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

Matthew C. Canver, [1,12] Pratibha Tripathi, [2,3,12] Michael J. Bullen, [2,3,12] Yogesh Kumar, [2,3] Moshe Olshansky, [4,5] Stephen J. Turner, [4,6] Samuel Lessard, [7,8,11] Luca Pinello, [9] Stuart H. Orkin, [1,10] 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 Microbiology and Immunology, University of Melbourne, Parkville, Victoria 3010, Australia

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

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

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

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

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

[12]These authors contributed equally to this work

*Corresponding authors:
stuart_orkin@dfci.harvard.edu
partha.das@monash.edu

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 of the *Oct4* gene locus, as well as genome-wide 2,613 high-confidence *trans*-REs (TREs), in mESCs. The ctSCAN-SMS identified 10 CREs and 12 TREs, as novel candidate REs of the *Oct4* gene in mESCs. Furthermore, deletion of these REs confirmed that the majority of the CREs and TREs are functionally active, and involved in regulating *Oct4* gene expression. Additionally, a subset of the functional CREs and TREs physically interact with the *Oct4* promoter to varying degrees through intra- and inter-chromosomal interactions, respectively. Comparative genomics analysis reveals that functional CREs are more conserved in terms of their regulatory sequence conservation between mouse and primates (including humans) than TREs. Notably, a few active CREs are devoid of canonical regulatory features. Taken together, our work demonstrates the reliability and robustness of ctSCAN-SMS screening to identify critical REs, and probe their roles in the regulation of transcriptional output of a target gene (in this case *Oct4)*.

Introduction

Large-scale genomic studies reveal that ~80% of the human genome may be involved in gene regulation, whereas only ~2% of the genome codes for proteins (ENCODE Project Consortium 2012). The functional non-coding genome can be broadly divided into regulatory elements (REs) and regions that encode non-coding RNAs (ncRNAs) (ENCODE Project Consortium 2012; Cech and Steitz 2014). 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(s) on the same chromosome, whereas TREs are located distally relative to their target genes on different chromosomes (Miele and Dekker 2008; Elkon and Agami 2017). Putative REs have been identified using various methods, including transcription factor binding, particular enhancer histone marks, DNA accessibility (open chromatin regions), enhancer-promoter interactions, and gene expression (Visel et al. 2009; ENCODE Project Consortium 2012; Mouse ENCODE Consortium et al. 2012; Roadmap Epigenomics Consortium et al. 2015; Ernst et al. 2011; Thurman et al. 2012; Buenrostro et al. 2013). REs enriched for sequence variants are associated with diverse human traits and diseases (Maurano et al. 2012; Andersson et al. 2014; Farh et al. 2015). In addition, REs play crucial roles in evolutionary turnover and divergence (Vierstra et al. 2014; Stergachis et al. 2014; Villar et al. 2015; Siepel and Arbiza 2014).

Initial efforts have systematically evaluated RE function using reporter assays on a massive scale (Melnikov et al. 2012; Patwardhan et al. 2012); such approaches fail to interrogate REs within their native genomic contexts. Advances in CRISPR-Cas9 mediated genome editing technology (Mali et al. 2013; Mojica et al. 2009) transform the ability to examine protein-coding genes (Shalem et al. 2015; Wang et al. 2014), as well as REs *in situ* in chromatin. High-throughput CRISPR-Cas9 mediated functional genetic screens have been performed to characterize the CREs in mammalian cells (Canver et al. 2015; Korkmaz et al. 2016; Rajagopal et al. 2016; Diao et al. 2016; 2017; Sanjana et al. 2016). Prior screens to identify functional CREs focused on targeting putative CREs of gene(s) of interest (gene-

centric) or on targeting putative CREs bound by selected TFs (TF-centric). However, 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, ESSRB), mediators (MED1), enhancer histone marks (H3K4me1, H3K27ac) and DNase I hypersensitivity (Whyte et al. 2013). SEs contain multiple ENs; SEs are also more densely co-occupied with TFs, enhancer histone marks, and chromatin regulators as compared to ENs, with a higher magnitude of transcriptional output (Whyte et al. 2013). We undertook a high-throughput CRISPR-Cas9 mediated genome editing approach to target all putative REs. First, we generated a genome-wide map of open chromatin regions using ATAC-seq in mESCs, as ATAC-seq identifies most EN REs (Buenrostro et al. 2013). ATAC-seq peaks were then overlapped with all putative ENs (8,563) and SEs (231) to designate high-confidence REs (2,613) (Fig. 1A; Supplemental S1A; Supplemental Table 1). 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 TREs (Fig.1A; Supplemental Fig. S1A; Supplemental Table 1, 2). This analysis created 70,480 sgRNAs with a median gap of 5 bp between adjacent genomic cleavages (Fig. 1C, 1D). Similarly, 1,827

sgRNAs were created at the surrounding ~12kb (-10kb to +2kb of TSS of the *Oct4*) region of the mouse *Oct4* gene locus to systematically dissect the *cis*-REs (CREs) of *Oct4* (Fig. 1B, 1C). 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 together, cloned into a lentiviral vector, and deep sequenced. The deep sequencing result represents >95% sgRNAs that target TREs, >99% sgRNAs that target CREs and control sgRNAs in the pooled library (Fig.1C; Supplemental Fig.S1B-S1F; Supplemental Table 2).

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 (Yeom et al. 1996; Seruggia et al. 2019) (Supplemental Fig. S2B). 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 (Supplemental Fig. S2A). After drug selection, "GFP-low" cells were sorted using fluorescence-activated cell sorting (FACS) (Supplemental Fig. S2A, S2C). 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 cell population (Supplemental Fig. S2A). The screen was performed in triplicate.

We calculated an "enrichment score" of each sgRNA by comparing its frequency in GFP-low over pre-sort cells. The enrichment scores were built based on the two best replicates (Supplemental Table 2). As expected, highest and lowest enrichment scores were obtained from GFP-targeting sgRNAs (mean log2FC 4.87 ± 1.16, $P<0.0001$) and NT-sgRNAs (mean log2FC 0.44 ± 0.74, $P<0.0001$) respectively, indicating that the screen was technically

successful (Fig. 2A, 3A). We ranked all sgRNAs based on their enrichment scores (Supplemental Table 2), and analyzed their off-target scores (ranged between 0-100) (Hsu et al. 2013) (Supplemental Fig. S3A, S4A; Supplemental Table 2). A higher off-target score signifies fewer off-targets for a particular sgRNA. We found that the majority of the evaluated sgRNAs (87.6% sgRNAs for CREs, and 84.5% sgRNAs for TREs) have off-target scores >10 (Supplemental Table 2).

To identify candidate CREs of the *Oct4*, we considered all 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, P<0.0001, which was higher than the mean enrichment score of NT-sgRNAs (mean log2FC 0.44 ± 0.74, P<0.0001) (Fig. 2A, 2B). Among 16 CREs, CREs-10 and 12 have been recognized previously as distal and proximal enhancers, respectively (Yeom et al. 1996); CREs-13 to 16 were present within the promoter region of *Oct4* (+/-2kb of TSS) (Yeom et al. 1996). The remaining 10 CREs were newly identified candidate CREs of the *Oct4* gene (Fig. 2A, 2B).

