



# The interactomes of *POU5F1* and *SOX2* enhancers in human embryonic stem cells

Fan Gao<sup>1,2</sup>, Zong Wei<sup>1</sup>, Woojin An<sup>3</sup>, Kai Wang<sup>2</sup> & Wange Lu<sup>1</sup>

<sup>1</sup>Eli and Edythe Broad Center for Regenerative Medicine and Stem Cell Research, Department of Biochemistry and Molecular Biology, University of Southern California, 1425 San Pablo Street, Los Angeles, CA 90033, USA, <sup>2</sup>Zilkha Neurogenetic Institute, University of Southern California, 1501 San Pablo Street, Los Angeles, CA 90033, USA, <sup>3</sup>Norris Comprehensive Cancer Center, Department of Biochemistry and Molecular Biology, University of Southern California, 1441 Eastlake Avenue, Los Angeles, CA 90033, USA.

SUBJECT AREAS:  
EMBRYONIC STEM CELLS  
EPIGENOMICS  
EPIGENETICS  
EPIGENETIC MEMORY

Received  
4 February 2013

Accepted  
15 March 2013

Published  
3 April 2013

Correspondence and requests for materials should be addressed to K.W. (kaichop@gmail.com) or W.L. (wange.lu@med.usc.edu)

The genes *POU5F1* and *SOX2* are critical for pluripotency and reprogramming, yet the chromosomal organization around these genes remains poorly understood. We assayed long-range chromosomal interactions on putative enhancers of *POU5F1* and *SOX2* genes in human embryonic stem cells (hESCs) using 4C-Seq technique. We discovered that their frequent interacting regions mainly overlap with early DNA replication domains. The interactomes are associated with active histone marks and enriched with 5-hydroxymethylcytosine sites. In hESCs, genes within the interactomes have elevated expression. Additionally, some genes associated with the *POU5F1* enhancer contribute to pluripotency. Binding sites for multiple DNA binding proteins, including ATF3, CTCF, GABPA, JUND, NANOG, RAD21 and YY1, are enriched in both interactomes. The *RARG* locus, frequently interacting with the *POU5F1* locus, has abundant RAD21 binding sites co-localized with other protein binding sites. Thus the interactomes of these two pluripotency genes could be an important part of the regulatory network in hESCs.

Embryonic stem (ES) cells are pluripotent stem cells characterized by unlimited self-renewal capacity and the ability to differentiate into all three germ layers. These cells are considered promising in cell replacement therapy. The development of induced pluripotent stem (iPS) cell technology, in which overexpression of four transcription factors (OCT4/*POU5F1*, *SOX2*, *KLF4*, and *MYC*) reverts somatic cells such as skin fibroblasts to the pluripotent state and generates ES-like iPS cells, further propelled the field of regenerative medicine and stem cell therapy<sup>1</sup>. These reprogramming factors eventually activate the expression of a class of pluripotency genes. Extensive efforts have been made to understand the epigenetic regulation of genes related to stem cell self-renewal, pluripotency and reprogramming. Currently, DNA methylation<sup>2</sup>, histone modification<sup>3–7</sup> and transcription factor binding<sup>8–11</sup> are recognized as critical factors in chromatin structure modulation and gene regulation<sup>12,13</sup>, and pluripotent stem cells (PSCs) exhibit distinct DNA methylation patterns and histone modifications compared to differentiated cells. In addition, a characteristic nuclear architecture has been established as a unique epigenetic feature of PSCs<sup>14</sup>; thus long-range chromatin interactions have been postulated to play critical roles in maintaining pluripotent cell function. However, detailed chromatin–interaction maps remain to be characterized to understand the functional importance of the chromatin interactomes, such as the crosstalk between enhancer elements and distally targeted genes.

In the last several years, chromosome conformation capture (3C) has emerged as a high-throughput molecular biology approach to explore long-range chromatin interactions. Newer techniques such as Hi-C and 5C detect global chromatin interactions<sup>15,16</sup>. In combination with next generation sequencing, Hi-C unveiled interesting chromosomal organization such as open and closed chromatin compartments in the nucleus. A recent study revealed topological domains of the mammalian genomes from 3C based high-resolution chromatin interaction maps<sup>17</sup>. Other methods like ChIA–PET uncovered interactions mediated by particular DNA binding proteins<sup>18</sup>. All of these methods provide novel insights into global higher-order chromosomal structure; however, they only detect limited interacting partners for a specific regulatory element. Circular chromosome conformation capture (4C) with restriction enzyme fragmentation and proximity-based ligation was designed to detect all unknown sites interacting with a known “bait” sequence<sup>19–22</sup>. In addition, a sonication-based 4C library preparation method was developed to capture strong chromatin interactions of the bait sequence<sup>18,23</sup>. Together with next-generation sequencing, this technology enables genome-wide identification of strong interacting partners of a regulatory element such as an enhancer.



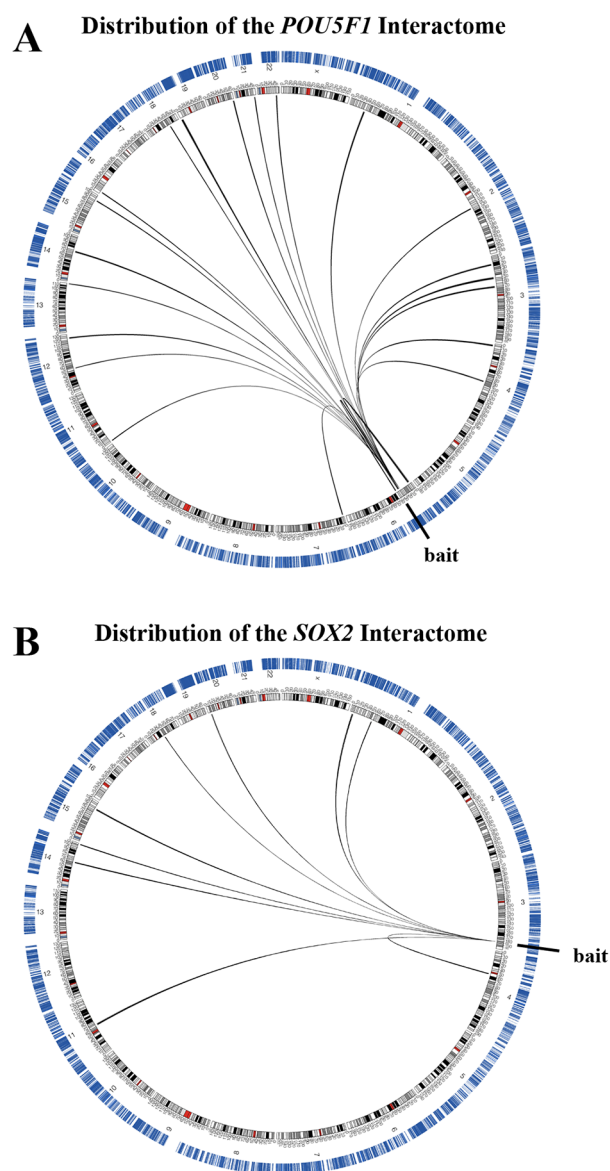
During early embryogenesis, enhancer elements marked with different chromatin signatures either activate or suppress transcription of nearby genes<sup>6</sup>. In hESCs, upstream putative enhancers of the *POU5F1* and *SOX2* genes are associated with p300 and H3K27ac<sup>7</sup>. The upstream distal enhancer of the *POU5F1* gene is a well-known regulator of OCT4 expression in PSCs via looping<sup>13,24</sup>. It has been reported that enhancer elements can regulate genes several hundred-thousand base pairs away<sup>25,26</sup>; thus, we hypothesize that *POU5F1* and *SOX2* enhancers may play long-range regulatory roles in addition to activating nearby genes. To test this hypothesis, we utilized a high-throughput assay to detect *POU5F1* and *SOX2* enhancer-associated long-range interactions. With extensive data analysis, we aim to understand the roles of these long-range interactions in the regulatory transcriptional circuitry that governs stem cell self-renewal and pluripotency.

