Title

A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells.

Journal

Nature methods, 14(6)

ISSN

1548-7091

Authors

Diao, Yarui
Fang, Rongxin
Li, Bin
et al.

Publication Date

2017-06-01

DOI

10.1038/nmeth.4264

Peer reviewed

A tiling-deletion-based genetic screen for *cis*-regulatory element identification in mammalian cells

Yarui Diao[1,10], Rongxin Fang[1,2,10], Bin Li[1,10], Zhipeng Meng[3,4], Juntao Yu[1,5], Yunjiang Qiu[1,2], Kimberly C Lin[3,4], Hui Huang[1,6], Tristin Liu[1], Ryan J Marina[6], Inkyung Jung[7], Yin Shen[8], Kun-Liang Guan[3,4] & Bing Ren[1,4,9]

Millions of *cis*-regulatory elements are predicted to be present in the human genome, but direct evidence for their biological function is scarce. Here we report a high-throughput method, *cis*-regulatory element scan by tiling-deletion and sequencing (CREST-seq), for the unbiased discovery and functional assessment of *cis*-regulatory sequences in the genome. We used it to interrogate the 2-Mb *POU5F1* locus in human embryonic stem cells, and identified 45 *cis*-regulatory elements. A majority of these elements have active chromatin marks, DNase hypersensitivity, and occupancy by multiple transcription factors, which confirms the utility of chromatin signatures in *cis*-element mapping. Notably, 17 of them are previously annotated promoters of functionally unrelated genes, and like typical enhancers, they form extensive spatial contacts with the *POU5F1* promoter. These results point to the commonality of enhancer-like promoters in the human genome.

Millions of candidate *cis*-regulatory elements have been annotated in the human genome on the basis of histone modification, transcription factor (TF) binding, and DNase I hypersensitivity[1–6]. These putative regulatory sequences harbor a disproportionately large number of sequence variants that are associated with diverse human traits and diseases, supporting the hypothesis that noncoding sequence variants contribute to common traits and diseases by disrupting transcriptional regulation[7–9]. However, research on the role of these putative functional elements in human development and disease has been hindered by a dearth of direct evidence for their biological function in the native genomic context.

High-throughput CRISPR–Cas9-mediated mutagenesis by single guide RNAs (sgRNAs) has been used to functionally characterize *cis*-regulatory elements in mammalian cells[10–15]. However, current approaches are limited because (1) not all sequences are suitable for CRISPR–Cas9-mediated genome editing, owing to the lack of protospacer-adjacent motifs (PAMs), which are required for targeting and DNA cutting by CRISPR–Cas9 (refs. 16–18); (2) CRISPR–Cas9-mediated genome editing with individual sgRNAs tends to cause point mutations or short insertions or deletions, thus necessitating the use of an unrealistically large number of sgRNAs to interrogate the human genome; and (3) it has been challenging to distinguish *cis*- and *trans*-regulatory elements. To overcome these limitations, we developed CREST-seq, which allows the efficient discovery and functional characterization of *cis*-regulatory elements through the introduction of massively parallel kilobase-long deletions in the genome. Here we provide evidence in support of the utility of CREST-seq for the large-scale identification of *cis*-regulatory elements in human embryonic stem cells (hESCs). We report the discovery of 45 regulatory sequences of *POU5F1*, and a surprisingly large number of enhancer-like promoters.

RESULTS

CREST-seq identified *cis*-regulatory elements of *POU5F1*

In a CREST-seq experiment, a large number of overlapping genomic deletions are first introduced to a genomic locus by CRISPR–Cas9-mediated genome editing with paired sgRNAs[16] (**Fig. 1a**). Cells with reduced expression of the gene of interest (**Fig. 1b**) are then isolated, and the enriched sgRNA pairs are determined by high-throughput sequencing. From the enriched sgRNA-pair sequences, one can infer the functional *cis*-regulatory sequences of the gene of interest (**Fig. 1a**). We applied CREST-seq to the 2-Mb *POU5F1* locus in an hESC line in which one *POU5F1* allele was genetically tagged with eGFP, which allowed us to monitor the transcription level of this allele on the basis of eGFP expression[19] (**Fig. 1b**).

We designed a total of 11,570 sgRNA pairs (**Supplementary Table 1**) to introduce the same number of genomic deletions (**Fig. 1a** and **Supplementary Fig. 1a**) to the *POU5F1* locus. The average size of each deletion was ~2 kb, with an overlap of 1.9 kb between two adjacent deletions (**Supplementary Fig. 1b**) such

[1]Ludwig Institute for Cancer Research, La Jolla, California, USA. [2]Bioinformatics and Systems Biology Graduate Program, University of California, San Diego, La Jolla, California, USA. [3]Department of Pharmacology, University of California, San Diego, La Jolla, California, USA. [4]Moores Cancer Center, University of California, San Diego, La Jolla, California, USA. [5]School of Life Sciences, University of Science and Technology of China, Hefei, China. [6]Biomedical Sciences Graduate Program, University of California, San Diego, La Jolla, California, USA. [7]Biological Science, KAIST, Daejeon, South Korea. [8]Institute for Human Genetics and Department of Neurology, University of California, San Francisco, San Francisco, California, USA. [9]Department of Cellular and Molecular Medicine, Institute of Genomic Medicine, University of California, San Diego, La Jolla, California, USA. [10]These authors contributed equally to this work. Correspondence should be addressed to B.R. (biren@ucsd.edu).

