1 **Robust enhancer-gene regulation identified by single-cell transcriptomes and**

2 **epigenomes**

3

4 Fangming Xie[1]*, Ethan J. Armand[2]*, Zizhen Yao[3], Hanqing Liu[4], Anna Bartlett[4], M. Margarita

5 Behrens[5], Yang Eric Li[6], Jacinta D. Lucero[5], Chongyuan Luo[7], Joseph R. Nery[4], Antonio Pinto-

6 Duarte[5], Olivier Poirion[6], Sebastian Preissl[6], Angeline C. Rivkin[4], Bosiljka Tasic[3], Hongkui

7 Zeng[3], Bing Ren[6], Joseph R. Ecker[4,8], Eran A. Mukamel[2,9]

8

9 *These authors contributed equally.

10

11 [1]Department of Physics, [2]Department of Cognitive Science, [6]Department of Cellular and

12 Molecular Medicine, University of California San Diego, La Jolla, CA 92037, USA

13 [3]Allen Institute for Brain Science, Seattle, WA, 98109, USA

14 [4]Genomic Analysis Laboratory, [5]Computational Neurobiology Laboratory, [8]Howard Hughes

15 Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA, 92037, USA

16 [7]Department of Human Genetics, University of California Los Angeles, Los Angeles, CA, 90095,

17 USA

18

19 [9]Correspondence: emukamel@ucsd.edu

20 **Abstract**

21 **Integrating single-cell transcriptomes and epigenomes across diverse cell types can link**

22 **genes with the *cis*-regulatory elements (CREs) that control expression. Gene co-**

23 **expression across cell types confounds simple correlation-based analysis and results in**

24 **high false prediction rates. We developed a procedure that controls for co-expression**

25 **between genes and integrates multiple molecular modalities, and used it to**

26 **identify >10,000 gene-CRE pairs that contribute to gene expression programs in different**

27 **cell types in the mouse brain.**

28

29 **Main text**

30 Single-cell epigenome sequencing techniques, including snATAC-seq and snmC-seq, can

31 identify cell-type-specific candidate *cis*-regulatory elements (cCREs), such as enhancers[1,2]. To

32 validate putative enhancers and elucidate their function, it is important to identify the genes they

33 directly regulate[3]. This can be accomplished by simultaneously perturbing enhancer activity and

34 measuring gene expression in the same cells[4,5]. However, perturbation experiments are

35 complex and to date have been used to screen pre-selected enhancers in cell types that could

36 be cultured *in vitro*[4,5]. By contrast, single-cell transcriptomes and epigenomes from complex

37 tissues, such as the brain, contain distinct genome-wide profiles from several hundred cell

38 types[6,7]. Correlating enhancer epigenetic profiles with transcription across cell types can identify

39 potential cell-type-specific enhancer-gene links[1,2,8]. However, genes with related functions often

40 have correlated expression patterns, leading to incidental associations that could confound co-

41 expression analyses with false-positives that do not reflect genuine enhancer-target gene

42 interactions[1,2,8–10].

43　　　　To separate spurious from genuine associations, *trans* enhancer-gene correlations can

44　be used as a negative control[11–15]. However, a principled analysis and validation of the most

45　appropriate null model has not been performed. Moreover, different epigenetic assays, such as

46　snATAC-seq and snmC-seq, measure distinct aspects of enhancer activity. It is unclear how the

47　differences between these data modalities affect the sensitivity and specificity for detecting

48　enhancer-gene correlations. Furthermore, correlation results may be strongly influenced by

49　clustering analysis of single cell data, which in turn depends on multiple unconstrained

50　parameters and algorithmic choices.

51　　　　To address these gaps, we identify high-confidence, robust enhancer-gene links using a

52　non-parametric permutation-based procedure to control for gene co-expression (Fig. 1a,

53　Supplementary Fig. 1a). We first integrate single-cell transcriptomes (scRNA-seq) and

54　epigenomes (open chromatin, snATAC-seq, and DNA methylation, snmC-Seq) to generate

55　multi-modality profiles using a dataset with over 200,000 single cells from the mouse primary

56　motor cortex[6]. We correlate the epigenetic state of putative enhancers with expression of nearby

57　genes, and compare the observed correlation with two null distributions. A conventional

58　shuffling procedure that randomly permutes cell labels effectively controls for noise present in

59　single-cell sequencing measurements[1,2,10]. However, as we discuss below, this null distribution

60　is confounded by gene co-expression and leads to spurious enhancer-gene associations. This

61　challenge can be addressed statistically using generalized least squares regression[16] (GLS),

62　which transforms data matrices to decorrelate observations. We used a more general non-

63　parametric approach, shuffling genomic regions to create an appropriate null distribution[11–15].

64　Moreover, we leveraged three complementary data modalities to cross-validate enhancer-gene

65　links with independent data. Finally, we validated the predicted links with multimodal 3D

66　chromatin conformation (snm3C-seq) data[17].
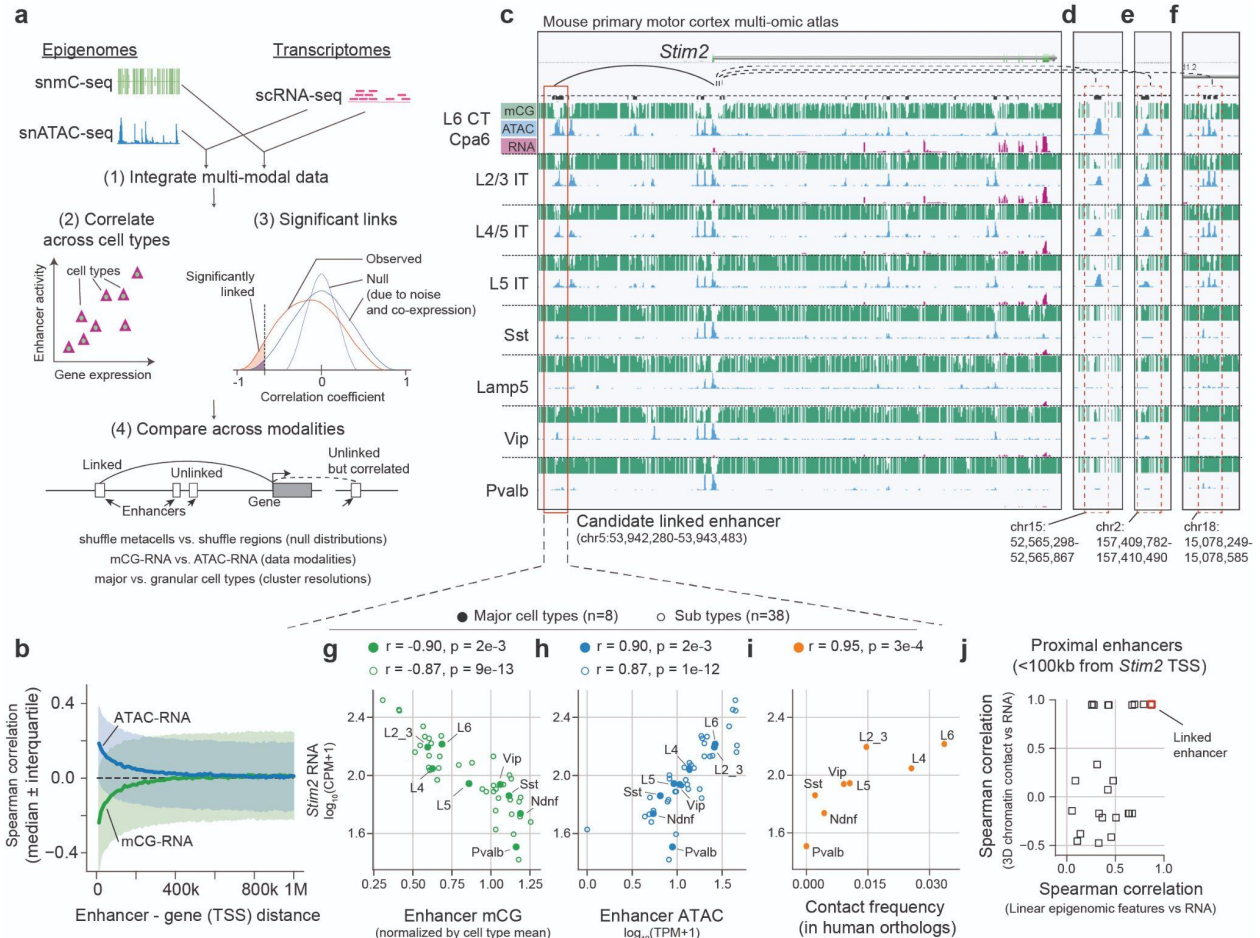
Fig. 1 | Identifying enhancer-gene links through integrated analysis of single-cell transcriptomes and epigenomes. a. Our proposed method links enhancers with target genes by (1) integrating single-cell transcriptomes (scRNA-seq) and epigenomes (snmC-seq and snATAC-seq), (2) correlating enhancer activity with gene expression across metacells, (3) identifying significant links compared with a shuffled null distribution, and (4) evaluating predicted links across null models, data modalities, and metacell resolutions. **b.** Strength of enhancer-gene association as a function of genomic distance. The wide interquartile range (shading) indicates high variability in enhancer-gene associations. **c-f.** Correlation of the gene *Stim2* with nearby (**c**) and distal (**d-f**) enhancer regions. **g-i.** Scatter plots of *Stim2* expression versus enhancer mCG (**g**), ATAC-seq signal (**h**), and enhancer-TSS chromatin contact frequency in human orthologs (**i**). **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 value between mCG-RNA and ATAC-RNA. Enhancer mCG level is normalized by the global mean mCG level of each cell type; RNA is $\log_{10}(\text{CPM}+1)$ normalized; ATAC is $\log_{10}(\text{TPM}+1)$ normalized.

To illustrate the risk of false associations due to gene co-expression, we analyzed a large set of single-cell transcriptome and epigenome data from the mouse primary motor cortex[6]. Putative enhancers (see Methods; Table S1, Supplementary Fig. 1b) within ~100 kb of

3

87   a gene promoter were enriched in associations with gene expression, including positive

88   correlations for chromatin accessibility and negative correlations for enhancer DNA methylation

89   (mCG) (Fig. 1b, Supplementary Fig. 1c,d). However, these associations were highly variable:

90   We observed many weak correlations for proximal enhancers (<100 kb), and relatively strong

91   correlations for some distal enhancers (>500kb) (Fig. 1b, interquartile range ~0.4). The broad

