Supplementary Text

A correlation-based strategy to rescue highly correlated CREs

Taking marker genes that had only one predicted CRE on GM12878 dataset as an example, we examined 82 among 4911 marker genes that had only one predicted CRE, and detected their highly correlated CREs with the absolute Pearson correlation coefficients higher than 0.25. This analysis resulted in an average of 7 highly correlated CREs for each gene. Moreover, these highly correlated CREs showed strong signals of H3K27ac (Supplementary Figure S10 E). These results indicated that post-hoc analysis could help to rescue the highly correlated CREs based on the correlation analysis. We provided an option in the codes of our package to allow the identification of such correlated CREs of all markers.

Evaluation of cellular heterogeneity on PBMC and A549 datasets

ROGUE has been introduced recently for assessing the purity of cell states(62). A cell state with ROGUE value equaling 1 has no cellular heterogeneity, while a cell state with ROGUE value equaling 0 has very large cellular heterogeneity. Taking PBMC dataset with paired scRNA-seq and scATAC-seq data as an example, we found that each cell state has larger ROGUE value (0.8-0.9) than the ROGUE value of all cells (0.4-0.6) (Supplementary Figure S17A), thus confirming that all cells have higher cellular heterogeneity than cells in one cell state. For most cases, ROGUE was shown to accurately assess the purity of cell states for a wide range of simulated and real datasets(62). However, the threshold for determining whether a cell state has sufficient cellular heterogeneity may vary across different protocols and platforms. For example, in the continuous A549 data from sci-CAR technique, while cells in each time point still exhibit higher ROGUE values than all cells across the three time points (Supplementary Figure S17B), the specific ROGUE values are larger than those in the PBMC dataset.

Evaluation of the overfitting performance of DIRECT-NET

In DIRECT-NET, we adopted ridge regularization term in each step to reduce overfitting. We also provided the following early stopping strategy to reduce overfitting. First, we divided the data into five folds with the proportion of training, validation, and testing set equaling 3:1:1. We first computed the differences of RMSEs between the training and validation data and selected the iteration at which the difference of RMSE had minimum change. Then we used the selected iteration to determine the early stop and tested the performance on testing data. With this early stopping strategy, we found that there is a small difference between RMSEs of the training and
testing data (Supplementary Figure S19A), suggesting the early stop indeed help to avoid overfitting problem. Finally, we compared the AUC values of DIRECT-NET with this early stop strategy with DIRECT-NET without this strategy and found that the AUC values were comparable (Supplementary Figure S19B). Because this strategy takes more time, we added an option for users to choose whether to use the early stop strategy or not in the updated package.

Calculation of ratios of connections validated by PCHiC, HiC, ChIA-PET or HiChiP and ratios of PCHiC, HiC, ChIA-PET or HiChiP links recalled

Here we take the naïve B cell state as an example to illustrate how we calculated ratios of connections validated by PCHiC and ratios of recalled PCHiC links.

First, for each marker gene, the distal peaks (chromatin regions outside of the promoter region but within 250kb both sides of TSS) were weighted by the importance scores from DIRECT-NET. The distal peaks of each marker gene were divided into five groups (> Q3, Q2-Q3, Q1-Q2, 0-Q1, =0) based on their assigned weights.

Second, for each marker in each group, we computed ratios of connections (between distal peaks and promoters) validated using PCHiC and ratios of recalled PCHiC links as follows. Suppose $A_{i1}, A_{i2}, A_{i3}, A_{i4}$ and $A_{i5}$ represent links between promoter and distal peaks for marker gene $i$ in these five groups, $B_i$ represents PCHiC links of marker gene $i$. Then for marker gene $i$ of the $j$-th group, ratio of connections validated by PCHiC was obtained by $r_i = \frac{|A_{ij} \cap B_i|}{|A_{ij}|}$, where $|A_{ij} \cap B_i|$ is the number of overlapped links. When the links of PCHiC and the distal peaks are within 1kb, we considered them as overlapped links. Ratio of recalled PCHiC links for marker gene $i$ of the $j$-th group is calculated by $recall_{ij} = \frac{|A_{ij} \cap B_i|}{|B_i|}$. Supplementary Figure S1A-B shows the boxplot of results of Naïve B, CD4 T Naïve and CD8 T Naïve cell states. Each element represents ratio of connections validated by PCHiC or ratio of recalled PCHiC links.

