

1 **crisprQTL mapping as a genome-wide association framework**
2 **for cellular genetic screens**
3

4 Molly Gasperini^{1*}, Andrew J. Hill¹, José L. McFaleine-Figueroa¹, Beth Martin¹,
5 Cole Trapnell^{1,2}, Nadav Ahituv³, Jay Shendure^{1,2,4*}
6

7 ¹Department of Genome Sciences, University of Washington, Seattle, Washington, USA
8

9 ²Brotman Baty Institute for Precision Medicine, University of Washington, Seattle, Washington,
USA

10 ³Department of Bioengineering and Therapeutic Sciences, Institute for Human Genetics,
11 University of California San Francisco, San Francisco, California, USA

12 ⁴Howard Hughes Medical Institute, Seattle, Washington, USA
13

14 *Correspondence to Jay Shendure (shendure@uw.edu) and Molly Gasperini (gasperim@uw.edu)
15
16
17

18

Abstract

19

20 Expression quantitative trait locus (eQTL) and genome-wide association studies (GWAS) are
21 powerful paradigms for mapping the determinants of gene expression and organismal
22 phenotypes, respectively. However, eQTL mapping and GWAS are limited in scope (to naturally
23 occurring, common genetic variants) and resolution (by linkage disequilibrium). Here, we
24 present crisprQTL mapping, a framework in which large numbers of CRISPR/Cas9 perturbations
25 are introduced to each cell on an isogenic background, followed by single-cell RNA-seq
26 (scRNA-seq). crisprQTL mapping is analogous to conventional human eQTL studies, but with
27 individual humans replaced by individual cells; genetic variants replaced by unique
28 combinations of ‘unlinked’ guide RNA (gRNA)-programmed perturbations per cell; and
29 tissue-level RNA-seq of many individuals replaced by scRNA-seq of many cells. By randomly
30 introducing gRNAs, a single population of cells can be leveraged to test for association between
31 each perturbation and the expression of any potential target gene, analogous to how eQTL
32 studies leverage populations of humans to test millions of genetic variants for associations with
33 expression in a genome-wide manner. However, crisprQTL mapping is neither limited to
34 naturally occurring, common genetic variants nor by linkage disequilibrium. As a
35 proof-of-concept, we applied crisprQTL mapping to evaluate 1,119 candidate enhancers with no
36 strong *a priori* hypothesis as to their target gene(s). Perturbations were made by a nuclease-dead
37 Cas9 (dCas9) tethered to KRAB, and introduced at a mean ‘allele frequency’ of 1.1% into a
38 population of 47,650 profiled human K562 cells (median of 15 gRNAs identified per cell). We
39 tested for differential expression of all genes within 1 megabase (Mb) of each candidate
40 enhancer, effectively evaluating 17,584 potential enhancer-target gene relationships within a
41 single experiment. At an empirical false discovery rate (FDR) of 10%, we identify 128 *cis*
42 crisprQTLs (11%) whose targeting resulted in downregulation of 105 nearby genes. crisprQTLs
43 were strongly enriched for proximity to their target genes (median 34.3 kilobases (Kb)) and the
44 strength of H3K27ac, p300, and lineage-specific transcription factor (TF) ChIP-seq peaks. Our
45 results establish the power of the eQTL mapping paradigm as applied to programmed variation
46 in populations of cells, rather than natural variation in populations of individuals. We anticipate
47 that crisprQTL mapping will facilitate the comprehensive elucidation of the *cis*-regulatory
48 architecture of the human genome.

49

50

51 **Main Text**

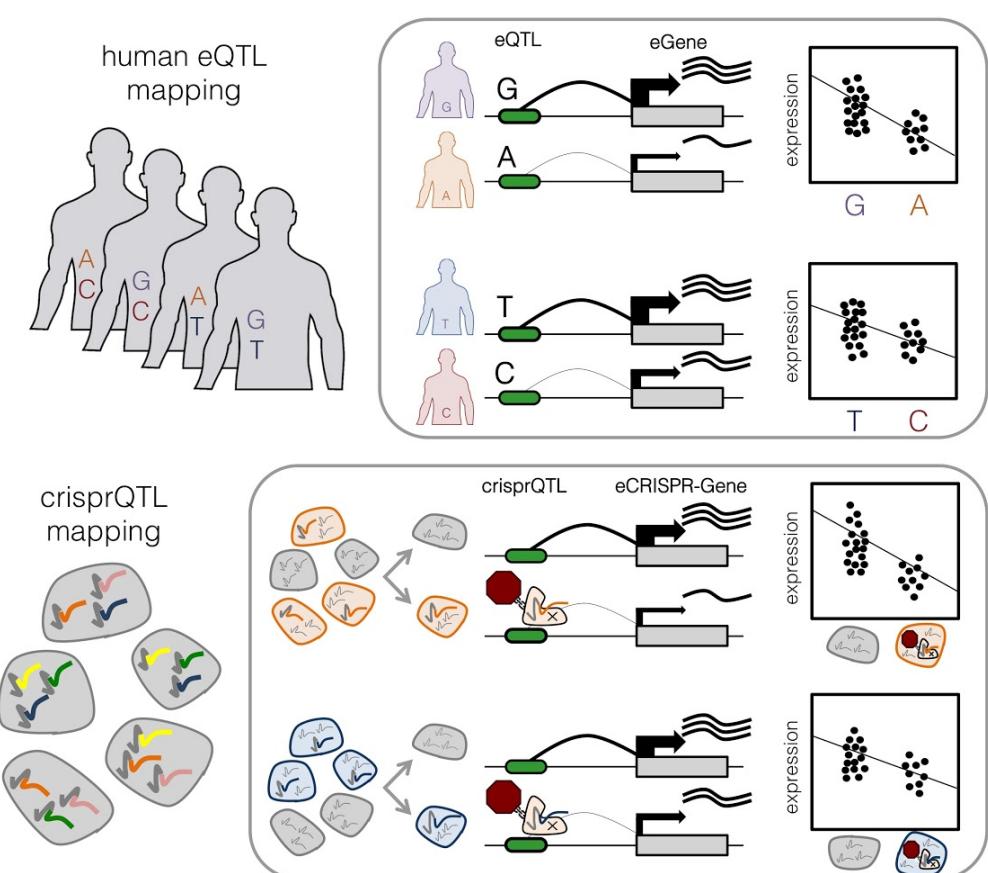
52
53 Consequent to an era of biochemical surveys of the human genome (such as the ENCODE
54 Project, (1)) and ‘common variant’ human genetics (*i.e.* GWAS, eQTL, (2, 3)), we are currently
55 awash in candidate regulatory elements and phenotype-linked haplotypes, respectively. It is clear
56 that determining whether and how each candidate regulatory element is truly functional, as well
57 as pinpointing which noncoding variants are causal for genetic associations, will require
58 functional characterization of vast numbers of sequences, *e.g.* through massively parallel reporter
59 assays and/or CRISPR/Cas9-directed perturbations.
60

61 We and other groups have recently adapted cell-based CRISPR/Cas9 genetic screens to evaluate
62 candidate regulatory sequences in their native genomic context (4–12). However, two key
63 aspects of the methods used in these studies limit their scalability. First, each study focuses on
64 the regulation of a single gene per experiment, typically entailing the development of a
65 gene-specific assay. Second, each cell is a vehicle for a single perturbation, *e.g.*
66 CRISPR-mediated repression, mutation or deletion of one candidate regulatory sequence per
67 cell, with the specificity-conferring gRNAs usually introduced via lentivirus at a low multiplicity
68 of infection (MOI). With millions of candidate regulatory elements and ~20,000 regulated genes
69 in the human genome, these limitations preclude the comprehensive dissection of the
70 *cis*-regulatory architecture even within a single cell line.
71

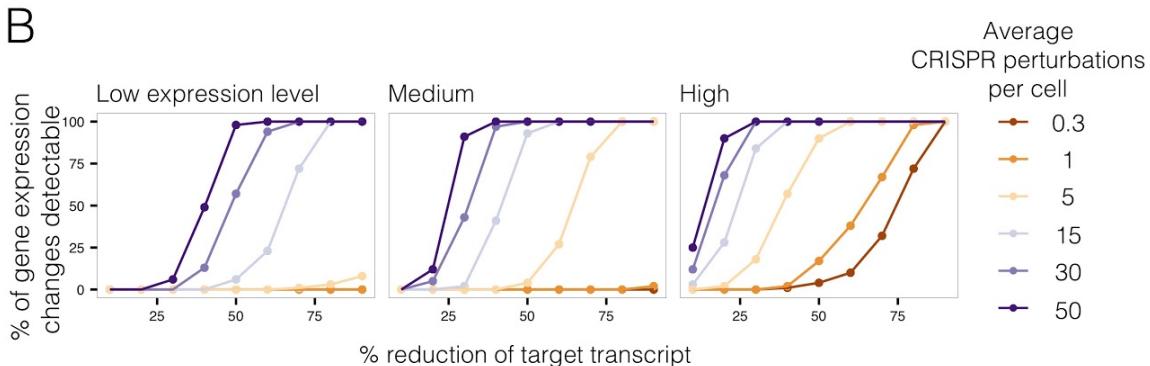
72 crisprQTL mapping (**Fig. 1A**) is designed to overcome both limitations. First, by using
73 scRNA-seq instead of gene-specific assays, a single experiment can globally capture
74 perturbations to gene expression, with no strong *a priori* hypothesis as to the target gene of each
75 regulatory element tested. Second, by introducing gRNAs at a high MOI, each individual cell
76 acquires a unique combination of ‘unlinked’ perturbations against the isogenic background of a
77 cell line. Introducing multiple perturbations per cell greatly increases the power to detect changes
78 in gene expression and reduces the number of cells required per experiment (**Fig. 1B**). As in
79 human eQTL studies (13, 14), an association framework can be used to map *cis*- and *trans*-
80 effects by comparing gene expression in the subsets of cells that contain each gRNA to the cells
81 that lack that guide (rather than comparing the subsets of people that do vs. don’t harbor each
82 genetic variant). However, unlike eQTL studies, the resolution of crisprQTL mapping is not
83 constrained by linkage disequilibrium, nor is it limited to studying the effects of naturally
84 occurring, common genetic variants.
85
86

87

1A



B



88
89
90
91
92
93
94
95
96
97
98

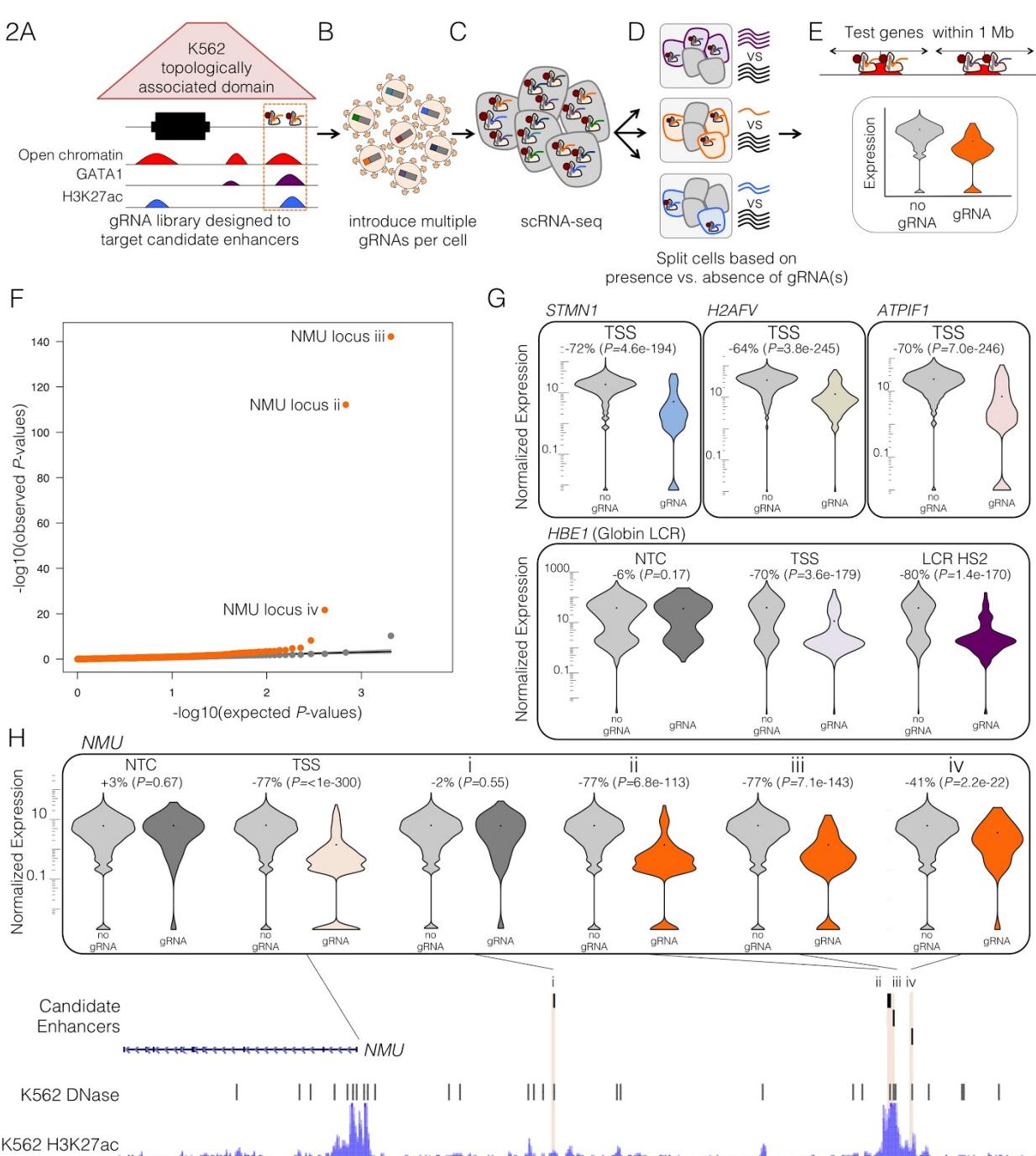
Fig. 1 | crisprQTL mapping. (A) crisprQTL mapping uses the same framework as human eQTL studies, but with a population of human individuals replaced by a population of individual cells, natural genetic variation replaced by diverse combinations of gRNA-programmed perturbations in each cell, and tissue-level RNA-seq of each person replaced by scRNA-seq. (B) Multiplex perturbations increase power to detect changes in gene expression in single-cell genetic screens while greatly reducing the number of cells necessary to profile. Simulated power calculations show that increasing the average number of perturbations per cell (e.g., by increasing MOI in lentiviral delivery of gRNAs) strongly increases power to detect changes in gene expression, including for genes with low (0.10 mean UMIs per cell), medium (0.32) or high (1.00) levels of mean expression. X-axis corresponds to the simulated % change of transcript repressed by targeting CRISPRi to the associated enhancer. Calculations assume a fixed number of 45,000 cells profiled by scRNA-seq.