92   distribution of correlation strength makes it difficult to reliably link specific enhancers with their
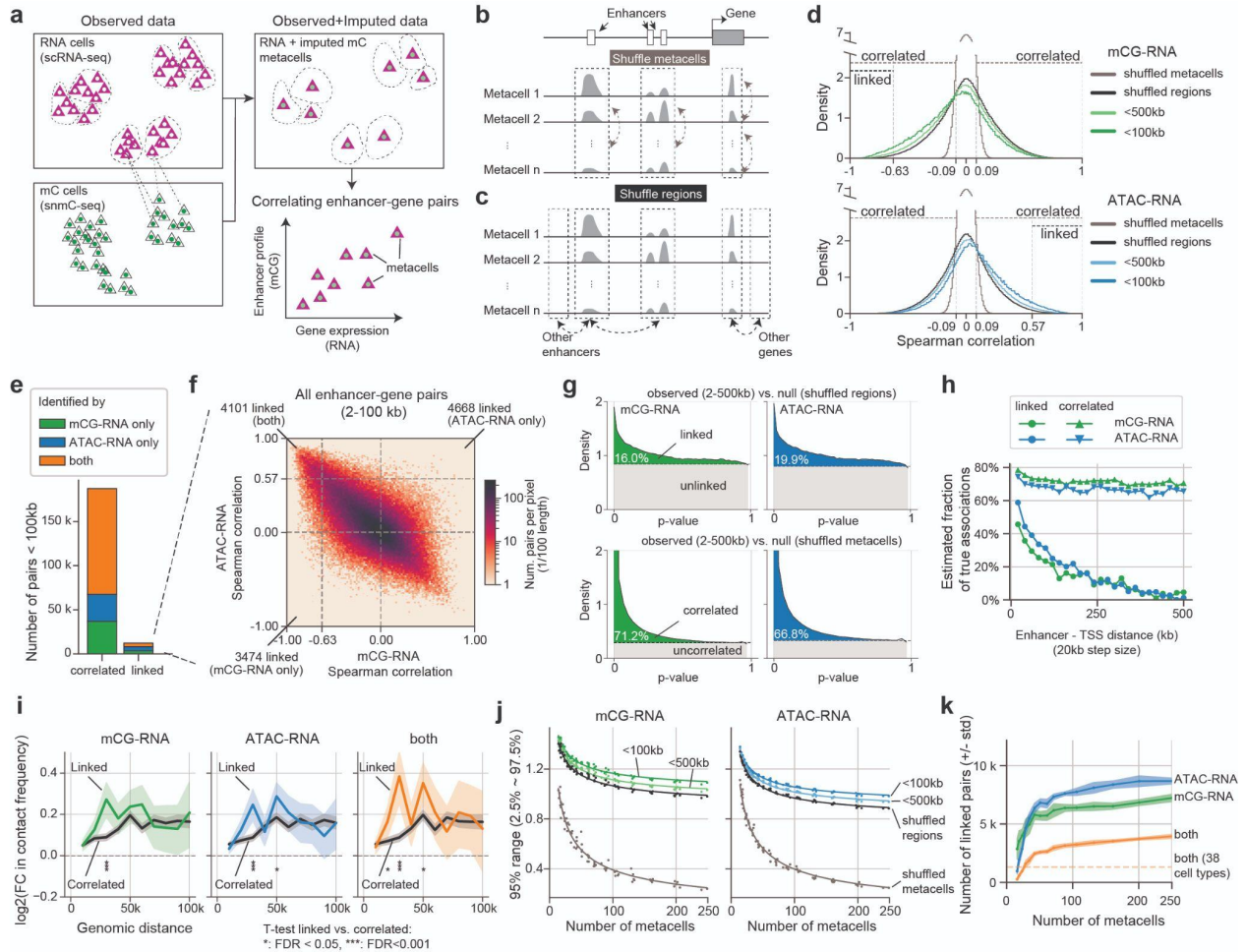
93   target genes.

94   A representative example is the gene *Stim2,* encoding a calcium sensor that helps

95   maintain basal Ca^{2+} levels in pyramidal neurons[18]. In cortical neurons, we identified 33

96   enhancers within 100 kb of the *Stim2* promoter. *Stim2* expression correlates with low mCG (r =

97   –0.87, p=9e-13, n=38 cell types) and high chromatin accessibility (r=0.87, p=1e-12) at a nearby

98   enhancer (Fig. 1c,g,h). By contrast, 15 other nearby enhancers have weaker, though still

99   significant (FDR<0.05), correlation with *Stim2* expression (|r|=0.46~0.85). Moreover, *Stim2*

100   expression also correlated significantly with 25,027 other enhancers located throughout the

101   genome (FDR<0.05; both mCG-RNA and ATAC-RNA), most of which (n = 23,526) were on

102   different chromosomes (Fig. 1d-f). Such numerous correlations with *trans*-enhancers likely

103   reflect gene co-expression, rather than direct causal links with the *Stim2* gene. For example,

104   these *trans*-enhancers might directly regulate nearby genes whose expression patterns across

105   cell types are similar to *Stim2* (Supplementary Fig. 1e-h,j,k).

106   Next, we used three-dimensional genome conformation data to test whether putative

107   enhancer-gene links correspond to bona fide physical interactions[19]. We analyzed the 3D

108   chromatin contact frequency of the predicted enhancer-gene pair (Fig. 1c) across homologous

109   human brain cell types, using multi-omic snm3C-seq data[17]. Chromatin contact frequency for

110   this enhancer was strongly correlated with *Stim2* expression (r=0.95, p=3e-4; Fig. 1i;

111   Supplementary Fig. 1i). By contrast, other proximal enhancers were less correlated (Fig. 1j).

112   In addition to the challenge of widespread spurious correlations, the case of *Stim2* also

113 illustrates the challenges associated with defining cell types[20]. For example, the same set of

114 cells can be grouped into either 8 major types or 38 fine-grained sub-types, leading to different

115 correlation values (Fig. 1g,h; Supplementary Fig. 1j,k).

116   To address these issues, we developed a procedure that controls the risk of false

117 positives from gene co-expression, and compares predicted links across data modalities and

118 cell type resolutions (Fig. 2a, Supplementary Fig. 2). We first integrate single-cell transcriptomes

119 (RNA) and epigenomes (DNA methylation or chromatin accessibility) using correlated gene-

120 level features across data modalities (SingleCellFusion)[6,21,22]. This allows us to build a neighbor

121 graph connecting cells within and across data modalities (see Methods). Next, we define

122 metacells[23], which aggregate the transcriptomic and epigenomic profiles from groups of similar

123 cells. Each metacell has a complete bi-modal (transcriptomic and epigenomic) profile, which

124 then allows us to correlate enhancer epigenetic features with gene expression. These metacells

125 represent cells with an adjustable resolution, capturing both discrete and continuous patterns of

126 variation.

Fig. 2 | Stringent statistical criteria capture enhancer-gene links with consistent signatures across data modalities and cell type resolutions. a. Method for linking enhancers to target genes using metacells with bi-modality profiles. **b-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 pairs <100kb). Two null models induce two different types of significance: linked (black bar; shuffle regions) and correlated (gray bar; 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 panels) or shuffled metacells (bottom panels). The estimated fraction of true positives is shown[24]. **h.** Estimated fraction of true associations vs. 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). Tracks are aggregated across all contacts from 8 neuronal cell types. **j.** The 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.)

We reasoned that genuine enhancer-gene interactions should correspond to stronger correlations than the background induced by co-expression. Correlations mediated by co-

6

148   expression are inherently limited in their strength by the magnitude of gene-gene correlations,

149   whereas direct enhancer-gene interactions can produce stronger associations. Importantly, this

150   assumption applies to the strongest enhancer-gene interactions; weak interactions that don't

151   exceed the background of gene co-expression cannot be detected by correlation-based

152   methods.

153   To test whether the observed correlations exceed what is expected due to noise and

154   gene co-expression, we compared the observed correlation coefficients with two null

155   distributions: shuffling metacells[1,2,10] and shuffling regions[11–13] (Fig. 2b-d). Shuffling metacells

156   decouples epigenetic and transcriptomic signatures across metacells, removing both enhancer-

157   gene correlation and gene co-expression (Fig. 2b). The significance arising from this distribution

158   is inflated by gene co-expression, potentially leading to false positives in which an enhancer-

159   gene pair may be correlated due to shared upstream regulation rather than direct interaction.

160   Shuffling regions retains the gene co-expression structure imposed by the hierarchical

161   organization of cell types, but it correlates each gene's expression with distant, randomly

162   selected enhancers (Fig. 2c)[11–13].

163   As expected, the distribution obtained by shuffling regions was wider than that derived

164   from shuffling metacells (Fig. 2d), reflecting incidental correlations due to gene co-expression.

165   Enhancer-gene pairs within 500kb of the TSS are significantly enriched in both positive and

166   negative correlations when compared with shuffling metacells. However, when compared with

167   shuffling regions, enrichment is only present in positive correlation for ATAC-RNA, and in

168   negative correlation for mC-RNA. Thus, shuffling regions is a more stringent null distribution for

169   calling significant enhancer-gene links, as it effectively controls for spurious enhancer-gene

170   correlations due to gene co-expression.

171   We call an enhancer-gene pair significantly "correlated" if it passes an FDR-adjusted

172   threshold based on shuffling metacells, whereas we reserve the term significantly "linked" for

173   pairs that pass the criteria set by shuffling regions. We used a relatively lenient FDR threshold

174   of 0.2 to reduce the risk of false negatives from our stringent null distribution. Linked pairs

175   (n=12,243 within 100kb, FDR<0.2) are a subset of correlated pairs (187,343 within 100kb,

176   FDR<0.2) (Fig. 2e,f), but they have a stronger association that rises above the background from

177   gene co-expression. Lowering the FDR threshold to 0.1 or 0.05 reduced the number of linked

178   pairs to 3,142 and 489, respectively.

179   Notably, we found that removing sample covariance using GLS abolished the difference

180   between shuffling regions and shuffling cells (Supplementary Fig. 3a-b). This manipulation thus

181   removes the distinction between correlated pairs and linked pairs (Supplementary Fig. 3c). In

182   addition, the shuffling-regions null distribution was robust with respect to differences in enhancer

183   GC content and an enhancer's distance to its nearest gene (Supplementary Fig. 4a-d).

184   We compared our results with two alternative strategies for estimating enhancer-gene

185   interactions using single-cell epigenomes. Using open chromatin data, CICERO[8] identified

186   1,869 significant enhancer-gene associations located within 100kb. These significantly overlap

187   with a subset of the correlated pairs we identified, and to a lesser degree with linked pairs

188   (Supplementary Fig. 5a,b). Notably, the mean CICERO co-accessibility scores are 4.8-5.9 fold

189   higher (p<2e-8) for linked pairs than for correlated pairs (Supplementary Fig. 5c). A second

190   strategy, the activity-by-contact (ABC) model[5], identified enhancer-gene links using both

191   chromatin accessibility and chromatin conformation data. This model identified enhancer-gene

192   links for each cell type independently, without considering correlated variability in expression

193   across cells. The ABC model identified 150,228 associations within 100kb, which significantly

194   overlap with our correlated and linked pairs (Supplementary Fig. 5d,e). In addition, the ABC

195   scores are 1.09-1.22 fold higher (p<1e-8) for linked pairs than for correlated pairs

196   (Supplementary Fig. 5f). These results show that linked pairs have stronger associations than

197   correlated pairs, and are more likely to capture genuine enhancer-gene associations.

198   A potential pitfall of our stringent enhancer-gene linking procedure is a higher risk of

199   false-negatives, i.e. failure to detect true interactions. We next empirically compared correlated

200   versus linked pairs from several biological and statistical perspectives, to test whether the

201   correlations filtered out by our method are likely false positives arising from gene co-expression.

202   First, we observed that correlated pairs include many enhancer-gene links with a non-

203   canonical direction of association (Fig. 2d; Supplementary Fig. 6a). For example, we found

204   about a third (47,137/150,285) of these pairs had a negative correlation of gene expression with

205   chromatin accessibility, and a similar proportion (53,687/156,932) had a positive correlation with

206   mCG. Non-canonical associations were also reported in recent large-scale studies of brain cell

207   epigenomes[1,2]. These correlations could suggest novel biological mechanisms such as

208   methylcytosine-preferring transcription factors[25]. However, they may also include false-positive

209   associations due to gene co-expression. Indeed, none of the non-canonical associations passed

210   our threshold for linked pairs (Fig. 2d). This is consistent with the canonical understanding of

211   enhancer activity associating with low DNA methylation and open chromatin.