We explored how the validation ratios varied across different distances to TSS. We first computed the distance of these peaks to the TSS of corresponding markers. Then the peaks of all markers were divided into bins with every 1k bp for each group. Next, we computed the ratio of validated connections using PCHiC and the ratio of PCHiC links recalled for each bin using the same procedure. Mean values of every 10 bins with each bin size of 1k bp were computed.
The result is shown in Supplementary Figure S1D. We also computed the ratios for every 10k bp directly (Supplementary Figure S1C). Those calculations support our conclusion in original manuscript that connections with high importance scores indicate high concordance with the PChIC links.
Figure S1. Validation of links with different importance scores on paired scRNA-seq and scATAC-seq data of PBMC. (A) Boxplot of ratios of links validated by PChIC. Each element in the box represents
one gene. There are 158, 71 and 148 markers for Naïve B, CD4 T Naïve and CD8 T Naïve, which have at least one PChIC links. P-value is calculated by paired student’s t test. (B) Boxplot of ratios of PChIC links recalled with different threshold values of importance scores of Naïve B, CD4 naïve T and CD8 naïve T cell states individually. (C) Ratios of links validated by PChIC (top) and ratios of PChIC links recalled (bottom) with different threshold values of importance scores. Each dot represents the ratio for every 10kb distance to the TSS. (D) Ratios of links validated by PChIC (top) and ratios of PChIC links recalled (bottom) with different threshold values of importance scores. Each dot represents the ratio for mean values of every 1kb distance to the TSS within every window in 10 kb size. Error bars indicates standard deviation of ratios within 10 kb. (E) Histogram of numbers of validation links across different distance to TSS of PChIC data of Naïve B, CD4 naïve and CD8 naive.
Figure S2. The performance of DIRECT-NET on paired scRNA-seq and scATAC-seq data of PBMC. (A) Ratios of DIRECT-NET-identified high-confidence, medium-confidence and low-confidence links validated by PChIC data over the total number of identified links of marker genes of CD4 naïve T and CD8 naïve T cells. *P*-values are computed from one-sided paired student's *t* test. (B) Ratios of PChIC links found in the DIRECT-NET-predicted high-confidence, medium-confidence and low-confidence links. (C) Comparison of ratios of links validated by merged PChIC data and cell state specific data for markers of Naïve B, Naïve CD4 T and Naïve CD8 T cell states separately. Each element in the boxplot represents one gene. (D) Comparison of ratios of recalled merged PChIC links and ratios of recalled cell state specific PChIC links. (E) Comparison the performance of single cell methods (DIRECT-NET, Cicero, ArchR, SnapATAC), bulk methods (Spearman, lasso, ridge, elastic net) and base methods (Distance, CloseGene) according to AUC value of each marker genes of CD4 naïve T and CD8 naïve T cells on PBMC dataset.
Figure S3. Comparison of overlap between DIRECT-NET and Cicero, ArchR and SnapATAC separately on PBMC dataset. (A) Venn plots of DIRECT-NET identified high importance CREs and Cicero identified high co-accessibility CREs, ArchR identified high co-accessibility CREs, SnapATAC identified high significance CREs and PCHiC links individually of marker genes of naïve B cells, CD4 naïve T cells (B) and CD8 naïve T cells (C). As links within 1kb are treated as overlapped links, the relationship is not one-to-one. (D) Overlap ratios between CREs identified by DIRECT and Cicero, ArchR and SnapATAC of marker genes of naïve B, CD4 naïve T and CD8 naïve T cells. Each element in the boxplot represents the ratio of one gene.
Figure S4. Validation of HC CREs, MC and LC regions by ATAC and aggregated scATAC-seq signals. (A) Average ATAC-seq signals of 1000 bp upstream and 1000 downstream from the middle base of HC CREs, MC and LC regions for Memory B, Naïve B and NK cells using bigWigAverageOverBed. Each value is computed by averaging over 20 bp bin. (B) Average signals of 1000 bp upstream and 1000 downstream from the middle base of HC CREs, MC and LC regions for CD4 TEM, CD14 Mono and NK cells. The signals are aggregated scATAC-seq signals of each corresponding cell state. (C) Average signals of 1000 bp upstream and 1000 downstream from the middle base of HC CREs, MC and LC regions for CD4 TEM, CD14 Mono and NK cells. The signals are aggregated scATAC-seq signals of all cells.
