Nucleic Acids Research, 2026, **54**, gkag437
https://doi.org/10.1093/nar/gkag437
Methods

Predicting enhancer–gene links from single-cell multi-omics data by integrating prior Hi-C information

Xuan Liang [1], Yuanyuan Miao[1], Dongmei Han [2], Yurun Li[1], Wenwen Zhang[1], Zhen Wang [2,3,*]

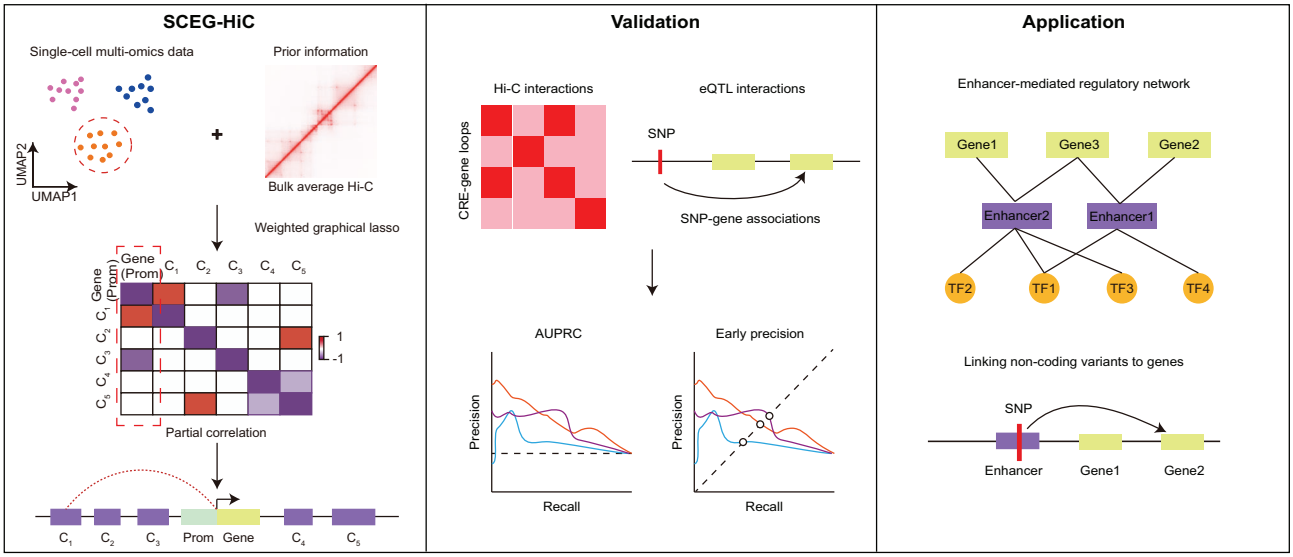
[1]Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China
[2]Key Laboratory of Systems Health Science of Zhejiang Province, School of Life Science, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou 310024, China
[3]Yuhang Institute of Medical Science Innovation and Transformation, Hangzhou 311100, China

*To whom correspondence should be addressed. Email: zwang01@sibs.ac.cn

Abstract

Enhancers play an important role in transcriptional regulation by modulating gene expression from distal genomic locations. Although single-cell ATAC and RNA sequencing (scATAC/RNA-seq) data have been leveraged to infer enhancer–gene links, establishing regulatory links between enhancers and their target genes remains a challenge due to the absence of chromatin conformation information. Here, we present SCEG-HiC, a machine learning method based on weighted graphical lasso, which decodes enhancer–gene links from single-cell multi-omics data by integrating bulk average Hi-C as prior knowledge. SCEG-HiC supports both paired scATAC/RNA-seq and scATAC-only inputs, improving prediction accuracy while retaining context-specific correlations and enabling the discovery of biologically relevant links. Comprehensive evaluation across 10 human and mouse single-cell multi-omics datasets shows that SCEG-HiC outperforms existing single-cell models. Application of SCEG-HiC to COVID-19 datasets illustrates its capacity to more reliably reconstruct gene regulatory networks underlying disease severity, and elucidate functional associations between noncoding variants and their putative target genes. SCEG-HiC is freely available as an open-source and user-friendly R package, facilitating broad applications in regulatory genomics research.

Graphical abstract



Introduction

In eukaryotes, transcriptional regulation is an essential biological process that enables cells or organisms to respond to various intra- and extra-cellular signals, maintain differentiated cell types, and coordinate cellular activities [1]. Promoters and enhancers, two major types of *cis*-regulatory elements (CREs), are short noncoding sequences bound by transcription factors (TFs) to govern transcriptional regulation of ad-

Received: September 21, 2025. **Revised:** April 10, 2026. **Accepted:** April 19, 2026

jacent genes [2, 3]. While promoters initiate transcription at or near the transcription start site (TSS), enhancers can augment transcription from more distal genomic locations. Genome-wide association studies (GWAS) have shown that >90% of single nucleotide polymorphisms (SNPs) associated with complex diseases are in noncoding regions of the genome far from promoters, implicating a potential regulatory role mediated by enhancers [4]. However, linking enhancers and their target genes (or equivalently, promoters) remains a significant challenge. First, enhancers are position-independent; they can be located upstream or downstream of their target promoter. Multiple studies have shown that enhancers can regulate genes located hundreds of thousands to millions of base pairs away, and they often do not regulate the closest gene [5–7]. Second, enhancers regulate the target gene expression in a cell- and tissue-specific manner, meaning that the same enhancer may target different genes under different biological contexts [3]. Third, enhancer–gene links are many-to-many: one enhancer can regulate multiple genes, and one gene may be controlled by multiple enhancers [8].

Several experimental approaches have been developed to identify enhancer–gene links. Chromatin conformation capture (3C) [9] and its derivative genomic assays, such as high-throughput chromatin conformation capture (Hi-C) [10], detect physical interactions between genomic regions. Other approaches, such as those based on expression quantitative trait loci (eQTL) [11] and CRISPR–Cas9 perturbation assays [12], infer regulatory links by associating genetic variants in enhancer regions with gene expression change. However, these assays are often technically challenging and costly, and have only been conducted at high resolution in a limited range of cell types.

As a supplement to experimental methods, various computational methods have been proposed. The simplest approach assigns an enhancer to its closest gene, but it overlooks the fact that enhancers can regulate gene expression over long genomic distances. As CRE activities can be measured via histone modification or chromatin accessibility profiling (e.g. H3K27ac ChIP-seq, ATAC-seq), multi-omics signals from diverse tissues and cell types have been modelled to infer enhancer–gene links [13–15]. Notably, the activity-by-contact (ABC) model combines enhancer activity with chromatin contact data (Hi-C) to assign enhancers to target genes [16]. This approach outperformed several other methods and showed comparable performance when using either cell type-specific Hi-C or average Hi-C map across multiple cell types. Yet, bulk sequencing technologies measure the average signals across biological tissues, underestimating cellular heterogeneity and lacking the ability to capture cell- and state-specific information.

The advent of single-cell technologies, such as single-cell RNA sequencing (scRNA-seq) [17] and single-cell ATAC sequencing (scATAC-seq) [18], has made it possible to uncover cellular heterogeneity at single-cell resolution. Notably, techniques such as SHARE-seq [19] and 10x Genomics Multiome ATAC + RNA [20] enable simultaneous measurement of gene expression and chromatin accessibility within the same cell. In principle, computational strategies developed for bulk sequencing data can be adopted for single-cell sequencing data by replacing bulk samples with individual cells [21]. Correlation-based models, such as Signac [22] and FigR [23], compute Pearson or Spearman correlations between enhancer accessibility and gene expression or promoter accessibility. Regression-based models, including DIRECT-NET [24], SCENIC+ [25], SCARlink [26], and Enhlink [27], quantify the relationship between one or more enhancers and target gene expression using multivariate regression. Latent variable-based models, such as scREG [28] and scMultiMap [29], integrate single-cell multi-omics data in latent space and infer enhancer–gene links based on the covariance in latent space. Supervised learning-based models (e.g. scE2G [30]), infer enhancer–gene links directly from single-cell data using experimentally validated labels or weakly supervised strategies with biologically informed priors. Hi-C structure-based models identify enhancer–promoter links using chromatin contact information from single-cell Hi-C [31, 32] or bulk Hi-C, such as SnapHiC2 [33], SnapHiC-G [34], and STARE (gABC) [35]. Despite their strengths, these approaches have inherent limitations: correlation-, regression-, and latent variable-based models rely solely on single-cell data and are sensitive to spurious correlations that lack regulatory relationships. Supervised learning-based models are limited by the scarcity of high-confidence enhancer–gene links, and Hi-C structure-based models often fail to adequately integrate multi-omics data at the single-cell level.

Here, we propose SCEG-HiC, a computational method that predicts **E**nhancer–**G**ene links from **S**ingle-**C**ell multi-omics data by integrating prior Hi-C information. SCEG-HiC employs the weighted graphical lasso (wglasso) model [36] to incorporate an average bulk Hi-C map, effectively regularizing the correlation matrix with the prior Hi-C contact matrix as a penalty term. SCEG-HiC can use either scATAC-seq alone or paired scATAC/RNA-seq data. Incorporation of the Hi-C prior enhances prediction accuracy, maintains cellular context-specific correlations, and enables the identification of biologically meaningful links. Through extensive benchmarking using independent functional genomics data, including cell type-specific Hi-C and eQTL data, we demonstrate that SCEG-HiC consistently outperforms existing enhancer–gene linking models across diverse human and mouse datasets. Moreover, we illustrate the biological relevance of SCEG-HiC by constructing cell type-specific gene regulatory networks (GRNs) and prioritizing target genes associated with noncoding GWAS variants. SCEG-HiC is implemented as an R package and is freely available at: https://github.com/wuwei77lx/SCEGHiC.

Materials and methods

Data preprocessing of single-cell data

Aggregation and single-cell retention approaches are developed separately for preprocessing single-cell omics data. We employed an aggregation method analogous to DIRECT-NET [24], which involved constructing a low-dimensional k-nearest neighbor (KNN) graph (default $k = 50$) from sparse single-cell omics data and subsequently averaging signals from similar cells. For paired scATAC/RNA-seq data, the KNN graph is constructed using the weighted nearest neighbor (WNN) approach implemented in the Seurat v4 package, which computes a joint neighbor graph based on a weighted integration of chromatin accessibility and gene expression profiles [37]. For scATAC-seq data alone, the KNN graph is built through singular value decomposition (SVD) on a transformed peak count matrix using term frequency-inverse document frequency (TF-IDF) [22]. The aggregation process fol-

lows a modified DIRECT-NET approach, selecting cells with minimal KNN overlap (default 0.8). Finally, the aggregated chromatin accessibility and gene expression data are normalized using scaling factors estimated by DESeq2 [38]. The single-cell retention approach requires normalization of both scATAC-seq and scRNA-seq data. We employed the SCTransform method in Seurat to normalize gene expression data, correcting for sequencing depth differences [39]. Chromatin accessibility data is normalized using the TF-IDF method in Signac, adjusting for sequencing depth and peak accessibility biases [22].

In this study, we evaluated the impact of six different data preprocessing methods on model accuracy, considering three factors including cell aggregation, scATAC-seq binarization and cell type scope: (i) aggregation using raw scATAC-seq data; (ii) aggregation using binarized scATAC-seq data; (iii) single-cell retention using TF-IDF normalized scATAC-seq data for all cell types; (iv) single-cell retention using binarized scATAC-seq data for all cell types; (v) single-cell retention using TF-IDF normalized scATAC-seq data for a specified cell type; and (vi) single-cell retention using binarized scATAC-seq data for a specified cell type.

Constructing and normalizing the contact matrix of bulk average Hi-C

To obtain the bulk average Hi-C map for constructing the enhancer–gene contact matrix, we adopted a similar method to the ABC model [16]. First, we extracted rows corresponding to the TSS of each gene from the standardized Hi-C matrix for each cell type (KR or SCALE normalization, at 5 kb resolution) [40]. Considering that the strength of intrachromosomal interactions follows a power-law distribution, we corrected for differences in the power-law decay of Hi-C signals across different cell types [41]. This is achieved by adjusting the signal based on the cell type-specific power-law gamma parameter, using the formula:

$$d^{(\gamma_{\text{ref}}-\gamma_{\text{celltype}})}, \tag{1}$$

where γ_{ref} is the reference gamma parameter, ensuring a consistent decay pattern across all cell types that matches the reference. To estimate γ_{celltype}, we aggregated Hi-C interactions within 1 Mb of all gene promoters and performed a linear regression in log-log space, where the regression slope corresponded to γ_{celltype}. After scaling, each Hi-C profile was normalized so that the total signal summed to one, ensuring equal weighting across all cell types. Finally, we averaged the normalized profiles across all cell types to construct a consensus Hi-C interaction landscape anchored at gene TSSs.

We extracted a contact matrix from the bulk average Hi-C map using the TSS of genes and the midpoints of all candidate enhancers. The coordinates were divided by 5 kb (the Hi-C resolution) and then floor-rounded using R's floor function. Given the low contact frequency of the resulting matrix, we standardized the data to the [0, 1] range and used it as a penalty matrix for subsequent wglasso modeling, with larger contact values assigned smaller penalties. Four standardization methods were evaluated to generate contact penalty matrix: (i) binarization, where contact penalties are assigned to 0 (contact present) or 1 (contact absent); (ii) min–max normalization, where contact values are scaled to the [0, 1] range by subtracting each value from their maximum and dividing by their range (max–min); (iii) -log transformation, where contact values are -log-transformed and then min-max normalized; (iv) rank score, where contact values are ranked in descending order and scaled to the [0, 1] range by relative ranks.