212   Second, as enhancer-gene interactions are mostly concentrated within ~100-500 kb

213   around gene promoters[4,5], we compared the distance dependence of linked and correlated

214   pairs. Using a p-value histogram method[24], we estimated 16.0-19.9% of enhancers that are 2-

215   500kb away from a promoter are linked (Fig. 2g, Supplementary Fig. 6b). A much larger fraction

216   (66.8-71.2%) were correlated. Notably, the proportion of correlated pairs remains high even for

217   distal pairs (e.g. >60% for pairs >1 Mb or on other chromosomes), whereas <5% of these pairs

218   are linked (Fig. 2h, Supplementary Fig. 6c). These correlated pairs contradict the biological

219   understanding that most enhancers activate genes in *cis*; the linked pairs are more coherent

220   with this canonical framework.

221   Third, we validated our predicted links with independent chromatin conformation data

222   from the human brain[17]. We reasoned that linked enhancer-gene pairs which are conserved

223   across species should have higher chromatin contact frequency compared with random regions.

224   Indeed, we found enrichment of contact frequency for both linked (mean fold change (FC) =

225   1.15, p=2e-4) and correlated pairs (mean FC = 1.10, p=1e-5). Moreover, linked pairs located

226   10-30 kb apart have higher levels of contact enrichment than correlated pairs (FDR<0.05; Fig.

227   2i, Supplementary Fig. 6d,e).

228   A key parameter for our analysis is the cell type granularity, as determined by the

229   number of metacells. The sparse genomic coverage of single-cell sequencing and the limited

230   number of profiled cells create a tradeoff between the number of metacells and the quality of

231   each metacell--i.e. between fine-grained resolution and signal/noise ratio. As the number of

232   metacells (N) increases, the width of the null distribution for the shuffled metacells approaches

233   zero as $\frac{1}{\sqrt{N}}$, which is consistent with independent random signals for each metacell (see

234   Methods; Fig. 2j; Supplementary Fig. 7a-c). By contrast, the range of the null distribution for

235   shuffled regions does not vanish for large N, but instead asymptotes at a non-zero value that

236   reflects gene co-expression (Supplementary Fig. 7c). Notably, the shuffling-regions null

237   distribution is less sensitive to the number of metacells, and more closely reflects the behavior

238   of the observed correlations. This suggests enhancer-gene link calling using shuffling-regions is

239   less sensitive to the choice of cell type granularity than using shuffling-metacells. We found

240   more linked pairs as the number of metacells increases, but with diminishing returns after $N >$

241   50. (Fig. 2k; Supplementary Fig. 7d).

242
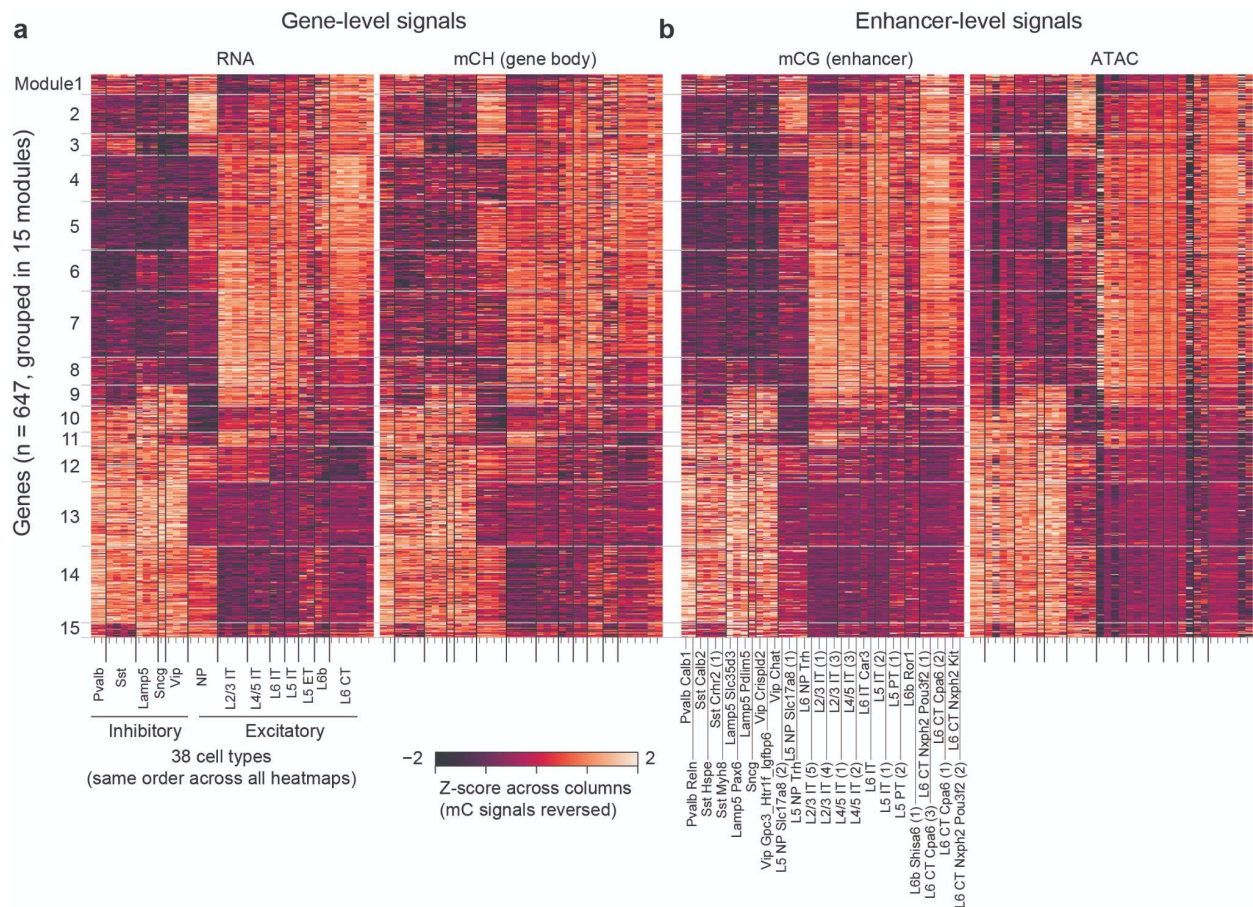
Fig. 3 | Consistent gene- and enhancer-level signatures for hundreds of enhancer-gene links. a-b.
Gene expression (**a**), gene body DNA methylation (**b**), and enhancer mCG (**c**) and ATAC signal (**d**)
across cell types. Genes are organized into 15 modules by K-means clustering. Enhancers are ordered
according to the genes they are linked to (FDR < 0.2 for both mCG-RNA and ATAC-RNA across n=38 cell
types). Signals from multiple enhancers linked to the same gene were averaged. The colormap for the
mC modalities (gene body mCH and enhancer mCG) are reversed.

250   We used our procedure to comprehensively examine regulatory interactions in neurons

251   of the mouse primary motor cortex[6]. Linked enhancer-gene pairs formed 15 modules that

252   capture diverse cell-type-specific signatures (Fig. 3a,b). For example, genes in module 13 are

253   specifically expressed in pan-inhibitory neurons, with corresponding low CG methylation level

254   and open chromatin at linked enhancers. Module 9 is most active in caudal ganglionic eminence

255   (CGE) derived inhibitory neurons (Lamp5, Sncg, and Vip) and in superficial-layer excitatory

11

256   neurons (L2/3 IT and L4/5 IT). These consistent gene- and enhancer-level signals integrated

257   from three data modalities provide strong support for our identified enhancer-gene associations.

258   Our analyses highlight the challenge of distinguishing genuine enhancer-gene

259   interactions from spurious correlations due to gene co-expression. We addressed this by

260   empirically estimating the expected correlations for unlinked enhancer-gene pairs under co-

261   expression, and comparing results across different epigenetic assays and cell type granularities.

262   Notably, mCG-RNA and ATAC-RNA associations show striking similarities (Fig. 2d,f-k; Fig. 3),

263   despite measuring distinct epigenetic features with opposite effects on gene expression.

264   Predicted enhancer-gene links are robust with respect to a wide range of cell type granularities

265   (Fig. 2k). We identified hundreds of genes and thousands of linked cCREs with highly

266   coordinated gene- and enhancer-level activities (Fig. 3a,b).

267   Correlation-based analysis has notable limitations. First, this approach cannot identify

268   constitutive enhancer-gene links that are present in all cell types. Larger datasets including

269   more diverse tissues or cell types may partly address this limitation. Second, rigorous control for

270   spurious correlations limits the power of detecting genuine but weak enhancer-gene

271   interactions. Finally, true causal interactions cannot be inferred from correlational analysis

272   alone. The links we identified (Fig. 3a,b) are strong candidates for causal enhancer-gene

273   interactions, which must be tested by perturbative experiments[26,27]. Future experimental

274   validation, including large-scale assays[4,5,28], will be needed to test correlation-based predictions.

275   By bringing together multiple data modalities to define robust enhancer-gene links, these

276   analyses can reveal the regulatory principles of cell-type-specific gene expression.

277

12

Supplementary Figures
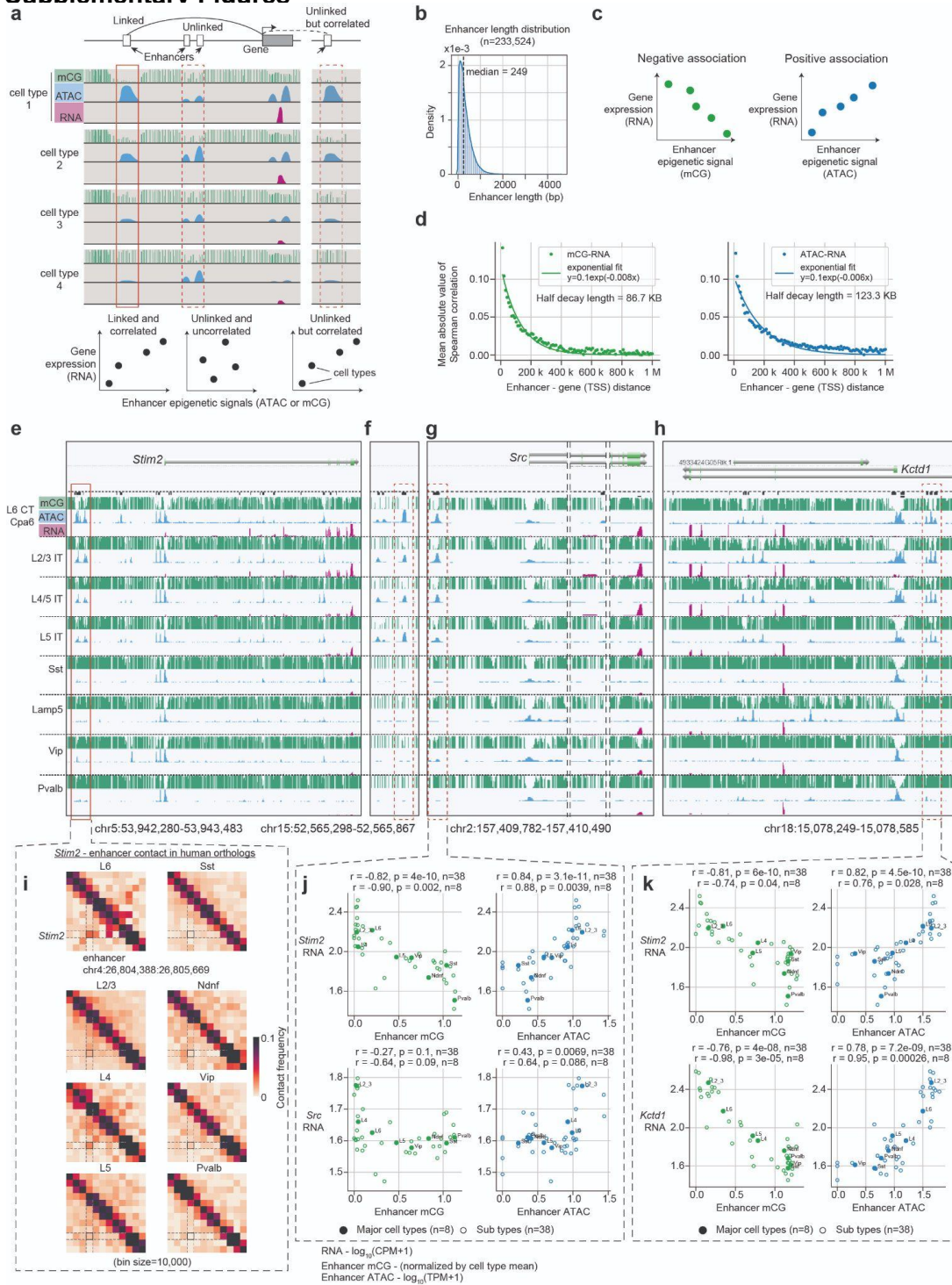
Supplementary Figure 1. Examples of enhancer-gene links (Related to Fig. 1). a. Approach for linking enhancers to target gene(s) by correlating enhancer activities and gene expression across cell types. Statistically significant correlation alone may not distinguish genuine vs.

