Parallel Characterization of *cis*-Regulatory Elements for Multiple Genes Using CRISPRpath

Authors: Xingjie Ren[1], Mengchi Wang[2], Bingkun Li[1], Kirsty Jamieson[1], Lina Zheng[2], Ian R. Jones[1], Bin Li[3], Maya Asami Takagi[1], Jerry Lee[1], Lenka Maliskova[1], Tsz Wai Tam[1], Miao Yu[3], Rong Hu[3], Lindsay Lee[4], Armen Abnousi[4], Gang Li[5], Yun Li[6,7,8], Ming Hu[4], Bing Ren[3,9,10], Wei Wang[2,11], Yin Shen[1,12,#]

Affiliation:

[1]Institute for Human Genetics, University of California, San Francisco, San Francisco, CA, USA.

[2]Bioinformatics and Systems Biology Graduate Program, University of California, San Diego, La Jolla, CA, USA.

[3]Ludwig Institute for Cancer Research, La Jolla, California, USA.

[4]Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic Foundation, Cleveland, OH, USA.

[5]Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC, USA.

[6]Department of Biostatistics, University of North Carolina, Chapel Hill, NC, USA.

[7]Department of Genetics, University of North Carolina, Chapel Hill, NC, USA.

[8]Department of Computer Science, University of North Carolina, Chapel Hill, NC, USA.

[9]Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, CA, USA.

[10]Moores Cancer Center, University of California, San Diego, La Jolla, California, USA.

[11]Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA, USA.

[12]Department of Neurology, University of California, San Francisco, San Francisco, CA, USA.

[#]Corresponding Author: yin.shen@ucsf.edu

Abstract

Current pooled CRISPR screens for cis-regulatory elements (CREs), based on transcriptional output changes, are typically limited to characterizing CREs of only one gene. Here, we describe CRISPRpath, a scalable screening strategy for parallelly characterizing CREs of genes linked to the same biological pathway and converging phenotypes. We demonstrate the ability of CRISPRpath for simultaneously identifying functional enhancers of six genes in the 6-thioguanine-induced DNA mismatch repair pathway using both CRISPR interference (CRISPRi) and CRISPR nuclease (CRISPRn) approaches. 60% of the identified enhancers are known promoters with distinct epigenomic features compared to other active promoters, including increased chromatin accessibility and interactivity. Furthermore, by imposing different levels of selection pressure, CRISPRpath can distinguish enhancers exerting strong impact on gene expression from those exerting weak impact. Our results offer a nuanced view of *cis*-regulation and demonstrate that CRISPRpath can be leveraged for understanding the complex gene regulatory program beyond transcriptional output at scale.

Teaser

A new CRISPR screening strategy allows characterizing enhancers for multiple genes associated with converging phenotypes.

Introduction

Cis-regulatory elements (CREs) are key regulators for spatial-temporal control of gene expression. Mutations in CREs can contribute to complex diseases by modulating gene expression over long genomic distances (*1-3*). Thus, functionally characterizing CREs can provide important insight into gene regulation mechanisms and enable us to better interpret non-coding genetic variants associated with diseases. Despite the fact that tremendous numbers of candidate CREs have been mapped by biochemical signature (*4*), our knowledge of whether, how, and how much these putative CREs are functional on gene expression remain scarce in the human genome. Pooled CRISPR screens have been developed for testing CREs in their native chromatin context by monitoring the transcriptional levels for the gene of interest (*5-11*). Although results from these studies have made significant contributions to the annotation of functional DNA elements, challenges remain in pooled CRISPR screens of CREs. First, CRISPR

61 screens for enhancers based on gene expression levels largely depend on generating reporter
62 knock-in cell lines (*7*) or using FlowFISH signals (*8*). These procedures, involving generation of
63 reporter lines and selection of cells with positive hits by flow cytometry, are time-consuming and
64 difficult to scale up to multiple genes in the same experiment. Second, the approaches of using
65 gene expression as the screening phenotype (*9, 10*) fail to connect the functions of DNA
66 elements from transcriptional regulation at the molecular level to interpretable cellular and
67 physiological functions. Third, in cases of CRE screens using phenotypes such as cell
68 proliferation and survival (*11, 12*), they fail to quantify the effect sizes of enhancers on
69 transcriptional output.
70
71 To address these limitations, we developed CRISPRpath, a pooled CRISPR screening approach
72 to simultaneously characterizing CREs for multiple target genes involved in the same biological
73 pathway. CRISPRpath allows us to screen functional DNA elements based on phenotypes
74 associated with well-defined biological pathways. We demonstrate the capacity of CRISPRpath
75 by performing CRISPR interference (CRISPRi) and nuclease (CRISPRn) screens for six genes
76 in human induced pluripotent stem cells (iPSCs), and reveal different strengths of enhancer
77 functions by imposing varying levels of selection pressure on the cells.
78
79 **Results**
80 **Leveraging CRISPRpath for parallel characterization of CREs for multiple genes in iPSCs.**
81 To characterize candidate *cis*-regulatory elements (CREs) for multiple genes within the same
82 pooled CRISPR screening, we designed and applied CRISPRpath to six genomic loci containing
83 six genes (*HPRT1, MSH2, MSH6, MLH1, PMS2, PCNA*) involved in the 6-thioguanine (6TG)-
84 induced mismatch repair (MMR) (Fig. 1A). The MMR pathway is highly conserved and essential
85 for the maintenance of genome stability (*13*). The MMR pathway recognizes DNA mismatches
86 caused by 6TG treatment and induces cell apoptosis (*14, 15*). On the other hand, cells with a
87 malfunctioning MMR pathway, due to aberrant expression levels of 6TG metabolism genes or
88 MMR genes, may survive during 6TG treatment. Employing the properties of the MMR pathway,
89 we used cell survival for selecting cells with the reduced expression of MMR genes due to
90 defects in enhancer activities (Fig. 1B). To design the screening library, we first identified open
91 chromatin regions by performing Assay for Transposase Accessible Chromatin using

92 sequencing (ATAC-Seq) in WTC11 iPSCs. We included all open chromatin regions defined by
93 ATAC-seq peaks located 1Mb upstream and 1Mb downstream of each of the six genes
94 (spanning a total of 10.6 Mb genomic regions) as candidate CREs for functional characterization
95 (fig. S1, A and B; and table S1). We then designed a sgRNA library with 32,383 distal sgRNAs
96 targeting 294 distal ATAC-seq peaks, 2,755 proximal sgRNAs targeting 81 ATAC-seq peaks
97 overlapped with transcription start site (TSS) and coding regions of the six genes, and 625 non-
98 targeting sgRNAs with genomic sequences in the same genomic loci but are not followed by
99 PAM sequences (fig. S1C and table S2). In total, we included 35,763 sgRNAs in the library with
100 an average of 110 sgRNA per ATAC-seq peak (fig. S1D, and fig. S2A and S2B). We generated
101 a lentiviral library expressing these sgRNAs and transduced this library into two engineered
102 WTC11 iPSC lines, one expressing doxycycline-inducible dCas9-KRAB (CRISPRi) and the other
103 doxycycline-inducible Cas9 (CRISPRn) (*16*), both at a multiplicity of infection (MOI) of 0.5 (Fig.
104 1B).
105
106 To carry out the screening, we pre-determined the minimal lethal concentration of 6TG at 80
107 ng/mL for CRISPRi and CRISPRn iPSC lines (see Materials and Methods for details), and
108 applied three different 6TG concentrations (1X: 80 ng/mL, 2X: 160 ng/mL, 3X: 240 ng/mL) in
109 both CRISPRi and CRISPRn screens. We extracted and sequenced DNA samples from the
110 survival cells seven days after 6TG treatment to determine enriched sgRNAs by comparing the
111 results to that of the control cells taken after sgRNA library infection before the 6TG treatment
112 (Fig. 1B). To avoid confounding signals generated by off-target effects of low-quality sgRNAs
113 (*17*), we only used sgRNAs with high specificity (defined as specificity score > 0.2 (*18*), and
114 without any off-target sites with sequence similarity of ≤2 mismatches) for data analysis. This
115 led to the use of a total of 12,702 high-quality sgRNAs with an average of 38 sgRNAs per ATAC-
116 seq peak for analysis (fig. S1, E and F; and fig. S2, C and D). We performed each screen in two
117 biological replicates with each pair of replicates exhibiting high reproducibility (fig. S2E). We
118 compared the abundance of each sgRNA between the 6TG treated population and the control
119 population using a negative binomial model, and computed the fold change and *P* value to
120 quantify the effect size and the significance of enrichment of each sgRNA. We used the 5%
121 percentile of the *P* values from non-targeting control sgRNAs as the empirical significance
122 threshold to achieve a false discovery rate of 5%. sgRNAs with *P* values less than the empirical

123 significance threshold and with fold changes > 2 were defined as enriched (fig. S3). As expected,
124 sgRNA targeting TSS and coding region were identified as positive hits from both CRISPRi and
125 CRISPRn screens exhibiting greater fold changes in CRISPRn screens compared to the
126 CRISPRi screens (fig. S4A). We also observed enrichment of sgRNA bias towards coding
127 regions over TSS regions for CRISPRn screen (fig. S4B). These results are consistent with
128 CRISPRi functioning best near TSS by inhibiting transcription, and CRISPRn can disrupt gene
129 function by generating indels downstream of TSS (*19, 20*).
130
131 Further sgRNA fold-change ranking analysis revealed strong positive correlation between the
132 screens with 2X and 3X 6TG treatment for both CRISPRi and CRISPRn screen (Spearman
133 correlation, CRISPRi = 0.97, CRISPRn = 0.84) (Fig. 1C) with the correlations for proximal
134 sgRNAs being higher than for distal sgRNAs (fig. S4C). On the contrary, results from the 1X
135 screen correlated poorly with either 2X or 3X screens (Fig. 1C), suggesting more substantial
136 selection pressure (2X and 3X) can reduce background noise in CRISPRpath screens. Thus,
137 we used sgRNAs enriched from 2X and 3X screens data for identifying active enhancers in the
138 following section (Fig. 1, D and E).
139
140 **CRISPRi is more efficient than CRISPRn in pooled CRISPR screens of CREs.**
141 Performing CRISPRpath with CRISPRi and CRISPRn in the same genetic background with an
142 identical sgRNA library offers a unique opportunity for comparing the efficacies of CRISPRi and
143 CRISPRn in pooled CRISPR screens of CREs. We noticed that CRISPRn screens recovered
144 fewer enriched distal sgRNAs than CRISPRi screens (Fig. 1F). This is possibly due to the fact
145 that CRISPRi-mediated heterochromatin formation can more effectively perturb CREs compared
146 to CRISPRn-mediated genetics perturbations. We then called an candidate element as an
147 enhancer if there are at least 3 enriched sgRNAs in that CRE. Based on this criterion, we
148 identified 62 and 33 enhancers from the 2X and 3X CRISPRi screen, respectively, and 19
149 enhancers from the 2X CRIPSRn screen (Fig. 1G and table S3). However, no enhancer was
150 identified from the 3X CRISPRn screen, indicating either the CRISPRn induced mutations did
151 not lead to any strong effect on gene expression to make the cells survive the 3X 6TG treatment
152 or there are insufficent numbers of sgRNAs exhibiting deleterious effects on the tested DNA
153 elements to satisfy our criterion of calling functional enhancers. In total, 66 unique enhancers

154 were identified for the six target genes with CRISPRpath under different 6TG treatments (Fig.
155 1G). Together, we demonstrate CRISPRpath can simultaneously identify enhancers for multiple
156 target genes with CRISPRi outperforming CRISPRn. For the following analysis, we focused on
157 the 63 enhancers identified from the 2X and 3X CRISPRi screens (Fig. 1G).
158
159 **Genomic feature of CRISPRpath identified enhancers.**
160 To determine the genomic feature of the enhancers, we plotted all the tested elements by their
161 genomic locations and enrichment scores (average of $\log_2$(fold change) of enriched sgRNAs of
162 each element) (Fig. 2A). Not surprisingly, our data suggest that each gene can be regulated by
163 multiple enhancers with the identified functional enhancers having no position bias relative to
164 the TSS. The average distance between an enhancer and its paired TSS is about 530 Kb (Fig.
165 2B) with an average of 10 interval genes between an identified enhancer and its target gene
166 pairs (Fig. 2C). Interestingly, we observed a weak negative correlation between the enhancer
167 enrichment score and the distance between an enhancer and its paired TSS (Fig. 2D, Pearson
168 correlation, ρ = -0.36, P = 0.01), suggesting enhancers near to TSS tend to have higher
169 regulatory activity compared to enhancers further away from their target genes. It is worth noting,
170 the relative positions for the enriched sgRNAs exhibited no preference relative to ATAC-seq
171 peaks (Fig. 2E and fig. S5A) and no preference for the strand on which the sgRNAs were
172 designed (fig. S5B), consistent with our knowledge that CRISPRi mediated heterochromatin
173 spreads over hundreds of base pairs in distance (*21*).
174
175 Previous studies have revealed that promoters can function as enhancers (*7, 22*). Indeed, 60%
176 (38 out of 63) of the functional enhancers identified in CRISPRi screens overlapped with
177 annotated promoters, providing an excellent opportunity to further explore the genomic features
178 of these enhancer-like promoters. To validate whether these promoters function as bona fide
179 enhancers, we targeted three enhancer-like promoters with CRISPRi. We confirmed significant
180 downregulation of their target genes including *MSH6*, *MSH2*, and *PCNA* (Fig. 3, A to C; and fig.
181 S6). In contrast, shRNAs against the transcripts from these promoters (*SOCS5*, *FOXN2*, and
182 *TMEM230*) only led to a significant downregulation of its own transcripts and did not affect their
183 target gene expression (Fig. 3, A to C). These results confirm that these promoter sequences
184 identified by CRISPRpath can function as enhancers. Although it has been shown that enhancer-

like protomors are enriched with active chromatin marks and physically close to target genes (*7*), it is not clear whether enhancer-like promoters have unqiue genome features that can differentiate them from other regular active promoters. To this end, we compared chromatin accessibility, occupancy of histone 3 lysine 4 trimethylation (H3K4me3), histone 3 lysine 27 acetylation (H3K27ac) and CTCF, transcription, and chromatin interactivity levels between enhancer-like promoters and all other active promoters that did not show enhancer activity in our CRISPRi screens. We show that enhancer-like promoters exhibit higher chromatin accessibility, higher level of transcription, stronger H3K4me3 and H3K27ac signals than those at other active promoters (Fig. 3D). On the other hand, we did not observe a significant difference for CTCF binding signals between enhancer-like promoters and control promoters (Fig. 3D). Furthermore, by evaluating chromatin interaction data using H3K4me3 Proximity Ligation-Assisted ChIP-seq (PLAC-seq), we show enhancer-like promoters have significantly more and stronger interactions compared to control promoters (Fig. 3E).

CRISPRpath is capable of distinguishing enhancers with distinct effect sizes.

Gene expression often is a result of combinatorial regulatory effects from multiple *cis*-regulatory elements (*11, 23*). Understanding how individual enhancers contribute to gene expression in a quantitative manner is an important first step in dissecting how enhancers orchestrate precise transcriptional control. We seek a new strategy to differentiate enhancers based on their effect sizes on gene expression using CRISPRpath. We hypothesized that cells with drastic down-regulation of MMR genes have a fitness advantage under higher 6TG concentration than cells with modest down-regulation of MMR genes. Consistent with this hypothesis, proximal sgRNAs exhibit larger fold changes than distal sgRNAs (Fig. 1E) because perturbing proximal regions has more profound effects on gene down-regulation than perturbing distal regulatory regions. Based on these observations, we hypothesize that enhancers identified under different selection pressure represent distinct regulatory strengths on transcriptional activation. We noticed that the enriched sgRNAs for enhancers identified under strong selection pressure (3X), have bigger fold changes compared to those for enhancers uniquely identified under weak selection pressure (2X) (fig. S7A). Similarly, enrichment scores for these two groups of enhancers are significantly different, with the TSS regions manifesting the highest enrichment scores (Fig. 4A). sgRNA fold changes (fig. S7B) and element enrichment scores (fig. S7C) both have no significant

differences between distal enhancers and enhancer-like promoters (fig. S7, B and C). Therefore, enhancers identified in the 3X screen are strong enhancers (n = 33), while enhancers uniquely identified in the 2X screen are weak enhancers (n = 30) (Fig. 1G).