## Results

**Identification of *POU5F1* and *SOX2* enhancer interactomes.** The putative *POU5F1* and *SOX2* enhancer regions are evolutionary conserved across vertebrates (44 species), with active enhancer signatures defined by epigenetic marks such as p300 histone acetyltransferase and H3K27ac but not H3K27me3 in hESCs<sup>7</sup> (Supplementary Fig. 1A). We applied the 4C technique to identify the interacting partners of the “bait” putative enhancers of *POU5F1* and *SOX2* in the pluripotent H9 cell line. Briefly, the cells were fixed to crosslink chromosomes within proximity. The chromosomes were then fragmented by sonication. Interacting chromatin fragments were ligated and the DNA pieces were purified. Genomic regions interacting with the “bait” were then amplified by nested PCR (Supplementary Fig. 1B, Supplementary Table 1). We designed inverse PCR primers to target putative enhancers of the *POU5F1* and *SOX2* genes. The constructed 4C library could be visualized by DNA electrophoresis, whereas the control without ligation showed almost no PCR products (Supplementary Fig. 1B), indicating that the 4C library was amplified from ligation products.

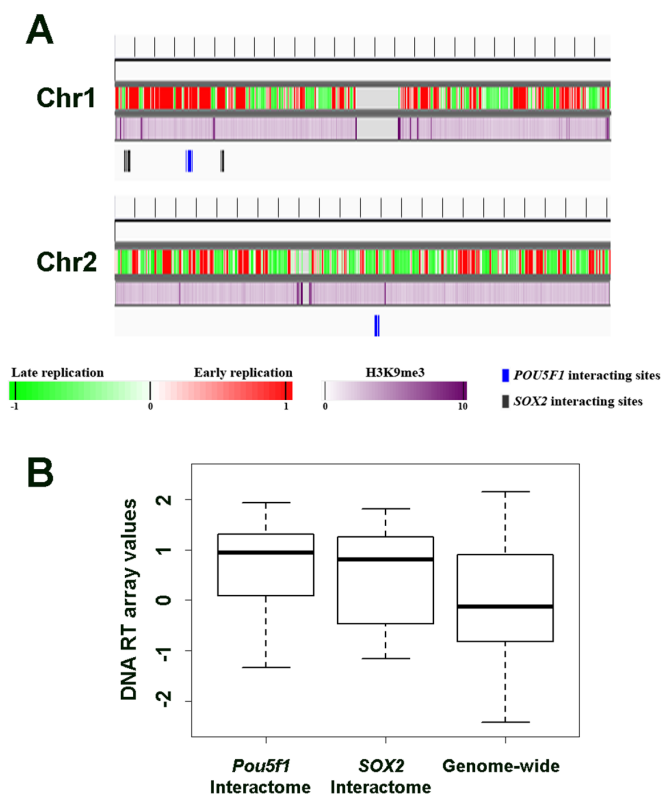
We performed next-generation DNA sequencing using an Illumina HiSeq sequencer, and classified the identified 4C interactions as either proximal or distal interactions (see Methods). Distal interactions include inter-chromosomal interactions and intra-chromosomal interactions over genomic distances greater than 20 kb, whereas proximal interactions cover intra-chromosomal regions with genomic distances between 300 bp and 20 kb. Consistent with the results of 3C-based studies, our distal interaction reads account for only a small portion of the total interactions, roughly 10% ~ 20% for the 4C libraries (Supplementary Table 2). We used two different batches of H9 cells to construct the 4C libraries for *POU5F1* and *SOX2* independently for next-generation sequencing. The two replicates of *POU5F1* and *SOX2* distal interactions overlap by 35% and 25%, respectively. This is consistent with the moderate overlap observed in biological replicates of CTCF-mediated chromatin interactomes in mouse ES cells<sup>27</sup>, and may reflect the diversity and dynamics of enhancer interactions, ligation efficiency, complexity of PCR amplification as well as sequencing depth. To filter out interactions that likely resulted from stochastic effects, we used a false discovery rate (FDR)-based statistical model to identify the enriched interacting domains across the whole genome. We further merged the enriched interacting domains that overlap between the biological replicates as being high-confidence frequent interacting domains. In total, 23 and 9 high-confidence frequent interacting domains were identified for the *POU5F1* and *SOX2* interactomes, respectively (Figure 1, Supplementary Table 3, 4), and most of them are inter-chromosomal interactions that involve gene-rich regions, similar to the e4C results<sup>20</sup>.

**The interactomes of *POU5F1* and *SOX2* enhancers overlap with early DNA replication domains but not with heterochromatic**



**Figure 1 |** Circos maps of the distal interactions are presented using Circos software<sup>52</sup>. The blue outer ring represents the gene density profile.

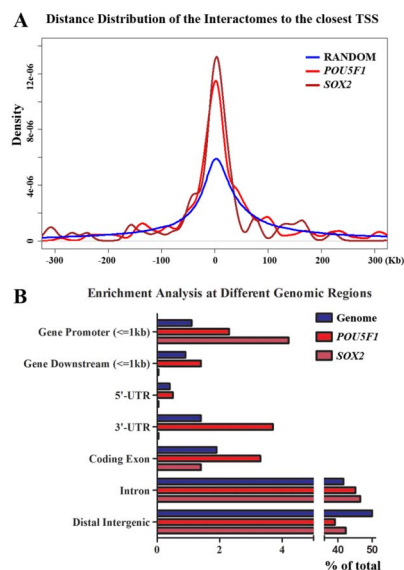
**regions.** We next examined whether the *POU5F1* and *SOX2* enhancer interacting domains (size ranging from 1 Mb to 4 Mb, Supplementary Table 3, 4) have any unique epigenetic features. As Ryba et al. showed<sup>28</sup>, DNA replication timing correlates with the spatial proximity of chromatin as measured by Hi-C analysis, suggesting that early and late DNA replication occur in spatially distinct compartments in the nucleus. We noticed that the *POU5F1* and *SOX2* gene loci are within early DNA replication domains in hESCs, and wondered whether their interacting domains have a similar DNA replication timing. As shown in Figure 2A, the *POU5F1* and *SOX2* enhancer interacting domains mainly overlap with early DNA replication domains in hESCs. The distribution of assayed replication timing values within the genomic regions surrounding the identified *POU5F1* and *SOX2* enhancer-interacting sites (50 kb range) shows a shift toward positive values compared to genome-wide array values ( $P < 2.2 \times 10^{-16}$  for both, by unpaired Wilcoxon rank-sum test; Figure 2B). This observation revealed an interesting epigenetic feature, that higher-order structures surrounding the *POU5F1* and *SOX2* enhancers have synchronized DNA replication timing. As early DNA replication domains have been connected to



**Figure 2** | Relationship of the interactomes with DNA replication domains and heterochromatic regions. Interacting sites within the frequent interacting domains are presented in (2A) (only Chr1 and Chr2 are shown). (2B), Comparison of distribution of DNA replication timing (RT) array values for the regions within  $\pm 50$  kb of the interacting sites to the whole genome.

active gene transcription in hESCs<sup>28</sup>, it is likely that the interactions of *POU5F1* and *SOX2* enhancers with the identified distal regions have functional significance. This observation is consistent with what we have seen in *POU5F1* enhancer interactome in mouse ES cells (data not shown). Also revealed in **Figure 2A**, *POU5F1* and *SOX2* interacting domains do not overlap with heterochromatic regions labeled with strong H3K9me3 signals in hESCs<sup>29</sup>, suggesting that the interactomes are within an active portion of the genome in terms of gene regulation. In addition, we explored a potential relationship with the nuclear lamina-associated domains (LADs) of human chromosomes<sup>30</sup>, but did not observe any connection, possibly due to the fact that the LADs were determined in human fibroblasts.