To classify the candidate TREs, we applied a Hidden Markov Model (HMM) to the sgRNAs enrichment scores (Canver et al. 2017), which initially identified 263 candidate TREs. Furthermore, we applied stringent criteria to select candidate TREs for validation, as follows: i) TREs must have sgRNAs with off-target scores>10 (Fig. 3A; Supplemental Fig. S4A); ii) TREs must possess at least 4 sgRNAs with mean enrichment scores (mean log2FC) >0.5, P<0.0001 (Fig. 3B); iii) TREs should co-occupy with ESC-TFs (OCT4, NANOG, SOX2), enhancer histone marks (H3K27ac, H3K4me1) (Fig. 3C; Supplemental Fig. S4B); and iv) they must contain “dynamic” open chromatin regions; i.e. open chromatin regions present at the undifferentiated state (0 hr) but gradually become closed with the progression of differentiation (8, 24, 96hr) of mESCs (Fig. 3D; Supplemental Fig. S4C). Based on these criteria, we selected 12 potential candidate TREs of the *Oct4* gene (Fig. 3A, 3B).

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

We selected a total of 33 REs, including 20 CREs – 16 CREs and 4 control CREs (CREs without any sgRNAs), and 13 TREs – 12 TREs and one control TRE (without any sgRNAs) of the *Oct4* gene locus for validation. Paired sgRNAs (5' and 3' sgRNAs with mCherry) were used to target the flanking ends of each selected candidate RE to create a deletion. The paired sgRNAs tagged with mCherry were transfected to the wild-type mESCs; mCherry-positive cells were sorted and endogenous Oct4 mRNA expression levels were measured (Fig. 2F, 3E). We observed significant reduction in Oct4 expression to different extents upon deletion of CREs- 1, 3, 5, 7, 10, 12, and 13-16 (Fig. 2F). Deletions of newly identified CREs- 1, 3, 5 and 7 showed greater reduction in Oct4 expression, compared to deletions of known distal and proximal enhancers (CREs- 10 and 12) of *Oct4* (Fig. 2F). However, co-occupancy of ESC-TFs (OCT4, NANOG, SOX2– ONS), enhancer histone marks (H3K27ac, H3K4me1) and dynamic open chromatin regions (ATAC-seq peaks at 0, 8 hr compared to 24, 96hr) were more prominent at CREs- 7, 10, 12 compared to CREs- 1, 3, 5 (Fig. 2C-2E; Supplemental Fig. S3B, S3C). Moreover, deletion of CREs- 13 to 16 (present at the promoter region of *Oct4*) showed 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; compared to other CREs present at the *Oct4* promoter (Fig. 2C-2E; Supplemental Fig. S3B, S3C). In contrast, deletion of control CREs (Control CREs- 1 to 4) displayed no significant changes in Oct4 expression (Fig. 2F); moreover, they exhibited low-level of co-occupancy of ONS, H3K27ac, H3K4me1, and without any dynamic open chromatin regions (Fig. 2D, 2E; Supplemental Fig. S3B, S3C). These data confirm the existence of "multiple" active CREs (including newly identified CREs) of the *Oct4*. Yet some active CREs fail to display canonical regulatory features (i.e. without any co-occupancy of TFs, enhancer histone marks, and open chromatin regions) as described recently (Diao et al. 2017; Rajagopal et al. 2016).

Deletion of several TREs (TREs- 1, 2, 3, 4, 7, 9, 11, 12) exhibited a substantial reduction in Oct4 expression to various extents (Fig. 3E). Nonetheless, all TREs revealed co-

occupancy with ONS, H3K27ac and H3K4me1 marks, as well as with dynamic open chromatin regions (Fig. 3C, 3D; Supplemental Fig. S4B, S4C). This is an agreement with all the candidate TREs (TREs 1-12) that were short-listed for validation based on their regulatory features. Conversely, control TRE did not contain any regulatory features (Fig. 3C, 3D); and deletion of the control TRE did not affect Oct4 expression (Fig. 3E). Interestingly, we observed that the majority of the neighbouring genes of validated TREs were lowly expressed in mESCs (Supplemental Fig. S4D). This suggests that – i) the active TREs are not the actual REs of their neighbouring genes; and/or ii) these genes are unrelated to the mESC state.

Taken together, these data validate the function of a subset of the newly identified candidate CREs and TREs of the *Oct4* gene. Also, our findings support the reliability and robustness of the ctSCAN-SMS screen to identify critical REs of the *Oct4* in a high-throughput manner.

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

REs (particularly enhancers) physically interact with the promoter of a gene, and control transcription (Murakawa et al. 2016). Several 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) in order to evaluate the significance of the REs (Elkon and Agami 2017; Mumbach et al. 2016). 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 (van de Werken et al. 2012). We used Oct4-234 (a region at 1.5kb upstream of TSS of *Oct4*) as a viewpoint, as previously (Supplemental Fig. S5A); calculated contact frequencies between the viewpoint and CREs (using 1kb resolution window, surrounding 30kb region of the *Oct4* gene locus) (Fig. 4A, 4C), as well as contact frequencies between the viewpoint and TREs (using 50kb resolution window, surrounding each of the TRE) (Fig. 4B). This analysis revealed

ranges of contact frequencies between functionally validated active CREs/TREs and the *Oct4* promoter. For example, newly identified CREs- 3, 5, 7, as well as CREs- 10 and 12 (known distal and proximal enhancers of *Oct4*), and CREs- 13 to16 (residing at the promoter region of *Oct4*) demonstrated significant intra-chromosomal interactions with the *Oct4* promoter (Fig. 4A, 4C). In contrast, we failed to detect significant interactions of CRE-1 with the *Oct4* promoter, similar to the four control CREs (Fig. 4A, 4C). Likewise, validated active TREs- 1, 2, 3, 4, 7, 9, 11, 12 also exhibited a degree of inter-chromosomal interactions with the *Oct4* promoter (Fig. 4B). Taken together, our data suggest that active CREs and TREs physically interact with the *Oct4* promoter to different extent as they influence *Oct4* expression.

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 (Villar et al. 2015; Long et al. 2016). Also, evidence implies that loss and gain of REs (called turnover) takes place during evolution (Siepel and Arbiza 2014). To understand the importance of validated active CREs and TREs of mouse *Oct4* in evolutionary turnover, we analyzed their regulatory sequence conservation along with evolved species, such as primates including human. 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 (Fig. 5A); whereas active CREs- 1, 7 did not show appreciable sequence conservation with primates (Fig. 5A). In comparison to CREs, only a few active TREs (TREs- 3 and 4) showed significant sequence conservation between mouse and primates, including human (Fig. 5B, 5C).

