Supplementary Information for

Genome-wide chromatin interactions of the *Nanog* locus in pluripotency, differentiation and reprogramming.

Effie Apostolou[*], Francesco Ferrari[*], Ryan M. Walsh, Ori Bar-Nur, Matthias Stadtfeld, Sihem Cheloufi, Hannah Taylor Stuart, Jose M. Polo, Toshiro K. Ohsumi[6], Mark L. Borowsky, Peter V. Kharchenko, Peter J. Park[#] and Konrad Hochedlinger[#]

* These authors contributed equally

Corresponding authors: K.H., khochedlinger@helix.mgh.harvard.edu; P.J.P., peter_park@harvard.edu

This file contains:

1. Supplementary Figures (S1-S5)
2. Supplementary Tables (S1-S6)
3. Supplementary Experimental Procedures
4. Supplementary References

Supplementary Figures.

Figure S1, related to Figure 1

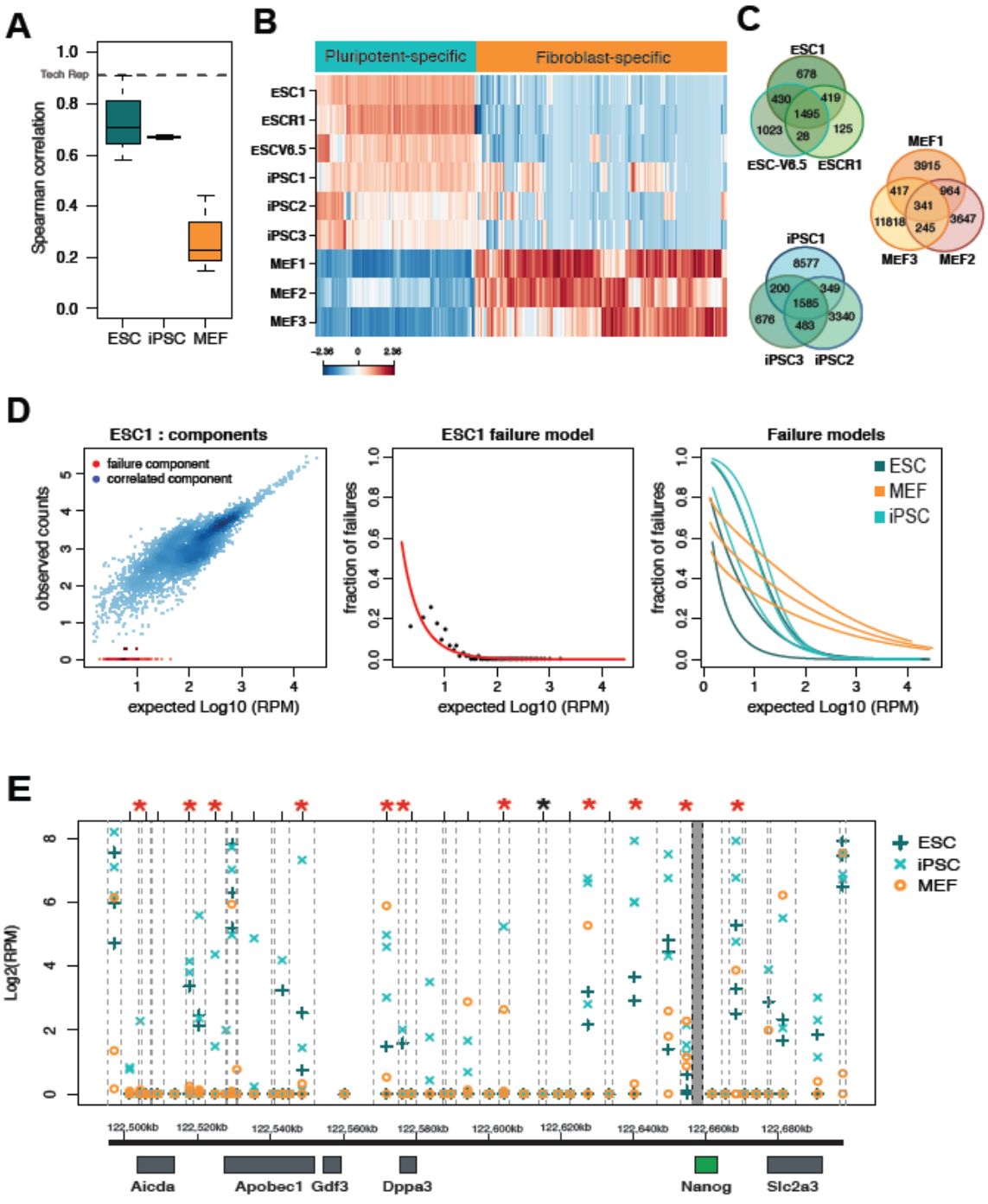

Figure S1.

(A) Boxplot showing the distribution of pairwise Spearman correlation coefficients, for the samples within each group of cells, after filtering noise from fragments with low intensity signal ("failure" component as estimated by our

modeling of interactions read count, see panel D). The dashed line indicates the level of correlation detected in 2 technical replicates (0.91). Whiskers extend to most extreme values within 1.5 times the inter quartile range (IQR) from the upper or lower quartile.

(B) Heatmap showing the relative change in normalized 4C-seq signal for each HindIII fragment selected as differential interaction between ESCs and MEFs. Columns refer to individual HindIII fragments and rows represent different 4C-seq samples. The iPSC samples clearly show a pattern concordant to ESCs. Color refers to scaled relative change across samples (z-score) of log transformed normalized 4C read counts.

(C) Venn diagrams showing the overlap of *Nanog*-interacting HindIII fragments within replicates for each cell type. Only the common interactions across all three replicates are used as conserved set in other analyses.

(D) Estimated reproducibility of interactions in different samples. The left panel compares one ESC sample (ESC1 line) to the expected results (average of other replicates) for single fragments read counts (Reads Per Million on $\log_{10}$ scale). A "Failure component" is modeled as a Poisson distribution (central panel for ESC1 line) to represent fragments not detected in one replicate but detected in others, and it provides an estimation of variability between replicates (right panel). As expected, the reproducibility of the interaction signals depends on the signal magnitude – fragments with high interaction signals tend to be more reproducible (*e.g.* >80% are reproducible at RPM=1000 level in all samples). Reproducibility varies considerably between samples, especially at lower signal magnitudes (*e.g.* 90%-20% at RPM=10). Some of this variability is likely due to experimental noise (e.g. variability in fixation, digestion and ligation efficiencies), however consistently lower reproducibility of MEF samples at high signal magnitudes suggests that biological heterogeneity is also an important factor in some cell types.

(E) 4C-seq normalized signal (Log transformed RPM counts) for individual HindIII fragments in the gene cluster around the *Nanog* locus. Vertical dashed grey lines mark the boundaries of individual fragments. One mark for each replicate is plotted for ESC, iPSC and MEF cell types. The red stars mark the positions that have been detected in a previous study (Lavasseur et al) by 3C where we observe 4C signal in at least one of our ESCs or iPSCs. The black star is a previously identified fragment where we don't see any signal.