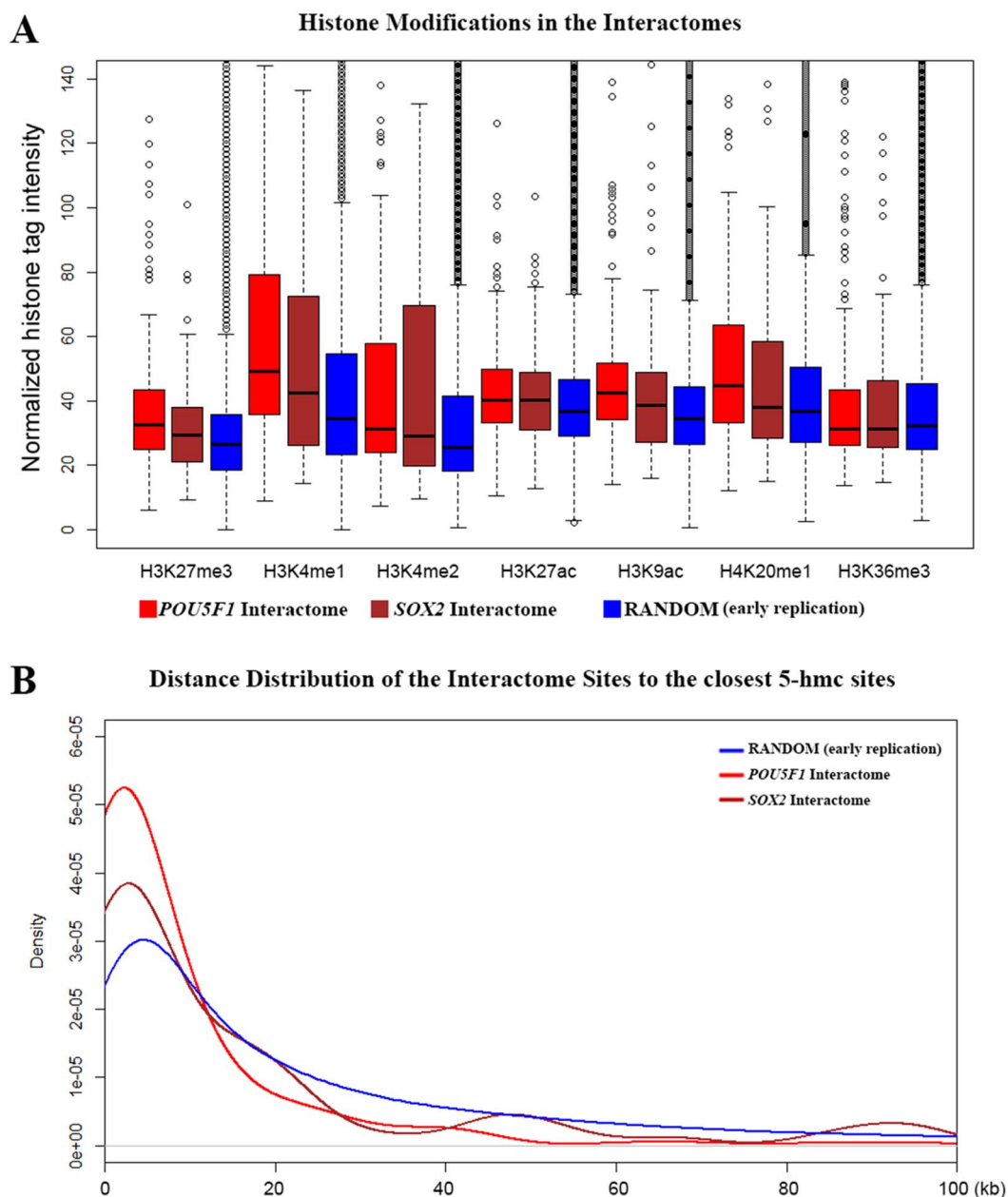
**The enhancer interactomes are adjacent to transcription start sites and possess active epigenetic features.** To explore the correlation of *POU5F1* and *SOX2* enhancer interactomes with annotated gene locations, we analyzed the distribution of the genomic distance of the enhancer interacting sites to the closest gene transcription start site (TSS) (**Figure 3A**). The kernel density plots of both *POU5F1* and *SOX2* enhancer-interacting sites clearly show a sharp peak centered at TSS. When compared to the distribution peak of randomly simulated sites in the whole genome, the peaks observed in the enhancer interactomes are significantly higher within a narrow window around TSS ( $P = 0.0018$ ,  $P = 0.0047$ , respectively, Kolmogorov-Smirnov test). Enrichment of the interacting sites around TSS suggests that *POU5F1* and *SOX2* enhancers prefer interacting with genomic regions containing genes. We further investigated the distribution of *POU5F1* and *SOX2* enhancer-interacting sites at different genomic regions<sup>31</sup>. Shown in **Figure 3B**, gene promoter



**Figure 3** | (A) Kernel density estimation of genomic distance from the interacting sites to the nearest TSS sites. Distances were calculated using HOMER software<sup>34</sup>. Density maps of 100,000 random sites in the whole genome are also included. (B) Enrichment analysis of the interactomes at different genomic regions. Distribution analysis was performed using CEAS software<sup>31</sup>.

sequences are one of the enriched regions for both the *POU5F1* and *SOX2* enhancer interactomes. Additionally, the fraction of distal intergenic regions is consistently reduced for the two interactomes. Therefore, our observation suggests that the higher-order chromosome structure surrounding the active enhancers may be directly involved in gene regulation. Also noted, when random sites in the early DNA replication areas (see Methods for details) are selected to compare distance distribution to TSS, the *POU5F1* and *SOX2* interactomes are still more enriched around TSS, although statistically not significant ( $P = 0.390$ ,  $P = 0.2098$ , respectively, Kolmogorov-Smirnov test, **Supplementary Fig. 2**).

Histone modification is known to be involved in transcriptional regulation by decondensing compact chromatin structures for transcription factor binding. 3C-Based genome-wide interactome studies have identified active and inactive chromatin compartments in the nucleus, with active compartments enriched with histone marks that activate gene transcription, such as H3K9ac<sup>32</sup>. We therefore asked whether active enhancers of *POU5F1* and *SOX2* selectively interact with distal regions showing active gene transcription. ChIP-Seq tag profiles of H3K27me3, H3K4me1, H3K4me2, H3K27ac, H3K9ac, H4K20me1, and H3K36me3 in H1 ES cell line<sup>33</sup> were used for analyzing histone patterns associated with the interactomes. ChIP-Seq tags around the enhancer interacting sites were counted and normalized<sup>34</sup>, and visualized as boxplot. As noted, *POU5F1* and *SOX2* interactomes have higher intensity profiles of H3K4me1, H3K4me2 and H4K20me1, which are marks for gene active regulation (**Figure 4A**). Compared to the random sites in early DNA replication areas, the investigated histone marks are in general more enriched in the *POU5F1* interactome than in the *SOX2* interactome, with H3K4me1 and H3K9ac are the most significantly enriched ones in the *POU5F1* interactome ( $P < 2.2 \times 10^{-16}$  for both marks, unpaired Wilcoxon rank-sum test). Interestingly, Polycomb-mediated gene silencing mark H3K36me3<sup>35</sup> is not enriched in either interactome ( $P = 0.485$ ,  $P = 0.922$ , respectively). We also investigated the relationship of *POU5F1* and *SOX2* enhancer interactomes with 5-hydroxymethylcytosine (5-hmC) sites in the genome. As a previous study revealed, genomic 5-hmC sites are enriched at active enhancers in hESCs<sup>36</sup>. Because *POU5F1* and *SOX2* upstream