To confirm the quantitative effect of enhancers on target gene expression, we tested 11 strong and 10 weak enhancers using CRISPRi followed by RT-qPCR measurement of the corresponding target gene expression (Fig. 4B and fig. S8A). We show perturbations of strong enhancers led to significantly more down-regulation of target gene expression (mean down-regulation of target gene by 21%) than perturbations of weak enhancers (mean down-regulation of target gene by 6%), with the perturbations of TSS regions achieving the strongest down-regulation of target genes, by an average of 68% reduction in gene expression (Fig. 4B, and fig. S8, A and B). These quantitative effects on target gene expression are consistent with the enrichment scores from our CRISPRpath screens (fig. S8C) and demonstrate the capacity of distinguishing enhancers with different effect sizes by imposing different levels of selection pressures. These CRISPRi validated enhancers also demonstrated enhancer activities in luciferase reporter assays with only one exception (fig. S9A). Luciferase activitives did not show the same level of strong anti-correlation with the levels of CRISPRi-mediated transcriptional repression compared to element enrichment scores (fig. S8, C and fig. S9, B and C), and can't distinguish weak and strong enhancers (fig. S9D), possibly because it tests enhancers out of their native genomic context with a heterogeneous promoter.

We further explored chromatin features of strong and weak enhancers by analyzing chromatin accessibility, H3K4me3, H3K27ac, and CTCF binding signals in these regions. At individual chromatin mark level, while CRISPRpath-identified enhancers were more accessible, and enriched with active chromatin marks, such as H3K4me3 and H3K27ac, and CTCF binding compared to negative elements or random elements (Fig. 4C), we did not observe significant differences between strong and weak enhancers in the chromatin features we individually examined. However, strong enhancers tend to have more active chromatin signatures than weak enhancers (Fig. 4D), suggesting combined signatures of active chromatin can be a better indicator of enhancer strength. Strong enhancers tend to have higher distance normalized PLAC-seq contact frequencies with their target promoters than weak enhancers, though not

statistically significant, possibly due to the small sample size in this study (Fig. 4E). We obtained similar results by expanding this analysis for characterized enhancers in K562 cells and mouse embryonic stem cells (*6, 8*) (fig. S10), which reinforces the idea that enhancers with larger effects on gene expression tend to have higher chromatin interactions with their cognate promoters. To explore the possible mechanisms that drive enhancer activities in a quantitative manner, we evaluated potential transcription factors (TFs) binding motifs in strong and weak enhancer sequences. Both strong and weak enhancers are enriched with CTCF binding motif (Fig. 4F). Indeed, most of strong and weak enhancers are bound by CTCF (Fig. 4D), consistent with the notion that CTCF-mediated chromatin loops are essential for gene activation (*24*). Furthermore, strong enhancers and weak enhancers have differential enrichment with TFs binding motifs. For example, the binding motifs for SP/KLF family (*25*) and E2F family (*26, 27*) appear more frequently in strong enhancers compared to weak enhancers, suggesting these strong enhancers could be major docking sites for master regulators in iPSCs (Fig. 4F).

Discussion

CRISPR-mediated high throughput screening using bulk cells allows for the functional characterization of regulatory elements in their native genomic context. However, current approaches are limited to validating a small number of regulatory elements for a single gene (*5, 7, 9, 12, 28, 29*). To overcome this bottleneck, we developed CRISPRpath, a strategy for functional characterization of enhancers for multiple genes simultaneously by leveraging the genes involved in the same biological pathway so that the effects can be measured via a define phenotype. For example, alpha-toxin resistance phenotype can be used to identify CREs for 17 genes in glycosylphosphatidylinositol (GPI)-anchor synthesis pathway (*30*). CRISPRpath can also be leveraged to identify CREs for protein folding regulators that contribute to the endoplasmic reticulum stress-response pathway (*31*) using UPRE reporter in mammalian cells. Since CRISPR screen technology is widely used, the CRISPRpath strategy is readily applicable to simultaneously identifying enhancers for genes converging in the defined biological processes and pathways across different cell types. Compared to the existing pooled CRISPR screens of CREs (*5, 7, 8, 10-12, 28, 29, 32*), CRISPRpath is scalable with additional benefits of connecting DNA elements to cellular function, beyond the most standard molecular phenotype of gene expression.

Multiple factors can contribute to the observed low effects on gene expression upon enhancer perturbation in this study. First, enhancers tend to have big impacts on genes with cell-type-specific expression patterns. For example, perturbing *HBG1*/*2* enhancers at the locus control region of the human β-globin locus leads to the drastic reduction of *HBG1*/*2* expression in K562 cells (*33*). On the other hand, enhancer disruptions for broadly expressed genes such as *GATA1*, *HDAC6*, and *MYC* genes only resulted in 9-62% expression reduction in K562 cells (*11*), comparable to our results on perturbing MMR gene enhancers. Second, on the technical level, it has been demonstrated that CRISPRi can not completely silence enhancers compared to the CRISPR deletion approach (*6*).

Promoters can function as enhancers more widespread than expected, with more than half of the enhancers identified for MMR genes in our study being previously annotated promoters. This is consistent with previous reports that enhancer-like promoters are more prevalent for ubiquitously expressed genes (*34*). Enhancer-like promoters are more accessible compared to other promoters, possibly because these regions are required to be more open to accommodate additional transcriptional machinery such as TF for activating target gene expression besides their own transcription (*35*). Enhancer-like promoters also exhibit significantly higher levels of chromatin interactions with distal regions compared to other active promoters. This observation can be explained by the fact that enhancer-like promoters will not only form chromatin loops with their distal target genes, but also with CREs for controlling the expression of their own genes.

Genomic studies of chromatin marks have revealed hundreds of thousands candidate CREs in the human genome but with very little quantitative information regarding how CREs contribute to gene regulation (*36, 37*). Using CRISPRpath, we can systematically classify enhancers based on their effect sizes on transcription. Identifying and charaterizing the effect size for each individual enhancer is the critical first step to future studies of their combinatory effects on target gene expression. Interestingly, strong and weak enhancers can not be distinguished by individual epigenetic marks we examined. One possible explanation for this observation is that chromatin features only mark enhancer's identity but do not quantify enhancer activity. On the other hand, the strong and weak enhancers we identified may regulate other genes differently

from regulating the MMR gene. Interestingly, strong enhancers tend to harbor more than one active chromatin signature, which indicates that enhancer activities are regulated by multiple epigenetic factors, for example, TF mediated transcriptional regulation. Differential TFs binding motifs observed within strong and weak enhancers suggest that enhancer strength is modulated by TF binding. Future studies that further integrating TF binding datasets with functional data of enhancers will shed light on the molecular mechanisms that drives enhancers' effect sizes on gene regulation.

Acknowledgements

This work was supported by the National Institutes of Health grants UM1HG009402 (to Y.S., B.R., and W. W) and U54DK107977 (to B. R. and M. H.).

Data and materials availability

The CRISPRpath screen datasets used in this study are available at the ENCODE portal (www.encodeproject.org, accession number: ENCSR617AZY (sgRNA plasmid library), ENCSR427OPP (CRISPRi control), ENCSR900AXT, ENCSR254SJU (CRISPRi 2X), ENCSR793DSE (CRISPRi 3X), ENCSR250ZWC (CRISPRn control), ENCSR117YGQ (CRISPRn 1X), ENCSR071ZGB (CRISPRn 2X), ENCSR482PHH (CRISPRn 3X)). WTC11 iPSCs H3K4me3 PLAC-seq datasets are available at the 4DN data portal (data.4dnucleome.org, accession number: 4DNESIZ5TTHO and 4DNESDRL4ZKM). ATAC-seq, ChIP-seq, CUT&Tag, and RNA-seq datasets in WTC11 iPSCs are available at the Gene Expression Omnibus under the accession number GSE166839. Data can be visualized on the WashU Epigenome Browser using the session: https://epigenomegateway.wustl.edu/browser/?genome=hg38&sessionFile=https://shen-xren.s3-us-west-1.amazonaws.com/CRISPRpath/eg-session-QRXJ0218-4d710b60-6ea7-11eb-8d8d-03c7189570c0.json. Tracks include ATAC-seq, H3K27ac, H3K4me3, and CTCF signals, and the identified enhancers from CRISPRi 2X and 3X screens. RefGene 38 genes are also displayed. The plasmids generated in this study are available from Addgene (#167186, #167188, # 167189). The computer code used for analyzing CRISPRpath datasets is available at https://github.com/MichaelMW/crispy. A copy of the custom code used for data analysis and

figure generation in this study is available upon request. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Competing interests

B.R. is co-founder and shareholder of Arima Genomics and Epigenome Technologies. The other authors declare that they have no competing interests.

Author contributions

X.R. and Y.S. designed the study. X.R. and B.L. designed the sgRNA library under the supervision of Y.S. and B.R. X.R., K.J., I.R.J., M.A.T., J.L., L.M., and T.W.T. performed the experiments. M.W. designed the CRISPY under the supervision of W.W. X.R., B.L., L.Z., G.L., and Y.L. performed data analysis. M.Y. and R.H. constructed the H3K4me3 PLAC-seq datasets. L.L., A.A. and M.H. analyzed PLAC-seq and HiChIP data. X.R. and Y.S. prepared the manuscript with input from all other authors.

353 **Fig. 1. CRISPRpath for identifying enhancers of multiple genes.**
354 (**A**) Six genes (*HPRT1*, *MSH2*, *MSH6*, *MLH1*, *PMS2* and *PCNA*) in the 6TG-induced mismatch
355 repair process were used for CRISPRpath screen in this study. (**B**) Schematic of the
356 CRISPRpath screening strategy with 6TG treatment in iPSCs. Cell survival was used as readout
357 for the screen. (**C**) Spearman correlation analysis of sgRNA ranking based on fold changes for
358 CRISPRpath screens with different 6TG concentrations (1X, 2X and 3X). (**D**) Venn diagram
359 shows the overlapping enriched sgRNAs identified from the screens with 2X and 3X 6TG
360 treatments. (**E**) Box plots show the fold changes of the enriched distal and proximal sgRNAs
361 from 2X, 3X CRISPRi and CRISPRn screens. Star indicates no enriched distal sgRNA was
362 identified from 3X CRISPRn screen. Boxplots indicate the median, interquartile range (IQR), Q1
363 − 1.5 × IQR and Q3 + 1.5 × IQR. (**F**) Bar plot shows the number of enriched distal and proximal
364 sgRNAs from 2X, 3X CRISPRi and CRISPRn screens. Star indicates no enriched distal sgRNA
365 was identified from 3X CRISPRn screen. (**G**) Venn diagram shows the identified enhancers from
366 each CRISPRpath screen.
367
368 **Fig. 2. Genomic features of identified enhancers from CRISPRpath using CRISPRi.**
369 (**A**) Genomic locations of identified enhancers relative to TSS. Circles indicate enhancers
370 identified from the CRISPRi 3X screen (red), enhancers uniquely identified from the CRISPRi
371 2X screen (blue), tested CREs that are not identified as enhancers (grey). Purple lines label the
372 location of each target gene. (**B**) Histogram shows the distance distribution between identified
373 enhancers and their paired TSS. (**C**) Histogram shows the number of interval genes between
374 enhancers and their target gene TSS. Mean is indicated with an orange dashed line in B and C.
375 (**D**) A weak negative correlation is observed between enrichment score and genomic distance
376 between enhancers and their target genes (Pearson correlation, r = -0.36, P = 0.01). Black
377 circles indicate TSS regions. The red and blue circles are enhancers showing in A. Only the
378 enhancers for *HPRT1*, *MLH1*, *PMS2*, *PCNA* are included in B to D. (**E**) Density plot shows no
379 significant difference (two-tailed two-sample Kolmogorov–Smirnov test) for the distribution of all
380 distal sgRNAs (gray), enriched distal sgRNAs from 2X (blue) and 3X screen (red) CRISPRi
381 screens.
382
383 **Fig. 3. Enhancer-like promoters act as functional enhancers.**

384 (**A**, **B**, **C**) Three examples of promoters function as enhancers. CRISPRi silencing of the
385 promoter region of *SOCS5*, *FOXN2* and *TMEM230* results in significant downregulation of
386 *MSH6*, *MSH2* and *PCNA*, respectively. shRNA knockdown of *SOCS5*, *FOXN2* and *TMEM230*
387 can only downregulate *SOCS5*, *FOXN2* and *TMEM230* expression. Three independent
388 replicates per condition and two independent sgRNAs or shRNAs per replicate were used for
389 each experiment. *P* values are from two-tailed two-sample *t*-test. (**D**) Average signal enrichment
390 of ATAC-seq, gene transcription, H3K4me3, H3K27ac and CTCF binding for enhancer-like
391 promoters (n = 38) and control promoters (n = 47). *P* values are from Wilcoxon test. Boxplots
392 indicate the median, IQR, Q1 − 1.5 × IQR and Q3 + 1.5 × IQR. (**E**) Number of H3K4me3 mediated
393 chromatin interactions and cumulative interaction score for enhancer-like promoters (n = 31) and
394 control promoters (n = 43). Boxplots indicate median, IQR, Q1 − 1.5 × IQR and Q3 + 1.5 × IQR.
395 *P* values are calculated from Wilcoxon test.
396
397 **Fig. 4. CRISPRpath can distinguish weak and strong enhancers by imposing different**
398 **selection pressures.**
399 (**A**) Box plots show the enrichment score of the tested elements. TSS regions (black circles)
400 show highest enrichment scores. Enhancers uniquely identified from the lower selection
401 pressure (CRISPRi 2X, blue circles) exhibit lower enrichment scores compared to the enhancers
402 identified from the higher selection pressure (CRISPRi 3X, red circles). Each circle represents
403 an individual element. *P* values are from Wilcoxon test. (**B**) Box plots show the CRISPRi
404 perturbation at enhancers induced various degrees of transcriptional repression of target genes
405 measured with RT-qPCR. Each dot represents the average value from three biological
406 replicates. CRISPRi targeting TSS regions (dark gray) achieved the highest transcriptional
407 repression. CRISPRi targeting strong enhancers (pink) leads a more substantial transcription
408 silencing on target gene compared to CRISPRi targeting weak enhancers (cyan). *P* values are
409 from Wilcoxon test. (**C**) Enrichment analysis of ATAC-seq, H3K27ac, H3K4me3, CTCF binding
410 signals for strong (n = 33) and weak (n = 30) enhancers. *P* values for the difference between
411 strong and weak enhancers are from Wilcoxon test; see table S7 for *P* values of all pairwise
412 comparisons. Boxplots (A to C) indicate the median, IQR, Q1 − 1.5 × IQR and Q3 + 1.5 × IQR.
413 (**D**) Intersection of genomic features for weak enhancers (blue bar) and strong enhancers (red
414 bar). (**E**) Distance normalized H3K4me3 PLAC-seq contact frequency for strong (n = 23) and