Figure S2, related to Figures 1 and 2

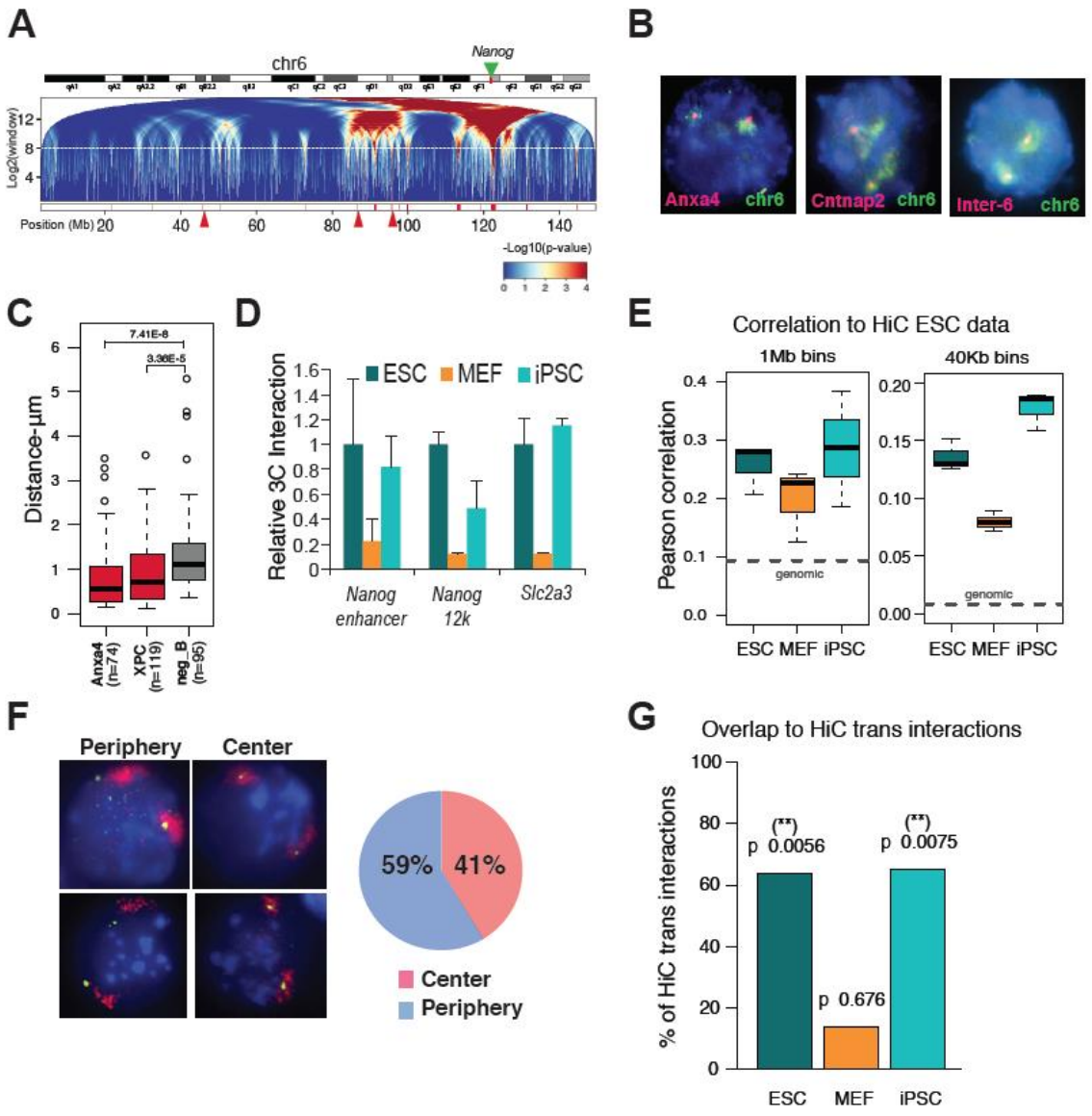

Figure S2.

(A) Broad intra-chromosomal interactions are represented as a domainogram for chromosome 6. The green arrow indicates the position of *Nanog*. The red ticks at the bottom of the domainogram mark the regions selected as center of broad interacting domains. The dashed horizontal white line indicates the maximum window size cutoff used to select interacting domains (p-value < 0.0001). The red arrows mark three of the regions we validated by DNA FISH (see panel E and

Figure 1D). The chromosome 6 domainogram for ESC1 is reported as representative of ESCs.

(B) DNA FISH confirming the specificity and the chromosomal position of each of the intra-chromosomal regions (Magenta Signals), which interact with *Nanog* in Figure 1D. All probes are indeed located within the chromosome 6, which is labeled in green.

(C) Boxplot for distances between the *Nanog* locus and three of the domains also shown in Figure 1E and 2D (n= number of measured nuclei). Whiskers extend to most extreme values within 1.5 times the inter quartile range (IQR) from the upper or lower quartile. For this boxplot, the distances were measured in higher resolution, allowing more accurate estimation of the closeness between colocalized signals (<0.25µm), which was usually overestimated in low resolution (>0.45 µm). Importantly, the distances between the tested interacting loci and *Nanog* were still significantly smaller compared to a negative control locus. P-values for Wilcoxon test are reported.

(D) 3C-PCR assay using primers specific for the *Nanog* promoter and three cis loci found to interact with it in our 4C-seq results. For each primer pair the PCR signal was calculated relative to the corresponding signal in ESCs ("Relative 3C Interaction"), after normalization with the PCR signal of primers designed on the bait (Table S6). Error bars indicate standard deviations (n=3 technical replicates). All 3C-PCR products were isolated and analyzed by Sanger sequencing.

(E) Correlation values between public mouse ESC HiC data (Dixon et al., 2012) and individual 4C replicates are reported as a boxplot. Whiskers extend to most extreme values within 1.5 times the inter quartile range (IQR) from the upper or lower quartile. A trend with higher correlation values in ESC/iPSC and lower values in MEFs is observed. Nanog centered cis interactions were extracted from published HiC data. Log transformed normalized data (observed over expected ratios) were compared to similarly normalized 4C data. Results with original

published bin size (40Kb) or larger bin size (1Mb) are shown. The actual correlation coefficients values are higher when using larger bin size, as expected.

(F) DNA FISH experiment on ESCs showing the position of the *Nanog* locus (Green signals) relative to chromosome 6 territory (Magenta cloud). Representative photos displaying *Nanog* at the periphery or outside of its chromosome territory (2 photos on the left) or in the center of the territory (2 photos on the right) are shown. A pie diagram quantifying the % of *Nanog* alleles located in the center or outside of the chromosome 6 territory (n=160 alleles) is reported.

(G) Barplot showing the overlap of *Nanog* trans interactions detected by m4C-Seq analysis and the published Hi-C data (Dixon et al., 2012). As published Hi-C data analyses were focused only on cis interactions, the normalized data matrix for trans interactions was re-computed using published normalization procedure by ((Yaffe and Tanay, 2011) and summarization at larger bin size (1Mb) as few reads are detected in Hi-C data for trans interactions. A set of trans interacting regions was identified using the top 100 bins with highest Hi-C signal for *Nanog* bait centered interactions. We found that more than 60% of these Hi-C trans interactions were overlapping m4C-Seq selected interacting fragments for ESCs and iPSCs. This overlap is statistically significant as assessed by random sampling of detectable 4C fragments (reported empirical p-values). In contrast, the overlap with our MEF 4C-Seq interacting fragments was low (15%) and not statistically significant.

Figure S3, related to Figure 3

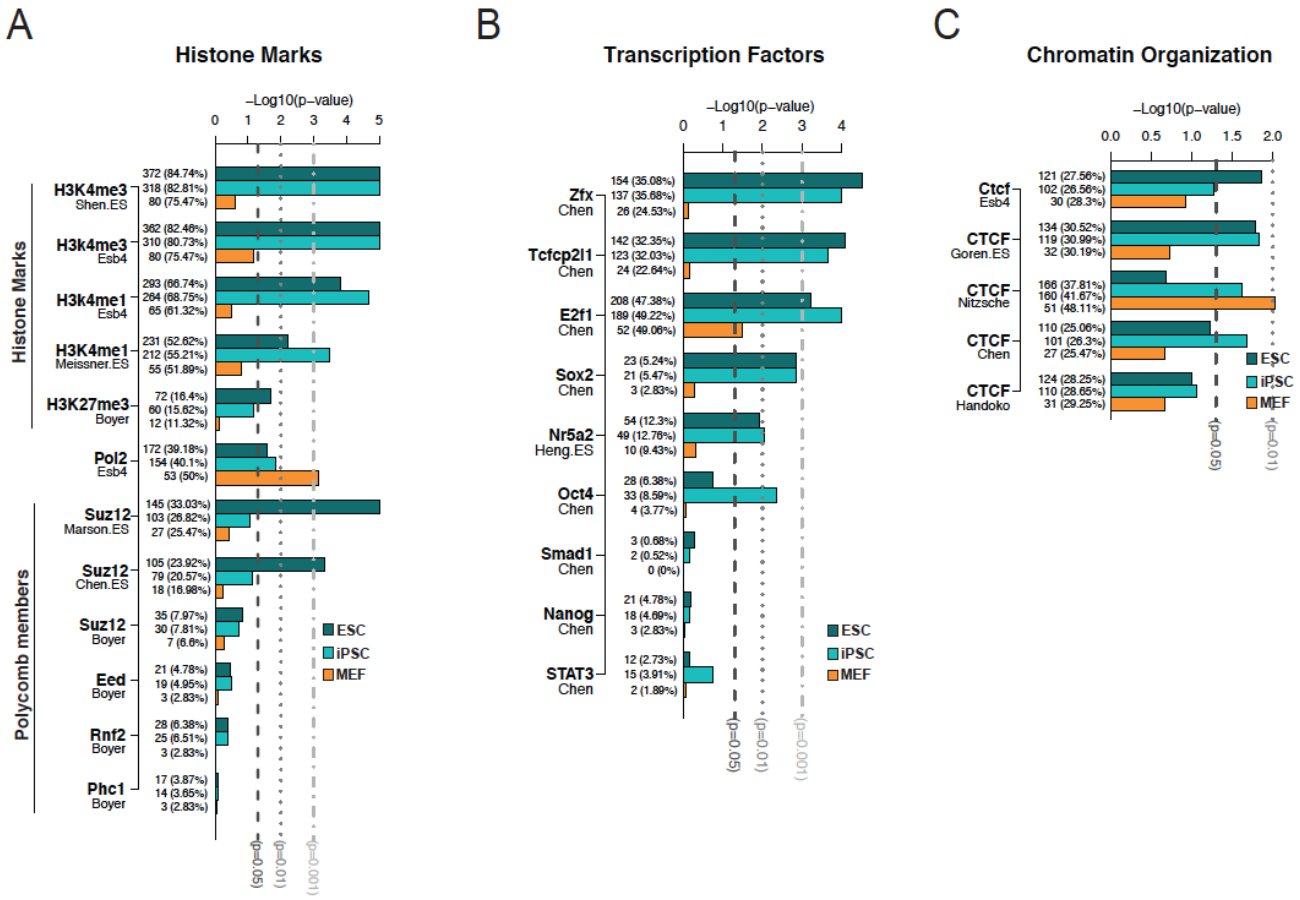

Figure S3.

