Title: A saturating mutagenesis CRISPR-Cas9 mediated functional genomic screen identifies *cis*- and *trans*- regulatory elements of *Oct4* in murine ESCs

Matthew C. Canver, [1,13] **Pratibha Tripathi,** [2,3,13] **Michael J. Bullen,** [2,3,13] **Moshe Olshansky,** [4,5] **Yogesh Kumar,** [2,3] **Lee H. Wong,** [6] **Stephen J. Turner,** [4,7] **Samuel Lessard,** [8,9,12] **Luca Pinello,** [10] **Stuart H. Orkin,** [1,11] **Partha Pratim Das** [2,3]

[1]Division of Hematology/Oncology, Boston Children's Hospital and Department of Pediatric Oncology, Dana-Farber Cancer Institute (DFCI), Harvard Stem Cell Institute, Harvard Medical School, Boston, MA 02115, USA

[2]Department of Anatomy and Developmental Biology, Monash University, Wellington Road, Clayton, Victoria 3800, Australia

[3]Development and Stem Cells Program, Monash Biomedicine Discovery Institute, Wellington Road, Clayton, Victoria 3800, Australia

[4]Department of Microbiology, Monash Biomedicine Discovery Institute, Monash University, Clayton, Victoria 3800, Australia

[5]Computational Biology & Bioinformatics, Baker Heart & Diabetes Institute, 75 Commercial Rd, Melbourne, Victoria 3004, Australia

[6]Department of Biochemistry and Molecular Biology, Cancer Program, Biomedicine Discovery Institute, Monash University, Wellington Road, Clayton, Victoria, 3800, Australia

[7]Department of Microbiology and Immunology, University of Melbourne, Parkville, Victoria 3010, Australia

[8]Research Center, Montreal Heart Institute, Montréal, QC H1T 1C8, Canada

[9]Department of Medicine, Faculty of Medicine, Université de Montréal, Montréal, QC H3T 1J4, Canada

[10]Molecular Pathology & Cancer Center, Massachusetts General Hospital & Harvard Medical School, Boston, MA 02114, USA

[11]Howard Hughes Medical Institute, Boston, MA 02115, USA

[12]Present address: Rare Blood Disorders, Sanofi, 225 Second Avenue, Waltham, MA 02451, USA

[13]These authors contributed equally to this work

*Correspondence to:
stuart_orkin@dfci.harvard.edu
partha.das@monash.edu

Running tile: ctSCAN-SMS identifies *cis*- and *trans*-regulatory elements of *Oct4*

Keywords: ctSCAN-SMS, regulatory elements (REs), CRISPR-Cas9

Abstract

Regulatory elements (REs) consist of enhancers and promoters that occupy a significant portion of the non-coding genome and control gene expression programs either in –*cis* or in –*trans*. Putative REs have been identified largely based on their regulatory features (co-occupancy of ESC-specific transcription factors, enhancer histone marks and DNase hypersensitivity) in mouse embryonic stem cells (mESCs). However, less has been established regarding their regulatory functions in their native context. We deployed *cis*- and *trans*-regulatory elements scanning through saturating mutagenesis and sequencing (ctSCAN-SMS) to target elements within the ~12kb *cis*-region (*Cis*-REs; CREs) of the *Oct4* gene locus, as well as genome-wide 2,613 high-confidence *trans*-REs (TREs), in mESCs. ctSCAN-SMS identified 10 CREs and 12 TREs, as novel candidate REs of the *Oct4* gene in mESCs. Furthermore, deletions of these candidate REs confirmed that the majority of the REs are functionally active, and CREs are more active than TREs in controlling *Oct4* gene expression. A subset of active CREs and TREs physically interact with the *Oct4* promoter to varying degrees; specifically, a greater number of active CREs compared to active TREs, physically interact with the *Oct4* promoter. Moreover, comparative genomics analysis reveals that more number of active CREs than active TREs are evolutionary conserved between mouse and primates, including human. Taken together, our study demonstrates the reliability and robustness of ctSCAN-SMS screening to identify critical REs, and investigate their roles in the regulation of transcriptional output of a target gene (in this case *Oct4)* in their native context.

Introduction

Large-scale genomic studies reveal that ~80% of the human genome may be involved in gene regulation, whereas only ~2% of the genome encodes proteins (1). The functional non-coding genome can be broadly divided into regulatory elements (REs) and regions that encode non-coding RNAs (ncRNAs) (1, 2). Furthermore, REs can be sub-divided into *cis*-REs (CREs) and *trans*-REs (TREs), based on their position relative to their target gene(s). CREs are present proximally or distally relative to their target gene on the same chromosome, whereas TREs are located distally relative to their target gene on different chromosomes (3, 4). Putative REs have been identified using various methods, including transcription factor binding, enhancer histone marks, DNA accessibility (open chromatin regions), enhancer-promoter interactions, and gene expression (1, 5-10). REs usually enriched with sequence variants that are associated with diverse human traits and diseases (11-13). In addition, REs play crucial roles in evolutionary turnover and divergence (14-17).

Initial efforts systematically evaluated RE functions using reporter assays on a massive scale (18, 19); however, such approaches fail to interrogate REs functions within their native genomic contexts. Advances in CRISPR-Cas9 mediated genome editing technology (20, 21) have transformed analysis of protein-coding genes (22, 23), as well as REs *in situ* in chromatin. A few high-throughput CRISPR-Cas9 mediated functional genetic screens were performed to characterize CREs in mammalian cells (24-29). Prior screens to identify functional CREs were focused on – either targeting putative CREs of gene(s) of interest (gene-centric) or targeting putative CREs bound by selected TFs (TF-centric). Nevertheless, identification of functional TREs presents a challenge that has attracted less attention.

Here, we deployed genome-wide *cis*- and *trans*-regulatory elements scanning through saturating mutagenesis and sequencing (ctSCAN-SMS) in mouse embryonic stem cells (mESCs) to identify critical CREs and TREs of the *Pou5f1/Oct4* gene (a master pluripotency regulator of mESCs). We uncovered new functionally active CREs and TREs, and how they regulate *Pou5f1/Oct4* gene expression in mESCs.

Results

Design of a saturating CRISPR-Cas9 pooled library for ctSCAN-SMS

In mESCs, several putative REs, including 8,563 enhancers (ENs) and 231 super-enhancers (SEs) have been identified based on co-occupancy of ESC-specific TFs (OCT4, NANOG, SOX2, KLF4, ESRRB), mediators (MED1), enhancer histone marks (H3K4me1, H3K27ac) and DNase I hypersensitivity (30). SEs contain multiple ENs; SEs are also more densely co-occupied with TFs, enhancer histone marks, and chromatin regulators as compared to ENs, and associate with greater transcriptional output (30). We undertook a high-throughput CRISPR-Cas9 mediated genome editing approach to comprehensively target putative REs in mESCs. First, we generated a genome-wide map of open chromatin regions using ATAC-seq in mESCs, as ATAC-seq identifies most EN REs (10). ATAC-seq peaks were then overlayed within all putative ENs (8,563) and SEs (231) to designate “high-confidence” REs (2,613) (Fig. 1a; Fig. S1a; Table S1). As these REs are distributed genome-wide and on different chromosomes relative to *Oct4* gene locus (in *trans*-), we termed these REs as TREs. All possible single guide RNAs (sgRNAs) (20 nt) were designed (within the TREs for tilling) upstream of the *S. pyogenes* Cas9 NGG-protospacer adjacent motif (PAM) sequences to target the high-confidence 2,613 TREs (Fig.1a; Fig. S1a; Table S1 and S2). This analysis yielded 70,480 sgRNAs with a median gap of 5 bp between adjacent genomic cleavages (Fig. 1, c and d). Likewise, 1,827 sgRNAs were designed at the surrounding ~12kb (-10kb to +2kb of TSS of the *Oct4*) region of the mouse *Oct4* gene locus to dissect the *cis*-REs (CREs) of *Oct4* (Fig. 1, b and c). In addition, the library included 2,000 non-targeting (NT) sgRNAs as negative controls, 119 sgRNAs targeting GFP (of the *Oct4*-GFP reporter that used for the screen) and 150 sgRNAs targeting coding sequence of mESC-TFs as positive controls (Fig. 1c). In total, the REs CRISPR-Cas9 pooled library contained 74,576 sgRNAs (Fig. 1c). These sgRNAs were synthesized, pooled, and cloned into a lentiviral vector for deep sequencing. The deep sequencing of the pooled library confirmed the presence of >95% sgRNAs that target TREs, >99% sgRNAs that target CREs and control sgRNAs (Fig. S1, b-f; Table S2).

Candidate CREs and TREs of the *Oct4* gene identified by ctSCAN-SMS

The pooled library was transduced into an *Oct4*-GFP reporter mESC line, which constitutively expresses Cas9 (31-33). The *Oct4*-GFP reporter was used as a "readout" for the screen to measure the reduction in GFP levels upon perturbation of any targeted RE regions by their corresponding sgRNAs. Lentiviral transduction of the pooled library was performed at low multiplicity (MOI) to ensure that each cell contained predominantly one sgRNA (Fig. S2a). After puromycin drug selection (since sgRNA constructs carrying puromycin drug resistance gene), "GFP-low" cells were sorted using fluorescence-activated cell sorting (FACS) (Fig. S2, a and b). As a control, cells were collected before FACS (the "pre-sort" sample). Genomic DNA was isolated from both "GFP-low" and "pre-sort" cell populations, and next-generation sequencing (NGS) was employed to enumerate the sgRNAs in each population (Fig. S2a). The screen was performed in triplicate.

We calculated an "enrichment score" of each sgRNA by comparing its frequency (presence) in GFP-low over pre-sort cells. Enrichment scores were determined based on the two best replicates (Table S2). Highest and lowest enrichment scores were obtained from GFP-targeting sgRNAs (mean log2FC 4.87, P-value <0.0001) and NT-sgRNAs (mean log2FC 0.44, P-value <0.0001), indicated that the screen was technically successful (Fig. 2a and 3a). We ranked all sgRNAs based on their enrichment scores (Table S2) and analyzed their off-target scores (ranged between 0-100) (34) (Fig. S3a and S4a; Table S2). A higher off-target score signifies fewer off-targets of a sgRNA. We found that the majority of the evaluated sgRNAs (87.6% sgRNAs for CREs, and 84.5% sgRNAs for TREs) had off-target scores >10 (Table S2); these sgRNAs are statistically significant based on the HMM (Hidden Markov Model) analysis (Fig. S3a).

To identify candidate CREs of the *Oct4*, we considered all HMM sgRNAs with off-target scores >10 and mapped them within the ~12kb surrounding region (-10kb to +2kb of TSS) of the *Oct4* gene locus. This yielded 16 candidate CREs (1-16), based on the mean enrichment score (mean log2FC) of sgRNAs per CRE. Each of the candidate CREs had a mean

log2FC>0.5 (range of mean log2FC 0.93-3.95), with P-values <0.0001, which was higher than the mean enrichment score of non-targeting (NT)-sgRNAs (mean log2FC 0.44, with P-value <0.0001) (Fig. 2, a and b). Among 16 CREs, CREs-10 and 12 have been recognized previously as distal and proximal enhancers, respectively (31); CREs-13 to 16 were present within the promoter region of *Oct4* (+/-2kb of TSS) (31). The remaining 10 CREs were newly identified candidate CREs of the *Oct4* gene (Fig. 2, a and b). Of note, CRE-12 displayed a relatively lower mean enrichment score (log2FC 0.77, P-values <0.0001) compared to others candidate CREs, because the original *Oct4*-GFP reporter lacks a portion of CRE-12 (33). Nonetheless, we included CRE-12 for further validation.