415 weak (n = 21) enhancers. Only the enhancers for *HPRT1*, *MLH1*, *PMS2*, *PCNA* are included
416 (see Materials and Methods for details). Boxplots indicate the median, IQR, Q1 − 1.5 × IQR and
417 Q3 + 1.5 × IQR. *P* value is from Wilcoxon test. (**F**) Heatmap shows the normalized frequency of
418 transcription factor motifs found in strong and weak enhancers.
419
420 **Fig. S1. Features of sgRNA library for CRISPRpath screen.**
421 (**A**) Bar graph shows the number of distal ATAC-seq peaks used as candidate CREs for six
422 MMR genes. (**B**) Histogram shows size distribution of distal ATAC-seq peaks. The average size
423 is 488 bp (blue dash line). (**C**, **E**) The composition of the sgRNA library. In total, 35,763 sgRNAs
424 were included in the library (C), and 12,702 sgRNAs are high quality sgRNAs (E). (**D**, **F**)
425 Distribution of the number of total sgRNAs (D) or high quality sgRNAs (F) per distal ATAC-seq
426 peak. Average numbers of sgRNA per ATAC-seq peak are indicated with blue dash lines.
427
428 **Fig. S2. Quality of the sgRNA library and CRISPRpath screen libraries.**
429 (**A, C**) Distribution of sgRNA read counts for all the sgRNAs (A) and high quality sgRNAs (C) in
430 the constructed sgRNA plasmid library. (**B, D**) Cumulative frequency of all the sgRNAs (B) and
431 high quality sgRNAs (D) in the constructed sgRNA plasmid library. The constructed sgRNA
432 plasmid library recorvered all the designed sgRNAs (A, C) with the copy number difference less
433 than five fold for at least 97% designed sgRNAs (B, D). (**E**) PCA analysis shows the high
434 reproducibility of the CRISPRpath screen libraries between biological replicates.
435
436 **Fig. S3. *P* value cutoff used for identifying enriched sgRNAs from each screens.**
437 (**A-F**) Distribution of *P* value for tested distal, proximal and non-targeting control sgRNA groups.
438 Orange dash lines indicate 5% percentile of the *P* values from non-targeting control sgRNAs to
439 achieve a false discovery rate of 5%.
440
441 **Fig. S4. Enriched proximal sgRNAs and sgRNA ranking analysis.**
442 (**A**) Number and fold changes of the enriched proximal sgRNAs for the six target genes from
443 CRISPRi and CRISPRn screens. The color indicates fold changes, and the size of circle
444 indicates the number of enriched sgRNAs. (**B**) Enrichment analysis shows the enriched proximal
445 sgRNAs bias towards to the TSS region for CRISPRi screens and the protein coding region

446 (CDS) for CRISPRn screens. Color represents the number of enriched sgRNAs. (**C**) Spearman
447 correlation analysis of the distal and proximal sgRNAs ranking shows proximal sgRNAs
448 exhibiting higher correlation between each screen compared to distal sgRNAs.
449
450 **Fig. S5. Enriched sgRNAs identified from CRISPRi screens exhibit no position and strand**
451 **preference.**
452 (**A**) Enriched sgRNAs from CRISPRi 2X (red dots, n = 448) and CRISPRi 3X (red dots, n = 260)
453 screens showed similar distributions across candidate CREs. (**B**) Odds ratio analysis of the fold
454 changes of enriched sgRNAs shows enriched sgRNAs have no strand preference. Enhancers
455 with enriched sgRNAs only targeting one strand were exclude for the analysis. Odds ratio was
456 calculated for each element with the equation of ave($\log_2$(fold change of sgRNA targeting plus
457 strand)) / ave($\log_2$(fold change of sgRNA targeting minus strand)). Violin plots show the
458 distributions of odds ratio values within each screen, and boxplots indicate the median, IQR,
459 Q1 − 1.5 × IQR and Q3 + 1.5 × IQR.
460
461 **Fig. S6. *SOCS5* promoter and *FOXN2* promoter do not regulate *MSH2* and *MSH6***
462 **expression.**
463 (**A**) CRISPRi silencing of *SOCS5* and *FOXN2* promoters do not change *MSH2* and *MSH6*
464 expression, respectively. Three independent replicates per condition and two independent
465 sgRNAs per replicate were used for each experiment. *P* values are from two-tailed two-sample
466 *t*-test. Error bars represent the s.d.
467
468 **Fig. S7. sgRNA fold changes and element enrichment scores comparison for the**
469 **identified enhancers.**
470 (**A**) Box plots show the sgRNA fold changes of the enriched sgRNAs for the tested elements.
471 The enriched sgRNAs in TSS regions of MMR genes show highest fold changes. Enriched
472 sgRNAs of the enhancers uniquely identified from the lower selection pressure (CRISPRi 2X)
473 exhibit lower fold changes compared to enriched sgRNAs of the enhancers identified from the
474 higher selection pressure (CRISPRi 3X). Boxplots indicate the median, IQR, Q1 − 1.5 × IQR and
475 Q3 + 1.5 × IQR. The dots represent individual sgRNA. *P* values are from Wilcoxon test. (**B**) Box
476 plots show the sgRNA fold changes of the enriched sgRNAs for different type enhancers. For

477 the enhancers identified from CRISPRi 3X screen and the enhancers uniquely identified from
478 CRISPRi 2X screen, sgRNA fold changes of the enriched sgRNAs did not show significant
479 difference between distal enhancers and enhancer-like promoters. Boxplots indicate the median,
480 IQR, Q1 − 1.5 × IQR and Q3 + 1.5 × IQR. The dots represent individual sgRNA. *P* values are
481 from Wilcoxon test. (**C**) Box plots show the element enrichment scores for enhancers. For the
482 enhancers identified from CRISPRi 3X screen and the enhancers uniquely identified from
483 CRISPRi 2X screen, element enrichment score did not show significant difference between distal
484 enhancers and enhancer-like promoters. Boxplots indicate the median, IQR, Q1 − 1.5 × IQR and
485 Q3 + 1.5 × IQR. The dots represent individual sgRNA. *P* values are from Wilcoxon test.
486
487 **Fig. S8. Validation of CIRSPRpath identified enhancers.**
488 (**A**) Validation of the strong (black) and weak (grey) enhancers with CRISPRi followed by RT-
489 qPCR. Three independent replicates per condition. The significance was calculated with two-
490 tailed two-sample *t*-test. Data are mean and s.d. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. (**B**)
491 CRISPRi-mediated transcriptional repression of six target genes by targeting TSS of each gene.
492 Three independent replicates per condition. The significance was calculated with two-tailed two-
493 sample *t*-test. Data are mean and s.d. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. (**C**) Pearson
494 correlation analysis reveals element enrichment scores from CRISPRpath screens correlate with
495 element effect sizes on transcription from CRISPRi (Pearson correlation, PCC = -0.68, $P = 3.2$
496 $\times 10^{-5}$).
497
498 **Fig. S9. Characterization of enhancers with dual-luciferase reporter assay.**
499 (**A**) Relative luciferase reporter activities for the weak and strong enhancers. The relative
500 luciferase activities were normalized to the average of the luciferase activity of minimal promoter.
501 The significance was compared to minimal promoter and two negative control elements, and
502 calculated with two-tailed two sample *t*-test. Data are mean and s.d. (**B**) Pearson correlation
503 analysis reveals weak correlation between relative luciferase activities and element effect sizes
504 on transcription from CRISPRi (Pearson correlation, PCC = -0.45, $P = 0.026$). (**C**) Pearson
505 correlation analysis reveals weak correlation between relative luciferase activities and element
506 enrichment scores from CRISPRpath screens (Pearson correlation, PCC = 0.26, $P = 0.21$). (**D**)
507 Dotplot shows the distribution of relative luciferase activities of weak and strong enhancers. Each

508 dot represents one tested enhancer. The black line labels the average of relative luciferase
509 activity. The significance was calculated with two-tailed two sample *t*-test.
510
511 **Fig. S10. Chromatin contract frequency analysis for the enhancers in K562 cells and**
512 **mESCs.**
513 (**A**) Distance normalized H3K27ac HiChIP contact frequency for strong (n = 34) and weak (n =
514 82) enhancers identified with crisprQTL mapping in K562 cells. (**B**) Distance normalized
515 H3K27ac HiChIP contact frequency for strong (n = 2) and weak (n = 20) enhancers identified
516 with CRISPRi-FlowFISH screen in K562 cells. (**C**) Distance normalized H3K4me3 PLAC-seq
517 contact frequency for strong (n = 10) and weak (n = 3) enhancers identified in mouse embryonic
518 stem cells. Boxplots indicate the median, IQR, Q1 − 1.5 × IQR and Q3 + 1.5 × IQR. *P* values are
519 from Wilcoxon test.
520
521 **Table S1. ATAC-seq peak regions used to design the sgRNA library.**
522
523 **Table S2. List of sgRNA sequences used for CRISPRpath screen.**
524
525 **Table S3. Enhancers identified from CRISPRn 2X, CRISPRi 2X and CRISPRi 3X screens.**
526
527 **Table S4. List of primers used for RT-qPCR.**
528
529 **Table S5. List of shRNA sequences used for RNA interference experiments.**
530
531 **Table S6. List of sgRNA sequences used for CRISPRi-mediated enhancer validation**
532 **experiments.**
533
534 **Table S7. Pairwise comparisons of data in Figure 4c.**
535
536 **Table S8. List of primers used for dual-luciferase reporter assay.**
537

538 **Materials and Methods**

539

540 **Cell culture**

541 Doxycycline inducible CRISPRi and CRISPRn WTC11 iPSC lines were purchased from
542 Gladstone Stem Cell Core. Both CRISPRi and CRISPRn WTC11 iPS cells were cultured on
543 Matrigel-coated (Corning, 354277) plates with Essential 8™ Medium (Life Technologies,
544 A1517001). iPSCs were passaged using Accutase (STEMCELL Technologies, 07922) and 10
545 μM ROCK inhibitor Y-27632 (STEMCELL Technologies, 72302). HEK293T cells were cultured
546 in Dulbecco's modified Eagle's medium (Gibco, 11995065) with 10% fetal bovine serum
547 (CPSSerum, FBS-500). HEK293T cells were passaged with Trypsin-EDTA (Gibco, 25200072).
548 All the cells were grown with 5% CO_2 at 37°C and verified mycoplasma free using the MycoAlert
549 Mycoplasma Detection Kit (Lonza, LT07-218).

550

551 **sgRNA library design**

552 CRISPRpath sgRNA library was designed to screen cis-regulatory elements for *HPRT1*, *MSH2*,
553 *MSH6*, *MLH1*, *PMS2* and *PCNA*. ATAC-seq peaks within the region of 1 Mb upstream and 1 Mb
554 downstream of each target gene including TSS and coding regions were selected as targeting
555 regions for the sgRNA library design (table S1). We generated a genome-wide sgRNA database
556 containing all the available unique sgRNAs, each followed by a 'NGG' PAM sequence. All the
557 designed unique sgRNAs in the target regions were added in the sgRNA library, excluding
558 sgRNAs containing AATAAA, AAAAA, TTTTT or TTTTTT sequences. Unique 20-bp sequences
559 in the target regions that were not followed by the 'NGG' or 'NAG' PAM sequences were taken
560 as non-targeting control sgRNAs, excluding non-targeting sgRNAs containing TTT, TTNTT,
561 AATAAA, AAAAA, TTTTT or TTTTTT sequences. Then, a guanine nucleotide was added to all
562 the sgRNAs if the sequence did not start with G to increase efficiency of transcription from U6
563 promoter. Final sgRNA oligos adhered to the following template: 5'-
564 ATATCTTGTGGAAAGGACGAAACACC-[20- or 21-bp sgRNA sequence]-
565 GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGC-3'. In total, 35,763 sgRNAs were
566 included in the library (fig. S1 and table S2). We retrieved specificity score and off-target site for
567 each sgRNA from GuideScan (*18*) (www.guidescan.com) and assigned the specificity score of

568 sgRNAs not existed in the GuideScan database to 0. The high-quality sgRNAs were filtered with
569 specificity score >0.2 and without perfectly matched or 1-2 mismatches off-target sites.
570
571 **Oligo synthesis and library cloning**
572 sgRNA library oligos were synthesized by TWIST BIOSCIENCE and amplified with the forward
573 primer 5'- TCGATTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACAC-3' and the reverse
574 primer
575 5'-AACGGACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAAC-3'. We replaced the Cas9
576 sequence in lentiCRISPR v2 plasmid (Addgene, 52961) with blasticidin S deaminase sequence
577 to construct the lentiCRISPR-v2-Blast-Puro plasmid (Addgene, 167186). The PCR products
578 were purified via gel excision and column purification (Promega, A9282), and then inserted into
579 the BsmBI-digested lentiCRISPR-v2-Blast-Puro vector by Gibson assembly (New England
580 Biolabs, E2621L). The assembled products were transformed into NEB 5-alpha
581 electrocompetent *E. coli* cells (New England Biolabs, C2989K) by electroporation. About 40
582 million independent bacterial colonies were cultured, and sgRNA library plasmids were extracted
583 with the Qiagen EndoFree Plasmid Mega Kit (Qiagen, 12381). The recovery rate and distribution
584 of the sgRNA library were checked with next generation sequencing (fig. S2, A to D).
585
586 **Lentivirus production and titration**
587 To make the lentiviral library, 5 μg of sgRNA plasmid library was co-transfected with 3 μg of
588 psPAX (Addgene, 12260) and 1 μg of pMD2.G (Addgene, 12259) lentivirus packaging plasmids
589 into 8 million HEK293T cells in a 10-cm dish with PolyJet (SignaGen Laboratories, SL100688).
590 For each individual sgRNA, 3.75 μg of sgRNA plasmid was co-transfected with 2.25 μg of psPAX
591 (Addgene, 12260) and 0.75 μg of pMD2.G (Addgene, 12259) plasmids into 4 million HEK293T
592 cells in a T25 flask with PolyJet (SignaGen Laboratories, SL100688). Media was replaced 12 h
593 after transfection, and harvested every 24 h for a total of three harvests. Harvested media
594 containing the desired virus were filtered through Millex-HV 0.45-μm PVDF filters (Millipore,
595 SLHV033RS) and further concentrated with 100,000 NMWL Amicon Ultra-15 centrifugal filter
596 units (Amicon, UFC910008).
597

598 The titer of lentivirus was determined by transducing 500,000 cells with varying amount (0, 0.5,
599 1.0, 2.0, 4.0 and 8.0 μL) of concentrated virus and polybrene (Millipore, TR-1003-G, 8 μg/mL,).
600 Viral transduction was performed by centrifuging the lentivirus and cell combination at 1000 RCF
601 for 90 min at 37°C. 3 to 4 h later, virus containing media was replaced with fresh media. 24 h
602 after the transduction, transduced cells were dissociated with Accutase, and seeded as
603 duplicates. One replicate was treated with blasticidin (Gibco, A1113903, 4 μg/mL), and the other
604 replicate was not treated with blasticidin. Four days later, the blasticidin resistant cells and
605 control cells were counted to calculate the ratio of infected cells and the viral titer.
606
607 **Determining 6TG concentration via killing curve titration**
608 Both CRISPRi and CRISPRn WTC11 iPSCs were used to determine the minimal lethal
609 concentration of 6TG. Cells were seeded in 24 well plates. When the cells reached around 50%
610 confluence (Day 0), they were treated with 6TG concentrations of 0 (control), 20, 40, 60, 80,
611 100, 120, 140, and 160 ng/mL. Two wells were allocated for each condition. The cells were
612 examined daily and cultured for 7 days. The media was replaced daily with the specified 6TG
613 concentration. After 3 days, wells with 6TG concentration greater than or equal to 100 ng/mL
614 had no surviving cells. On Day 4 of treatment, the wells with 80 ng/mL 6TG treatment had no
615 surviving cells. On the last day of treatment, the wells with 40 and 60 ng/mL treatments had very
616 few surviving cells, while the 20 ng/mL treatment had many surviving cells. Based on these
617 results, we set 80 ng/mL as the minimal lethal concentration for 6TG.
618
619 **CRISPRpath screening and sequencing library preparation**
620 CRISPRpath screens were carried out with 72 million doxycycline inducible CRISPRi or
621 CRISPRn iPSCs in biological replicate. The cells for lentiviral transduction were seeded into 6
622 well plates with 1 million cells per well, and the lentiviral library (MOI = 0.5) was transduced into
623 the iPSCs with 8 μg/mL Polybrene (Millipore, TR-1003-G) and spun at 1000 RCF at 37°C for 90
624 min. The transduced cells were treated with doxycycline (Sigma, D9891, 2 μM) and blasticidin
625 (Gibco, A1113903, 4 μg/mL) for 4 days. After this doxycycline and blasticidin treatment, 10
626 million cells were reserved as a control population, and 100 million cells were used for
627 CRISPRpath screen with doxycycline and 6TG (Sigma, A4660) treatment for 7 days. Finally,
628 survival cells were collected from the 6TG treated population.

