be matched to the metacell in green, and we would compute ΔX_t, the difference in peak t accessibilities of the pair. In either case, we would compute ΔY_g, the difference in gene g expressions between the pair. We would then mask those metacell-pairs whose values, for the variable being matched, are too low (i.e., those pairs where the TF is off or the peak is closed). Then, ΔX_t or ΔY_j, together with ΔY_g, would be submitted for level 1 test.

Even stronger evidence for a regulatory trio could be claimed if the *degree* of association between the pairwise differences depends on the matched variable. For example, we would tend to believe that TF j binds to peak t to regulate gene g if, in cells with high expression of TF j, an increase in peak t accessibility yields a much larger

increase in gene g expression, as compared to in cells with low expression of TF j. One could screen for such interactions by matching by either TF j or peak t accessibility. TRIPOD screens for such interaction effects through a "conditional level 2 test", which assesses the association between ΔY_g and the product of the matched variable with the difference in the unmatched variable, after taking partial residuals on the difference in the unmatched variable.

For significant trios, TRIPOD further carries out a sampling-based influence analysis, where phenotypically contiguous sets of metacells are held out to measure their influence on the estimated coefficients. The corresponding cell types/states that lead to significant deviations from the null upon their removal have high influence scores, which can be used to identify cell types/states that drive a regulatory relationship.

To highlight the differences between TRIPOD and existing methods based on marginal associations, we show two canonical examples where the two approaches disagree. Fig. 2c outlines a significant trio detected by TRIPOD's level 2 testing, yet the marginal peak-gene and TF-gene associations were insignificant. It turns out that a subset of cells with high peak accessibility $\{X_t\}$ have close-to-zero TF expressions $\{Y_j\}$, and, meanwhile, another subset of cells with high TF expressions $\{Y_j\}$ have close-to-zero peak accessibilities $\{X_t\}$. In these cells, either the peak is closed, or the TF is not expressed, and this leads to the target gene not being expressed, which masks the marginal associations. The high peak accessibility and TF expression in these cells, which act through other regulatory trios, cancel out when we consider the interaction $\{X_t \times Y_j\}$, leading to a significant interaction term detected by TRIPOD. Conversely, Fig. 2d outlines another trio, whose marginal associations were significant, yet TRIPOD did not detect significant conditional associations from either level 1 or level 2 testing. In this case, with almost constant TF expression, the large difference in peak accessibility leads to a small difference in target gene expression. Meanwhile, the cells that drive the significantly positive correlation between $\{Y_g\}$ and $\{Y_j\}$ have almost zero values for $\{X_t\}$. Both observations suggest that this peak has little to do with the regulation of the target gene *FGL2* by this specific TF MAFK. Notably, we do not claim that the significantly linked peaks and TFs through marginal association are false positives, but rather this specific trio is insignificant (i.e., the peak and TF may act through other TF and peak, respectively).

In summary, TRIPOD puts peak-TF-gene trios into one unified model, complementing existing methods based on marginal associations and allowing for simultaneous identification of all three factors and prioritization of a different set of regulatory relationships.

TRIPOD identifies three-way regulatory relationships in PBMCs with orthogonal validations. We first applied TRIPOD to identify regulatory trios in the 10k PBMC dataset. Cell-type labels for this dataset were transferred from a recently released CITE-seq reference of 162,000 PBMC cells measured with 228 antibodies[14]. After quality control, we kept 7790 cells from 14 cell types pooled into 80 metacells, 103,755 peaks, 14,508 genes, and 342 TFs; the UMAP reduced dimensions are shown in Supplementary Fig. 1a. Distribution of the number of peaks 100kb/200kb upstream and downstream of the TSS per gene, as well as distribution of the number of motifs per peak, are shown in Supplementary Fig. 5.

Results from TRIPOD and marginal association tests overlap but exhibit substantial differences (Supplementary Fig. 6). The previous section showed example trios where the two frameworks disagree. As a proof of concept, we now illustrate two trios where the frameworks agree, identified by level 1 conditional testing (regulation of *CCR7* by LEF1, Fig. 3a) and level 2 interaction testing (regulation of *GNLY* by TBX21, Fig. 3b). From the influence analyses, TRIPOD identified B and T cells as the cell types where LEF1 regulates *CCR7*, and natural killer (NK) cells as where TBX21 regulates *GNLY*. These cell type-specific regulatory relationships are corroborated by motif's deviation scores using chromVAR[22] (Fig. 3a,b) and the enrichment of Tn5 integration events in the flanking regions using DNA footprinting analyses[12] (Supplementary Fig. 7). Unlike chromVar and DNA footprinting analyses, which only give genome-wide average enrichments, TRIPOD significantly enhances the resolution by identifying the specific *cis*-regulatory regions that the TFs bind for the regulation of target genes.

To our best knowledge, no experimental technique can directly validate three-way regulatory relationships at high resolution with high throughput. Therefore, we performed validation and benchmarking by harnessing existing databases and orthogonal

sequencing experiments that interrogate each pairwise relationship among the three factors (Table 1).

First, to validate the *cis*-linkage between peak region and target gene, we used the enhancer databases of blood and non-cancerous cells from FANTOM5[23] (from HACER[24]), 4DGenome[25] (from HACER[24]), and EnhancerAtlas 2.0[26], as well as *cis*-eQTLs in the whole blood reported by the GTEx consortium[9]. We collapsed TRIPOD's trio calls into peak-gene relationships and benchmarked against Signac's LinkPeaks[12] on single cells and marginal testing on metacells; for each target gene, we performed a hypergeometric test for enrichment of the peak-gene linkages in the regulatory databases and annotations (see Methods for details). For all four databases, TRIPOD's p-values for enrichment are substantially significant (Fig. 3c).

Second, to validate the TF-gene edge in the TRIPOD-identified trios, we referred to knockTF[7], a TF knockdown/knockout gene expression database, and hTFtarget[27], a database of known TF regulatory targets. Specifically, in knockTF, we found seven TF knockdown/knockout RNA-seq experiments in the peripheral blood category. For these TFs, we identified significantly linked genes by marginal association and by TRIPOD and found TRIPOD's results to have significantly higher precision and recall (Fig. 3d). For hTFtarget, we obtained, for each highly variable gene, its blood-specific TFs, and calculated the gene-specific precision-recall rates – TRIPOD is more sensitive compared to marginal association testing, although both suffered from inflated "false positives," which are likely due to the low sensitivity in the "gold-standard" *in silico* calls by hTFtarget (Fig. 3e).

Third, to validate the TF-peak edge representing TF binding to peak regions, in addition to the DNA footprinting analysis shown in Supplementary Fig. 7e, we downloaded from the Cistrome portal[28] non-cancerous ChIP-seq data from sorted human blood cells (B lymphocyte, T lymphocyte, and monocyte (Supplementary Table 2). The peaks identified by TRIPOD had a substantially higher percentage of overlap with the ChIP-seq peaks (Fig. 3f). In summary, existing databases and public data of different types from a wide range of studies extensively support each of the three pairwise links in the trios reported by TRIPOD, demonstrating its effectiveness in uncovering true regulatory relationships.

TRIPOD identifies known and novel putative regulatory relationships during mouse embryonic brain development. We next applied TRIPOD to single-cell multiomic data of 5k mouse embryonic brain cells at day 18 by 10X Genomics. The cell type labels were transferred from an independent scRNA-seq reference[29] using both Seurat V3[11] and SAVERCAT[30]. We kept 3,962 cells that had consistent transferred labels from seven major cell types: radial glia, neuroblast, GABAergic neuron, glutamatergic neuron, glioblast, oligodendrocyte, and Cajal−Retzius neuron (Supplementary Fig. 1b). We applied TRIPOD to 633 TFs, 1000 highly variable genes, and ATAC peaks 200kb up/downstream of the genes' TSSs.

First, we investigate a known regulatory trio involving target gene *Sox10*, TF Olig2 (binding motif CAGCTG), and a *cis*-regulatory element known as the U2 enhancer (chr15: 79201691-79201880 from mm10) that has been experimentally validated[31] The expression of *Sox10* and *Olig2* across the seven cell types, and the ATAC-seq profiles of the region containing the U2 enhancer and *Sox10*, are shown in Fig. 4a. This known regulatory trio was found to be conditional level 2 significant by TRIPOD (Fig. 4b). Importantly, the U2 enhancer resides in one of three *cis*-regulatory elements that were identified by TRIPOD to be enhancers for *Sox10* involving TF Olig2; all three significantly linked peaks were validated by Olig2 ChIP-seq data (Fig. 4a).

