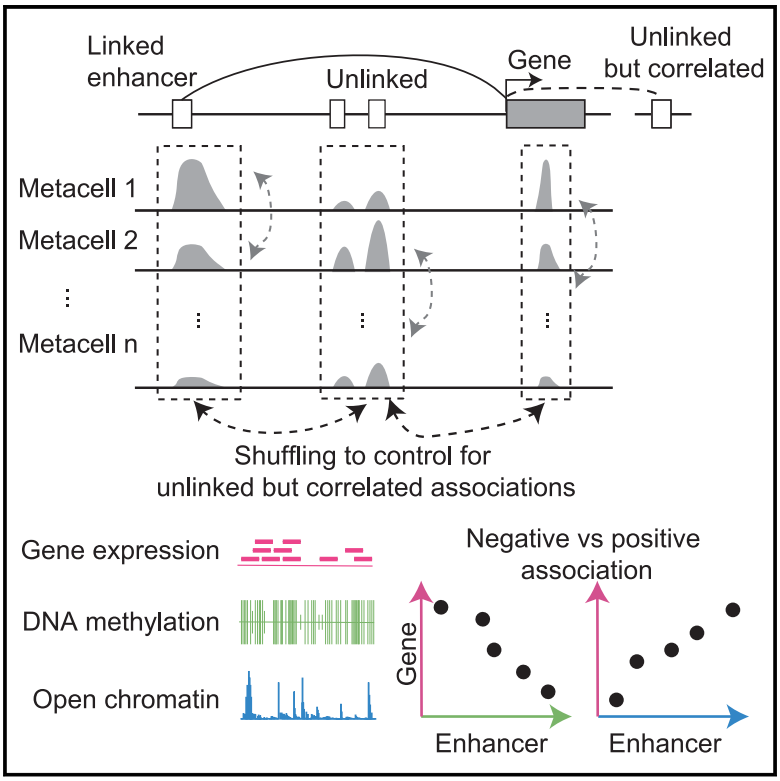
Cell Genomics

Short Article

Robust enhancer-gene regulation identified by single-cell transcriptomes and epigenomes

Graphical abstract



Highlights

- Link enhancers with genes using single-cell transcriptomic and epigenomic signatures
- Control false-positive associations caused by gene co-expression
- Enhancer methylation and accessibility are associated with target gene expression
- Linked enhancer-gene pairs are enriched in chromatin contact frequency

Authors

Fangming Xie, Ethan J. Armand, Zizhen Yao, ..., Bing Ren, Joseph R. Ecker, Eran A. Mukamel

Correspondence

emukamel@ucsd.edu

In brief

Integrated analysis of single-cell transcriptomes and epigenomes with a simple, permutation-based procedure detects cell-type-specific enhancer-gene associations. This procedure establishes stringent statistical criteria that control the risk of false-positive associations due to gene co-expression.

Xie et al., 2023, Cell Genomics 3, 100342
July 12, 2023
https://doi.org/10.1016/j.xgen.2023.100342

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Short Article

Robust enhancer-gene regulation identified by single-cell transcriptomes and epigenomes

Fangming Xie,[1,10,14] Ethan J. Armand,[2,14] Zizhen Yao,[3] Hanqing Liu,[4] Anna Bartlett,[4] M. Margarita Behrens,[5] Yang Eric Li,[6] Jacinta D. Lucero,[5,12] Chongyuan Luo,[7] Joseph R. Nery,[4] Antonio Pinto-Duarte,[5,13] Olivier B. Poirion,[6,11] Sebastian Preissl,[6,9] Angeline C. Rivkin,[4] Bosiljka Tasic,[3] Hongkui Zeng,[3] Bing Ren,[6] Joseph R. Ecker,[4,8] and Eran A. Mukamel[2,15,*]

[1]Department of Physics, University of California San Diego, La Jolla, CA 92037, USA
[2]Department of Cognitive Science, University of California San Diego, La Jolla, CA 92037, USA
[3]Allen Institute for Brain Science, Seattle, WA 98109, USA
[4]Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA
[5]Computational Neurobiology Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA
[6]Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla, CA 92037, USA
[7]Department of Human Genetics, University of California Los Angeles, Los Angeles, CA 90095, USA
[8]Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA
[9]Institute of Experimental and Clinical Pharmacology and Toxicology, Faculty of Medicine, University of Freiburg, Freiburg, Germany
[10]Department of Biological Chemistry, University of California Los Angeles, Los Angeles, CA 90095, USA
[11]The Jackson Laboratory, Farmington, CT, USA
[12]Present address: Department of Neurosciences, University of California, San Diego, La Jolla, CA 92037, USA
[13]Present address: Ionis Pharmaceuticals, Carlsbad, CA 92037, USA
[14]These authors contributed equally
[15]Lead contact
*Correspondence: emukamel@ucsd.edu
https://doi.org/10.1016/j.xgen.2023.100342

SUMMARY

Single-cell sequencing could help to solve the fundamental challenge of linking millions of cell-type-specific enhancers with their target genes. However, this task is confounded by patterns of gene co-expression in much the same way that genetic correlation due to linkage disequilibrium confounds fine-mapping in genome-wide association studies (GWAS). We developed a non-parametric permutation-based procedure to establish stringent statistical criteria to control the risk of false-positive associations in enhancer-gene association studies (EGAS). We applied our procedure to large-scale transcriptome and epigenome data from multiple tissues and species, including the mouse and human brain, to predict enhancer-gene associations genome wide. We tested the functional validity of our predictions by comparing them with chromatin conformation data and causal enhancer perturbation experiments. Our study shows how controlling for gene co-expression enables robust enhancer-gene linkage using single-cell sequencing data.

INTRODUCTION

Enhancer-gene association studies (EGAS) seek to define cell-type-specific gene regulatory networks via single-cell sequencing.[1–7] These studies have used a variety of experimental and computational strategies, leveraging diverse data modalities including single-cell transcriptomics and epigenomics. The gold standard for linking enhancers with target genes are experiments that perturb enhancer activity and measure gene expression in the same cells.[8,9] However, perturbation experiments are complex and, to date, have been limited to screening pre-selected enhancers in cell culture.[8,9] By contrast, single-cell transcriptomes and epigenomes from complex tissues, such as the brain, contain distinct genome-wide profiles from dozens to hundreds of cell types.[10,11] Correlating enhancer epigenetic profiles with transcription across cell types or cell states can identify potential cell-type-specific enhancer-gene links,[1,3–5,12] but the statistical validity and biological significance of these associations are unclear.

An obstacle to EGAS analysis is the widespread correlation of gene expression patterns across distinct cell types. Gene co-expression arises from shared functions in related cell types, such as the common expression of synaptic proteins in neurons, but not glial cells in the brain. Such co-expression reflects the hierarchical organization of cell types in terms of their functional and developmental relatedness.[13] In the context of EGAS, co-expression can create incidental associations between a gene and enhancers that are not directly linked in a regulatory interaction. Instead, those enhancers may regulate other genes whose expression pattern across cells is similar.

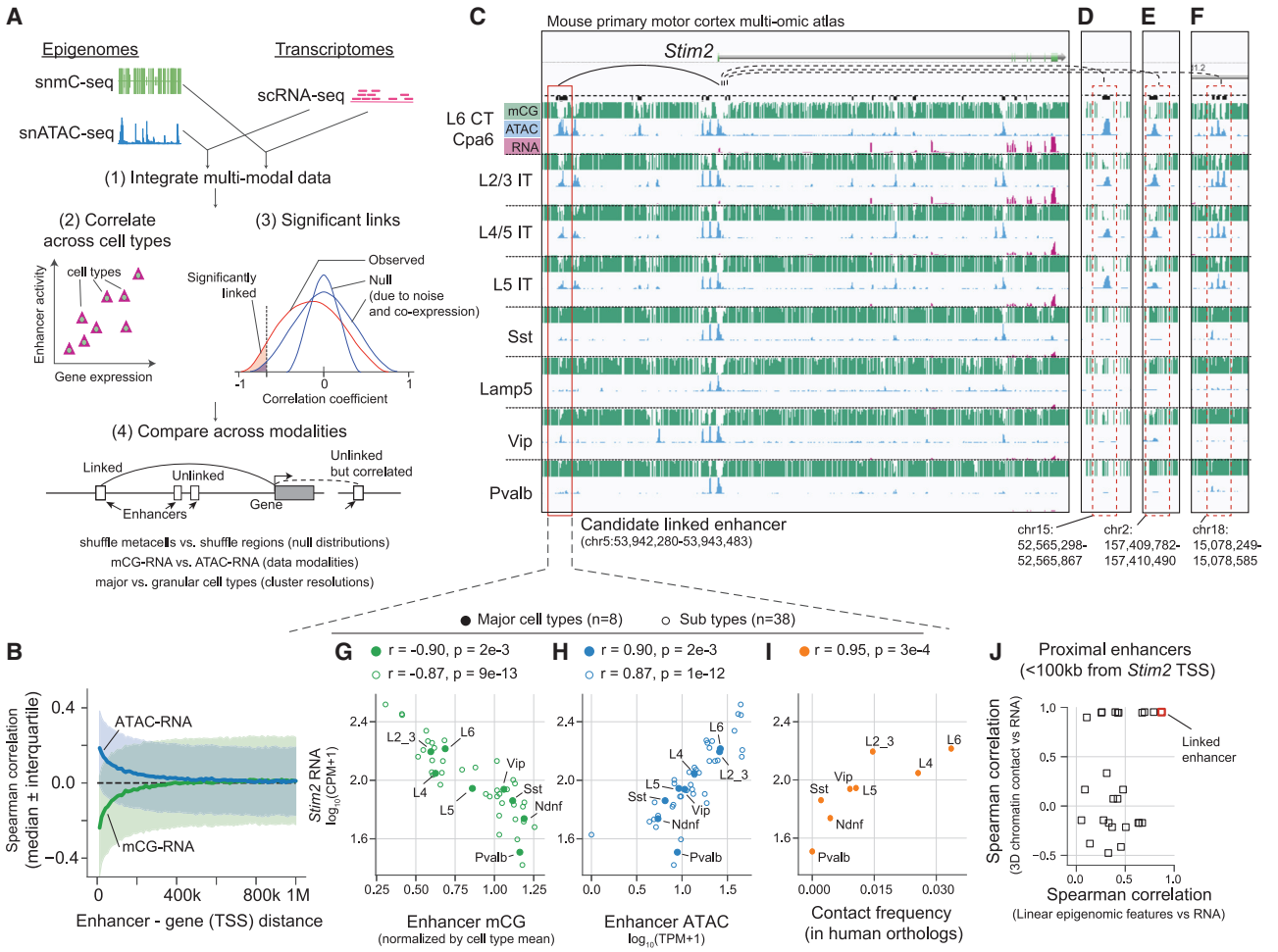


Figure 1. Identifying enhancer-gene links through integrated analysis of single-cell transcriptomes and epigenomes

(A) A procedure to link enhancers with target genes.