Construction of a predictive enhancer–gene link model

SCEG-HiC employs the wglasso method to predict enhancer–gene links, effectively handling the high dimensionality of single-cell datasets, where the number of potential element pairs far exceeds the cell number [36]. Moreover, wglasso integrates multilevel biological information, enhancing model accuracy and robustness. Like the glasso method, wglasso estimates the inverse covariance matrix to compute partial correlation coefficients, thereby removing indirect links [42]. The key difference is that while glasso applies a single penalty parameter, wglasso leverages prior information to specify different amounts of penalties to different element pairs, enabling adaptive shrinkage of partial correlations for improved accuracy. For each gene, wglasso maximizes the following objective function:

$$\log\det\Theta - tr\,(S\Theta) - \rho\|P * \Theta\|_1, \tag{2}$$

where Θ is the inverse covariance matrix to be estimated, S is the observed covariance matrix, P is the prior information matrix ($P \in [0, 1]$), ρ is the penalty parameter ($\rho \in [0, 1]$, with a step size of 0.01), and $*$ represents element-wise multiplication. The inverse covariance matrix Θ, observed covariance matrix S, and prior matrix P are constructed over the gene and its candidate enhancers, with matrix dimensions corresponding to the gene plus the total number of candidate enhancers. To ensure consistent scale across data modalities, we employed a correlation matrix (i.e. standardized covariance matrix) instead of the raw covariance matrix. Specifically, with scATAC-seq data alone, the matrix is computed based on the chromatin accessibility of promoters (peaks within ± 1 kb from the TSS) and candidate enhancers (peaks within ± 250 kb from the TSS). When paired scATAC/RNA-seq data are available, the matrix is computed using gene expression and enhancer accessibility. The standardized contact penalty matrix derived from the bulk average Hi-C map serves as the prior information matrix P, where larger values indicate stronger penalties and weaker prior support for corresponding enhancer–gene links. To mitigate self-interaction effects during optimization, a small constant (1×10^{-4}) is added to the diagonal of P. The optimal penalty parameter ρ is determined by cross-validation using the Bayesian information criterion (BIC), which tends to select highly sparse networks and performs well in scale-free network simulations [43]. The BIC for the wglasso model takes the form

$$\text{BIC} = -2l_n\,(\Theta) + |E|\log\,(n), \tag{3}$$

where $|E|$ denotes the number of edges, i.e. the number of nonzero elements in Θ corresponding to the gene and its candidate enhancers, n is the sample size, corresponding to the number of single cells or pseudo-bulk samples, and l_n is the log likelihood function simplified from

$$l_n\,(\Theta) = \frac{n}{2}\left[\log\left(\det\,(\Theta)\right) - tr\,(S\Theta)\right]. \tag{4}$$

After estimating the inverse covariance matrix, partial correlation coefficients are calculated to infer enhancer–gene links, with a default threshold of 0.01 applied to select significant links. The partial correlation coefficient R is calculated

as:

$$R = -\frac{\Theta}{DD^T}. \quad (5)$$

D is a diagonal matrix composed of the square roots of the diagonal elements of Θ, covering the gene and its candidate enhancers.

Preprocessing paired scRNA-seq and scATAC-seq datasets

Single-cell paired multi-omics data were processed using Seurat (scRNA-seq) and Signac (scATAC-seq). Quality control (QC) was applied to both the RNA and ATAC data, assessing RNA counts, ATAC fragment counts, nucleosome signal, TSS enrichment, and optionally blacklist fraction. QC thresholds were defined based on the distribution of these metrics within each dataset. For scRNA-seq data, RNA counts were normalized using Seurat's SCTransform function, followed by dimensionality reduction with RunPCA. For scATAC-seq data, MACS2 was used via Signac's CallPeaks function to identify peaks from fragment files [44]. After filtering out peaks on nonstandard chromosomes and overlapping with blacklisted genomic regions, a peak-by-cell matrix was generated using Signac's FeatureMatrix function for downstream analyses. ATAC peaks were normalized using FindTopFeatures and RunTFIDF in Signac, and dimensionality reduction was performed using RunSVD. If the dataset contained multiple samples, batch effects across different samples were corrected using the Harmony method with the RunHarmony function [45]. The WNN graph was constructed using latent semantic indexing dimensions 2–40 for scATAC-seq (excluding the first singular component due to its strong correlation with technical factors like sequencing depth) and PCA dimensions 1–50 for scRNA-seq. Finally, high-resolution clustering was performed using the Louvain algorithm on the joint WNN graph via Seurat's FindNeighbors and FindClusters functions.

Existing methods for paired scATAC-seq/scRNA-seq data

We compared SCEG-HiC with several existing methods for paired scATAC/scRNA data, including two classical correlation-based approaches (Pearson and Spearman), as well as eight computational tools. (i) CellOracle (v0.14.0), which integrates Cicero results with TSS peak information to infer peak–gene links from scATAC-seq data [46]; (ii) DIRECT-NET (v0.0.1), which applies eXtreme Gradient Boosting (XGBoost) to identify high-confidence functional CRE–gene links from parallel scATAC/RNA-seq data [24]; (iii) SCENIC+ (v1.0.1), which utilizes gradient boosting machine regression to compute CRE–gene importance scores from scATAC/RNA-seq data [25]; (iv) FigR (v0.1.0), which employs Spearman correlation coefficients to connect distal CREs to genes from paired scATAC/RNA-seq data [23]; (v) Signac (v1.12.0), which applies Pearson correlation with background correction to compute linkage scores for peak–gene pairs from scATAC/RNA-seq data [22]; (vi) SCARlink (v1.0.0), which employs a regularized Poisson regression model using tile-level chromatin accessibility to predict gene expression and leverages Shapley values to infer enhancer–gene links from paired scATAC/RNA-seq data [26]; (vii) Enhlink (v0.21.4), which adopts a random forest-like framework incorporating biological and technical covariates to infer enhancer–gene co-accessibility from multi-omics scATAC/RNA-seq data [27]; (viii) scMultiMap (v1.0.1), which utilizes moment-based fast estimation within a joint latent variable model and adjusts for technical confounders to infer enhancer–gene links from highly sparse multimodal count data [29]. One notable difference among these models is that SCARlink, Enhlink, and scMultiMap predict enhancer–gene links by analyzing data from each cell type separately, whereas the remaining models are designed to infer links by correlating across all cell types in a dataset simultaneously. For all models, promoters are defined as regions within ±1 kb from the TSS, while candidate enhancers are defined as regions within ±250 kb from the TSS, excluding promoter regions. A common set of marker genes was selected for each cell type, and all resulting links were retained for benchmarking purposes. For detailed implementation of these methods, see Supplementary Methods.

Existing methods for scATAC-seq data alone

We compared SCEG-HiC with several existing methods for scATAC-seq data alone, including two statistical approaches: Pearson correlation and chi-squared test with false discovery rate correction (Chi^2 + FDR), as well as four computational tools. (i) Cicero (v1.3.9), which uses the glasso model to infer co-accessibility relationships between CREs from scATAC-seq data alone [47]; (ii) scEChIA (v0.1.0), which utilizes signal imputation and L1 regularization to infer chromatin interaction links from scATAC-seq data [48]; (iii) DIRECT-NET (v0.0.1), which applies XGBoost to identify high-confidence functional CRE–gene links from scATAC-seq data alone [24]; (iv) Enhlink (v0.21.4), which adopts a random forest-like framework to infer enhancer–promoter co-accessibility from scATAC-seq data [27]. One notable difference among these models is that Chi^2 + FDR, Enhlink, and scEChIA predict enhancer–gene links by analyzing data from each cell type separately, whereas the remaining models infer links by correlating across all cell types in a dataset simultaneously. All methods followed the same promoter and enhancer definitions as previously described to ensure consistency across evaluations. For detailed implementation of these methods, see Supplementary Methods.

ABC model

Additionally, we compared SCEG-HiC with the ABC method for bulk ATAC-seq, which integrates chromatin accessibility and an average Hi-C contact map to predict enhancer–gene links [16]. Pseudobulk profiles were generated by averaging gene expression and chromatin accessibility for each cell type. To maintain consistency, ABC analysis employed the same promoter and enhancer definitions as described above (see Supplementary Methods for details).

Validation with Hi-C and eQTL data

To evaluate the accuracy of predicted enhancer–gene links, we used published cell type-specific Hi-C profiles. We used Juicer Tools to process raw sequencing data into a list of Hi-C contacts (.hic files) [49]. The chromatin interaction scores were normalized using the SCALE or KR methods and extracted in text format at 5 kb resolution. An enhancer–gene link is considered a true positive (TP) if: (i) the gene TSS is within a bait fragment, (ii) the midpoint of the peak is within

an other-end fragment, and (iii) a high interaction score is observed between the bait and other-end fragment in Hi-C data, determined by a threshold to select top-enriched chromatin interactions. Note that the cell type-specific Hi-C data used for validation are independent and not included in the bulk average Hi-C map. Even when the validation data corresponded to the same cell type, they were obtained from independent experimental sources, ensuring that model evaluation was independent of the prior. Additionally, we used eQTL data from matched tissues in GTEx (v8) to assess the accuracy of enhancer–gene links [50]. Specifically, we used *cis*-eQTL data, which link genetic variants to the expression of nearby genes. An enhancer–gene link is considered a TP if (i) the peak overlaps with an eQTL locus, and (ii) the locus is associated with the target gene's expression.

Evaluation metrics

To derive a common gene set for benchmarking, we first identified 3000 highly variable genes for each dataset. We then used the Presto package to select marker genes for each cell type with AUROC > 0.5 [51]. The evaluation metrics for comparison included area under the precision–recall curve (AUPRC), AUPRC ratio, early precision (EP), and EP ratio (EPR) [52]. AUPRC was computed using the pr.curve function from the PRROC package, and EP was defined as the fraction of TPs among the top-k predicted positives, where k represents the total number of TPs. Both the AUPRC ratio and EPR were calculated by dividing the AUPRC and EP values by the precision of a random predictor, which is determined by the total fraction of TPs.

Analysis of experimentally validated enhancers in the VISTA Enhancer Browser

To further evaluate the accuracy of predicted enhancer–gene links, we conducted an additional analysis using a curated set of enhancers experimentally validated in the VISTA Enhancer Browser. This database contains data from over 4500 experiments, with most evaluated at embryonic stage E11.5, primarily covering human and mouse elements [53]. Among the enhancers with positive validation results, nearly 80% exhibit regulatory activity in the brain, limb, or heart. We selected experimentally validated brain-associated enhancers (midbrain, hindbrain, forebrain, or neural tube) with positive results from the VISTA Enhancer Browser, including 738 human elements and 432 mouse elements. These curated enhancers were then intersected with scATAC-seq peaks called by MACS2 using bedtools [54], and target genes were predicted for the overlapped peaks.

eQTL enrichment analysis

Cis-eQTLs from matched tissues in GTEx (v8) were used as independent functional evidence to evaluate the predicted links under different prior matrices [50]. For each dataset, the lost or gained links were defined as the foreground set, respectively, with all candidate enhancer–gene links serving as the background set. Overlaps between foreground links and *cis*-eQTL loci associated with the target gene were counted, and enrichment significance was calculated using a two-sided Fisher's exact test.

Correlations between cumulative enhancer activity and gene expression

To calculate cumulative enhancer activity, we first normalized the peak counts by the total number of unique fragments in each cell's peaks. After normalization, the cumulative enhancer activity for each gene is defined as the sum of the normalized counts from all significantly associated peaks, resulting in a cell$\times$ cumulative enhancer activity matrix. The thresholds for significant peaks were model-specific (ABC: 0.02, SCEG-HiC: 0.001). We then computed the Spearman correlations between cumulative enhancer activity and gene expression values across cells for each gene.

Selection and differential expression analysis of COVID-19 scRNA-seq data

The scRNA-seq dataset was processed and annotated by Wilk *et al.* following the Seurat workflow [55]. Since our primary focus was comparing peripheral blood mononuclear cell (PBMC) samples from mild and severe COVID-19 patients, we excluded samples derived from red blood cell-lysed whole blood and removed individuals annotated as "healthy" or "convalescent." To compare mild and severe COVID-19 patients, differential expression tests were performed for each cell type using Seurat's FindMarkers function with the Wilcoxon rank-sum test. Genes were considered significantly differentially expressed if they met all the following criteria: (i) absolute $\log_2$-fold change (FC) ≥ 0.25, (ii) Bonferroni-corrected P-value $< .05$, and (iii) expression in at least 10% of cells in either group.

TF motif enrichment analysis

TF motif enrichment analysis was performed using motif position weight matrices from the JASPAR2020 database [56]. Core TF motifs were retrieved using the getMatrixSet function from the TFBSTools package [57] with parameters species = "Homo sapiens," collection = "CORE," and all_versions = FALSE, resulting in 633 core TF motifs. Genomic region statistics—including GC content, sequence length, and dinucleotide frequencies—were computed using the RegionStats function. Motif annotations were then added to the object using AddMotifs. Finally, we selected relevant peaks and performed motif enrichment analysis within these regions with the FindMotifs function. Motifs with a fold enrichment greater than 2 and an adjusted P-value $<.05$ were considered significant.

Construction of GRNs

To construct GRNs, we used the matchMotifs function from the motifmatchr package to identify TFs associated with peak regions, leveraging motif data from the JASPAR2020 database [58]. Only TF motifs showing significant enrichment in these regions were considered. We integrated these TF–peak links with peak–gene links derived from SCEG-HiC to establish TF–peak–gene relationships. In this framework, peaks serve as intermediaries linking TFs to their putative target genes, forming a GRN structure.

Results

Overview of SCEG-HiC framework

SCEG-HiC is a computational framework designed to predict enhancer–gene links using single-cell multi-omics data (paired scATAC/RNA-seq or scATAC-seq alone), combined with three-dimensional chromatin conformation information derived from a bulk average Hi-C map. Gene promoters are defined as ± 1 kb regions around TSS, while candidate enhancers are identified as accessible chromatin regions outside promoters but within a user-specified genomic window (default: ± 250 kb). The SCEG-HiC framework comprises three major steps (Fig. 1 and see the "Materials and methods" section): (i) preprocessing CRE accessibility and gene expression matrices; (ii) constructing and normalizing bulk average Hi-C to generate the contact penalty matrix; (iii) inferring enhancer–gene links using the wglasso model.

The preprocessing step mainly involves two considerations: the sparse nature of single-cell sequencing data (especially scATAC-seq), and spurious peak–gene associations confounded by cell type heterogeneity [59, 60]. To systematically evaluate these issues, we compared six preprocessing strategies (Supplementary Fig. S1, Supplementary Table S1, and see the "Materials and methods" section), considering either aggregation or single-cell retention approach, along with the impact of binarizing scATAC-seq data. The aggregation approach, applied across cell types, aggregates signals from small clusters of similar cells to alleviate data sparsity. The single-cell retention approach, tailored to specific cell types, preserves individual cell signals. Our results showed that the single-cell retention strategy with normalized scATAC-seq data achieved the highest prediction accuracy, albeit identifying fewer enhancer–gene links. In contrast, the aggregation approach with binarized scATAC-seq data resulted in slightly lower accuracy but captured a broader range of enhancer–gene links. To balance accuracy and coverage, we implemented both preprocessing strategies in SCEG-HiC, with aggregation designated as the default.