A) Association of conserved *Nanog* interacting genes within each cell type (ESCs, iPSCs or MEFs) to active and repressive chromatin features. Additional datasets to those presented in Figure 3C.

(B) Association of conserved *Nanog* interacting genes within each cell type (ESCs, iPSCs or MEFs) to pluripotency transcription factors binding. Additional datasets to those presented in Figure 3D.

(C) Association of conserved *Nanog* interacting genes within each cell type (ESCs, iPSCs or MEFs) to binding of CTCF. Additional datasets to those presented in Figure 3E.

Figure S4, related to Figure 4

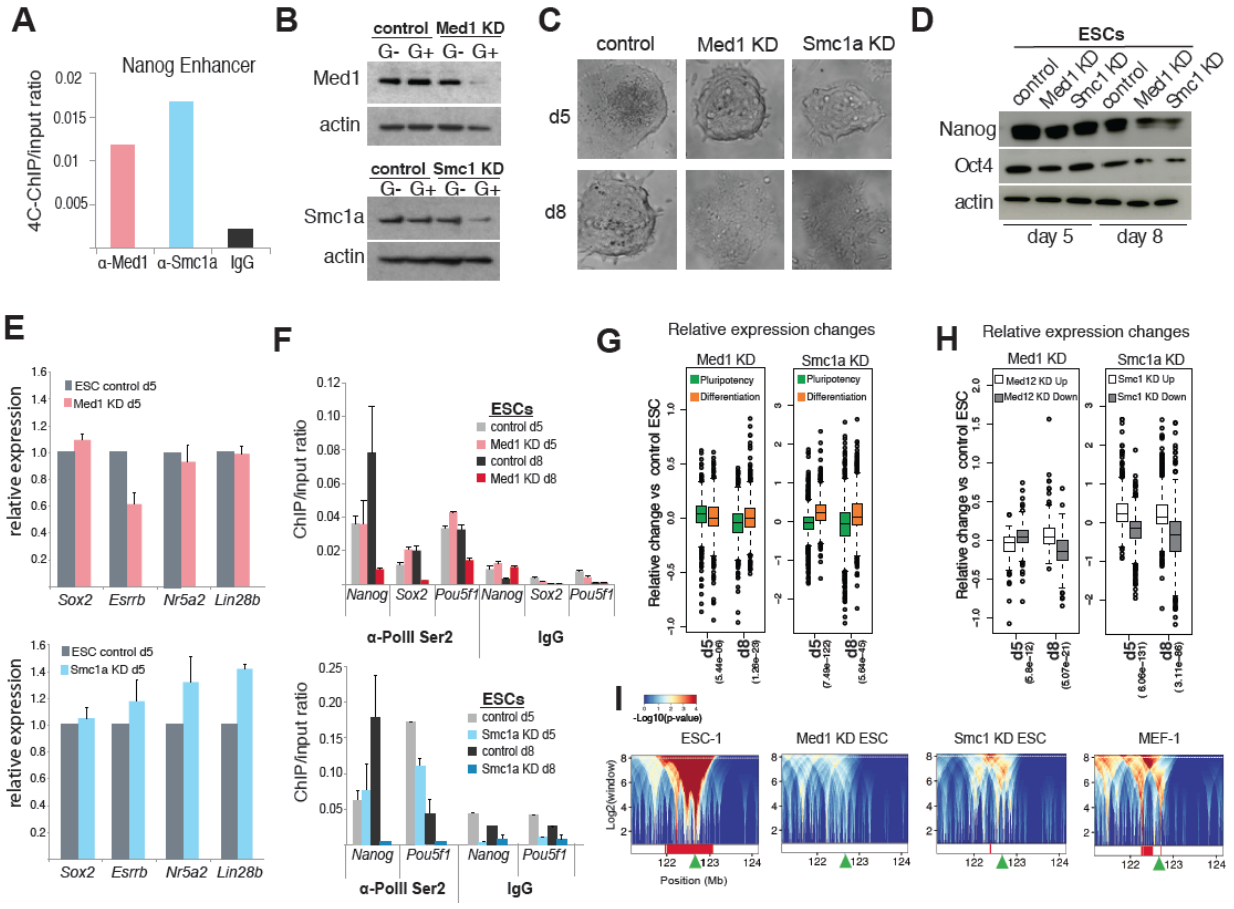

Figure S4.

(A) Efficiency of the 4C-ChIPs measured by qPCR immediately after the HindIII digestion and the Chromatin Immunoprecipitation step and before the intramolecular ligation. Primers specific for the *Nanog* enhancer were used, since both of these proteins have been described to occupy this region in ESCs. Immunoprecipitation with IgG was used as negative control.

(B) Western Blot testing the efficiency of knock-down of *Med1* and *Smc1a* five days after infection ES cells with pSico-shRNA-GFP. GFP positive and negative cells (control uninfected) were sorted by FACS and used for Western blot analysis with antibodies against Smc1a and Med1. Beta-actin was used as loading control.

(C) Representative brightfield images showing the morphology of ESC1 colonies 5 and 8 days, respectively, after infection with lentiviruses carrying *Med1* or *Smc1a* shRNAs or an empty control vector. Note that cells were still mostly undifferentiated when isolated on day 5 for 4C-Seq analysis.

(D) Western blot analysis of Nanog and Oct4 protein in extracts isolated from down ESC1 cells cultured for 5 or 8 days after infections with empty virus or shRNA or virus expressing shRNAs against either *Med1* or *Smc1a.* Beta-actin was used as loading control.

(E) RT-PCR analysis on control and Med1 (top graph) or Smc1a (bottom graph) knocked-down ESC1 cells on day 5 after infection. The mRNA levels of 4 pluripotency genes are tested and normalized to Gapdh levels. The graphs show the signal of each gene relative to the corresponding in control ESC1. Error bars indicate standard deviation (n=3 technical replicates)

(F) Chromatin immunoprecitiation (ChIP) experiments on control and Med1 (top graph) or Smc1a (bottom graph) knocked-down ESC1 cells (KD) after 5 or 8 days with an antibody against the phosphoSer2 RNA Polymerase II (anti-PolII Ser2), which indicates ongoing transcription. Primers spanning the Nanog, Oct4 and Sox2 (only for the Med1 KD) promoters were used. The qPCR signal was normalized to the input. IgG was used as negative control. Error bars indicate standard deviations of 2 experiments.

(G) RNA-Seq data after Med1 or Smc1a silencing (day 5 and day 8). Log fold change relative to control ESC is plotted for pluripotency and differentiation associated genes. Whiskers extend to most extreme values within 1.5 times the inter quartile range (IQR) from the upper or lower quartile. Wilcoxon test p-values are reported for the difference between each box plot pair.

(H) RNA-Seq data after Med1 or Smc1a silencing (day 5 and day 8). Log fold change relative to control ESC is plotted for previously published Med12 or Smc1a targets by (Kagey et al, 2010). Whiskers extend to most extreme values within 1.5 times the inter quartile range (IQR) from the upper or lower quartile.

Targets are separated in 2 lists expected to be down- or up-regulated, respectively, according to original publication. Wilcoxon test p-values are reported for the difference between each box plot pair.

(I) Domainogram detail for a 3Mb region around Nanog bait (green arrow) in ESC1, Med1/Smc1 KD in ESC1 and representative MEF sample. The broad interacting domain detected in ESCs was lost after Med1 or Smc1 KD, resembling the weaker interaction pattern in MEFs. Note that in all samples the m4C-seq reads from saturated bait fragments (self-ligation and undigested fragments) were.