RECEIVED 13 DECEMBER 2016; ACCEPTED 15 MARCH 2017; PUBLISHED ONLINE 17 APRIL 2017; DOI:10.1038/NMETH.4264

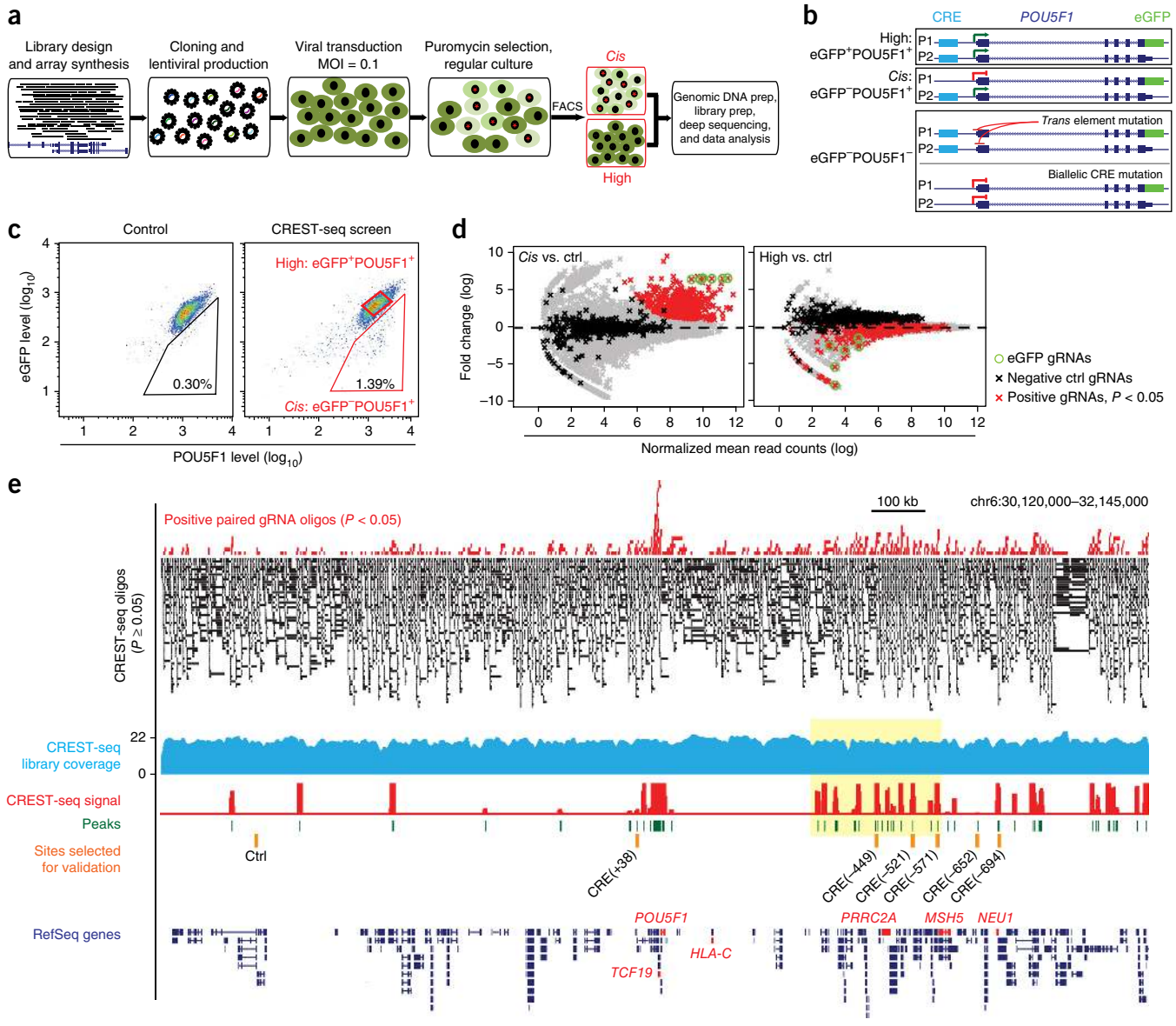

Figure 1 | CREST-seq experimental design and application to the *POU5F1* locus in hESCs. (**a**) CREST-seq workflow. MOI, multiplicity of infection. (**b**) Schematic of monoallelic or biallelic deletion of *cis*-regulatory elements of *POU5F1*. The eGFP-tagging allele was designated as P1, and the wild-type allele as P2. (**c**) FACS analysis of H1 POU5F1–eGFP cells transduced with control lentivirus expressing Cas9 but not sgRNA (left) or with the CREST-seq lentiviral library (right) 14 d after transduction. (**d**) The read counts of sgRNA pairs from the *cis* (left) and high (right) cell populations compared with those from a non-sorted control population (ctrl). Fold changes represent the ratios between read counts in the *cis* or high population and the control population. The significance of enrichment was calculated by negative binomial test. Gray crosses denote sgRNA pairs that were not significantly enriched. (**e**) Genome browser screenshot showing CREST-seq positive sgRNA pairs ($P < 0.05$; top) and CREST-seq negative sgRNA pairs ($P \geq 0.05$; black bars), genomic coverage of the CREST-seq library (blue track), the computed CREST-seq signals (red bars) (Online Methods), the genomic regions identified as *cis*-regulatory sequences of *POU5F1* (green bars) and the CRE sites selected for further in-depth validation (orange bars). The yellow shaded region highlights a region enriched for CREs; a close-up view is shown in **Figure 2b**. Data in **c**–**e** are representative of five independent experiments.

that each nucleotide in the locus was covered by ~20 distinct genomic deletions on average. As negative controls, we included 424 sgRNA oligos that lacked the PAM sequence necessary for effective double-stranded DNA breaks. As positive controls, we included six sgRNA pairs that target the eGFP gene coding sequence (**Supplementary Table 1**). We constructed a lentiviral library that expressed these sgRNA pairs (**Supplementary Fig. 2**) and transduced it into the hESC line at a low multiplicity of infection (0.1), which ensured that the majority of cells received one or no lentiviral particle (detailed in the **Supplementary Protocol**).

To isolate mutant cells with deletions in *POU5F1*'s *cis*-regulatory sequences, we used fluorescence-activated cell sorting (FACS) to sort out cells that showed reduced *POU5F1* expression from the eGFP-tagged allele but relatively unchanged expression from the non-tagged allele (**Fig. 1c**). We refer to this eGFP$^-$POU5F1$^+$ subpopulation as the '*cis*' population (**Fig. 1b**,**c**). As a control, we also collected a sample of cells before FACS. Finally, we collected the eGFP$^+$POU5F1$^+$ ('high') population (**Fig. 1b**,**c** and **Supplementary Notes 1**–**4**). We purified genomic DNA from each cell population, and then used massively parallel sequencing

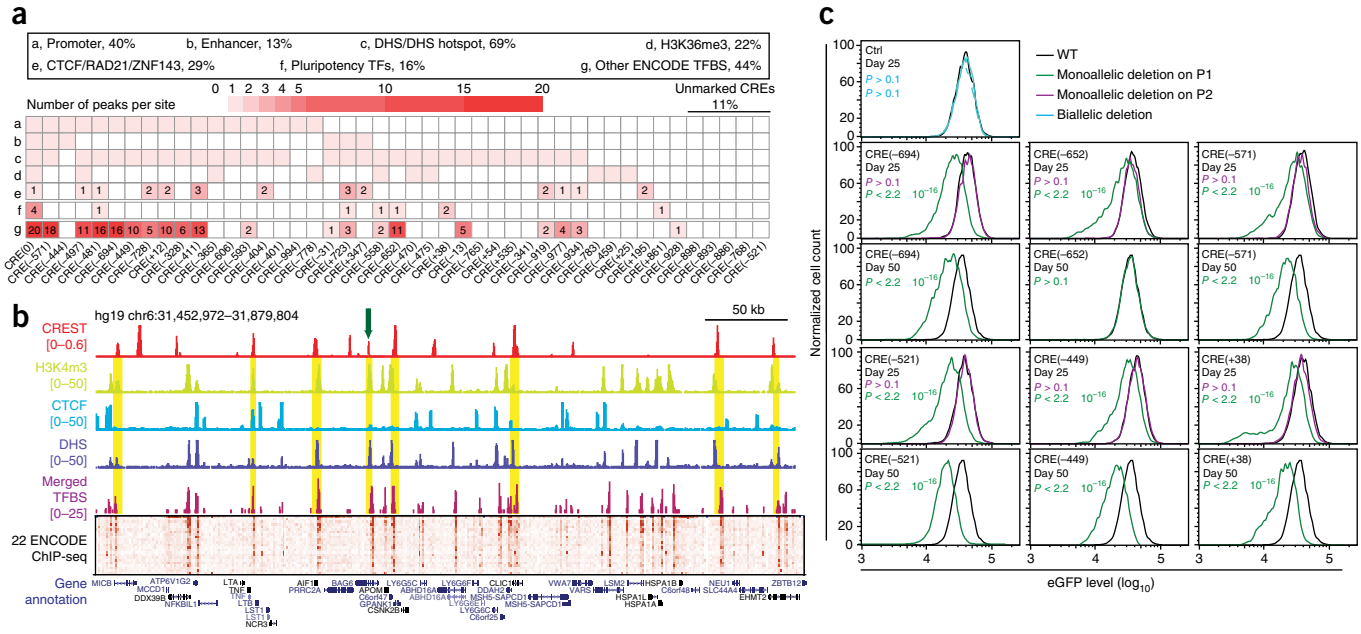

Figure 2 | CREs tend to be associated with canonical active chromatin markers of *cis*-regulatory elements and dense TF clusters. (**a**) The chromatin features and TF binding sites (TFBS) at the 45 CREs. "Pluripotency TFs" includes POU5F1, SOX2, NANOG and PRDM14 (see **Supplementary Table 5** for detailed features). (**b**) A close-up view of the region highlighted in yellow in **Figure 1e**, with tracks corresponding to the indicated chromatin modifications. The height of the merged TF-binding site bars indicates the number of bound TFs. Yellow bars highlight regions where CREs overlap with active chromatin marks and TF-binding site clusters. The green arrow at the top points to the CREs shown in **Supplementary Figure 6a**. (**c**) Six CREs and one CREST-seq negative site (control) were selected (orange bars in **Fig. 1e**) for individual validation. We generated mutant clones harboring biallelic deletion (ctrl), monoallelic deletion on the P1 allele (eGFP-containing allele), or monoallelic deletion on the P2 allele (non-eGFP allele) at the indicated genomic loci. We carried out FACS analysis of all the mutant clones and wild-type (WT) cells at day 25 and day 50 after CRISPR–Cas9 transfection. We quantified the FACS data with FlowJo, and calculated *P* values by two-sample *t*-test. *P* values are color-coded according to the key to correspond to the P1 and P2 deletion mutants. Data in **a** and **b** are from five independent experiments; data in **c** are representative of 27 independent CRE clones (listed in **Supplementary Table 6**).

to determine which sgRNA pairs were present in each subpopulation (**Supplementary Table 2**). We carried out the experiment in multiple replicates (**Supplementary Table 2** and **Supplementary Fig. 3a**), and observed that the abundance of sgRNA pairs was highly reproducible between replicates (Pearson correlation coefficient $R = 0.90$ for *cis*, 0.92 for control and 0.97 for high; **Supplementary Fig. 3b**).