On the genome scale, the union of TRIPOD's level 1 and 2 tests gave a larger number of unique peak-gene pairs and TF-gene pairs than LinkPeaks[12] and marginal metacell association tests (Supplementary Fig. 8a). To evaluate results, we first examined whether the peak-gene links were enriched in previously reported enhancer-promoter chromatin contacts using PLAC-seq data of mouse fetal brain[32] (Table 1, Fig. 4c, Supplementary Fig. 8b). We observed that the regulatory links detected by both marginal association and TRIPOD showed significant enrichment in those supported by PLAC-seq (Fig. 4d, Supplementary Fig. 8c). Importantly, TRIPOD detected sets of trio relationships that were overlapping but distinct from the sets obtained by the marginal model, and a substantial fraction of the links identified by TRIPOD but not by the marginal method were validated by PLAC-seq. This suggests that TRIPOD identifies real regulatory relationships that complement those detected by existing methods. To validate the links between TFs

and peaks, we used publicly available ChIP-seq data of mouse embryonic brain for Olig2[33], Neurog2[34], Eomes[34], and Tbr1[35], TFs that play key roles in embryonic brain development (Table 1). TF binding peaks identified by TRIPOD were significantly enriched in the TF ChIP-seq peaks across all embryonic brain datasets; Olig2 ChIP-seq data of mature oligodendrocytes (mOL) serves as a negative control and has insignificant enrichment (Fig. 4e).

The validations and global benchmarking demonstrate TRIPOD's effectiveness in finding real regulatory relationships. Next, we focused on a set of TFs known to play essential roles during mouse embryonic brain development. Specifically, we chose Pax6, Neurog2, Eomes, Neurod1, and Tbr1, major TFs mediating glutamatergic neurogenesis[36], and Olig2, Sox10, Nkx2-2, Sox9, Nfia, and Ascl1, which initiate and mediate gliogenesis[37]; the known regulatory cascade is shown in Fig. 4f. TRIPOD's level 1 and level 2 testing successfully captured six out of the seven known regulatory links; interestingly, TRIPOD's results also suggest substantial crosstalk between the two cascades, where neurogenesis-specific TFs activate gliogenesis-specific TFs (Fig. 4g). ChIP-seq data of Neurog2, Eomes, and Tbr1 supported four of the crosstalk links: regulation of *Sox9* by Neurog2 and regulation of *Nfia* by Neurog2, Eomes, and Tbr1, respectively (Supplementary Fig. 9). These crosstalk links that were validated by ChIP-seq are highly significant by both marginal and conditional associations. Thus, we think it is highly plausible that neurogenesis TFs activate gliogenesis genes at day 18 of embryonic mouse brain development, which is exactly when the switch is being made from neurogenesis to gliogenesis. To our best knowledge, these possible links between neurogenesis and gliogenesis pathways have not been systematically explored and thus warrant future investigation. Finally, for each of the neurogenesis and gliogenesis TFs, we performed a gene ontology (GO) analysis of their significantly linked target genes using DAVID[38]; the enriched terms were largely consistent with the regulatory functions of the TFs during neurogenesis and gliogenesis (e.g., negative regulation of neuron differentiation and oligodendrocyte differentiation) (Fig. 4h).

So far, we have taken advantage of the cross-cell-type variation to identify the trio regulatory relationships. To dissect cell-type-specific regulation, we next applied the influence analysis framework (see Methods for details) to the significant trios involving

neurogenesis and gliogenesis TFs. For a given TF, the number of trios, for which a given cell type was influential (FDR < 0.01), is summarized in Fig. 4i, with details for specific example trios given in Supplementary Fig. 10. The analyses underpinned the cell types in which the transcriptional regulation was active, and, reassuringly, the neurogenesis and gliogenesis TFs have the most regulatory influence in neuroblasts and glioblasts, respectively. Additionally, Ascl1 is active in GABAergic neurons in addition to neuroblasts and glioblasts, consistent with its role as a GABAergic fate determinant[39]. Notably, the highly influential cell types that lead to the significant trios involving several neurogenesis-specific TFs include not only neuroblast but also glioblast, supporting our previous findings on the crosstalk between the two cascades. Overall, TRIPOD allows fine characterization of cell-type- and cell-state-specific functions of the TFs during neurogenesis and gliogenesis.

TRIPOD infers lineage-specific regulatory relationships in differentiating mouse hair follicle cells. As a last example, we applied TRIPOD to SHARE-seq[13] data (Supplementary Fig. 1c) of mouse hair follicle cells, consisting of four broadly defined cell types – transit-amplifying cells (TAC), inner root sheath (IRS), hair shaft, and medulla cells – along a differentiation trajectory. The cell-type labels were curated based on marker genes, TF motifs, and ATAC peaks from the original publication[13]; pseudotime was inferred using Palantir[40] and overlaid on the cisTopic[41] reduced dimensions of the ATAC domain. Cells were partitioned using both the pseudotime and the UMAP coordinates to construct metacells (Fig. 5a). Due to the low RNA coverage (Supplementary Fig. 4), we focused on 222 highly-expressed TFs, 794 highly expressed genes reported to have more than ten linked *cis*-regulatory peaks[13], and peaks 100kb up/downstream of the genes' TSSs.

For validation, we used H3K4me1 and H3K27ac ChIP-seq data from an isolated mouse TAC population[42] (Table 1). H3K4me1 and H3K27ac are marks for active enhancers and are used to benchmark TRIPOD's linked peaks against previously reported domains of regulatory chromatin (DORCs)[13], as well as randomly sampled peaks. The linked peaks by TRIPOD have higher scores for both H3K4me1 and H3K27ac, than DORCs, the latter identified through marginal associations (Fig. 5b). To further validate

the regulatory effects of the linked peaks, we obtained previously characterized super-enhancers (SEs) in mouse TACs[42]. Target genes of the 381 SEs were assigned based on the gene's proximity to the SE, as well as the correlation between loss of the SE and loss of the gene transcription[42]. TRIPOD was able to successfully recapitulate the SE regions for the genes considered, with four examples shown in Fig. 5c, where significantly linked peaks mostly reside in the SEs.

To demonstrate, Fig. 5d shows regulatory trios that are specific to the IRS lineage, the hair shaft lineage, and the medulla lineage. These trios also show significant pairwise marginal associations (Fig. 5e), lending confidence that they are real. The cell types where the regulation happens are identified by influence analysis, for which the p-values are smoothed along the differentiation trajectory and overlaid on the UMAP embedding (Fig. 5f). DNA footprinting analyses surveyed the enrichment of Tn5 integration events surrounding the corresponding motif sites and showed cell-type-specific enrichment (Fig. 5g), corroborating TRIPOD's results.

Discussion

We have considered the detection of regulatory trios, consisting of a TF binding to a regulatory region to activate or repress the transcription of a nearby gene, using single-cell RNA and ATAC multiomic sequencing data. The presented method, TRIPOD, is a new nonparametric approach that goes beyond marginal relationships to detect conditional associations and interactions on peak-TF-gene trios. We applied TRIPOD to three single-cell multiomic datasets from different species and protocols with extensive validations and benchmarks. We started our analyses with predicting gene expression from both peak accessibility and TF expression. Supervised frameworks have been proposed to predict gene expression from DNA accessibility[43], and vice versa[44], using matched bulk transcriptomic and epigenomic sequencing data. Blatti *et al.*[45] showed that joint analysis of DNA accessibility, gene expression, and TF motif binding specificity allows reasonably good prediction of TF binding as measured by ChIP-seq. However, none of these methods incorporate TF expression. By selecting peaks near the genes' TSSs and TFs with high motif scores in the selected peaks, we constructed biologically

meaningful peak-TF pairs as predictors and showed that such a mechanistic model significantly boosts the prediction accuracy of gene expression.