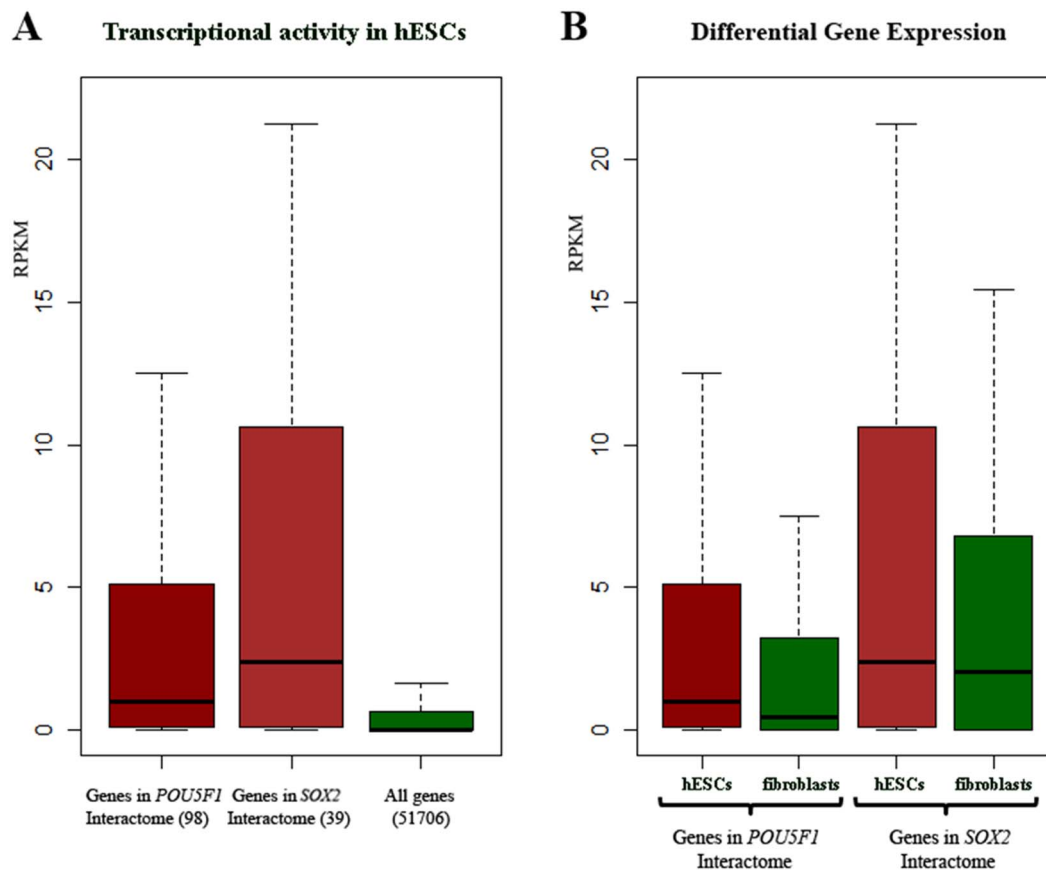
**Figure 4** | (A) Boxplot of different histone marks around the interacting sites and random sites. ChIP-Seq tags within  $\pm 5$  kb from an interacting site were counted and normalized. (B) Distance distribution of the interacting sites and random sites to the closest 5-hmC site were compared. 100,000 random sites were generated from early DNA replication areas for comparison.

enhancers are adjacent to 5-hmC sites (within 5 kb), we asked whether the enhancer interactomes are also enriched with 5-hmC sites. Shown in **Figure 4B**, 5-hmC sites are enriched in the proximity of *POU5F1* and *SOX2* interactomes compared to the random sites in early DNA replication areas ( $P < 2.2 \times 10^{-16}$ ,  $P = 0.047$ , respectively, Kolmogorov-Smirnov test). Thus, the two enhancer interactomes in hESCs possess epigenetic features of active enhancers, which may facilitate active regulation of gene transcription.

**Transcriptional status of *POU5F1* and *SOX2* enhancer interacting genes.** To understand the transcriptional status of genes associated with *POU5F1* and *SOX2* enhancers, we analyzed RNA-Seq transcriptome data in hESCs and fetal lung fibroblasts<sup>37</sup>, focusing on genes that have a TSS within 10 kb of the enhancer-interacting sites. In hESCs, expression of *POU5F1* and *SOX2* enhancer-interacting genes is significantly higher than that of all genes in

hESCs ( $P = 7.5 \times 10^{-6}$  and  $P = 1.2 \times 10^{-6}$ , respectively, by unpaired Wilcoxon rank-sum test; **Figure 5A**), indicating that the interactomes are enriched with actively transcribed genes. Thus, active *POU5F1* and *SOX2* enhancers could be involved in mediating transcription of the associated distal genes. RNA-Seq transcriptome analysis also revealed that the genes originally associated with active *POU5F1* and *SOX2* enhancers in hESCs are down-regulated in differentiated lung fibroblasts ( $P = 1.2 \times 10^{-5}$  and  $P = 0.013$ , respectively, by paired Wilcoxon rank-sum test; **Figure 5B**). These results indicate that the enhancers of these two stemness-related genes interact with a group of genes that is highly expressed in hESCs but repressed in fibroblasts.

**Distal genes associated with the *POU5F1* enhancer are involved in the regulatory circuitry of pluripotency.** To explore whether *POU5F1* and *SOX2* enhancer-interacting genes contribute to



**Figure 5** | (A) Comparison of RPKM values for the interacting genes (genes with a TSS  $\leq 10$  kb from the interacting sites) with all human Ensembl genes (average RPKM values from replicate RNA-Seq data in ENCODE were used). (B) Differential expression was analyzed to compare interactome genes in hESCs and fetal lung fibroblasts.

