Supplementary Information, Data S1

Detailed Materials and Methods

Reagents

Inducers and inhibitors were used at the following concentrations: BMP4, 50 ng/ml (R&D System, Minneapolis, MN, USA); WNT3A, 10, 30, 50, 100 ng/ml (StemRD, Empowering Stem Cell R&D); ChIR99021, 10 μM (Stemgent); Noggin, 200 ng/ml (R&D System); IWR1-e, 10μM (Merck Millipore); IWP2, 10 μM (Stemgent); DKK1, 100 ng/ml (StemRD). Recombinant SFRP2 protein (R&D) was used at 200ng/ml.

Western blot and Co-IP

For Western blot and Co-IP experiments with endogenous proteins, cells were collected in the Co-IP buffer containing 50 mMTris-HCl pH 7.5 (Sangon, Shanghai, China), 150 mM NaCl (Sangon), 5 mM EDTA, pH 8.0 (Sangon), 10% glycerol (Sangon), 0.5% nonylphenoxypolyethoxylethanol (NP-40, Sangon), 0.1 mM PMSF and protease inhibitor cocktail tablets (EDTA-free) (Roche). After thorough lysis by vortex for 30 min in 4 ^{o}C, insoluble materials were precipitated by centrifugation at 20,000 g for 30 min at 4 ^{o}C, the supernatant cellular lysate was subjected to standard denaturing or immunoprecipitation of endogenous proteins with anti-SOX2 antibody and goat IgG, or anti-H2A.Z antibody and rabbit IgG at 4 ^{o}C overnight. Antibodies were then coated to magnetic dynabeads (Life Technologies) according to manufacture's recommendations for 2 hrs at 4 ^{o}C. One mg of total proteins (about 2 μg/ml) and 5μg antibody (about 10μg/ml) was used for each Co-IP. Immuno-precipitated complexes were washed with the cold Co-IP lysis buffer for 3

times, resuspended in SDS sample buffer, and subjected to SDS-PAGE and western blot analysis. A list of antibodies used in Western blot and Co-IP are listed in Table S3.

RNA extraction, cDNA synthesis and real-time quantitative PCR (qPCR)

Total RNA was isolated in TRIzol reagent (Life Technologies) and reverse transcribed into cDNA using a FastQuant RT kit according to manufacture's instructions (Tiangen). Real-time qPCR was performed on an ABI PRISM 7900 Fast Real-Time PCR system (Applied Biosystems), using a FastStart Universal SYBR Green Master mix (Roche) according to manufacture's recommendations. GAPDH was used for normalization. A list of primers used in RT-qPCR is provided in Table S4.

ChIP-qPCR assays

ChIP assays were performed as previously described[1]. Briefly, cells were cross-linked with 1% formaldehyde for 10 min at room temperature and quenched by 2.5 M glycine. Chromatin DNA was sheared by sonication to an average size of 100-1000bp. Chromatin fragments were then immune-precipitated with antibodies listed in Table S3. Approximately 6 million cells and 5-10 μg antibody was used for each ChIP. Input DNA and antibody enriched chromatin DNA were purified using a QIAquick PCR purification kit (QIAGEN) and quantified using a Quant-iTPicoGreen dsDNA Assay Kit (Invitrogen/Life Technologies, Grand Island, NY, USA). Purified DNA was subjected to library construction and subsequent sequencing or examination of site specific enrichment. To evaluate site-specific enrichment, qPCR

assays were carried out using the same amount of input DNA and immune-precipitated DNA, results were shown as relative enrichment to input. All primers used in ChIP assays are listed in Table S5.

