

1 **Endogenous fine-mapping of functional regulatory elements in complex**
2 **genetic loci**
3

4 Ke Zhao^{1, #}, Yao Zhou^{1, #}, Chengyue Wu², Jianhua Wang¹, Hongcheng Yao³, Xin Cheng¹, Lin
5 Zhao¹, Wei Wang⁴, Xinlei Chu⁴, Xianfu Yi¹, Yupeng Chen², Miaoxin Li⁵, Wange Lu⁶, Kexin
6 Chen⁴, Pak Chung Sham³, Mulin Jun Li^{1,*}

7

8 ¹Department of Bioinformatics, The Province and Ministry Co-sponsored Collaborative Innovation
9 Center for Medical Epigenetics, Key Laboratory of Prevention and Control of Human Major Diseases
10 (Ministry of Education), Tianjin Medical University Cancer Institute and Hospital, Tianjin Medical
11 University, Tianjin, China.

12 ²Department of Biochemistry and Molecular Biology, School of Basic Medical Sciences, Tianjin Medical
13 University, Tianjin, China.

14 ³Centre for PanorOmic Sciences-Genomics and Bioinformatics Cores, The University of Hong Kong,
15 Hong Kong.

16 ⁴Department of Epidemiology and Biostatistics, Tianjin Key Laboratory of Molecular Cancer
17 Epidemiology, National Clinical Research Center for Cancer, Tianjin Medical University Cancer Institute
18 and Hospital, Tianjin Medical University, Tianjin, China.

19 ⁵Program in Bioinformatics, Zhongshan School of Medicine, Sun Yat-Sen University, Guangzhou, China

20 ⁶Institute of Precision Medicine, The First Affiliated Hospital of Sun Yat-Sen University, Guangzhou,
21 China.

22

23 [#]The authors contributed equally to this work

24 *Correspondence: mulinli@connect.hku.hk (M.J.L.)

25

26 **Abstract**

27 The vast majority of genetic loci associated with polygenic complex traits are located in
28 non-coding regions of the human genome. However, many of these regions exhibit high-
29 order gene regulatory relationships and complicated linkage disequilibrium (LD)
30 configurations, which bring challenges to accurately identify causal variants and their
31 target genes controlling specific molecular processes or traits. We employed multiplexed
32 single-cell CRISPR interference and activation perturbations to explore the links between
33 *cis*-regulatory element (CRE) and target gene expression within tight LD in the
34 endogenous chromatin context. We validated the prevalence of multiple causality in
35 perfect LD (pLD) for independent expression quantitative trait locus (eQTL), and revealed
36 fine-grained genetic effects on gene expression within pLD. These effects are difficult to
37 decipher using conventional eQTL fine-mapping or to predict via existing computational
38 methods. We found that nearly half of the casual CREs lack classical epigenetic markers,
39 potentially affecting gene expression through hidden regulatory mechanisms. Integrative
40 analysis on different types of perturbation effects suggested a high regulatory plasticity of
41 the human genome. These findings will propel further in-depth exploration of functional
42 genomic elements, facilitating a more comprehensive understanding of gene expression
43 regulatory patterns and the development of complex traits.

44

45

46 **Introduction**

47 Identifying fine-grained regulatory elements and complex trait/disease causal regulatory
48 variants on the human genome is a significant challenge in the current fields of functional
49 genomics and genetics. Years of functional genomic profiling and expression quantitative
50 trait locus (eQTL) studies have identified numerous *cis*-regulatory elements (CREs) and
51 expression-associated alleles in nearly a hundred human tissues/cells (1-3). However,
52 accurately pinpointing which CREs or even single allele(s) can modulate gene
53 expression under specific biological conditions remains difficult. This is especially true for
54 complex genetic loci, where complicated CRE-target gene relationships and linkage
55 disequilibrium (LD) contamination make it computationally and experimentally
56 challenging to precisely locate all causal elements and their target genes (4).
57 Furthermore, genome-wide association studies (GWASs) have shown that the majority of
58 genetic loci associated with complex traits and diseases are located in non-coding
59 regions of the genome; however, colocalization analysis with eQTLs revealed only a
60 limited (8%-25%) proportion of shared genetic loci (5-7). This has sparked debate within
61 the field over whether a substantial portion of trait/disease-causal variants may not cause
62 phenotypic development by affecting gene expression.

63

64 A variety of large-scale, high-throughput experimental methods have been used to
65 systematically evaluate the regulatory potential of human genomic sequences and allele-
66 specific effects. Firstly, exogenous and episomal massively parallel reporter assays
67 (MPRAs) are highly efficient in identifying functional regulatory sites and alleles in various
68 cells (8, 9). Consistent with computational simulations (4), a recent study employed
69 MPRA to deeply assess high LD variants within independent eQTL signals,
70 systematically validating the widespread presence of multiple allelic effects in tight LD
71 (10). However, these experiments struggle to evaluate genetic effects and relationships
72 with specific molecular phenotypes (e.g., target gene expression) in the endogenous
73 chromatin environment of local genetic loci. Additionally, using CRISPR base editing
74 technology, researchers have been able to study the links between variants and complex
75 phenotypes in targeted genomic regions (11-14). However, due to the restrictions of
76 editing preference and efficiency in mammalian cells, most of these strategies are not
77 suitable for fine-mapping causal genetic loci affecting molecular phenotypes. Finally,
78 some studies employ CRISPR-based chromatin perturbations to exhaustively
79 characterize regulatory sites and their causal relationships with particular gene
80 expression(s) through tiling screening (15-18), but these studies have focused solely on a
81 small number of target genes.

82

83 Multiplexed single-cell CRISPR perturbations provide technical support for systematically
84 studying the regulatory relationships between genetic loci and fine-grained molecular
85 phenotypes in an endogenous cellular environment (19-21). However, these studies
86 heavily rely on prior knowledge (e.g., specific epigenetic markers) for the selection of
87 genetic loci and employ solely a single type of CRISPR perturbation. Whether such
88 technologies can accurately and unbiasedly pinpoint causal CREs that regulate gene

89 expression within complex genetic loci (e.g., ultra-high LD region) is a scientific question
90 worth exploring. In this study, we leveraged multiplexed single-cell perturbations, through
91 both CRISPR interference and activation, to validate the ubiquity of multiple causal CREs
92 in perfect LD (pLD) within independent eQTL signal and to interrogate their regulatory
93 relationships with target genes. We have demonstrated that endogenous perturbations
94 can reveal intricate genetic effects and gene expression regulatory patterns, which are
95 challenging to identify through conventional eQTL mapping approaches. Moreover, we
96 found that current computational methods struggle to accurately predict the effects of
97 many endogenous perturbations on gene expression, and further investigation suggested
98 that nearly half of the causal CREs lack classical epigenetic markers, potentially
99 influencing gene expression through unique regulatory mechanisms. Lastly, by
100 comparing various types of CRISPRi/a perturbations, we have shed light on the
101 regulatory plasticity of the human genome through a distinct perspective.
102

103 **Results**

104 **Comprehensive discrimination of independent eQTL effect at perfect LD (pLD)** 105 **using multiplexed single-cell CRISPR perturbations**

106 Given fine-mapping causal allele(s) among perfectly correlated variants is either
107 computationally intractable or experimentally costly, we designed a series of extreme
108 scenarios and applied a single-cell perturbation strategy to provide unbiased
109 interrogation of expression-modulating causal effects at native genomic settings.
110 Generally, we first leveraged statistical fine-mapping on GTEx V8 whole-genome
111 sequencing (WGS)-based eQTLs (670 whole blood samples) to nominate independent
112 signals. Additionally, we collected and fine-mapped six WGS-based blood-derived eQTL
113 datasets from Geuvadis (445 lymphoblastoid cell samples), BLUEPRINT (190 monocyte
114 samples, 196 neutrophil samples and 165 CD4+ T cell samples), and TwinsUK (523
115 lymphoblastoid cell samples and 246 whole blood samples) (Table S1), and required that
116 independent signal is reproducible in at least one additional dataset. Ultimately, we
117 integrated all fine-mapped variants in seven eQTL datasets to retain independent signals
118 which contain at least two undistinguishable lead eQTL variants (LEVs, with equal causal
119 probability) in pLD (Fig. 1A and Fig. S1A, see details in Methods). As large-scale base
120 editing screens, such as prime editing (13), coupled with transcriptome readouts are
121 currently impractical, we performed multiplexed regional perturbations at the single-cell
122 level to investigate the causal relationships and multiplicity underlying eQTL effects within
123 these complex signals (Fig. 1A, see details in Methods).
124

125 To achieve effective discrimination of single-cell expression perturbations via CRISPR,
126 we focused on eQTL genes (eGenes) which are highly (top 20%) expressed in a human
127 near-haploid leukemia cell line (HAP1), and excluded unqualified LEVs (e.g., proximal
128 variants or protein-coding/splicing-altering variants) in each signal (Fig. S1A and Fig. S1B,
129 see details in Methods). Thus, 81 independent eQTL signals (reproducible in 2 to 7
130 additional datasets) were finally selected, and each contains more than a single qualified
131 LEV within pLD (range of 2 to 6). Finally, the perturbation library incorporates 217 LEVs
132 (Fig. 1B and Table S2). Two or three single guide RNAs (sgRNAs) were designed to

133 target each LEV-tagged region, and additional two sets of control sgRNAs were also
134 used, including 39 non-targeting control (NTC) sgRNAs as negative control and 11
135 previously validated sgRNAs as positive control (21). The sequence of 472 designed
136 sgRNAs was cloned into a lentiviral CROP-seq-opti vector (Table S3) (22). The quality of
137 the constructed sgRNA plasmid library was validated and was found to be of high quality,
138 as evidenced by the coverage rate of greater than 99% (Fig. S1C) and a low degree of
139 uniformity, with the top 90th to 10th sgRNA having a representation difference of less than
140 10-fold (Fig. S1D). Consistency in sgRNA distribution within plasmid library was also
141 observed upon viral transduction at varied multiplicity of infection (MOI) (Fig. S1E).

142

143 To maximize the detection power of CRISPR-based perturbations for fine-mapped LEV-
144 tagged CRE (eCRE, \pm 50 bp region of LEVs) discovery, we used both CRISPR
145 interference (CRISPRi, a nuclease-deactivated Cas9 tethered to the KRAB repressor
146 domain, dCas9-KRAB) and CRISPR activation (CRISPRa, a dCas9 tethered to the
147 transcriptional activator VP64, dCas9-VP64) systems (Fig. S1F). Previous studies have
148 demonstrated that the introduction of multiple perturbations per cell can substantially
149 augment the statistical power to identify causal relationships between CREs and their
150 target genes in a cost-efficient manner (21, 23, 24). Thus, lentiviral vectors containing our
151 sgRNA library were then transduced into selected monoclonal cell lines at either
152 moderate or high MOI (Fig. S1F-S1I). This approach allowed us to comprehensively
153 investigate the causal eCREs and their target genes among multiple eQTL signals, while
154 minimizing genetic heterogeneity in our experimental system. After a 14-day cell culture
155 for effective CRISPRi/a perturbation, the transcriptomes of 14,481/15,296 (CRISPRi) and
156 17,303/17,709 (CRISPRa) single cells were profiled with four 10x single-cell RNA
157 sequencing (scRNA-seq) libraries. Targeted amplification of sgRNAs from cDNA in these
158 perturbation libraries (25) suggested that the number of sgRNAs per cell and the number
159 of cells per perturbation was increased as MOI increases in both CRISPRi and CRISPRa
160 perturbations (Fig. S1J and S1K). Joint analysis of data under different MOIs revealed a
161 median of 13 sgRNAs per cell and a median of 850 cells bearing each perturbation with
162 CRISPRi, and a median of 6 sgRNAs per cell and a median of 245 cells bearing each
163 perturbation with CRISPRa, respectively (Fig. 1C).

164

165 We applied a unified normalization-association framework, Normalisr (26), to analyze the
166 relationship between each eCRE and nearby expressed genes (\pm 1 Mb) (see details in
167 Methods). Quantile-quantile plots indicated an excess of significant associations of
168 sgRNAs targeting eCRE compared with NTCs in all conditions (132 and 34 significant
169 pairs of eCRE and its perturbed gene (perturbGene) in merged CRISPRi (Table S4) and
170 CRISPRa (Table S5) screenings respectively, $FDR < 0.2$). The perturbation with high MOI
171 achieved higher power than moderate MOI (Fig. 1D), and the perturbation effects from
172 significant hits between moderate and high MOIs showed good agreement (Fig. S2A and
173 Fig. S2B). Besides, we observed significant consistency between the results of the
174 perturbation experiments under different MOI conditions ($R = 0.23$, $P = 0.0034$) (Fig.
175 S2C). To further explore the influence factors determining the statistical power of
176 multiplexed single-cell CRISPR perturbations, we simulated several scRNA-seq datasets

177 with various perturbation conditions, including effective sgRNAs per cell and total
178 captured cells in each perturbation (see details in Methods). The results demonstrated
179 that, consistent with the real experiments, increasing the number of perturbations in each
180 cell led to better statistical power. However, simply increasing the number of captured
181 cells in scRNA-seq only slightly promoted statistical power (Fig. S2D). These findings
182 suggest that, for perturbation effects that are sparse in terms of their downstream
183 consequences, multiplexed sgRNAs will offer powerful and cost-efficient solution in
184 single-cell CRISPR perturbations.

185

186 To validate the screening results, we first confirmed the effects of all positive control
187 sgRNAs, which were highly concordant with the literature report (Fig. 1E). Additionally,
188 we randomly selected 20 groups of sgRNAs targeting causal eCRE and performed
189 individual CRISPRi or CRISPRa perturbations, followed by reverse-transcription qPCR
190 (RT-qPCR), to confirm the effect of sgRNAs on their pertubGenes, with no-targeting
191 sgRNAs as control. As expected, a positive correlation ($R = 0.75$, $P = 5.70e-12$, T-test)
192 between the RT-qPCR results and the effect size observed in multiplexed single-cell
193 CRISPR perturbations was found (Fig. 1F and Fig. S3). This results further demonstrate
194 the credibility of our single-cell perturbation screens in testing the effects of eCREs on
195 their potential target genes.

196

197 **Multiple causal *cis*-effects and target gene configurations underlie complex 198 genetic associations on gene expression**

199 Based on the significant hits in both CRISPRi/CRISPRa screens, we first inquired
200 whether the presence of multisite *cis*-regulation and multiplicity of target genes in pLD is
201 prevalent. Our results showed that, over 70% of the investigated pLD signals contained
202 at least one significant causal eCRE. Among these significant pLD signals, 49.1% of
203 them showed multiple (2 to 5) causal eCRE (44.2% in CRISPRi and 11% in CRISPRa,
204 respectively), suggesting a high proportion of multisite *cis*-regulation. Besides, over half
205 (52.6%) of these pLD signals had multiple target genes, with the maximum being greater
206 than four, while standalone CRISPRi and CRISPRa analyses revealed proportions of
207 51.9% and 14.8%, respectively. Additionally, we found that the perturbGenes from 39%
208 pLD signals were the closest to the associated causal eCRE, 53% were located distally,
209 and the remaining 9% could link to both closest and distal target genes (Fig. 2A and Fig.
210 S4A). These results emphasize the pervasive existence of multisite *cis*-regulation
211 affecting various target genes in pLD, which complicates the identification of true causal
212 variants in both eQTL and GWAS fine-mapping.