To identify *cis*-regulatory elements of *POU5F1*, we first compared the abundance of sgRNA pairs between the *cis* population and the control population (**Supplementary Table 2**) by using a negative binomial test, and computed the fold enrichment and *P* value of each sgRNA pair (**Supplementary Table 3** and **Supplementary Fig. 3c**). We found 495 sgRNA pairs that were significantly enriched ($P < 0.05$ and log(fold change) > 1) in the *cis* samples (**Fig. 1d**,**e** and **Supplementary Table 3**). As expected, all six sgRNA pairs that targeted the eGFP gene sequence were strongly enriched in the *cis* population (**Fig. 1d**). By contrast, only 2 of the 424 negative control sgRNAs were enriched, corresponding to an empirical false discovery rate (FDR) of <0.5%. Further supporting the effectiveness of our experimental design, the sgRNA pairs with significant enrichment in the *cis* population were generally depleted in the high-population samples (**Supplementary Table 3** and **Fig. 1d**). Next, we sought to identify *cis*-regulatory sequences by taking full advantage of the tiling-deletion design (**Fig. 1e**). We began by ranking all sgRNA pairs on the basis of their enrichment levels in the *cis* population relative to the control (**Supplementary Table 3**). We then partitioned the 2-Mb *POU5F1* locus into 50-bp bins, and used robust rank aggregation[20] to calculate a score for each bin to indicate whether the ranks of deletions spanning that bin were skewed toward the top of the sorted list (Online Methods and **Supplementary Table 4**). In total, we identified 45 genomic regions with a significant score (**Fig. 1e** and **Supplementary Table 5**). Using the same criteria, we did not identify any genomic region as positive in the high population (**Supplementary Fig. 4a**). We named each of the 45 CREST-positive elements (referred to hereinafter as CREs) according to its relative genomic distance (in kilobases) from the transcription start site (TSS) of *POU5F1*, with a negative sign used to denote elements upstream of *POU5F1*, and a positive sign used to indicate downstream elements (**Supplementary Table 5**). The 45 CREs included 4 previously identified *POU5F1*-regulatory elements that act in *cis*: its promoter (**Supplementary Fig. 4b**), an upstream enhancer[21] (**Supplementary Fig. 4b**) and two temporarily phenotypic enhancers[13] (DHS_65 and DHS_108; **Supplementary Fig. 4c**). The remaining 41 CREs were *POU5F1*-regulatory sequences newly discovered in this study (**Supplementary Note 5**).

CREs cluster with active chromatin marks and TFs

To determine the chromatin features of the CREs, we examined the publically available chromatin-accessibility data, TF-binding profiles and chromatin-modification data sets for the H1 hESC line[3,5]. We also carried out assays for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq)[22]

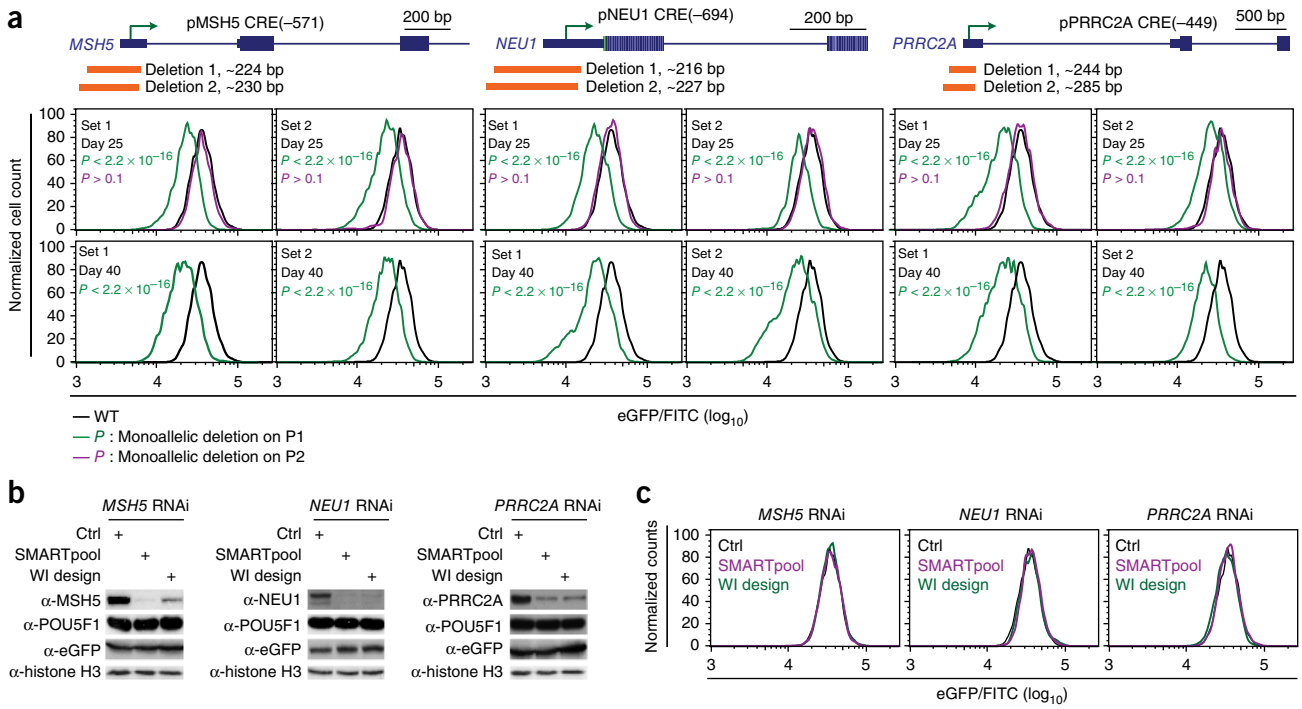

Figure 3 | The core promoter regions of *MSH5*, *NEU1* and *PRRC2A* are required for optimal *POU5F1* expression in hESCs. (**a**) The core promoter regions of *MSH5*, *NEU1* and *PRRC2A* were deleted by two sets of distinct sgRNAs (deletions 1 and 2). Mutant cell clones harboring monoallelic deletions on the P1 allele (green curves) or P2 allele (magenta curves) were identified after genotyping and sequencing of the phased single-nucleotide polymorphisms. FACS analysis was performed for all the mutant clones and wild-type cells (WT; black curves) at day 25 and day 40 after transfection. The FACS data were quantified with FlowJo, and represent 40 independent core promoter mutant clones (listed in **Supplementary Table 6**). We computed *P* values by two-sample *t*-test. (**b**,**c**) H1 POU5F1–eGFP cells were transfected with either control scrambled siRNA or siRNAs targeting the indicated genes. Each gene was targeted by two sets of siRNAs (SMARTpool and WI design) with different sequences. The cells were analyzed 48 h after transfection. (**b**) Whole-cell extract was collected and subjected to western blotting analysis with the indicated antibodies. (**c**) An aliquot of cells was dissociated into single cells for FACS analysis. Black, magenta and green curves represent the data from cells treated with scrambled siRNA (ctrl), SMARTpool siRNA and WI-designed (http://sirna.wi.mit.edu/) siRNA, respectively. Data in **b** and **c** are representative of three independent experiments.

and CTCF chromatin immunoprecipitation followed by sequencing (ChIP-seq) with the cell line used in the present study, and ensured that the data closely resembled previously obtained data sets for the same parental cell line[5] (**Supplementary Fig. 5**). As expected, a majority of CREs were associated with biochemical features characteristic of *cis*-regulatory elements, including DNase hypersensitivity (69%); TF occupancy; and active chromatin marks such as acetylation of histone H3 on Lys27 (H3K27ac; 22%), methylation of histone H3 on Lys4 (H3K4me3; 31%) and H3K4me1 (22%) (**Supplementary Table 5**)[5]. Notably, CREs were also enriched for binding sites of CTCF/RAD21 (29%), which have been linked to DNA looping and topologically associating domain boundaries[23,24] (**Fig. 2a**,**b** and **Supplementary Table 5**). It has been reported that TF binding in human cells tends to lead to the formation of dense clusters[25–27]. Accordingly, we found that the CREST-positive regions overlapped with dense clusters of TF-binding sites (16% of CREs were bound by essential pluripotency master regulators, and 44% by other TFs; **Fig. 2a**,**b** and **Supplementary Fig. 6a**) and were bound by more TFs on average than DNase hypersensitive sites (DHSs) (**Supplementary Fig. 6b**; $P < 6 \times 10^{-11}$). In general, CREST-positive regions were significantly associated with TF binding and the active histone modifications H3K4me1, H3K4me3 and H3K27ac, and were depleted for the repressive chromatin marks H3K9me3 and H3K27me3 (ref. 28) (**Supplementary Fig. 6c**; $P < 0.01$; other features are described in **Supplementary Fig. 6d**), consistent with the findings of previous studies highlighting the role of clustered TF-binding sites in gene regulation[25,29]. Interestingly, five CREs lacked any canonical chromatin signatures associated with active *cis*-regulatory sequences (**Fig. 2a**; unmarked region, 11%), which suggests the existence of *cis*-regulatory elements without canonical epigenetic signatures, as recently reported[12].