Figure S5, related to Figure 5

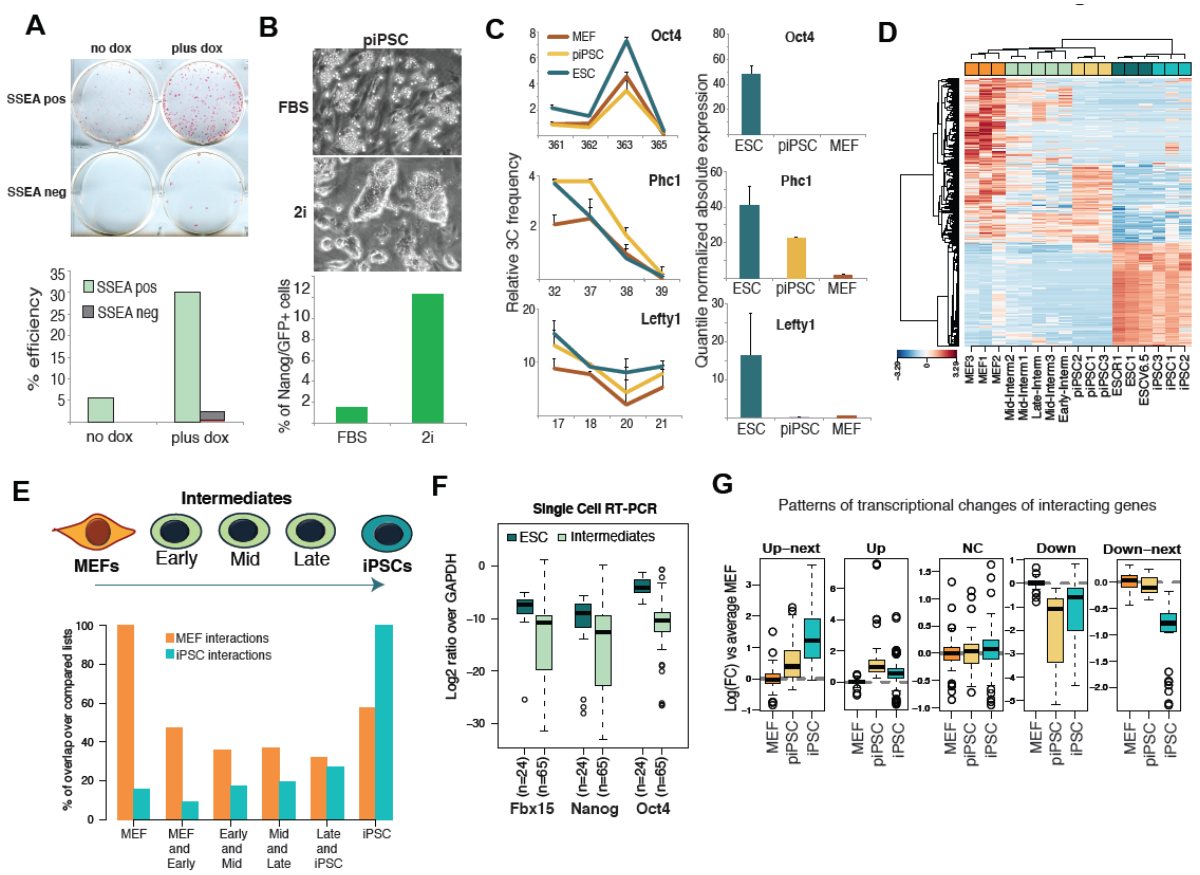

Figure S5.

(A) Alkaline phosphatase (AP) staining of iPSC colonies derived by either SSEA positive or negative cells isolated by MACs on day 6 of reprogramming. The sorted cells were plated on gelatin and cultured for 6 additional days on the presence (plus dox) or absence (nodox) of doxycyclin. The number of AP positive colonies divided by the number of plated cells gives the % efficiency plotted at the graph below.

(B) Representative brightfield images of partial iPSCs (piPSCs) under normal FBS conditions or under 2i (Gsk3+Mek1 inhibitors) conditions. Graph showing the percentage of Nanog-GFP expressing cells under each condition.

(C) *Left panels*: 3C-assay on ESCs, MEFs and partial iPSC (piPSC) testing the promoter-enhancer looping frequency of *Phc1*, *Oct4* and *Lefty1* genes. The protocol and the primers used for this analysis are described in Kagey et al (2010). The 3C-PCR signal was normalized to a gene desert control (See Experimental procedures). *Right Panels*: Graphs showing the normalized absolute expression of the 3 genes in the indicated cell types as were reported in (REF). Error bars represent the standard deviation of 3 different biological replicates.

(D) Heatmap showing the relative change for the set of 4C fragments selected as differential interactions between ESCs and MEFs across five groups of samples: ESCs, iPSCs, SSEA1+ intermediates, partially reprogrammed iPSCs (piPSC) and MEFs. Colors in the heatmap represent scaled relative change across samples (z-score) of log transformed normalized 4C read counts.

(E) *Nanog* interacting genes common between sequential time points of reprogramming time course are selected. In addition, *Nanog* interacting genes conserved in MEFs or iPSCs are used as references for start and ending points of the reprogramming process. The percentage over reference start and end interactions lists are plotted for each intermediate related lists.

(F) Boxplot showing the distribution of mRNA levels of 3 different pluripotency genes in single ESCs or single SSEA positive cells after normalization to the average *Gapdh* levels for each cell type. The number of examined cells is indicated. Whiskers extend to most extreme values within 1.5 times the inter quartile range (IQR) from the upper or lower quartile.

(G) Gene expression data in microarray dataset for MEF, piPSC and iPSC cells. Relative change VS average of MEF replicates is plotted for selected gene sets reporting data from individual replicates. Gene sets are defined from the 4C-gene level interactions gained in the transition between MEF and piPSC as described in Figures 5E-F. Whiskers extend to most extreme values within 1.5 times the inter quartile range (IQR) from the upper or lower quartile.

Figure S6, related to Figure 6

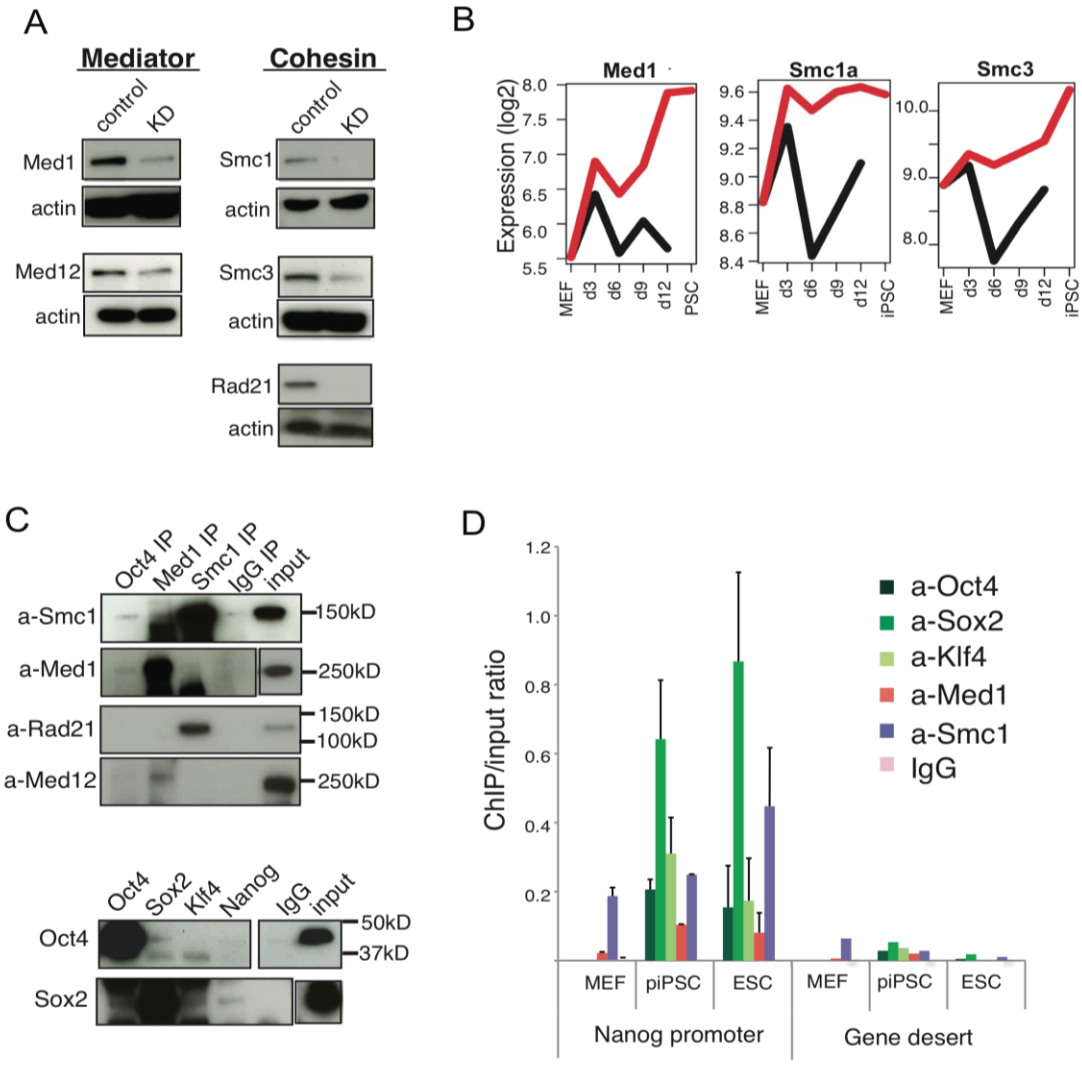

Figure S6.

(A) Western Blot testing the efficiency of knock-down of 2 mediator subunits (*Med1* and *Med12*) and 3 Cohesin subunits (*Smc1a, Smc3 and Rad21)* three days after infection of reprogrammable OKSM-MEF cells with pSico-shRNA-GFP or empty vector (control). Beta-actin was used as loading control.

(B) mRNA levels of Med1 and Smc1a during reprogramming as quantified by microarray analysis in SSEA1+ intermediates (Polo et al., 2012). Note that even

though Med1 is expressed both in MEFs and iPSCs, its significantly upregulated during reprogramming.