213

214 To investigate the target gene configurations underlying the complex genetic associations
215 on gene expression, we partitioned the pLD signals containing significant eCRE-
216 perturbGene pairs into two regulatory patterns (single causal and multiple causal) (Fig.
217 2B, Fig. S4B). For pLD signals with unique causal eCRE, some only associated with the
218 nearest gene (12.3% of the total significant pLD signals) (Fig. 2B). For example, two
219 highly linked LEVs, rs3782235 (GRCh37: chr12:56915547-G-A) and rs4759247
220 (12:56918834-T-G) located more than 1 Kb apart from each other. Our CIRSPRa

221 perturbation screens revealed that only rs3782235-tagged CRE significantly affected the
222 expression of the adjacent gene *RBMS2*, which was confirmed by RT-qPCR. Specifically,
223 the sgRNA targeting rs3782235 significantly up-regulated the expression of *RBMS2* with
224 a greater effect size than sgRNAs targeting rs475924, consistent with the screening
225 results (Fig. 2C). Interestingly, a recent GWAS of blood traits (27) found rs3782235 was
226 significantly associated with hematocrit percentage, suggesting that rs3782235 could be
227 a causal variant modulating *RBMS2* expression in haematocrit-related traits. Besides,
228 instead of the nearest gene, 29.8% significant pLD signals connected to a single distal
229 gene via corresponding causal eCRE (Fig. 2B). Moreover, causal eCREs in a small
230 fraction of pLD signals (8.8%) can regulate multiple target genes (Fig. 2B). For instance,
231 CRISPRa perturbations revealed that rs1049359-tagged CRE, rather than other highly
232 linked variant-marked eCRE in pLD, affected the expression of two distal target genes,
233 *RNF181* and *TMSB10*, which were independently verified by RT-qPCR (Fig. 2C).

234
235 As for multiple causal patterns, 19.3% of the total significant pLD signals incorporated
236 more than one causal eCREs targeting the same target gene(s), indicating the common
237 phenomenon of multisite constraints on eQTL fine-mapping (4, 10). For example, we
238 identified two significant eCREs (tagged by rs4930698 (chr11:64085063-G-C) and
239 rs79423518 (chr11:64105454-G-A)) in a pLD signal that were both associated with two
240 target genes *PRDX5* and *TRMT112*. Consistent with the CRISPRi screening results, the
241 effect of two eCREs was confirmed through RT-qPCR, which showed that the sgRNAs
242 targeting the corresponding eCRE significantly decreased the expression levels of both
243 *PRDX5* and *TRMT112* (Fig. 2C). The rs4930698-tagged CRE is located in the upstream
244 promoter region of *PRDX5* and downstream promoter region of *TRMT112*, and overlays
245 several active chromatin marks including H3K4me3, H3K27ac and open chromatin,
246 indicating its high regulatory potential. Besides, rs79423518-tagged CRE lies at
247 intergenic region ~20 Kb downstream of *PRDX5*, and obtains weak enhancer marks such
248 as H3K4me1 and H3K27ac (Fig. 2D). To investigate the regulatory function of these two
249 eCREs, we first performed chromosome conformation capture combined with high-
250 throughput sequencing (4C-seq) that anchored at rs79423518-tagged CRE, and
251 observed a strong interaction between the CRE and the promoter region of *PRDX5* and
252 *TRMT112*. This suggests that a direct regulation between rs79423518-tagged CRE and
253 the promoter of two target genes. Then, luciferase reporter assay revealed that
254 rs4930698-tagged CRE exhibited both promoter and enhancer activities, and showed an
255 allele-specific effect, while rs79423518-tagged CRE was also found to have regulatory
256 functions (Fig. 2E). By contrast, more pLD signals (29.8%) received multiple causal
257 eCREs that regulated different target genes (Fig. 2B), highlighting the complexity of
258 genetic regulation in highly linked loci. In summary, our endogenous perturbation screen
259 in pLD serves as a valuable method to facilitate the identification of true causal variants
260 and their associated CREs in cases where statistical fine-mapping faces challenges,
261 while also nominating potential target genes regulated by functional CREs. Consistent
262 with recent exogenous research (10), our findings question the assumption that a single
263 variant typically accounts for the causality of an independent association locus. Overall,
264 this result underscores the importance of recognizing the complexity of genetic regulation

265 when interpreting GWAS signals.

266

267 **Comparison of regulatory effects between eCRE perturbation and conventional**
268 **eQTL mapping**

269 To assess the physiological relevance of the eCRE-perturbGene pairs uncovered through
270 our multiplexed single-cell CRISPR perturbations, we compared them with the whole
271 blood eQTL mapping results from GTEx V8 and eQTLgen (28). Among all significant
272 eCRE-perturbGene pairs identified in pLD regions, approximately 30% associated eQTL-
273 eGene pairs were found to exist in either GTEx or eQTLgen (Fig. 3A). Similar trends
274 were also observed for perturbGenes (Fig. 3B), suggesting that a large proportion of
275 significant hits from CRISPR perturbations were not captured by current eQTL mapping
276 in whole blood. To explore this further, we compared our findings to a larger eQTL
277 dataset from QTLbase (3), which integrates multiple eQTL studies from various
278 tissues/cell types and conditions. We found that a large proportion of additional overlaps
279 could be recovered (Fig. S5A-S5D), suggesting different types of CRISPR perturbation
280 could capture some eCREs that are specific to the cell type or environmental condition
281 not being well studied.

282

283 Previous statistical methods for causal eQTL fine-mapping mostly assumed that
284 functional regulatory variants are sparsely distributed and the tight linkage among them
285 has limited the ability to accurately estimate the magnitude of genetic effects (4). Despite
286 the positive overall effect correlation between eCRE-perturbGene pairs from CRISPR
287 perturbations and corresponding eQTL-eGene pairs from GTEx whole blood tissue ($R =$
288 0.14, $P = 0.039$), we observed many discrepancies at same pLD signal with multiple
289 causal eCREs (Fig. 3C). For examples, independent CRISPR perturbations showed
290 similar or varied effects for two causal eCREs in pLD respectively (Fig. 3D and Fig. 3E).
291 However, at these multisite regulation pLD regions, the true effect sizes of individual
292 causal loci were indistinguishable (e.g., overestimation or underestimation) using
293 conventional eQTL mapping. These suggest that our multiplexed single-cell CRISPR
294 perturbations offer a more comprehensive assessment of the magnitude of expression-
295 modulating causal effects at endogenous genomic environment.

296

297 The high enrichment of functional eQTLs near the transcriptional start site (TSS) had
298 been extensively documented (29, 30). By evaluating the target gene TSS distances of
299 eQTL-associated LEVs from GTEx and causal eCRE-associated LEVs from CRISPR
300 perturbation, we found that the LEVs from CRISPR screen hits lie at greater distances
301 from the nearest TSS (median 26 Kb) compared to LEVs in conventional eQTL mapping
302 (median 400 Kb) ($P < 2.2e-16$, T test, Fig. 3F and Fig. 3G, see details in Methods). For
303 instance, rs59508494 (19:16211630-A-G) and rs143558304 (19:16213697-T-TA) are
304 located close to each other (~2 Kb) within pLD region. The two LEVs are significantly
305 associated with *TPM4* gene expression in GTEx whole blood tissue. Interestingly, Both
306 CRISPRi perturbation screen and RT-qPCR revealed that rs59508494 could regulate a
307 distal gene *EPS15L* (390 Kb) and rs143558304 could regulate another distal gene *ILVBL*
308 (900 Kb), respectively (Fig. 3I and 3J). The long-distance interactions between the two

309 LEVs and their previously unknown targets were validated by 4C-seq that anchored at a
310 genomic fragment containing the two variants (Fig. 3H). Given the evidence that GWAS
311 hits are further from TSSs than eQTLs and show limited overlaps with them (5-7), the
312 multiplexed CRISPR perturbations would provide a comprehensive supplement to study
313 the shared genetic effect between disease/trait-causal variants and functional regulatory
314 sites.

315

316 Topologically associating domains (TAD) and LD are two measurements of chromosomal
317 interaction and genome genetic structure respectively, by which the genome is divided
318 into different segments. Boundaries and ranges of both measurements are important for
319 exploring the relationship between genetic architecture and gene regulation. Previous
320 findings revealed that genomic architectures of genetic and physical interactions are
321 generally independent, and the regulation range of eQTL-eGene is irrelevant with LD (31).
322 Here, using the eCRE-perturbGenes identified by CRISPR perturbation, we reassessed
323 such relevance and found that the causal eCREs regulating their target genes within the
324 same TAD are more likely to locate in a highly linked LD block (Odds ratio = 14.2, $P =$
325 9.4e-11, Fisher's exact test, Fig. 3K, see details in Methods). This also suggests that
326 artificial genetic perturbation would capture additional layers of gene regulation against
327 the traditional eQTL mapping in homeostatic conditions.

328

329 **Endogenous perturbation effects are poorly predicted via computational
330 predictions and functional annotations**

331 Given statistical fine-mapping faced the tremendous challenge to accurately identify true
332 causal eQTL variants in pLD, we next sought to evaluate the performance of the existing
333 computational methods and functional annotations in distinguishing the regulatory
334 potential of eCREs through endogenous CRISPR perturbation. First, we leveraged 20
335 functional/pathogenic variant scores to test their abilities of causal eCRE/LEV
336 classification (Table S6 and Table S7, see details in Methods). Consistent with the
337 previous benchmarks using massively parallel reporter assays (MPRAs) data (10, 32),
338 existing prediction tools showed restricted performance in discriminating significant
339 causal eCREs or corresponding LEVs from non-significant ones in our CRISPR
340 perturbation screens (Fig. 4A, Fig. 4B, Fig. S6A and Fig. S6B). Specifically, the results
341 showed that among the 20 methods for predicting functional eCRE/LEV, DVAR (33),
342 RegBase_REG (34) and Eigen-PC (35) achieved a better performance (Area Under the
343 ROC Curves (AUCs) are close to 0.6) than others in the majority of benchmarks. Notably,
344 these top-performed tools were either learned from unsupervised algorithms (such as
345 DVAR and Eigen-PC) or presented an ensemble score by integrating existing prediction
346 methods (like RegBase), implying they could capture unknown features that explain the
347 endogenous activity of regulatory sites. Second, we applied a recent deep learning model,
348 Enformer (36), which predicts sequence effects on gene expressions and chromatin
349 states, to investigate the agreement between predicted effects and perturbation effects at
350 tested eCRE/LEV sites (Table S6 and Table S7, see details in Methods). We noted
351 similar enrichment patterns of investigated eCREs and the associated LEVs on Enformer
352 scores, in which significantly causal eCREs or tagging LEVs enriched at the top

353 percentiles of Enformer scores from the first two components, compared with non-
354 significant ones (Fig. 4C, Fig. 4D, Fig. S6C and Fig. S6D). Nevertheless, the differences
355 between the eCRE groups remain small.

356

357 Furthermore, we illustrated a case wherein computational predictions and functional
358 annotations poorly worked. Specifically, we observed that rs75446625 and rs80159064
359 were two perfectly linked variants located in the intronic region of *NUDT7*. While the
360 rs75446625-tagged CRE harbored several active chromatin states (including open
361 chromatin, H3K27ac and H3K4me3), the rs80159064-tagged CRE was completely
362 depleted from classical markers. As expected, rs75446625 was highly scored by four top-
363 performed scores (including DVAR, CDTs (37), RegBase_REG and Eigen-PC) and two
364 Enformer components. In contrast, rs80159064 showed very low scores for most of these
365 tools (Fig. 4E). However, our CRISPRi perturbation screens identified that the CREs
366 tagged by the two LEVs can modulate the expression of a common target gene *NUDT7*,
367 which was also successfully validated through RT-qPCR (Fig. 4F). Luciferase reporter
368 assays revealed that the two eCREs both had regulatory functions as an enhancer and
369 showed allelic-specific effects (Fig. 4G and Fig. 4H). Particularly, compared to
370 rs75446625-tagged CRE, rs80159064-tagged CRE exhibited larger effect in CIRSPR
371 screen (Fig. 4E) and equivalent effects at *in vitro* reporter assays, respectively. Taken
372 together, current computational methods for statistical fine-mapping and functional
373 prediction are less actionable for the identification of true causal regulatory variants in
374 high LD.

375

376 **Unbiased endogenous perturbation reveals many unmarked regulatory elements**

377 The majority of previous Perturb-seq studies targeting regulatory DNA sequences
378 typically rely on prior knowledge of specific types of CREs, such as enhancers, open
379 chromatin regions, and transcription factor binding sites (38). However, in our CRISPRi/a
380 perturbation screen, we did not use any chromatin marks or sequence features to select
381 LEVs and associated eCREs. This approach provided a unique opportunity to unbiasedly
382 learn the regulatory potential of genomic sequences. By integrating seven well-
383 characterized epigenetic marks in HAP1 cells, we were able to classify the identified
384 eCREs into two major groups: marked CREs and unmarked CREs (UREs) (see details in
385 Methods). Active but not repressive chromatin signals, such as open chromatin (ATAC-
386 seq), enhancer/promoter (H3K27ac, H3K4me1, and H3K4me3), and actively transcribed
387 genomic regions (H3K36me3), were prominent at marked CREs. Surprisingly, we found
388 that over 40% of significant eCREs and their associated LEVs lacked classical epigenetic
389 marks almost entirely (Fig. 5A). This suggests the pervasive existence of UREs across
390 the whole genome that may be driven by specific biological conditions.

391

392 To demonstrate the regulatory potential of URE in gene regulation, we performed several
393 functional assays on rs73156934-tagged CRE which showed a significant effect in
394 CRISPRi screen. rs73156934 is an intronic variant at *EXOC4* gene, and its surrounding
395 genomic region is not marked by any epigenetic signals in HAP1 (Fig. 5B) or rarely
396 occupied with H3K36me3 in other tissues/cell types (query from VannoPortal (39)),

397 suggesting the regulatory activity of the rs73156934-tagged CRE is preserved in most
398 cellular contexts. Compared to the candidate CREs tagged by other highly linked LEVs
399 (e.g. rs10428917) in pLD region, our CRISPRi perturbation screen and RT-qPCR
400 revealed that this URE could be manipulated to regulate a distal gene *CALD1* (800 Kb)
401 instead of its associated eGene *EXOC4* in GTEx (Fig. 5C and Fig. 5D). The long-
402 distance interactions between the eCRE and *CALD1* were further confirmed by 4C-seq
403 anchored at rs73156934-containing region (Fig. 5B). In addition, luciferase reporter
404 assays indicated that the eCRE had significant regulatory functions and was affected by
405 different alleles of rs73156934 (Fig. 5E). Moreover, the regulatory relationship between
406 rs73156934 and its target gene *CALD1* was supported by several blood single-cell eQTL
407 studies (Fig. 5B).