To validate the function of the novel *POU5F1* CREs, we selected six for in-depth analysis (**Fig. 1e**). We chose regions that met three criteria: (1) they were located at a wide range of genomic distances (38–694 kb) from the *POU5F1* TSS; (2) they were surrounded by phased single-nucleotide polymorphisms so that allelic analysis of gene expression could be performed; and (3) they represented a wide range of CREST-seq signals, ranking 9th, 13th, 23rd, 24th and 37th out of 45 (**Supplementary Table 5**). Additionally, whereas five CREs—CRE(–694), CRE(–652), CRE(–571), CRE(–449) and CRE(+38)—were marked by canonical chromatin marks (**Fig. 2a** and **Supplementary Fig. 7a**), CRE(–521) was unmarked (**Fig. 2a** and **Supplementary Fig. 7a**). As a control, we tested a CREST-negative region (**Fig. 1e** and **Supplementary Fig. 7a**). We applied CRISPR–Cas9 genome editing in the hESC line to introduce monoallelic deletions 2–4 kb in length to remove these regions (**Supplementary Fig. 7a**). All cell clones with monoallelic deletion on the P1 (eGFP-tagged) allele

showed a significant reduction in eGFP gene expression (**Fig. 2c** and **Supplementary Fig. 7b**; $P < 2.2 \times 10^{-16}$). By contrast, clones bearing monoallelic deletions on the P2 (wild-type) allele showed normal eGFP gene expression (**Fig. 2c**), which indicates that these sequences act in *cis* to regulate *POU5F1* expression. We did not observe any change in eGFP gene expression in clones that contained biallelic deletions of the negative control region (**Fig. 2c**). Notably, deletion of CRE(−521), which lacked any canonical marks of regulatory sequences (**Supplementary Fig. 7a**), also led to a decrease in *POU5F1* expression in the *cis* population. Interestingly, whereas the deletion of each of the five CREs resulted in a durable reduction of *POU5F1* expression, deletion of CRE(−652) led to only a temporary reduction of eGFP gene expression that recovered fully by day 50 (**Fig. 2c** and **Supplementary Fig. 7b**), which suggests that it is the type of temporarily phenotypic enhancer that we recently reported[13]. Taken together, these results provided strong evidence that CREST-seq can be used to identify *cis*-regulatory sequences of a specific target gene in an unbiased and high-throughput manner.

Promoters acting as distal enhancers

Results from the CREST-seq experiments showed that 18 gene promoters, including the *POU5F1* promoter, are necessary for optimal *POU5F1* expression in hESCs (**Supplementary Table 5**). This is surprising because promoters have traditionally been thought to mediate the transcription of their target genes' immediate downstream sequences. Although recent reports indicate that some long noncoding RNA and mRNA promoters may act as enhancers of their adjacent genes[12,30,31], definitive evidence illustrating a causative role of promoters acting as distal enhancers is still lacking. The identification of CRE(−449), CRE(−571) and CRE(−694) as *cis*-regulatory elements of *POU5F1* suggests that promoters of *PRRC2A*, *MSH5* and *NEU1* may act as distal enhancers of *POU5F1* in hESCs (**Supplementary Fig. 7a**). To rule out the possibility that promoter-proximal elements in these genes were responsible for *POU5F1* regulation, we deleted 216–285-bp core promoter sequences containing the TSS of each gene and carried out allelic expression analysis in the resulting cell clones (**Fig. 3a** and **Supplementary Fig. 8**). To avoid potential off-target effects, we used two sets of sgRNA pairs (deletion 1 and deletion 2; **Fig. 3a** and **Supplementary Fig. 8**) for the genome editing, and we recovered a total of 37 independent clones carrying monoallelic deletions for in-depth analysis (**Supplementary Fig. 9** and **Supplementary Table 6**). We found that all mutants with the P1 monoallelic deletion showed long-lasting reductions in eGFP gene expression, whereas in mutant clones with the P2 monoallelic deletion, eGFP levels were indistinguishable from those in the wild type (**Fig. 3a** and **Supplementary Fig. 9a**,**b**; quantified in **Supplementary Table 6** and **Supplementary Fig. 9c**). The reduced eGFP gene expression could not be due to the loss of the *PRRC2A*, *MSH5* or *NEU1* gene products, because knockdown of each gene using two sets of short interfering RNAs (siRNAs) (**Fig. 3b**,**c**) and short hairpin RNAs (**Supplementary Fig. 10a**–**c**) did not affect levels of *POU5F1* mRNA (**Supplementary Fig. 10d**) or POU5F1 protein (**Fig. 3b** and **Supplementary Fig. 10e**). Thus, the core promoter sequences of *PRRC2A*, *MSH5* and *NEU1*, but not their gene products, are required for optimal *POU5F1* expression.

To further show whether these gene promoters could function as enhancers in a traditional reporter assay, we constructed

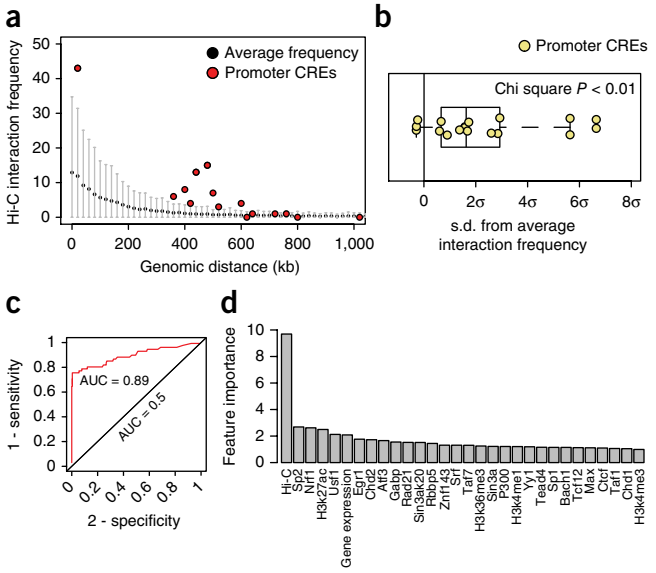

Figure 4 | Analysis of chromatin interactions between enhancer-like promoters and the *POU5F1* promoter in hESCs. (**a**) The distribution of pairwise Hi-C contact frequencies within the 2-Mb locus, and between the *POU5F1* TSS and the 17 *POU5F1*-regulating promoters. The gray bars represent the s.d. of normalized Hi-C read counts at a given genomic distance. (**b**) The s.d. of the Hi-C read counts between the *POU5F1* TSS and the promoter CREs compared with the expected value (0). (**c**) Receiver operating characteristic curve showing that *POU5F1*-regulating promoters could be separated from the other promoters in the 2-Mb region with high accuracy (AUC = 0.89) with a random forest model built from binding sites of 52 TFs, seven histone-modification profiles, gene expression profiles, and maps of long-range chromatin interactions (**Supplementary Table 7**; additional details are provided in the Online Methods). (**d**) The relative importance of each feature to the random forest classifier for predicting enhancer-like promoters.

reporter plasmids that contained the 360-bp *POU5F1* core promoter sequence driving a luciferase reporter gene, with the core promoter fragments of *PRRC2A*, *MSH5* or *NEU1* inserted downstream of the reporter[13,32]. We transfected these plasmids into H1 hESCs and assayed for luciferase activity 3 d after transfection. All elements showed significant enhancer activity compared with the control vector (**Supplementary Fig. 10f**).

To rule out the possibility that CRISPR–Cas9-mediated genome editing affects *POU5F1* expression through locus-wide, nonspecific mechanisms, we carried out FACS analysis of the CRE-deletion mutant clones to monitor levels of both POU5F1–eGFP and HLA-C, located 100 kb upstream of the *POU5F1* TSS. We found that deletion of a CRE resulted in downregulation of POU5F1–eGFP expression without any observable effect on levels of HLA-C (**Supplementary Fig. 11**). To further rule out the possibility that CRISPR–Cas9 leads to transcriptional silencing induced by double-stranded DNA breaks in cells, we assessed the presence of phosphorylated H2AX (γH2AX; a DNA-damage marker) in the mutant clones[33–35]. We found that none of the mutant clones stained positive for γH2AX at the time of the experiments when downregulation of *POU5F1* was detected (25 d after transfection) (**Supplementary Fig. 11a**). Therefore, it is not likely that our CREST-seq identification of multiple promoters serving as distal enhancers of *POU5F1* was due to artifacts of the experimental system.

