Title

DIRECT-NET: An efficient method to discover cis-regulatory elements and construct regulatory networks from single-cell multiomics data

Journal
Science Advances, 8(22)

ISSN
2375-2548

Authors
Zhang, Lihua
Zhang, Jing
Nie, Qing

Publication Date
2022-06-03

DOI
10.1126/sciadv.abl7393

Peer reviewed

LIFE SCIENCES

DIRECT-NET: An efficient method to discover cis-regulatory elements and construct regulatory networks from single-cell multiomics data

Lihua Zhang[1,2,3], Jing Zhang[4]*, Qing Nie[2,3,5]*

The emergence of single-cell multiomics data provides unprecedented opportunities to scrutinize the transcriptional regulatory mechanisms controlling cell identity. However, how to use those datasets to dissect the cis-regulatory element (CRE)–to–gene relationships at a single-cell level remains a major challenge. Here, we present DIRECT-NET, a machine-learning method based on gradient boosting, to identify genome-wide CREs and their relationship to target genes, either from parallel single-cell gene expression and chromatin accessibility data or from single-cell chromatin accessibility data alone. By extensively evaluating and characterizing DIRECT-NET's predicted CREs using independent functional genomics data, we find that DIRECT-NET substantially improves the accuracy of inferring CRE-to-gene relationships in comparison to existing methods. DIRECT-NET is also capable of revealing cell subpopulation–specific and dynamic regulatory linkages. Overall, DIRECT-NET provides an efficient tool for predicting transcriptional regulation codes from single-cell multiomics data.

INTRODUCTION

In eukaryotes, transcriptional regulation is essential to maintaining proper cell identity during differentiation, determining the appropriate responses to intra- and extracellular signals, and coordinating the myriad of cellular activities at all times (*1*). It undergoes a precise spatial and temporal control within the cell via complex interactions of various cis-regulatory elements (CREs; e.g., enhancers and promoters), transcription factors (TFs), and chromatin remodelers (*2*–*4*). The task to uncover the transcriptional regulation code orchestrating gene activities within a cell includes identification of functional CREs, characterizing their molecular functions, linking them to genes, and finding TF–to–CRE–to–gene regulatory interactions.

Until recently, high-throughput sequencing data at the bulk tissue level have been the main resource for identification of CREs and constructing TF regulatory networks (TRNs). For example, several methods, with unsupervised/supervised approaches, used the combinatory patterns of various epigenetic features within a genomic region [e.g., 200–base pair (bp) bins] to infer the existence of enhancers (*5*–*7*). Other machine learning models also used different (or combinatory) functional genomics data, such as chromatin immunoprecipitation sequencing (ChIP-seq), high-throughput chromosome conformation capture (HiC), RNA sequencing (RNA-seq), and HiChIP to infer the CRE (especially enhancers) to gene linkages. TF binding profiles (*8*) or coexpression patterns of genes (*9*) were also used to construct TRNs. At the tissue scale, those methods are effective in annotating the noncoding genome, uncovering transcriptional regulation, and interpreting variant impacts (*10*, *11*). However, biological tissues often consist of multiple cell types or states with intimate connections among them, different types of cells, or cells at different spatial and temporal states, with potentially different transcriptional regulation codes. Methods based on averaged genomics signals arising from thousands to millions of cells in a tissue invariably limit their ability to identify complex transcriptional regulations that may vary among different cell types or states [hereafter, we will use "cell state" to describe a group of defined subpopulation of cells without distinguishing the differences between "cell type" and cell state (*12*, *13*)].

Recent advances in single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq) have enabled chromatin accessibility landscape profiling across tens of thousands of single cells (*14*, *15*), providing new opportunities for determining regulation codes in individual cells or different cell states. However, very few methods are available for addressing this critical yet challenging task (*16*–*18*). One promising approach, Cicero, was recently developed to infer the links between distal CREs and target genes from scATAC-seq data (*16*). Using a linear graphical lasso model, Cicero quantifies how changes in chromatin accessibility relate to changes in the expression of nearby genes. In ArchR, co-accessibility only uses ATAC-seq data to look for correlations in accessibility between two peaks, while peak-to-gene linkage leverages integrated single-cell RNA-seq (scRNA-seq) data to look for correlations between peak accessibility and gene expression (*17*). SnapATAC predicts gene-enhancer pairs by estimating the significance of the association between binarized chromatin accessibility and gene expression using a logistic regression (*18*).

More recently, several single-cell multiomics technologies such as sci-CAR-seq (*19*), scCAT-seq (*20*), Paired-seq (*21*), SHARE-seq (*22*), and 10X Genomics Multiome (ATAC + RNA) have emerged, enabling simultaneous measurement of gene expression and chromatin accessibility in the same individual cells. The sparse nature of single-cell multiomics data, in particular the chromatin accessibility data, introduces a major challenge for computational analysis of those datasets. Computational tools such as Seurat (*23*), LIGER (*24*), MAESTRO (*25*), MATCHER (*26*), coupled NMF (*27*), scAI (*28*), and BABEL (*29*) were designed to integrate single-cell transcriptomic and epigenomic data. However, these methods are unable to reveal functional CREs and their target genes.

[1]School of Computer Science, Wuhan University, Wuhan 430072, China. [2]Department of Mathematics, University of California, Irvine, Irvine, CA 92697, USA. [3]NSF-Simons Center for Multiscale Cell Fate Research, University of California, Irvine, Irvine, CA 92697, USA. [4]Department of Computer Science, University of California, Irvine, Irvine, CA 92697, USA. [5]Department of Developmental and Cell Biology, University of California, Irvine, Irvine, CA 92697, USA.
*Corresponding author. Email: jingz31@uci.edu (J.Z.); qnie@uci.edu (Q.N.)

Here, we present a computational method, named DIRECT-NET (Discover cis-Regulatory Elements and Construct TF regulatory NETwork), using eXtreme Gradient Boosting (XGBoost) machine learning (*30*). Specifically, DIRECT-NET can accommodate either scATAC-seq or sc-multiomics data with gene expression and chromatin accessibility profiles measured in the same cell to answer two key questions on transcriptional regulation. First, it identifies functional CREs among all accessible chromatin regions based on the expectation that the on-and-off status of truly functional CREs should markedly alter the expression patterns of their target genes. Their relationship is described by a nonlinear predictive model between chromatin accessibility scores and gene expression values (or promoter accessibility scores) using XGBoost, and thus, we turn the CRE identification problem into a model selection problem. Second, DIRECT-NET infers the TF binding footprints using known motif patterns from public databases and constructs the TF–to–CRE–to–gene regulatory networks. Four public single-cell datasets, including two scATAC-seq datasets and two parallel scRNA-seq and scATAC-seq datasets generated by two different protocols, are used to evaluate DIRECT-NET. Extensive benchmark analyses using independent functional genomics data such as ChIA-PET, HiC, HiChIP, and ChIP-seq as well as disease-associated genetic variants from genome-wide association studies (GWAS) show that DIRECT-NET is able to characterize the transcriptional regulation code and reveal cell state–specific regulatory mechanisms.

RESULTS

Overview of DIRECT-NET

DIRECT-NET is a tool that detects functional CREs, links CREs to their target genes, and constructs TF–to–CRE–to–gene regulatory networks (Fig. 1). For each gene (the yellow rectangle), any peak within the 500 bp upstream of its transcription start site (TSS) is defined as the promoter (i.e., the green rectangular), while other open chromatin regions outside of the promoter region but within a user-specified neighborhood (default at 250 kb both sides) are defined as distal candidate functional regions. DIRECT-NET takes either parallel scRNA-seq and scATAC-seq data or scATAC-seq data as input. To account for the sparse nature of single-cell sequencing data, especially for the nearly binary scATAC-seq data, aggregation is often used as an efficient countermeasure (*28*). Therefore, DIRECT-NET first aggregates the binary epigenomic signals or both transcriptomic and epigenomic profiles across similar cells, which are inferred on the basis of nearest neighbors in a learned low-dimensional representation of the input data (see details in the "Aggregation of sparse single-cell data" section in Materials and Methods). Second, DIRECT-NET identifies CREs by regressing either the expression level of a particular gene (parallel scRNA-seq and scATAC-seq) or the accessibility score of a promoter (scATAC-seq), using the accessibility scores of all possible peaks in its neighborhood via the efficient ensemble gradient boosting machine-based model XGBoost. Third, our DIRECT-NET model then selects functional CREs via the importance scores learned from the XGBoost model (see the "Identification of functional CREs" section in Materials and Methods). Last, we build a TF–gene regulatory network by integrating CRE-TF with CRE-gene relationships, where TFs bind to the predicted functional CREs inferred using known motif patterns from public databases.

DIRECT-NET uncovers cell state–specific regulatory elements from parallel scRNA-seq and scATAC-seq data

Identification of cell state–specific CREs is crucial to dissecting cell fate decisions. To evaluate the ability of DIRECT-NET to detect cell state–specific CREs, we first used the 11,909 human peripheral blood mononuclear cells (PBMCs) of 19 cell clusters, with measurements of transcriptomic and chromatin accessibility profiles in the same cells (see "Data and materials availability") (*23*).

By classifying the regulatory links into five groups based on their inferred weights from high to low, we observed that regulatory links with higher importance scores exhibited higher concordance with the promoter capture HiC (PCHiC) (*31*) connections (fig. S1; see the "Validation of inferred connections by PCHiC, ChIA-PET, HiC, and HiChIP data" section in Materials and Methods). Peaks with importance scores equaling zero had nearly no connections in PCHiC data. On the other hand, we found that PCHiC connections were most recalled by regulatory links with higher importance scores (fig. S1, A and B), suggesting the ability of DIRECT-NET in differentiating nonfunctional open chromatin regions from functional CREs.

To investigate the ability of DIRECT-NET to uncover cell state–specific regulatory elements, we first divided peaks into high-confidence (HC) CREs, medium-confidence (MC) regions, and low-confidence (LC) regions. In this study, we treated HC CREs as functional CREs (see the "Identification of functional CREs" section in Materials and Methods). We found that HC CREs are able to differentiate each cell state, while MC regions show some mixed cell states such as T_{reg} (regulatory T cells), CD4 T_{CM} (central memory T cells), and CD4 T_{EM} (effector memory T cells), and cells are completely mixed with LC regions on the UMAP (Uniform Manifold Approximation and Projection) space (Fig. 2A). We next compared the performance quantitatively based on the Local Inverse Simpson's Index (LISI) and Silhouette metrics. HC CREs show higher LISI and Silhouette values than MC and LC regions with all P values less than 1×10^{10} by one-sided Wilcoxon rank test (Fig. 2B). To examine whether HC CREs are enriched in the most accessible peaks, we identified the most accessible peaks using the logistic regression test in Seurat (*32*). Among the identified 26,139 most accessible peaks, 80,740 HC CREs, and 29,162 pure HC CREs (removing the overlapped peaks with MC or LC regions from all HC CREs), we found that 85% most accessible peaks are HC CREs and 38% most accessible peaks are pure HC CREs. Moreover, pure HC CREs are significantly enriched in the most accessible peaks by using Fisher exact test ($P < 2.2 \times 10^{-16}$). Then, we applied the available PCHiC data of native CD4 T cells, native CD8 T cells, and native B cells from a previous study to validate predicted cell state–specific CREs (*31*). Here, the links between HC CREs and promoters are called HC functional links, while the links between LC regions and promoters are called LC links. For each of these three cell states, the identified HC functional links were highly validated by the corresponding PCHiC data compared to LC links (native CD4 T: HC versus LC, 0.24 versus 0.11, $P = 6.9 \times 10^{-4}$; native CD8 T: HC versus LC, 0.2 versus 0.07, $P = 6.5 \times 10^{-9}$; native B: HC versus LC, 0.18 versus 0.05, $P = 4.8 \times 10^{-10}$ via one-sided Student's t test; Fig. 2C and fig. S2A). On the other hand, the connections in cell state–specific PCHiC data were highly recalled in the HC functional links compared to both LC links and MC links (native CD4 T: HC versus MC, 0.66 versus 0.46, $P = 0.02$, HC versus LC, 0.66 versus 0.13, $P = 3.4 \times 10^{-11}$; native CD8 T: HC versus MC, 0.72 versus 0.42, $P = 3.9 \times 10^{-6}$, HC versus LC, 0.72 versus 0.08,