To classify the candidate TREs, we applied a statistically relevant Hidden Markov Model (HMM) to the sgRNAs enrichment scores (35), which initially identified 263 candidate TREs. Furthermore, we applied stringent criteria to select candidate TREs for validation, as follows: i) TREs must have HMM sgRNAs with off-target scores>10 (Fig. 3a and Fig. S4a); ii) TREs must possess at least 4 sgRNAs with mean log2FC >0.5 (range of mean log2FC 0.86-2.52), with P-values <0.0001 (Fig. 3b); iii) TREs must co-occupy with ESC-TFs (OCT4, NANOG, SOX2), enhancer histone marks (H3K27ac, H3K4me1) (Fig. 3c and Fig. S4b); and iv) they must contain “dynamic” open chromatin regions; i.e. “open” chromatin regions present in the undifferentiated state (0 hr) that gradually become “closed” with differentiation (24, 96hr) of mESCs (Fig. 3d and Fig. S4c). Based on these criteria, we selected 12 candidate TREs of the *Oct4* gene (Fig. 3, a and b).

Dissection of functionally active CREs and TREs of the *Oct4* gene

We selected a total of 33 REs, including 20 CREs – 16 candidate CREs and 4 control CREs (CREs with no significant enrichment scores of sgRNAs), and 13 TREs – 12 candidate TREs and one control TRE (with no significant enrichment scores of sgRNAs) of the *Oct4* gene locus for validation using wild-type mESCs. “Paired” sgRNAs (5’ and 3’ sgRNAs) tagged with mCherry were used to target both the flanking ends of each selected candidate RE to create a deletion. Following transfection of paired sgRNAs-mCherry into the wild-type mESCs,

mCherry-positive cells were sorted to measure the endogenous Oct4 mRNA expression and to confirm the genomic deletions of CREs/TREs (Fig. 2f and 3e; Fig. S3d; Fig. S4e; Table S3). We observed a significant reduction in Oct4 expression to different extents upon deletions of CREs- 1, 3, 5, 7, 10, 12, and 13-16 (out of 16 candidate CREs) (Fig. 2f). Deletions of newly identified CREs- 1, 3, 5 and 7 showed a greater reduction in Oct4 expression, compared to deletions of known distal and proximal enhancers (CREs- 10 and 12) of *Oct4* (Fig. 2f). However, "regulatory features", such as co-occupancy of ESC-TFs (OCT4, NANOG, SOX2– ONS), enhancer histone marks (H3K27ac, H3K4me1) and dynamic open chromatin regions (ATAC-seq peaks at 0hr compared to 24, 96hr), were more prominent at CREs- 7, 10, 12 compared to CREs- 1, 3, 5 (Fig. 2, c-e; Fig. S3, b and c). Moreover, deletions of CREs- 13 to 16 (present at the promoter region of *Oct4*) showed a significant reduction in Oct4 expression, as expected (Fig. 2f). Nonetheless, only CREs- 13, 14 showed substantial co-occupancy of ONS, H3K27ac and dynamic open chromatin regions, as compared to other CREs present at the *Oct4* promoter (Fig. 2, c-e; Fig. S3, b and c). In contrast, deletions of control CREs (C1-C4) led to no significant changes in Oct4 expression (Fig. 2f). Additionally, control CREs exhibited low-level co-occupancy of ONS, H3K27ac, H3K4me1, and dynamic open chromatin regions (Fig. 2, d and e; Fig. S3, b and c). These data confirm the existence of multiple "active" CREs ("active" REs depict – REs that reduce the Oct4 expression, upon their deletions), including newly identified active CREs of the *Oct4*. Yet some "active" CREs fail to display recognized regulatory features (i.e. without of any co-occupancy of TFs, enhancer histone marks, and open chromatin regions) as described recently (26, 28), suggesting that a subset of functionally "active" REs may lack typical regulatory features.

Deletions of 8 TREs (TREs- 1, 2, 3, 4, 7, 9, 11, 12) of 12 candidates showed a significant reduction in Oct4 expression to various extents (Fig. 3e). Nonetheless, all 12 TREs were co-occupied with ONS, H3K27ac and H3K4me1 enhancer marks, and exhibited dynamic open chromatin regions (Fig. 3, c and d; Fig. S4, b and c); as all the candidate TREs (TREs 1-12) were short-listed for validation based on their regulatory features. Conversely, control TRE lack typical regulatory features (Fig. 3, c and d); and deletion did not affect Oct4

expression (Fig. 3e). Moreover, neighbouring genes of most of the validated active TREs were lowly expressed in mESCs (Fig. S4d), indicating that these genes may not be critical for the mESC state maintenance. Importantly, we observed reduction of Oct4 expression was greater for the majority of active CREs than active TREs, upon their deletions (Fig. 2f and 3e). This suggests that CREs generally contribute more than TREs to control overall *Oct4* gene expression. Taken together, these data demonstrate that reliability and robustness of the ctSCAN-SMS screen to identify new candidate REs of the *Oct4* in a high-throughput manner. Moreover, we confirm that a subset of these newly identified candidate CREs and TREs are functionally "active" REs of the *Oct4* gene, and CREs are more active compared to TREs.

Cis- **and** ***trans-*** **regulation of the** ***Oct4*** **gene expression**

REs (particularly enhancers) physically interact with gene promoters and control transcription (36). Chromosome conformation capture (3C) based methods – 4C, Hi-C, capture Hi-C, ChiA-PET and HiChIP-- have been utilized to identify physical contacts between promoters and REs (enhancers) (4). To interrogate the potential mechanisms by which candidate CREs and TREs regulate *Oct4* gene expression, we examined interactions between REs (CREs and TREs) and the *Oct4* promoter using published 4C-seq data. These data were generated to study intra-chromosomal and inter-chromosomal interactions between REs and the *Oct4* promoter at a genome-wide scale (37). We used Oct4-234 (a region at ~1.5kb upstream of TSS of *Oct4*) as a "viewpoint", as previously (Fig. S5a) (37). Next, "contact frequencies" were calculated between the viewpoint and CREs (using 1kb resolution window, surrounding 30kb region of the *Oct4* gene locus) (Fig. 4, a and c), as well as between the viewpoint and TREs (using 50kb resolution window, surrounding each of the TREs) (Fig. 4b). This analysis revealed ranges of contact frequencies between functionally validated "active" CREs/TREs and the *Oct4* promoter (Fig. 2f and 3e; Fig. 4, a and b). For example, functionally validated "active" CREs, including CREs- 3, 5, 7 (newly identified CREs), CREs- 10 and 12 (previously known as distal and proximal enhancers of *Oct4*), and CREs- 13 to16 (residing at the promoter region of *Oct4*) demonstrated intra-chromosomal interactions with the *Oct4* promoter (Fig. 4, a and c).

Established enhancers, such as CREs-10 and 12, showed higher contact frequencies compared to newly identified CREs- 3, 5, 7 (Fig. 4, a and c). However, active CRE-1 and control CREs (C1-C4) did not show significant interactions with the *Oct4* promoter (Fig. 4, a and c). In contrast, among functionally validated "active" TREs- 1, 2, 3, 4, 7, 9, 11, 12; only a minority (TREs- 2, 3, 4, 11) displayed higher inter-chromosomal interactions (compared to control TRE) with the *Oct4* promoter (Fig. 4b).

Likewise, a high-resolution Micro-C (Micrococcal nuclease (MNase) based Hi-C assay that captures genome-wide 3D chromatin organization/contact frequencies at single nucleosome resolution (~100-200bp)) from mESCs (38), displayed ranges of intra-chromosomal interactions between active CREs and the *Oct4* promoter (Fig. S5, b and c). Further, these data evaluated all other intra-chromosomal interactions between any two genomic loci around ~20kb region of the *Oct4* gene locus at 200bp resolution (Fig. S5b). In addition, inter-chromosomal interactions were also observed between a few active TREs (TREs- 4, 11) and the *Oct4* promoter (Fig. S5d). Both 4C-seq and Micro-C data showed that greater contact frequencies of active CREs compared to active TREs with the *Oct4* promoter, and a greater number of active CREs compared to active TREs physically interacted with the *Oct4* promoter (Fig. 4, a-c; Fig. S5, b-d).

Taken together, our data suggest that a subset of active CREs and TREs physically interact with the *Oct4* promoter at different frequencies as they influence the *Oct4* transcriptional output.

Conserved functionally active CREs and TREs of the *Oct4* gene

Recent studies demonstrate that the majority of species-specific REs/ ENs evolved *de novo* from ancestral DNA regulatory sequences (16, 39). Also, evidence implies that loss or gain of REs (called RE turnover) takes place during evolution (17). To understand the importance of validated active CREs and TREs of the mouse *Oct4* gene in evolutionary turnover, we analyzed their regulatory sequence conservation. Active CREs- 3, 5; CREs- 10 and 12 (known distal and proximal ENs); CREs- 13 to 16 (present within the promoter) demonstrated

significant conservation between mouse and primates, including human (Fig. 5a); whereas active CREs- 1, 7 did not show appreciable sequence conservation between mouse and primates (Fig. 5a). In comparison to active CREs, only a small subset of active TREs (TREs-3 and 4) showed significant sequence conservation between mouse and primates (Fig. 5, b and c). These observations suggest that active CREs are more critical than active TREs for *Oct4* expression in mouse and primates (including human) during evolution.

Next, we analyzed regulatory sequence conservation of previously identified high-confidence CREs (-449, -571, -694) of human *OCT4*. These CREs are located distally between ~450 to 700kb upstream of the human *OCT4* TSS, and physically interact with the *OCT4* gene (28). We found that these human CREs are also evolutionary conserved at the upstream regions of the mouse *Oct4* locus (Fig. S6).

Taken together, our comprehensive comparative genomic analysis shows that a larger subset of active CREs compared to active TREs of *Oct4* are well-conserved between mouse and primates including human. These observations support the existence of both conserved, as well as non-conserved REs of *Oct4* among mouse and primates (including human), which is consistent with “RE turnover/ divergence” (gain and loss of REs) at the *Oct4* locus during evolution that may be critical for positive selection, as proposed earlier (17, 39).