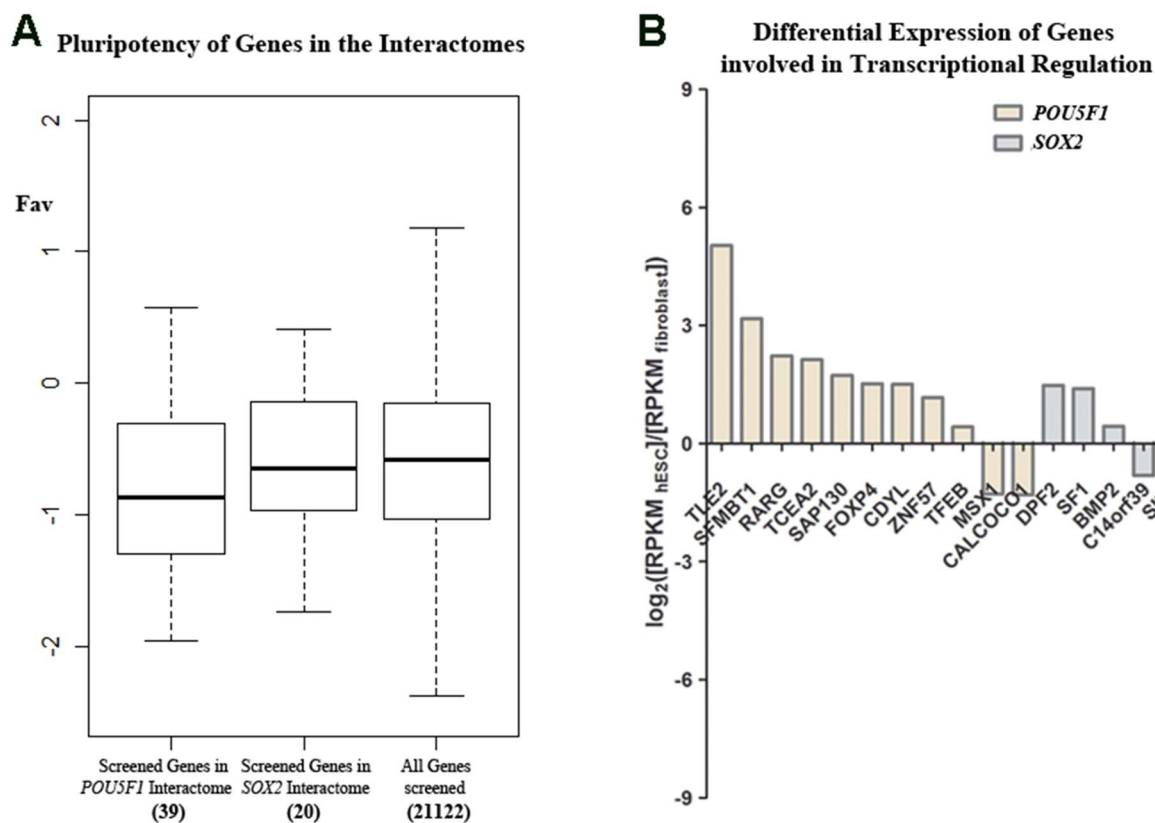
self-renewal and pluripotency of hESCs, we analyzed a data from a genome-wide RNA interference screen using a *POU5F1* promoter reporter assay in hESCs<sup>38</sup>. Interference of transgenic *POU5F1*-GFP expression is quantified with Fav values reflecting GFP fluorescence intensity, which represents the propensity to be stemness-related (Figure 6A). In comparison to all the genes screened, *POU5F1* enhancer-interacting genes (genes closest to the interacting sites, with genomic distance from TSS to the interacting sites less than 10 kb) show a trend of decreasing Fav values, although statistically not significant ( $P = 0.08$ , unpaired Wilcoxon rank-sum test). However, for *SOX2*-targeting genes, the Fav values are similar to all the genes screened ( $P = 0.98$ , unpaired Wilcoxon rank-sum test). Therefore, based on these data, genes in the *POU5F1* interactome appear to be more important for stem cell pluripotency than genes in the *SOX2* interactome.

Gene ontology analysis<sup>39</sup> of 39 *POU5F1*-interacting genes and 20 *SOX2*-interacting genes revealed that a significant portion of them to be involved in transcriptional regulation (30.8% and 35.0% for the *POU5F1* and *SOX2* interactomes, respectively; Supplementary Table 5, 6), and distributed to 8 and 4 different interacting domains in the *POU5F1* and *SOX2* interactomes, respectively, with no more than 3 genes in one interacting domain. Differential expression analysis of these transcriptional regulators in hESCs and fetal lung fibroblasts revealed 9 of those in the *POU5F1* interactome (11 of 12 identified transcriptional regulators have RNA-Seq data) to have higher expression in hESCs (Figure 6B), suggesting active roles for these genes in the pluripotency regulatory network in hESCs. For those transcriptional regulators in the *SOX2* interactome, 3 of 5 showed increased expression in hESCs. Therefore, the *POU5F1* enhancer

interactome may play a greater role in stem cell pluripotency than the *SOX2* enhancer interactome.

**Several transcription factor binding sites are enriched in *POU5F1* and *SOX2* enhancer interactomes.** Transcription factor binding is one of the characteristics that enhancers possess, and it may facilitate long-range chromatin–chromatin interactions of the enhancer with distal genes. *POU5F1* and *SOX2* putative enhancers are known hot spots for association of multiple DNA binding proteins in hESCs. To understand whether certain transcription factors are enriched in *POU5F1* and *SOX2* enhancer-interacting regions, we systematically analyzed ChIP-Seq profiles of 32 DNA binding proteins in hESCs (see Methods). Normalized ChIP-Seq tag counts around the center of the 4C interacting sites were calculated, and visualized using boxplot (Figure 7A). As a control, the counts around the center of the random sites in the early DNA replication areas were also calculated. Statistical analysis identified ATF3, CTCF, GABPA, JUND, NANOG, RAD21 and YY1 sites were the most enriched proteins in both enhancer interactomes compared to the control ( $P < 0.01$ , unpaired Wilcoxon rank-sum test). However, pluripotency-related transcription factor OCT4 is not enriched in either *POU5F1* or *SOX2* interactome. Therefore, ATF3, CTCF, GABPA, JUND, NANOG, RAD21 and YY1 are the characteristic proteins in *POU5F1* and *SOX2* enhancer interactomes in hESCs, and their presence may be functionally important.

**RAD21 is potentially in complex with other transcription factors to mediate the enhancer interactomes.** As shown in Figure 7A, RAD21 is one of the enriched proteins identified in *POU5F1* and



**Figure 6** | (A) Correlation of the interactome genes with pluripotency. The interacting genes screened in a siRNA assay using *POU5F1*-GFP reporter gene in hESCs were included for analysis. Distribution of Fav values (change of fluorescence) for the interacting genes are compared to all 21122 genes screened in the original assay. (B) Analysis of differential expression of identified transcriptional regulators in hESCs and fetal lung fibroblasts.