Figure S5. Gene regulatory analysis of CD4 TEM and CD4 TCM cell states on PBMC dataset. (A) The schematic diagram of building cell state specific CRE-gene linkages. G1 is a marker gene of cell state A, while G2 is a marker gene of cell state B. R1 is a high-confidence CRE of G1, while R3 is a high-confidence CRE of G2. Since G1 is not active in cell state B, thus the linkage TF2->R2->G1 does not appear in cell state B. The resultant network will be TF1->R1-G1 and TF3->R3-G2. (B) The inferred gene regulatory networks of CD4 TEM and CD4 TCM (C), respectively. (D) Gene regulatory network of CD4 TEM and CD4 TCM with marker genes and TFs bounded to HC CREs. Genes represented by circles, while TFs represented by rectangles. As we focused on the differential regulations between CD4 TEM and CD4 TCM, we thus extracted HFs that were overlapped with corresponding cell state specific marker peaks. Then we identified TFs bounded to these HC CREs by ChromVAR. Next, the regulatory network was built together to compare these two cell states. Genes or TFs highly expressed in CD4 TEM (fold change > 2) were colored, while TFs highly expressed in CD4 TCM (fold change > 2) were colored differently. Others were colored by grey.
Figure S6. The performance of DIRECT-NET on paired scRNA-seq and scATAC-seq data of A549. (A) Boxplot of ratios of links validated by HiC (left) and Boxplot of ratios of HiC links recalled with different threshold values of importance scores. Each element in the box represents one gene. There are 139 genes, which have at least one HiC links. P-value is calculated by paired student’s t test. (B) Ratios of links validated by HiC (top) and ratios of HiC links recalled (bottom) with different threshold values of importance scores. Each dot represents the ratio for every 10kb distance to the TSS. (C) Ratios of links validated by HiC (top) and ratios of HiC links recalled (bottom) with different threshold values of importance scores. Each dot represents the ratio for mean values of every 1kb distance to the TSS within every window in 10 kb size. Error bars indicates standard deviation of ratios within 10 kb. (D) Boxplot of ratios of DIRECT-NET-identified high-confidence, medium-confidence and low-confidence links validated by HiC data. Each element in the boxplot represents one gene. P-values are computed from one-sided paired student’s t test. (E) Comparison the performance of single cell methods (DIRECT-NET, Cicero), bulk methods (Spearman, lasso, ridge, elastic net) and base methods (Distance, CloseGene) according to AUC value of marker genes of A549. (F) Venn plots of DIRECT-NET identified high importance CREs and Cicero identified high co-accessibility CREs and HiC links. (G) Overlap ratios between CREs identified by DIRECT and Cicero. Each element in the boxplot represents the ratio of one gene. (H) Histogram of numbers of validation links across different distance to TSS of A549 HiC data.
Figure S7. Average NR3C1 signals of 1000 bp upstream and 1000 bp downstream from middle base of HC CREs, MC and LC regions across the three time points using bigWigAverageOverBed. Each value is computed by averaging 20 bp bin. (A) Similar to (B) except for H3K27ac signals.
Figure S8. Gene regulatory network of differently expressed genes and TFs bounded to HC CREs on scRNA-seq and scATAC-seq data of A549.
Figure S9. Validation of links with different importance scores on scATAC-seq data of GM12878 using ChIA-PET data. (A) Boxplot of ratios of links validated by ChIA-PET (left) and Boxplot of ratios of ChIA-PET links recalled (right) with different threshold values of importance scores. Each element in the box represents one gene. There are 1540 markers which have at least one ChIA-PET links. *P*-value is calculated by paired student's t test. (B) Boxplot of ratios of ChIA-PET links recalled by HC CREs, MC and LC regions. Each dot represents the ratio of one gene. (C) Ratios of links validated by ChIA-PET (left) and ratios of ChIA-PET links recalled (right) with different threshold values of importance scores. Each dot represents the ratio for every 10kb distance to the TSS. (D) Ratios of links validated by ChIA-PET (left) and ratios of ChIA-PET links recalled (right) with different threshold values of importance scores. Each dot represents the ratio for mean values of every 1kb distance to the TSS within every window in 10 kb size. Error bars indicates standard deviation of ratios within 10 kb.
Figure S10. Validation of links with different importance scores on scATAC-seq data of GM12878 using HiC data. (A) Boxplot of ratios of links validated by HiC (left) and Boxplot of ratios of HiC links recalled (right) with different threshold values of importance scores. Each element in the box represents one gene. There are 758 markers which have at least one HiC links. P-value is calculated by paired student’s t test. (B) Boxplot of ratios of links validated by HiC (left) and Boxplot of ratios of HiC links recalled (right) by HC CREs, MC and LC regions. Each dot represents the ratio of one gene. (C) Ratios of links validated by HiC (left) and ratios of HiC links recalled (right) with different threshold values of importance scores. Each dot represents the ratio for every 10kb distance to the TSS. (D) Ratios of links validated by HiC (left) and ratios of HiC links recalled (right) with different threshold values of importance scores. Each dot represents the ratio for mean values of every 1kb distance to the TSS within every window in 10 kb size. Error bars indicates standard deviation of ratios within 10 kb. (E) Average H3K27ac
signals of 1000 bp upstream and 1000 downstream from middle base of rescued CREs using bigWigAverageOverBed.