13

283   spurious links. **b.** Distribution of putative enhancer length (list adapted from Ref[6]; see Methods).

284   **c.** Illustration of two modes of enhancer-gene associations: enhancer mCG typically have

285   positive correlation with gene expression, while enhancer ATAC-seq signals typically have

286   negative correlation. **d.** Median Spearman correlation as a function of enhancer-TSS distance.

287   Both mCG-RNA and ATAC-RNA decay exponentially, with a half decay length of 86.7 kb and

288   123.3 kb, respectively (Related to Fig. 1b). **e-h.** Genome browser views across cell types and

289   data modalities near the gene *Stim2* (**e**), as well as other regions (**f-h**) with strongly correlated

290   enhancer signals. Note that the highlighted enhancers in (**g-h**) are also correlated with the

291   expression of their nearby genes (*Src* and *Kctd1*) (Related to Fig. 1c-f). **i.** Heatmaps of

292   chromatin contact frequency in human brain cells near *Stim2* and the human ortholog of the

293   highlighted enhancer across 8 human neuronal cell types. **j-k.** Scatter plot of *Stim2* expression

294   (upper row) / local genes expression (lower row) versus the highlighted enhancers. Enhancer

295   mCG level is normalized by the global mean mCG level of each cell type; RNA is log10(CPM+1)

296   normalized; ATAC is log10(TPM+1) normalized.

297
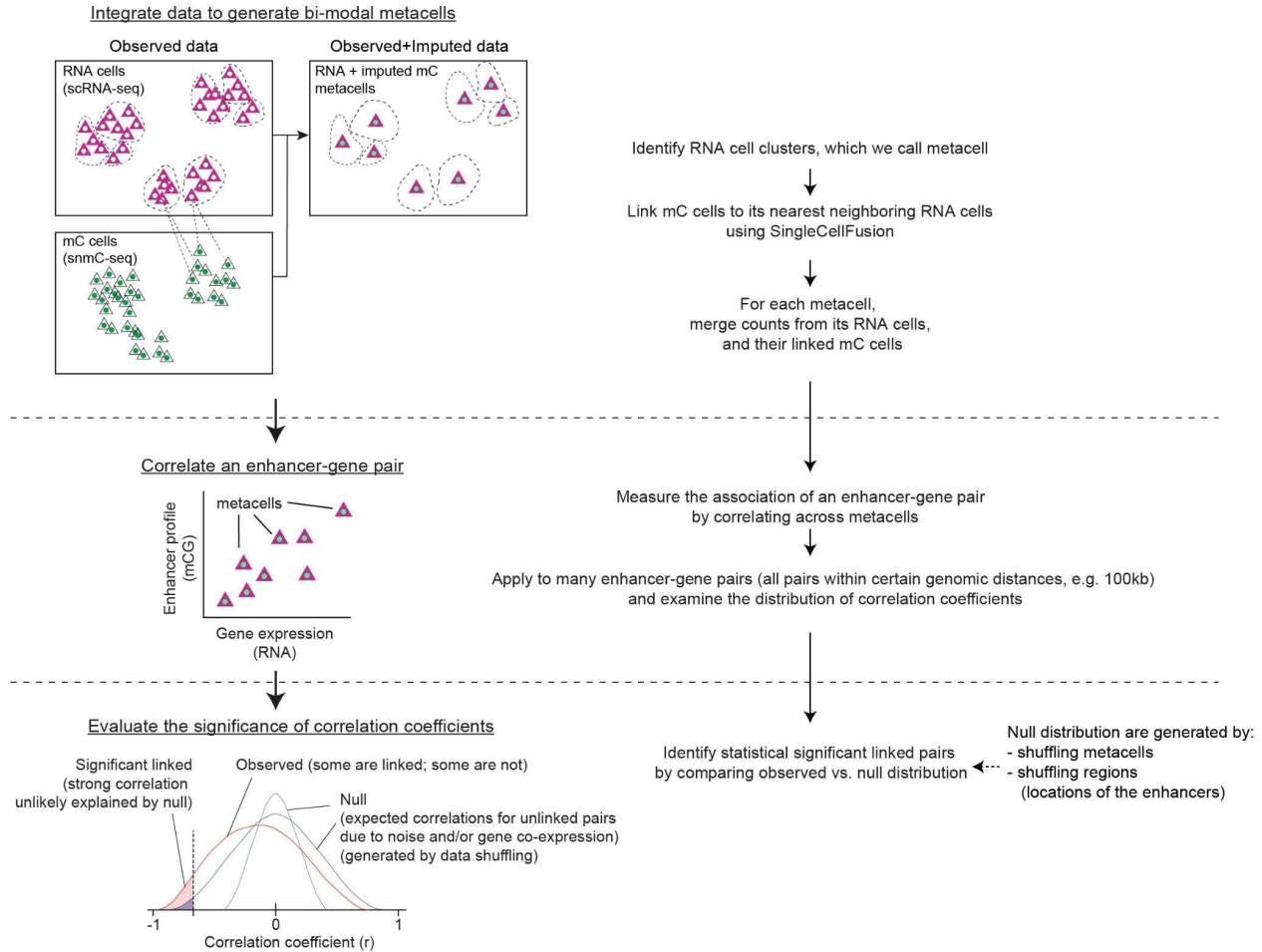
Supplementary Figure 2. Method overview. The analysis involves three main steps. 1. Integrate transcriptomic and epig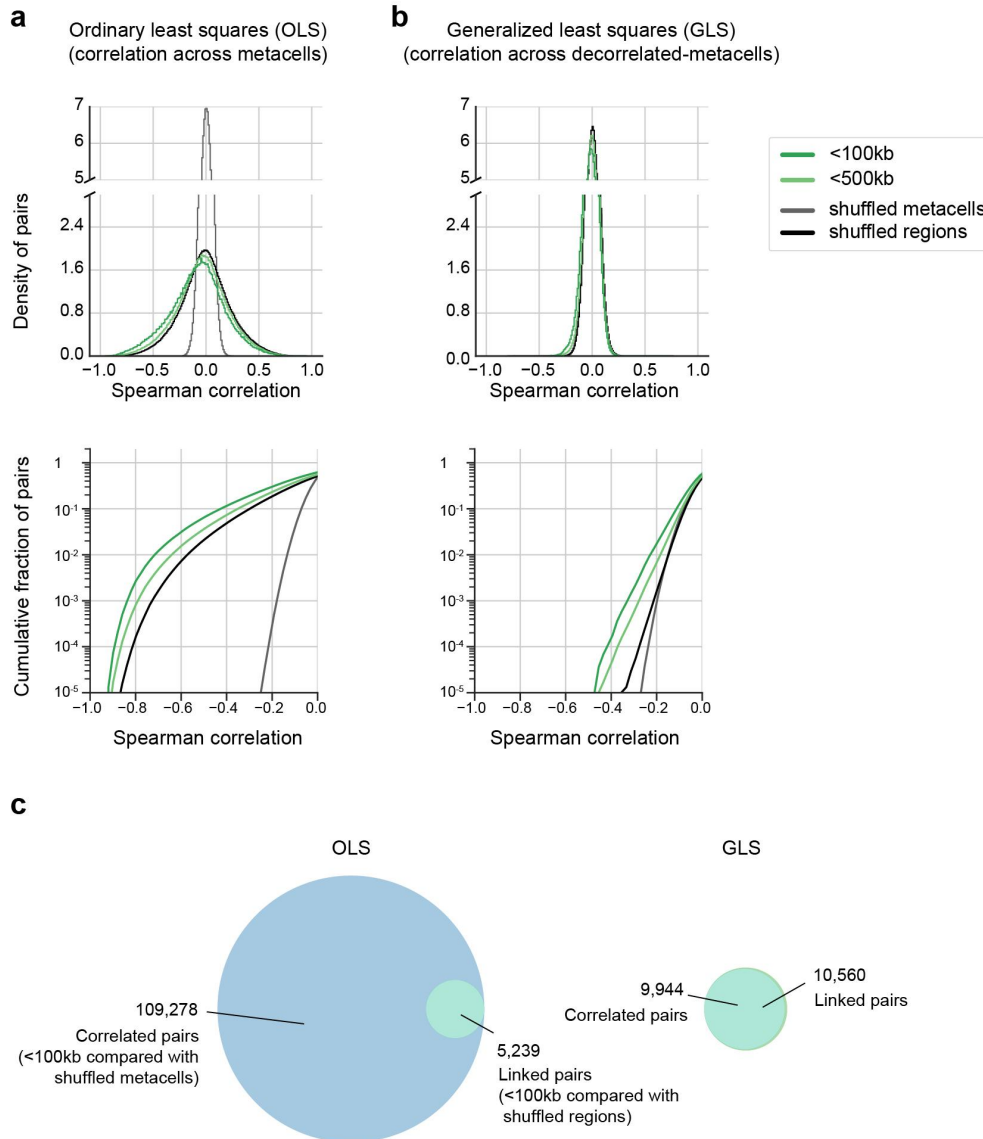enomic data to generate metacells with bi-modal profiles. 2. Correlate enhancer-gene pairs to get correlation coefficients for individual enhancer-gene pairs. 3. Evaluate the statistical significance of correlations by comparing the observed correlations with null distributions generated by data shuffling.

a Ordinary least squares (OLS)
(correlation across metacells)

b Generalized least squares (GLS)
(correlation across decorrelated-metacells)

c

OLS

GLS

109,278
Correlated pairs
(<100kb compared with
shuffled metacells)

5,239
Linked pairs
(<100kb compared with
shuffled regions)

9,944
Correlated pairs

10,560
Linked pairs

Supplementary Figure 3. Generalized Least Squares (GLS) transformation abolishes the difference between shuffling metacells and shuffling regions (Related to Fig. 2). a-b. Density distribution (top) and cumulative distribution (bottom) of enhancer-gene correlations across metacells using OLS (**a**), and across decorrelated metacells using GLS (**b**). GLS transformation decouples the covariance across metacells, making the shuffling-regions distribution similar to the shuffling-metacell distribution (see Methods). **c**. Venn diagram showing the degree of overlap between correlated pairs and linked pairs, using OLS (left) and GLS (right) approaches. GLS abolishes the difference between correlated and linked pairs.