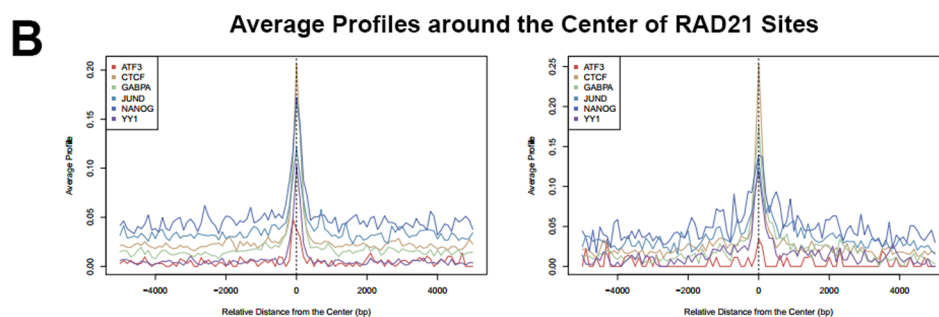
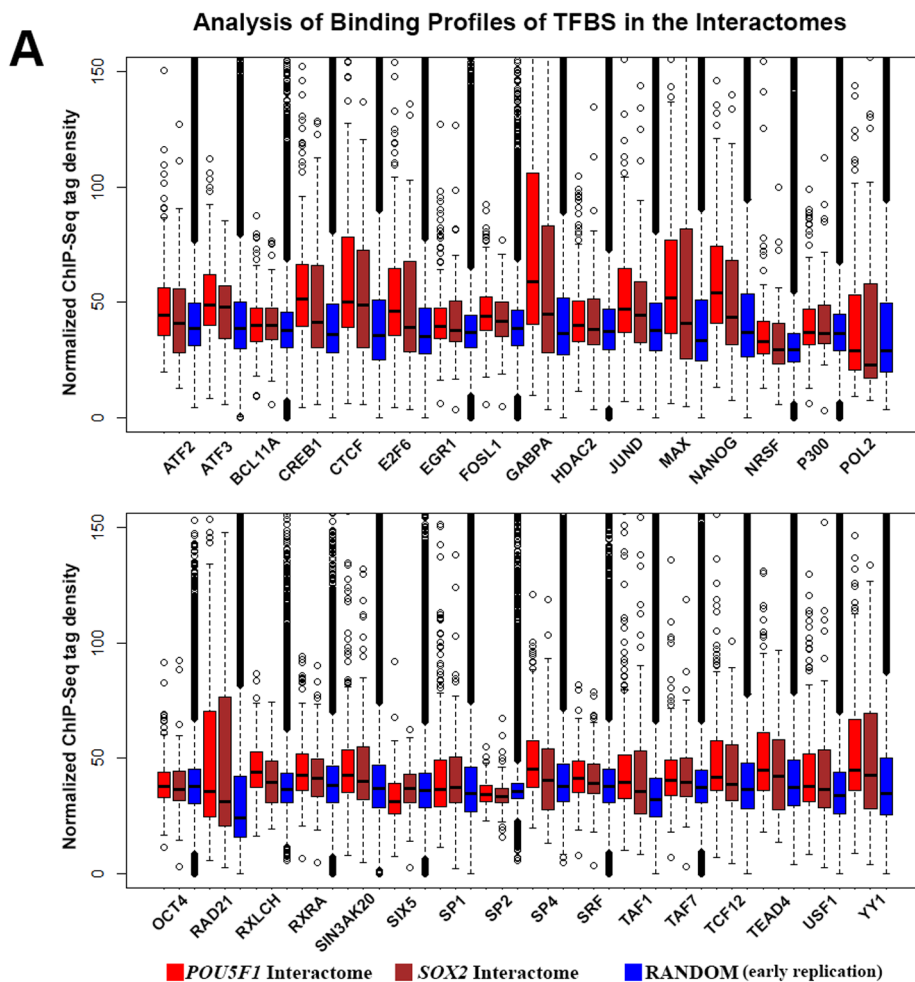
*SOX2* enhancer interactomes in hESCs. RAD21 is part of a cohesion complex known as a DNA chain crosslinker. Cohesin has been shown to mediate looping of the *Pou5f1* distal enhancer with the *Pou5f1* gene promoter in mouse ES cells<sup>40</sup>. Consistent with a previous report that Rad21 cooperates with Ctf and other transcription factors in maintaining mouse ES cell identity<sup>40</sup>, RAD21 sites in both *POU5F1* and *SOX2* interactomes in hESCs are co-occupied by transcription factors. Aggregation plots clearly illustrate the peak signals of ATF3, CTCF, GABPA, JUND, NANOG and YY1 binding profiles at the center of RAD21 sites in *POU5F1* and *SOX2* interactomes (Figure 7B). Knockdown of *Gabpa* in mouse ES cells is known to reduce *Oct3/4* expression, whereas overexpression of *Gabpa* maintains expression of *Oct3/4* even without LIF in mouse ES cell culture<sup>41</sup>. In a genome-wide siRNA screen in hESCs<sup>38</sup>, knockdown of most of the enriched proteins, such as RAD21, CTCF, GABPA, JUND, NANOG and YY1, significantly reduced the expression of a transgenic *POU5F1*-promoter linked to a GFP reporter, with Fav values of  $-1.50421$ ,  $-1.05828$ ,  $-1.44161$ ,  $-1.07731$ ,  $-1.82749$ ,  $-2.57204$ , respectively (within the top 25% of all of the genes screened). To further understand the chromatin–chromatin interactions, we applied DNA fluorescent *in-situ* hybridization (FISH) to visualize co-localization of two genomic loci at the single cell level. The *RARG* gene locus on chromosome 12 has been identified in the *POU5F1* interactome in hESCs. *RARG* was recently shown to play a very important role in pluripotency and can enhance iPS cell induction by two magnitudes<sup>42</sup>. Another locus on chromosome 12 that is not part of the *POU5F1* interactome was used as a control. The co-localization frequency between the *RARG* locus (chr12: 51751589–51956291) and the *POU5F1* locus (chr6: 31169844–31340561) was found to be 1.67-fold higher than the frequency between the control locus (chr12: 80380971–80564119) and the *POU5F1* locus (9.35% versus 5.61%,  $P < 0.05$ ,

Supplementary Fig. 3A). The higher contact frequency between the *RARG* and *POU5F1* loci is consistent with our sequencing data, and suggests that more stable interactions are formed between the two loci. Examination of the *RARG* and control loci (Supplementary Fig. 3B) revealed that there are more RAD21 and other transcription factor binding sites in the *RARG* locus with frequent co-localization. Coincidence of chromosome interaction frequency with RAD21-containing binding sites for multiple transcription factors suggests that RAD21 may cooperate with other transcription factors to organize long-range chromatin–chromatin interactions. Thus, RAD21, together with multiple transcription factors, is likely to contribute to higher-order chromosomal structure surrounding *POU5F1* and *SOX2* enhancers in hESCs as part of a pluripotency network. However, additional molecular studies are needed to confirm this hypothesis derived from 4C-Seq study.

## Discussion

To our knowledge, interactome maps of the *POU5F1* and *SOX2* enhancers in hESCs are presented for the first time, obtained by combining 4C with next-generation sequencing. The interactome data show that the putative enhancers of *POU5F1* and *SOX2* can physically associate with distal active genomic regions, which could be an important part of the regulatory network in hESCs.

Previous studies revealed that putative enhancer elements are marked with H3K27ac histone modification<sup>6</sup> and binding of p300 histone acetyltransferase<sup>7</sup>. The binding sites of p300 have accurately identified novel enhancers with tissue-specific activities in transgenic mouse assays<sup>43</sup>. Histone modifications at the enhancers are cell type-specific, and regulate genome-wide gene expression<sup>7</sup>. Classical enhancers are located immediately upstream of a promoter for regulating transcription, but more evidence indicates that enhancer



**Figure 7** | (A) Boxplots of different DNA-binding proteins around the enhancer interacting sites and random sites (from early DNA replication areas). ChIP-Seq tags within  $\pm 5$  kb from an interacting site were counted and normalized. (B) RAD21 co-localizes with other transcription factors in the interactomes. Average intensity profiles around RAD21 peak sites in *POU5F1* and *SOX2* interactomes in hESCs are plotted for transcription factors ATF3, CTCF, GABPA, JUND, NANOG and YY1.

elements distal to the target gene can loop with the target gene<sup>44</sup>. A recent study also revealed extensive promoter-promoter interactions, both intra-chromosomally and inter-chromosomally<sup>45</sup>. For *POU5F1* and *SOX2* putative enhancers marked with p300 in hESCs, we observed their preferential interactions with distal gene-rich regions that are associated with enhancer histone marks, suggesting that the regulatory effects of the two enhancers are not restricted to their nearby *POU5F1* and *SOX2* genes, and may involve long-range contact with other distal genes. Consistently, the interactomes are enriched with 5-hmC sites that are associated with active enhancers in hESCs<sup>36,46</sup>. Our observations of enhancer

interactomes support genome-wide long-range interactions between the enhancers and the targeted regions. These long-range interactions are likely to be dynamic and transient. For example, regarding the *RARG* locus on chromosome 12 that interacts with the *POU5F1* locus on chromosome 6, only around 10% of total cells exhibit a specific interaction between the two loci.

Chromatin–chromatin interactions construct a higher-order chromosomal structure that is part of a pluripotency network in ES cells. The *Nanog* gene locus in mouse ES cells is enriched with DNaseI hypersensitivity sites, facilitating pluripotency-related transcription factor binding which serves to organize a higher-order



chromosomal structure that may have functional roles<sup>47</sup>. DNA replication domains are also structured in the nucleus, with early and late DNA replication domains correlating nicely with different compartments of higher-order chromosomal structure in ES cells<sup>28</sup>. For *POU5F1* and *SOX2* enhancer mediated higher-order chromosomal structure, we observed strong association with early DNA replication domains, active enhancer epigenetic marks, and activated transcription of interacting genes. Furthermore, a portion of the genes targeted by the *POU5F1* enhancers is pluripotency-related and involved in transcriptional regulation in hESCs.