Figure S11. Comparison the performance of DIRECT-NET on scATAC-seq data of GM12878. (A) (C) Comparison of ratios of links validated by HiC (left) and ratios of HiC links recalled (right) by HC CREs,
MC and LC regions. Each dot represents the ratio for every 10kb distance to the TSS. (B) Venn plots of DIRECT-NET identified high importance CREs, Cicero identified high co-accessibility CREs, ArchR identified high co-accessibility CREs and ChIA-PET links. (C) Overlap ratios between CREs identified by DIRECT, Cicero and ArchR across genes which have at least one ChIA-PET link. Each element in the boxplot represents the ratio of one gene. (D) Venn plots of DIRECT-NET identified high importance CREs, Cicero identified high co-accessibility CREs, ArchR identified high co-accessibility CREs and HiC links. (E) Overlap ratios between CREs identified by DIRECT, Cicero and ArchR across genes which have at least one HiC link. (F) Histogram of numbers of validation links across different distance to TSS of GM12878 ChIA-PET and HiC data.

Figure S12. Validation of links with different importance scores on scATAC-seq data of Brain using HiChip data. (A) Boxplot of ratios of links validated by HiChIP with different threshold values of importance scores. P-value is calculated by paired student’s t test. (B) Projection of cells onto UMAP space using the scATAC-seq data of HC CREs (left panel), MC regions (middle panel) and LC regions
(right panel). Cells are colored by annotated cell labels. (C) Evaluation of cell state separation on the UMAP space using HC CREs, MC and LC regions via LISI (left panel) and Silhouette metrics (right panel). (D) Average H3K27ac signals of nigral, astrocytes cells of 1000 bp upstream and 1000 bp downstream from middle base of HC CREs, MC and LC regions on Nigral astrocytes, Isocortical astrocytes and Striatal astrocytes using bigWigAverageOverBed. (E) Gene regulatory network of marker genes and TFs bounded to HC CREs on scATAC-seq data of Brain.
Figure S13. The performance of DIRECT-NET scATAC-seq data of Brain. (A) Boxplot of ratios of links of HC CREs, MC and LC regions validated by HiChIP. Each element represents one gene. There are 58 markers which have at least one HiChIP links. $P$-value is calculated by paired student’s t test. (B) Projection of cells onto UMAP space using the scATAC-seq data of HC CREs (left panel), MC (middle panel) and LC regions (right panel). Cells are colored by annotated cell labels. (C) Evaluation of cell state separation on the UMAP space using HC CREs, MC and LC regions via LI$^2$SI (left panel) and Silhouette metrics (right panel). (D) Average H3K27ac signals of nigral, astrocytes cells of 1000 bp upstream and 1000 bp downstream from middle base of HC CREs, MC and LC regions on Nigral astrocytes, Isocortical astrocytes and Striatal astrocytes using bigWigAverageOverBed. (E) Ratios of links of HC CREs, MC and LC regions validated by HiChIP data. Each dot represents the ratio for every 10kb distance to the TSS. (F) Histogram of numbers of validation links across different distance to TSS of Brain HiChIP data. (G) Venn plots of DIRECT-NET identified high importance CREs, Cicero identified high co-accessibility CREs, ArchR identified high co-accessibility CREs and HiChIP links. (H) Overlap ratios between CREs identified by DIRECT, Cicero and ArchR across genes which have at least one HiChIP link. Each element in the boxplot represents the ratio of one gene.