ChIP-seq and RNA-seq

For ChIP-seq library preparation, ChIP-enriched DNA samples were treated with end-repair of the DNA, adding ‘‘A” bases to the DNA, ligating sequencing adapters to DNA fragments, amplifying adapter-modified DNA by PCR and gel purification for ChIP-seq using a NEB Next Ultra DNA Library Prep Kit for Illumina. Purified libraries were sequenced on Illumina HiSeq 2500 platforms. For RNA-seq, cDNA libraries were constructed following the TruSeq™ RNA Sample Preparation Guide (Illumina). Briefly, total RNA was isolated with TRIzol reagent, and polyA RNA was isolated using RNA Purification Beads (Illumina). The mRNA was fragmented by incubation in the Elute, Prime, Fragment Mix at 94 ^{o}C for 8 min to obtain 120-200 bp inserts. The first strand cDNA was synthesized with SuperScript II Reverse Transcriptase (Invitrogen) using random primers, and Ampure XP beads were used to isolate the double-strand (ds) cDNA synthesized by the Second Strand Master Mix. The adapter was ligated to the A-Tailing fragment, and 12 cycles of PCR was performed to enrich those DNA fragments that have adapter molecules on both ends and to amplify the amount of DNA in the library. Purified libraries were quantified by Qubit® 2.0 Fluorometer and validated by Agilent 2100 bioanalyzer to confirm the insert size and calculate the mole concentration. Cluster was generated by cBot with the library diluted to 10 pM and then was sequenced on the Illumina Genome Analyzer IIx for 75

cycles. The library construction and sequencing was performed at the Shanghai Biotechnology Corporation.

Bioinformatics analysis

ChIP-Seq data for SOX2 in mouse ESCs and mouse NPCs (GSE35496), H3K4me1, H3K4me3, H3K27me3 and H3K27ac in human ESCs and human NPCs (GSE62193) and histone variant H2A.Z (GSE39237) in human ESCs were obtained from Gene Expression Omnibus (GEO), respectively, and reanalyzed using methods described below. Sequences were aligned using bowtie software (version: 1.0.1) to the human (UCSC hg19) and mouse (UCSC mm9) reference genome [2]. Only sequences that mapped uniquely to the genome with no more than two mismatches were kept for further analyses. When multiple reads mapped to the same genomic position, a maximum of one read mapping to the same position were used. Peak calling was performed using MACS2 (version: 2.0.10), a new version of MACS (Model-based analysis of ChIP-Seq) [3]. For the ChIP-seq data of transcription factor, a cutoff of q=0.01 was applied, while for the other datasets, a cutoff of q=0.1 with an additional parameter "--broad" were used. Bedtools (version: 2.20.1) was employed to perform overlap analysis [4]. Promoters were defined as the regions from 2kb upstream of the TSS to 2kb downstream of the TSS, whilst the remaining loci were classified as promoter distal regions. ToppFun function in Toppgene suite [5] was utilized to perform gene set enrichment analyses on the nearest genes matched to each SOX2 peaks by ChIP-Enrich [6] and P-values corrected by Benjamini-Horchberg method were used to indicate the level of enrichment. Cis-regulatory element annotation

system (CEAS) was used to provide statistics on ChIP enrichment at important genome features, such as promoter, exon, intron [7].

The cross species comparison of SOX2 binding sites in mouse and human were performed according to the previous methods[8]. The UCSC LiftOver tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver) was applied with minMatch= 0.1 to align the binding sites of SOX2 from mm9 to hg19 genome assemblies. Then the binding sites were lifted back to mm9 again and tested if they overlapped a minimum of 90% of the original region. Peaks met this criteria were kept in further analysis. If > 10% of the lifted mouse SOX2 peaks were overlapped by a human peak, it would be considered a retained peak in the mouse and *vice versa*. Mouse SOX2 binding sites that were not retained were then considered as specific sites to the mouse (lost in the human). Similarly, human SOX2 peaks not overlapped for 10% by one lifted mouse SOX2 peak were then considered specific to the human (lost in the mouse). For those binding sites lost in the human, if there was a human SOX2 binding site located ±10 kb (max 10 kb between the midpoints of the two peaks) from the lifted mouse peak, it was considered as a turnover event. The rest human peaks and lifted mouse peaks were identified as the unique peaks within their own respective species. The rest human peaks and lifted mouse peaks were identified as the unique peaks within their own respective species. The Conservation Plot tool from Cistrome [9] was used to calculate the UCSC PhastCons conservation score [10] and show average conservation score profiles around the centers of different batches of peaks. Targets of retained peaks were those nearest genes matched to each retained peaks. Genes

with targeting turnover peaks, but without retained peaks, were regarded as targets of turnover peaks, while those targeted by unique peaks, but without retained and turnover peaks, were defined as targets of unique peak.