(B) Strength of enhancer-gene association as a function of genomic distance. The wide interquartile range (shading) indicates high variability.

(C–F) Correlation of the gene *Stim2* with nearby (C) and *trans* (D–F)-enhancers.

(G–I) *Stim2* expression versus enhancer mCG (G), ATAC (H), and enhancer-TSS chromatin contact frequency in human orthologs (I). The pseudobulk profiles are computed using major types and subtypes.

(J) Enhancer-gene association from linear-genome features (mCG, ATAC) versus 3D-genome features (chromatin contact frequency) for *Stim2* proximal enhancers. The x axis shows the minimum absolute correlation between mCG-RNA and ATAC-RNA.

These spurious associations are analogous to the effect of linkage disequilibrium in genome-wide association studies (GWAS),[14] in which many non-causal but statistically significant associations arise from their genetic correlation with a causal SNP within a risk locus.

To separate spurious from genuine associations, *trans*-enhancer-gene correlations can be used as a negative control.[2,6,7,15–17] However, major questions about the best practices for EGAS remain open. Different epigenome assays, such as single-nucleus assay of transposase-accessible chromatin (snATAC-seq)[5] and single-nucleus methylcytosine sequencing (snmC-seq),[4] measure distinct signatures of enhancer activity and may have different sensitivity and specificity for EGAS. In addition, correlation results may be strongly influenced by clustering analysis of single-cell data, which in turn depends on multiple unconstrained parameters and algorithmic choices.[18]

To address these gaps, we identify high-confidence, robust enhancer-gene links using a non-parametric procedure to control for gene co-expression by shuffling genomic regions[2,6,7,15–17] (Figures 1A and S1A). We leveraged three complementary data modalities (RNA, DNA methylation, and open chromatin) to cross-validate enhancer-gene links with independent data. We further validated our predictions with chromatin conformation data[19] (single-nucleus methyl-3C sequencing [snm3C-seq]) and with large-scale functional perturbation data.[8] Our study shows that single-cell sequencing can identify significant enhancer-gene links across diverse cell types despite the background of spurious associations from gene co-expression.

RESULTS

Gene co-expression confounds EGAS

To illustrate the risk of false associations due to gene co-expression, we analyzed a large set of single-cell sequencing data from the mouse primary motor cortex.[10] We integrated single-cell transcriptomes (single-cell RNA sequencing [scRNA-seq]) and epigenomes (open chromatin, snATAC-seq, and DNA methylation, snmC-seq), to generate multimodal profiles and estimate enhancer locations in over 50 cell types.[10] The integrated data showed that putative enhancers (see STAR Methods; Table S1; Figure S1B) within ~100 kb of a gene promoter were enriched in associations with gene expression, including positive correlations for chromatin accessibility and negative correlations for enhancer DNA methylation (mCG) (Figures 1B, S1C, and S1D). However, these associations were highly variable: we observed many weak correlations for proximal enhancers (<100 kb) and relatively strong correlations for some distal enhancers (>500 kb) (Figure 1B; interquartile range ~0.4). This broad distribution of correlation strength across all enhancer-gene pairs makes it difficult to identify the subset of enhancers that directly regulate their target gene.

A representative example is the gene *Stim2,* encoding a calcium sensor that helps maintain basal Ca^{2+} levels in pyramidal neurons.[20] In cortical neurons, we identified 33 putative enhancers within 100 kb of the *Stim2* promoter. *Stim2* expression correlates with low mCG ($r = -0.87$, $p = 9e{-}13$, $n = 38$ cell types) and high chromatin accessibility ($r = 0.87$, $p = 1e{-}12$) at a nearby enhancer (Figures 1C, 1G, and 1H). By contrast, 15 other nearby enhancers have weaker, though still significant (false discovery rate [FDR] < 0.05), correlation with *Stim2* expression ($|r| = 0.46$–0.85). Moreover, *Stim2* expression also correlated significantly with 25,027 other enhancers located throughout the genome (FDR < 0.05; both mCG-RNA and ATAC-RNA), most of which (n = 23,526) were located on different chromosomes (Figures 1D–1F). The majority of these *trans*-associations likely reflect gene co-expression rather than direct causal links with the *Stim2* gene. For example, many of these enhancers might directly regulate nearby genes whose expression patterns across cell types are similar to *Stim2* (Figures S1E–S1H, S1J, and S1K).

Next, we used 3D genome conformation data to test whether putative enhancer-gene links correspond to *bona fide* physical interactions.[21] We analyzed the 3D chromatin contact frequency of the predicted enhancer-gene pair (Figure 1C) across homologous human brain cell types using multiomics snm3C-seq data.[19] Chromatin contact frequency for this enhancer was strongly correlated with *Stim2* expression ($r = 0.95$, $p = 3e{-}4$; Figures 1I and S1I). By contrast, other proximal enhancers were less correlated (Figure 1J).

The *Stim2* locus also illustrates the challenges associated with defining cell types.[22] The same set of cells can be grouped into either 8 major types or 38 fine-grained subtypes, leading to different correlation values (Figures 1G, 1H, S1J, and S1K).

Permutation-based control for linking enhancers to genes

To address these issues, we developed a procedure that controls the risk of false positives from gene co-expression and compares predicted links across data modalities and cell-type resolutions (Figures 2A and S2). Our analysis can use multiomics data[23] or separate measurements of each data modality that are computationally combined through data fusion. To demonstrate the general applicability of our approach, we first integrated single-cell transcriptomes (RNA) and epigenomes (DNA methylation or chromatin accessibility) using correlated gene-level features to link cells of the same type across data modalities (SingleCellFusion).[10,24,25] Next, we defined metacells,[26] which aggregate the transcriptomic and epigenomic profiles from groups of similar cells. Each metacell has a complete multimodal profile, enabling analysis of correlated enhancer epigenetic features with gene expression. When working with multiomics data, we created metacells by aggregating single cells without a data fusion step. Importantly, metacells have a controllable resolution determined by the number of cells contributing to each metacell. This adjustable resolution allowed us to capture both discrete and continuous patterns of variation and to characterize the impact of cell-type resolution on enhancer-gene links.

We reasoned that genuine enhancer-gene interactions should have stronger correlations than the background induced by co-expression. Correlations mediated by co-expression are inherently limited in their strength by the magnitude of gene-gene correlations, whereas direct enhancer-gene interactions can produce stronger associations. Importantly, this assumption applies to the strongest enhancer-gene interactions; weak interactions that do not exceed the background of gene co-expression may be present but cannot be detected by correlation-based methods. This is a fundamental limitation of any correlation-based analysis method, which could be potentially overcome using perturbation experiments.

To test the significance of observed correlations, we compared them with two null distributions: shuffling metacells[3–5] and shuffling regions[2,6,7] (Figures 2B–2D). Shuffling metacells decouple epigenetic and transcriptomic signatures, removing both enhancer-gene correlation and gene co-expression (Figure 2B). The significance arising from this distribution is inflated by gene co-expression, potentially leading to false positives in which an enhancer-gene pair may be correlated due to shared upstream regulation rather than direct interaction. Shuffling regions, by randomly swapping the locations of genes across the genome, retains the gene co-expression structure imposed by the hierarchical organization of cell types, but it correlates each gene's expression with distant, randomly selected enhancers (Figure 2C).[2,6,7]

As expected, the null distribution of correlations after shuffling regions was wider than after shuffling metacells (Figure 2D) due to gene co-expression. Enhancer-gene pairs within 500 kb of the transcription start site (TSS) are significantly enriched in both positive and negative correlations when compared with shuffling metacells. However, when compared with shuffling regions, we only found significant correlations with a positive sign for the ATAC-RNA comparison or a negative sign for the mC-RNA comparsion. Shuffling regions is thus a more stringent null distribution that effectively controls for spurious enhancer-gene correlations due to gene co-expression.