629

630 The genomic DNA was extracted from each sample via cell lysis and digestion (100 mM Tris-
631 HCl pH8.5, 5 mM EDTA, 200 mM NaCl, 0.2% SDS, 100 μg/mL proteinase k), phenol:chloroform
632 (Thermo Scientific, 17908) extraction and isopropanol (Fisher Scientific, BP2618500)
633 precipitation. To amplify the sgRNA sequences from each sample, thirty-two 50 μl PCR reactions
634 were performed using 500 ng genomic DNA for each reaction and NEBNext® High-Fidelity 2X
635 PCR Master Mix (New England Biolabs, M0541S). The purified libraries were sequenced on the
636 NovaSeq 6000 with 150-bp paired-end sequencing. The detail protocol is available on ENCODE
637 portal (https://www.encodeproject.org/documents/2e6451a9-3b98-4d95-922e-a3d8d2100ddf/).
638

639 **CRISPRpath data analysis**

640 The sequence files were down-sampled to the same amount of total reads, and then mapped to
641 the sgRNA library with the requirement of exact match of designed sgRNA sequences in the
642 following patten 5'-CCG-[N19 or N20]-GTT-3'. Only the highly specific sgRNAs (specificity score
643 >0.2, without perfectly matched or 1-2 mismatches off-target sites) were used for downstream
644 data analysis. The sgRNA enrichment for each screen was calculated by comparing 6TG treated
645 samples with the associated control samples with edgeR and TMM normalization. We first used
646 edgeR (*38*) to calculate *P* value based on negative binomial model for both targeting sgRNAs
647 and non-targeting control sgRNAs. To achieve empirical false discovery rate less than 5%, we
648 then selected a *P* value cutoff corresponding to the 5% percentile of *P* values from non-targeting
649 control sgRNAs. Finally, we defined enriched sgRNAs with a *P* value less than the selected *P*
650 value cutoff, and a fold change >2. The ATAC-seq peaks were identified as functional enhancers
651 for the six MMR genes by having at least 3 significant enriched sgRNAs. Analysis scripts are
652 available at https://github.com/MichaelMW/crispy.
653

654 **Analysis of genomic feature and chromatin signature of identified enhancers**

655 Genomic distances between enhancer and TSS pairs were calculated based on the distance
656 from the center of enhancers to the transcription start sites of the target genes. The number of
657 interval genes is the number of all the RefSeq annotated genes between each enhancer and
658 paired target gene. The signal of chromatin signatures, including ATAC-seq, H3K27ac,
659 H3K4me3, CTCF binding and RNA-seq, were calculated by deeptools (v3.4.3) (*39*). The

660 enhancer-like promoters are the enhancers overlap with the region 500 bp upstream and
661 downstream of a RefSeq annotated TSS.
662
663 **Validation of identified enhancers using CRISPRi**
664 We cloned lentiCRISPR-v2-HygR-EGFP (Addgene, 167188) and lentiCRISPR-v2-HygR-
665 mCherry (Addgene, 167189) vectors by replacing the Cas9 and puromycin N-acetyltransferase
666 sequences in lentiCRISPR v2 plasmid (Addgene, 52961) with hygromycin B phosphotransferase
667 and EGFP or mCherry sequences. To validate the identified enhancers, individual sgRNAs
668 targeting identified enhancers were cloned into the lentiCRISPR-v2-HygR-GFP or lentiCRISPR-
669 v2-HygR-mCherry vector. The doxycycline inducible CRISPRi WTC11 iPSCs were infected with
670 the lentivirus expressing sgRNAs for three replicates per sgRNA. The sgRNA infected cells were
671 grown with hygromycin (Gibco, 10687010, 150 μg/mL) and doxycycline (Sigma, D9891, 2 μM)
672 containing media. Seven days later, the cells were collected and total RNA was extracted from
673 the cells using the Qiagen RNeasy® Plus Kit (Qiagen, 74134). One μg of RNA was then used
674 to synthesize cDNA using the Bio-RAD iScript cDNA Synthesis Kit (Bio-RAD, 1708840). qPCR
675 reactions for targeted genes were performed with the Luminaris HiGreen qPCR Master Mix
676 (Thermo Scientific, K0993) on the Roche LightCycler 96 System. The qPCR primers are listed
677 in table S4 and the sgRNA sequences are listed in table S6. For each tested element in Fig. 3,
678 A to C and fig. S6B, we performed CRISPRi experiments with two independent sgRNAs and
679 used the results from the sgRNA with stronger transcriptional repression in Fig. 4B.
680
681 **shRNA mediated RNA interference**
682 shRNAs were designed by using DSIR tool (http://biodev.extra.cea.fr/DSIR/DSIR.html) targeting
683 *SOCS5*, *FOXN2* and *TMEM230*. The sequences of shRNAs are listed in table S5. The shRNAs
684 were cloned into lentiCRISPR-v2-HygR-mCherry vector under the control of human U6 promoter
685 and packaged into lentivirus for cell transduction. The WTC11 iPSCs transduced with shRNA
686 lentivirus were treated with hygromycin (Gibco, 10687010, 150 μg/mL) for 7 days and then
687 collected for RNA extraction and qPCR.
688
689 **ATAC-seq**

690 ATAC-seq was carried out using the Nextera DNA Library Prep Kit (Illumina, FC-121-1030) as
691 previously described (*40*). The detailed protocol is available on the ENCODE portal
692 (https://www.encodeproject.org/documents/0317894c-5a42-4f03-b865-c2a2d08708ef/). Briefly,
693 each library started with 100,000 fresh iPSCs, and the cells were incubated with ice cold nuclei
694 extraction buffer (10 mM TrisHCl pH 7.5, 10 mM NaCl, 3 mM MgCl2, 0.1% Igepal CA630, and
695 1x protease inhibitor) for 5 min on ice, then centrifuged at 500 RCF for 5 min. 50,000 resulting
696 nuclei were treated with tagmentation buffer (25 µL Buffer TD with 50,000 nuclei, 22.5 µL water,
697 2.5 µL TDE1) for 30 min at 37°C. The transposed DNA was purified using Qiagen MinElute PCR
698 purification kit (Qiagen, 28006), amplified using Nextera primers, then size-selected for
699 fragments between 150 and 1000 bp using SPRISelect beads (Beckman Coulter, B23319).
700 Libraries were sent for single-end sequencing on the HiSeq 4000 (50 bp single-end reads).
701 Reads were mapped to hg38/GRCh38 and processed using the ENCODE pipeline
702 (https://github.com/kundajelab/atac_dnase_pipelines, V1.8.0), which ran on the default settings.
703 The ATAC-seq peaks were filtered with FDR cutoff of 0.1%, and adjacent peaks were merged if
704 they are less than 1 kb apart.
705
706 **RNA-seq**
707 RNA was extracted from fresh cells using the RNeasy Plus Mini Kit (Qiagen, 74134).
708 Approximately 1000 ng of extracted RNA was used to prepare libraries for sequencing using the
709 TruSeq Stranded mRNA Library Prep Kit (Illumina, 20020594). Libraries were sent for paired-
710 end sequencing on the NovaSeq 6000 (100 bp paired-end reads). Reads were aligned to
711 hg38/GRCh38 using STAR 2.7.0f (*41*) with the standard ENCODE settings, and transcript
712 quantification was performed in a strand-specific manner using RSEM 1.3.1 (*42*) with the
713 annotation from GENCODE v32. Only the first read was used, and all reads were trimmed to
714 51bp using TrimGalore 0.4.5 running the following options: -q 20 --length 20 -- stringency 3 --
715 trim-n. The edgeR package in R (3.20.9) (*38*) was used to calculate TMM-normalized FPKM
716 values for each gene based on the expected counts and gene lengths for each library. The mean
717 gene expression across all replicates was used for analysis.
718
719 **ChIP-seq**

720 ChIP-seq libraries were constructed from 2 million WTC11 iPSCs. Cells were crosslinked in 1%
721 formaldehyde at room temperature for 20 min and then quenched with 2.5 M glycine at room
722 temperature for 5 min. Fixed cells were lysed and chromatin was sonicated by Covaris S220
723 focused-ultrasonicator with the following parameters: Duty Factor: 2%, Peak Incident Power:
724 105W, Cycles per Burst: 200, for 30 min. Input chromatin was removed and stored at -20°C for
725 later processing. Magnetic beads (Invitrogen, Dynabeads Protein A, 10001D) were preincubated
726 with H3K27ac antibody (Active Motif, 39133, Lot 22618011) for 2 hours at 4°C before being
727 added to sheared chromatin. Samples were incubated overnight at 4°C. Beads were washed 3
728 times and chromatin was then eluted. Samples were incubated at 65°C overnight to reverse the
729 crosslinking. DNA was treated with RNase A for 1 hr at 37°C and Proteinase K (New England
730 Biolabs, 8107) for 1 hr at 55°C. DNA was purified by phenol-chloroform extraction and ethanol
731 precipitation. Libraries were prepared using Tru-seq adapters and size-selected using
732 SPRIselect beads prior to amplification and paired-end sequencing. Libraries were sent for
733 paired-end sequencing on the NovaSeq 6000 (150 bp paired-end reads). Sequencing reads
734 were trimmed to 50 bp and mapped to hg38 using bowtie2 with the following options: --local --
735 very-sensitive-local --no-unal --no-mixed --no-discordant --phred33 -I 10 -X 700. Picard Tools
736 was used to remove blacklisted regions and duplicate reads and MACS2 was used to call peaks
737 on merged replicates at an FDR cutoff of 1%.
738
739 **CUT&Tag**
740 CUT&Tag libraries were constructed from 150,000 WTC11 iPSCs according to previously
741 described methods (*43*). Cells were lysed in nuclei extraction buffer (20 mM HEPES-KOH pH
742 7.9, 10 mM MgCl2, 0.1% Triton X-100, 20% glycerol and 1x protease inhibitor) on ice for 10 min.
743 The samples were spun and resuspended in 100 µl nuclei extraction buffer. Meanwhile, 10 µl of
744 BioMag Plus Concanavalin A (Bangs Laboratories, BP531) were equilibrated in binding buffer
745 (1x PBS, 1 mM $CaCl_2$, 1 mM $MgCl_2$ and 1 mM $MnCl_2$). The equilibrated beads were added to
746 the samples and incubated with rotation for 15 min at 4°C. Nuclei-bound beads were washed
747 with Buffer 1 (20 mM HEPES-KOH pH 7.9, 150 mM NaCl, 2 mM EDTA, 0.5 mM spermidine,
748 0.1% BSA and 1x protease inhibitor) and Buffer 2 (20 mM HEPES-KOH pH 7.9, 150 mM NaCl,
749 0.5 mM spermidine, 0.1% BSA and 1x protease inhibitor). After washing nuclei-bound beads
750 were resuspended in 50 µl Buffer 2 with 0.5 µl antibody (H3K4me3 from Millipore, 04-745, Lot

751 3543820 and CTCF from Millipore, 07-729, Lot 3059608) and incubated with rotation overnight
752 at 4°C. Samples were washed twice with Buffer 2 and resuspended in 50 µl Buffer 2 with
753 antibody (antibodies-online Inc., Guinea Pig anti-Rabbit IgG, ABIN101961, Lot 42323) and
754 incubated for 1 hr at room temperature with rotation. Samples were washed again with Buffer 2
755 and resuspended in 100 µl Buffer 3 (20 mM HEPES-KOH pH 7.9, 300 mM NaCl, 0.5 mM
756 Spermidine, 0.1% BSA and 1x proteinase inhibitor) containing 0.04 µM pA-Tn5. Samples were
757 incubated for 1 hr at room temperature, washed three times with Buffer 3 and resuspended in
758 tagmentation buffer (20 mM HEPES-KOH pH 7.9, 300 mM NaCl, 0.5 mM Spermidine, 10 mM
759 $MgCl_2$, 0.1% BSA and 1x proteinase inhibitor). Samples were incubated for 1 hr at 37°C.
760 Samples were treated with Proteinase K (New England Biolabs, 8107) for 1 hr at 50°C. DNA
761 was purified by phenol-chloroform extraction and ethanol precipitation. Libraries were prepared
762 using Tru-seq adapters and size-selected using SPRIselect beads prior to amplification and
763 paired-end sequencing. Libraries were sent for paired-end sequencing on the Mini-seq (37 bp
764 paired-end reads, H3K4me3 libraries) or NovaSeq 6000 (150 bp paired-end reads, CTCF
765 libraries). Sequencing reads (CTCF libraries were trimmed to 50 bp) were mapped to hg38 using
766 bowtie2 with the following options: --local --very-sensitive-local --no-unal --no-mixed --no-
767 discordant --phred33 -I 10 -X 700. Picard Tools was used to remove blacklisted regions and
768 duplicate reads and SEACR (*44*) was used to call peaks on merged replicates.
769
770 **H3K4me3 PLAC-seq**
771 H3K4me3 PLAC-seq data in WTC11 cells were generated as previously described (*45*) in
772 biological replicates (clone 6 and clone 28) (https://data.4dnucleome.org/experiment-set-
773 replicates/4DNESDRL4ZKM/ and https://data.4dnucleome.org/experiment-set-
774 replicates/4DNESIZ5TTHO/). We combined the two biological replicates, and applied the MAPS
775 pipeline (*46*) to identify significant long-range chromatin interactions at 5 kb bin resolution for the
776 genomic distance 10 kb ~ 1 Mb. The reference genome is GRCh38. In addition, for each 5 kb
777 bin pair anchored at H3K4me3 peaks, the MAPS pipeline outputs the normalized contact
778 frequency, which adjusts for the biases from effective fragment length, GC content, sequence
779 mappability, H3K4me3 enrichment level and 1D genomic distance effect.
780