Distinct from previous methods, SCEG-HiC integrates prior knowledge of chromatin conformation to improve prediction accuracy. Although chromatin interactions can be cell type-specific, previous studies have shown that many chromatin interactions are highly similar across different cell types [61]. Notably, an average Hi-C map generated from diverse cell types, termed "bulk average Hi-C," represents the average of normalized cell type-specific maps rather than data from a pooled experiment, and can capture potential chromatin interactions in new cell types as validated by the ABC model [16] (see the "Materials and methods" section). Due to the low contact frequency in the bulk average Hi-C map, we normalized the contact values to generate the contact penalty matrix. Among the four evaluated normalization methods, the rank score method significantly outperformed the others in terms of prediction accuracy (Supplementary Fig. S2, Supplementary Table S1, and see the "Materials and methods" section). A small constant is added to the diagonal of the normalized contact penalty matrix to mitigate self-interaction effects, and the choice of this constant is robust with respect to predictive performance (Supplementary Fig. S3).

Finally, SCEG-HiC employs the wglasso model to infer enhancer–gene links, effectively addressing the high-dimensionality of single-cell data while integrating prior information [36]. Specifically, SCEG-HiC integrates the normalized Hi-C contact penalty matrix with the correlation matrix of gene expression (or promoter accessibility) and candidate enhancer accessibility. The penalty parameter is optimized using BIC [43] cross-validation, with the value minimizing BIC selected as the final parameter (Supplementary Fig. S4A). Analysis across 2000 highly variable genes revealed gene-specific optimal penalty values, while subsampling 70% of cells for 100 randomly selected genes over 100 iterations showed consistently low interquartile range, indicating both gene-level specificity and overall model stability (Supplementary Fig. S4B and C). Partial correlation coefficients derived from the model quantify the strength of enhancer–gene links, with positive values indicating biologically meaningful links. A default threshold of 0.01 for the partial correlation coefficient is applied as needed, providing a balance between precision and recall (Supplementary Fig. S5).

Benchmarking SCEG-HiC with human paired single-cell multi-omics data

We analyzed five paired scATAC/RNA-seq datasets from different human tissues generated by 10x Genomics, including PBMCs, skin stromal cells [62], fetal retinal tissue [63], brain gray matter [64], and developing cerebral cortex [65] (Fig. 2A and Supplementary Table S2). For each dataset, we applied Seurat [37] and Signac [22] to perform QC, cell clustering, cell type annotation, and uniform manifold approximation and projection (UMAP) visualization based on WNN integration of gene expression and chromatin accessibility data (Fig. 2A, Supplementary Fig. S6, and Supplementary Table S3; see the "Materials and methods" section and Supplementary Methods). We applied SCEG-HiC to each dataset to conduct data aggregation and scATAC-seq binarization, and integrated a human bulk average Hi-C map to predict enhancer–gene links for marker genes of each cell type. The bulk average Hi-C map for humans was obtained from the ABC model and constructed by preprocessing 34 Hi-C datasets from the ENCODE project (Supplementary Table S4) [16, 66].

To evaluate the accuracy of predicting enhancer–gene links in the human paired single-cell multi-omics data, we compared SCEG-HiC against two classical correlation-based methods (Pearson and Spearman), and eight dedicated computational tools: CellOracle [46], DIRECT-NET [24], SCENIC+ [25], FigR [23], Signac [22], SCARlink [26], Enhlink [27], and scMultiMap [29] (see the "Materials and methods" section and Supplementary Methods). We collected experimentally supported interactions from cell type-specific Hi-C and eQTL data as the benchmarks (Supplementary Table S5). Given the highly imbalanced fraction of verified enhancer–gene links, we used the AUPRC and EP as evaluation metrics [52, 67], which are inherently threshold-independent (see the "Materials and methods" section). Using CD4 + T cells from the PBMC dataset, we first validated the performance of SCEG-HiC against CD4 + T cell Hi-C data from ENCODE [68] and whole blood eQTL data from GTEx [50]. In Hi-C-based validation, SCEG-HiC achieved an AUPRC of 0.683 and an EP of 0.689; in eQTL-based validation, it reached an AUPRC of 0.219 and an EP of 0.255, substantially outperforming all other methods (Fig. 2B). As a case study, we examined enhancer–gene links involving *PDK1*, a key regulator of T cell development and activation [69, 70], using the PBMC dataset (Fig. 2C). SCEG-HiC accurately recovered validated links using CD4 + T cell-specific Hi-C data from ENCODE

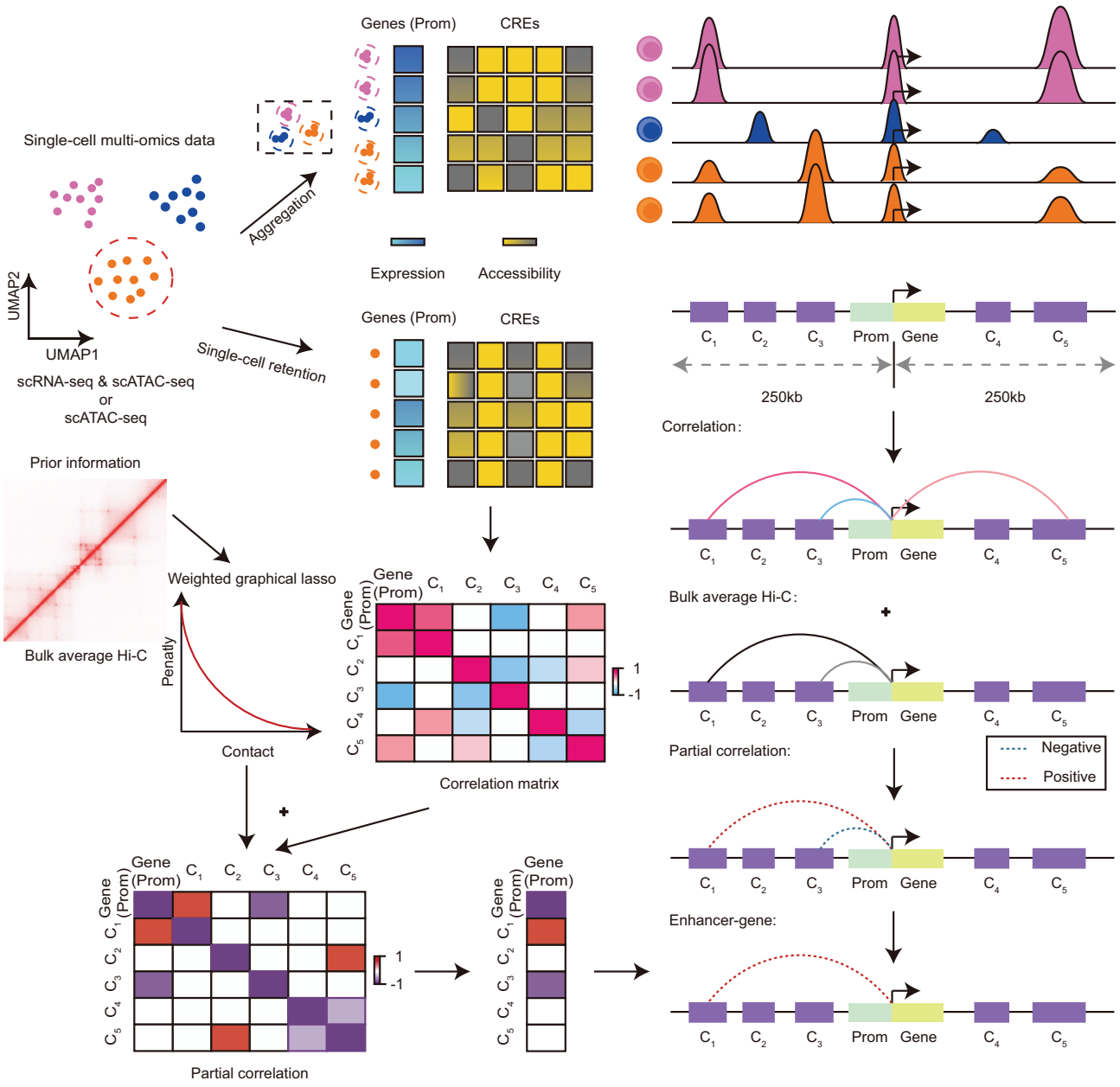


Figure 1. The workflow of SCEG-HiC. SCEG-HiC integrates single-cell multi-omics data and bulk average Hi-C map to infer enhancer–gene links. The left panel illustrates the detailed workflow of SCEG-HiC. Taking paired scATAC/RNA-seq or scATAC-seq data alone as input, SCEG-HiC first computes enhancer–gene correlation matrices based on gene expression (or promoter accessibility) and distal CRE accessibility across aggregated cells or single cells from a specific cell type. Second, it incorporates chromatin interaction priors from a bulk average Hi-C map to construct a normalized contact penalty matrix between genes and CREs. Third, SCEG-HiC applies a wglasso model to estimate partial correlation coefficients, preserving enhancer–gene links with positive coefficients. The right panel shows a schematic genomic region illustrating how SCEG-HiC integrates Hi-C priors with single-cell omics correlations to calculate partial correlation coefficients for enhancer–gene links. For each gene, the ±1 kb region around the TSS is defined as the promoter, while CREs within ±250 kb regions (excluding promoters) are defined as candidate enhancers (e.g. C_1-C_5). Prom: promoter.

and whole blood eQTL data from GTEx. Importantly, we observed that integrating the bulk average Hi-C map substantially reduces false positives compared to models relying solely on correlations between gene expression and chromatin accessibility.

To further illustrate the broad applicability of SCEG-HiC, we evaluated its performance across nine additional representative cell types from the five multi-omics datasets. These included cell types covered and uncovered by the bulk average Hi-C map, allowing a comprehensive assessment of model robustness across diverse cellular contexts. To ensure comparability across datasets and cell types, we calculated the AUPRC ratio and EPR relative to the fraction of verified interactions (i.e. random baseline) for each cell type within each dataset. Using Hi-C-validated enhancer–gene links from all 10 cell types, SCEG-HiC consistently yielded significantly higher AUPRC ratio and EPR than all competing methods (Fig. 2D, *P*-values from paired *t*-tests are provided in Supplementary Table S6). A similar trend was observed when using eQTL-derived links from eight cell types as benchmarks: SCEG-HiC again significantly outperformed most other methods in terms of both AUPRC ratio and EPR, with no significant

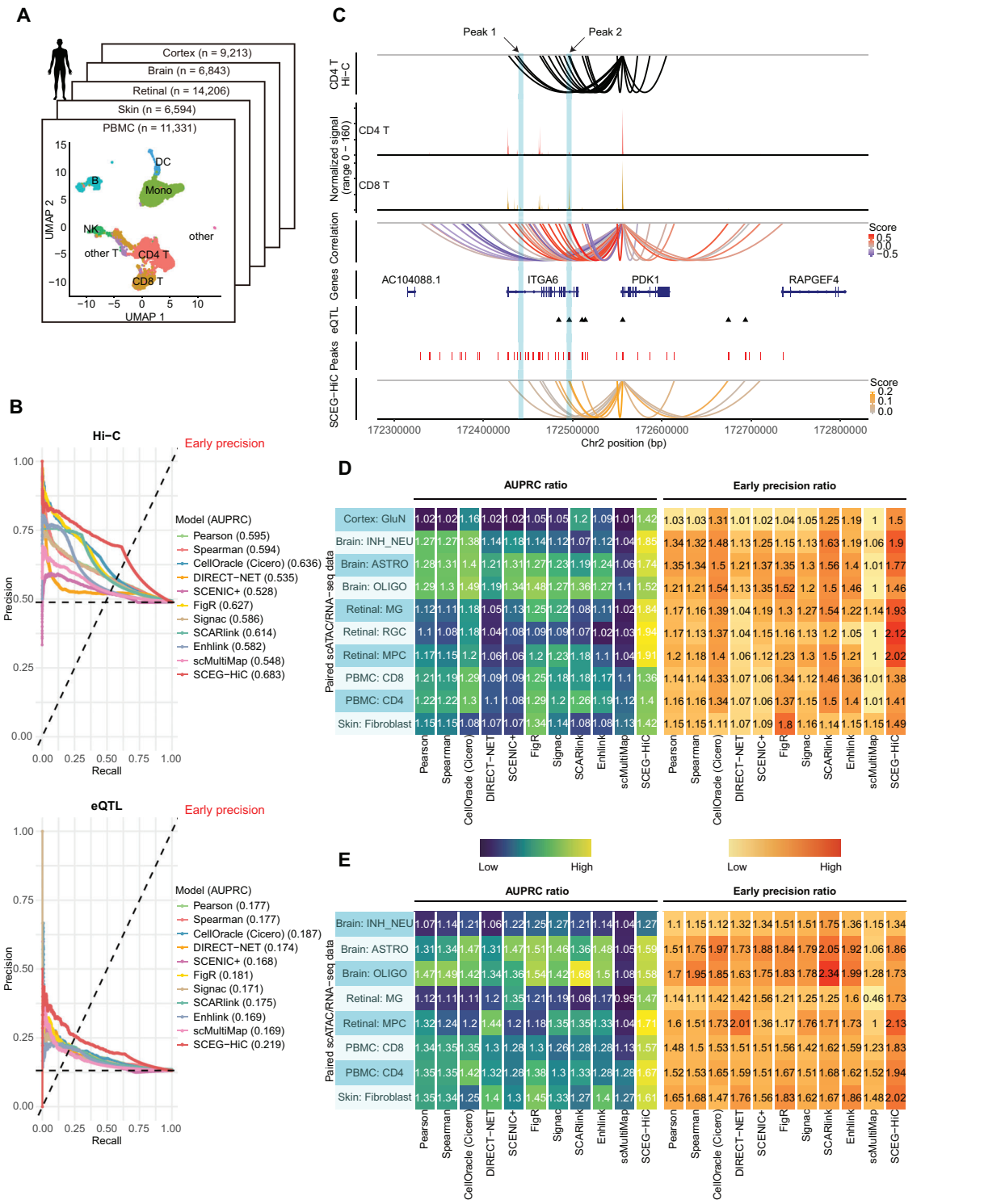


Figure 2. SCEG-HiC can accurately predict enhancer–gene links using human paired single-cell multi-omics data. (**A**) Summary of human paired scATAC/RNA-seq datasets used for benchmarking (*n*= cell number), along with UMAP visualizations of a representative PBMC dataset. Cell types: CD4 T, CD4 + T cell; CD8 T, CD8 + T cell; Mono, monocyte; DC, dendritic cell; NK, natural killer cell; other T: other T cell. Datasets: skin, skin stromal cells; retinal, fetal retinal tissue; brain, brain gray matter; cortex, developing cerebral cortex. (**B**) Precision–recall curves evaluating enhancer–gene link prediction performance for CD4 + T cells, validated using CD4 + T cell-specific Hi-C (top) and whole blood eQTL data (bottom). Models compared include SCEG-HiC, CellOracle (Cicero-based), DIRECT-NET, SCENIC+, FigR, Signac, SCARlink, Enhlink, scMultiMap, and correlation methods (Pearson, Spearman). The intersection of the dashed diagonal line with the precision–recall curve indicates EP, while the horizontal dashed line represents the random prediction baseline. AUPRC, area under the precision–recall curve. (**C**) Representative example of predicted enhancer–gene links for *PDK1*. Integration of CD4 + T cell Hi-C interactions, scATAC-seq signals, aggregated ATAC-RNA correlations, eQTL SNPs, and enhancer–gene links predicted by SCEG-HiC around *PDK1*. Peak 1 represents an enhancer not predicted by SCEG-HiC, exhibiting strong correlations but lacking Hi-C and eQTL support. Peak 2 corresponds to an enhancer predicted by SCEG-HiC, supported by both Hi-C and eQTL evidence. (D, E) Performance of enhancer–gene link predictions across different human cell types, measured by AUPRC ratio (left) and EPR (right). Validation was performed using cell type-specific Hi-C (**D**) and eQTL data (**E**), respectively.

difference from FigR, SCARlink, and Enhlink only in terms of EPR (Fig. 2E, *P*-values from paired *t*-tests are provided in Supplementary Table S6).