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 (Diao et al. 2017). Surprisingly, we found that these human CREs are evolutionary conserved at the upstream regions of the mouse *Oct4* locus as well (Supplemental Fig. S6A). Taken together, our comprehensive comparative genomic analysis shows that several

functional CREs and TREs (but not all REs) of *Oct4* are well-conserved between mouse and primates including human. These data support the occurrence of REs turnover/divergence (gain and loss of REs) and its activity at the *Oct4* locus during evolution, which may be critical for positive selection as proposed earlier (Siepel and Arbiza 2014; Long et al. 2016).

Discussion

A handful of small to large 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 (Canver et al. 2015; Korkmaz et al. 2016; Rajagopal et al. 2016; Sanjana et al. 2016; Fulco et al. 2016). All these screens were successful in identifying 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 show “transient” enhancer regulatory activity in *OCT4* gene expression (Diao et al. 2016). 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 was applied to design 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 (Diao et al. 2017). 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 identified mostly 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 *Oct4* gene, as potential candidate REs (Fig. 2A, 3A). Deletion studies confirmed that the

majority of these CREs and TREs are functionally “active” for controlling *Oct4* expression; however, CREs are more active than TREs (Fig. 2F, 3E). In addition, we showed that a subset of active CREs and TREs physically interacts with the *Oct4* promoter to different extents through intra- and inter-chromosomal interactions, respectively (Fig. 4A, 4B); as described previously in other gene regulatory contexts (Miele and Dekker 2008). Nonetheless, “enhancer activity” of REs is not directly correlated to their physical contact intensity with the *Oct4* promoter (Fig. 2F, 3E, 4A-4C). Interestingly, we found a few active CREs (CREs- 1, 3) that lack typical regulatory features (Fig. 2D-2F). These observations support earlier studies that identified unmarked REs (UREs) with no typical regulatory features, yet play critical roles in transcriptional output (Rajagopal et al. 2016; Diao et al. 2017). Moreover, comparative genomics analysis revealed that several active CREs and a few TREs of *Oct4* are evolutionarily conserved (regarding their regulatory sequences) between mouse and primates, including human (Fig. 5). Though, we observed divergence of *Oct4* REs among mouse and primates (including human), which may account for the vital roles of RE turnover during evolution for positive selection (Miele and Dekker 2008; Long et al. 2016).

Several studies demonstrate that “multiple” REs act either in a co-operative or competitive fashion to control transcriptional output (Long et al. 2016). 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, 3E). Further systematic studies will be required to elucidate how multiple REs function combinatorially to control the transcriptional output of the *Oct4* locus. In conclusion, we have demonstrated the utility of ctSCAN-SMS as an approach to identify functional CREs and TREs of a gene locus, and dissect their regulatory contributions to the transcriptional output of a target within its normal chromosomal setting.

Methods

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 (Seruggia et al. 2019; Das et al. 2014). mESCs were cultured at 37°C, 5% CO_2.

Differentiation of mouse embryonic stem cells (mESCs)

ZHBTc4 mESCs (Niwa et al. 2000) were cultured in mESC media with all the supplements, including LIF. This was able to maintain the undifferentiated mESC state. Upon addition of doxycycline (2µg/ml) with mESC media+LIF in ZHBTc4 mESCs, they lead to decrease in Oct4 expression level and undifferentiated mESCs facilitate towards the differentiation (Whyte et al. 2012). For ATAC-seq experiments, ZHBTc4 mESCs were treated with mESC media containing LIF and doxycycline, and the cells were collected after 0, 8, 24, 96 hr. These cells were washed with 1X PBS, counted, and proceeded to ATAC-seq library preparation.

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 (Whyte et al. 2013). First, we generated ATAC-seq (Buenrostro et al. 2013) 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 (Seruggia et al. 2019) 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 (Seruggia et al. 2019). 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-puro plasmid that carries Cas9 and mCherry, using Golden Gate Cloning approach as previously described (Seruggia et al. 2019). 100,000 wild-type (J1) mESCs were transfected with 500ng of each 5' and 3' sgRNA-Cas9-mCherry-puro plasmids using Lipofectamine 2000 (Thermo Fisher Scientific). After 24hrs of transfection, 10µg/ml blasticidin (InvivoGen) and 1µg/ml puromycin (Sigma-Aldrich) drugs were added to select the cells for the next 3-4 days. Next, drug-resistant cells were used to sort the mCherry-positive cells; at least 50,000 mCherry-positive cells were collected to isolate the total RNA and measure the Oct4 mRNA expression levels by quantitative RT-PCR.

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 (Buenrostro et al. 2013), with some modifications. We used J1 wild-type mESCs; and ZHBTc4 mESCs from 0, 8, 24, 96hr after the doxycycline treatment. 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 (Bolger et al. 2014). The resulting fastq files were aligned with Bowtie2 (Langmead and Salzberg 2012) with the following options --local -X 2000. Peaks were called with MACS2 (Zhang et al. 2008) 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 (van de Werken et al. 2012), 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.

Data access

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 GSE (pending during the submission time). ChIP-seq and RNA-seq data used from (Seruggia et al. 2019) GSE113335, and (Das et al. 2014) 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. P.P.D. is supported by National Health and Medical Research Council (NHMRC) of Australia (GNT1159461). L.P. is supported by the National Human Genome Research Institute (NHGRI) Career Development Award (R00HG008399), and Genomic Innovator Award (R35HG010717). S.H.O is an Investigator of the Howard Hughes Medical Institute (HHMI). The authors declare no conflicts of interest.

Author contributions: M.C.C., P.T., M.J.B., 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.

References

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* **507**: 455–461.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* 1–8.

Canver MC, Lessard S, Pinello L, Wu Y, Ilboudo Y, Stern EN, Needleman AJ, Galactéros F, Brugnara C, Kutlar A, et al. 2017. Variant-aware saturating mutagenesis using multiple Cas9 nucleases identifies regulatory elements at trait-associated loci. *Nature Publishing Group* **49**: 625–634.

Canver MC, Smith EC, Sher F, Pinello L, Sanjana NE, Shalem O, Chen DD, Schupp PG, Vinjamur DS, Garcia SP, et al. 2015. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* **527**: 192–197.

Cech TR, Steitz JA. 2014. The noncoding RNA revolution-trashing old rules to forge new ones. *Cell* **157**: 77–94.

Das PP, Shao Z, Beyaz S, Apostolou E, Pinello L, De Los Angeles A, O'Brien K, Atsma JM, Fujiwara Y, Nguyen M, et al. 2014. Distinct and Combinatorial Functions of Jmjd2b/Kdm4b and Jmjd2c/Kdm4c in Mouse Embryonic Stem Cell Identity. *Molecular Cell* **53**: 32–48.

Diao Y, Fang R, Li B, Meng Z, Yu J, Qiu Y, Lin KC, Huang H, Liu T, Marina RJ, et al. 2017. A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nature Methods* **14**: 629–635.

Diao Y, Li B, Meng Z, Jung I, Lee AY, Dixon J, Maliskova L, Guan K-L, Shen Y, Ren B. 2016. A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening. *Genome Res* **26**: 397–405.