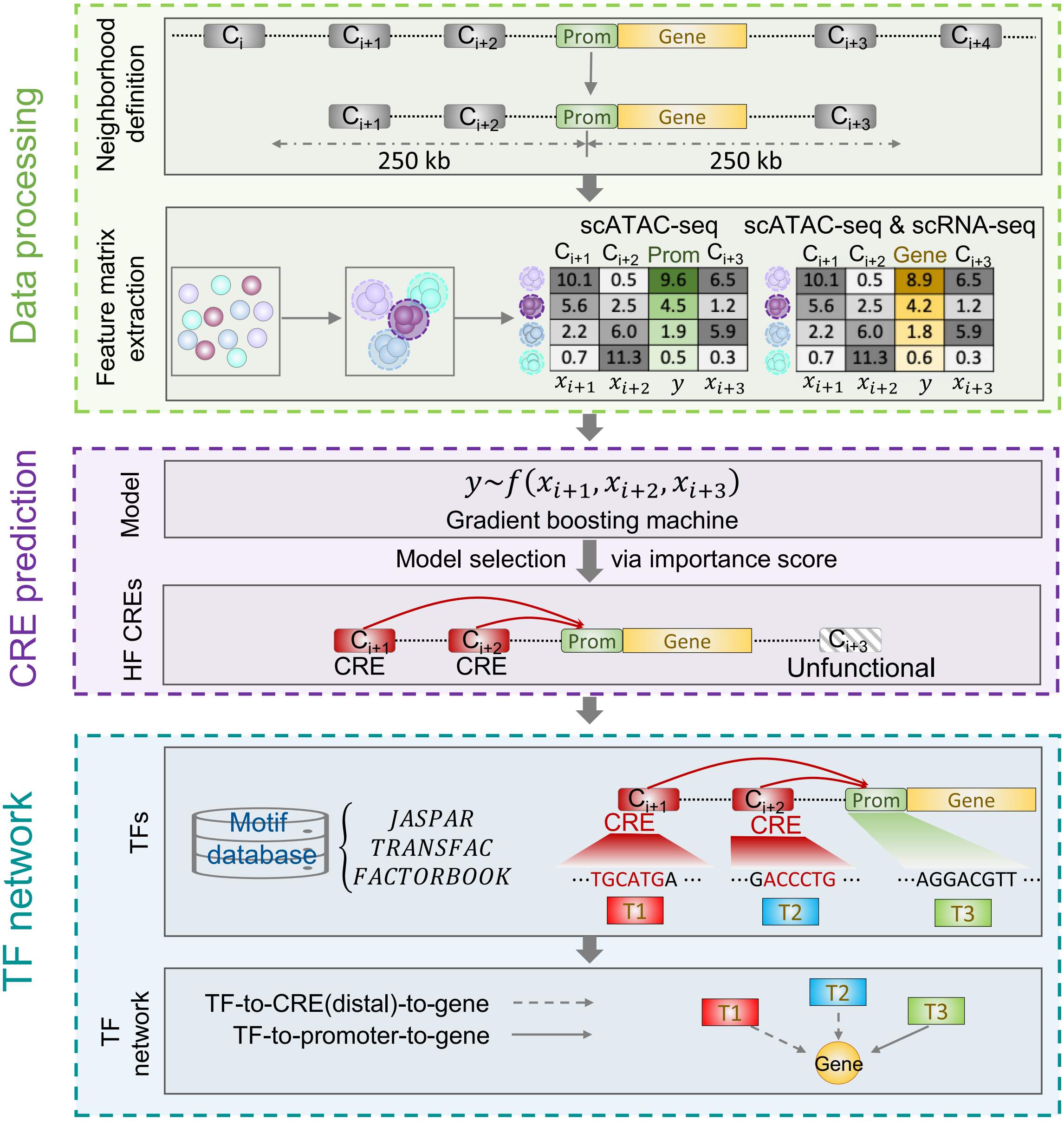

Fig. 1. Overview of DIRECT-NET. For each gene (yellow), the 500 bp upstream of its TSS is treated as promoter (green), and the 250 kbp upstream and downstream of its TSS are treated as distally candidate functional regions (i.e., C2, C3, and C4). Taking parallel scRNA-seq and scATAC-seq data measured at the same cells or scATAC-seq data as input, DIRECT-NET first constructs a new feature matrix by aggregating signals across similar cells. Second, DIRECT-NET infers CREs using gradient boosting machine models. For each promoter or gene, this model regresses the accessibility of promoter or gene expression level with the accessibility of distally candidate functional regions. Last, by identifying TFs bounded to the promoters and identified CREs using motif enrichment analysis, DIRECT-NET reconstructs TF–gene regulatory networks.

$P = 6.8 \times 10^{-31}$; native B: HC versus MC, 0.78 versus 0.39, $P = 6.3 \times 10^{-11}$, HC versus LC, 0.78 versus 0.04, $P = 1.3 \times 10^{-45}$ via one-sided Student's *t* test; Fig. 2C and fig. S2B). We merged the significant interaction of these three PCHiC. Then, we computed ratios of the links validated by PCHiC connections and ratios of recalled PCHiC connections by the links of markers for these three different cell states based on merged PCHiC data and naïve B cell–, CD4 naïve T cell–, and CD8 naïve T cell–specific PCHiC data individually. We found that ratios of the links validated by merged PCHiC data were significantly increased than cell state–specific PCHiC data of the corresponding cell state–specific markers. However, ratios of recalled merged PCHiC connections and ratios of recalled cell state–specific PCHiC connections were comparable (fig. S2, C and D).

We next compared DIRECT-NET with two baseline methods (CloseGene and Distance), four bulk methods [Spearman correlation coefficients (SCCs), Lasso, Ridge, and ElasticNet], and three single-cell methods [Cicero (*16*), ArchR (*17*), and SnapATAC (*18*)] (Materials and Methods). To reduce the influence of different thresholds on performance of different methods, we computed the area under the ROC curve (AUC) for each marker gene and compared these AUCs across methods using a paired, right-tailed Student's *t* test. By using PCHiC links of naïve B, CD4 naïve T, and

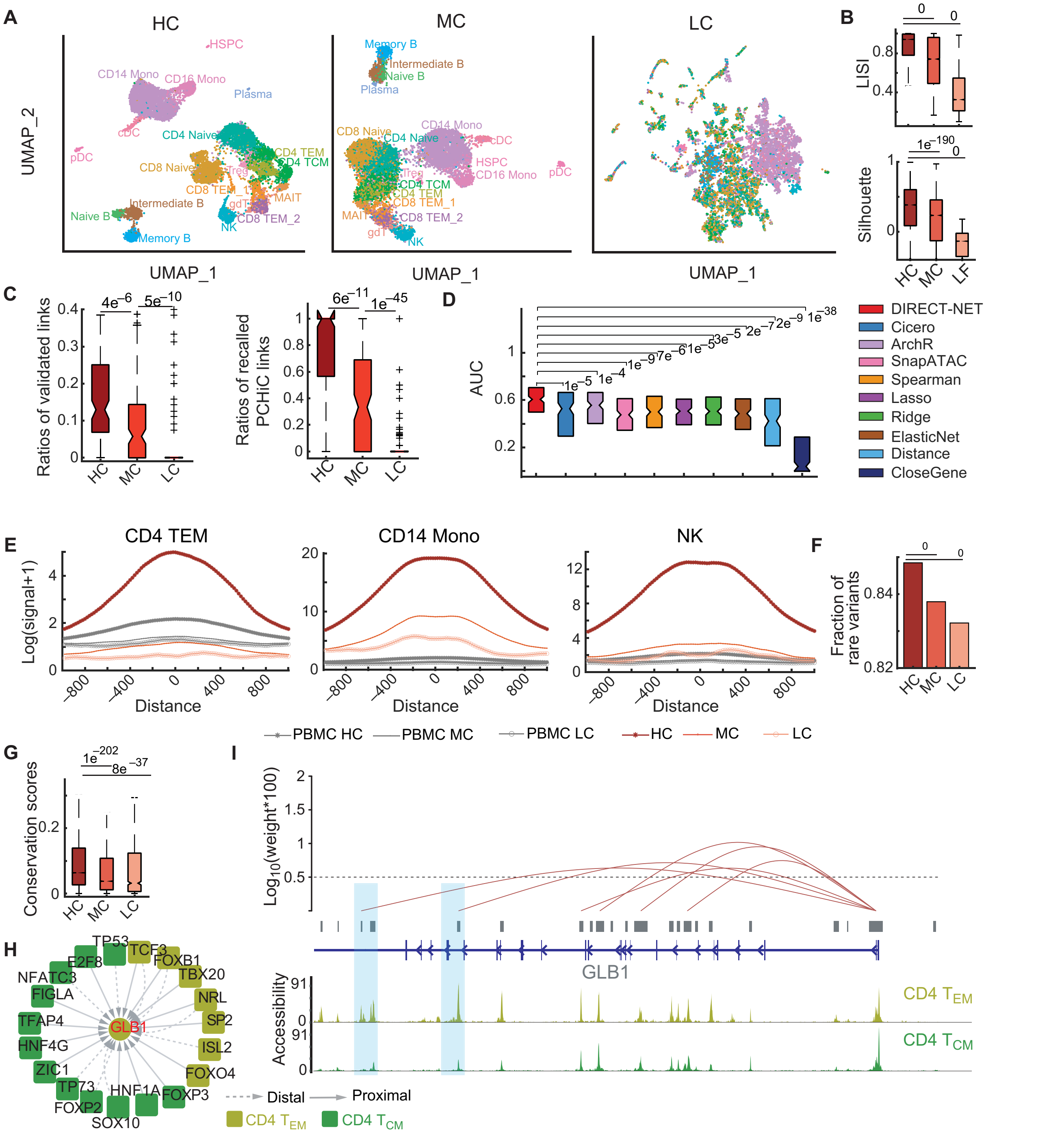

Fig. 2. DIRECT-NET detects cell state–specific regulatory links on PBMC dataset. (**A**) Projection of cells onto UMAP space using the scATAC-seq data of pure HC CREs (left), pure MC regions (middle), and pure LC regions (right). We identified pure HC CREs by removing the overlapped peaks with MC or LC regions from all HC CREs. Similarly, we obtained pure MC regions and pure LC regions separately. (**B**) Evaluation of cell state separation on the UMAP space using HC CREs, MC regions, and LC regions via LISI (top) and Silhouette metrics (bottom). (**C**) Ratios of links of HC CREs, MC regions, and LC regions validated by PCHiC data on markers of naïve B cells. *P* values are from one-sided Student's *t* test. (**D**) Comparison of performance of DIRECT-NET, Cicero, ArchR, SnapATAC, Spearman, Lasso, Ridge, ElasticNet, Distance, and CloseGene using the AUC value of each marker gene of naïve B cells. (**E**) Average H3K27ac signals of 1000 bp upstream and 1000 downstream from the middle base of HC CREs, MC regions, and LC regions for NK, CD4 T_{EM}, and CD14 Mono cells using bigWigAverageOverBed. Each value is computed by averaging over 20-bp bin. (**F**) Comparison of enrichment scores of rare variants between HC CREs, MC regions, and LC regions using PCAWG data. The enrichment score of rare variants is defined by the ratio of overlapped rare variants to the sum of overlapped rare and common variants. *Nr* and *Nc* represent the number of overlapped rare and common variants. *P* values are from right-sided Binormal tests with the enrichment score of LC regions as probability. (**G**) Comparison of phastCons conservation scores between HC CREs, MC regions, and LC regions. *P* value is from a right-sided Wilcoxon rank sum test. (**H**) Subnetwork of the gene regulatory network of differentially expressed TFs and target genes between CD4 T_{EM} and CD4 T_{CM}. (**I**) Links for the GLB1 locus with H3K27ac signals of CD4 T_{EM} and CD4 T_{CM}.