Additionally, we assessed the accuracy of SCEG-HiC predictions on a curated set of enhancers experimentally validated in the VISTA Enhancer Browser [53]. Considering the tissue specificity of enhancers, we focused on brain-associated elements (see the "Materials and methods" section). Due to the limited overlap between eQTL and VISTA enhancers, the inferred enhancer–gene links were validated using brain midfrontal cortex Hi-C data from ENCODE. In the human brain gray matter dataset, SCEG-HiC consistently achieved higher AUPRC and EP than all other methods (Supplementary Fig. S7A). These evaluation results collectively confirm that SCEG-HiC achieves superior performance in predicting enhancer–gene links from human paired single-cell multi-omics data.

We also assessed the computational cost and scalability of SCEG-HiC across varying dataset sizes. Using randomly sampled PBMC data according to cell type proportions with different numbers of cells (3 000, 5 000, and 10 000) and highly variable genes (500, 1 000, and 3 000), we measured runtime (CPU seconds) and peak memory usage (MB) and compared SCEG-HiC with other methods. The results showed that the runtime of SCEG-HiC was primarily influenced by the number of genes, with more genes leading to longer runtimes, while memory usage mainly increased with the number of cells, with more cells resulting in higher memory usage (Supplementary Fig. S8 and Supplementary Table S7). SCEG-HiC comprises two steps: precomputing Hi-C weights and executing wglasso. We observed that wglasso exhibited moderate runtime and memory usage relative to other methods under all tested conditions. While precomputing Hi-C weights required additional runtime, this step only needs to be performed once for a given gene set.

Impact of Hi-C prior on enhancer–gene link predictions in SCEG-HiC

Given that SCEG-HiC incorporates an averaged bulk Hi-C prior constructed from a limited number of cell lines, an important question is how the Hi-C prior influences model behavior and predictive performance. We first assessed the representativeness of the bulk average Hi-C map by comparing it with validated pairs from cell type-specific Hi-C maps. We found that the coverage of these pairs shows an approximately linear decline with increasing contact percentile thresholds in the bulk average Hi-C map, suggesting that bulk Hi-C broadly reflects the contact distribution of validated pairs rather than being limited to the strongest contacts (Supplementary Fig. S9A and see Supplementary Methods).

To confirm that the improvement of SCEG-HiC stemmed from incorporating the bulk average Hi-C map, we performed an ablation analysis comparing SCEG-HiC with its no-prior version under the same model framework. Across both cell type-specific Hi-C and eQTL validation data, SCEG-HiC consistently outperformed the no-prior model in AUPRC ratio and EPR (Fig. 3A and B). To further assess whether the model depends on true Hi-C structure rather than only distance constraints, we performed distance-stratified randomization of the bulk Hi-C prior and reran SCEG-HiC (Supplementary Fig. S10 and see Supplementary Methods). Predictions based on real Hi-C data were significantly higher than those using the shuffled prior for AUPRC and EPR with cell type-specific Hi-C validation, and for AUPRC with eQTL validation (Fig. 3A and B). These results indicate that the performance gains of SCEG-HiC are not solely driven by linear genomic distance, highlighting the important role of local Hi-C contact structure.

We next systematically evaluated the impact of Hi-C prior strength by exponentiating the contact penalty matrix as P^{τ}, where the temperature parameter τ ranged from 0 to 2 at a fixed step of 0.5. It is important to emphasize that as τ increases, the proportion of Hi-C contacts with relatively low penalty increases, thereby strengthening the prior. Based on five paired multi-omics datasets, most links identified at $\tau = 0$ remained unchanged with increasing τ, and the number of lost links was relatively small and declined further as τ increased (Fig. 3C and Supplementary Fig. S11). However, increasing τ yielded a substantial number of newly gained links. Analysis of the characteristics of lost, unchanged, and gained links across different τ values revealed that lost links at higher τ tended to show significantly lower prior contact values, and their correlations were also generally lower than those of unchanged links (Fig. 3D and Supplementary Fig. S12). In contrast, gained links at higher τ exhibited higher prior contact values than unchanged links, despite significantly lower correlations. Across both cell type-specific Hi-C and eQTL validation data, SCEG-HiC achieved optimal predictive accuracy at $\tau = 1$. Further increasing τ to 2 did not improve performance and may even slightly reduce accuracy (Supplementary Fig. S13).

As increasing τ may remove a proportion of links with high correlations but low prior support, we compared the observed (especially those lost and gained) links at $\tau = 1$ with $\tau = 0$ across the datasets. First, we compared the cell type-specific correlations of lost links with those of unchanged links. The analysis showed that, except in some cell types where no significant difference was observed due to the limited number of cells, lost links generally exhibited significantly lower correlations than unchanged links (Supplementary Fig. S14). Second, we calculated the overlap of lost and gained links with the consensus predictions shared by SCARlink [26] and Enhlink [27] (thresholds: SCARlink, Z-score > 0.05 and adjusted $P < .05$; Enhlink, adjusted $P < .05$). Our results indicated that the overlap of lost links was substantially lower than that of gained links, suggesting that the lost links were not more strongly supported by these Hi-C free methods. (Fig. 3E and Supplementary Fig. S15A–E). Third, we performed orthogonal functional validation by assessing eQTL and TF motif enrichment. In the eQTL enrichment analysis, lost links showed no significant enrichment, whereas gained links were significantly enriched across four datasets (Fig. 3F and see the "Materials and methods" section). In the TF motif enrichment analysis of candidate cell type-specific motifs, enhancers associated with gained links also showed significantly stronger motif enrichment than those associated with lost links (Supplementary Fig. S15F and see Supplementary Methods). Collectively, these results demonstrate that the removed links are few in number and lack robust support from independent functional evidence, and indicate that incorporating the Hi-C prior can effectively complement single-cell statistical signals to improve the accuracy and biological relevance of predicted enhancer–gene links.

Given the availability of Hi-C maps more closely matched to the tissues or cell types of the analyzed samples, we com-

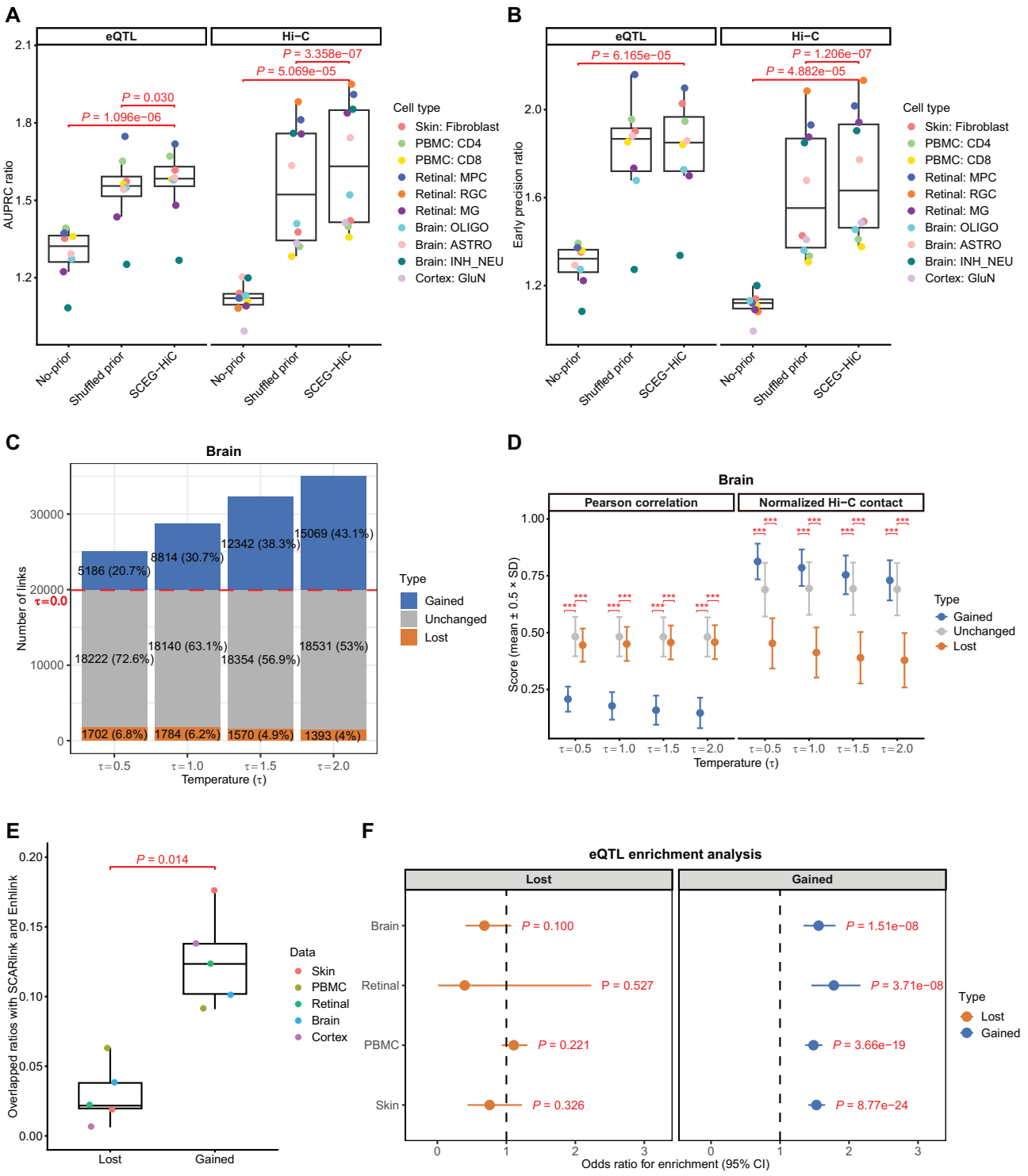


Figure 3. Incorporation of Hi-C priors significantly improves predictive accuracy. Boxplots comparing performance of no-prior, shuffled prior, and SCEG-HiC across human cell types, measured by AUPRC ratio (**A**) and EPR (**B**). Validation was performed using cell type-specific Hi-C and eQTL data, respectively. Each dot represents a human cell type. *P*-values were calculated using two-sided paired *t*-tests. (**C**) Number of predicted enhancer–gene links at different prior temperatures τ based on the paired multi-omics data from brain gray matter. Predicted links are categorized as lost, gained, and unchanged at a given τ relative to $\tau = 0$. Four temperature parameters were tested ($\tau = 0.5$, 1.0, 1.5, 2.0). (**D**) Distribution of Pearson correlation coefficients and normalized prior Hi-C contact values at different prior temperatures τ, for the paired multi-omics data from brain gray matter. Each dot represents the mean value for a given τ, with error bars indicating $\pm$ 0.5 $\times$ standard deviation. *P*-values were calculated using two-sided Wilcoxon rank-sum tests between link types (***$P < .001$). (**E**) Overlap ratios between lost and gained enhancer–gene links predicted by SCEG-HiC and the intersection of significant predictions from SCARlink and Enhlink. Each point represents one paired multi-omics dataset. Lost and gained links are defined at $\tau = 1$ relative to $\tau = 0$. *P*-values were calculated using two-sided paired *t*-tests. (**F**) eQTL enrichment of lost and gained enhancer–gene links predicted by SCEG-HiC across four paired single-cell multi-omics datasets. Odds ratios are shown as points with 95% confidence intervals as bars. *P*-values were calculated using two-sided Fisher's exact tests.

pared the effects of two different Hi-C prior sources on predictive performance within the SCEG-HiC framework using PBMC data: the bulk average Hi-C map and a cell type-matched Hi-C map (CD8 + T cell Hi-C dataset from ENCODE: ENCFF009ONH). Across independent CD8 + T cell Hi-C data (ENCODE: ENCFF063KHP, different from the prior source) and whole blood eQTL data from GTEx [50] used for validation, predictive performance based on the cell type-matched Hi-C map was overall comparable to that based on the averaged bulk Hi-C map, albeit slightly lower in terms of AUPRC and EP (Supplementary Fig. S16). This likely arises because the bulk average Hi-C map generally exhibits higher genomic coverage and resolution.

Finally, we systematically evaluated the impact of penalty strength by multiplicatively scaling the parameter as $\alpha\rho$, with α ranging from 0 to 2 at a fixed step of 0.5. Compared to the no-prior model, the prior contact values of lost, unchanged, and gained links all increased with rising penalty strength, with gained links consistently showing the highest contact values, followed by unchanged and lost links (Supplementary Fig. S17). These results indicated that increasing the penalty strength preferentially led to the prediction of enhancer–gene links with higher prior support. Therefore, higher penalty strength (greater Hi-C reliance) is suitable for tissues well represented in the datasets used to construct the bulk average Hi-C map, whereas lower penalty strength (less Hi-C reliance but greater reliance on single-cell data) is preferable for more unique tissues or conditions.