99 As an initial pilot of crisprQTL mapping, we targeted 100 candidate enhancers in the chronic
100 myelogenous leukemia cell line K562, with CRISPR-interference (CRISPRi) as our mode of
101 perturbation. For CRISPRi, we used a nuclease-inactive Cas9 tethered to the KRAB repressor
102 domain to induce heterochromatin across a 1-2 kilobase (Kb) window around a gRNA's target
103 site (15). The 100 candidate enhancers were all intergenic sites of coincident DNase I
104 hypersensitivity (DHS), H3K27 acetylation and GATA1 binding, that resided within the same
105 topologically associated domain (TAD) as at least one gene with high K562-specific expression
106 (**Fig. 2A**). Two gRNAs were designed to target each candidate enhancer. Additional pairs of
107 gRNAs served as positive (targeting the transcription start sites (TSSs) of 25 highly expressed
108 genes, or hypersensitivity sites of the β -globin locus control region (LCR)) or negative controls
109 (13 non-targeting controls or 'NTC' that either fall nowhere in the genome or in a gene desert)
110 (**Table S1**).

111
112 This pilot gRNA library was cloned into the lentiviral CROP-seq vector modified to include a
113 CRISPRi-optimized backbone (16, 17) and K562 cells were transduced at a high MOI (**Fig. 2B**).
114 After 10 days to allow for effective CRISPRi, the transcriptomes of 6,806 single cells were
115 sequenced on the 10x Genomics platform (**Fig. 2C**). With a targeted enrichment protocol (18),
116 we identified a median of 8 +/- 4.5 gRNAs per cell (interquartile range = 7, **Fig. S1A**). The mean
117 'allele frequency' of perturbations in this experiment – that is, the proportion of cells in which a
118 given candidate enhancer or control was targeted – was 4.1% +/- 2.2% (**Fig. S1B**). For each
119 targeted element, we partitioned the 6,806 cells based on whether they did or did not contain
120 gRNA(s) targeting it (**Fig. 2D**). We then tested for reductions in expression of all genes within 1
121 Mb of that element (**Fig. 2E**), similar to *cis* eQTL identification in conventional eQTL studies
122 (14).
123

124 The 13 NTC controls were tested against all 1,196 genes that fell within 1 Mb of any targeted
125 element (**Fig. 2F**). We defined a 10% empirical FDR threshold based on these NTC tests as they
126 are subject to the same sources of error as the element-targeting gRNAs. With this threshold,
127 68% (17 of 25) of TSS-targeting positive controls repressed their associated genes, as did the
128 β -globin LCR controls (**Fig. 2G**). In contrast, targeting of only 3% of the candidate enhancers (3
129 of 100) significantly decreased expression of any gene within 1 Mb (**Fig. 2F**, **Fig. S1C**). All
130 three decreased expression of the same gene, *NMU*, which encodes neuromedin U, a
131 neuropeptide that plays roles in inflammation as well as erythropoiesis (19–22). These three
132 crisprQTLs were relatively closely located: 93.4, 94.1 and 97.6 Kb upstream of the *NMU* TSS
133 (ii-iv in **Fig. 2H**; region denoted 'e-NMU' as potentially too close to one another to resolve by
134 CRISPRi). In contrast, targeting of a different candidate enhancer located just 34.4 Kb upstream
135 of the *NMU* TSS (i in **Fig. 2H**) did not significantly reduce its expression.
136
137

138



139

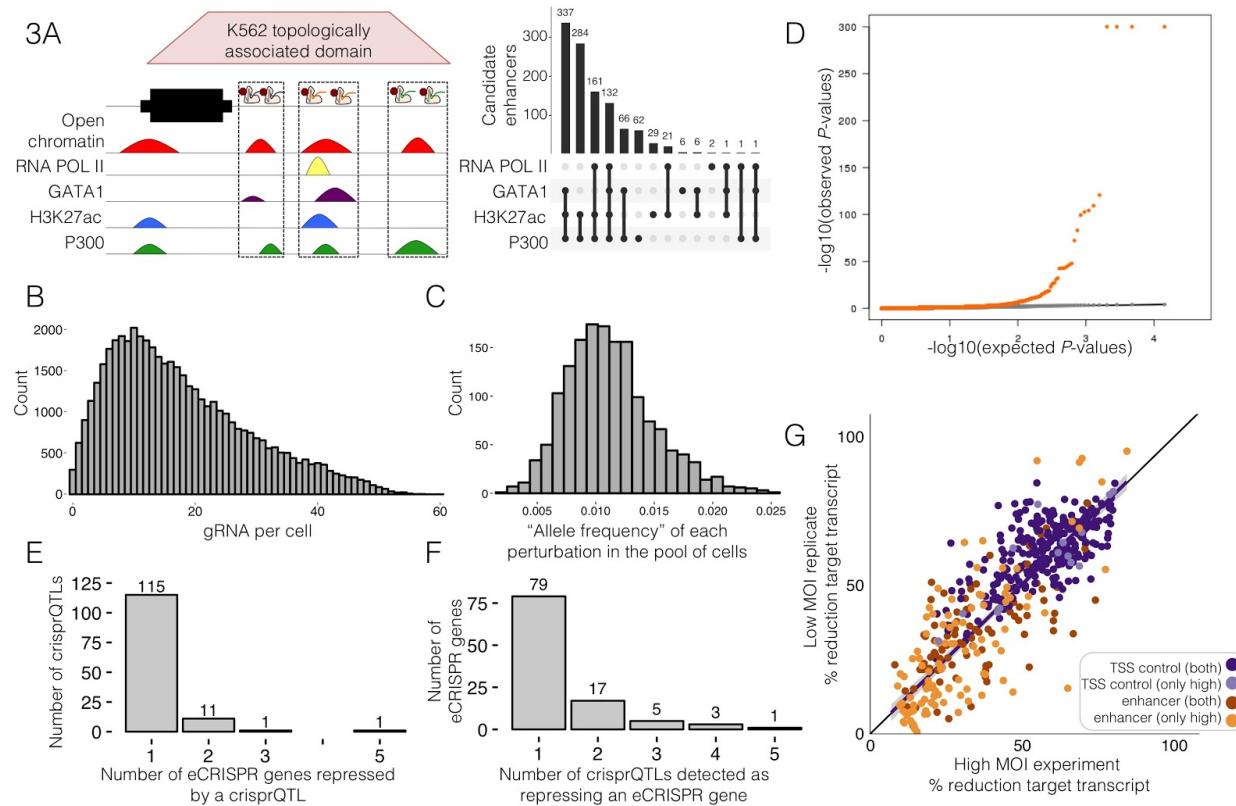
Fig. 2 | Pilot crisprQTL mapping study testing 100 candidate enhancers in K562 cells. (A-E), Schematic of crisprQTL mapping. **A**, Candidate enhancers were chosen based on enhancer-associated features, and were each targeted by 2 gRNAs. **B**, gRNAs were cloned into the CRISPRi-optimized CROP-seq lentiviral vector, and delivered to K562 cells at a high MOI. **C**, scRNA-seq was performed on these cells, with concurrent identification of the combination of gRNAs present in each cell. **D**, For each candidate enhancer, cells were partitioned based on whether or not they contained a gRNA targeting it. **E**, For each such partition, we used Monocle (23) to test for differential expression between the two populations for any gene within 1 Mb of the candidate enhancer. **(F)** Quantile-quantile (Q-Q) plot of the

147 differential expression tests. Distributions of observed vs. expected *P*-values for candidate
148 enhancer-targeting gRNAs (orange) and NTC gRNAs (gray; downsampled) are shown. **(G)** Expression of
149 selected TSS and β -globin LCR positive controls. Positive controls showed significant differential
150 expression of the expected target genes between cells *with* versus *without* targeting gRNAs, in contrast
151 with NTCs. **(H)** Distal enhancers of *NMU* identified as crisprQTLs. Targeting three candidate enhancers
152 (labeled *ii-iv*) located 93.4, 94.1 and 97.6 Kb upstream of *NMU*, significantly reduced expression of
153 *NMU*, but targeting an intervening candidate enhancer (labeled *i*) located 34.4 Kb upstream did not.
154

155
156 To demonstrate crisprQTL mapping at scale, we designed gRNAs targeting 1,119 candidate
157 enhancers throughout the genome (2 gRNAs each; 62/100 elements from pilot experiment were
158 re-targeted, including the three crisprQTLs upstream of *NMU*). For this experiment, we adjusted
159 criteria for selection, focusing on intergenic DHS that exhibited various combinations of
160 H3K27ac, p300 binding, GATA1 binding and RNA Pol II binding (**Fig. 3A**). Candidate
161 enhancers were required to fall within the same TAD (24) as at least one gene exhibiting strong
162 K562-specific expression (top 10% most highly expressed genes in the 6,806 cells of the pilot
163 dataset), and were collectively distributed across 510 TADs on every chromosome. Almost half
164 of all genes that are expressed in K562 cells fell within 1 Mb of at least one targeted candidate
165 enhancer (5,751 of the 12,984 expressed genes; expressed = observed in at least 0.525% of cells
166 in this experiment). In addition, pairs of gRNAs targeting 381 TSSs (of genes sampled from the
167 top decile of K562 expression), the globin LCR and 50 NTCs were again included as positive or
168 negative controls (**Table S2**).
169

170 We cloned these gRNAs into the CRISPRi-optimized CROP-seq vector, transduced K562 cells at
171 a high MOI, and then generated 47,650 scRNA-seq profiles. With greater input material to the
172 targeted amplification than was used in the pilot study, we identified a median of 15 +/- 11.3
173 gRNAs per cell (interquartile range 15, **Fig. 3B**). The mean ‘allele frequency’ of perturbations in
174 this experiment was 1.1% +/- 0.06% (**Fig. 3C**). We then iteratively partitioned the cells based on
175 whether they did or did not contain gRNA(s) targeting each TSS or candidate enhancer, and
176 tested for reductions in expression of all genes within 1 Mb of it (**Fig. S2**). We also tested the 50
177 NTCs against all 5,751 genes within 1 Mb of any targeted candidate enhancer. For perspective,
178 in a ‘one gRNA per cell’ framework, achieving equivalent power would require generating
179 ~715,000 single cell transcriptomes.
180

181



182 **Fig. 3 | crisprQTL mapping at scale in K562 cells.** (A) gRNAs were designed to target 1,119 candidate
 183 enhancers, as defined by DNase hypersensitive sites intersecting with 1+ enhancer-associated biochemical
 184 marks. (B) gRNAs were delivered to K562 cells at a high MOI (median 15 +/- 11.3 gRNAs identified per
 185 cell). (C) The median ‘allele frequency’, *i.e.* the proportion of the 47,650 cells in which a given element
 186 was targeted, was 1.1% +/- 0.06%. (D) Q-Q plot of the differential expression tests. Distributions of
 187 observed vs. expected *P*-values for candidate enhancer-targeting gRNAs (orange) and NTC gRNAs (gray;
 188 downsampled) are shown. (E) Histogram of the number of eCRISPR-genes significantly impacted by
 189 crisprQTLs. (F) Histogram of the number of crisprQTLs detected as regulating an eCRISPR-gene. (G)
 190 Comparison of effect sizes from the original high MOI study and a low MOI replicate (41,284 cells,
 191 median 1 gRNA/cell). Only TSS or candidate enhancers that met thresholds in the high MOI study are
 192 plotted (Spearman’s rho 0.826; slope 0.9995, y-intercept -0.011; both = identified in both low and high
 193 MOI study; only high = only in the high MOI study).