781 **Comparison between strong enhancers and weak enhancers using H3K4me3 PLAC-seq**
782 **data**
783 For 4 genes *HPRT1*, *MLH1*, *PMS2* and *PCNA*, there are 23 enhancer-promoter pairs between
784 strong enhancers and their target genes, and 21 enhancer-promoter pairs between weak
785 enhancers and their target genes. We mapped each enhancer and promoter of target gene into
786 5 kb bins, and obtained the distance normalized H3K4me3 PLAC-seq contact frequency for 5
787 kb bin pairs containing the enhancer-promoter pairs. Since *MSH2* and *MSH6* are located within
788 407 kb linear genomic distance with each other and we can't assign enhancers to either gene
789 reliably, enhancers identified near *MSH2* and *MSH6* were excluded from this analysis.
790
791 **Comparison between enhancer-like promoters and control promoters using H3K4me3**
792 **PLAC-seq data**
793 For this analysis, control promoters are active promoter regions with annotated ATAC-seq peaks
794 and tested negative as enhancers for the for MMR genes. We mapped each promoter into a 5
795 kb bin that was used in the PLAC-seq analysis. We only choose the bins with one annotated
796 active promoter, which gave us 31 enhancer-like promoters and 43 control promoters in this
797 analysis. We counted the number of significant H3K4me3 PLAC-seq interactions anchored at
798 the 5 kb bins with these promoter sequences. In addition, as described in our previous study
799 (*23*), for promoters with at least one significant interaction, we calculated the summation of –
800 log10 FDR of significant interactions, which is a measure of the overall interaction strength.
801
802 **Chromatin contact frequency comparison between strong enhancers and weak**
803 **enhancers in K562 cells and mESCs**
804 For the chromatin contact frequency comparison of enhancers in K562 cells and mouse
805 embryonic stem cells (mESCs), we downloaded the identified enhancers from each publication
806 (*6, 8*), and defined strong enhancer with cutoff of 50% ≤ transcriptional contribution ≤ 100%,
807 weak enhancer with cutoff of 0% < transcriptional contribution ≤ 20%. H3K27ac HiChIP data in
808 K562 cells(*47*) and H3K4me3 PLAC-seq data in mESCs (*46*) were used for comparison. The
809 comparisons were performed in 10 kb resolution.
810
811 **Motif scan and transcription factor identification**

812 The fasta files were first generated in the hg38 genome for the identified strong enhancers and
813 weak enhancers separately. For each strong enhancer and weak enhancer, the FIMO software
814 (version 5.1.0) (*48*) with human motif database HOCOMOCO (v11 FULL) (*49*) was used to scan
815 the motifs. All the FIMO motif scans were in default settings. We then filtered the transcription
816 factors (TFs) in each strong and weak enhancer loci by FDR cutoff of 0.05 and p-value cutoff of
817 0.0001 and gene expressions cutoff of FPKM > 1. By taking the TFs with TF motif appeared in
818 more than 80% enhancers, 47 TFs were considered as commonly appearing in the strong
819 enhancers, and 35 TFs were in the weak enhancers.
820
821 **Dual-luciferase reporter assay**
822 Dual-luciferase reporter assay system (Progema, E1910) was used to test the enhancer activity
823 of weak and strong enhancers. The weak enhancers, strong enhancers and negative control
824 elements (table S8) were PCR amplified from WTC11 iPSCs genomic DNA with NEBNext®
825 High-Fidelity 2X PCR Master Mix (New England Biolabs, M0541S). The amplified DNA elements
826 and synthesized minimal promoter were cloned into XhoI and NcoI-digested pGL4.13 vector
827 (Progema, E6681) by Gibson assembly (New England Biolabs, E2621L). After validation of the
828 sequence by Sanger sequencing, the vectors were co-transfected with pRL-CMV-Renilla
829 luciferase vector (Promega, E2261) in WTC11 iPSCs with FuGENE HD (Promega, E2311) at a
830 3:1 reagent to DNA ratio. pGL4.13-SV40-Firefly luciferase vector (Progema, E6681) was used
831 as a positive control. The luciferase activity was measured 48 hours after transfection with
832 BioTek Synergy 2 multi-mode microplate reader. The relative firefly luciferase activity of each
833 sample was normalized to the average of the activities of minimal promoter.
834

835 **References**

836 1. F. Soldner *et al.*, Parkinson-associated risk variant in distal enhancer of alpha-synuclein
837 modulates target gene expression. *Nature* **533**, 95-99 (2016).
838 2. A. Visel, E. M. Rubin, L. A. Pennacchio, Genomic views of distant-acting enhancers. *Nature*
839 **461**, 199-205 (2009).
840 3. H. K. Long *et al.*, Loss of Extreme Long-Range Enhancers in Human Neural Crest Drives a
841 Craniofacial Disorder. *Cell Stem Cell* **27**, 765-783 e714 (2020).
842 4. E. P. Consortium *et al.*, Expanded encyclopaedias of DNA elements in the human and mouse
843 genomes. *Nature* **583**, 699-710 (2020).
844 5. N. E. Sanjana *et al.*, High-resolution interrogation of functional elements in the noncoding
845 genome. *Science* **353**, 1545-1549 (2016).
846 6. M. Gasperini *et al.*, A Genome-wide Framework for Mapping Gene Regulation via Cellular
847 Genetic Screens. *Cell* **176**, 377-390 e319 (2019).
848 7. Y. Diao *et al.*, A tiling-deletion-based genetic screen for cis-regulatory element identification
849 in mammalian cells. *Nat Methods* **14**, 629-635 (2017).
850 8. C. P. Fulco *et al.*, Activity-by-contact model of enhancer-promoter regulation from thousands
851 of CRISPR perturbations. *Nat Genet* **51**, 1664-1669 (2019).
852 9. D. R. Simeonov *et al.*, Discovery of stimulation-responsive immune enhancers with CRISPR
853 activation. *Nature* **549**, 111-115 (2017).
854 10. T. S. Klann *et al.*, CRISPR-Cas9 epigenome editing enables high-throughput screening for
855 functional regulatory elements in the human genome. *Nat Biotechnol* **35**, 561-568 (2017).
856 11. C. P. Fulco *et al.*, Systematic mapping of functional enhancer-promoter connections with
857 CRISPR interference. *Science* **354**, 769-773 (2016).
858 12. M. Gasperini *et al.*, CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for
859 HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions. *Am J Hum Genet*
860 **101**, 192-205 (2017).
861 13. N. Pecina-Slaus, A. Kafka, I. Salamon, A. Bukovac, Mismatch Repair Pathway, Genome Stability
862 and Cancer. *Front Mol Biosci* **7**, 122 (2020).
863 14. T. Yan, S. E. Berry, A. B. Desai, T. J. Kinsella, DNA mismatch repair (MMR) mediates 6-
864 thioguanine genotoxicity by introducing single-strand breaks to signal a G2-M arrest in
865 MMR-proficient RKO cells. *Clin Cancer Res* **9**, 2327-2334 (2003).
866 15. G. M. Li, Mechanisms and functions of DNA mismatch repair. *Cell Res* **18**, 85-98 (2008).
867 16. M. A. Mandegar *et al.*, CRISPR Interference Efficiently Induces Specific and Reversible Gene
868 Silencing in Human iPSCs. *Cell Stem Cell* **18**, 541-553 (2016).
869 17. J. Tycko *et al.*, Mitigation of off-target toxicity in CRISPR-Cas9 screens for essential non-
870 coding elements. *Nat Commun* **10**, 4063 (2019).
871 18. A. R. Perez *et al.*, GuideScan software for improved single and paired CRISPR guide RNA
872 design. *Nat Biotechnol* **35**, 347-349 (2017).
873 19. A. Radzisheuskaya, D. Shlyueva, I. Muller, K. Helin, Optimizing sgRNA position markedly
874 improves the efficiency of CRISPR/dCas9-mediated transcriptional repression. *Nucleic Acids*
875 *Res* **44**, e141 (2016).
876 20. J. Rosenbluh *et al.*, Complementary information derived from CRISPR Cas9 mediated gene
877 deletion and suppression. *Nat Commun* **8**, 15403 (2017).
878 21. K. Li *et al.*, Interrogation of enhancer function by enhancer-targeting CRISPR epigenetic
879 editing. *Nat Commun* **11**, 485 (2020).

880 22. J. M. Engreitz *et al.*, Local regulation of gene expression by lncRNA promoters, transcription
881 and splicing. *Nature* **539**, 452-455 (2016).
882 23. M. Song *et al.*, Cell-type-specific 3D epigenomes in the developing human cortex. *Nature* **587**,
883 644-649 (2020).
884 24. N. Kubo *et al.*, Promoter-proximal CTCF binding promotes distal enhancer-dependent gene
885 activation. *Nat Struct Mol Biol*, (2021).
886 25. M. E. Fernandez-Zapico *et al.*, A functional family-wide screening of SP/KLF proteins
887 identifies a subset of suppressors of KRAS-mediated cell growth. *Biochem J* **435**, 529-537
888 (2011).
889 26. B. Ren *et al.*, E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M
890 checkpoints. *Genes Dev* **16**, 245-256 (2002).
891 27. S. Polager, Y. Kalma, E. Berkovich, D. Ginsberg, E2Fs up-regulate expression of genes involved
892 in DNA replication, DNA repair and mitosis. *Oncogene* **21**, 437-446 (2002).
893 28. M. C. Canver *et al.*, BCL11A enhancer dissection by Cas9-mediated in situ saturating
894 mutagenesis. *Nature* **527**, 192-197 (2015).
895 29. N. Rajagopal *et al.*, High-throughput mapping of regulatory DNA. *Nat Biotechnol* **34**, 167-174
896 (2016).
897 30. H. Koike-Yusa, Y. Li, E. P. Tan, C. Velasco-Herrera Mdel, K. Yusa, Genome-wide recessive
898 genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat*
899 *Biotechnol* **32**, 267-273 (2014).
900 31. B. Adamson *et al.*, A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic
901 Dissection of the Unfolded Protein Response. *Cell* **167**, 1867-1882 e1821 (2016).
902 32. Y. Diao *et al.*, A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-
903 mediated genetic screening. *Genome Res* **26**, 397-405 (2016).
904 33. S. Xie, J. Duan, B. Li, P. Zhou, G. C. Hon, Multiplexed Engineering and Analysis of Combinatorial
905 Enhancer Activity in Single Cells. *Mol Cell* **66**, 285-299 e285 (2017).
906 34. M. A. Zabidi *et al.*, Enhancer-core-promoter specificity separates developmental and
907 housekeeping gene regulation. *Nature* **518**, 556-559 (2015).
908 35. L. T. M. Dao, S. Spicuglia, Transcriptional regulation by promoters with enhancer function.
909 *Transcription* **9**, 307-314 (2018).
910 36. O. Mikhaylichenko *et al.*, The degree of enhancer or promoter activity is reflected by the
911 levels and directionality of eRNA transcription. *Genes Dev* **32**, 42-57 (2018).
912 37. H. F. Thomas *et al.*, Temporal dissection of an enhancer cluster reveals distinct temporal and
913 functional contributions of individual elements. *Mol Cell*, (2021).
914 38. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: a Bioconductor package for differential
915 expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).
916 39. F. Ramirez *et al.*, deepTools2: a next generation web server for deep-sequencing data
917 analysis. *Nucleic Acids Res* **44**, W160-165 (2016).
918 40. J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf, Transposition of native
919 chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding
920 proteins and nucleosome position. *Nat Methods* **10**, 1213-1218 (2013).
921 41. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
922 42. B. Li, C. N. Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or
923 without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
924 43. H. S. Kaya-Okur *et al.*, CUT&Tag for efficient epigenomic profiling of small samples and single
925 cells. *Nat Commun* **10**, 1930 (2019).

926 44. M. P. Meers, D. Tenenbaum, S. Henikoff, Peak calling by Sparse Enrichment Analysis for
927 CUT&RUN chromatin profiling. *Epigenetics Chromatin* **12**, 42 (2019).
928 45. R. Fang *et al.*, Mapping of long-range chromatin interactions by proximity ligation-assisted
929 ChIP-seq. *Cell Res* **26**, 1345-1348 (2016).
930 46. I. Juric *et al.*, MAPS: Model-based analysis of long-range chromatin interactions from PLAC-
931 seq and HiChIP experiments. *PLoS Comput Biol* **15**, e1006982 (2019).
932 47. M. R. Mumbach *et al.*, Enhancer connectome in primary human cells identifies target genes
933 of disease-associated DNA elements. *Nat Genet* **49**, 1602-1612 (2017).
934 48. C. E. Grant, T. L. Bailey, W. S. Noble, FIMO: scanning for occurrences of a given motif.
935 *Bioinformatics* **27**, 1017-1018 (2011).
936 49. I. V. Kulakovskiy *et al.*, HOCOMOCO: towards a complete collection of transcription factor
937 binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res* **46**,
938 D252-D259 (2018).
939

Fig. 1

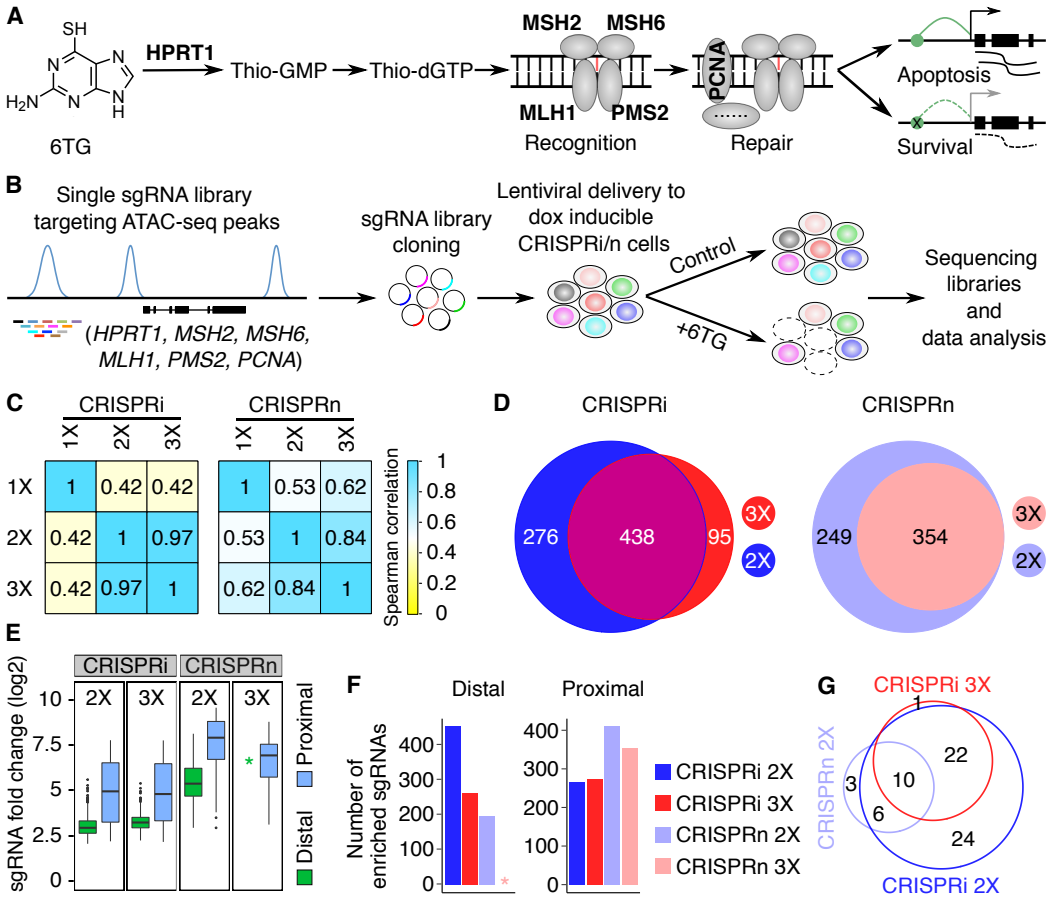

Fig. 1. CRISPRpath for identifying enhancers of multiple genes.