Benchmarking SCEG-HiC with human scATAC-seq data alone

To assess the accuracy of predicting enhancer–gene links using human scATAC-seq data alone, we compared SCEG-HiC with two statistical methods, Pearson correlation and Chi^2 + FDR, as well as four computational tools: Cicero [47], DIRECT-NET [24], scEchIA [48], and Enhlink [27] (see the "Materials and methods" section and Supplementary Methods). DIRECT-NET and Enhlink are suitable for both paired scATAC/RNA-seq data and scATAC-seq data alone. scEchIA shares similarities with our approach, using average chromatin interactions from other cell types to predict links between distal sites (25 kb) in a specific cell type. Using the same five paired single-cell multi-omics datasets, we restricted our analysis to the scATAC-seq data to predict enhancer–gene links. Data preprocessing and integration of prior bulk average Hi-C map in SCEG-HiC were consistent with the approach used for paired multi-omics data.

Similarly, we validated the inferred enhancer–gene links against both cell type-specific Hi-C and eQTL data, using AUPRC ratio and EPR as evaluation metrics (Supplementary Table S5 and see the "Materials and methods" section). In Hi-C-based validation across 10 cell types, SCEG-HiC significantly outperformed all other methods in both AUPRC ratio and EPR (Fig. 4A, *P*-values from paired *t*-tests or Wilcoxon tests are provided in Supplementary Table S6). Notably, scEchIA demonstrated the second-best performance, closely following SCEG-HiC. This observation further supports the notion that incorporating chromatin interaction information can substantially enhance model prediction accuracy. In eQTL-based validation across eight cell types, SCEG-HiC again showed significantly higher AUPRC ratio and EPR across most cell types (Fig. 4B, *P*-values from paired *t*-tests are provided in Supplementary Table S6). However, in contrast to the Hi-C validation results, scEchIA did not consistently rank second in accuracy, possibly due to its lack of Hi-C data from a broader set of cell types.

We also evaluated the accuracy of SCEG-HiC and other methods in predicting the target genes of a curated set of brain-associated enhancers experimentally validated in the VISTA Enhancer Browser [53]. Using brain midfrontal cortex Hi-C data for validation, SCEG-HiC consistently outperformed all other methods in both AUPRC and EP (Supplementary Fig. S7B). In summary, these results demonstrate that SCEG-HiC maintains robust predictive performance when operating solely on human scATAC-seq data.

Finally, we observed that the runtime and memory usage of SCEG-HiC showed similar trends to the paired scATAC/RNA-seq data: runtime primarily increased with the number of genes, while memory usage mainly increased with the number of cells (Supplementary Fig. S18 and Supplementary Table S7). Compared with other methods, SCEG-HiC also maintained a moderate runtime and memory usage, excluding the runtime for precomputing Hi-C weights.

Benchmarking SCEG-HiC with mouse paired single-cell multi-omics data

We analyzed five paired single-cell multi-omics datasets from different mouse tissues generated by different platforms, including mouse embryonic brain cells at day 18 (10x Genomics), mouse skin (SHARE-seq) [19], adult mouse cerebral cortex (SHARE-seq) [71], thymic epithelial cells (10x Genomics) [72], and mouse liver (10x Genomics) [73] (Fig. 5A and Supplementary Table S2). Each dataset was processed using Seurat and Signac for QC, dimensionality reduction, UMAP visualization, clustering, and cell annotation (Fig. 5A, Supplementary Fig. S19, and Supplementary Table S3, and see the "Materials and methods" section and Supplementary Methods). Given the absence of a pre-existing average Hi-C map for mice, we collected seven Hi-C datasets from the ENCODE project, including two embryonic stem cell lines (mESC1, mESC2), CH12LX, CH12F3, fiber, epithelium, and B cells (Supplementary Table S4) [68]. Following the ABC model's approach for constructing the human average Hi-C map, we scaled each Hi-C dataset using dataset-specific power-law distributions and computed their average to generate the mouse average Hi-C map (Fig. 5B and see the "Materials and methods" section). Although fewer datasets are included for mice than for humans, the bulk average Hi-C map in mice well represents validated pairs from independent cell type-specific Hi-C data, similar to that in humans (Supplementary Fig. S9B and see Supplementary Methods). Subsequently, we applied SCEG-HiC to each mouse single-cell dataset with the aggregation and scATAC-seq binarization approach, integrating the constructed mouse average Hi-C map to predict enhancer–gene links.

We benchmarked SCEG-HiC against 10 methods or tools for paired scATAC/RNA-seq data, as mentioned before. Due to the lack of eQTL data for mice, we only used cell type-specific Hi-C data for benchmarks, with AUPRC ratio and EPR as evaluation metrics (Supplementary Table S5). Across the five datasets, we selected eleven representative cell types for evaluation, including those covered and uncovered by the mouse bulk average Hi-C map. We observed that SCEG-HiC consistently achieved significantly higher AUPRC ra-

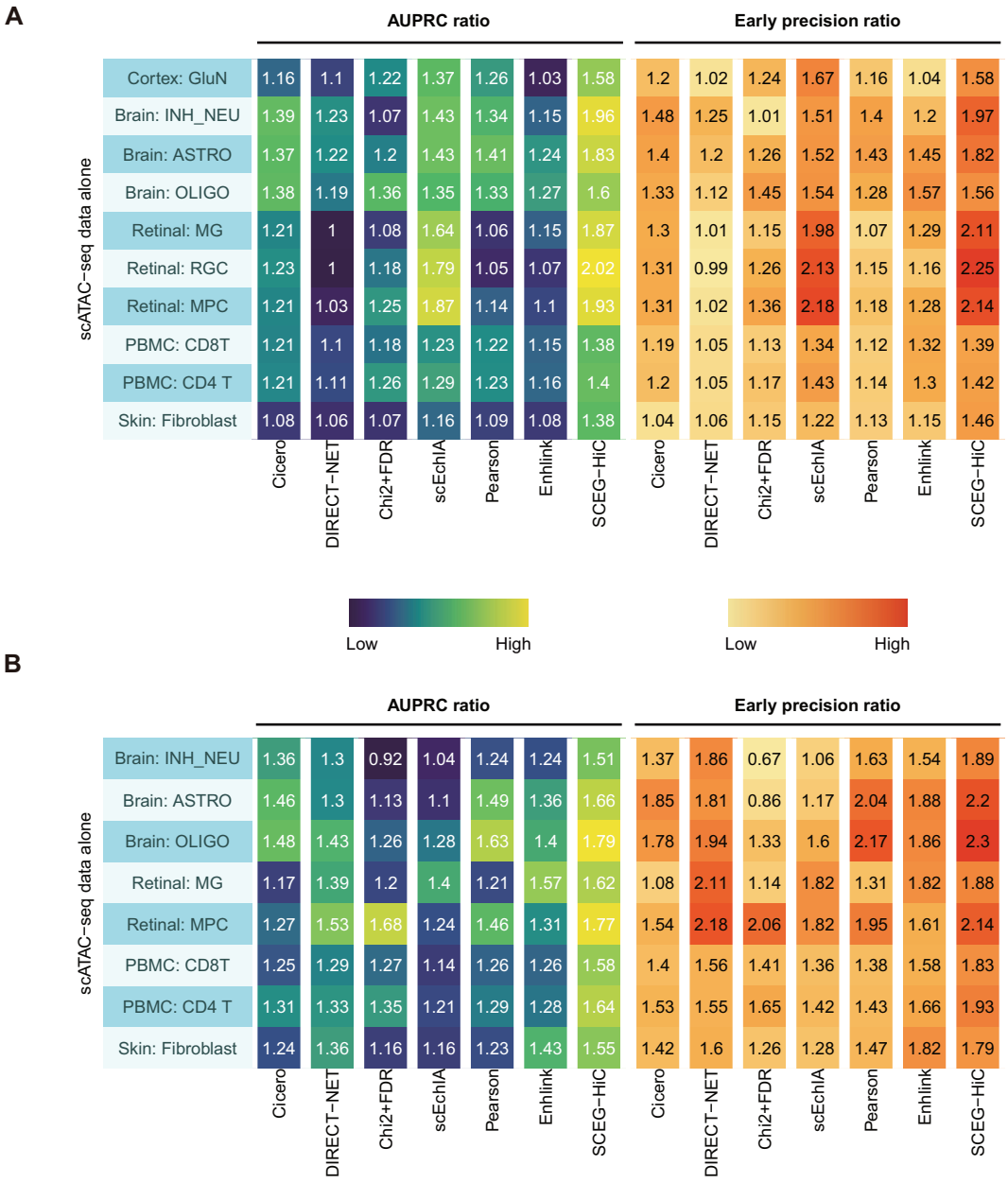


Figure 4. SCEG-HiC can accurately predict enhancer–gene links using human scATAC-seq data alone. (A, B) Performance of enhancer–gene link predictions across different human cell types, measured by AUPRC ratio (left) and EPR (right). Validation was performed using cell type-specific Hi-C (**A**) and eQTL data (**B**). Six competing methods were considered for comparison, including Cicero, DIRECT-NET, Chi^2 + FDR, scEchIA, Pearson correlation, and Enhlink.

tio and EPR compared to all other methods (Fig. 5C, *P*-values from paired Wilcoxon tests or *t*-tests are provided in Supplementary Table S6). Furthermore, the performance improvements observed in mouse datasets were even more pronounced than those in human datasets, highlighting the robustness and cross-species applicability of SCEG-HiC.

To further evaluate the accuracy of SCEG-HiC predictions, we applied the same analysis pipeline used for human datasets to a curated set of brain-associated enhancers experimentally validated in the VISTA Enhancer Browser [53], comparing SCEG-HiC with other methods in predicting their target genes (see the "Materials and methods" section). We extended the analysis to mouse embryonic brain cells at day 18, using mouse embryonic forebrain Hi-C data for validation. While the EP of SCEG-HiC was slightly lower than that of SCARlink in this dataset, it outperformed all other models across most remaining metrics (Supplementary Fig. S7C). In conclusion, we validated SCEG-HiC as a superior method in predicting enhancer–gene links using mouse paired single-cell multiomics data.

Comparison of SCEG-HiC and ABC

Although the ABC model was originally designed for bulk sequencing data, it has also been adopted in scATAC-seq analysis by aggregating cells into pseudobulk profiles (see the "Materials and methods" section and Supplementary Methods). Given that both SCEG-HiC and ABC utilize a bulk aver-

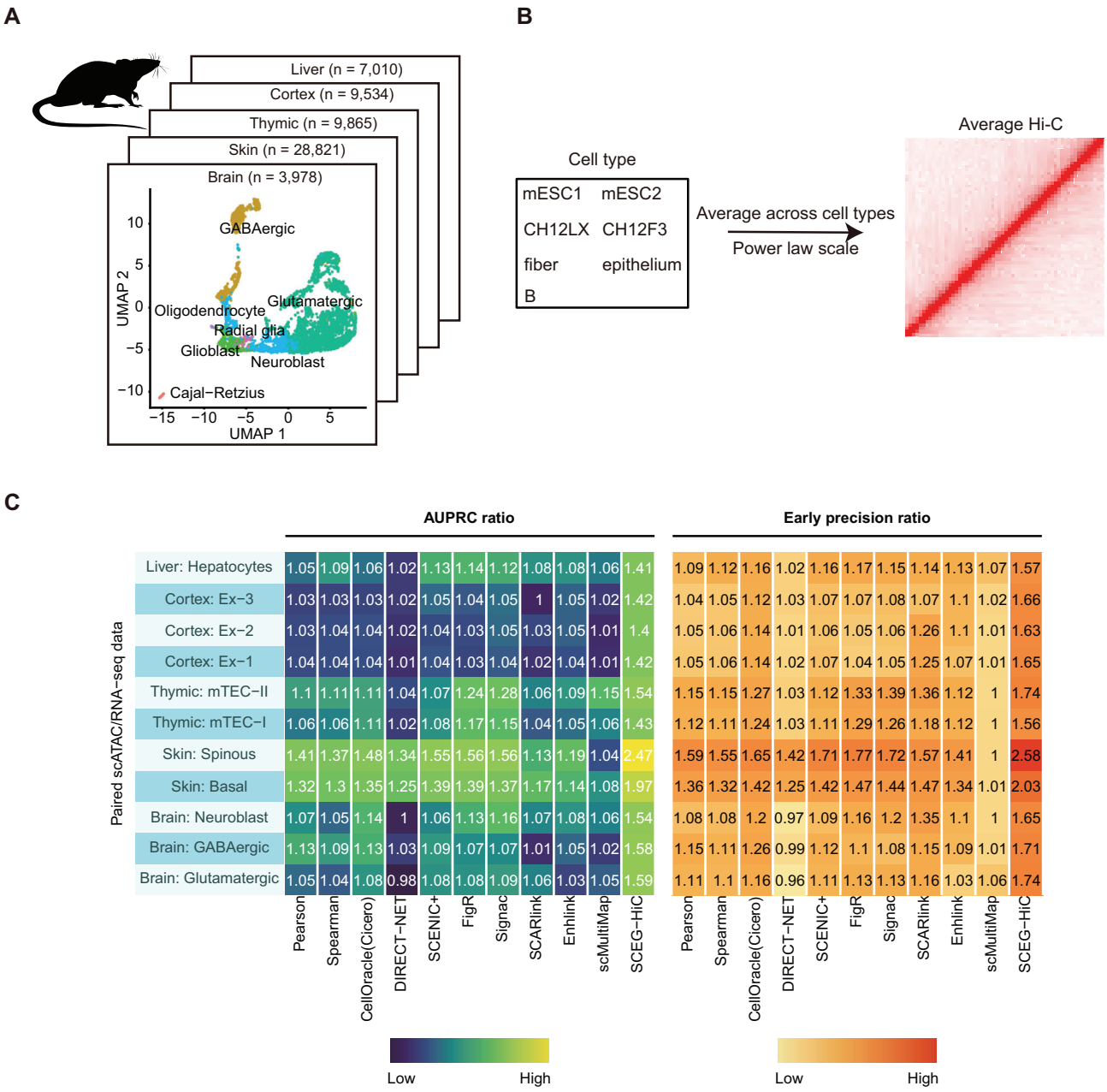


Figure 5. SCEG-HiC can accurately predict enhancer–gene links using mouse paired single-cell multi-omics data. (**A**) Summary of mouse paired scATAC/RNA-seq datasets used for benchmarking (*n*= cell number), along with UMAP visualizations of a representative dataset of mouse embryonic brain cells at day 18. Datasets: brain, mouse embryonic brain cells at day 18; skin, mouse skin; thymic, thymic epithelial cells; cortex, adult mouse cerebral cortex; liver, mouse liver. (**B**) A schematic illustration for generating a mouse bulk average Hi-C map from seven datasets. (**C**) Performance of enhancer–gene link predictions across different mouse cell types, assessed by AUPRC ratio (left) and EPR (right) using cell type-specific Hi-C data. Ten competing methods were considered for comparison, including Pearson correlation, Spearman correlation, CellOracle (Cicero-based), DIRECT-NET, SCENIC+, FigR, Signac, SCARlink, Enhlink, and scMultiMap.

age Hi-C map, we systematically compared the two models to evaluate the improvements attained by SCEG-HiC [16]. First, we assessed the prediction accuracy using both cell type-specific Hi-C and eQTL data. In Hi-C-based validation, the ABC model showed a significantly higher AUPRC ratio in both human and mouse datasets (Fig. 6A–D). In contrast, SCEG-HiC consistently outperformed the ABC model in terms of EPR. In eQTL-based validation, no significant differences were observed between the two models in terms of AUPRC ratio, though the ABC model demonstrated superior performance in the EPR metric.