We next considered the detection and significance assessment for individual peak-TF-gene trios, comprehensively comparing our detections with those made by tissue- and cell-type-matched PLAC-seq and ChIP-seq experiments, by *cis*-eQTL and TF knockdown/knockout studies, and by those recorded in the main enhancer databases. Our current study is limited in several ways. A study in *Drosophila*[46] modeled motif binding specificities and chromatin accessibilities in bulk RNA and ATAC sequencing data to predict the cooperative binding of pairs of TFs, using *in vitro* protein-protein binding experiments for validation. The detection of synergies between multiple TFs and peaks on the genome-wide scale and in a cell-type-specific manner needs further investigation. Additionally, while we have not differentiated between positive and negative regulation, TRIPOD reports both types of relationships and categorizes them by sign. While we describe the trios with a positive sign to be enhancers, it is not clear how to interpret the trios with negative signs, the latter having lower overlap with other benchmarking datasets. Transcription activation and repression have been active research areas in biology, with a lot yet unknown[47]. TRIPOD's results provide potential targets for experimental follow-up and detailed characterization.

Our analysis focused on three datasets where the RNA and ATAC modalities have sufficient depths of coverage. For the SHARE-seq data, the sequencing depth for RNA is very low, and thus we focused only on highly expressed genes and TFs (Fig. 5). For SNARE-seq data, whose coverage in both modalities is even lower, we focused on prediction models and not trio detection, where we saw only marginal improvement beyond existing methods[19] (Supplementary Fig. 2). For data where the coverage is even lower, e.g., PAIRED-seq, cross-modality metacells could not be stably formed, making such analyses impossible (Supplementary Table 1, Supplementary Fig. 4). With rapidly increasing sequencing capacity and technological advancement, TRIPOD, applied to more cells sequenced at higher depth, can uncover novel regulatory relationships at a finer resolution.

Methods

Data input and construction of metacells. Denote X_{it} as the peak accessibility for peak t $(1 \leq t \leq T)$ in cell i $(1 \leq i \leq N)$, Y_{ig} as the gene expression for gene g $(1 \leq g \leq G)$, and Y_{ij} as the TF expression for TF j $(1 \leq j \leq M)$. The TF expression matrix is a subset of the gene expression matrix, and for single-cell multiomic data, the cell entries are matched. To mitigate the effect of ATAC sparsity[48] and RNA expression stochasticity[49], as a first step, TRIPOD performs cell-wise smoothing by pooling similar cells into "metacells." This, by default, is performed using the weighted-nearest neighbor method by Seurat V4[14] to jointly reduce dimension and identify cell clusters/states across different modalities. In practice, the metacells can also be inferred using one modality – for example, RNA may better separate the different cell types[29], and in other cases, chromatin accessibility may prime cells for differentiation[13]. To account for peaks overlapping with other genes (Supplementary Fig. 5b), TRIPOD has the option to either remove the overlapped peaks or to adjust the peak accessibilities by the expressions of the overlapped genes, in a similar fashion to MAESTRO[17]. Library size is adjusted for both the RNA and ATAC domain by dividing all counts by a metacell-specific size factor (total read counts divided by 10^6).

RNA prediction by TF expression and peak accessibility. To predict RNA from ATAC, Signac[12] and Cicero[16] take the sum of peak accessibilities in gene bodies and promoter regions to construct a pseudo-gene activity matrix: $\hat{Y}_{ig} = \sum_{t \in E_g} X_{it}$, where E_g is the set of peaks within gene bodies and upstream regions of TSSs. Instead of directly taking the sum, MAESTRO[17] adopts a "regulatory potential" model by taking the weighted sum of accessibilities across all nearby peaks: $\hat{Y}_{ig} = \sum_{t \in E_g} w_t^g X_{it}$, with weights $\{w_t^g\}$ pre-calculated based on existing gene annotations. Specifically, the method weighs peaks by exponential decay from TSS, sums all peaks on the given gene exons as if they are on the TSS, normalizes the sum by total exon lengths, and excludes the peaks from promoters and exons of nearby genes. The strategy to take the unweighted/weighted sum of accessibility as a proxy for expression has been adopted to align the RNA and ATAC modalities when scRNA-seq and scATAC-seq are sequenced in parallel from the same cell population but not the same cells[11]. For single-cell multiomic data, sci-CAR[18] performs

feature selection to identify *cis*-linked peaks via a LASSO regression: $Y_{ig} \sim \sum_{t \in E_g} \beta_t^g X_{it}$, where an L1 regularization is imposed on β_t^g. Compared to MAESTRO, which pre-fixes the weights $\{w_t^g\}$, $\{\beta_t^g\}$ are estimated from the data by regressing RNA against matched ATAC data. What we propose is a feature selection model involving both peak accessibility and TF expression: $Y_{ig} \sim \sum_{t \in E_g} \sum_{j \in f_t} \beta_{tj}^g X_{it} Y_{ij}$, where f_t contains the set of TFs with high-scoring binding motifs in peak t inferred from the JASPAR database[50].

TRIPOD model and trio regulatory relationship. For a given target gene g, a peak t within a window centered at the gene's TSS, and a TF j whose binding motif is high-scoring in the peak, TRIPOD infers the relationship between a regulatory trio (t, j, g). TRIPOD focuses on one trio at a time and goes beyond the marginal associations to characterize the function $Y_g = f(X_t, Y_j)$. In what follows, we first describe TRIPOD's matching-based nonparametric approach and then describe a linear parametric approach, followed by a discussion on the connections and contrasts between the two approaches.

For each cell i whose TF expression is above a threshold δ (we only carry out testing in cells that express the TF), we carry out a minimum distance pairwise cross-match based on $\{Y_{ij} | Y_{ij} > \delta\}$. Let $\{(i_p, i_{p^*})\}$ be the optimal matching, after throwing away those pairs that have $|Y_{i_p j} - Y_{i_{p^*} j}| > e$. For each pair p, i_p and i_{p^*} are two metacells with matched TF expression, for which we now observe two, possibly different, values $\{X_{i_p t}, X_{i_{p^*} t}\}$ for peak t, as well as two corresponding values $\{Y_{i_p g}, Y_{i_{p^*} g}\}$ for gene g. We then compute the following auxiliary differentials within each pair:

$$\Delta X_{pt} = X_{i_p t} - X_{i_{p^*} t},$$

$$\Delta Y_{pg} = Y_{i_p g} - Y_{i_{p^*} g},$$

as well as

$$\bar{Y}_{pj} = (Y_{i_p j} + Y_{i_{p^*} j})/2.$$

For level 1 testing of conditional association, we estimate $\hat{r}_t^g = \rho(\Delta X_{pt}, \Delta Y_{pg})$, where ρ is Spearman correlation, and test $H_1: r_t^g = 0$. For level 2 testing of interaction, we perform a regression $\Delta Y_{pg} = \alpha \Delta X_{pt} + \gamma \bar{Y}_{pj} \times \Delta X_{pt}$, set $\hat{\gamma}_{tj}$ to be the least-squares solution for γ, and test $H_2: \gamma_{tj} = 0$. For visualization of the model fitting, we take the partial residuals of ΔY_{pg}

and $\bar{Y}_{pj} \times \Delta X_{pt}$ on ΔX_{pt}, respectively. Note that even though TF expression is not included in the model as a main term, it is controlled for (and not just in the linear sense) by the matching. Similarly, we can also perform this procedure matching by peak accessibility. As a summary, for level 1 testing of conditional association, we have:

$$\text{Match by } Y_j,\ \alpha = \rho\left(\Delta Y_g, \Delta X_t\right),$$

$$\text{Match by } X_t,\ \beta = \rho\left(\Delta Y_g,\ \Delta Y_j\right).$$

For level 2 testing of (TF expression)×(peak accessibility) interaction effects, we have:

$$\text{Match by } Y_j,\ \Delta Y_g = \alpha^* \Delta X_t + \gamma_1\left(\bar{Y}_j \times \Delta X_t\right),$$

$$\text{Match by } X_t,\ \Delta Y_g = \beta^* \Delta Y_j + \gamma_2\left(\bar{X}_t \times \Delta Y_j\right).$$

To test for the conditional associations and interactions, we can also use apply a parametric method, such as multiple linear regression:

$$Y_g = \mu + \alpha_L X_t + \beta_L Y_j,$$

$$Y_g = \mu + \alpha_L^* X_t + \beta_L^* Y_j + \gamma_L X_t Y_j.$$

See Supplementary Fig. 11 for linear testing results for trios shown in Fig. 3 and Fig. 5. The estimated coefficients from the nonparametric and parametric methods are correlated on the global scale (Supplementary Fig. 12), and their interpretations are similar: α and α_L estimate the change in gene expression per change in peak accessibility, fixing TF expression; β and β_L estimate the change in gene expression per change in TF expression, fixing peak accessibility; γ_1 and γ_L measure how the change in gene expression per change in peak accessibility at each fixed TF expression relies on the TF expression; γ_2 and γ_L measure how the change in gene expression per change in TF expression at each fixed peak accessibility relies on the peak accessibility. However, the underlying models and assumptions are different. Matching controls for not just the linear variation in the matched variable, but also any nonlinear variation. This contrasts with adding the variable as a covariate in the linear regression, where we simply remove linear dependence. The main motivation for using the matching model above is our reluctance to assume the simple linear relationship. Additionally, we use the rank-based Spearman correlation, which will not be driven by outliers – a "bulk" association between ranks is needed for significance. Thus, the nonparametric model of TRIPOD is more stringent (Supplementary Fig. 13) and more robust to outliers.