194

195

196 A quantile-quantile plot showed an excess of significant associations involving the targeting of
 197 candidate enhancers, relative to NTC controls (**Fig. 3D**). Setting a threshold corresponding to a
 198 10% empirical FDR (based on the NTCs), 94% (357 of 381) of TSS-targeting positive controls
 199 repressed their associated genes, as did the β -globin LCR controls. In addition, targeting of 11%
 200 (128 of the 1,119) of candidate enhancers repressed 1+ gene(s) within 1 Mb. As there were 13
 201 candidate enhancers whose targeting impacted more than one gene (**Fig. 3E**), our data identified
 202 a total of 145 crisprQTL enhancer-gene relationships (**Table S3**). Of the 105 downregulated

203 genes ('eCRISPR genes'), 26 were impacted by targeting of more than one candidate enhancer in
204 the library (**Fig. 3F**).
205

206 The three candidate enhancers implicated as crisprQTLs of *NMU* in the pilot experiment were
207 again identified here, as was an enhancer 3.6 Kb upstream of *GATA1* that was previously found
208 in a tiling CRISPRi-noncoding screen (11). Additional examples are highlighted in **Fig. 4**. Our
209 data support three elements as crisprQTLs of *FEZ1*, a gene implicated in schizophrenia (25, 26)
210 that is nonetheless expressed in K562 cells (**Fig. 4A**). One falls 37 Kb upstream (i; **Fig. 4A**),
211 while the other two are clustered 56 Kb and 57 Kb upstream (ii-iii; **Fig. 4A**). Targeting these
212 latter two elements also decreased expression of nearby *EI24*, which is 17 Kb away in the
213 opposite direction. Our data also support a crisprQTL within the fatty acid desaturase (*FADS*)
214 cluster. The *FADS* locus is thought to have experienced multiple independent episodes of local
215 adaptation in geographically diverse human populations in response to differences in diet
216 (27–29). The candidate enhancer lies between *FADS2* and *FADS3* (which are oriented
217 tail-to-tail), but downregulation is only detected for the more distal *FADS1* (**Fig. 4B**).
218

219 A longstanding question in gene regulation is by what distances are endogenous enhancers
220 separated from their target genes in the human genome? We find crisprQTLs were separated
221 from the TSS of their target genes by a median distance of 34.3 Kb (upstream crisprQTLs only,
222 to avoid biasing of this calculation by our avoidance of intronic candidate enhancers), supporting
223 a strong role for proximity in governing enhancer-promoter choice, given that we tested
224 association of all genes within 1 Mb (**Fig. 4C**). However, the majority of crisprQTL regulatory
225 relationships did not involve the most proximal gene, and not even the most proximal
226 K562-expressed gene (**Fig. 4D**; only 34/145 (23%) of the crisprQTLs detected in this experiment
227 involved the K562-expressed gene most proximal to the targeted candidate enhancer).
228

229 Of our 357 "positive control" TSSs whose targeting successfully repressed the expected gene, 55
230 (15%) reduced expression of 1+ additional genes (72 apparent promoter-promoter relationships
231 in total). 16 of these 72 involved overlapping promoters (TSSs within 1 Kb), such that the
232 observed effect of CRISPRi is likely direct. As for the remaining 56, one possibility is that these
233 represent examples of promoters acting as enhancers, as has been recently reported (11, 12).
234 However, these 56 are *not* enriched for proximity to affected genes (median distance of upstream
235 genes = 448.6 Kb; median distance to all genes within 1 Mb = 441 Kb) (**Fig. S3**), in contrast
236 with the 145 enhancer-TSS crisprQTL relationships (median distance = 34.3 Kb) (**Fig. 4C**). We
237 therefore hypothesize that these are more likely *trans* effects of repressing the primary target.
238

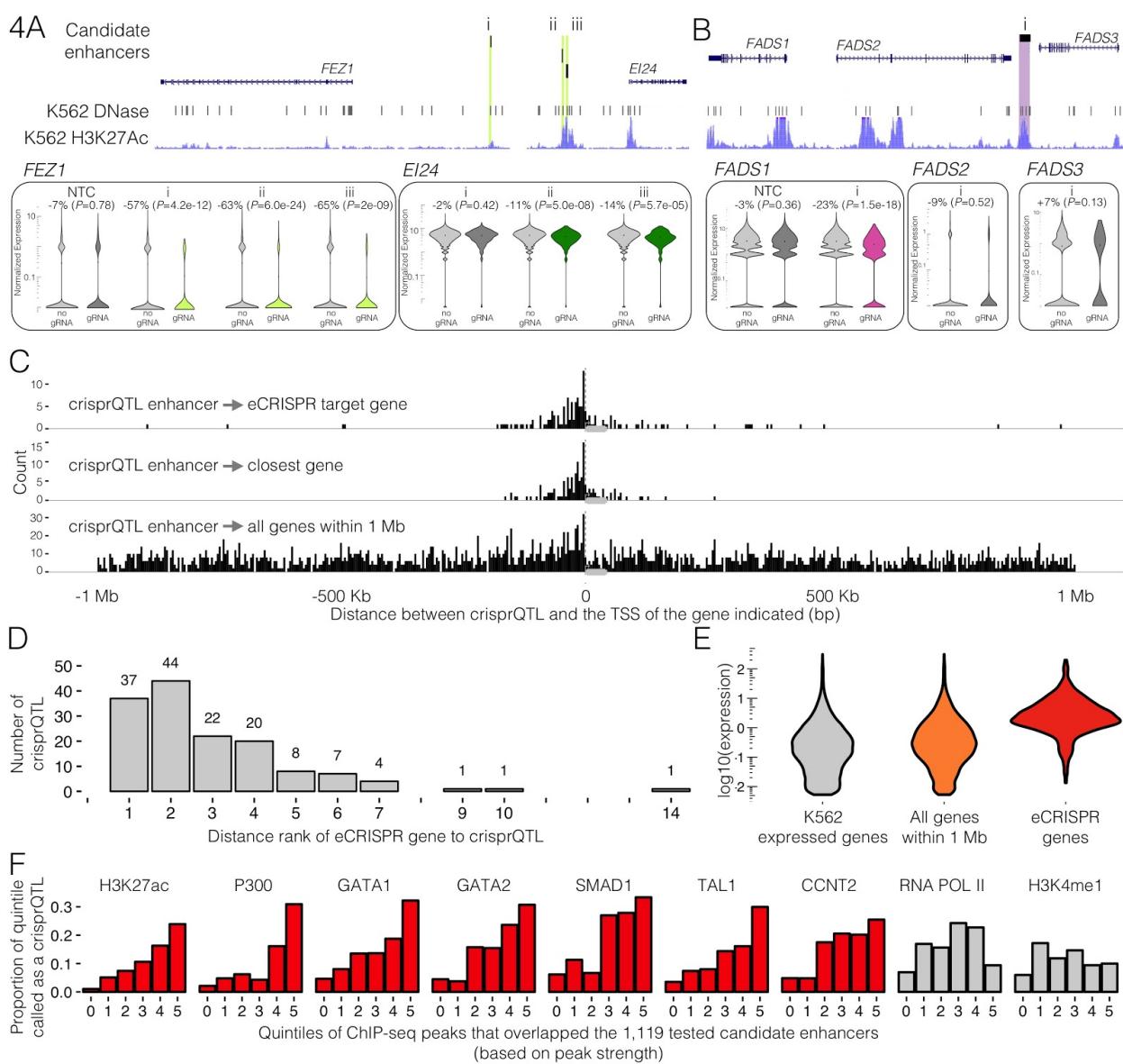
239 The set of 105 eCRISPR genes identified here, *i.e.* genes regulated by 1+ crisprQTL(s), had
240 several notable characteristics. First, their expression levels largely fall in the top half of all
241 5,751 genes against which we tested (**Fig. 4E**), suggesting that we are likely underpowered to

242 detect regulatory effects on more lowly expressed genes. Second, we detected no depletion of
243 housekeeping genes (hypergeometric test; P -value = 0.95, 34 of 105 (32%) genes annotated as
244 housekeeping (30) vs. 35% of all genes expressed in K562), arguing against the view that a
245 prevailing characteristic of housekeeping genes may be a relative dearth of distal regulatory
246 elements (5, 31). Finally, functional annotation enrichment of the 105 eCRISPR genes was most
247 significant for hematopoiesis and regulation of erythrocyte differentiation, consistent with distal
248 enhancers primarily shaping the expression of K562 cell type-specific genes (**Table S4**) (32, 33).
249 These terms were not enriched in a random expression-level matched control gene set.
250

251 We also examined the characteristics of crisprQTLs whose targeting significantly impacted
252 expression of 1+ genes in *cis*. Among the 1,119 total candidate enhancers targeted, the 128
253 crisprQTLs were positively correlated with ChIP-seq peak strength (based on average
254 enrichment in ChIP-seq peak region) for enhancer-associated histone modifications (H3K27ac,
255 logistic regression P -value = 4e-05), co-activators (p300, P -value = 4e-16) and lineage-specific
256 TFs (GATA1 P -value = 2e-07, GATA2 P -value = 3e-10, SMAD1 P -value = 1e-06, TAL1
257 P -value = 6e-06, CCNT2 P -value = 3e-07), whereas RNA Pol II and H3K4me1 were not
258 associated (**Fig. 4F**). Because many of these biochemical features are correlated, we trained a
259 multivariate logistic regression classifier to ask whether we could distinguish the 128 crisprQTLs
260 from the 991 candidate enhancers for which we did not identify an eCRISPR target gene,
261 achieving an AUPR of 0.31 (area under precision-recall curve; median from five-fold cross
262 validation, **Fig. S4**). H3K27ac, p300 and distance from the closest TSS were the only significant
263 predictors in the multivariate model (**Table S5**).
264

265 To replicate a small subset of our crisprQTLs outside of the multiplexed screen, we separately
266 transduced small pools of gRNA that targeted individual crisprQTL candidate enhancers, and
267 investigated repression of eCRISPR gene expression levels via bulk RNA-seq. For such
268 experiments targeting eNMU, ePRKCB, ePTGER3, and eGYPC, we observed effect sizes on the
269 expected eCRISPR target gene that matched the effect size from the pooled mapping experiment
270 (crisprQTL repression ranging from -79% to -42%, P -values from 6.8e-113 to 6.2e-11, **Fig. S5**).
271 For these four cases, the expected eCRISPR gene was the top differentially expressed target
272 gene. Although eHMGA1 failed to replicate, we note that of the five, it was the crisprQTL
273 closest to our 10% empirical FDR threshold, with only -11% repression and P -value of 0.00033.

274



275 **Fig. 4 | Examples and characteristics of K562 crisprQTLs. (A)** crisprQTLs of *EI24* and *FEZ1*. When
 276 targeted in our screen, three candidate enhancers (i-iii) decreased expression of the
 277 schizophrenia-associated gene *FEZ1* (25), with the latter two (ii-iii) also decreasing nearby *EI24*, a
 278 putative tumor suppressor (34) (shown: chr11:125,363,009-125,443,355, hg19). **(B)** Insight into
 279 regulation of the *FADS* locus, a region under positive selection in geographically-diverse human
 280 populations. A candidate enhancer (i) lies between the fatty acid desaturases *FADS2* and *FADS3*, which
 281 are oriented tail-to-tail, but targeting it only decreased expression of the neighboring *FADS1* (shown:
 282 chr11:61,566,537-61,660,119, hg19). **(C)** crisprQTLs fall close to target genes. Distribution of distances
 283 between the crisprQTLs' candidate enhancers and: their target eCRISPR gene's TSS (top), the TSS of
 284 whatever K562-expressed gene is closest (middle), or the TSS of any gene within 1 Mb (bottom). Plotted
 285 with respect to gene orientation. **Fig. S8** presents a zoomed in view of +/- 20 Kb from the TSS. Of note,

candidate enhancers within 1 Kb of any gene body, *i.e.* including all exonic or intronic candidate enhancers, were excluded in library design. **(D)** 111 of 148 crisprQTLs do not target the most proximal K562-expressed gene. Genes are ranked by their distance to the candidate enhancer (1 = closest, 2 = second closest, *etc.*). **(E)** crisprQTL mapping is better powered to detect regulatory effects on highly expressed genes (expression = mean reads/cell in the entire 47,650 cell dataset). **(F)** crisprQTLs tend to fall in enhancer-associated ChIP-seq peaks that show stronger signals. All ChIP-seq peaks that overlap the 1,119 candidate enhancers were divided into quintiles of strength, defined as the average enrichment in ChIP-seq peak region (0 = no such peak overlaps the candidate enhancer, 1 = lowest, 5 = highest). Histograms of the proportion of each 1,119-quintile that were called as crisprQTLs are shown. Red = P -value < 0.005 for independent logistic regression for predicting a candidate enhancer as a crisprQTL based on this peak type.