16
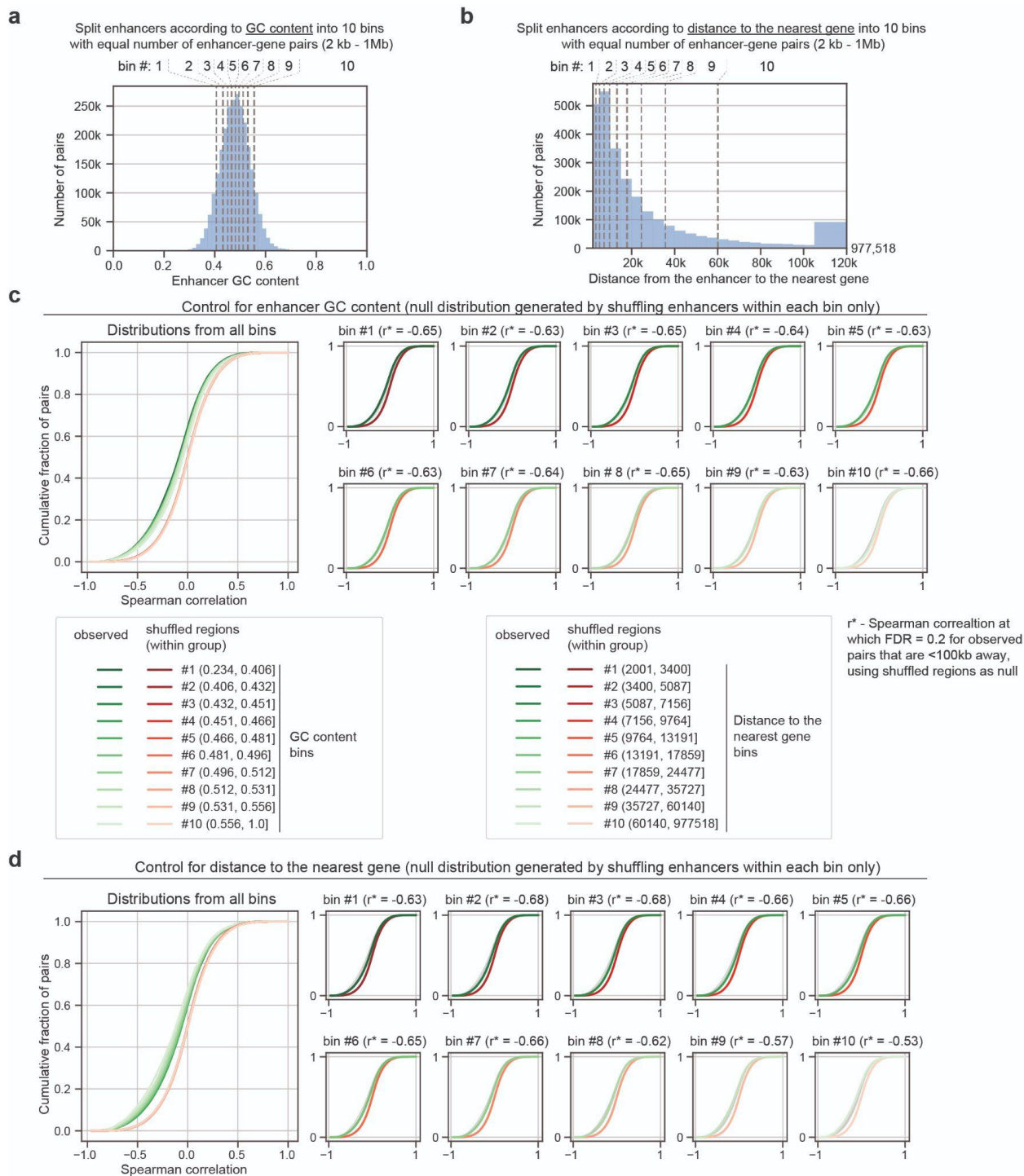
Supplementary Figure 4. The shuffling-regions null distribution is robust with respect to enhancer GC content and distance to the nearest gene (Related to Fig. 2). a-b. Distribution of GC content (**a**) and distance to the nearest gene (**b**) for enhancers that are in all enhancer-gene pairs (2kb - 1Mb). In each case, they are grouped into 10 bins (deciles) with an equal number of enhancer-gene pairs. **c-d.** Cumulative distribution of enhancer-gene correlation (mCG-RNA; observed (<100kb) vs. null (shuffling regions)). The same analyses are applied to

17

324   each of the 10 bins by enhancer GC content (**c**) and by distance to the nearest gene (**d**),

325   respectively. Null distributions from different bins highly overlap.
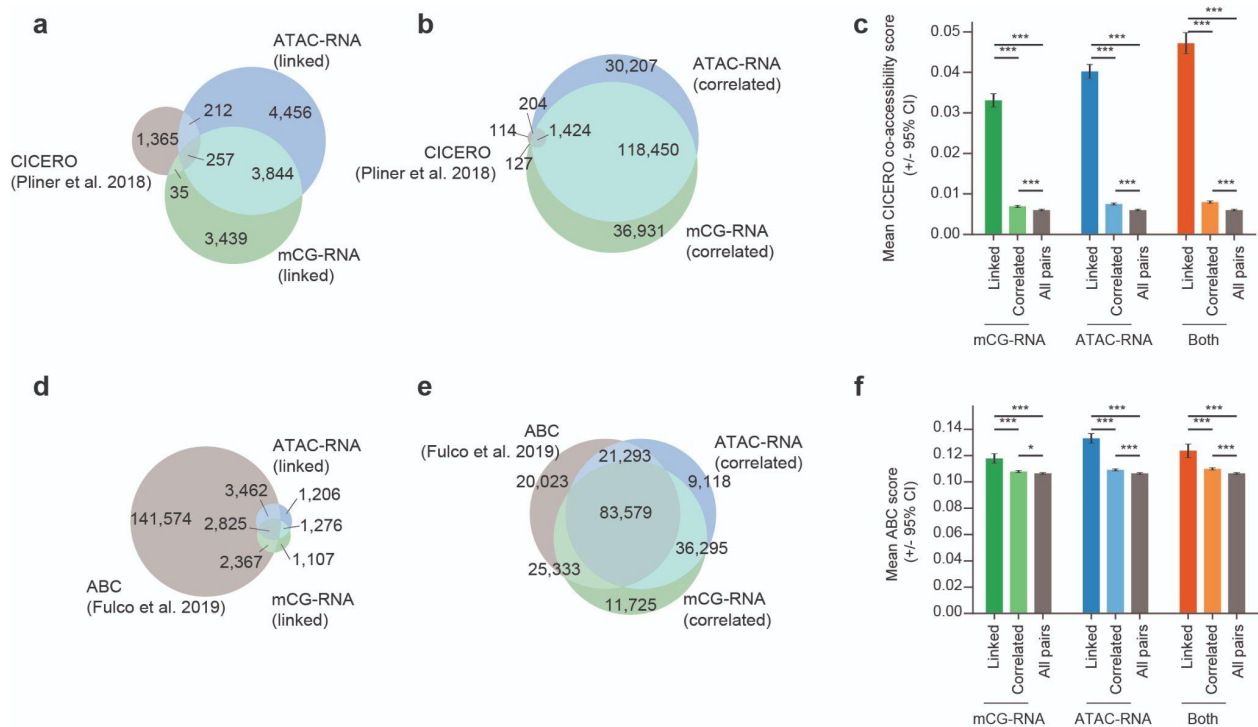
326
Supplementary Figure 5. Comparison with CICERO[9] and the activity-by-contact (ABC)
model[5] (Related to Fig. 2). a-b. Venn diagram comparing the enhancer-gene associations
identified by applying CICERO[8] to the mouse MOp data[6] versus linked pairs (**a**) and correlated
pairs (**b**) found in this study. **c.** Barplots comparing the mean CICERO scores across different
groups of enhancer-gene pairs identified in this study. Error bars indicate 95% confidence
intervals. Independent t-test are used to compare between groups (* p<0.05, *** p<0.001). **d-e.**
Venn diagram comparing the enhancer-gene associations identified by applying ABC model[5] to
the mouse MOp data[6] versus linked pairs (**d**) and correlated pairs (**e**) found in this study. **f.**
Barplots comparing the mean ABC scores across different groups of enhancer-gene pairs
identified by this study. ABC scores are generated for each enhancer-gene pair and cell type
(n=38). We first took the maximum across cell types, followed by taking the mean of each group
of enhancer-gene pairs. Error bars indicate 95% confidence intervals. Independent t-test are
used to compare between groups (* p<0.05, *** p<0.001).

340

341
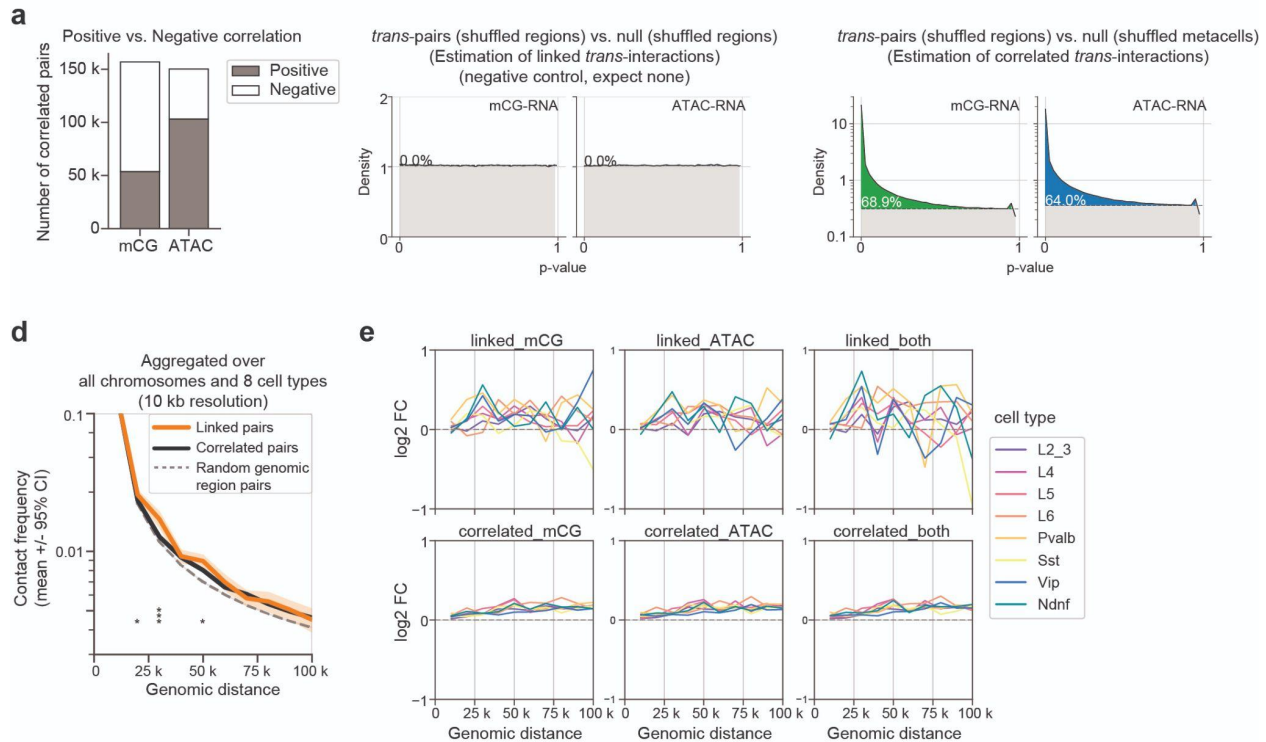
Supplementary Figure 6. Linked vs. correlated enhancer-gene pairs have distinct characteristics (Related to Fig. 2). a. The number of positively or negatively correlated enhancer-gene pairs for mCG-RNA and ATAC-RNA, respectively. **b.** P-value histograms[24] of *trans*-enhancer-gene pairs using shuffling-regions as the null distribution. The histogram closely follows a uniform distribution, indicating trans-enhancer-gene pairs are linked. This serves as a negative control for Fig. 2g. **c.** P-value histograms[24] of *trans*-enhancer-gene pairs, using shuffling-metacells as the null distribution. The numbers mark the fraction of p-values that deviate from the uniform distribution, which estimates the fraction of correlated *trans*-enhancer-gene pairs. **d.** Chromatin contact frequencies of pairs of genomic bins as a function of genomic distance. Linked and correlated pairs (lifted over from mm10 to hg38[29]) are compared with random genomic pairs. Results are aggregated over all chromosomes (autosomes + chrX) and 8 different human neuronal cell types (L2/3, L4, L5, L6, Pvalb, Sst, Vip, Ndnf) at 10kb resolution of chromatin contact maps[17]. **e.** Enrichment of contact frequency of linked and correlated enhancer-gene pairs compared with random genomic region pairs across 8 human neuronal cell types.