Discussion

A handful of modest scale CRISPR-Cas9 mediated functional screens (using hundreds to thousands of sgRNAs) have been performed to target specific non-coding CREs of gene(s) of interest (24-26, 29, 40). These screens successfully identified functional CREs of the target gene(s). In the context of identification of CREs of the *OCT4* gene, a previous CRISPR-Cas9 mediated screen was performed to target 174 candidate CREs of *OCT4* within its 1MB topological associated domain (TAD) in human ESCs; it revealed 4 temporary CREs and 2 known proximal CREs. The temporary CREs showed “transient” enhancer regulatory activity in *OCT4* gene expression (27). However, the functional relevance of these temporary CREs is uncertain in human *OCT4* gene regulation. Furthermore, another CRISPR-Cas9 mediated

screen was performed by the same group using a different strategy, called CREST-seq. This method employed 11,570 paired sgRNAs to introduce deletions to target 2Mb surrounding the *OCT4* locus in human ESCs (hESCs), which created 2kb deletions on average with an overlap of 1.9kb between two adjacent deletions. This screen identified total 45 CREs, of which 17 CREs (with regulatory features) reside at the promoters of "unrelated" genes (intra-chromosomally) that act as typical enhancers of the *OCT4* gene (28). Our study employed ctSCAN-SMS – an unbiased, high-resolution, high-throughput screening approach using 1,827 sgRNAs to target CREs and 70,480 sgRNAs to target TREs of the mouse *Oct4* gene (Fig. 1).

Previous CRISPR-Cas9 screens were focused on CREs of the target gene(s). In contrast, our screen was designed to identify both CREs and TREs. Indeed, we discovered 16 CREs (including 10 novel CREs) and 12 novel TREs of the murine *Oct4* gene, as potential candidate REs (Fig. 2a and 3a). Furthermore, deletion studies confirmed that the majority of these CREs (10 CREs out of 16 CREs) and TREs (8 TREs out of 12 TREs) are functionally "active" in controlling the *Oct4* expression. Overall, CREs are more functionally active than TREs (Fig. 2f and 3e). In addition, we showed that a subset of these active CREs and TREs physically interacts with the *Oct4* promoter to different extents through intra- and inter-chromosomal interactions, respectively (Fig. 4; Fig. S5). Notably, a greater number of active CREs, compared with active TREs, physically interact with the *Oct4* promoter. Moreover, the interactions between active CREs (compared to active TREs) and the *Oct4* promoter are more prominent (Fig. 4, Fig. S5). Nonetheless, "enhancer activities" of several active CREs/TREs are not directly correlated to their physical contact frequencies with the *Oct4* promoter (Fig. 2f, 3e; Fig. 4; Fig. S5). Interestingly, we found several active CREs (CREs- 1, 3, 5) lack typical regulatory features (Fig. 2, d-f; Fig. 6); this supports earlier studies that identified unmarked REs (UREs) without typical regulatory features may play critical roles in transcriptional output (26, 28). Moreover, comparative genomics analysis revealed that numerous active CREs and active TREs (more number of active CREs than active TREs) of *Oct4* are evolutionarily conserved between mouse and primates (including human) (Fig. 5). However, we also

observed divergence of *Oct4* REs among mouse and primates (including human), which may account for the vital roles of "RE turnover" during evolution (3, 39).

Several studies demonstrate that "multiple REs" act either in a co-operative or competitive fashion to control transcriptional output (39). Our study identified multiple active REs of *Oct4* and revealed a spectrum of regulatory activities of "individual" CREs and TREs in *Oct4* gene expression (Fig. 2f and 3e). Further systematic studies will be required to elucidate how multiple REs function combinatorially to control *Oct4* transcriptional output . In conclusion, we have demonstrated the utility of ctSCAN-SMS as an approach to identify functional REs of a gene locus, and dissect their regulatory contributions to the transcriptional output of a target gene within its normal chromosomal context.

Experimental procedures

Mouse embryonic stem cells (mESCs)

Mouse ESCs (mESCs) were cultured in mouse ESC media that contains DMEM (Dulbecco's modified Eagle's medium) (Thermo Fisher Scientific) supplemented with 15% fetal calf serum (FCS) (Merck Millipore), 0.1mM b-mercaptoethanol (Sigma-Aldrich), 2mM L-glutamine (Thermo Fisher Scientific), 0.1mM nonessential amino acid (Thermo Fisher Scientific), 1% of nucleoside mix (Merck Millipore), 1000 U/ml recombinant leukemia inhibitory factor (LIF/ESGRO) (Merck Millipore), and 50U/ml Penicillin/Streptomycin (Thermo Fisher Scientific), as described previously (32, 41). mESCs were cultured at 37°C, 5% CO_2.

Mouse REs CRISPR-Cas9 pooled library design for the ctSCAN-SMS

For this study, we selected a list of putative 8,563 enhancer (EN) and 231 super-enhancer (SE) REs from the mESCs, as described previously (30). First, we generated ATAC-seq (10) data from wild-type mESCs, and mapped within all the putative EN and SE REs to identify open chromatin regions, as well as high-confidence REs. Next, ±100 bp (200 bp) around the

centre of the ATAC-seq peaks were obtained from the high-confidence REs. In total, we identified 2,613 REs for targeting. All possible single guide RNAs (sgRNAs) (20nt) were designed upstream of the *S. pyogenes* Cas9 NGG-protospacer adjacent motif (PAM) sequences at these defined REs (2,613), which created 70,480 sgRNAs with a median gap 5 bp between adjacent genomic cleavages. Since these EN and SE REs distributed in *trans*- of the *Oct4* gene locus, we called these REs as *trans*-REs (TREs). Likewise, 1,827 sgRNAs were designed prior to all possible NGG-PAM sequences at the adjacent ~12kb (-10kb to +2kb of TSS of *Oct4*) region of the mouse *Oct4* gene locus to systematically dissect the REs of *Oct4*. As these REs reside adjacent to the *Oct4* gene locus, they are called *cis*-REs (CREs). We also included 2,000 non-targeting (NT) sgRNAs as negative controls; 119 sgRNAs targeting GFP of the *Oct4*-GFP reporter and 150 sgRNAs targeting coding sequence of mESC-TFs as positive controls. Altogether, the REs CRISPR-Cas9 pooled library contained total 74,576 sgRNAs.

REs CRISPR-Cas9 pooled library construction for the ctSCAN-SMS

All the sgRNA oligonucleotides of the library were synthesized as previously described (32) using a B3 synthesizer (CustomArray, Inc.), pooled together, PCR amplified and cloned into Esp3I-digested plentiGuide-Puro (Addgene plasmid ID: 52963) lentiviral vector, using a Gibson assembly master mix (New England Biolabs).Gibson assembly products were transformed into electrocompetent cells (*E. cloni*, Lucigen) and plated on 245mm x 245mm square LB-agar plates to obtain the sufficient number of bacterial colonies at a ~50× library coverage. Bacterial colonies were collected from the plates, genomic DNA was isolated and plasmid libraries were prepared for high-throughput sequencing to confirm the representation of individual sgRNA in the REs CRISPR-Cas9 pooled library.

Lentiviral library production

HEK293T cells were seeded onto 15cm dishes ~24hrs prior to transfection. Cells were transfected at 80% confluence in 16ml of media with 8.75μg of VSVG, 16.25μg of psPAX2,

and 25µg of the REs CRISPR-Cas9 pooled lentiviral plasmids, using 150µg of linear polyethylenimine (PEI) (Sigma-Aldrich). Media was changed with fresh media, 16–24hrs after transfection. Lentiviral supernatant was collected at 48 and 72hrs post-transfection and subsequently concentrated by ultracentrifugation (24,000 rpm, 4ºC, 2hrs) (Beckman Coulter SW32).

CRISPR-Cas9 mediated ctSCAN-SMS in mESCs

Oct4-GFP reporter mESCs with stably expressed Cas9 were transduced with REs CRISPR-Cas9 pooled lentiviral library at low multiplicity of infection (MOI) to avoid more than one lentiviral integration per cell. Test transductions were performed to estimate the viral titration and transduction rate. Briefly, 300,000 *Oct4*-GFP+Cas9 mESCs were plated per well of a 12-well plate. After 24hrs, different amounts of (1, 2, 4, 6, 8µl) of the lentiviral library was added to the cells. 10µg/ml blasticidin (InvivoGen) and 1µg/ml puromycin (Sigma-Aldrich) were added 24hrs after the transduction to select for lentiviral library integrants (puromycin resistant) in cells with Cas9 (blasticidin resistant). Cells were selected for the next 3-4 days. The same number of cells were seeded as a control; but not infected with lentiviral library and not treated with blasticidin and puromycin. The number of blasticidin and puromycin resistant cells, and control cells were counted to calculate the viral titre and transduction rate (to achieve 30%).

For the actual screen, we seeded ~112 million (~75K sgRNAs in the pooled library, with 500X coverage, for 30% transduction rate) *Oct4*-GFP+Cas9 mESCs in the same format (i.e. 300K cells/ well of the 12-well plate) for each independent screening replicate. Lentiviral library was added to each well of 12-well plate to achieve 30% transduction rate with low MOI (MOI 0.1) to make sure each infected cell obtained one viral particle. 24hrs post-transduction, fresh mESC media was added to the cells with 10µg/ml blasticidin (InvivoGen) and 1µg/ml puromycin (Sigma-Aldrich) and selected for 4 days. These selected cells were used to sort the “GFP-low” cells. The “pre-sort” cells were collected before sorting and used as a control. Genomic DNA was isolated from both the “GFP-low” and “pre-sort” cell populations, libraries

were prepared for deep sequencing to enumerate the presence of sgRNAs in these cell populations as previously described (32). The screening was performed in biological triplicates.

Deletion of CREs and TREs of *Oct4* in mESCs

Paired sgRNAs (5' and 3' sgRNAs) were designed to target both the ends of each selected candidate RE to create a deletion. Both the sgRNAs were cloned into lentiguide-plasmid that carries Cas9 and mCherry, using Golden Gate Cloning approach as previously described (32). 1,000,000 wild-type (J1) mESCs were transfected with 1µg of each 5' and 3' sgRNA-Cas9-mCherry plasmids using Lipofectamine 2000 (Thermo Fisher Scientific). After 24hrs of transfection, mCherry-positive cells were sorted; at least 50,000-100,000 mCherry-positive cells were collected either to isolate the total RNA for measuring the endogenous Oct4 mRNA expression levels by quantitative RT-PCR or to perform the genotyping PCR for checking deletions of targeted REs.

Flow cytometry

Cells were dissociated using trypsin, washed with 1XPBS, followed by sorting. i) During the CRISPR-Cas9 screening, *Oct4*-GFP reporter mESCs were sorted based on GFP-low intensity after the transduction with REs CRISPR-Cas9 pooled library; ii) to quantify the Oct4 mRNA expressions upon deletions of CREs and TREs, their targeting sgRNAs-mCherry-positive cells were sorted.

RNA isolation and RT-qPCR

DNA-free total RNA was isolated from mESCs using RNeasy Mini Kit (Qiagen), and cDNA was prepared using iScript cDNA Synthesis Kit (Bio-Rad). RT-qPCR was performed using iQ SYBR Green Supermix (Bio-Rad) on Bio-Rad iCycler RT-PCR detection system.

ctSCAN-SMS screen data analysis

sgRNA sequences present in the GFP-low and pre-sort pools were enumerated. Enrichment was determined by the log2 transformation of the median number of occurrences of a particular sgRNA in the GFP-low pool divided by the median number of occurrences of the same sgRNA in the pre-sort pool across the best two biological screen replicates.