One concern was whether multiple gRNAs within a single cell would dilute the efficacy of CRISPRi. To evaluate this, we conducted a biological replicate experiment with the same gRNA library but at a low MOI (41,284 cells, median 1 gRNA/cell, standard deviation=1.6, interquartile range = 1, **Fig. S6A**). The effect sizes identified in the high MOI experiment were well correlated with the effect sizes for the same gRNAs in the low MOI experiment (Spearman's rho: 0.826, slope: 0.9995, **Fig. 3G**). However, at the same threshold corresponding to a 10% empirical FDR, only 265 TSSs and 20 crisprQTLs were identified in the low MOI experiment (**Fig. S6B-D**), validating the substantial increase in power resulting from our multiplexed perturbation framework (**Fig. 1B**).

Even though this experiment solely targeted regulatory elements with CRISPRi, *trans* effects are still possible, *e.g.* as the secondary consequences of *cis* downregulation of a transcription factor. We therefore repeated our association testing, but evaluated each perturbation (TSS or candidate enhancer) against all K562-expressed genes located >1 Mb from the target gene or on another chromosome. Once again setting a threshold at an empirical FDR of 10% based on equivalent tests on NTCs, we identified 15 *cis* crisprQTLs that are also associated with downregulation of 1+ genes >1 Mb away or on other chromosomes (**Fig. S7A**). In addition, targeting 130 of the TSSs (targeting of which we previously associated with downregulation of the cognate gene) also resulted in apparent *trans* effects, with 81 (62%) of these influencing >1 distally located gene (**Fig. S7A**). For example, targeting the TSS of *MRPL33* successfully repressed both *MRPL33* and 112 other genes. As *MRPL33* encodes the 39S large subunit of mitochondrial ribosomes, affected genes were correspondingly most enriched for electron transport chain and mitochondrial pathways (strongest ten q -values).

Understanding the regulatory landscape of the human genome requires that we validate and identify the target genes of vast numbers of candidate enhancers, as nominated by biochemical marks and eQTL mapping. crisprQTL mapping has the potential to meet this challenge. In a

326 single experiment, we evaluated 17,584 potential *cis* regulatory relationships involving 1,119
327 candidate enhancers and 5,751 expressed genes. In contrast, nine recently published CRISPR
328 screens of noncoding sequences cumulatively studied regulatory effects on a total of 17 genes
329 (4–12). By delivering a median of 15 perturbations to each of 47,650 cells, this experiment was
330 powered equivalently to a ‘one gRNA per cell’ experiment involving 715,000 single cell
331 transcriptomes. Such an experiment would be amongst the largest scRNA-seq datasets generated
332 to date and therefore our framework represents enormous savings in cost and labor. Of note, one
333 recent study used scRNA-seq as a readout for the effects of CRISPR-based perturbations of 71
334 candidate regulatory elements on ~100 genes in seven genomic regions (35). However, its power
335 and scope was limited by a low MOI (**Fig. 1B**) and a gRNA barcoding strategy that suffers from
336 a ~50% rate of template switching (18).

337

338 Although it may be surprising that *cis* changes in gene expression were identified for only 11%
339 of the 1,119 candidate enhancers tested, there are several potential caveats to bear in mind: a)
340 Not all enhancers may be susceptible to CRISPRi-KRAB perturbation; b) gRNAs may be
341 variably effective in targeting enhancers; c) We are underpowered to detect effects on lowly
342 expressed genes; d) Shadow enhancers or other redundancies might mask some effects; e) Some
343 enhancers might be required for the initial establishment rather than maintenance of gene
344 expression; f) We did not comprehensively survey the noncoding landscape surrounding each
345 gene, and the marks we used to define candidate enhancers may be excluding some classes of
346 distal regulatory elements. These caveats are respectively addressable in the future by using other
347 epigenetic modifiers or nuclease-active Cas9, by using more gRNAs per candidate enhancer, by
348 performing better powered experiments via more scalable scRNA-seq methods (36), by
349 combinatorial perturbation of selected loci (35), by using cell models of differentiation, and by
350 densely tiling selected loci with perturbations.

351

352 However, the fact that our crisprQTLs are predicted by the strength of enhancer-associated marks
353 (e.g. H3K27ac, p300) supports the assertion that we are identifying *bona fide* enhancers, and
354 simultaneously weakens the case for elements that were negative. Also, our study provides
355 relatively unbiased insight into key properties of mammalian enhancers, e.g. the distribution of
356 distances between enhancers and their target genes. A full understanding of the precise rules
357 governing enhancer-promoter choice is a topic of great interest, and will be facilitated by the
358 identification of more crisprQTLs.

359

360 A limitation of crisprQTL mapping as implemented here relates to the resolution of CRISPRi (on
361 the order of a 1-2 Kb distribution of heterochromatin silencing as induced by KRAB (15)). In the
362 future, this can potentially be improved upon by adapting crisprQTL mapping to use single or
363 pairs of gRNAs with nuclease-active Cas9 to disrupt or delete candidate enhancers at the
364 sequence level. A separate concern is whether high MOI transduction is inducing a cellular

365 inflammatory response, and therefore biasing discovery. However, although some genes with
366 roles in inflammation are amongst our eCRISPR genes (*e.g.* *NMU*, *IL6*), we did not observe
367 pathway-level enrichment. Moreover, effect sizes in our high MOI and low MOI experiments
368 were very well-correlated.
369

370 A key strength of crisprQTL mapping over conventional eQTL studies is that our framework is
371 not limited by natural genetic variants nor by linkage disequilibrium. However, a weakness is
372 that we are only achieving tens of perturbations in each cell, as opposed to the millions of genetic
373 variants in each human. What are the limits to greater multiplexing of CRISPR-mediated
374 perturbations? It will likely be possible to push the effective MOI per cell higher than we have
375 have here, *e.g.* by repeated rounds of transduction. Coupled with further engineering (*e.g.*
376 multiple gRNAs per lentivirus, or alternative delivery systems), it may be possible to introduce
377 hundreds of ‘unlinked’ perturbations per cell. The other avenue to boost power is to simply
378 profile more cells, which may be facilitated by adapting crisprQTL mapping to more scalable
379 single cell transcriptome profiling schemes such as combinatorial indexing (36).
380

381 Overall, we anticipate that crisprQTL mapping will enable the validation and target gene
382 discovery of large numbers of candidate enhancers in cell lines and primary tissues. It may also
383 facilitate the fine mapping of causal regulatory variants and their target genes within haplotypes
384 implicated in common disease. To date, ENCODE has catalogued 1.3 million candidate
385 regulatory elements based on biochemical marks (<http://screen.umassmed.edu/>). Although
386 directly testing all of these elements might not be practical in the near term, the characteristics of
387 enhancers validated here will inform predictive model-driven selection of additional candidates
388 to enrich for crisprQTLs in future screens. Such iterative testing and prediction may be critical
389 for fully elucidating *cis*-regulatory architecture in diverse cell types and throughout the human
390 genome.
391
392

393 **References and Notes:**

- 394
1. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature*. **489**, 57–74 (2012).
 - 395
 - 396 2. E. E. Schadt *et al.*, Genetics of gene expression surveyed in maize, mouse and man. *Nature*. **422**, 297–302 (2003).
 - 397
 - 398 3. N. Risch, K. Merikangas, The future of genetic studies of complex human diseases. *Science*. **273**, 1516–1517 (1996).
 - 399
 - 400 4. N. E. Sanjana *et al.*, High-resolution interrogation of functional elements in the noncoding genome. *Science*. **353**, 1545–1549 (2016).
 - 401
 - 402 5. M. Gasperini *et al.*, CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions. *Am. J. Hum. Genet.* **101**, 192–205 (2017).
 - 403
 - 404
 - 405 6. M. C. Canver *et al.*, BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature*. **527**, 192–197 (2015).
 - 406
 - 407 7. G. Korkmaz *et al.*, Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat. Biotechnol.* **34**, 192–198 (2016).
 - 408
 - 409 8. Y. Diao *et al.*, A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening. *Genome Res.* **26**, 397–405 (2016).
 - 410
 - 411 9. N. Rajagopal *et al.*, High-throughput mapping of regulatory DNA. *Nat. Biotechnol.* **34**, 167–174 (2016).
 - 412
 - 413 10. T. S. Klann *et al.*, CRISPR-Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nat. Biotechnol.* **35**, 561–568 (2017).
 - 414
 - 415 11. C. P. Fulco *et al.*, Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science*. **354**, 769–773 (2016).
 - 416
 - 417 12. Y. Diao *et al.*, A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat. Methods*. **14**, 629–635 (2017).
 - 418
 - 419 13. M. Morley *et al.*, Genetic analysis of genome-wide variation in human gene expression. *Nature*. **430**, 743–747 (2004).
 - 420
 - 421 14. B. E. Stranger *et al.*, Patterns of cis regulatory variation in diverse human populations. *PLoS Genet.* **8**, e1002639 (2012).
 - 422
 - 423 15. P. I. Thakore *et al.*, Highly specific epigenome editing by CRISPR-Cas9 repressors for

- 424 silencing of distal regulatory elements. *Nat. Methods.* **12**, 1143–1149 (2015).
- 425 16. P. Datlinger *et al.*, Pooled CRISPR screening with single-cell transcriptome readout. *Nat.*
426 *Methods.* **14**, 297–301 (2017).
- 427 17. B. Chen *et al.*, Dynamic imaging of genomic loci in living human cells by an optimized
428 CRISPR/Cas system. *Cell.* **155**, 1479–1491 (2013).
- 429 18. A. J. Hill *et al.*, On the design of CRISPR-based single-cell molecular screens. *Nat.*
430 *Methods.* **15**, 271–274 (2018).
- 431 19. A. Wallrapp *et al.*, The neuropeptide NMU amplifies ILC2-driven allergic lung
432 inflammation. *Nature.* **549**, 351–356 (2017).
- 433 20. C. S. N. Klose *et al.*, The neuropeptide neuromedin U stimulates innate lymphoid cells and
434 type 2 inflammation. *Nature.* **549**, 282–286 (2017).
- 435 21. V. Cardoso *et al.*, Neuronal regulation of type 2 innate lymphoid cells via neuromedin U.
436 *Nature.* **549**, 277–281 (2017).
- 437 22. J. E. Gambone, S. S. Dusaban, R. Loperena, Y. Nakata, S. E. Shetzline, The c-Myb target
438 gene neuromedin U functions as a novel cofactor during the early stages of erythropoiesis.
439 *Blood.* **117**, 5733–5743 (2011).
- 440 23. X. Qiu *et al.*, Single-cell mRNA quantification and differential analysis with Census. *Nat.*
441 *Methods.* **14**, 309–315 (2017).
- 442 24. S. S. P. Rao *et al.*, A 3D map of the human genome at kilobase resolution reveals principles
443 of chromatin looping. *Cell.* **159**, 1665–1680 (2014).
- 444 25. E. Kang *et al.*, Interaction between FEZ1 and DISC1 in regulation of neuronal development
445 and risk for schizophrenia. *Neuron.* **72**, 559–571 (2011).
- 446 26. D. T. Bryant *et al.*, Disruption to schizophrenia-associated gene Fez1 in the hippocampus of
447 HDAC11 knockout mice. *Sci. Rep.* **7**, 11900 (2017).
- 448 27. K. Ye, F. Gao, D. Wang, O. Bar-Yosef, A. Keinan, Dietary adaptation of FADS genes in
449 Europe varied across time and geography. *Nat Ecol Evol.* **1**, 167 (2017).
- 450 28. M. Fumagalli *et al.*, Greenlandic Inuit show genetic signatures of diet and climate
451 adaptation. *Science.* **349**, 1343–1347 (2015).
- 452 29. A. Ameur *et al.*, Genetic adaptation of fatty-acid metabolism: a human-specific haplotype
453 increasing the biosynthesis of long-chain omega-3 and omega-6 fatty acids. *Am. J. Hum.*
454 *Genet.* **90**, 809–820 (2012).
- 455 30. E. Eisenberg, E. Y. Levanon, Human housekeeping genes, revisited. *Trends Genet.* **29**,