Transcriptome reads from RNA-Seq experiments were mapped to the reference genome (hg19) by TopHat (version 2.0.12) [11] with default parameters. Differential expression analyses were performed using GFOLD (generalized fold change, version: 1.1.0) algorithm [12], exactly as outlined in the GFOLD manual. Briefly, differentially expressed genes were identified by the cutoff of an absolute GFOLD value at 0.5.

RNA interference with oligonucleotides and shRNA

For RNA interference with oligonucleotides, transfection was carried out using DharmaFECT1 Transfection Reagent (GE, Dharmacon) according to manufacture's instructions. For reverse transfection in hESCs, cells were dissociated into single cells and seeded in 6 well plates at a density of one million cells per well in Y27632 (Stemcell Technologies) containing mTeSR1. For forward transfection in hNPCs, cells were seeded in 6 well plates at a density of 800,000 cells per well. siRNAs used to knock down *SOX2* in hESCs and hNPCs were named si2-1: 5'-GCCCUGCAGUACAACUCCAUGACCA-3', purchased from Stealth RNA interference (RNAi) Duplex Oligonucleotides (Invitrogen); si2-2: 5'-CCATGGGTTCGGTGGTCAA-3' purchased from GenePharma. For *SOX2* and *SOX3* knock down in hESCs, we used two sets of oligos, including one specially designed Stealth RNA interference (RNAi) Duplex Oligonucleotides (Invitrogen) that simultaneously targets both *SOX2*and *SOX3*, named si2/3-1:

5’-CCUCCGGGACAUGAUCAGCAUGUAU-3’, and another set named si2/3-2 using si2-1 plus one SOX3 oligo (5’-AGCCAAGGAGUGAAUGGGAGAAACA-3’). Other siRNAs include siSFRP2-1: 5’-CGACATAATGGAAACGCTT-3’; siSFRP2-2: 5’-GCTCCAAAGGTATGTGAAG-3’; siNT: 5’-TAAGGCTATGAAGAGATAC-3’

Lentiviral packaging and virus infection, stable knock down cell line generation

For RNA interference with shRNA, two sets of shRNA duplexes against *WLS*, pGIPZ-shWLS-1: 5’-TGCTGTTTGGTGACATCCG-3’ and pGIPZ-shWLS-2: 5’-TGGACCTGGATGCTGCTGT-3’ were purchased fromGIPZ Human Rest of Genome shRNA Library Rev2012 (GE, Dharmacon). Non-targeting (NT) shRNA(pGIPZ-shNT) was used as a control. The sequences of all constructs were verified by DNA sequencing. Lentiviral packaging was conducted as previously described [13]. Colonies of hESCs maintained in the mTeSR on Matrigel-coated dishes were dissociated into single cells using Accutase one day before lentiviral infection and seeded in 6 well plate in Y27632 containing mTeSR1 at a density of 1 million cells per well. On day 2, hESCs were transfected with PGIPZ-shWLS and pGIPZ-shNT lentiviruses, respectively. The medium was changed 6 hrs post infection. Puromycin selection (1 μg/ml) was conducted 3 days post infection, uninfected cells with no puromycin resistance were eliminated 2 days after selection and the remaining cell population was observed for green fluorescence, pooled together and expanded to generate NT and *WLS* stable knock down cell lines. Puromycin selection was maintained throughout the experiment.

Immunostaining

For neuronal induction, H9 hESCs derived hNPCs were plated onto poly-L ornithine/laminincoated coverslips (Falcon). On neuronal induction day 7, cells were fixed with 4% paraformaldehyde and immunostaining was performed as previously described [14]. The confocal images were captured using a confocal microscope (TCS SP5; Leica Microsystems, Wetzlar, Germany). Antibodies used are listed in Table S3.

Flow Cytometry Analysis

Cells were incubated with Brdu (10 μM; BD Biosciences) for 2 hrs. Cell cycle analysis was performed using a BD Pharmingen™ FITC BrdU Flow Kit according to manufacture's instructions on an Accuri C6 flow cytometer (BD Accuri Cytometers, Ann Arbor, MI, USA).