CD8 naïve T cells as ground-truth, we found that DIRECT-NET consistently exhibited significantly higher AUC values than other methods (Fig. 2D and fig. S2E). Compared to Cicero, ArchR, and SnapATAC, more regulatory links are recalled by DIRECT-NET according to PCHiC data of naïve B, CD4 naïve T, and CD8 naïve T cells (fig. S3, A to C). Cicero shows higher overlap ratios with DIRECT-NET than do other methods. Around 60% of links are overlapped between links with high importance and high co-accessibility (fig. S3D).

To investigate whether DIRECT-NET's detected HC CREs can recover cell state–specific signals, we restricted the HC CREs, MC regions, and LC regions associated with marker genes of CD4 T_{EM}, CD14 Mono, and natural killer (NK) cells. The HC CREs of cell state–specific marker genes are found to exhibit the strongest H3K27ac signals to their corresponding cell states, while LC regions exhibited nearly no H3K27ac signals (Fig. 2E). In contrast, the H3K27ac signals of bulk PBMC samples on the DIRECT-NET's detected cell state–specific CREs are lower than the cell state–specific H3K27ac signals (Fig. 2E), which is likely because the bulk PBMC samples are composed of different immune cell state and the measured H3K27ac signals are the average across different cell states. Moreover, HC CREs show strong cell state–specific ATAC-seq signals and strong aggregated scATAC-seq signals of each corresponding cell state (fig. S4, A and B), but the signals of aggregated cells from different immune cell states do not have cell state–specific characteristics on CREs of CD4 T_{EM} and CD14 Mono (fig. S4C). These results indicate that DIRECT-NET succeeds in revealing cell state–specific functional CREs and that single-cell genomic data are invaluable in deciphering cell state–specific regulatory mechanisms.

Lines of previous studies have demonstrated that truly functional CREs are highly conserved across both populations and species when compared to random genomic regions, as reflected by their depleted patterns in common variants and higher conservation scores in comparative genomics analysis (*33*, *34*). Therefore, we further compared the DIRECT-NET's predicted HC CREs with the MC and LC regions using both cross-population and cross-species conservation. Specifically, we first intersected PCAWG variants with HC CREs, MC regions, and LC regions and then calculated the fraction of rare variants, respectively. As expected, HC CREs show significantly higher rare variant fractions than do MC and LC regions (HC versus MC: 0.849 versus 0.839, $P = 0$; HC versus LC: 0.849 versus 0.834, $P = 0$ by one-sided binomial test; Fig. 2F). Adding on, HC CREs are significantly more conserved than the MC and LC regions using the PhastCons method (*35*) on placental mammals species (HC versus MC: $P = 1 \times 10^{-202}$; HC versus LC: $P = 8 \times 10^{-37}$ via one-sided Wilcoxon rank test) (Fig. 2G), demonstrating once again DIRECT-NET's capacity in identifying functional CREs.

To study the cell state–specific gene regulatory mechanisms, we next built the gene regulatory network of TFs and target genes for CD4 T_{EM} and CD4 T_{CM} (see the "Reconstruction of gene regulatory networks" section in Materials and Methods; Fig. 1H, and fig. S5). Cell state–specific enhancer-gene links were extracted on the basis of active peaks and marker genes of the corresponding cell state (fig. S5A). On the basis of the cell state–specific enhancer-gene links, we built TRNs of CD4 T_{EM} and CD4 T_{CM} separately (fig. S5, B and C). To investigate different TFs and targets as well as their regulations between CD4 T_{EM} and CD4 T_{CM}, we combined the two cell state–specific networks into a single regulatory network (see the "Reconstruction of gene regulatory networks" section in Materials and Methods) and presented genes or TFs highly expressed in CD4 T_{EM} or CD4 T_{CM} with fold change > 2. The gene regulatory network exhibits different regulatory links between CD4 T_{EM} and CD4 T_{CM} (fig. S5D). One interesting CD4 T_{EM}–specific marker gene is GLB1, which is regulated by TFs such as TF T cell factor 3 (TCF3; Fig. 1H). TCF3 has been illustrated previously as a constitutive enhancer (*36*) and transcriptional activator (*37*). Next, we studied whether there are increased signals in CREs of GLB1. As expected, two sites underwent significantly increased accessibility in CD4 T_{EM} than in CD4 T_{CM} around GLB1 (Fig. 2I). Together, DIRECT-NET shows strong capability in capturing cell state–specific regulatory connections.

Constructing dynamic gene regulatory networks from time-course parallel scRNA-seq and scATAC-seq data

We next evaluated the ability of DIRECT-NET to learn CREs and reconstruct gene regulatory networks across a dynamic process from parallel scRNA-seq and scATAC-seq data. We applied DIRECT-NET to 2641 lung adenocarcinoma–derived A549 cells after 0, 1, and 3 hours of 100 nM dexamethasone (DEX) treatment, which includes both gene expression and chromatin accessibility profiles measured in the same cells (*19*). To detect regulatory changes during the DEX treatment process, we focused on CREs of differentially expressed genes across three time points (see the "Datasets and gene selection" section in Materials and Methods). Similar to the observation above in the PBMC dataset, regulatory links with high importance scores show high concordance with the HiC connections downloaded from ENCODE, while links with zero importance scores exhibit nearly no concordance with HiC data (fig. S6, A to C). The links between HC CREs and target genes were more highly validated than those of MC and LC regions by HiC data (mean values: HC versus MC: 0.09 versus 0.05, $P = 4 \times 10^{-7}$; HC versus LC: 0.05 versus 0.03, $P = 1 \times 10^{-11}$ by one-sided Student's *t* test; Fig. 3A). Moreover, HiC connections were highly recalled by HC CREs than MC and LC regions (HC versus MC: 0.84 versus 0.32, $P = 3 \times 10^{-16}$; HC versus LC: 0.84 versus 0.19, $P = 4 \times 10^{-31}$ by one-sided Student's *t* test; fig. S6D), illustrating the ability of DIRECT-NET to differentiate functional and less functional regions near marker genes. Moreover, DIRECT-NET has significantly higher AUC values than Cicero, Spearman, Lasso, Ridge, ElasticNet, Distance, and CloseGene methods (fig. S6E). Compared to Cicero, more regulatory links are recalled by DIRECT-NET (fig. S6F) according to the HiC data. More than 60% of links are overlapped between links with high importance and high co-accessibility (fig. S6G).

Next, we explored whether DIRECT-NET could capture the biological significance of CREs by the following three aspects. First, by performing dimensional reduction on the chromatin accessibility data of HC CREs using PHATE (*38*), we found that cells from the three time points are well separated in the low-dimensional space. In contrast, the cells from 0 and 1 hour are mixed, with the peak importance values equaling 0 (Fig. 3B), suggesting the ability of CREs to dissect cellular heterogeneity. Second, similar to previous results, we further compared cross-population and cross-species conservation of the discovered HC CREs and LC regions. As expected, HC CREs are significantly enriched in rare variants as compared to the MC and LC regions using PCAWG variants (HC versus MC: 0.851 versus 0.845, $P = 3 \times 10^{-7}$; HC versus LC: 0.851 versus 0.847, $P = 1 \times 10^{-15}$ by one-sided binomial test; Fig. 3C). Third, the HC CREs show significantly higher PhastCons scores than MC and LC regions

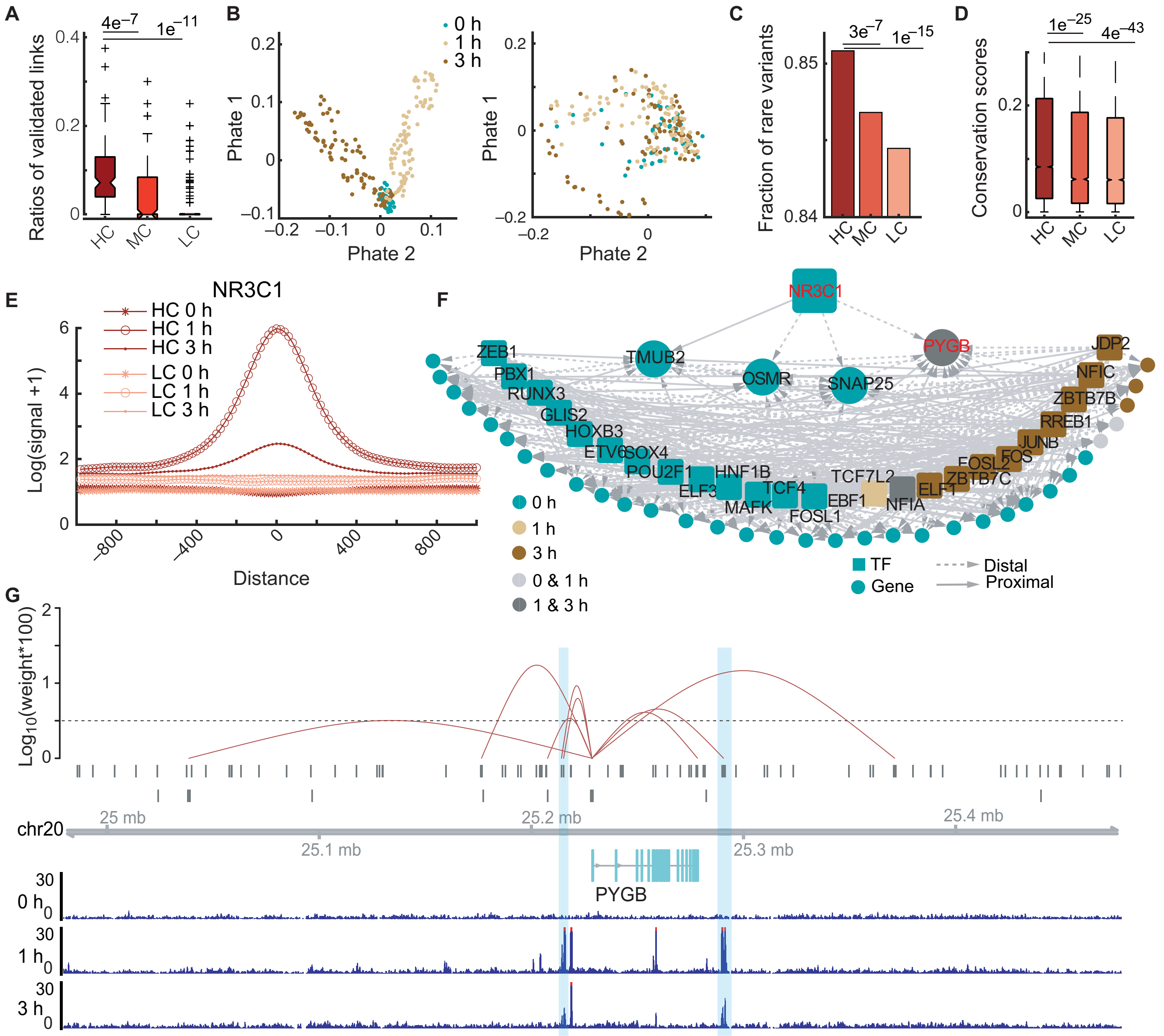

Fig. 3. DIRECT-NET detects time-series regulatory links on DEX-treated A549 cell line. (**A**) Ratios of DIRECT-NET–identified links validated by promoter HiC data for HC, MC, and LC links over the total number of identified links. (**B**) Projection of cells onto Phate space using the scATAC-seq data of HC peaks and peaks with importance scores equaling 0. Cells are colored by time points. (**C**) Comparison of enrichment scores of rare variants between HC CREs, MC regions, and LC regions using PCAWG data. (**D**) Comparison of phastCons conservation scores between HC CREs, MC regions, and LC regions. (**E**) Average NR3C1 signals of 1000 bp upstream and 1000 bp downstream from the middle base of HC CREs and LC regions across the three time points using bigWigAverageOverBed. (**F**) Subnetwork of gene regulatory network of NR3C1, four target genes of NR3C1, regulated TFs of these four genes, and all genes regulated by these TFs. (**G**) DIRECT-NET–predicted connections for the PYGB locus with NR3C1 signals across 0, 1, and 3 hours.