ATAC-seq experiment and data analysis

ATAC-seq was performed according to the previously described protocol (10), with some modifications. Briefly, each library was started with 50,000 cells, which were washed with 1X PBS and permeabilized with 50µl of lysis buffer (10mM Tris, pH 7.4; 10mM NaCl; 3mM MgCl2; 0.1% IGEPAL) at 4^0C by resuspension. Cells were centrifuged at 500g for 10 min at 4^0C to pellet the nuclei. The resulting nuclei were resuspended in 50µl of transposition reaction buffer (25µl of 2x TD buffer from the Nextera kit, Illumina; 2.5µl of Tn5 transposase enzyme from the Nextera kit, Illumina; 22.5µl of nuclease free water), and incubated at 37^0C for 90 min for chromatin tagmentation. Next, DNA was purified using Qiagen MinElute PCR purification kit, and eluted in 10µl of nuclease-free water. PCR amplification was performed using Nextera primers (Illumina) to make the libraries for deep sequencing.

The obtained deep sequencing data in .FASTQ format was inspected first by FASTQC. Next, reads were trimmed for adapters using Trimmomatic (42). The resulting fastq files were aligned with Bowtie2 (43) with the following options --local -X 2000. Peaks were called with MACS2 (44) with the following options callpeak --gsize mm --nomodel --shift -100 --extsize 200 --call-summit.

4C-seq data analysis

The normalised interaction frequencies between CREs and *Oct4* promoter measured based on 4C-seq pipeline (37), with some modification. Normalised interaction frequencies between CREs and *Oct4* promoter (Oct-234 used as a view point) was quantified at a higher resolution (1kb resolution window compared to previous analysis used 7kb resolution window). We used

the following command:

perl 4cseqpipe.pl -dopipe -ids 1 -fastq_fn Oct4/fastq/Oct4_234.fastq -convert_qual 1 -calc_from 35620000 -calc_to 35660000 -stat_type median -trend_resolution 1000 -figure_fn Oct4_234_1K.pdf -feat_tab rawdata/Oct4_234_features.txt

The contact frequencies between TREs and *Oct4* promoter was calculated based on the number of contacts between the *Oct4* promoter (Oct4-234 used as a view point) and a 50kb region centred at each TRE, using bedtools intersect command.

Micro-C data analysis

Raw Micro-C data from mouse ESC cells (mESCs) was downloaded from GEO: GSE130275 (38). Samples were processed with a standard juicer pipeline (45), using mm9 mouse genome. Reads from all 37 samples were merged together (after duplicates removal). This resulted in more than 4 billion contacts, allowing the creation of Hi-C file at a very high (200 bp) resolution. For CRE contacts, the chr17 intra-chromosomal contacts map was processed using in house tool to produce normalized observed/expected counts. These counts were used for the creation of heatmap/matrix and barplot. For each CRE (including *Oct4* TSS), all 200bp bins overlapping the loci were used. For TRE contacts, whole-genome 1kb resolution contacts map was created and balanced using our in house C++ code (this is the GW_SCALE normalization available in juicer_tools_1.22.01.jar – *manuscript in preparation*). These counts were used for the creation of the barplot. In this case, *Oct4* promotor was defined as -/+ 3kb of TSS, and TREs were used 200kb bin centered at the particular TRE.

The heatmap was generated using a heatmap.2 R function from gplots package and barplots were generated using a barplot R function.

Data availability

All the high-throughput sequencing data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO: https://www.ncbi.nlm.nih.gov/geo/) under the accession

number of GSE140911. ChIP-seq and RNA-seq data used from (32) GSE113335, and (41) GSE43231.

Acknowledgements

We thank Monash University FACS core facility. We thank Xiaofeng Wang for Illumina HiSeq2500 high-throughput Sequencing at Harvard Medical School; as well as the Genewiz high-throughput sequencing facility, China.

Author contributions

M.C.C., P.T., M.J.B., L.H.W. and P.P.D. performed experiments and analyzed the data. M.C.C., P.T., M.J.B., Y.K., S.J.T., S.L., L.P., S.H.O., and P.P.D. interpreted the data. M.C.C., Y.K., M.O., S.L., and L.P. performed bioinformatics analyses. M.C.C., P.T., S.H.O., and P.P.D. wrote the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by National Health and Medical Research Council (NHMRC) of Australia (APP1159461 and APP1182804) to P.P.D., National Human Genome Research Institute (NHGRI) Career Development Award (R00HG008399) and Genomic Innovator Award (R35HG010717) to L.P., and S.H.O. is an Investigator of the Howard Hughes Medical Institute (HHMI).

Conflict of interests

The authors declare that they have no competing interests.

References

1. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*. **489**, 57–74
2. Cech, T. R., and Steitz, J. A. (2014) The noncoding RNA revolution-trashing old rules to forge new ones. *Cell*. **157**, 77–94
3. Miele, A., and Dekker, J. (2008) Long-range chromosomal interactions and gene regulation. *Mol Biosyst*. **4**, 1046–1057
4. Elkon, R., and Agami, R. (2017) Characterization of noncoding regulatory DNA in the human genome. *Nat. Biotechnol.* **35**, 732–746
5. Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Ren, B., Rubin, E. M., and Pennacchio, L. A. (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*. **457**, 854–858
6. Mouse ENCODE Consortium, Stamatoyannopoulos, J. A., Snyder, M., Hardison, R., Ren, B., Gingeras, T., Gilbert, D. M., Groudine, M., Bender, M., Kaul, R., Canfield, T., Giste, E., Johnson, A., Zhang, M., Balasundaram, G., Byron, R., Roach, V., Sabo, P. J., Sandstrom, R., Stehling, A. S., Thurman, R. E., Weissman, S. M., Cayting, P., Hariharan, M., Lian, J., Cheng, Y., Landt, S. G., Ma, Z., Wold, B. J., Dekker, J., Crawford, G. E., Keller, C. A., Wu, W., Morrissey, C., Kumar, S. A., Mishra, T., Jain, D., Byrska-Bishop, M., Blankenberg, D., Lajoie, B. R., Jain, G., Sanyal, A., Chen, K.-B., Denas, O., Taylor, J., Blobel, G. A., Weiss, M. J., Pimkin, M., Deng, W., Marinov, G. K., Williams, B. A., Fisher-Aylor, K. I., Desalvo, G., Kiralusha, A., Trout, D., Amrhein, H., Mortazavi, A., Edsall, L., McCleary, D., Kuan, S., Shen, Y., Yue, F., Ye, Z., Davis, C. A., Zaleski, C., Jha, S., Xue, C., Dobin, A., Lin, W., Fastuca, M., Wang, H., Guigo, R., Djebali, S., Lagarde, J., Ryba, T., Sasaki, T., Malladi, V. S., Cline, M. S., Kirkup, V. M., Learned, K., Rosenbloom, K. R., Kent, W. J., Feingold, E. A., Good, P. J., Pazin, M., Lowdon, R. F., and Adams, L. B. (2012) An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol.* **13**, 418
7. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T., and Kellis, M. (2015) Integrative analysis of 111 reference human epigenomes. *Nature*. **518**, 317–330
8. Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., and Bernstein, B. E. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*. **473**, 43–49
9. Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutyavin, T., Lajoie, B., Lee, B.-K., Lee,

K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. A., Stamatoyannopoulos, G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E., and Stamatoyannopoulos, J. A. (2012) The accessible chromatin landscape of the human genome. *Nature*. **489**, 75–82

10. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*. 10.1038/nmeth.2688
11. Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., Bates, D., Hansen, R. S., Neph, S., Sabo, P. J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S. R., Kaul, R., and Stamatoyannopoulos, J. A. (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science*. **337**, 1190–1195
12. Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., Bagger, F. O., Jørgensen, M., Andersen, P. R., Bertin, N., Rackham, O., Burroughs, A. M., Baillie, J. K., Ishizu, Y., Shimizu, Y., Furuhata, E., Maeda, S., Negishi, Y., Mungall, C. J., Meehan, T. F., Lassmann, T., Itoh, M., Kawaji, H., Kondo, N., Kawai, J., Lennartsson, A., Daub, C. O., Heutink, P., Hume, D. A., Jensen, T. H., Suzuki, H., Hayashizaki, Y., Müller, F., Forrest, A. R. R., Carninci, P., Rehli, M., and Sandelin, A. (2014) An atlas of active enhancers across human cell types and tissues. *Nature*. **507**, 455–461
13. Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., Shoresh, N., Whitton, H., Ryan, R. J. H., Shishkin, A. A., Hatan, M., Carrasco-Alfonso, M. J., Mayer, D., Luckey, C. J., Patsopoulos, N. A., De Jager, P. L., Kuchroo, V. K., Epstein, C. B., Daly, M. J., Hafler, D. A., and Bernstein, B. E. (2015) Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*. **518**, 337–343
14. Vierstra, J., Rynes, E., Sandstrom, R., Zhang, M., Canfield, T., Hansen, R. S., Stehling-Sun, S., Sabo, P. J., Byron, R., Humbert, R., Thurman, R. E., Johnson, A. K., Vong, S., Lee, K., Bates, D., Neri, F., Diegel, M., Giste, E., Haugen, E., Dunn, D., Wilken, M. S., Josefowicz, S., Samstein, R., Chang, K.-H., Eichler, E. E., De Bruijn, M., Reh, T. A., Skoultchi, A., Rudensky, A., Orkin, S. H., Papayannopoulou, T., Treuting, P. M., Selleri, L., Kaul, R., Groudine, M., Bender, M. A., and Stamatoyannopoulos, J. A. (2014) Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science*. **346**, 1007–1012
15. Stergachis, A. B., Neph, S., Sandstrom, R., Haugen, E., Reynolds, A. P., Zhang, M., Byron, R., Canfield, T., Stelhing-Sun, S., Lee, K., Thurman, R. E., Vong, S., Bates, D., Neri, F., Diegel, M., Giste, E., Dunn, D., Vierstra, J., Hansen, R. S., Johnson, A. K., Sabo, P. J., Wilken, M. S., Reh, T. A., Treuting, P. M., Kaul, R., Groudine, M., Bender, M. A., Borenstein, E., and Stamatoyannopoulos, J. A. (2014) Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature*. **515**, 365–370
16. Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., Park, T. J., Deaville, R., Erichsen, J. T., Jasinska, A. J., Turner, J. M. A., Bertelsen, M. F., Murchison, E. P., Flicek, P., and Odom, D. T. (2015) Enhancer evolution across 20 mammalian species. *Cell*. **160**, 554–566
17. Siepel, A., and Arbiza, L. (2014) Cis-regulatory elements and human evolution. *Current Opinion in Genetics & Development*. **29**, 81–89
18. Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C. G., Kinney, J. B., Kellis, M., Lander, E. S., and Mikkelsen, T. S. (2012) Systematic dissection and optimization of inducible enhancers in human cells