408

409

410 **Combinatory analysis of CRISPRi/a effects suggests regulatory plasticity of the**
411 **human genome**

412 Applying both CRISPRi and CRISPRa to the same sgRNA library in the multiplexed
413 single-cell screens enabled us to systematically compare the regulatory effects by
414 different types of perturbations. We found that a large number of perturbations
415 unexpectedly affected their target gene expression in proximity (± 1 Mb surrounding TSS),
416 and the casual eCREs received varied regulatory outcomes under the same perturbation
417 (Fig. 6A). Specifically, CRISPRi can up-regulate nearby target gene expression among
418 one-third of the significant eCRE-perturbGene pairs, and CRISPRa also could down-
419 regulate local gene expression occasionally, although we cannot figure out which hits are
420 from *trans*-effects of perturbation. Besides, over 35% of casual eCRE showed opposite
421 effects on different target genes through either CRISPRi or CRISPRa perturbations.
422 Interestingly, we also revealed 13 eCRE-perturbGene pairs were significant in both
423 CRISPRi and CRISPRa screens (Fig. 6B).

424

425 By partition all significant eCRE-perturbGene pairs into four categories according to the
426 effect direction of perturbGenes in CRISPRi/a screens (Fig. 6A), we observed that, for
427 the pairs whose gene expression were unexpectedly up-regulated via CRISPRi
428 perturbations (group I and group IV), the distances of casual eCRE from target gene TSS
429 displayed a polarized trend, and most of which were more than 100 Kb away from the
430 TSSs (Fig. 6C). Such phenomenon was also observed for those repressive effects from
431 CRISPRa perturbations (group III) (Fig. 6C). While some effects could be attributed to
432 *trans*-gene regulation, these observations highlight the existence of several hidden but
433 unique mechanisms underlying the distal gene regulation in 3D genome (Fig. 6C).
434 However, the casual eCREs were almost located at the promoter region of their
435 perturbGenes for the 13 significant eCRE-perturbGene pairs identified by both CRISPRi
436 and CRISPRa (Fig. 6C). Besides, compared with casual eCREs detected in single type
437 of perturbation or insignificant ones, more active chromatin signals (including H3K27ac,
438 H3K4me1, H3K4me3 and open chromatin) were enriched at casual eCREs identified by
439 both perturbations. (Fig. 6D). Together, these results indicate that the human genome
440 and chromatin display high plasticity in response to various stimulations at different

441 genomic positions, warranting further research to elucidate the role of sequence
442 variations and chromatin dynamics in shaping functional molecular phenotypes.

443

444

445 **Discussion**

446 Investigating the molecular phenotypes influenced by genetic loci that subsequently lead
447 to the development of complex traits and diseases is a crucial scientific question in the
448 current field of genetics research (3). Gene expression, as an essential molecular
449 process for transmitting genetic effects, has made the integrated analysis of eQTL a
450 standard approach to exploring the causal genetic mechanisms of complex traits and
451 diseases (40, 41). However, challenges persist in identifying the true causal regulatory
452 variants and their target genes in highly linked regions due to factors such as LD
453 contamination and the complexity of CRE-target gene regulation. In this study, we
454 innovatively employed multiplexed single-cell CRISPRi/a perturbations to investigate the
455 regulatory patterns of genetic loci on target genes in endogenous cellular environments
456 and under pLD conditions. We systematically demonstrated the widespread presence of
457 multiple causal regulatory loci within pLD regions and the intricate nature of their
458 regulatory interactions with target genes. Additionally, our findings revealed that
459 endogenous perturbations can unveil elusive genetic effects and gene expression
460 regulatory patterns not easily detected by traditional eQTL mapping, including evidence
461 for long-range regulatory relationships and high-resolution analysis of regulatory effects.
462 Furthermore, we identified that existing computational methods face difficulties in
463 precisely predicting the influences and consequences of numerous endogenous
464 perturbations on gene expression. We also discovered that approximately half of the
465 causal eCREs lack conventional epigenetic markers, potentially affecting gene
466 expression via distinct regulatory mechanisms. Lastly, through a comparative analysis of
467 CRISPRi/a perturbation effects, we expounded upon the regulatory plasticity of the
468 human genome from a novel perspective. We propose that incorporating multiplexed
469 single-cell CRISPR perturbations into molecular trait QTL and genome-wide GWAS
470 causal variant fine-mapping could complement the limitations of traditional diverse MPRA
471 approaches in assessing the magnitude of genetic effects in endogenous chromatin
472 environments, and their target genes (8-10). Our novel findings and supporting evidence
473 will also promote the development of new technologies and theories in functional
474 genomics and related computational biology, ultimately leading to a more comprehensive
475 understanding of gene expression regulatory patterns.

476

477 The functional evaluation of the genetic effects for GWAS/eQTL causal regulatory
478 variants by *in situ* modulation of the genomic sequence under endogenous chromatin
479 environments remains challenging. First, unbiased high-throughput screening at single-
480 base resolution remains limited due to the characteristics of editing technologies. For
481 instance, while cytosine base editors (CBEs) and adenine base editors (ABEs) have
482 been widely employed for screening of allele effects under complex phenotypes or at the
483 single-cell level (11, 12, 42, 43), their base editing types and uncertain editing outcomes
484 within the editing window hinder their application in functional fine-mapping studies.

485 Precise editing tools, such as prime editing (PE) and other retron-based systems (13, 44,
486 45), have been used for high-throughput screening of functional single-base variants, but
487 their relatively low editing efficiency in mammalian cells significantly restricts large-scale
488 genome-wide screening. Second, most of the current Perturb-seq-based functional
489 genomic effect assessments are based on diploid cells with heterozygous genetic
490 background (21, 46), such as K562 cells. However, in polyploid mammalian cells,
491 incomplete editing or interference of other non-homozygous alleles may mask the
492 expected genetic effects and phenotypes, potentially leading to reduced statistical power.
493 To balance the advantages and disadvantages of existing technologies, we combined
494 CRISPRi and CRISPRa to perform multiplexed single-cell perturbation screening on
495 diploid HAP1 cells, which have a relatively homozygous genetic background.
496

497 Previous high-density GWAS, WGS-based eQTL studies, and simulation analyses have
498 indicated that multiple causal effects within specific genomic loci are not uncommon, and
499 multi-target regulation, along with LD contamination, further complicates the fine-mapping
500 of true causal variants and estimation of true effect sizes (2, 4, 47). By merely interfering
501 with a large number of potentially independent eQTL regions harboring multiple causal
502 effects using CRISPRi/a, we discovered that 19.3% of these regions contain multiple
503 causal eCREs targeting the same target gene. Consistent with previous MPRA in high LD
504 (17.7% of eQTLs exhibit more than one major allelic effect) (10), our results confirm the
505 widespread presence of multisite regulation of gene expression under endogenous
506 genetic systems, emphasizing the serious need to consider such situations when fine-
507 mapping causal variants in different LD regions. Additionally, through CRISPRi/a
508 screening, we found that causal eCRE(s) within the majority of pLD signals (around 80%)
509 can distally impact the expression levels of multiple target genes and are not necessarily
510 always the nearest genes. Compared to traditional eQTL mapping, which more readily
511 detects signals near TSS (5, 6), single-cell CRISPR-based QTL mapping approach may
512 be better suited for interpreting missed signals in GWAS-eQTL colocalization studies.
513 Consequently, the systematic integration of large-scale single-cell CRISPR-based eQTLs
514 and traditional context-specific eQTLs will further unravel the 'missing regulation'
515 phenomenon in non-coding regions of GWAS (7).
516

517 Fine-mapping computational methods based on functional annotation have been widely
518 used to explore GWAS/QTL causal variants and their potential functions (48). However,
519 both our work and several current studies have observed that existing computational
520 methods and bioinformatics tools struggle to accurately predict the functional
521 consequences (e.g., expression effects) of a given DNA sequence or variant allele (10,
522 32), and the reasons for these inconsistencies remain unexplained. The most plausible
523 explanation might be that the non-coding regulatory features used in current prediction
524 models are largely similar, lacking a novel perspective on how regulatory sequences
525 exert their functions; moreover, the functional genomic features used for computational
526 modeling are predominantly measured in normal tissue or cellular environments, with
527 insufficient data under diverse biological conditions. By integrating epigenetic information,
528 we have revealed that approximately half of the significant causal eCRE genomic regions

529 are severely lacking traditional epigenetic markers (referred to as UREs), and some loci
530 even lack epigenetic markers across all known biological conditions. UREs have been
531 considered functional in previous GWAS and functional genomic studies (9, 15, 16, 49,
532 50), which drives researchers to further investigate universal regulatory mechanisms to
533 explain their regulatory potential, and provides a direction for improving the performance
534 of prediction methods.

535

536 In various types of perturbations, we found that CRISPRi has better effects than
537 CRISPRa in driving changes in gene expression. Recent research has discovered that
538 the efficacy of CRISPRa depends on basal expression and chromatin state, and bivalent
539 genes are more sensitive to this perturbation (51). However, CRISPRi may be more
540 capable of modifying and altering chromatin structure, inducing heterochromatin
541 formation in any genomic region where transcription occurs, thereby causing changes in
542 the expression of multiple genes in close chromosomal proximity. Therefore, introducing
543 a combination of dCas9 fused to different activation domains, such as enCRISPRa
544 technology (50), may enhance the power of functional eCRE detection. On the other
545 hand, we found that a portion of CRISPRi/a perturbations exhibited opposite trends to the
546 expected gene expression disruption effects. Presumably, the effects of CRISPRi and
547 CRISPRa on gene expression can depend on the specific regulatory elements and
548 factors present within the targeted genomic region, such as the location and orientation
549 of the targeted site relative to regulatory elements, the activity and accessibility of
550 chromatin and epigenetic modifiers, the availability and activity of other transcriptional
551 regulators, as well as the competition among different CREs (18, 52, 53). For example, a
552 genomic region containing both an enhancer and a silencer element might exhibit
553 different responses to CRISPRi and CRISPRa depending on which element is targeted.
554 Additionally, some regions may contain a composite enhancer/silencer element that is
555 responsive to both positive and negative regulatory signals, or they may incorporate
556 multiple distinct CREs that are differentially responsive to transcriptional activators and
557 repressors.

558

559 Our study has the following limitations and unresolved issues. First, in order to select
560 reproducible independent eQTL signals, we systematically integrated seven blood-
561 derived WGS-based eQTL datasets and used LD data from 1000 Genomes Project
562 European population to screen for associated pLD LEVs. Although most of the used
563 eQTL samples have European ancestry and were derived from blood tissues, factors
564 such as heterogeneity of WGS variant calling, discrepancies in LD structure among
565 subpopulations, and differences in blood cell-specific gene expression levels may lead to
566 incomplete selection of pLD LEVs in each independent eQTL signal. Second, as
567 emphasized before, we still have difficulty in screening the genetic effects of causal
568 variants on specific phenotypes at the single-base level in a high-throughput and
569 unbiased manner. CRISPRi/a-based genomic perturbation cannot accurately assess the
570 genetic effects at the variant and allelic levels, nor can it distinguish between closely
571 located genetic variations (e.g., less than 1kb). Additionally, some potential functional
572 CREs cannot be driven by CRISPRi/a. Therefore, we need to develop efficient single-cell

573 single-base perturbation technologies to accurately measure the true effects of genetic
574 variations on gene expression. Finally, this experiment only evaluates the effects of
575 genetic loci using limited gene expression as the readout due to genomic distance and
576 expression level constraints. However, many functional regulatory variations may affect
577 gene expression levels through moderate or *trans*-effects or directly influence other
578 transcriptional-level molecular phenotypes without changing gene expression conditions.
579 Despite these limitations, our study reveals several unique patterns for the complexity of
580 gene expression regulation.

581

582

583 **Methods**

584

585 **Plasmids**

586 The dCas9-VP64-blast (Addgene, #61425) and dCas9-KRAB-blast (Addgene, #89567)
587 plasmids were separately used to perform CRISPRi and CRISPRa experiments. The
588 sgRNA oligos were annealed and cloned into the CROP-seq-opti plasmid (Addgene,
589 #106280) after *BsmBI* (NEB, R0580L) digestion. The sgRNA library was synthesized by
590 Synbio Technologies. The psPAX2 (Addgene, #12260) and pMD2.G (Addgene, #12259)
591 were used to pack lentiviruses. The plasmids were used for dual-luciferase reporter
592 assay, including the pRL-TK Renilla luciferase control vector (Promega, E2241), pGL3-
593 Promoter (Promega, E1761), and pGL3-basic (Promega, E1741).

594

595 **Cell lines and cell culture**

596 293FT (ThermoFisher, R70007) cells were cultured in Dulbecco's Modified Eagle's
597 Medium (DMEM; ThermoFisher, 11965092) containing 10% FBS. HAP1 (Horizon
598 Discovery) cells were cultured in Iscove's Modified Eagle's Medium (IMDM;
599 ThermoFisher, 31980030). Both cells were supplemented with 10% fetal bovine serum in
600 5% CO₂ at 37°C. Diploid HAP1 cells were isolated from Hoechst (MCE, Y-15559)-stained
601 HAP1 cells by flow sorting.

602

603 **WGS-based blood eQTL fine-mapping**

604 Seven WGS-based blood-derived eQTL datasets were used to identify independent
605 eQTL signals and fine-mapped LEVs based on CaVEMaN method (54). Specifically, the
606 fine-mapped eQTL variants for GTEx V8 whole blood eQTLs (670 whole blood samples)
607 were obtained from GTEx portal (2), and the fine-mapped eQTL variants for several other
608 cohorts with blood-derived samples, including Geuvadis (445 lymphoblastoid cell
609 samples), and TwinsUK (523 lymphoblastoid cell samples and 246 whole blood samples),
610 were obtained from the original CaVEMaN publication (54). Additionally, for BLUEPRINT
611 samples (190 monocyte samples, 196 neutrophil samples and 165 CD4+ T cell samples),
612 we conducted eQTL mapping using FastQTL (55) based on the individual WGS
613 genotypes and normalized RNA-seq quantifications. Permutation test was applied to
614 estimate the nominal *P* thresholds required for the conditional analysis, Then, we used
615 CaVEMaN to perform eQTL fine-mapping for each significant eQTL signal, and extracted
616 the best eQTL that was most likely to be causal.

617

618 **Nomination of reproducible pLD signals with multiple LEVs**

619 Based on eQTL fine-mapping results from GTEx V8 whole blood samples, we sought to
620 identify independent eQTL signals that are reproducible in at least one additional blood-
621 derived eQTL dataset. Generally, we measured LD using genotypes of European
622 samples from 1000 Genomes Project phase3 (56) and searched perfectly correlated
623 LEVs ($R^2 = 1$) with each GTEx LEV among fine-mapped eQTL signals of other blood-
624 derived datasets. Ultimately, we integrated all fine-mapped variants in seven eQTL
625 datasets to retain independent signals which contains at least two undistinguishable
626 LEVs in pLD. To ensure that the genes being tested were measurable in 10x scRNA-seq,

627 we only included top 20% expressed genes in HAP1 cells. Then, LEVs located in protein-
628 coding/splicing-altering region of the human genome or in the chr15 diploid region of
629 HAP1 cell line were excluded. Due to the limitations of CRISPRi/a in the scope of
630 genomic targeting, we only retained signals with a distance greater than 1 kb between
631 each pairwise LEVs. Thus, pLD signals that were distant from these highly expressed
632 genes (> 1 Mb) in HAP1 cells were also excluded. All of these analyses, including the
633 following, were based on the human genome assembly GRCh37/hg19.