(HC versus MC: 1×10^{-25}, HC versus LC, $P = 4 \times 10^{-43}$ by one-sided Wilcoxon rank test; Fig. 3D) (*35*). Together, these results suggest that DIRECT-NET can yield interpretable and biologically meaningful CREs in the dynamic process from parallel scRNA-seq and scATAC-seq data.

To further demonstrate DIRECT-NET's power to predict CREs associated with the dynamic process, we examined the ChIP-seq signals of the identified HC CREs associated with NR3C1, a well-known marker of early events after treatment (*19*). As expected, HC CREs have no NR3C1 signals at 0 hours but the highest NR3C1 signals at 1 hour, consistent with the role of NR3C1 as an early activation marker after treatment (Fig. 3E) (*39*). Moreover, HC CREs show higher NR3C1 signals than MC and LC regions at 1 and 3 hours (fig. S7A), and exhibit higher H3K27ac signals than MC and LC regions at all time points (fig. S7B). These results illustrate that DIRECT-NET successfully identified functional CREs associated with the dynamic process.

Last, we built the gene regulatory network to explore the key factors of the glucocorticoid receptor (GR) activation process along the DEX treatment (Fig. 3F). Here, we focus on TFs bound to the differentially accessible HC peaks of the differentially expressed marker genes (see the "Reconstruction of gene regulatory networks" section in Materials and Methods). In this network, one notable TF is NR3C1, showing the highest expression level at 0 hours, although the motif of NR3C1 is mostly activated at 1 hour (Fig. 3E). This result is consistent with a previous study that the motif of NR3C1 is activated, while its expression is decreased during GR activation (*19*). Four genes (TMUB, OSMR, SNAP25, and PYGB) are regulated by NR3C1, and all these genes are highly expressed in 0 hours except for PYGB (glycogen phosphorylase B), which is highly expressed at both 1 and 3 hours (Fig. 3F). The peak intensity of ChIP-seq data of NR3C1 of two sites that overlapped with HC CREs at 1 and 3 hours is stronger than that overlapped at 0 hours, implying the potential regulation of NR3C1 on PYGB associated with the DEX treatment process (Fig. 3G). In addition to these results on NR3C1, DIRECT-NET reveals other important TFs such as FOSL2, FOS, and JUNB (fig. S8), which are the AP-1 TF family members that commonly colocalize with GR (*40*). DIRECT-NET is highly effective in identifying CREs and TF-to-target links, which are important in regulating the dynamic process.

DIRECT-NET accurately predicts functionally distinct co-accessibility connections between CRE and promoter regions from scATAC-seq data

We applied DIRECT-NET to a widely used scATAC-seq data of 889 human lymphoblastoid GM12878 cells and benchmarked its performance against Cicero in predicting CREs and their target genes (*16*). Regulatory links with higher importance scores exhibit higher concordance with polymerase II (Pol II) ChIA-PET connections (figs. S9 and S10). We found that HC CREs show the highest concordance with the Pol II ChIA-PET connections than MC and LC regions (Fig. 4A). Moreover, HC CREs are significantly enriched with rare variants than LC regions (HC versus LC: 0.846 versus 0.838, $P = 1 \times 10^{-13}$ by one-sided binomial test; Fig. 4B). HC CREs show significantly stronger signals of H3K27ac than do the MC and LC regions (Fig. 4C). These results illustrated the ability of DIRECT-NET to uncover CREs.

We next compared DIRECT-NET with other competitive methods. DIRECT-NET and Distance produce regulatory links that are more likely to be found in ChIA-PET and HiC than other methods, as reflected by the significantly higher AUC values of DIRECT-NET and Distance on highly variable genes (Fig. 4D), suggesting the DIRECT-NET and Distance's higher accuracy to predict functional CREs. When ChIA-PET links show the strongest proximal characteristics, the genomic distance baseline method Distance has the highest AUC values compared to the other methods (fig. S11). Consistently higher number of proximal ChIA-PET is found in DIRECT-NET's predictions (Fig. 4E and fig. S11). Moreover, by comparing the predicted CREs with GM12878 ChIA-PET connections, we found that the ratios of HC CREs that are validated by ChIA-PET are larger than those of MC and LC regions (Materials and Methods and Fig. 4D). On the other hand, up to 80% ChIA-PET connections are detected by DIRECT-NET for these variable genes (Fig. 4E). Together, DIRECT-NET exhibits better performance in predicting biologically meaningful CREs.

Detecting cell state–specific neurological disease–associated regulatory elements from scATAC-seq data of adult brain

CRE identification is crucial to understanding disease mechanisms. To evaluate the ability of DIRECT-NET in detecting CREs associated with brain diseases, we used scATAC-seq data of 70,631 individual cells across seven brain regions of adult brain. Eighteen distinct clusters were identified on the basis of unbiased iterative clustering and Harmony-based batch correction method in the previous study (Materials and Methods) (*41*). Links with higher importance values tend to exhibit higher concordance with HiChIP connections (fig. S12).

To begin, we demonstrated the biological significance of the identified HC CREs by the following four criteria. First, most cell states such as microglia, isocortical inhibitory, and isocortical excitatory are differentiated by both HC and MC regions on the UMAP space. HC CREs have better performance than MC regions in differentiating hippocampal excitatory and OPCs (oligodendrocyte progenitor cells), while cells are completely mixed with LC regions (fig. S13B). Furthermore, HC CREs have higher LISI and Silhouette values than MC and LC regions, with all P values less than 1×10^{-10} by one-sided Wilcoxon rank test (fig. S13C). Second, HC CREs show higher validation ratios based on HiChIP data than do MC and LC regions (fig. S13A). Moreover, more HiChIP links are recalled by HC CREs than by MC and LC regions (HC versus MC: 0.8 versus 0.66, $P = 0.03$, HC versus LC: 0.8 versus 0.14, $P = 6 \times 10^{-15}$, one-tailed Student's t test; Fig. 5A). Third, rare variants are significantly enriched in HC CREs than in MC and LC regions (HC versus MC: 0.852 versus 0.848, $P = 2 \times 10^{-15}$; HC versus LC: 0.852 versus 0.845, $P = 0$ by one-sided binomial test; Fig. 5B). Moreover, HC CREs were found to be more conserved than MC and LC regions (HC versus MC, $P = 3 \times 10^{-215}$; HC versus MC, $P = 7 \times 10^{-71}$ by one-sided Wilcoxon rank test; Fig. 5C). Fourth, HC CREs of the cell state–specific marker genes exhibit the highest H3K27ac signals compared to MC and LC regions. For example, HC CREs of the nigral OPCs cluster's marker genes show stronger H3K27ac signals of substantia nigra than MC and LC regions (Fig. 5D). HC CREs of nigral astrocytes and striatal astrocytes clusters had strong H3K27ac signals of astrocytes (fig. S13D). Next, we compared DIRECT-NET with other methods in identifying loops overlapped with HiChIP (Fig. 5E). Distance has the highest AUC value compared to the other methods, consistent with the fact that regulatory links tend to be proximal (fig. S13F). DIRECT-NET, Spearman, and ArchR have higher AUC values than ElasticNet, Circero, and CloseGene (Fig. 5E). Moreover, more proximal HiChIP connections are recalled by DIRECT-NET (Fig. 5F). DIRECT-NET has higher number of links validated by HiChIP than Cicero and ArchR (fig. S13G).

To explore whether HC CREs can uncover neurological disease variants, we computed linkage disequilibrium (LD) score regression of GWAS single-nucleotide polymorphism (SNPs) of psychiatric disease autism-associated conditions in these CREs (Materials and Methods). We found that HC CREs of the striatal inhibitory 1 and striatal inhibitory 2 clusters are significantly enriched in variants associated with autism (Fig. 5G). FEV (ETS family member) is required for both development and function of serotonergic neurons, and it is a key transcriptional factor associated with autism (*42*). We noticed that FEV is highly expressed in striatal inhibitory 2 and regulated some striatal inhibitory 2–specific markers (NAT14, RGS14, DEGS2, PCP4L1, and CCDC39) from the gene regulatory network (Fig. 5H and fig. S14). The previous study reveals that two brothers with autism carry homozygous stop-gain mutations in FEV

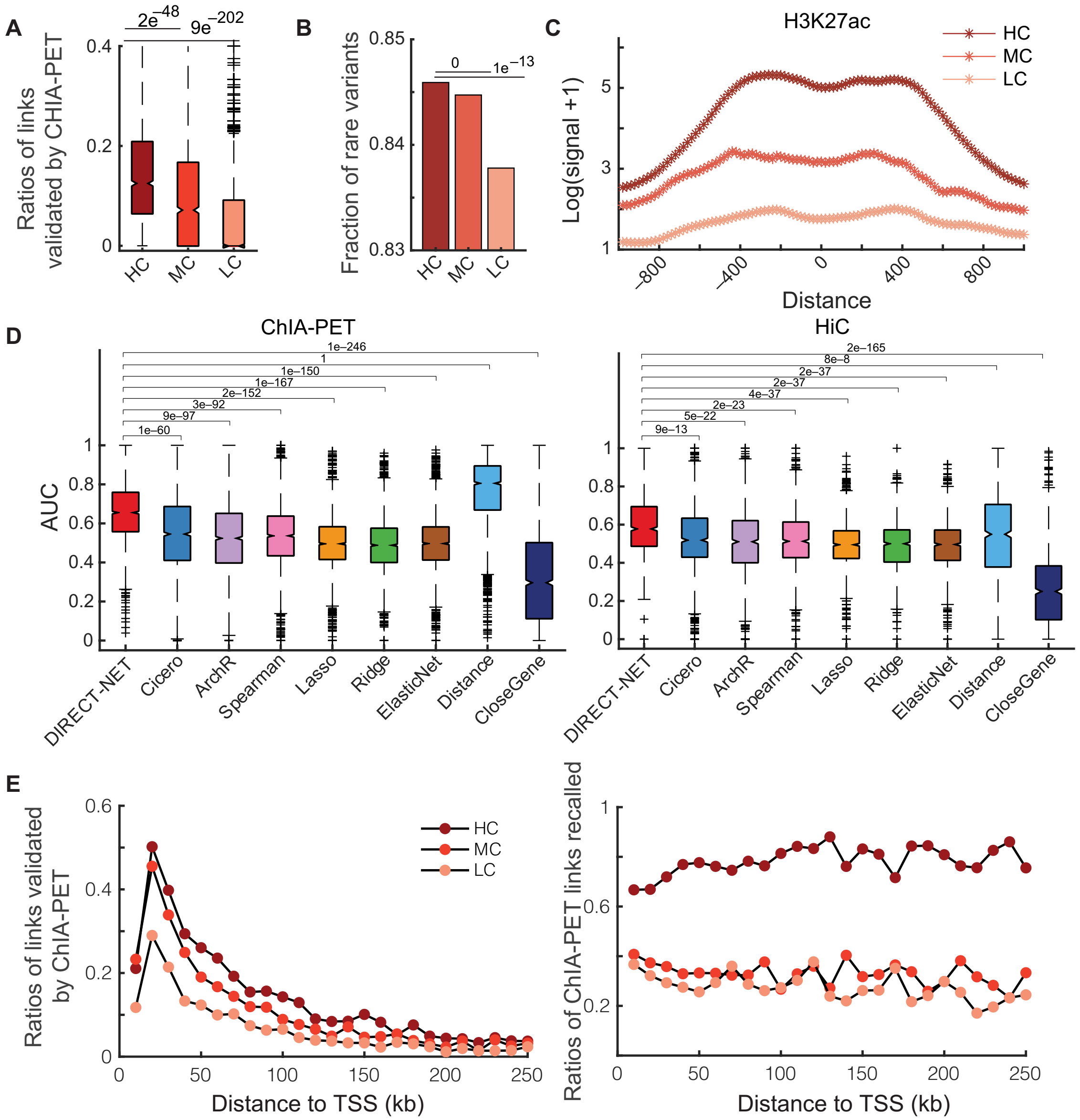

Fig. 4. DIRECT-NET detects accurate regulatory links on GM12878 cell line. (**A**) Ratios of DIRECT-NET–identified links verified by ChIA-PET data over the total number of identified links for three groups of regulatory links, which are defined by the HC, MC, and LC. (**B**) Comparison of enrichment scores of rare variants between HC CREs, MC regions, and LC regions using PCAWG data. (**C**) Average H3K27ac signals of 1000 bp upstream and 1000 bp downstream from the middle base of HC CREs, MC regions, and LC regions using bigWigAverageOverBed. (**D**) Comparison of performance of single-cell methods (DIRECT-NET, Cicero, and ArchR), bulk methods (Spearman, Lasso, Ridge, and ElasticNet), and base methods (Distance and CloseGene) using the AUC value of each marker gene using ChIA-PET (right) and HiC data (right). (**E**) Ratios of links of HC CREs, MC regions, and LC regions validated by ChIA-PET (left) and ratios of ChIA-PET links recalled (right) across different distances to TSS. Each dot represents the ratio for every 10-kb distance to the TSS.