using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271–277
19. Patwardhan, R. P., Hiatt, J. B., Witten, D. M., Kim, M. J., Smith, R. P., May, D., Lee, C., Andrie, J. M., Lee, S.-I., Cooper, G. M., Ahituv, N., Pennacchio, L. A., and Shendure, J. (2012) Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30**, 265–270
20. Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., Norville, J. E., and Church, G. M. (2013) RNA-guided human genome engineering via Cas9. *Science*. **339**, 823–826
21. Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J., and Almendros, C. (2009) Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology (Reading, Engl.)*. **155**, 733–740
22. Shalem, O., Sanjana, N. E., and Zhang, F. (2015) High-throughput functional genomics using CRISPR–Cas9. *Nature Publishing Group*. **16**, 299–311
23. Wang, T., Wei, J. J., Sabatini, D. M., and Lander, E. S. (2014) Genetic screens in human cells using the CRISPR-Cas9 system. *Science*. **343**, 80–84
24. Canver, M. C., Smith, E. C., Sher, F., Pinello, L., Sanjana, N. E., Shalem, O., Chen, D. D., Schupp, P. G., Vinjamur, D. S., Garcia, S. P., Luc, S., Kurita, R., Nakamura, Y., Fujiwara, Y., Maeda, T., Yuan, G.-C., Zhang, F., Orkin, S. H., and Bauer, D. E. (2015) BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature*. **527**, 192–197
25. Korkmaz, G., Lopes, R., Ugalde, A. P., Nevedomskaya, E., Han, R., Myacheva, K., Zwart, W., Elkon, R., and Agami, R. (2016) Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat. Biotechnol.* **34**, 192–198
26. Rajagopal, N., Srinivasan, S., Kooshesh, K., Guo, Y., Edwards, M. D., Banerjee, B., Syed, T., Emons, B. J. M., Gifford, D. K., and Sherwood, R. I. (2016) High-throughput mapping of regulatory DNA. *Nat. Biotechnol.* **34**, 167–174
27. Diao, Y., Li, B., Meng, Z., Jung, I., Lee, A. Y., Dixon, J., Maliskova, L., Guan, K.-L., Shen, Y., and Ren, B. (2016) A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening. *Genome Res.* **26**, 397–405
28. Diao, Y., Fang, R., Li, B., Meng, Z., Yu, J., Qiu, Y., Lin, K. C., Huang, H., Liu, T., Marina, R. J., Jung, I., Shen, Y., Guan, K.-L., and Ren, B. (2017) A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nature Methods*. **14**, 629–635
29. Sanjana, N. E., Wright, J., Zheng, K., Shalem, O., Fontanillas, P., Joung, J., Cheng, C., Regev, A., and Zhang, F. (2016) High-resolution interrogation of functional elements in the noncoding genome. *Science*. **353**, 1545–1549
30. Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I., and Young, R. A. (2013) Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell*. **153**, 307–319
31. Yeom, Y. I., Fuhrmann, G., Ovitt, C. E., Brehm, A., Ohbo, K., Gross, M., Hübner, K., and Schöler, H. R. (1996) Germline regulatory element of Oct-4 specific for the totipotent cycle of embryonal cells. *Development*. **122**, 881–894
32. Seruggia, D., Oti, M., Tripathi, P., Canver, M. C., LeBlanc, L., Di Giammartino, D. C., Bullen, M. J., Nefzger, C. M., Sun, Y. B. Y., Farouni, R., Polo, J. M., Pinello, L., Apostolou, E., Kim, J., Orkin, S. H., and Das, P. P. (2019) TAF5L and TAF6L Maintain Self-Renewal of Embryonic Stem Cells via the MYC Regulatory Network. *Molecular Cell*. 10.1016/j.molcel.2019.03.025
33. Szabó, P. E., Hübner, K., Schöler, H., and Mann, J. R. (2002) Allele-specific expression of imprinted genes in mouse migratory primordial germ cells. *Mech. Dev.* **115**, 157–160
34. Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., Li, Y., Fine, E. J., Wu, X., Shalem, O., Cradick, T. J., Marraffini, L. A., Bao, G., and Zhang, F. (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832
35. Canver, M. C., Lessard, S., Pinello, L., Wu, Y., Ilboudo, Y., Stern, E. N., Needleman, A. J., Galactéros, F., Brugnara, C., Kutlar, A., McKenzie, C., Reid, M., Chen, D. D., Das,

P. P., Cole, M. A., Zeng, J., Kurita, R., Nakamura, Y., Yuan, G.-C., Lettre, G., Bauer, D. E., and Orkin, S. H. (2017) Variant-aware saturating mutagenesis using multiple Cas9 nucleases identifies regulatory elements at trait-associated loci. *Nature Publishing Group*. **49**, 625–634

36. Murakawa, Y., Yoshihara, M., Kawaji, H., Nishikawa, M., Zayed, H., Suzuki, H., Fantom Consortium, and Hayashizaki, Y. (2016) Enhanced Identification of Transcriptional Enhancers Provides Mechanistic Insights into Diseases. *Trends Genet.* **32**, 76–88
37. van de Werken, H. J. G., Landan, G., Holwerda, S. J. B., Hoichman, M., Klous, P., Chachik, R., Splinter, E., Valdes-Quezada, C., Oz, Y., Bouwman, B. A. M., Verstegen, M. J. A. M., de Wit, E., Tanay, A., and de Laat, W. (2012) Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nature Methods*. **9**, 969–972
38. Hsieh, T.-H. S., Cattoglio, C., Slobodyanyuk, E., Hansen, A. S., Rando, O. J., Tjian, R., and Darzacq, X. (2020) Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding. *Molecular Cell*. **78**, 539–553.e8
39. Long, H. K., Prescott, S. L., and Wysocka, J. (2016) Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell*. **167**, 1170–1187
40. Fulco, C. P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S. R., Perez, E. M., Kane, M., Cleary, B., Lander, E. S., and Engreitz, J. M. (2016) Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science*. **354**, 769–773
41. Das, P. P., Shao, Z., Beyaz, S., Apostolou, E., Pinello, L., De Los Angeles, A., O'Brien, K., Atsma, J. M., Fujiwara, Y., Nguyen, M., Ljuboja, D., Guo, G., Woo, A., Yuan, G.-C., Onder, T., Daley, G., Hochedlinger, K., Kim, J., and Orkin, S. H. (2014) Distinct and Combinatorial Functions of Jmjd2b/Kdm4b and Jmjd2c/Kdm4c in Mouse Embryonic Stem Cell Identity. *Molecular Cell*. **53**, 32–48
42. Bolger, A. M., Lohse, M., and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. **30**, 2114–2120
43. Langmead, B., and Salzberg, S. L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*. **9**, 357–359
44. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137
45. Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S. P., Huntley, M. H., Lander, E. S., and Aiden, E. L. (2016) Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst*. **3**, 95–98

Figure legends

Figure. 1 Design of a saturating CRISPR-Cas9 pooled library for ctSCAN-SMS. (a) Genomic tracks show co-occupancy of ESC-TFs (OCT4, NANOG, SOX2), mediator (MED1), enhancer histone marks (H3K27ac, H3K4me1), and RNA Pol2 at a putative TRE of the *Oct4* gene locus in mESCs. ATAC-seq track represents open chromatin regions and enhancers (ENs); RNA-seq track represents gene expression. Highlighted region displays as a putative TRE. All possible sgRNAs (shown with red dashed lines) were designed upstream of PAM sequences (NGG) within the putative TRE. (b) Genomic tracks illustrate co-occupancy of ESC-TFs (OCT4, NANOG, SOX2), mediator (MED1), enhancer histone marks (H3K27ac, H3K4me1) and RNA Pol2 at putative CREs of the *Oct4* gene locus in mESCs. ATAC-seq track characterises open chromatin regions and enhancers; RNA-seq track represents gene expression. sgRNAs (shown with red dashed lines) tilled upstream of PAM sequences (NGG) at the ~12kb surrounding region (-*cis* region) of the *Oct4* locus. (c) Mouse REs CRISPR-Cas9 pooled library distribution. (D) Gaps between adjacent genomic cleavages of NGG PAM sgRNAs targeting CREs and TREs of the *Oct4*.

Figure. 2 Identification and dissection of active CREs of the *Oct4* gene. (a) Dot plot analysis demonstrates the enrichment score of each sgRNAs by comparing their frequency in the GFP-low cells to the pre-sort cells. 16 candidate CREs were identified based on the mean enrichment score (mean log2FC) of sgRNAs per CRE. Four control CREs do not contain any sgRNAs. sgRNAs targeting GFP (green in color) and non-targeting sgRNAs (black in color) were used as positive and negative controls, respectively. (b) A list of identified candidate CREs of the *Oct4* gene. (c) Genomic tracks exhibit open chromatin regions/ENs by ATAC-seq at different time points (0, 8, 24, 96 hr) from undifferentiated to differentiated mESC state; co-occupancy of ESC-TFs (OCT4, NANOG, SOX2), enhancer histone marks (H3K27ac, H3K4me1) are also displayed at the mouse *Oct4* locus. RNA-seq shows the *Oct4* expression. (d) Violin plots outlining the binding dynamics of OCT4, NANOG, SOX2, H3K27ac, and H3K4me1 within the different CREs of *Oct4*. (e) Dynamic changes of open chromatin

regions/ENs measured by ATAC-seq (using 0, 8, 24, 96 hr time points from undifferentiated to differentiated mESC state) within the CREs of *Oct4*. (f) Endogenous Oct4 mRNA expression levels quantified upon deletion of individual CREs of *Oct4.* Oct4 mRNA levels normalized to Gapdh. Data represented as mean ± SEM (n = 3); p-values calculated using ANOVA. *p < 0.05; **p < 0.01; ***p < 0.001; ****p < 0.0001; and ns (non-significant).

Figure. 3 Identification and validation of active TREs of the *Oct4* gene. (a) Dot plot analysis displays the enrichment score of each sgRNAs by comparing their frequency in the GFP-low cells to the pre-sort cells at the selected candidate 12 TREs. (b) A list of identified candidate TREs of the *Oct4* gene. (c) Binding dynamics of OCT4, NANOG, SOX2, H3K27ac, and H3K4me1 represented within the TREs of *Oct4*. (d) ATAC-seq demonstrates the changes in open chromatin regions/ENs from undifferentiated to differentiated mESC state (0, 8, 24 and 96 hr) within the TREs. (e) Quantitative RT-PCR data illustrates the mRNA expression changes of endogenous Oct4 upon deletion of individual TREs of *Oct4*. Oct4 mRNA levels normalized to Gapdh. Data represented as mean ± SEM (n = 3); p-values calculated using ANOVA. *p < 0.05; **p < 0.01; ***p < 0.001; and ns (non-significant).

Figure. 4 Physical interactions between CREs, TREs, and the *Oct4* promoter in *Oct4* gene regulation. (a) 4C-seq data represent normalized interaction frequencies between CREs and the *Oct4* promoter. The interaction/contact frequencies between CREs and the *Oct4* promoter measured at a 1kb resolution window. (b) The interaction/contact frequencies between TREs and the *Oct4* promoter quantified at a 50 kb resolution window. (c) The contact profile of CREs and the *Oct4* promoter at a 1kb resolution window. Bottom triangle is a heat map of normalized contact frequencies between 1kb bins represented with the color codes (the highest is 1 with red; the lowest is 0 with white in color). At the upper part, the black line (within the grey region) represents the normalized median contact frequencies between a locus and the viewpoint. Grey region exhibits 20th-80th percentile of the normalized contact frequencies.