(C) Oct4, Med1 and Smc1a protein immunoprecipitation (IP) on piPSC protein extracts. IgG was used as negative control. Smc1, Med1, Rad21 and Med12 were tested as potential protein interaction partners by Western blot. Input was loaded as positive control. Oct4 seems to interact with both Med1 and Smc1. Rad21 and Med12 were pulled down, as expected, together with Smc1a (Cohesin complex) and Med1 (Mediator complex), respectively. The efficiency and specificity of IPs in ESCs were confirmed by the detection of known pluripotency protein-protein interactions (bottom panel), like Oct4-Sox2 interaction.

(D) Chromatin Immunoprecipitation (ChIP) experiments showing the binding of the reprogramming factors Oct4, Sox2 and Klf4 as well as the Med1 and Smc1a on the Nanog promoter in reprogrammable fibroblasts (MEFs), partial iPSCs (piPSC) and ESCs. All the ChIP-qPCR signals are normalized to the input. IgG and primers spanning a gene desert region were used as negative controls. Error bars indicate standard deviation (n=3)

Supplementary Tables.

Table S1, related to Figure 1: Common *Nanog*-interactions within each cell type (excel file)

Table S2, related to Figure 2: Differential *Nanog*-interactions between each pair of cell types (excel file)

Table S3, related to Figure 3: Lists of datasets used for the association studies (excel file)

Table S4, related to Figure 4: Med1 and Smc1a dependent *Nanog*-interactions (excel file)

Table S5, related to Figure 5: List of HindIII fragments found to interact with Nanog locus specifically in partial iPSC (excel file)

Table S6, related to Figures 1, 2, 4, 5 and 6: List of Primers used in this study (excel file)

Supplementary Experimental Procedures

Cell Culture

ESCs, established iPSCs and partial iPSCs were cultured on irradiated feeder cells (Global Stem) in KO-DMEM (Invitrogen) supplemented with L-Glutamine, penicillin-streptomycin, nonessential amino acids, β-mercaptoethanol, 1000 U/ml LIF ("ESC media") and 15% fetal bovine serum (FBS) (HyClone). MEF cultures were established by trypsin-digestion of midgestation (E13.5-15.5) embryos isolated from the "reprogrammable" mouse strain (Stadtfeld et al., 2010b) followed by culture in DMEM supplemented with 10% FBS, L-Glutamine, penicillin-streptomycin, non-essential amino acids and β-mercaptoethanol.

Cellular reprogramming

MEFs harboring M2-rtTA in the ROSA26 locus and tetO-OKSM in the collagen Ia1 locus (Stadtfeld et al., 2010b) were seeded on gelatinized plates in ESC media containing 15% FBS, 1ug/ml doxycycline and 50ug/ml ascorbic acid. On day 6 (mid) and 9 (late) after dox induction, SSEA+ cells were isolated by magnetic-activated cell sorting (MACS) after incubation with anti-SSEA1 microbeads (Miltenyi biotech) using the positive selection program on an AutoMACS cell separator according to the manufacturer's instructions. The purity of all isolated cell fractions was confirmed by flow cytometric analysis using an LSRII machine (BD). For the 48hr time point (early) bulk population was used instead. The partial iPSC were generated by infection of MEFs with constitutive retroviruses bearing the reprogramming factors (Maherali et al., 2007)

shRNA virus production and infection

The shRNA lentiviruses for Med1 and Smc1a, were designed according to a previous study (Kagey et al., 2010) and cloned into a different vector (Addgene-pSicoR-GFP). Briefly, 293T cells were cotransfected with packaging plasmids and either emplty pSicoR-GFP as control or vector containing the shRNA

sequence against Med1 or Smc1a using Fugene (Roche) transfection reagent. Viral supernatants were harvested between 48 and 72 hours after transfection and concentrated by ultracentrifugation at 20,000 rpm for 1.5 hours at 4^{o}C. Viral concentrates were resuspended in PBS and stored at -80^{o}C. Transduction of MEFs and ESCs was carried out in the corresponding media containing 5µg/ml polybrene followed by centrifugation for 2150rpm for 35min. GFP positive cells (infected) were isolated by FACS after 3 days and tested for the shRNA efficiency by western blot. All the shRNA sequences used for this study are shown in Table S6.

Reprogramming of knockdown MEFs

Reprogrammable (tetO-OKSM) MEFs were transduced with Lenti-virus containing shRNA for each of the tested Mediator and cohesin factors (see Table S6 for shRNA sequences). After 2 days 20K of the transduced MEFs were plated on gelatin in ESC medium supplemented with Vitamin C and doxycyclin. Nine days later some of the cells were analyzed by LSRII for the presence of the following surface markers (Thy1, SSEA1 and EpCam). On day 12 doxycyclin was removed and 4 days later Alkaline Phosphatase (AP)-staining was performed for the scoring of iPSC formed colonies.

RNA-seq library preparation

Total RNA from fibroblast infected with scrambled control shRNA or shRNAs targeted against the cohesin complex subunits was DNAse treated and purified using Qiagen RNeasy mini kit (according to manufacturer's instructions). RNA quality was assessed using an Agilent 2100 Bioanalyzer. RNA-seq libraries were prepared with the Truseq RNA sample preparation v2 kit. Briefly, total RNA was polyA selected using poly-T oligo-attached magnetic beads, fragmented and subject to first strand cDNA synthesis using random primers and superscript II (Invitrogen). This is followed by a second strand cDNA synthesis using polymeraseI and RNAseH. The resulting cDNA fragment were end repaired, dA-tailed at their 3'end and ligated to illumina's multiple indexing adapters. Libraries were then PCR amplified, quantified using Qubit fluorometer and their quality

assessed using an Agilent 2100 Bioanalyzer. Libraries were pooled and sequenced on Illumina HiSeq 2000 platform. On average 20 million reads were generated per library.

Protein Co-immunoprecipation (IP).

10-20 millions reprogrammable MEFs before and after 48hr of doxycyclin induction were harvested and used for the protein pull downs. 30-50 millions of partial iPSCs and ESCs V6.5 were used after preplating of cells to eliminate feeder contamination. The cells were lysed to 100-400 IP low salt buffer (50mM This-HCl pH 7.4, 100mM NaCl, 0.05% Triton, 5% glycerol, 1mM EDTA, 1mM DTT and protease inhibitors) for 20 min on ice, followed by 5 cycles of sonication (30'' on, 30'' off). The debris was removed by centrifugation and the protein concentration was estimated by Qbit. 1-3ug/ul. The protein extracts were diluted in a final concentration of 1-3ug/ul and precleared with Protein-G agarose beads for an hour at 4oC. 1-3mg of precleared protein extracts were then incubated with 5ug of antibody over night at 4oC. Next day, 10ul of preblocked (with BSA) Protein-G DynaBeads were added per reaction for 2-3 hours, followed by extensive washes 2 times with the IP buffer and 2 times with IP medium salt Buffer (300mM NaCl). The beads with the Immunoprecipitated proteins were finally resuspended in Leammli Buffer and used for Western blot. The antibodies used for this study were: Med1 (Bethyl Laboratories), Smc1 (Bethyl Laboratories), Oct4 (Santa cruz for Western and R&D for IP), Sox2 (R&D), Klf4 (R&D), Nanog (Bethyl Laboratories), actin-HPRT (abcam), Med12 (Bethyla Laboratories), Smc3 (abcam), Rad21 (Santa-Cruz).

Chromatin Immunoprecipitation

Cells were fixed with 1% formaldeyde for 10 minutes at room temperature (RT) and then lysed in 1ml lysis buffer (50mM Tris-HCl, pH 8.0, 10mM EDTA, 1% SDS, protease inhibitors) for 20 minutes on ice. The lysate was split into three tubes and sonicated using Bioruptor for five times five minutes at high intensity (30'' on/30'' off). After 10 minutes centrifugation, the supernatant was precleared

for 1 hour at 4ºC with agarose beads preblocked with BSA (1ug BSA for 10ul beads) in IP Buffer (50mM Tris-HCl, pH8, 150mM NaCl, 2mM EDTA, 1% NP-40, 0.5% Sodium Deoxycholate, 0.1% SDS, protease inhibitors). 100ul of precleared chromatin per reaction diluted in 1ml IP Buffer in presence of 2ug antibody were used for each reaction according to manufacturer's protocol. The antibodies used were: Oct4 (R&D), Sox2 (R&D), Klf4 (R&D), Med1 (Bethyl Laboratories), Smc1 (Bethyl Laboratories), IgG (abcam), PolII phospho-Ser2 (abcam). The primers used for the qPCR analysis are listed in Table S6.

Western blot analysis

Cells were harvested and lysed in 1x RIPA buffer. Protein concentration was measured using BCA protein Assay reagent (ThermoScientific). 10µg of cell extract was loaded and western blot analysis was performed using standard protocols and probed with the following antibodies: anti-Med1 and anti-Smc1a (Bethyl laboratories), anti-Med12 (Bethyl laboratories), anti-Smc3 (abcam), anti-Rad21 (santacruz)

RNA isolation and RT-PCR analysis

Cells were harvested and used for RNA isolation with the miRNeasyMini Kit . cDNA was produced with the Transcriptor First Strand cDNA Synthesis Kit (Roche). Real-time quantitative PCR reactions were set up in triplicate with the Brilliant III SYBR Green QPCR Master Mix (Stratagene) using the primers in Table S6. Reactions were run on a Mx3000P QPCR System (Stratagene) with 40 cycles of 30 seconds at 95^{o}C, 30 seconds at 58^{o}C and 30 seconds at 72^{o}C.