Elkon R, Agami R. 2017. Characterization of noncoding regulatory DNA in the human genome. *Nat Biotechnol* **35**: 732–746.

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**: 43–49.

Farh KK-H, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, Shoresh N, Whitton H, Ryan RJH, Shishkin AA, et al. 2015. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**: 337–343.

Fulco CP, Munschauer M, Anyoha R, Munson G, Grossman SR, Perez EM, Kane M, Cleary B, Lander ES, Engreitz JM. 2016. Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* **354**: 769–773.

Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, Li Y, Fine EJ, Wu X, Shalem O, et al. 2013. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol* **31**: 827–832.

Korkmaz G, Lopes R, Ugalde AP, Nevedomskaya E, Han R, Myacheva K, Zwart W, Elkon R, Agami R. 2016. Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat Biotechnol* **34**: 192–198.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357–359.

Long HK, Prescott SL, Wysocka J. 2016. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* **167**: 1170–1187.

Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. 2013. RNA-guided human genome engineering via Cas9. *Science* **339**: 823–826.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**: 1190–1195.

Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG, Kinney JB, et al. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30**: 271–277.

Miele A, Dekker J. 2008. Long-range chromosomal interactions and gene regulation. *Mol Biosyst* **4**: 1046–1057.

Mojica FJM, Díez-Villaseñor C, García-Martínez J, Almendros C. 2009. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology (Reading, Engl)* **155**: 733–740.

Mouse ENCODE Consortium, Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, Gilbert DM, Groudine M, Bender M, Kaul R, et al. 2012. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol* **13**: 418.

Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, Chang HY. 2016. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nature Methods* **13**: 919–922.

Murakawa Y, Yoshihara M, Kawaji H, Nishikawa M, Zayed H, Suzuki H, Fantom Consortium, Hayashizaki Y. 2016. Enhanced Identification of Transcriptional Enhancers Provides Mechanistic Insights into Diseases. *Trends Genet* **32**: 76–88.

Niwa H, Miyazaki J, Smith AG. 2000. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nature Genetics* **24**: 372–376.

Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee S-I, Cooper GM, et al. 2012. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* **30**: 265–270.

Rajagopal N, Srinivasan S, Kooshesh K, Guo Y, Edwards MD, Banerjee B, Syed T, Emons BJM, Gifford DK, Sherwood RI. 2016. High-throughput mapping of regulatory DNA. *Nat Biotechnol* **34**: 167–174.

Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317–330.

Sanjana NE, Wright J, Zheng K, Shalem O, Fontanillas P, Joung J, Cheng C, Regev A, Zhang F. 2016. High-resolution interrogation of functional elements in the noncoding genome. *Science* **353**: 1545–1549.

Seruggia D, Oti M, Tripathi P, Canver MC, LeBlanc L, Di Giammartino DC, Bullen MJ, Nefzger CM, Sun YBY, Farouni R, et al. 2019. TAF5L and TAF6L Maintain Self-Renewal of Embryonic Stem Cells via the MYC Regulatory Network. *Molecular Cell* 1–24.

Shalem O, Sanjana NE, Zhang F. 2015. High-throughput functional genomics using CRISPR–Cas9. *Nature Publishing Group* **16**: 299–311.

Siepel A, Arbiza L. 2014. Cis-regulatory elements and human evolution. *Current Opinion in Genetics & Development* **29**: 81–89.

Stergachis AB, Neph S, Sandstrom R, Haugen E, Reynolds AP, Zhang M, Byron R, Canfield T, Stelhing-Sun S, Lee K, et al. 2014. Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature* **515**: 365–370.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. 2012. The accessible chromatin landscape of the human genome. *Nature* **489**: 75–82.

van de Werken HJG, Landan G, Holwerda SJB, Hoichman M, Klous P, Chachik R, Splinter E, Valdes-Quezada C, Oz Y, Bouwman BAM, et al. 2012. Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nature Methods* **9**: 969–972.

Vierstra J, Rynes E, Sandstrom R, Zhang M, Canfield T, Hansen RS, Stehling-Sun S, Sabo PJ, Byron R, Humbert R, et al. 2014. Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science* **346**: 1007–1012.

Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R, Erichsen JT, Jasinska AJ, et al. 2015. Enhancer evolution across 20 mammalian species. *Cell* **160**: 554–566.

Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al. 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**: 854–858.

Wang T, Wei JJ, Sabatini DM, Lander ES. 2014. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**: 80–84.

Whyte WA, Bilodeau S, Orlando DA, Hoke HA, Frampton GM, Foster CT, Cowley SM, Young RA. 2012. Enhancer decommissioning by LSD1 during embryonic stem cell differentiation. *Nature* **482**: 221–225.

Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA. 2013. Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell* **153**: 307–319.

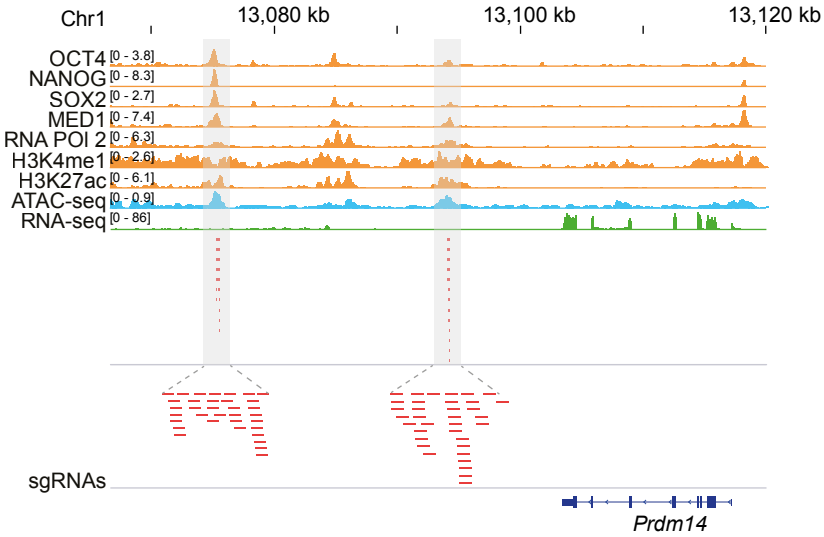
Yeom YI, Fuhrmann G, Ovitt CE, Brehm A, Ohbo K, Gross M, Hübner K, Schöler HR. 1996. Germline regulatory element of Oct-4 specific for the totipotent cycle of embryonal cells.

Development **122**: 881–894.