357

358
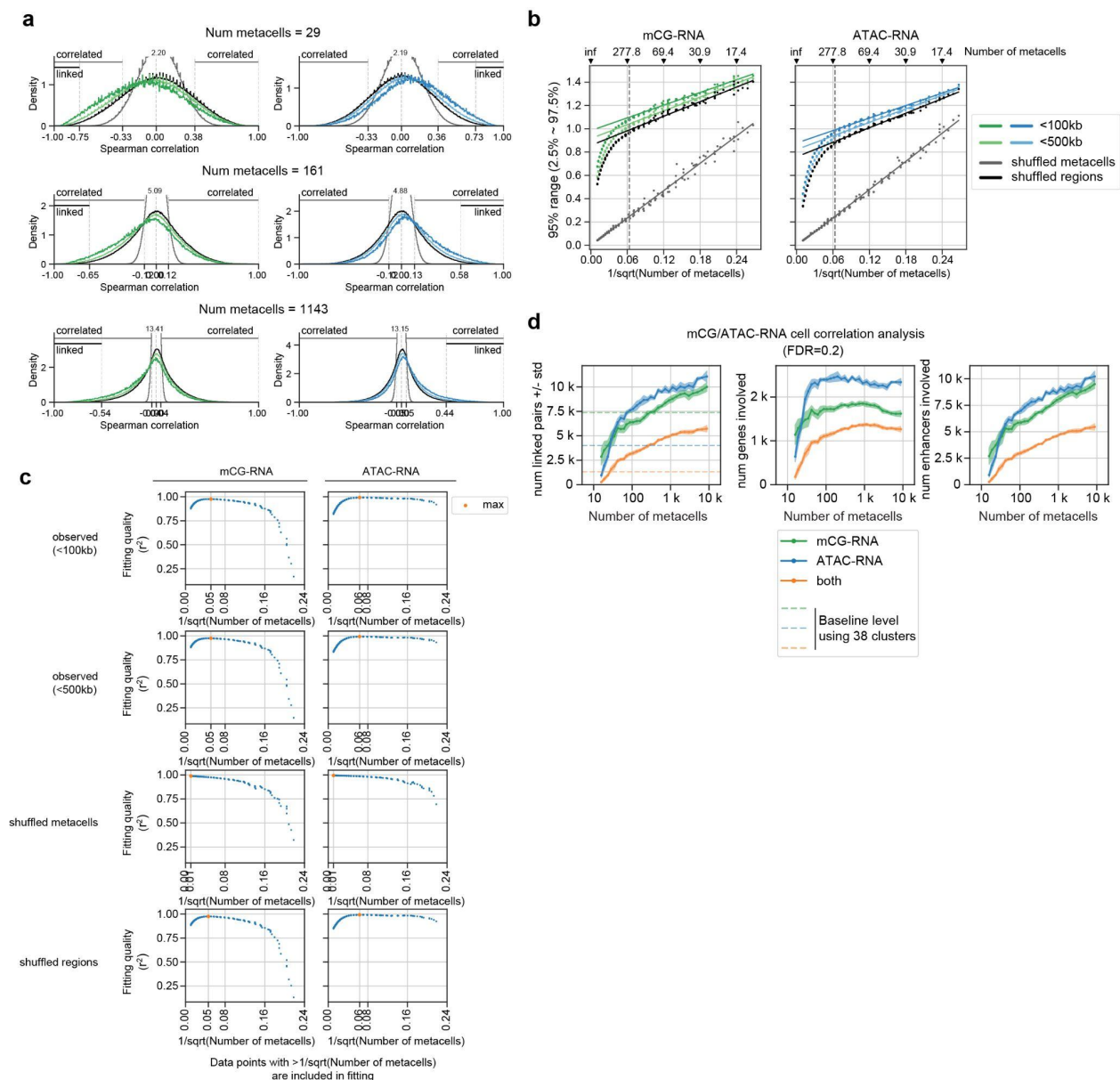
Supplementary Figure 7. Effect of the granularity of metacells on enhancer-gene correlations (Related to Fig. 2). a. Distributions of correlation coefficients for different numbers of metacells. The distributions become narrower as the number of metacells increases. Here the number of metacells controls cell type granularity. **b.** Range of correlation coefficients (2.5%~97.5% range) as a function of $1/\sqrt{N}$, where N is the number of metacells. Data points are well fitted by a straight line for $N < 277$. **c.** Fitting quality, as measured by r^2, as a function of fitting cutoff--range of data points in (**b**) used for fitting. The fitting quality peaks at $1/\sqrt{N} = 0.05$, i.e., $N = 278$. **d.** The number of linked pairs (left), number of genes involved (middle), and number of enhancers involved (right) as a function of the number of metacells.

21

368 **Methods**

369 **Datasets** We used three single-cell sequencing datasets from the mouse primary motor cortex

370 (MOp)[6]. They are scRNA-seq (single cell; 10x genomics V3; Allen Institute for Brain Science),

371 snmC-seq (single nucleus; DNA methylation; Ecker lab from the Salk Institute), and snATAC-

372 seq (single nucleus; chromatin accessibility; Ren lab from UCSD). Only high-quality neuronal

373 cells, as determined in Ref[6] (from its Supplementary Table 2; column SCF/SingleCellFusion),

374 are retained for our analysis. These datasets are publicly available and provided by a previous

375 study (Ref[6]; https://assets.nemoarchive.org/dat-ch1nqb7). The starting point of all analyses are

376 gene-by-cell matrices and/or enhancer-by-cell matrices depending on the data modality. For the

377 scRNA-seq dataset, we start from the gene-by-cell count matrix. For the snATAC-seq dataset,

378 we quantified both enhancer-by-cell and gene-by-cell count matrices. For the snmC-seq

379 dataset, we quantified enhancer-by-cell CG DNA methylation profiles and gene-by-cell non-CG

380 (CH) DNA methylation profiles. The DNA methylation profile for a particular region and cell can

381 be summarized by two numbers: the number of methylated cytosines (mC) and the total number

382 of cytosines covered (C). The DNA methylation level is the ratio of mC to C (mC/C). Please see

383 sections below for dataset specific procedures of normalizations. The mouse gene annotation

384 file is downloaded from gencode (vM16). The enhancer list is adapted from the putative

385 enhancer list from Ref[6] (see below).

386

387 **Calling putative enhancers** We constructed our putative enhancer list based on the mouse

388 MOp neuronal cell type-specific putative enhancers from Ref[6] (from its Supplementary Table 7).

389 In that study, the enhancers are called using REPTILE[30], an algorithm that uses the DNA

390 methylation and ATAC-seq profiles of 13 mouse neuronal cell types, as well as mouse

391 embryonic stem cells, as input. Starting from this list, we first selected regions with enhancer

392 score >0.5 and merged overlapping regions using bedtools[31]. We subsequently removed

393   regions overlapping any gene promoter regions (transcription start site +/- 2kb; all transcripts

394   from gencode vM16), exons (vM16), and ENCODE blacklist[32]. This leaves us with 233,524

395   enhancers in total, with a median size of ~250 bp (Supplementary Fig. 1b; Table S1).

396

397   **Curated cell types** For analyses related to Figure 1, we curated a list of 38 neuronal cell

398   clusters based on the SingleCellFusion clusters (L1 and L2, with n=29 to 56 cell types

399   respectively) in Ref[6]. We aimed to merge small clusters to increase pseudo bulk coverage at

400   enhancers, while retaining as much cell type diversity as possible. To achieve this, we first call

401   an enhancer *covered* in a cluster if it has at least 20 sequenced CpG sites in that cluster, where

402   the cluster-level coverage is the sum of cell-level coverages. Next, we call an enhancer

403   *common*, if it is covered in more than half of the L2 clusters. We call a cluster *covered*, if more

404   than half of the common enhancers are covered in that cluster. For each L1 cluster we then

405   evaluate 3 cases:

406   1. If the cluster itself is not covered, we drop it along with all its child (L2) clusters.

407   2. Else if less than 2 (n<2) of its child (L2) clusters are covered, we retain the L1 cluster

408   itself, but drop all its child (L2) clusters.

409   3. Else if at least 2 (n≥2) of its child (L2) clusters are covered, we retain the covered L2

410   clusters, but drop the uncovered L2 clusters and the L1 cluster.

411   This procedure resulted in 38 clusters with adequate coverage. Table S4 summarized the

412   correspondence between the 38 clusters we get from this procedure and the cell types defined

413   in Ref[6].

414   To compare with the cell types in snm3C-seq data[17], we further merged these 38 fine-

415   grained clusters into 8 major clusters based on the well-established neuronal cell type

416   taxonomy[33]. Table S4 summarized the correspondence between the 38 fine grained and the 8

417   major cell clusters defined in this study and those defined in Ref[6,17].

23

418

419 **Clustering and defining metacells.** For analyses related to Figure 2, we generated cell

420 clusterings with a range of cluster resolutions. We start by normalizing the scRNA-seq count

421 matrix with $\log_{10}$(CPM+1), where CPM stands for counts per million mapped reads. We then

422 calculated the top 50 principal components (PCs), and built a k-nearest neighbor graph (k = 30)

423 connecting cells according to the Euclidean distance in the PC space. We used Leiden

424 community detection to generate clusters[34]. Different resolution parameters (r = 1 ~ 794) were

425 chosen to generate clusters with different granularity (n = 13 ~ 8850 metacells). The pseudo

426 bulk profiles from each of the individual clusters were used as metacells.

427

428 **Feature selection and normalization.** We preprocessed the data matrices separately for each

429 data modality. The starting point is always cell-level matrices containing counts (RNA and

430 ATAC) or methylation level (mC). To get cluster-level (metacell) matrices, we summed counts

431 from cells in the same clusters (metacells) to create pseudo-bulk samples. For methylation

432 data, we summed methylated counts and total counts (coverage) separately. Next, we

433 normalized matrices as follows:

434 - For an RNA matrix (gene-by-cluster/metacell), we normalize the raw count matrix with

435 $\log_{10}$(CPM+1).