(**A**) Six genes (*HPRT1*, *MSH2*, *MSH6*, *MLH1*, *PMS2* and *PCNA*) in the 6TG-induced mismatch repair process were used for CRISPRpath screen in this study. (**B**) Schematic of the CRISPRpath screening strategy with 6TG treatment in iPSCs. Cell survival was used as readout for the screen. (**C**) Spearman correlation analysis of sgRNA ranking based on fold changes for CRISPRpath screens with different 6TG concentrations (1X, 2X and 3X). (**D**) Venn diagram shows the overlapping enriched sgRNAs identified from the screens with 2X and 3X 6TG treatments. (**E**) Box plots show the fold changes of the enriched distal and proximal sgRNAs from 2X, 3X CRISPRi and CRISPRn screens. Star indicates no enriched distal sgRNA was identified from 3X CRISPRn screen. Boxplots indicate the median, interquartile range (IQR), Q1 − 1.5 × IQR and Q3 + 1.5 × IQR. (**F**) Bar plot shows the number of enriched distal and proximal sgRNAs from 2X, 3X CRISPRi and CRISPRn screens. Star indicates no enriched distal sgRNA was identified from 3X CRISPRn screen. (**G**) Venn diagram shows the identified enhancers from each CRISPRpath screen.

Fig. 2

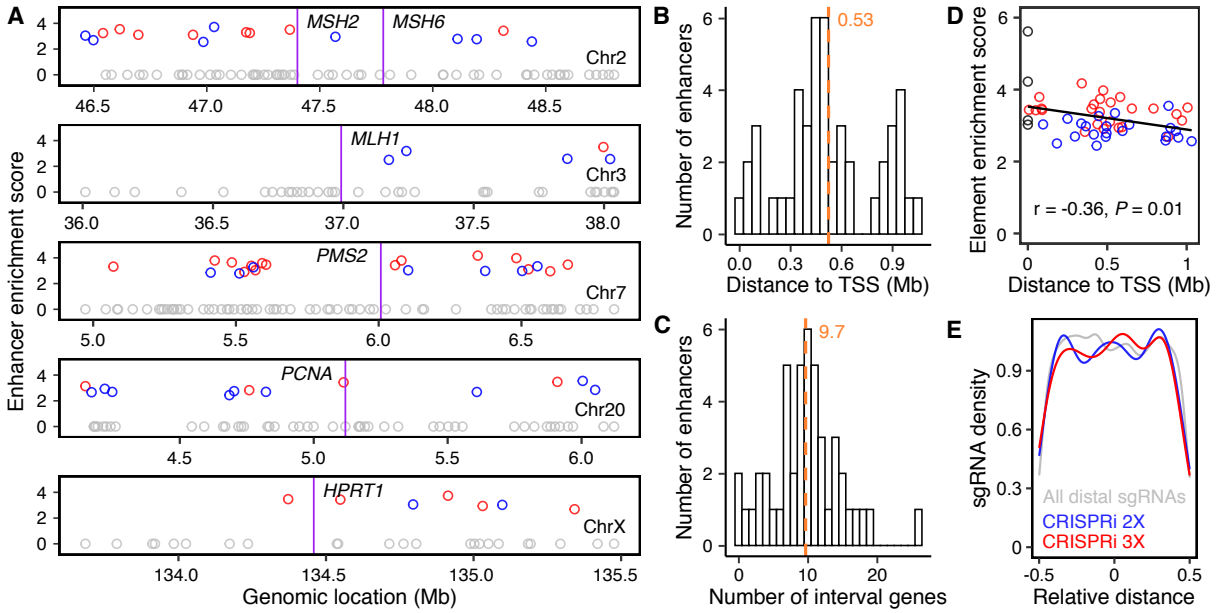

Fig. 2. Genomic features of identified enhancers from CRISPRpath using CRISPRi.

(**A**) Genomic locations of identified enhancers relative to TSS. Circles indicate enhancers identified from the CRISPRi 3X screen (red), enhancers uniquely identified from the CRISPRi 2X screen (blue), tested CREs that are not identified as enhancers (grey). Purple lines label the location of each target gene. (**B**) Histogram shows the distance distribution between identified enhancers and their paired TSS. (**C**) Histogram shows the number of interval genes between enhancers and their target gene TSS. Mean is indicated with an orange dashed line in B and C. (**D**) A weak negative correlation is observed between enrichment score and genomic distance between enhancers and their target genes (Pearson correlation, r = -0.36, *P* = 0.01). Black circles indicate TSS regions. The red and blue circles are enhancers showing in A. Only the enhancers for *HPRT1*, *MLH1*, *PMS2*, *PCNA* are included in B to D. (**E**) Density plot shows no significant difference (two-tailed two-sample Kolmogorov–Smirnov test) for the distribution of all distal sgRNAs (gray), enriched distal sgRNAs from 2X (blue) and 3X screen (red) CRISPRi screens.

Fig. 3

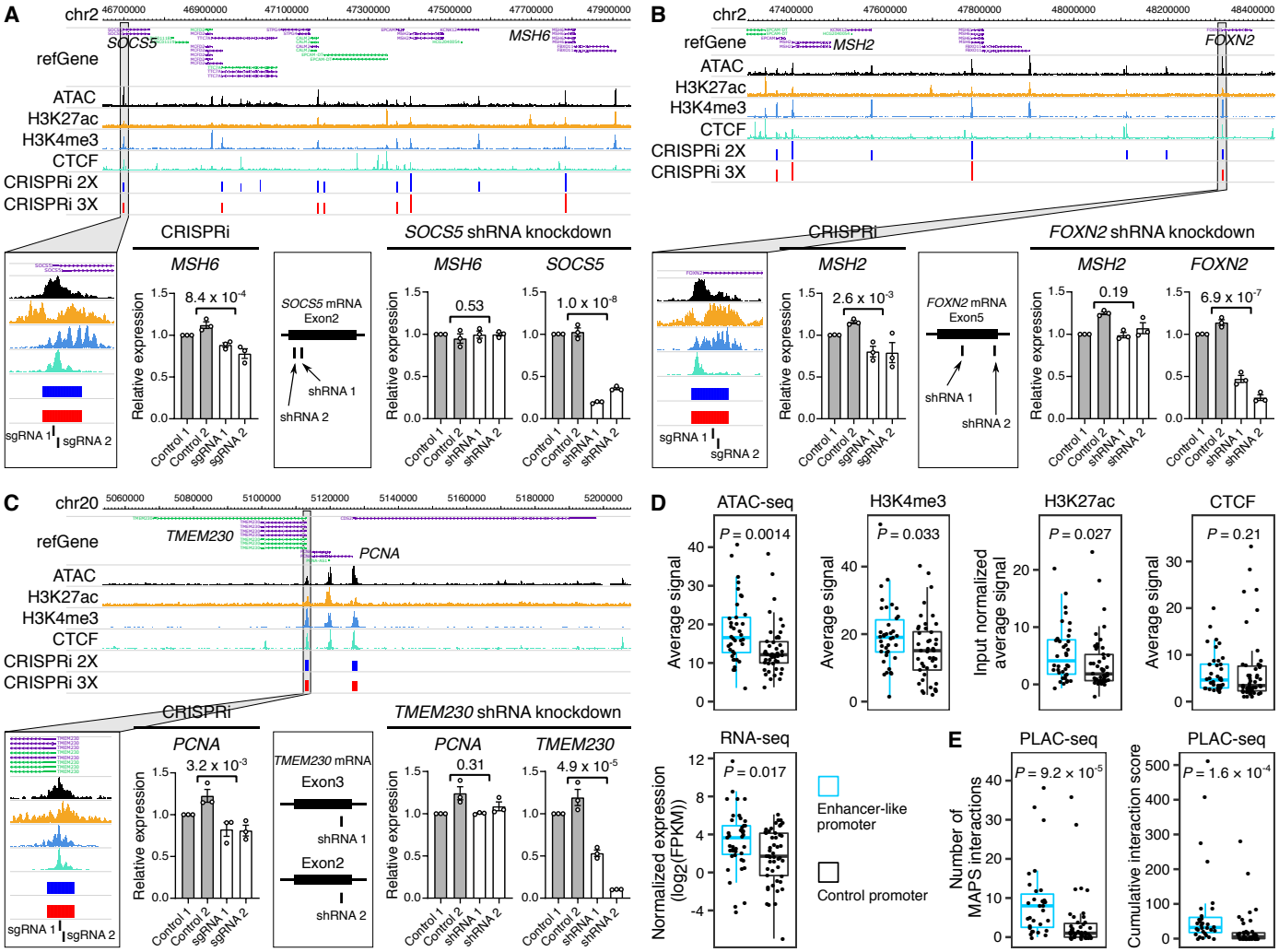

Fig. 3. Enhancer-like promoters act as functional enhancers.

(**A**, **B**, **C**) Three examples of promoters function as enhancers. CRISPRi silencing of the promoter region of *SOCS5*, *FOXN2* and *TMEM230* results in significant downregulation of *MSH6*, *MSH2* and *PCNA*, respectively. shRNA knockdown of *SOCS5*, *FOXN2* and *TMEM230* can only downregulate *SOCS5*, *FOXN2* and *TMEM230* expression. Three independent replicates per condition and two independent sgRNAs or shRNAs per replicate were used for each experiment. *P* values are from two-tailed two-sample *t*-test. (**D**) Average signal enrichment of ATAC-seq, gene transcription, H3K4me3, H3K27ac and CTCF binding for enhancer-like promoters (n = 38) and control promoters (n = 47). *P* values are from Wilcoxon test. Boxplots indicate the median, IQR, Q1 − 1.5 × IQR and Q3 + 1.5 × IQR. (**E**) Number of H3K4me3 mediated chromatin interactions and cumulative interaction score for enhancer-like promoters (n = 31) and control promoters (n = 43). Boxplots indicate median, IQR, Q1 − 1.5 × IQR and Q3 + 1.5 × IQR. *P* values are calculated from Wilcoxon test.

Fig. 4

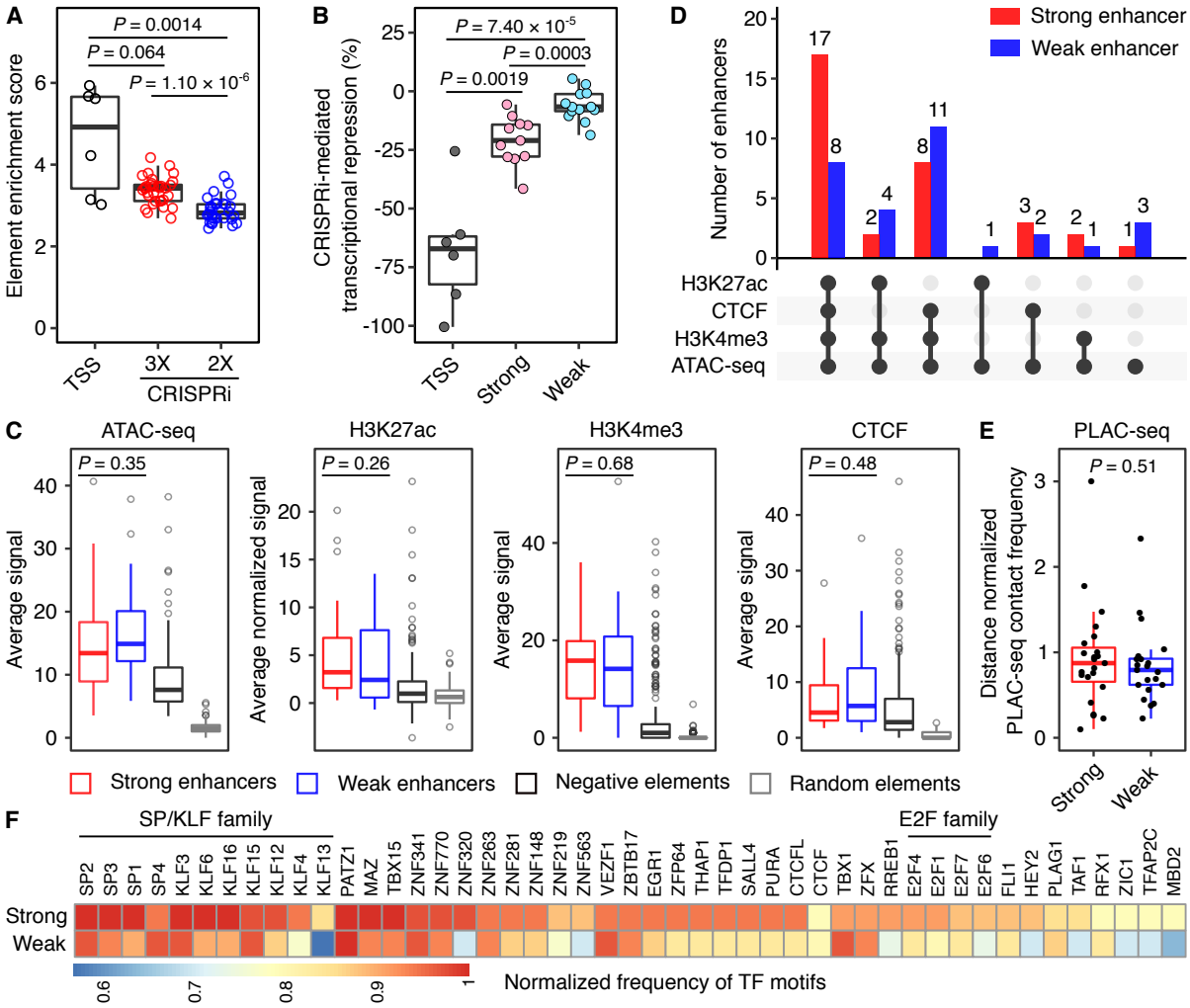

Fig. 4. CRISPRpath can distinguish weak and strong enhancers by imposing different selection pressures.

(**A**) Box plots show the enrichment score of the tested elements. TSS regions (black circles) show highest enrichment scores. Enhancers uniquely identified from the lower selection pressure (CRISPRi 2X, blue circles) exhibit lower enrichment scores compared to the enhancers identified from the higher selection pressure (CRISPRi 3X, red circles). Each circle represents an individual element. *P* values are from Wilcoxon test. (**B**) Box plots show the CRISPRi perturbation at enhancers induced various degrees of transcriptional repression of target genes measured with RT-qPCR. Each dot represents the average value from three biological replicates. CRISPRi targeting TSS regions (dark gray) achieved the highest transcriptional repression. CRISPRi targeting strong enhancers (pink) leads a more substantial transcription silencing on target gene compared to CRISPRi targeting weak enhancers (cyan). *P* values are from Wilcoxon test. (**C**) Enrichment analysis of ATAC-seq, H3K27ac, H3K4me3, CTCF binding signals for strong (n = 33) and weak (n = 30) enhancers. *P* values for the difference between strong and weak enhancers are from Wilcoxon test; see table S7 for *P* values of all pairwise comparisons. Boxplots (A to C) indicate the median, IQR, Q1 − 1.5 × IQR and Q3 + 1.5 × IQR. (**D**) Intersection of genomic features for weak enhancers (blue bar) and strong enhancers (red bar). (**E**) Distance normalized H3K4me3 PLAC-seq contact frequency for strong (n = 23) and weak (n = 21) enhancers. Only the enhancers for *HPRT1*, *MLH1*, *PMS2*, *PCNA* are included (see Materials and Methods for details). Boxplots indicate the median, IQR, Q1 − 1.5 × IQR and Q3 + 1.5 × IQR. *P* value is from Wilcoxon test. (**F**) Heatmap shows the normalized frequency of transcription factor motifs found in strong and weak enhancers.