Our results also suggest that several DNA binding proteins might mediate the *POU5F1* and *SOX2* enhancer interactomes. We observed abundant inter-DNA chain linker cohesion protein RAD21 and looping protein CTCF together with pluripotency-related transcription factors such as GABPA, JUN, NANOG and YY1. RAD21 mediates DNA looping, particularly between enhancer and promoter regions in ES cells<sup>40</sup>. CTCF, one of the most studied insulator proteins, also mediates long-range interactions that are functionally important to pluripotent cells<sup>27</sup>. For the enhancer interactomes in pluripotent cells, RAD21 may complex with CTCF and other transcription factors, thus potentially bridging distal chromatin and mediating interactions between regions like enhancers and promoters. Co-localized transcription factors such as YY1 may facilitate ordered recruitment of chromatin modifying enzymes to regulate higher-order chromosomal structure<sup>48,49</sup>. Thus, pluripotency-related transcription factors may coordinate with structural proteins RAD 21 and CTCF to mediate *POU5F1* and *SOX2* enhancer interactomes in hESCs.

Collectively, we characterized *POU5F1* and *SOX2* putative enhancer interactomes in hESCs to understand long-range enhancer interactions with the targeted regions. However, the data from the next-generation sequencing are merely average observations across a large group of cells, therefore, the identified interacting regions should not be viewed as interacting with *POU5F1* or *SOX2* enhancer simultaneously. The low concordance between biological replicate data from sonication based 4C approach is similar to the results from ChIA-PET. Also consistent with other sonication based 4C studies reported in Yijun Ruan group, the identified sites are either in close proximity to the bait or distributed across the whole genome, which is different from enzyme-based 4C approach. Most likely, the sonication based 4C approach only captures strong chromatin-chromatin interactions. Distal chromosomal interactions are infrequent compared to proximal interactions in the nucleus, thus many of the detected distal sites associated with the “bait” might not be biologically meaningful. Therefore, statistical analysis is necessary to filter out random noise and identify enriched interacting regions. It is expected that with increasing sequencing depth, higher concordance between biological replicates can be achieved.

## Methods

**Cell culture and chromatin preparation.** H9 hESCs were obtained from the USC stem cell core center under feeder-free conditions. Around 2 million cells were cross-linked with 1% freshly prepared formaldehyde for 10 min at room temperature. The reaction was quenched by addition of ice-cold glycine to a final concentration of 0.125 M and incubated on ice for 5 min. The cell pellets were collected and re-suspended in 10 mL Triton X100 buffer (0.25% Triton X100, 10 mM EDTA pH = 8.0, 10 mM Tris-HCl pH = 8.0, 100 mM NaCl, 1× protease inhibitor cocktail) and rotated at 4°C for 30 min. The mixture was then centrifuged at 3000 rpm for 5 min at room temperature. The collected chromatin pellets were re-suspended in 600 μL SDS lysis buffer (1% SDS, 5 mM EDTA pH = 8.0, 50 mM Tris-HCl pH = 8.0, 1× protease inhibitor cocktail) and sonicated to an average size of 500 bp using a Branson Digital Sonifier. After a final centrifugation at 14000 rpm for 10 min at 4°C, the supernatant was collected and saved at −80°C.

**4C library preparation.** The chromatin fragments were diluted 5-fold in 10 mM Tris-HCl pH = 8, mixed with Triton X100 (1.7% final concentration) and rotated at 37°C for 2 h. The mixture was blunt-end repaired (NEB) for 30 min at room temperature and then ligated under diluted conditions (chromatin corresponding to ~70,000 cells per mL of ligation mixture) for 24 h at 4°C. Reversal of crosslinking was carried out at 65°C for 20 h with proteinase K (0.7 mg/mL final concentration). Protein-free DNA was extracted with phenol:chloroform (1:1) and precipitated in

70% EtOH, 0.3 M NaAcetate pH = 5.6 at −80°C. After centrifugation at 14000 rpm for 45 min, the DNA pellets were washed with 70% EtOH, air-dried and dissolved in 10 mM Tris-HCl pH = 8.0. The resuspended DNA was further treated with RNase A before purification using a MinElute Reaction Cleanup Kit.

The 4C libraries were amplified by nested inverse PCR using 2 sets of PCR primers and KOD hot-start DNA polymerase. Each of the two rounds of PCR ran 20 cycles for amplification. PCR products were purified using a MinElute PCR Purification Kit. In the final step, the DNA fragments were sonicated to an average size of 200 bp using a Covaris sonicator before Illumina next-generation sequencing.

**Generation and processing of 4C-Seq data.** The paired-end sequencing data from the Illumina HiSeq platform were processed as follows. 20 bp tags from the paired-ends were aligned to the hg18 assembly separately using BWA aligner<sup>50</sup>. Uniquely mapped paired tags were processed to capture 4C junction sites formed between the bait putative enhancer region and its interacting partners, classified as distal and proximal interactions. Distal interactions included intra-chromosomal junctions with genomic distance greater than 20 kb and inter-chromosomal junctions, whereas proximal interactions covered intra-chromosomal junctions with a genomic distance between 300 bp and 20 kb. One additional filter was used to remove the tags close to the reads enclosed by simple repeating elements (<http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/>). Finally, tags within a 100 bp range were merged as a unique interacting site. Singlet sites were removed before further analysis, following previous studies<sup>18,21</sup>.

**Statistical model to identify enriched interacting domains.** A statistical model was built to identify interacting regions that exhibit a higher frequency of interactions than expected from random collision<sup>21</sup>. For every interacting site  $i$  on chromosome  $W$  (length  $L_W$ ), the number of interacting sites within a certain window  $w$  with length  $l_w$  (we chose  $\pm 1$  Mb in the study) from the analyzed site was counted as  $C_{i,w}$ , and a z-score was calculated as an enrichment score:

$$z_i = \frac{(C_{i,w} - \mu_w)}{\sqrt{\mu_w \cdot (1 - p_w)}}$$

in which  $p_w = \mu_w/l_w$  ( $\mu_w$  is the expected number of interacting sites in window  $w$  on chromosome  $W$ ).

We further assigned statistical significance to each interacting site by a FDR-based approach. Briefly, we randomly permuted calculated z-score data for every chromosome 100 times, and chose interacting sites with a false discovery rate (FDR)  $\leq 5\%$  as positively interacting sites. FDR for each site was calculated by counting the number of randomly permuted Z-scores that are above experimentally determined Z-score. All the interacting sites within  $\pm 1$  Mb range of a positively interacting site were clustered as an enriched interacting domain. Overlapping interacting domains were further merged together.

**DNA-FISH.** FISH probes labeled with Cy5 and Alexa-488 dyes were prepared in-house. Probes mixed with human Cot-1 DNA and salmon sperm DNA were denatured for 7 min at 75°C. Competition was carried out for 30 min at 37°C. Formaldehyde fixed H9 cells grown on glass covers were dehydrated with 80%, 90% and 100% ice-cold ethanol sequentially, each for 5 min on ice. The air-dried samples were permeabilized in 0.2 N HCl/0.2% Tween 20 for 10 min at room temperature, followed by neutralization with CMF-PBS/0.2% Tween. The cells were then denatured in 70% formamide/2× SSC (pH = 7.2) for 10 min. at 80°C and dehydrated again with 80%, 90% and 100% ice-cold ethanol sequentially. The hybridization reaction was carried out at 37°C overnight and fluorescent images were taken on a Zeiss 510 Meta two-photon system. The percentage of co-localization between *POU5F1* and potentially interacting loci was compared with that between the *POU5F1* locus and a negative control region on chromosome 12 with roughly four-hundred cells counted in each case. Statistical analysis was performed using Fisher's exact test with two-tails.