Modified 4C-seq analysis (m4C-seq)

4C was performed as has been previously described (Schoenfelder et al., 2010) with some modifications. Briefly, 5-10*10^6 cells were fixed in 2% formaldeyde for 10 min in RT. After quenching with 0.125M glycine and washes with ice-cold PBS the cells were lysed in 5ml Lysis Buffer (10 mM Tris-HCl pH 7.5; 10 mM NaCl; 5 mM $MgCl_2$; 0.1 mM EGTA; 1x complete protease inhibitor, 0.2% NP-40)

for 10 min on ice. The pellet was resuspended in 500ul 1.2 X Buffer 2 (NEB) with 0.3% SDS for 1hr in 37*o*C and then 2% Triton X-100 for an additional hour, before the addition of 400U HindIII and incubation >16hr at 37^{o}C. After confirming the digestion efficiency , the enzyme was inactivated in 1.2% SDS in 65^{o}C for 20min. Then, the HindIII digested samples were diluted in a final volume of 7 ml of 1X T4 ligase Buffer with 1% Triton-X-100 for 1hr in 37^{o}C. Ligation followed with the addition of 7ul of T4 ligase (20U/ul) and incubation for >16hrs at 16^{o}C and 1hr at RT. After testing the ligation efficiency we reversed the crosslinks by adding 300ug of proteinase K and incubating for 6hr at 65^{o}C. Subsequently the RNA was removed using 300ug of RNase for 1hr at 37^{o}C. Extensive phenol/chloroform extraction was followed by EtOH precipitation with 2ug glycogen and 0.3M acetic sodium in -80^{o}C for more than 20hr. The pellet was resuspended in 1X Buffer 4 (NEB) plus BSA and it was digested using 7ul NlaIII enzyme for 4-5hr at 37^{o}C, followed by 1hr incubation with CIP anzyme (NEB). Phenol/phloroform exctraction and ethanol precipitation followed after confirmation of the difestion efficiency. The pellet was resuspended in water and shared into 5 tubes for NlaIII adapter (Illumina) ligation at 16oC for 16hrs. After ethanol precipitation the pellets were resuspended in 50ul total and 10ul were used for LM-PCR for 15 cycles with the NlaIII big adapter and a biotinylated Nanog primer (See Table S6). The primers were removed with Qiagen PCR miniElute kit and the biotinylated products were purified using Streptavidin M-280 Dynabeads (Invitrogen) according to the manufacturer's instructions. After elution in 25ul, 1/5 was used as template for the final PCR with the Illumina adapters (See Table S6) for 25 cycles. The PCR reactions were loaded in 1.8% agarose gel in 1XTAE and the products between 150-500 were isolated and exctracted form the gel (Qiagen Minelute) creating the 4C libraries which were sequenced in GAII or HiSeq sequencers.

Bioinformatics analyses of 4C-seq and associations to public datasets: See Supplementary Experimental Procedures.

4C-ChIP-seq analysis

Cells were treated as for 4C-seq with the exception that an immunoprecipitation step was included. Specifically, the HindIII digested chromatin was diluted in IP Buffer (50mM Tris-HCl, pH8, 150mM NaCl, 2mM EDTA, 1% NP-40, 0.5% Sodium Deoxycholate, 0.1% SDS, protease inhibitors) in the presence of 2ug antibody (anti-Med1, anti-Smc1a, Bethyl laboratories). After overnight incubation at 4^{o}C protein A Dynabeads were added for 3 hours followed by washes according to Millipore's ChIP protocol. 1/10 of the beads were used to validate the ChIP efficiency. The rest of the beads were then diluted in 1ml 1X T4 ligation buffer with the addition of 3ul T4 ligase (20U/ul) and incubated for >16hrs at 16^{o}C, followed by all the steps described in "4C-seq" method.

3C analysis

Independent cell preparations were treated as for the 4C-seq protocol until the NlaIII adapter ligation step. Then PCR followed using the Nanog specific primer and NlaIII Adapter for 18 cycles. The purified PCR products were diluted 1:10 in water and 5ul were used as template for quantitave PCR (qPCR) using Brilliant III SYBR Green QPCR Master Mix (Stratagene) and primers specific for Nanog and each of the candidate loci. The primers for each candidate locus were designed between the closest NlaIII site and the HindIII site(s), which were found to interact with Nanog based on the 4C-seq results. Primers for the amplification of the "bait" sequence were used as an internal normalization control for each of the samples. The primers used for this study are shown in Table S6. Reactions were run on a Mx3000P QPCR System (Stratagene) with 40 cycles of 30 seconds at 95^{o}C, 30 seconds at 58^{o}C and 30 seconds at 72^{o}C.

For the Oct4, Phc1 and Lefty1 3C-looping assays 10 million cells (MEFs, piPSCs and ESC V6.5) were digested with HaeIII and MspI respectively. The protocol and the primers used for this analysis are described at Kagey et al, 2010 and Table S6.

3D-DNA FISH and image analysis

ES and MEF cells were trypsinized into single cells and cytospun on glass slides prior to paraformaldehyde fixation. 3D-DNA FISH analysis was performed as described (Xu et al., 2006) using BAC clones labeled by Digoxygenin or Cy3 nick translation (Roche). For this study we used the following BACs: RP23-19018 (Nanog), RP23-192A14 (Cntnap2), RP23-451F4 (Anxa4), RP24-335C14 (intergenic chr6), RP24-226F3 (intrachromosomal negative control A), RP23-393F8 (intrachromosomal negative control B), RP23-358E22 (Nfya), RP23-129I5 (Ncor2), RP23-26D8 (Rprml), RP23-268C14 (negative control for chr 17), RP23-307K22 (negative control for chr5), RP23-451D17 (negative control for chr11), RP23-53E13 (XPC), RP23-68B4 (Uggt2) and RP23-343L5 (negative control for chr14). Z-sections were captured with 0.2 mm intervals using a Nikon 90i eclipse microscope. ND software was used to analyze probe-probe distances.

Data Availability. All sequencing data will be made publicly available in SRA dataset with accession number SRA051554.

m4C-seq data analysis

Sequencing data processing and quality control

All m4C-seq data were 40bp or 50bp single-end Illumina sequencing reads with 17.5-30 million per sample. The distributions of FASTQ quality scores and of ambiguous base calls (“N”) were examined for each sequencing sample to make sure that the quality was in the acceptable range. 4%-20% of reads per sample included some portion of the 3’-end sequencing adapter. We hypothesized this might occur when the distance between HindIII and NlaIII restriction sites is short. Therefore, we trimmed end sequences matching the 3’-end adapter, which resulted in improved mappability of reads to the reference genome. The trimmed sequences were relatively short (<10bp for most reads).

Reads were aligned to the mouse reference genome (mm9 version) using Bowtie (Langmead et al., 2009), keeping only the reads with unique alignment (“-m 1”

parameter) and using default options for other parameters. Between 77% and 91% of reads for each sample were uniquely aligned to the reference genome, resulting in 15 to 28 million aligned reads per sample. In each sample, at least 99% of the aligned reads were confirmed to be within 2 bp of an annotated NlaIII restriction site, as expected from the sample preparation protocol. It was also verified that in nearly all cases the reads were found near the first annotated NlaIII restriction site neighboring an annotated HindIII restriction site, with the average and median read counts dropping to zero or almost zero in the second or third NlaIII site relative to the HindIII site location. For the 3 NlaIII sites neighboring each HindIII site, we observe that in most of the samples (90% of samples), the reads are mapped only to the first NlaIII site for > 84% of the fragments. We therefore adopted a conservative approach by considering only reads associated with the first NlaIII site, immediately adjacent to an annotated HindIII site.

Since the initial digestion in our 4C protocol was performed using HindIII, the HindIII fragments were used as the basic unit to quantify the interaction signal obtained from the 4C experiments. Specifically, we examined each genomic fragment defined by adjacent HindIII sites and calculated its signal as the number of reads mapping next to the leftmost and rightmost NlaIII sites located within the HindIII fragment, taking into account the orientation of reads at the NlaIII sites to assign them to the correct HindIII fragment.

Although the total read count after filtering was similar among the samples, the number of fragments with non-zero read count varied widely, with ESCs, iPSCs and partial iPSCs having a smaller number ($7x10^3$ - $30x10^3$) and MEFs and reprogramming intermediates showing a higher number ($56x10^3$-$186x10^3$). We believe this reflects different specificity of interactions as observed in distinct cell types. Indeed, this is consistent with our model of Nanog showing less functional interaction in more differentiated cells (MEFs) where the locus is inactive. These interactions would be less stable and reproducible. Under these circumstances we may expect the inactive Nanog locus to engage in not functional co-

localization with a larger variety of loci. In the 4C-ChIP samples, where the immunoprecipitation selects a subset of interactions mediated by one specific protein, we observed even a smaller number of fragments with non-zero read count ($2.2x10^3$ – $2.4x10^3$), despite having higher total read counts (18-23 million). This indicates that the difference in read distribution reflects true differences in the interaction patterns.