We call an enhancer-gene pair significantly correlated if it passes an FDR-adjusted threshold based on shuffling metacells,

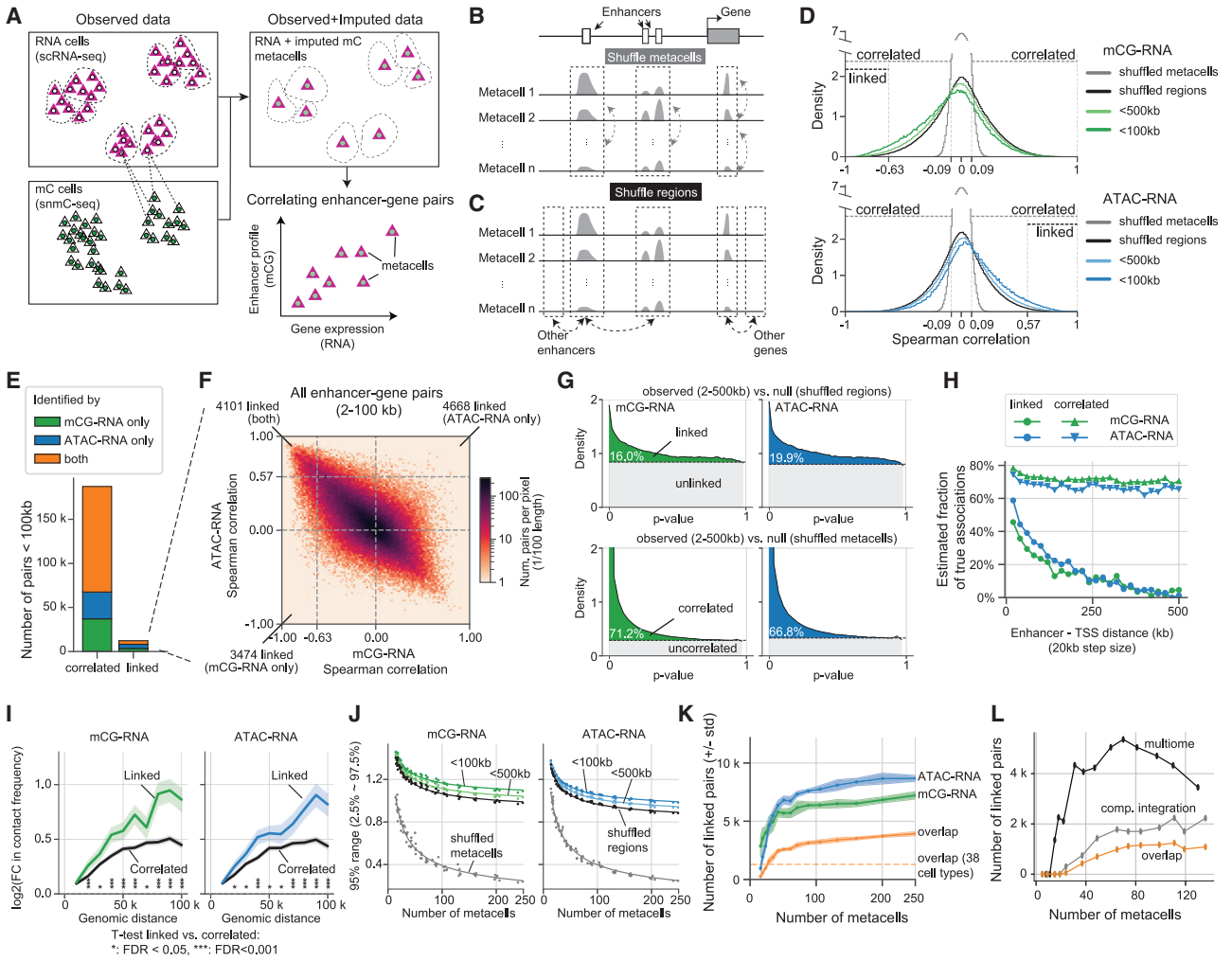


Figure 2. Systematic identification and characterization of enhancer-gene links across null models, data modalities, and cell-type granularity

(A) Method for linking enhancers to target genes using metacells with bimodality profiles.

(B and C) Null distributions derived from shuffling metacells (B) or shuffling regions (C).

(D) Distribution of enhancer-gene correlations. Bars indicate regions of statistical significance (FDR <0.2 for <100 kb pairs). Two null models induce two different bars: linked (black; shuffle regions) and correlated (gray; shuffle metacells).

(E) The number of significantly linked or correlated pairs using mCG-RNA, ATAC-RNA, or both.

(F) Joint distribution of mCG-RNA correlation versus ATAC-RNA correlation for enhancer-gene pairs (2–100 kb).

(G) p value histograms of enhancer-gene pairs (2–500 kb) using shuffled regions (top) or shuffled metacells (bottom). Numbers show estimated fraction of true positives.[27]

(H) Estimated fraction of true associations versus enhancer-TSS distance.

(I) Enrichment of chromatin contact frequency of linked and correlated enhancer-gene pairs compared with random genomic region pairs (mean ± 95% confidence interval). Enrichment profiles are aggregated across sites and across 8 neuronal cell types in Lee et al.[19]

(J) Spread (95% range) of correlation coefficients as a function of the number of metacells. Dots represent observed data; lines represent inverse square root fit ($y \sim a/\sqrt{x}+b$).

(K) Number of linked pairs as a function of the number of metacells (FDR = 0.2; mean ± standard deviation across 5 bootstrap samples with 80% of cells).

(L) Number of linked pairs called using either multiome information (10× multiome PBMCs[28]) or computational integration.

whereas we reserve the term "linked" for pairs that pass the criteria set by shuffling regions. We used a relatively lenient FDR threshold of 0.2 to reduce the risk of false negatives from our stringent null distribution. Linked pairs (n = 12,243 within 100 kb, FDR < 0.2) are a subset of correlated pairs (187,343 within 100 kb, FDR < 0.2) (Figures 2E and 2F) that rise above the background from gene co-expression. Lowering the FDR threshold to 0.1 or 0.05 reduced the number of linked pairs to 3,142 and 489, respectively.

Our shuffling procedure is a non-parametric analog of generalized least-squares (GLS) regression,[29] which transforms data matrices to decorrelate observations. We found that removing

sample covariance using GLS abolished the difference between shuffling regions and shuffling cells (Figures S3A and S3B) and between correlated and linked pairs (Figure S3C). In contrast with GLS, our procedure does not rely on parametric assumptions about gene co-expression.

Differences in the genomic context of different enhancers could impact the strength of their effect on gene expression. To control for this, we verified that our findings were robust when we randomly shuffled enhancers across the genome while controlling for their GC content and distance to the nearest gene (Figures S4A–S4D).

Comparison with alternative strategies

We compared our results with two alternative strategies for estimating enhancer-gene interactions using the same single-cell epigenome dataset.[10] CICERO uses chromatin accessibility (snATAC-seq) data to identify co-accessible regions near each gene's promoter, without transcriptome data.[1] The procedure does not explicitly control for the significance of associations but instead uses an arbitrary threshold (default 0.2) to link regions with the strongest co-accessibility. We used CICERO to identify 1,869 enhancer-gene associations exceeding the default threshold and located within 100 kb, far fewer than our set of linked enhancers (7,575 using mCG-RNA; 8,769 using ATAC-RNA). Most of the CICERO links (76%) overlap with a subset of the correlated pairs we identified (11%) and to a lesser degree with linked pairs (26.9% of CICERO links, 4.1% of our linked pairs) (Figures S5A and S5B). Mean CICERO co-accessibility scores were 4.8- to 5.9-fold higher ($p < 2e-8$) for linked than correlated pairs (Figure S5C). This indicates that CICERO had lower power for detecting linked pairs compared with our analysis of combined snATAC+scRNA-seq data.

The activity-by-contact (ABC) model[9] identifies enhancer-gene links using both chromatin accessibility (snATAC) and chromatin conformation (e.g., Hi-C) data. This model processes each cell type independently without considering correlated variability in expression across cells. The ABC model can use snATAC-seq without matched Hi-C data by substituting a power-law function of distance to estimate chromatin contact frequency.[30] Using the ABC model, we identified 150,228 associations within 100 kb, closely matching the number identified by our snATAC-RNA (150,285) and mCG-RNA analyses (156,932). The ABC linked pairs largely overlap our correlated (70%) and linked pairs (68%–73%) (Figures S5D and S5E). The ABC scores are 1.09- to 1.22-fold higher ($p < 1e-8$) for linked pairs than for correlated pairs (Figure S5F). This comparison shows that ABC largely recapitulates our analysis of correlated enhancers without control for gene co-expression. Linked enhancers represent a distinct set of strongly correlated candidate enhancer-gene pairs that are not fully captured by CICERO or ABC. These candidate enhancer-gene pairs must be validated by perturbation experiments performed in the same cell types.

Biological and statistical validation of EGAS links

A potential pitfall of our stringent procedure is a higher risk of false negatives, i.e., failure to detect genuine interactions. We next empirically compared correlated versus linked pairs on both biological and statistical criteria to test whether the correlations filtered out by our method are likely false positives arising from gene co-expression.

First, we observed that correlated pairs frequently had a non-canonical direction of association (Figures 2D and S6A). For example, we found that about a third (47,137/150,285) of correlated pairs had a negative association of gene expression with chromatin accessibility and that a similar proportion (53,687/156,932) had a positive association with mCG. Non-canonical associations were also reported in recent large-scale studies of brain cell epigenomes.[4,5] These correlations could suggest novel biological mechanisms such as methylcytosine-preferring transcription factors.[31] However, they may also include false-positive associations due to gene co-expression. Indeed, none of the non-canonical associations passed our threshold for linked pairs (Figures 2D–2F). This is consistent with the canonical understanding of enhancer activity associating primarily with low DNA methylation and open chromatin.

Second, as enhancer-gene interactions are mainly within ~100–500 kb of the TSS,[8,9] we compared the distance dependence of linked and correlated pairs. Using a p value histogram method,[27] we estimated that 16%–19.9% of enhancers detected within 2–500 kb from a promoter are linked (Figures 2G and S6B). By contrast, a much larger fraction (66.8%–71.2%) are estimated to be correlated. Notably, the proportion of correlated pairs remains high (>60%) even for distal pairs (>1 Mb) or *trans*-pairs on different chromosomes. In contrast, <5% of these pairs are linked (Figures 2H and S6C). The correlated pairs contradict the biological understanding that most enhancers activate genes in *cis*; the linked pairs are more coherent with this canonical framework.

Third, we validated the predicted links with independent chromatin conformation data, generated by snm3C-seq from the human brain.[19] We reasoned that linked enhancer-gene pairs that are conserved across species should have higher chromatin contact frequency compared with random regions. Indeed, we found enrichment of contact frequency for both linked (mean fold change [FC] = 1.51, $p = 2e-4$) and correlated pairs (mean FC = 1.28, $p = 2e-5$). Moreover, linked pairs located 20–100 kb apart have higher levels of contact enrichment than correlated pairs ($p = 0.047$; Figures 2I and S6D–S6F).