634

635 **sgRNA library design**

636 We used FlashFry 1.9.3 (57) to design sgRNAs targeting each eCRE. We excluded
637 sgRNAs whose splicing sites located more than 50 bp away from the eCRE-associated
638 LEVs and with IN_GENOME \geq 2. Then we ranked the remaining sgRNAs based on
639 Doench2014OnTarget, Hsu2013, and Doench2016CDFScore, as well as otCount. We
640 selected the top two sgRNAs targeting each eCRE (except for cases where only 1 or no
641 sgRNA met the criteria). To include appropriate controls, we incorporated 39 non-
642 targeting control gRNAs from the Human CRISPR Knockout Pooled Library (GeCKO V2)
643 (58) as negative controls and nine sgRNAs targeting promoters of 4 genes (*EZH2* (59),
644 *CANX* (60), *NEAT1* (61), *PARK7* (20), among which *EZH2* was targeted by three sgRNAs)
645 and two sgRNAs targeting an enhancer of *NMU* (60) from different studies as positive
646 controls. The oligos of the sgRNA library were synthesized and cloned to the CROP-seq-
647 opti vector after *BsmBI* digestion by Synbio Technologies, according to the GeCKO V2
648 (58). Additionally, sgRNAs targeting the same eCRE or positive control site are referred
649 to as a "sgRNA group". In the following bioinformatics analysis, all sgRNA groups that
650 target eCRE are referred to as "perturbative sgRNA groups", whereas all other sgRNA
651 groups are referred to as "control sgRNA groups".

652

653 **Quality control of synthetic sgRNA library**

654 To assess the quality and potential bias of the sgRNA library, the sgRNA sequences were
655 amplified using PCR from either the plasmid library or genomic DNA extracted from
656 HAP1 cells 4 days post-transduction, using qsgRNA-F and qsgRNA-R primers and 2x
657 Phanta Max Master Mix (Vazyme, P515-01). The resulting PCR products were purified
658 using the QIAquick PCR Purification Kit (Qiagen, 28106) and then used to generate a
659 next-generation library using the VAHTS Universal DNA Library Prep Kit for Illumina V3
660 (Vazyme, N607-01). The library was purified using AMPure XP beads (Beckman Coulter,
661 A63880) and sequenced on an Illumina NovaSeq PE150. The sgRNAs were identified by
662 matching the sequence to "CACCG[sgRNA]GTTT" and compared to the designed
663 sgRNAs to determine the correct rate of the sgRNA plasmid library. The potential bias
664 was evaluated by calculating the ratio of 90th percentile to 10th percentile sgRNAs that
665 had at least one sequencing read.

666

667 **Production of lentivirus**

668 293FT cells were seeded 24h prior to lentivirus packaging. The lentivirus was produced
669 by co-transfecting the backbone plasmid with viral packaging plasmid (psPAX2) and viral
670 envelope plasmid (pMD2.G) at a ratio of 4:3:1 into 293FT cells using LipoFiter (Hanbio,

671 HB-LF-1000) according to the manufacturer's instructions. The cell culture supernatant
672 was collected 48h post-transfection and filtered using a 0.45 μ m filter.

673

674 **Construction of dCas9-KRAB and dCas9-VP64 stably expressed HAP1 cells**

675 Lentivirus containing dCas9-KRAB-blast and dCas9-VP64-blast were separately used to
676 construct stably-expressing HAP1 cells. Diploid HAP1 cells were seeded into a six-hole
677 plate supplemented with 8 μ g/mL polybrene (Beyotime, ST1380) 24h prior to lentivirus
678 infection. The lentivirus was added to the cells and 24 hours post-infection, blasticidin
679 (ThermoFisher, 461120) was added to the culture supernatant to a final concentration of
680 10 μ g/mL. Selection was maintained for 3 days to obtain stably-expressing cell lines. The
681 Anti-CAS9 Antibody (BOSTER, BM5120) and Anti- β -Actin antibody (ABclonal, AC02)
682 were used to verify the expression of the dCas9-KRAB and dCas9-VP64 by Western blot.
683 The stably-expressing cell lines were then plated in a 96-well plate by limiting dilution and
684 cultured for 2 weeks to obtain single clones. The activation and inhibition efficiency of
685 single clones were verified by infection with lentivirus containing sgRNA targeting the
686 TSS of *EZH2*. The two most efficient clones of each cell line were selected for
687 perturbation.

688

689 **Infection of lentivirus with different MOIs**

690 The lentivirus supernatant of the sgRNA library was concentrated using the ViraTrapTM
691 Lentivirus Concentration Reagent (Biomiga, BW-V2001-01) and titrated using the Lenti-
692 PacTM HIV qRT-PCR Titration Kit (GeneCopoeia, LT005). The dCas9-KRAB and dCas9-
693 VP64 stably expressed HAP1 cell lines were seeded into a 24-well plate and transfected
694 with the concentrated sgRNA library at MOI=500 (moderate MOI). To increase the
695 infection efficiency, 8 μ g/mL polybrene was added to the cell culture. After 24 hours, the
696 cells were treated with 0.3 μ g/mL puromycin (Sigma-Aldrich, P7255) for three days. For
697 high MOI infections, a second round of infection was performed to achieve a higher MOI.

698

699 **scRNA-seq and sgRNA-transcript enrichment**

700 To prepare multiplexed CROP-seq libraries, adherent cells were digested with 0.25%
701 trypsin and collected in a 15mL tube containing serum-containing medium, followed by
702 centrifugation and washing of the cell pellet with serum-free basal medium to obtain a
703 single-cell suspension. The cell density was adjusted with a cell counter and the 10x
704 Genomics Chromium Single Cell 3' Library reagents V3 were used according the
705 manufacturer's instructions. To enrich for sgRNA-transcripts, PCR was performed on 15
706 ng of cDNA from the 3' single-cell RNA libraries using SI-PCR primer and 10x-sgRNA i7-
707 N720 primer in each 50 μ L reaction with an annealing temperature of 60°C and 2 \times
708 Phanta Max Master Mix. The enriched sgRNA libraries were sequenced on Illumina
709 NovaSeq6000 PE150 with the same configuration as the standard 10x libraries.

710

711 **scRNA-seq data processing**

712 The sequencing data from 10x Genomics Chromium 3' scRNA-seq underwent initial
713 processing with Cell Ranger v5.0.1, which involved sequence alignment, filtering,
714 barcode counting, and UMI counting. The resulting data were further analyzed using

715 Seurat v4.0.3 (62), where a series of quality control measures were applied. Specifically,
716 cells with mitochondrial percentage exceeding 10% or having less than 200 gene UMIs
717 were removed. Additionally, genes that were expressed in less than 0.525% of cells were
718 filtered out. Droplets were also identified and removed using scDblFinder (63).

719

720 **sgRNA and single cells assignment**

721 With the fore-mentioned amplification protocol (25), we enriched sgRNAs and then
722 calculated the distribution of sgRNAs in different perturbation. Firstly, we aligned the
723 enrichment reads to reference genome (GRCh37/hg19) using Cell Ranger v5.0.1,
724 following the same procedure as for scRNA-seq data processing. We then extracted the
725 reads that failed to align to the reference genome using Samtools (64). Next, we mapped
726 each unmapped read to the sequences of the sgRNA library and its reverse complement.
727 To limit the sequence before and after the sgRNA sequence, we used ACCG as the left
728 anchor (which is at the end of the U6 promoter) and GTTT as the right anchor. During the
729 sgRNA detection process, we also extracted the cell barcode and sgRNA UMI count for
730 each perturbed cell. The number of sgRNAs per cell and the number of cells bearing
731 each perturbation was calculated based on data from multiplexed single-cell CRISPRi
732 and CRISPRa perturbations with different MOIs.

733

734 **eCRE-gene association analysis**

735 For each targeted eCRE, we divided cells into two groups based on whether they contain
736 specific sgRNAs targeting observed eCRE, and evaluated the differential expression of
737 candidate genes within 1 Mb of the eCRE between perturbative sgRNA group and control
738 sgRNA group using Normalisr (26). Normalisr is a unified normalization-association
739 framework for statistical inference of gene regulation. It uses an ingenious normalization
740 strategy followed by a regular linear regression model. The normalization step estimates
741 the pre-measurement mRNA frequencies from the scRNA-seq UMI counts and regresses
742 out the nonlinear effect of library size on expression variance. And then a linear
743 regression model tests the associations between eCRE and perturbGene, where two-
744 sided P were computed from Beta distribution, and log fold change was estimated using
745 maximum likelihood. We also included additional covariates including mitochondrial
746 percentage, unique gene count, and sgRNA count in the association testing to improve
747 the accuracy of the results. To avoid false positive results, we estimated the P of the
748 associations between genes and NTC sgRNAs and considered them background P . An
749 empirical P was calculated based on the background P and raw P , which was then
750 adjusted by FDR (q-value) for multiple testing corrections. Finally, we defined a 0.2
751 threshold of q-value based on the NTC tests as they are subject to the same sources of
752 error as the eCRE-targeting sgRNAs.

753

754 **Simulation for power estimation**

755 Based on Splatter (65), we simulated synthetic scRNA-seq datasets with various
756 group.prob parameters to mimic CROP-seq cells under different perturbation conditions.
757 The Splatter simulation process consisted of two steps. First, we estimated the

758 necessary parameters for simulation using *splatEstimate* from the CROP-seq dataset.
759 Subsequently, we utilized the estimated parameters to simulate synthetic scRNA-seq
760 datasets with *splatSimulate*, where the average expression of each gene was randomly
761 sampled from a gamma distribution and the cell's experimentally measured count was
762 sampled from a Poisson distribution. To simulate perturbation under different MOIs, we
763 used the *group.prob* parameter, which is a vector containing two values. The first value
764 was obtained by dividing the number of perturbations per cell (ranging from 200 to 3,800)
765 by the total number of cells, and the second value was obtained by subtracting the first
766 value from 1. We used *nCells* parameter to simulate perturbation under different cells
767 (ranging from 5,000 to 20,000). Finally, we assessed the power of each simulation by
768 computing the ratio of recovered CROP-seq results to all CROP-seq results.

769 **Validation of individual hits**

770 For validation of positive eCRE-perturbGene pairs identified by multiplexed CRISPRi/a
771 perturbations, the expression of perturbGenes was measured by RT-qPCR following
772 individual perturbation with sgRNAs targeting the eCREs. Briefly, oligonucleotides of
773 sgRNAs targeting causal eCREs and random sgNTC were synthesized and annealed,
774 and then cloned into CROP-seq-opti plasmids that were digested by *Bsm*BI. The
775 resulting plasmids containing sgRNAs targeting the same eCRE were mixed in equal
776 amounts and packaged into lentiviruses. The dCas9-KRAB and dCas9-VP64 stably
777 expressed HAP1 cell lines were infected with the lentiviruses and selected with
778 puromycin. RNA was extracted from cells, and cDNA was generated using HiScript II Q
779 Select RT SuperMix for qPCR (+gDNA wiper) (Vazyme, R223-01). The cDNA was
780 amplified with 2x SYBR Green qPCR Master Mix (Bimake, B21202) using ACTB as a
781 positive control. Data were analyzed using the $2^{-\Delta\Delta Ct}$ method, with sgNTC as the control.
782 The effect on gene expression of sgRNAs targeting eCREs was calculated after
783 normalization.

784

785 **Luciferase reporter assay**

786 Genomic sequences containing causal eCREs with different alleles of LEVs were
787 amplified from HAP1 cell genomic DNA using overlapping PCR. The resulting fragments
788 were integrated upstream of the luciferase gene in pGL3-Basic and pGL3-Promoter
789 plasmids, and the concentration of the recombinant plasmids was determined using the
790 Qubit dsDNA HS Assay Kit (ThermoFisher, Q32851). 293FT cells were transfected with 1
791 μ g of recombinant plasmids or blank vectors with 40 ng of pRL-TK Renilla luciferase
792 control vector using Lipofiter. After 24 h, cells were lysed with 200 μ L lysis buffer and
793 shaken slowly on ice for 10 min. Relative luminescence signals were measured using the
794 Dual-Luciferase Reporter Assay System (Promega, E1960) and GloMax® 20/20
795 Luminometer by normalizing firefly luciferase signal with renilla luciferase signal.

796

797 **Circularized chromosome conformation capture (4C) assay**

798 We used the 4C-seq method as described previously (66) to validate eCREs tagged by
799 specific LEVs. To prepare the 4C-seq samples, approximately 1×10^7 HAP1 cells was

800 collected to crosslink by formaldehyde for 10min, and quenched with glycine at a final
801 concentration of 125 mM. Cell pellets was washed twice with cold PBS and lysis on ice
802 for 15 min with 5 mL lysis buffer. The nuclei were digested with the first enzyme (CvQI
803 (NEB, R0639L) or *N*laIII (NEB, R0125S)) and ligated with T4 ligase (NEB, M0202)
804 overnight at 16°C. Proteinase K (Transgene, 20 mg/ml) was added to the sample, which
805 was then placed in a 65°C water bath overnight to reverse the cross-links. After de-
806 crosslinking, DNA was extracted using phenol-chloroform-isoamylalcohol (Invitrogen,
807 25:24:1) and finally dissolved in 150 μ L of 10 mM Tris-HCl (pH 7.5). The sample was
808 digested with the second enzyme (*Eco*RI (NEB, R0101S) or *Dpn*II (NEB, R0543L)) and
809 ligated with T4 ligase (NEB, M0202) overnight at 16°C. DNA precipitation was obtained
810 by ethanol precipitation, and the DNA was dissolved in 150 μ L of 10 mM Tris-HCl (pH
811 7.5). The 4C template was obtained by adding 750 μ L of Buffer PB solution (QIAGEN,
812 #28106) to the sample and dividing it into three spin columns for centrifugation and
813 washing. Each spin column was washed with 50 μ L of 10 mM Tris-HCl (pH 7.5). The
814 eluates from the three tubes were collected as 4C template. Viewpoint-specific
815 amplification was performed with 8 \times 200 ng 4C template using 2 \times Phanta Max Master
816 Mix. The next-generation library was constructed using the VAHTS Universal DNA Library
817 Prep Kit for Illumina V3 and purified by AMPure XP beads following the manufacturer's
818 instructions. The 4C-seq library were sequenced on the Illumina NovaSeq 6000 platform,
819 producing pair-end reads of 150 bp. Processing and visualization of 4C-seq data was
820 done using pipe4C and peakC pipeline as previously described (66, 67).

821

822 **Measurement of eCRE regulatory range across LD block and TAD**

823 We defined the LD blocks as multiple variants in strong LD across a region. Specifically,
824 we measured LD between variants surrounding each fine-mapped signals (500 Kb)
825 based on genotypes of Europeans from 1000 Genomes Project phase3 (56), using the
826 LDproxy module of LDlink (68). We then merged all genomic positions for variants in
827 strong LD with the fine-mapped eQTL LEVs ($R^2 \geq 0.8$) to form the LD blocks. To define
828 a gene-regulating region, we considered the region from the CRE (± 1 Kb of LEVs) to 5
829 Kb upstream and 3 Kb downstream of the gene body of the target gene. We obtained
830 TADs from *in situ* Hi-C of the HAP1 cell line (69). By constructing a 2 \times 2 contingency
831 table, we performed a Fisher's exact test to evaluate the nonrandom association between
832 the number of gene-regulating regions that lie in Hi-C TADs and those that lie in LD
833 blocks.