- 456 569–574 (2013).
- 457 31. M. Ganapathi *et al.*, Comparative analysis of chromatin landscape in regulatory regions of
458 human housekeeping and tissue specific genes. *BMC Bioinformatics*. **6**, 126 (2005).
- 459 32. L. Väremo, J. Nielsen, I. Nookaew, Enriching the gene set analysis of genome-wide data by
460 incorporating directionality of gene expression and combining statistical hypotheses and
461 methods. *Nucleic Acids Res.* **41**, 4378–4391 (2013).
- 462 33. D. Merico, R. Isserlin, O. Stueker, A. Emili, G. D. Bader, Enrichment map: a network-based
463 method for gene-set enrichment visualization and interpretation. *PLoS One*. **5**, e13984
464 (2010).
- 465 34. C. N. Mork, D. V. Faller, R. A. Spanjaard, Loss of putative tumor suppressor EI24/PIG8
466 confers resistance to etoposide. *FEBS Lett.* **581**, 5440–5444 (2007).
- 467 35. S. Xie, J. Duan, B. Li, P. Zhou, G. C. Hon, Multiplexed Engineering and Analysis of
468 Combinatorial Enhancer Activity in Single Cells. *Mol. Cell.* **66**, 285–299.e5 (2017).
- 469 36. J. Cao *et al.*, Comprehensive single-cell transcriptional profiling of a multicellular
470 organism. *Science*. **357**, 661–667 (2017).
- 471 37. C. Trapnell *et al.*, Differential analysis of gene regulation at transcript resolution with
472 RNA-seq. *Nat. Biotechnol.* **31**, 46–53 (2013).
- 473 38. T. Wang *et al.*, Identification and characterization of essential genes in the human genome.
474 *Science*. **350**, 1096–1101 (2015).
- 475 39. M. A. Horlbeck *et al.*, Compact and highly active next-generation libraries for
476 CRISPR-mediated gene repression and activation. *eLife*. **5** (2016), doi:10.7554/eLife.19760.
- 477 40. A. McKenna, J. Shendure, FlashFry: a fast and flexible tool for large-scale CRISPR target
478 design (2017), , doi:10.1101/189068.
- 479 41. N. L. Bray, H. Pimentel, P. Melsted, L. Pachter, Erratum: Near-optimal probabilistic
480 RNA-seq quantification. *Nat. Biotechnol.* **34**, 888 (2016).
- 481 42. H. Pimentel, N. L. Bray, S. Puente, P. Melsted, L. Pachter, Differential analysis of RNA-seq
482 incorporating quantification uncertainty. *Nat. Methods*. **14**, 687–690 (2017).
- 483
- 484

485 **Acknowledgements:** We thank the entire Shendure, Trapnell, and Ahituv labs, in particular
486 Darren Cusanovich, Anh Leith, Dana Jackson, Gregory Findlay, Lea Starita, Delasa
487 Aghamirzaie, Aaron McKenna, Malte Spielmann, Vikram Agarwal, Jason Klein, Seungsoo Kim,
488 Xiaojie Qui, Jonathan Packer, Rajiv McCoy, Jacob Tome, Silvia Domcke, Fumitaka Inoue, Orry
489 Elor, and Nadja Makki. We additionally thank James Ousey, Kyuho Han, Michael Bassik, Martin
490 Wohlfahrt, Madison Norton, the staff of CustomArray, Inc. and Martin Kircher. This work was
491 supported by awards from the NIH (UM1HG009408 to NA and JS and DP1HG007811 to JS)
492 and training awards from the National Science Foundation and National Institutes of Health
493 (Graduate Research Fellowship and 5T32HG000035 to MG). JS is an Investigator of the Howard
494 Hughes Medical Institute.
495
496

497 **Materials and Methods**

498

Cell Lines and Culture

499 K562 cells expressing dCas9-BFP-KRAB (Addgene #46911) were a gift of the Bassik lab, and
500 cultured in RPMI 1640 + L-Glutamine (Gibco), supplemented with 10% fetal bovine serum
501 (Rocky Mountain Biologicals) and 1% penicillin-streptomycin (GIBCO).

502

gRNA-library design

503 *Note about terminology:* A gRNA-group is defined as all the gRNA that are targeting the same
504 candidate enhancer or positive control site (usually 2). To note, all novel TSS and candidate
505 enhancer targeting gRNA-groups are referred to as “perturbative gRNA-groups”, whereas all
506 others are referred to as “control gRNA-groups. The 100 candidate enhancer pilot library will
507 be referred to as the “100plex library”, and the 1,119 candidate enhancer at-scale library will be
508 referred to as the “1,119plex library”.

509

510 *100 Candidate Enhancer Pilot Library (100plex library, Table S1):*

511

512 *Picking candidate enhancers:* Regions of open chromatin were defined from K562 DNase-seq
513 narrowPeaks (ENCSR000EKS) more than 1 Kb away from any gene (GENCODE March 2017
514 v26lift37). Using bedtools, these regions were intersected with K562 GATA1 ChIP-seq
515 narrowPeaks (ENCSR000EFT, lifted to hg19), H3K27ac ChIP-seq narrowPeaks
516 (ENCSR000AKP, lifted to hg19), and K562 HiC domains (24) that contained at least 1 gene in
517 the top 10% FPKM of K562 most expressed genes that were also more highly expressed in K562
518 (ENCSR000EYO) than in H1-hESC (ENCSR000EYP) or HepG2 (ENCSR000EYR) as called by
519 Cuffdiff (q -value < 0.05)(37) but were not K562 essential genes (38).

520

521 *TSS positive control gRNA:* Each of the ‘top expressed’ K562 genes in the utilized HiC domains
522 were targeted by two TSS-proximal gRNA, from protospacers as best scored by the empirical
523 and predicted scores of the hCRISPRiv2 library (39). To note, in further analysis, though its TSS
524 was targeted ‘TOMM6’ was excluded from analysis as it was not detected in the K562
525 scRNA-seq data, though it was in the bulk RNA-seq dataset (ENCSR000EYO).

526

527 *Candidate enhancer gRNA:* NGG-protospacers within these candidate enhancers were scored
528 using default parameters of FlashFry (40), and the two top quality-scoring per region were used
529 as spacers in the 100plex gRNA library (scores prioritized by Doench2014OnTarget > Hsu2013
530 > Doench2016CDFScore > otCount, two gRNA per gRNA-group).

531

532 *Non-targeting control (NTC) gRNA:* 11 scrambled-sequence spacers with no targets in the
533 genome and 11 protospacers targeting 6 gene-devoid regions of the genome (hg19
534 chr4:25697737-25700237, chr5:12539119-12541619, chr6:23837183-23839683,
535 chr8:11072736-11075236, chr8:23768553-23771053, chr9:41022164-41024664) were chosen as
536 evaluated by Benchling's CRISPR tool. These were randomly paired to create a gRNA-group.
537 Additionally, a gRNA targeting the CAG promoter was included for a total of 13 gRNA groups.
538

539 *Distal enhancer positive control gRNA:* 15 gRNA targeting the *HBE1* TSS, and HS1-4 of the
540 Globin LCR were chosen as validated from (10, 35). These were paired based on target-site to
541 create a gRNA-group.
542

543 *Additional gRNA:* We also included a small set of gRNA that targeted variants that were
544 functional in massively parallel reporter assays (Ulirsch et al. 2016). However, as only
545 intragenically targeting gRNA created a significant effect on its host gene in this dataset (and
546 thus were potentially confounded by CRISPRi interfering with transcription via a steric rather
547 than epigenetic mechanism), we did not report on these results here.
548

549 *1,119 Candidate Enhancer At-Scale Library (1,119plex library, Table S2):*
550

551 *Picking candidate enhancer regions:* K562 DNase-seq narrowPeaks (ENCSR000EKS) < 1 Kb
552 away from any gene (GENCODE March 2017 v26lift37) were bedtools-intersected with K562
553 HiC domains (24) that contained at least one of the top 10% highest express genes in the 6,806
554 cell pilot study K562 data set. The remaining regions were tested for intersections with K562
555 GATA1 ChIP-seq narrowPeaks (ENCSR000EFT, lifted to hg19), H3K27ac ChIP-seq
556 narrowPeaks (ENCSR000AKP, lifted to hg19), RNA Pol II ChIP-seq narrowPeaks
557 (ENCSR000AKY), and EP300 ChIP-seq narrowPeaks (ENCSR000EHI) (**Fig. 3A**).
558

559 *Candidate enhancer gRNA:* NGG-protospacers within these candidate enhancers were again
560 scored using default parameters of FlashFry (40), and the two top quality-scoring gRNA per
561 region were chosen as spacers to be used in the gRNA library (scores prioritized by
562 Doench2014OnTarget > Hsu2013 > Doench2016CDFScore > otCount). The same spacers from
563 the 100plex library that were putative candidate enhancers of *NMU* and *HBG1* were also added.
564

565 *TSS positive control gRNA:* 380 genes were randomly sampled from the highly-expressed genes
566 within the same HiC domain (as described above) and 2 gRNA were again chosen per gene from
567 spacers with the best empirical and predicted scores of the hCRISPRiv2 library (39). The gRNAs
568 targeting the TSS of *NMU* was again included for a total of 381 TSS targeted.
569

570 *NTC gRNA*: All original NTC were again included from the 100plex study. More were chosen
571 from 6 random regions of the hg19 genome (chr4:25697737-25700237,
572 chr5:12539119-12541619, chr6:23837183-23839683, chr8:11072736-11075236,
573 chr8:23768553-23771053, chr9:41022164-41024664) using FlashFry (40) to total 50 targeting
574 these gene-devoid regions of the genome. A further 50 were derived from the 11 scrambled
575 sequences used in the 100plex library and 39 NTC sampled from those recommended by (39).
576

577 *Distal enhancer positive control gRNA*: Also included were 14 of the gRNA from 100plex
578 library that targeted the globin LCR.
579

580 Of note, our initial FlashFry quality annotations did not label a small number of protospacers
581 with perfect repeat off-targets, permitting their inclusion in our library (81 of 2,238 spacers
582 ordered in the 1,119plex library; only 9 gRNA-groups with both spacers affected). gRNA-groups
583 with an impacted spacer were rare in our significant crisprQTLs. We also note that we still
584 expect these guides to target their intended site.
585

586 gRNA-library cloning, and transduction

587

588 The lentiviral CROP-seq gRNA-expression vector (16) was modified by Q5-Site Directed
589 Mutagenesis (New England BioLabs, F:5-acagcatagcaagtAAATAAGGCTAGTCCGTATC-3
590 R:5-ttccagcatagcttAAACAGAGACGTACAAAAAAG-3) to incorporate the previously
591 described gRNA-(F+E)-combined backbone optimized for CRISPRi (17)(18). Prepared vector
592 was digested with *BsmBI* (FastDigest Esp3I, Thermo Fisher Scientific), “filler” sequence
593 removed by gel extraction, and cleaned (Zymo Research DNA Clean & Concentrator-5) vector
594 without “filler” was used for all downstream cloning.
595

596 Protospacer libraries were ordered single stranded from Integrated DNA Technologies (100plex
597 library: as separate oligos and then pooled) or as a pool (1,119plex library: CustomArray Inc.,
598 each protospacer synthesized in triplicate in the pool). Pools were amplified
599 (5-atcttGTGGAAAGGACGAAACA-3, 5-acttgctaTGCTGTTCCAGC-3, Kapa Biosystems
600 HiFi Hotstart ReadyMix (KHF)) and purified amplicons (Zymo Research DNA Clean &
601 Concentrator-5) were cloned into CRISPRi-optimized CROP-seq vector prepared as described
602 above (NEBuilder® HiFi DNA Assembly Cloning Kit, NEB). Products were transformed into
603 Stable Competent *E. coli* (NEB C3040H) to produce >20 colonies per gRNA in the library.
604

605 The Fred Hutchinson Co-operative Center for Excellence in Hematology Vector Production core
606 produced all virus. Cells were transduced with varying titers of virus to achieve differing MOI
607 with 8 µg/mL polybrene. At 24 hours post-transduction, cells were spun and resuspended with
608 virus-free media. At a total 48 hours post-transduction, 2 µg/mL puromycin was added to the

609 culture, and changed to 1 µg/mL puromycin at the next passage for maintenance. A total of 10
610 days post transduction, cells were collected for scRNA-seq.
611

Single cell RNA sequencing

612
613 ~4000-8000 cells were captured per lane of a 10X Chromium device using 10X V2 Single Cell
614 3' Solution reagents (10X Genomics, Inc). A single lane was used for the 100plex experiment,
615 while six lanes were used for both the low and high MOI 1,119plex experiments. All protocols
616 were performed as per manufacturer's recommendations, except prior to the enzymatic shearing
617 step, a portion of full length cDNA was taken for PCR enrichment of gRNA-sequences off the
618 CRISPRi-optimized CROP-seq transcripts as described below.
619

620 Final libraries were sequenced on a NextSeq 500 with 75-cycle high-output kits (R1:26 I1:8 I2:0,
621 R2:57). The 100plex library lane was sequenced to 24.1 thousand median UMI per cell, and
622 44.4% sequencing saturation. The low MOI 1,119plex library lanes were sequenced to 16.7,
623 20.9, 21.8, 22.7, 23.4, 25.6 thousand median UMI per cell, at 16-25% sequencing saturation. The
624 high MOI 1,119plex library lanes were sequenced to 16.1, 18.5, 19.5, 19.8, 20.3, 20.6 thousand
625 median UMI per cell, at 16-20% sequencing saturation. All quality statistics are reported from
626 Cell Ranger defaults.
627

gRNA-transcript enrichment PCR

628
629 A three-step hemi-nested PCR reaction was performed to enrich gRNA sequences from the 3'
630 UTR of puromycin resistance gene transcripts produced by the CRISPRi-optimized CROP-seq
631 integrant. PCR was monitored by qPCR to avoid overamplification, and each reaction was
632 stopped immediately before it reached saturation.
633

PCR I:

634 10-13 ng of full-length 10x scRNA-seq cDNA were amplified in each 50 µl KHF reaction,
635 spiked with SYBR-Gold (Invitrogen) for qPCR monitoring (3% of all cDNA for the 100plex
636 experiment versus 10% of all cDNA for 1,119-plex experiments). Greater input appeared to lead
637 to less cells with incomplete gRNA assignments, as indicated by **Fig. 3B** versus the more
638 zero-inflated distribution of **Fig. S1A**.
639

640
641 F: U6_OUTER 5- TTTCCCATGATTCTTCATATTGC -3
642 R: R1_PCR1 5- ACACCTTTCCCTACACGACG-3
643

644 PCR 2: 4% of each previous reaction was brought forward, sample replicates pooled, cleaned
645 with 1x Agencourt AMPure XP beads (Beckman Coulter), and 1/25th of the cleaned pooled
646 product was amplified in each of six 50 µl reactions.
647

648 F: U6_INNER_with_P7_adapter
649 5-GTCTCGTGGCTCGGAGATGTGTATAAGAGACAGcTTGTGGAAAGGACGAAACAC
650 -3
651 R: R1-P5 5-AATGATAACGGCGACCACCGAGATCTACACTCTTCCCTACACGACG-3
652

653 PCR 3: The PCR 2 replicate reactions were pooled, 1x AMPure cleaned, and 1/25th of the
654 cleaned pooled product was amplified in each of six 50 µl reactions and products cleaned once
655 again via 1x Ampure.
656

657 F: 5- CAAGCAGAAGACGGCATACGAGATIIIIIIiiGTCTCGTGGCTCGG-3 (standard
658 NEXTERA P7 indexing primer)
659 R: R1-P5 again
660

661 Digital gene expression quantification

662 Sequencing data from each sample was processed using the Cell Ranger software package as
663 provided by 10x Genomics, Inc., to generate sparse matrices of UMI counts for each gene across
664 all cells in the experiment.
665

666 Each lane of cells was processed independently using cellranger count, aggregating data from
667 multiple sequencing runs. The 100plex library experiment was processed with cellranger 1.3.1
668 and the 1,119plex library experiments were each processed with cellranger 2.0.2.
669

670 *Definition of genes well-expressed or ‘detectably expressed’ in K562:* Unless otherwise noted,
671 genes were defined as well expressed or detectably expressed in K562 if they had at least one
672 read in 0.525% of cells (250 cells) in the high-MOI 47,650-cell dataset (12,984 total genes).
673

674 *Assigning genotypes to cells:* gRNAs were assigned to cells as performed in Hill,
675 McFaline-Figueroa et al. (18); briefly, sequences corresponding to the gRNA-containing
676 CRISPRi-optimized CROP-seq transcripts are extracted from the cellranger position sorted BAM
677 file after running our custom indexed libraries through the cellranger pipeline to tag reads with
678 corrected cell barcodes and UMIs. gRNA sequences are extracted and corrected to the library
679 whitelist within an edit distance of two ,and gRNA-cell pairs are tracked when a valid cell
680 barcode and UMI are both assigned to the read. Likely chimeric reads are detected and removed
681 to reduce noise in the assignments as previously described. As in Hill, McFaline-Figueroa *et al.*
682

683 (18), we utilized thresholds to set minimum acceptable values for the total reads for a gRNA-cell
684 pair and for the proportion of all CROP-seq transcript reads accounted for by each gRNA
685 observed in a cell to distinguish noise from real assignments. Here, given the larger number of
686 guides contained in each cell, we find that UMI counts provide a much cleaner distribution than
687 read counts and have used UMI counts in all calculations. For the 100plex library, we used
688 cutoffs of 0.035 and 5 UMIs as the thresholds for the fraction of CRISPRi-optimized CROP-seq
689 transcripts attributed to a guide and the total number of observed UMIs for a gRNA-cell pair,
690 respectively. For the 1,119plex library experiments we used 0.01 and 5 UMI in both our low and
691 high MOI for each of these thresholds. Only cell barcodes that appear in the set of passing cells
692 output by cellranger, which imposes an automated threshold on the total UMIs observed in cells,
693 are carried forward in downstream analysis.
694

695 *Differential expression tests:* In our *cis* analyses, we tested each perturbing gRNA-group against
696 a gene within 1 Mb of the gRNA. These gRNA-gene pairs were identified by using bedtools to
697 intersect the DHSs targeted by the gRNA library with 1 Mb windows in either direction of TSS
698 annotations from GENCODE March 2017 v26lift37 (total of 2 Mb, centered around the TSS). In
699 our *trans* analysis, all gRNA-groups were paired with all genes that were defined as expressed in
700 K562. In both *cis* and *trans* analyses, NTCs were tested against any genes used to test
701 perturbing-gRNAs.
702

703 For each gRNA-group we assigned a label of “1” to cells that contained a gRNA belonging to
704 that group and a label of “0” to all other cells in the dataset. Monocle2 was used to perform a
705 differential expression test, using the negative binomial family, over this categorical label to find
706 differentially expressed genes between these two groups. Due to its support of complex model
707 formulas, Monocle2 does not provide model coefficients as part of the differential expression
708 results. We created a modified version of the differentialGeneTest function and associated helper
709 functions that return both the intercept term and the coefficient of the group assignment to
710 facilitate more robust prioritization and characterization of hits from our screen. The negative
711 binomial family uses log as the link-function, so we can calculate the initial expression level as
712 $\exp(\text{intercept})$, and the fold change in expression between the two groups as
713 $\exp(\text{group_coefficient} + \text{intercept}) / \exp(\text{intercept})$. We verified data from our power simulations
714 that the appropriate effect sizes can be obtained with this method using the coefficients output by
715 VGAM.
716

717 Calling hits from differential expression test results

718
719 In both the *cis* and *trans* analyses, all differential expression test results were filtered to only
720 K562 expressed genes.
721

722 Tests with two sources of potential false positives were excluded:

- 723 1) We identified inflation of NTCs when testing them against genes highly impacted by
724 perturbing-gRNA in our library (for example, NTCs associated with targets of our TSS
725 and globin LCR controls). This was due to subtle yet detectable nonrandom associations
726 of gRNA-groups with other gRNA-groups across cells. To exclude this source of
727 inflation, we used Fischer's exact test to identify when an NTC was nonrandomly
728 assorted with a perturbing-gRNA (q -value < 0.01 & odds ratio > 1). Then, any test of an
729 NTC against a gene within 1 Mb of that gRNA's gRNA-group was excluded from further
730 analytical steps. This was used to exclude such interactions in both the *cis* and *trans*
731 analyses.
- 732 2) We noted Monocle was susceptible to inflating P -values when a gene was highly
733 expressed but only in few cells. To avoid this problem, we excluded 23 outlier genes that
734 were expressed in $< 20,000$ cells in the high-MOI 47,650-cell dataset and with
735 $\log_{10}(\text{total reads / cells with a read}) > 0.2$ greater than predicted by a spline fit generated
736 via `smooth.spline()` with `spar=0.85` to limit overfitting (**Fig. S9**).

737 Remaining tests were filtered to those that decreased expression of the target gene. Then, a 10%
738 empirical FDR was defined as the P -value at which the proportion of passing NTC-tests/total
739 NTC-tests was 10% of the proportion of passing candidate enhancer tests/total candidate
740 enhancer tests. *cis* pilot P -value = 5.52e-11; *cis* high MOI P -value = 5.82e-05; *trans* high MOI
741 P -value = 8.45e-13; *cis* low MOI P -value = 7.87e-05.

742

Aggregate analysis of crisprQTLs

743

744 *Distance between enhancer and eCRISPR-gene:* Distance was calculated between the
745 GENCODE March 2017 v26lift37 annotated TSS of the perturbed gene and the leftmost gRNA
746 (if targeting a candidate enhancer) or the GENCODE-annotated TSS of the originally targeted
747 transcript (if targeting a TSS).

748

749 *ChIP-seq strength quintile analysis and logistic regression classifier:* 1,119 candidate enhancers
750 were bedtools-intersected with ChIP-seq of histone-associated marks (all narrowPeak; CCNT2
751 ENCSR000DOA, H3K27ac ENCSR000AKP, ENCSR038DJJ SMAD1, H3K4me1
752 ENCSR000AKS, RNA Pol II ENCSR388QZF, TAL1 ENCSR106FRG, EP300 ENCSR000EGE,
753 ENCSR000EFT GATA1, ENCSR000EWG GATA2 (1)), broken into quintiles of the 7th
754 "signalValue" column (representing peak strength, usually representing overall average
755 enrichment in the region), and the rates of crisprQTL identified in each quintile were used. The
756 signalValue column of each dataset was also used to fit both independent and multivariate
757 logistic regression classifiers using the `glm()` function with binomial family. Fivefold
758 cross-validation was

759

760

761 *Functional annotation enrichment:* We used the Piano package (32) to perform functional
762 annotation enrichment from the 'all pathways' Gene Ontology
763 (http://download.baderlab.org/EM_Genesets/June_20_2014/Human/June_20_2014_versions.txt).

765 In the *cis* analysis, the 5,751 K562-expressed genes within 1 Mb of a perturbing-gRNA were
766 used as our background dataset, and 105 eCRISPR-genes randomly sampled from genes with
767 expression greater than the mean of our 105 eCRISPR-genes was used as the comparison set of
768 “expression matched controls”. In the *trans* analysis, K562 expressed genes were used as a
769 background. In both analyses, the most significant terms were described in the text.
770

771 Replication of crisprQTL outside of multiplexed experiment

772
773 To replicate the CRISPRi phenotype outside of the single-cell pooled screen, we prepared small
774 pools of gRNAs targeting e-NMU (hg19 chr4:56595855-56597095), e-PRKCB
775 (chr16:23833191-23833694), e-PTGER3 (chr1:71140581-71141997), e-GYPC
776 (chr2:127402697-127403360), e-HMGA1 (chr6:34191315-34192504), or the TSSs of their
777 respective eCRISPR target genes (**Table S6**). These small gRNA pools were cloned into
778 CRISPRi-optimized-CROP-seq (as described above). Lentiviral preps from these gRNA pools
779 were transduced at low MOI into the K562-dCas9-BFP-KRAB line, and cultured for 10 days
780 under puromycin selection before two replicates of RNA were collected. Bulk RNA-seq libraries
781 were prepared from each replicate via a TruSeq RNA kit (400 ng input), and sequenced on a
782 NextSeq 500 with two 150-cycle kits in mid output mode. Gene-level quantifications and
783 differential expression tests were performed via kallisto (41) and sleuth (42). The quantification
784 uncertainty for each sample was estimated by the kallisto and sleuth bootstrapping procedure
785 (n=100 bootstraps).

786

787

788 Power Simulations

789

790 In order to predict the impact of multiplexing on the power of crisprQTL experiments, we
791 developed a simulation framework. First, using single-cell RNA-seq data collected from the
792 1,119plex 47,650 K562 cells, we estimated a dispersion function that relates the mean expression
793 of a gene to its dispersion estimate (one of the two parameters required for the negative binomial
794 distribution) calling the Monocle2 functions estimateSizeFactors and estimateDispersions. This
795 function is typically used in differential expression testing to shrink dispersion estimates, but
796 here we use it to estimate dispersion values for simulated transcripts. This dispersion function is
797 then extracted from the CellDataset object output by Monocle2 and used as input to our
798 simulations.