Data Normalization

Removing outliers. We examined the distribution of read counts in the fragments around the bait position and verified that the fragment at or adjacent to the bait location has very high read counts (typically a few million reads, at least an order of magnitude higher than other neighboring fragments). This is an expected consequence of self-ligation and presence of undigested fragments. We observe read enrichment around the bait region in our large domains analysis past the two outlier fragments immediately adjacent to the bait (Figure S2A), but the read count around the bait decreases more rapidly in m4C-seq than what has been reported by groups using alternative 4C-seq protocols (Splinter et al., 2011). Before normalization procedures and downstream analyses, we excluded reads from the two bait-neighboring fragments, as well as reads from the two non-adjacent fragments located on chromosome 8 that showed extremely large number of reads, e.g., more than one million, in one or more samples. Further examination of the bait probe sequence excluded similarities with these regions, so we removed these outlier fragments from all of the analyses as potential amplification artifacts.

Read count quantification. The read count per fragment was scaled over total library size and quantified as reads per million (RPM). We also evaluated the effect of applying additional normalization procedures such as the Trimmed Median of M-values (TMM) normalization (Robinson et al., 2010a), which had been proposed for RNA-seq data to account for differences in library complexity. TMM, however, was inappropriate for m4C-seq data as it appeared to assign too much weight to fragments in cell types with more variable set of interactions.

Genomic control. We also sequenced a genomic control sample, where all steps of the 4C-library construction protocol were followed except for the initial crosslinking step. The distribution of reads in the genomic control is expected to account for amplification and sequencing biases inherent to the experimental procedure. Thus, the log ratio over the genomic control of RPM normalized read count was used as one of the parameters for filtering interactions.

Bias estimation. We also performed additional filtering to account for GC content, fragment length and mappability biases, similar to the procedure proposed by Yaffe et al. (2011) for HiC data. For m4C-seq data, the normalization problem is simpler compared to HiC, as we are measuring “one to all” interactions between one fixed position and all of the other genomic regions, whereas HiC measures “all to all” interactions. Moreover, in our 4C protocol, reads originate from precise positions immediately adjacent to NlaIII restriction sites. We estimated mappability for each fragment end with a score equal to 1 or 0 if we observed, respectively, any or no reads from the fragment end across all of the dataset samples including genomic control and additional test samples (not included in the final dataset). To estimate GC content and fragment length biases, we divided fragment ends as in the original procedure: 20 bins defined using GC content over a 200bp window downstream the restriction site from where sequencing reads originate, and 20 independent bins defined using the length of the NlaIII-HindIII fragments. We also tried binning fragment ends according to the HindIII-HindIII fragment length, as in the original procedure, but noted that in our case the reads were almost uniformly distributed across bins obtained with this parameter, whereas they show more marked bias when considering fragment ends length. The procedure proposed by Yaffe et al. (2011) estimates the bias correction factors using unique pairs, *i.e.,* counting only once each fragment pair without taking into account the actual number of observed reads from each pair. This is suitable for HiC as the number of possible fragment end pairs is much larger compared to the number of sequencing reads and the actual read count per fragments pair would be small. In our case, the number of possible HindIII fragments is about $8x10^5$, and the number of the fragment ends is twice that

number. This is orders of magnitude smaller than the number of sequenced reads. Because of this, we took into account the actual number of observed reads per fragment when scoring the strength of the interactions. We, therefore, followed the bias estimation procedure described by Yaffe et al. (2011), but re-defined the bias factor estimates for fragment length bin *i* as $S_{len}[i] = (1/P_{prior}[i]) \cdot \frac{O_{len}[i]}{O}$, where O is the total reads in the sample, and $O_{len}[i]$ is the total number of observed reads in the fragments belonging to the fragment length bin *i*. The $P_{prior}[i]$ is the bin specific prior (different from the global prior used in original method definition) with $P_{prior} = \frac{T[i]}{T}$, where T is the total number of fragments, and $T[i]$ is the total number of fragments belonging to bin *i*. Similar modifications are adopted for S_{gc} estimates of GC content bias. Then we used code from the original procedure to build a similar log likelihood function from bias estimation scores and to implement the optimization procedure. The bias estimates for each HindIII fragment end were used to estimate the expected distribution of read counts at the individual fragments level. We used only the genomic control sample for the bias estimation, as this sample is supposed to provide the best representation of the biases stemming from sequencing and other sample preparation steps. As cross-linking does not occur in the genomic control sample, we did not separate the *cis* and trans fragments during the bias estimation procedure.

Selection of interacting positions

Selecting large domains with domainograms. For multi-scale analysis and visualization of the m4C-seq data we used a custom adaptation of the "domainogram" approach (Bantignies et al., 2011). This method is based on scoring the interaction strengths over multiple sizes of sliding windows $W_{i,j}$ where the window size *i* is the number of consecutive restriction fragments (with $i \geq 2$) and *j* is the position of the fragment where the window is centered, with $j \geq i/2$ and $j \leq (N - i/2)$. N is the total number of restriction fragments in the chromosome. For each window $W_{i,j}$ the interaction score $S_{i,j}$ is computed as a

sum of log transformed and normalized read counts C_k observed in each fragment k included in the window: $S_{i,j} = \sum_{k=j-(i/2)}^{k=j+(i/2)} C_k$, with $C_k = \left(Log_2(F_k+1) - Log_2(G_k+1)\right)$, where F_k is the read count for the fragment k in the m4C-seq sample, and G_k is the read count for fragment k in the genomic control sample. If $C_k \leq 0$, it is replaced with zero. We also examined domainograms using read count normalized over expected distribution of reads computed with the bias estimation described above, which yielded similar results (not shown).

The significance of the $S_{i,j}$ score is assessed by comparing its value to the values obtained from random permutations of C_k values, grouped and summed to compute expected distribution of S scores in windows of size i. The observed S score is compared to the upper tail of the sampled distribution to determine an empirical p-value. The obtained empirical p-values are log transformed and mapped to a color gradient to visualize the domainograms. To identify the fragments involved in large domain interactions, all of the windows of sizes ranging from 2 to 256 fragments centered at each position j are considered. Fragment j is considered the center of a large domain interaction if for any of these windows the p-value is ≤ 0.0001. The maximum window size of 256 fragments corresponds to 0.79Mb on average and was chosen to avoid selecting very large windows with unclear biological significance.

Selecting individual fragments. We determined individual interacting fragments using a combination of three filtering criteria on RPM normalized read counts per fragment: i) at least 2-fold enrichment over the genomic control sample; ii) at least 10-fold enrichment over the expected read count as estimated from the bias estimation procedure above; and iii) read count ≥5RPM at each fragment. Then for each cell type, we considered interaction as “conserved” if the fragment was found in all of the biological replicates or alternative cell lines.

Selecting interacting genes. We summarized m4C-seq signal per gene locus by summing read counts at individual fragments overlapping the gene locus or a

20kb window upstream the annotated transcription start site. RefSeq annotations for protein coding genes were retrieved from UCSC Genome Browser database (mm9 genome) and used as reference annotations. Normalization of gene level read count was performed using the same procedures described above for fragment-level quantification of m4C-seq signal. We selected interacting genes using a combination of three filtering criteria on RPM normalized read counts per gene: i) at least 2-fold enrichment over the genomic control sample; ii) at least 2-fold enrichment over the expected read count as estimated from the bias estimation procedure above; and iii) read count ≥5RPM. Then for each cell type, we considered interaction as "conserved" if the fragment was found in all of the biological replicates or alternative cell lines.

Selection of differential interactions

We have employed a Bayesian approach in evaluating the statistical significance of the differential interactions. Briefly, each replicate was modeled as a mixture of a negative binomial distribution representing detected fragments, and a Poisson distribution describing signal for the fragments that were not detected. The mixing of the two distributions was controlled by a binomial process with the probability of failure modeled as a logistic regression on the signal magnitude (on log scale). The parameters of the model were fit using an iterative EM procedure. The initial signal estimates for the model fitting were determined by cross-fitting all possible pairs of biological replicates for a given cell type. The signal intensities observed for a given pair were modeled as a mixture of two Poisson "failure" components and one "correlated" negative binomial component. The median of the signals observed in the correlated components was then used as initial fragment signal estimate during model fitting.

The models derived for each individual experiment were then used in determining the statistical significance of the difference in the 4C signal observed between different cell types. Given the signal counts observed for a given fragment in each experiment, the Bayes factor was calculated as a ratio of the joint likelihood that the true 4C signal intensity differed between cell types to the likelihood that the underlying 4C signals were identical. The corresponding

fragment-level z-score was used in further calculations (see below). To avoid selecting differences involving signals below our thresholds for interactions, for fragments where the three threshold criteria for the single fragment interaction were not satisfied in at least 2 out of 3 replicates of the compared cell types, the computed z-score was replaced with zero.