834

835 **Benchmarks with computational prediction scores**

836 We conducted an evaluation on perturbation effect of causal eCREs against various
837 computational methodologies for predicting base-wise regulatory potentials of DNA
838 sequence. Multiplexed single-cell CRISPRi/a perturbation results were used as the
839 golden standard to measure prediction performance. Positive samples were significant
840 causal eCRE-associated LEVs, while negative samples were non-significant ones. To
841 conduct a comprehensive evaluation of the performance of these algorithms, we used
842 the AUC metric of the ROC to distinguish positive and negative samples at both LEV-
843 level and eCRE-level. First, we obtained pre-computed base-wise scores of 20 existingg

844 computational methods from regBase V1.1.1 (34). Due to the class imbalance of positive
845 and negative samples, we randomly under sampled negative eCREs to match the
846 number of positive eCREs. We repeated each sampling ten times. Second, Enformer
847 (36), a deep neural network to predict gene expression levels given genomic sequence,
848 was employed in our comparisons. Two Enformer principal components (PCs),
849 representing a summary of the most important features that contribute to the prediction of
850 gene expression levels, were calculated and used to generate empirical cumulative
851 probability distributions of the first and second PC scores. By comparing these
852 distributions, we determined if the PCs for significant and non-significant eCRE-
853 associated LEVs differed. For eCRE-level benchmarks, we estimated median, mean, and
854 max scores for ± 1 Kb of each eCRE-associated LEVs.
855

856 **Epigenomics analysis**

857 The raw sequencing reads of HAP1 ChIP-seq (H3K27ac, H3K27me3, H3K36me3,
858 H3K4me1, H3K4me3, H3K9me3) and ATAC-seq were analyzed by the nf-core pipeline
859 (70). Meta-profiles of LEV-centered regions (± 2.5 Kb) were generated from the bigWig
860 files by deepTools (71). Heatmaps were generated using EnrichedHeatmap (72), and
861 narrow peaks were called using MACS2 (73). The causal eCREs that intersected with
862 peak(s) for at least one of fore-mentioned marks were classified as marked eCREs. In
863 contrast, the causal eCREs that do not intersect with peak(s) for any marks are referred
864 to as URE.
865

866 **Statistical analyses**

867 Statistical analyses were carried out with GraphPad Prism 8.0 (GraphPad Software). All
868 experiments were performed at least three replicates, unless otherwise noted.
869 Differences in means were compared using an unpaired two-tailed Student's t-test, and
870 graphed as the means \pm standard deviations (SD). Statistical significance denoted as
871 follows: *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$; ns, no significant.
872

873 **Acknowledgements**

874 The work was supported by the following grants: the National Natural Science
875 Foundation of China (32270717 and 32070675 to M.J.L.).
876

877 **Author contributions**

878 M.J.L. designed the studies and wrote the manuscript. K.Z., and Y.Z. performed the
879 experiments, analyzed data and wrote the manuscript. C.Y.W., J.H.W., H.C.Y., X.C., L.Z.,
880 W.W., X.L.C., and X.F.Y. conducted the bioinformatics analysis or valuation experiment.
881 Y.P.C., M.X.L., W.G.L., K.X.C., and P.C.S. contributed scientific expertise and reviewed
882 the manuscript. All the authors read and approved the manuscript.
883

884 **Competing interests**

885 The authors declare no competing interests.
886

887 **Data availability**

888 All sequencing data, including scRNA-seq and 4C-seq, generated in this study have
889 been deposited in the Gene Expression Omnibus (GEO) database under accession
890 GSEXXXX. All public data sources and primer sequences used in this study are listed in
891 Supplementary Tables. The average expression of each gene in HAP1 cells in this study
892 was calculated from several bulk RNA-seq data from GEO database under accession
893 number GSE75515, GSE110142, and GSE111272. TAD information of *in situ* Hi-C for
894 HAP1 cells was obtained from GEO under accession number GSE74072. The raw
895 sequencing reads of HAP1 ChIP-seq (H3K27ac, H3K27me3, H3K36me3, H3K4me1,
896 H3K4me3, H3K9me3) were downloaded from ENCODE repositories. Raw data of HAP1
897 ATAC-seq was downloaded from GEO under accession number GSE111047.

898

899

900 **Reference**

901 1. E. P. Consortium *et al.*, Expanded encyclopaedias of DNA elements in the human
902 and mouse genomes. *Nature* **583**, 699-710 (2020).

903 2. G. T. Consortium, The GTEx Consortium atlas of genetic regulatory effects across
904 human tissues. *Science* **369**, 1318-1330 (2020).

905 3. D. Huang *et al.*, QTLbase2: an enhanced catalog of human quantitative trait loci on
906 extensive molecular phenotypes. *Nucleic Acids Res* **51**, D1122-D1128 (2023).

907 4. B. Zeng *et al.*, Constraints on eQTL Fine Mapping in the Presence of Multisite Local
908 Regulation of Gene Expression. *G3 (Bethesda)* **7**, 2533-2544 (2017).

909 5. S. Chun *et al.*, Limited statistical evidence for shared genetic effects of eQTLs and
910 autoimmune-disease-associated loci in three major immune-cell types. *Nat Genet* **49**,
911 600-605 (2017).

912 6. H. Mostafavi, J. P. Spence, S. Naqvi, J. K. Pritchard, Limited overlap of eQTLs and
913 GWAS hits due to systematic differences in discovery. *bioRxiv*,
914 2022.2005.2007.491045 (2022).

915 7. N. J. Connally *et al.*, The missing link between genetic association and regulatory
916 function. *Elife* **11**, (2022).

917 8. J. C. Klein *et al.*, A systematic evaluation of the design and context dependencies of
918 massively parallel reporter assays. *Nat Methods* **17**, 1083-1091 (2020).

919 9. N. Rajagopal *et al.*, High-throughput mapping of regulatory DNA. *Nat Biotechnol* **34**,
920 167-174 (2016).

921 10. N. S. Abell *et al.*, Multiple causal variants underlie genetic associations in humans.

922 *Science* **375**, 1247-1254 (2022).

923 11. L. Cheng *et al.*, Single-nucleotide-level mapping of DNA regulatory elements that
924 control fetal hemoglobin expression. *Nat Genet* **53**, 869-880 (2021).

925 12. R. E. Hanna *et al.*, Massively parallel assessment of human variants with base editor
926 screens. *Cell* **184**, 1064-1080 e1020 (2021).

927 13. S. Erwood *et al.*, Saturation variant interpretation using CRISPR prime editing. *Nat
928 Biotechnol* **40**, 885-895 (2022).

929 14. M. Brandt, A. Gokden, M. Ziosi, T. Lappalainen, A polyclonal allelic expression assay
930 for detecting regulatory effects of transcript variants. *Genome Med* **12**, 79 (2020).

931 15. Y. Diao *et al.*, A tiling-deletion-based genetic screen for cis-regulatory element
932 identification in mammalian cells. *Nat Methods* **14**, 629-635 (2017).

933 16. D. R. Simeonov *et al.*, Discovery of stimulation-responsive immune enhancers with
934 CRISPR activation. *Nature* **549**, 111-115 (2017).

935 17. Z. Dai *et al.*, Inducible CRISPRa screen identifies putative enhancers. *J Genet
936 Genomics* **48**, 917-927 (2021).

937 18. S. K. Reilly *et al.*, Direct characterization of cis-regulatory elements and functional
938 dissection of complex genetic associations using HCR-FlowFISH. *Nat Genet* **53**,
939 1166-1176 (2021).

940 19. J. A. Morris *et al.*, Discovery of target genes and pathways of blood trait loci using
941 pooled CRISPR screens and single cell RNA sequencing. *bioRxiv*,
942 2021.2004.2007.438882 (2021).

943 20. Y. Pan, R. Tian, C. Lee, G. Bao, G. Gibson, Fine-mapping within eQTL credible

944 intervals by expression CROP-seq. *Biol Methods Protoc* **5**, bpaa008 (2020).

945 21. M. Gasperini *et al.*, A Genome-wide Framework for Mapping Gene Regulation via
946 Cellular Genetic Screens. *Cell* **176**, 377-390 e319 (2019).

947 22. A. J. Hill *et al.*, On the design of CRISPR-based single-cell molecular screens. *Nat
948 Methods* **15**, 271-274 (2018).

949 23. S. Xie, D. Armendariz, P. Zhou, J. Duan, G. C. Hon, Global Analysis of Enhancer
950 Targets Reveals Convergent Enhancer-Driven Regulatory Modules. *Cell Rep* **29**,
951 2570-2578 e2575 (2019).

952 24. D. Yao *et al.*, Compressed Perturb-seq: highly efficient screens for regulatory circuits
953 using random composite perturbations. *bioRxiv*, 2023.2001.2023.525200 (2023).

954 25. S. Xie, G. C. Hon, Experimental and Computational Approaches for Single-Cell
955 Enhancer Perturbation Assay. *Methods Mol Biol* **1935**, 203-221 (2019).

956 26. L. Wang, Single-cell normalization and association testing unifying CRISPR screen
957 and gene co-expression analyses with Normalisr. *Nature Communications* **12**, 6395
958 (2021).

959 27. M. H. Chen *et al.*, Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667
960 Individuals from 5 Global Populations. *Cell* **182**, 1198-1213 e1114 (2020).

961 28. M. van der Wijst *et al.*, The single-cell eQTLGen consortium. *Elife* **9**, (2020).

962 29. J. B. Veyrieras *et al.*, High-resolution mapping of expression-QTLs yields insight into
963 human gene regulation. *PLoS Genet* **4**, e1000214 (2008).

964 30. C. D. Brown, L. M. Mangravite, B. E. Engelhardt, Integrative modeling of eQTLs and
965 cis-regulatory elements suggests mechanisms underlying cell type specificity of

966 eQTLs. *PLoS Genet* **9**, e1003649 (2013).

967 31. S. Whalen, K. S. Pollard, Most chromatin interactions are not in linkage disequilibrium.
Genome Res **29**, 334-343 (2019).

969 32. M. Kircher *et al.*, Saturation mutagenesis of twenty disease-associated regulatory
970 elements at single base-pair resolution. *Nat Commun* **10**, 3583 (2019).

971 33. H. Yang *et al.*, De novo pattern discovery enables robust assessment of functional
972 consequences of non-coding variants. *Bioinformatics* **35**, 1453-1460 (2019).

973 34. S. Zhang *et al.*, regBase: whole genome base-wise aggregation and functional
974 prediction for human non-coding regulatory variants. *Nucleic Acids Res* **47**, e134
975 (2019).

976 35. I. Ionita-Laza, K. McCallum, B. Xu, J. D. Buxbaum, A spectral approach integrating
977 functional genomic annotations for coding and noncoding variants. *Nat Genet* **48**,
978 214-220 (2016).

979 36. Z. Avsec *et al.*, Effective gene expression prediction from sequence by integrating
980 long-range interactions. *Nat Methods* **18**, 1196-1203 (2021).

981 37. J. di Iulio *et al.*, The human noncoding genome defined by genetic diversity. *Nat
982 Genet* **50**, 333-337 (2018).

983 38. L. Przybyla, L. A. Gilbert, A new era in functional genomics screens. *Nat Rev Genet*
984 **23**, 89-103 (2022).

985 39. D. Huang *et al.*, VannoPortal: multiscale functional annotation of human genetic
986 variants for interrogating molecular mechanism of traits and diseases. *Nucleic Acids
987 Res* **50**, D1408-D1416 (2022).

988 40. E. Cano-Gamez, G. Trynka, From GWAS to Function: Using Functional Genomics to
989 Identify the Mechanisms Underlying Complex Diseases. *Front Genet* **11**, 424 (2020).

990 41. N. Kerimov *et al.*, A compendium of uniformly processed human gene expression and
991 splicing quantitative trait loci. *Nat Genet* **53**, 1290-1299 (2021).

992 42. R. Cuella-Martin *et al.*, Functional interrogation of DNA damage response variants
993 with base editing screens. *Cell* **184**, 1081-1097 e1019 (2021).

994 43. S. Jun, H. Lim, H. Chun, J. H. Lee, D. Bang, Single-cell analysis of a mutant library
995 generated using CRISPR-guided deaminase in human melanoma cells. *Commun Biol*
996 **3**, 154 (2020).

997 44. E. Sharon *et al.*, Functional Genetic Variants Revealed by Massively Parallel Precise
998 Genome Editing. *Cell* **175**, 544-557 e516 (2018).

999 45. M. G. Schubert *et al.*, High-throughput functional variant screens via in vivo
1000 production of single-stranded DNA. *Proc Natl Acad Sci U S A* **118**, (2021).

1001 46. N. K. Tuano *et al.*, CRISPR screens identify gene targets at breast cancer risk loci.
1002 *Genome Biol* **24**, 59 (2023).

1003 47. H. Huang *et al.*, Fine-mapping inflammatory bowel disease loci to single-variant
1004 resolution. *Nature* **547**, 173-178 (2017).

1005 48. D. J. Schaid, W. Chen, N. B. Larson, From genome-wide associations to candidate
1006 causal variants by statistical fine-mapping. *Nat Rev Genet* **19**, 491-504 (2018).

1007 49. B. Ding *et al.*, Noncoding loci without epigenomic signals can be essential for
1008 maintaining global chromatin organization and cell viability. *Sci Adv* **7**, eabi6020
1009 (2021).

1010 50. M. C. Canver *et al.*, A saturating mutagenesis CRISPR-Cas9-mediated functional
1011 genomic screen identifies cis- and trans-regulatory elements of Oct4 in murine ESCs.
1012 *J Biol Chem* **295**, 15797-15809 (2020).

1013 51. Q. Wu *et al.*, Massively parallel characterization of CRISPR activator efficacy in
1014 human induced pluripotent stem cells and neurons. *Mol Cell* **83**, 1125-1139 e1128
1015 (2023).

1016 52. C. P. Fulco *et al.*, Activity-by-contact model of enhancer-promoter regulation from
1017 thousands of CRISPR perturbations. *Nat Genet* **51**, 1664-1669 (2019).

1018 53. R. Tian *et al.*, Genome-wide CRISPRi/a screens in human neurons link lysosomal
1019 failure to ferroptosis. *Nat Neurosci* **24**, 1020-1034 (2021).

1020 54. A. A. Brown *et al.*, Predicting causal variants affecting expression by using whole-
1021 genome sequencing and RNA-seq from multiple human tissues. *Nat Genet* **49**, 1747-
1022 1751 (2017).

1023 55. H. Ongen, A. Buil, A. A. Brown, E. T. Dermitzakis, O. Delaneau, Fast and efficient
1024 QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479-1485
1025 (2016).

1026 56. C. Genomes Project *et al.*, A global reference for human genetic variation. *Nature*
1027 **526**, 68-74 (2015).

1028 57. A. McKenna, J. Shendure, FlashFry: a fast and flexible tool for large-scale CRISPR
1029 target design. *BMC Biol* **16**, 74 (2018).

1030 58. N. E. Sanjana, O. Shalem, F. Zhang, Improved vectors and genome-wide libraries for
1031 CRISPR screening. *Nat Methods* **11**, 783-784 (2014).

1032 59. S. Konermann *et al.*, Genome-scale transcriptional activation by an engineered
1033 CRISPR-Cas9 complex. *Nature* **517**, 583-588 (2015).

1034 60. N. C. Yeo *et al.*, An enhanced CRISPR repressor for targeted mammalian gene
1035 regulation. *Nat Methods* **15**, 611-616 (2018).