(*42*). Here, we focused on the regulatory links of FEV and found that the two autism SNPs rs78146971 and rs111227810 are located near one HC CRE. Moreover, the CRE has weak accessibility in striatal inhibitory 1 than striatal inhibitory 2 ($\log_2$Fold-change = 1.57; Fig. 5I). Therefore, DIRECT-NET holds the potential to reveal tiny cell subpopulation–specific CREs associated with brain diseases.

DISCUSSION

In this study, we presented a machine learning method, named DIRECT-NET, to identify HC functional CREs, link them with their target genes, and build multimodal TF–to–CRE–to–gene regulatory networks using single-cell transcriptomic and epigenomic data. DIRECT-NET is applicable to parallel scRNA-seq and scATAC-seq data as well as scATAC-seq data alone. We demonstrated the capability of DIRECT-NET using four public datasets from four different scenarios, including two parallel scRNA-seq and scATAC-seq datasets with one involved in discrete cell states from multiple immune cell states and the other one involved in continuous cell states from three time points, and two scATAC-seq datasets from a widely used cell line and a complex tissue. Using the available bulk ChIA-PET, promoter HiC data, and HiChIP data, DIRECT-NET was shown to

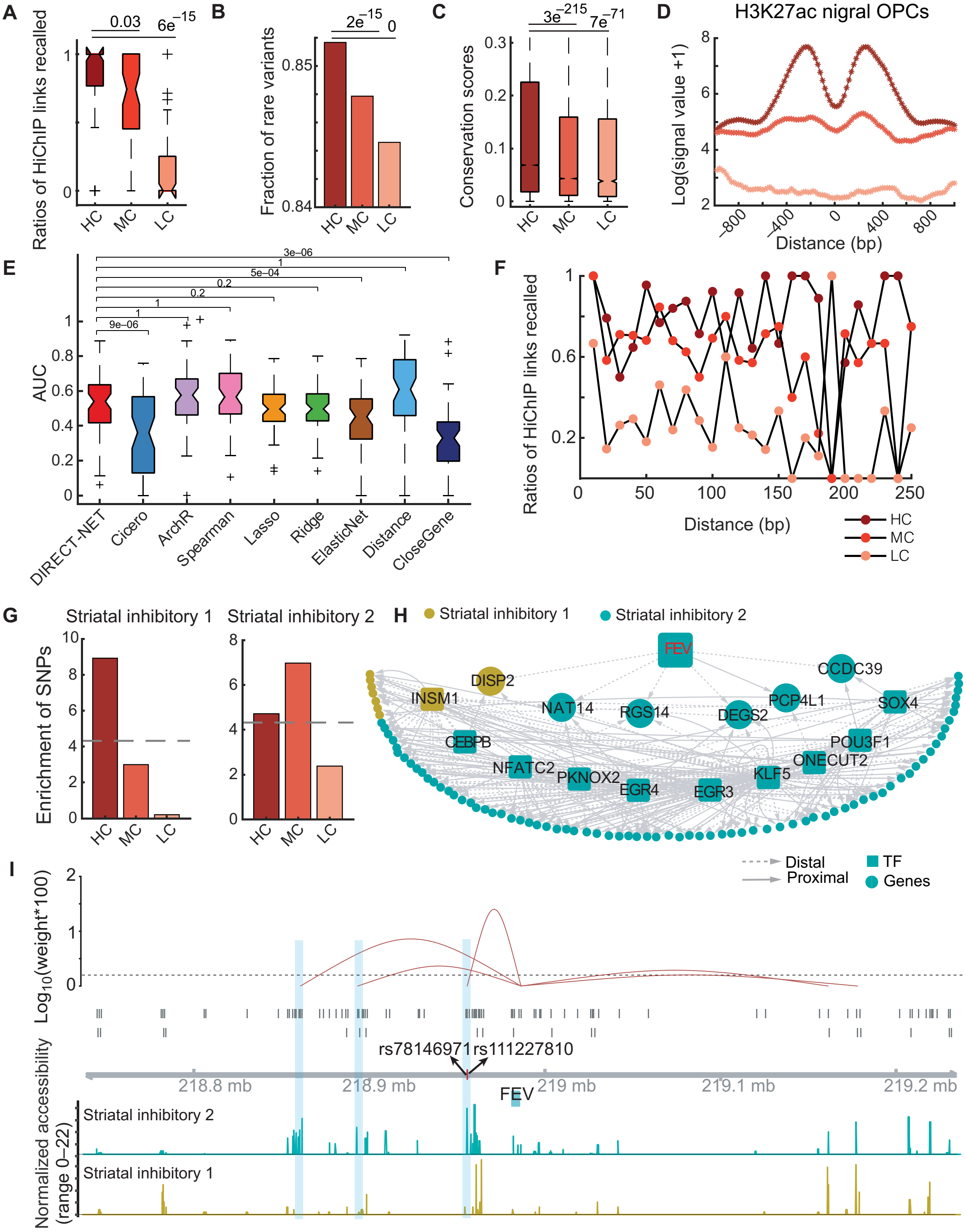

Fig. 5. DIERCT-NET detects cell subpopulation–specific disease-associated CREs on Brain data. (**A**) Ratio of HiChIP links recalled by connections of HC, MC, and LC across marker genes. Each element represents the ratio of one gene. (**B**) Comparison of enrichment scores of rare variants between HC CREs, MC regions, and LC regions using PCAWG data. (**C**) Comparison of phastCons conservation scores between HC CREs, MC regions, and LC regions. (**D**) Average H3K27ac signals of nigral obtained from ENCODE of 1000 bp upstream and 1000 bp downstream from the middle base of HC CREs, MC regions, and LC regions using bigWigAverageOverBed. (**E**) Comparison of performance of single-cell methods (DIRECT-NET, Cicero, and ArchR), bulk methods (Spearman, Lasso, Ridge, and ElasticNet), and base methods (Distance and CloseGene) according to AUC value of each marker gene using HiChIP data. (**F**) Ratios of HiChIP links recalled by the links of HC CREs, MC regions, and LC regions validated across different distances to TSS. Each dot represents the ratio for every 10-kb distance to the TSS. (**G**) LD score regression identifying the enrichment of GWAS SNPs from psychiatric disease autism-associated conditions in the peak regions of HC CREs, MC regions, and LC regions of striatal inhibitory 1 and striatal inhibitory 2. (**H**) Subnetwork of gene regulatory network of FEV, three target genes of FEV, regulated TFs of these three genes, and all genes regulated by these TFs. (**I**) DIRECT-NET–inferred connections of FEV.

exhibit superior performance in discriminating HC functional CREs from MC and LC regions. Using PBMC parallel single-cell transcriptomic and epigenomic data, we showed that DIRECT-NET is able to reveal cell state–specific regulatory elements and gene regulatory networks. Application of DIRECT-NET to the time-course single-cell multiomics data produces a dynamical gene regulatory network and identifies both known and novel TFs and their regulated target genes that are specific to the different stages during DEX treatment.

Promoter accessibility usually serves as a proxy of gene expression, which has been widely used in many approaches such as Seurat and LIGER when integrating scRNA-seq and scATAC-seq data. We adopted two strategies to evaluate whether promoter accessibility is a good proxy of gene expression. First, we computed the SCC between each gene's expression level and its promoter accessibility on the paired scATAC-seq and scRNA-seq of PBMC and A549 datasets. To reduce the effect of sparsity of single-cell multiomics data, we aggregated the nearest 50, 100, and 500 neighbor cells within each cell state as well as all cells within each cell state on PBMC and A549 datasets, with cell state information extracted from previous studies (*19*, *23*). SCC values increase with the decrease of sparsity, and the gene's expression level is highly correlated with promoter accessibility after aggregating cells within each cell state (SCC > 0.8; fig. S15A). Second, we implemented DIRECT-NET on the paired scRNA-seq and scATAC-seq PBMC dataset using promoter accessibility and gene expression as modeling targets, respectively. To test whether these two different modeling targets produce consistent HC CREs, we computed the Jaccard index using the top 10, 15, and 20 inferred CREs and found that the medians of Jaccard indexes are 0.89, 0.9, and 0.89 on PBMC dataset and 0.54, 0.62, and 0.68 on A549 dataset (fig. S15B), suggesting relative high consistence of inferred regulatory elements and interactions when using promoter accessibility and gene expression as the modeling targets. While these results indicate that the promoter accessibility can largely serve as a good proxy of gene expression, this proxy may not work well in certain complicated regulatory relationships.

To study potential differences using DIRECT-NET on the paired scRNA-seq and scATAC-seq data compared to scATAC-seq data only, we compared the precision and recall of inferred links validated by HiC data. The precision and recall are defined as the ratio of inferred connections in HiC and the ratio of HiC in the inferred connections for each gene, respectively. We found that precision and, to a lesser extent, recall are higher when using paired scRNA-seq and scATAC-seq compared to scATAC-seq only (fig. S16). In addition, integrating scRNA-seq and scATAC-seq data can improve the power of finding cellular heterogeneity compared to using only one of them, as shown in our previous study (*8*). Moreover, the paired scRNA-seq and scATAC-seq data allow inference of the context-based regulations, while using scATAC-seq alone may lead to inferred regulations without knowing the causal relationship. The paired data usually enable more accurate identification of cell states and context-based regulations when studying gene regulations in cell fate decisions.

Although machine learning methods have been proposed on bulk level for identifying CREs and constructing TRNs (*2*, *5*, *7*, *8*, *43*), directly applying them to single-cell data might be challenging due to the extremely sparse and heterogeneous characteristics of single-cell data. DIRECT-NET first generates data of "pseudo-cells" to alleviate the sparse effects and then uses the fast ensemble machine model XGBoost, which has been successfully applied to many problems (*44*, *45*). We treat the identification of functional CREs as a prediction problem by assessing the contribution of each CRE's chromatin accessibility to the promoter's accessibility or target gene's expression. In contrast to the pairwise linear correlation–based method Cicero, our nonlinear regression–based method shows superior performance in detecting functional CREs in terms of AUC.

Covariability-based methods (e.g., Cicero) (*16*–*18*) and DIRECT-NET have one underlying assumption: The chromatin accessibility needs to vary across cells significantly. To evaluate whether a cell state has sufficient heterogeneity, computational methods such as an entropy-based metric ROGUE could be used (Supplementary Text and fig. S17) (*46*). The ROGUE values of individual cell states are much larger than those of PBMCs, indicating less heterogeneity in individual cell states than in all cells (fig. S17A). In the continuous A549 data from sci-CAR technique, cells in each time point still exhibit higher ROGUE values than all cells across the three time points (fig. S17B). However, when a subset of enhancers and promoters has relative high heterogeneity among the cells of a cell state, the inference may become feasible. We implemented DIRECT-NET and Cicero on each cell state of the A549 data (0, 1, and 3 hours). Using HiC to validate the findings, we found that, even for individual cell state, some AUCs were still high, which likely correspond to the subset of enhancers and promoters with higher heterogeneity of activity among the cells of the same cell state (fig. S17, C and D). For these covariability-based methods and DIRECT-NET, it is difficult to identify target genes for peaks consistently accessible across all cells. Applying these methods to CREs that have similar activity levels in different cells of the same cell state or even across different cell states could be failed. However, most of these consistently accessible peaks are often associated with highly expressed genes for all cell states, such as the housekeeping genes. In addition, those genes and peaks are most likely not cell state–specific and less critical in determining cell fates. As for those consistently accessible chromatin regions, one can add additional information (e.g., distance to TSS and HiC if available) to supplement the predictions made by DIRECT-NET, such as those used in Cicero.