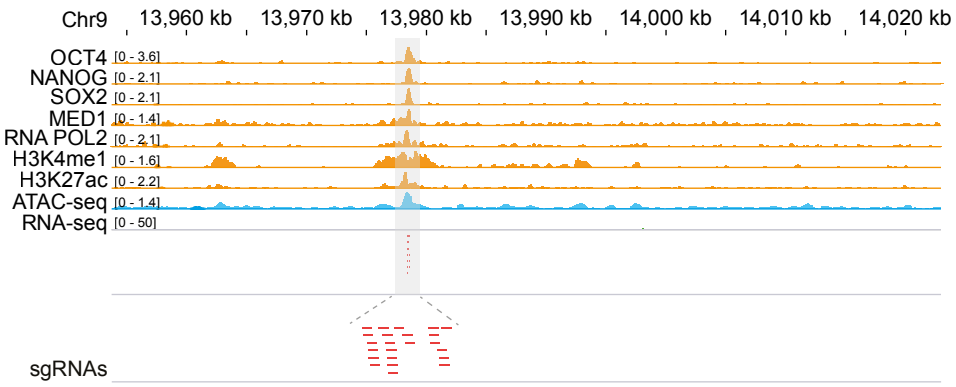
Figure. 5 Conserved active CREs and TREs of the *Oct4* gene. (a) Orthologous sequences from the representative primates (including human) are listed around the ~12kb region of the mouse *Oct4* locus. CREs of the mouse *Oct4* are labelled with solid red bars. PhyloP and PhastCons estimate evolutionary conservation among 60 vertebrates. (b-c) Orthologous sequences from the representative primates (including human) are listed at the active TRE-3 (b) and TRE-4 (c) regions of the mouse *Oct4*. PhyloP and PhastCons estimate evolutionary conservation among 60 vertebrates.

Figure. 6 Model representing the detailed functions of CREs and TREs in *Oct4* gene regulation. The proposed model describes the existence of multiple "active" CREs (red in color) and TREs (blue in color) of the *Oct4* locus in mESCs. However, not all of the active REs have regulatory features (i.e. co-occupancy of ESC-TFs (OCT4, NANOG, SOX2), active enhancer histone marks (H3K27ac, H3K4me1), and open chromatin regions), which are termed as unmarked REs (UREs). Also, a subset of active REs physically interacts with the *Oct4* promoter through intra-chromosomal (for CREs) and inter-chromosomal (for TREs) interactions. Taken together, it suggests that active REs act beyond their regulatory features and physically contact with the *Oct4* promoter to control the transcriptional output of *Oct4* gene.

Fig. 1

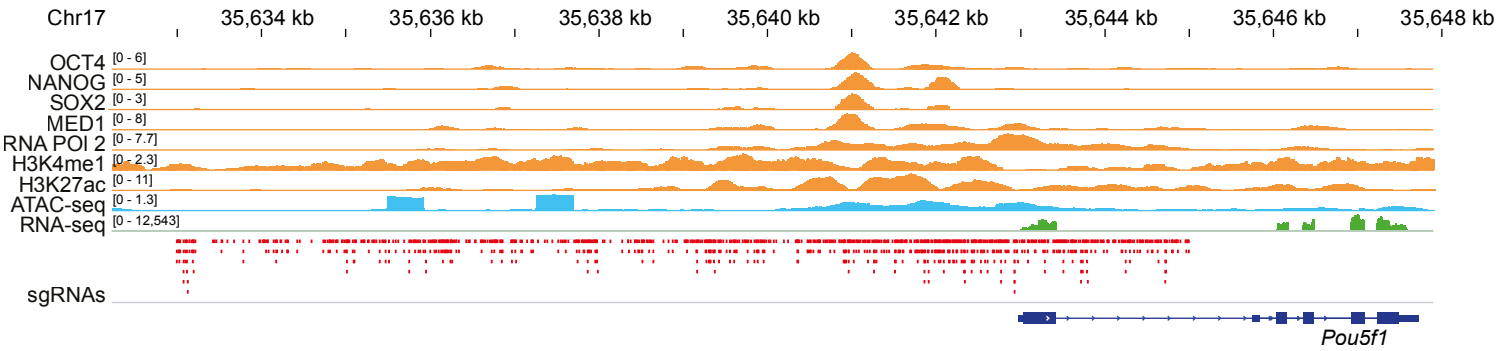
a

A putative TRE of the *Oct4* gene locus:



b

Putative CREs of the *Oct4* gene locus:



c

Mouse regulatory elements (REs) CRISPR-Cas9 pooled library

List of sgRNAs	No. of sgRNAs (%)
EN and SE TREs of *Oct4*	70,480 (94.50%)
ESC-specific TFs (positive controls)	150 (0.20%)
GFP (positive controls)	119 (0.16%)
CREs of *Oct4*	1,827 (2.44%)
Non-targeting (negative controls)	2,000 (2.68%)
Total	74,576

d

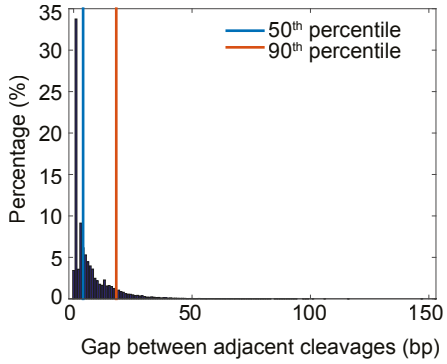
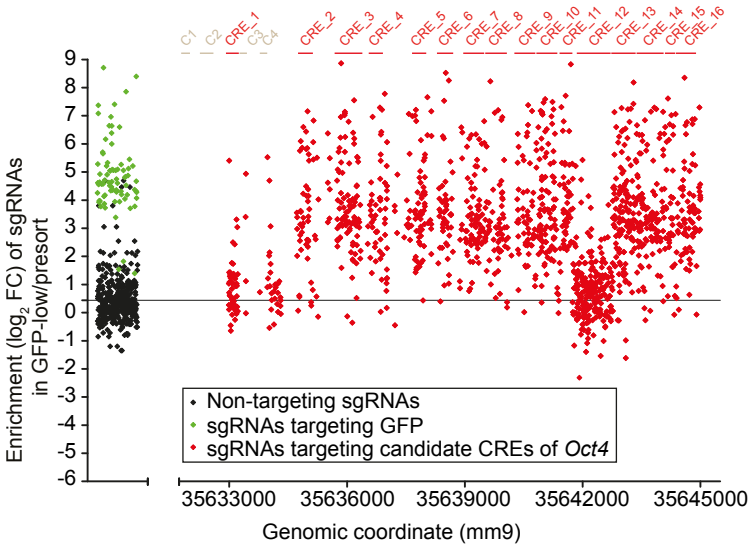


Fig. 2

a



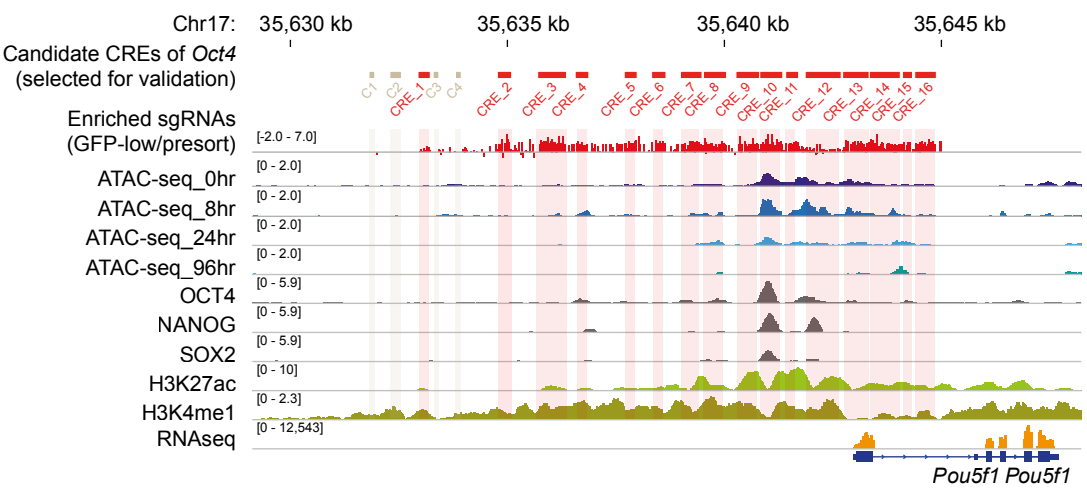
b

List of candidate CREs of the *Oct4* gene

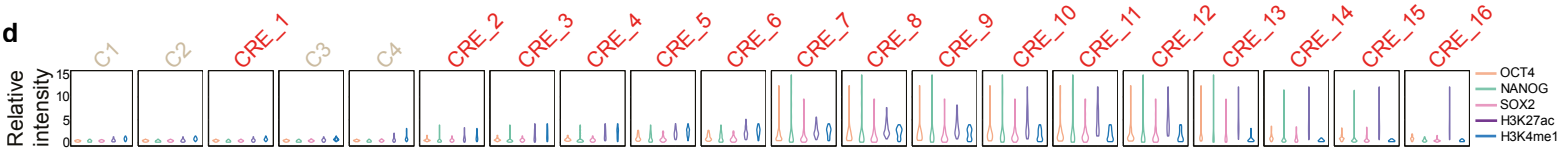
Based on mean enrichment scores (mean log2FC) of sgRNAs (in GFP-low/pre-sort) per CRE. Candidate CREs have mean log2FC >0.5, P-values <0.0001. All the HMM sgRNAs per CRE have off-target score >10.

Candidate CREs	Genomic locations of REs (mm9) (enriched sgRNAs)	Length (bp)	No. of sgRNAs
CRE_1	chr17:35,632,944-35,633,217	273	51
CRE_2	chr17:35,634,773-35,635,108	335	40
CRE_3	chr17:35,635,706-35,636,370	664	98
CRE_4	chr17:35,636,568-35,636,884	316	35
CRE_5	chr17:35,637,687-35,637,998	311	53
CRE_6	chr17:35,638,316-35,638,663	347	45
CRE_7	chr17:35,638,981-35,639,479	498	73
CRE_8	chr17:35,639,516-35,640,048	532	60
CRE_9	chr17:35,640,274-35,640,811	537	56
CRE_10	chr17:35,640,811-35,641,342	531	97
CRE_11	chr17:35,641,411-35,641,715	304	50
CRE_12	chr17:35,641,862-35,642,696	834	187
CRE_13	chr17:35,642,726-35,643,342	616	140
CRE_14	chr17:35,643,343-35,644,057	714	113
CRE_15	chr17:35,644,097-35,644,345	248	35
CRE_16	chr17:35,644,380-35,644,880	500	70
Control CRE_1 (C1)	chr17:35,631,818-35,631,951	133	0
Control CRE_2 (C2)	chr17:35,632,289-35,632,557	268	0
Control CRE_3 (C3)	chr17:35,633,285-35,633,418	133	0
ControlCRE_4 (C4)	chr17:35,633,801-35,633,934	133	0
sgRNAs targeting GFP (positive controls)	NA	NA	118
Non-targeting sgRNAs (negative controls)	NA	NA	1984