Impact of cell cluster granularity

A key parameter for our analysis is the cell-type granularity, determined by the number of metacells or clusters. The sparse genomic coverage of single-cell sequencing and the limited number of profiled cells create a trade-off between the number of metacells and the quality of each metacell—i.e., between fine-grained resolution and signal-to-noise ratio. As the number of metacells (N) increases, the width of the null distribution for the shuffled metacells approaches zero as $\frac{1}{\sqrt{N}}$ (see STAR Methods; Figures 2J and S7A–S7C). By contrast, the range of the null distribution for shuffled regions does not vanish for large N but instead asymptotes at a non-zero value that reflects gene co-expression (Figure S7B). Notably, the shuffled region's null distribution is less sensitive to the number of metacells and more closely reflects the behavior of the observed correlations. This suggests that linked enhancer-gene pairs are less sensitive to the choice of cell-type granularity correlated pairs. We found

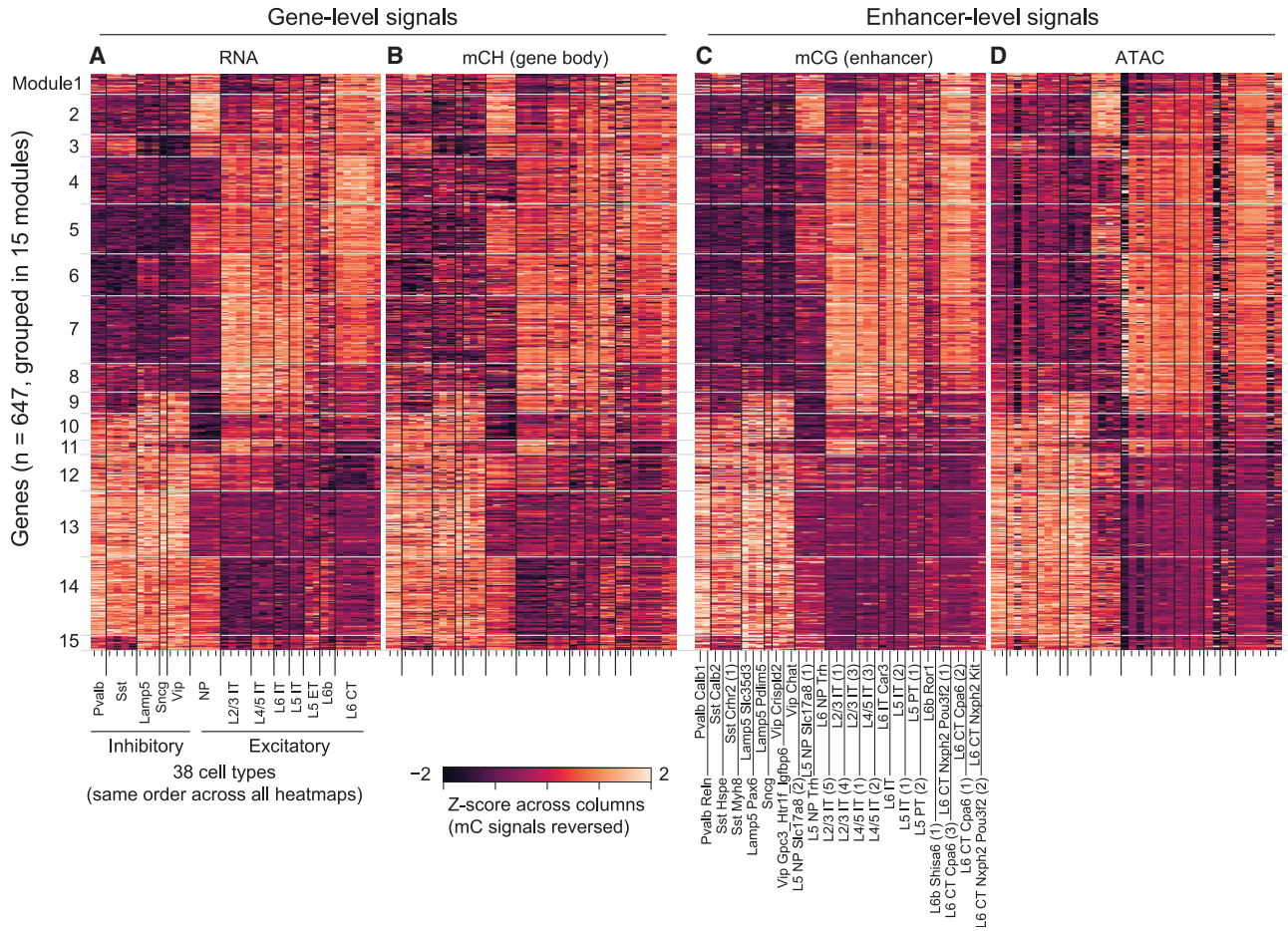


Figure 3. Consistent gene- and enhancer-level signatures across data modalities

Gene expression (A), gene body non-CG DNA methylation (mCH) (B), enhancer mCG (C), and enhancer accessibility (ATAC) (D) across cell types. Signals from multiple enhancers linked to the same gene were averaged. The colormaps for the mC modalities (mCH and mCG) are reversed.

more linked pairs as the number of metacells increases but with diminishing returns after N > 50. (Figures 2K and S7D).

Multiome data improve link detection

Our methods can also be applied to multiomics data, which profile transcriptome and epigenome signatures in the same cells and do not require computational data fusion. Using ATAC-RNA multiomics data (10× Genomics, human cerebellum[32] and peripheral blood mononuclear cells [PBMCs][28]; Table S5), we identified thousands of linked enhancer-gene pairs (6,182 in cerebellum, 5,452 in PBMCs; Figures S8A and S8C) with strong positive associations (Figures S8B and S8D). Using mCG-RNA multiomics data (snmCAT-seq, human frontal cortex[25]; Table S5), we identified 616 linked pairs with strong negative associations (Figure S8E). Linked pairs were also more enriched in chromatin contacts (snm3C-seq, human frontal cortex[19]; Table S5) than correlated pairs (Figure S8F).

We compared the effectiveness of enhancer-gene linkage using multiome data with computational integration of single-modality data using the PBMC dataset.[28] As the number of metacells increased from a few to ~70, the accuracy of computational integration (using SingleCellFusion[25]) decreased from over 80% to less than 20% (Figure S8G). More than twice as many links were found using multiome information than using computational integration (Figures 2L and S8H). Thus, when multiome data are available, they have the potential to improve the sensitivity and specificity of enhancer-gene link calling.

Functional validation of predicted enhancer-gene links

To validate our predicted links, we used data from perturbations of enhancer activity in human K562 cells using CRISPR-dCas9 followed by scRNA-seq.[8] Functionally validated enhancer-gene pairs from causal CRISPR-dCas9-based perturbations had stronger correlations in the multiomics datasets compared with other proximal (<100 kb) enhancer-gene pairs (median r = 0.18 for cerebellum, 0.23 for PBMCs; Figures S9A and S9C). We then directly compared our predicted links with the functional study at enhancers that were tested in both datasets (1,606 enhancers in cerebellum; 2,345 in PBMCs). Notably, this comparison was limited by the different cell types used for the functional study compared with our analyses. Despite this, we found that predicted linked pairs overlapped with functionally

validated pairs 7.1- to 8.5-fold more than expected by chance (Figures S8B and S8D; $p < 0.001$). Moreover, a >5-fold higher proportion of the predicted linked pairs than correlated pairs was validated in the functional assay (Figures S9E and S9F). The precision of our predicted linked pairs was notable given the differences in the cell types between the datasets, as well as the limited power of both the functional validation and single-cell enhancer-gene linkage assays. Future perturbation experiments, including *in vivo* measurements in matched cell types,[33] will better evaluate our predicted links.

Thousands of enhancer-gene links in mouse brain cells with multimodal validation

We comprehensively examined regulatory interactions in neurons of the mouse primary motor cortex.[10] Linked enhancer-gene pairs (discovered using mCG-RNA correlation) formed 15 modules that capture diverse cell-type-specific signatures. Notably, the complementary data types (gene body DNA methylation [mCH] and enhancer ATAC), which were not used in the discovery of enhancer-gene links, showed consistent correlated signatures (Figures 3A–3D). For example, genes in module 13 are specifically expressed in GABAergic neurons, with corresponding low CG methylation levels and open chromatin at linked enhancers. Module 9 is most active in caudal ganglionic eminence (CGE)-derived inhibitory neurons (Lamp5, Sncg, and Vip) and in superficial-layer excitatory neurons (L2/3 IT and L4/5 IT). These consistent gene- and enhancer-level signals from three data modalities provide support for the robustness of our enhancer-gene links.

DISCUSSION

Our procedure shares some features with Signac,[15] a single-cell chromatin analysis tool that also evaluates statistical significance by shuffling regions (enhancers). However, unlike Signac, which tests each enhancer-gene pair separately, we efficiently test millions of enhancer-gene pairs at the same time with robust empirical p values and multiple comparison correction. In addition, by applying our procedure to different data modalities (gene expression, DNA methylation, and chromatin accessibility), we found that mCG-RNA and ATAC-RNA associations are strikingly consistent, despite measuring distinct epigenetic features with opposite effects on gene expression. Predicted enhancer-gene links are robust with respect to the granularity used to analyze cell types across a broad range of parameters. Our method can be applied to computationally integrated single-modality datasets or to multiomics data. Single-modality data are more widely available and generally offer higher quality and throughput, but their use for enhancer-gene linkage analysis is limited by the accuracy and resolution of computational integration. By contrast, multiomics data can directly correlate enhancer activity with gene expression at the level of fine-grained metacells or single cells, taking full advantage of the natural variation across cells to find enhancer-gene links.

Limitations of the study

Correlation-based analysis has notable limitations for linking enhancers with genes. This approach cannot identify constitutive enhancer-gene links that are present in all cell types. Larger datasets including more diverse tissues or cell types may partly address this limitation. Rigorous control for spurious correlations limits the power for detecting genuine but weak enhancer-gene interactions, leading us to potentially underestimate the number of genuine links. Finally, causal interactions cannot be inferred from correlational analysis alone; our linked pairs are strong candidates that must be tested by perturbative experiments.[34,35] Improved experimental techniques, including large-scale assays[8,9,36] *in vivo*[33] on complex tissues with diverse cell types, will help to test correlation-based predictions. Our study shows that single-cell transcriptomic and epigenomic data can identify statistically and biologically robust enhancer-gene links in complex tissues, a prerequisite for elucidating the regulatory principles of cell-type-specific gene expression.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Datasets
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Calling putative enhancers
  - Curated cell types
  - Clustering and defining metacells
  - Feature selection and normalization
  - Correlating enhancer-gene pairs across cell types
  - Correlating enhancer-gene pairs across metacells
  - Estimating the statistical significance of enhancer-gene links
  - Enrichment of 3D chromatin contact frequencies
  - Comparison with CICERO
  - Comparison with the ABC model
  - Multiomics (ATAC + RNA) and functional validation analysis
  - Human snmCAT-seq multiomics and chromatin contact validation analysis
  - Generalized least squares (GLS) analysis to decouple covariance across metacells
  - Expected range of correlation coefficients for independent variables

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2023.100342.