This method is an extension of the methods using a negative binomial distribution for modeling differences in read counts between samples with replicates (Anders et al. 2011; Robinson et al., 2010b). These methods are not appropriate for m4C-seq data in their current implementation as they do not allow a proper control of outliers when a large fraction of features have zero read count in one replicate and non-zero count in another. Such detection failures are frequent in the m4C-seq data, and are more prevalent at lower signal magnitude. The method we employ here takes such stochasticity into account.

Selecting differential fragments. The set of fragment-level differential interactions was selected as those with the z-score magnitude ≥ 1.96 (or ≤ -1.96).

Selecting differential large domains with domainograms. The z-scores computed and filtered at the single fragment level were also used as an input for the differential domainogram analysis which aims to select larger domains of either increased or decreased interaction frequency. In this case for each window $W_{i,j}$ the score for differential interactions D is computed as $D_{i,j} = \sum_{k=j-(i/2)}^{k=j+(i/2)} Z_k$, where Z_k is the z-score associated to each fragment *k*. The observed D score is compared to D score computed after random permutation of Z values and either upper tail (for up-regulation) or lower tail (for down regulation) of the random distribution is used to assess two single tail p-values for either up or down regulation for each window $W_{i,j}$. For each position *j* all the windows of size between 2 and 256 fragments, centered at position *j* are considered. Fragment j is considered the center of a large differential interaction domain, if for any of these windows the p-value is ≤ 0.001, for either up- or down-regulation analysis, respectively.

Correlation to genomic features

m4C-seq data were correlated with various publicly available datasets on chromatin marks, transcription factor and other protein binding sites, DNase hypersensitivity sites and replication timing (see Supplementary Table 3).

The co-localization of m4C interaction sites and specific chromatin features is measured as log ratio enrichment of observed over expected, where the "observed" is the number of the selected HindIII fragments overlapping a given chromatin feature and the "expected" is derived from the fraction of the HindIII fragments overlapping the considered feature over the whole genome. Where the gene level association is reported instead, the selection of m4C-seq interactions was based on gene-level analysis, as above described. Fragments for which no associated reads were observed across all datasets (including genomic control and test datasets) were considered "non-detectable" and were excluded from the calculations.

For the ChIP-seq data, the binding or enrichment regions reported in the original paper were used whenever genome-wide peak calls were available. Individual genes were linked to a specific mark or transcription factor if the ChIP enrichment peaks overlap with a -5Kb/+1Kb window around RefSeq annotated transcription start sites for protein coding genes. When necessary, conversion of enrichment peaks to the mm9 coordinates was performed using the liftover tool by UCSC Genome Browser. When the original genome-wide peak calls were not available, we used SPP package (Kharchenko et al., 2008) to process the data and call either broad regions of enrichment (for chromatin marks) or precise binding positions for transcription factors, such as the pluripotency associated factors, using default parameters. When calling precise binding positions with SPP, a broader 1kb region around the binding positions were considered as associated to the target protein.

To assess the significance of the "observed over expected" ratios, empirical p-values based on random sampling ($1x10^5$ samplings) of gene lists were calculated. The enrichment p-value was assessed separately for each

considered list (e.g. for ESC conserved interactions or for MEF conserved interactions) and presented in the plots of Figures 3 and S3.

For the replication time data (Hiratani et al., 2010) the original segmented regions and associated replication time scores were used. For each segment, the median replication time score across replicates for each cell type was used. Fragments overlapping each genomic segment were associated to the corresponding replication time score, and then divided into 5 quantiles from early to late replication times.

Analysis of microarray gene expression datasets

Gene expression datasets generated on Affymetrix microarrays (Polo et al., 2012; Stadtfeld et al., 2010a) were used to derive a list of pluripotency associated genes. The expression data were preprocessed using up-to-date custom probeset definitions based on the Entrez gene database (Dai et al., 2005) and normalized by the RMA procedure. A linear model implemented in the Limma Bioconductor package was used to select differentially expressed genes in pairwise comparisons with false discover rate (FDR) ≤ 0.001 (Smyth, 2004). Pluripotency associated genes were identified as genes up-regulated both in ESC vs MEF and in iPSC vs MEF pairwise comparisons (pluripotency genes). Conversely, differentiation-associated genes were defined as genes down-regulated in both pairwise comparisons (differentiation genes).

To identify expression pattern of gene level 4C-seq interactions activated in partial-iPSC data from different Affymetrix microarrays datasets were used (Polo et al., 2012; Sridharan et al., 2009; Stadtfeld et al., 2010a). The expression data were preprocessed using up-to-date custom probeset definitions based on the Entrez gene database (Dai et al., 2005) for each updated probeset only probes with the same sequence in the two microarray versions are kept (Fallarino et al., 2010) and used to compute expression signal with RMA normalization (Irizarry et al., 2003). To remove remaining dataset specific biases, we applied also ComBat procedure for batch effect correction (Johnson et al., 2007)

RNA-seq data processing

Illumina single end 101 bp long reads for RNA-seq data were processed using tophat and cufflinks software for reads alignment and gene expression quantification (Trapnell et al., 2012). Each sample yielded between 29 and 40 million raw reads. Illumina igenome annotation freeze for mm9 UCSC genome version was used as reference (http://cufflinks.cbcb.umd.edu/igenomes.html). Tophat 2.0.4 was used with options "-G --bowtie1" (default for others) and cufflinks 2.0.2 with options "--frag-bias-correct --multi-read-correct --compatible-hits-norm". Gene level FPKM values were used as estimated expression values.

Supplementary References

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol *11*, R106.

Bantignies, F., Roure, V., Comet, I., Leblanc, B., Schuettengruber, B., Bonnet, J., Tixier, V., Mas, A., and Cavalli, G. (2011). Polycomb-dependent regulatory contacts between distant Hox loci in Drosophila. Cell *144*, 214-226.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376-380.

Dai, M., Wang, P., Boyd, A.D., Kostov, G., Athey, B., Jones, E.G., Bunney, W.E., Myers, R.M., Speed, T.P., Akil, H.*, et al.* (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res *33*, e175.

Fallarino, F., Volpi, C., Fazio, F., Notartomaso, S., Vacca, C., Busceti, C., Bicciato, S., Battaglia, G., Bruno, V., Puccetti, P., *et al.* (2010). Metabotropic glutamate receptor-4 modulates adaptive immunity and restrains neuroinflammation. Nat Med *8*, 897-902.

Hiratani, I., Ryba, T., Itoh, M., Rathjen, J., Kulik, M., Papp, B., Fussner, E., Bazett-Jones, D.P., Plath, K., Dalton, S.*, et al.* (2010). Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. Genome Res *20*, 155-169.

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B., Speed, T.P. (2003). Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res *31*, e15

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118-127.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S.*, et al.* (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430-435.

Kharchenko, P.V., Tolstorukov, M.Y., and Park, P.J. (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol *26*, 1351-1359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Maherali, N., Sridharan, R., Xie, W., Utikal, J., Eminli, S., Arnold, K., Stadtfeld, M., Yachechko, R., Tchieu, J., Jaenisch, R.*, et al.* (2007). Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. Cell stem cell *1*, 55-70.

Polo, J.M., Anderssen, E., Walsh, R.M., Schwarz, B.A., Nefzger, C.M., Lim, S.M., Borkent, M., Apostolou, E., Alaei, S., Cloutier, J.*, et al.* (2012). A Molecular Roadmap of Reprogramming Somatic Cells into iPS Cells. Cell *151*, 1617-1632.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.

Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol *11*, R25.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. Cell *148*, 458-472.

Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol *3*, Article 3

Sridharan, R., Tchieu, J., Mason, M.J., Yachechko, R., Kuoy, E., Horvath, S., Zhou, Q., and Plath, K. (2009). Role of the murine reprogramming factors in the induction of pluripotency. Cell *136*, 364-377.

Splinter, E., de Wit, E., Nora, E.g., Klous, P., van de Werken, H., Zhu, Y., Kaaij, L., van Ijcken, W., Gribnau, J., Heard, E.*, et al.* (2011). The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. Genes & development *25*, 1371-1454.

Stadtfeld, M., Apostolou, E., Akutsu, H., Fukuda, A., Follett, P., Natesan, S., Kono, T., Shioda, T., and Hochedlinger, K. (2010). Aberrant silencing of imprinted genes on chromosome 12qF1 in mouse induced pluripotent stem cells. Nature *465*, 175-181.

Stadtfeld, M., Maherali, N., Borkent, M., and Hochedlinger, K. (2010b). A reprogrammable mouse strain from gene-targeted embryonic stem cells. Nat Methods *7*, 53-55.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc *7*, 562-578.

Tolhuis, B., Blom, M., Kerkhoven, R.M., Pagie, L., Teunissen, H., Nieuwland, M., Simonis, M., de Laat, W., van Lohuizen, M., and van Steensel, B. (2011). Interactions among Polycomb domains are guided by chromosome architecture. PLoS Genet *7*, e1001343.

Xu, N., Tsai, C.-L., and Lee, J. (2006). Transient homologous chromosome pairing marks the onset of X inactivation. Science (New York, NY) *311*, 1149-1201.

Yaffe, E., and Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet *43*, 1059-1065.

Yaffe, E., and Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet *43*, 1059-1065.