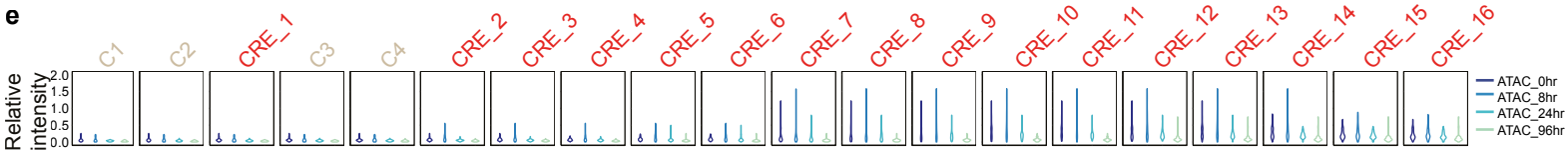
c



d



e



f

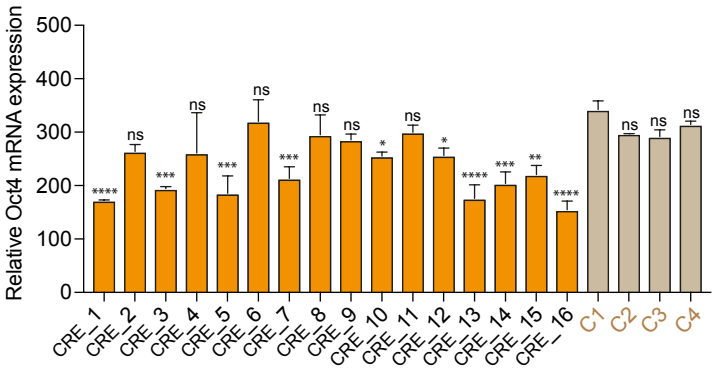
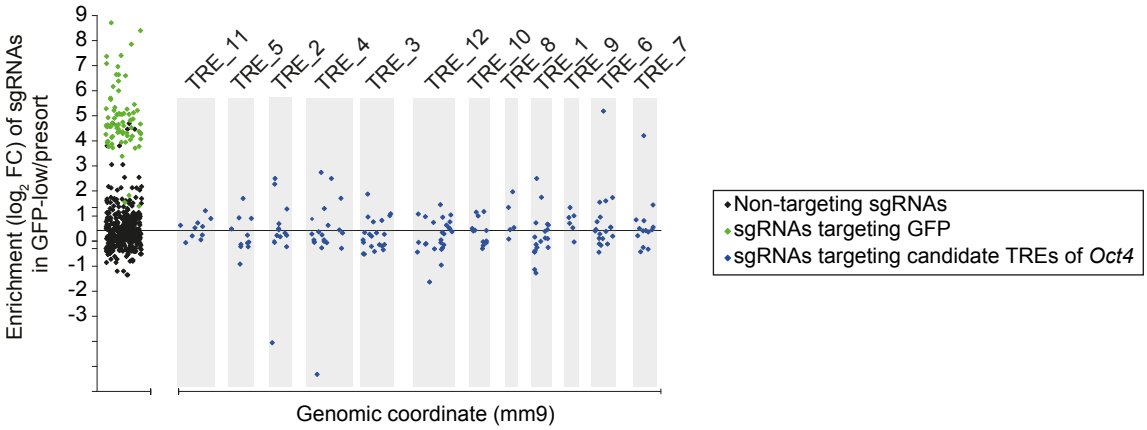


Fig. 3

a



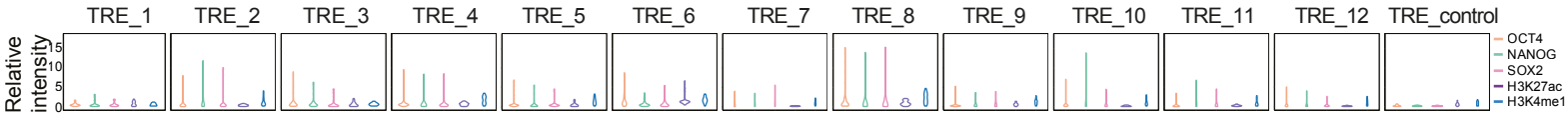
b

List of candidate TREs of the *Oct4* gene

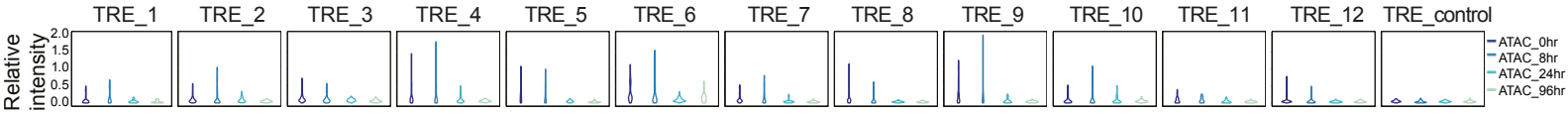
Based on HMM based enrichment scores of sgRNAs.
All the sgRNAs must have off-target score >10.
TREs must possess at least 4 sgRNAs with mean enrichment scores (mean log2FC) >0.5, P-values <0.0001.
TREs should co-occupy with ESC-TFs (OCT4, NANOG, SOX2), enhancer histone marks (H3K27ac, H3K4me1) and dynamic open chromatin regions.

Candidate TREs	Genomic locations	Length (bp)	No. of sgRNAs
TRE_1	chr14:59,989,540-59,989,621	81	20
TRE_2	chr4:71,807,347-71,807,416	69	12
TRE_3	chr9:74,696,403-74,696,575	172	21
TRE_4	chr9:13,978,361-13,978,560	199	20
TRE_5	chr4:13,023,503-13,023,629	126	11
TRE_6	chr19:53,534,741-53,534,849	108	19
TRE_7	chrX:130,208,023-130,208,140	117	14
TRE_8	chr12:42,828,460-42,828,512	52	5
TRE_9	chr18:82,062,626-82,062,657	31	6
TRE_10	chr11:63,347,393-63,347,468	75	11
TRE_11	chr1:7,722,856-7,723,033	177	10
TRE_12	chr10:7,365,884-7,366,087	203	25
TRE_control	chr1:120,957,901-120,958,247	346	0
sgRNAs_GFP (positive controls)	NA	NA	118
sgRNAs_non-targeting (negative controls)	NA	NA	1984

c



d



e

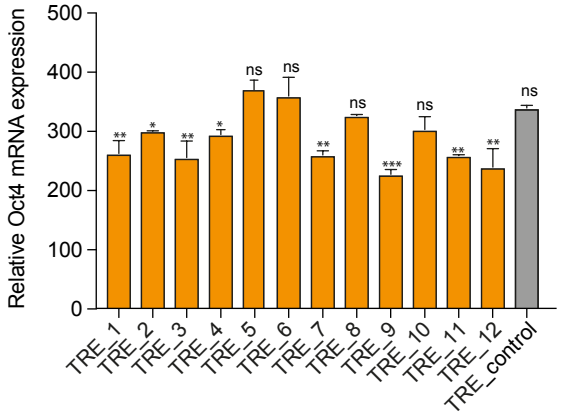


Fig. 4

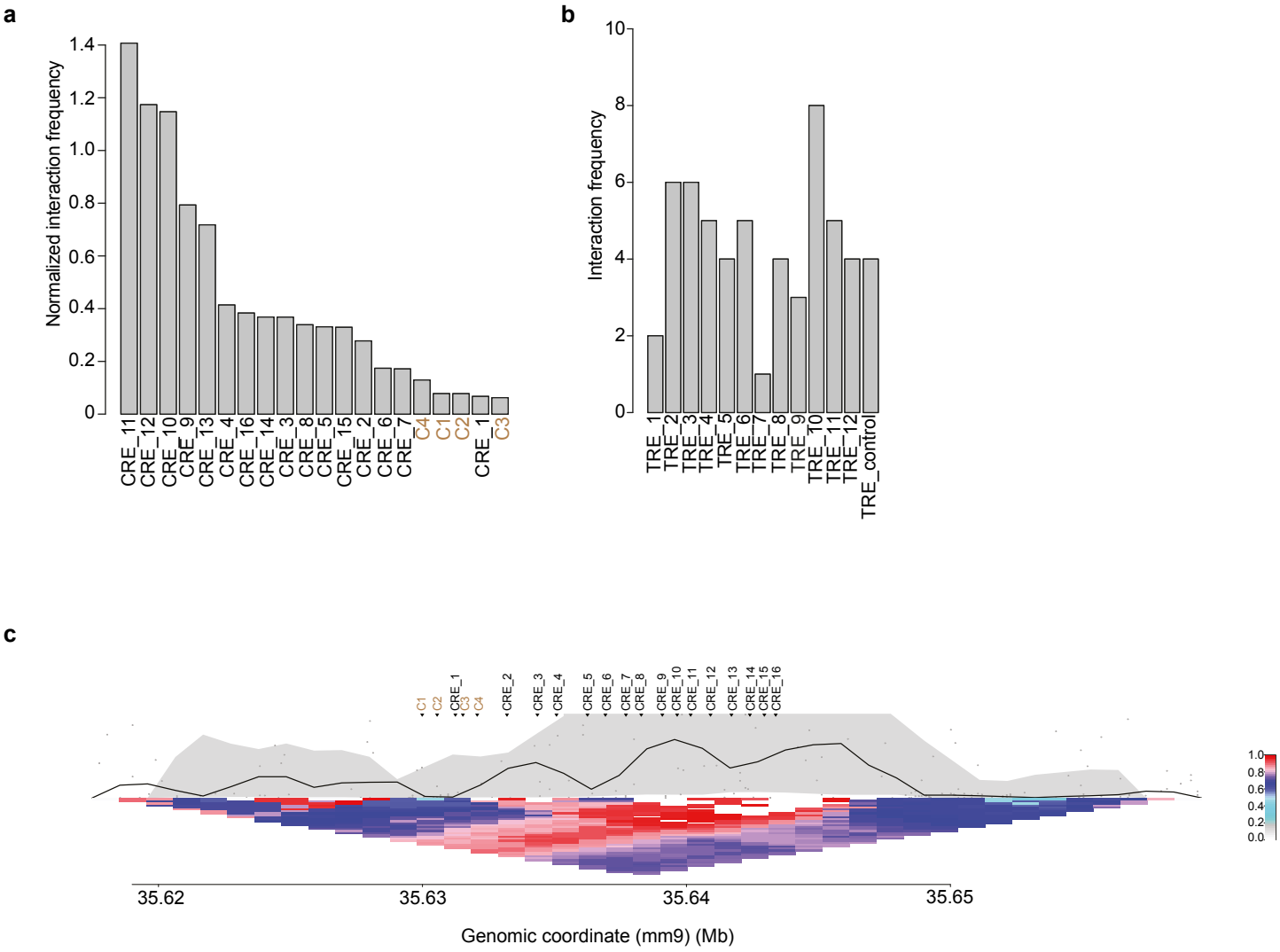
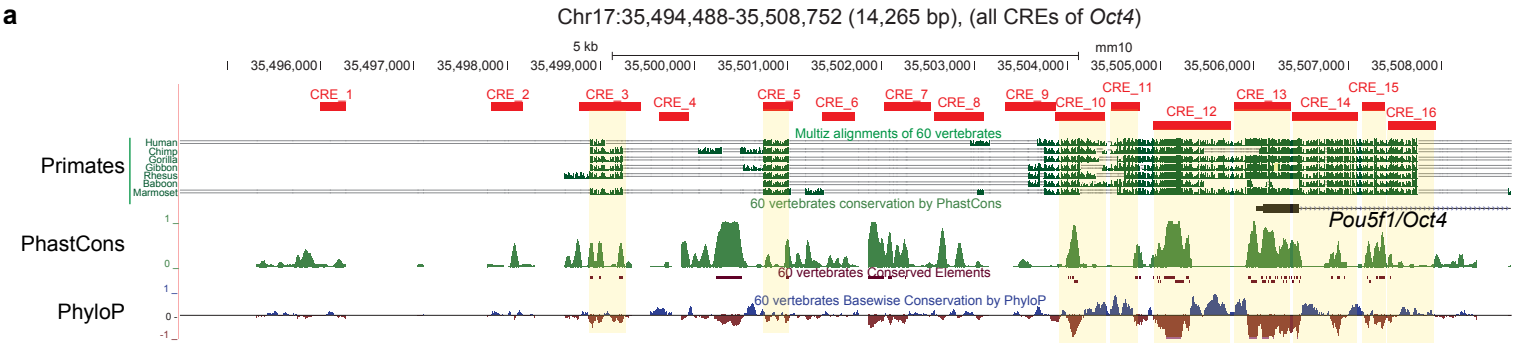
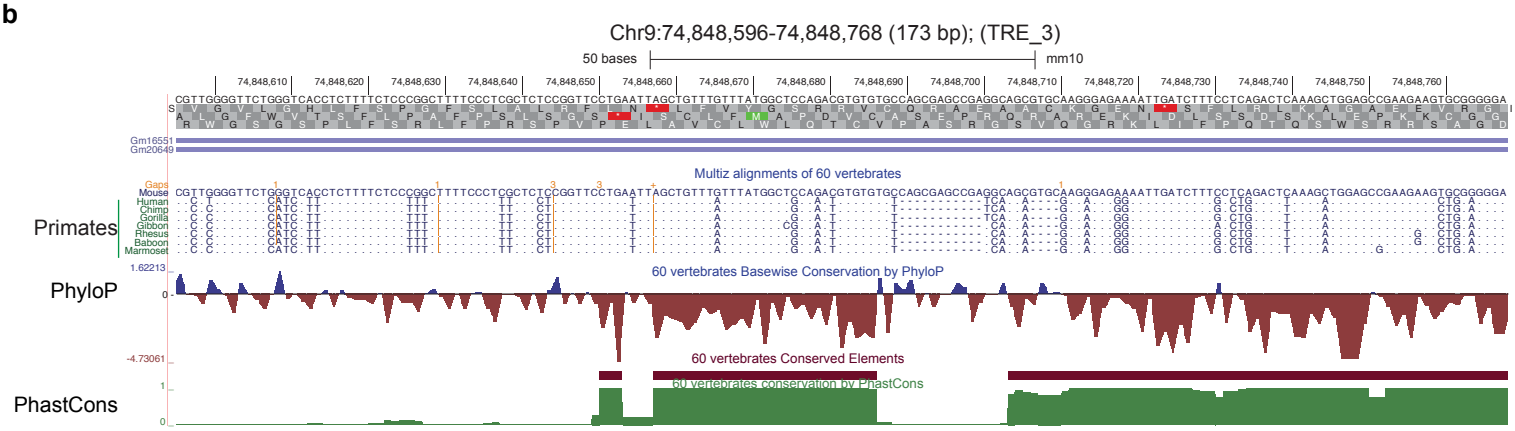