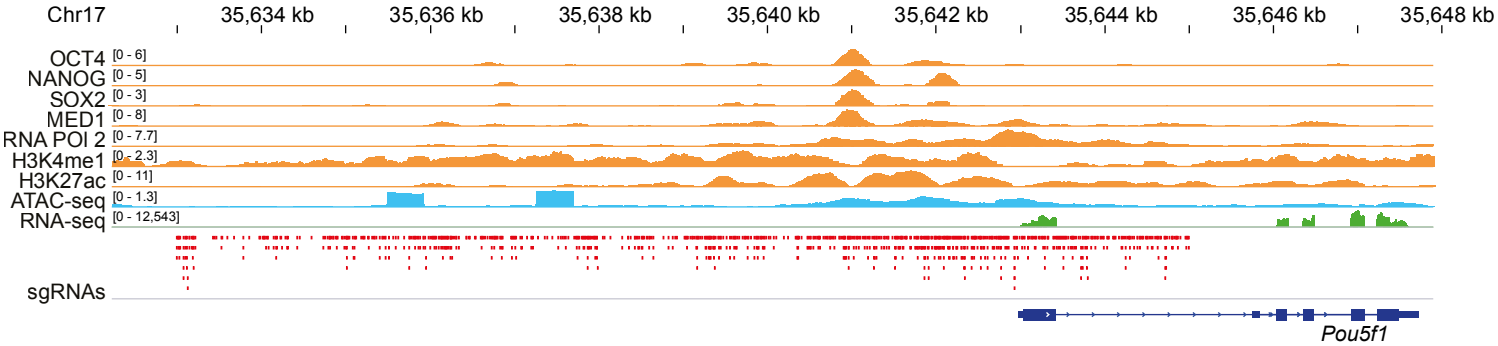
Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.

Canver_Fig1

A



B



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

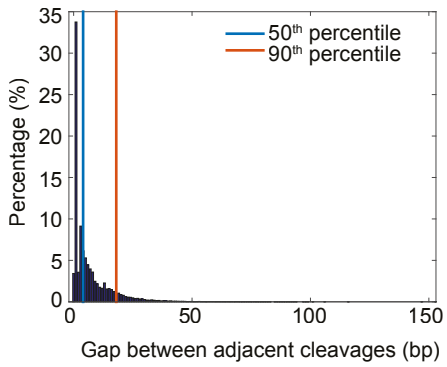
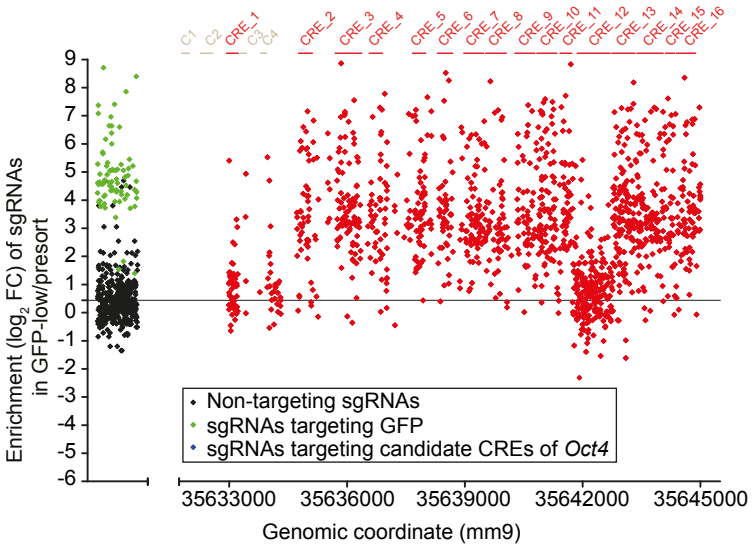


Figure 1. Design of a saturating CRISPR-Cas9 pooled library for ctSCAN-SMS. (A) Genomic tracks show co-occupancy of ESC-TFs (OCT4, NANOG, SOX2), mediator (MED1), enhancer histone marks (H3K27ac, H3K4me1), and RNA Pol2 at the *Prdm14* gene locus in mESCs. ATAC-seq track represents open chromatin regions and enhancers (ENs); RNA-seq track represents gene expression. Highlighted regions display as putative EN REs. All possible sgRNAs (shown with red dashed lines) upstream of PAM sequences (NGG) within the putative REs. These REs termed as *trans*-REs (TREs). (B) Genomic tracks illustrate co-occupancy of ESC-TFs (OCT4, NANOG, SOX2), mediator (MED1), enhancer histone marks (H3K27ac, H3K4me1) and RNA Pol2 at 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 *Oct4*.

Canver_Fig2

A



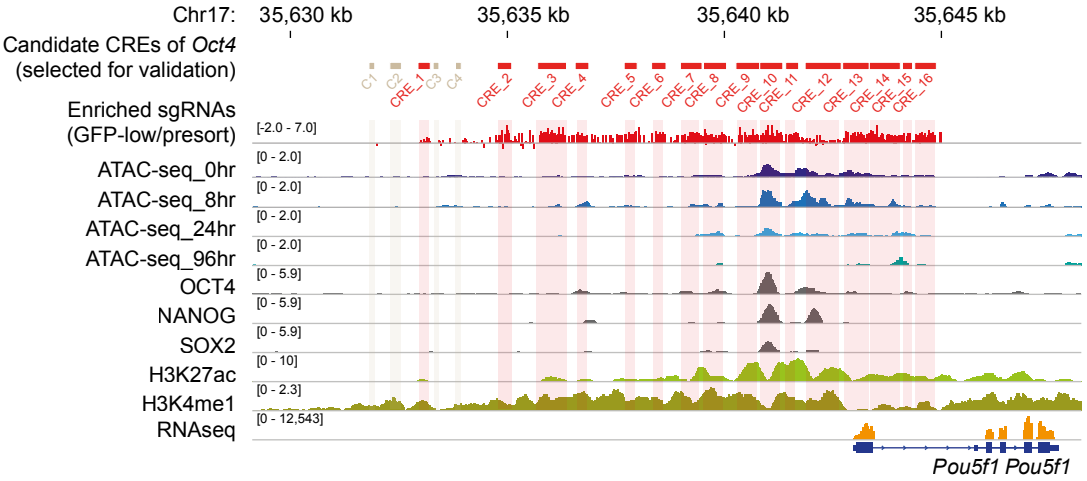
B **List of candidate CREs *of the Oct4* gene**

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

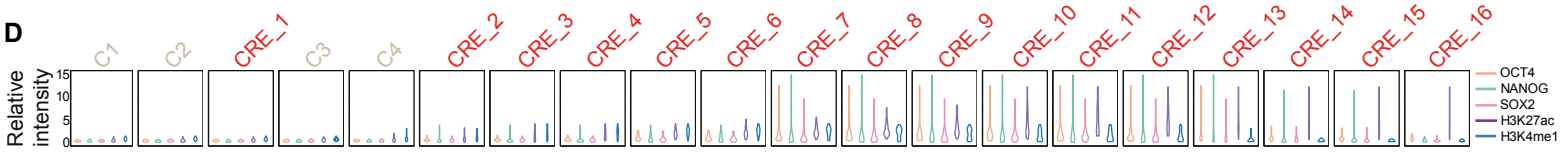
Candidate CREs	Genomic locations of REs (mm9) (enriched sgRNAs)	Length (bp)	No. of sgRNAs
CRE_1	chr17:35,632,944-35,633,217	273	51
CRE_2	chr17:35,634,773-35,635,108	335	40
CRE_3	chr17:35,635,706-35,636,370	664	98
CRE_4	chr17:35,636,568-35,636,884	316	35
CRE_5	chr17:35,637,687-35,637,998	311	53
CRE_6	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 (postive controls)	NA	NA	118
Non-targeting sgRNAs (negative controls)	NA	NA	1984