Identifying regulatory cell type(s) and cell state(s). For the significant trios detected by TRIPOD, we next seek to identify the underlying regulatory cell type(s). Specifically, we carry out a cell-type-specific influence analysis to identify cell types that are highly influential in driving the significance of the trio. Traditional approaches (e.g., the Cook's distance and the DFFITs) delete observations one at a time, refit the model on remaining observations, and measure the difference in the predicted value from the full model and that from when the point is left out. While they can be readily adopted to detect "influential" metacells one at a time (Supplementary Fig. 7a,b), these methods do not adjust for the degree of freedom properly when deleting different numbers of metacells from different cell types. That is, they do not account for the different numbers of observations that are simultaneously deleted. Additionally, both methods adopt a thresholding approach to determine significance, without returning p-values that are necessary for multiple testing correction. We, therefore, develop a sampling-based approach to directly test for the influence of multiple metacells and to return p-values (Supplementary Fig. 7c).

Here, we focus on the linear model for its ease of computation: $\hat{Y}_g = \hat{\mu} + \hat{\alpha}X_t + \hat{\beta}Y_j + \hat{\gamma}X_tY_j$. Given a set of observations $I = \{i: i\text{th metacell belongs to a cell type}\}$, we remove these metacells, fit the regression model, and make predictions: $\hat{Y}_g^{(I)} = \hat{\mu}^{(I)} + \hat{\alpha}^{(I)}X_t + \hat{\beta}^{(I)}Y_j + \hat{\gamma}^{(I)}X_tY_j$. The test statistics are the difference in the fitted gene expressions $\left|\hat{Y}_g - \hat{Y}_g^{(I)}\right|$. We generate the null distribution via sampling. Specifically, within each sampling iteration, we sample without replacement the same number of metacells, denoted as a set of I^*, delete these observations, and refit the regression model on the remaining observations: $\hat{Y}_g^{(I^*)} = \hat{\mu}^{(I^*)} + \hat{\alpha}^{(I^*)}X_t + \hat{\beta}^{(I^*)}Y_j + \hat{\gamma}^{(I^*)}X_tY_j$. The p-value is computed across K sampling iterations as $p_{Y_g} = \sum_{I^*} 1\left(\sum|\hat{Y}_g - \hat{Y}_g^{(I)}| \geq \sum|\hat{Y}_g - \hat{Y}_g^{(I^*)}|\right)/K$, where 1() is the indicator function. In addition to testing each cell type separately, the framework can be extended to test for the influence of cell-type groups. For example, in Fig. 3a,b, we reconstruct the cell-type hierarchy using expression levels of highly variable genes from the RNA domain and carry out the aforementioned testing scheme at each split for its descendent cell types in the hierarchical structure.

For transient cell states, TRIPOD first identifies the neighbors of each metacell along the trajectory and then carries out metacell-specific testing by simultaneously removing each metacell and its neighbors using the framework described above. The resulting p-values are, therefore, smoothed and can be visualized in the UMAP plot, as shown in Fig. 5f and Supplementary Fig. 10, to identify the underlying branches/segments that are key in defining the significant regulatory trio. This approach can be directly applied to cells with branching dynamics without the need to isolate cell subsets or to identify cell types.

Validation resources and strategies. Resources for validating the trio regulatory relationships are summarized in Table 1. To validate the peak-gene relationships, we referred to existing enhancer databases: FANTOM5[23] links enhancers and genes based on enhancer RNA expression; 4DGenome[25] links enhancers and genes based on physical interactions using chromatin-looping data including 3C, 4C, 5C, ChIA-PET, and Hi-C; EnhancerAtlas 2.0[26] reports enhancers using 12 high-throughput experimental methods including H3K4me1/H3K27ac ChIP-seq, DNase-seq, ATAC-seq, and GRO-seq. We only focused on blood and non-cancerous cells from these databases (Fig. 3c). A list of *cis*-eQTLs within the whole blood mapped in European-American subjects was downloaded from the GTEx consortium[9] (Fig. 3c). For the mouse embryonic brain dataset, we additionally adopted H3K4me3-mediated PLAC-seq data[32], which reported enhancer-promoter chromatin contacts mapped in mouse fetal forebrain (Fig. 4c,d). For the mouse skin dataset, we adopted TAC-specific ChIP-seq data of H3K4me1 and H3K27ac[42], both of which are histone marks for active enhancers (Fig. 5b); we also obtained previously reported super-enhancers in mouse TACs from *in vivo* studies[42] (Fig. 5c). Genomic coordinates were lifted over from mm9 to mm10 when necessary.

To validate the TF-gene relationships, we utilized the knockTF[7] and the hTFtarget[27] databases. knockTF interrogates the changes in gene expression profiles in TF knockdown/knockout experiments to link the TFs to their target genes in a tissue- or cell-type-specific manner. We downloaded 12 experiments, corresponding to 12 TFs (BCL11A, ELK1, GATA3, JUN, MAF, MYB, NFATC3, NFKB1, STAT3, STAT6, TAL1, and ZNF148) in the peripheral blood category, and focused on seven TFs that have at least

one linked gene by any model benchmarked (Fig. 3d). hTFtarget computationally predicts TF-gene relationships using ChIP-seq data, and we manually downloaded the TFs associated with each of the top 100 highly variable genes in the blood tissue (Fig. 3e).

To validate the peak-TF relationships, we downloaded non-cancerous cell-type-specific ChIP-seq data of human blood (B lymphocyte, T lymphocyte, and monocyte) from the Cistrome[28] portal for the PBMC data (Fig. 3f, Supplementary Table 2), and ChIP-seq data of Olig2[33], Neurog2[34], Eomes[34], and Tbr1[35] for the mouse embryonic brain data. The Olig2 ChIP-seq data were generated in three types of rat cells: data from oligodendrocyte precursor cells (OPC) and immature oligodendrocytes (iOL) were used for validation, while data from mature oligodendrocytes (mOL) serve as a negative control[33]. Genomic coordinates were converted from rn4 to mm10. The Neurog2 and Eomes ChIP-seq data were generated in mouse embryonic cerebral cortices at day 14.5[34]; the Tbr1 ChIP-seq data was generated in the whole cortex dissected from embryos at day 15.5[35]. In addition, DNA footprinting signatures were corrected for Tn5 sequence insertion bias and stratified by cell types using the Signac package and can be used to validate the identified TFs/motifs in a cell-type-specific manner (Fig. 5g, Supplementary Fig. 7e).

For peak enrichment analysis compared to the existing enhancers, *cis*-eQTLs, and enhancer-promoter contacts, we carried out a hypergeometric test as follows. Let k be the number of significantly linked peaks, q be the number of significantly linked peaks that overlap with annotations (e.g., annotated enhancers), m be the number of peaks that overlap with the annotations, and n be the number of peaks that do not overlap with annotations. The p-value of enrichment is derived from the hypergeometric distribution using the cumulative distribution function, coded as phyper(q, m, n, k, lower.tail=F) in R. We used this hypothesis testing framework to validate and benchmark the reported peak-gene links, with results shown in Fig. 3c and Fig. 4d. A similar analysis was carried out to test for peak enrichment in TF binding sites by ChIP-seq, thus validating the peak-TF relationships (Fig. 4e).

Data availability

This study analyzed existing and publicly available single-cell RNA and ATAC multiomic data. 10X Genomics single-cell multiomic datasets of PBMC (10k and 3k) and mouse

embryonic brain were downloaded https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets. SNARE-seq data of adult mouse brain and SHARE-seq data of mouse skin are available from the Gene Expression Omnibus (GEO) database with accession numbers GSE126074 and GSE140203. A detailed data summary is provided in Supplementary Table 1. Validation resources based on existing databases and high-throughput sequencing data are summarized in Table 1 and Supplementary Table 2.