ACKNOWLEDGMENTS

We gratefully acknowledge members of the Mukamel, Ecker, Ren, and Zeng laboratories and collaborators within the BRAIN Initiative Cell Census Network (BICCN). This work was funded by the NIH BRAIN Initiative (RF1 MH120015 to

E.A.M., U19MH114830 to H.Z., and U19MH121282 to J.R.E.; J.R.E is an Investigator of the Howard Hughes Medical Institute) and by CZI Collaborative Computational Tools for the Human Cell Atlas (to E.A.M.).

AUTHOR CONTRIBUTIONS

E.A.M. and F.X. designed the study. Z.Y., B.T., and H.Z. generated scRNA-seq data. H.L., A.B., M.M.B., J.D.L., C.L., J.R.N., A.P.-D., A.C.R., and J.R.E. generated DNA methylation (snmC-seq) data. H.L., M.M.B., Y.E.L., J.D.L., A.P.-D., O.B.P., S.P., and B.R. generated snATAC-seq data. F.X. led the computational analysis. F.X. and E.J.A. developed code and performed analysis. F.X., E.J.A., and E.A.M. wrote and edited the manuscript. All authors approved the manuscript.

DECLARATION OF INTERESTS

J.R.E. serves on the scientific advisory board of Zymo Research, Inc. B.R. is a shareholder of Arima Genomics, Inc. and Epigenome Technologies, Inc. H.Z. is on the scientific advisory board of MapLight Therapeutics, Inc.

Received: August 11, 2022
Revised: March 9, 2023
Accepted: May 17, 2023
Published: June 19, 2023

REFERENCES

1. Pliner, H.A., Packer, J.S., McFaline-Figueroa, J.L., Cusanovich, D.A., Daza, R.M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A., et al. (2018). Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. Mol. Cell *71*, 858–871.e8.
2. Corces, M.R., Granja, J.M., Shams, S., Louie, B.H., Seoane, J.A., Zhou, W., Silva, T.C., Groeneveld, C., Wong, C.K., Cho, S.W., et al. (2018). The chromatin accessibility landscape of primary human cancers. Science *362*, eaav1898.
3. Zhu, C., Yu, M., Huang, H., Juric, I., Abnousi, A., Hu, R., Lucero, J., Behrens, M.M., Hu, M., and Ren, B. (2019). An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. Nat. Struct. Mol. Biol. *26*, 1063–1070.
4. Liu, H., Zhou, J., Tian, W., Luo, C., Bartlett, A., Aldridge, A., Lucero, J., Osteen, J.K., Nery, J.R., Chen, H., et al. (2021). DNA methylation atlas of the mouse brain at single-cell resolution. Nature *598*, 120–128.
5. Li, Y.E., Preissl, S., Hou, X., Zhang, Z., Zhang, K., Qiu, Y., Poirion, O.B., Li, B., Chiou, J., Liu, H., et al. (2021). An atlas of gene regulatory elements in adult mouse cerebrum. Nature *598*, 129–136.
6. Trevino, A.E., Sinnott-Armstrong, N., Andersen, J., Yoon, S.-J., Huber, N., Pritchard, J.K., Chang, H.Y., Greenleaf, W.J., and Pașca, S.P. (2020). Chromatin accessibility dynamics in a model of human forebrain development. Science *367*, eaay1645.
7. Ma, S., Zhang, B., LaFave, L.M., Earl, A.S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V.K., Tay, T., et al. (2020). Chromatin potential identified by shared single-cell profiling of RNA and chromatin. Cell *183*, 1103–1116.e20.
8. Gasperini, M., Hill, A.J., McFaline-Figueroa, J.L., Martin, B., Kim, S., Zhang, M.D., Jackson, D., Leith, A., Schreiber, J., Noble, W.S., et al. (2019). A genome-wide framework for mapping gene regulation via cellular genetic screens. Cell *176*, 1516.
9. Fulco, C.P., Nasser, J., Jones, T.R., Munson, G., Bergman, D.T., Subramanian, V., Grossman, S.R., Anyoha, R., Doughty, B.R., Patwardhan, T.A., et al. (2019). Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. Nat. Genet. *51*, 1664–1669.
10. Yao, Z., Liu, H., Xie, F., Fischer, S., Adkins, R.S., Aldridge, A.I., Ament, S.A., Bartlett, A., Behrens, M.M., Van den Berge, K., et al. (2021). A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. Nature *598*, 103–110.
11. Yao, Z., van Velthoven, C.T.J., Nguyen, T.N., Goldy, J., Sedeno-Cortes, A.E., Baftizadeh, F., Bertagnolli, D., Casper, T., Chiang, M., Crichton, K., et al. (2021). A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. Cell *184*, 3222–3241.e26.
12. Cusanovich, D.A., Hill, A.J., Aghamirzaie, D., Daza, R.M., Pliner, H.A., Berletch, J.B., Filippova, G.N., Huang, X., Christiansen, L., DeWitt, W.S., et al. (2018). A single-cell atlas of in vivo mammalian chromatin accessibility. Cell *174*, 1309–1324.e18.
13. Zeng, H., and Sanes, J.R. (2017). Neuronal cell-type classification: challenges, opportunities and the path forward. Nat. Rev. Neurosci. *18*, 530–546.
14. Slatkin, M. (2008). Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. Nat. Rev. Genet. *9*, 477–485.
15. Stuart, T., Srivastava, A., Madad, S., Lareau, C.A., and Satija, R. (2021). Single-cell chromatin state analysis with Signac. Nat. Methods *18*, 1333–1341.
16. Gorkin, D.U., Barozzi, I., Zhao, Y., Zhang, Y., Huang, H., Lee, A.Y., Li, B., Chiou, J., Wildberg, A., Ding, B., et al. (2020). An atlas of dynamic chromatin landscapes in mouse fetal development. Nature *583*, 744–751.
17. Sarropoulos, I., Sepp, M., Frömel, R., Leiss, K., Trost, N., Leushkin, E., Okonechnikov, K., Joshi, P., Giere, P., Kutscher, L.M., et al. (2021). Developmental and evolutionary dynamics of cis-regulatory elements in mouse cerebellar cells. Science *373*, eabg4696. https://doi.org/10.1126/science.abg4696.
18. Kiselev, V.Y., Andrews, T.S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. Nat. Rev. Genet. *20*, 273–282.
19. Lee, D.-S., Luo, C., Zhou, J., Chandran, S., Rivkin, A., Bartlett, A., Nery, J.R., Fitzpatrick, C., O'Connor, C., Dixon, J.R., and Ecker, J.R. (2019). Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. Nat. Methods *16*, 999–1006.
20. Serwach, K., and Gruszczynska-Biegala, J. (2019). STIM proteins and glutamate receptors in neurons: role in neuronal physiology and neurodegenerative diseases. Int. J. Mol. Sci. *20*, 2289.
21. Schoenfelder, S., and Fraser, P. (2019). Long-range enhancer-promoter contacts in gene expression control. Nat. Rev. Genet. *20*, 437–455.
22. Endersby, J. (2009). Lumpers and splitters: Darwin, Hooker, and the search for order. Science *326*, 1496–1499.
23. Zhu, C., Preissl, S., and Ren, B. (2020). Single-cell multimodal omics: the power of many. Nat. Methods *17*, 11–14.
24. Welch, J.D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., and Macosko, E.Z. (2019). Single-cell multi-omic integration compares and contrasts features of brain cell identity. Cell *177*, 1873–1887.e17.
25. Luo, C., Liu, H., Xie, F., Armand, E.J., Siletti, K., Bakken, T.E., Fang, R., Doyle, W.I., Stuart, T., Hodge, R.D., et al. (2022). Single nucleus multiomics identifies human cortical cell regulatory genome diversity. Cell Genom. *2*, 100107.
26. Baran, Y., Bercovich, A., Sebe-Pedros, A., Lubling, Y., Giladi, A., Chomsky, E., Meir, Z., Hoichman, M., Lifshitz, A., and Tanay, A. (2019). MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. Genome Biol. *20*, 206.
27. Nettleton, D., Hwang, J.T.G., Caldo, R.A., and Wise, R.P. (2006). Estimating the number of true null hypotheses from a histogram of p values. J. Agric. Biol. Environ. Stat. *11*, 337–356.
28. 10X Multiome human PBMC. https://www.10xgenomics.com/resources/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-10-k-1-standard-2-0-0.
29. Aitken, A.C. (1936). On least squares and linear combination of observations. Proc. R. Soc. Edinb. *55*, 42–48.

30. Nasser, J., Bergman, D.T., Fulco, C.P., Guckelberger, P., Doughty, B.R., Patwardhan, T.A., Jones, T.R., Nguyen, T.H., Ulirsch, J.C., Lekschas, F., et al. (2021). Genome-wide enhancer maps link risk variants to disease genes. Nature *593*, 238–243.

31. Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P.K., Kivioja, T., Dave, K., Zhong, F., et al. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science *356*, eaaj2239.

32. 10X Multiome human cerebellum. https://www.10xgenomics.com/resources/datasets/frozen-human-healthy-brain-tissue-3-k-1-standard-2-0-0.

33. Jin, X., Simmons, S.K., Guo, A., Shetty, A.S., Ko, M., Nguyen, L., Jokhi, V., Robinson, E., Oyler, P., Curry, N., et al. (2020). In vivo Perturb-Seq reveals neuronal and glial abnormalities associated with autism risk genes. Science *370*, eaaz6063. https://doi.org/10.1126/science.aaz6063.