CRE_10 and CRE_12 are the known distal and proximal enhancers
CRE_13 consist TSS; CREs_13-16 present within the promoter of *Oct4*

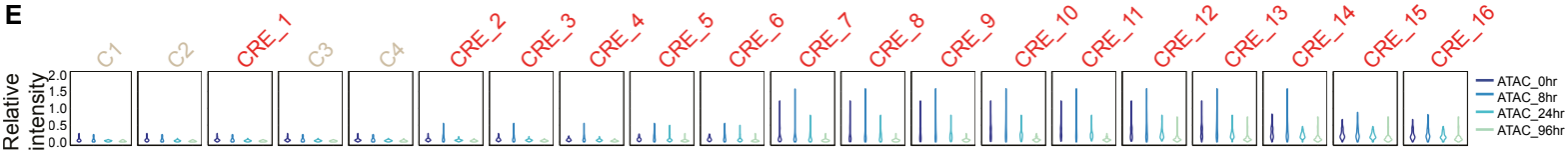
C



D



E



F

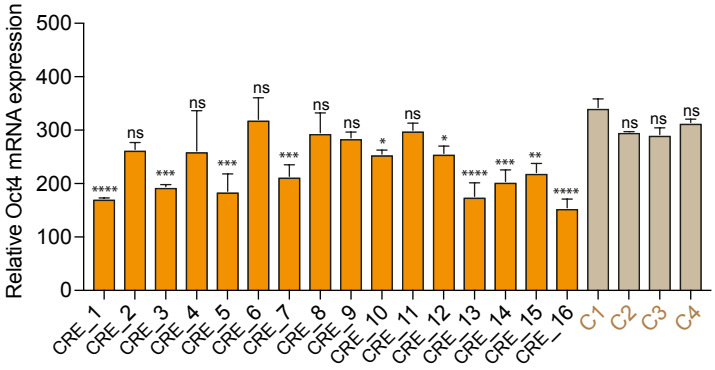
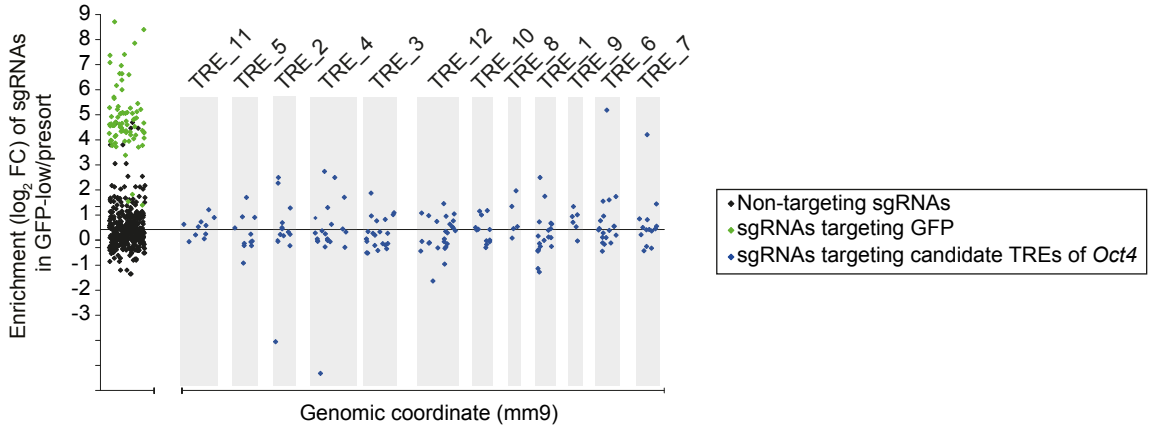


Figure 2. Identification and dissection of active CREs of the *Oct4* gene. (A) Dot plot analysis demonstrates the enrichment score of each sgRNA by comparing their frequency in the GFP-low cells to the pre-sort cells. 16 candidate CREs were identified based on the mean enrichment score (mean log2FC) 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 *Oct4.* (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 changes 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 level quantified upon deletion of individual CRE 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).

Canver_Fig3

A



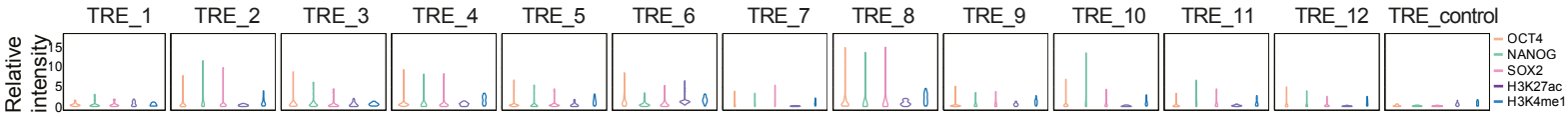
B

List of candidate TREs of the *Oct4* gene

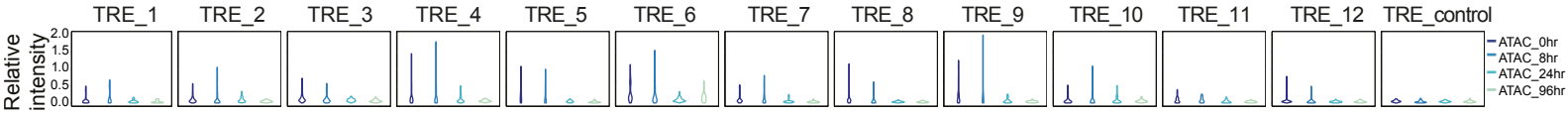
Based on HMM based enrichment scores of sgRNAs.
All the sgRNAs must have off-target score >10.
TREs must possess at least 4 sgRNAs with mean enrichment scores (mean log2FC) >0.5, P<0.0001.
TREs should co-occupy with ESC-TFs (OCT4, NANOG, SOX2), enhancer histone marks (H3K27ac, H3K4me1) and dynamic open 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 (postive controls)	NA	NA	118
sgRNAs_non-targeting (negative controls)	NA	NA	1984