Code availability

TRIPOD is compiled as an open-source R package available at https://github.com/yharigaya/TRIPOD. Scripts used for analyses carried out in this paper are deposited in the GitHub repository.

Acknowledgments

This work was supported by the National Institutes of Health (NIH) grant R35 GM133712 (to C.Z.), R01 HG006137 (to N.R.Z), R01 GM125301 (to N.R.Z.), and R35 GM138342 (to Y.J.). The authors thank Dr. Sai Ma for support and guidance on the SHARE-seq data, Manas Tiwari for help on accessing the hTFtarget database, and Drs. Yun Li, Michael Love, Li Qian, and Jason Stein for helpful discussions and comments.

Author contributions

N.R.Z. and Y.J. initiated and envisioned the study. Y.H., N.R.Z., and Y.J. formulated the model, developed the algorithm, and performed data analysis. Z.Z. processed reference datasets and performed cell-type label transfer. H.Z. and C.Z. provided support on validation, offered consultation, and contributed to result interpretation. Y.H., N.R.Z., and Y.J. wrote the manuscript, which was read and approved by all authors.

Competing Interests

The authors declare no competing interests.

Figure Legends

Fig. 1 | Interaction between TF expression and peak accessibility improves RNA prediction accuracy. a, UMAP embedding of 10x Genomics PBMC (left), 10x Genomics embryonic mouse brain (center), and SHARE-seq mouse skin (right) cells from single-cell RNA and ATAC multiomic sequencing. Cell-type labels were transferred from existing single-cell references or curated based on marker genes, motifs, and peaks; metacells were constructed to mitigate sparsity and stochasticity. **b**, Genome-wide distributions of Pearson correlations between observed and leave-one-out predicted RNA expression levels, with varying window sizes. **c**, Predicted and observed RNA expression levels for highly variable genes, *CCR7*, *Adamts6*, and *Ano7*, from the three datasets, respectively.

Fig. 2 | TRIPOD infers peak-TF-gene trio regulatory relationships using single-cell multiomic data. a, Data input and schematic on a peak-TF-gene trio. **b**, Overview of TRIPOD for inferring regulatory relationships. TRIPOD complements existing methods based on marginal associations by identifying conditional associations through matching by TF expression or peak accessibility. **c**, An example trio identified by TRIPOD, but not by the marginal associations due to heterogeneity of cell-type-specific regulations. **d**, An example trio identified by the marginal associations, but not by TRIPOD. The peak and TF are significantly linked to the gene, yet they act through other TF and peak, and thus the regulatory trio is insignificant. The points represent metacells (left two panels) and pairs of matched metacells (right two panels). Genomic coordinates for the peaks are from hg38.

Fig. 3 | TRIPOD identified trio regulatory relationships in PBMC single-cell multiomic dataset. a-b, Example trios identified by TRIPOD. Violin plots show cell-type-specific distributions of gene expression, peak accessibility, and TF expression. Scatterplots show TRIPOD's level 1 and level 2 testing, respectively. Inner and outer circles around the points are color-coded based on the cell types of the matched metacells. Hierarchical clustering is performed on RNA expression levels of highly variable genes. Red/gray circles indicate whether removal of the corresponding branches of metacells significantly changes the model fitting; crosses indicate that removal of the groups of

metacells resulted in inestimable coefficients. Genomic coordinates for the peaks are from hg38. **c**, Peak-gene validation based on enhancer databases (FANTOM5, 4DGenome, and EnhancerAtlas) and tissue-specific *cis*-eQTL data from the GTEx Consortium. Box plots show distributions of p-values from gene-specific hypergeometric tests. **d**, TF-gene validation based on lists of TF-gene pairs from the knockTF database. **e**, Precision and recall rates for TF-gene pairs using ground truths from the hTFtarget database. **f**, Peak-TF validation based on eight cell-type-specific TF ChIP-seq datasets (B lymphocytes, monocytes, and T lymphocytes). Percentages of significantly linked peaks and all peaks that overlap with the ChIP-seq peaks are shown.

Fig. 4 | TRIPOD identified known and novel regulatory relationships during mouse neurogenesis and gliogenesis. a, TRIPOD identified a previously reported regulatory trio with gene *Sox10*, TF Olig2, and *cis*-regulatory U2 element. TRIPOD identified two additional linked peaks; all three *cis*-regulatory elements were validated by Olig2 ChIP-seq data. **b**, TRIPOD's level 2 testing matching peak accessibility for the *Sox10* gene, the Olig2 TF, and the U2 enhancer. **c**, Venn diagram of the number of peak-gene pairs captured by PLAC-seq, the marginal model, and the union set of TRIPOD's level 1 and level 2 testing matching TF expression. **d**, Enrichment of peak-gene pairs captured by LinkPeaks, marginal association, and TRIPOD in enhancer-promoter contacts by PLAC-seq. **e**, Peak-TF validation by ChIP-seq data. Olig2 ChIP-seq data of precursor/immature oligodendrocytes (OPC/iOL) were used for validation; data from mature oligodendrocytes (mOL) served as a negative control. **f**, A schematic of well-characterized TF regulatory cascades during neurogenesis and gliogenesis. **g**, Trio examples from known regulatory relationships, as well as from crosstalks supported by ChIP-seq data, captured by TRIPOD. **h**, GO analysis of putative target genes of the neurogenesis and gliogenesis TFs. The number of TRIPOD-identified target genes in the GO categories is shown. The background heatmap shows negative log p-values (FDR adjusted) from hypergeometric tests examining enrichment of ATAC peaks in ChIP-seq peaks. **i**, Bar plots showing the number of putative cell-type-specific trios mediated by the neurogenesis- and gliogenesis-specific TFs.

Fig. 5 | TRIPOD identified regulatory relationships in mouse hair follicles with transient cell states. a, UMAP embedding of hair follicle cells from the mouse skin data. Cells are colored by cell types (TAC, IRS, hair shaft, and medulla) and pseudotime. **b**, H3K4me1 and H3K27ac ChIP-seq scores for linked peaks identified by TRIPOD, DORCs (regulatory domains identified by gene-peak correlations), and randomly sampled peaks. **c**, TRIPOD's linked peaks for four representative genes were significantly enriched in previously annotated super-enhancers in the mouse TAC population. **d**, Trios detected by TRIPOD that were active in IRS (top), medulla (middle), and hair shaft (bottom), respectively. **e**, Dot plots of gene expressions, peak accessibilities, and TF expressions across different cell types. **f**, Influence analyses identified segments along the differentiation trajectory where the regulation took effect. The colors in the UMAP embedding correspond to the smoothed p-values from a sampling-based approach. **g**, DNA footprinting assays showed cell-type-specific enrichments of Tn5 integration events. The findings were consistent with those from the influence analyses.

Table 1 | Resources for validating peak-TF-gene regulatory relationship. While there is no existing experimental approach to validate all three factors in a trio at high resolution with high throughput, we resort to existing databases and orthogonal sequencing data to validate peak-gene, peak-TF, and TF-gene pairs, completing the loop.

References

1. Gasperini, M., Tome, J.M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat Rev Genet* **21**, 292-310 (2020).
2. Hobert, O. Gene regulation by transcription factors and microRNAs. *Science* **319**, 1785-1786 (2008).
3. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-628 (2008).
4. Johnson, D.S., Mortazavi, A., Myers, R.M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497-1502 (2007).
5. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213-1218 (2013).
6. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293 (2009).
7. Feng, C. et al. KnockTF: a comprehensive human gene expression profile database with knockdown/knockout of transcription factors. *Nucleic Acids Res* **48**, D93-D100 (2020).