436 - For an ATAC matrix (enhancer-by-cluster/metacell), we normalize the raw count matrix

437 with $\log_{10}$(TPM+1), where TPM stands for transcripts per million mapped reads.

438 Enhancers that are covered in <50% of clusters are removed.

439 - For a gene body mCH matrix (gene-by-cluster/metacell), we first removed low coverage

440 genes if the gene has <50% clusters surpassing 1000 counts in the gene body (or <

441 80% metacells surpassing 20 counts). We then take the ratio of the number of

442 methylated to the number of coverage to get the methylation fraction. All the steps here

443 consider cytosines in non-CG (CH) dinucleotide context only.

24

444   -   For an enhancer mCG matrix (enhancer-by-cluster/metacell), we first removed low

445   coverage enhancers if the gene has <50% clusters surpassing 20 counts (or <80%

446   metacells surpassing 5 counts) in the enhancer region. We then take the ratio of the

447   number of methylated to the number of coverage to get the methylation fraction. All the

448   steps here consider cytosines in CG dinucleotide context only.

449   After normalization and filtering of individual matrices, we then consider only enhancers that are

450   shared in both ATAC and mCG matrices for downstream analyses.

451

452   **Correlating enhancer-gene pairs across cell types.** We calculate the Spearman correlation

453   coefficient between any pair of enhancer and gene that are within 1 Mbp (enhancer center to

454   gene TSS) across curated cell types (n=38 or n=8). This was done separately for enhancer

455   mCG vs. RNA and enhancer ATAC vs. RNA. Enhancer mCG signals are normalized by the

456   global mean mCG levels of each cell type; enhancer ATAC signals are $\log_{10}(TPM+1)$

457   normalized; RNA expression levels are $\log_{10}(CPM+1)$ normalized.

458   To assess the statistical significance of the enhancer-gene correlations, we repeated the

459   correlation analysis with 2 types of data shuffling control, as explained in the main text. To

460   control for random noise, we shuffled cell cluster labels of the gene-by-cluster RNA matrix,

461   followed by calculating correlation coefficients. To control for background co-expression across

462   enhancer-gene pairs, we shuffled gene labels of the gene-by-cluster RNA matrix, followed by

463   calculating correlation coefficients.

464

465   **Correlating enhancer-gene pairs across metacells.** Given a transcriptomic dataset (scRNA-

466   seq) and an epigenetic dataset (e.g. snmC-seq) collected from the same tissue, we first

467   generate a constrained k-nearest neighbor network linking cells across the two modalities

468   (SingleCellFusion; Ref[6,22]). This network allows us to impute the DNA methylation profiles (mC)

469   for each RNA cell. We then cluster scRNA-seq cells using Leiden community detection [34] (see

25

470   section **Clustering/Generating metacells**). We call these clusters *metacells*, to emphasize that

471   they do not necessarily correspond to discrete cell types, but could also capture continuous

472   changes among cell populations. These preparations allow us to construct bi-modal profiles for

473   each metacell, by aggregating counts--either observed or imputed--from cells in the same

474   metacells. Finally, we evaluate the correlations between enhancer-gene pairs across metacells.

475   To be specific, the starting point of this analysis involves 4 matrices: an enhancer-by-cell

476   mCG (or ATAC) matrix E_{ec}, a gene-by-cell RNA matrix $R_{gc'}$, a cross-modal cell-to-cell k nearest

477   neighbor matrix: $K_{cc'}$, and a metacell assignment matrix of RNA cells $K_{c'z}$. Here we use c, c' and

478   z to denote an mC cell, an RNA cell, and a metacell, respectively. A metacell is a group of RNA

479   cells generated by Leiden clustering. We use g and e to denote an enhancer and a gene,

480   respectively. All matrices contain unnormalized raw counts. $K_{cc'}$ is generated by

481   SingleCellFusion[6,22] with default settings and cross-modal k=30. $K_{c'z}$ is generated by Leiden

482   clustering on the RNA-seq dataset as mentioned in previous sections.

483   To get bi-modal profiles for a metacell, we aggregate counts from the cells belonging to

484   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

485   normalized as mentioned in previous sections to adjust for metacell size, library size, and gene

486   length. Finally, normalized R_{gz} and E_{ez} allow us to correlate a specific pair of gene $g^{(i)}$ and

487   enhancer $e^{(i)}$ across metacells (z). We calculated Spearman correlation coefficients for all

488   enhancer-gene pairs with distance between 2kb to 1Mb (enhancer center - TSS).

489

490   **Estimating the statistical significance of enhancer-gene links.** To assess the statistical

491   significance of a correlation coefficient r, we constructed two null distributions by shuffling

492   metacells (Fig. 2b) and shuffling regions (Fig. 2c). In the first case, we shuffle metacell labels

493   independently for transcriptomic and epigenetic data, such that the two data modalities become

494   independent of each other. In the second case, we permute and enhancers randomly from their

495   original genomic location to the locations of other genes and enhancers, while retaining the

496   linked bi-modal profiles of each metacell.

497   Either null distribution can be used to get empirical p-values and false discovery rate

498   (FDR). The empirical p-value of a correlation coefficient r is defined as the cumulative fraction

499   of the null distribution that has more extreme (stronger) correlation coefficients than r. We

500   calculated two-sided p-values when using the shuffled metacells distribution, and single-sided p-

501   values when using the shuffled regions distribution. FDRs are then calculated using the

502   Benjamini-Hochberg procedure[35]. We call an enhancer-gene pair significantly *linked (correlated)*

503   if its empirical FDR is <0.2 using shuffling regions (metacells) as the null.

504   To see if the shuffled regions distribution depends on enhancer properties such as its

505   sequence GC content and distance to the nearest gene, we also performed stratified shuffling

506   analyses (Supplementary Figure 4). We first grouped enhancers into 10 bins (deciles) according

507   to their GC content or distance to the nearest gene. We then shuffled enhancers within each bin

508   and compared observed enhancer-gene correlations with shuffled ones for each bin separately.

509

510   **Enrichment of 3D chromatin contact frequencies.** We validated the predicted enhancer-gene

511   links using single-cell measurements of 3D-chromatin contact frequency in human prefrontal

512   cortex [17]. Raw contact matrices of 8 neuronal cell types were downloaded as mcool files [17]. We

513   calculated contact frequencies from raw counts using matrix balancing using Cooler[36,37]. We

514   then focused on analyzing these contact frequency matrices at a resolution of 10kb non-

515   overlapping genomic bins across the genome.

516   To compare our enhancer-gene links predicted in the mouse brain with the chromatin

517   contact data from human brain, we lifted genes (gencode vM16 whole genes) and putative

518   enhancers from mm10 to hg38 using LiftOver[29] with parameters -minMatch=0.8 and -

519   minBlocks=1.00.

520   To calculate enrichment, we first assigned enhancers (center) and genes (TSS) to their

521   corresponding genomic bins (non-overlapping 10kb bins genomewide). We compared the

522   contact frequencies of the predicted enhancer-gene pairs with random genomic region pairs

523   with similar genomic distance. We separately tested the enrichment of contact frequencies of 6

524   groups of predicted enhancer-gene pairs: mCG-RNA linked, ATAC-RNA linked, pairs linked by

525   both modalities, mCG-RNA correlated, ATAC-RNA correlated, and pairs correlated in both

526   modalities. For each of the 8 neuronal cell types, we only include pairs that are active in the

527   specific cell type, i.e. whose gene expression is greater than the median across all 8 cell types.

528

529   **Comparison with CICERO**. We installed the R package CICERO[8] from the Bioconductor

530   following the instructions from the authors' tutorial (https://cole-trapnell-lab.github.io/cicero-

531   release/docs_m3/#constructing-cis-regulatory-networks). We ran CICERO on MOp ATAC-seq

532   data using default parameters. The program takes as input a peak-by-cell ATAC-seq matrix,

533   where peaks include both putative enhancers we specified and gene promoters (500 bp

534   upstream of TSS). The program returns co-accessibility scores for peak pairs. We filtered the

535   output down to enhancer-promoter pairs only, removing enhancer-enhancer and promoter-

536   promoter pairs. We also focused on analyzing enhancer-gene pairs that are within 100kb apart,

537   to compare with our correlation-based analysis. We used a threshold = 0.2 following Ref[9] to call

538   positive enhancer-gene pairs.

539

540   **Comparison with the ABC model.** We downloaded code from the github repository of the ABC

541   model[5] (https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction) and followed

542   instructions. We ran ABC for each MOp cell type (n=38) using our identified putative enhancer

543   list (n=233,524) and pseudo-bulk ATAC-seq and RNA-seq data as input. We used genomic-

544   distance based power law estimation to model chromatin contacts (--score_column

545   powerlaw.Score). The software returns a score (ABC score) for each enhancer-gene pair and

546   cell type. We excluded the expressed genes from the results, as suggested by the authors. We

547   also focused on analyzing enhancer-gene pairs that are within 100kb. We used a threshold =

548   0.022 as recommended by the authors to call positive enhancer-gene pairs.

549

550   **Generalized least squares (GLS) analysis to decouple covariance across metacells.** We

551   used GLS [16] to test the association between gene expression and enhancer activity across cell

552   types (metacells). We will focus on only one given enhancer-gene pair (g, e), as the same

553   procedure applies to all enhancer-gene pairs independently. Given an enhancer e and gene g,

554   Let y_{cg} be the mRNA expression in cell type c, x_{ce} be the enhancer activity (e.g., mC or ATAC).

555   Let C be the number of cell types. A linear model associating g and e can be written as:

556    $$y_c = a + \beta x_c + \varepsilon_c \qquad\qquad \text{(eq. 1)}$$

557   where c is the index for cell types, β is the association strength, and ε is a noise term. In

558   addition, a is an intercept term that can be omitted after data centering (x and y can be pre-

559   centered to ensure $E[y_c] = E[x_c] = 0$). In matrix notation, (eq. 1) can be simply noted as $y =$

560    $\beta x + \varepsilon$.

561   In ordinary least squares (OLS), we assume ε is uncorrelated across cell types: $E[\varepsilon_c] =$

562    $0, E[\varepsilon_c \varepsilon_{c'}] = \sigma^2 \delta_{c,c'}$. The correlation coefficient $r = E[xy]/\sigma_x \sigma_y$ is then a measure of the linear

563   association, and it has an associated p-value calculated using the t distribution. Alternatively,

564   inference can be performed by permutation analysis to get an empirical p-value.

565   However, in our case we have correlated noise: $E[\varepsilon_c \varepsilon_{c'}] = \Omega_{c,c'}$, which reflects the

566   correlation between cell types due to gene co-expression. That is, $\Omega_{c,c'}$ represents the

567   background of correlated variability in gene expression due to the hierarchical structure of cell

568   types in complex tissues. We can estimate the correlation using the genome-wide covariance,

569    $\hat{\Omega}_{c,c'} = Cov[y]_{c,c'}$. In this case, generalized least squares[16] (GLS) can be used to give an

570   estimate of the coefficient β. This corresponds to transforming the variables x, y from the

571   original basis (cell types/metacells, denoted c) to an decorrelated basis (denoted r), and then

572   performing OLS on the decorrelated variables.