C



D



E

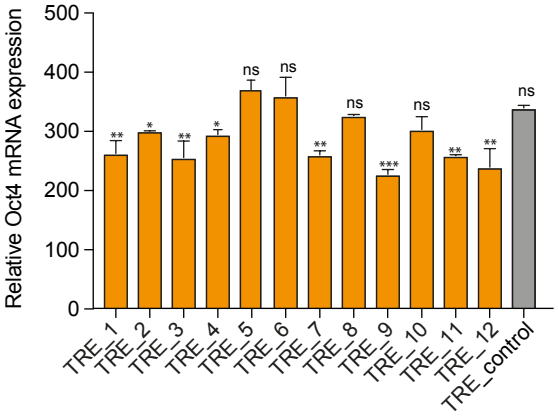
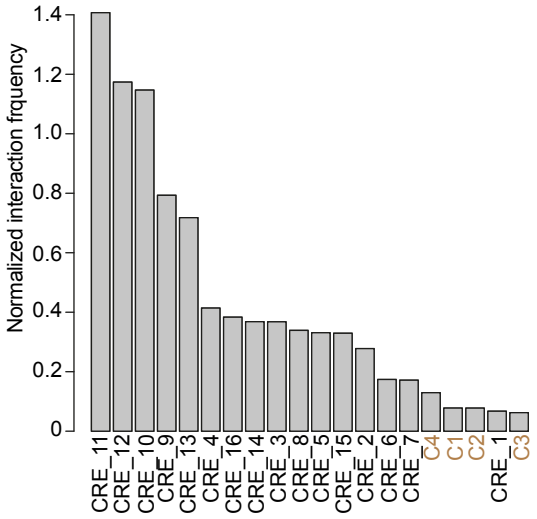


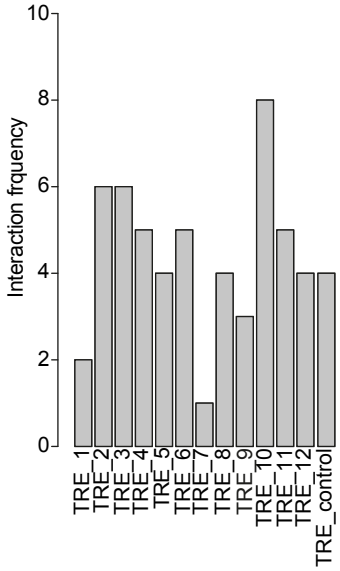
Figure 3. Identification and validation of active TREs of the *Oct4* gene. (A) Dot plot analysis displays the enrichment score of each sgRNA by comparing their frequency in the GFP-low cells to the pre-sort cells at the selected candidate 12 TREs. (B) A list of identified candidate TREs of *Oct4*. (C) Binding changes 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 deletions of individual TRE 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).

Canver_Fig4

A



B



C

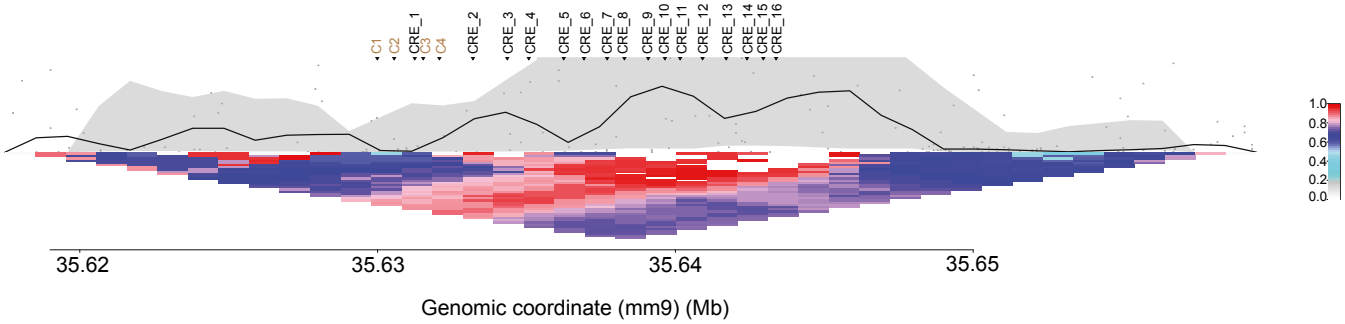
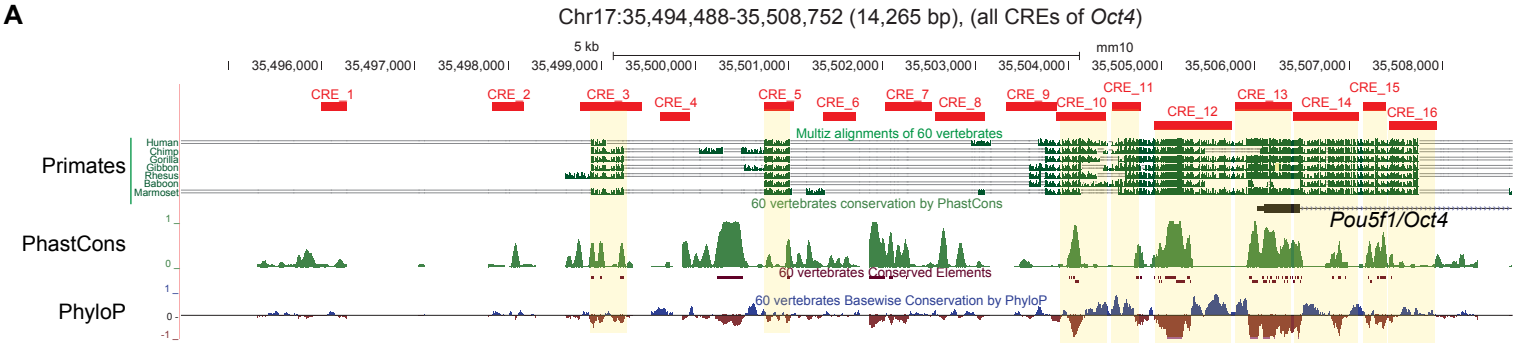


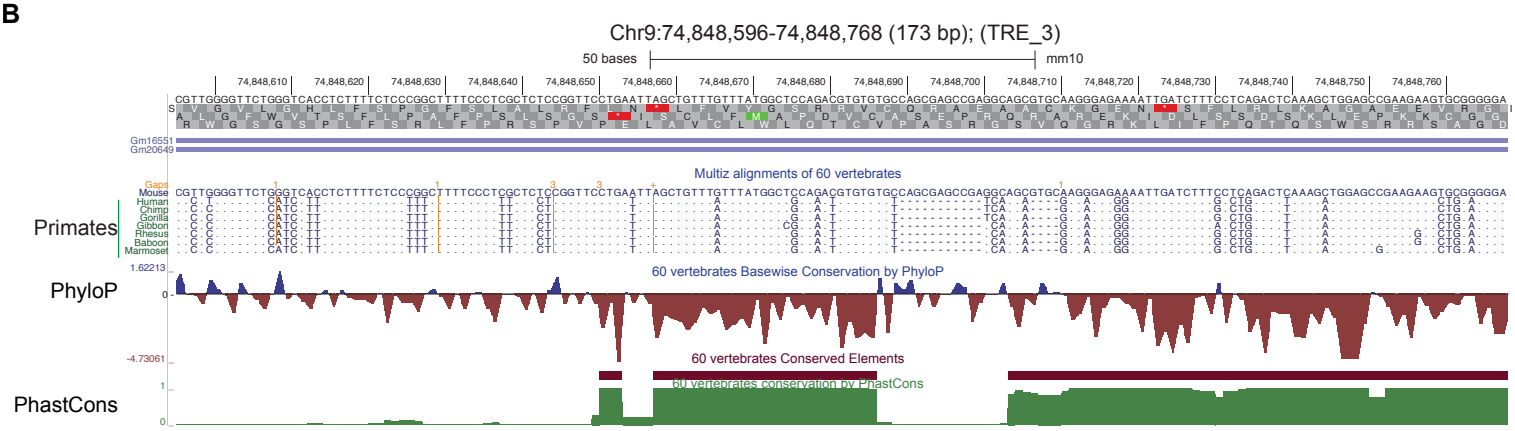
Figure 4. Physical interactions between CREs, TREs, and the *Oct4* promoter in *Oct4* gene regulation. (A) 4C-seq data represent normalized interaction frequencies between CREs and the *Oct4* promoter. The interaction/contact frequencies between CREs and *Oct4* promoter measured at 1 kb resolution window. (B) The interaction/contact frequencies between TREs and *Oct4* promoter quantified at 50 kb resolution window. (C) The contact profile of CREs and *Oct4* promoter at 1 kb resolution window. Bottom triangle is a heat map of normalized contact frequencies between 1 kb bins represented with the color codes (highest is 1 with red in color; 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 displays 20th-80th percentile of the normalized contact frequencies.