Next, considering that gene expression is typically regulated by the combined activity of multiple enhancers, we evaluated the functional relevance of predicted enhancer–gene links from both models. Specifically, we summed the ATAC-seq signals across all enhancers associated with each gene at the single-cell level, then calculated the Spearman correlations between the cumulative enhancer activity and corresponding gene expression (see the "Materials and methods" section). This metric assesses the functional plausibility of predicted enhancer–gene links, under the assumption that accurate regulatory relationships should manifest as a strong posi-

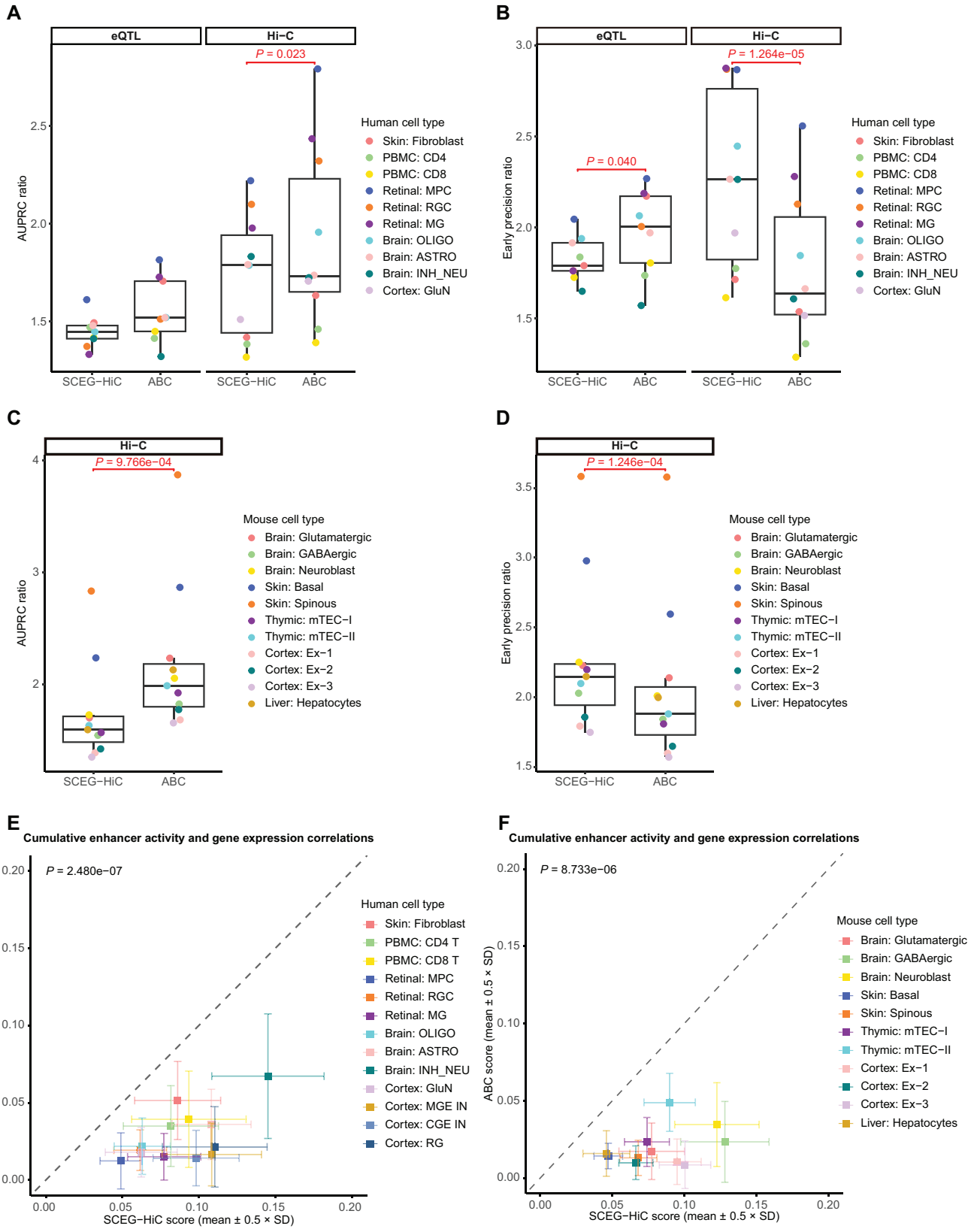


Figure 6. SCEG-HiC demonstrates higher expression consistency compared to ABC. Boxplots comparing performance between SCEG-HiC and ABC across various human (**A**, **B**) and mouse (**C**, **D**) cell types, measured by AUPRC ratio (A, C) and EPR (B, D). Validation was performed using cell type-specific Hi-C data for both species, and eQTL data for humans only. Each dot represents a distinct cell type. *P*-values were calculated using two-sided paired *t*-tests. Cumulative enhancer activity and gene expression correlations compared between SCEG-HiC and ABC across human (**E**) and mouse (**F**) cell types. The *x*- and *y*-axes represent Spearman correlations from SCEG-HiC and ABC predictions, respectively. Each dot corresponds to the mean correlation across genes for a specific cell type, with error bars representing ± 0.5 × standard deviation. Dashed diagonal line indicates equal correlations. *P*-values were calculated between the two models using two-sided paired *t*-tests.

tive correlation between cumulative enhancer activity and target gene expression. Across diverse human and mouse cell types, SCEG-HiC consistently outperformed the ABC model, showing significantly stronger correlations between predicted cumulative enhancer activity and gene expression (Fig. 6E and F, and Supplementary Fig. S20). This expression concordance analysis highlights a critical distinction between single-cell methods that account for cell-to-cell variability and the ABC model, which does not, thereby complementing validations based on Hi-C or eQTL data.

Finally, under the same gene and cell number conditions, SCEG-HiC required less runtime and memory usage than the ABC model, excluding the runtime for precomputing Hi-C weights (Supplementary Fig. S21 and Supplementary Table S7). In summary, ABC and SCEG-HiC exhibit relative advantages in Hi-C and eQTL-based benchmarks, with performance dependent on specific metrics. Notably, SCEG-HiC combines moderate computational cost with markedly stronger correlations between predicted cumulative enhancer activity and gene expression, providing a key advantage in efficiency and regulatory interpretability for inferring enhancer–gene links.

Analysis of COVID-19 unpaired single-cell multi-omics data

We applied SCEG-HiC to unpaired PBMC scRNA-seq and scATAC-seq datasets from patients infected with SARS-CoV-2, focusing on enhancer-mediated regulations underlying severe and mild COVID-19 symptoms [55]. The patients were stratified into mild and severe categories based on the World Health Organization (WHO) severity scores (severe: 5–7, mild: 2–4). We utilized the scRNA-seq data preprocessed and clustered by Wilk *et al*., selecting 18 samples that met our study requirements. This dataset comprised a total of 61 486 cells spanning 14 distinct cell types (Supplementary Fig. S22A–C and see the "Materials and methods" section). For scATAC-seq analysis, we selected 17 COVID-19 samples (9 mild and 8 severe) and processed the data using Signac. After QC, we retained 52 817 high-quality cells, performed peak calling on each sample using MACS2, and generated a unified peak set comprising 195 424 peaks (Supplementary Table S3) [44]. Following data integration, clustering, and annotation, 10 cell types were identified (Supplementary Fig. S22D–F and see Supplementary Methods). Differential expression analysis per cell type between mild and severe groups revealed that CD14 + monocytes exhibited the highest number of differentially expressed genes (DEGs), highlighting this cell type's pronounced response to SARS-CoV-2 infection (Supplementary Fig. S23A and see the "Materials and methods" section). In total, 871 genes showed significant differential expression in CD14 + monocytes ($|\log_2\text{-FC}| > 0.25$, adjusted $P < .05$; Fig. 7A and Supplementary Table S8). We then applied SCEG-HiC to the scATAC-seq data of CD14 + monocytes using the single-cell retention approach and benchmarked its performance against Cicero [47], a classic method that predicts enhancer–gene links using scATAC-seq data alone.

We first evaluated the accuracy of predicting enhancer–DEG links using AUPRC and EP. We benchmarked against three external validation datasets: CD14 + monocyte Hi-C data from ENCODE [68], whole blood eQTL data from eQTLGen [74], and GTEx [50]. Across all benchmarks, SCEG-HiC consistently outperformed Cicero in both AUPRC and EP, demonstrating its superior accuracy in identifying enhancer–gene links for a specific cell type (Fig. 7B). In addition, we separately used CD14 + monocyte scATAC-seq data from severe and mild COVID-19 patients to predict enhancer–DEG links. The results showed that SCEG-HiC consistently achieved higher AUPRC and EP than Cicero in both patient groups, reinforcing its robustness across varying disease states (Supplementary Fig. S23B and C).

Next, we constructed GRNs using TFs linked to enhancer–DEG links predicted by the model, focusing on the 41 most significant DEGs in CD14 + monocytes ($|\log_2\text{-FC}| > 1$, adjusted $P < .05$). Using both SCEG-HiC and Cicero, we predicted enhancers associated with these DEGs and selected the top 80 enhancer–DEG links with the highest prediction scores from each model. Motif enrichment analysis was then performed on the predicted enhancer regions to identify significantly enriched TFs (see the "Materials and methods" section). Among the top 80 predicted links, SCEG-HiC identified significantly more links supported by external datasets compared to Cicero (Hi-C: $P = 6.640 \times 10^{-7}$; eQTLGen: $P = 0.037$; GTEx: $P = 0.017$ via one-sided Fisher's exact test). Motif enrichment analyses based on both models revealed significant enrichment of C/EBP and AP-1 TF families (Supplementary Table S9). Previous studies have shown that AP-1 and C/EBP are associated with the immune activity of classical monocytes in COVID-19 patients [75]. After viral infection, host cells secrete interferons, activating signaling pathways such as JAK-STAT, which in turn regulate AP-1 and C/EBP TFs and further influence the expression of inflammatory genes like *IL1B* [76]. SCEG-HiC more reliably recapitulated enhancer-mediated relationships between these TFs and their target genes, as shown in GRNs inferred from predicted links (Fig. 7C, Supplementary Fig. S23D, and Supplementary Table S10, and see the "Materials and methods" section).

Finally, we explored the application of SCEG-HiC in linking noncoding variants to their target genes. We retrieved 265 fine-mapped SNPs from [77], which were derived from GWAS analyses comparing hospitalized COVID-19 patients with infected nonhospitalized individuals [78]. Given that these SNPs were shown to be enriched in the accessible chromatin regions of CD14 + monocytes [77], we leveraged CD14 + monocyte scATAC-seq data to predict their target genes. SCEG-HiC and Cicero identified 20 and 17 SNP–gene links, respectively, which were validated against CD14 + monocyte Hi-C and eQTLGen data. SCEG-HiC validated 20/20 (100%) SNP–gene links with Hi-C data and 13/20 (65.0%) with eQTLGen data (Fig. 7D and Supplementary Table S11). In contrast, Cicero validated only 7/17 (41.2%) links with Hi-C data and 5/17 (29.4%) with eQTLGen data. These results demonstrated that SCEG-HiC significantly outperformed Cicero in the accuracy of predicting SNP–gene links.

As shown in Fig. 7E, we illustrated a potential regulatory program indicating the effect of a fine-mapped SNP, rs71327024, in the context of COVID-19. Specifically, SCEG-HiC linked this SNP to *CCR1* based on CD14 + monocyte scATAC-seq data from COVID-19 patients, an association supported by both CD14 + monocyte Hi-C and whole blood eQTL data. In contrast, Cicero linked it to *CCRL2*, a prediction lacking external validation. Besides, previous studies have confirmed that *CCR1* is a critical mediator of monocyte/macrophage polarization and tissue infiltration, which are pathogenic hallmarks of severe COVID-19 [79, 80]. Mul-