Figure S14. Gene regulatory network of marker genes of and TFs bounded to HC CREs on scATAC-seq data of Striatal inhibitory 1 and Striatal inhibitory 2 cells.
Figure S15. Evaluation of whether promoter accessibility serves as a good proxy of gene expression on PBMC and A549 datasets. (A) Spearman correlation coefficients (SCC) between gene expression level and its promoter accessibility of aggregated cells using the nearest 50, 100 and 500 neighbors within each cell state as well as all cells within each cell state on PBMC (left) and A549 dataset (right). P-values are computed from one-tailed t-test. (B) The overlap ratios of inferred CREs across all genes are based on Jaccard index. Suppose $A_i$ and $B_i$ represent CREs detected based on gene expression and promoter accessibility separately for gene $i$. The ratio of overlapped CREs equals $\frac{|A_i \cap B_i|}{|A_i \cup B_i|}$, where $|A_i \cap B_i|$ is the number of overlapped CREs.

Figure S16. Comparison of the validation of predicted links of DIRECT-NET on paired scRNA-seq + scATC-seq data and scATAC-seq only via density-based scatter plot. Left panel is the distribution of the ratios of functional connections detected in HiC, while the right panel is the distribution of the ratios of HiC links recalled by functional connections. On each two-dimensional plot, x-axis and y-axis represent the ratios calculated using the predicted links by paired data and scATAC-seq only, respectively. Each colored dot represents the density of genes with the corresponding ratios. In the right panel, most genes have the same ratios, as reflected by the few
dots. P-values were calculated by paired student t test with right tail.

Figure S17. Evaluation of cellular heterogeneity on PBMC (A) and A549 datasets (B) based on ROGUE metric. ROGUE metrics are computed on both scRNA-seq and scATAC-seq with UMI counts for both cell states and all cells. (C) Comparison of AUC values of DIRECT-NET in identifying connections on a subset of cells only from each time point and all cells using HiC links. P-value is calculated by paired student's t test. (D) Comparison of AUC values of Cicero in identifying connections on a subset of cells only from each time point and all cells using HiC links.
Figure S18. Workflow of identifying the maximum number of cells with its neighbors in which the overlap ratio of any two cells is less than an overlap_rate. \( n \) is the number of cells, \( |A_{\text{step}}| \) is the number of set \( A_{\text{step}} \).
Figure S19. Assessment of overfitting performance of DIRECT-NET on Brain dataset. (A) RMSE (Root Mean Square Error) over the iterations, which measures the error of the regression model in predicting promotor’s accessibility of each gene using training, validation, and testing data, respectively. The colored dots and shading areas are the mean values and the standard deviations of RMSE over all computed genes at each iteration. (B) Comparison of the accuracy of predicted links using AUC metric from DIRECT-NET vs. DIRECT-NET with early stop strategy. AUC value was calculated by evaluating the DIRECT-NET-identified regulatory links against the true links in HiChIP data.

Figure S20. Schematic diagram showing the identification of high-confidence CREs, medium-confidence regions and low-confidence regions by applying different thresholds to different genes. Red dashed lines indicate the threshold for discriminating high-confidence CREs and medium-confidence regions, and light red dashed lines indicate the threshold for discriminating low-confidence regions and medium-confidence regions.
Supplementary Tables

Table S1. Summary of the numbers of validation links on all datasets and the number of markers of which there were at least one validation link.

| Dataset        | Type      | No. of markers | No. of links |
|----------------|-----------|----------------|--------------|
| PBMC-naïve B   | PCHiC     | 135            | 775          |
| PBMC-CD4 naïve T | PCHiC   | 58             | 452          |
| PBMC-CD8 naïve T | PCHiC   | 126            | 779          |
| A549           | HiC       | 139            | 209          |
| GM12878        | CHIA-PET  | 1540           | 5309         |
| GM12878        | HiC       | 758            | 3549         |
| Brain          | HiChip    | 58             | 281          |