Correlated CRE selection is a multicollinearity problem, a common challenge to many current methods such as Cicero, ArchR, and SnapATAC. DIRECT-NET uses the gradient boosting machine model that is relatively better immune to the multicollinearity problem compared to other methods such as Pearson correlation, Gaussian graphical model, and random forest (*30*). Different from random forest where each tree is independent from the others, the tree in gradient boosting machine focuses its learning on what has not been well modeled by its predecessor. Thus, DIRECT-NET usually does not assign lower importance scores for all correlated CREs, leading to identifying a subset of these correlated CREs that have important roles in the prediction. A post hoc analysis could help to rescue the highly correlated CREs based on the correlation analysis (Materials and Methods). Moreover, because phenotype is robust to loss-of-function mutations in individual correlated enhancers or even redundant enhancers (*47*, *48*), the identified subset of correlated CREs will unlikely affect biological discovery.

In DIRECT-NET, peaks were called by MACS2 on merged cells. Some packages such as SnapATAC (*3*) and ArchR (*2*) call peaks by MACS2 on aggregated data of each cluster, which is identified on the basis of scATAC-seq or (scATAC-seq and scRNA-seq) at first. However, this may depend on the clustering performance and the rare subpopulation may also be missed. It is more natural that peak calling on aggregated data of each cell state is used when the cell state information is available. Of course, high quality of peak calling will improve the overall performance of the method.

Understanding how different CREs and TFs control gene activation or increase levels of expression is critically important. Compared to the existing methods such as WGCAN (*49*), which takes correlation-based methods on gene expression data to construct gene regulatory network, DIRECT-NET uses a context-based method that allows explicit inclusion of regulatory elements. DIRECT-NET takes advantage of the scATAC-seq data to build cell state–specific regulatory networks. Specifically, for datasets with medium number of cells, one constant CRE-gene linkage across different cell states can be inferred to maximize statistical power (fig. S5A). However, TRNs are only constructed on cell state–specific active peaks, allowing the cell state specificity of TRNs. This is a common and reasonable assumption often used in many other methods, such as Cicero, ArchR, and SnapATAC (*1*–*3*). For datasets with enough number of cells for each cell state, one can run a separate DIRECT-NET model in each cell state, allowing construction of differential CRE-gene linkages even when these CREs are all in different cell states. Other genomic features, such as genetic variants, DNA methylation, and spatial location, can be included in this prediction model to account for more complex regulatory mechanisms.

In this study, we have aimed to extract validation data from individual cell states, such as naïve B cells, naïve CD4 T cells, and naïve CD8 T cells, rather than analyzing only unsorted PBMC data. We observed that the specificity of DIRECT-NET in identifying these validation links exceeded 0.8 in all of these datasets. With the rapid development of technologies for mapping single-cell 3D genome organization, such as single-cell HiC, more data will become available to benchmark and optimize methods. While there is still much room for improvement, DIRECT-NET already provides an effective way to detect functional CREs and build gene regulatory networks, addressing an urgent need for extracting information on cell heterogeneity and regulatory mechanisms from single-cell data.

MATERIALS AND METHODS

Framework of DIRECT-NET

The computational method for DIRECT-NET, a tool to predict HC functional CREs from parallel single-cell chromatin accessibility data and gene expression data or single-cell chromatin accessibility data only, consists of three main steps: (i) aggregation of sparse single-cell data, (ii) identification of HC functional CREs, and (iii) reconstruction of gene regulatory networks. Below, we describe each of these steps in detail.

Aggregation of sparse single-cell data

The sparse scATAC-seq data and scRNA-seq data are aggregated by averaging signals of similar cells, which are learned from a *k*-nearest neighbor (KNN) graph (default $k = 50$) constructed from a low-dimensional representation of the data. When only scATAC-seq data are available, the low-dimensional representation is learned by performing singular value decomposition on a transformed single-cell chromatin accessibility peak count matrix via the term frequency–inverse document frequency (TF-IDF) (*50*). Notably, the first singular component is removed because of its high correlation with technical factors such as sequencing depths. The first 40 components are used for constructing a KNN graph. When parallel scRNA-seq and scATAC-seq data are available, a single unified low-dimensional representation of single-cell multiomics data is learned by a weighted nearest neighbor (WNN) analysis, which has been implemented in Seurat V4 package (*23*). WNN analysis produces a single similarity metric between any two cells based on a weighted combination of chromatin accessibility and gene expression similarities. We use a similar aggregation approach as in Cicero after constructing the KNN or WNN graph by obtaining the KNNs of each cell. We then identify the maximum number of cells with their KNNs in which the overlap ratio of any two cells is less than *over_rate* (default 0.8). The overlap ratio is defined as the number of common neighbors divided by *k*. The detailed workflow for finding such a set of cells is shown in fig. S18. We obtain the aggregated data of each cell in this identified set by summarizing the epigenomic profiles of its KNNs. Last, the aggregated chromatin accessibility data and gene expression data are normalized using a set of scaling factors, which are estimated by the estimateSizeFactors function implemented in DESeq2 package (*51*).

Identification of functional CREs

To identify functionally distinct CREs from single-cell chromatin accessibility data, DIRECT-NET builds a regression model based on the framework of XGBoost, which is an improved gradient boosting machine model with more accurate approximations for finding the optimal decision trees (*30*). When only scATAC-seq data are available, for each promoter (peak within 500 bp upstream of TSS), the XGBoost model is used to regress its accessibility using the accessibility values of candidate distal CREs (peaks within 250 kbp upstream and downstream of TSS). Similarly, when parallel scRNA-seq and scATAC-seq data are available, the XGBoost model is used to regress a gene's expression using the accessibility values of candidate distal CREs. Specifically, let X represent scATAC-seq data matrix with q loci across n cells, and Y represent scRNA-seq data matrix with p genes across n cells. Both X and Y have been aggregated and normalized based on the above procedure described in the previous section. In more detail, $\mathbf{y_i}$ represents gene expression levels of gene i or the accessibility values of the promoter of gene i if scRNA-seq is not available. If there are more than one peak defined as promoters, we use the average values across all promoters of gene i. Let $\mathbf{x_i^1}, \mathbf{x_i^2}, \cdots, \mathbf{x_i^k}$ represent k open chromatin regions within the 250 kilo–base pairs (kbp) upstream and downstream of the TSS of gene i. Then, DIRECT-NET learns the model $\widehat{F}_i$ for gene i with

$$\widehat{F}_i = \underset{F}{\operatorname{argmin}} \|\mathbf{y_i} - F(\mathbf{x_i^1}, \mathbf{x_i^2}, \cdots, \mathbf{x_i^k})\|_2.$$

The XGBoost model, an efficient implementation of the Gradient Boost Trees algorithm (*30*), is used here. Specially, the objective function at the tth iteration is as follows

$$L^{(t)} = \|\mathbf{y_i} - (\hat{\mathbf{y}}_\mathbf{i}^{(\mathbf{t-1})} + F_t(\mathbf{x_i^1}, \mathbf{x_i^2}, \cdots, \mathbf{x_i^k}))\|_2 + \frac{1}{2}\|\mathbf{w^t}\|^2,$$

where $\mathbf{w^t}$ represents the weights in the tth step linear regression model. The first term represents the difference between true gene expression or promoter accessibility and predicted value with peaks within 250 kbp upstream and downstream of TSS. The second term is the regularization term on peaks within 250 kbp upstream and downstream of TSS. The complexity of regression model in DIRECT-NET is higher than lasso regression. While the XGBoost model is likely overfitting in terms of predicting promotor's accessibility or gene expression, an early stopping strategy may reduce the overfitting phenomenon (Supplementary Text). DIRECT-NET still has good generalization performance in predicting links when applying to the data it never interacts with before (Supplementary Text and fig. S19).

The model produces a set of co-accessibility connections from distal CREs to a gene or a promoter region, which are weighted by importance scores. The importance score of the *j*th variable is computed as $\hat{I}_j^2(T) = \sum_{t=1}^{J-1} \hat{i}_t^2 1(v_t = j)$, where T is a J-terminal node tree of the XGBoost model, v_t is the splitting variable associated with node t, and $\hat{i}_t^2$ is the corresponding improvement in squared error as a result of split (*52*). Then, the importance scores are normalized with the sum of importance scores of all variables equaling 1. It indicates how useful or valuable a distal CRE is in the construction of the boosted decision trees for predicting a gene's expression or promoter's accessibility (*53*). In DIRECT-NET, the functional CREs for each gene are detected independently based on the importance scores of that gene (fig. S20). In DIRECT-NET, HC CREs are defined as a set of peaks with the importance scores higher than the maximum of the median of importance scores and 0.001. LC regions are defined on the basis of their importance scores less than the maximum of the first quantile of importance scores and 0.001. HC CREs are functional CREs that are critical in inferring the expression level of target genes, implying that these HC CREs are likely the important regulators of their target genes. Conversely, LC regions have nearly no regulatory activities of their target genes. MC regions are the remaining regions that are accessible but usually have weak regulatory activities for their target genes. That is, these three categories are defined on the basis of the degree of their influence on the expression levels of their target genes.

DIRECT-NET uncovers CREs on all marker genes (described in the "Datasets and gene selection" section). Functional CREs are more likely to be varying across cell states; however, not every open chromatin region is a functional CRE, as demonstrated by lines of bulk tissue analysis where ATAC-seq peaks only partially overlap with active histone ChIP-seq data and functional validation assays (e.g., MPRA and STARR-seq). In this study, a varying accessible peak is not considered as a functional CRE unless it affects the expression/promoter activities of genes. DIRECT-NET does not assign lower importance scores to all highly correlated CREs, and it usually emphasizes on one of them. When there is a group of correlated CREs, we can use a post hoc strategy to rescue the additional highly correlated CREs. The post hoc strategy is based on the correlation analysis where we treat regions whose absolute Pearson correlation coefficients with identified CREs are higher than 0.25 as rescued highly correlated CREs (Supplementary Text).

Fraction of rare variants

The rare variant fraction r is computed as $r = Nr/(Nr + Nc)$, where Nr is the number of rare variants in functional CREs and Nc is the number of common variants in functional CREs. Thus, this number represents the proportion of rare variants over the union of common and rare variants.

LD score regression

The stratified LD score regression method is applied to sets of cluster-specific HC CREs, MC regions, and LC regions to identify disease-relevant clusters for Alzheimer's disease and some neural degenerative diseases, downloaded from the PGC website (www.med.unc.edu/pgc/results-and-downloads/). The LD method implementation is based on the LD score regression tutorial (https://github.com/bulik/ldsc/wiki).

Reconstruction of gene regulatory networks

To reconstruct cell state–specific gene regulatory networks, we first identify differentially accessible HC CREs of the cell state–specific marker genes. The differentially accessible HC CREs are the overlapped peaks between HC CREs and differentially accessible peaks, identified by a logistic regression model as suggested by a previous study (*54*). Then, we identify the enriched TFs in the differentially accessible HC CREs using the motifmatchr function in ChromVAR package (*55*). For each cell state, we consider the linkages only if their target genes are the markers for the corresponding cell state. As shown in fig. S5A, G1 is a marker gene of cell state *A*, while G2 is a marker gene of cell state *B*. Because G1 is not active in cell state *B*, the linkage TF2→R2→G1 does not appear in cell state *B*. Our resultant network will be TF1→R1→G1 and TF3→R3→G2. Last, TRN was built by integrating the relationships between CREs and their target genes with the relationships between CREs and TFs. The network was visualized by Cytoscape software (*56*).