Table 1 | Comparison of CREST-seq data to published functional screens of noncoding regulatory sequences

Reference	Target region	Total oligos	Oligo density (per kb)	Coverage	Able to distinguish *trans* or *cis*?
Canver *et al.*[10]	4.2 kb, 3 DHSs and 1 exon	582	137	~1×	No
Korkmaz *et al.*[11]	685 p53 ChIP-seq peaks	1,116	N.A.	1.3–1.6 oligos per ChIP-seq peak	No
	73 ChIP-seq peaks for ER-α expressing enhancer RNA	97	N.A.		
	2-kb *CDKN1A* locus	197	98.5	<93.6%	
Rajagopal *et al.*[12]	40-kb *Tdgf1* locus	3,908	98	<93.1%	No
	Rpp25, *Nanog* and *Zfp42* loci	3,908	N.A.	N.A.	
Diao *et al.*[13]	37.6 kb, 174 putative enhancers in 1-Mb *POU5F1* locus	1,964	52	<49.4%	No
Sanjana *et al.*[14]	200-kb *NF1* locus	6,682	33.4	<31.7%	No
	200-kb *NF2* locus	6,934	34.6	<32.9%	
	200-kb *CUL3* locus	4,699	23.5	<22.3%	
Fulco *et al.*[15]	1.29-Mb *GATA1* and *MYC* loci	98,000	76	~64×	No
CREST-seq	2-Mb *POU5F1* locus	11,600	5.7	20×	Yes

Here CREST-seq is compared with published screens of noncoding regulatory elements. The following aspects are compared: the size of the screen region, the total number of oligos required to construct the library, the average number of oligos per kilobase in each screen, and the estimated coverage of the target region. To estimate the coverage of the target region, we assumed that the PAMs were equally distributed across the genome and that each gRNA created a mean insertion/deletion size of 9.5 ± 13.7 bp. To compute the coverage of the CRISPRi screen using dCas9–KRAB, we assumed that the average size of H3K9me3 peaks introduced by dCas9–KRAB was about 850 bp. N.A., not available.

Enhancer-like promoters are spatially close to the *POU5F1* TSS

To understand the potential mechanisms that allow the 17 CREST-positive promoters, among promoters of ~120 genes in this 2-Mb locus, to specifically regulate *POU5F1*, we examined the 3D chromatin organization of the locus, reasoning that long-range chromatin interactions may allow these enhancer-like promoters to act as distal *cis*-regulatory sequences. Indeed, analysis of H1 hESC Hi-C data[36] indicated that 14 of the 17 *POU5F1*-regulating promoters had significantly higher levels of chromatin interactions with the *POU5F1* TSS than would be expected to occur by chance (**Fig. 4a**,**b**; $P < 0.01$). The enhancer-like promoters were also characterized by other chromatin features that distinguished them from other promoters in the region, such as high levels of POL2 binding, H3K4me3 and H3K27ac (**Supplementary Fig. 12a**,**b**; $P < 0.01$). In addition, mRNA transcription from these promoters was significantly higher than that of other genes in the same region (**Supplementary Fig. 12c**; $P < 0.01$).

To further characterize the features of enhancer-like promoters, we developed a random-forest-based classifier capable of predicting which promoters are *cis*-regulatory sequences of *POU5F1*. As input, we used data sets of TF binding sites (**Supplementary Table 7**), histone-modification[5] profiles, gene expression profiles, and the long-range chromatin contacts centered at *POU5F1* (ref. 36). We evaluated the performance of the classifier by using leave-one-out cross-validation. Strikingly, our model was able to distinguish *POU5F1*-regulating promoters from control promoters in the 2-Mb screening region with high accuracy (**Fig. 4c**; area under the curve (AUC), 0.89; error rate, 6.3%; positive predictive value (PPV), 97.2%). We next determined feature importance by estimating the average decrease in node impurity after permuting each predictor variable, and found that the chromatin-interaction frequency was the single most important predictor (**Fig. 4d** and **Supplementary Fig. 13**). This result provides strong evidence that the enhancer-like promoters specifically affect *POU5F1* expression through chromatin interactions. This observation prompted us to use spatial proximity alone to make a single-variable random forest model, which also achieved high-accuracy predictions (AUC, 0.93; error rate, 9.0%) but yielded a lower PPV (74.5%), thus suggesting that although physical proximity is an important predictor for regulatory relationships, other factors are also crucial (**Supplementary Note 6**).

DISCUSSION

Our finding that nearly 40% of the *cis*-regulatory sequences of *POU5F1* correspond to promoters of other genes reveals the commonality and widespread use of promoters as distal enhancers. Previous studies have suggested that promoters and enhancers share common properties in terms of TF binding and the ability to produce RNA transcripts[37]. Recently, it was shown that the promoters of long noncoding RNAs and mRNAs can act as enhancers of adjacent genes[12,31,38]. The current study adds to the accumulating literature on the idea that distal promoters can regulate the expression of a gene other than the gene immediately downstream. Our results further show that one potential mechanism by which promoters could act as enhancers is long-range chromatin interaction (**Supplementary Note 7**). This is consistent with previous studies that show extensive promoter–promoter interactions in mammalian cells[30,36,39–46], and reports that many promoters indeed show enhancer activity in heterologous ectopic luciferase reporter assays[30,47].

CREST-seq is a highly scalable tool for the unbiased discovery of *cis*-regulatory sequences in the human genome. Compared with previous CRISPR–Cas9 screens, which have typically required more than 100 gRNA-expressing oligos to 'saturate' a target region, CREST-seq achieved 20× coverage for the entire 2-Mb *POU5F1* locus, with fewer than six sgRNAs per kilobase (**Table 1**). CREST-seq also compared favorably to the dCas9–KRAB-based CRISPRi (CRISPR interference) screen[15] in which the size of H3K9me3 peaks generated by dCas9–KRAB is less than 850 bp (ref. 48). Although the positive hits identified by CREST-seq are usually larger in size than the elements or motifs identified by single sgRNA approaches, by generating overlapping deletions in a massively parallel fashion, CREST-seq allows the functional interrogation of a large fraction of the genome with high sensitivity and specificity. More important, CREST-seq can distinguish *cis*- and *trans*-regulatory sequences by enabling researchers to monitor the allelic expression of a reporter gene without knowledge of the haplotypes of the genome (**Supplementary Figs. 14** and **15**). Finally, it is feasible to design nested tiling deletions across a whole chromosome or even across the genome. The combination of CREST-seq and single sgRNA screen approaches would allow for both high coverage and high resolution, thereby

enabling truly comprehensive discovery of transcriptional regulatory sequences in the human genome.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank D. Gorkin and J. Yan for feedback on previous versions of the manuscript. We thank Z. Ye and S. Kuan for technical assistance. This work was supported by the US National Institutes of Health (NIH) (grants U54 HG006997, U01 DK105541, R01HG008135, 1UM1HG009402 and 2P50 GM085764 to B.R.), the Ludwig Institute for Cancer Research (to B.R.) and the Human Frontier Science Program (HFSP) (Long Term Postdoctoral Fellowship to Y.D.).