a
Normalized interaction frequency
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0
CRE_11
CRE_12
CRE_10
CRE_9
CRE_13
CRE_4
CRE_16
CRE_14
CRE_3
CRE_8
CRE_5
CRE_15
CRE_2
CRE_6
CRE_7
C4
C1
C2
CRE_1
C3
b
Interaction frequency
10
8
6
4
2
0
TRE_1
TRE_2
TRE_3
TRE_4
TRE_5
TRE_6
TRE_7
TRE_8
TRE_9
TRE_10
TRE_11
TRE_12
TRE_control
c
C1
C2
CRE_1
C3
C4
CRE_2
CRE_3
CRE_4
CRE_5
CRE_6
CRE_7
CRE_8
CRE_9
CRE_10
CRE_11
CRE_12
CRE_13
CRE_14
CRE_15
CRE_16
35.62
35.63
35.64
35.65
Genomic coordinate (mm9) (Mb)

Fig. 5

a


Chr17:35,494,488-35,508,752 (14,265 bp), (all CREs of *Oct4*)
5 kb
mm10
CRE_1
CRE_2
CRE_3
CRE_4
CRE_5
CRE_6
CRE_7
CRE_8
CRE_9
CRE_10
CRE_11
CRE_12
CRE_13
CRE_14
CRE_15
CRE_16
Multiz alignments of 60 vertebrates
Primates
Human
Chimp
Gorilla
Gibbon
Rhesus
Baboon
Marmoset
60 vertebrates conservation by PhastCons
PhastCons
Pou5f1/Oct4
60 vertebrates Conserved Elements
PhyloP
60 vertebrates Basewise Conservation by PhyloP

b


Chr9:74,848,596-74,848,768 (173 bp); (TRE_3)
50 bases
mm10
Gm16551
Gm20649
Multiz alignments of 60 vertebrates
Primates
Mouse
Human
Chimp
Gorilla
Gibbon
Rhesus
Baboon
Marmoset
PhyloP
60 vertebrates Basewise Conservation by PhyloP
60 vertebrates Conserved Elements
PhastCons
60 vertebrates conservation by PhastCons

c

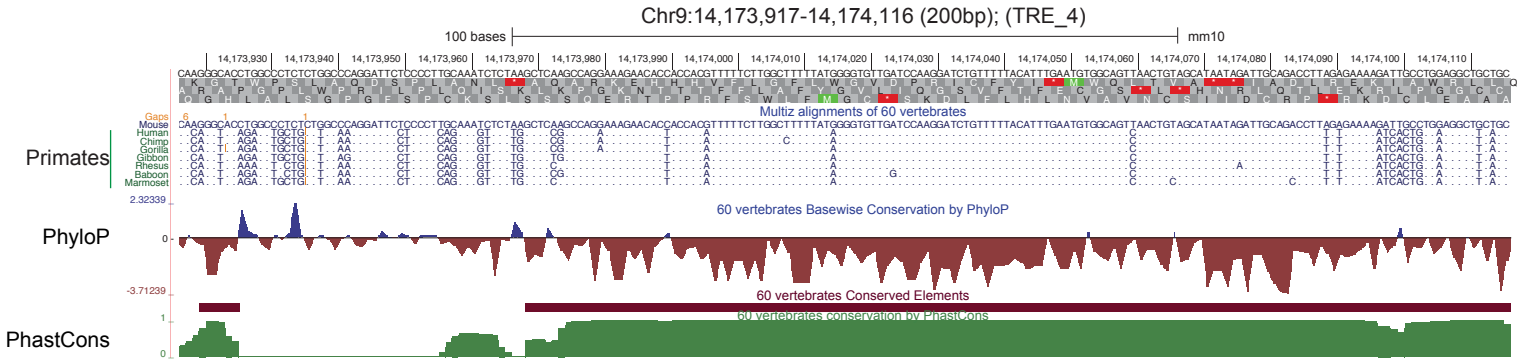
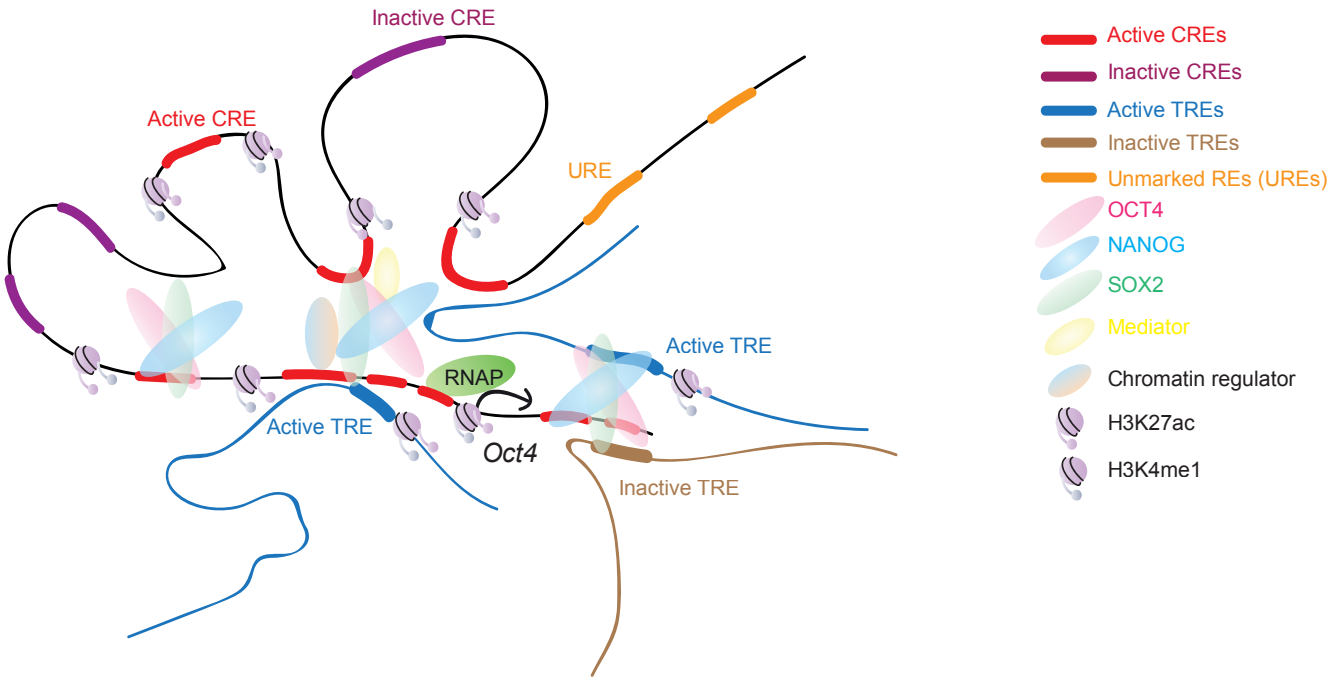

Chr9:14,173,917-14,174,116 (200bp); (TRE_4)
100 bases
mm10
Multiz alignments of 60 vertebrates
Primates
Mouse
Human
Chimp
Gorilla
Gibbon
Rhesus
Baboon
Marmoset
PhyloP
60 vertebrates Basewise Conservation by PhyloP
60 vertebrates Conserved Elements
PhastCons
60 vertebrates conservation by PhastCons

Fig. 6


Inactive CRE
Active CRE
URE
Active TRE
RNAP
Active TRE
Oct4
Inactive TRE
Active CREs
Inactive CREs
Active TREs
Inactive TREs
Unmarked REs (UREs)
OCT4
NANOG
SOX2
Mediator
Chromatin regulator
H3K27ac
H3K4me1