1036 61. T. Yamazaki, C. Fujikawa, A. Kubota, A. Takahashi, T. Hirose, CRISPRa-mediated
1037 NEAT1 lncRNA upregulation induces formation of intact paraspeckles. *Biochem
1038 Biophys Res Commun* **504**, 218-224 (2018).

1039 62. Y. Hao *et al.*, Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587
1040 e3529 (2021).

1041 63. P. L. Germain, A. Lun, C. Garcia Meixide, W. Macnair, M. D. Robinson, Doublet
1042 identification in single-cell sequencing data using scDblFinder. *F1000Res* **10**, 979
1043 (2021).

1044 64. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,
1045 2078-2079 (2009).

1046 65. L. Zappia, B. Phipson, A. Oshlack, Splatter: simulation of single-cell RNA sequencing
1047 data. *Genome Biology* **18**, 174 (2017).

1048 66. P. H. L. Krijger, G. Geeven, V. Bianchi, C. R. E. Hilvering, W. de Laat, 4C-seq from
1049 beginning to end: A detailed protocol for sample preparation and data analysis.
1050 *Methods* **170**, 17-32 (2020).

1051 67. G. Geeven, H. Teunissen, W. de Laat, E. de Wit, peakC: a flexible, non-parametric
1052 peak calling package for 4C and Capture-C data. *Nucleic Acids Res* **46**, e91 (2018).

1053 68. M. J. Machiela, S. J. Chanock, LDlink: a web-based application for exploring

1054 population-specific haplotype structure and linking correlated alleles of possible

1055 functional variants. *Bioinformatics* **31**, 3555-3557 (2015).

1056 69. A. L. Sanborn *et al.*, Chromatin extrusion explains key features of loop and domain

1057 formation in wild-type and engineered genomes. *Proc Natl Acad Sci U S A* **112**,

1058 E6456-6465 (2015).

1059 70. P. A. Ewels *et al.*, The nf-core framework for community-curated bioinformatics

1060 pipelines. *Nat Biotechnol* **38**, 276-278 (2020).

1061 71. F. Ramirez, F. Dundar, S. Diehl, B. A. Gruning, T. Manke, deepTools: a flexible

1062 platform for exploring deep-sequencing data. *Nucleic Acids Res* **42**, W187-191 (2014).

1063 72. Z. Gu, R. Eils, M. Schlesner, N. Ishaque, EnrichedHeatmap: an R/Bioconductor

1064 package for comprehensive visualization of genomic signal associations. *BMC*

1065 *Genomics* **19**, 234 (2018).

1066 73. Y. Zhang *et al.*, Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137

1067 (2008).

1068

1069

1070 **Figure Legends**

1071

1072 **Fig. 1 | Identification of causal eQTL-tagged CREs in pLD regions using**
1073 **multiplexed single-cell CRISPRi/a perturbations.** **a**, Schematic of the multiplexed
1074 single-cell CRISPRi/a perturbations used to screen for causal eCREs tagged by eQTL
1075 variants in pLD signals and to systematically study the regulatory relationships between
1076 genetic loci and gene expression levels. LEV: lead eQTL variants with equal fine-mapped
1077 causal probability; eCRE: LEV-tagged CRE (± 50 bp region of LEVs); pLD: perfect LD
1078 ($R^2=1$); MOI: multiplicity of infection; Blueprint-Tcel, Blueprint-neut and Blueprint-mono:
1079 fine-mapped eQTL data of CD4+ T cell samples, neutrophil samples and monocyte
1080 samples from BLUEPRINT project; TwinsUK-LCL and TwinsUK-blood: fine-mapped
1081 eQTL data of lymphoblastoid cell samples and whole blood samples from UK10K
1082 TwinsUK project; Geuvadis-LCL: fine-mapped eQTL data of lymphoblastoid cell samples
1083 from Geuvadis project; GTEx V8-blood: fine-mapped eQTL data of whole blood samples
1084 from GTEx V8 project. **b**, Summary information for the selected qualified LEVs and their
1085 corresponding pLD signals, including the number of independent eQTL datasets that
1086 support corresponding pLD signals, the number of fine-mapped LEVs within
1087 corresponding pLD signals, and the causal probability (mean score of different datasets,
1088 estimated via CaVEMaN) distribution of selected fine-mapped LEVs. **c**, The number of
1089 gRNAs per cell and the number of cells bearing each perturbation in CRISPRa/i screen,
1090 based on integration of data from both high or moderate MOIs. **d**, Quantile-quantile plot
1091 comparing observed versus expected P of eCRE-targeting sgRNAs (blue) and non-
1092 targeting control (NTC) sgRNAs (gray; down-sampled) associated with gene expression.
1093 **e**, Results of positive control sgRNAs in the CRISPRi and CRISPRa perturbations.
1094 sgRNAs targeting *EZH2*, *CANX*, *NEAT1*, *PARK7* are positive control targeting promoter
1095 of genes, and sgRNA effect on *NMU* is associated with an enhancer. **f**, Consistence
1096 analysis of perturbation results and RT-qPCR results for the 20 randomly selected causal
1097 eCREs-perturbGene pairs.

1098

1099 **Fig. 2 | CRISPR perturbations identify diverse regulatory patterns between causal**
1100 **eCREs and their target genes in pLD signals.** **a**, Distribution of the pLD signals
1101 containing different numbers of causal eCREs (top), regulating different numbers of
1102 perturbGenes (middle), and showing varied distances between eCREs and their
1103 perturbGenes (bottom). **b**, Patterns of causal eCREs targeting their perturbGenes in
1104 corresponding pLD signals. Genes are shown in dark blue or gray (possible nearest
1105 gene). The direction and position of TSS are indicated by arrows. pLD signals are labeled
1106 in light blue. Yellow dots represent eCREs. Arcs indicate regulatory relationships between
1107 causal eCREs and perturbGenes, while dashed lines indicate possible regulatory
1108 relationships. **c**, Validations of several representative eCRE-perturbGene pairs.
1109 Corresponding eCRE-perturbGene regulatory patterns (top), RT-qPCR analysis of
1110 perturbGene expression upon perturbation with sgRNAs targeting eCREs (middle) and
1111 differential perturbGene expressions in multiplexed CRISPR perturbations (bottom). **d**,
1112 4C assay and epigenomic evidence indicating chromatin interactions between
1113 rs79423518-located DNA fragment (yellow shaded region) and promoter of *PRDX5* and

1114 *TMRT112* (blue shaded region). **e**, Luciferase assay results of rs4930698/rs79423518-tagged CREs and allele-specific effects of them. Error bars represent standard deviation of the mean. P are indicated by asterisks, with $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$, and ns indicating not significant.

1118

1119 **Fig. 3 | Illustration and comparison of regulatory effects between eCRE**
1120 **perturbations and eQTL mapping.** **a**, Venn diagram comparing eCRE-perturbGene
1121 pairs identified by the multiplexed single-cell perturbations with eQTL-eGene pairs
1122 identified in GTEx V8 and eQTLgen datasets. **b**, Gene-level comparison including
1123 perturbGene and eGene, similar with **a**. **c**, Consistency of the estimated effects between
1124 eCRE-perturbGene pairs identified by the multiplexed perturbations and eQTL-eGene
1125 pairs derived from the conventional eQTL datasets. The red and green dots represent
1126 eCRE-perturbGene pairs in CRISPRa and CRISPRi perturbations, respectively, and the
1127 lines between the dots indicate that the eCREs are located within the same pLD signal. **d**
1128 and **e**, Examples of the varied regulatory effects on the target gene expression for the
1129 causal eCREs in two LD signals. **f**, Distribution of the distance between the LEVs and
1130 TSS of target genes in multiplexed perturbations (yellow) and GTEx V8 dataset (blue),
1131 with the median indicated by dashed lines. **g**, Comparison of the distance distribution
1132 between multiplexed perturbation results and GTEx V8 dataset in **f**, with statistical
1133 analysis by two-tailed t-test. **h**, 4C results showing the chromatin interaction frequency
1134 between the fragment containing rs143558304 and rs59508494 (top red arrow, 4C
1135 viewpoint) and the promoter regions (blue shaded region) of the two perturbGenes
1136 (bottom red arrows). **i**, RT-qPCR validation results for the sgRNAs targeting eCREs
1137 tagged by rs59508494 and rs143558304, with sgNTC as control. Error bars represent
1138 standard deviation of the mean. $^{****}P < 0.0001$. **j**, Differential perturbGene expressions in
1139 the multiplexed CRISPR perturbations for the causal eCREs tagged by rs59508494 and
1140 rs143558304. **k**, Schematic of patterns for causal eCREs regulating their target genes
1141 when considering the ranges of TAD and LD block. The symbols representing TAD, LD
1142 block, eCRE and perturbGene are shown in the figure.

1143

1144 **Fig. 4 | Evaluation of the endogenous perturbation effects via computational**
1145 **predictions and functional annotations.** **a**, Benchmarks of existing computational
1146 methods in predicting causal eCRE-associated LEVs, non-significant eCRE-associated
1147 LEVs were used as negative samples. **b**, Benchmarks of existing computational methods
1148 in predicting causal eCRE by taking median prediction scores of all possible ± 1 Kb
1149 variants surrounding corresponding LEVs, non-significant eCREs were used as negative
1150 samples. **c**, Empirical cumulative probability distribution of the first and second Enformer
1151 principal component scores for significant and non-significant eCRE-associated LEVs. **d**,
1152 Empirical cumulative probability distribution of the first and second median Enformer
1153 principal component scores for significant and non-significant eCRE, by taking median
1154 prediction scores of all possible ± 1 Kb variants surrounding corresponding LEVs. **e**,
1155 Epigenomic and computational evidence for two causal eCRE-associated LEVs,
1156 rs75446625 and rs80159064. **f**, RT-qPCR validation results for the sgRNAs targeting
1157 eCREs tagged by rs75446625 and rs80159064, with sgNTC as control. **g**, Differential

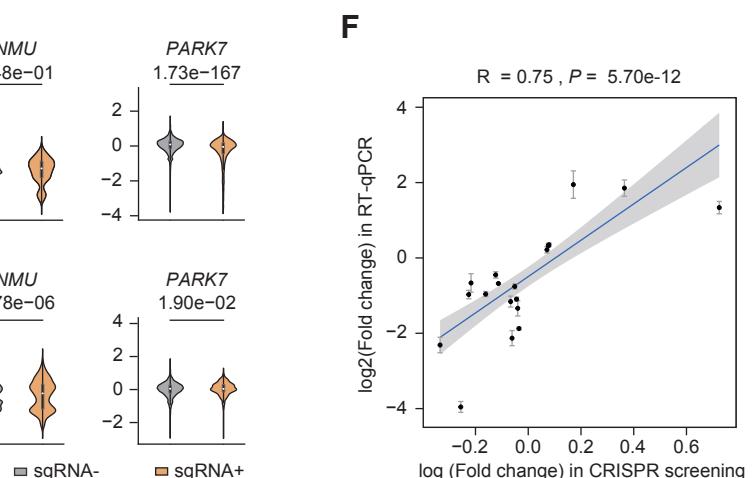
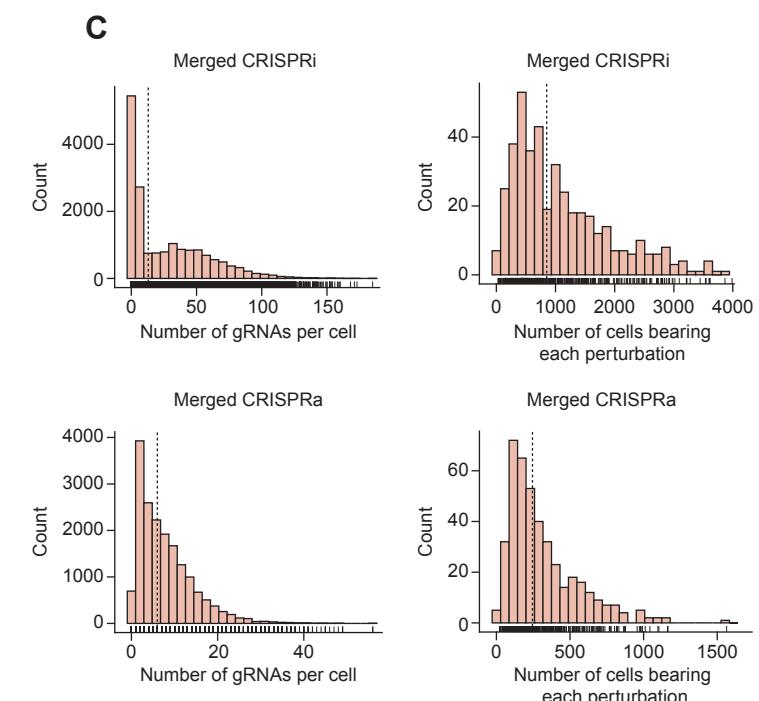
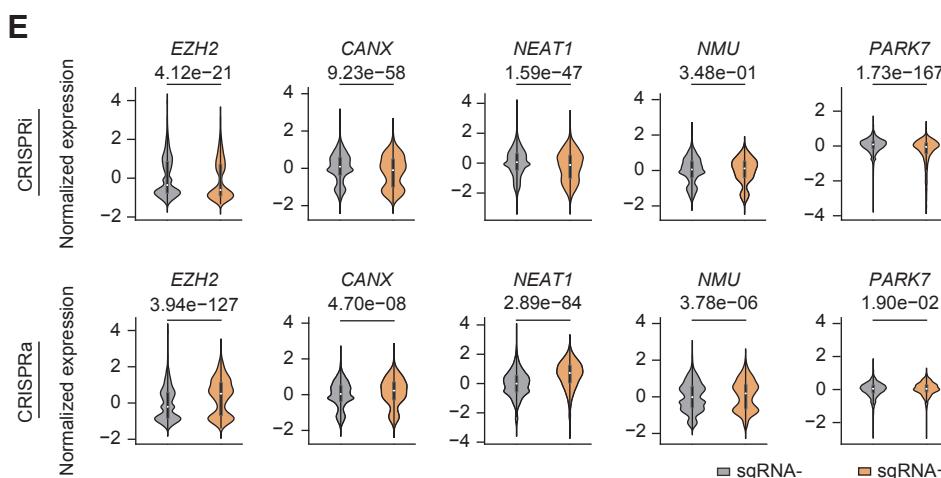
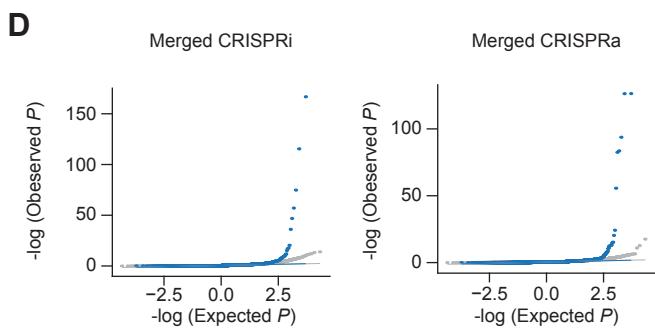
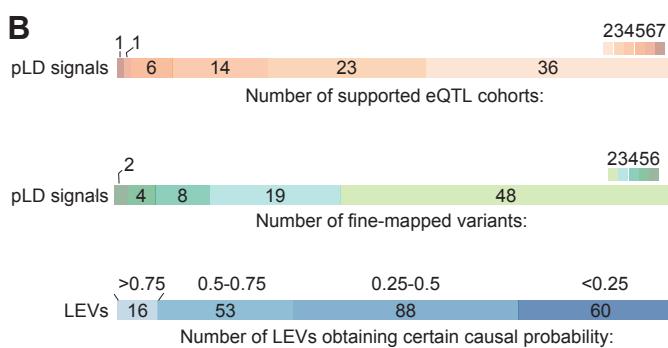
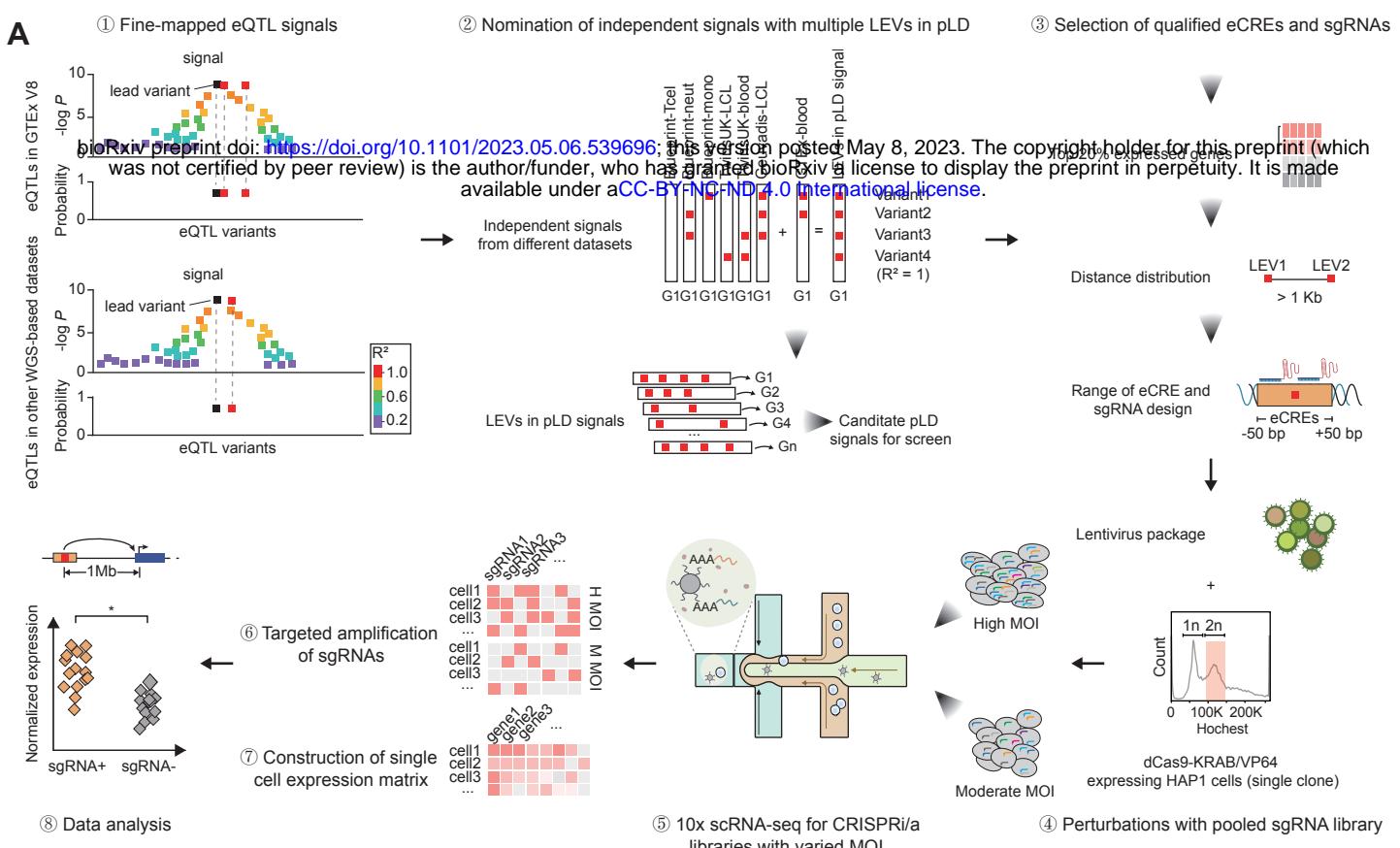
1158 perturbGene expressions in the multiplexed CRISPR perturbations for the causal eCREs
1159 tagged by rs75446625 and rs80159064. **h**, Luciferase assay results of
1160 rs75446625/rs80159064-tagged CREs and allele-specific effects of them. * $P < 0.05$; ** P
1161 < 0.01; *** $P < 0.001$; **** $P < 0.0001$; ns, no significant.