AUTHOR CONTRIBUTIONS

Y.D. and B.R. conceived the idea for CREST-seq; R.F., Y.D. and B.L. conducted integrative data analysis with help from Y.Q., H.H. and I.J.; B.L. and Y.D. designed paired sgRNA libraries; Y.D., Z.M., J.Y., K.C.L., T.L., H.H., R.J.M. and Y.S. performed the experiment; Z.M., K.C.L. and K.-L.G. packaged the lentiviral library; and Y.D., R.F., B.L. and B.R. wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Gerstein, M.B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91–100 (2012).
2. Shen, Y. *et al.* A map of the *cis*-regulatory sequences in the mouse genome. *Nature* **488**, 116–120 (2012).
3. Xie, W. *et al.* Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134–1148 (2013).
4. Roadmap Epigenomics Consortium. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
5. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
6. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
7. Thurman, R.E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
8. Farh, K.K. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
9. Gjoneska, E. *et al.* Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature* **518**, 365–369 (2015).
10. Canver, M.C. *et al.* BCL11A enhancer dissection by Cas9-mediated *in situ* saturating mutagenesis. *Nature* **527**, 192–197 (2015).
11. Korkmaz, G. *et al.* Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat. Biotechnol.* **34**, 192–198 (2016).
12. Rajagopal, N. *et al.* High-throughput mapping of regulatory DNA. *Nat. Biotechnol.* **34**, 167–174 (2016).
13. Diao, Y. *et al.* A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening. *Genome Res.* **26**, 397–405 (2016).
14. Sanjana, N.E. *et al.* High-resolution interrogation of functional elements in the noncoding genome. *Science* **353**, 1545–1549 (2016).
15. Fulco, C.P. *et al.* Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* **354**, 769–773 (2016).
16. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
17. Mojica, F.J., Diez-Villasenor, C., Garcia-Martinez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733–740 (2009).
18. Sternberg, S.H., Redding, S., Jinek, M., Greene, E.C. & Doudna, J.A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62–67 (2014).
19. Zwaka, T.P. & Thomson, J.A. Homologous recombination in human embryonic stem cells. *Nat. Biotechnol.* **21**, 319–321 (2003).
20. Li, W. *et al.* MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol.* **15**, 554 (2014).
21. Ware, C.B. *et al.* Derivation of naive human embryonic stem cells. *Proc. Natl. Acad. Sci. USA* **111**, 4484–4489 (2014).
22. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
23. Ghirlando, R. & Felsenfeld, G. CTCF: making the right connections. *Genes Dev.* **30**, 881–891 (2016).
24. Dixon, J.R., Gorkin, D.U. & Ren, B. Chromatin domains: the unit of chromosome organization. *Mol. Cell* **62**, 668–680 (2016).
25. Yan, J. *et al.* Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell* **154**, 801–813 (2013).
26. MacArthur, S. *et al.* Developmental roles of 21 *Drosophila* transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biol.* **10**, R80 (2009).
27. Yip, K.Y. *et al.* Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.* **13**, R48 (2012).
28. Chandra, T. *et al.* Independence of repressive histone marks and chromatin compaction during senescent heterochromatic layer formation. *Mol. Cell* **47**, 203–214 (2012).
29. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
30. Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84–98 (2012).
31. Engreitz, J.M. *et al.* Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* **539**, 452–455 (2016).
32. Chia, N.Y. *et al.* A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. *Nature* **468**, 316–320 (2010).
33. Rogakou, E.P., Boon, C., Redon, C. & Bonner, W.M. Megabase chromatin domains involved in DNA double-strand breaks in vivo. *J. Cell Biol.* **146**, 905–916 (1999).
34. Downs, J.A., Lowndes, N.F. & Jackson, S.P. A role for *Saccharomyces cerevisiae* histone H2A in DNA repair. *Nature* **408**, 1001–1004 (2000).
35. Burma, S., Chen, B.P., Murphy, M., Kurimasa, A. & Chen, D.J. ATM phosphorylates histone H2AX in response to DNA double-strand breaks. *J. Biol. Chem.* **276**, 42462–42467 (2001).
36. Dixon, J.R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331–336 (2015).
37. Core, L.J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46**, 1311–1320 (2014).
38. Paralkar, V.R. *et al.* Unlinking an lncRNA from its associated cis element. *Mol. Cell* **62**, 104–110 (2016).
39. Handoko, L. *et al.* CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat. Genet.* **43**, 630–638 (2011).
40. DeMare, L.E. *et al.* The genomic landscape of cohesin-associated chromatin interactions. *Genome Res.* **23**, 1224–1234 (2013).
41. Kieffer-Kwon, K.R. *et al.* Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. *Cell* **155**, 1507–1520 (2013).
42. Ji, X. *et al.* 3D chromosome regulatory landscape of human pluripotent cells. *Cell Stem Cell* **18**, 262–275 (2016).
43. Tang, Z. *et al.* CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* **163**, 1611–1627 (2015).
44. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290–294 (2013).
45. Rao, S.S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
46. Dixon, J.R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
47. Arnold, C.D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
48. Thakore, P.I. *et al.* Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nat. Methods* **12**, 1143–1149 (2015).

ONLINE METHODS

Step-by-step protocol. A protocol for CREST-seq is available as a **Supplementary Protocol** and also in ref. 49.

Cell culture. The POU5F1–eGFP H1 hESC line was purchased from WiCell (DL-02) and was described previously[19]. The cells were cultured on Matrigel-coated (Corning; 354277) plates and maintained in TeSR-E8 media (STEMCELL Technologies; 05940), and passaged by Accutase (STEMCELL Technologies; A1517001) with 10-μM ROCK inhibitor Y-27632 (STEMCELL Technologies; 72302) supplement. The cells were tested by WiCell Research Institute and the UCSD Human Stem Cell Core facility to confirm that there was no mycoplasma contamination.

Design of sgRNA pairs for CREST-seq. The CREST-seq library design is available online (http://crest-seq.ucsd.edu/web/) and included the following steps: (1) all 20-bp potential sgRNA sequences followed by the PAM 'NGG' within the 2-Mb screened region were first identified; (2) Bowtie[50] was used to map these 20-bp sgRNA sequences to the reference genome (hg19) with the parameter -t -a -f -m 1000 --tryhard -v 3, which outputs alignments for up to 1,000 candidates with fewer than four mismatches; (3) to prevent off-target binding, an sgRNA sequence was filtered out if it (a) mapped perfectly to another region on the genome, (b) had suboptimal alignment with one or two mismatched bases outside the sgRNA 'seed' region (i.e., the 10-bp sequence adjacent to the PAM)[51] or (c) had suboptimal alignment with three mismatches, but all three mismatched bases were 17 bp farther from the PAM sequence; and (4) the identified sgRNA sites were paired to generate 2-kb deletions evenly across the 2-Mb region. On the basis of the distribution of the filtered sgRNAs, we selected a chain of unique sgRNAs as follows: first, the initial sgRNA was picked, and the next sgRNA was chosen according to a predetermined distance cutoff (*D*; e.g., 100 bp) and an odd-number step size (*S*; e.g., 15) such that the distance between the target sequences of the two sgRNAs was no less than *D*; the procedure was repeated until no more unique sgRNAs were found. Next, we designed the first sgRNA pair using the 1st sgRNA and the 16th (1 + *S*) sgRNA, then the second pair using the 3rd and 18th (3 + *S*) sgRNAs; we repeated this procedure to the end of the chain. The distance cutoff *D* and step size *S* were both adjustable to allow for different deletion sizes and genomic coverage. For example, with $D = 100$ and $S = 15$, the deletion size would be a minimum of 1,500 bp, and an average of 2,000 bp in the current design. The average coverage was $(1 + S)/2$, or eight times with $S = 15$, as there were eight sgRNAs (1st, 3rd, …, 15th) with crossover to eight guide RNAs on other side (16th, 18th, …, 30th) for any region in the middle. Three different sets of deletions/steps were used: 100/15, 200/13 and 500/13. A unique guide RNA was not used if it had been used in a previous selection. After a pair of dual CRISPR guide RNAs—namely, {a, b}—had been selected, we used the following template to link two guide RNAs: TGTGGAAAGGACGAAACACC{a}GTTTAGAGACG{rnd}CGTCTCACCTT{b}GTTTTAGAGCTAGAAATAGCAAGTT.

Note that if a guide RNA started with A, C or T, a G was added in front. {rnd} represents the random bases that were selected from all combinations of 8-bp nucleotide sequences excluding numbers of GC segments less than 4 or more than 6, or including any subsequence within: {AAAA, CCCC, TTTT, GGGG, GAGACG, or CGTCTC}.

Oligo synthesis and library cloning. The CREST-seq oligo library with sequences shown in **Supplementary Figure 2a** was amplified with the following primers:

Forward primer: CTTGTGGAAAGGACGAAAC
Reverse primer: TTTTAACTTGCTATTTCTAGCTCTAAAAC

The PCR product was size-selected and gel-purified with NucleoSpin gel and a PCR clean-up kit (Clontech; 740609), and then inserted into BsmbI-digested lentiCRISPRv2 plasmid by Gibson assembly (Addgene, 52961). The end product was electrotransformed into 5-α electrocompetent *Escherichia coli* (NEB; C2989K) and grown on agar plates. About 20 million independent bacterial colonies were collected, and the plasmids were extracted with the Qiagen Plasmid Giga Kit (12191). The resulting plasmid DNA was linearized by BsmbI digestion, gel-purified, and ligated with a DNA fragment (the complete IDT gBlocks sequence is in **Supplementary Table 8**) containing tracRNA(E/F) and the mouse U6 promoter (mU6). The ligate was electrotransformed into 5-α electrocompetent *E. coli* and plated on agar plates. About 20 million bacterial colonies were collected and purified with the EndoFree Plasmid Giga Kit (Qiagen; 12391).