799

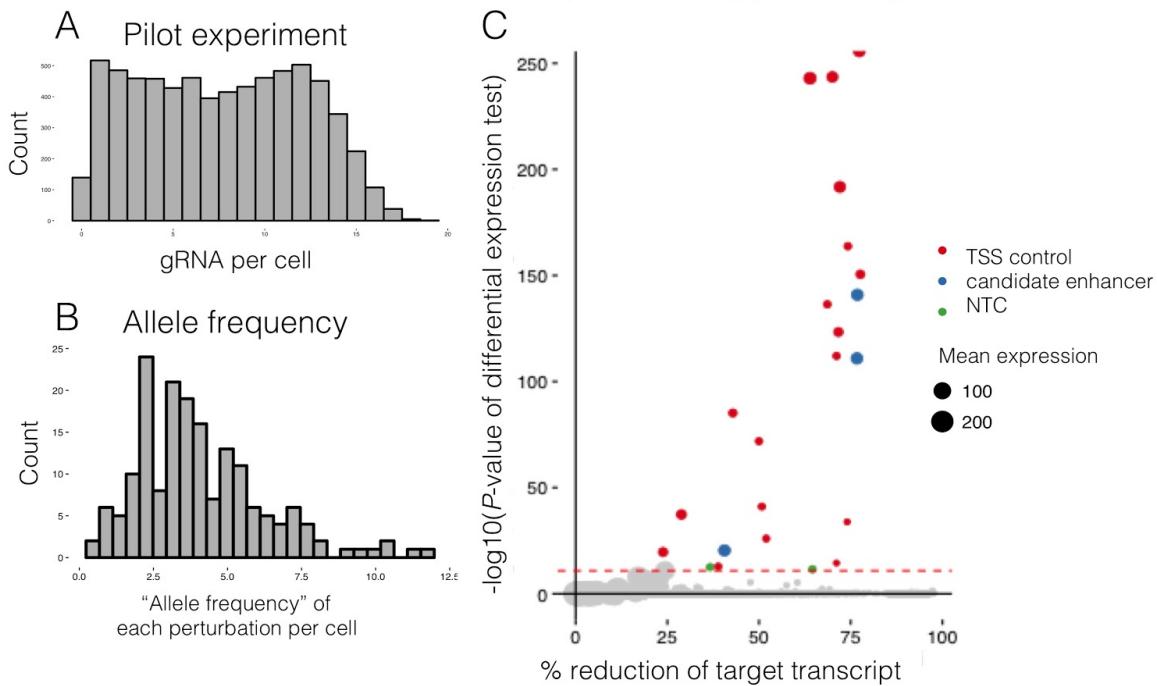
800 Next, we chose relevant ranges for each of the parameters varied in our simulation: the MOI,
801 total cell count, effect size (fraction repressed by CRISPRi), and mean expression level of the
802 gene being tested. By examining the range of expression values observed in our data, we chose
803 to simulate expression data for genes having mean expression values (size parameter of the
804 negative binomial distribution) of 0.01, 0.1, 0.32, 1.0, 3.16, and 10.0 UMIs (0.10, 0.32 and 1.00

805 used respectively as low, medium, and high in **Fig. 1B**) to provide a range of representative
806 values.
807

808 We simulated MOIs at several values from 0.3 to 50, a range which includes the MOIs estimated
809 from our own crisprQTL mapping screens. For each MOI, we calculate the expected number of
810 cells containing a given guide by assuming a Poisson distribution of lentiviral delivery,
811 zero-truncating the distribution to account for drug selection for cells that contain a guide
812 transcript, and rescaling the probability distribution of guide counts accordingly. Perfect library
813 uniformity was assumed to obtain the expected number of cells containing a given guide and the
814 number of cells that do not contain that guide. Effect sizes of CRISPRi repression were chosen
815 using estimates from the literature and were simulated at several values between 10% to 90%
816 percent repression of the average expression level of the target transcript (size parameter input to
817 the negative binomial distribution).
818

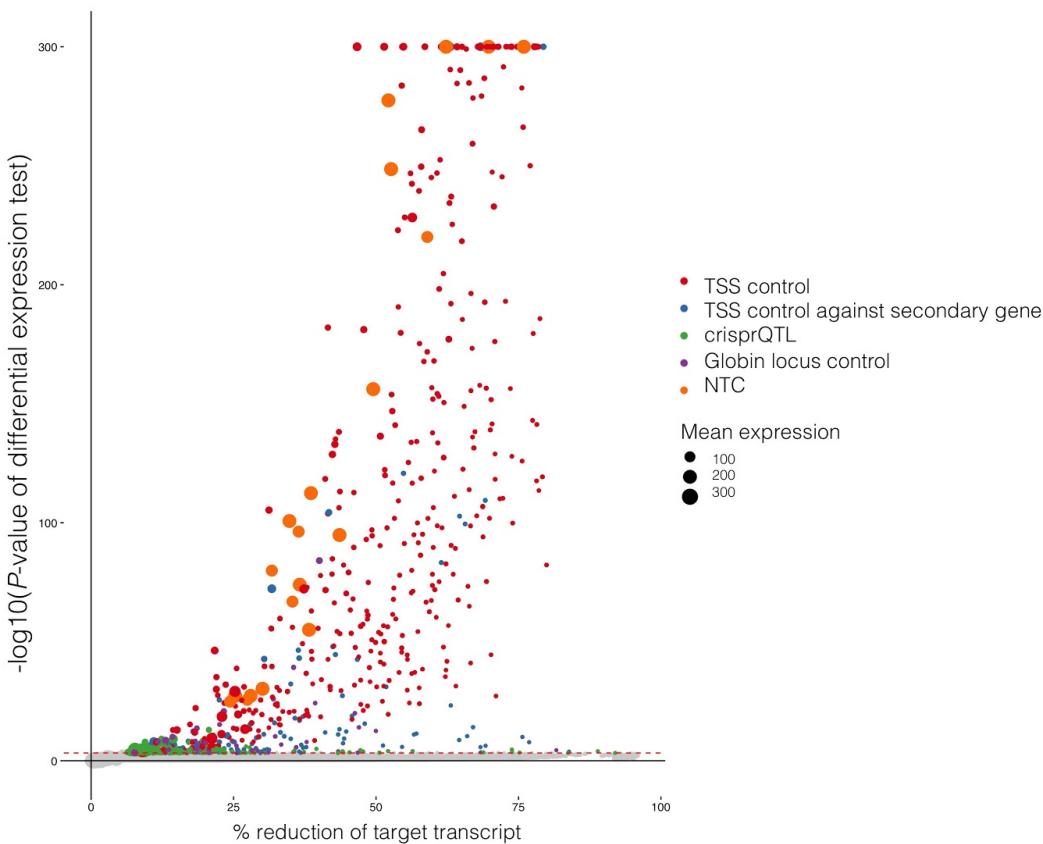
819 Finally, we simulated several values of total cells included in the experiment ranging from
820 35,000 to 300,000 cells (45,000 cells shown in **Fig. 1A**). Expression data from transcripts
821 corresponding to 100 samplings per set of parameters were generated for the populations of cells
822 containing the gRNA and not containing the gRNA respectively. Our expression data simulation
823 assumed a negative binomial distribution with the appropriate size parameter for the cells with
824 and without the gRNA, and a dispersion value estimated using the dispersion function described
825 above given the starting mean expression level being simulated. For each set of parameters, the
826 simulated transcripts were subjected to a differential expression test performed between cells
827 with and without the gRNA assigned using our modified version of the Monocle2 function
828 differentialGeneTest as described above (see Differential Expression Tests). *P*-values were
829 obtained and corrected assuming an average number of 20 tests per group in the library to
830 approximate the number of genes contained within 1 Mb on either side of each gRNA-group and
831 the impact of multiple testing. The rate of tests falling below a *q*-value of 0.05 were tabulated at
832 each set of parameters to make power curves.
833
834

835



836 **Fig. S1 | Extended data for 100 candidate enhancer pilot experiment.** (A) Number of gRNAs
837 identified per cell. Median 8 gRNA +/- 4.5 gRNA were identified per cell from targeted
838 amplification of the pre-fragmentation 10x cDNA library. However, the amount of cDNA input
839 used for this amplification was less than used in later experiments. (B) “Allele frequency” of
840 each perturbation in the cell population, *i.e.* the number of cells in which a given site was
841 targeted. Mean 4.1% +/- 2.2%. Sites targeted by >2 gRNAs not included here. (C) Distribution
842 of all differential expression tests (significance x effective size) that decrease test-gene
843 expression. $\log_{10}(q\text{-val})$ versus % the target transcript is reduced. Point size corresponds to
844 relative expression level. Dotted red line = *P*-value threshold used to call hits.
845

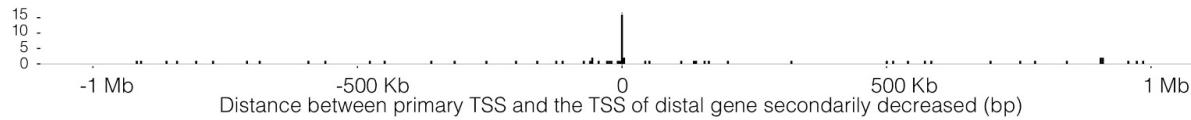
846
847
848



849 **Fig. S2 | Distribution of all differential expression tests for the 1,119 candidate enhancer**
850 **crisprQTL experiment.** Effect size vs. significance is plotted for all tests that showed a
851 decrease in test gene expression. $\log_{10}(P\text{-value})$ versus % target transcript is reduced. Point size
852 corresponds to relative expression level. Dotted red line = P -value threshold used to call hits.
853

854

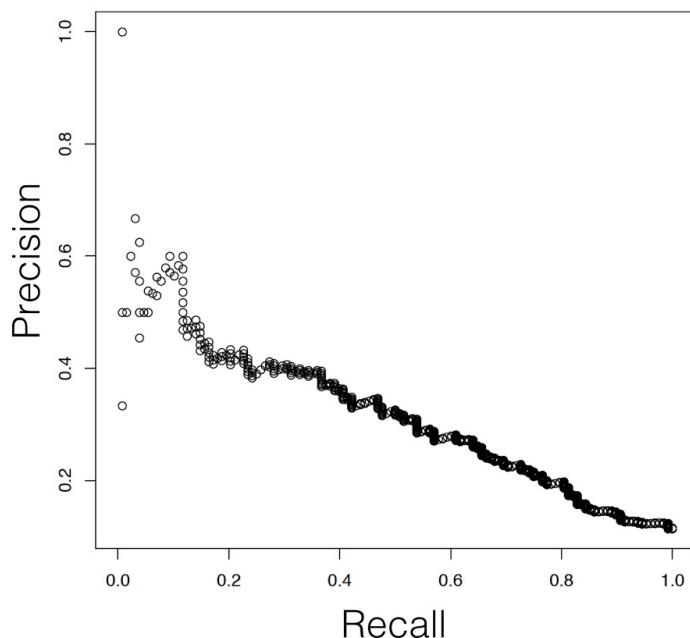
primary TSS “positive control” gene → secondary distal gene also reduced in expression



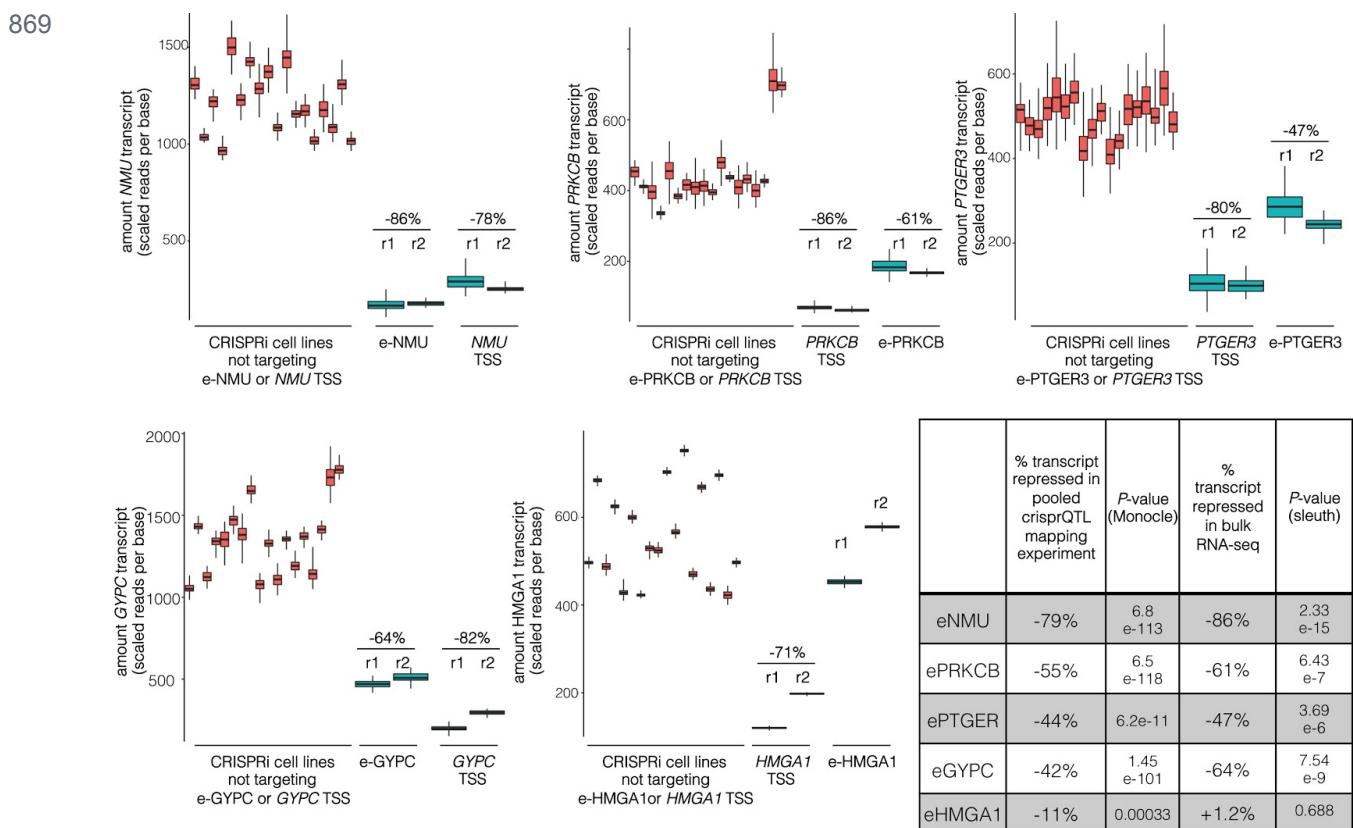
855

Fig. S3 | Distribution of distances between “positive control” TSSs and secondarily repressed genes. Distance between the 56 “positive control” targeted TSSs and the TSS of the secondary gene (over 1 Kb away) whose expression was simultaneously also decreased. Plotted with respect to primary target gene’s orientation.