34. Daigle, T.L., Madisen, L., Hage, T.A., Valley, M.T., Knoblich, U., Larsen, R.S., Takeno, M.M., Huang, L., Gu, H., Larsen, R., et al. (2018). A suite of transgenic driver and reporter mouse lines with enhanced brain-cell-type targeting and functionality. Cell *174*, 465–480.e22.

35. Graybuck, L.T., Daigle, T.L., Sedeño-Cortés, A.E., Walker, M., Kalmbach, B., Lenz, G.H., Morin, E., Nguyen, T.N., Garren, E., Bendrick, J.L., et al. (2021). Enhancer viruses for combinatorial cell-subclass-specific labeling. Neuron *109*, 1449–1464.e13.

36. de Boer, C.G., Vaishnav, E.D., Sadeh, R., Abeyta, E.L., Friedman, N., and Regev, A. (2020). Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. Nat. Biotechnol. *38*, 56–65.

37. Abdennur, N., and Mirny, L.A. (2020). Cooler: scalable storage for Hi-C data and other genomically labeled arrays. Bioinformatics *36*, 311–316.

38. He, Y., Gorkin, D.U., Dickel, D.E., Nery, J.R., Castanon, R.G., Lee, A.Y., Shen, Y., Visel, A., Pennacchio, L.A., Ren, B., and Ecker, J.R. (2017). Improved regulatory element prediction based on tissue-specific local epigenomic signatures. Proc. Natl. Acad. Sci. USA *114*, E1633–E1640.

39. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

40. Amemiya, H.M., Kundaje, A., and Boyle, A.P. (2019). The ENCODE blacklist: identification of problematic regions of the genome. Sci. Rep. *9*, 9354.

41. Traag, V.A., Waltman, L., and van Eck, N.J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep. *9*, 5233.

42. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B Stat. Methodol. *57*, 289–300.

43. Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J., and Mirny, L.A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat. Methods *9*, 999–1003.

44. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.

45. Schraivogel, D., Gschwind, A.R., Milbank, J.H., Leonce, D.R., Jakob, P., Mathur, L., Korbel, J.O., Merten, C.A., Velten, L., and Steinmetz, L.M. (2020). Targeted Perturb-seq enables genome-scale genetic screens in single cells. Nat. Methods *17*, 629–635.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
scRNA-seq from mouse primary motor cortex (MOp)	Yao et al.[10]	nemo:dat-ch1nqb7
snmC-seq from mouse MOp	Yao et al.[10]	nemo:dat-ch1nqb7
snATAC-seq from mouse MOp	Yao et al.[10]	nemo:dat-ch1nqb7
Integrated genome browser of mouse MOp data	Yao et al.[10]	https://brainome.ucsd.edu/BICCN_MOp
snm3C-seq from human cortex	Lee et al.[19]	https://salkinstitute.app.box.com/s/fp63a4j36m5k255dhje3zcj5kfuzkyj1
Multiome (RNA+ATAC; 10X Genomics) from human PBMC	10X Genomics	https://www.10xgenomics.com/resources/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-10-k-1-standard-2-0-0
Multiome (RNA+ATAC; 10X Genomics) from human cerebellum	10X Genomics	https://www.10xgenomics.com/resources/datasets/frozen-human-healthy-brain-tissue-3-k-1-standard-2-0-0
snmCAT-seq from human cortex	Luo et al.[25]	Raw data: GEO GSE140493 Clustering and DMRs: Luo et al.[25] Table S6 and Table S9.
Functional validation data (CRISPER-dCas9 + scRNA-seq) in K562 cell lines	Gasperini et al.[8]	Gasperini et al.[8] Table S2
Software and algorithms		
robustlink	This paper	https://github.com/mukamel-lab/robustlink Zenodo: https://doi.org/10.5281/zenodo.7911853
SingleCellFusion	Luo et al.[25]	https://github.com/mukamel-lab/SingleCellFusion
ABC-Enhancer-Gene-Prediction	Fulco et al.[9]	https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction
CICERO	Pliner et al.[1]	https://www.bioconductor.org/packages/release/bioc/html/cicero.html
Cooler	Abdennur and Mirny[37]	https://github.com/open2c/cooler

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to the lead contact, Eran A. Mukamel (emukamel@ucsd.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Sequencing data used in this project are all from previous studies and are publicly available. Accession numbers and links to the datasets are listed in the key resources table.
- All original code has been deposited at GitHub, archived at Zenodo and is publicly available as of the date of publication. GitHub URLs and DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Datasets

We used three single-cell sequencing datasets from the mouse primary motor cortex (MOp)[10] to demonstrate our enhancer-gene association analysis. They are scRNA-seq (single cell; 10x genomics V3; Allen Institute for Brain Science), snmC-seq (single nucleus; DNA methylation; Ecker lab at the Salk Institute), and snATAC-seq (single nucleus; chromatin accessibility; Ren lab at UCSD). Only

high-quality neuronal cells, as determined in Ref.[10] (from its Table S2; column SCF/SingleCellFusion), are retained for our analysis. These datasets are publicly available with identifiers listed in the key resources table. The starting point of all analyses are gene-by-cell matrices for transcriptomes and enhancer-by-cell matrices for epigenomes. For the scRNA-seq dataset, we used the gene-by-cell count matrix. For the snATAC-seq dataset, we quantified the enhancer-by-cell count matrix by counting the number of reads overlapping with each enhancer region in each cell. For the snmC-seq dataset, we quantified both enhancer-by-cell CG DNA methylation profiles and gene-by-cell non-CG (CH) DNA methylation profiles. The DNA methylation profile for a particular region and cell can be summarized by two numbers: the number of methylated cytosines (mC) and the total number of cytosines covered (C). The DNA methylation level is the ratio of mC to C (mC/C). Please see sections below for dataset specific procedures of normalizations. The mouse gene annotation file is downloaded from gencode (vM16). The enhancer list is adapted from the putative enhancer list from Ref.[10] (see below).

We also used single-cell multiomics data, multimodal chromatin conformation data, and functional perturbation data to validate our methods. See Table S5 for a summary of all the datasets used in this study. See sections below for dataset-specific processing procedures.

QUANTIFICATION AND STATISTICAL ANALYSIS

Calling putative enhancers

We constructed our putative enhancer list based on the mouse MOp neuronal cell type-specific putative enhancers from Ref.[10] (from its Table S7). In that study, the enhancers are called using REPTILE,[38] an algorithm that uses the DNA methylation and ATAC-seq profiles of 13 mouse neuronal cell types, as well as mouse embryonic stem cells, as input. Starting from this list, we first selected regions with enhancer score >0.5 and merged overlapping regions using bedtools.[39] We subsequently removed regions overlapping any gene promoter regions (transcription start site +/− 2kb; all transcripts from gencode vM16), exons (vM16), and ENCODE blacklist.[40] This leaves us with 233,524 enhancers in total, with a median size of ~250 bp (Figure S1B; Table S1).

Curated cell types

For analyses related to Figure 1, we curated a list of 38 neuronal cell clusters based on the SingleCellFusion clusters (L1 and L2, with n = 29 to 56 cell types respectively) in Ref.[10]. We aimed to merge small clusters to increase pseudo bulk coverage at enhancers, while retaining as much cell type diversity as possible. To achieve this, we first call an enhancer *covered* in a cluster if it has at least 20 sequenced CpG sites in that cluster, where the cluster-level coverage is the sum of cell-level coverages. Next, we call an enhancer *common*, if it is covered in more than half of the L2 clusters. We call a cluster *covered*, if more than half of the common enhancers are covered in that cluster. For each L1 cluster we then evaluate 3 cases.

1. If the cluster itself is not covered, we drop it along with all its child (L2) clusters.
2. Else if less than 2 ($n < 2$) of its child (L2) clusters are covered, we retain the L1 cluster itself, but drop all its child (L2) clusters.
3. Else if at least 2 ($n \geq 2$) of its child (L2) clusters are covered, we retain the covered L2 clusters, but drop the uncovered L2 clusters and the L1 cluster.

This procedure resulted in 38 clusters with adequate coverage. Table S4 summarized the correspondence between the 38 clusters we get from this procedure and the cell types defined in Ref.[10].

To compare with the cell types in snm3C-seq data,[19] we further merged these 38 fine-grained clusters into 8 major clusters based on the well-established neuronal cell type taxonomy.[13] Table S4 summarized the correspondence between the 38 fine grained and the 8 major cell clusters defined in this study and those defined in Ref.[10,19].

Clustering and defining metacells

For analyses related to Figure 2, we generated cell clusterings with a range of cluster resolutions. We start by normalizing the scRNA-seq count matrix with $\log_{10}$(CPM+1), where CPM stands for counts per million mapped reads. We then calculated the top 50 principal components (PCs), and built a k-nearest neighbor graph (k = 30) connecting cells according to the Euclidean distance in the PC space. We used Leiden community detection to generate clusters.[41] Different resolution parameters (r = 1–794) were chosen to generate clusters with different granularity (n = 13–8850 metacells). The pseudo bulk profiles from each of the individual clusters were used as metacells.

Feature selection and normalization

We preprocessed the data matrices separately for each data modality. The starting point is always cell-level matrices containing counts (RNA and ATAC) or methylation level (mC). To get cluster-level (metacell) matrices, we summed counts from cells in the same clusters (metacells) to create pseudo-bulk samples. For methylation data, we summed methylated counts and total counts (coverage) separately. Next, we normalized matrices as follows.

- For an RNA matrix (gene-by-cluster/metacell), we normalize the raw count matrix with $\log_{10}$(CPM+1).

- For an ATAC matrix (enhancer-by-cluster/metacell), we normalize the raw count matrix with $\log_{10}$(TPM+1), where TPM stands for transcripts per million mapped reads. Enhancers that are covered in <50% of clusters are removed.
- For a gene body mCH matrix (gene-by-cluster/metacell), we first removed low coverage genes if the gene has <50% clusters surpassing 1000 counts in the gene body (or <80% metacells surpassing 20 counts). We then take the ratio of the number of methylated to the number of coverage to get the methylation fraction. All the steps here consider cytosines in non-CG (CH) dinucleotide context only.
- For an enhancer mCG matrix (enhancer-by-cluster/metacell), we first removed low coverage enhancers if the gene has <50% clusters surpassing 20 counts (or <80% metacells surpassing 5 counts) in the enhancer region. We then take the ratio of the number of methylated to the number of coverage to get the methylation fraction. All the steps here consider cytosines in CG dinucleotide context only.