Lentiviral library production. The CREST-seq lentiviral library was prepared as previously described[52], with minor modifications. Briefly, 5 μg of lentiCRISPR plasmid library was cotransfected with 4 μg of PsPAX2 and 1 μg of pMD2.G (Addgene, 12260 and 12259) into a 10-cm dish of HEK293T cells in DMEM (Life Technologies) containing 10% FBS (Life Technologies) by PolyJet transfection reagents (Signagen; SL100688). Growth medium was replaced 6 h after transfection. The supernatant of the cell culture media was harvested at 24 h and 48 h after transfection, and filtered through Millex-HV 0.45-μm PVDF filters (Millipore; SLHV033RS). The viruses were further concentrated with 100,000 NMWL Amicon Ultra-15 centrifugal filter units (Amicon; UFC910008).

For viral titration, 0.5 million POU5F1–eGFP hESCs were seeded per well on a six-well plate. 12 h later, different amounts (1, 2, 4 and 8 μl) of concentrated virus-containing media were added to the cell culture media to infect the hESCs according to the same protocol described in the section on lentiviral screening. The same number of noninfected cells was seeded and not treated with puromycin as the control. 24 h post-infection, the infected cells were treated with 500 ng/ml puromycin (Life Technologies; A1113802) for another 72 h. We counted the number of puromycin-resistant cells and control cells to calculate the ratio of infected cells and the viral titer. In the screening, about 10 million POU5F1–eGFP hESCs were used in each independent screening replicate and infected with viral particles at a low multiplicity of infection (MOI; 0.1) to make sure each infected cell got one viral particle.

Lentiviral transduction and FACS. Briefly, the screening was carried out according to a previously described protocol[13], with minor modifications. In each independent screen, about 10 million cells per 12-well plate were spin-infected with the CREST-seq lentiviral library at an MOI of 0.1. 24 h post-infection, the cells were dissociated with Accutase and plated into a 15-cm culture dish coated with Matrigel (4 million cells per dish). The cells

doi:10.1038/nmeth.4264

were treated with E8 media containing 250 ng/ml puromycin for 7 d, and then cultured for another 7-d without puromycin treatment. For CREST-seq screen FACS sorting, the cells were dissociated and coimmunostained with phycoerythrin (PE)-conjugated anti-POU5F1 and allophycocyanin (APC)-conjugated anti-eGFP. The eGFP$^-$POU5F1$^+$, eGFP$^+$POU5F1$^+$ and non-sorted control cells were collected by FACS for further analysis.

Sequencing library construction. Genomic DNA was extracted from eGFP$^-$POU5F1$^+$, eGFP$^+$POU5F1$^+$ and non-sorted control cell populations. The sgRNA inserts were then amplified from genomic DNA by PCR using the following primers:

Forward: AATGGACTATCATATGCTTACCGTAACTTGAAAGTATTTCG
Reverse: GGACTGTGGGCGATGTGCGCTCTG

The PCR products were gel-purified and subjected to a second PCR reaction to add the Illumina TruSeq adaptor sequence with the following primers:

Forward: AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTctTGTGGAAAGGACGAAAC
Reverse ("N" indicates the index sequence): CAAGCAGAAGACGGCATACGAGANNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTTTAACTTGCTATTTCTAGCTCTAAAAC

Sequencing and processing of CREST-seq libraries. CREST-seq libraries were sequenced using HiSeq 4000 in pair-ended mode with 100-bp read length. An sgRNA pair {a, b} was considered valid if it matched the initial sgRNA design and met the following criteria: (1) a subsequence of read 1 matched GGACGAAACACCG, followed by 19 or 20 nt (namely, {a′}), and GTTTAAGAGCTATGCTG; (2) a subsequence of read 2 matched AAAC, followed by 19 or 20 nt (namely, {b′}), and CAA; (3) {a} exactly matched {a′} if the length of {a′} was 20 nt, or {a} exactly matched G + {a′} if the length of {a′} was 19 nt; and (4) {b} exactly matched the reverse complementary sequence of {b′} if the length of {b′} was 20 nt, or {b} exactly matched the G+ reverse complementary sequence of {b′} if the length of {b′} was 19. Those sgRNA pairs with total read counts less than 30 among all samples were filtered out. In the end, we kept 10,159 sgRNA pairs for further analysis (**Supplementary Table 4**).

Peak-calling in CREST-seq data. For each sgRNA pair, the MAGeCK algorithm[20] was used to estimate the statistical significance (using a negative binomial test) of enrichment in the cell population relative to the control population. Next, we ranked sgRNA pairs in increasing order, using the equation log (NB *P*) × sign (log(exp/control)) (NB, negative binomial). Third, we partitioned the 2-Mb screened region into a set of non-overlapping 50-bp bins, $B = (b_1, \ldots, b_n)$; a bin was considered positive if many of the sgRNA pairs spanning it ranked near the top of the sorted list. A robust rank aggregation (RRA) algorithm[53] was then used to identify the positive bins. Specifically, we let $R_i = (r_{i1}, \ldots, r_{ik})$ be the vector of ranks of sgRNA pairs that spanned bin b_i, and we normalized R_i into percentiles $U_i = (u_{i1}, \ldots, u_{ik})$, where $u_{ij} = r_{ij}/M$ (where M is the total number of sgRNA pairs). The goal was to identify the bins for which the normalized rank vector U_i skewed strongly toward zero. Under the null hypothesis where the normalized ranks follow a uniform distribution between 0 and 1, the jth smallest value among $(u_{i1}, \ldots, u_{ik})$ is an order statistic $\rho(u_{ij})$ that can be calculated by the β-distribution $\beta(j, k + 1 - j)$ We defined the final score for the rank vector U_i as the minimum of the negative score:

$$\rho(U_i) = \min_{i=1\to k} \rho(u_{ij})$$

The $\rho(U_i)$ score was converted to a *P* value by permutation test, as proposed by Li *et al.*[20], and finally the *P* value was adjusted to an FDR by the Benjamini–Hochberg procedure. A bin was considered as significant if its FDR was smaller than a custom threshold.

Calculation of enrichment test score. We downloaded DHSs and peaks of ChIP-seq data sets from H1 hESCs from the ENCODE data portal[5]. Enhancers were predicted by RFECS[54], and promoter coordinates were based on RefSeq gene annotation. The observed overlap ratio o_i of feature i was computed as the fraction of CREST-seq peaks that overlapped with that feature. We then randomly shuffled CREST-seq peaks in the region using shuffleBed[55], and counted the expected overlap rate e_i as the fraction of shuffled peaks that overlapped with feature i. Fold enrichment was computed as o_i/e_i. We repeated this process 1,000 times for each feature and defined the enrichment test score as the fraction of tests in whcih the fold enrichment was >1. The significance of enrichment was determined by χ^2 test.

Analysis of chromatin signatures of *POU5F1*-regulating promoters. We randomly shuffled CREST-seq peaks in the 2-Mb *POU5F1* region using shuffleBed[55] and kept only those permutations with 18 peaks overlapping promoter regions. The expected overlap rate for each shuffle was counted as the fraction of permutations that contained an active promoter signature (Pol2/H3k4m3/H3k27ac). We repeated this process 1,000 times and calculated the permutation *P* value as the percentage of tests in which the overlap rate was >0.78.

Classification of *POU5F1*-regulating promoters by random forest. We downloaded RefSeq-annotated promoters (2,000 bp upstream from the TSS) within the screened region from the UCSC genome browser. Promoters were divided into positive and control groups on the basis of their overlap with CREs. RNA-seq data were taken from previously published work, and gene expression was estimated with the software Cufflinks for each transcript. A random forest implemented by the R package "randomForest" was applied to classify positive promoters from the negative ones with default parameter settings, without further model selection. Prediction performance was evaluated by leave-one-out cross-validation. We estimated feature importance on the basis of the average decrease of node purity by permuting each variable.

CRISPR–Cas9-mediated deletion. CRISPR–Cas9 constructs targeting the genomic loci indicated in **Supplementary Figure 6a** were made according to a previously described protocol[13]. The oligos used for cloning are listed in **Supplementary Table 8**.