Canver_Fig5

A



B



C

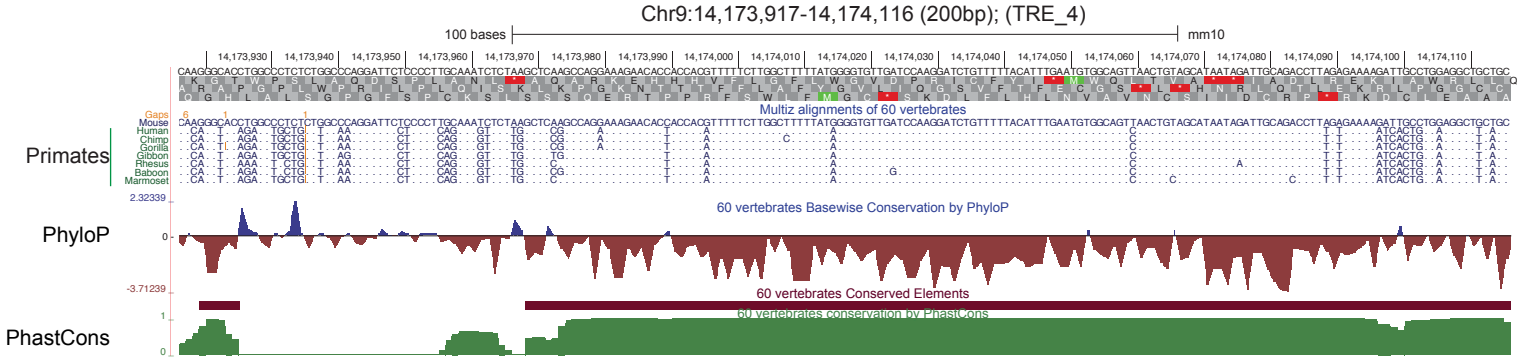


Figure 5. Conserved active CREs and TREs of the *Oct4* gene. (A) Orthologous sequences from the representative primates (including human) are listed around the ~12kb region of the mouse *Oct4* locus. CREs of mouse *Oct4* are labelled with solid red bars. PhyloP and PhastCons estimate evolutionary conservation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

Canver_Fig6

A

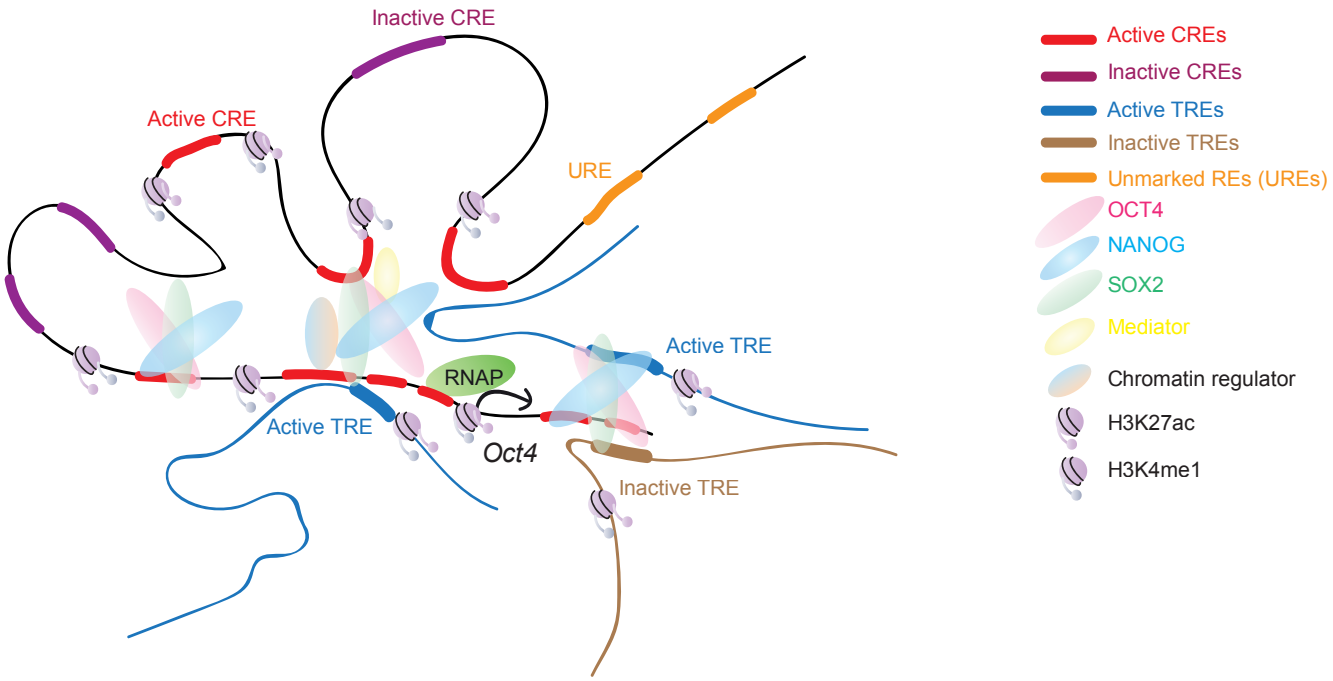


Figure 6. Model representing the detailed functions of CREs and TREs in *Oct4* gene regulation. The proposed model describes the existence of multiple active CREs (red in color) and TREs (blue in color) of *Oct4* locus in mESCs. However, not all of the active REs have regulatory features (i.e. co-occupancy 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 few active REs physically interacts with the *Oct4* promoter through intra-chromosomal (for CREs) and inter-chromosomal interactions (for TREs). Taken together, it suggests that active REs act beyond their regulatory features and physical contact with the *Oct4* promoter to control the transcriptional output of *Oct4* gene.