After normalization and filtering of individual matrices, we then consider only enhancers that are shared in both ATAC and mCG matrices for downstream analyses.

Correlating enhancer-gene pairs across cell types

We calculate the Spearman correlation coefficient between any pair of enhancer and gene that are within 1 Mbp (enhancer center to gene TSS) across curated cell types (n = 38 or n = 8). This was done separately for enhancer mCG vs. RNA and enhancer ATAC vs. RNA. Enhancer mCG signals are normalized by the global mean mCG levels of each cell type; enhancer ATAC signals are $\log_{10}$(TPM+1) normalized; RNA expression levels are $\log_{10}$(CPM+1) normalized.

To assess the statistical significance of the enhancer-gene correlations, we repeated the correlation analysis with 2 types of data shuffling control, as explained in the main text. To control for random noise, we shuffled cell cluster labels of the gene-by-cluster RNA matrix, followed by calculating correlation coefficients. To control for background co-expression across enhancer-gene pairs, we shuffled gene labels of the gene-by-cluster RNA matrix, followed by calculating correlation coefficients.

Correlating enhancer-gene pairs across metacells

Given a transcriptomic dataset (scRNA-seq) and an epigenetic dataset (e.g. snmC-seq) collected from the same tissue, we first generate a constrained k-nearest neighbor network linking cells across the two modalities (SingleCellFusion; Ref.[10,25]). This network allows us to impute the DNA methylation profiles (mC) for each RNA cell. We then cluster scRNA-seq cells using Leiden community detection[41] (see section Clustering/Generating metacells). We call these clusters *metacells*, to emphasize that they do not necessarily correspond to discrete cell types, but could also capture continuous changes among cell populations. These preparations allow us to construct bimodal profiles for each metacell, by aggregating counts–either observed or imputed–from cells in the same metacells. Finally, we evaluate the correlations between enhancer-gene pairs across metacells.

To be specific, the starting point of this analysis involves 4 matrices: an enhancer-by-cell mCG (or ATAC) matrix E_{ec}, a gene-by-cell RNA matrix $R_{gc'}$, a cross-modal cell-to-cell k nearest neighbor matrix: $K_{cc'}$, and a metacell assignment matrix of RNA cells $K_{c'z}$. Here we use c, c' and z to denote an mC cell, an RNA cell, and a metacell, respectively. A metacell is a group of RNA cells generated by Leiden clustering. We use e and g to denote an enhancer and a gene, respectively. All matrices contain unnormalized raw counts. $K_{cc'}$ is generated by SingleCellFusion[10,25] with default settings and cross-modal k = 30. $K_{c'z}$ is generated by Leiden clustering on the RNA-seq dataset as mentioned in previous sections.

To get bimodal profiles for a metacell, we aggregate counts from the cells belonging to that metacell: $R_{gz} = \sum_{c'} R_{gc'} K_{c'z}$, and $E_{ez} = \sum_{c} E_{ec} K_{cc'} K_{c'z}$. The metacell profiles are then normalized as mentioned in previous sections to adjust for metacell size, library size, and gene length. Finally, normalized R_{gz} and E_{ez} allow us to correlate a specific pair of gene $g^{(i)}$ and enhancer $e^{(i)}$ across metacells (z). We calculated Spearman correlation coefficients for all enhancer-gene pairs with distance between 2kb and 1Mb (enhancer center - TSS).

Estimating the statistical significance of enhancer-gene links

To assess the statistical significance of a correlation coefficient r, we constructed two null distributions by shuffling metacells and shuffling regions. In the first case, we shuffle metacell labels independently for transcriptomic and epigenetic data, such that the two data modalities become independent of each other. In the second case, we permute genes by randomly swapping the location of each gene with that of another gene, while keeping the labels of metacells. Permuting genes randomizes the spatial relationship of enhancer-gene pairs, and is equivalent to permuting enhancers but are more computationally efficient.

Either null distribution can be used to get empirical p values and false discovery rate (FDR). The empirical p value of a correlation coefficient r is defined as the cumulative fraction of the null distribution that has more extreme (stronger) correlation coefficients than r. We calculated two-sided p values when using the shuffled metacells distribution, and single-sided p values when using the shuffled regions distribution. FDRs are then calculated using the Benjamini-Hochberg procedure.[42] We call an enhancer-gene pair significantly *linked* (*correlated)* if its empirical FDR is < 0.2 using shuffling regions (metacells) as the null.

To see if the shuffled regions distribution depends on enhancer properties such as its sequence GC content and distance to the nearest gene, we also performed stratified shuffling analyses (Figure S4). To control for GC content, we first grouped enhancers into 10 bins (deciles) according to their GC content. We then shuffled enhancers within each bin, i.e., randomly swapping each enhancer's

location with that of another enhancer that has similar GC content. The same procedure was separately done to control for enhancers' distance to the nearest gen as well.

Enrichment of 3D chromatin contact frequencies

We validated the predicted enhancer-gene links using single-cell measurements of 3D-chromatin contact frequency in human prefrontal cortex.[19] Raw contact matrices of 8 neuronal cell types were downloaded as mcool files.[19] We calculated contact frequencies from raw counts using matrix balancing using Cooler.[37,43] We then focused on analyzing these contact frequency matrices at a resolution of 10kb non-overlapping genomic bins across the genome.

To compare our enhancer-gene links predicted in the mouse brain with the chromatin contact data from human brain, we lifted genes (gencode vM16 whole genes) and putative enhancers from mm10 to hg19 using LiftOver[44] with parameters -minMatch = 0.8 and -minBlocks = 1.00.

To calculate enrichment, we first assigned enhancers (center) and genes (TSS) to their corresponding genomic bins (non-overlapping 10kb bins genomewide). We compared the contact frequencies of the predicted enhancer-gene pairs with random genomic region pairs with similar genomic distance. We separately tested the enrichment of contact frequencies of 6 groups of predicted enhancer-gene pairs: mCG-RNA linked, ATAC-RNA linked, pairs linked by both modalities, mCG-RNA correlated, ATAC-RNA correlated, and pairs correlated in both modalities. For each of the 8 neuronal cell types, we only include pairs that are active in the specific cell type, i.e. whose gene expression is greater than the median across all 8 cell types.

Comparison with CICERO

We installed the R package CICERO[1] from the Bioconductor following the instructions from the authors' tutorial (https://cole-trapnell-lab.github.io/cicero-release/docs_m3/#constructing-cis-regulatory-networks). We ran CICERO on MOp ATAC-seq data using default parameters. The program takes as input a peak-by-cell ATAC-seq matrix, where peaks include both putative enhancers we specified and gene promoters (500 bp upstream of TSS). The program returns co-accessibility scores for peak pairs. We filtered the output down to enhancer-promoter pairs only, removing enhancer-enhancer and promoter-promoter pairs. We also focused on analyzing enhancer-gene pairs that are within 100kb apart, to compare with our correlation-based analysis. We used a threshold = 0.2 following ref.[12] to call positive enhancer-gene pairs.

Comparison with the ABC model

We downloaded code from the github repository of the ABC model[9] (https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction) and followed instructions. We ran ABC for each MOp cell type (n = 38) using our identified putative enhancer list (n = 233,524) and pseudo-bulk ATAC-seq and RNA-seq data as input. We used genomic-distance based power law estimation to model chromatin contacts (–score_column powerlaw.Score). The software returns a score (ABC score) for each enhancer-gene pair and cell type. We excluded the expressed genes from the results, as suggested by the authors. We also focused on analyzing enhancer-gene pairs that are within 100kb. We used a threshold = 0.022 as recommended by the authors to call positive enhancer-gene pairs.

Multiomics (ATAC + RNA) and functional validation analysis

We downloaded *single cell Multiome ATAC and gene expression* datasets,[28,32] which were generated from human cerebellum and peripheral blood mononuclear cells (PBMCs). Please see Table S5 for detailed descriptions of the datasets. Notably, the cells in 10X multiome datasets are pre-filtered by 10X Genomics to exclude the low-quality cells. We did not impose extra cell-level quality control on top of the existing ones. For each dataset, we were provided with a cell-by-gene RNA count matrix, and a cell-by-peak ATAC count matrix for the same cells.

For each of the two 10X multiome dataset, we generated metacells by clustering single cells based on their gene expression information using a off-the-shelf workflow of Scanpy. This includes library size normalization using *scanpy.pp.normalize_total*, log transformation of normalized counts using *scanpy.pp.log1p*, highly-variable gene selection using *scanpy.pp.highly_variable_genes*, reducing dimensions using *scanpy.tl.pca* (n_comps = 50), building a cell-cell neighboring graph using *scanpy.pp.neighbors* (n_neighbors = 10), and Leiden clustering using *scanpy.tl.leiden* (resolution = 1–10). We adopted default parameters unless otherwise noted. A range of resolution (1–10) was used to generate clusterings of different granularities.

For each clustering, we merged single-cell count matrices, for both the cell-by-gene RNA matrix and the cell-by-peak ATAC matrix, into metacell level count matrices. We then removed lowly expressed genes ($\leq$100 counts in total across all metacells), followed by normalization and correlation across metacells with shuffling controls to identify enhancer-gene links (see Section estimating the statistical significance of enhancer-gene links). As a result, we identified thousands of linked enhancer-gene pairs (FDR <0.2) for a range of clustering resolution. For each dataset, we focused on reporting the results using the clustering that generates the most number of significant pairs.

We next seeked to validate our predicted links by comparing with functional experimental data. In the past few years, high-throughput functional examinations of enhancer-gene links have been developed, by combining enhancer activity perturbation using

CRISPR-dCas9 and gene expression readouts.[8,9,45] To the best of our knowledge, ref.[8] contains the largest functional enhancer-gene validation experiment to date, which perturbed 5,779 putative enhancers in the K562 human leukemia cell line, and found 664 positive enhancer-gene links (FDR<0.1).