Gene selection

Nineteen cell states were detected in previous study on the PBMC dataset (*23*). We adopt presto package (*57*) to perform fast differential expressional expression analysis, which returns an AUC statistic value to represent the power of each gene as a marker of cell state. We selected markers of each cell state, with AUC values being higher than 0.5. In total, there are 13,016 marker genes across all 19 cell states.

Because the time point information of the A549 dataset is known, we focused on differentially expressed genes across time points. A total of 1185 differentially expressed genes were identified using Wilcoxon rank test with $\log_2$-transformed fold change higher than 0.1 and the percent of expressed cells greater than 5%.

For the GM12878 dataset, we first converted the peaks from hg19 to hg38. We then adopted MAESTRO package (*25*) to convert scATAC-seq data to gene activity score data and further detected 4911 highly variable genes.

For the Brain dataset, there are 70,631 individual cells and 24 distinct clusters, identified based on unbiased iterative clustering and Harmony-based batch correction method in the previous study (*41*). There are 66,982 cells left after removing cells in the unclassified clusters. Then, we aggregated scATAC-seq based on the remaining 18 cell clusters and obtain 12,259 pseudo-bulk cells. A total of 5181 marker genes across all clusters were detected on gene activity matrix transformed by MAESTRO package (*25*).

Validation of inferred connections by PCHiC, ChIA-PET, HiC, and HiChIP data

For the GM12878 dataset, we used public Pol II CHIA-PET data and promoter-capture HiC data (GSE72816) (*58*) to validate DIRECT-NET's predicted connections. For Brain data, we downloaded loops of HiChIP data from the supplementary table of the previous study (*41*). For A549 dataset, we downloaded HiC data of A549 cell line from ENCODE and transformed the data from hg38 to hg19 using UCSC liftover tool. For PBMC dataset, we downloaded HiC data of naïve CD4 T cells, naïve CD8 T cells, and native B cells from a previous study (*31*).

To compare the predicted connections with HiC and CHIA-PET data, we identified the overlapped peaks with HiC and CHIA-PET separately. When the two peaks are within 1 kb of each other, we call them overlapped peaks. To focus on distal-proximal connections, we only retained the pairs of scATAC-seq peaks where at least one peak is a promoter and another peak is an enhancer represented in the HiC or CHIA-PET data. The overlapped anchors in CHIA-PET data were merged to create comparable CHIA-PET peaks using the loopsMake function in diffloops package (*59*), with parameter mergegap being zero. If one peak of a HiC link is within 500 bp upstream of the marker gene's TSS and another peak is within 251 kb both sides of the TSS, then this link is used for validation. The same criterion is applied to PCHiC and ChIA-PET data. The numbers of validated links of ChIA-PET, HiC, or PCHiC used on each data are listed in table S1. The procedures of Calculating the ratios of connections validated by PCHiC, HiC, ChIA-PET, or HiChIP and ratios of PCHiC, HiC, ChIA-PET, or HiChIP links recalled are in Supplementary Text ("Calculation of ratios of connections validated by PCHiC, HiC, ChIA-PET or HiChIP and ratios of PCHiC, HiC, ChIA-PET or HiChIP links recalled").

ChIP-seq signals

For A549 data, we downloaded bigwig files of NR3C1 ChIP-seq data on 0, 1, and 3 hours of 100 nM DEX treatment from ENCODE and converted the available hg38 version to hg19 using CrossMap tool (*60*). We then detected the NR3C1 signals of HC CREs, MC regions, and LC regions across the three time points using bigWigAverageOverBed from UCSC Genome Browser.

Comparison with other methods

CloseGene treats the closest gene as the peak's target gene. Distance treats genomic distance as a simple predictor of CRE-gene interaction. Let d_i represent the distance between TSS of a gene and the ith peak within the neighborhood window (500 kb) of TSS, and d be the union of d_i. In the Distance method, we treated $(\max(d) - d_i)$ as the weight of the link between the ith peak and the corresponding gene. We implemented CloseGene and Distance methods on all peaks that are within the regions that are 250 kb upstream and downstream of marker genes' TSS but not 500 bp upstream of marker genes' TSS. We computed SCC between accessible signals of promoters and peaks within the regions that are 250 kb upstream and downstream of marker genes' TSS but not 500 bp upstream of marker genes' TSS on GM12878 and Brain datasets, and we used gene expression levels instead of accessible signals of promoters on PBMC and A549 datasets. We regressed the accessible signals of promoters (or gene expression) using accessible signals of peaks by Lasso, Ridge, and ElasticNet regression methods on the data after the aggregation for the inputs.

Cicero (*16*) was run on the basis of the tutorial (https://cole-trapnell-lab.github.io/cicero-release/docs_m3/#constructing-cis-regulatory-networks). To ensure that the predicted CREs are more comparable, we retained connections predicted by Cicero between peaks (250 kb upstream and downstream of TSS) and promoters (500 bp upstream of TSS) of each gene.

ArchR was run based on the tutorial (www.archrproject.com/bookdown/index.html).

To ensure that the predicted CREs are more comparable, we set the parameter maxDist equal 500000. The addPeak2GeneList function was used on PBMC and A549 datasets, while the AddCoAccessibility function was used on GM12878 and Brain datasets.

SnapATAC needs both gene expression and chromatin accessibility data to predict gene-enhancer pairs. We applied it on A549 and PBMC datasets based on the tutorial (https://github.com/r3fang/SnapATAC/blob/master/examples/10X_PBMC_15K/README.md#gene_peak_pair). All the comparison codes are available from the GitHub link (https://github.com/zhanglhbioinfor/DIRECT-NET) and Zenodo link (DOI: 10.5281/zenodo.5821082).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at https://science.org/doi/10.1126/sciadv.abl7393

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. A. Casamassimi, A. Ciccodicola, Transcriptional regulation: Molecules, involved mechanisms, and misregulation. *Int. J. Mol. Sci.* **20**, 1281 (2019).
2. L. Tang, M. C. Hill, J. Wang, J. Wang, J. F. Martin, M. Li, Predicting unrecognized enhancer-mediated genome topology by an ensemble machine learning model. *Genome Res.* **30**, 1835–1845 (2020).
3. R. Jothi, S. Balaji, A. Wuster, J. A. Grochow, J. Gsponer, T. M. Przytycka, L. Aravind, M. M. Babu, Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture. *Mol. Syst. Biol.* **5**, 294 (2009).
4. S. L. Klemm, Z. Shipony, W. J. Greenleaf, Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* **20**, 207–220 (2019).
5. A. Sethi, M. Gu, E. Gumusgoz, L. Chan, K. K. Yan, J. Rozowsky, I. Barozzi, V. Afzal, J. A. Akiyama, I. Plajzer-Frick, C. Yan, C. S. Novak, M. Kato, T. H. Garvin, Q. Pham, A. Harrington, B. J. Mannion, E. A. Lee, Y. Fukuda-Yuzawa, A. Visel, D. E. Dickel, K. Y. Yip, R. Sutton, L. A. Pennacchio, M. Gerstein, Supervised enhancer prediction with epigenetic pattern recognition and targeted validation. *Nat. Methods* **17**, 807–814 (2020).
6. Z. Chen, J. Zhang, J. Liu, Y. Dai, D. Lee, M. R. Min, M. Xu, M. Gerstein, DECODE: A deep-learning framework for condensing enhancers and refining boundaries with large-scale functional assays. *Bioinformatics* **37**, i280–i288 (2021).
7. J. Zhang, D. Lee, V. Dhiman, P. Jiang, J. Xu, P. McGillivray, H. Yang, J. Liu, W. Meyerson, D. Clarke, M. Gu, S. Li, S. Lou, J. Xu, L. Lochovsky, M. Ung, L. Ma, S. Yu, Q. Cao, A. Harmanci, K.-K. Yan, A. Sethi, G. Gürsoy, M. R. Schoenberg, J. Rozowsky, J. Warrell, P. Emani, Y. T. Yang, T. Galeev, X. Kong, S. Liu, X. Li, J. Krishnan, Y. Feng, J. C. Rivera-Mulia, J. Adrian, J. R. Broach, M. Bolt, J. Moran, D. Fitzgerald, V. Dileep, T. Liu, S. Mei, T. Sasaki, C. Trevilla-Garcia, S. Wang, Y. Wang, C. Zang, D. Wang, R. J. Klein, M. Snyder, D. M. Gilbert, K. Yip, C. Cheng, F. Yue, X. S. Liu, K. P. White, M. Gerstein, An integrative ENCODE resource for cancer genomics. *Nat. Commun.* **11**, 3696 (2020).
8. S. R. Kulkarni, D. Vaneechoutte, J. Van de Velde, K. Vandepoele, TF2Network: Predicting transcription factor regulators and gene regulatory networks in Arabidopsis using publicly available binding site information. *Nucleic Acids Res.* **46**, e31 (2018).
9. V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, P. Geurts, Inferring regulatory networks from expression data using tree-based methods. *PLOS ONE* **5**, e12776 (2010).
10. E. Rojano, P. Seoane, J. A. G. Ranea, J. R. Perkins, Regulatory variants: From detection to predicting impact. *Brief. Bioinform.* **20**, 1639–1654 (2019).
11. E. Giacopuzzi, N. Popitsch, J. Taylor, GREEN-DB: A framework for the annotation and prioritization of non-coding regulatory variants from whole-genome sequencing data. *Nucl. Acids Res.* **50**, 2522–2535 (2020).
12. C. Trapnell, Defining cell types and states with single-cell genomics. *Genome Res.* **25**, 1491–1498 (2015).
13. H. Clevers, What is your conceptual definition of "cell type" in the context of a mature organism? *Cell Syst.* **4**, 255–259 (2017).
14. J. D. Buenrostro, B. Wu, U. M. Litzenburger, D. Ruff, M. L. Gonzales, M. P. Snyder, H. Y. Chang, W. J. Greenleaf, Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
15. D. A. Cusanovich, R. Daza, A. Adey, H. A. Pliner, L. Christiansen, K. L. Gunderson, F. J. Steemers, C. Trapnell, J. Shendure, Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
16. H. A. Pliner, J. S. Packer, J. L. McFaline-Figueroa, D. A. Cusanovich, R. M. Daza, D. Aghamirzaie, S. Srivatsan, X. Qiu, D. Jackson, A. Minkina, A. C. Adey, F. J. Steemers, J. Shendure, C. Trapnell, Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol. Cell* **71**, 858–871.e8 (2018).
17. J. M. Granja, M. R. Corces, S. E. Pierce, S. T. Bagdatli, H. Choudhry, H. Y. Chang, W. J. Greenleaf, Author Correction: ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 935 (2021).
18. R. Fang, S. Preissl, Y. Li, X. Hou, J. Lucero, X. Wang, A. Motamedi, A. K. Shiau, X. Zhou, F. Xie, E. A. Mukamel, K. Zhang, Y. Zhang, M. M. Behrens, J. R. Ecker, B. Ren,

Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat. Commun.* **12**, 1337 (2021).