**ChIP-Seq, RNA-Seq and DNA replication timing data.** The ChIP-Seq data for histone modifications and DNA binding proteins in H1 hESCs were retrieved from<sup>7</sup> and the ENCODE database (<http://genome.ucsc.edu/ENCODE>, GSE32465). The transcriptome data for H1 hESCs and AG04450 fetal lung fibroblasts were retrieved from the ENCODE database (<http://genome.ucsc.edu/ENCODE>). DNA replication timing array data for H9 hESCs was downloaded from the Replication Domain Database (<http://www.replicationdomain.com>). After data smoothing, segmentation of replication timing was performed by DNACopy (Bioconductor) using the parameters defined in the previous study<sup>51</sup>. The segments with mean replication timing ratio above zero are defined as early replication areas.

**Statistical analysis.** Statistical tests were executed using the R software suite (<http://www.r-project.org/>).

1. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861–72 (2007).
2. Bock, C. *et al.* Reference Maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. *Cell* **144**, 439–52 (2011).
3. Pan, G. *et al.* Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. *Cell Stem Cell* **1**, 299–312 (2007).





4. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–26 (2006).
5. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–60 (2007).
6. Creighton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**, 21931–6 (2010).
7. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–83 (2011).
8. Boyer, L. A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947–56 (2005).
9. Kim, J., Chu, J., Shen, X., Wang, J. & Orkin, S. H. An extended transcriptional network for pluripotency of embryonic stem cells. *Cell* **132**, 1049–61 (2008).
10. Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106–17 (2008).
11. Orkin, S. H. *et al.* The transcriptional network controlling pluripotency in ES cells. *Cold Spring Harb Symp Quant Biol* **73**, 195–202 (2008).
12. Bernstein, B. E., Meissner, A. & Lander, E. S. The mammalian epigenome. *Cell* **128**, 669–81 (2007).
13. Young, R. A. Control of the embryonic stem cell state. *Cell* **144**, 940–54 (2011).
14. Meshorer, E. & Misteli, T. Chromatin in pluripotent embryonic stem cells and differentiation. *Nat Rev Mol Cell Biol* **7**, 540–6 (2006).
15. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–93 (2009).
16. Dostie, J. & Dekker, J. Mapping networks of physical interactions between genomic elements using 5C technology. *Nat Protoc* **2**, 988–1002 (2007).
17. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–80 (2012).
18. Fullwood, M. J. *et al.* An oestrogen-receptor- $\alpha$ -bound human chromatin interactome. *Nature* **462**, 58–64 (2009).
19. Zhao, Z. *et al.* Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet* **38**, 1341–7 (2006).
20. Schoenfelder, S. *et al.* Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat Genet* **42**, 53–61 (2010).
21. Splinter, E. *et al.* The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. *Genes Dev* **25**, 1371–83 (2011).
22. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet* **38**, 1348–54 (2006).
23. Huang, P. Y. *et al.* Protocol: Sonication-based Circular Chromosome Conformation Capture with next-generation sequencing analysis for the detection of chromatin interactions. *Protocol Exchange* (2010).
24. Yeom, Y. I. *et al.* Germline regulatory element of Oct-4 specific for the totipotent cycle of embryonic cells. *Development* **122**, 881–94 (1996).
25. Liu, Z. & Garrard, W. T. Long-range interactions between three transcriptional enhancers, active  $\kappa$  gene promoters, and a 3' boundary sequence spanning 46 kilobases. *Mol Cell Biol* **25**, 3220–31 (2005).
26. Duan, H., Xiang, H., Ma, L. & Boxer, L. M. Functional long-range interactions of the IgH 3' enhancers with the bcl-2 promoter region in t(14;18) lymphoma cells. *Oncogene* **27**, 6720–8 (2008).
27. Handoko, L. *et al.* CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet* **43**, 630–8 (2011).
28. Ryba, T. *et al.* Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res* **20**, 761–70 (2010).
29. Rosenbloom, K. R. *et al.* ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic Acids Res* **38**, D620–5 (2010).
30. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948–51 (2008).
31. Shin, H., Liu, T., Manrai, A. K. & Liu, X. S. CEAS: cis-regulatory element annotation system. *Bioinformatics* **25**, 2605–6 (2009).
32. Kalhor, R., Tjong, H., Jayatilaka, N., Alber, F. & Chen, L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotechnol* **30**, 90–8 (2012).
33. Ram, O. *et al.* Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. *Cell* **147**, 1628–39 (2011).
34. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576–89 (2010).
35. Abed, J. A. & Jones, R. S. H3K36me3 key to Polycomb-mediated gene silencing in lineage specification. *Nat Struct Mol Biol* **20**, 244 (2013).
36. Szulwach, K. E. *et al.* Integrating 5-hydroxymethylcytosine into the epigenomic landscape of human embryonic stem cells. *PLoS Genet* **7**, e1002154 (2011).
37. Deng, X. *et al.* Evidence for compensatory upregulation of expressed X-linked genes in mammals, *Caenorhabditis elegans* and *Drosophila melanogaster*. *Nat Genet* **43**, 1179–85 (2011).
38. Chia, N. Y. *et al.* A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. *Nature* **468**, 316–20 (2010).
39. Huang da, W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**, 1–13 (2009).
40. Kagey, M. H. *et al.* Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**, 430–5 (2010).
41. Kinoshita, K. *et al.* GABP $\alpha$  regulates Oct-3/4 expression in mouse embryonic stem cells. *Biochem Biophys Res Commun* **353**, 686–91 (2007).
42. Wang, W. *et al.* Rapid and efficient reprogramming of somatic cells to induced pluripotent stem cells by retinoic acid receptor gamma and liver receptor homolog 1. *Proc Natl Acad Sci U S A* **108**, 18283–8 (2011).
43. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–8 (2009).
44. Nolis, I. K. *et al.* Transcription factors mediate long-range enhancer-promoter interactions. *Proc Natl Acad Sci U S A* **106**, 20222–7 (2009).
45. Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84–98 (2012).
46. Stroud, H., Feng, S., Morey Kinney, S., Pradhan, S. & Jacobsen, S. E. 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. *Genome Biol* **12**, R54 (2011).
47. Levasseur, D. N., Wang, J., Dorschner, M. O., Stamatoyannopoulos, J. A. & Orkin, S. H. Oct4 dependence of chromatin structure within the extended Nanog locus in ES cells. *Genes Dev* **22**, 575–80 (2008).
48. Lee, J. S. *et al.* Relief of YY1 transcriptional repression by adenovirus E1A is mediated by E1A-associated protein p300. *Genes Dev* **9**, 1188–98 (1995).
49. Austen, M., Luscher, B. & Luscher-Firzlaff, J. M. Characterization of the transcriptional regulator YY1. The bipartite transactivation domain is independent of interaction with the TATA box-binding protein, transcription factor IIB, TAFII55, or cAMP-responsive element-binding protein (CPB)-binding protein. *J Biol Chem* **272**, 1709–17 (1997).
50. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–60 (2009).
51. Hiratani, I. *et al.* Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS Biol* **6**, e245 (2008).
52. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639–45 (2009).

## Acknowledgments

We thank Dr. Gerhard Coetzee for stimulating discussions on this work. We thank all members of the Lu Laboratory for help on this work. This work was partially supported by CIRM grant to W. L. (RB1-01353).

## Author contributions

F.G., K.W. and W.L. designed the study; F.G. performed the experiments and analyzed the data; Z.W. and W.A. contributed to reagents/materials/analysis; F.G., K.W. and W.L. prepared the manuscript; All authors reviewed the manuscript.

## Additional information

**Supplementary information** accompanies this paper at <http://www.nature.com/scientificreports>

**Competing financial interests:** The authors declare no competing financial interests.

**License:** This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/3.0/>

**How to cite this article:** Gao, F., Wei, Z., An, W., Wang, K. & Lu, W. The interactomes of *POU5F1* and *SOX2* enhancers in human embryonic stem cells. *Sci. Rep.* **3**, 1588; DOI:10.1038/srep01588 (2013).