1162

1163 **Fig. 5 | Evaluation of unmarked regulatory elements revealed by endogenous**
1164 **perturbations and epigenomic marks.** **a**, The epigenomic profile of HAP1 cells for
1165 seven classical marks, including ATAC-seq, H3K27ac, H3K36me3, H3K4me1, H3K4me3,
1166 H3K27me3 and H3K9me3, within a ± 2.5 Kb range surrounding causal eCRE-associated
1167 LEVs. **b**, 4C results of the interaction frequency between the causal eCRE tagged by
1168 rs73156934 (viewpoint, red triangle and yellow shaded region) and the promoter region
1169 of CALD1 (blue shaded region). Epigenomic and computational evidence for two causal
1170 eCRE-associated LEVs, including rs73156934 and rs10428917, are shown below. **c**, RT-
1171 qPCR validation results for the sgRNAs targeting eCREs tagged by rs73156934 and
1172 rs10428917, with sgNTC as control. **d**, Differential perturbGene expressions in the
1173 multiplexed CRISPR perturbations for the causal eCREs tagged by rs73156934 and
1174 rs10428917. **e**, Luciferase assay results of rs73156934/rs10428917-tagged eCREs and
1175 allele-specific effects of them. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$; **** $P < 0.0001$; ns, no
1176 significant.

1177

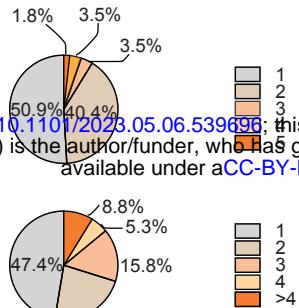
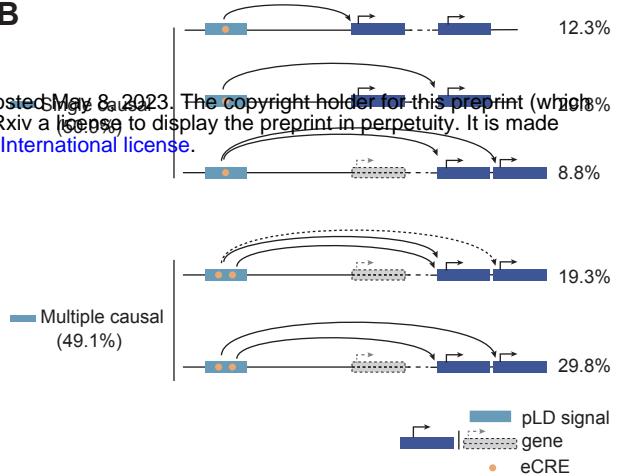
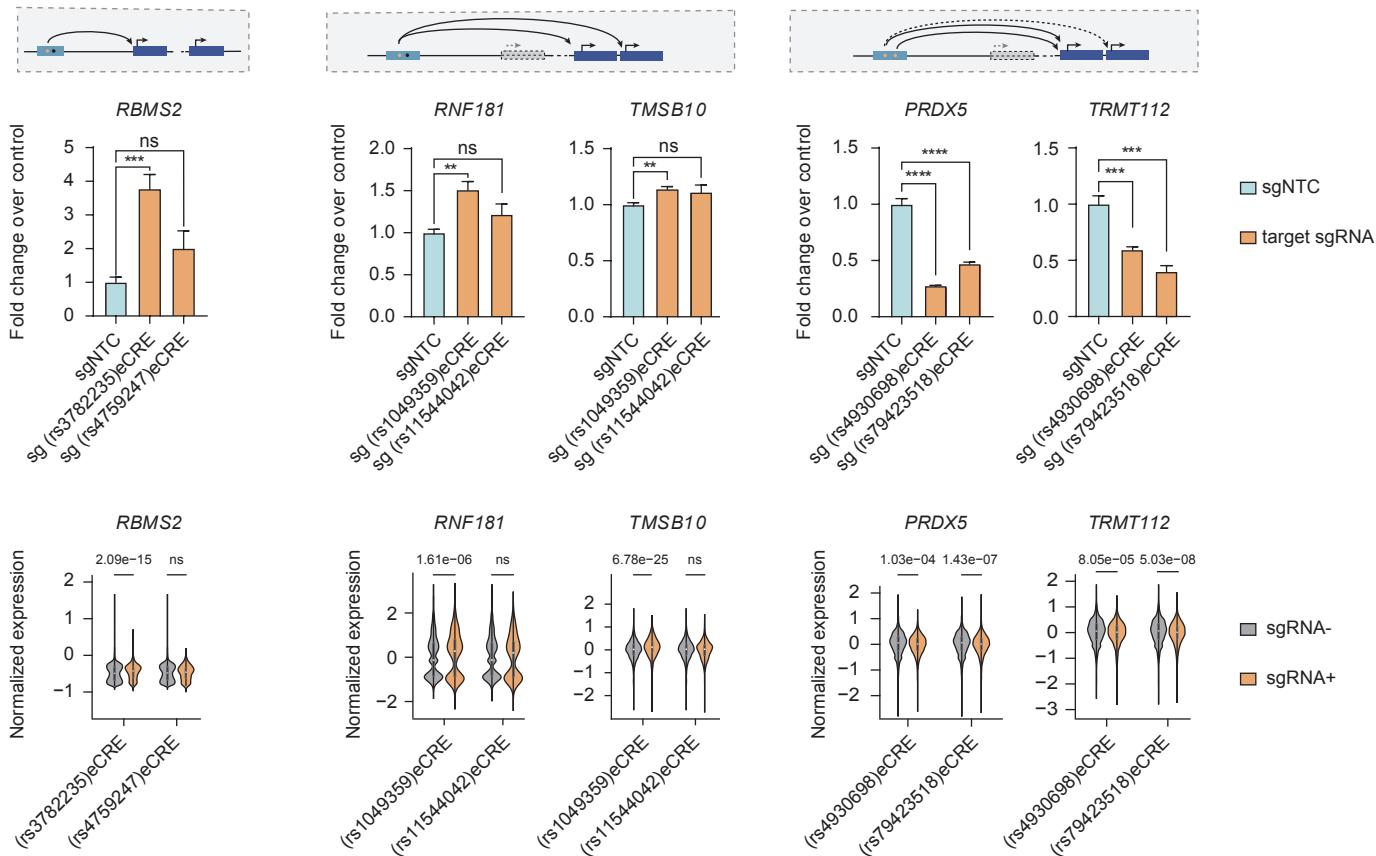
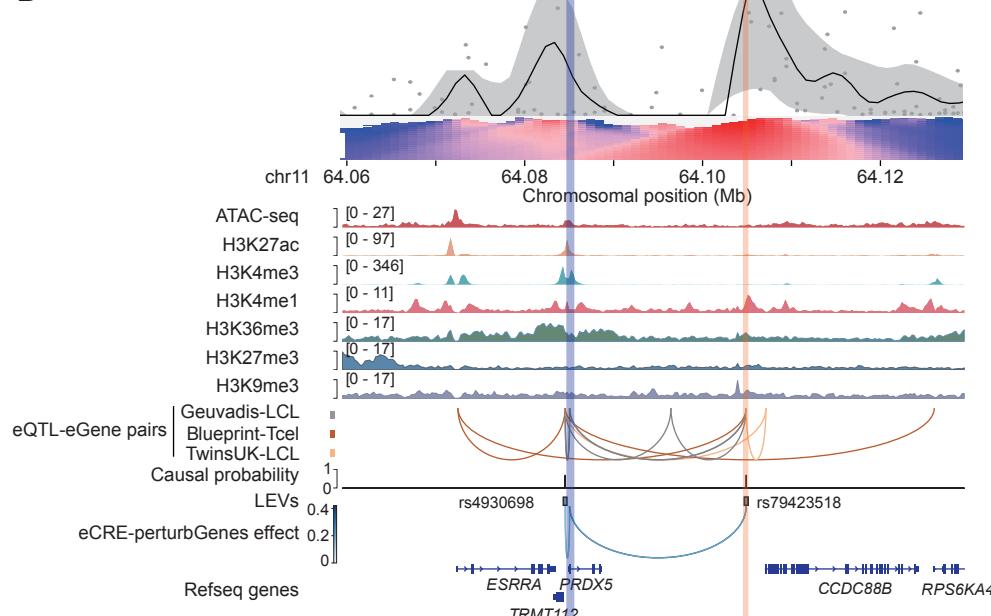
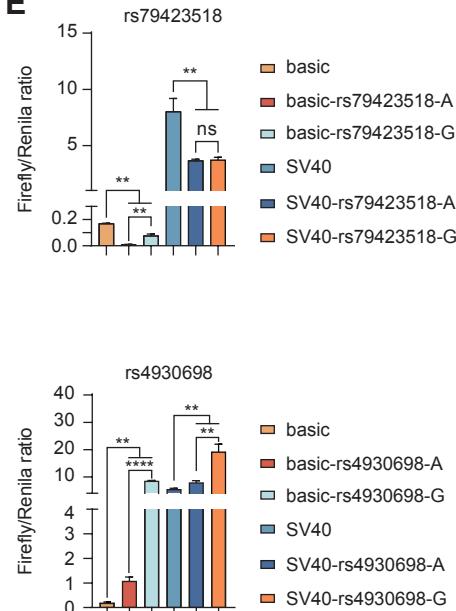
1178 **Fig. 6 | Genome-wide comparison of various CRISPRi and CRISPRa effects.** **a**,
1179 Magnitude and direction of perturbation effects for all tested eCRE-gene pairs in
1180 CRISPRi and CRISPRa perturbations, with blue dots indicating significance only in
1181 CRISPRi, yellow dots indicating significance only in CRISPRa, red dots indicating
1182 significance in both CRISPRa and CRISPRi, gray dots indicating no significance in both
1183 perturbations. Significance threshold is $q < 0.2$. **b**, LocusZoom-like plot for all tested hits
1184 and epigenomic marks for the causal eCREs identified via both CRISPRi and CRISPRa
1185 perturbations, with blue indicating significance in CRISPRi, yellow indicating significance
1186 in CRISPRa, and the vertical purple line indicating causal eCRE-perturbGene pairs
1187 identified by both CRISPRa and CRISPRi perturbations. **c**, Distribution of the distances
1188 between causal eCRE-associated LEVs and TSS of perturbGenes in hits at four
1189 quadrants obtained from **a**. Left y-axis represents the type of quadrant, with colors
1190 consistent in **a**. **d**, Enrichment of epigenomic marks in the four perturbation effect groups,
1191 tested by Fisher's exact test.