8. Tompa, M. et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* **23**, 137-144 (2005).
9. Consortium, G. Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213 (2017).
10. Kim-Hellmuth, S. et al. Cell type-specific genetic regulation of gene expression across human tissues. *Science* **369** (2020).
11. Stuart, T. et al. Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821 (2019).
12. Stuart, T., Srivastava, A., Lareau, C. & Satija, R. Multimodal single-cell chromatin analysis with Signac. *bioRxiv*, 2020.2011.2009.373613 (2020).
13. Ma, S. et al. Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**, 1103-1116 e1120 (2020).
14. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587 e3529 (2021).
15. Duren, Z., Chen, X., Jiang, R., Wang, Y. & Wong, W.H. Modeling gene regulation from paired expression and chromatin accessibility data. *Proc Natl Acad Sci U S A* **114**, E4914-E4923 (2017).
16. Pliner, H.A. et al. Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol Cell* **71**, 858-871 e858 (2018).
17. Wang, C. et al. Integrative analyses of single-cell transcriptome and regulome using MAESTRO. *Genome Biol* **21**, 198 (2020).
18. Cao, J. et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380-1385 (2018).
19. Chen, S., Lake, B.B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat Biotechnol* **37**, 1452-1457 (2019).
20. Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* (2018).
21. Taylor, J. & Tibshirani, R.J. Statistical learning and selective inference. *Proc Natl Acad Sci U S A* **112**, 7629-7634 (2015).
22. Schep, A.N., Wu, B., Buenrostro, J.D. & Greenleaf, W.J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods* **14**, 975-978 (2017).
23. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455-461 (2014).
24. Wang, J. et al. HACER: an atlas of human active enhancers to interpret regulatory variants. *Nucleic Acids Res* **47**, D106-D112 (2019).
25. Teng, L., He, B., Wang, J. & Tan, K. 4DGenome: a comprehensive database of chromatin interactions. *Bioinformatics* **32**, 2727 (2016).
26. Gao, T. & Qian, J. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res* **48**, D58-D64 (2020).
27. Zhang, Q. et al. hTFtarget: A Comprehensive Database for Regulations of Human Transcription Factors and Their Targets. *Genomics Proteomics Bioinformatics* **18**, 120-128 (2020).
28. Mei, S. et al. Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res* **45**, D658-D662 (2017).
29. La Manno, G. et al. Molecular architecture of the developing mouse brain. *Nature* **596**, 92-96 (2021).
30. Huang, M., Zhang, Z. & Zhang, N.R. Dimension reduction and denoising of single-cell RNA sequencing data in the presence of observed confounding variables. *bioRxiv*, 2020.2008.2003.234765 (2020).
31. Kuspert, M., Hammer, A., Bosl, M.R. & Wegner, M. Olig2 regulates Sox10 expression in oligodendrocyte precursors through an evolutionary conserved distal enhancer. *Nucleic Acids Res* **39**, 1280-1293 (2011).

32. Zhu, C. et al. An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat Struct Mol Biol* **26**, 1063-1070 (2019).
33. Yu, Y. et al. Olig2 targets chromatin remodelers to enhancers to initiate oligodendrocyte differentiation. *Cell* **152**, 248-261 (2013).
34. Sessa, A. et al. The Tbr2 Molecular Network Controls Cortical Neuronal Differentiation Through Complementary Genetic and Epigenetic Pathways. *Cereb Cortex* **27**, 5715 (2017).
35. Notwell, J.H. et al. TBR1 regulates autism risk genes in the developing neocortex. *Genome Res* **26**, 1013-1022 (2016).
36. Mira, H. & Morante, J. Neurogenesis From Embryo to Adult - Lessons From Flies and Mice. *Front Cell Dev Biol* **8**, 533 (2020).
37. Emery, B. & Lu, Q.R. Transcriptional and Epigenetic Regulation of Oligodendrocyte Development and Myelination in the Central Nervous System. *Cold Spring Harb Perspect Biol* **7**, a020461 (2015).
38. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).
39. Achim, K., Salminen, M. & Partanen, J. Mechanisms regulating GABAergic neuron development. *Cell Mol Life Sci* **71**, 1395-1415 (2014).
40. Setty, M. et al. Characterization of cell fate probabilities in single-cell data with Palantir. *Nat Biotechnol* **37**, 451-460 (2019).
41. Bravo Gonzalez-Blas, C. et al. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat Methods* **16**, 397-400 (2019).
42. Adam, R.C. et al. Pioneer factors govern super-enhancer dynamics in stem cell plasticity and lineage choice. *Nature* **521**, 366-370 (2015).
43. Natarajan, A., Yardimci, G.G., Sheffield, N.C., Crawford, G.E. & Ohler, U. Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res* **22**, 1711-1722 (2012).
44. Zhou, W. et al. Genome-wide prediction of DNase I hypersensitivity using gene expression. *Nat Commun* **8**, 1038 (2017).
45. Blatti, C., Kazemian, M., Wolfe, S., Brodsky, M. & Sinha, S. Integrating motif, DNA accessibility and gene expression data to build regulatory maps in an organism. *Nucleic Acids Res* **43**, 3998-4012 (2015).
46. Kazemian, M., Pham, H., Wolfe, S.A., Brodsky, M.H. & Sinha, S. Widespread evidence of cooperative DNA binding by transcription factors in Drosophila development. *Nucleic Acids Res* **41**, 8237-8252 (2013).
47. Panigrahi, A. & O'Malley, B.W. Mechanisms of enhancer action: the known and the unknown. *Genome Biol* **22**, 108 (2021).
48. Urrutia, E., Chen, L., Zhou, H. & Jiang, Y. Destin: toolkit for single-cell analysis of chromatin accessibility. *Bioinformatics* **35**, 3818-3820 (2019).
49. Jiang, Y., Zhang, N.R. & Li, M. SCALE: modeling allele-specific gene expression by single-cell RNA sequencing. *Genome Biol* **18**, 74 (2017).
50. Fornes, O. et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **48**, D87-D92 (2020).

Figures

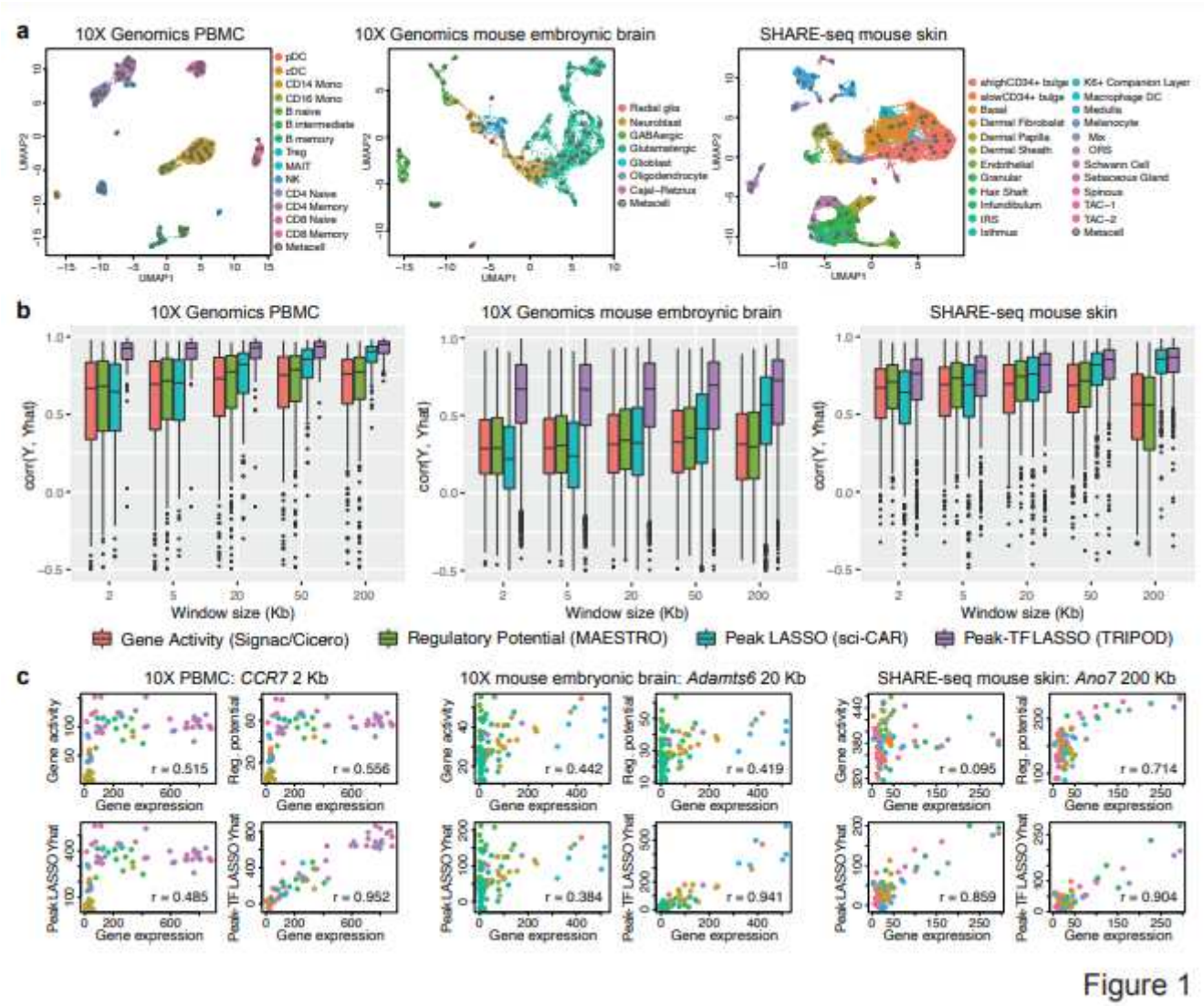

Figure 1

Figure 1

Interaction between TF expression and peak accessibility improves RNA prediction accuracy.