Fig. S1

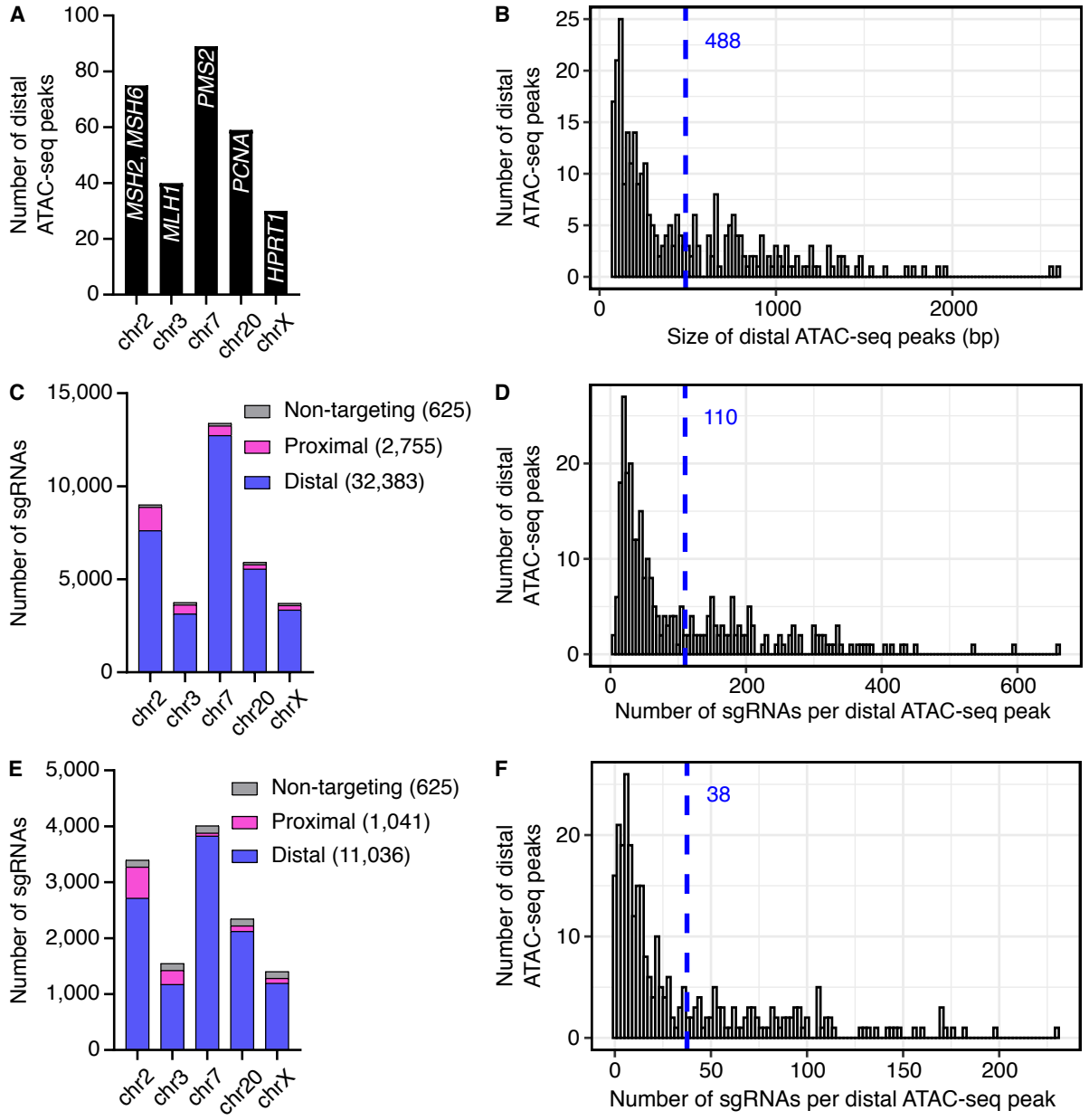

Fig. S1. Features of sgRNA library for CRISPRpath screen.

(**A**) Bar graph shows the number of distal ATAC-seq peaks used as candidate CREs for six MMR genes. (**B**) Histogram shows size distribution of distal ATAC-seq peaks. The average size is 488 bp (blue dash line). (**C**, **E**) The composition of the sgRNA library. In total, 35,763 sgRNAs were included in the library (C), and 12,702 sgRNAs are high quality sgRNAs (E). (**D**, **F**) Distribution of the number of total sgRNAs (D) or high quality sgRNAs (F) per distal ATAC-seq peak. Average numbers of sgRNA per ATAC-seq peak are indicated with blue dash lines.

Fig. S2

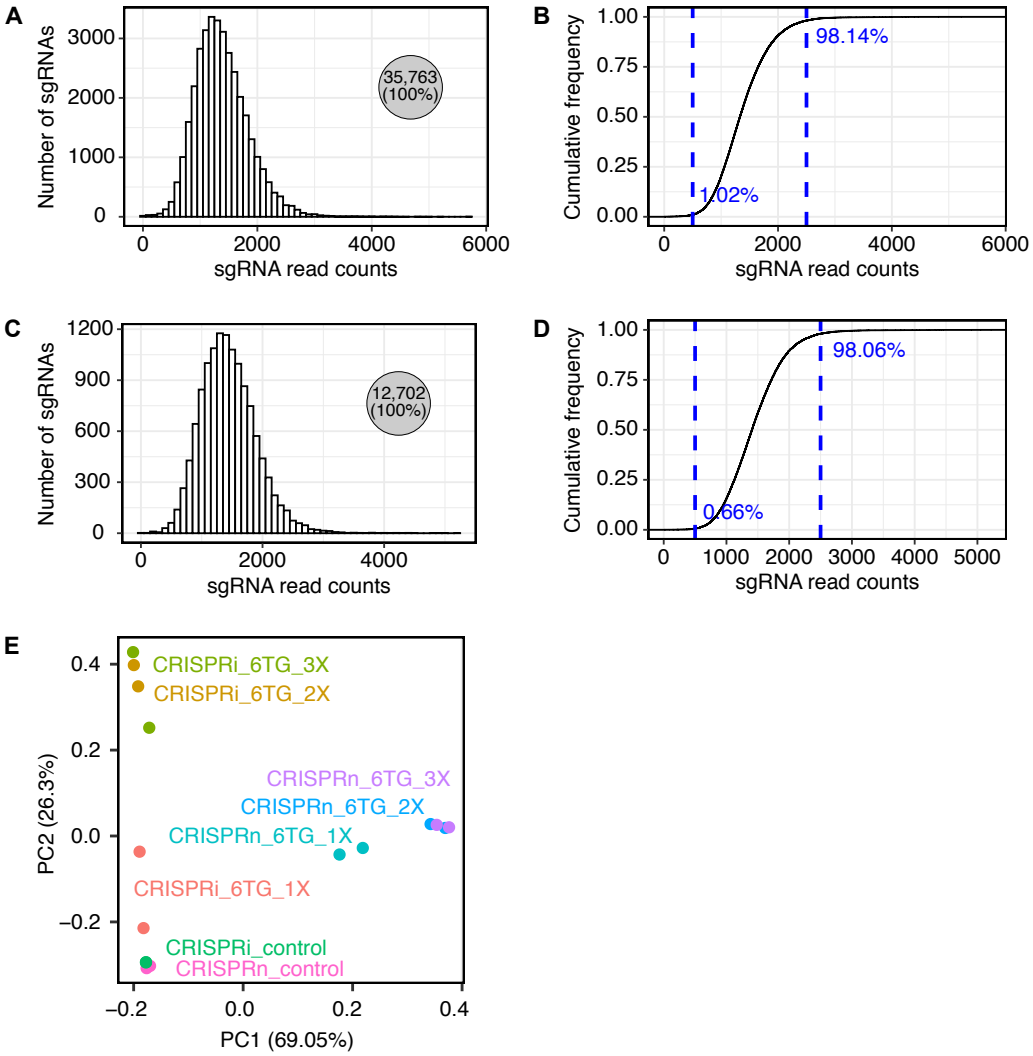

Fig. S2. Quality of the sgRNA library and CRISPRpath screen libraries.

(**A**, **C**) Distribution of sgRNA read counts for all the sgRNAs (A) and high quality sgRNAs (C) in the constructed sgRNA plasmid library. (**B**, **D**) Cumulative frequency of all the sgRNAs (B) and high quality sgRNAs (D) in the constructed sgRNA plasmid library. The constructed sgRNA plasmid library recorvered all the designed sgRNAs (A, C) with the copy number difference less than five fold for at least 97% designed sgRNAs (B, D). (**E**) PCA analysis shows the high reproducibility of the CRISPRpath screen libraries between biological replicates.

Fig. S3

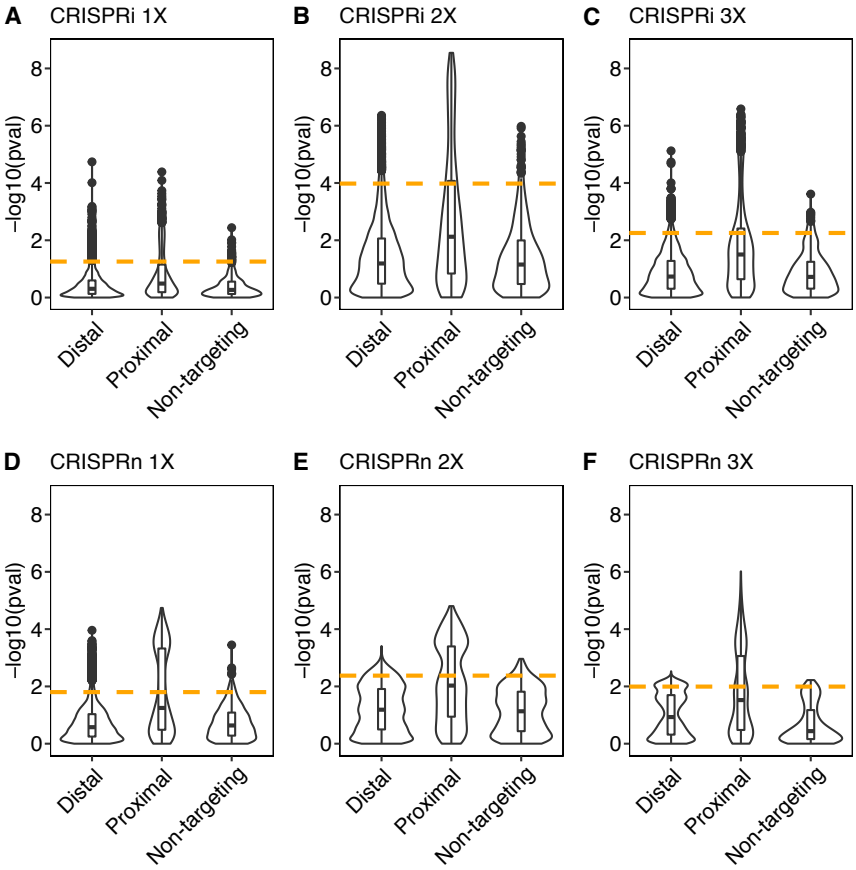

Fig. S3. *P* value cutoff used for identifying enriched sgRNAs from each screens.

(**A-F**) Distribution of *P* value for tested distal, proximal and non-targeting control sgRNA groups. Orange dash lines indicate 5% percentile of the *P* values from non-targeting control sgRNAs to achieve a false discovery rate of 5%.

Fig. S4

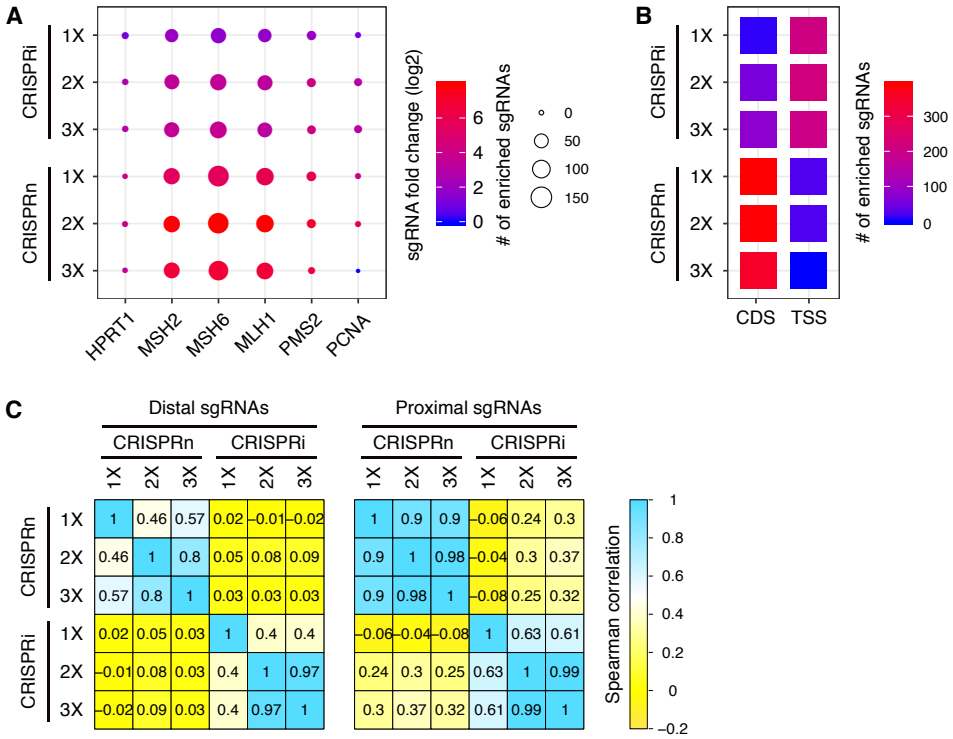

Fig. S4. Enriched proximal sgRNAs and sgRNA ranking analysis.

(**A**) Number and fold changes of the enriched proximal sgRNAs for the six target genes from CRISPRi and CRISPRn screens. The color indicates fold changes, and the size of circle indicates the number of enriched sgRNAs. (**B**) Enrichment analysis shows the enriched proximal sgRNAs bias towards to the TSS region for CRISPRi screens and the protein coding region (CDS) for CRISPRn screens. Color represents the number of enriched sgRNAs. (**C**) Spearman correlation analysis of the distal and proximal sgRNAs ranking shows proximal sgRNAs exhibiting higher correlation between each screen compared to distal sgRNAs.

Fig. S5

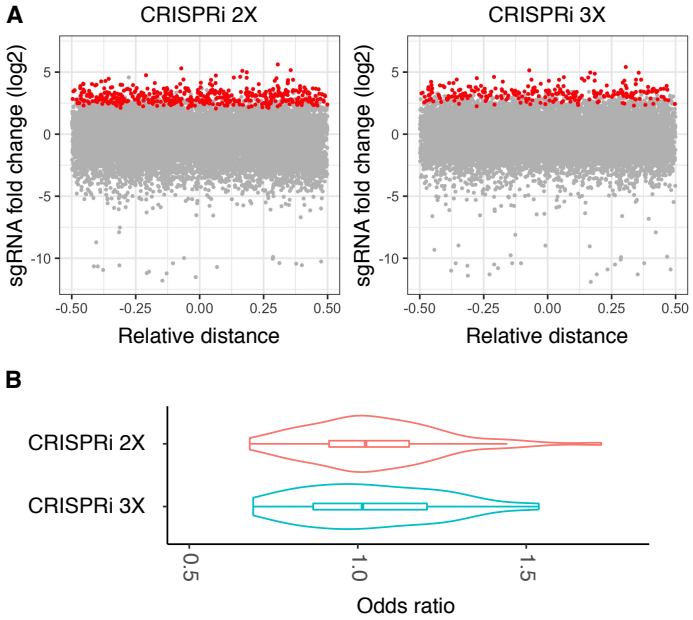

Fig. S5. Enriched sgRNAs identified from CRISPRi screens exhibit no position and strand preference.

(**A**) Enriched sgRNAs from CRISPRi 2X (red dots, n = 448) and CRISPRi 3X (red dots, n = 260) screens showed similar distributions across candidate CREs. (**B**) Odds ratio analysis of the fold changes of enriched sgRNAs shows enriched sgRNAs have no strand preference. Enhancers with enriched sgRNAs only targeting one strand were exclude for the analysis. Odds ratio was calculated for each element with the equation of ave($\log_2$(fold change of sgRNA targeting plus strand)) / ave($\log_2$(fold change of sgRNA targeting minus strand)). Violin plots show the distributions of odds ratio values within each screen, and boxplots indicate the median, IQR, Q1 − 1.5 × IQR and Q3 + 1.5 × IQR.