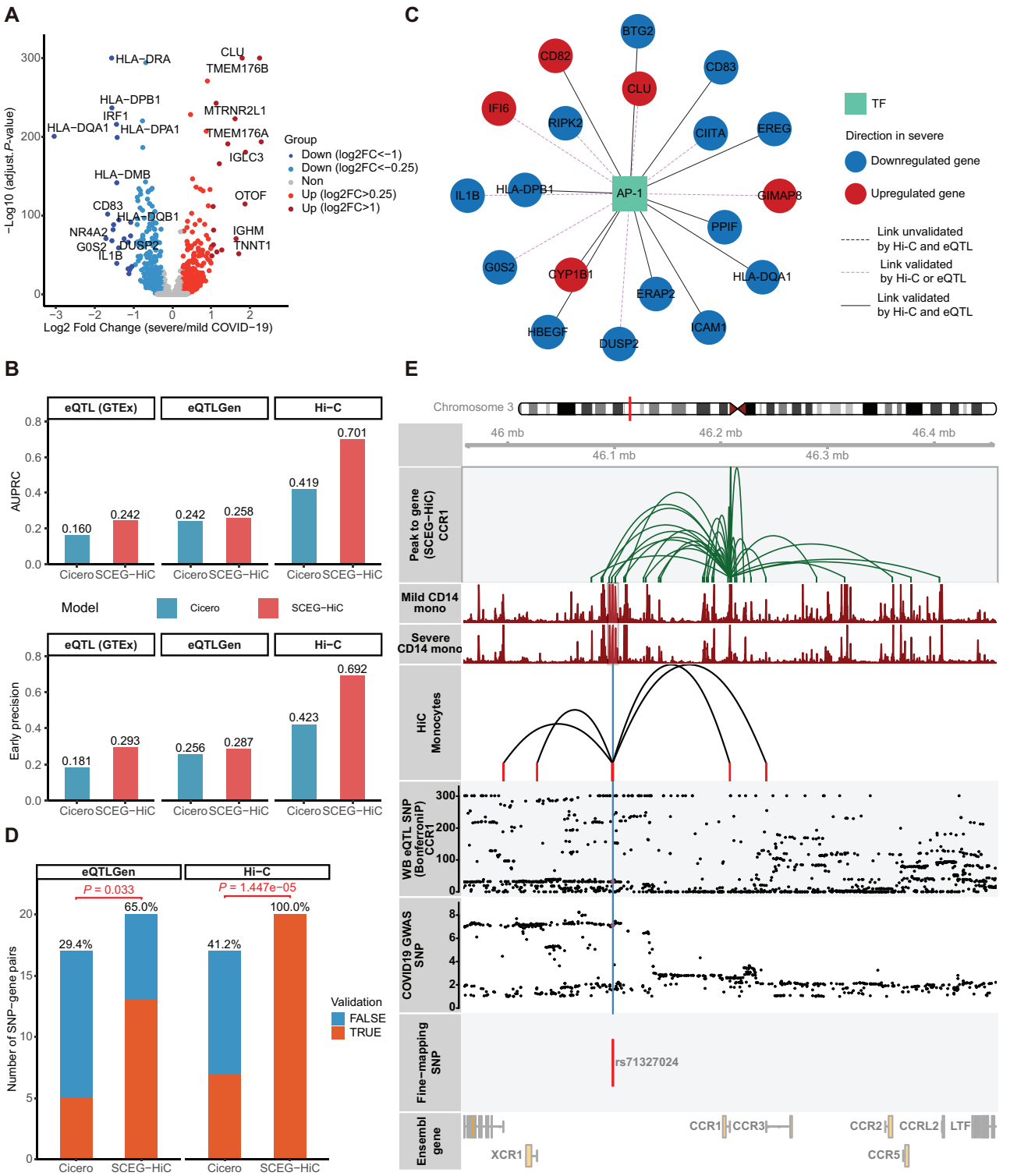


Figure 7. SCEG-HiC enables GRN construction and noncoding variant annotation in COVID-19 single-cell data. (**A**) Volcano plot showing DEGs in scRNA-seq data between CD14 + monocytes from severe and mild COVID-19 patients. Genes with adjusted *P*-value $< .05$ and $\log_2$-FC > 0.25 are shown in red; those with $\log_2$-FC < -0.25 are shown in blue. (**B**) Bar plots comparing performance between SCEG-HiC and Cicero in CD14 + monocytes, evaluated by AUPRC (top) and EP (bottom). The validation was performed using CD14 + monocyte Hi-C and whole blood eQTL data from GTEx and eQTLGen. (**C**) Subnetwork of the GRN between AP-1 TF families and target genes, inferred from TF motif enrichment analysis of SCEG-HiC-predicted enhancer–gene links. (**D**) Comparison of the number of SNP–gene pairs predicted by SCEG-HiC and Cicero, with the validation fractions shown using CD14 + monocyte Hi-C and eQTLGen data. *P*-values were calculated using Fisher's exact tests. (**E**) SCEG-HiC prediction of fine-mapped SNP rs71327024 linked to *CCR1*. Integration of SCEG-HiC-predicted enhancer–gene links, scATAC-seq signals, CD14 + monocyte Hi-C interactions, whole blood eQTL SNPs, and COVID-19 GWAS SNPs around rs71327024.

tiple risk SNPs have been shown to be associated with the *CCR* gene cluster (*CCR1*, *CCR2*, *CCR3*, and *CCR5*) in monocytes/macrophages, whereas *CCRL2* exhibits no such associations [81]. We also identified another example where *CXCR6* was associated with multiple fine-mapped SNPs, which were validated by Hi-C and eQTL data (Supplementary Fig. S23E). The chemokine receptor *CXCR6* and its ligand *CXCL16* have been found to play an important role in the immunopathogenesis of severe COVID-19 [82]. In contrast, Cicero failed to identify any genes associated with these SNPs. A PheWAS analysis also identified several monocyte-related phenotypes associated with SNPs in the *CXCR6* region, indicating a functional role of *CXCR6* in monocytes [83, 84]. These results underscored SCEG-HiC's superiority in identifying biologically plausible SNP–gene associations.

Discussion

In this study, we presented SCEG-HiC, a machine learning framework that integrates single-cell omics correlations with prior chromatin conformation information to predict enhancer–gene links. SCEG-HiC is applicable to both paired scATAC/RNA-seq data and scATAC-seq data alone. Incorporating the bulk Hi-C prior enhances the accuracy of enhancer–gene link predictions, while retaining context-specific correlations and facilitating the identification of biologically relevant links. Using 10 single-cell multi-omics datasets from both humans and mice, we systematically benchmarked SCEG-HiC across three scenarios: (i) comparison with methods for paired multi-omics in human datasets, (ii) comparison with methods for scATAC-seq-only in human datasets, and (iii) comparison with methods for paired multi-omics in mouse datasets. SCEG-HiC consistently demonstrated high accuracy and robust performance in predicting enhancer–gene links across diverse tissues and cell types, validated using available cell type-specific Hi-C and eQTL data.

The key innovation of SCEG-HiC lies in its effective incorporation of prior knowledge to predict enhancer–gene links. Current methods for inferring gene regulatory relationships from single-cell omics data primarily rely on inherent statistical associations between gene expression and chromatin accessibility. These approaches are sensitive to noise and assume that correlation implies causation, an assumption often violated in cases of co-regulation, indirect regulation, or combinatorial effects of TFs, chromatin structure, and cellular states [85, 86]. Integrating prior knowledge as a structural constraint helps increase confidence in biologically meaningful connections, suppress spurious correlations, and improve both model generalizability and interpretability.

Given that Hi-C data were employed as both prior information and validation benchmarks, this could potentially introduce bias favoring our Hi-C-informed method. To ensure the validity of the model evaluation, two complementary strategies were adopted. On one hand, the ENCODE datasets used for constructing bulk average Hi-C maps do not cover all cell types used for validation in our study. Notably, in cell and tissue types not included in the prior Hi-C construction (e.g. human skin fibroblasts, glutamatergic neurons, as well as mouse thymic epithelial cells and liver cells), SCEG-HiC still maintained stable and superior predictive performance. On the other hand, under eQTL-based evaluation, which is entirely independent of Hi-C data, SCEG-HiC also significantly outperformed other methods. These results collectively indicate that the predictive advantage of SCEG-HiC is not driven by potential bias from the Hi-C-based validation standard.

Importantly, the Hi-C prior is derived from continuous contact matrices rather than discrete enhancer–gene links, and thus primarily reflects general chromatin spatial proximity features (e.g. distance-dependent contact and topologically associating domain). While these structural features are largely conserved across cell types, finer-scale regulatory interactions are often highly cell type-specific. Thus, a key question is whether our model can effectively identify enhancer–gene links that exhibit strong signals in single-cell data but lack or have only weak Hi-C support. Our results show that, compared to a model with Hi-C prior strength $\tau = 0$, SCEG-HiC retains most enhancer–gene links with high correlation in single-cell data. The removed links are not only relatively small in number but also typically lack independent functional support. This indicates that, distinct from previous chromatin structure-based models (e.g. ABC) where Hi-C contact is a limiting factor, the Hi-C prior primarily serves as a soft constraint in SCEG-HiC. Notably, SCEG-HiC outperformed ABC in capturing concordance between predicted cumulative enhancer activity and gene expression. In practice, adjusting the penalty parameter in SCEG-HiC can reduce dependence on prior Hi-C contact, enabling its application to single-cell data from novel tissues and conditions.

Leveraging scATAC-seq data from PBMCs of COVID-19 patients, we demonstrated the capacity of SCEG-HiC to recover molecular mechanisms underlying complex diseases. On one hand, combining TF binding site prediction and TF motif enrichment analysis enables systematic construction of enhancer-mediated GRNs, offering insights into cell type-specific regulatory programs governing diverse biological processes. On the other hand, SCEG-HiC facilitates functional prioritization of noncoding genetic variants from GWAS by accurately mapping them to their putative target genes.

While SCEG-HiC demonstrates robust performance and broad applicability, several limitations merit consideration and point to directions for future development. First, it relies on a bulk average Hi-C map from a variety of cell lines and is constrained in species with scarce chromatin interaction resources. Second, while the linear structure of wglasso provides good interpretability, it may overlook complex nonlinear regulatory mechanisms. Additionally, SCEG-HiC constructs static regulatory links that do not explicitly account for dynamic regulatory changes, necessitating further extensions to capture time-dependent processes such as cell fate transitions or disease progression. Future extensions may incorporate high-quality tissue-matched or computationally predicted Hi-C maps, as well as enhancer–gene links inferred across more cell types and diverse data sources (e.g. ABC-Max [87] and ENCODE rE2G [66]), as structural priors. In addition, adaptive weighting could be introduced to more effectively integrate single-cell evidence by selectively reducing regularization penalties for enhancer–gene links with strong correlations.

In conclusion, SCEG-HiC establishes a computational framework that integrates prior Hi-C information to infer enhancer–gene links from single-cell omics data. It offers a powerful computational tool for constructing transcriptional regulatory maps and advancing our understanding of the molecular mechanisms underlying complex traits and major diseases.

Acknowledgements

We thank Prof. Sijia Wang and Prof. Tao Huang for helpful discussions on this work.

Author contributions: X.L. and Z.W. designed the research. X.L. performed data analysis. X.L., Y.M., and D.H. built and tested the software. X.L. wrote the manuscript. Y.M., D.H., Y.L., W.Z., and Z.W. revised the manuscript. Z.W. supervised the study.

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

This work was supported by the National Key R&D Program of China (2021YFA1100501), the Major Project of Guangzhou National Laboratory (GZNL2024A01003), and the National Natural Science Foundation of China (32370690).

Data availability

All datasets used in this study are publicly available without restrictions. The paired scATAC/RNA-seq datasets were obtained from the 10X Genomics website and the Gene Expression Omnibus (GEO) database: PBMC (https://www.10xgenomics.com/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-10-k-1-standard-1-0-0), skin stromal (GSE195452), fetal retina (GSE246169), brain gray matter (GSE193240), developing cerebral cortex (GSE162170), mouse embryonic brain cells at day 18 (https://www.10xgenomics.com/datasets/fresh-embryonic-e-18-mouse-brain-5-k-1-standard-2-0-0), mouse skin (GSE140203), adult mouse cerebral cortex (GSE126074), thymic epithelial (GSE236075), and mouse liver (GSE218468). Cell type-specific Hi-C data used for benchmarking, as well as mouse Hi-C datasets used to generate the average Hi-C, were obtained from the ENCODE project and GEO database: CD4 + T (ENCFF762HKR), CD8 + T (ENCFF063KHP), CD14 + monocyte (ENCFF684YCT), brain midfrontal cortex (ENCFF925QIF), skin fibroblasts (GSM5266524), retinal (GSE229683), glutamatergic neurons (GSE228118), mouse embryonic forebrain (GSE201186), mouse cortex (GSE34587), skin epidermis/suprabasal (GSE197024), mTEC (GSE180933), liver (GSE65126), mouse embryonic stem cell (ENCFF409ZRS, ENCFF584EDJ), CH12F3 (ENCFF909ODS), mature B cell (ENCFF026SNO), CH12.LX (ENCFF076GYR), and epithelium/fiber (GSE243851). The eQTL data used for benchmarking were mainly from the GTEx eQTL summary statistics (v8) (https://www.gtexportal.org/home/datasets), retinal eQTL summary statistics (https://www-huge.uni-regensburg.de/databases.html), and eQTLGen summary statistics (release 2019-12-11) (https://www.eqtlgen.org). The COVID-19 PBMC scATAC-seq and scRNA-seq data were downloaded from GEO (GSE174072). The human bulk average Hi-C map was obtained from the ENCODE project (ENCFF134PUN). Details of the data sources are provided in Supplementary Tables S2, S4, and S5. The processed single-cell data used in this study, along with the mouse bulk average Hi-C map, are available via Zenodo: https://zenodo.org/record/14849886. SCEG-HiC is available as an R package at https://github.com/wuwei77lx/SCEGHiC. A GitHub repository containing all original code used for running and evaluating existing models is available at https://github.com/wuwei77lx/compare_model. The specific versions of the code associated with this publication are archived in Zenodo and are accessible via https://doi.org/10.5281/zenodo.18874012 and https://doi.org/10.5281/zenodo.18874419.

References

1. Casamassimi A, Ciccodicola A. Transcriptional regulation: molecules, involved mechanisms, and misregulation. *Int J Mol Sci* 2019;20:1281. https://doi.org/10.3390/ijms20061281
2. Lenhard B, Sandelin A, Carninci P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet* 2012;13:233–45. https://doi.org/10.1038/nrg3163
3. Pennacchio LA, Bickmore W, Dean A *et al.* Enhancers: five essential questions. *Nat Rev Genet* 2013;14:288–95. https://doi.org/10.1038/nrg3458
4. Claringbould A, Zaugg JB. Enhancers in disease: molecular basis and emerging treatment strategies. *Trends Mol Med* 2021;27:1060–73. https://doi.org/10.1016/j.molmed.2021.07.012
5. Sanyal A, Lajoie BR, Jain G *et al.* The long-range interaction landscape of gene promoters. *Nature* 2012;489:109–13. https://doi.org/10.1038/nature11279
6. Zhang Y, Wong CH, Birnbaum RY *et al.* Chromatin connectivity maps reveal dynamic promoter–enhancer long-range associations. *Nature* 2013;504:306–10. https://doi.org/10.1038/nature12716
7. Dixon JR, Jung I, Selvaraj S *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* 2015;518:331–6. https://doi.org/10.1038/nature14222
8. van Arensbergen J, van Steensel B, Bussemaker HJ. In search of the determinants of enhancer–promoter interaction specificity. *Trends Cell Biol* 2014;24:695–702. https://doi.org/10.1016/j.tcb.2014.07.004
9. Dekker J, Rippe K, Dekker M *et al.* Capturing chromosome conformation. *Science* 2002;295:1306–11. https://doi.org/10.1126/science.1067799
10. van Berkum NL, Lieberman-Aiden E, Williams L *et al.* Hi-C: a method to study the three-dimensional architecture of genomes. *J Vis Exp* 2010;39:1869.
11. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* 2017;550:204–13. https://doi.org/10.1038/nature24277
12. Gasperini M, Hill AJ, McFaline-Figueroa JL *et al.* A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* 2019;176:377–90. https://doi.org/10.1016/j.cell.2018.11.029
13. Moore JE, Pratt HE, Purcaro MJ *et al.* A curated benchmark of enhancer–gene interactions for evaluating enhancer–target gene prediction methods. *Genome Biol* 2020;21:17. https://doi.org/10.1186/s13059-019-1924-8
14. Hariprakash JM, Ferrari F. Computational biology solutions to identify enhancers–target gene pairs. *Comput Struct Biotechnol J* 2019;17:821–31. https://doi.org/10.1016/j.csbj.2019.06.012
15. Xu H, Zhang S, Yi X *et al.* Exploring 3D chromatin contacts in gene regulation: the evolution of approaches for the identification of functional enhancer–promoter interaction. *Comput Struct Biotechnol J* 2020;18:558–70. https://doi.org/10.1016/j.csbj.2020.02.013
16. Fulco CP, Nasser J, Jones TR *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR

perturbations. *Nat Genet* 2019;51:1664–9. https://doi.org/10.1038/s41588-019-0538-0
17. Tang F, Barbacioru C, Wang Y *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* 2009;6:377–82. https://doi.org/10.1038/nmeth.1315
18. Buenrostro JD, Giresi PG, Zaba LC *et al.* Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 2013;10:1213–8. https://doi.org/10.1038/nmeth.2688
19. Ma S, Zhang B, LaFave LM *et al.* Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* 2020;183:1103–16. https://doi.org/10.1016/j.cell.2020.09.056
20. Stoeckius M, Hafemeister C, Stephenson W *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* 2017;14:865–8. https://doi.org/10.1038/nmeth.4380
21. Kim D, Tran A, Kim HJ *et al.* Gene regulatory network reconstruction: harnessing the power of single-cell multi-omic data. *NPJ Syst Biol Appl* 2023;9:51. https://doi.org/10.1038/s41540-023-00312-6
22. Stuart T, Srivastava A, Madad S *et al.* Single-cell chromatin state analysis with Signac. *Nat Methods* 2021;18:1333–41. https://doi.org/10.1038/s41592-021-01282-5
23. Kartha VK, Duarte FM, Hu Y *et al.* Functional inference of gene regulation using single-cell multi-omics. *Cell Genomics* 2022;2:100166. https://doi.org/10.1016/j.xgen.2022.100166
24. Zhang L, Zhang J, Nie Q. DIRECT-NET: an efficient method to discover *cis*-regulatory elements and construct regulatory networks from single-cell multiomics data. *Sci Adv* 2022;8:eabl7393. https://doi.org/10.1126/sciadv.abl7393
25. Bravo González-Blas C, De Winter S, Hulselmans G *et al.* SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat Methods* 2023;20:1355–67. https://doi.org/10.1038/s41592-023-01938-4
26. Mitra S, Malik R, Wong W *et al.* Single-cell multi-ome regression models identify functional and disease-associated enhancers and enable chromatin potential analysis. *Nat Genet* 2024;56:627–36. https://doi.org/10.1038/s41588-024-01689-8
27. Poirion OB, Zuo W, Spruce C *et al.* Enhlink infers distal and context-specific enhancer–promoter linkages. *Genome Biol* 2024;25:235. https://doi.org/10.1186/s13059-024-03374-9
28. Duren Z, Chang F, Naqing F *et al.* Regulatory analysis of single cell multiome gene expression and chromatin accessibility data with scREG. *Genome Biol* 2022;23:114. https://doi.org/10.1186/s13059-022-02682-2
29. Su C, Lee D, Jin P *et al.* scMultiMap: cell-type-specific mapping of enhancers and target genes from single-cell multimodal data. *Nat Commun* 2025;16:3941. https://doi.org/10.1038/s41467-025-59306-z
30. Sheth MU, Qiu WL, Rosa Ma X *et al.* Mapping enhancer–gene regulatory interactions from single-cell data. bioRxiv, https://doi.org/10.1101/2024.11.23.624931, 24 November 2024, preprint: not peer reviewed.
31. Nagano T, Lubling Y, Stevens TJ *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* 2013;502:59–64. https://doi.org/10.1038/nature12593
32. Dautle MA, Chen Y. Single-cell Hi-C technologies and computational data analysis. *Adv Sci* 2025;12:e2412232. https://doi.org/10.1002/advs.202412232
33. Li X, Lee L, Abnousi A *et al.* SnapHiC2: a computationally efficient loop caller for single cell Hi-C data. *Comput Struct Biotechnol J* 2022;20:2778–83. https://doi.org/10.1016/j.csbj.2022.05.046
34. Liu W, Zhong W, Giusti-Rodríguez P *et al.* SnapHiC-G: identifying long-range enhancer–promoter interactions from single-cell Hi-C data via a global background model. *Brief Bioinform* 2024;25:bbae426. https://doi.org/10.1093/bib/bbae426
35. Hecker D, Behjati Ardakani F, Karollus A *et al.* The adapted activity-by-contact model for enhancer–gene assignment and its application to single-cell data. *Bioinformatics* 2023;39:btad062. https://doi.org/10.1093/bioinformatics/btad062
36. Li Y, Jackson SA. Gene network reconstruction by integration of prior biological knowledge. *G3* 2015;5:1075–9. https://doi.org/10.1534/g3.115.018127
37. Hao Y, Hao S, Andersen-Nissen E *et al.* Integrated analysis of multimodal single-cell data. *Cell* 2021;184:3573–87. https://doi.org/10.1016/j.cell.2021.04.048
38. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010;11:R106. https://doi.org/10.1186/gb-2010-11-10-r106
39. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* 2019;20:296. https://doi.org/10.1186/s13059-019-1874-1
40. Knight PA, Ruiz D. A fast algorithm for matrix balancing. *IMA J Numer Anal* 2013;33:1029–47. https://doi.org/10.1093/imanum/drs019
41. Sarnataro S, Chiariello AM, Esposito A *et al.* Structure of the human chromosome interaction network. *PLoS One* 2017;12:e0188201. https://doi.org/10.1371/journal.pone.0188201
42. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 2008;9:432–41. https://doi.org/10.1093/biostatistics/kxm045
43. Foygel R, Drton M. Extended Bayesian information criteria for Gaussian graphical models. *Adv Neural Inf Process Syst* 2010;23:604–12.
44. Zhang Y, Liu T, Meyer CA *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 2008;9:R137. https://doi.org/10.1186/gb-2008-9-9-r137
45. Korsunsky I, Millard N, Fan J *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* 2019;16:1289–96. https://doi.org/10.1038/s41592-019-0619-0
46. Kamimoto K, Stringa B, Hoffmann CM *et al.* Dissecting cell identity via network inference and *in silico* gene perturbation. *Nature* 2023;614:742–51. https://doi.org/10.1038/s41586-022-05688-9
47. Pliner HA, Packer JS, McFaline-Figueroa JL *et al.* Cicero predicts *cis*-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol Cell* 2018;71:858–71. https://doi.org/10.1016/j.molcel.2018.06.044
48. Pandey N, Omkar C, Mishra S *et al.* Improving chromatin-interaction prediction using single-cell open-chromatin profiles and making insight into the *cis*-regulatory landscape of the human brain. *Front Genet* 2021;12:738194. https://doi.org/10.3389/fgene.2021.738194
49. Durand NC, Shamim MS, Machol I *et al.* Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst* 2016;3:95–8. https://doi.org/10.1016/j.cels.2016.07.002
50. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 2020;369:1318–30. https://doi.org/10.1126/science.aaz1776
51. Korsunsky I, Nathan A, Millard N *et al.* Presto scales Wilcoxon and auROC analyses to millions of observations. bioRxiv, https://doi.org/10.1101/653253, 29 May 2019, preprint: not peer reviewed.
52. Pratapa A, Jalihal AP, Law JN *et al.* Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat Methods* 2020;17:147–54. https://doi.org/10.1038/s41592-019-0690-6
53. Kosicki M, Baltoumas FA, Kelman G *et al.* VISTA enhancer browser: an updated database of tissue-specific developmental enhancers. *Nucleic Acids Res* 2025;53:D324–30. https://doi.org/10.1093/nar/gkae940
54. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26:841–2. https://doi.org/10.1093/bioinformatics/btq033
55. Wilk AJ, Lee MJ, Wei B *et al.* Multi-omic profiling reveals widespread dysregulation of innate immunity and hematopoiesis in COVID-19. *J Exp Med* 2021;218:e20210582. https://doi.org/10.1084/jem.20210582

56. Fornes O, Castro-Mondragon JA, Khan A *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2020;48:D87–92.
57. Tan G, Lenhard B. TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics* 2016;32:1555–6. https://doi.org/10.1093/bioinformatics/btw024
58. Korhonen J, Martinmäki P, Pizzi C *et al.* MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics* 2009;25:3181–2. https://doi.org/10.1093/bioinformatics/btp554
59. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* 2019;15:e8746. https://doi.org/10.15252/msb.20188746
60. Yan F, Powell DR, Curtis DJ *et al.* From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol* 2020;21:22. https://doi.org/10.1186/s13059-020-1929-3
61. Tang Z, Luo OJ, Li X *et al.* CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* 2015;163:1611–27. https://doi.org/10.1016/j.cell.2015.11.024
62. Gur C, Wang SY, Sheban F *et al.* LGR5 expressing skin fibroblasts define a major cellular hub perturbed in scleroderma. *Cell* 2022;185:1373–88.e20. https://doi.org/10.1016/j.cell.2022.03.011
63. Wohlschlegel J, Finkbeiner C, Hoffer D *et al.* ASCL1 induces neurogenesis in human Müller glia. *Stem Cell Rep* 2023;18:2400–17. https://doi.org/10.1016/j.stemcr.2023.10.021
64. Meijer M, Agirre E, Kabbe M *et al.* Epigenomic priming of immune genes implicates oligodendroglia in multiple sclerosis susceptibility. *Neuron* 2022;110:1193–210.e13. https://doi.org/10.1016/j.neuron.2021.12.034
65. Trevino AE, Müller F, Andersen J *et al.* Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. *Cell* 2021;184:5053–69.e23. https://doi.org/10.1016/j.cell.2021.07.039
66. Gschwind AR, Mualim KS, Karbalayghareh A *et al.* An encyclopedia of enhancer–gene regulatory interactions in the human genome. bioRxiv, https://doi.org/10.1101/2023.11.09.563812, 13 November 2023, preprint: not peer reviewed.
67. Davis JJ, Goadrich MH. The relationship between precision–recall and ROC curves. In: *Proceedings of the 23rd International Conference on Machine Learning, ICML '06*. New York: Association for Computing Machinery, 2006, 233–40.
68. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489:57–74.
69. Park SG, Schulze-Luehrman J, Hayden MS *et al.* The kinase PDK1 integrates T cell antigen receptor and CD28 coreceptor signaling to induce NF-kappaB and activate T cells. *Nat Immunol* 2009;10:158–66. https://doi.org/10.1038/ni.1687
70. Hinton HJ, Alessi DR, Cantrell DA. The serine kinase phosphoinositide-dependent kinase 1 (PDK1) regulates T cell development. *Nat Immunol* 2004;5:539–45. https://doi.org/10.1038/ni1062
71. Chen S, Lake BB, Zhang K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat Biotechnol* 2019;37:1452–7. https://doi.org/10.1038/s41587-019-0290-0
72. Givony T, Leshkowitz D, Del Castillo D *et al.* Thymic mimetic cells function beyond self-tolerance. *Nature* 2023;622:164–72. https://doi.org/10.1038/s41586-023-06512-8
73. Bravo González-Blas C, Matetovici I, Hillen H *et al.* Single-cell spatial multi-omics and deep learning dissect enhancer-driven gene regulatory networks in liver zonation. *Nat Cell Biol* 2024;26:153–67. https://doi.org/10.1038/s41556-023-01316-4
74. Võsa U, Claringbould A, Westra HJ *et al.* Large-scale *cis*- and *trans*-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet* 2021;53:1300–10.
75. Hwang N, Huh Y, Bu S *et al.* Single-cell sequencing of PBMC characterizes the altered transcriptomic landscape of classical monocytes in BNT162b2-induced myocarditis. *Front Immunol* 2022;13:979188. https://doi.org/10.3389/fimmu.2022.979188
76. Mishra B, Ivashkiv LB. Interferons and epigenetic mechanisms in training, priming and tolerance of monocytes and hematopoietic progenitors. *Immunol Rev* 2024;323:257–75. https://doi.org/10.1111/imr.13330
77. Yu F, Cato LD, Weng C *et al.* Variant to function mapping at single-cell resolution through network propagation. *Nat Biotechnol* 2022;40:1644–53. https://doi.org/10.1038/s41587-022-01341-y
78. COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. *Nature* 2021;600:472–7. https://doi.org/10.1038/s41586-021-03767-x
79. Mantovani A, Sica A, Sozzani S *et al.* The chemokine system in diverse forms of macrophage activation and polarization. *Trends Immunol* 2004;25:677–86. https://doi.org/10.1016/j.it.2004.09.015
80. Merad M, Martin JC. Pathological inflammation in patients with COVID-19: a key role for monocytes and macrophages. *Nat Rev Immunol* 2020;20:355–62. https://doi.org/10.1038/s41577-020-0331-4
81. Stikker BS, Stik G, van Ouwerkerk AF *et al.* Severe COVID-19-associated variants linked to chemokine receptor gene control in monocytes and macrophages. *Genome Biol* 2022;23:96. https://doi.org/10.1186/s13059-022-02669-z
82. Payne DJ, Dalal S, Leach R *et al.* The CXCR6/CXCL16 axis links inflamm-aging to disease severity in COVID-19 patients. bioRxiv, https://doi.org/10.1101/2021.01.25.428125, 25 January 2021, preprint: not peer reviewed.
83. Dai Y, Wang J, Jeong HH *et al.* Association of CXCR6 with COVID-19 severity: delineating the host genetic factors in transcriptomic regulation. *Hum Genet* 2021;140:1313–28. https://doi.org/10.1007/s00439-021-02305-z
84. Kamat MA, Blackshaw JA, Young R *et al.* PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. *Bioinformatics* 2019;35:4851–3. https://doi.org/10.1093/bioinformatics/btz469
85. Larcombe MR, Hsu S, Polo JM *et al.* Indirect mechanisms of transcription factor-mediated gene regulation during cell fate changes. *Advanced Genetics* 2022;3:2200015. https://doi.org/10.1002/ggn2.202200015
86. Cheng Y, Brunner AL, Kremer S *et al.* Co-regulation of invected and engrailed by a complex array of regulatory sequences in *Drosophila*. *Dev Biol* 2014;395:131–43. https://doi.org/10.1016/j.ydbio.2014.08.021
87. Nasser J, Bergman DT, Fulco CP *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* 2021;593:238–43. https://doi.org/10.1038/s41586-021-03446-x

Received: September 21, 2025. **Revised:** April 10, 2026. **Accepted:** April 19, 2026