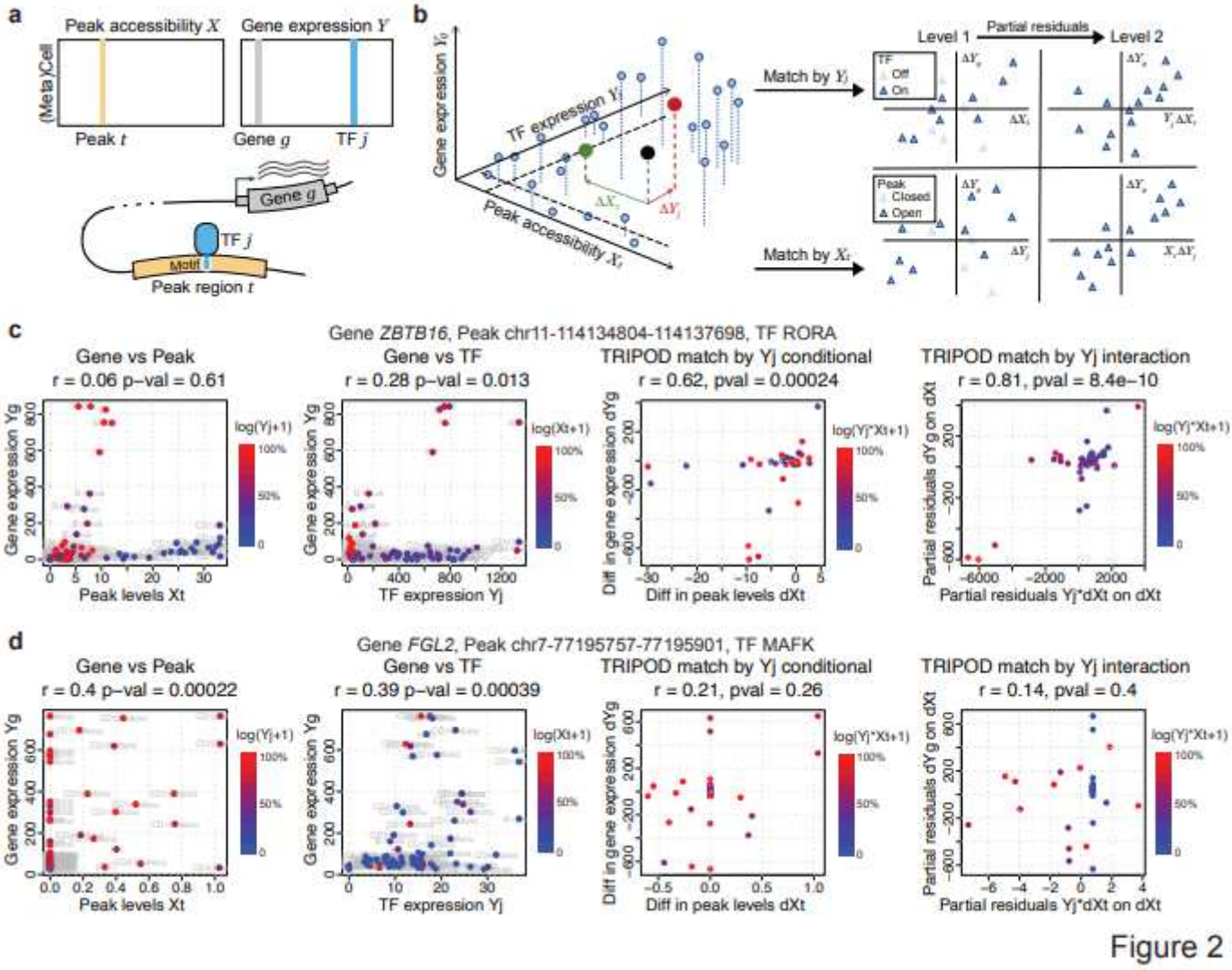

Figure 2

TRIPOD infers peak-TF-gene trio regulatory relationships using single-cell multiomic data.

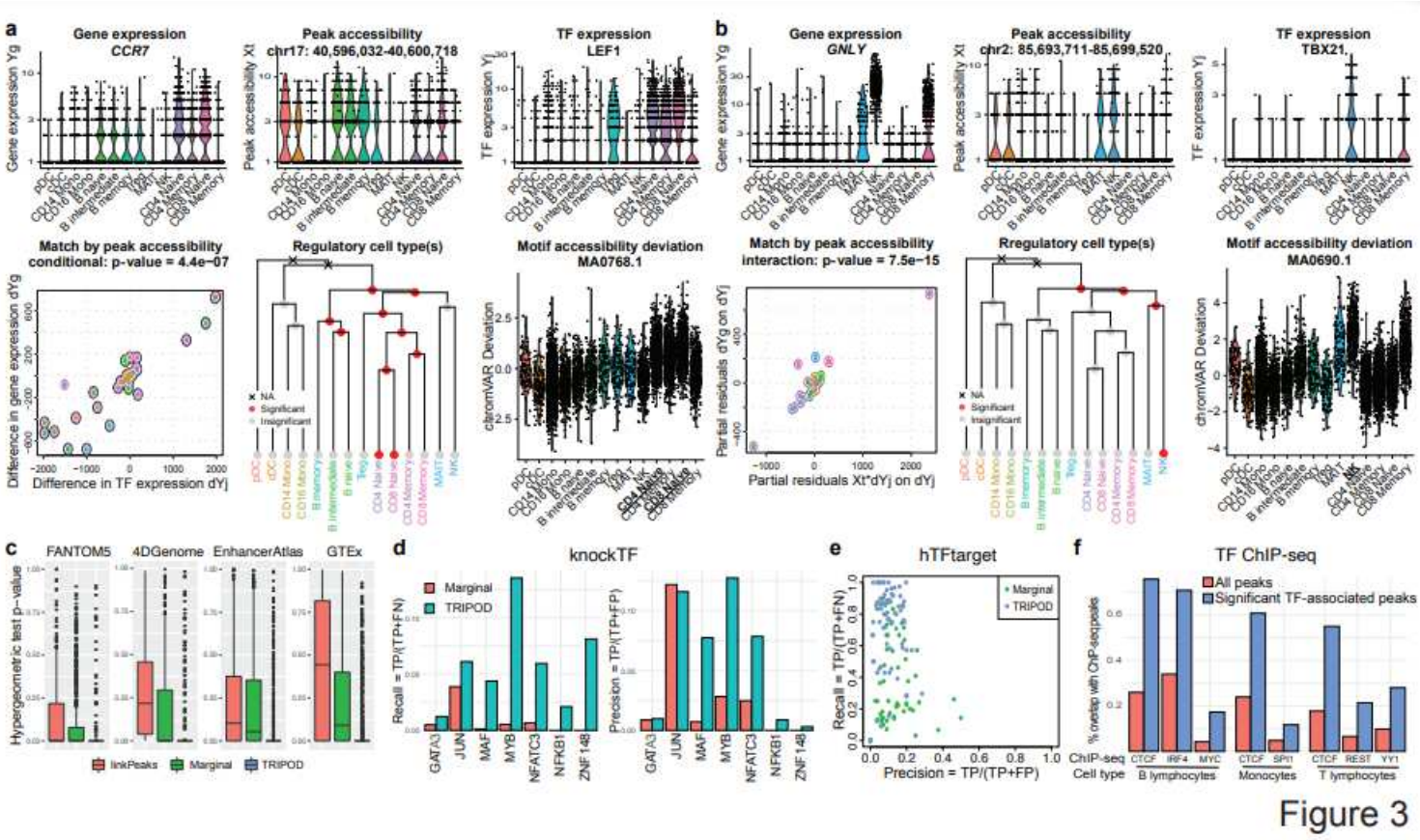

Figure 3

Figure 3

TRIPOD identified trio regulatory relationships in PBMC single-cell multiomic dataset.