573   We first use singular value decomposition (SVD) to decompose the mean-subtracted

574   gene expression matrix, $y_{cg} = \sum_r U_{cr} S_{rr} V^T{}_{rg}$, where $r = min(c,g)$. Defining $Z = US$, we have

575     $\Omega = ZZ^T$. Multiplying both sides of (eq. 1) by $Z^{-1} = S^{-1}U^T$ corresponds to a transformation from

576   correlated to decorrelated (or whitened) basis:

$$y' = \beta x' + \varepsilon' \qquad\qquad\qquad (eq.\ 2)$$

578

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

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

581   where I is the identity matrix. We can therefore use the correlation coefficient and its associated

582   test statistics on transformed data $y'$and x', as in the case of OLS.

583

584   **Expected range of correlation coefficients for independent variables.** Here we provide

585   theoretical justification on why we expect the range of correlation coefficients ($\hat{r}$) to scale as $\frac{1}{\sqrt{N}}$,

586   as seen in Fig. 2j and Supplementary Fig. 7b, where N is the number of metacells.

587   Let X and Y be two independent random variables. Let x_i and y_i be independent and

588   identically distributed samples of X and Y, where $i \in \{1,2,\ldots,N\}$. In our case, N represents the

589   number of metacells, and x_i and y_i are the transcriptomic and epigenetic signals for a given

590   enhancer-gene pair for metacell i. We require X and Y to be independent of each other as they

591   are unlinked, and x_i and y_i be independent samples as different metacells are also independent

592   observations of X and Y, such as in the case of null distribution created by shuffling cells.

593   To simplify the notation, we assume $E[X] = E[Y] = 0$, as the mean does not affect

594   correlation coefficient r. We also assume X and Y are symmetric, as in the case of normal

595   distribution. It is obvious that $r(X, Y) = 0$. However, we are interested in how the variance of $\hat{r}$

596   depends on N, where $\hat{r}$ is the sample estimate of r by $\{x_i\}$ and $\{y_i\}$.

597
$$var[\hat{r}] \sim E(\hat{r}^2) \sim E\left[\frac{(\sum_{i=1}^{N} x_i y_i)^2}{\sum_{i=1}^{N} x_i^2 \cdot \sum_{i=1}^{N} y_i^2}\right]$$

598
$$= E\left[\frac{\sum_{a=1}^{N} \sum_{b=1}^{N} x_a y_a x_b y_b}{\sum_{i=1}^{N} x_i^2 \cdot \sum_{i=1}^{N} y_i^2}\right]$$

599
$$= \sum_{a=1}^{N} \sum_{b=1}^{N} E\left[\frac{x_a y_a x_b y_b}{\sum_{i=1}^{N} x_i^2 \cdot \sum_{i=1}^{N} y_i^2}\right]$$

600
$$= \sum_{a=1}^{N} E\left[\frac{(x_a y_a)^2}{\sum_{i=1}^{N} x_i^2 \cdot \sum_{i=1}^{N} y_i^2}\right]$$

601   (eq.3)

602   The last equality holds, as only non-interaction terms ($a = b$) are nonzero. Moreover, as $(x_a y_a)^2$

603   are equivalent for different $a = \{1 \ldots N\}$, the above summation can be further simplified as:

604
$$\sum_{a=1}^{N} E\left[\frac{(x_a y_a)^2}{\sum_{i=1}^{N} x_i^2 \cdot \sum_{i=1}^{N} y_i^2}\right] = N \cdot E\left[\frac{(x_1 y_1)^2}{\sum_{i=1}^{N} x_i^2 \cdot \sum_{i=1}^{N} y_i^2}\right] = N \cdot E\left[\frac{x_1^2}{\sum_{i=1}^{N} x_i^2}\right] \cdot E\left[\frac{y_1^2}{\sum_{i=1}^{N} y_i^2}\right],$$

605   (eq. 4)

606   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

607   arrive at

608
$$var(\hat{r}) \propto N \cdot E\left[\frac{x_1^2}{\sum_{i=1}^{N} x_i^2}\right] \cdot E\left[\frac{y_1^2}{\sum_{i=1}^{N} y_i^2}\right] = N \cdot \frac{1}{N} \cdot \frac{1}{N} = \frac{1}{N},$$

609   (eq. 5)

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

611

612   **Supplementary tables**

613   **Table S1.** A list of putative enhancers (cCREs; n=233,524 in total)

614 **Table S2.** Significant linked enhancer-gene pairs by mCG-RNA correlation

615 **Table S3.** Significant linked enhancer-gene pairs by ATAC-RNA correlation

616 **Table S4.** Cell type correspondence between this study and Ref[6]

617

618

Acknowledgements

625

Author contributions

627 EAM and FX designed the study. ZY, BT, and HZ generated scRNA-seq data. HL, AB, MMB,

628 JDL, CL, JRN, APD, ACR and JRE generated DNA methylation (snmC-Seq) data. HL, MMB,

629 YEL, JDL, APD, OP, SP, and BR generated snATAC-Seq data. FX led the computational

630 analysis. FX and EA developed code and performed analysis. FX, EA, and EAM wrote and

631 edited the manuscript. All authors approved the manuscript.

632

Competing interests

634 The authors declare no competing interests.

635

636 **Data availability**

637 The scRNA-seq, snmC-seq, and snATAC-seq datasets from the mouse primary motor cortex

638 are generated by BICCN (RRID:SCR_015820) as reported previously[6]. The data can be

639 accessed via the NeMO archive (RRID:SCR_002001) at accession:

640 https://assets.nemoarchive.org/dat-ch1nqb7. Genome browser:

641 https://brainome.ucsd.edu/BICCN_MOp. The chromatin contact data generated by snm3C-seq

642 is downloaded from publicly available files (Ref[17];

643 https://salkinstitute.app.box.com/s/fp63a4j36m5k255dhje3zcj5kfuzkyj1).

644

645 **Code availability**

646 Analysis scripts used for this paper are at https://github.com/FangmingXie/scf_enhancer_paper.

647 SingleCellFusion is available at https://github.com/mukamel-lab/SingleCellFusion.

648 The code for ABC model analysis is downloaded from its github repository

649 (https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction; Ref[5]).

650 The code for CICERO analysis is downloaded as an R package

651 (https://www.bioconductor.org/packages/release/bioc/html/cicero.html; Ref[8]).

652

653 **References**

654 1. Liu, H. *et al.* DNA methylation atlas of the mouse brain at single-cell resolution. *Nature* **598**,

655   120–128 (2021).

656 2. Li, Y. E. *et al.* An atlas of gene regulatory elements in adult mouse cerebrum. *Nature* **598**,

657   129–136 (2021).

658 3. Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of

659   validated and target-linked human enhancers. *Nat. Rev. Genet.* **21**, 292–310 (2020).

660   4.   Gasperini, M. *et al.* A Genome-wide Framework for Mapping Gene Regulation via Cellular

661   Genetic Screens. *Cell* **176**, 1516 (2019).

662   5.   Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from

663   thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).

664   6.   Yao, Z. *et al.* A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex.

665   *Nature* **598**, 103–110 (2021).

666   7.   Yao, Z. *et al.* A taxonomy of transcriptomic cell types across the isocortex and hippocampal

667   formation. *Cell* **184**, 3222–3241.e26 (2021).

668   8.   Pliner, H. A. *et al.* Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell

669   Chromatin Accessibility Data. *Mol. Cell* **71**, 858–871.e8 (2018).

670   9.   Cusanovich, D. A. *et al.* A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility.

671   *Cell* **174**, 1309–1324.e18 (2018).

672   10. Zhu, C. *et al.* An ultra high-throughput method for single-cell joint analysis of open

673   chromatin and transcriptome. *Nat. Struct. Mol. Biol.* **26**, 1063–1070 (2019).

674   11. Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers.

675   *Science* **362**, eaav1898 (2018).

676   12. Trevino, A. E. *et al.* Chromatin accessibility dynamics in a model of human forebrain

677   development. *Science* **367**, eaay1645 (2020).

678   13. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and

679   Chromatin. *Cell* **183**, 1103–1116.e20 (2020).

680   14. Gorkin, D. U. *et al.* An atlas of dynamic chromatin landscapes in mouse fetal development.

681   *Nature* **583**, 744–751 (2020).

682   15. Sarropoulos, I. *et al.* Developmental and evolutionary dynamics of cis-regulatory elements

683   in mouse cerebellar cells. *Science* **373**, (2021).

684   16. Aitken, A. C. On Least Squares and Linear Combination of Observations. *Proceedings of*

685   *the Royal Society of Edinburgh* **55**, 42–48 (1936).

686   17.   Lee, D.-S. *et al.* Simultaneous profiling of 3D genome structure and DNA methylation in

687   single human cells. *Nat. Methods* **16**, 999–1006 (2019).

688   18.   Serwach, K. & Gruszczynska-Biegala, J. STIM Proteins and Glutamate Receptors in

689   Neurons: Role in Neuronal Physiology and Neurodegenerative Diseases. *Int. J. Mol. Sci.*

690   **20**, 2289 (2019).

691   19.   Schoenfelder, S. & Fraser, P. Long-range enhancer-promoter contacts in gene expression

692   control. *Nat. Rev. Genet.* **20**, 437–455 (2019).

693   20.   Endersby, J. Lumpers and splitters: Darwin, Hooker, and the search for order. *Science* **326**,

694   1496–1499 (2009).

695   21.   Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of

696   Brain Cell Identity. *Cell* **177**, 1873–1887.e17 (2019).

697   22.   Luo, C. *et al.* Single nucleus multi-omics links human cortical cell regulatory genome

698   diversity to disease risk variants. *bioRxiv* 2019.12.11.873398 (2019)

699   doi:10.1101/2019.12.11.873398.

700   23.   Baran, Y. *et al.* MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions.

701   *Genome Biol.* **20**, 206 (2019).

702   24.   Nettleton, D., Hwang, J. T. G., Caldo, R. A. & Wise, R. P. Estimating the number of true null

703   hypotheses from a histogram of p values. *J. Agric. Biol. Environ. Stat.* **11**, 337 (2006).

704   25.   Yin, Y. *et al.* Impact of cytosine methylation on DNA binding specificities of human

705   transcription factors. *Science* **356**, eaaj2239 (2017).

706   26.   Daigle, T. L. *et al.* A Suite of Transgenic Driver and Reporter Mouse Lines with Enhanced

707   Brain-Cell-Type Targeting and Functionality. *Cell* **174**, 465–480.e22 (2018).

708   27.   Graybuck, L. T. *et al.* Enhancer viruses for combinatorial cell-subclass-specific labeling.

709   *Neuron* **109**, 1449–1464.e13 (2021).

710   28.   de Boer, C. G. *et al.* Deciphering eukaryotic gene-regulatory logic with 100 million random

711   promoters. *Nat. Biotechnol.* **38**, 56–65 (2020).

712   29. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006

713   (2002).

714   30. He, Y. *et al.* Improved regulatory element prediction based on tissue-specific local

715   epigenomic signatures. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E1633–E1640 (2017).

716   31. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic

717   features. *Bioinformatics* **26**, 841–842 (2010).

718   32. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of

719   Problematic Regions of the Genome. *Sci. Rep.* **9**, 9354 (2019).

720   33. Zeng, H. & Sanes, J. R. Neuronal cell-type classification: challenges, opportunities and the

721   path forward. *Nat. Rev. Neurosci.* **18**, 530–546 (2017).

722   34. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-

723   connected communities. *Sci. Rep.* **9**, 5233 (2019).

724   35. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful

725   approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* 289–300 (1995).

726   36. Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome

727   organization. *Nat. Methods* **9**, 999–1003 (2012).

728   37. Abdennur, N. & Mirny, L. A. Cooler: scalable storage for Hi-C data and other genomically

729   labeled arrays. *Bioinformatics* **36**, 311–316 (2020).

730