859

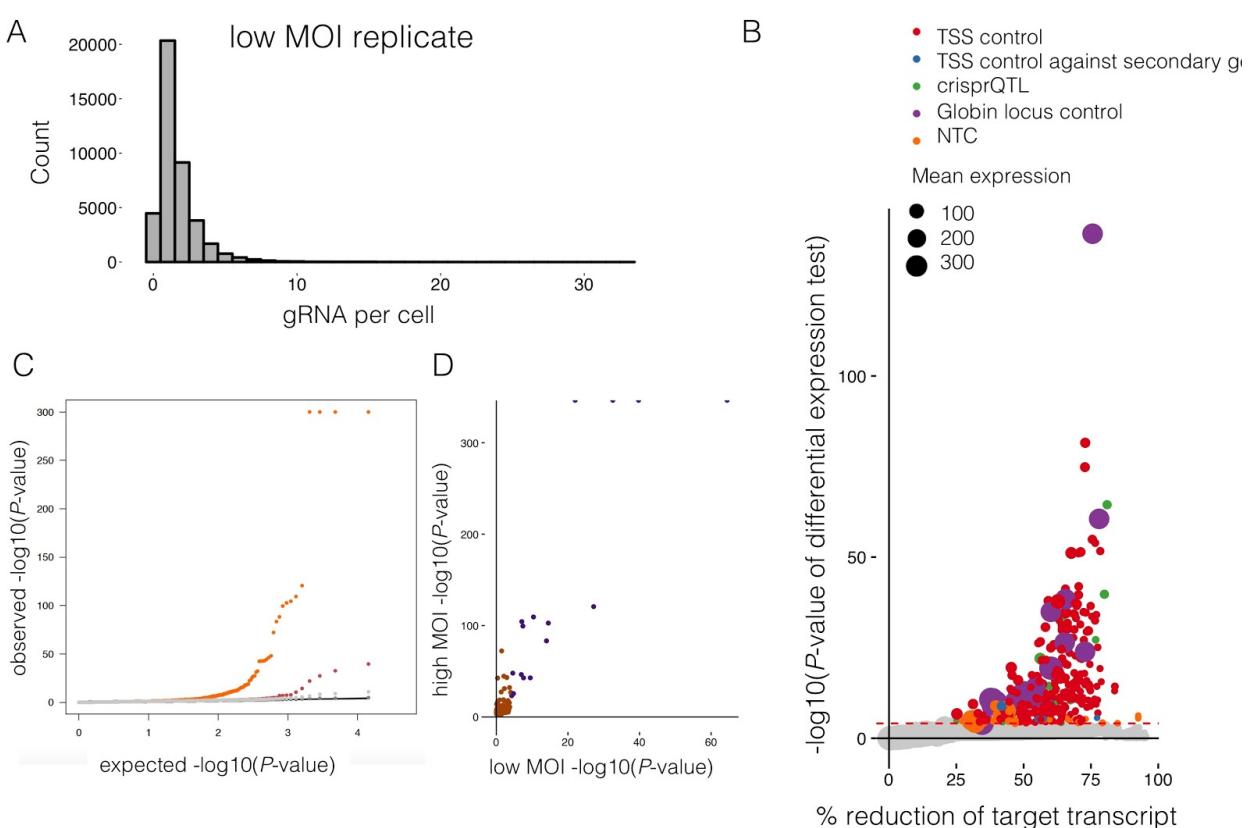


860 **Fig. S4 | Precision recall curve for a multivariate logistic regression classifier based on**
861 **ENCODE enhancer-associated biochemical features** that differentiates experimentally
862 identified crisprQTLs from the 1,119 candidate enhancer library. The median AUPR from
863 fivefold cross-validation was 0.31.
864
865
866
867
868



870 **Fig. S5 | Replication of crisprQTL outside of multiplexed mapping experiment.** Two
 871 replicates of bulk RNA-seq were performed on RNA extracted from CRISPRi-positive K562
 872 separately transduced with sgOPTI-CROP-seq gRNAs solely targeting e-NMU (hg19
 873 chr4:56595855-56597095), e-PRKCB (chr16:23833191-23833694), e-PTGER3
 874 (chr1:71140581-71141997), e-GYPC (chr2:127402697-127403360), e-HMGA1
 875 (chr6:34191315-34192504), or the TSSs of their respective eCRISPR target genes. Error bars
 876 represent the quantification uncertainty for each sample as estimated through the bootstrapping
 877 procedure implemented in kallisto and sleuth. Percent repression is calculated from Monocle
 878 differential expression test (in the case of the crisprQTL mapping experiment) or from the ratio
 879 of: (mean transcript in targeted RNA) / (mean transcript in non-targeted RNA) (in the case of
 880 bulk RNA-seq, as called by sleuth and kallisto).

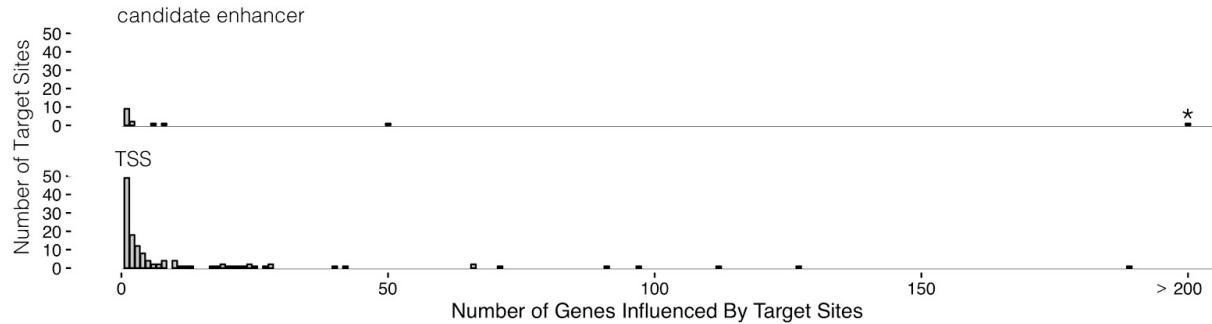
881



882
883
884
885
886
887
888
889
890

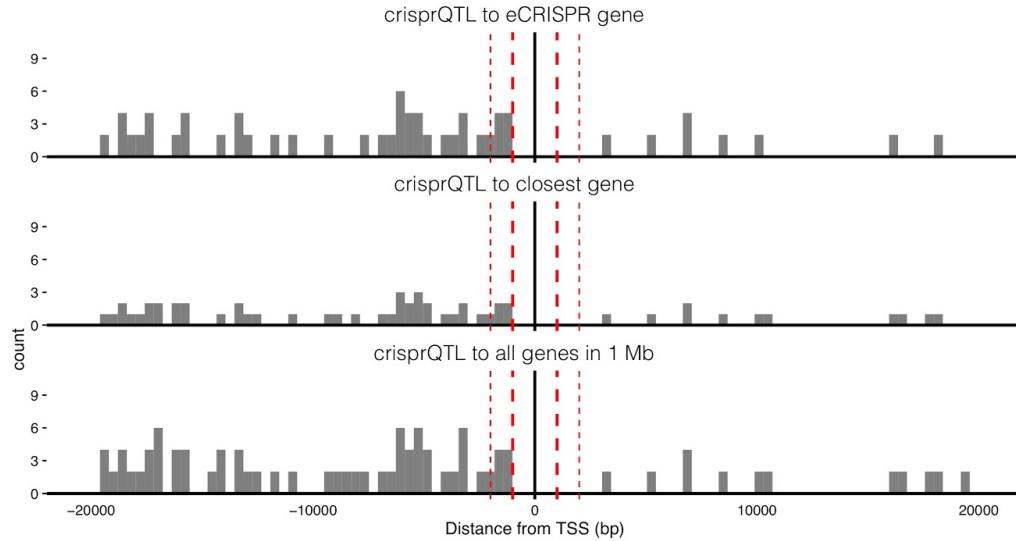
Fig. S6 | Extended data of the low MOI biological replicate. (A) gRNA/cell. The 1,119 K562-candidate enhancer library was transduced at an MOI of ~1. (B), Differential expression tests results for the low MOI experiment. Colored points correspond to perturbations called as ‘hits’ in this study, dashed red line shows the P -value threshold that corresponded to an empirical FDR of 10%. (C), Q-Q plot of the low (red) versus high (orange) MOI experiments shows an increase in power in the high MOI experiment, as seen by further inflated P -values. (D), P -values from the low versus high MOI studies. Only crisprQTLs found to be significant in the high MOI study are plotted (Pearson’s $r = 0.334$, purple = hit in both, orange = hit in just the high MOI study).

891



892 **Fig. S7 | Extended data for *trans* crisprQTLs.** Histogram of the number of eCRISPR-genes
893 significantly impacted by *trans* crisprQTLs. (* = one candidate enhancer located 75 Kb upstream
894 of *LMO2* was found to act as a *trans* crisprQTL of 1,177 eCRISPR genes).
895

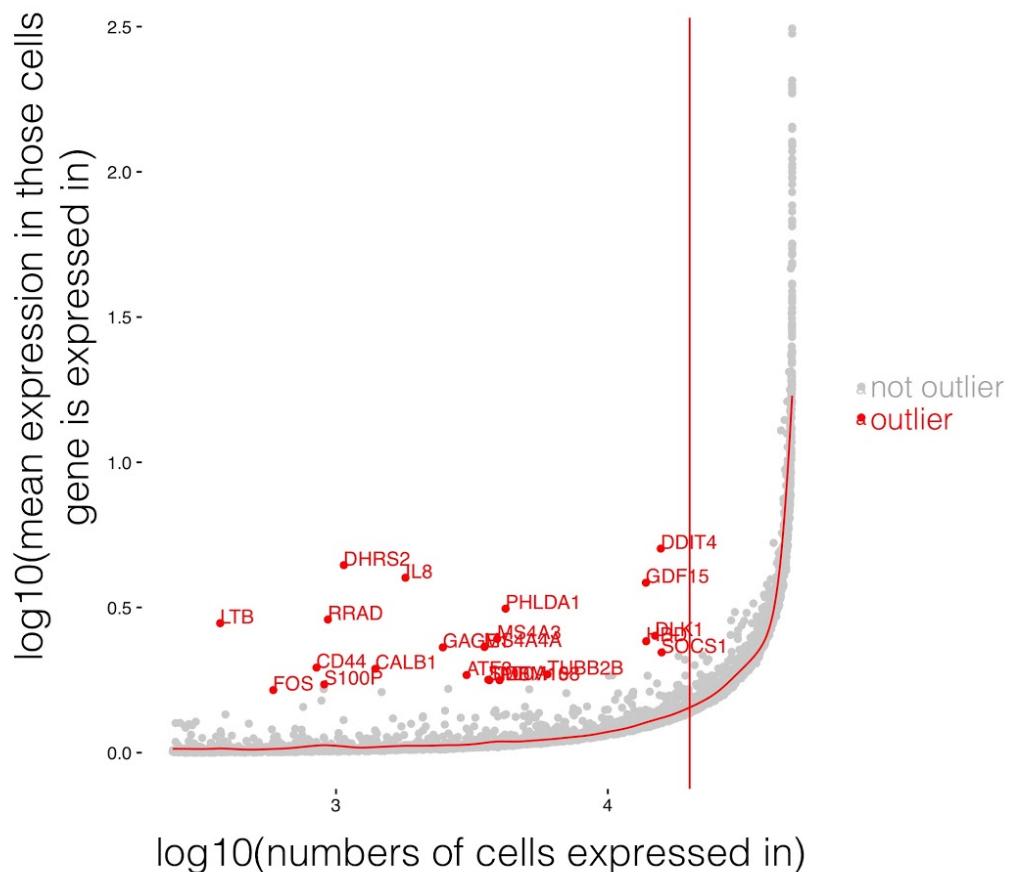
896



897 **Fig. S8 | +/- 20 Kb zoomed in view of distance distribution between target gene and**
898 **targeted site.** To note, gene bodies were excluded from targeting to avoid potential CRISPRi
899 interference with transcription. We also excluded any site within 1 Kb upstream of TSSs and 1
900 Kb downstream of TESs. (Thin dotted red line = 2 Kb from TSS, thick dotted red line = 1 Kb
901 from TSS, black line = TSS)

902

903



904 **Fig. S9 | Excluded 23 expression outliers.** We observed that differential expression test
905 P -values were inflated when a transcript was highly expressed but only in few cells. Thus, we
906 excluded 23 outlier genes (red) that were expressed in fewer than 20,000 cells in the high-MOI
907 47,650-cell dataset that also had >0.2 greater \log_{10} (mean expression in those cells gene is
908 expressed in) than predicted.
909

910

- 911 **Table S1 | 384 oligos used in pilot 100plex-library experiment.**
912 **Table S2 | 3,118 oligos used in scaled-up 1,119plex-library experiment.**
913 **Table S3 | Supplementary information of 145 cis crisprQTLs.**
914 **Table S4 | Ontological enrichment results of 105 eCRISPR genes.**
915 **Table S5 | Logistic regression P-values for independent and multivariate models.**
916 **Table S6 | Oligos used for replication of selected crisprQTLs outside of pool.**