Luciferase reporter assay

TOP/FOP luciferase reporter assay was performed to evaluate the activity of canonical Wnt signaling. H9 hESCs were transfected with 8XTOPFLASH or 8XFOPFALSH plasmids (Addgene) together with internal control pRL-TK plasmid (Promega) using the HP1 reagent (Roche). The medium was changed 8 hrs post transfection, and cells were then subjected to siRNA transfection for *SOX2* and *SOX3* knock down. The medium was changed 12 hrs post transfection. Cells were lysed 48 hrs later and luciferase activities were monitored with the Dual Glow Luciferase Assay System (Promega) according to manufacturer's instructions.TOP/FOP ratios were

calculated by the luciferase activity ratio of (TOP firefly /relina) /(FOP firefly /renila).

Cytoplasm and nuclear separation

Cells were dissociated by accutase and collected by centrifuging. Cell pellets were then subjected to the cytoplasm and nuclear separation using NE-PER™ Nuclear and Cytoplasmic Extraction Reagents (Thermo Scientific) according to manufacture's instructions.

Identification of SOX2-interacting proteins

Whole cell proteins from H9 hESCs and H9-derived NPCs were extracted using the Co-IP buffer, about 10 mg protein was used for each IP. Lysates were precleared using IgG before incubating with antibody against SOX2 or IgG at 4^{o}C overnight. The mixture was then incubated with Dynabeads protein G (Life Technologies) at 4 ^{o}C for 4 hrs to precipitate SOX2 and associated proteins. SOX2 containing protein complexes were separated by SDS-PAGE and visualized by Coomassie Brilliant Blue staining. The gels were cut into 10 pieces, and proteins in the gels were then sequenced by mass spectrometry in BIDMC Proteomics Center and Dana Farber/Harvard Cancer Center, Cancer Proteomics Core in Beth Israel Deaconess Medical Center at Harvard Medical School, USA. Protein Pilot 3.0 was used to acquire the mass spectrometry [15].

Supplemental References

1. Lee, TI, Johnstone, SE, and Young, RA. Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nat Protoc* 2006; **1**:729-48.

2. Langmead, B, Trapnell, C, Pop, M, and Salzberg, SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009; **10**:R25.

3. Zhang, Y, Liu, T, Meyer, CA*, et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 2008; **9**:R137.

4. Quinlan, AR and Hall, IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010; **26**:841-842.

5. Chen, J, Bardes, EE, Aronow, BJ, and Jegga, AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res* 2009; **37**:W305-11.

6. Welch, RP, Lee, C, Imbriano, PM*, et al.* ChIP-Enrich: gene set enrichment testing for ChIP-seq data. *Nucleic Acids Research* 2014.

7. Shin, H, Liu, T, Manrai, AK, and Liu, XS. CEAS: cis-regulatory element annotation system. *Bioinformatics* 2009; **25**:2605-6.

8. Schmidt, SF, Jorgensen, M, Chen, Y, Nielsen, R, Sandelin, A, and Mandrup, S. Cross species comparison of C/EBPalpha and PPARgamma profiles in mouse and human adipocytes reveals interdependent retention of binding sites. *BMC Genomics* 2011; **12**:152.

9. Liu, T, Ortiz, JA, Taing, L*, et al.* Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol* 2011; **12**:R83.

10. Siepel, A, Bejerano, G, Pedersen, JS*, et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 2005;

15:1034-50.

11. Trapnell, C, Pachter, L, and Salzberg, SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009; **25**:1105-1111.

12. Feng, J, Meyer, CA, Wang, Q, Liu, JS, Shirley Liu, X, and Zhang, Y. GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. *Bioinformatics* 2012; **28**:2782-2788.

13. Li, C, Yu, H, Ma, Y*, et al.* Germline-competent mouse-induced pluripotent stem cell lines generated on human fibroblasts without exogenous leukemia inhibitory factor. *PLoS One* 2009; **4**:e6724.

14. Xu, HM, Liao, B, Zhang, QJ*, et al.* Wwp2, an E3 ubiquitin ligase that targets transcription factor Oct-4 for ubiquitination. *J Biol Chem* 2004; **279**:23495-503.

15. Shilov, IV, Seymour, SL, Patel, AA*, et al.* The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. *Mol Cell Proteomics* 2007; **6**:1638-55.