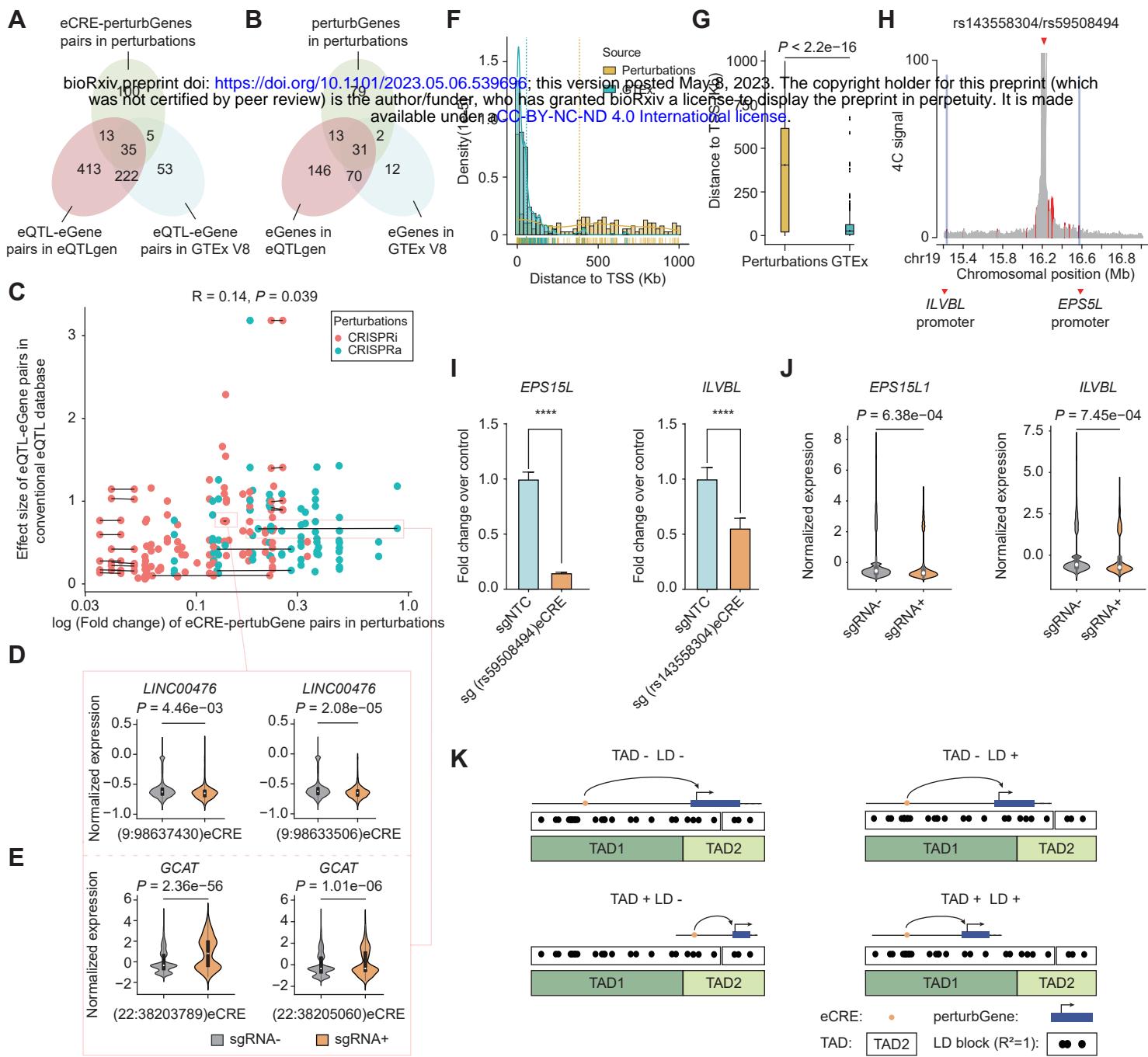


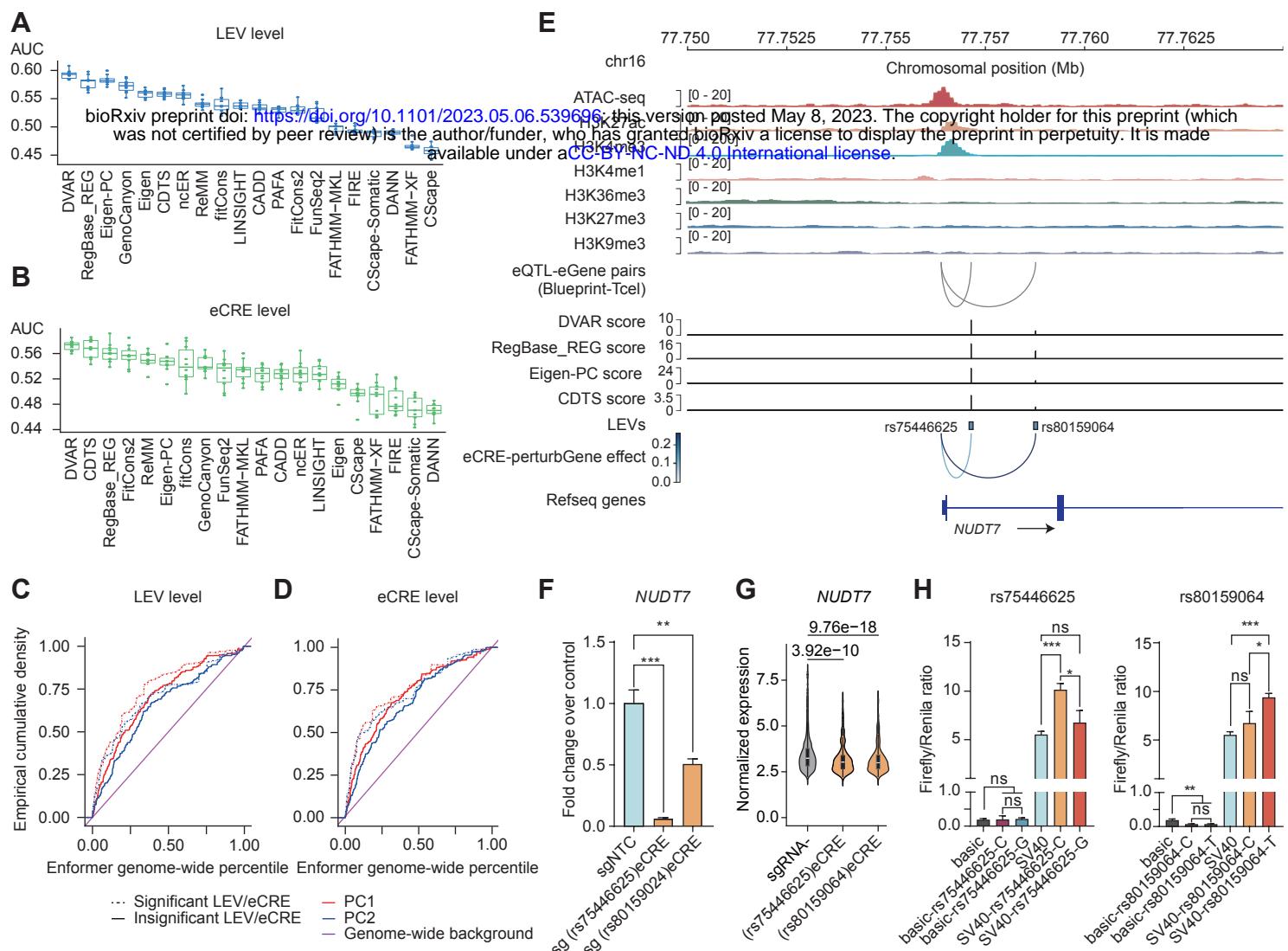
A

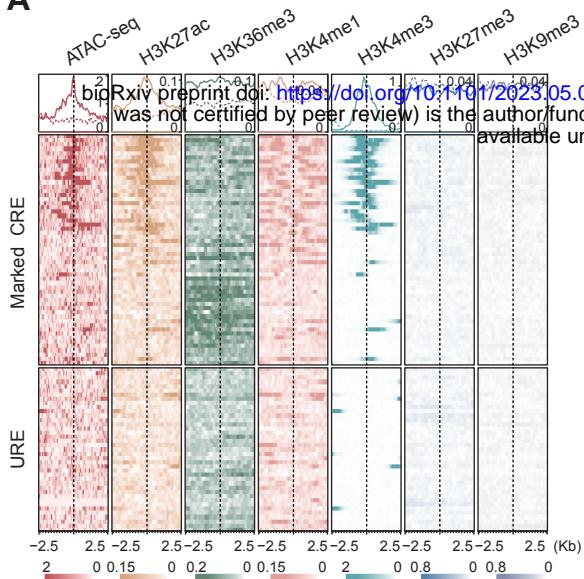
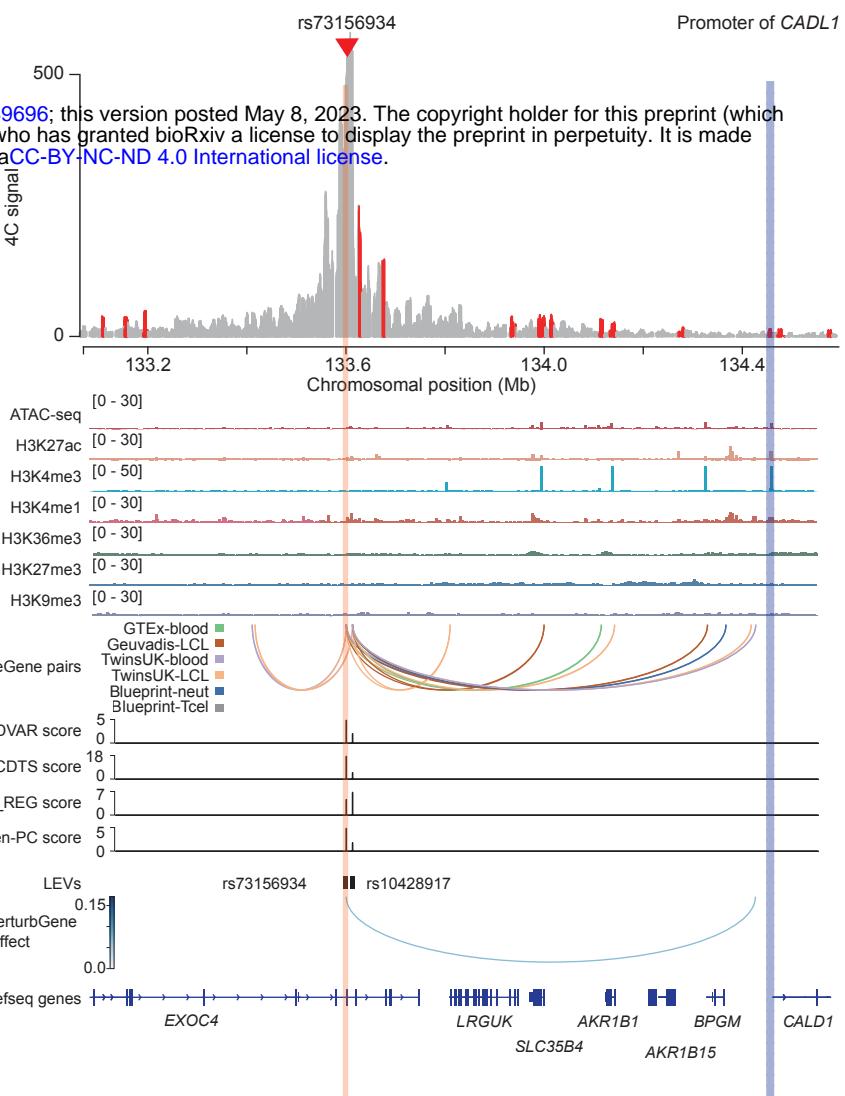
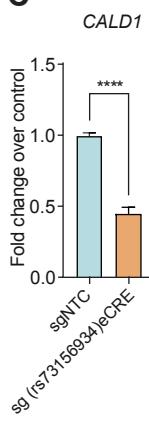
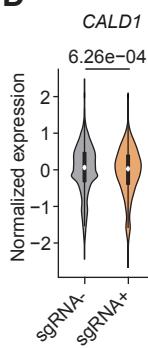
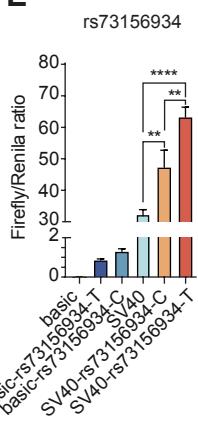
bioRxiv preprint doi: <https://doi.org/10.1101/2023.05.06.539696>; this version posted May 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

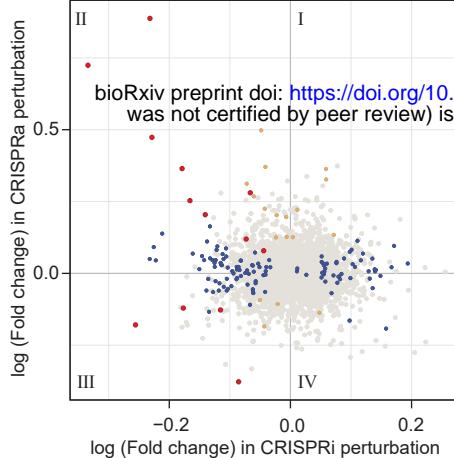
Propotion of pLD signals with different number of perturbGenes

**B****C****D****E**

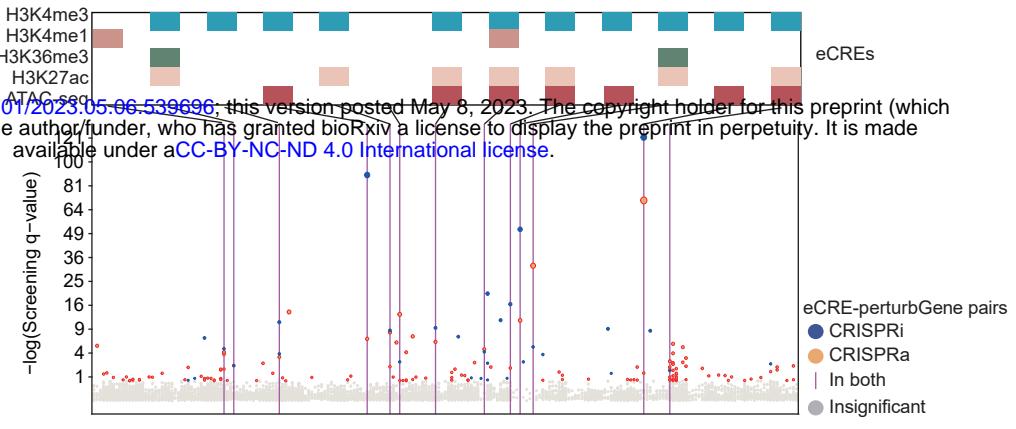
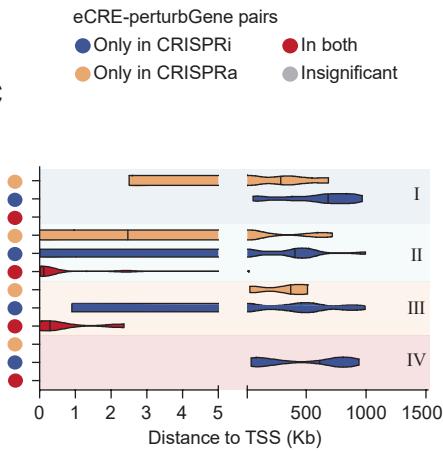




A**B****C****D****E**

A

bioRxiv preprint doi: <https://doi.org/10.1101/2023.05.06.539696>; this version posted May 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

B**C****D**