Fig. S6

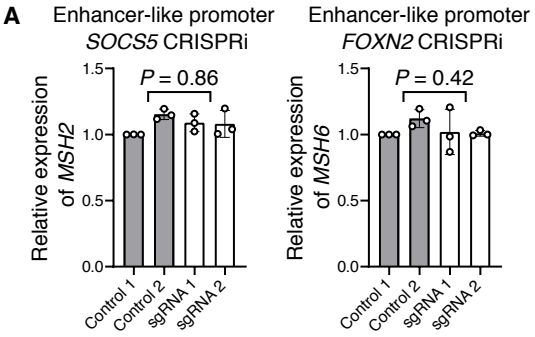

Fig. S6. *SOCS5* promoter and *FOXN2* promoter do not regulate *MSH2* and *MSH6* expression.
(**A**) CRISPRi silencing of *SOCS5* and *FOXN2* promoters do not change *MSH2* and *MSH6* expression, respectively. Three independent replicates per condition and two independent sgRNAs per replicate were used for each experiment. *P* values are from two-tailed two-sample *t*-test. Error bars represent the s.d.

Fig. S7

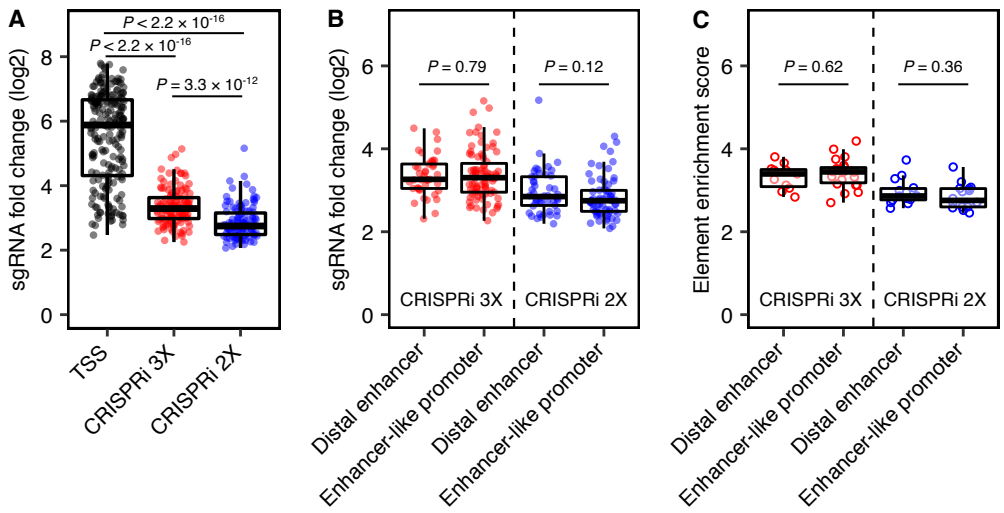

Fig. S7. sgRNA fold changes and element enrichment scores comparison for the identified enhancers.

(**A**) Box plots show the sgRNA fold changes of the enriched sgRNAs for the tested elements. The enriched sgRNAs in TSS regions of MMR genes show highest fold changes. Enriched sgRNAs of the enhancers uniquely identified from the lower selection pressure (CRISPRi 2X) exhibit lower fold changes compared to enriched sgRNAs of the enhancers identified from the higher selection pressure (CRISPRi 3X). Boxplots indicate the median, IQR, Q1 − 1.5 × IQR and Q3 + 1.5 × IQR. The dots represent individual sgRNA. *P* values are from Wilcoxon test. (**B**) Box plots show the sgRNA fold changes of the enriched sgRNAs for different type enhancers. For the enhancers identified from CRISPRi 3X screen and the enhancers uniquely identified from CRISPRi 2X screen, sgRNA fold changes of the enriched sgRNAs did not show significant difference between distal enhancers and enhancer-like promoters. Boxplots indicate the median, IQR, Q1 − 1.5 × IQR and Q3 + 1.5 × IQR. The dots represent individual sgRNA. *P* values are from Wilcoxon test. (**C**) Box plots show the element enrichment scores for enhancers. For the enhancers identified from CRISPRi 3X screen and the enhancers uniquely identified from CRISPRi 2X screen, element enrichment score did not show significant difference between distal enhancers and enhancer-like promoters. Boxplots indicate the median, IQR, Q1 − 1.5 × IQR and Q3 + 1.5 × IQR. The dots represent individual sgRNA. *P* values are from Wilcoxon test.

Fig. S8

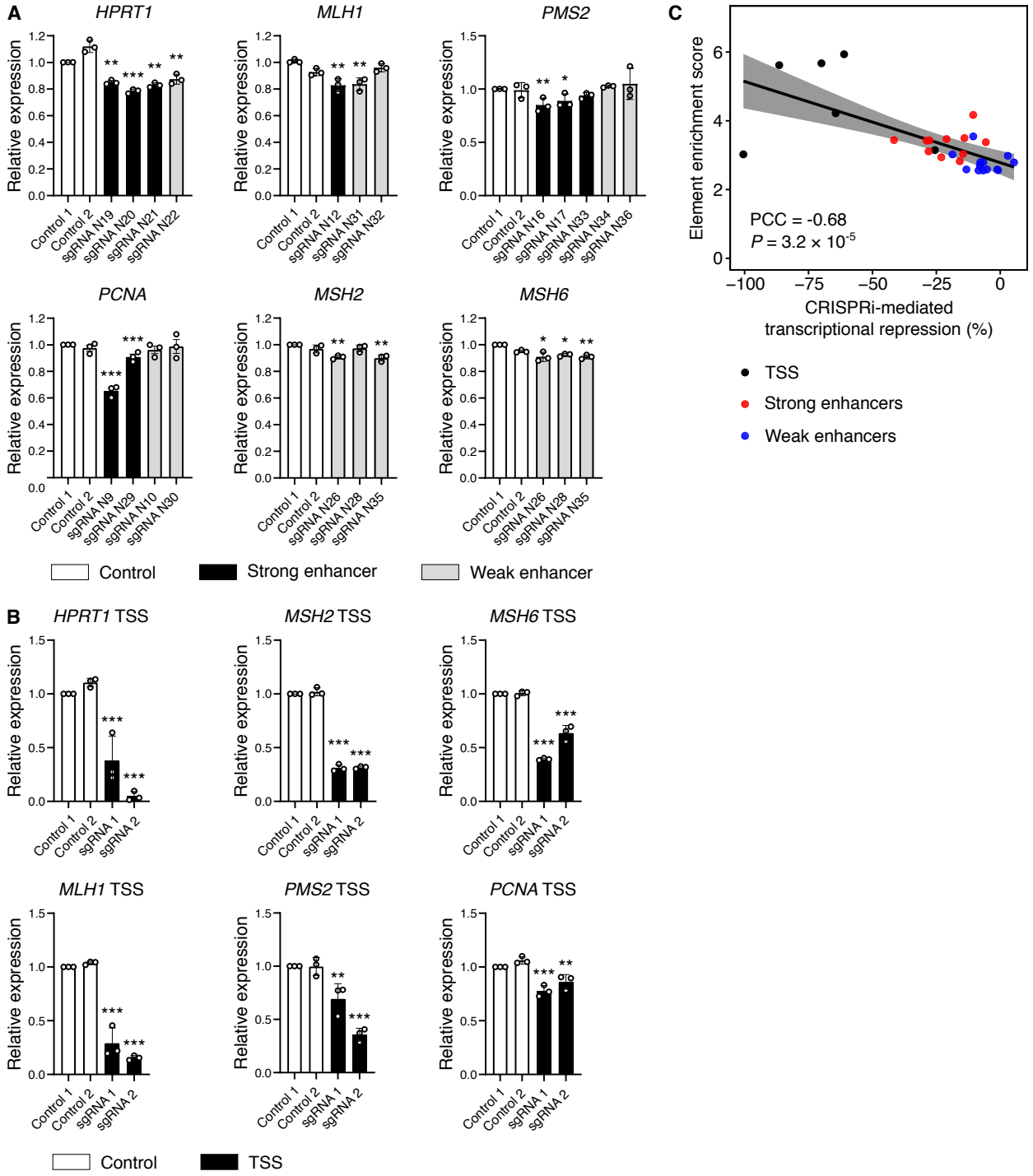

Fig. S8. Validation of CIRSPRpath identified enhancers.

(**A**) Validation of the strong (black) and weak (grey) enhancers with CRISPRi followed by RT-qPCR. Three independent replicates per condition. The significance was calculated with two-tailed two-sample *t*-test. Data are mean and s.d. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. (**B**) CRISPRi-mediated transcriptional repression of six target genes by targeting TSS of each gene. Three independent replicates per condition. The significance was calculated with two-tailed two-sample *t*-test. Data are mean and s.d. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. (**C**) Pearson correlation analysis reveals element enrichment scores from CRISPRpath screens correlate with element effect sizes on transcription from CRISPRi (Pearson correlation, PCC = -0.68, $P = 3.2 \times 10^{-5}$).

Fig. S9

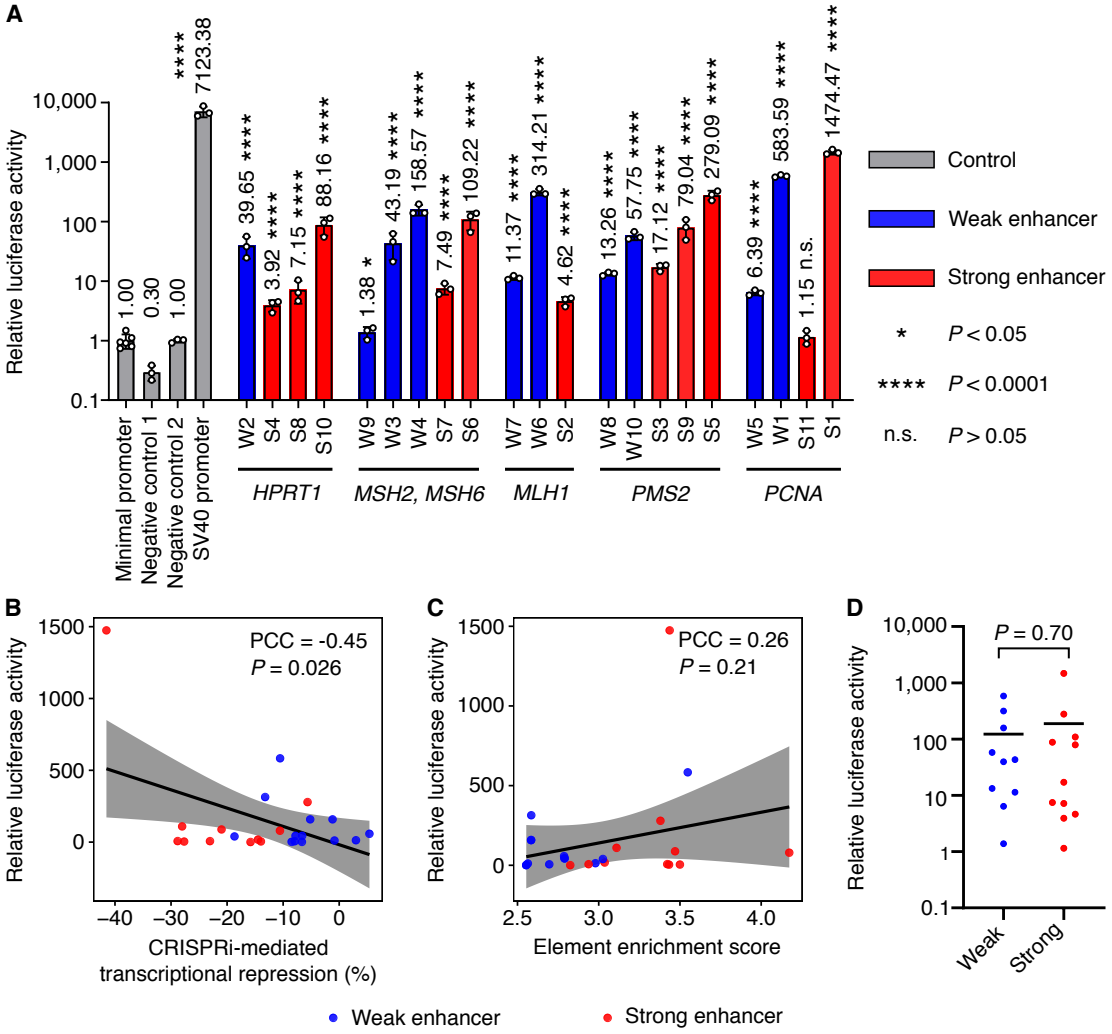

Fig. S9. Characterization of enhancers with dual-luciferase reporter assay.

(**A**) Relative luciferase reporter activities for the weak and strong enhancers. The relative luciferase activities were normalized to the average of the luciferase activity of minimal promoter. The significance was compared to minimal promoter and two negative control elements, and calculated with two-tailed two sample *t*-test. Data are mean and s.d. (**B**) Pearson correlation analysis reveals weak correlation between relative luciferase activities and element effect sizes on transcription from CRISPRi (Pearson correlation, PCC = -0.45, P = 0.026). (**C**) Pearson correlation analysis reveals weak correlation between relative luciferase activities and element enrichment scores from CRISPRpath screens (Pearson correlation, PCC = 0.26, P = 0.21). (**D**) Dotplot shows the distribution of relative luciferase activities of weak and strong enhancers. Each dot represents one tested enhancer. The black line labels the average of relative luciferase activity. The significance was calculated with two-tailed two sample *t*-test.

Fig. S10

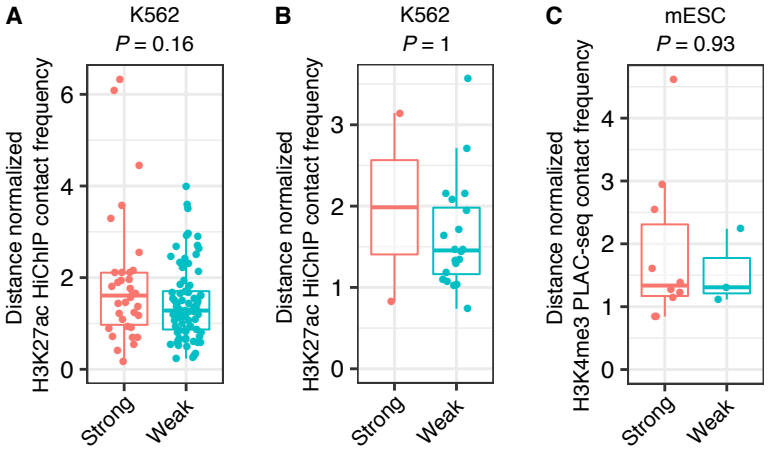

Fig. S10. Chromatin contract frequency analysis for the enhancers in K562 cells and mESCs. (**A**) Distance normalized H3K27ac HiChIP contact frequency for strong (n = 34) and weak (n = 82) enhancers identified with crisprQTL mapping in K562 cells. (**B**) Distance normalized H3K27ac HiChIP contact frequency for strong (n = 2) and weak (n = 20) enhancers identified with CRISPRi-FlowFISH screen in K562 cells. (**C**) Distance normalized H3K4me3 PLAC-seq contact frequency for strong (n = 10) and weak (n = 3) enhancers identified in mouse embryonic stem cells. Boxplots indicate the median, IQR, Q1 − 1.5 × IQR and Q3 + 1.5 × IQR. *P* values are from Wilcoxon test.