The designed sgRNA sequence was cloned into the pX330-U6-Chimeric_BB-CBh-hSpCas9 vector (Addgene, 42230). After validating the sgRNA sequences by Sanger sequencing, we mixed a pair of plasmids targeting the 5′ and 3′ boundaries of the same element at a 1:1 ratio and cotransfected plasmid expressing mCherry into POU5F1–eGFP cells with hESC Nuclearfector Kit 2 (Lonzo; VPH-5022) according to the manufacturer's instructions. To knock out *POU5F1*-regulatory core promoters, we used *in vitro*–synthesized CRISPR crRNA and CRISPR tracrRNA (IDT) with the sequence specified in **Supplementary Table 8**. The Cas9 recombinant protein was purchased from NEB (M0386M), and the Cas9/crRNA/tracRNA was assembled *in vitro* according to a published protocol[56]. The RNP complex was electrotransfected into the POU5F1–eGFP hESC reporter line with the Neon Transfection System 10-µl kit (Thermo Fisher Scientific; MPK1096) with default electrotransfection protocol #9.

72 h after transfection, the mCherry-positive cells were collected by FACS. mCherry-positive single cells were plated into a Matrigel-coated plate at low density (about 1,000 cells per 10-cm coated petri dish) and cultured in E8 media supplemented with 10 µM ROCK inhibitor. After 10–14 d, the surviving sorted single cells formed colonies. Individual colonies were picked and expanded, and then subjected to genotyping and in-depth analysis.

Genotyping of mutant clones. The cells from mutant clones were collected and treated with QuickExtract DNA extraction solution (Epicentre; QE0905T), and then subjected to genotyping PCR using the primers listed in **Supplementary Table 8**. We then carried out Topo cloning (Life Technologies; K2800-20) and Sanger sequencing to verify the sequences.

FACS analysis. To directly monitor eGFP expression levels, we dissociated wild-type or mutant POU5F1–eGFP cells with Accutase and subjected them to FACS analysis with a BD FACSAria II. To examine the levels of HLA-C protein, we stained the cells with PE-conjugated antibody that specifically recognized HLA-C (Millpore; MABF233). For immunostaining of eGFP, POU5F1 or γH2AX, the cells were fixed with 2% PFA for 30 min and then subjected to overnight permeabilization in methanol at −20 °C. The treated cells were stained with the appropriate antibodies. PerCP-cy5.5-conjugated mouse anti-H2AX(pS139) was purchased from BD Biosciences (564718), PE-conjugated anti-human OCT4(OCT3) was from STEMCELL Technologies (60093PE.1), and APC-conjugated anti-GFPuv/eGFP was from R&D Systems (IC4240A).

Luciferase reporter assays. Luciferase assays were conducted as previously described[57]. Briefly, for tests of the enhancer activity of CREs with the native *POU5F1* promoter, the 360-bp *POU5F1* minimal promoter[32] (hg18 Chr6: 31,246,377–31,246,736) was synthesized as gblock by IDT and cloned into the pGL3-promoter vector to replace the original SV-40 promoter. The core promoter regions of pPRRC2A, pMSH5, pNEU1 and pTFC19 were PCR-amplified from H1 hESC genomic DNA and cloned into a modified pGL3-POU5F1 vector (Promega) in which the SV-40 promoter had been replaced by a 360-bp minimal *POU5F1* promoter by in-fusion cloning. The primer sequences are listed in **Supplementary Table 8**. After validation by Sanger sequencing, the constructs were cotransfected with pRL-SV40 Renilla reporter vector in H1 hESCs with Fugene HD (Roche) at a 4:1 reagent-to-DNA ratio. The transfected cells were cultured for an additional 2 d before being harvested for reporter assay. We used a dual-luciferase reporter assay kit (Promega; E1960) according to the manufacturer's protocol. The adjusted firefly luciferase activity of each sample was normalized to the average of the activities of three negative control regions.

RNA interference. The siRNAs were purchased from Dharmacon in the format of ON-TARGETplusSMARTpool-Human targeting *MSH5*, *NEU1* and *PRRC2A*. We also designed siRNAs by using the WI siRNA selection program. siRNA sequences are listed in **Supplementary Table 8**. The siRNAs were transfected into hESCs with Human Stem Cell Nucleofector Kit 2 (Lonza) per the manufacturer's instructions.

Western blotting. We carried out western blotting according to a previously described protocol[58]. Briefly, whole cell extracts (WCEs) were collected and quantified with a Pierce BCA protein assay kit (23225). 30 µg of WCE of each sample was subjected to western blotting analysis with antibodies specifically recognizing NEU1 (Thermo Scientific; PA5-42552), PRRC2A (Abcam; ab188301), MSH5 (Abcam; ab130484), histone H3(Abcam; ab1791), POU5F1 (Abcam; ab19875) and eGFP (Abcam; ab190584).

ATAC-seq experiment and analysis. ATAC-seq was carried out according to a previously described protocol[22]. Briefly, each library started with 100,000 cells, which were permeabilized with NPB (0.2% NP-40, 5% BSA, 1 mM DTT in PBS with one complete proteinase inhibitor) at 4 °C for 10 min, and then spun down at 500*g* for 5 min at 4 °C. The resulting nuclei were resuspended in 20 µl of 1× DMF (33 mM Tris-acetate, pH 7.8, 166 mM potassium acetate, 10 mM magnesium acetate, 16% DMF). For chromatin tagmentation, 0.5 µl of Tn5 (provided in the Nextera DNA kit, Illumina) was added to 10 µl of solution for 30 min at 37 °C.

We processed our ATAC-seq data in the following steps: (1) ATAC-seq sequencing reads were mapped to the hg19 reference genome using Bowtie in paired-end mode; (2) poorly mapped, improperly paired, and mitochondrial reads were filtered; (3) PCR duplications were removed using Picards MarkDuplicates (http://broadinstitute.github.io/picard); (4) mapping positions of reads were adjusted to account for Tn5 insertion; (6) reads were shifted 75 bp, and peaks were called with MACS2 (ref. 59) with the parameters -q 0.01 --nomodel --shift 175 --B --SPMR --keep-dup all --call-summits; and (7) ATAC-seq signal was normalized into reads per kilobase of transcript per million mapped reads (RPKM), using deeptools[60] for visualization.

PCA. We first extracted all 478 H1 DHSs within the screened regions and counted the average RPKM for each site using 122 public DHS data sets (**Supplementary Table 8**) and our own ATAC-seq data set. Pairwise Pearson correlation between the data sets was calculated and used as input for principal component analysis. We found that the first two principal components accounted for 80% of the variance, and therefore we used them for 2D visualization as shown in **Supplementary Figure 5b**.

doi:10.1038/nmeth.4264

Code availability. The program and database for whole-genome CRISPR design is available at https://github.com/bil022/CRISPR-web. The computer code used in this study is available at https://github.com/r3fang/CRESTseq.

Data availability. Sequencing data have been deposited in the NCBI Gene Expression Omnibus (GEO) under accession number GSE81026. Additional materials, data, code and associated protocols are available upon request.

49. Diao, Y., Fang, R., Li, B. & Ren, B. A dual sgRNA mediated tiling-deletion based genetic screen to identify regulatory DNA sequence in mammalian cells. *Protoc. exch.* http://dx.doi.org/10.1038/protex.2017.037 (2017).
50. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
51. Wu, X., Kriz, A.J. & Sharp, P.A. Target specificity of the CRISPR-Cas9 system. *Quant. Biol.* **2**, 59–70 (2014).
52. Meng, Z. *et al.* Berbamine inhibits the growth of liver cancer cells and cancer-initiating cells by targeting Ca^{2+}/calmodulin-dependent protein kinase II. *Mol. Cancer Ther.* **12**, 2067–2077 (2013).
53. Kolde, R., Laur, S., Adler, P. & Vilo, J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* **28**, 573–580 (2012).
54. Rajagopal, N. *et al.* RFECS: a random-forest based algorithm for enhancer identification from chromatin state. *PLoS Comput. Biol.* **9**, e1002968 (2013).
55. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
56. Kelley, M.L., Strezoska, Ž., He, K., Vermeulen, A. & Smith, Av. Versatility of chemically synthesized guide RNAs for CRISPR-Cas9 genome editing. *J. Biotechnol.* **233**, 74–83 (2016).
57. Heintzman, N.D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).
58. Diao, Y. *et al.* Pax3/7BP is a Pax7- and Pax3-binding protein that regulates the proliferation of muscle precursor cells by an epigenetic mechanism. *Cell Stem Cell* **11**, 231–241 (2012).
59. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
60. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).