19. J. Cao, D. A. Cusanovich, V. Ramani, D. Aghamirzaie, H. A. Pliner, A. J. Hill, R. M. Daza, J. L. McFaline-Figueroa, J. S. Packer, L. Christiansen, F. J. Steemers, A. C. Adey, C. Trapnell, J. Shendure, Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).
20. L. Liu, C. Liu, A. Quintero, L. Wu, Y. Yuan, M. Wang, M. Cheng, L. Leng, L. Xu, G. Dong, R. Li, Y. Liu, X. Wei, J. Xu, X. Chen, H. Lu, D. Chen, Q. Wang, Q. Zhou, X. Lin, G. Li, S. Liu, Q. Wang, H. Wang, J. L. Fink, Z. Gao, X. Liu, Y. Hou, S. Zhu, H. Yang, Y. Ye, G. Lin, F. Chen, C. Herrmann, R. Eils, Z. Shang, X. Xu, Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. *Nat. Commun.* **10**, 470 (2019).
21. C. Zhu, M. Yu, H. Huang, I. Juric, A. Abnousi, R. Hu, J. Lucero, M. M. Behrens, M. Hu, B. Ren, An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat. Struct. Mol. Biol.* **26**, 1063–1070 (2019).
22. S. Ma, B. Zhang, L. M. La Fave, A. S. Earl, Z. Chiang, Y. Hu, J. Ding, A. Brack, V. K. Kartha, T. Tay, T. Law, C. Lareau, Y.-C. Hsu, A. Regev, J. D. Buenrostro, Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* **183**, 1103–1116.e20 (2020).
23. Y. Hao, S. Hao, E. Andersen-Nissen, W. M. Mauck III, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, P. Hoffman, M. Stoeckius, E. Papalexi, E. P. Mimitou, J. Jain, A. Srivastava, T. Stuart, L. M. Fleming, B. Yeung, A. J. Rogers, J. M. McElrath, C. A. Blish, R. Gottardo, P. Smibert, R. Satija, Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
24. J. D. Welch, V. Kozareva, A. Ferreira, C. Vanderburg, C. Martin, E. Z. Macosko, Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* **177**, 1873–1887.e17 (2019).
25. C. Wang, D. Sun, X. Huang, C. Wan, Z. Li, Y. Han, Q. Qin, J. Fan, X. Qiu, Y. Xie, C. A. Meyer, M. Brown, M. Tang, H. Long, T. Liu, X. S. Liu, Integrative analyses of single-cell transcriptome and regulome using MAESTRO. *Genome Biol.* **21**, 198 (2020).
26. J. D. Welch, A. J. Hartemink, J. F. Prins, MATCHER: Manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome Biol.* **18**, 138 (2017).
27. Z. Duren, X. Chen, M. Zamanighomi, W. Zeng, A. T. Satpathy, H. Y. Chang, Y. Wang, W. H. Wong, Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 7723–7728 (2018).
28. S. Jin, L. Zhang, Q. Nie, scAI: An unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. *Genome Biol.* **21**, 25 (2020).
29. K. E. Wu, K. E. Yost, H. Y. Chang, J. Zou, BABEL enables cross-modality translation between multiomic profiles at single-cell resolution. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2023070118 (2021).
30. T. Q. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in *KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, 2016), pp. 785–794.
31. B. M. Javierre, O. S. Burren, S. P. Wilder, R. Kreuzhuber, S. M. Hill, S. Sewitz, J. Cairns, S. W. Wingett, C. Várnai, M. J. Thiecke, F. Burden, S. Farrow, A. J. Cutler, K. Rehnström, K. Downes, L. Grassi, M. Kostadima, P. Freire-Pritchett, F. Wang, Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell* **167**, 1369–1384.e19 (2016).
32. T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert, R. Satija, Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).
33. J. Zhang, J. Liu, D. Lee, J. J. Feng, L. Lochovsky, S. Lou, M. Rutenberg-Schoenberg, M. Gerstein, RADAR: Annotation and prioritization of variants in the post-transcriptional regulome of RNA-binding proteins. *Genome Biol.* **21**, 151 (2020).
34. E. Khurana, Y. Fu, V. Colonna, X. J. Mu, H. M. Kang, T. Lappalainen, A. Sboner, L. Lochovsky, J. Chen, A. Harmanci, J. Das, A. Abyzov, S. Balasubramanian, K. Beal, D. Chakravarty, D. Challis, Y. Chen, D. Clarke, L. Clarke, F. Cunningham, U. S. Evani, P. Flicek, R. Fragoza, E. Garrison, R. Gibbs, Z. H. Gümüş, J. Herrero, N. Kitabayashi, Y. Kong, K. Lage, V. Liluashvili, S. M. Lipkin, D. G. MacArthur, G. Marth, D. Muzny, T. H. Pers, G. R. S. Ritchie, J. A. Rosenfeld, C. Sisu, X. Wei, M. Wilson, Y. Xue, F. Yu; 1000 Genomes Project Consortium, E. T. Dermitzakis, H. Yu, M. A. Rubin, C. Tyler-Smith, M. Gerstein, Integrative annotation of variants from 1092 humans: Application to cancer genomics. *Science* **342**, 1235587 (2013).
35. A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L.-D. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, D. Haussler, Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
36. B. He, S. Xing, C. Chen, P. Gao, L. Teng, Q. Shan, J. A. Gullicksrud, M. D. Martin, S. Yu, J. T. Harty, V. P. Badovinac, K. Tan, H.-H. Xue, $CD8^+$ T cells utilize highly dynamic enhancer repertoires and regulatory circuitry in response to infections. *Immunity* **45**, 1341–1354 (2016).
37. B. J. Laidlaw, J. G. Cyster, Transcriptional regulation of memory B cell differentiation. *Nat. Rev. Immunol.* **21**, 209–220 (2021).
38. K. R. Moon, D. van Dijk, Z. Wang, S. Gigante, D. B. Burkhardt, W. S. Chen, K. Yim, A. van den Elzen, M. J. Hirn, R. R. Coifman, N. B. Ivanova, G. Wolf, S. Krishnaswamy, Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.* **37**, 1482–1492 (2019).
39. N. Z. Lu, S. E. Wardell, K. L. Burnstein, D. Defranco, P. J. Fuller, V. Giguere, R. B. Hochberg, L. M. Kay, J.-M. Renoir, N. L. Weigel, E. M. Wilson, D. P. McDonnell, J. A. Cidlowski, International Union of Pharmacology. LXV. The pharmacology and classification of the nuclear receptor superfamily: Glucocorticoid, mineralocorticoid, progesterone, and androgen receptors. *Pharmacol. Rev.* **58**, 782–797 (2006).
40. S. C. Biddie, S. John, P. J. Sabo, R. E. Thurman, T. A. Johnson, R. L. Schiltz, T. B. Miranda, M.-H. Sung, S. Trump, S. L. Lightman, C. Vinson, J. A. Stamatoyannopoulos, G. L. Hager, Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding. *Mol. Cell* **43**, 145–155 (2011).
41. M. R. Corces, A. Shcherbina, S. Kundu, M. J. Gloudemans, L. Frésard, J. M. Granja, B. H. Louie, T. Eulalio, S. Shams, S. T. Bagdatli, M. R. Mumbach, B. Liu, K. S. Montine, W. J. Greenleaf, A. Kundaje, S. B. Montgomery, H. Y. Chang, T. J. Montine, Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. *Nat. Genet.* **52**, 1158–1168 (2020).
42. R. N. Doan, E. T. Lim, S. De Rubeis, C. Betancur, D. J. Cutler, A. G. Chiocchetti, L. M. Overman, A. Soucy, S. Goetze; Autism Sequencing Consortium, C. M. Freitag, M. J. Daly, C. A. Walsh, J. D. Buxbaum, T. W. Yu, Recessive gene disruptions in autism spectrum disorder. *Nat. Genet.* **51**, 1092–1098 (2019).
43. T. A. Hait, D. Amar, R. Shamir, R. Elkon, FOCS: A novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer–promoter map. *Genome Biol.* **19**, 56 (2018).
44. X. Ji, W. Tong, Z. Liu, T. Shi, Five-feature model for developing the classifier for synergistic vs. antagonistic drug combinations built by XGBoost. *Front. Genet.* **10**, 600 (2019).
45. B. Yu, W. Qiu, C. Chen, A. Ma, J. Jiang, H. Zhou, Q. Ma, SubMito-XGBoost: Predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting. *Bioinformatics* **36**, 1074–1081 (2020).
46. B. Liu, C. Li, Z. Li, D. Wang, X. Ren, Z. Zhang, An entropy-based metric for assessing the purity of single cell populations. *Nat. Commun.* **11**, 3155 (2020).
47. M. Osterwalder, I. Barozzi, V. Tissières, Y. Fukuda-Yuzawa, B. J. Mannion, S. Y. Afzal, E. A. Lee, Y. Zhu, I. Plajzer-Frick, C. S. Pickle, M. Kato, T. H. Garvin, Q. T. Pham, A. N. Harrington, J. A. Akiyama, V. Afzal, J. Lopez-Rios, D. E. Dickel, A. Visel, L. A. Pennacchio, Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554**, 239–243 (2018).
48. E. Z. Kvon, R. Waymack, M. Gad, Z. Wunderlich, Enhancer redundancy in development and disease. *Nat. Rev. Genet.* **22**, 324–336 (2021).
49. B. Zhang, S. Horvath, A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, Article17 (2005).
50. T. Stuart, A. Srivastava, C. Lareau, R. Satija, Multimodal single-cell chromatin analysis with Signac. bioRxiv 2020.11.09.373613 [Preprint]. 10 November 2020. https://doi.org/10.1101/2020.11.09.373613.
51. S. Anders, W. Huber, Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
52. J. H. Friedman, Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
53. J. Elith, J. R. Leathwick, T. Hastie, A working guide to boosted regression trees. *J. Anim. Ecol.* **77**, 802–813 (2008).
54. V. Ntranos, L. Yi, P. Melsted, L. Pachter, A discriminative learning approach to differential expression analysis for single-cell RNA-seq. *Nat. Methods* **16**, 163–166 (2019).
55. A. N. Schep, B. Wu, J. D. Buenrostro, W. J. Greenleaf, chromVAR: Inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
56. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
57. I. Korsunsky, A. Nathan, N. Millard, S. Raychaudhuri, Presto scales Wilcoxon and auROC analyses to millions of observations. bioRxiv 653253 [Preprint]. 29 May 2019.https://doi.org/10.1101/653253.
58. Z. Tang, O. J. Luo, X. Li, M. Zheng, J. J. Zhu, P. Szalaj, P. Trzaskoma, A. Magalska, J. Wlodarczyk, B. Ruszczycki, P. Michalski, E. Piecuch, P. Wang, D. Wang, S. Z. Tian, M. Penrad-Mobayed, L. M. Sachs, X. Ruan, C. L. Wei, E. T. Liu, G. M. Wilczynski, D. Plewczynski, G. Li, Y. Ruan, CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* **163**, 1611–1627 (2015).
59. C. A. Lareau, M. J. Aryee, diffloop: A computational framework for identifying and analyzing differential DNA loops from sequencing data. *Bioinformatics* **34**, 672–674 (2018).

60. H. Zhao, Z. Sun, J. Wang, H. Huang, J.-P. Kocher, L. Wang, CrossMap: A versatile tool fo coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006–1007 (2014).

Acknowledgments: We thank C. Lee for helpful discussions. **Funding:** This work was supported by NSF grant DMS1763272; Simons Foundation grant 594598 (to Q.N.); NIH grants U01AR073159, P30AR07504, K01MH123896, and U01DA053628; and an UCI ICS research exploration grant. **Author contributions:** J.Z. and Q.N. conceived the project. L.Z. and J.Z. conducted the research. Q.N. and J.Z. supervised the research. L.Z., J.Z., and Q.N. wrote the manuscript. **Competing interests:** The authors declare that they have no competing interests.

Data and materials availability: All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The datasets analyzed in this study are publicly available. The PBMC dataset was downloaded from the 10X Genomics website including 11,909 cells with gene expression and chromatin accessibility profiles measured in the same individual cells (https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k). The A549 dataset was downloaded from GEO (accession codes: GSM3271040 and GSM3271041), including 2641 lung adenocarcinoma–derived A549 cells after 0, 1, and 3 hours of 100 nM dexamethasone treatment, where parallel gene expression and chromatin accessibility profiles are available. The GM12878 dataset was accessed from a previous study, including single-cell chromatin accessibility data of 889 human lymphoblastoid GM12878 cells (*16*). The Brain dataset was downloaded from GEO (accession code: GSE147672) (*41*). DIRECT-NET is publicly available as an R package. Source codes and tutorials have been deposited at the GitHub repository (https://github.com/zhanglhbioinfor/DIRECT-NET) and Zenodo (DOI: 10.5281/zenodo.5821082).

Submitted 3 August 2021
Accepted 14 April 2022
Published 1 June 2022
10.1126/sciadv.abl7393