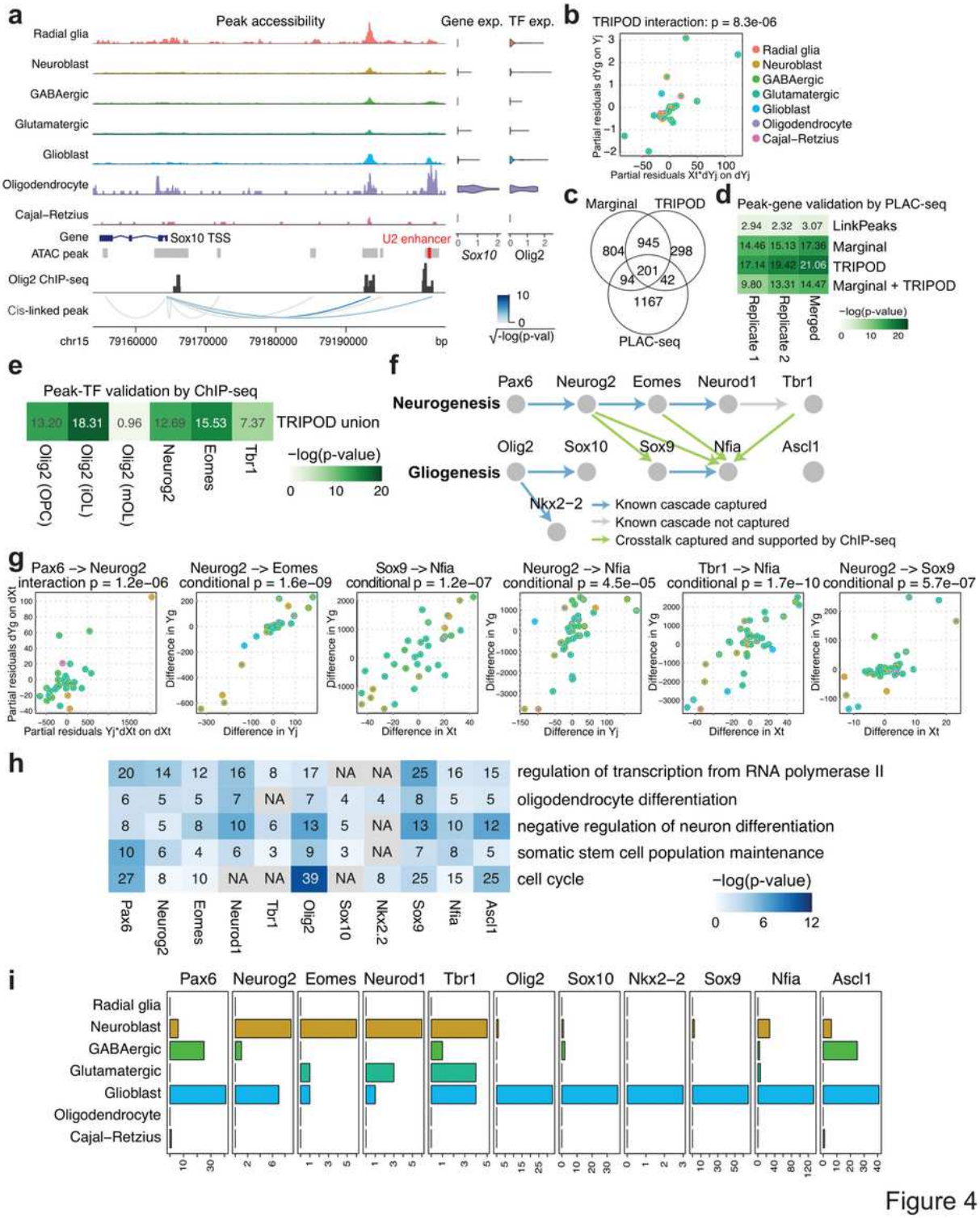

Figure 4

Figure 4

TRIPOD identified known and novel regulatory relationships during mouse neurogenesis and gliogenesis.

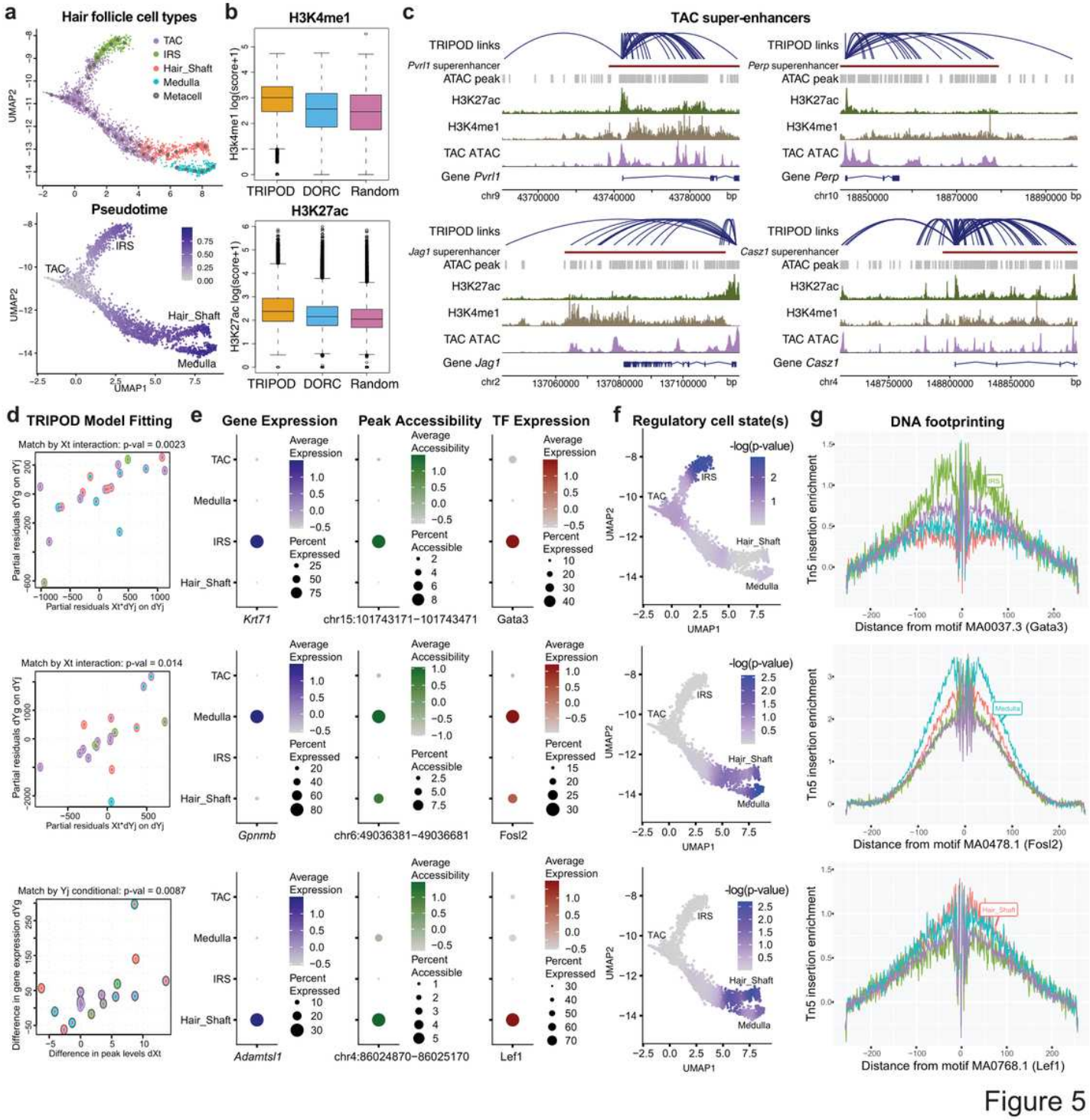

Figure 5

Figure 5

TRIPOD identified regulatory relationships in mouse hair follicles with transient cell states.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1validationresources.pdf
- TRIPODsupplementsv4.pdf