To make a direct comparison, we lifted over enhancer coordinates perturbed by ref.[8] from hg19 (n = 5,779) to hg38 (n = 5,778), including all 664 positive ones. We next overlapped them with the ATAC-peaks from the 10X multiome datasets. Requiring at least 20 bp overlap, we found 1,606/133,233 ATAC peaks in the human cerebellum dataset overlap with those enhancers, and 2,345/143,160 ATAC peaks in the human PBMC dataset overlap with those enhancers. For the overlapped enhancers, we then compared the overlap between functional validation results and multiome predictions by generating a 2-by-2 confusion matrix of enhancer-gene pairs, and tested its significance by Fisher exact test. Other common statistics, including precision and recall, were derived from the confusion matrix.

Human snmCAT-seq multiomics and chromatin contact validation analysis

We acquired a single-nucleus methylCytosine, chromatin Accessibility and Transcriptome sequencing (snmCAT-seq) dataset from human frontal cortex from the co-authors of ref.[25]. This assay measures both DNA methylomes and transcriptomes for the same cells. Notably, the DNA methylation profiles in this dataset is confounded by its chromatin accessibility component, which uses a GpC methyltransferase M.CviPI to methylate GpC sites at accessible chromatin regions. To remove this confounding factor, we characterized CG DNA methylation level using HCG sites (ACG, TCG, CCG) only, whenever it occurs below.

To get a list of putative enhancers for this dataset, we started from a list of cell type-specific differentially methylated regions (DMRs) provided by Table S9 in ref.[25]. We concatenated hypo-DMRs for all neuronal cell types (n = 2,172,541), removed any DMRs with less than 3 CpG sites (n = 1,439,125 remain), and merged overlapping regions using *bedtools merge* (n = 412,730 remain). We subsequently filtered the remaining non-overlapping DMRs by keeping only regions in autosomes and chromosome X, removing the top and bottom 1% of regions by length (>3,477 bp and <36 bp). This in the end left us with n = 402,665 regions, which we used as putative enhancers for the following analysis.

We quantified CG methylation profiles at enhancers and gene expression profiles for 52 neuronal subtypes identified in ref.[25] by summing over the counts from single-cell profiles (n = 3,898). To get robust enhancer-gene correlation, we then filtered out lowly expressed genes (total counts $\leq$100 summing over all subtypes), and included DMRs with coverage $\geq$10 in $\geq$90% cell types. We then ran correlation analysis across these 52 metacells to identify enhancer-gene links, following the same methods described in other parts of the methods.

We next compared our predicted links from snmCAT-seq with chromatin conformation data profiled by snm3C-seq,[19] following the same procedure as described in section Enrichment of 3D chromatin contact frequencies. Notably, in this case, both datasets were generated from the human frontal cortex and analyzed using the same genome version (hg19), therefore the regions are directly comparable with no genome liftover needed. We reconciled the cell type resolution difference (n = 52 for snmCAT-seq and n = 8 for snm3C-seq) based on cell type annotations provided by the original ref. [19,25]. Their exact correspondence was also documented in Table S4.

Generalized least squares (GLS) analysis to decouple covariance across metacells

We used GLS[29] to test the association between gene expression and enhancer activity across cell types (metacells). We will focus on only one given enhancer-gene pair (g, e), as the same procedure applies to all enhancer-gene pairs independently. Given an enhancer e and gene g, Let y_{cg} be the mRNA expression in cell type c, x_{ce} be the enhancer activity (e.g., mC or ATAC). Let C be the number of cell types. A linear model associating g and e can be written as:

$$y_c = a + \beta x_c + \varepsilon_c \quad \text{(Equation 1)}$$

where c is the index for cell types, β is the association strength, and ε is a noise term. In addition, a is an intercept term that can be omitted after data centering (x and y can be pre-centered to ensure $E[y_c] = E[x_c] = 0$). In matrix notation, (Equation 1) can be simply noted as $y = \beta x + \varepsilon$.

In ordinary least squares (OLS), we assume ε is uncorrelated across cell types: $E[\varepsilon_c] = 0, E[\varepsilon_c \varepsilon_{c'}] = \sigma^2 \delta_{c,c'}$. The correlation coefficient $r = E[xy]/\sigma_x \sigma_y$ is then a measure of the linear association, and it has an associated p value calculated using the t distribution. Alternatively, inference can be performed by permutation analysis to get an empirical p value.

However, in our case we have correlated noise: $E[\varepsilon_c \varepsilon_{c'}] = \Omega_{c,c'}$, which reflects the correlation between cell types due to gene co-expression. That is, $\Omega_{c,c'}$ represents the background of correlated variability in gene expression due to the hierarchical structure of cell types in complex tissues. We can estimate the correlation using the genome-wide covariance, $\widehat{\Omega}_{c,c'} = Cov[y]_{c,c'}$. In this case, generalized least squares[29] (GLS) can be used to give an estimate of the coefficient β. This corresponds to transforming the variables x, y from the original basis (cell types/metacells, denoted) to an decorrelated basis (denoted r), and then performing OLS on the decorrelated variables.

We first use singular value decomposition (SVD) to decompose the mean-subtracted gene expression matrix, $y_{cg} = \sum_r U_{cr} S_{rr} V^T_{rg}$, where $r = min(c, g)$. Defining $Z = US$, we have $\Omega = ZZ^T$. Multiplying both sides of (Equation 1) by $Z^{-1} = S^{-1}U^T$ corresponds to a transformation from correlated with decorrelated (or whitened) basis:

$$y' = \beta x' + \varepsilon' \quad \text{(Equation 2)}$$

where $y' = Z^{-1}y$, $x' = Z^{-1}x$, and $\varepsilon' = Z^{-1}\varepsilon$. The noise term is now uncorrelated, because

$$Cov[\varepsilon'] = E\left[\varepsilon'\varepsilon'^{T}\right] = E\left[\varepsilon\varepsilon^{T}\left(Z^{-1}\right)^{T}\right] = Z^{-1}\Omega\left(Z^{-1}\right)^{T} = Z^{-1}ZZ^{T}\left(Z^{-1}\right)^{T} = I$$

where I is the identity matrix. We can therefore use the correlation coefficient and its associated test statistics on transformed data y' and x', as in the case of OLS.

Expected range of correlation coefficients for independent variables

Here we provide theoretical justification on why we expect the range of correlation coefficients ($\hat{r}$) to scale as $\frac{1}{\sqrt{N}}$, as seen in Figures 2J and S7B, where N is the number of metacells.

Let X and Y be two independent random variables. Let x_i and y_i be independent and identically distributed samples of X and Y, where $i \in \{1,2,\ldots,N\}$. In our case, N represents the number of metacells, and x_i and y_i are the transcriptomic and epigenetic signals for a given enhancer-gene pair for metacell i. We require X and Y to be independent of each other as they are unlinked, and x_i and y_i be independent samples as different metacells are also independent observations of X and Y, such as in the case of null distribution created by shuffling cells.

To simplify the notation, we assume $E[X] = E[Y] = 0$, as the mean does not affect correlation coefficient r. We also assume X and Y are symmetric, as in the case of normal distribution. It is obvious that $r(X,Y) = 0$. However, we are interested in how the variance of $\hat{r}$ depends on N, where $\hat{r}$ is the sample estimate of r by $\{x_i\}$ and $\{y_i\}$.

$$var[\hat{r}] \sim E(\hat{r}^2) \sim E\left[\frac{\left(\sum_{i=1}^{N} x_i y_i\right)^2}{\sum_{i=1}^{N} x_i^2 \cdot \sum_{i=1}^{N} y_i^2}\right] = E\left[\frac{\sum_{a=1}^{N}\sum_{b=1}^{N} x_a y_a x_b y_b}{\sum_{i=1}^{N} x_i^2 \cdot \sum_{i=1}^{N} y_i^2}\right] = \sum_{a=1}^{N}\sum_{b=1}^{N} E\left[\frac{x_a y_a x_b y_b}{\sum_{i=1}^{N} x_i^2 \cdot \sum_{i=1}^{N} y_i^2}\right]$$
$$= \sum_{a=1}^{N} E\left[\frac{(x_a y_a)^2}{\sum_{i=1}^{N} x_i^2 \cdot \sum_{i=1}^{N} y_i^2}\right] \quad \text{(Equation 3)}$$

The last equality holds, as only non-interaction terms ($a = b$) are nonzero. Moreover, as $(x_a y_a)^2$ are equivalent for different $a = \{1\ldots N\}$, the above summation can be further simplified as:

$$\sum_{a=1}^{N} E\left[\frac{(x_a y_a)^2}{\sum_{i=1}^{N} x_i^2 \cdot \sum_{i=1}^{N} y_i^2}\right] = N \cdot E\left[\frac{(x_1 y_1)^2}{\sum_{i=1}^{N} x_i^2 \cdot \sum_{i=1}^{N} y_i^2}\right] = N \cdot E\left[\frac{x_1^2}{\sum_{i=1}^{N} x_i^2}\right] \cdot E\left[\frac{y_1^2}{\sum_{i=1}^{N} y_i^2}\right], \quad \text{(Equation 4)}$$

where $E\left[\frac{x_1^2}{\sum_{i=1}^{N} x_i^2}\right] = \frac{1}{N} E\left[\frac{\sum_{i=1}^{N} x_i^2}{\sum_{i=1}^{N} x_i^2}\right] = \frac{1}{N}$, due to the symmetry among indices. Therefore, we finally arrive at

$$var(\hat{r}) \propto N \cdot E\left[\frac{x_1^2}{\sum_{i=1}^{N} x_i^2}\right] \cdot E\left[\frac{y_1^2}{\sum_{i=1}^{N} y_i^2}\right] = N \cdot \frac{1}{N} \cdot \frac{1}{N} = \frac{1}{N}, \quad \text{(Equation 5)}$$

and thus the range of the distribution goes as $\frac{1}{\sqrt{N}}$.